



**QCon** 全球软件开发大会  
INTERNATIONAL SOFTWARE  
DEVELOPMENT CONFERENCE

BEIJING 2017

# 美团点评旅游推荐系统的演进

郑刚



促进软件开发领域知识与创新的传播



关注InfoQ官方信息  
及时获取QCon软件开发者  
大会演讲视频信息



扫码，获取限时优惠

**ArchSummit**  
全球架构师峰会 2017 [深圳站]

2017年7月7-8日 深圳·华侨城洲际酒店

咨询热线：010-89880682

**QCon**

全球软件开发大会 [上海站]

2017年10月19-21日

咨询热线：010-64738142

- 2015年至今 美团点评酒旅事业群

- 负责酒旅搜索排序推荐
- 负责酒旅数据仓库和数据产品建设

- 2014年之前 美团网技术部数据组

- 参与数据平台搭建
- 负责全平台数据仓库和数据产品建设

- 2011年 百度电子商务事业部

- 有啊商城的开发

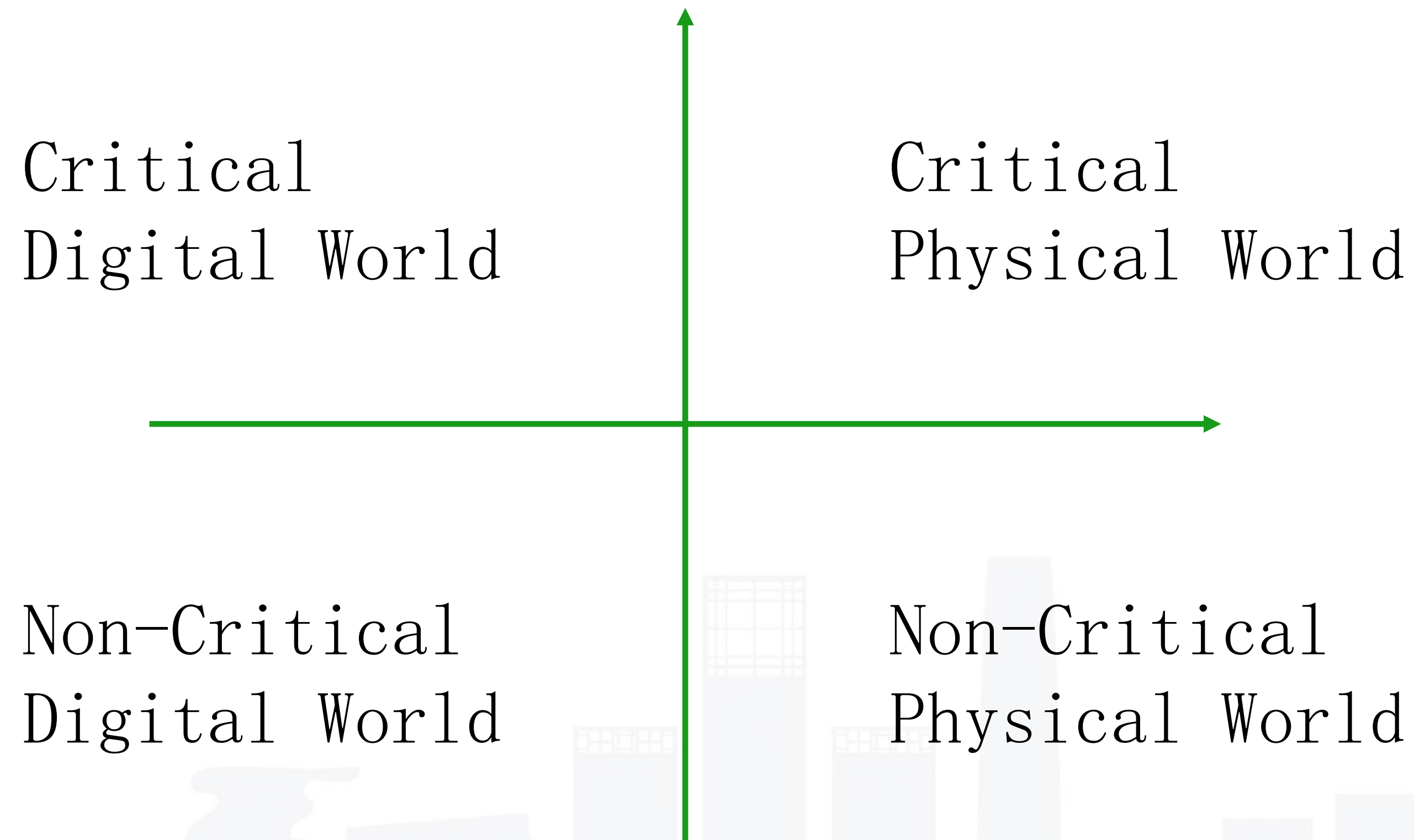
- 2010年毕业于中科院计算所



# Outline

- 美团点评酒旅业务简介
- 基于用户画像的召回策略演进
- 基于L2R的排序策略优化
- 从海量大数据的离线计算到高并发在线服务的推荐引擎架构设计
- 推荐在美团点评酒旅的应用实践

# 人工智能应用



# 新美大酒旅

国内发展最快的一站式综合住宿服务平台  
国内最大的在线门票交易平台

## • 2016年美团点评酒店覆盖量与间夜量



2016年全年，美团点评双平台酒店间夜量超**1.3亿**，高星酒店间夜单月增速最高达**300%**，酒店商家**32万**。

## • 2016年美团点评票务销售量



2016年全年，美团点评门票销售**6700万张**。

2016年全年，**机票销售200万张**。



2016年全年，**火车票销售800万张**。

# 新年新高度!

2017年1月31日

## 美团点评景点门票

单日交易额 新纪录

# 1.14

亿

单日入园人次超过

# 106

万

比去年春节同期增长超过**100%**

  美团点评

国内最大的在线门票交易平台  
为全国超过2万个景区景点提供业界领先的专业票务解决方案

# 酒旅搜索推荐

- 2015Q3 组建推荐团队
- 2015Q4 周边游频道内推荐
- 2016Q1 搜索少/无结果推荐
- 2016Q2 详情页推荐
- 2016Q3 酒旅交叉推荐
- 2016Q4 点评旅游推荐

酒店住宿

境内度假

境外度假

大交通

搜索/推荐

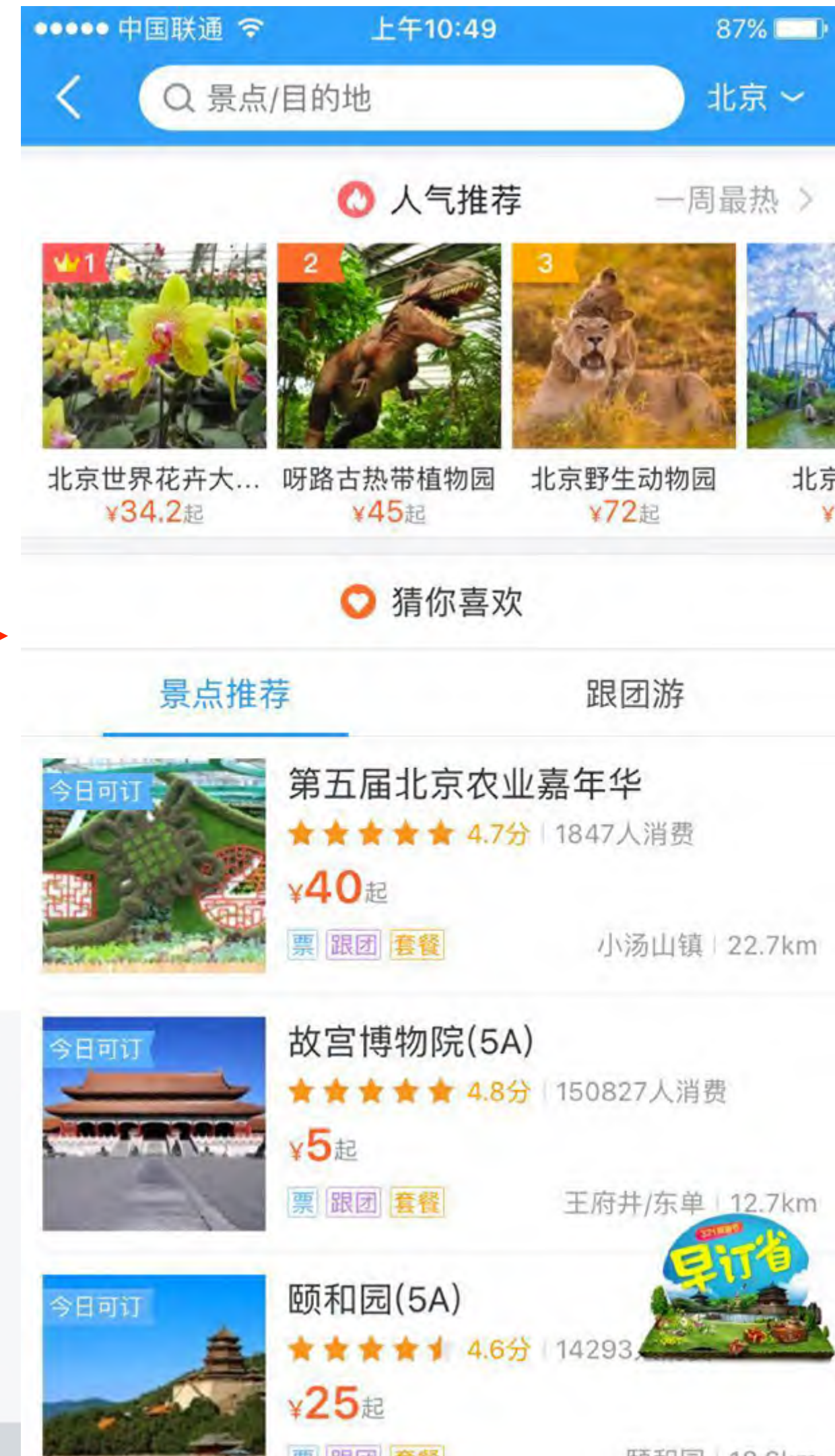
数据挖掘

数据产品

酒旅数据仓库

集团数据平台

# 旅游推荐产品形态



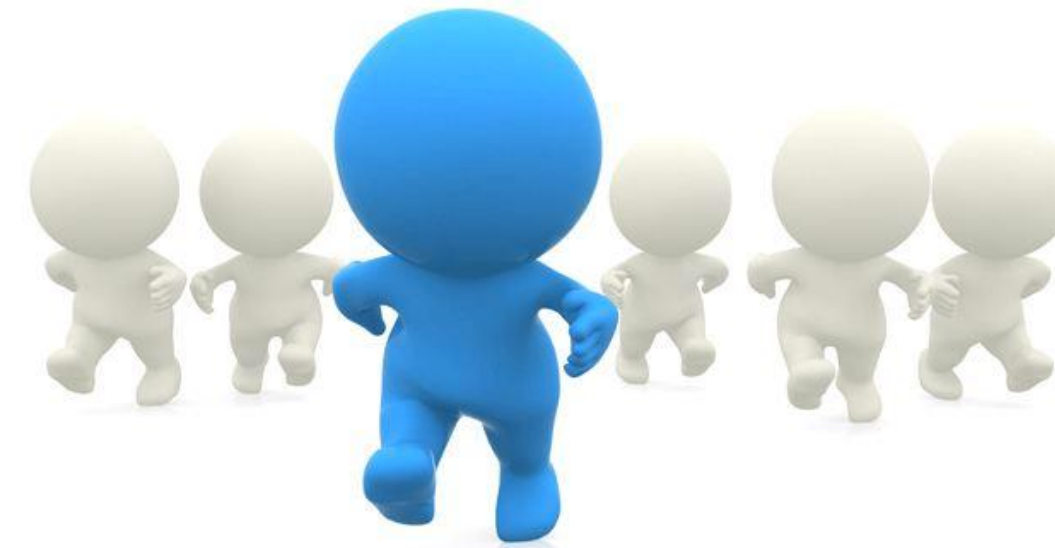
旅游场景下用户兴趣点不明确，频道内超过50%订单来自推荐



# 旅游推荐面临的问题



本异地差异大



推荐形式多样



季节性明显



需求个性化

# 本异地差异大

中国联通 上午11:55 74%

北京 第五届北京农业嘉年华

— 猜你喜欢 —

- 故宫博物院(5A)** 12.6km  
4.8分 | 25546人评价  
门票、套餐、线路游 等优惠  
¥5 起 已售150856
- 颐和园(5A)** 18.5km  
4.6分 | 6200人评价  
门票、套餐、线路游 等优惠  
¥25 起 已售14312
- 八达岭长城(5A)** 56.4km  
4.8分 | 4393人评价  
门票、套餐、线路游 等优惠  
¥35 起 已售79610
- 圆明园(4A)** 15.4km  
4.7分 | 6572人评价  
门票、套餐、线路游 等优惠  
¥21.9 起 立减1元 已售28552
- 呀路古热带植物园(3A)** 40.2km

首页 附近 逛一逛 订单 我的

中国联通 上午11:58 72%

北京 第五届北京农业嘉年华

— 猜你喜欢 —

- 北京欢乐谷(4A)** 15.6km  
4.7分 | 93655人评价  
门票、套餐、线路游 等优惠  
¥58 起 已售104885
- 呀路古热带植物园(3A)** 40.2km  
4.6分 | 6202人评价  
门票、套餐、线路游 等优惠  
¥45 起 已售54912
- 活的3D博物馆** 2.7km  
4.4分 | 11383人评价  
门票 等优惠  
¥38 起 已售25698
- 太平洋海底世界(3A)** 18.4km  
4.4分 | 25212人评价  
门票、线路游 等优惠  
¥66.7 起 立减10元 已售23442

首页 附近 逛一逛 订单 我的

超过30%订单来自异地请求

常驻城市!=浏览城市

# 推荐形式多样

中国联通 下午2:52 64%

Q 景点/目的地 北京

景点推荐 跟团游

今日可订 故宫博物院(5A) ★★★★★ 4.8分 | 153336人消费  
¥5起 返券5元 票 跟团 王府井/东单 | 12.7km

今日可订 北京欢乐谷(4A) ★★★★★ 4.7分 | 99754人消费  
¥58起 返券5元 票 跟团 套餐 北京欢乐谷 | 15.7km

今日可订 第五届北京农业嘉年华 暂无评分 | 3801人消费  
¥40起 限时减5元 票 跟团 套餐 小汤山镇 | 22.7km

今日可订 呀路古热带植物园(3A) ★★★★★ 4.6分 | 52422人消费  
¥45起 限时减5元 票 跟团 套餐 大兴区 | 22.7km

今日可订 圆明园(4A) ★★★★★ 4.6分 | 48245人消费  
¥20.8起

中国联通 下午2:52 64%

Q 景点/目的地 北京

景点推荐 跟团游

北京出发 八达岭长城纯玩1日跟团游\*只含交通费不排队挤车  
立减5元 ¥48 门市价:¥98 已售991

北京出发 八达岭长城、鸟巢、水立方纯玩1日跟团游\*不起早无购物天天发团  
立减10元 ¥84 门市价:¥169 已售3136

北京出发 古北水镇、张裕爱斐堡国际酒庄纯玩1日跟团游\*夜游古北水镇,品红酒  
¥238 门市价:¥298 已售1390

北京出发 八达岭长城、什刹海、鸟巢等纯玩1日跟团游\*真纯玩 无购物  
¥78 门市价:¥138

北京出发 八达岭长城、水立方、鸟巢纯玩1日跟团游\*不早起,深入八达岭  
立减10元

景点下有大量相似门票,不适合按Deal样式展现

跟团游、景酒套餐关联多个景点,不适合按POI样式展现

# 季节性明显



冬季温泉订单占比超过20%，  
而夏季不到7%

# 需求个性化



用户人群  
时间地域场景  
内容形态

北京亲子游玩群正在热烈讨论 点击加入

“亲子游”是宝宝成长中最温馨永久的记忆，也是父母与孩子间的良好交流渠道。小编特别整理“潮爸潮妈都说好”的亲子游好去处，赶快携子出游吧！

一个人吃狗粮，两个人玩浪漫。趁着阳光正好，微风不噪，带上你心爱的TA，一起去赏古迹、泡温泉、逛展览，解锁“城会玩”新技能！到游乐场里，撒泼打滚“嗨翻天”！来山水田园，花式“拍美照”！

全部 酒店+景点 嗨翻天 拍美照

全部 1岁及以上 4岁及以上 6岁及以上



# 基于用户画像的召回策略演进



# 热销策略

基于Deal所在城市统计分城市热销



分类	场景	召回策略
本地需求	常驻城市=浏览城市 (北京人浏览北京)	当地用户购买的热销POI
异地需求	常驻城市!=浏览城市 (重庆人浏览北京)	异地用户购买的热销POI (所有非北京人购买的热销POI)

$$deal\_score = \sum_{i=1}^{28} p_i * \alpha^i$$

销量按时间衰减

# 热销策略

## •精确统计POI销量

- Deal -> POI
- POI售卖数据不准
- 客户端埋点



POI详情页  
F\_poiid



DEAL详情页  
F\_poiid



下单页  
F\_poiid





# 用户画像

$$h_w(x) = P(y = 1|x; w) = \frac{1}{1 + e^{-wx}}$$



## •模型

- LR：预测常驻城市与某维度城市相等的概率

## •样本

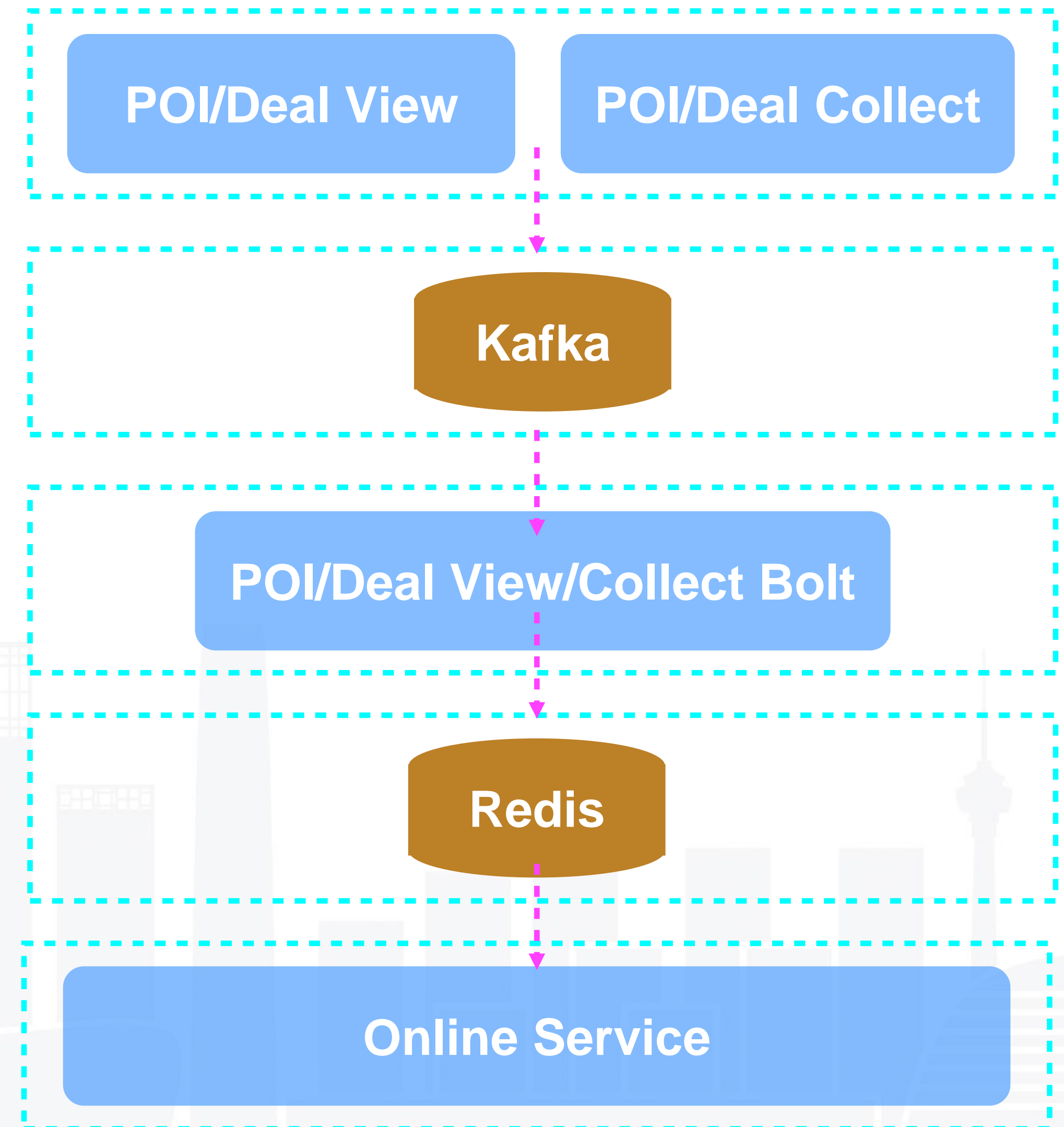
- 调查问卷

## •特征

- 注册城市
- 注册手机号
- 手机定位城市
- 浏览城市
- 消费城市：团购、电影、外卖
- 接受短信手机号

# 用户历史行为强相关策略

- 热销策略能区分本异地用户差异
  - 不能对具体用户个性化推荐
- 用户一个月内浏览、收藏的 POI/Deal
- 越实时权重越高



# Location-Based策略

## •冷启动

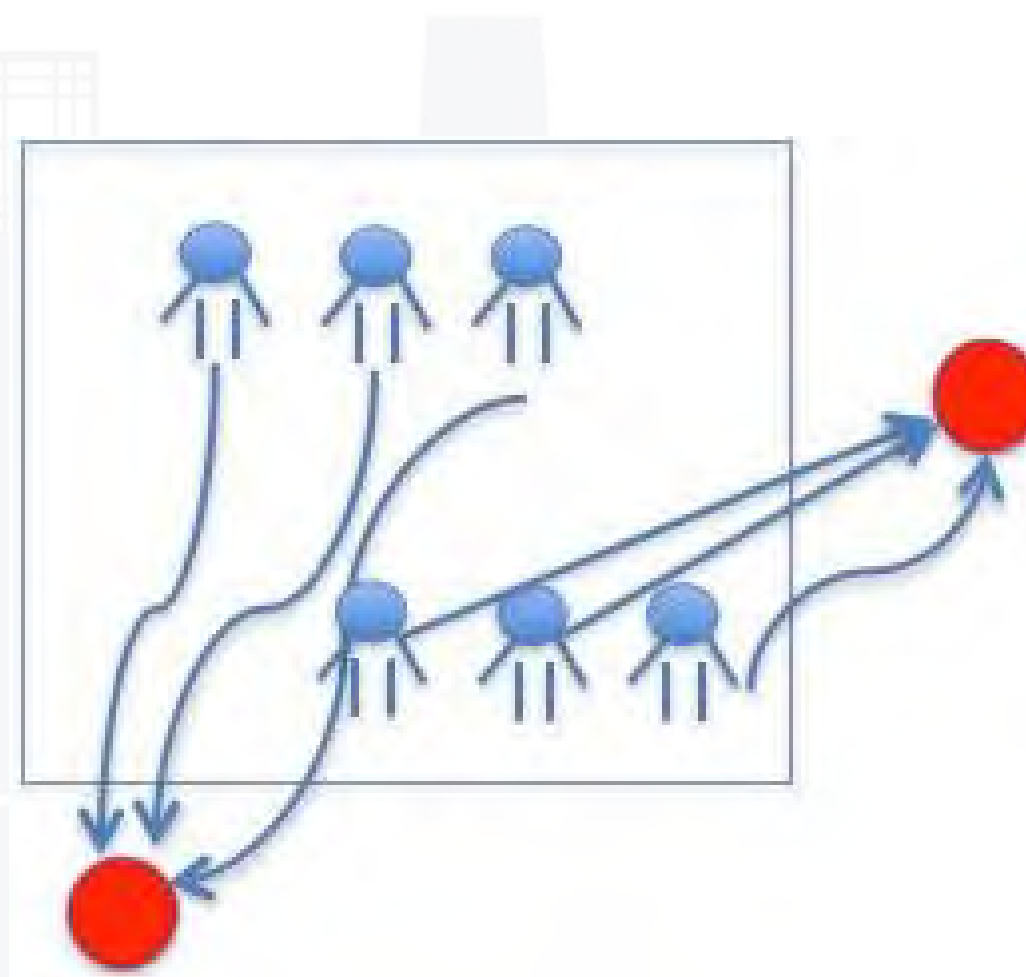
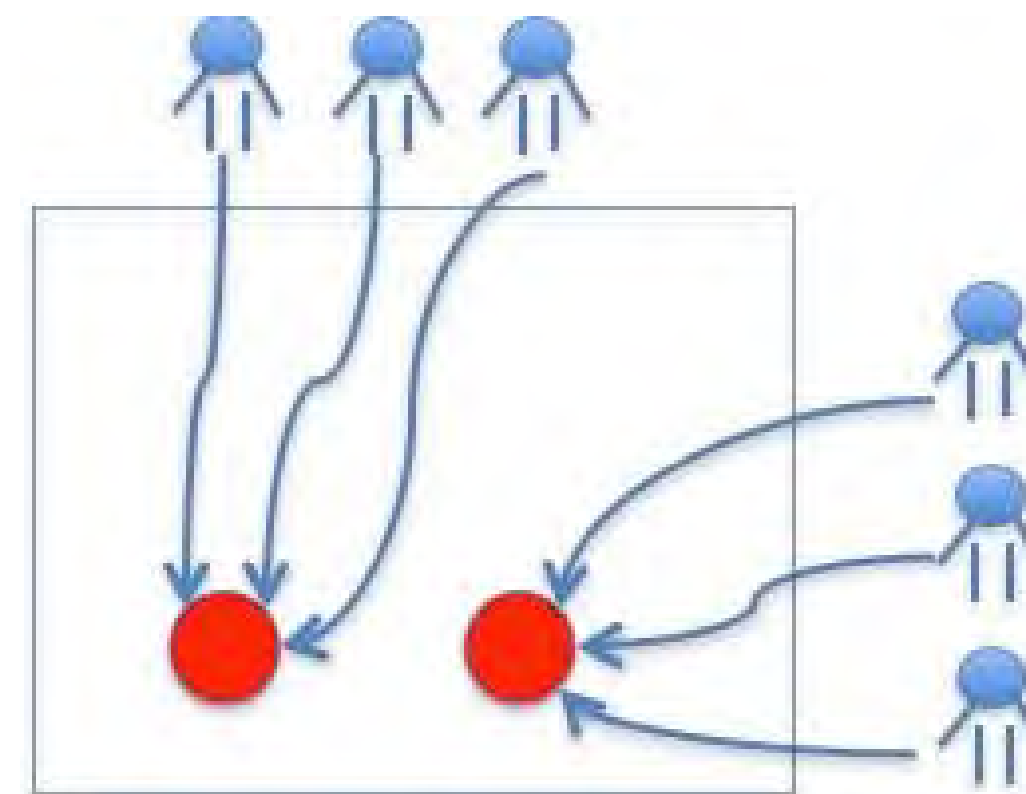
- 新POI
- 新用户

## •区域消费热门POI

- 5KM范围内的热销POI

## •区域购买热门POI

- 5KM范围内的用户购买的POI
- 回龙观附近没有POI



# 协同过滤

## •Item CF

### •基于POI浏览行为

•POI相似度每天离线更新，User浏览POI行为实时更新

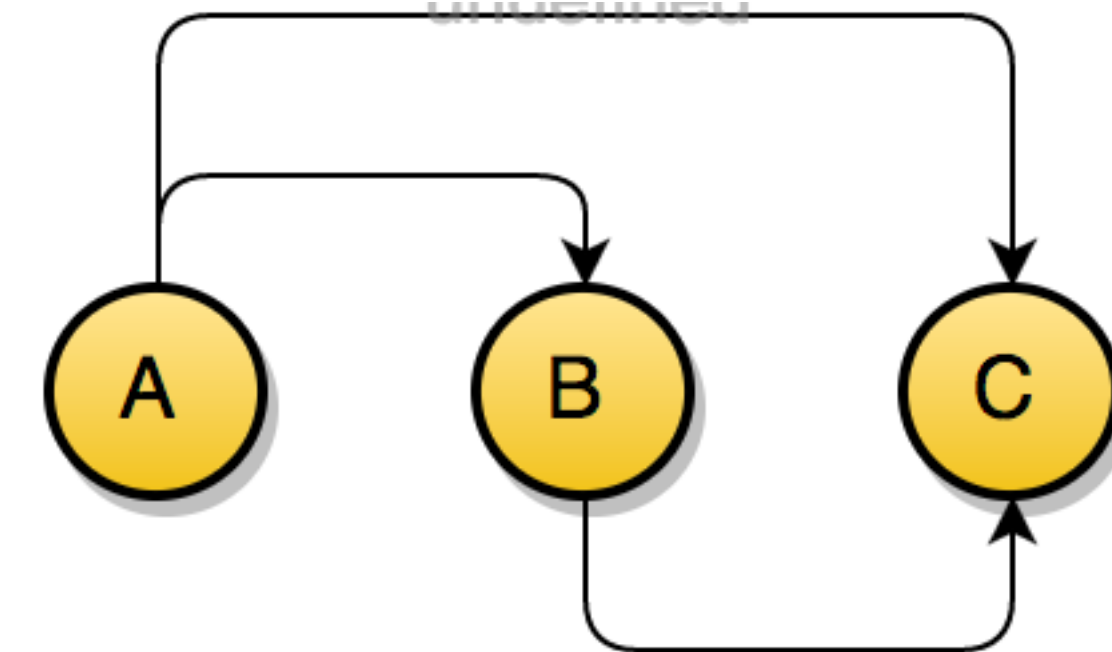
### •相似度改进：时间序列衰减

### •基于用户搜索行为

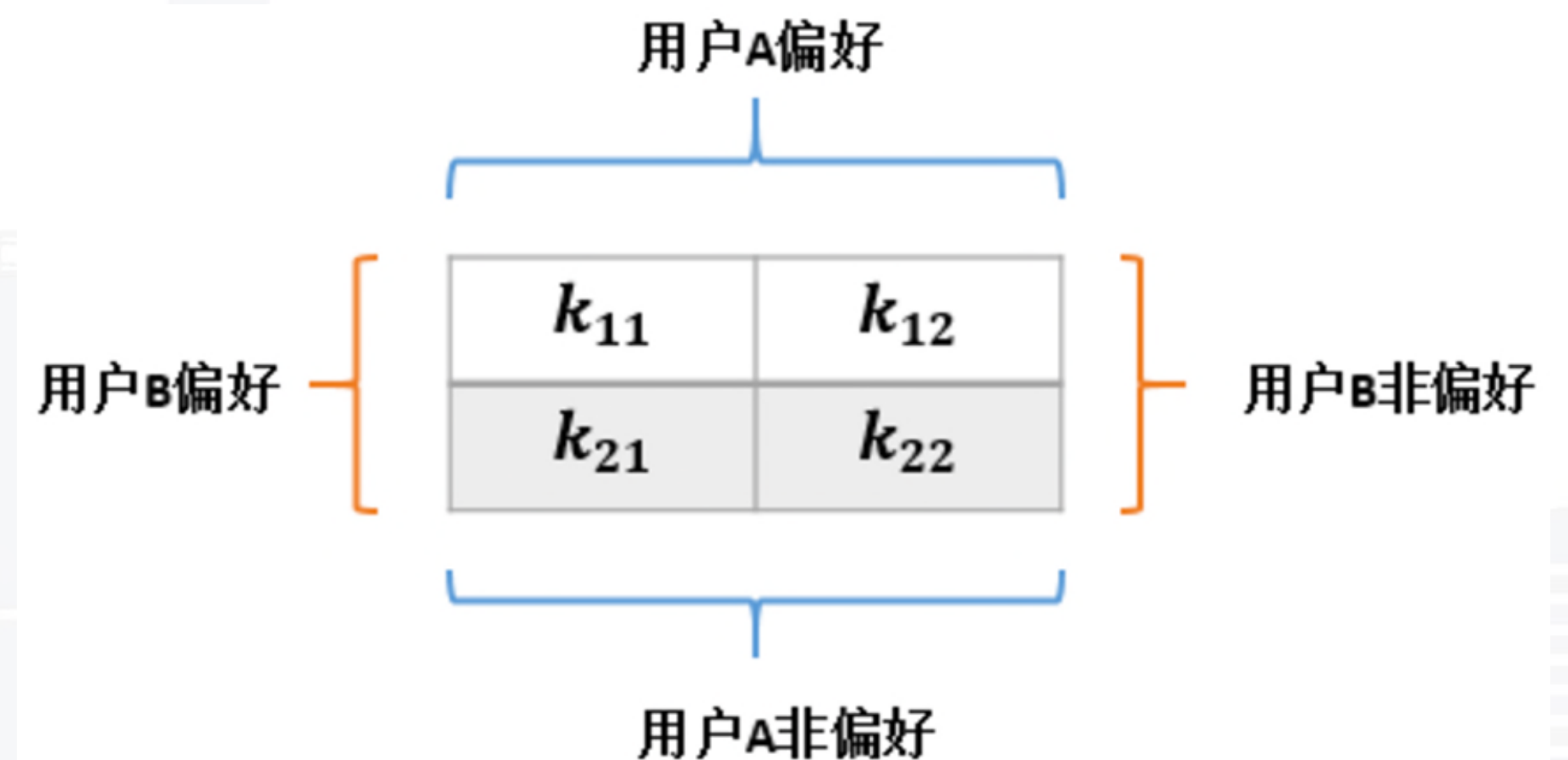
•基于搜索后浏览POI行为构造<Query,POI>矩阵，计算POI相似度

## •User CF

### •loglikelihood ratio



$$w_{ij} = \frac{\sum_t N_{ijt} * \alpha^t}{\sqrt{|N(i)||N(j)|}} (i < j)$$



# 基于用户画像的推荐

用户标签偏好\*标签权重\*POI标签偏好\*POI销量

## 基础属性

- 性别、年龄、职业

## 人群属性

- 有车：汽车保养
- 宅男：外卖
- 情侣：电影
- 亲子：儿童乐园
- 旅游达人：酒店旅游交通

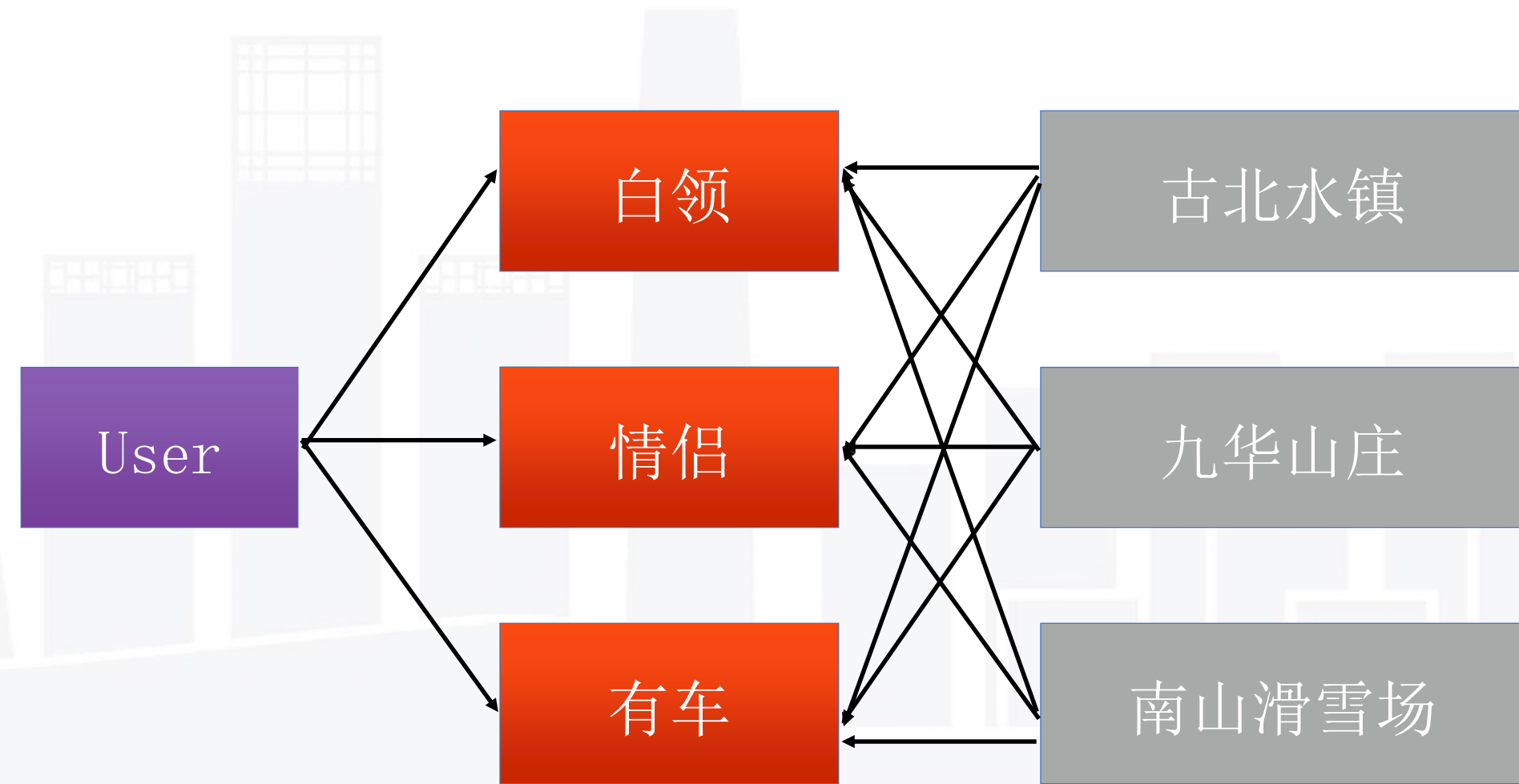
## 推荐

- 基于用户标签计算POI标签
- 精确匹配：给亲子类用户推荐亲子类POI
- 模糊匹配：基于标签计算用户和POI相似度

$$D(P \parallel Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

标签在POI维度的分布

标签在用户维度的分布



# 召回策略演进过程

本异地热销策略

历史行为强相关

- 浏览未购买
- 收藏未购买

Location-Based

- 区域热门消费
- 区域热门购买

协同过滤

- Item CF
- POI CF
- Query CF
- User CF
- LLR

基于用户画像的推荐

- 精确匹配
- 模糊匹配

# 基于L2R的排序策略优化



# 机器学习流程



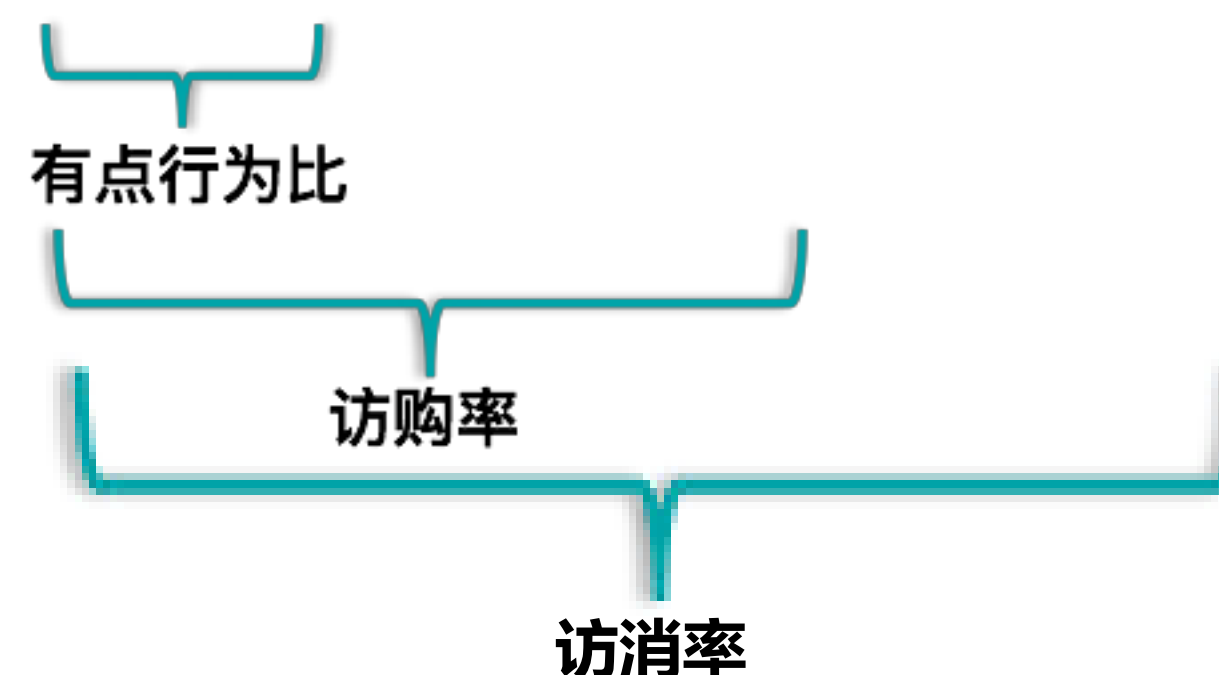


# 问题建模

## • 访购率为目标

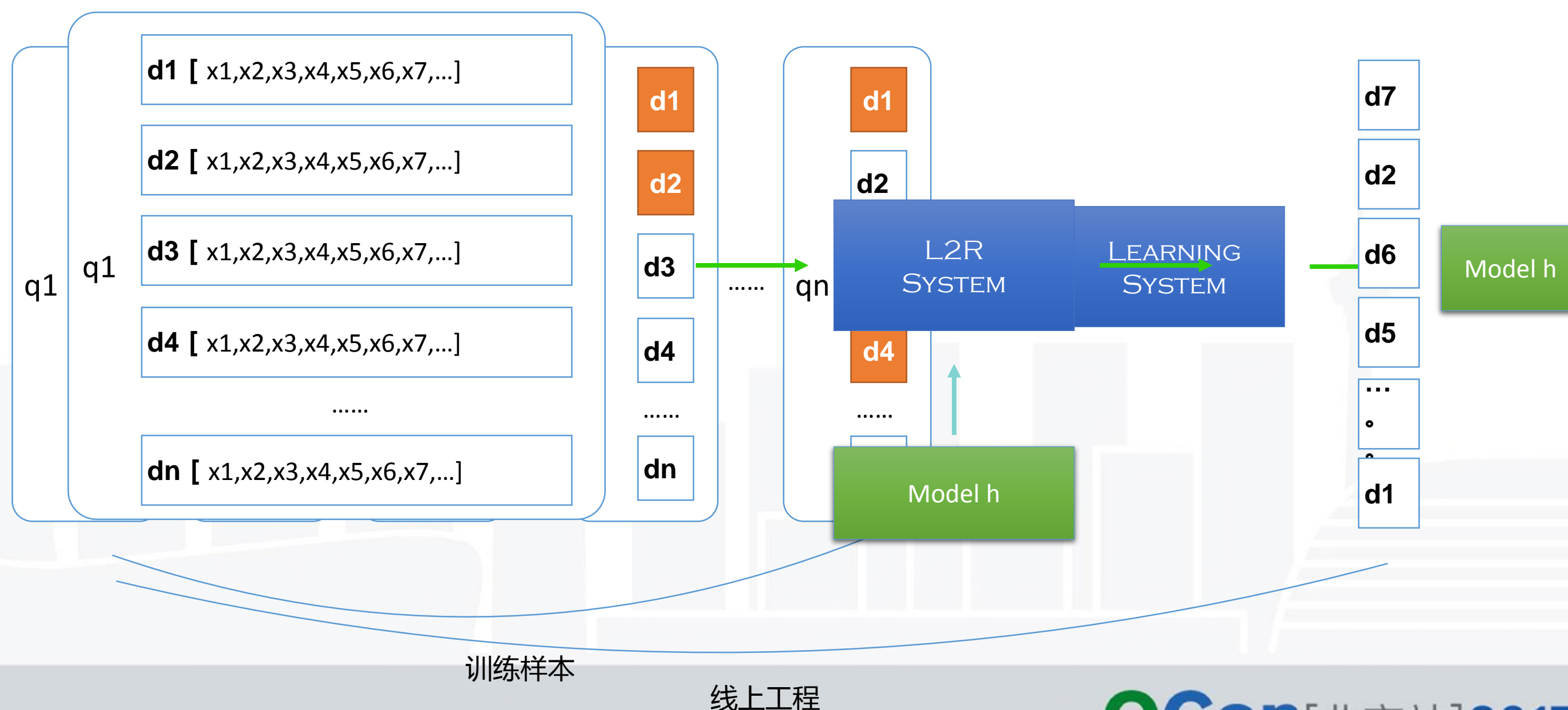
- 只看点击率没有反映出交易属性
- 看最终收入
  - 消费受购买限制、退款条件等影响
  - 收入跟BD谈单毛利相关

$$\text{每推荐用户收入} = \underbrace{\frac{\text{点击用户数}}{\text{推荐用户数}}}_{\text{有点行为比}} \times \underbrace{\frac{\text{支付用户数}}{\text{点击用户数}}}_{\text{访购率}} \times \underbrace{\frac{\text{消费用户数}}{\text{支付用户数}}}_{\text{访消率}} \times \text{每用户消费收入}$$



## • Pointwise L2R

- Pairwise性能问题
- NN做rank?



# 问题建模

## •GBDT

- 非线性
- High Level特征多

## •XGBoost

- 泰勒展开，利用了二阶导数信息
- 对数据预排序，性能更高

## •多模型融合

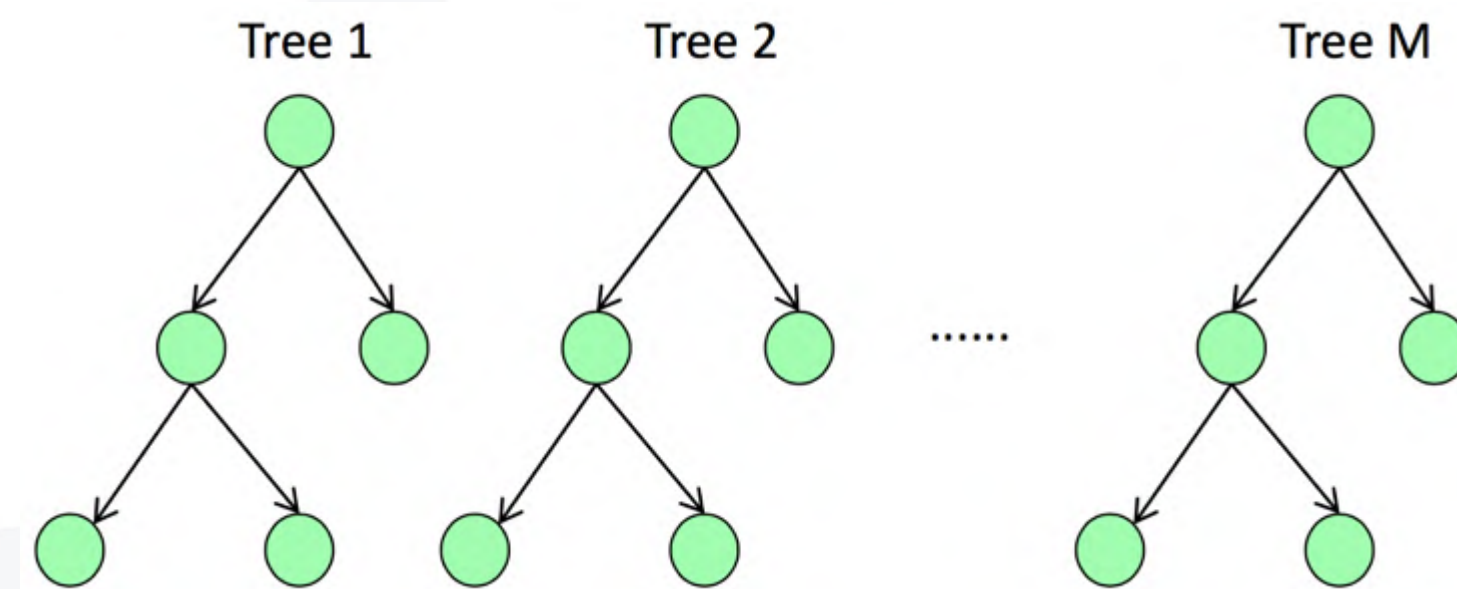
- GBDT模型+FFM模型

Hypothesis: 
$$h_M(x) = \sum_{m=1}^M \beta_m T(x; \Theta_m)$$

Loss Function: 
$$\min_{h \in H} L(h) \equiv \min_{h \in H} \frac{1}{2} \sum_{i=1}^N (y_i - h(x_i))^2$$

Update Function: 
$$\{\hat{\Theta}_m, \beta_m\} = \arg \min_{\Theta_m, \beta_m} \sum_{i=1}^N L(y_i, h_{m-1}(x) + \beta_m T(x; \Theta_m))$$

$$G_j = \sum_{i \in I_j} g_i \quad H_j = \sum_{i \in I_j} h_i$$
$$Obj = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T$$



# 问题建模

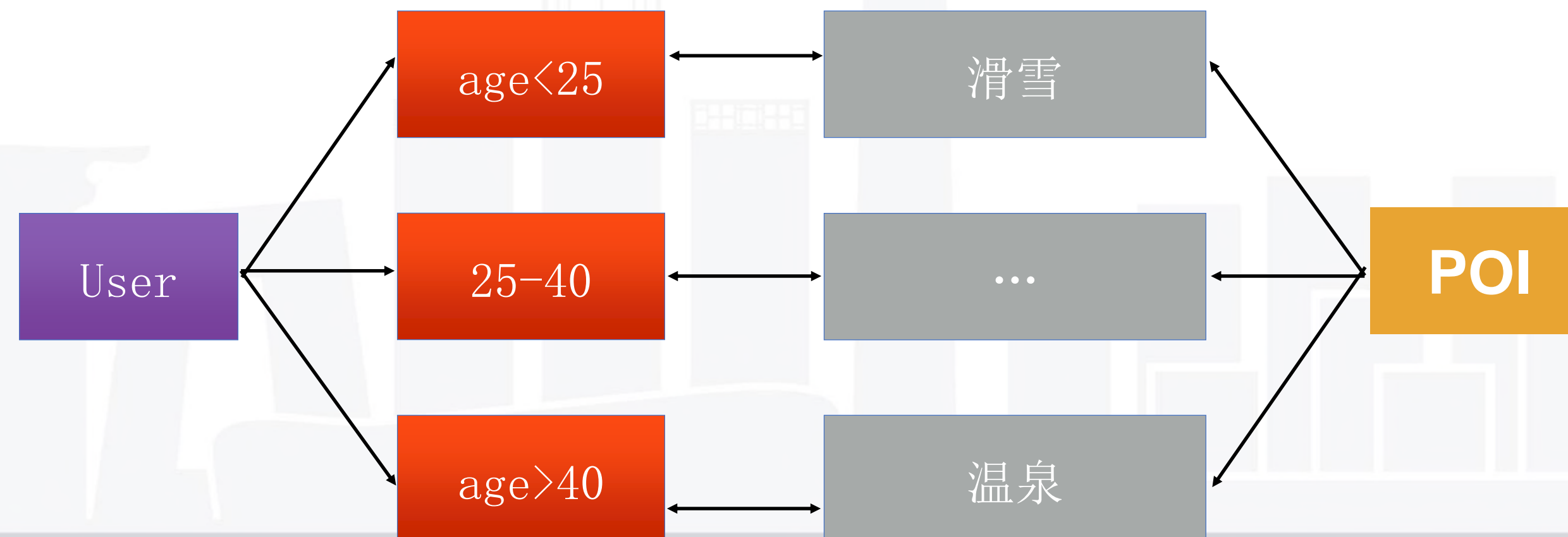
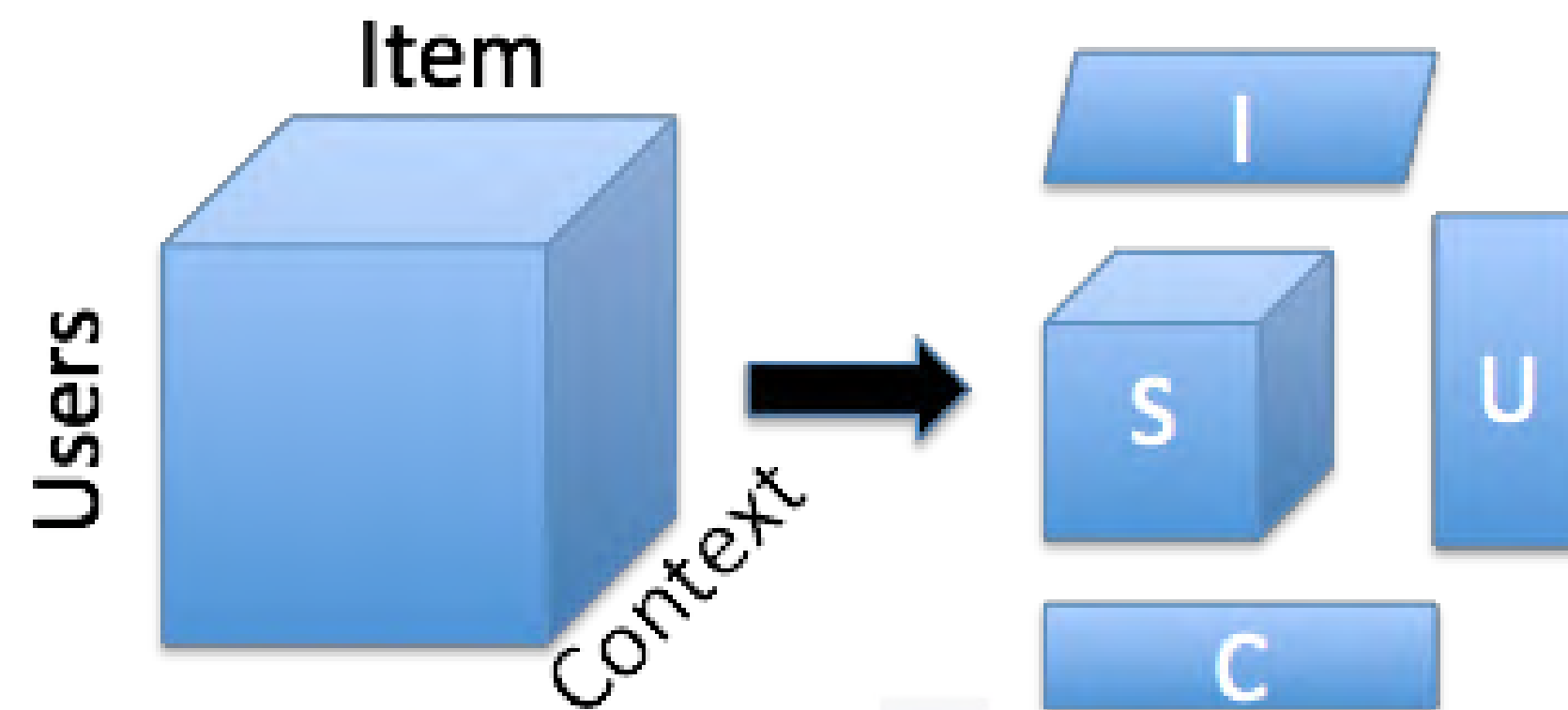
$$y(\mathbf{x}) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n \langle \mathbf{v}_{ij}, \mathbf{v}_{ji} \rangle x_i x_j$$

## • FFM

- 矩阵分解+回归
- Low Level特征多

## • 情景推荐

- 发现特征关联关系
- 用户画像
- 上下文
- POI ID&属性



# 数据标注

## 样本关联

- global id: 串联展示、点击、下单

## 样本选择

- 时间窗口
- 过滤
- 样本迁移
- Skip above

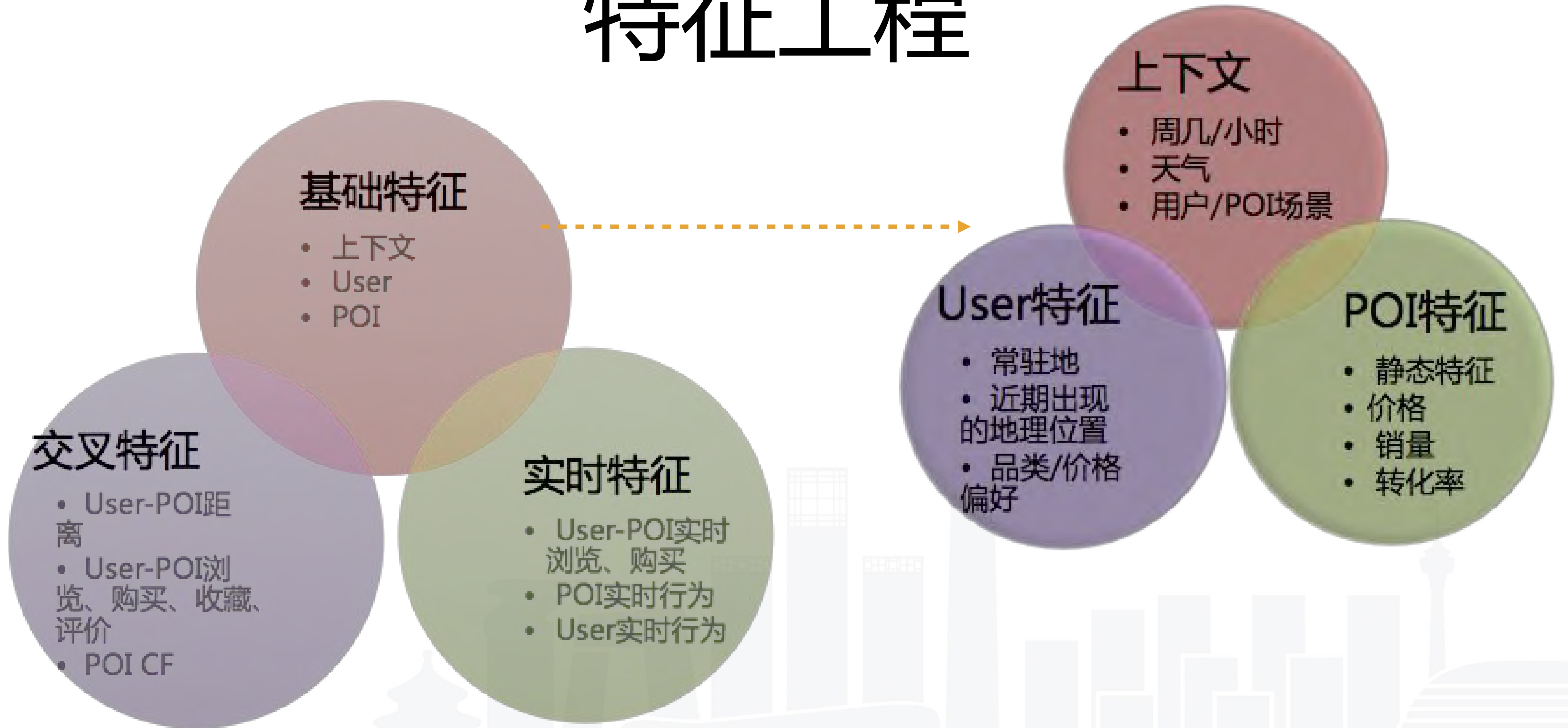
## 样本采样

- 访购模型: 减少负样本
- 点击模型: skip above+2

## 样本权重

- 支付>下单>点击

# 特征工程



# 特征工程

## •特征预处理

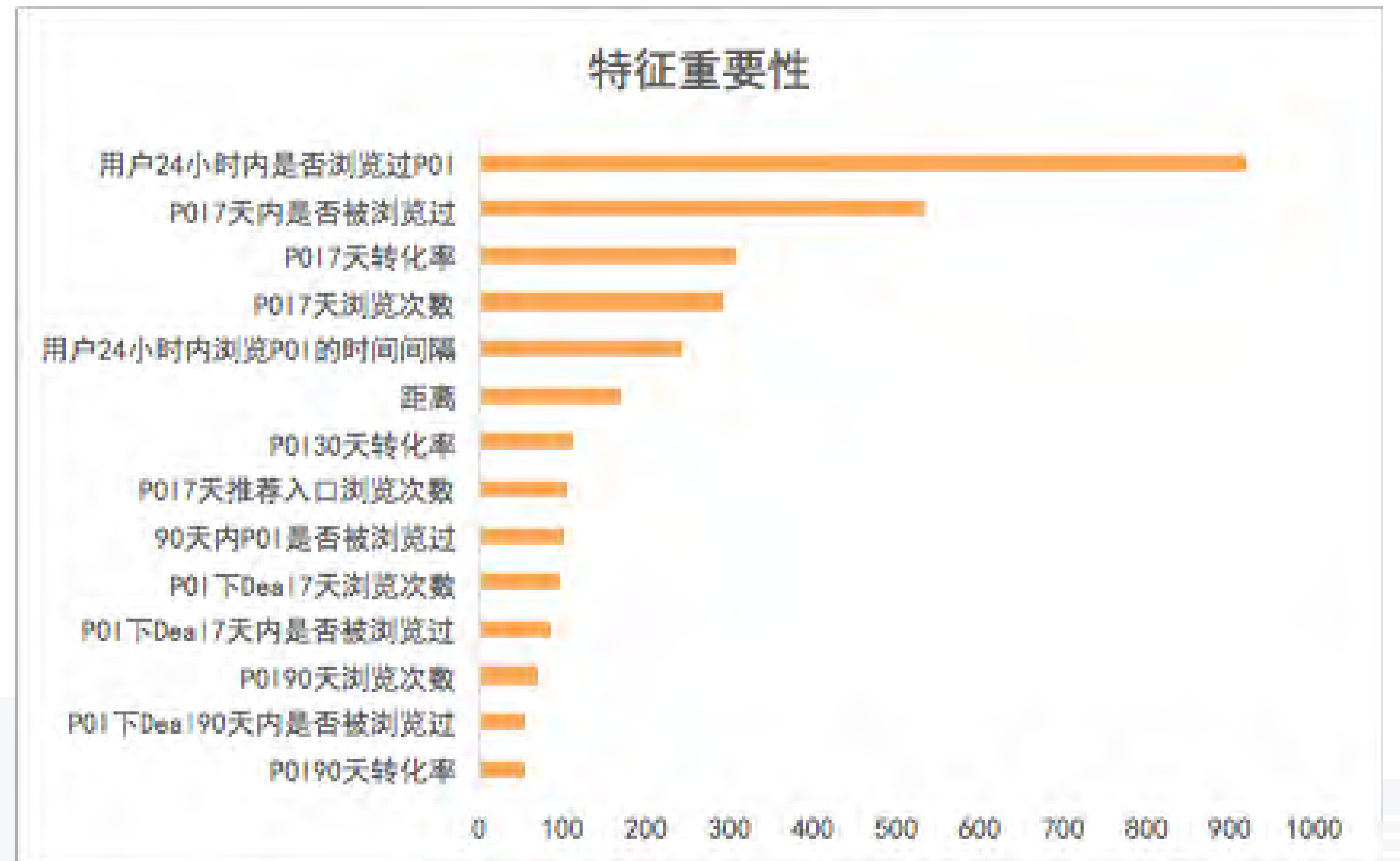
- missing value : 不需要处理
- position bias : COEC
- One-Hot Encoding ? 周几/小时/city id
- Normalize?

## •召回策略特征化

- 销量拆分本异地
- User-POI行为 : 实时/长期
- GeoHash热销
- POI CF

## •特征选择

- 特征在每棵树每个节点的信息增益之和



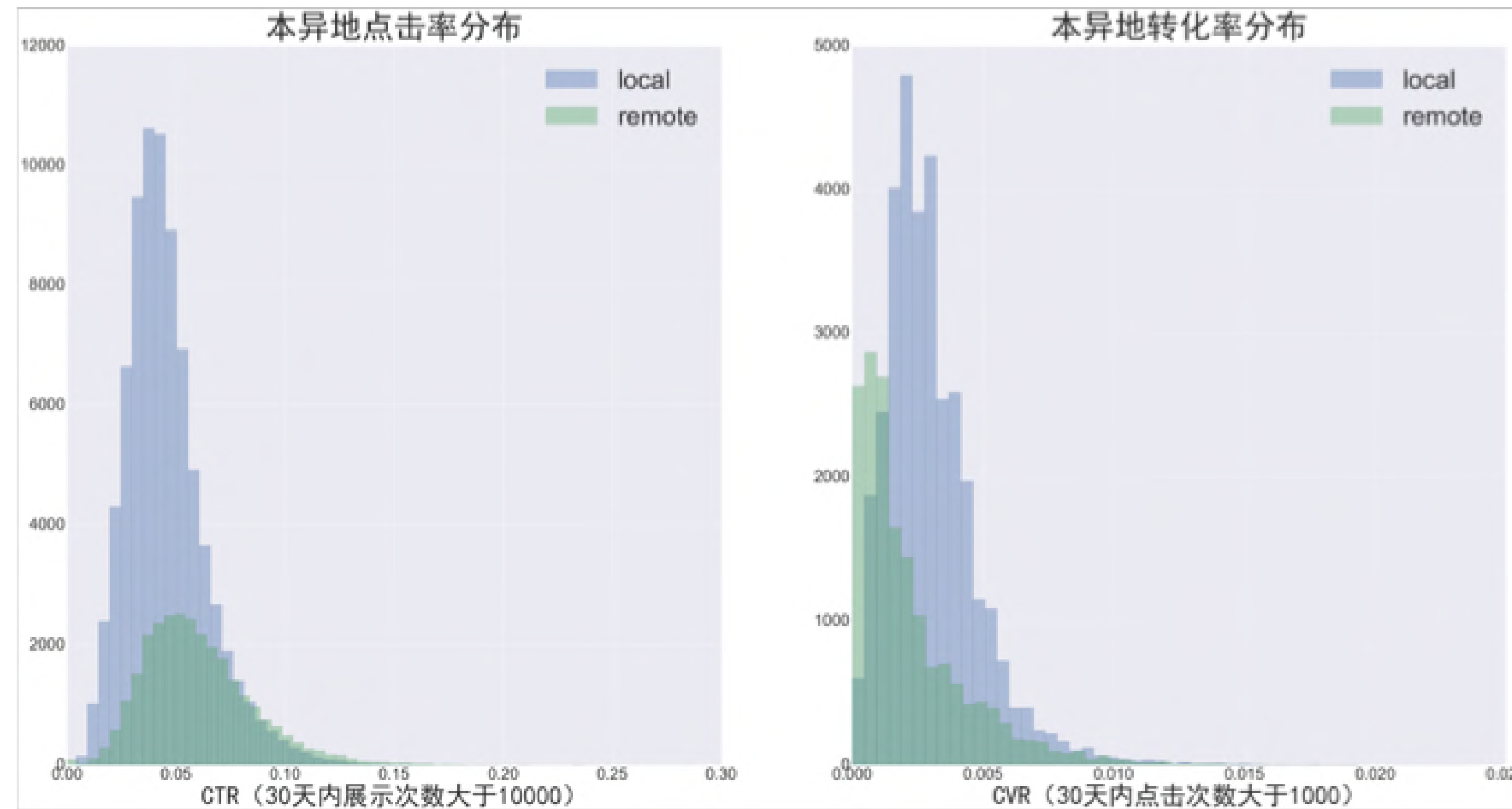
# 特征工程

## •特征分析

- 分本异地统计转化率、销量
- 天气

## •特征监控

- 覆盖率
- 值域范围
- 分布异常



特征	coverRatio			numValid		
	参考值	监控值	波动率	参考值	监控值	波动率
CLICKNET	0.931341	0.481258	0.483263	3271405	1745754	0.46636
DISTANCE	0.651660	0.033357	0.948812	13447	662	0.950770

# 模型训练

## •模型训练

- 单机VS分布式
- 目标函数：binary:logistic
- 过拟合VS欠拟合
  - 样本大小&树的棵数
  - 样本和特征随机采样
  - 模型复杂度：max\_depth, min\_child\_weight

## •通用离线训练工具

- 流程抽象化、组件化
- 提供公共组件，支持定制组件





# 效果评估&线上迭代

## • 离线评估

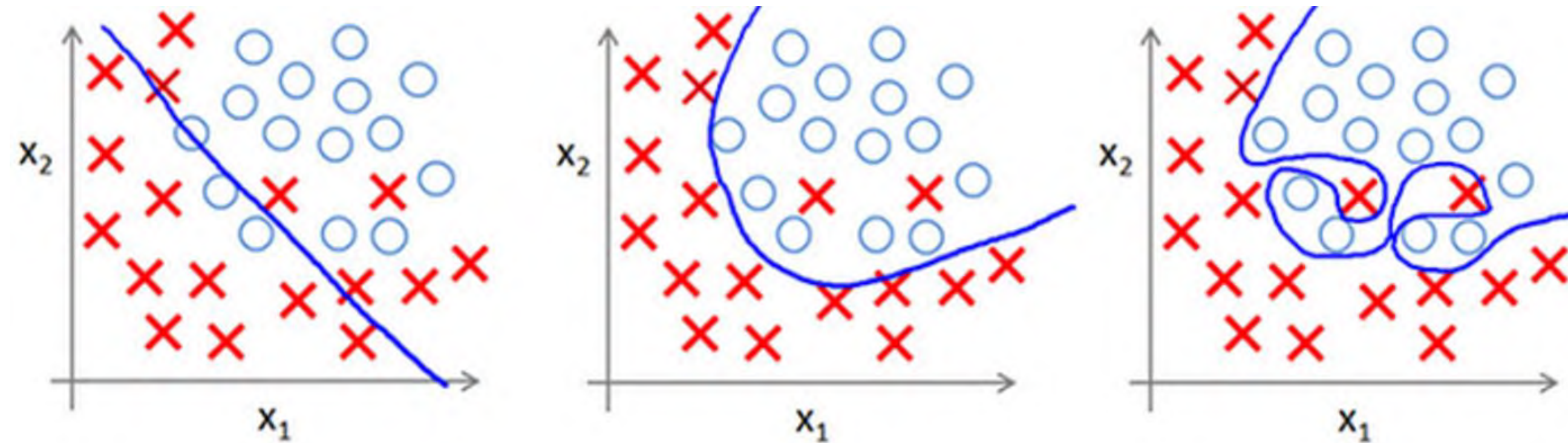
- AUC
- logloss

## • 在线评估

- ABTest：按UUID分流

## • 线上迭代

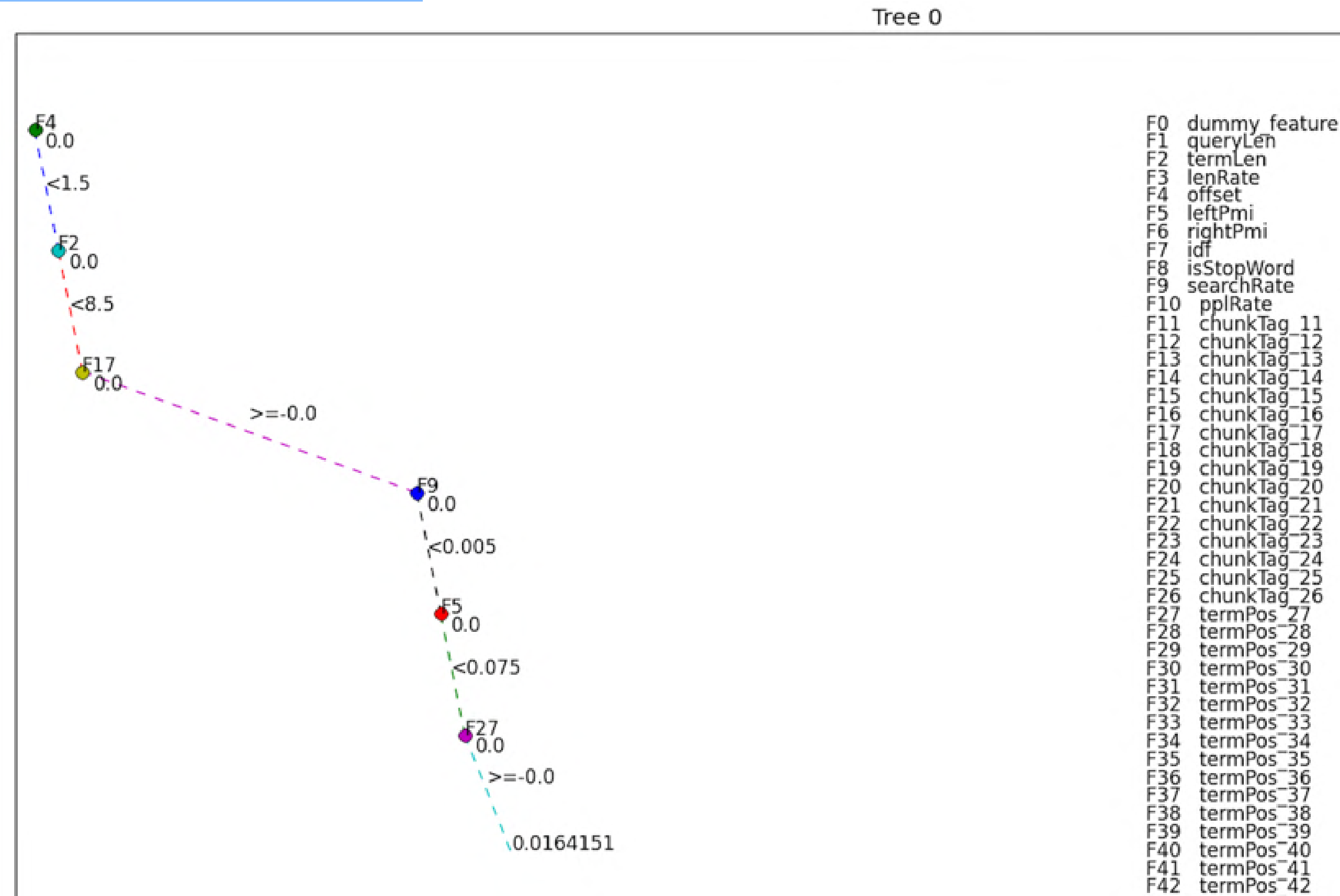
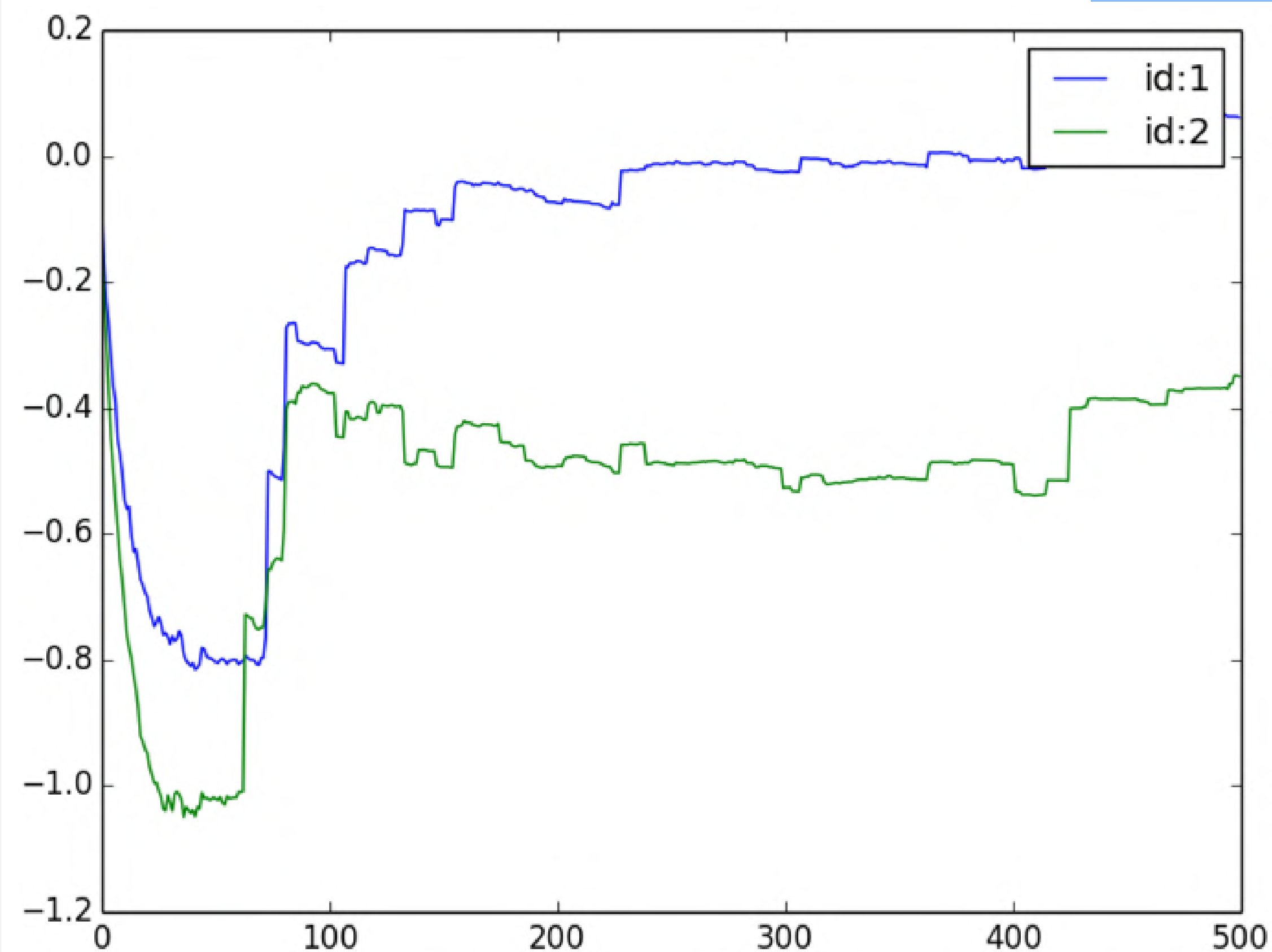
- 模型更新
- 特征漂移：更新延迟



训练集表现	测试集表现	问题
<期望目标值	<期望目标值	Underfitting
>期望目标值	接近或略逊于训练集	合适
>期望目标值	远差于训练集	Overfitting

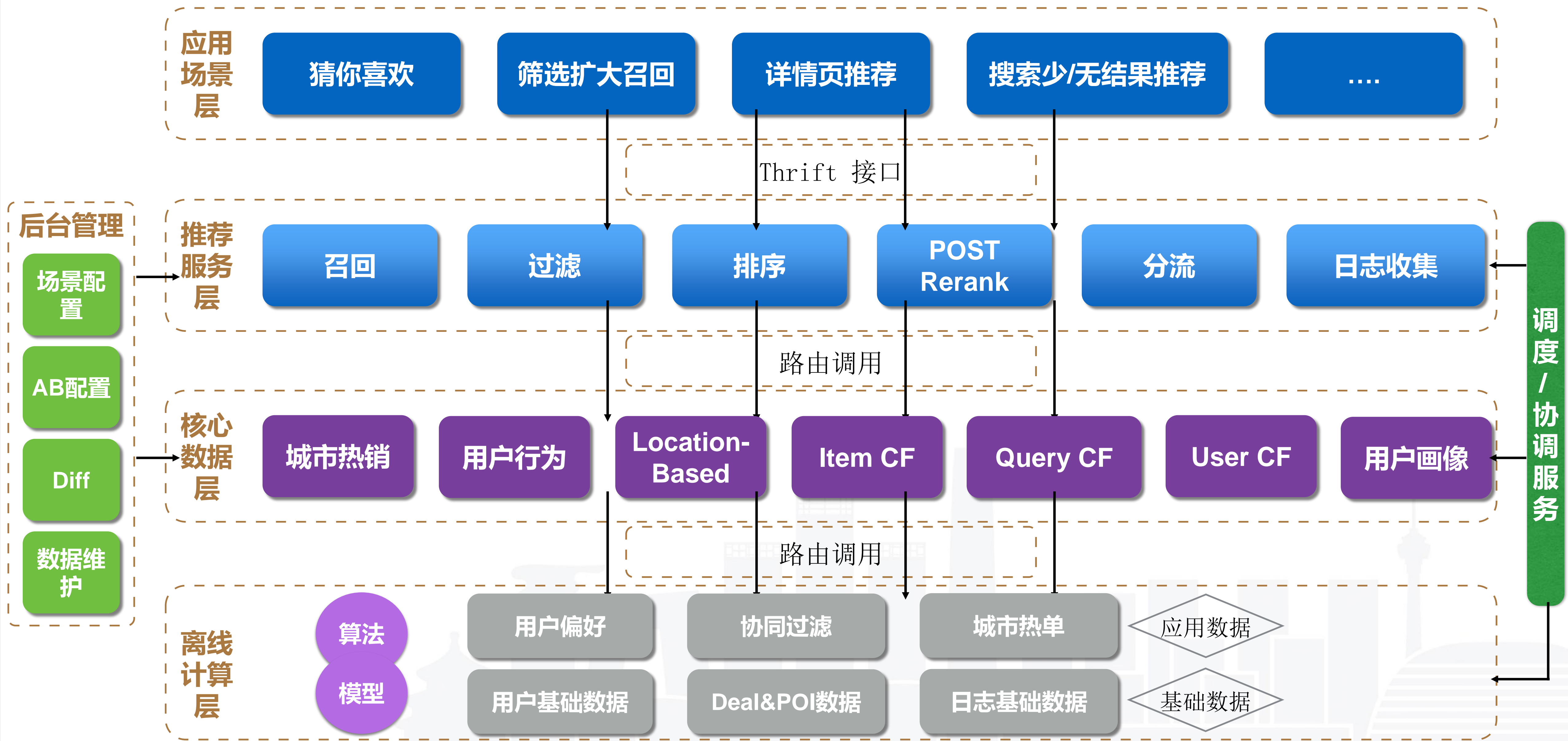
# 模型调试

## 模型Debug工具

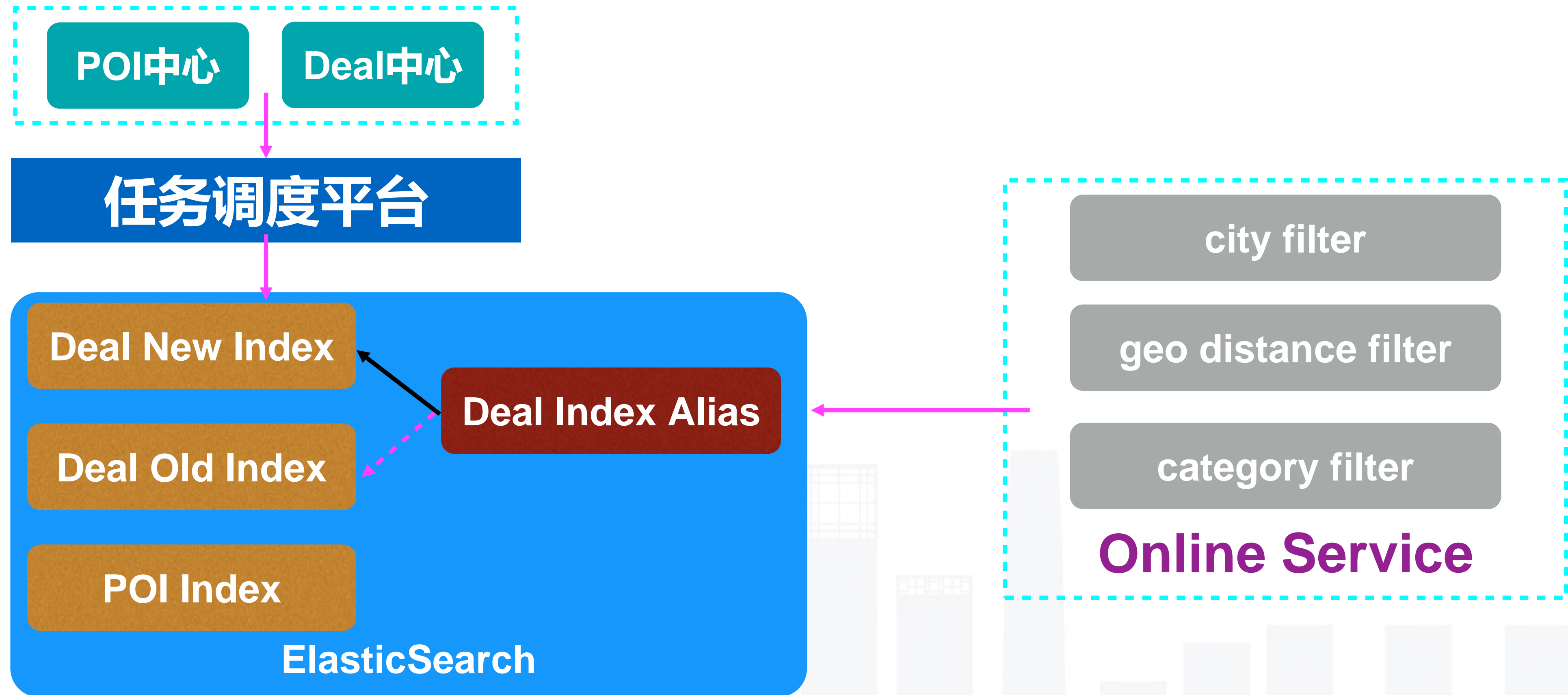


- F0 dummy feature
- F1 queryLen
- F2 termLen
- F3 lenRate
- F4 offset
- F5 leftPmi
- F6 rightPmi
- F7 idf
- F8 isStopWord
- F9 searchRate
- F10 pplRate
- F11 chunkTag\_11
- F12 chunkTag\_12
- F13 chunkTag\_13
- F14 chunkTag\_14
- F15 chunkTag\_15
- F16 chunkTag\_16
- F17 chunkTag\_17
- F18 chunkTag\_18
- F19 chunkTag\_19
- F20 chunkTag\_20
- F21 chunkTag\_21
- F22 chunkTag\_22
- F23 chunkTag\_23
- F24 chunkTag\_24
- F25 chunkTag\_25
- F26 chunkTag\_26
- F27 termPos\_27
- F28 termPos\_28
- F29 termPos\_29
- F30 termPos\_30
- F31 termPos\_31
- F32 termPos\_32
- F33 termPos\_33
- F34 termPos\_34
- F35 termPos\_35
- F36 termPos\_36
- F37 termPos\_37
- F38 termPos\_38
- F39 termPos\_39
- F40 termPos\_40
- F41 termPos\_41
- F42 termPos\_42

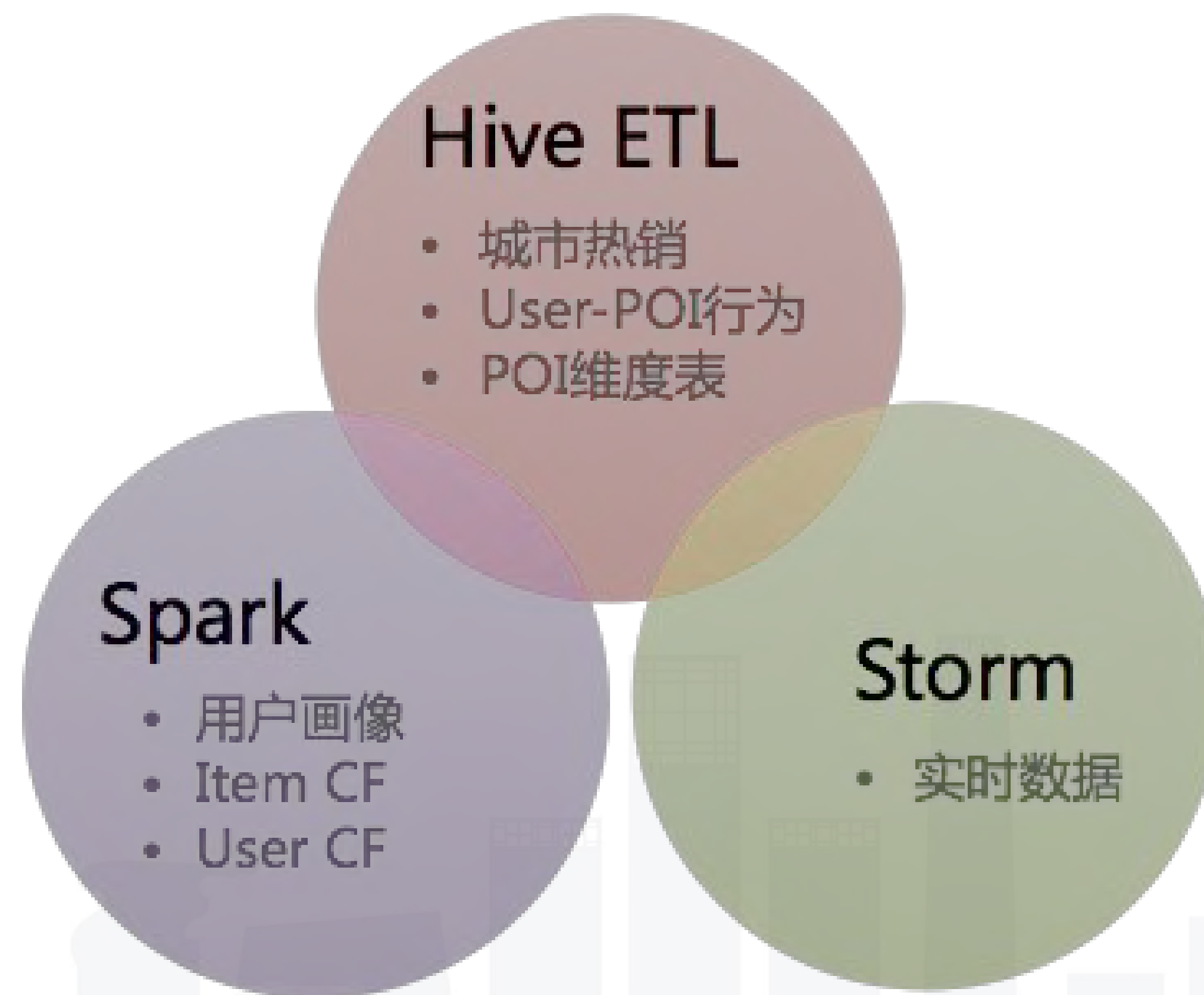
# 从海量大数据的离线计算到高并发在线服务的推荐引擎架构设计



# 基础数据



# 应用数据



# 应用数据线上化-DataHub

## • 特征抽取

- 统一特征抽取调度
- 精确控制数据导入速率，避免并发写压力过大

## • 特征存储

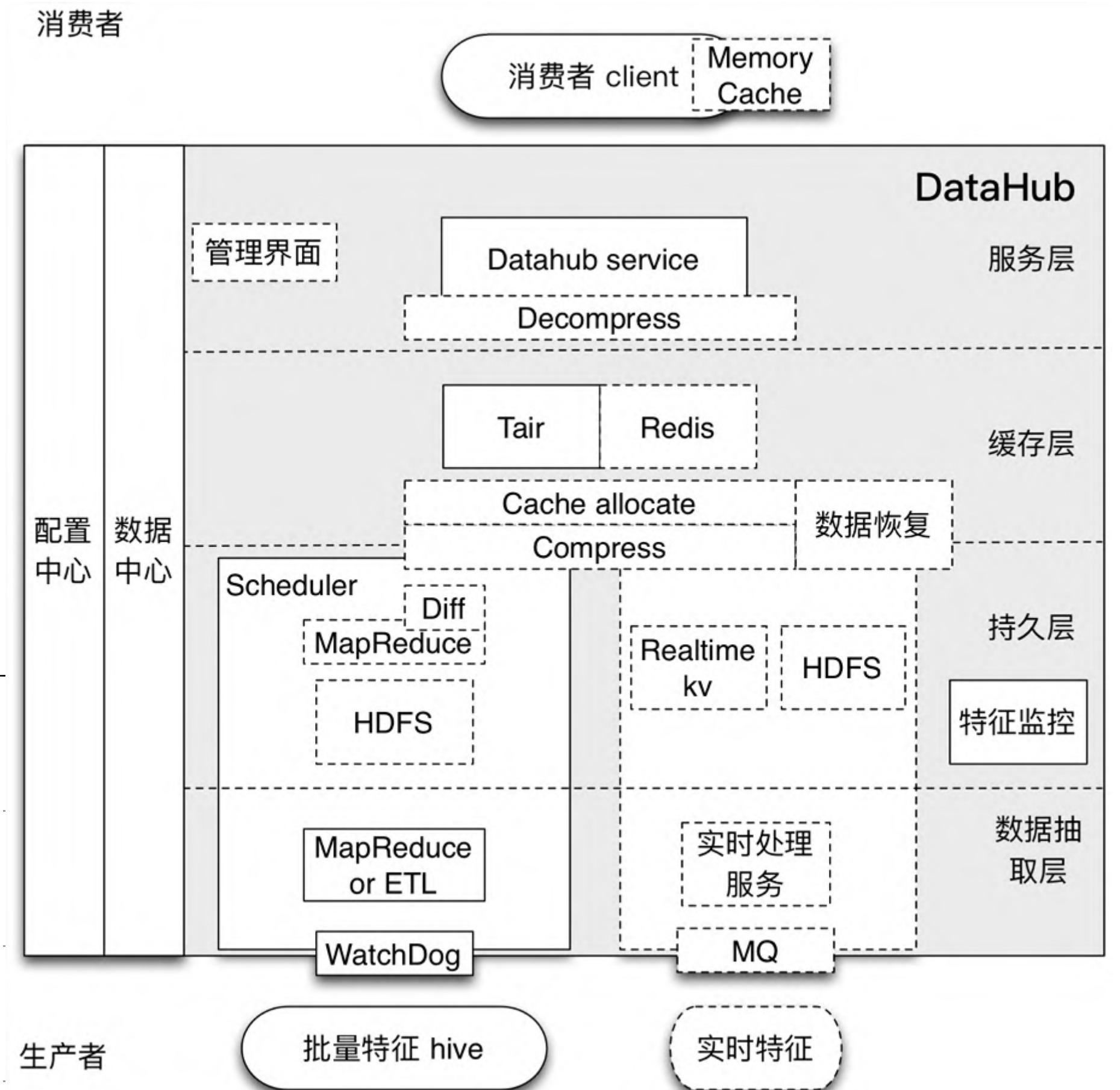
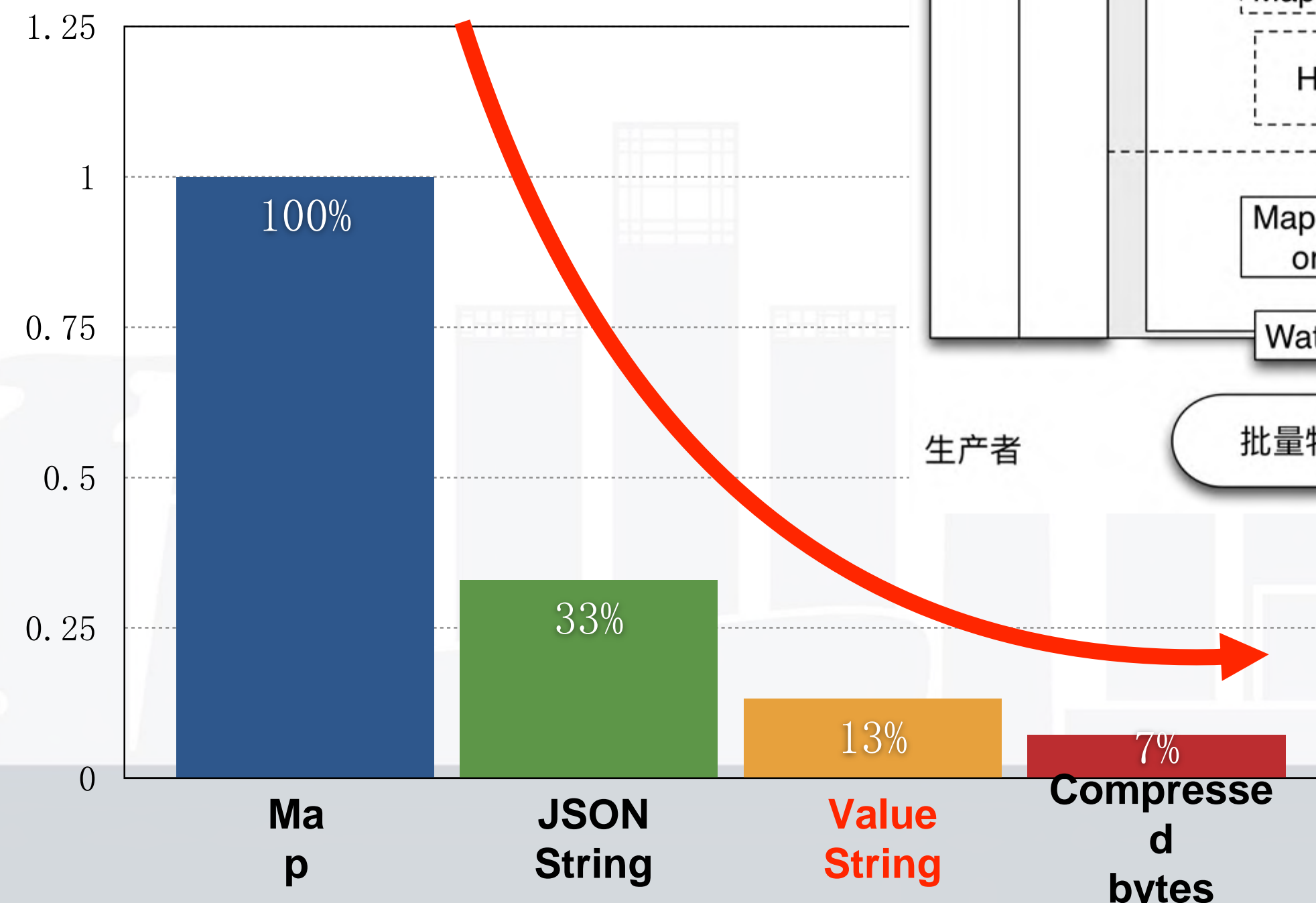
- 数据压缩：Value String

## • 特征管理

- 特征注册、特征监控

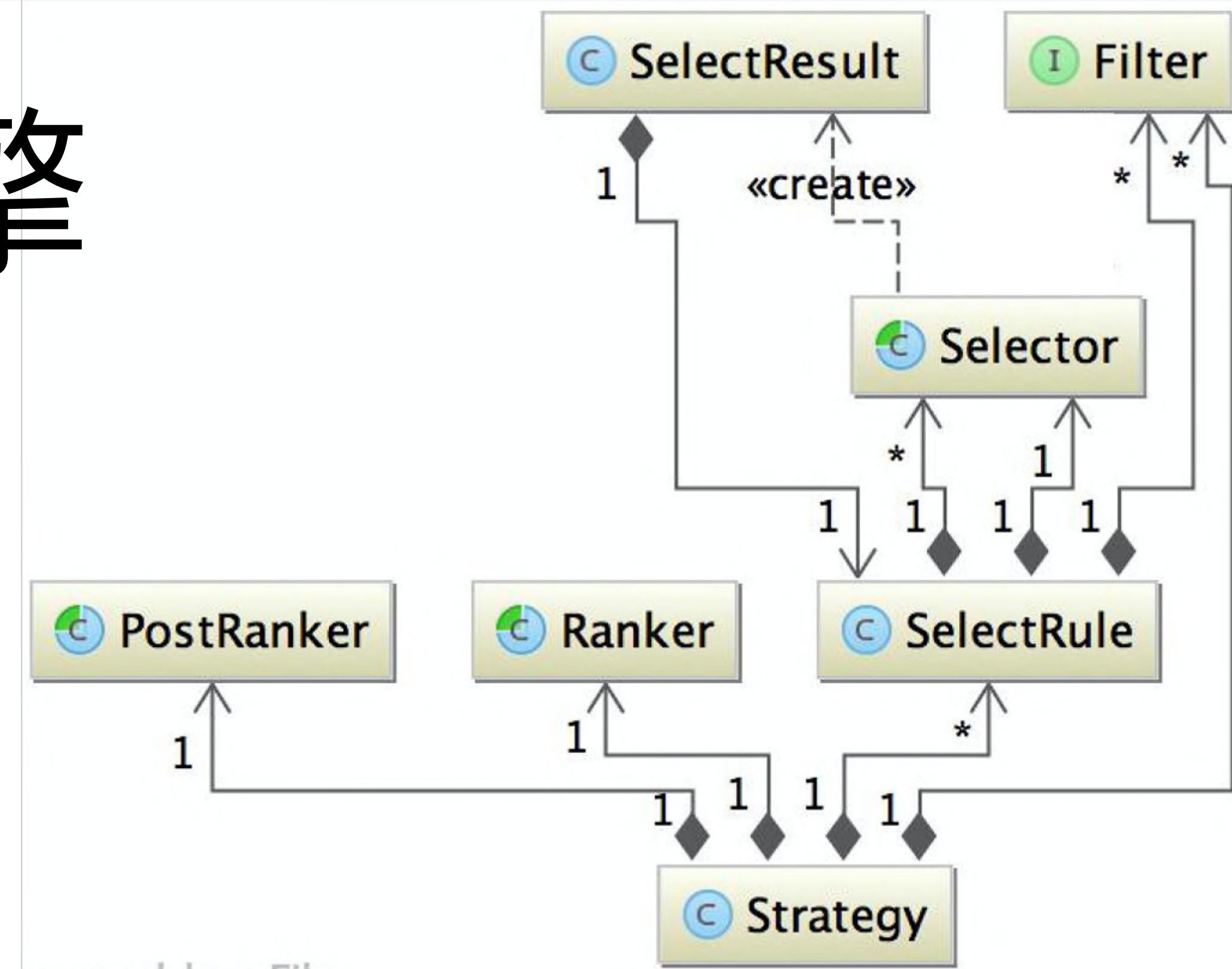
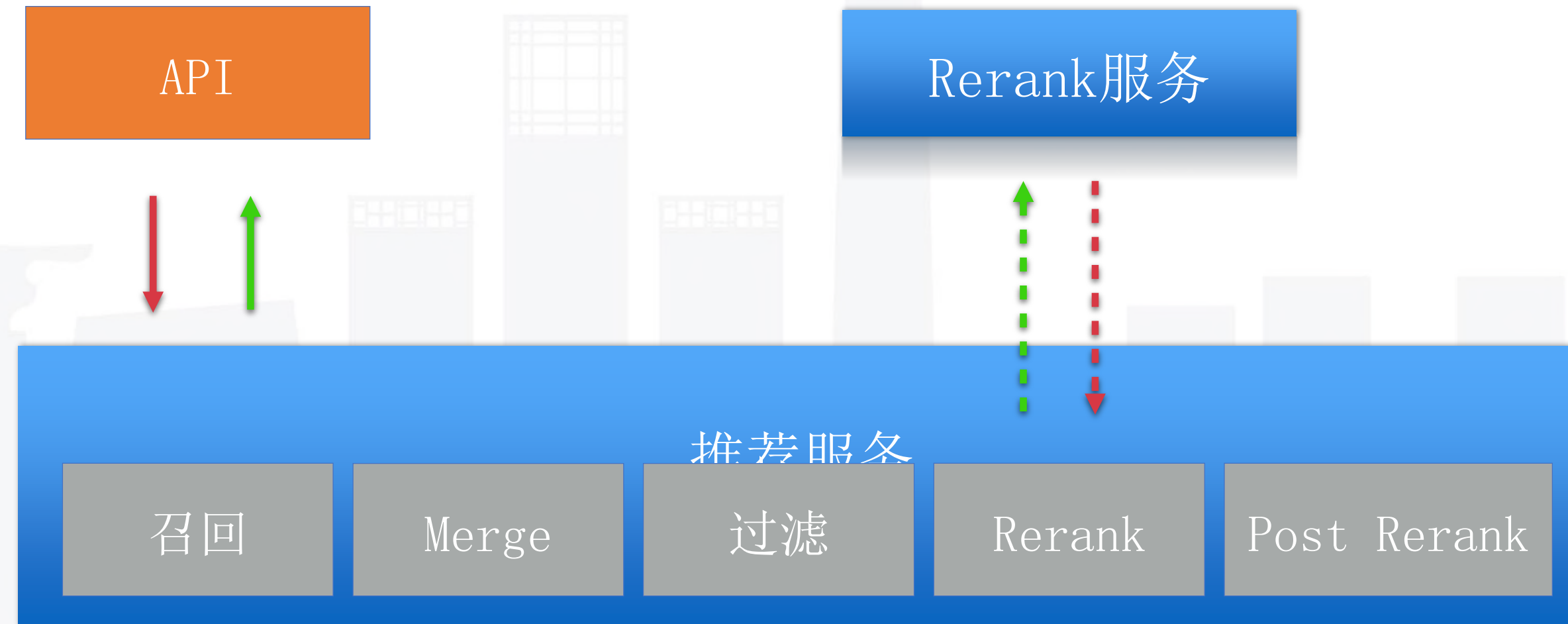
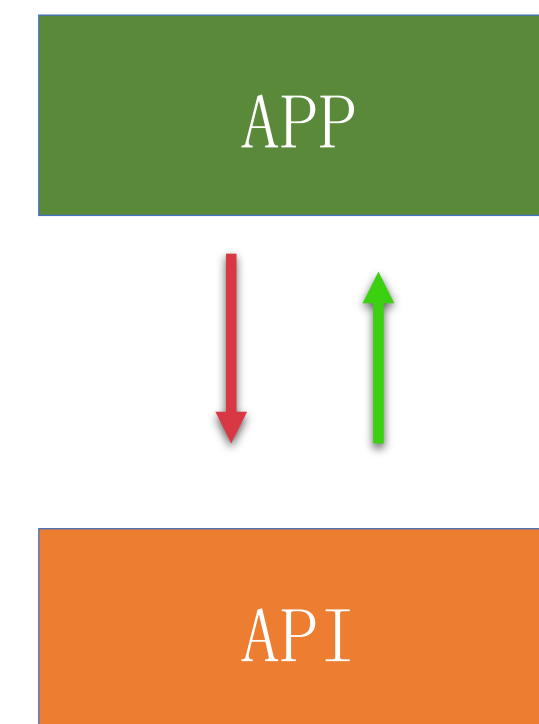
## • 特征消费

- Client缓存：Direct Momery
- 异步化：Thrift Async



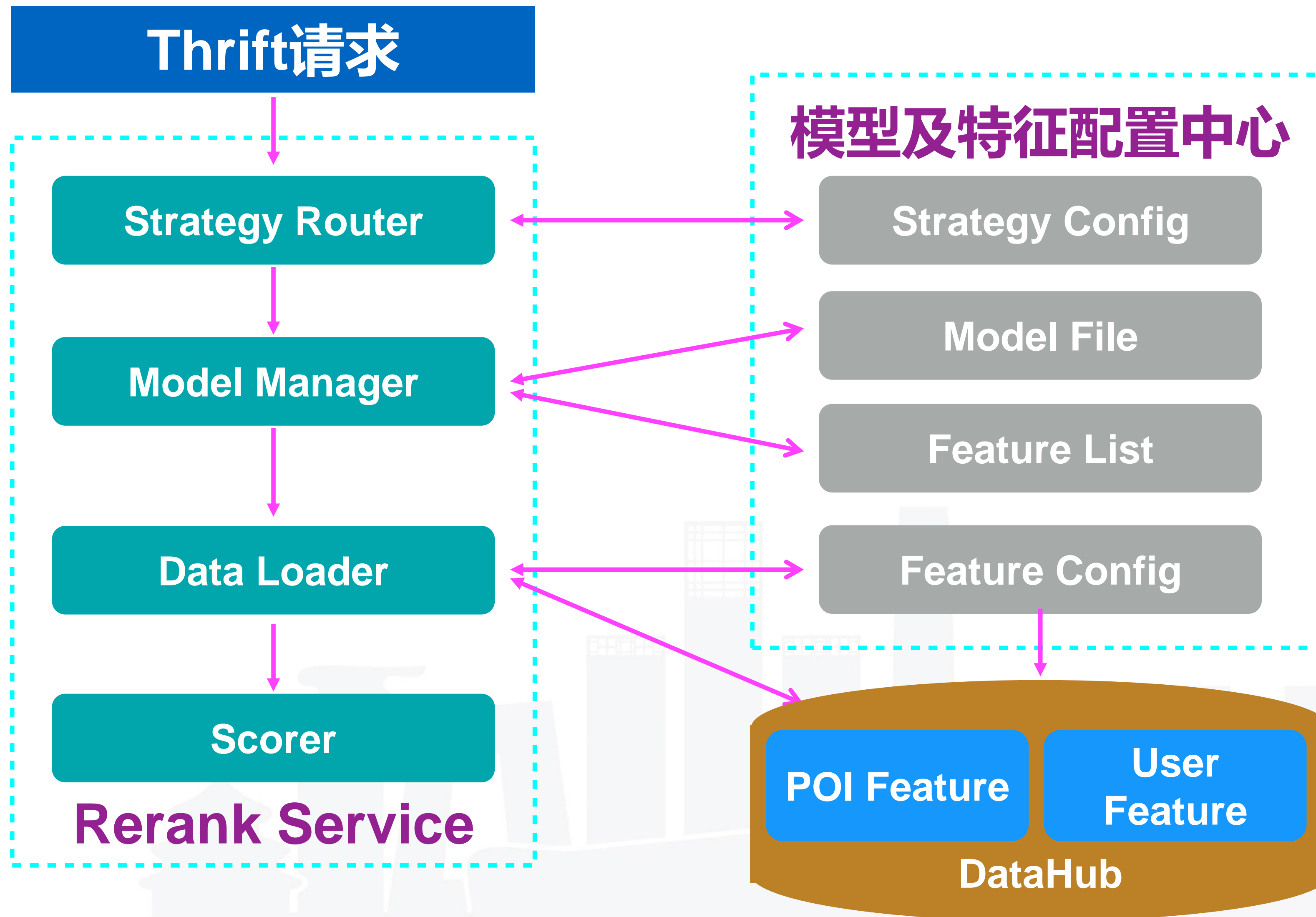
# 推荐引擎

- 召回
  - Booth : 场景
  - Strategy : Baseline
  - SelectRule : Location-Based
  - Selector : 区域热销POI
- Merge : 子策略融合
  - 调制
  - 分级
- 过滤
  - 通用过滤策略 : 黑名单
  - 针对某类召回策略 : 浏览未购买
- Rerank : 个性化排序
- Post Rerank





# 推荐引擎-Rerank



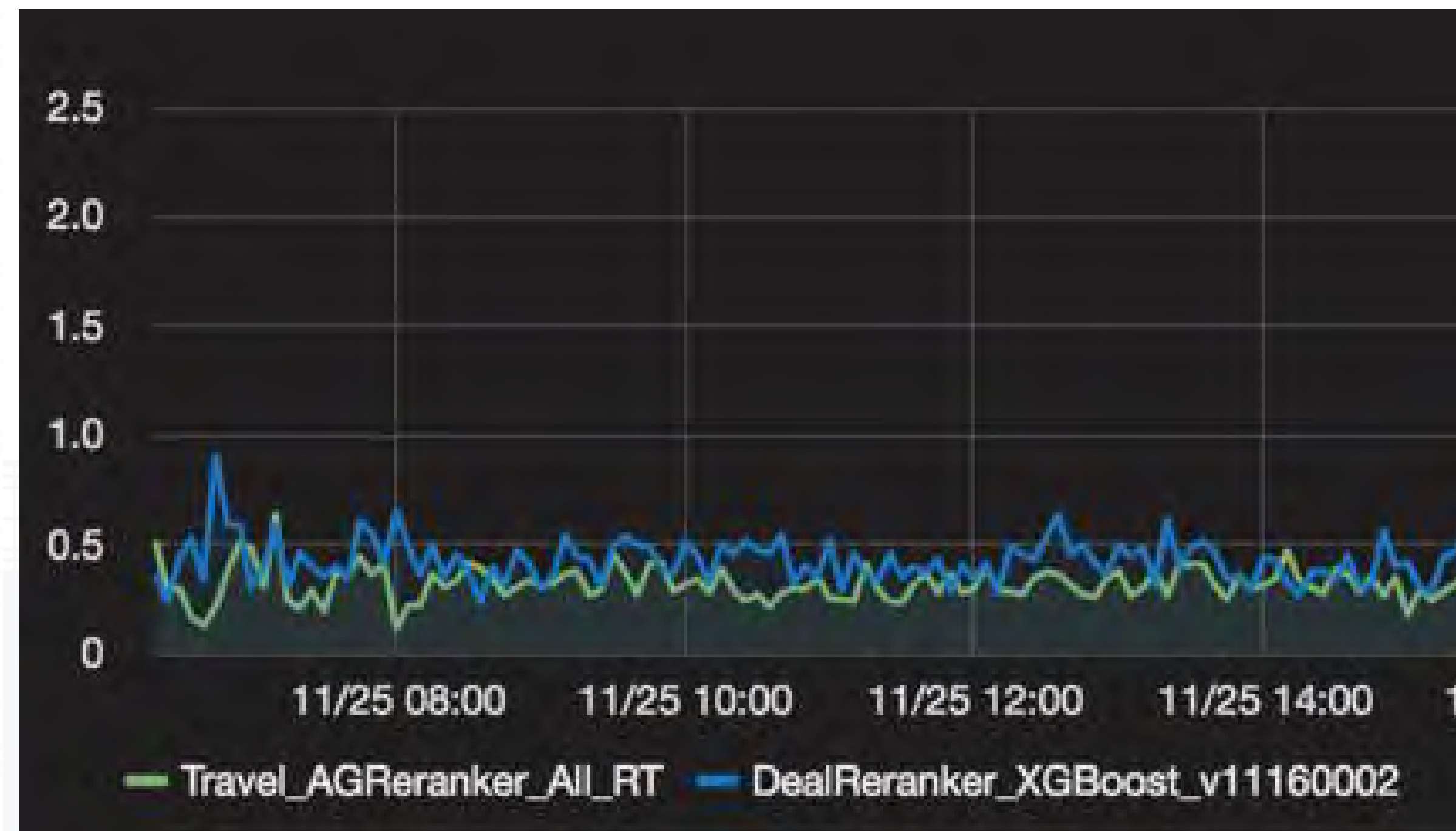
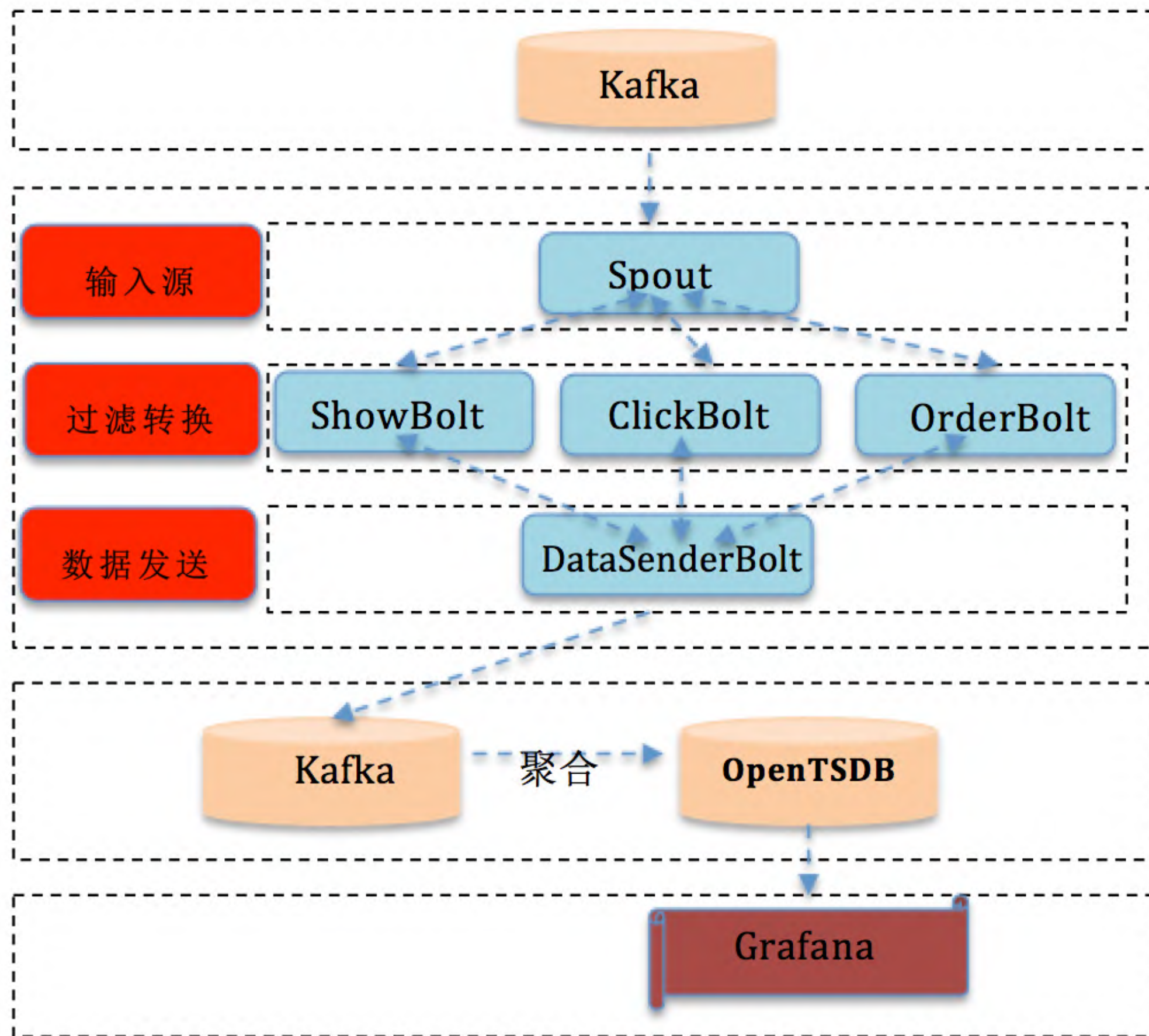
# 实时策略效果统计

数据源

Storm

数据存储

监控展示



# 推荐在美团点评酒旅的应用实践



# 推荐应用

## 美团/点评双平台

### 频道首页

- 品类区
- Banner
- 头条
- 人气区
- 猜你喜欢

### 搜索结果页

- 少结果推荐
- 无结果推荐

### 筛选结果页

- 筛选标签
- 异地召回

### 详情页

- 看了又看
- 附近景点
- 附近酒店



# 推荐应用

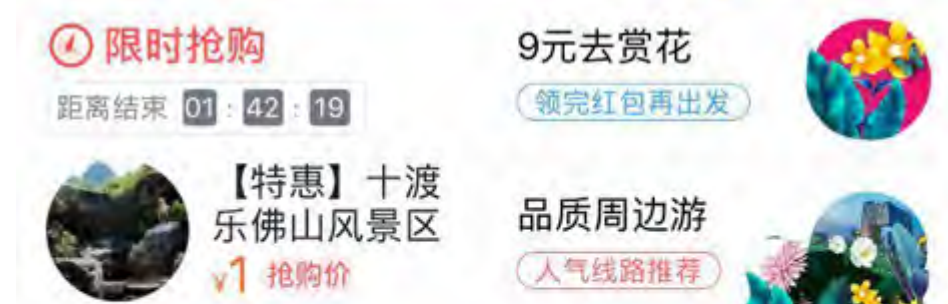
品类区



banner



周边游频道首页



人气区



角标



看了又看

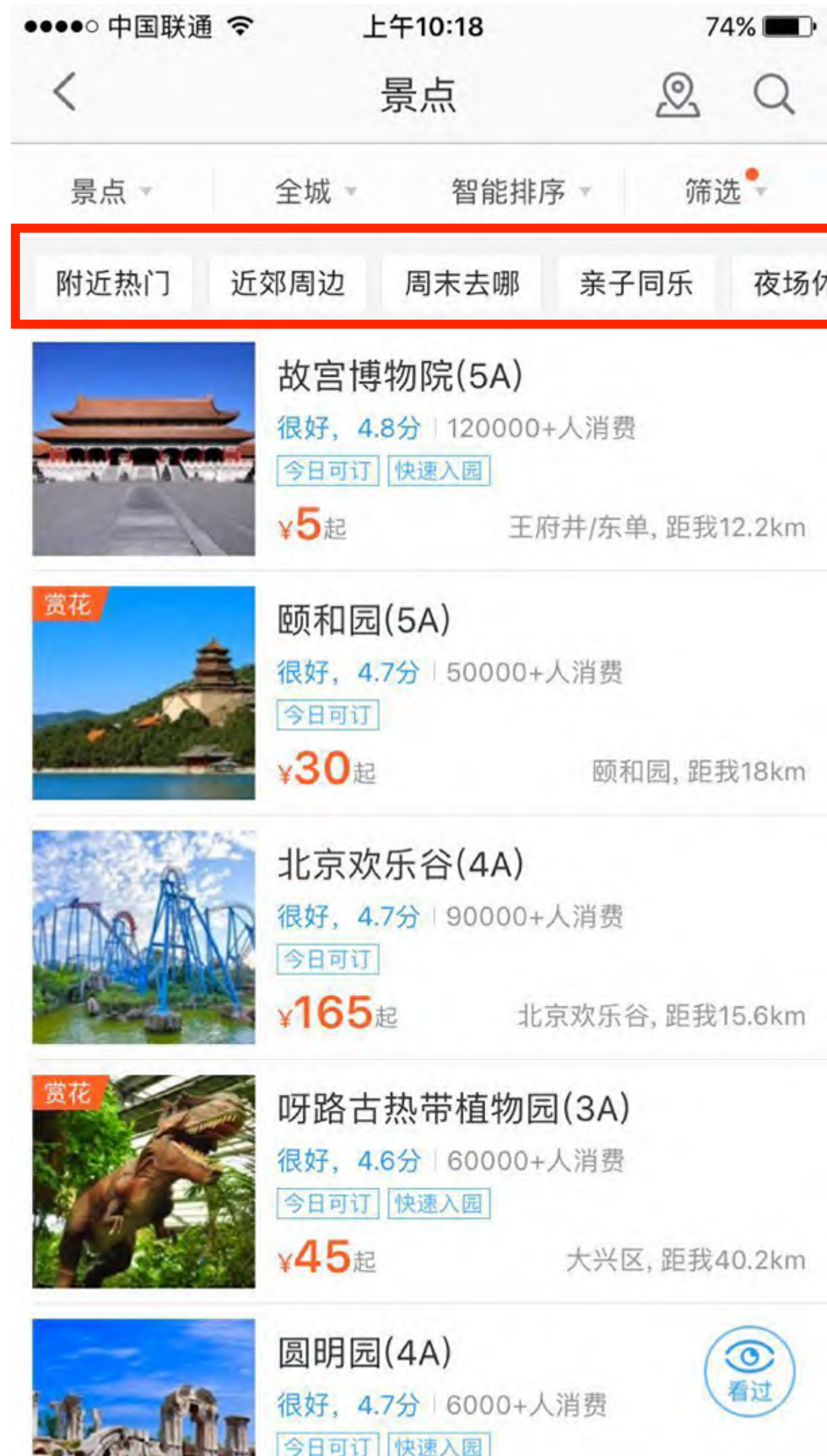
景点POI详情页



酒旅交叉推荐

# 推荐应用

POI标签



## 筛选列表页

筛选异地召回



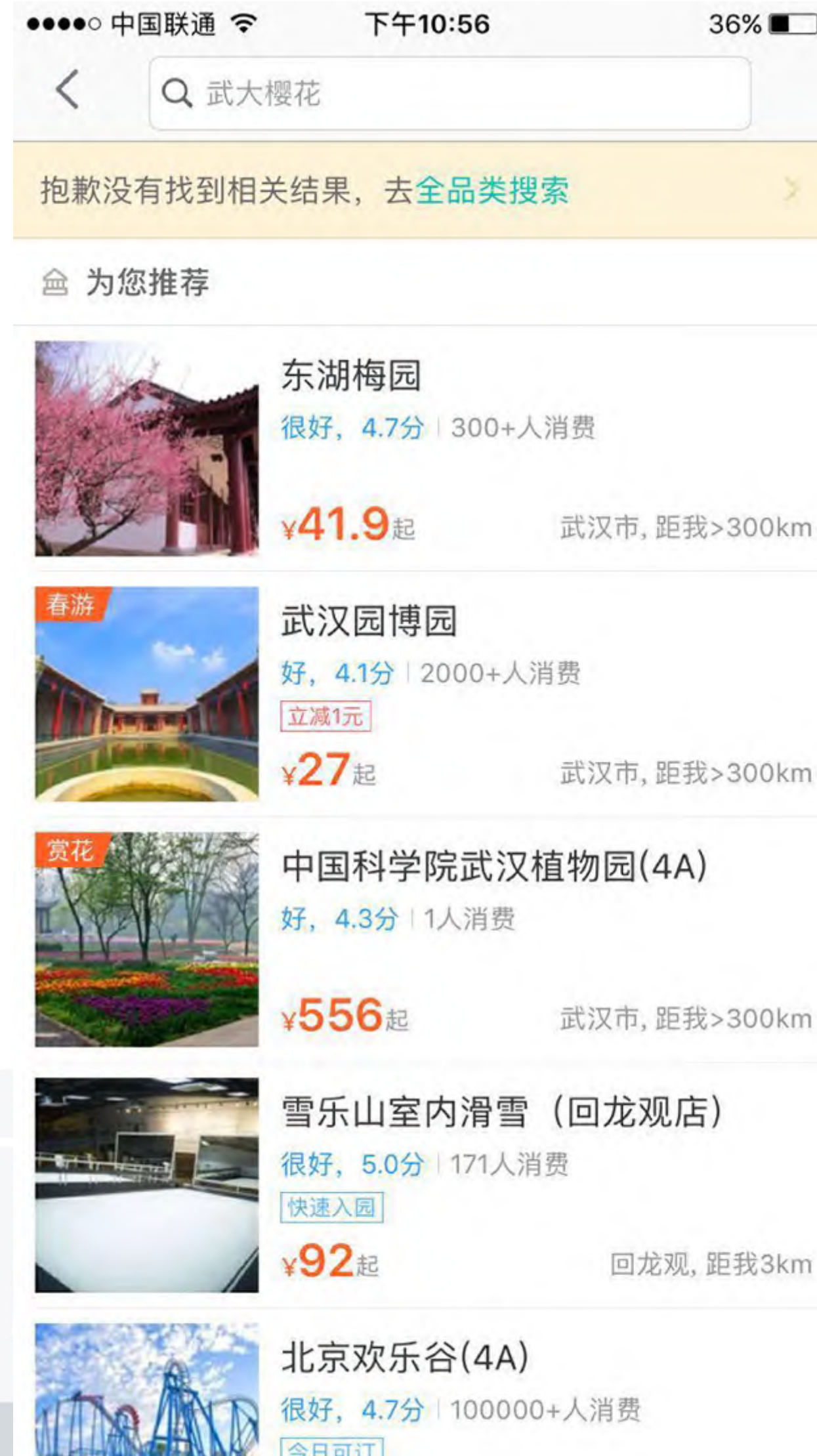
# 搜索少/无结果推荐

- 无结果推荐

- 查询改写
- Query CF
- Location-Based
- 热销POI

- 少结果推荐

- POI CF
- POI同品类推荐



# 总结

## 本异地差异大

- 基于常驻城市定义本异地
- 召回：热销区分本异地
- 排序：分本异地统计销量、转化率

## 推荐形式多样

- 基于埋点精确统计景点POI销量
- 游记攻略头条

## 季节性明显

- 销量加速
- POI标签

## 需求个性化

- 用户历史行为：浏览、收藏、搜索、下单
- 相似用户/相似Item
- 用户画像







关注QCon微信公众号，  
获得更多干货！

# Thanks!



主办方 **Geekbang** > **InfoQ**  
极客邦科技