
亿级视频广告事件预测系统构建之道

潘晓彤

4月 2017

FreeWheel



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

关于FreeWheel

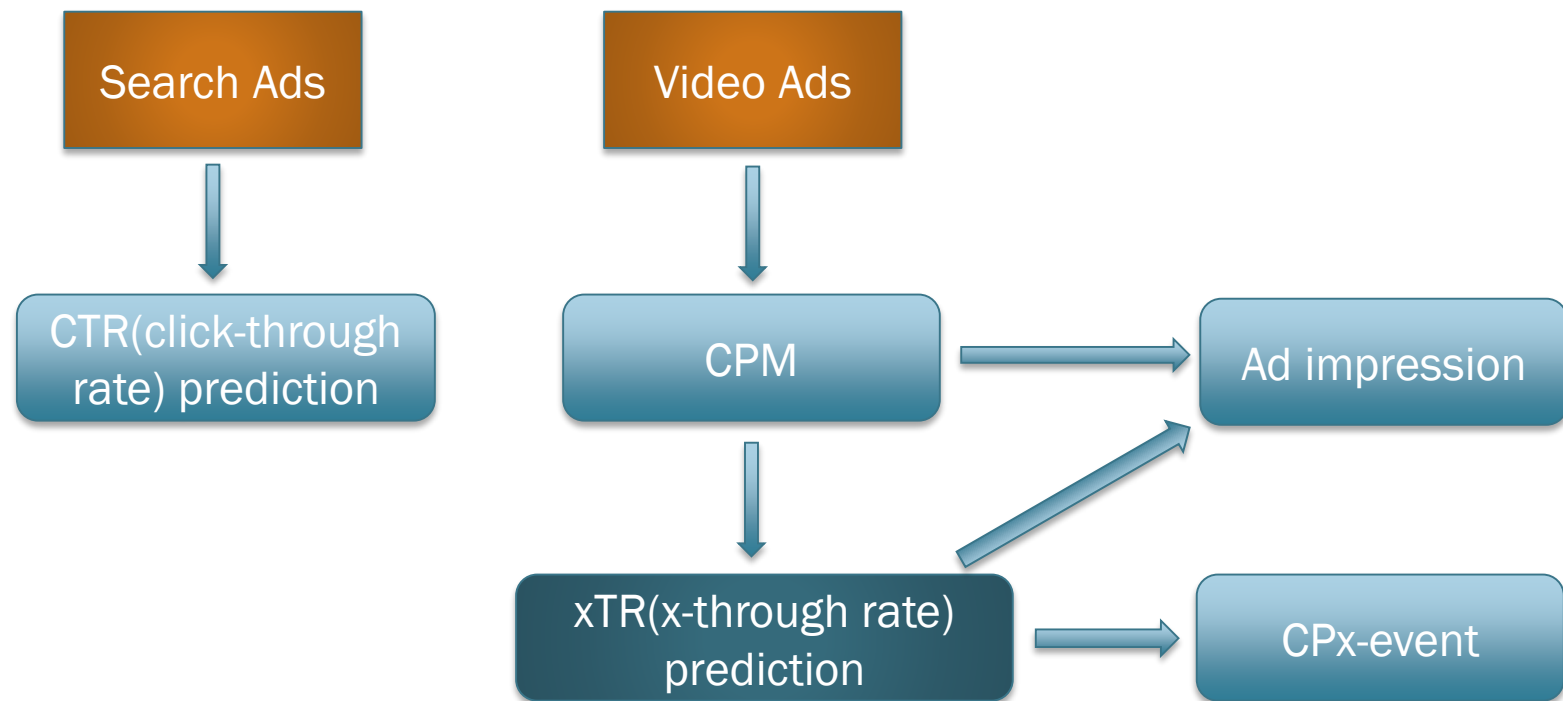
- 视频广告解决方案

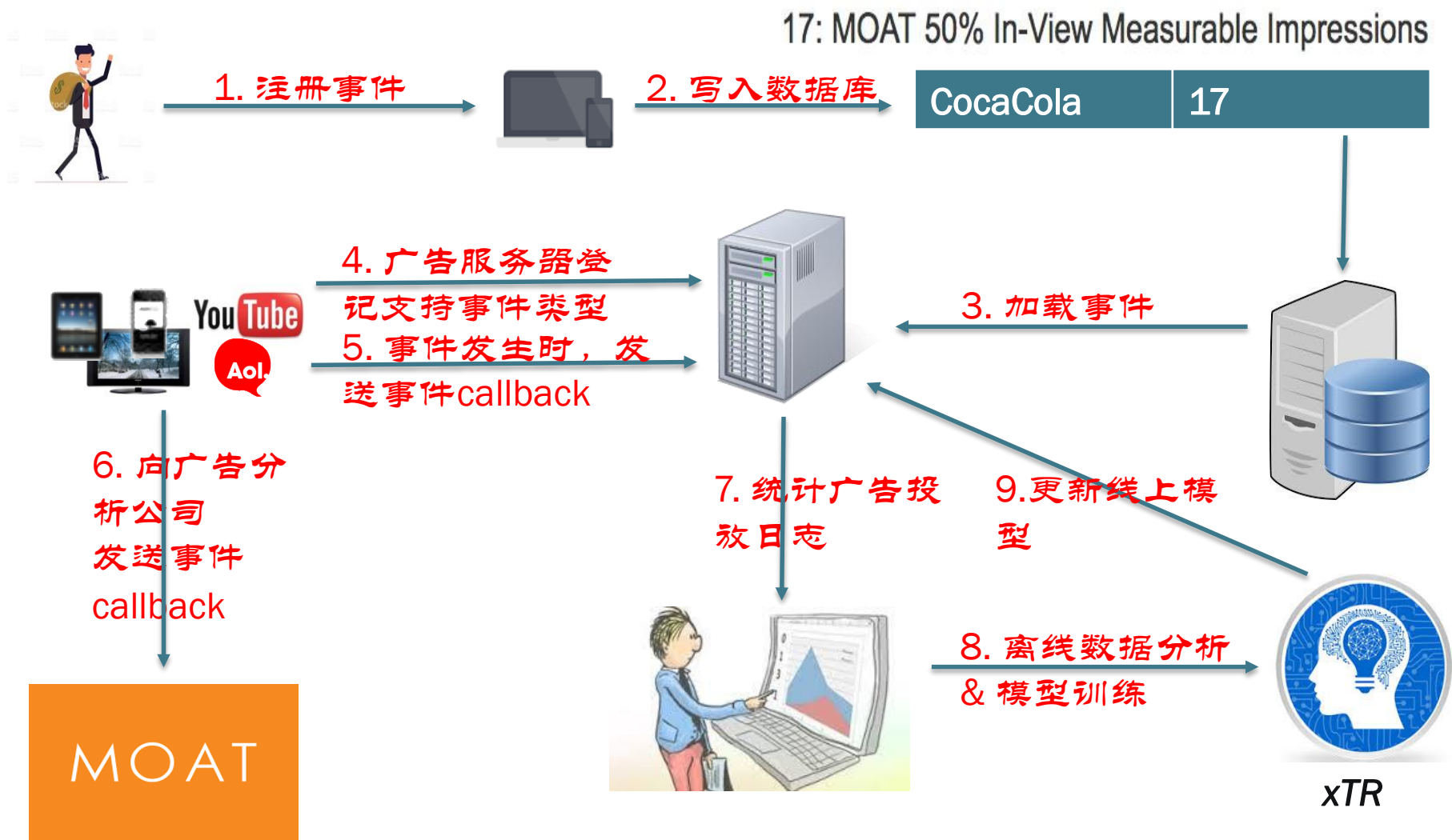
提纲

- xTR是什么
- xTR系统架构
- 特征提取
- 模型训练与优化

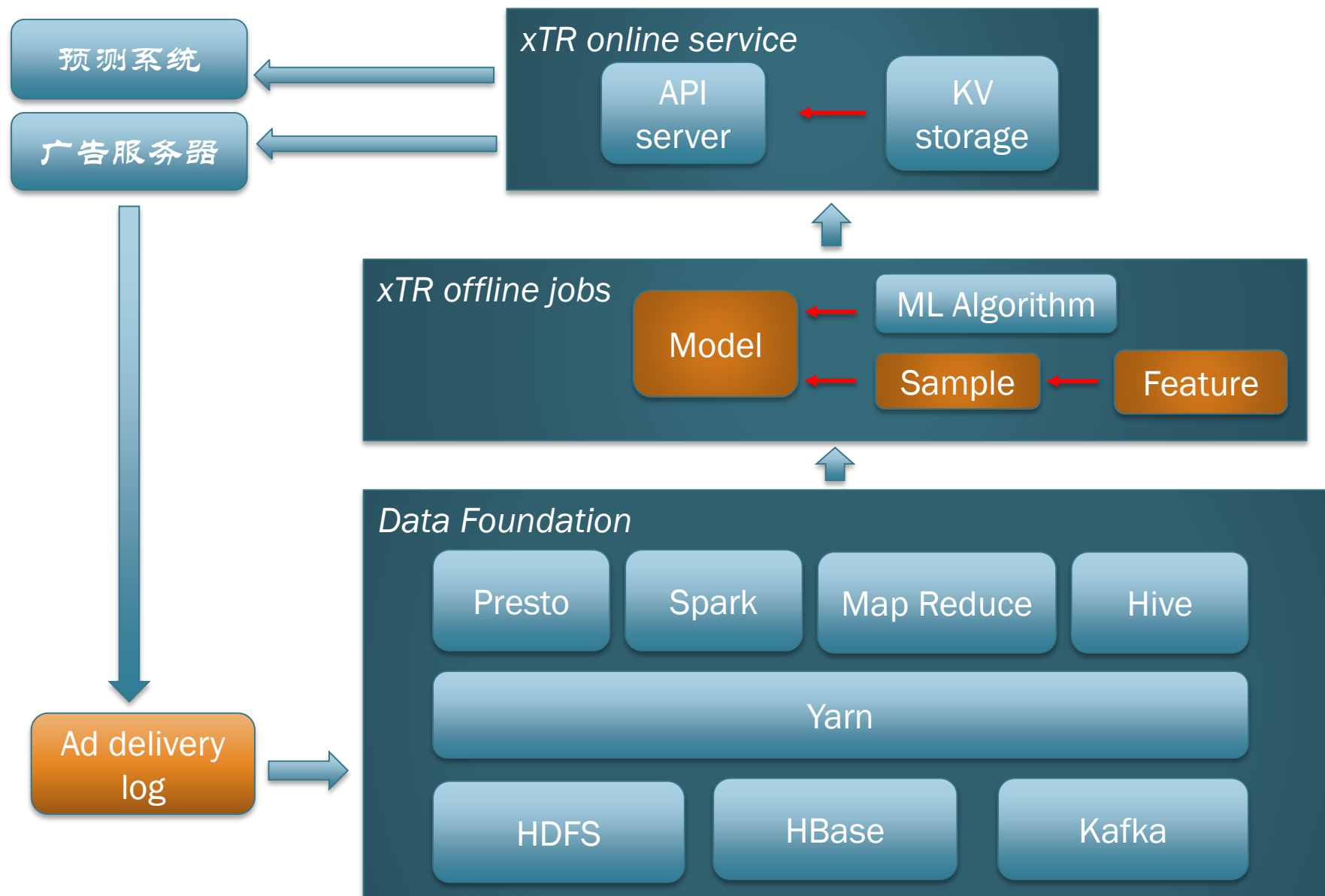
xTR是什么



CPx-event流程图



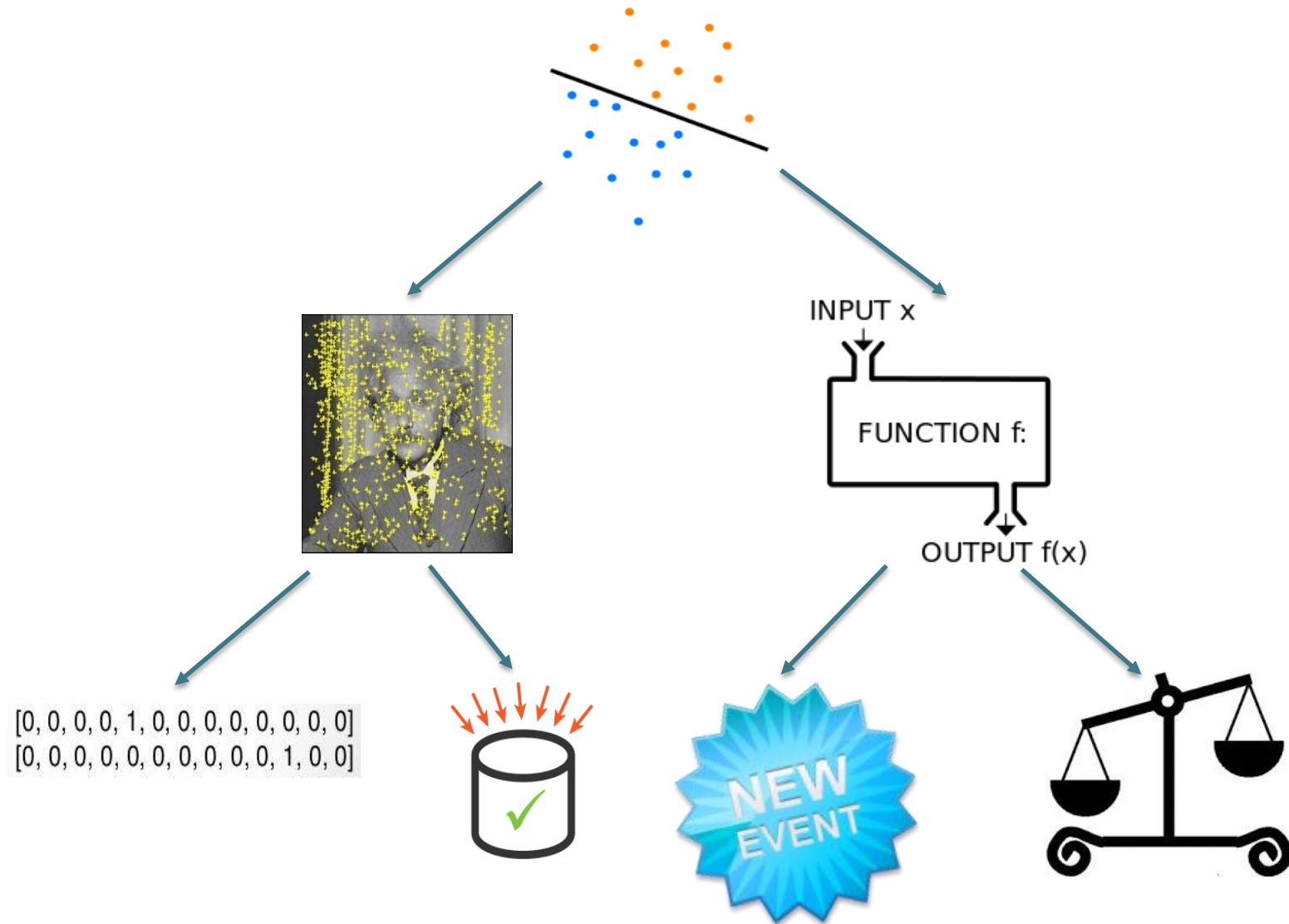
xTR系统架构



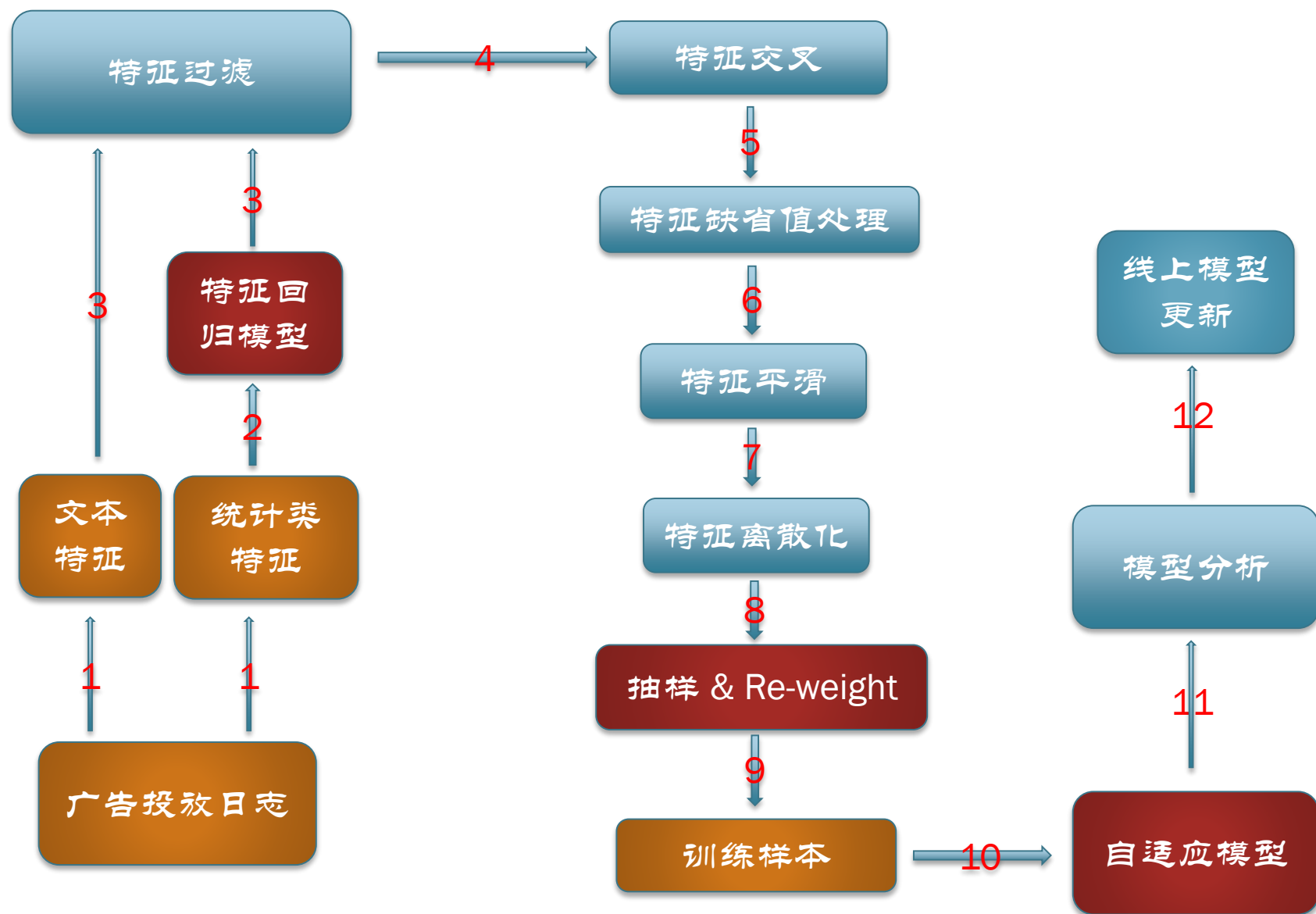
xTR数据规模

- 交易量: 1.5 billion (1 day)
- 特征量: 1.5 billion (30 days)
- 样本量: 250 million (7 days)
- 模型量: 50+ events
- 在线服务(Thrift RPC)
 - 100w QPS
 - 10ms timeout

xTR问题



xTR流程图



提取哪些信息作为特征？

- 上下文统计类特征/视频文本分类特征
- 上下文： 视频/网页/GEO(国家,洲,城市)/视频运营商/...

[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]

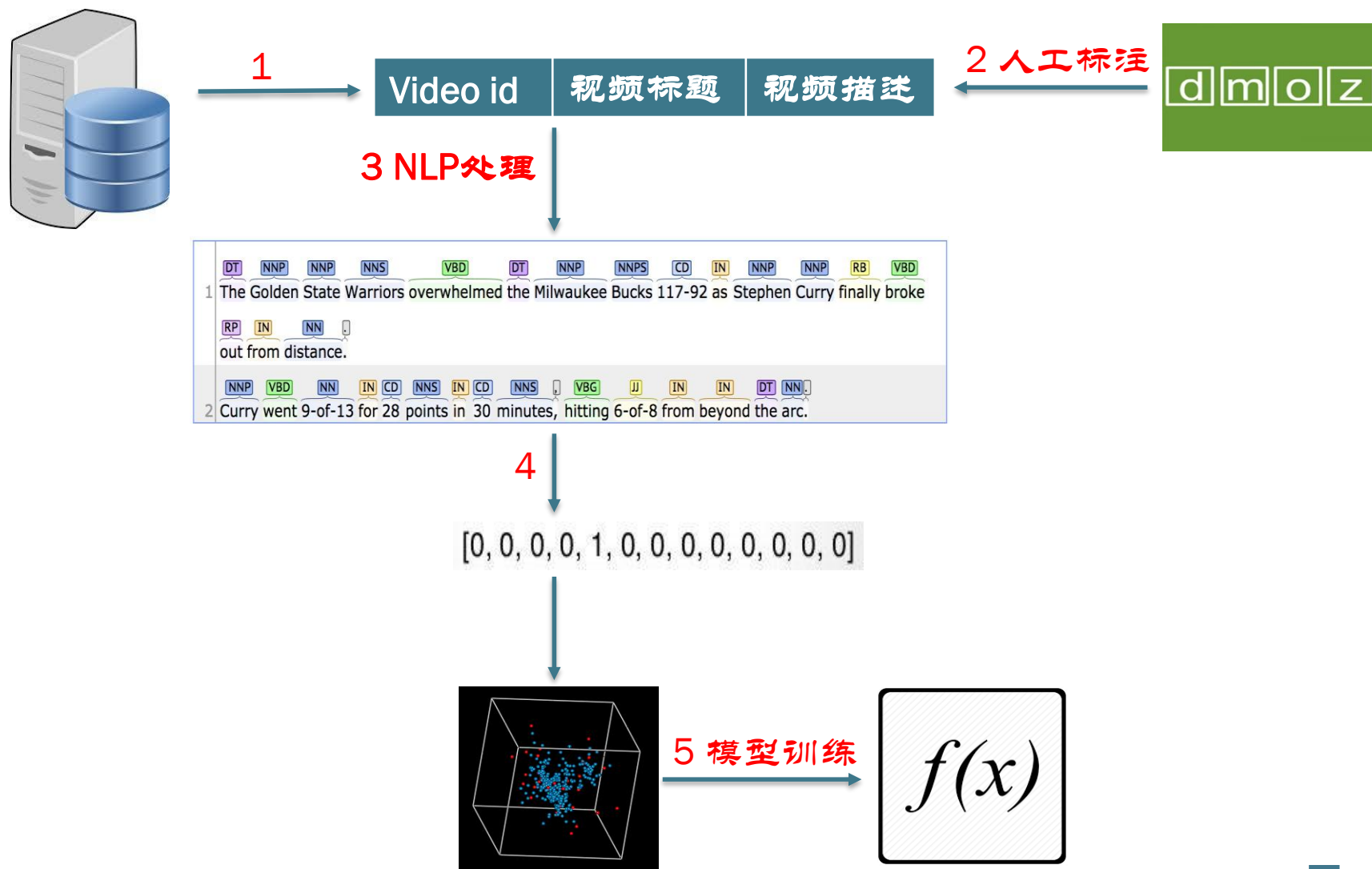
事件

纸牌屋第一季第一集 + 观看率 0.53

特征抽取方案	存储消耗
静态特征	n
动态特征	m (m为事件个数) << n

提取哪些信息作为特征？

- 上下文统计类特征/视频文本分类特征



特征组织方式

- **问题：所有特征存储在KV，难以满足高并发在线服务需求**
- **特征回归模型**
 - 学习目标：特征值
 - 特征：上下文信息one-hot representation，定义为“子特征”
 - Factorization machine回归模型

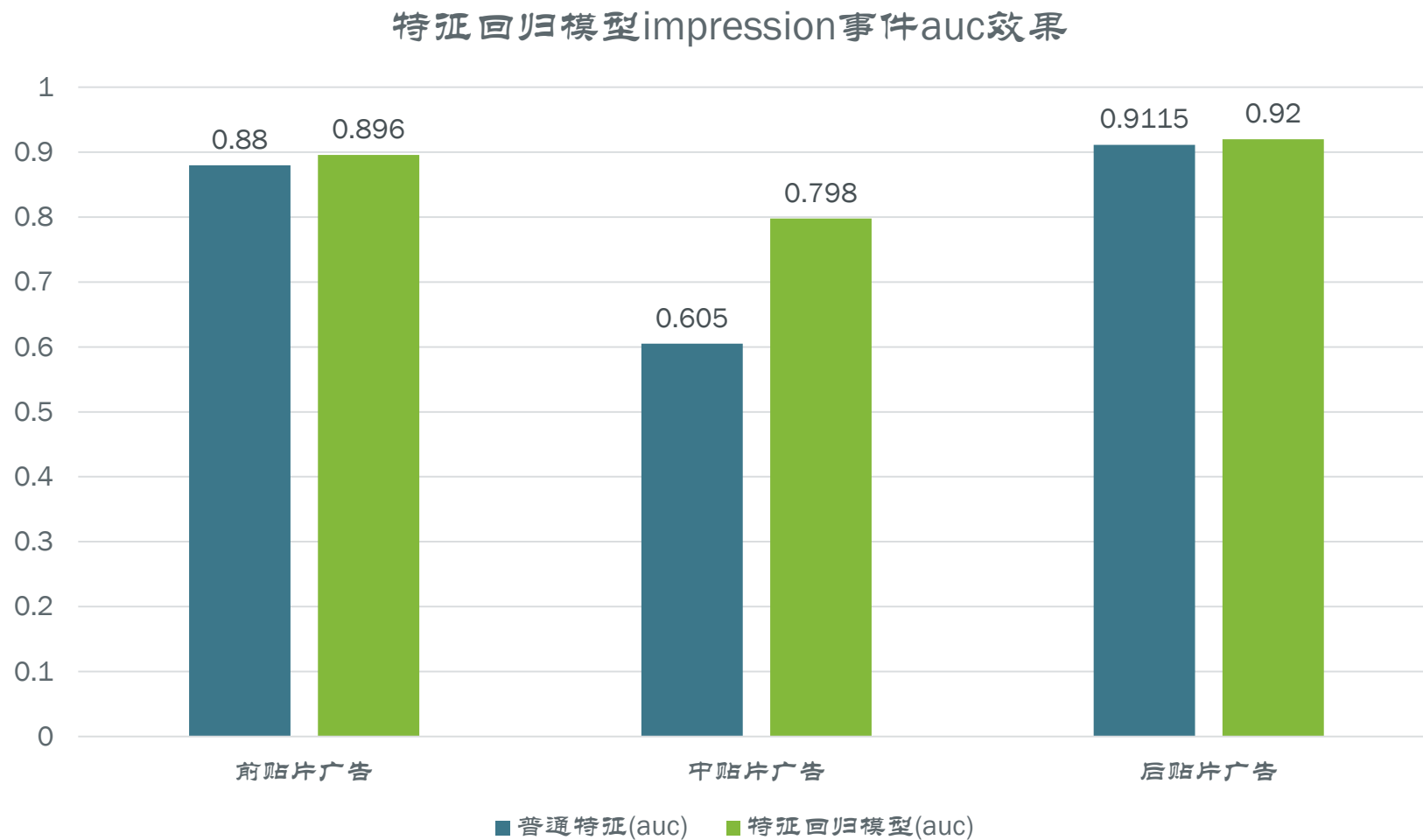
One hot vector : n

$$\bar{x} = (x_1, \dots, x_n)$$

$$\bar{V} = (v_1, \dots, v_k)$$

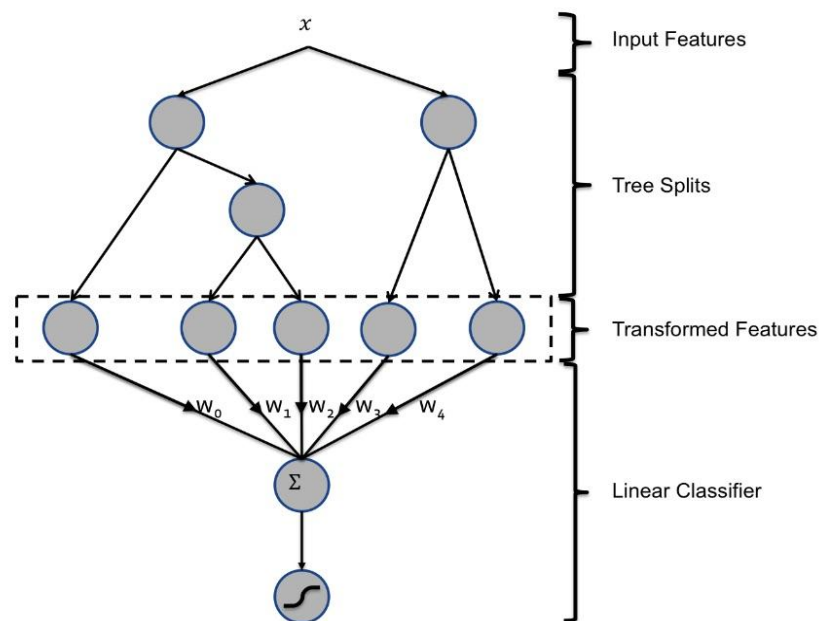
$$f(\bar{x}) = w_0 + \sum_{i=1}^n w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n (V_i^T V_j) x_i x_j$$

特征回归模型效果



特征筛选

- 选择最好的特征子集合
- Reference: 《Practical Lessons from Predicting Clicks on Ads at Facebook》



场景	基础特征(auc)	基础特征 + GBDT binary(auc)	提升效果
前贴片广告	0.88	0.9044	2.77%
中贴片广告	0.605	0.8432	39.37%
后贴片广告	0.9115	0.9227	1.23%

特征筛选

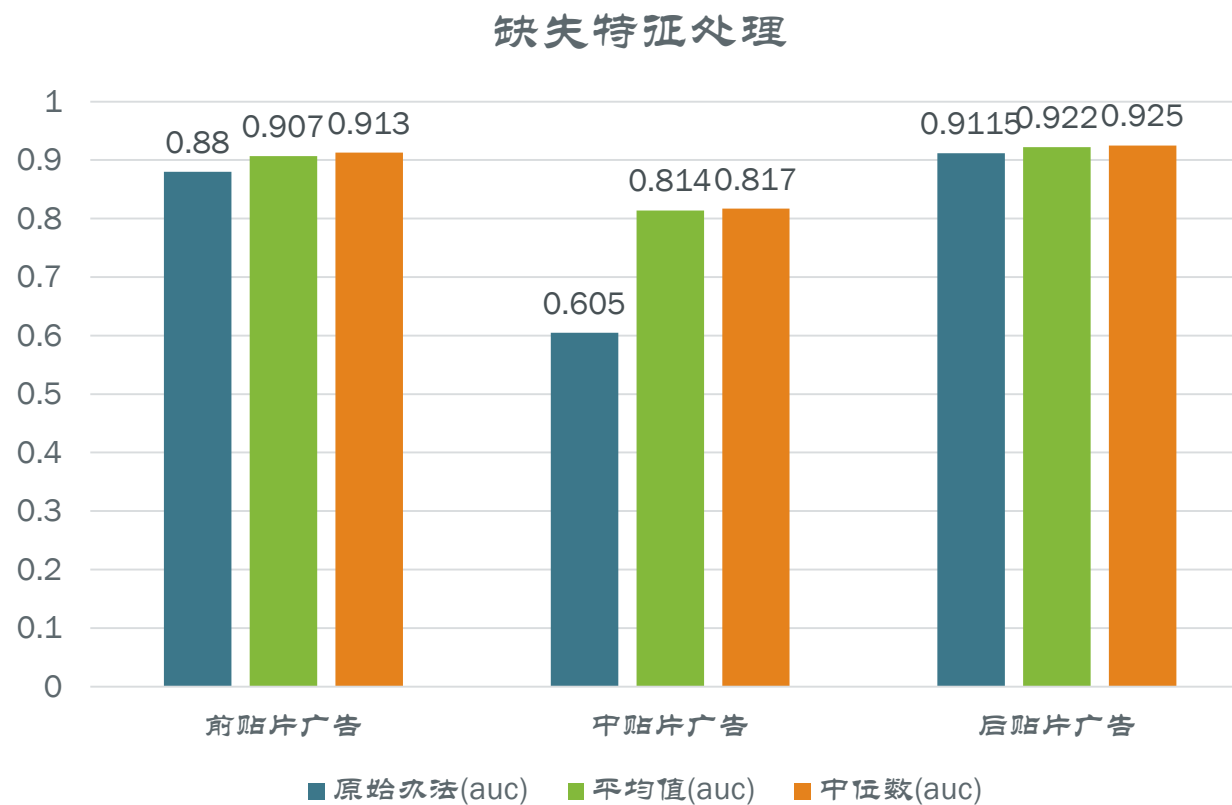
- 交叉特征：找出不同特征之间相关性

$$f(x) = \frac{1}{1 + e^{-\left(\sum_{i=1}^n w_i x_i + b\right)}}$$

$$f_{\text{crossfeature}}(x) = \frac{1}{1 + e^{-\left(\sum_{i=1}^n w_i x_i + \sum_{i=1}^n \sum_{j=i+1}^n w_{ij} x_i x_j + b\right)}}$$

怎样做特征工程

- 缺失特征处理：中位数/平均数



怎样做特征工程

- 特征平滑：Min-Max平滑/Gaussian平滑

Min – max normalization :

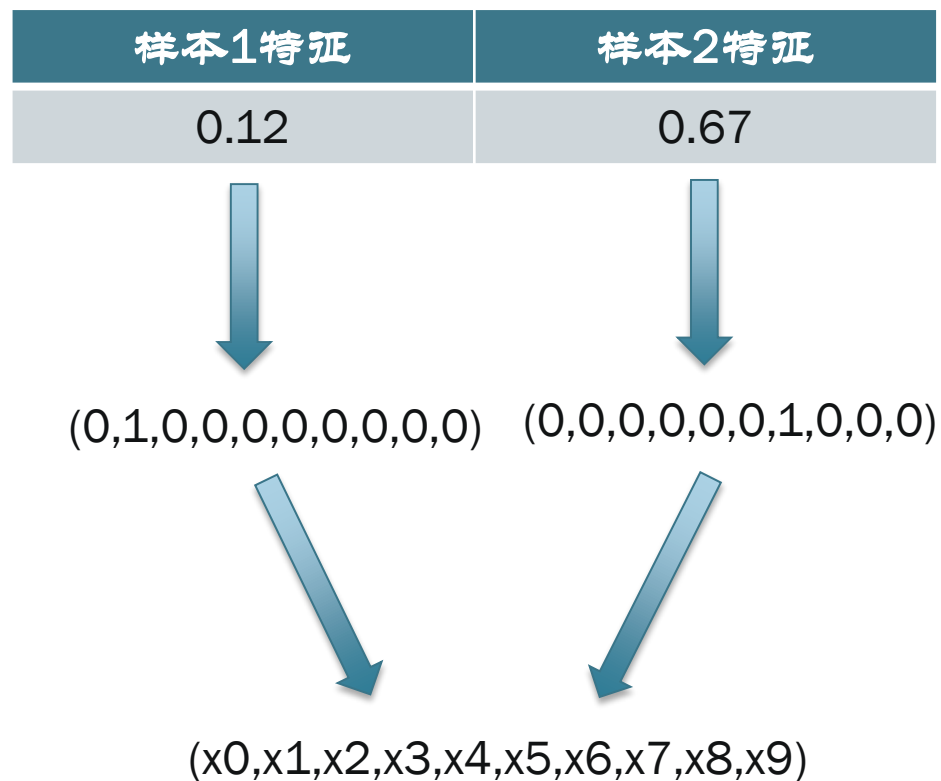
$$x_i(j) = \frac{x_i(j) - x_i^{\min}}{x_i^{\max} - x_i^{\min}}$$

Gaussian normalization :

$$x_i(j) = \frac{x_i(j) - \frac{1}{n} \sum_{k=1}^n x_i(k)}{\frac{1}{n} \sqrt{\sum_{k=1}^n (x_i(k) - \frac{1}{n} \sum_{k=1}^n x_i(k))^2}}$$

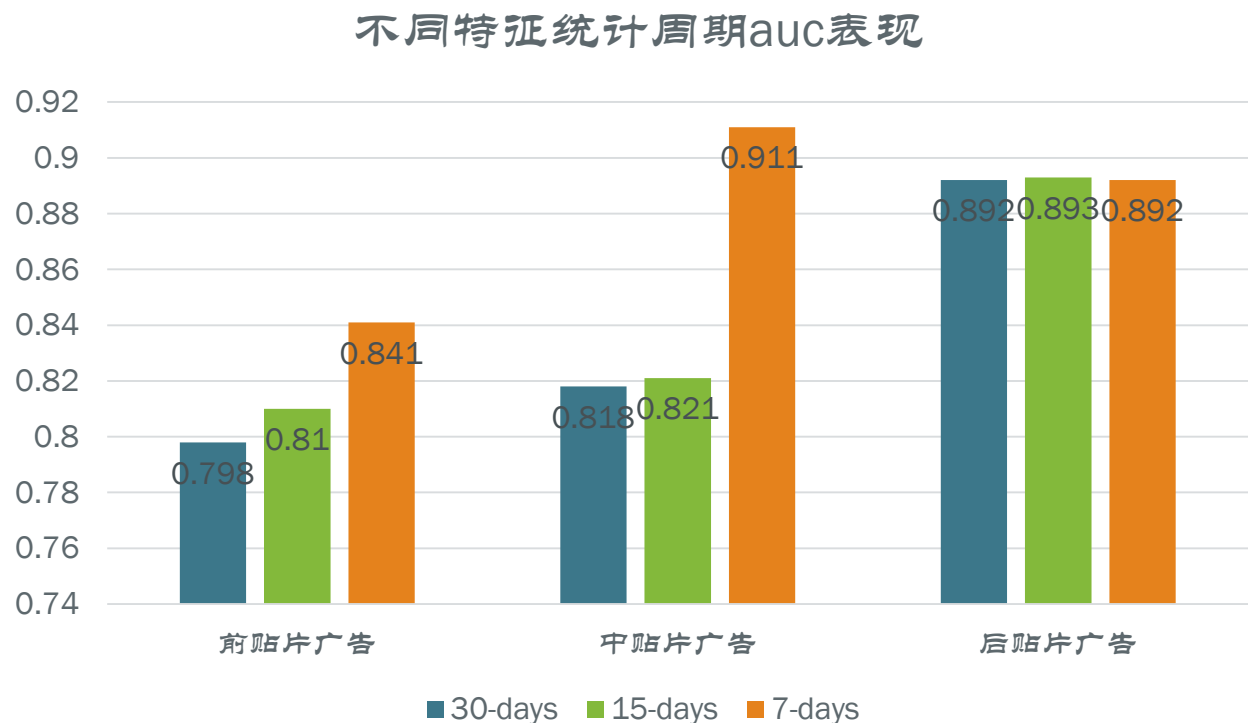
怎样做特征工程

- 连续特征离散化



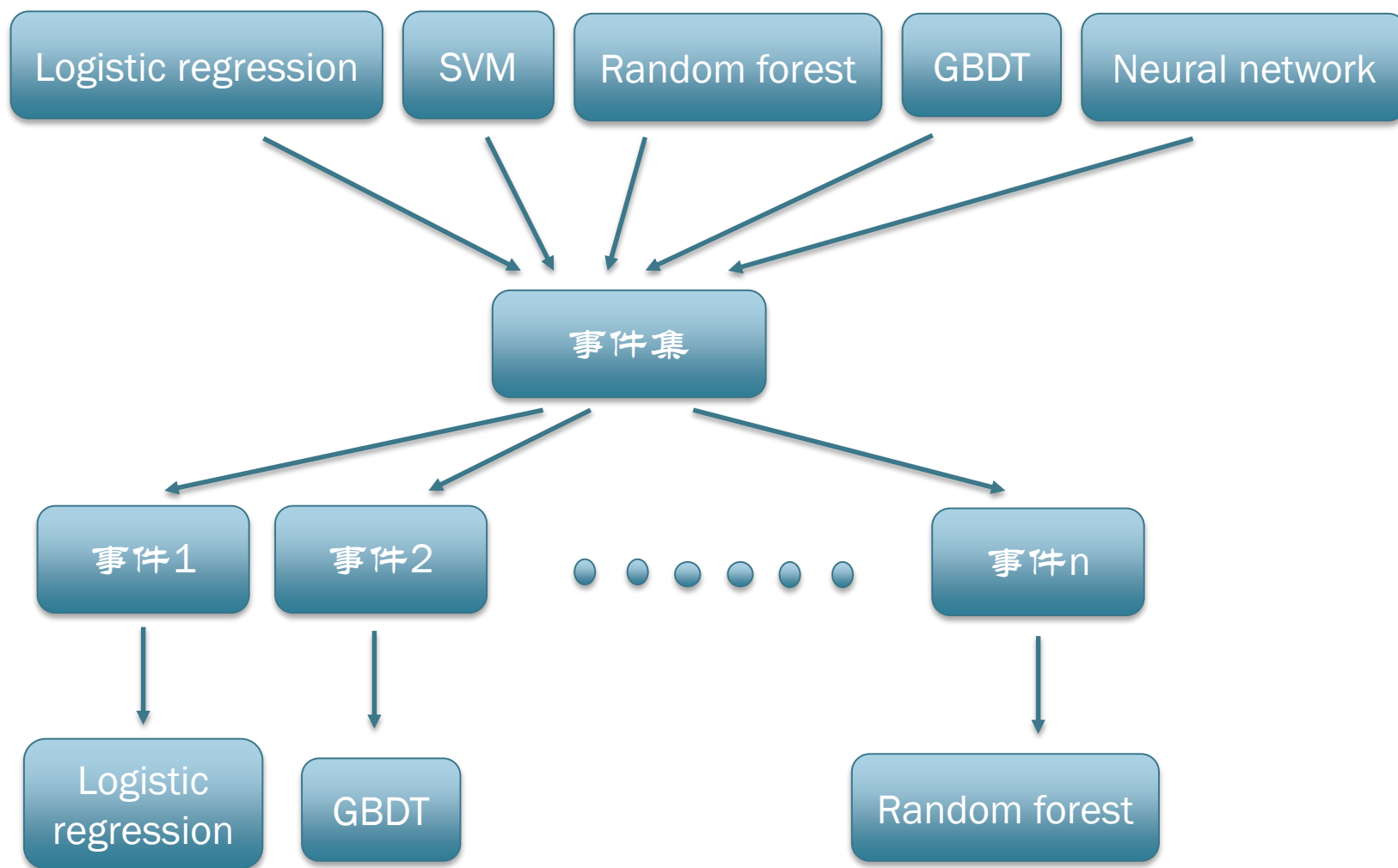
特征更新策略

- 加长统计周期，训练效果不一定变好



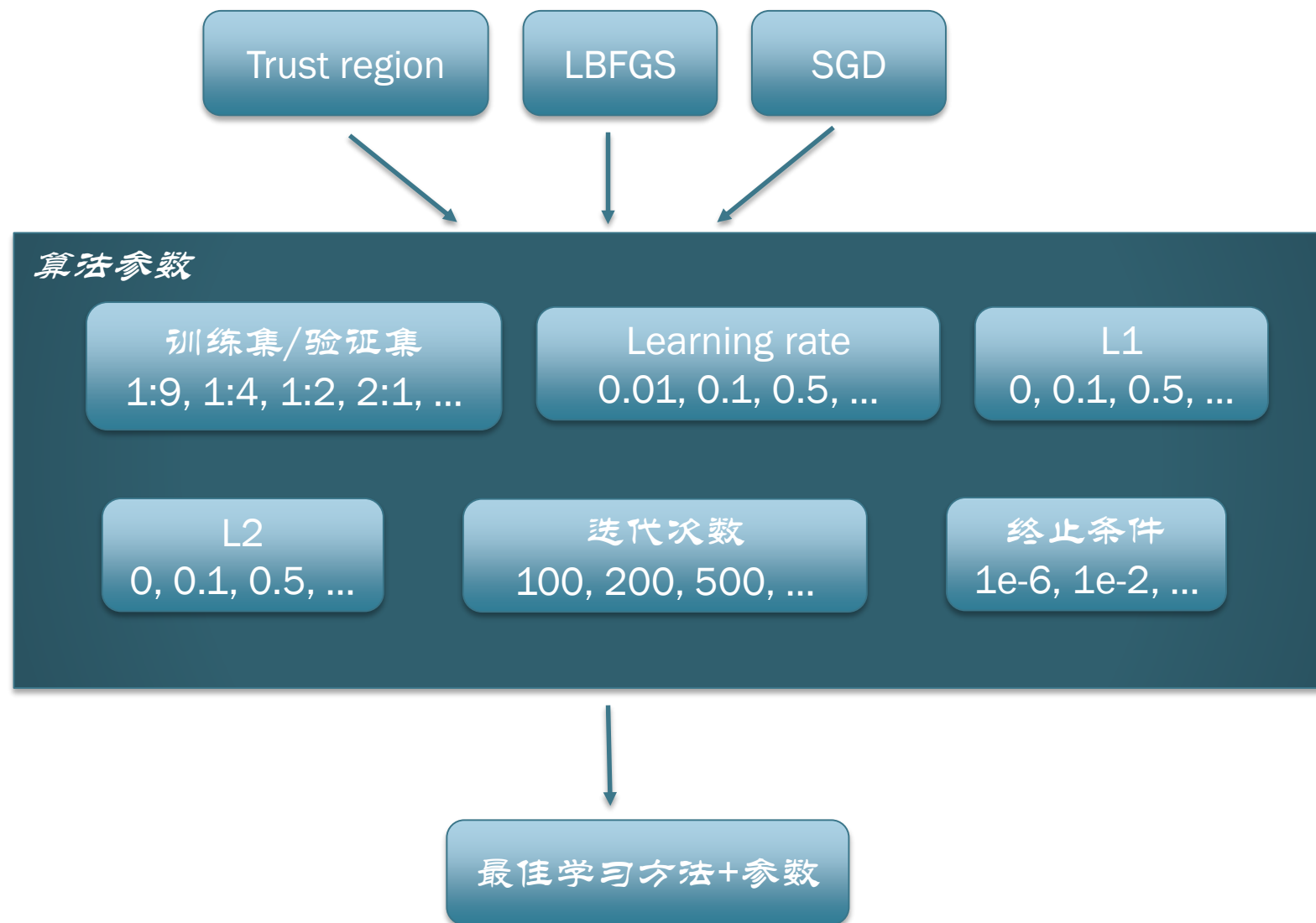
$$\frac{\alpha \times impression_{pre} + impression_{now}}{\alpha \times delivery_{pre} + delivery_{now}} \quad \alpha \in (0,1)$$

选择哪个模型训练



最优化模型参数

- 同一模型选择最佳学习方法与参数：枚举



自动化模型选择

- 在线学习 + 强化学习



在线强化学习模型效果

n model space

$$\gamma_{avg} = \frac{1}{t} \sum_{i=1}^t \gamma_i$$

$$\epsilon = \frac{e^{\gamma_{avg}}}{\sqrt{t}}$$

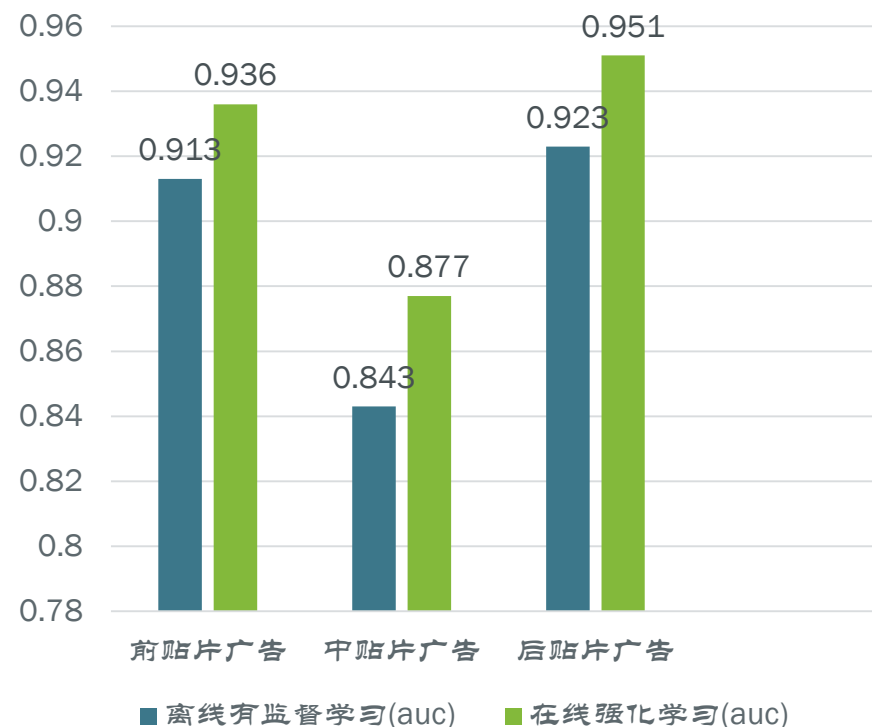
if $\epsilon \geq \mu_{threshold}$ then exploration
else exploitation

Exploration(soft max):

$$p(m_t^i) = \frac{e^{\frac{-\alpha \times \gamma_{avg}^i}{\lambda}}}{\sum_{j=1}^n e^{\frac{-\alpha \times \gamma_{avg}^j}{\lambda}}}$$

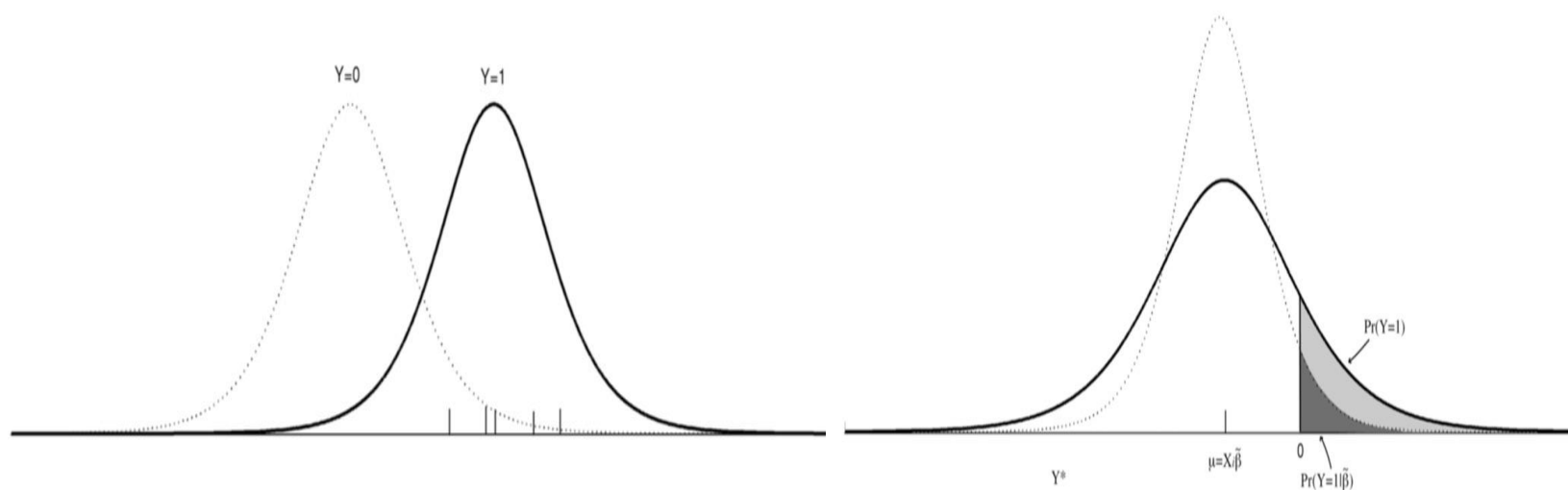
Policy: FTRL update

在线强化学习模型效果



解决不同事件中正负样本比例不均衡问题

场景	正负样本比例
前贴片广告	1:2
后贴片广告	1:40



解决不同事件中正负样本比例不均衡问题

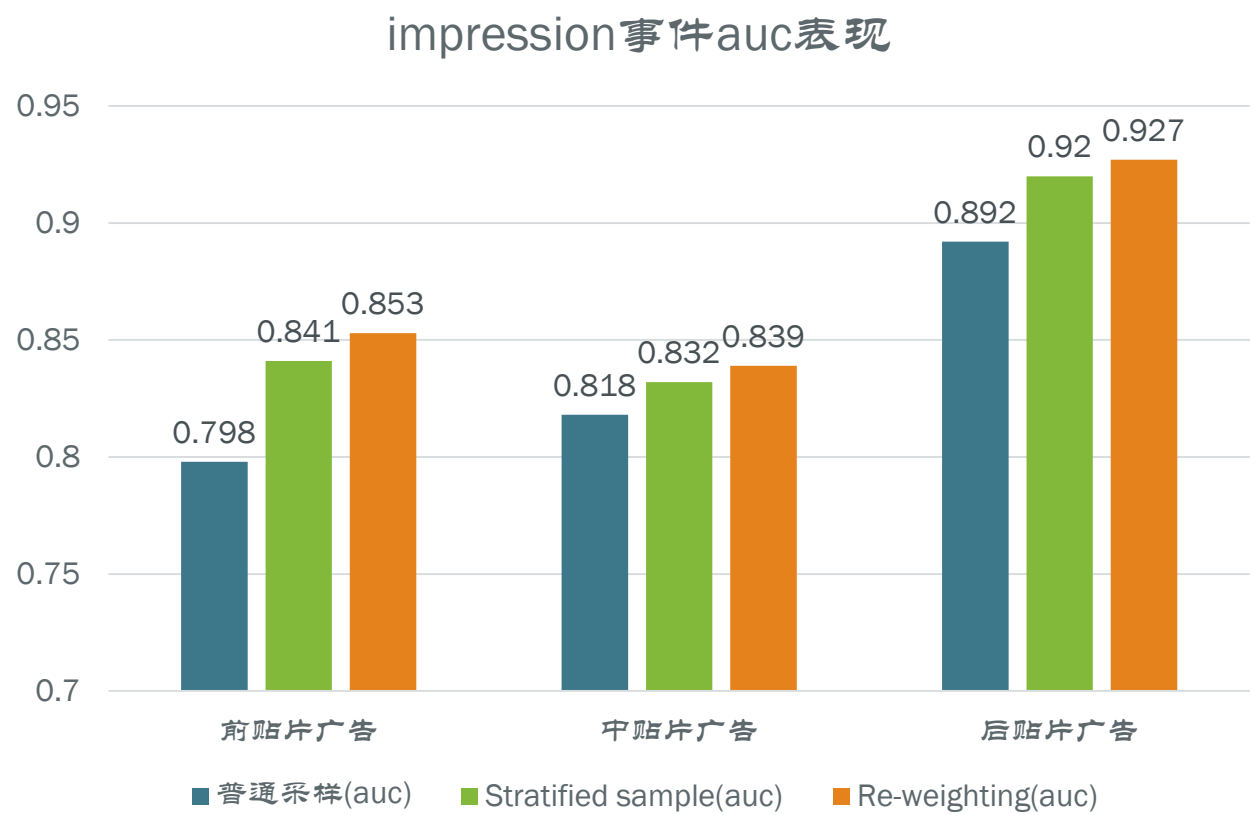
- 解决正负样本不均衡问题

- 小样本量数据全部采样，大样本量数据抽样采样
- Stratified sampling：根据特征来采样，使样本尽量覆盖特征空间
- Re-weighting：小样本被分到大样本时，在损失函数中的惩罚更大

$$\alpha > 1$$

$$L_i = \begin{cases} \frac{\alpha}{2}(y_i - p_i)^2, & y_i = 1 \\ \frac{1}{2}(y_i - p_i)^2, & y_i = 0 \end{cases}$$

解决正负样本不均衡方法效果



未来计划

- 用户属性特征
- 深度学习

Freewheel



- 潘晓彤
- Lead Researcher
- xtpan@freewheel.tv

Q & A

Thanks!