

# 特卖场景下的大数据平台和机器学习实践

周黄玲

huangling.zhou@husor.com.cn

# About me

- 2009 北京邮电大学
- 2012 搜狗
- 2014 天猫
- now 贝贝网

# 母婴特卖特点

商品  
周期短

需求  
变化快

移动化  
>80%

# 大纲

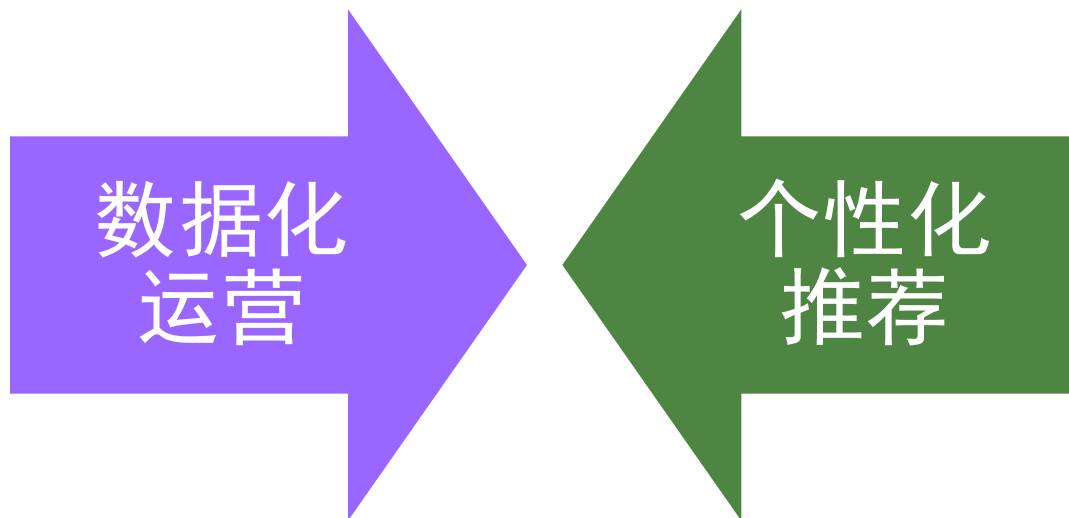
## 大数据平台

- 定位与架构
- 数据流程

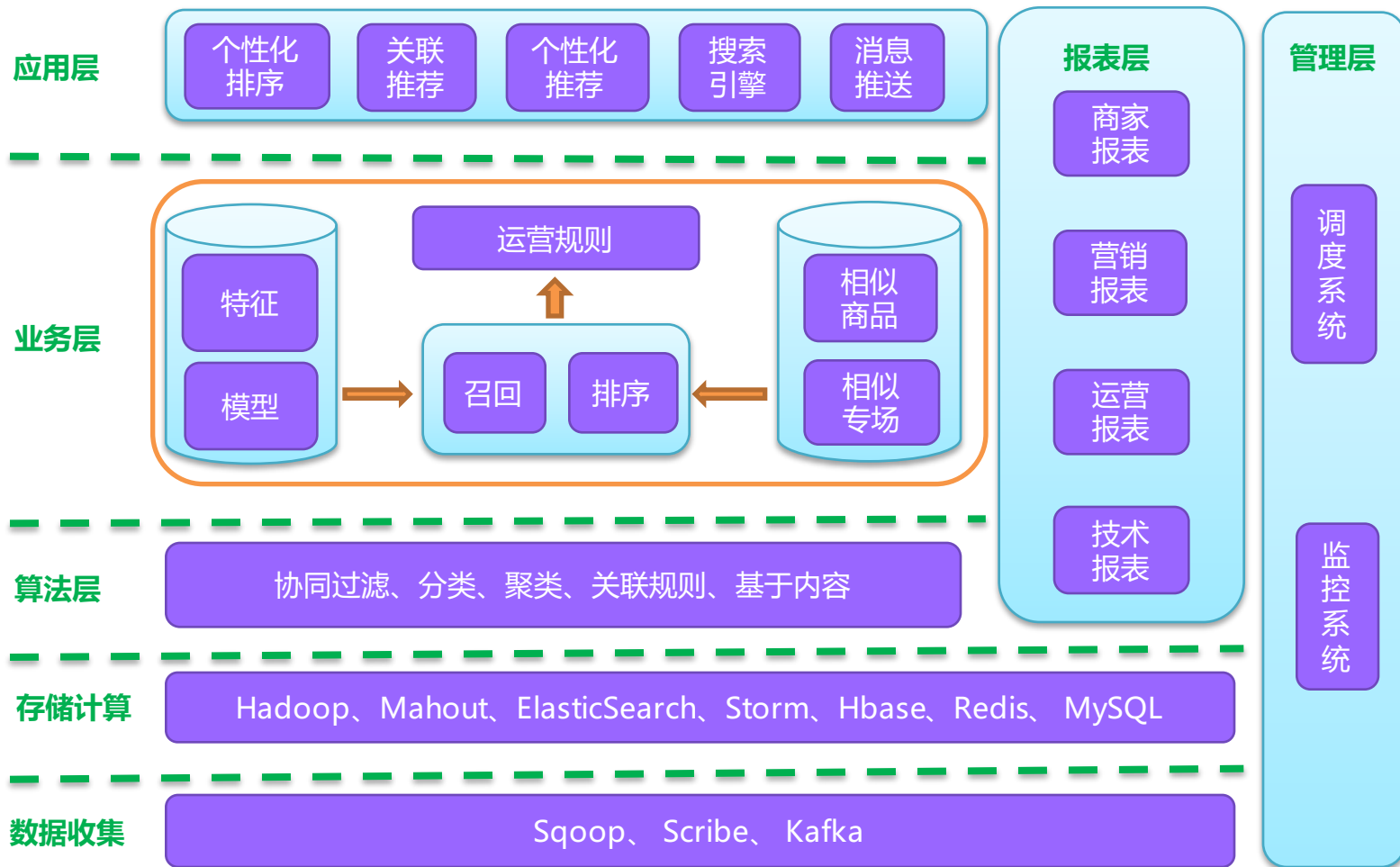
## 机器学习实践

- 推荐产品
- 技术方案

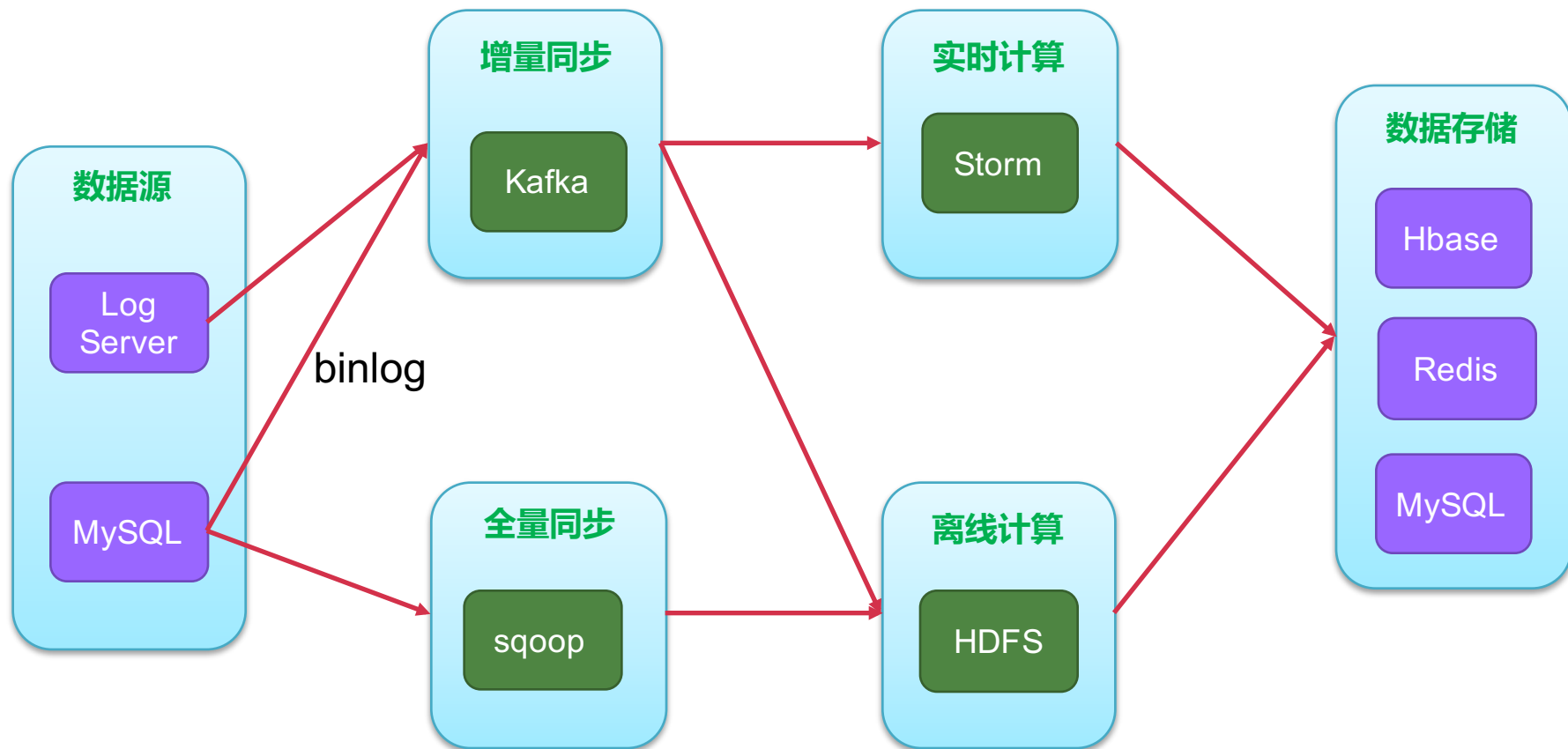
# 大数据平台定位



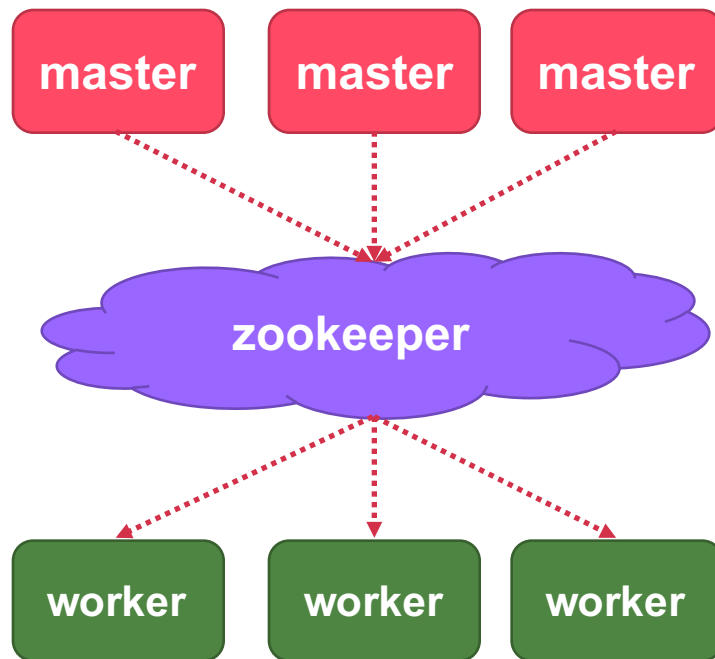
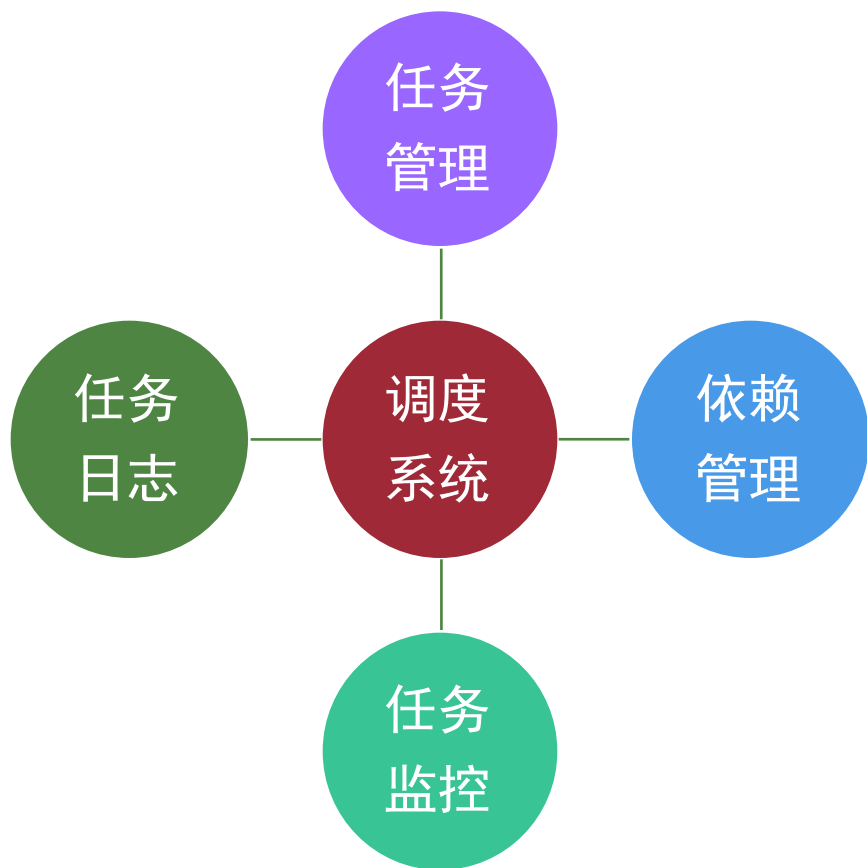
# 大数据平台架构



# 数据处理流程



# 分布式调度系统





# 大纲

## 大数据平台

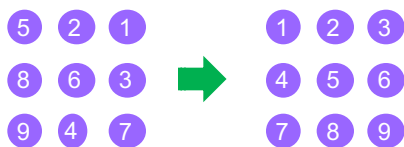
- 定位与架构
- 数据流程

## 机器学习实践

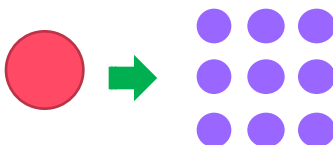
- 推荐产品
- 技术方案

# 推荐产品

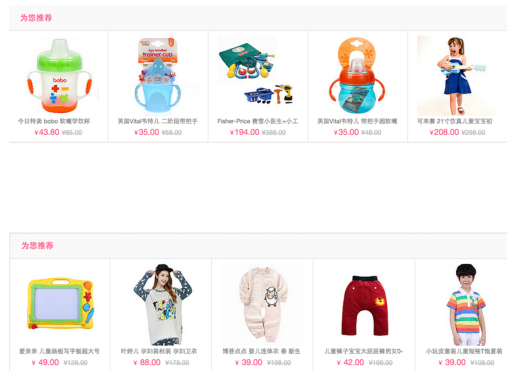
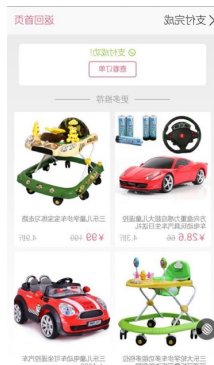
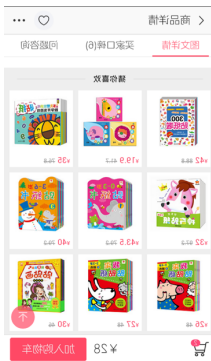
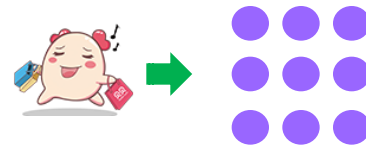
## 个性化排序



## 关联推荐



## 个性化推荐



# 全路径覆盖

## 流量导入

- 个性化短信
- 个性化APP推送

## 浏览

- 频道页：个性化专场列表
- 列表页：个性化商品列表
- 详情页：相似商品推荐

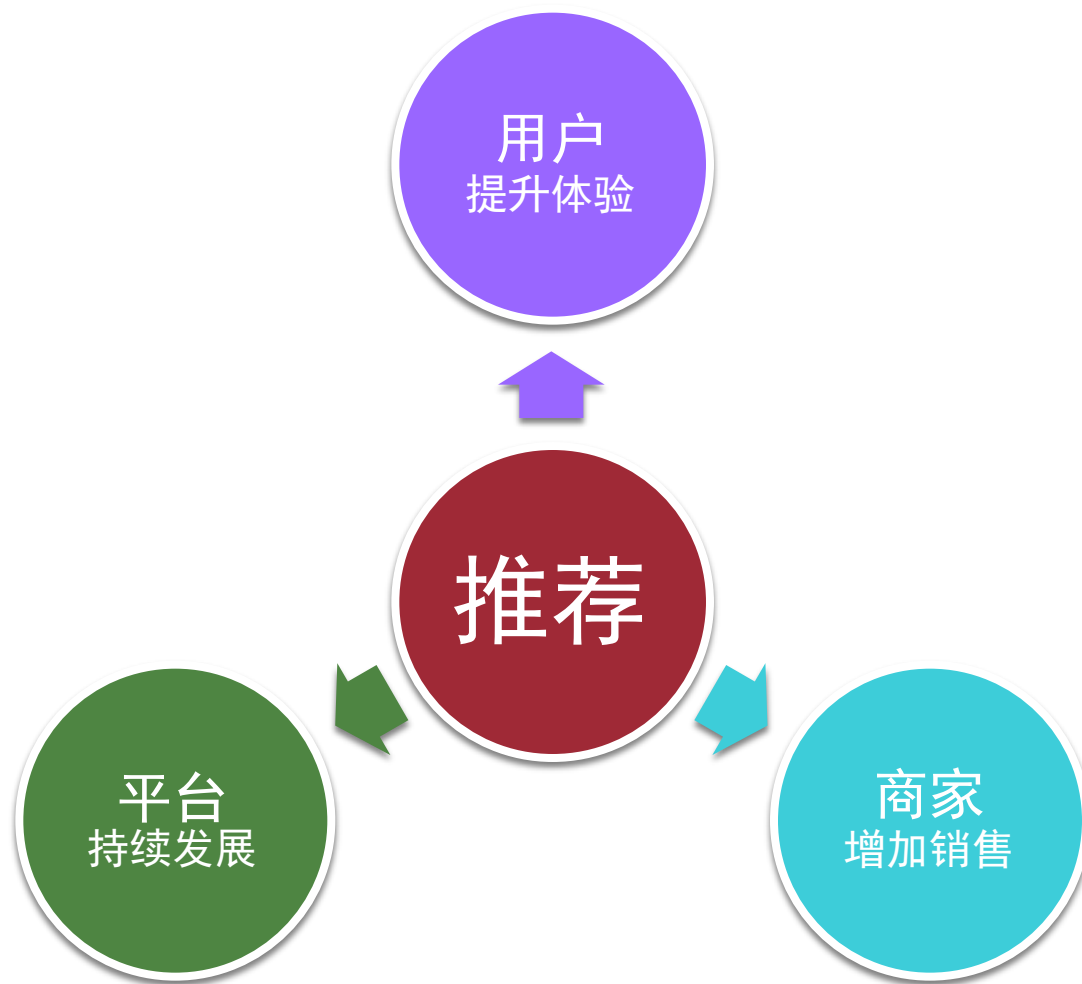
## 交易

- 购物车、订单：搭配商品推荐
- 风险控制

## 交易后

- 周期购买预测
- 客户流失预警

# 推荐的价值



# 用户画像



# 宝宝性别年龄预测

用户	特征：类目上的行为次数				宝宝性别	宝宝年龄
	浏览次数	搜索次数	收藏次数	购买次数		
u1	5	3	2	1	公主	0-1岁
u2	2	0	1	0	王子	3-6岁



机器学习模型

# 购物偏好

- 启发式

- 行为权重

- 浏览、点击、收藏、购物车、购买

- 行为次数

- 行为间隔

- 指数衰减

- 机器学习

- 特征

- 用户前一时间段内行为

- 目标

- 预估当前偏好程度

# 个性化专场排序

专场排序



规则

机器学习

无法千人千面

人工权重，无法  
科学定量

历史GMV，加  
剧马太效应

模型随着数据动  
态更新

千人千面

机器学习算法确  
定最优权重

多种维度特征综  
合考虑

首页到列表  
页转化率提  
升6%



# 个性化专场排序

离线

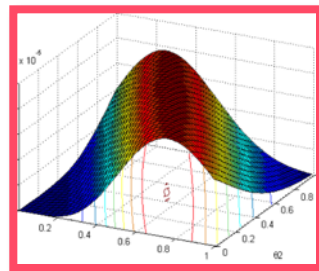
训练集

uid	bid	Y
101	211	0
101	212	1
.....		
102	211	0

特征提取

X	Y
[0,0.32,...,1.0]	0
[1,0.42,...,0.3]	1
.....	
[0.3,0.82,...,0]	0

模型训练



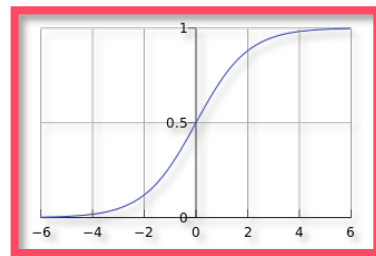
在线

uid	bid
101	211
101	212
.....	
102	211

排序请求

X
[0,0.32,...,1.0]
[1,0.42,...,0.3]
.....
[0.3,0.82,...,0]

特征提取



模型

uid	bid	P
101	211	0.17
101	212	0.22
.....		
102	211	0.13

预估结果

# 个性化专场排序

## 属性特征

- seller
- brand
- category
- 价格
- 折扣
- 上新率
- .....

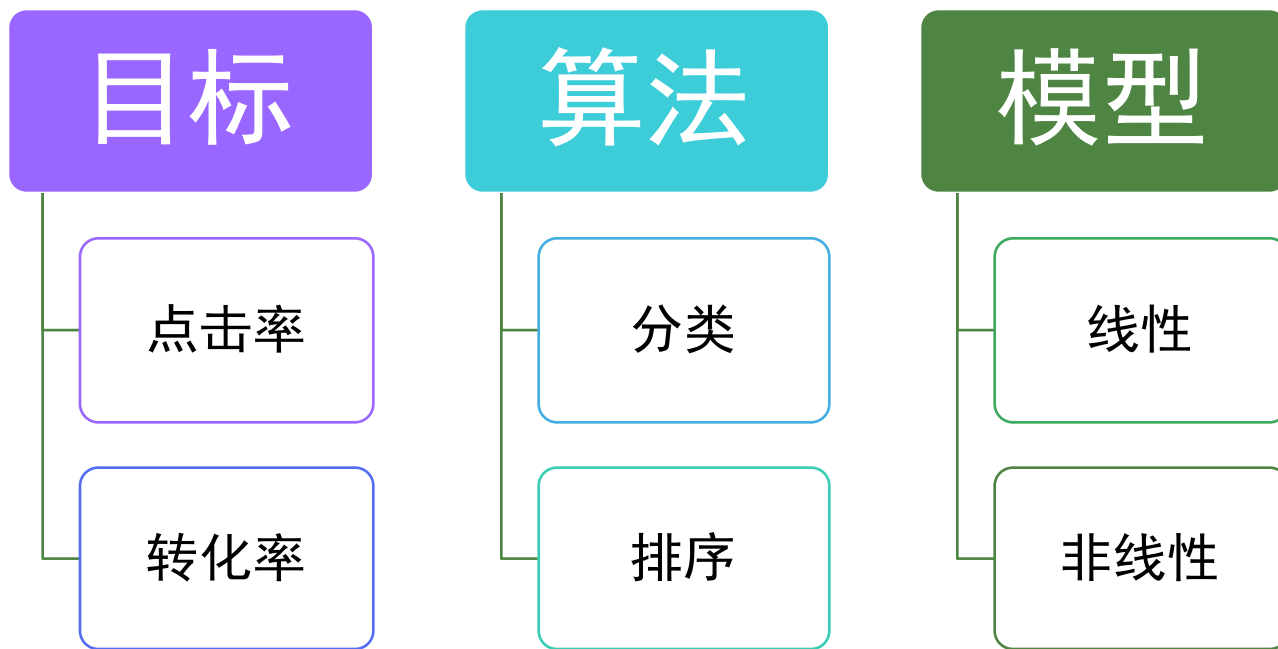
## 统计特征

- CTR
- CVR
- 销售额
- 客单价
- 退货率
- 评分
- .....

## 偏好特征

- 宝宝性别
- 宝宝年龄
- 价格段
- 类目
- 品牌
- 地域
- 终端
- .....

# 个性化专场排序



# 相似商品

特卖场景下的user-item矩阵

	week 1			week 2			week 3		
	i1	i2	i3	i4	i5	i6	i1	i4	i7
u1	1		1		1		1		1
u2		1	1		1	1	1		
u3	1			1		1		1	
u4	1		1		1		1		1
u5		1		1		1		1	

- 商品在线时间短
  - 相似商品不在线售卖
- 在线商品数量少
  - 数据比较不稀疏
- 领域知识
  - 宝宝性别、年龄
  - 时序性
    - 0-1岁->1-3岁 ✓ ✓ ✓
    - 1-3岁->0-1岁 × × ×

# 相似商品

## 协同过滤

- 时间衰减
- 热门打压

## 基于内容

- 属性相似
- 文本相似

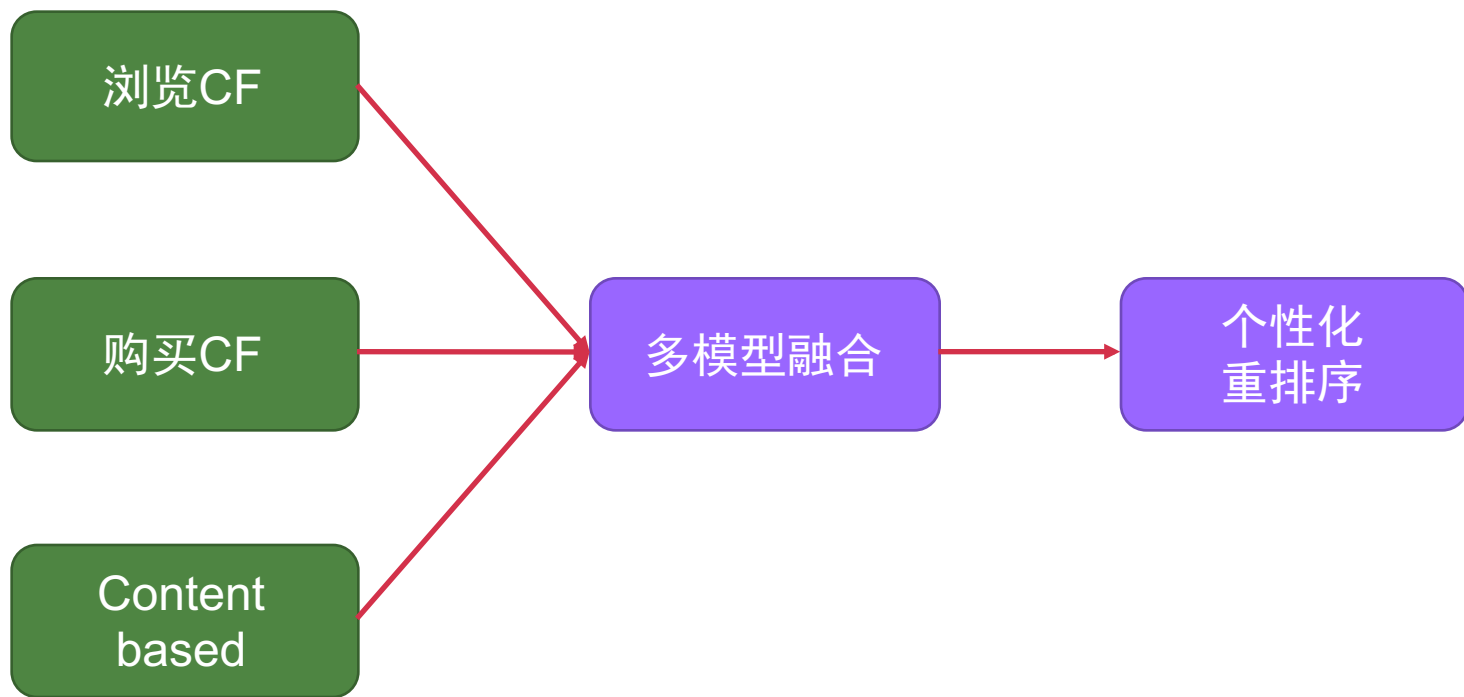
## 运营规则

- 宝宝性别
- 宝宝年龄
- 同品牌
- 跨品牌



妈妈的特卖会  
beibei.com

# 相似商品



# 关联推荐效果

资源位	提升
相似专场-品牌特卖	+96.83%
相似专场-海外购	+27.94%
相似商品-品牌特卖	+83.70%
相似商品-海外购	+541.4%
相似商品-限量购	+32.30%
猜你喜欢 ( PC ) -品牌特卖	+98.98%
猜你喜欢 ( APP ) -品牌特卖	+68.23%
购物车商品推荐	+52.61%

# 类目搭配

多项式分布，极大似然估计

$$P(c2|c1) = \frac{\#buy(c1, c2)}{\sum_i \#buy(c1, ci)}$$

买了	又买	比例
奶嘴	奶瓶	39%
奶粉	湿巾	25%
纸尿裤	湿巾	30%
孕妇帽	纸尿裤	37%
文胸	内裤	20%
烫衣板	毛球修剪器	40%



# 个性化推送

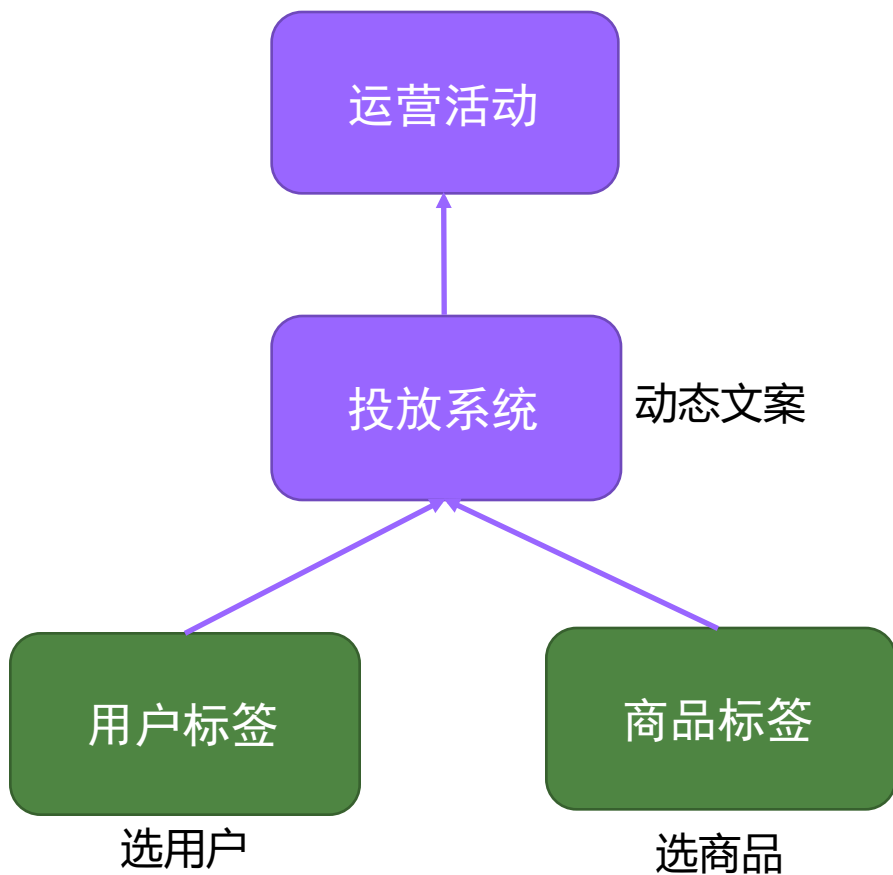
用户痛点：大量不感兴趣短信、推送消息

运营痛点：营销活动，如何找到目标用户

精准化营销，提升转化率

用户：	<input type="checkbox"/> 男用户 <input type="checkbox"/> 女用户 <input type="checkbox"/> 未知
宝宝性别：	<input type="checkbox"/> 男宝宝 <input type="checkbox"/> 女宝宝 <input type="checkbox"/> 未知 <input type="checkbox"/> 孕妈
宝宝年龄 / 月：	<input type="text"/> ---- <input type="text"/>
认证情况：	<input type="checkbox"/> 手机验证的用户 <input type="checkbox"/> 邮箱验证的用户 <input type="checkbox"/> 什么都没认证的用户
会员：	<input type="checkbox"/> 普通会员 <input type="checkbox"/> 铜牌会员 <input type="checkbox"/> 银牌会员 <input type="checkbox"/> 金牌会员 <input type="checkbox"/> 铂金会员 <input type="checkbox"/> 钻石会员
商品频道标签：	<input type="checkbox"/> 专场 <input type="checkbox"/> 限量购 <input type="checkbox"/> 海外购
是否有购买行为：	<input type="checkbox"/> 是 <input type="checkbox"/> 否
品类偏好：	<input type="checkbox"/> 童装 <input type="checkbox"/> 童鞋 <input type="checkbox"/> 文体 <input type="checkbox"/> 玩具 <input type="checkbox"/> 婴儿装 <input type="checkbox"/> 孕妈专区 <input type="checkbox"/> 奶粉 <input type="checkbox"/> 辅食
品牌偏好：	<input type="checkbox"/> 国际名牌 <input type="checkbox"/> 国内名牌
价格段偏好：	<input type="checkbox"/> 0-50元 <input type="checkbox"/> 50-100元 <input type="checkbox"/> 100-200元 <input type="checkbox"/> 200-300元 <input type="checkbox"/> 300-500元 <input type="checkbox"/> 500-1000元 <input type="checkbox"/> 1000以上
访问终端偏好：	<input type="checkbox"/> PC <input type="checkbox"/> APP <input type="checkbox"/> WAP
时间偏好：	<input type="text"/> ---- <input type="text"/>
节日偏好：	<input type="checkbox"/> 双十一 <input type="checkbox"/> 双十二 <input type="checkbox"/> 儿童节 <input type="checkbox"/> 双旦
注册时间：	<input type="text"/> ---- <input type="text"/>
最近一次消费时间 / 天(Recency)：	<input type="text"/> ---- <input type="text"/>
消费频率(Frequency)：	<input type="checkbox"/> F=1 <input type="checkbox"/> F=2 <input type="checkbox"/> F=3 <input type="checkbox"/> F=4 <input type="checkbox"/> F=5 <input type="checkbox"/> F=6
累计消费金额(Monetary)：	<input type="checkbox"/> M<=50 <input type="checkbox"/> 50-M<=100 <input type="checkbox"/> 100-M<=200 <input type="checkbox"/> 200-M<=500 <input type="checkbox"/> 500-M<=1000 <input type="checkbox"/> 1000-M<=5000 <input type="checkbox"/> M>5000
发送时间：	<input type="text"/>
发送方式：	<input type="checkbox"/> 短信 <input type="checkbox"/> 推送

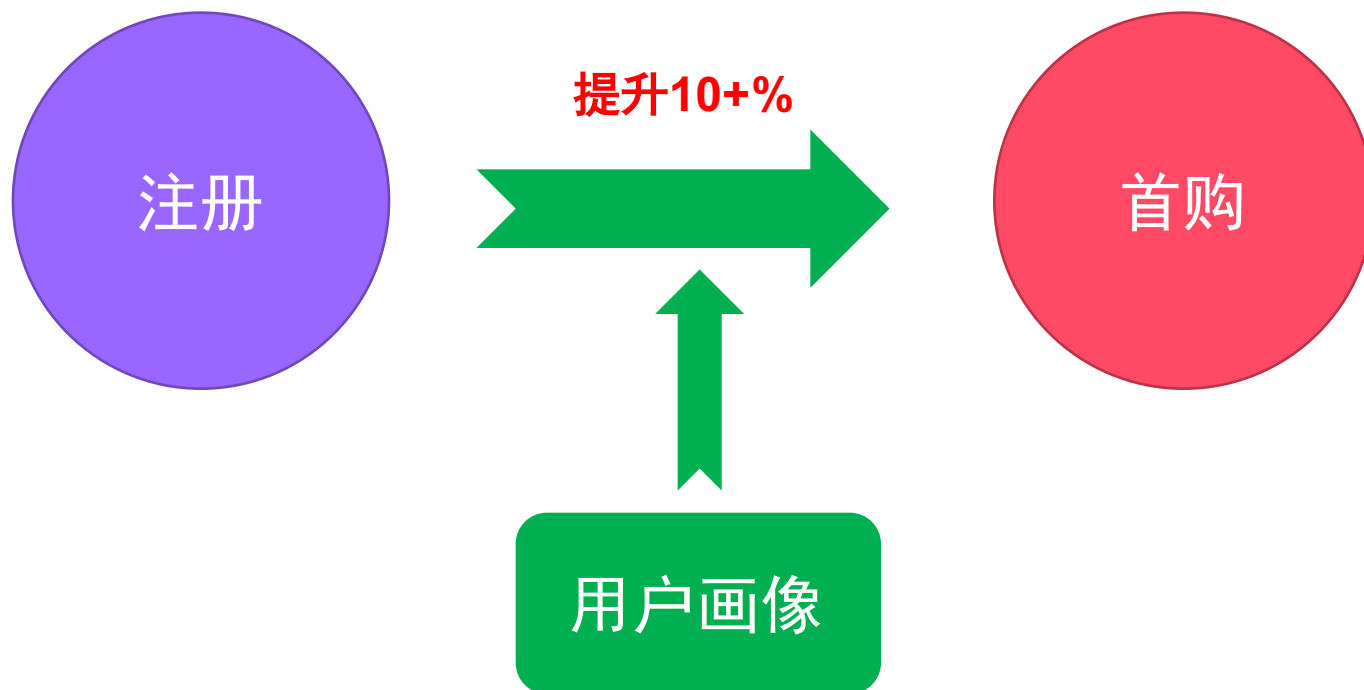
# 个性化推送



动态生成hive sql查询

JDBC提交到hive server执行

# 运营拉新



# 未来方向

- 数据
  - 精准用户画像
  - 实时偏好
- 模型
  - 特征工程
  - online learning
- 应用
  - 个性化大促
  - 商家端

# 贝贝



**beibei.com**



**母婴特卖**



**杭州**



**2014.4**



**C轮 10亿 \$**

# Thanks!



huangling.zhou@husor.com.cn