



IT大咖说  
知识共享平台

携程技术中心

# 携程技术沙龙

## 算法测试探秘

分享人：张海宁



姓名:张海宁



- 7年互联网老兵
- BAT转了一圈，最后回归阿里体系的怀抱
- 现在负责饿了么北京物流中心的质量工作



# 目录

CONTENTS

1

算法在外卖的应用

2

算法测试体系建设

3

算法实战案例解析



**场景：**非诚勿扰有个男嘉宾，互联网数据分析师，对近100期节目数据进行训练，想分析出节目的规律，用来预估女嘉宾的最终灭灯概率，当然结果非常准确，最后优雅的转身。其实这就是一个从数据挖掘再到算法应用的典型场景，可以映射为**灭灯概率预估模型**

- 大众体：AlphaGo、无人车、最强大脑、讯飞输入法、芯片
- 官方定义：算法=大数据+智能化
- 白话体：你看到的数据是不是你想要的（懂你）

# 目录

CONTENTS

1

算法在外卖的应用

2

算法测试体系建设

3

算法实战案例解析



## 用户A：今天中午吃什么

目标：脑中莜面，列表西贝，促进用户决策



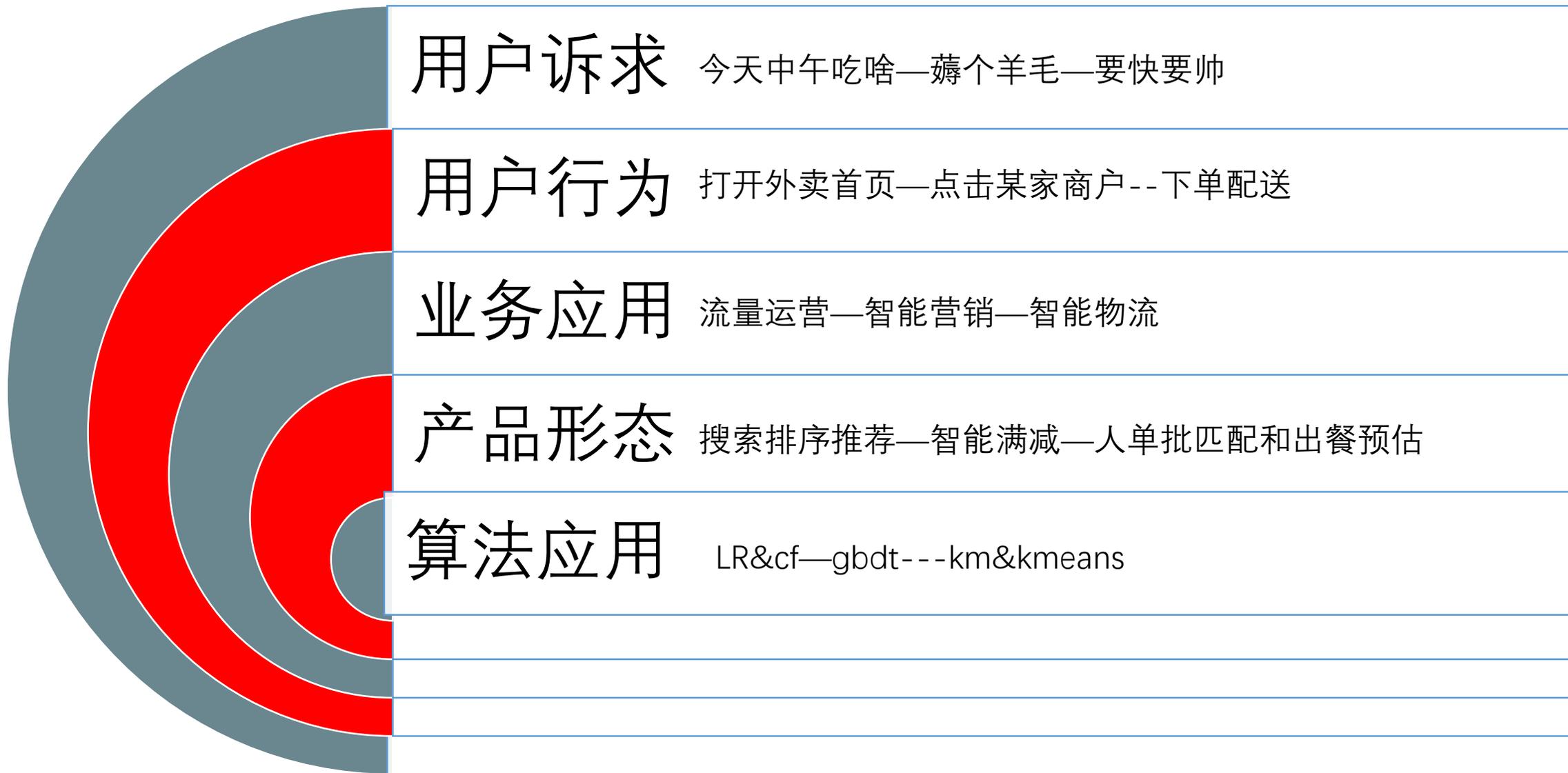
## 用户B：薅个羊毛

目标：优惠促进用户下单



## 用户C：要快、要帅

目标：智能派单、精准预估，提升用户体验



# 目录

CONTENTS

- 1 算法在外卖的应用
- 2 算法测试体系建设
- 3 算法实战案例解析

# 业务生态

平台  
服务

可视化和API赋能（波塞东、魔镜、达芬奇）

算法  
应用

流量运营、智能营销、智能物流

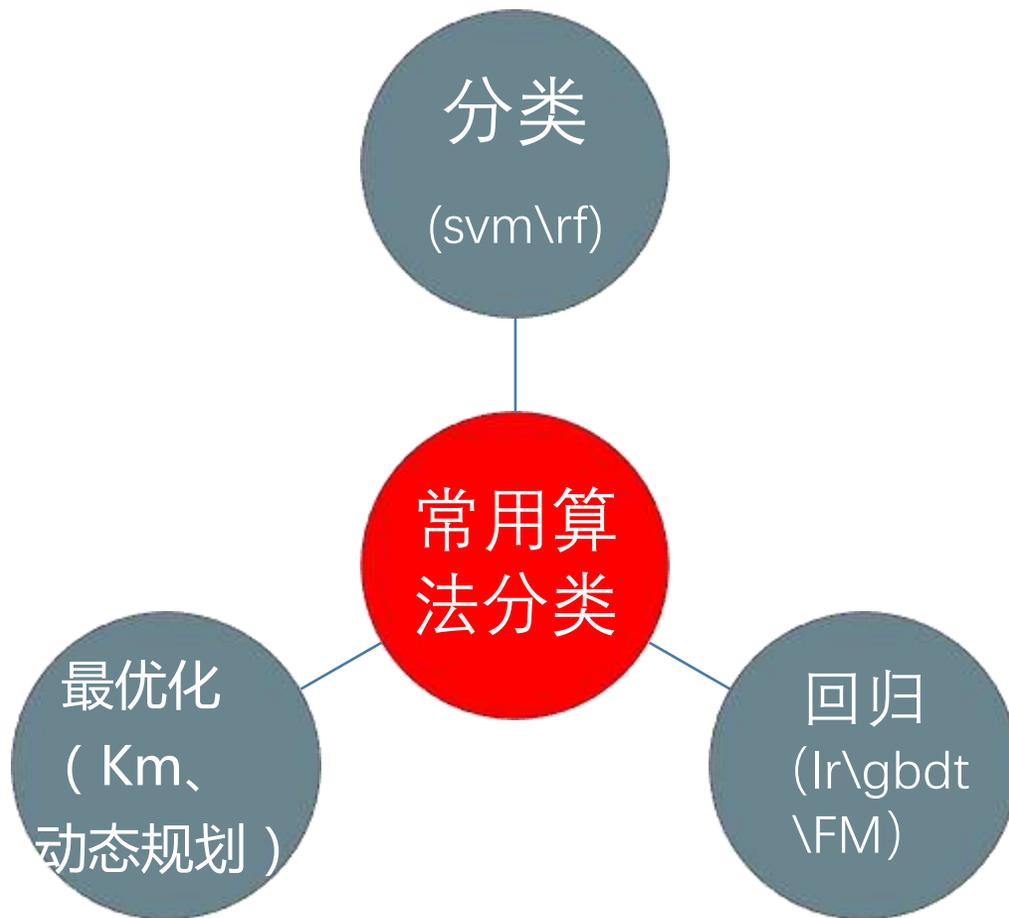
商业  
画像

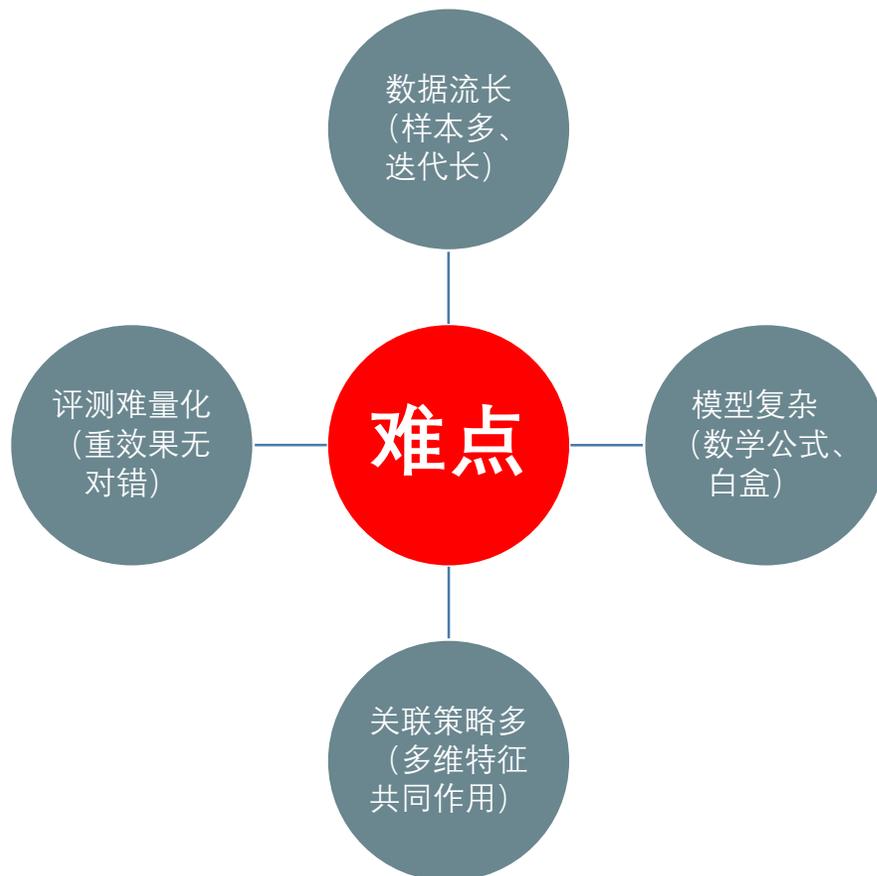
用户、商户、骑士画像，统一特征库

BDI

集群、生态产品、ETL







## 传统测试

- 上线前
- 功能
- 可用性
- 偏UI
- 用户使用

V  
S

## 算法测试

- 上线后
- 效果
- 数据集
- 偏后端
- 用户感知

从数据到流程再到深度—基于分层测试+测试前置+专业测试3大原则

### 数据层

- **数据量级** (大小、性能、范围)
- 存储模式 (redis、mysql、es、hadoop)
- 数据准确性(数据本身、异常测试)

### 模型层

- 特征校验 (抽取、重复、缺失、超范围)
- 训练集和测试集的选择 (7:3)
- **蜕变测试** (变化公式数据, diff结果)
- diff (汉明距离、simhash)

### 效果层

- 打分、命中策略
- **AB**(指标对比)
- 抽样 (对内、对外)

## 前置数据测试

- 性能：源数据拉取
- 监控：范围、量级、逻辑域
- 数据本身：空、**类型**、边界、异常、KV一致性、**顺序**、**重复**、归一化、样本均衡、浮点精度、**fuzz测试**

## 模型专业测试

- 分类指标：TPR（正确），FPR&TNR（误报或非匹配），**F值**、**AUC** or ROC
- 回归指标：**RMSE**（均方根误差）、MSE（方根误差）、MAE（平均绝对误差）、SCC（平方相关系数）、Rsquare（拟合系数）

## 上线后效果测试

- 业务指标：pv、uv、**ctr**、下单、流水预估、转化率等
- 策略指标：**打分**、**命中策略**
- AB：不同城市不同商圈逐步**灰度**
- 抽样：不同城市、不同人群

# 目录

CONTENTS

- 1 算法在外卖的应用
- 2 算法测试体系建设
- 3 算法实战案例解析

## 背景

Such as : 30-5 ; 50-8 ; 100-10 ;

## 流程



## 问题

1

一致性

Redis超时、前后端不一致、运营数据改了，生效地方乱了

2

跨天数据如何处理

到底算哪天的，昨天还是今天，如何处理

3

如何把感知量化

今明都是30-5，是未生效还是结果一样

- 多渠道拉取diff&监控  
(Redis与mysql一致性、策略结果、运营后台和C端一致性)

## 关联测试

- 数据拉取性能、数据本身(空、类型、档位关系)、跨天数据处理  
(约束+增量回溯)

## 数据测试

- 从范围、量级、趋势等阈值断言
- 感知量化(不同角色、时间段、假期、档位)

## 监控

多维测试

## 01 横向校验

- 流程---源数据获取、离线训练、调架构、服务拉取、展示
- 存储方式---依赖的存储格式、类型 ( redis、es、hadoop )
- 位置---B端、C端、后台、用户多角色监控

## 02 纵向校验

- 数据本身---类型、量级、一致性、重复、归一化、**shuffle**、特征抽取和组合
- 性能---数据拉取性能、运行时长、资源消耗
- 异常---**fuzz测试** ( 样本数据部分丢失或未更新 )

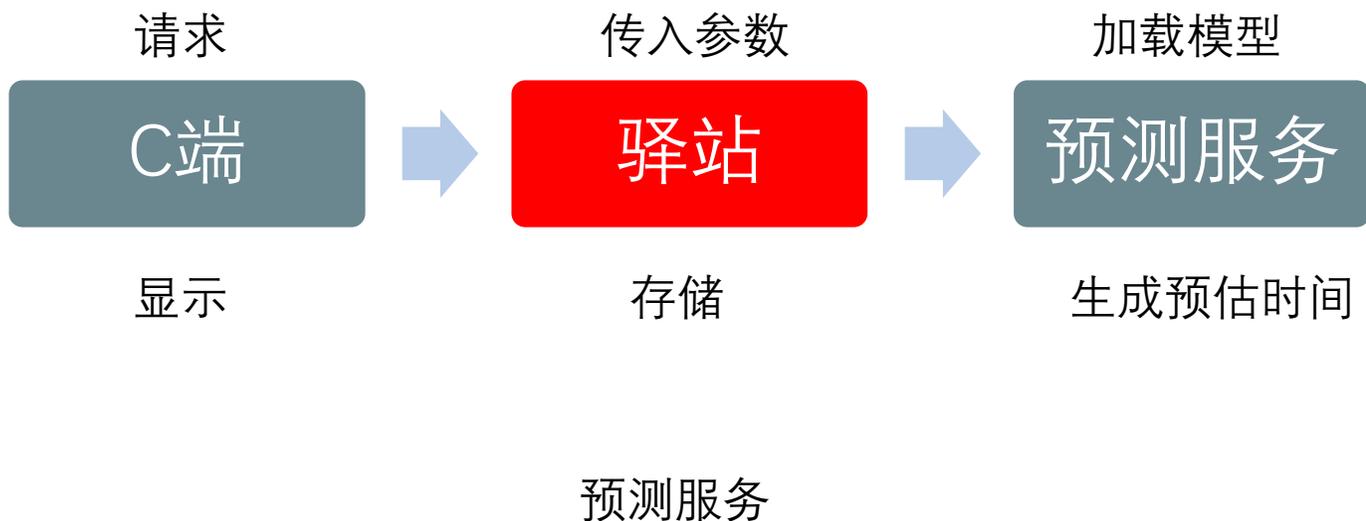
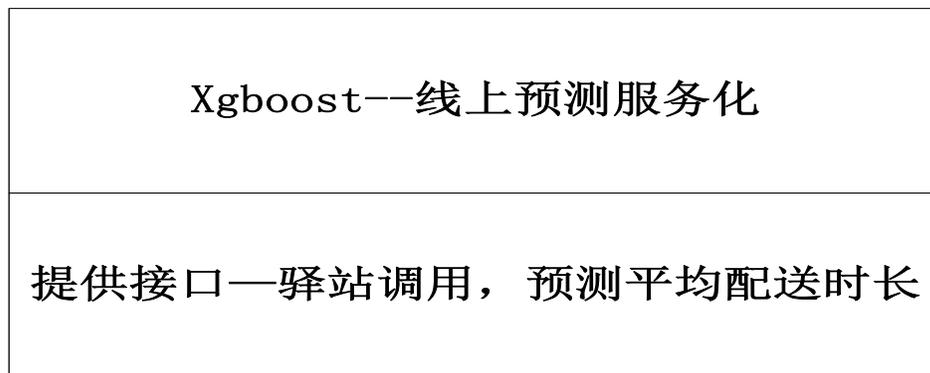
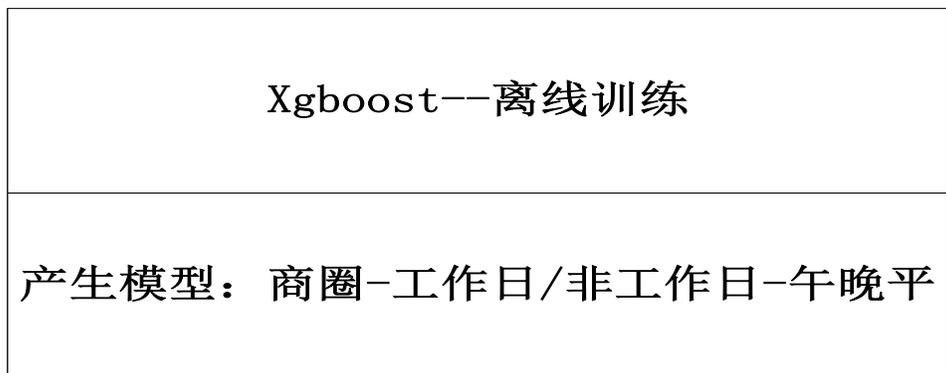
## 03 AI自动化 ( 12箴言 ) 实用化 ( 准入、上线、回归 )

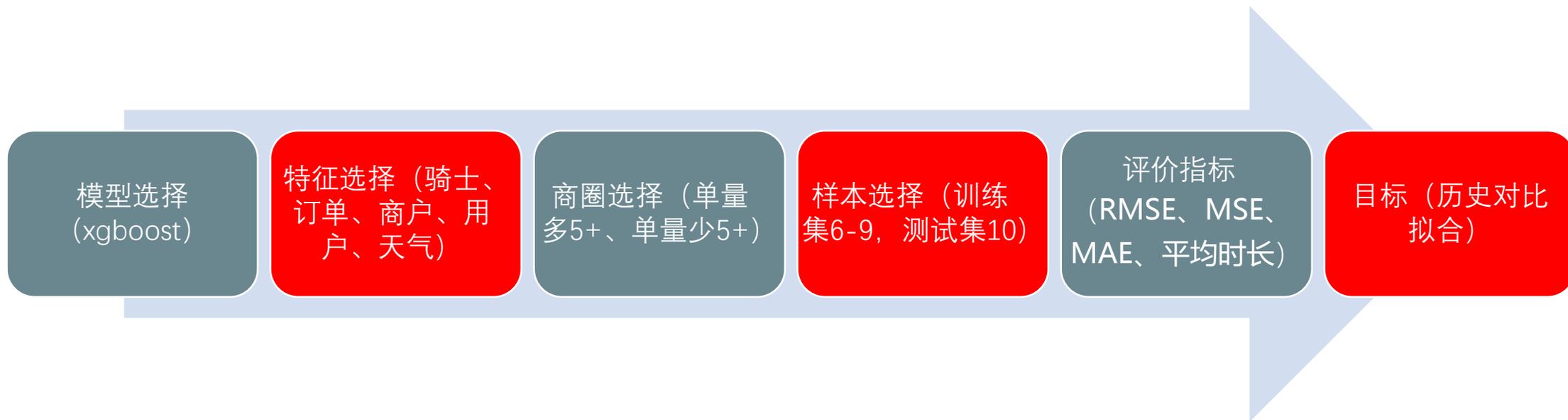
体系化 ( 标准一致，统一服务 )

精细化 ( 小改进 )

常维护 ( 保持年轻 )

基于体系化和效率原则进行三维立体保障





XGBOOST训练流程



双重视角

模型维度&业务维度

两个特点

多迭代&多模型

五种方法

样本：稀疏、不均匀、重复、删除，**调整训测集**

特征：去燥、归一化、删除或增加某一特征

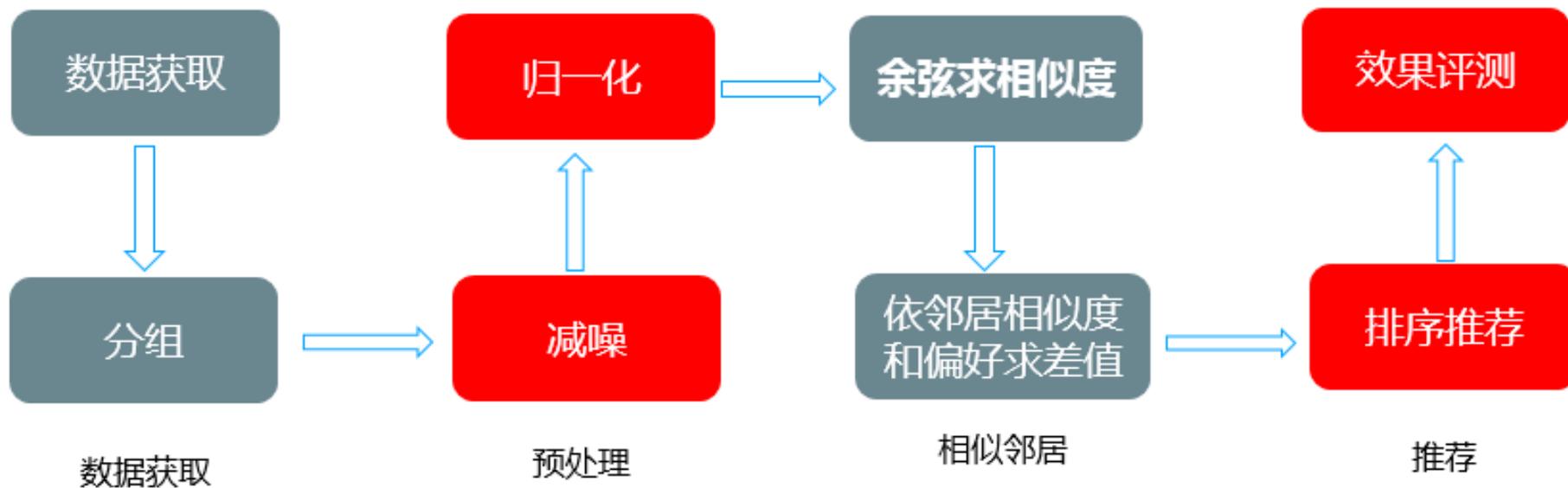
参数：大小、精度调整

模型：多模型**混搭** (gbdt训练特征, lr产出模型)

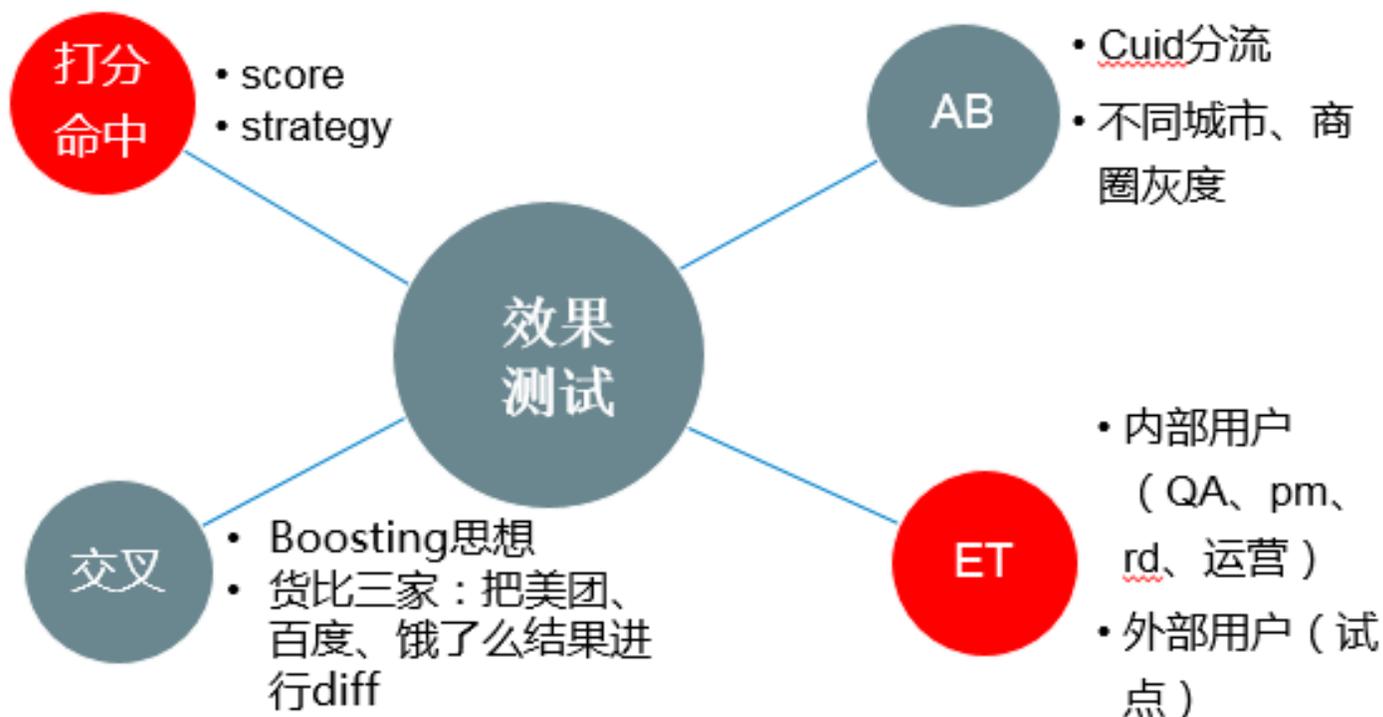
评估：目标、评价、损失函数等进行变量、模型、泛化、衰退评估

**背景：**午饭时间，天气好热，中午吃啥，点开app刷啊刷一看半小时过去；假期综合症，胃口不好，嘉禾、宏状元就出现在首页；核心就是把你最想点的优先展现，从而通过点击率来提升转化，这就是常见的ctr

用户\物品	喝粥	吃面	盖饭
张三	√		√
李四（随便）	√	√	√
王五	√		推荐



不同试点、量级、多角色--基于从小到大，从内到外扩散原则



基于spas原则，集成汇总、ab、debug、用户session

01

**多任务**

单个查看命中策略和原因  
两个进行新旧对比  
多个用户结果一起跑

02

**多场景**

多城市、多商圈、  
多试验、多配比

03

**一体化**

多消息协同作战  
环境使用率高  
智能监测  
自动化的定制化指标

模型  
测试

D : 数据  
维度

S : 模型维  
度

P : 评测维  
度

关注选菜

关注做法

关注好吃

效果  
评测

模型指标---AUC ( 祖传秘方 )

灰度---Abtest ( 北京开家店试试 )

内测---ET ( 试营业几天 )

抽样---RU ( 真实用户试吃评价 )

项目  
总结

**总结—改进—回顾**

排期长

PM：参数训练排3天，有点太多了吧  
RD：我有上亿维特征，要迭代N万次

测试长

QA:上完线回归没问题就完事了  
RD:上线后才是真正的开始，才完成一小半周期

何为对

QA：怎么才能判断算法效果的对错，要不然断言不好加，不好自动化  
RD：策略求的是全局最优，整体是正向还是负向，适合你的才是最好的



IT大咖说  
知识共享平台

携程技术中心

# THANK YOU!

---

## Q&A

本PPT来自携程技术沙龙，  
更多精彩分享请搜索关注“携程技术中心”微信公号~