算法测试探秘

张海宁

MTSC2018
第四届中国移动互联网测试开发大会
TesterHome | IT大咖说
TesterHome

自我介绍

京东

Baidu 百度

百度外卖

北京技术中心

阿里巴巴 Alibaba

引言：算法的魅力

场景A：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

什么是算法

- 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）

CONTENTS

- 算法在外卖的应用
- 算法测试体系建设
- 算法测试案例解析

外卖应用的3个核心场景

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

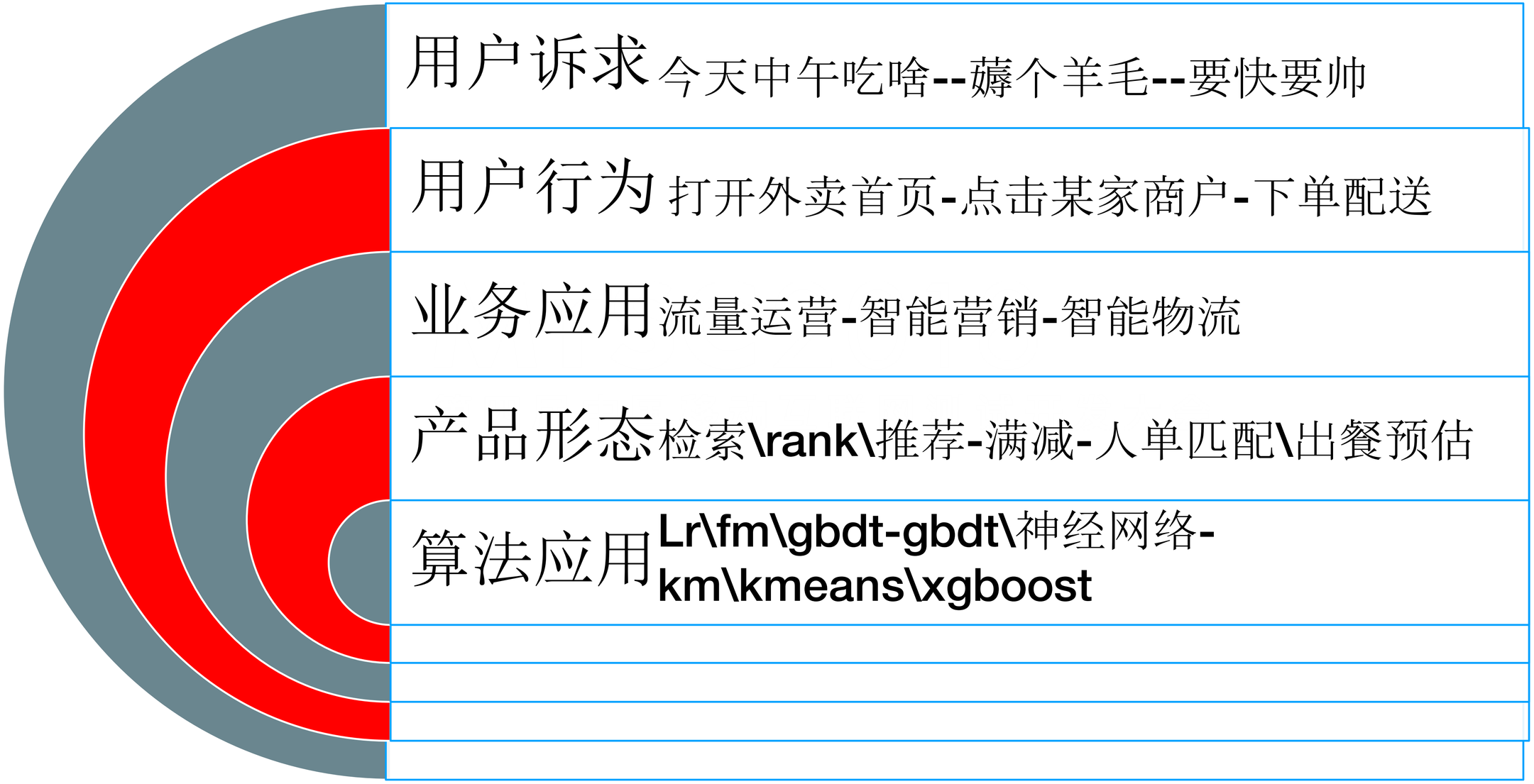
用户B：薅个羊毛

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

算法层级应用


用户诉求 今天中午吃啥--薅个羊毛--要快要帅
用户行为 打开外卖首页-点击某家商户-下单配送
业务应用 流量运营-智能营销-智能物流
产品形态 检索\rank\推荐-满减-人单匹配\出餐预估
算法应用 Lr\fm\gbdt-gbdt\神经网络-km\kmeans\xgboost

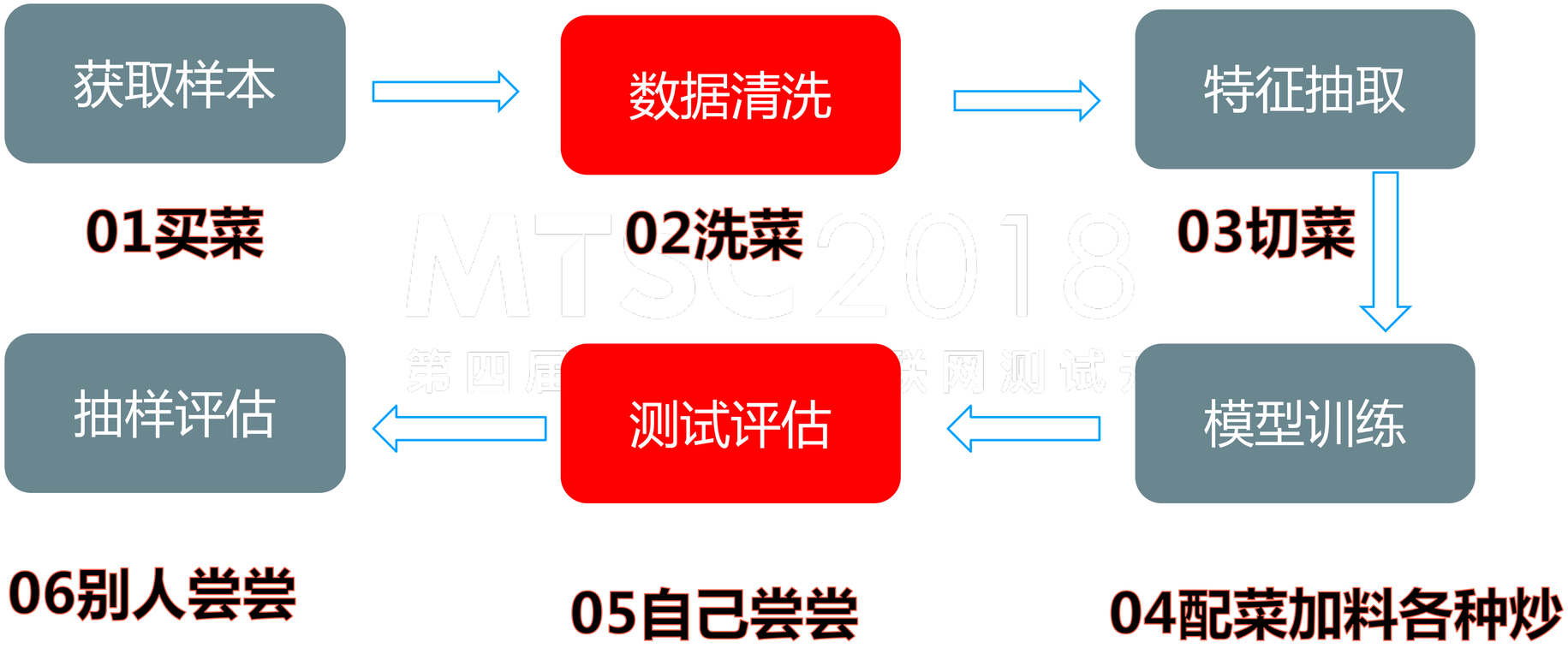
整体生态建设


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
大数据
集群、生态产品、ETL

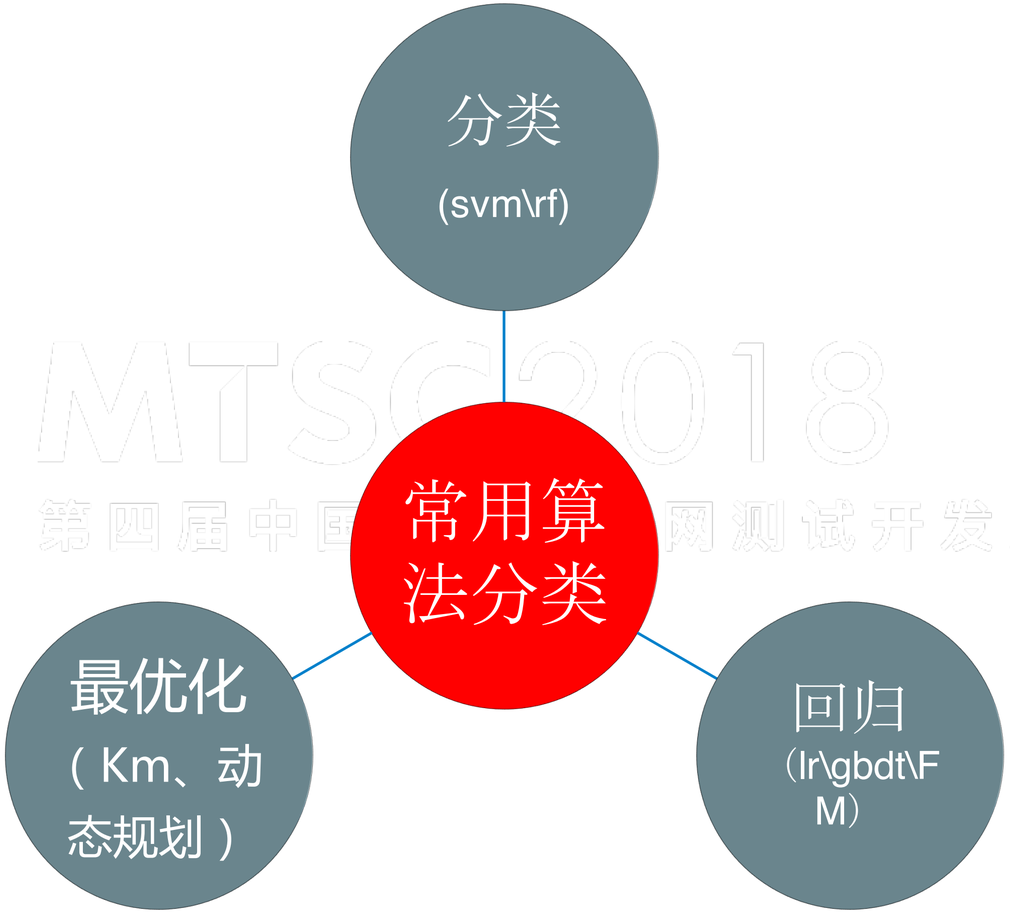
CONTENTS

- 算法在外卖的应用
- 算法测试体系建设
- 算法测试案例解析

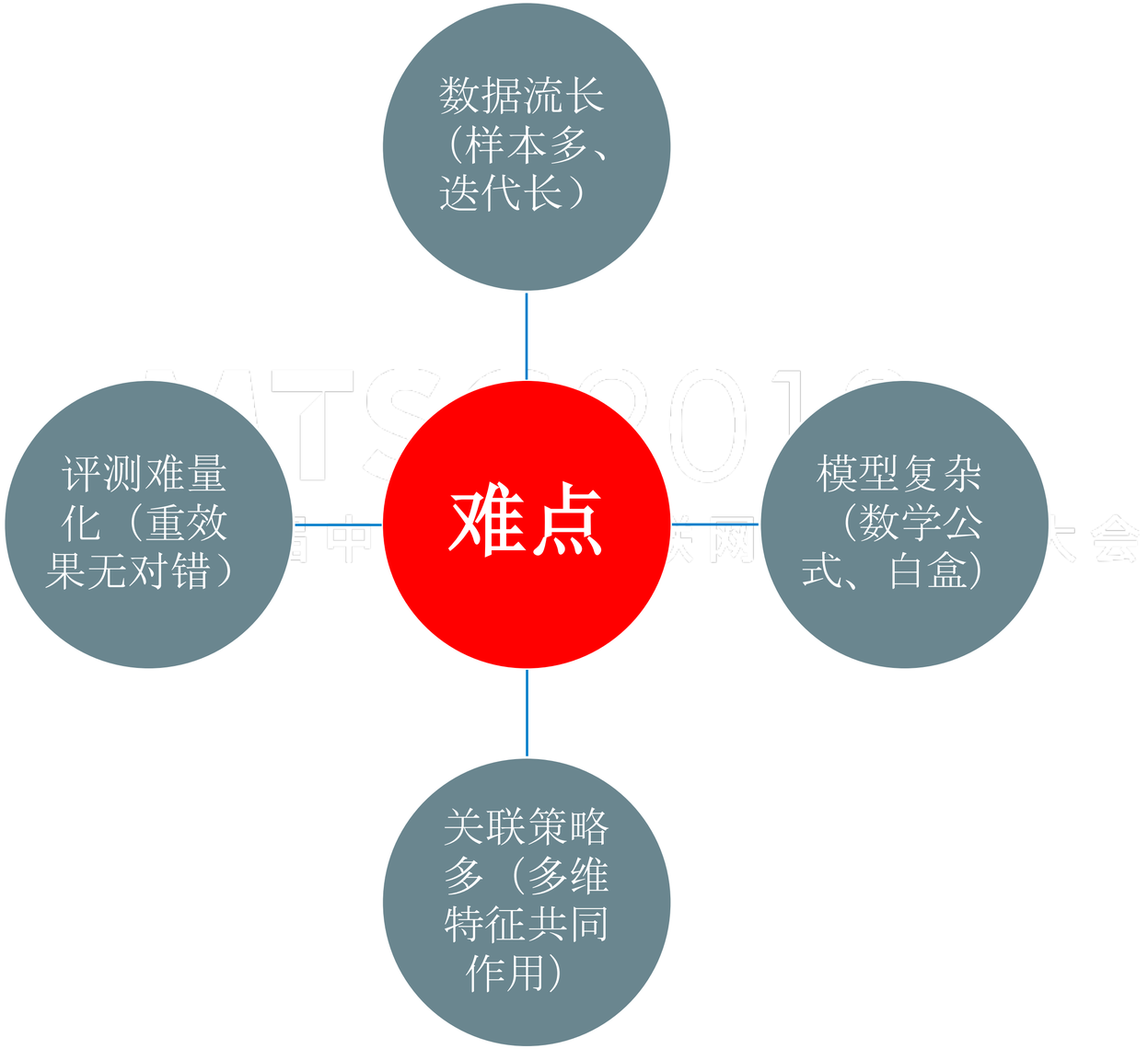
吃货视角下的数据建模流程



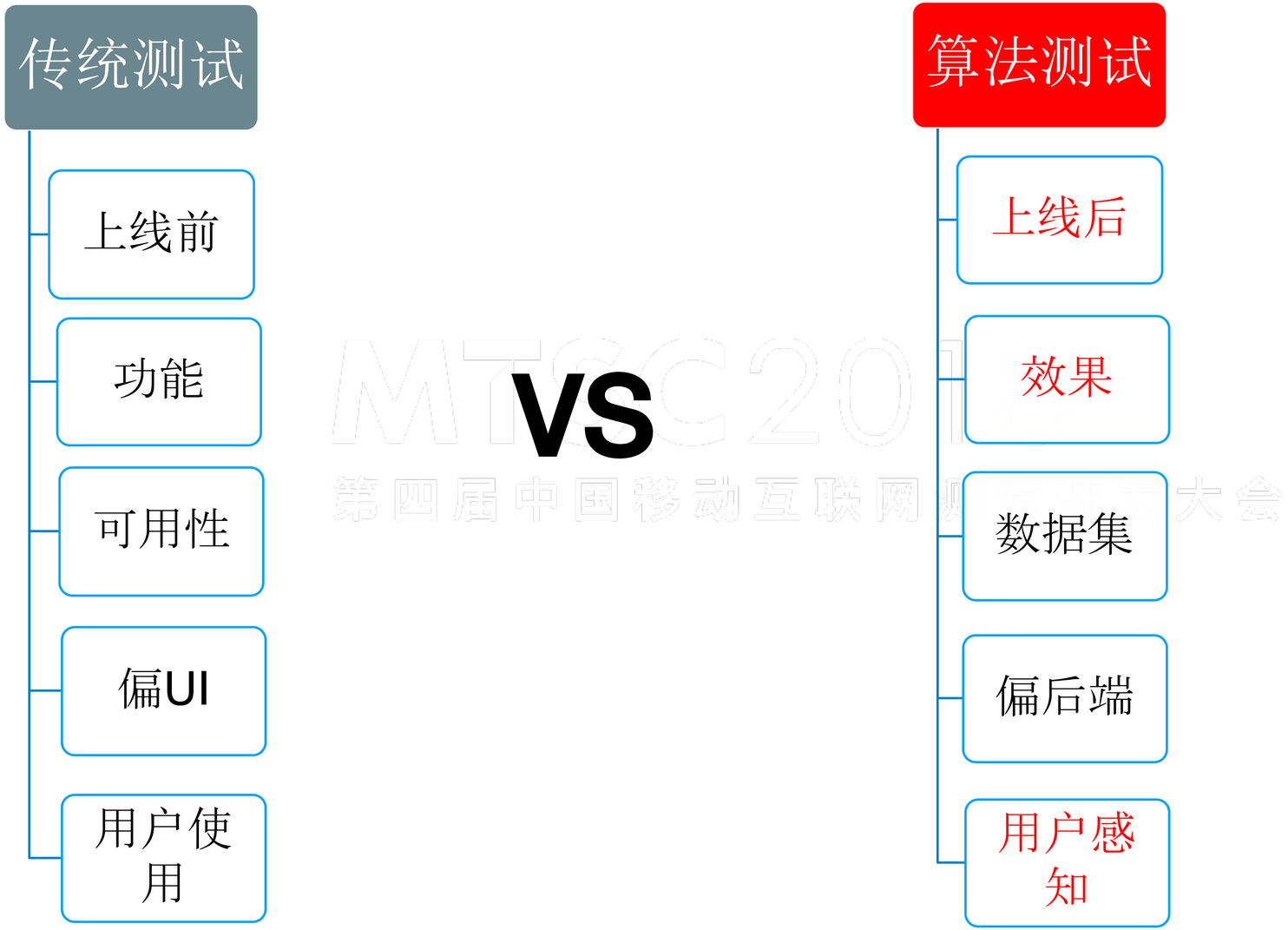
常用算法分类



算法测试难点



与传统测试diff



算法测试核心思路

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）

重要测试场景

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）

上线后的效果测试

- 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

CONTENTS

- 算法在外卖的应用
- 算法测试体系建设
- 算法测试案例解析

数据测试S 满减场景

背景

Such as：30-&；50-(；100-10；

流程

数据处理 → 离线训练 → 档位和偏移值 → 架构存储 → C端展现

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-&，是未生效还是结果一样

数据测试—如何测

关联测试	数据测试	监控
• 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）	• 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）	• 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）

多维测试

数据测试—总结

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 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）

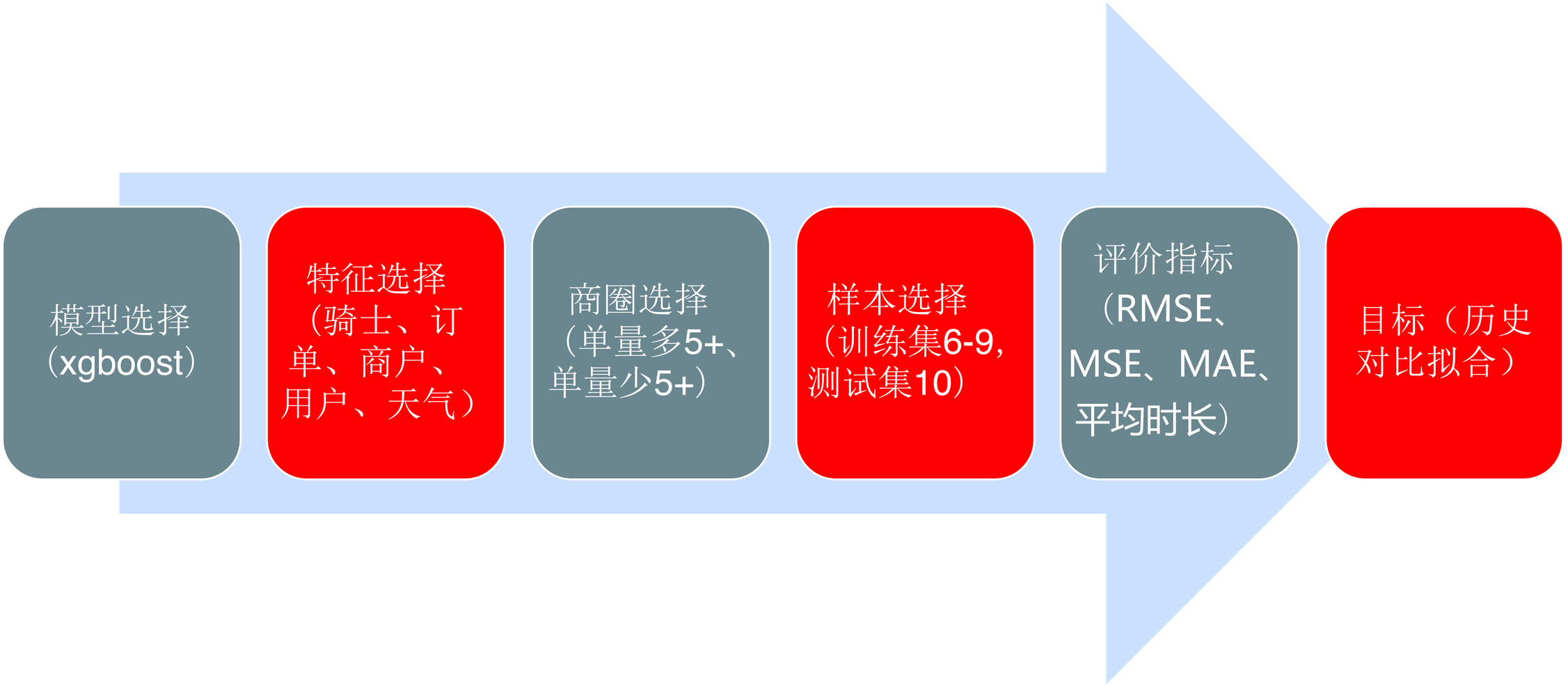
基于体系化和效率原则进行三维立体保障

模型测试—配送时长预估场景

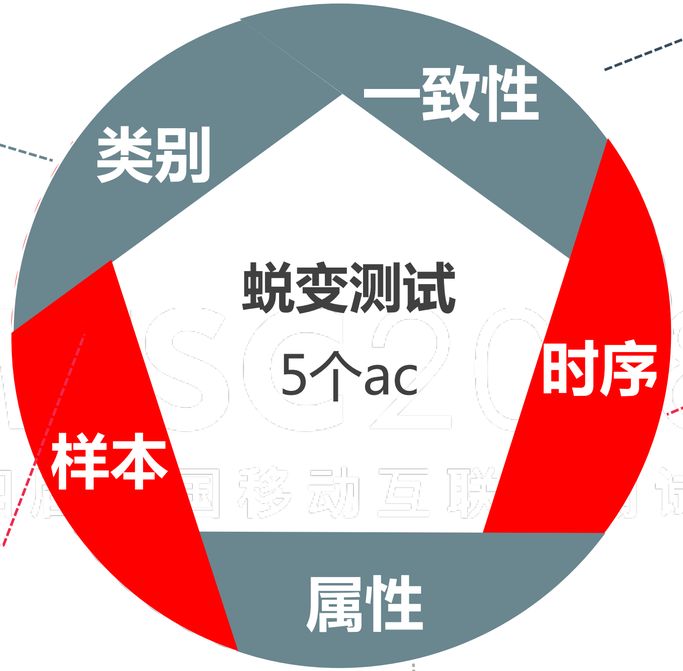
Xgboost - - 离线训练	Xgboost - - 线上预测服务化
产生模型 ： 商圈- 工作日/ 非工作日 - 午晚平	提供接口 —驿站调用， 预测平均配送时长

预测服务应用

模型测试-训练流程


模型选择（xgboost）
特征选择（骑士、订单、商户、用户、天气）
商圈选择（单量多5+、单量少5+）
样本选择（训练集6-9，测试集10）
评价指标（RMSE、MSE、MAE、平均时长）
目标（历史对比拟合）

模型测试—蜕变测试



复制样本转换成新的标签

Diff，保持预测和测试结果一致

利用random让类别、属性乱序

删减、重复、附加训练样本

增加有无信息属性，其他保持一致

模型测试—总结

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

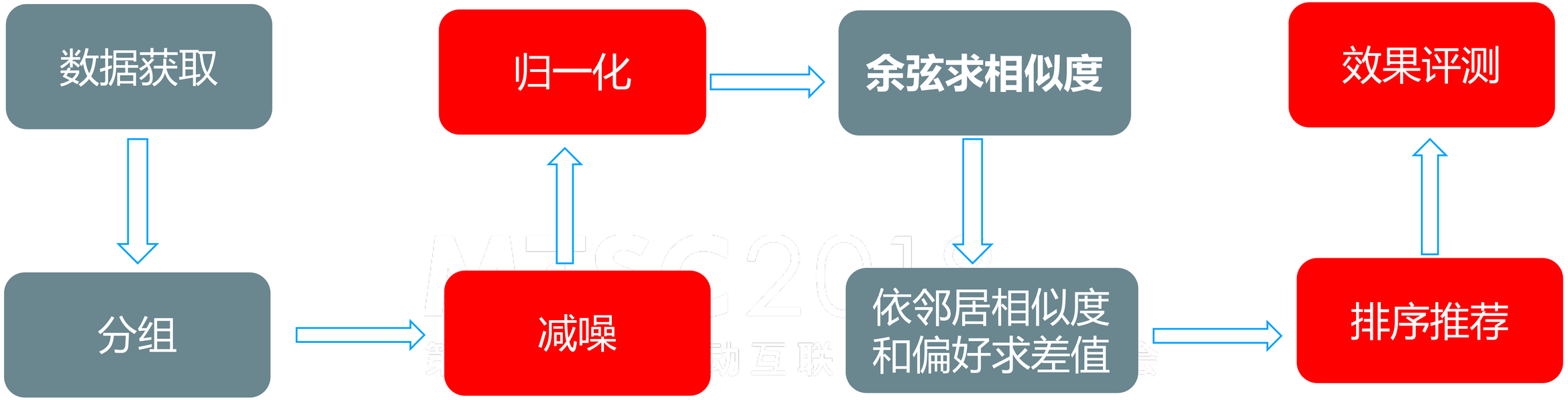
- 样本：稀疏、不均匀、重复、删除，调整训测集
- 特征：去燥、归一化、删除或增加某一特征
- 参数：大小、精度调整
- 模型：多模型混搭（gbdt训练特征，lr产出模型）
- 评估：目标、评价、损失函数等来进行变量、模型、泛化、衰退评估

效果测试-推荐排序场景

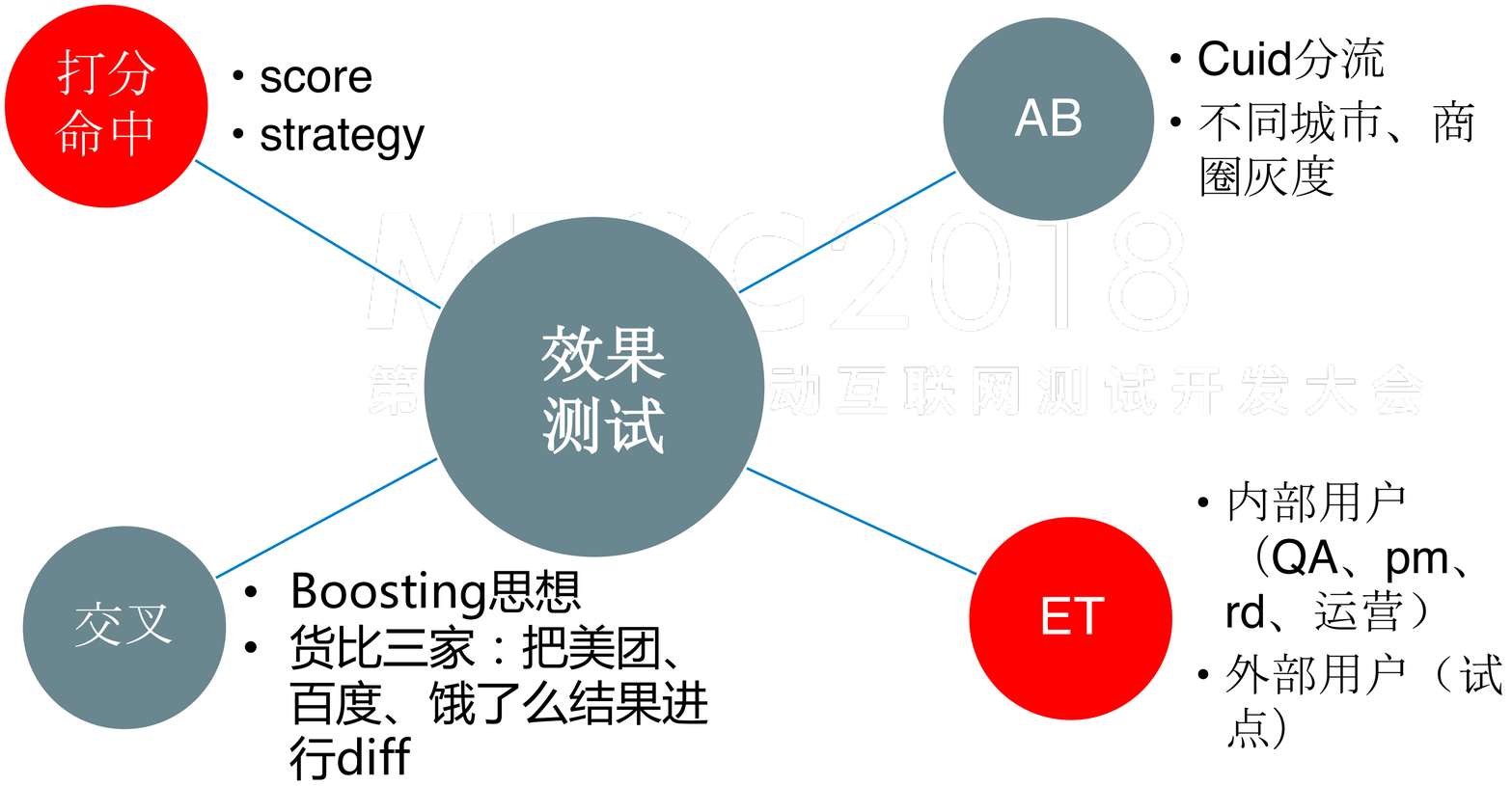
背景：午饭时间，天气好热，中午吃啥，点开app刷啊刷一看半小时过去；假期综合症，胃口不好，嘉禾、宏状元就出现在首页；核心就是把你最想点的优先展现，从而通过点击率来提升转化，这就是常见的ctr

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

效果测试--推荐流程


数据获取
分组
减噪
归一化
余弦求相似度
依邻居相似度
和偏好求差值
排序推荐
效果评测
数据获取
预处理
相似邻居
推荐

基于从小到大，从内到外扩散原则（不同试点、量级、多角色）



效果测试——一体化服务

基于SPAS原则，集成汇总、ab、debug、用户行为析、画像等

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

整体回顾-343

模型测试

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

About算法有3个梗

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

谢谢大家

MTSC2018

第四届中国移动互联网测试开发大会

TesterHome | IT大咖说

TesterHome