主办方：腾讯云 | 51CTO

大数据在教育行业的研究与应用

谭安林

腾讯　高级工程师

SPEAKER

谭安林

腾讯　高级工程师

2015年加入腾讯，8年互联网从业经历，从事大数据平台、产品开发相关工作；先后参与广告、金融等领域产品项目，目前负责行为预测解决方案；针对教育行业，提供行业定制的行为分析与预测建模，助力客户数据增长。

探索大数据赋能行业的方向

平台技术  数据服务

移动市场增长变缓，存量用户价值高

2017年移动设备增长趋势&2018年Q1进入稳定发展期

移动端设备稳定在12亿左右，移动端新增用户人口红利基本消失。争夺市场即争夺存量用户，APP市场竞争白热化。

在线教育用户规模持续增长

2013-2018年在线教育持续增长

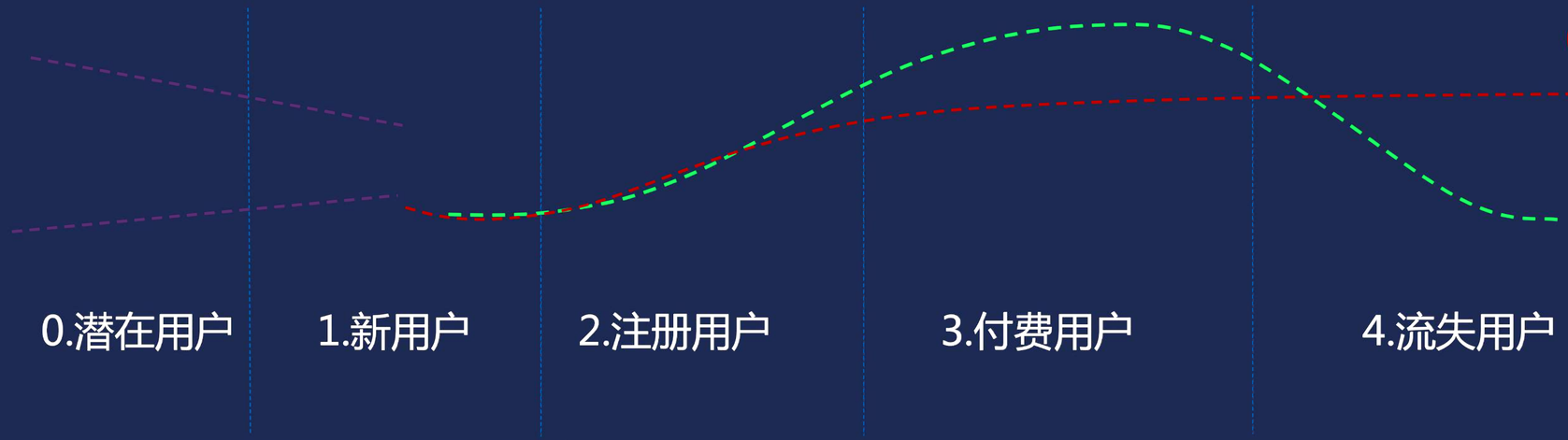
在线教育技术的持续升级，产品的丰富与成熟度，未来几年教育市场将进一步增长。

运营痛点— 谁会成为你的用户？谁价值高？谁容易流失？

A.精准营销
（获客管理、提升ROI）

B.用户价值提升
（用户分级、差异营销）

C.用户生命周期延长
（用户粘性、忠诚度管理）



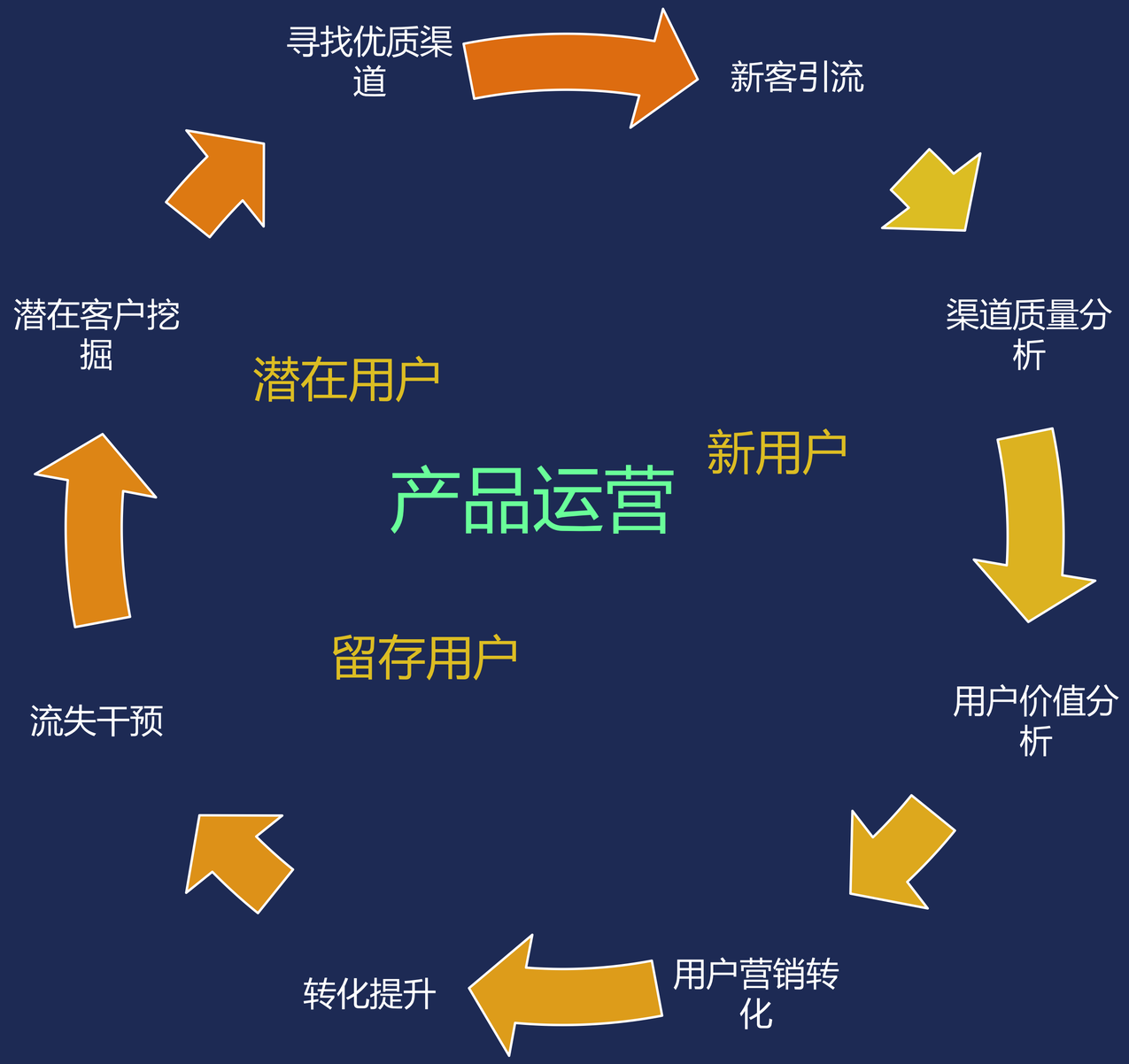
0.潜在用户　1.新用户　2.注册用户　3.付费用户　4.流失用户　5.回流

- 针对哪些用户营销，最有效？
- 如何识别高价值用户提升变现能力？
- 如何识别用户粘性高低，以定制化服务策略，延长用户生命周期
- 针对不同再付费概率的用户，定制不同的营销策略，如何区分？

潜在用户

既得用户

贯穿始终的数据工作


寻找优质渠道
新客引流
渠道质量分析
用户价值分析
用户营销转化
转化提升
流失干预
潜在客户挖掘
潜在用户
新用户
产品运营
留存用户

数据体系搭建思路

指标规划

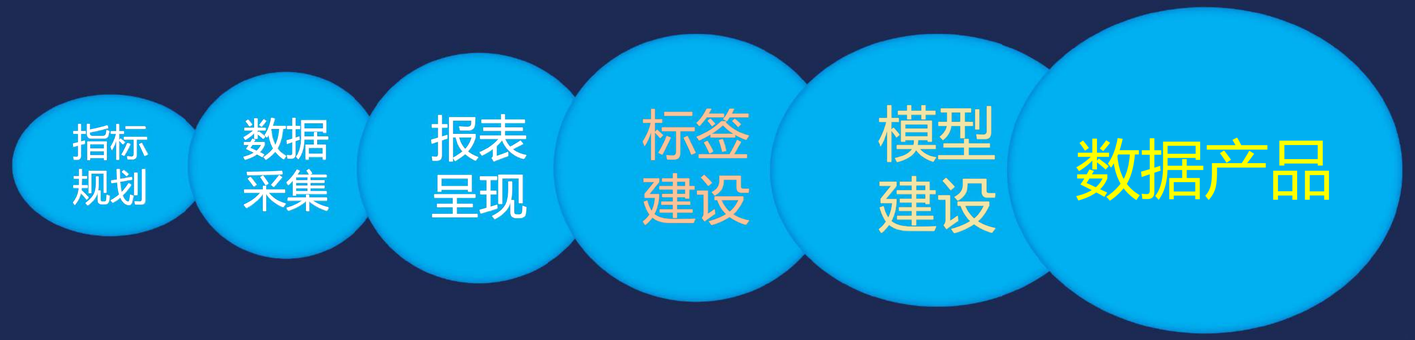
- 指标定义
- 维度设置
- 更新周期

数据采集

- 字段分类
- 数据埋点
- 数据上报

报表呈现

- 趋势图/列表
- 筛选控件
- 有效性、准确性验证



标签建设

- 标签定义
- 有效性评估
- 相关性分析

模型建设

- 样本选择
- 特征裁剪
- 调参优化

数据产品

- 数据可视化
- 迭代优化
- 新功能增加

数据覆盖少，实施成本大

- 推动难

--运营需联动产品、开发多岗位人员，需较高段位和职级才能推动

- 数据少

--受限于采集方式、用户使用频次等因素，存在新用户冷启动、留存用户数据少等问题。

- 成本高

--数据建设是一套复杂的流程，且最终实际应用需系统化，整体需较大的人力成本与资源

- 评估难

--怎么验证标签的有效性？

- 选择难

--标签体系大而全，不知哪些维度会真正有用？

- 使用难

--标签如何使用？如何落地在策略运营？

行为预测服务

化繁为简，助力攻克标签建设、模型建设2大难关

落地难

- 推动难
- 数据少
- 成本高

简化落地

20人 → 2人

- 上传数据即可，系统自动生成标签
- 融入互联网大盘脱敏数据，洞察客户更精准
- 节省繁琐的内部标签建设过程，效率高，成本低

应用难

- 选择难
- 使用难
- 评估难

模型预测

标签 → 概率

- 基于标签，系统建模预测行为概率，无需人工选择标签，定义阈值
- 自定义用户分群，一步到位

小步实验

人工 → 工具

- 基于用户分群，小步实验工具助力策略A/B test
- 跟踪用户行为，效果可视化

行为预测服务

服务于精细化运营：更少的成本，更好的效果

标签化用户数据

用户分群

自定义分群：

1.高危流失用户

2.一般粘性用户

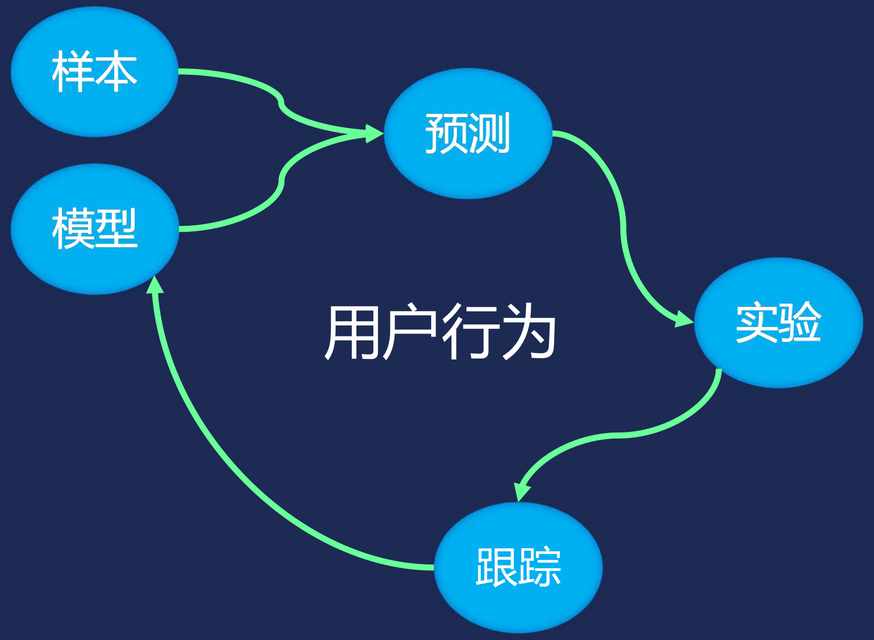
3.高粘性用户

指导精细化策略

- 短信触达 高付费用户
- 精准推送 高流失用户
- 建群服务 VIP用户
- 广告投放 优质潜客

行为预测服务

使用介绍：从有到无，寻找高价值用户



1.盘活留存用户

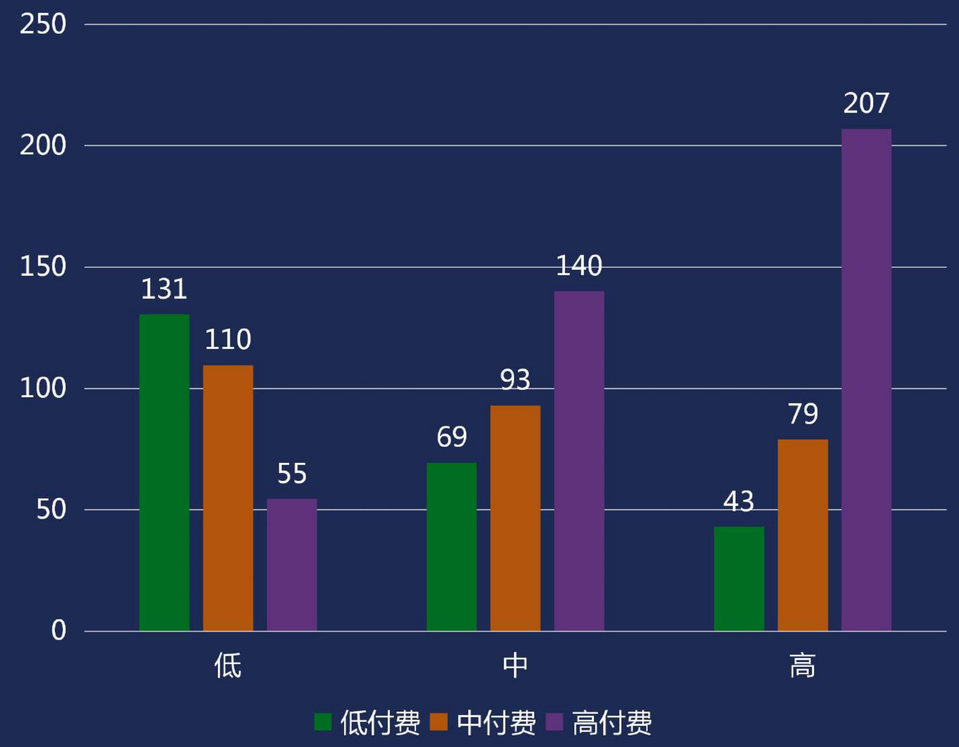
2.洞察新增用户

3.挖掘潜在用户

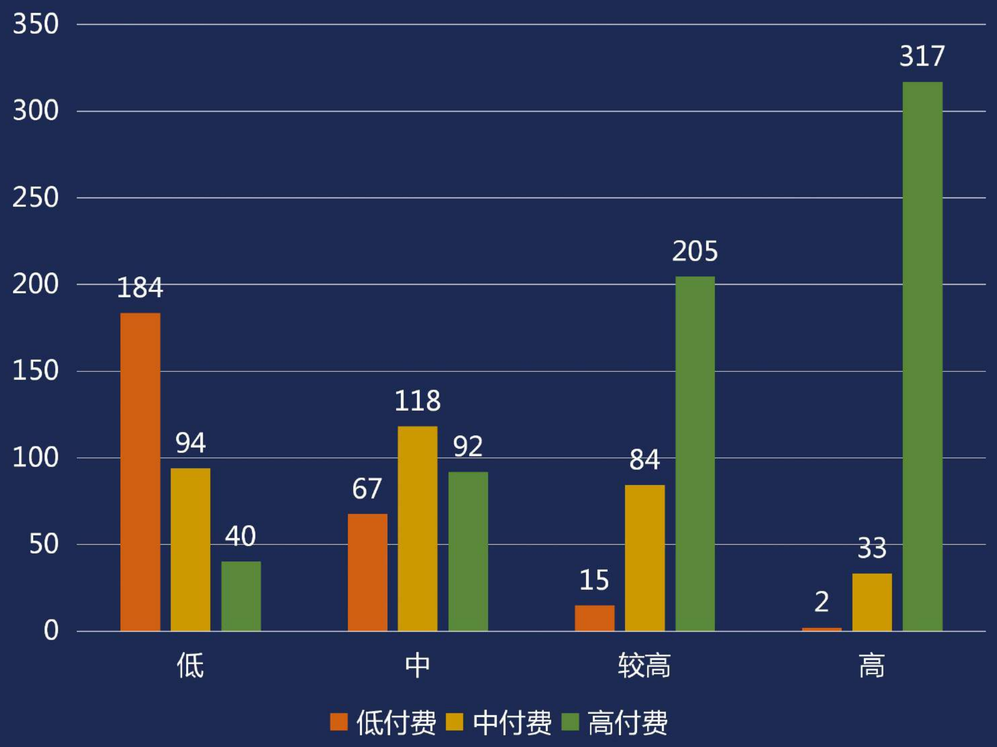
行为预测 服务模式1 盘活存留用户

分群洞察 案例分享

- **教育关注度**：关注度越高，越愿意付费



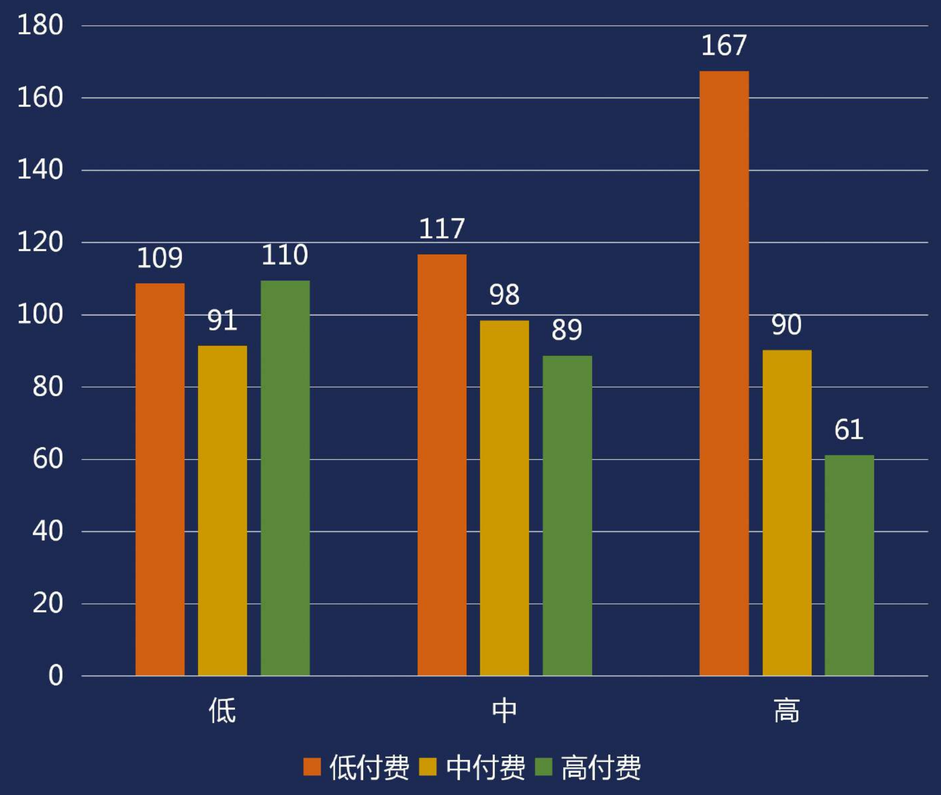
- **教育坚持度**：越能坚持的人付费概率越高



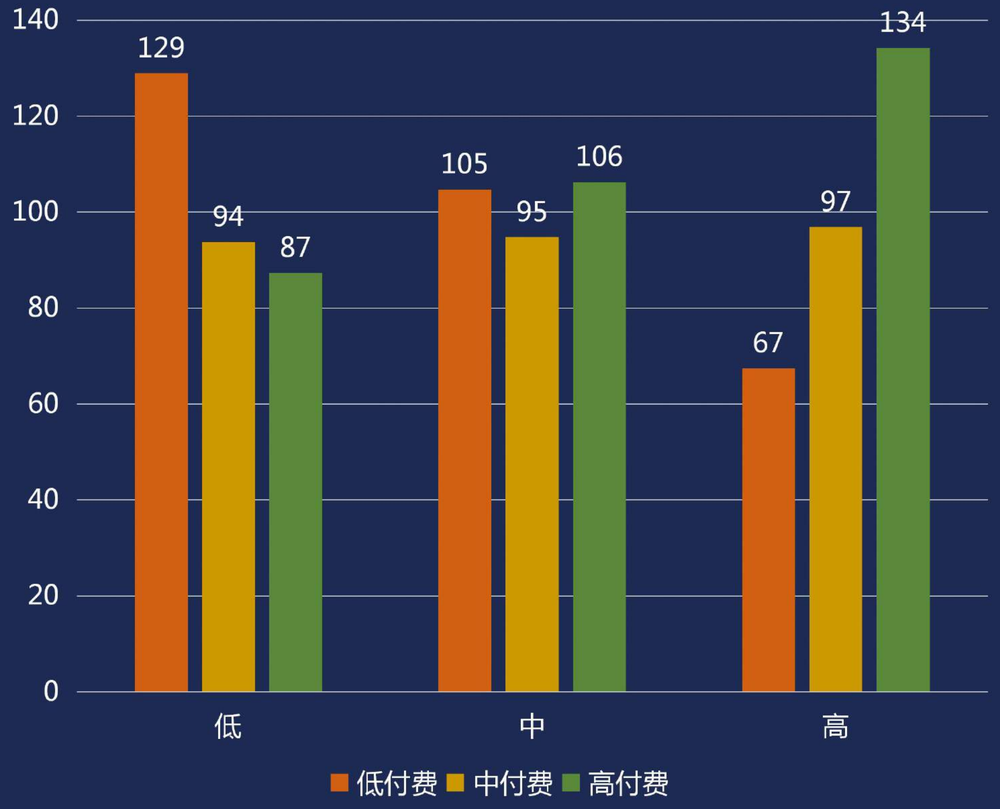
相对显著性 TGI指数= [目标群体中具有某一特征的群体所占比例/总体中具有相同特征的群体所占比例]*标准数100。

分群洞察 案例分享

- **游戏沉迷度**：越沉迷，越不可能在教育上付费

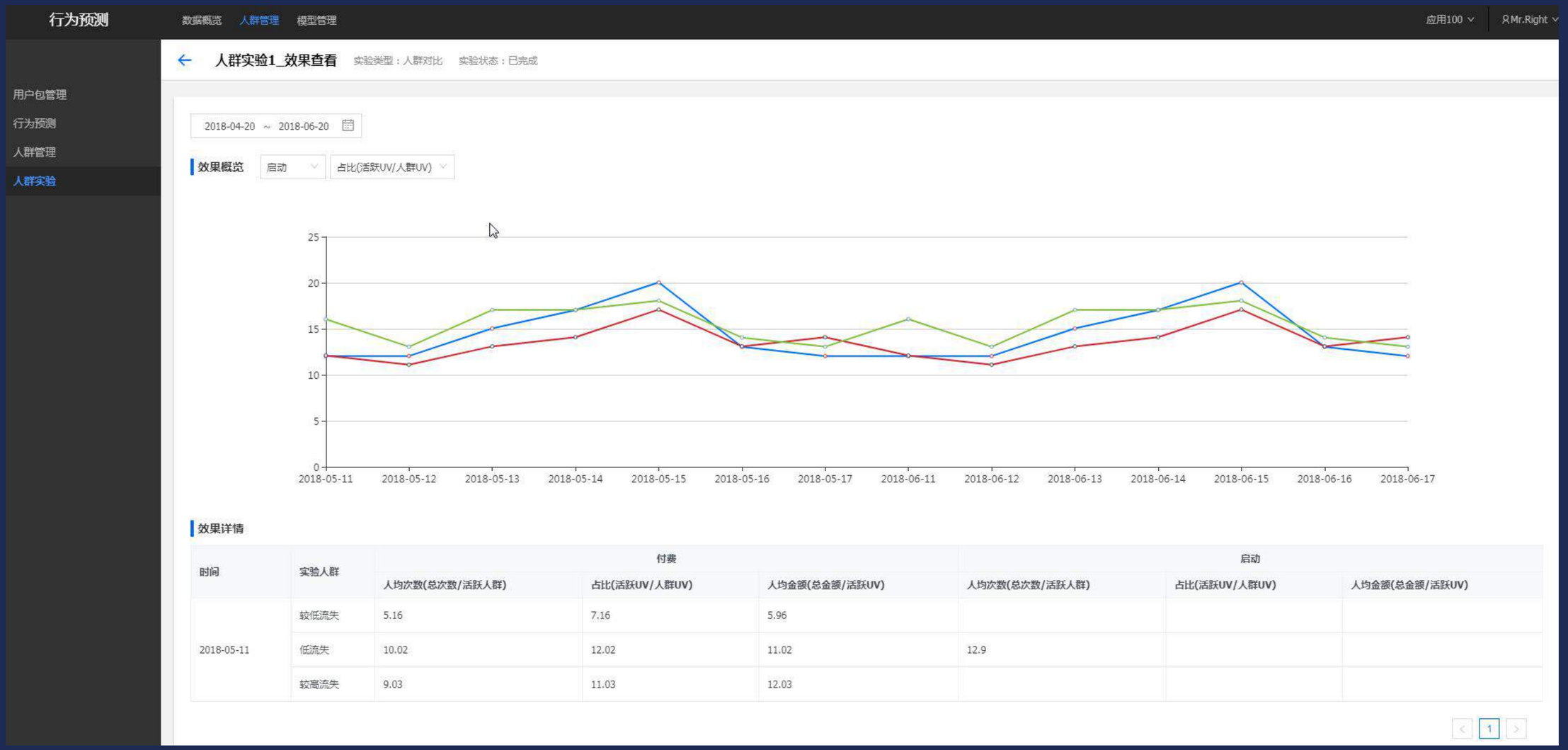


- **自我驱动力**：越上进的人付费可能性越大



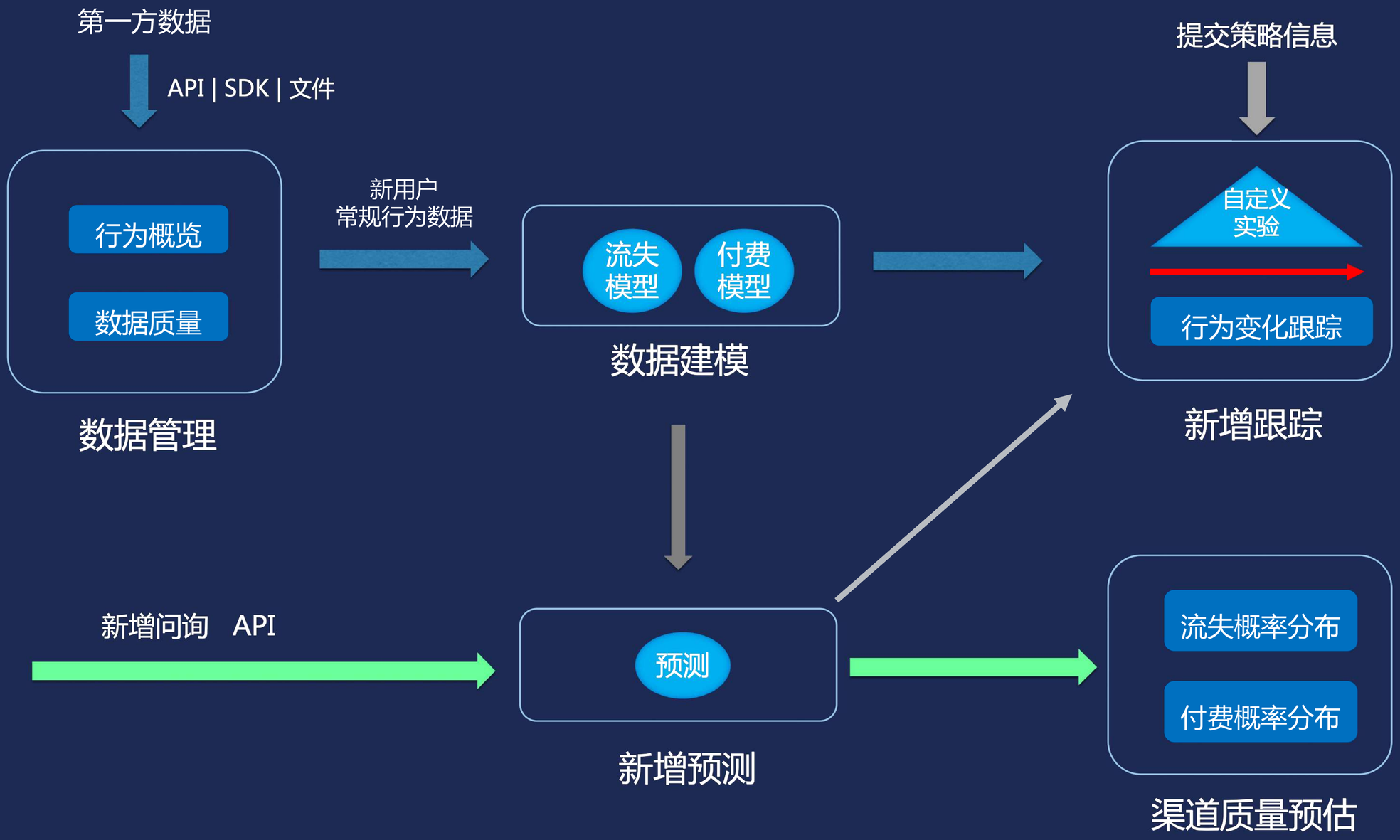
相对显著性 TGI指数= [目标群体中具有某一特征的群体所占比例/总体中具有相同特征的群体所占比例]*标准数100。

实验效果跟踪 页面demo

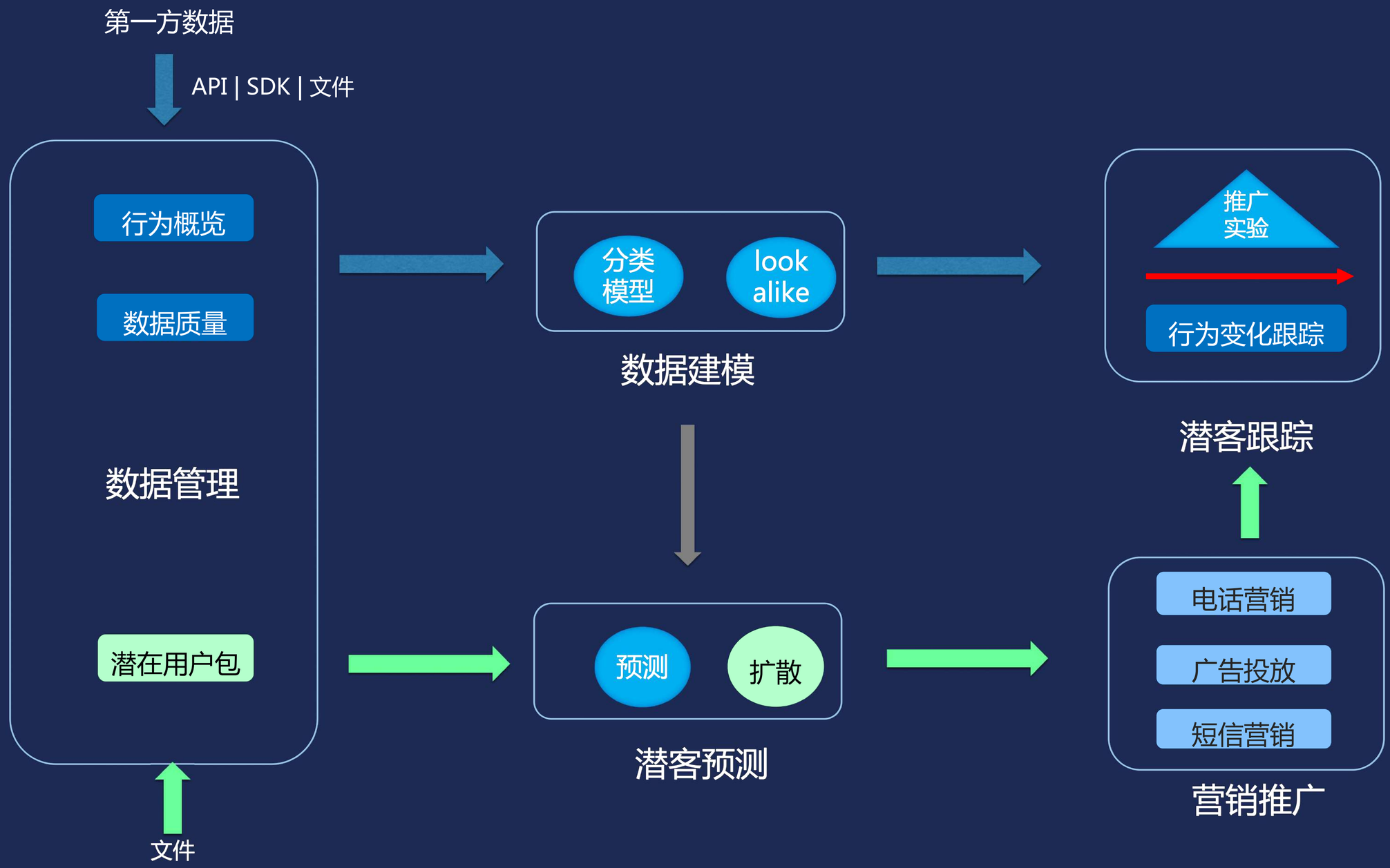


时间	实验人群	付费			启动		
		人均次数(总次数/活跃人群)	占比(活跃UV/人群UV)	人均金额(总金额/活跃UV)	人均次数(总次数/活跃人群)	占比(活跃UV/人群UV)	人均金额(总金额/活跃UV)
2018-05-11	较低流失	5.16	7.16	5.96			
	低流失	10.02	12.02	11.02	12.9		
	较高流失	9.03	11.03	12.03			

行为预测 服务模式2 洞察新增用户



行为预测 服务模式3 挖掘潜在客户



Lookalike

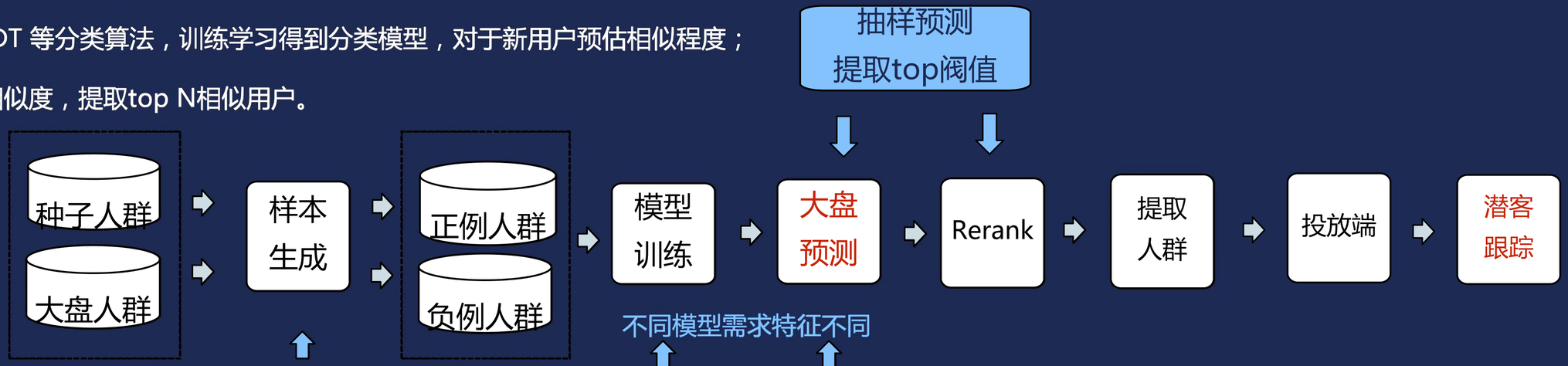
相似用户挖掘可简化为一个二元分类问题：

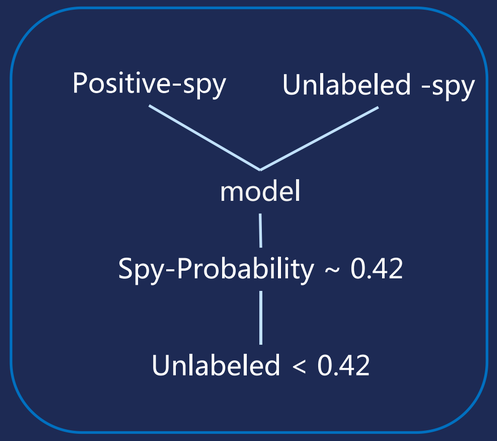
1.种子用户为正例（Positive data）；

2.从非种子用户中随机筛选出一定比例的“负例”（Unlabeled data）；

3.选用 LR 或 GBDT 等分类算法，训练学习得到分类模型，对于新用户预估相似程度；

4.预测大盘用户相似度，提取top N相似用户。



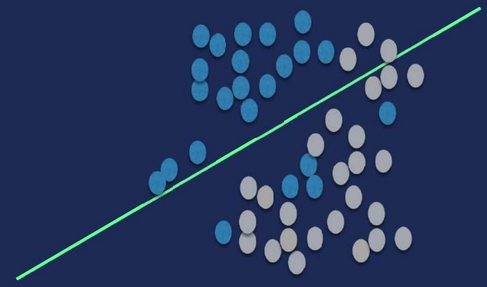


1.从大盘随机抽样本

2.标签抽取负样本

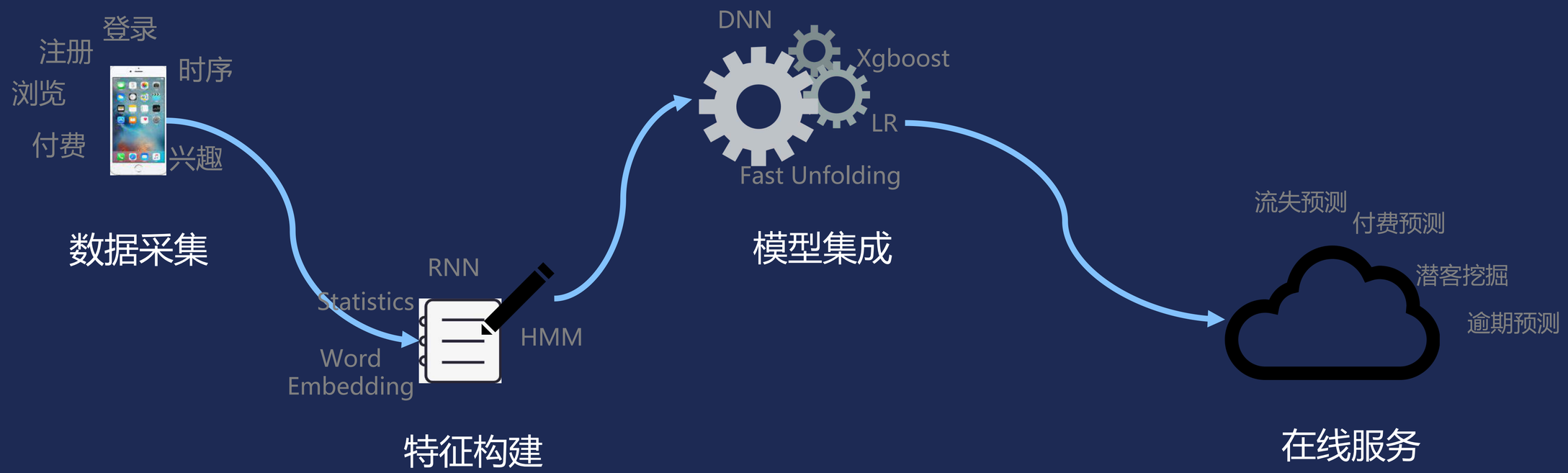
3.PULearing训练负样本

特征集市KV

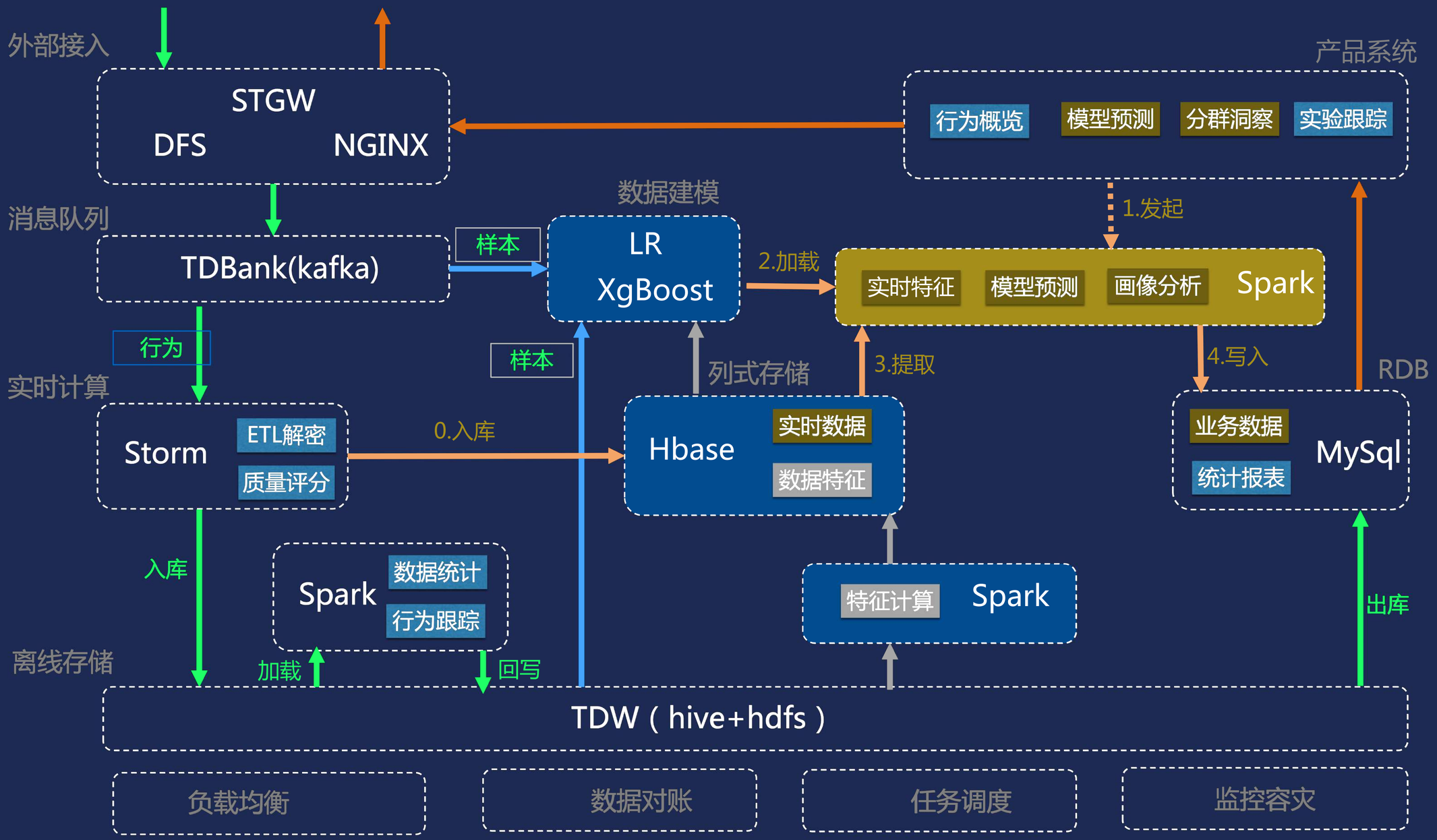


行为预测服务

脱敏行为数据，实践多种模型



整体技术架构


外部接入
STGW
DFS
NGINX
产品系统
行为概览
模型预测
分群洞察
实验跟踪
消息队列
TDBank(kafka)
样本
数据建模
LR
XgBoost
2.加载
1.发起
实时特征
模型预测
画像分析
Spark
行为
实时计算
样本
列式存储
3.提取
4.写入
RDB
Storm
ETL解密
质量评分
0.入库
Hbase
实时数据
数据特征
业务数据
统计报表
MySql
入库
Spark
数据统计
行为跟踪
特征计算
Spark
出库
离线存储
加载
回写
TDW（hive+hdfs）
负载均衡
数据对账
任务调度
监控容灾

整体产品解决方案


场景应用
盘活留存
流失先兆分析
差异化流失挽回
差异化付费营销
新增&潜客
渠道质量预估
新增用户营销
潜客预测与扩散
风险识别
信用风险
欺诈风险
服务落地
用户分群
基于概率分群
分群用户洞察
不同群体对比
小步实验
策略对比
人群对比
自定义策略
效果闭环
实验效果跟踪
模型迭代优化
数据建模
标签工具
行业标签
场景标签
特征工程
特征选择
特征组合
数据建模
监督学习/半监督学习
深度学习/增量学习
数据资源
第一方数据
样本数据
行为数据
注册信息
内外融合
内部画像
更精准的标签
更广的覆盖维度

模型建设与第一方数据

	模型子类	主要算法	模型AUC
教育行业 --流失模型	• 7天流失模型	LR	• 0.896
	• 14天流失模型		• 0.895
	• 30天流失模型		• 0.782
教育行业 --付费模型	• 会员付费模型	XGBoost	• 0.76
	• 商城付费模型		• 0.806
汽车行业 --邀约模型	• 邀约试驾模型	XGBoost	• 0.763

第一方数据的**完善程度、质量情况**,**与模型效果呈正相关**

教育类预测的线上特征库

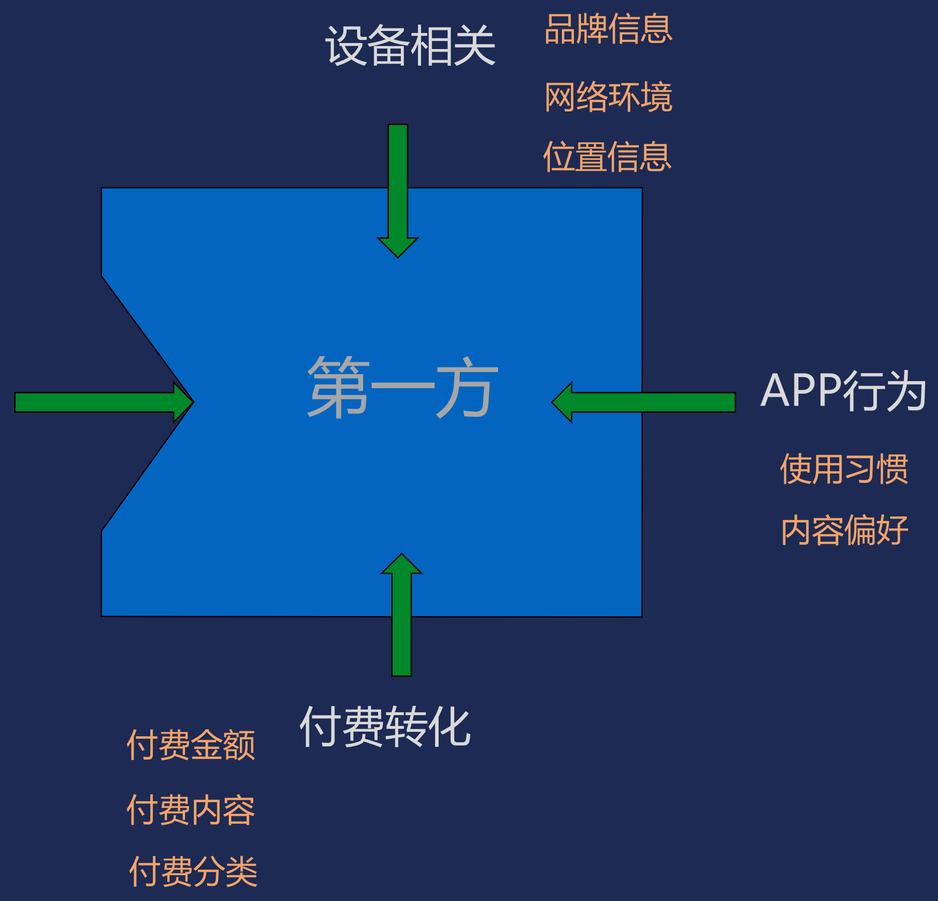
- 人口属性
- 设备属性
- app属性
- LBS属性
- …

- 教育属性
- 汽车属性
- 游戏属性
- …

- 产品关注
- 竞品使用
- 特定属性
- …

- 流失场景
- 付费场景
- …

第一方数据与安全机制



数据传输

非对称加密、对称加密

数据存储

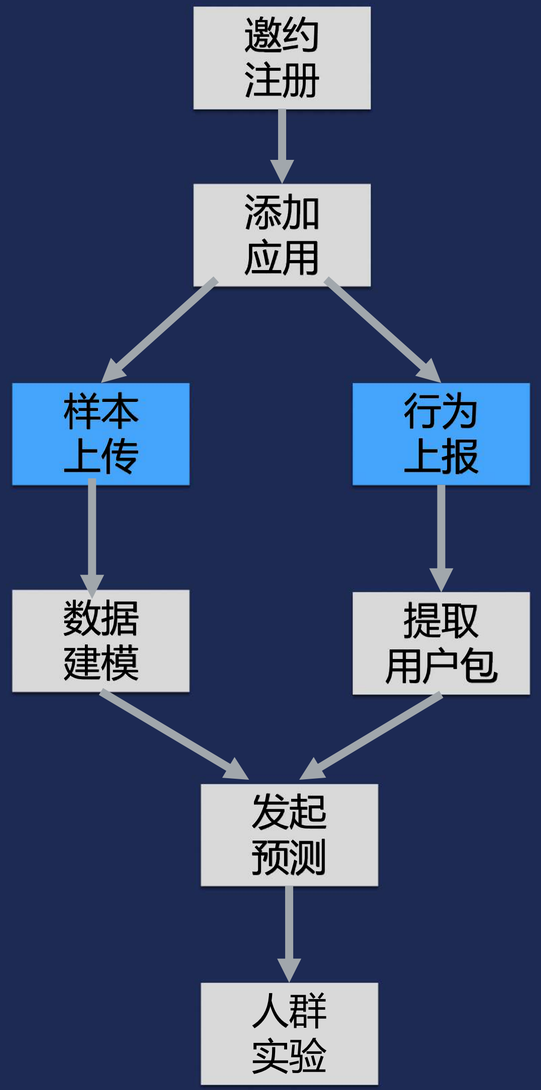
分表存储、权限管控

数据脱敏

帐号加密

内容脱敏

接入方式



- 行为数据通过API | SDL 进行上报

action	相关解释
install	安装
start	启动
reg	注册
login	登录
brow	浏览
buy	购买付费
member	开通会员
msg	消息通知授权
share	分享
complain	投诉
update	更新版本
uninstall	卸载

联系方式

合作洽谈

技术交流

技术支持

Thanks!

联合主办方：腾讯云 | 51CTO

直播支持：腾讯课堂
KE.QQ.COM 学习成就梦想