智能优化&A/B测试

实验驱动用户增长的理论与技术实践

陈冠诚

Testin云测CTO

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE
全球技术领导力峰会

Geekbang. 极客邦科技 | TGO 鲲鹏會

500+高端科技领导者与你一起探讨
技术、管理与商业那些事儿

2019年6月14–15日 | 上海圣诺亚皇冠假日酒店

扫码了解更多信息

自我介绍

01

IBM Research

高级科学家

创新、大数据、云计算、机器学习、OpenPOWER、学术论文、发明、专利

02

OneAPM

大数据首席架构师

SaaS创业、产品、APM，Spark，Druid，开源社区、机器学习、大数据

03

Testin
云测

副总裁,GM -> CTO（16年至今）

A/B测试、智能测试，众包用工，AI采集标注，产品、增长、数据分析、创业、创新

目录

- A/B 测试如何助力用户增长？
- 如何在团队中有效推进A/B测试？
- A/B 测试实际案例分享
- 智能优化 & A/B 测试系统技术实践

估值750亿美金的字节跳动的增长秘籍是什么？

“头条发布一个新APP，其名字都必须打N个包放到各大应用市场进行多次A/B测试而决定，张一鸣告诉同事：**哪怕你有99.9%的把握那是最好的一个名字，测一下又有神马关系呢？”**- 来源1

实验驱动，科学增长

数据分析：发现问题

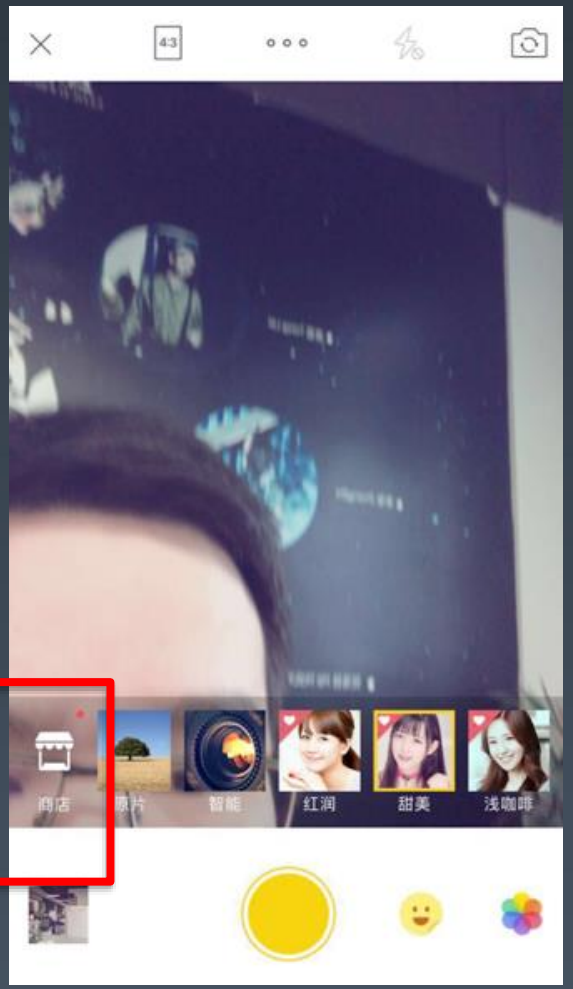
A/B测试：解决问题

实验驱动增长 = 数据分析 X A/B测试

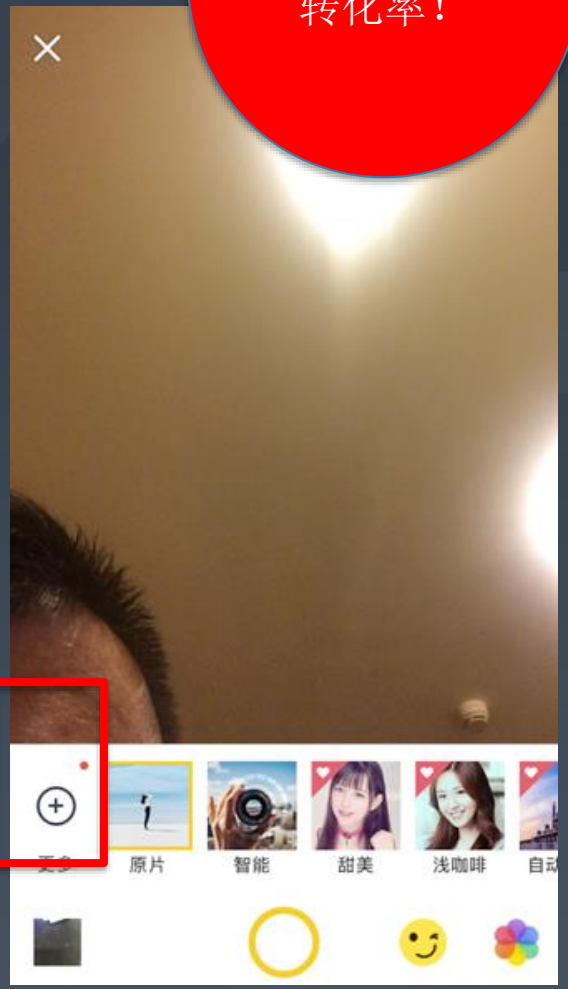
来源1：《你的时间，要么姓张，要么姓张——张小龙和张一鸣的对立统一》，https://36kr.com/p/5130129.html
来源2： 以今日头条为例，详述数据思维驱动产品设计的方法论，http://www.woshipm.com/pd/726436.html

猜一猜哪个转化率高？

A



提升80%
转化率！



B

猜一猜哪个转化率高？

-26%
转化率！

猜一猜哪个转化率高？

原始版本

版本1

版本2

版本3

版本4

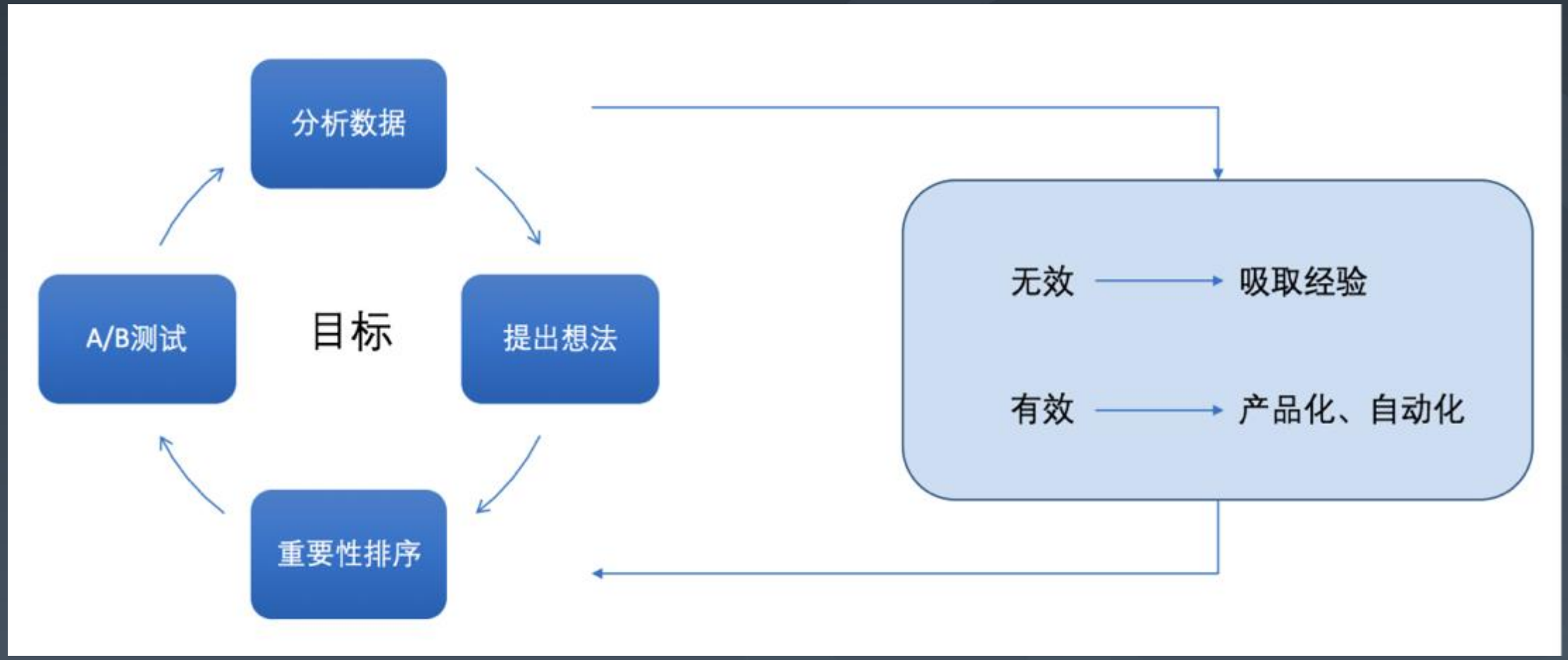
版本5

版本	PV	UV	downloadclick 转化人数 下载按钮被点击的次数	downloadclick 转化率 下载按钮被点击的次数	变化度[95%置信区间]		
原始版本 (3%)	8,978	4,730	220	4.65116279%			
版本一 (2%)	4,139	2,200	116	5.27272727%	+13.36% [-10.51%,+37.23%]	0.27 统计效果不显著	19.52% 功效不足
版本二 (92%)	52,645	27,855	2,032	7.29492012%	+56.84% [+42.36%,+71.32%]	0.01 效果明显优于原始版本	100.00% 功效良好
版本三 (1%)	1,238	658	26	3.95136778%	-15.05% [-49.56%,+19.46%]	0.39 统计效果不显著	13.65% 功效不足
版本四 (1%)	1,255	646	31	4.79876161%	+3.17% [-34.54%,+40.88%]	0.87 统计效果不显著	5.31% 功效不足
版本五 (1%)	4,906	2,595	113	4.35452794%	-6.38% [-27.63%,+14.87%]	0.56 统计效果不显著	9.05% 功效不足

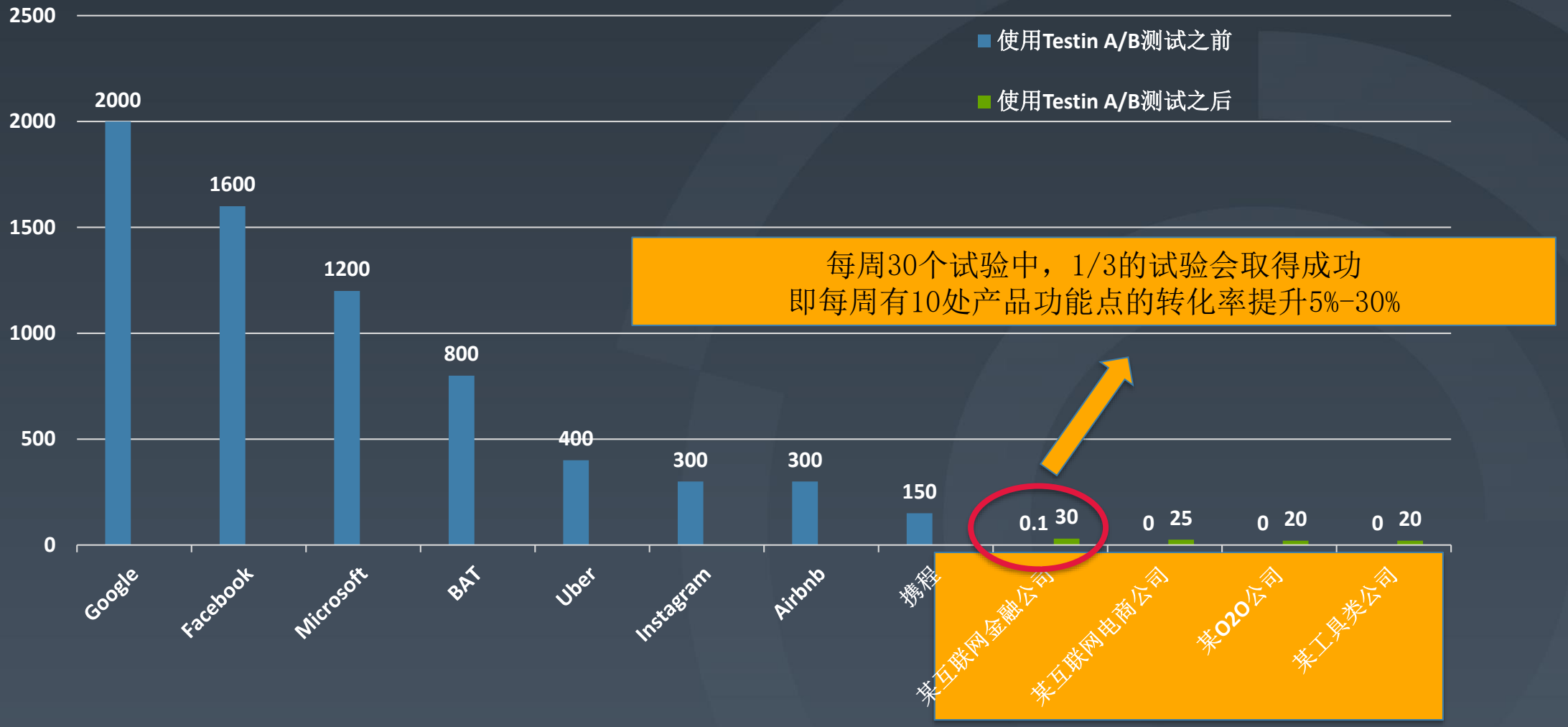
提升56%
转化率！

http://ab.testin.cn/blog/ex_c_10.html

A/B测试如何驱动用户增长？



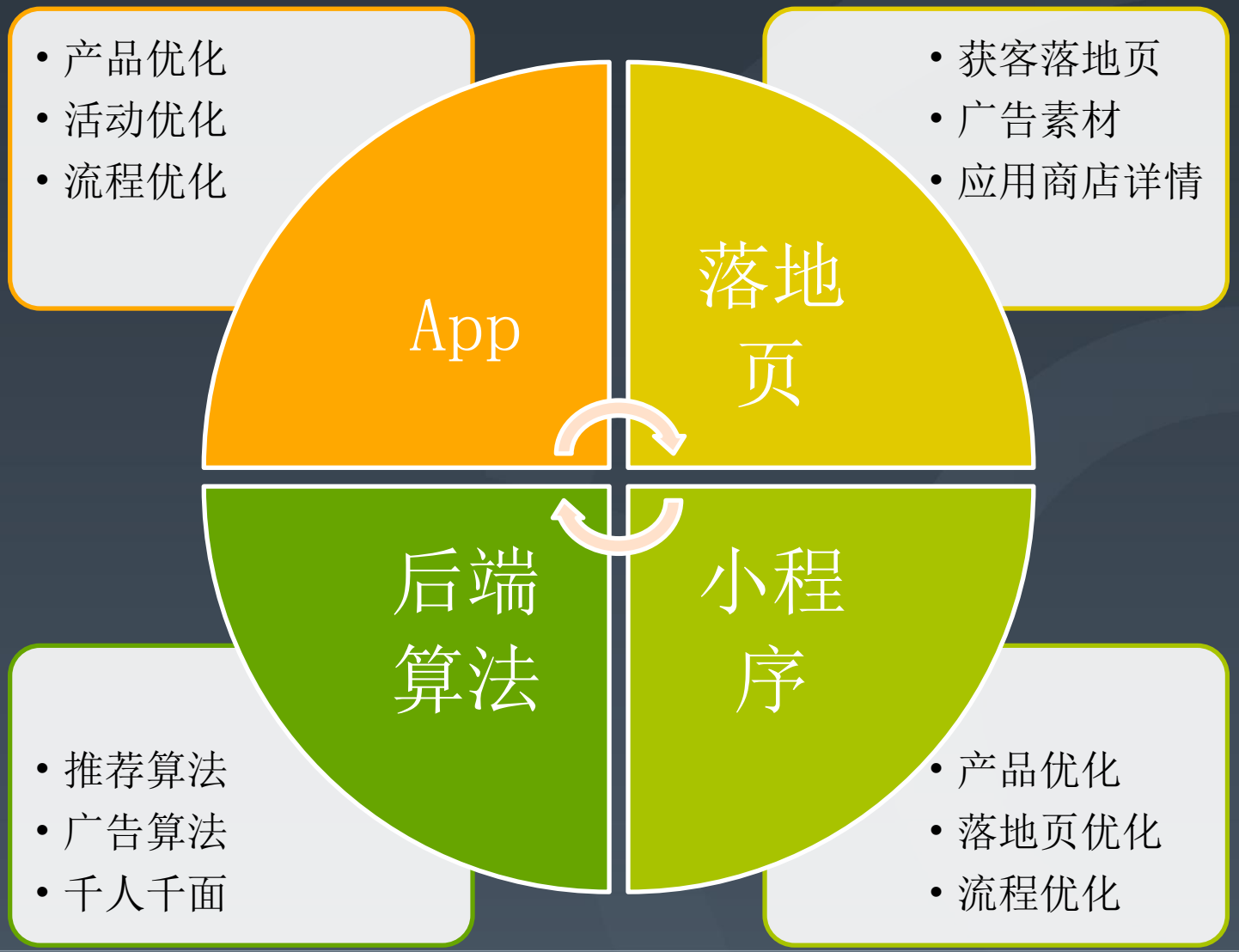
典型公司的A/B测试频率（每周）



目录

- A/B 测试如何助力用户增长？
- 如何在团队中有效推进A/B测试？
- A/B 测试实际案例分享
- 智能优化 & A/B 测试系统技术实践

移动应用A/B测试场景


• 产品优化
• 活动优化
• 流程优化
• 获客落地页
• 广告素材
• 应用商店详情
App
落地页
后端算法
小程序
• 推荐算法
• 广告算法
• 千人千面
• 产品优化
• 落地页优化
• 流程优化

移动应用A/B测试场景

- 获客渠道
- 应用商店海报
- 落地页
- 邀请流程

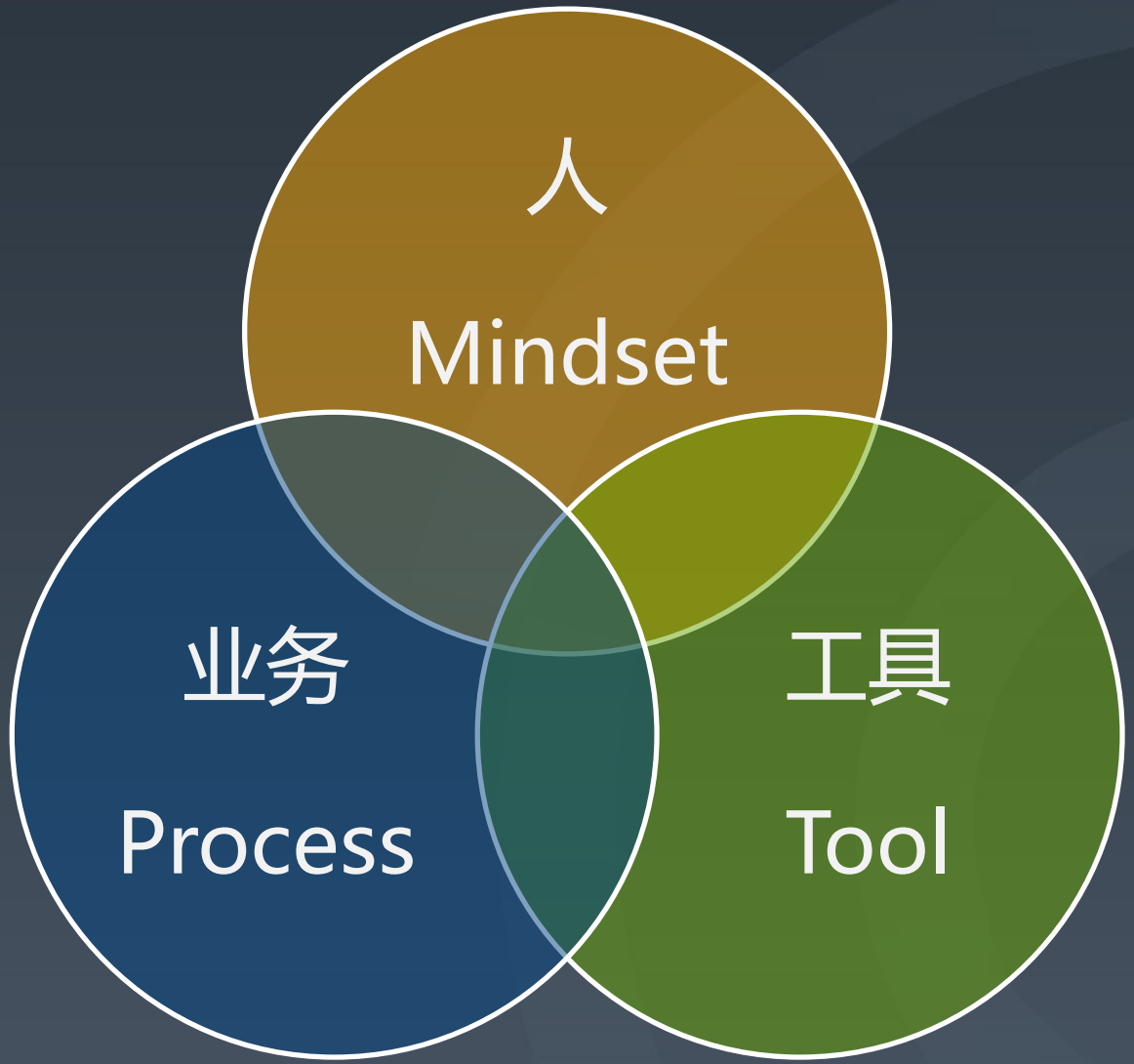
- 注册流程
- 新手引导
- 首页
- Feed流

- 商品详情页
- 补贴策略
- 支付流程
- 裂变机制

产品何时需要A/B测试？

To be or not to be?

如何做A/B测试？



A/B测试的坑在哪里？

好不容易做了一个实验，却没效果

在团队里没人听你的，推不动

研发资源紧张，做一个都费劲，咋做2个版本甚至4个8个？

随机分流，结果看上去挺好，上线后发现不是那么回事

A/B测试咋做啊？不会

结果统计不显著，咋整？

做实验的用户量不够咋办？

从没做过A/B测试，该如何入手？

从最简单的文案A/B测试开始：测试关键按钮中不同文案的转化率

多做团队间的经验分享：有效果的事情大家都愿意尝试

一把手工程：取得CEO认可，自顶向下推动

可以使用第三方的免费A/B测试工具：Testin AB测试

扫码使用Testin AB测试

可视化A/B测试

- iOS/安卓可视化编辑
  - 修改文案
  - 修改背景图片
  - 修改文本颜色
  - 修改字体大小
  - 修改背景颜色
  - 修改透明度
  - 隐藏控件
  - 修改是否可交互
  - 修改字体对齐
  - 修改字体

目录

- A/B 测试如何助力用户增长？
- 如何在团队中有效推进A/B测试？
- A/B 测试实际案例分享
- 智能优化 & A/B 测试系统技术实践

如何完整进行一次A/B测试？

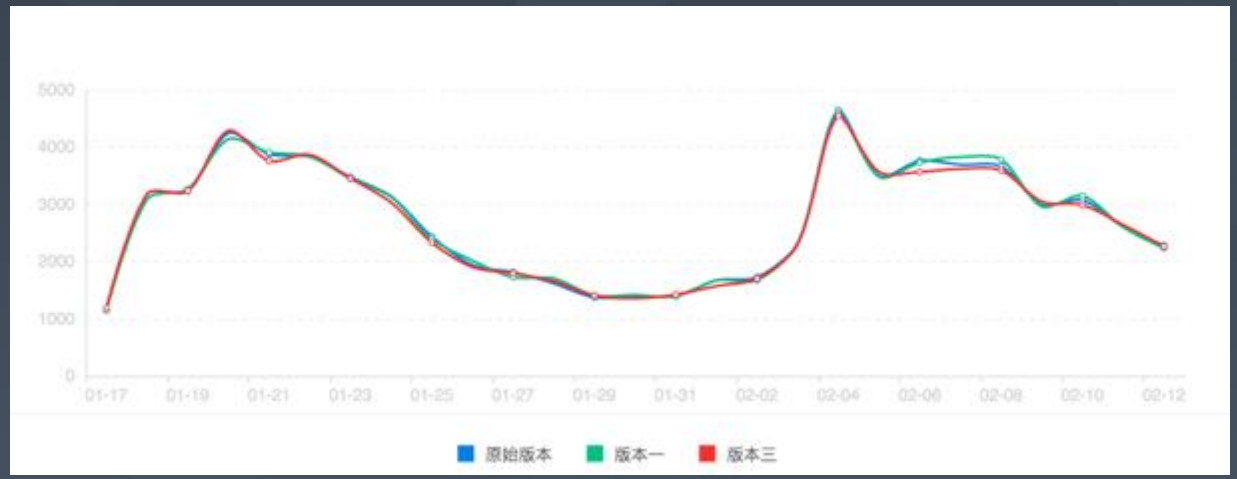
1. A/B测试目标：
将分享按钮点击量提升5%

2. 分析问题：
现在这个按钮人均点击次数是0.2次，通过改变“立即分享”这个文案，是否能有效地传递产品意图，从而提升分享按钮的点击量？

3. 产生新想法
再换几个更简单直接的文案

4. 排优先级
分享的数据指标对拉新很重要，而且开发成本低，先从H5页面上线

5. 进行”A/B测试”
每个版本每天分配约4000用户（各33%流量），运行2周，观察数据



如何完整进行一次A/B测试？（2）

1. A/B测试目标：
将分享按钮点击量提升5%

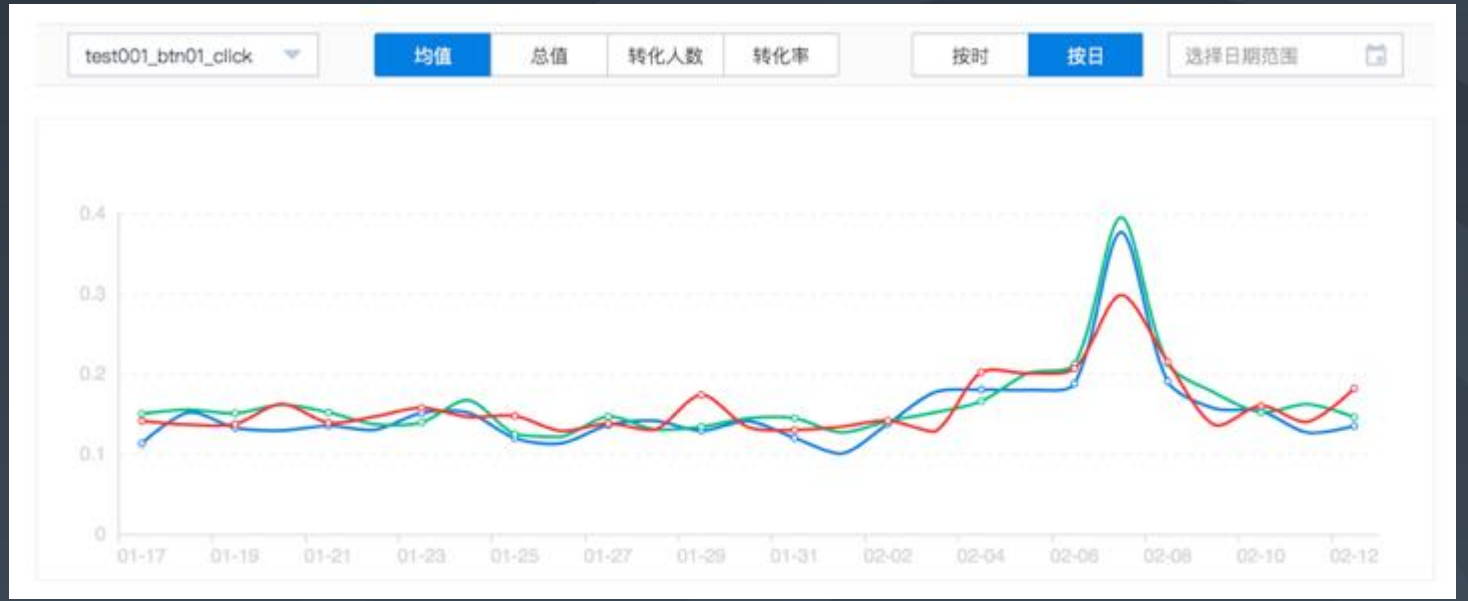
2. 分析问题：
现在这个按钮人均点击次数是0.2次，通过改变“立即分享”这个文案，是否能有效地传递产品意图，从而提升分享按钮的点击量？

3. 产生新想法
再换几个更简单直接的文案

4. 排优先级
分享的数据指标对拉新很重要，而且开发成本低，先从H5页面上线

5. 进行”A/B测试”
每个版本每天分配约4000用户（各33%流量），运行2周，观察数据

6. 分析”A/B测试数据”



统计分析

版本	PV	UV	test001_btn01_click总值	均值	变化度[95%置信区间]	p-value
原始版本	123234	52227	11703	0.2241		
版本一	121894	52366	12574	0.2401	7.14%[1.25%,13.07%]	0.01 效果明显优于原始版本
版本三	120550	52094	11967	0.2297	2.50%[-3.06%,8.09%]	0.19 统计效果不显著

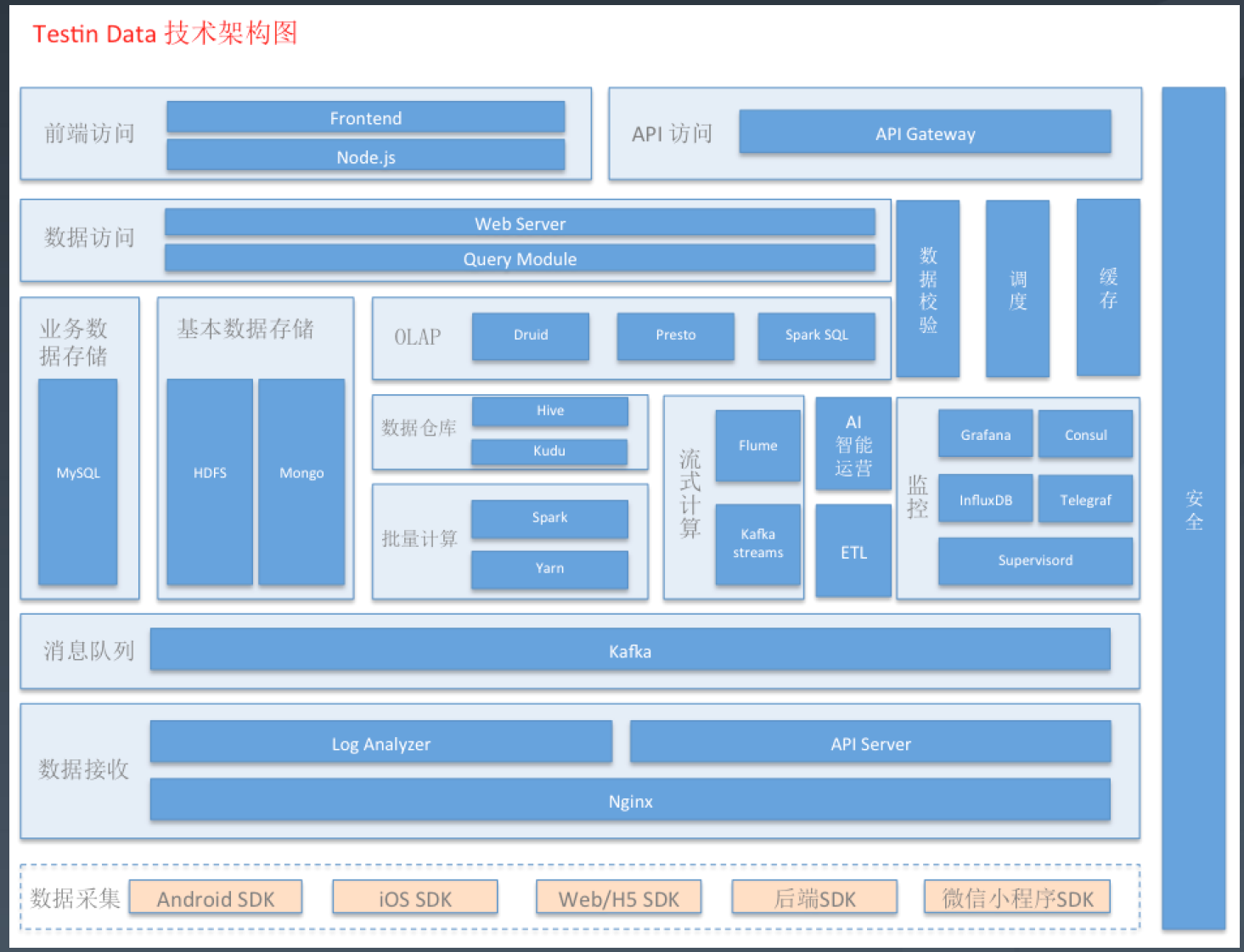
“我们起初做了一些小实验，并没有发现有明显的结果出现。后来在一次讨论中，基于我们产品经理的经验与直觉，第一判断【联系客服】按钮的位置可能有些隐蔽，不容易被用户看到。我们第一个实验选取了对【联系客服】按钮的位置调整进行A/B测试，原始版本与实验版本分别如下截图所示的位置：”

“实验进行约3周时间，我们可以看到新版本对原始版本提升了37.66%的转化率，同时该转化率在95%的置信区间是可信的。”
http://ab.testin.cn/blog/interview_leyou_1.html

目录

- A/B 测试如何助力用户增长？
- 如何在团队中有效推进A/B测试？
- A/B 测试实际案例分享
- 智能优化 & A/B 测试系统技术实践

Testin云测A/B测试整体系统架构

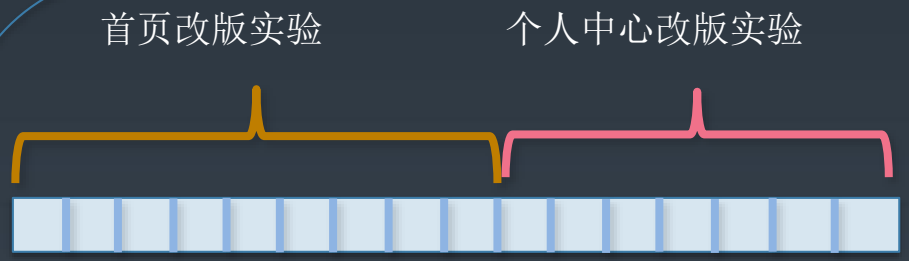


- 10亿API调用/天
- 基于Druid/Presto的查询引擎
- 基于强化学习的智能优化算法
- 支持安卓/iOS/Web/H5小程序/后端 A/B测试

A/B测试之科学流量分割

- 唯一性
  - » 通过精准且高效的Hash算法，确保单个用户每次登录应用时被分到的试验版本是唯一的
- 均匀性
  - » 分流的人群，各维度分配比例均匀
- 灵活性
  - » 用户可以随时在试验的进行过程中调节试验版本之间的流量分配比例
- 定向性
  - » 可以根据用户标签来实现精准定向分流
    - 设备标签（SDK自动采集）
    - 用户标签（需要上传）
- 分层分流
  - » 满足并行做大量A/B测试的需求

为什么需要分层流量分割机制？

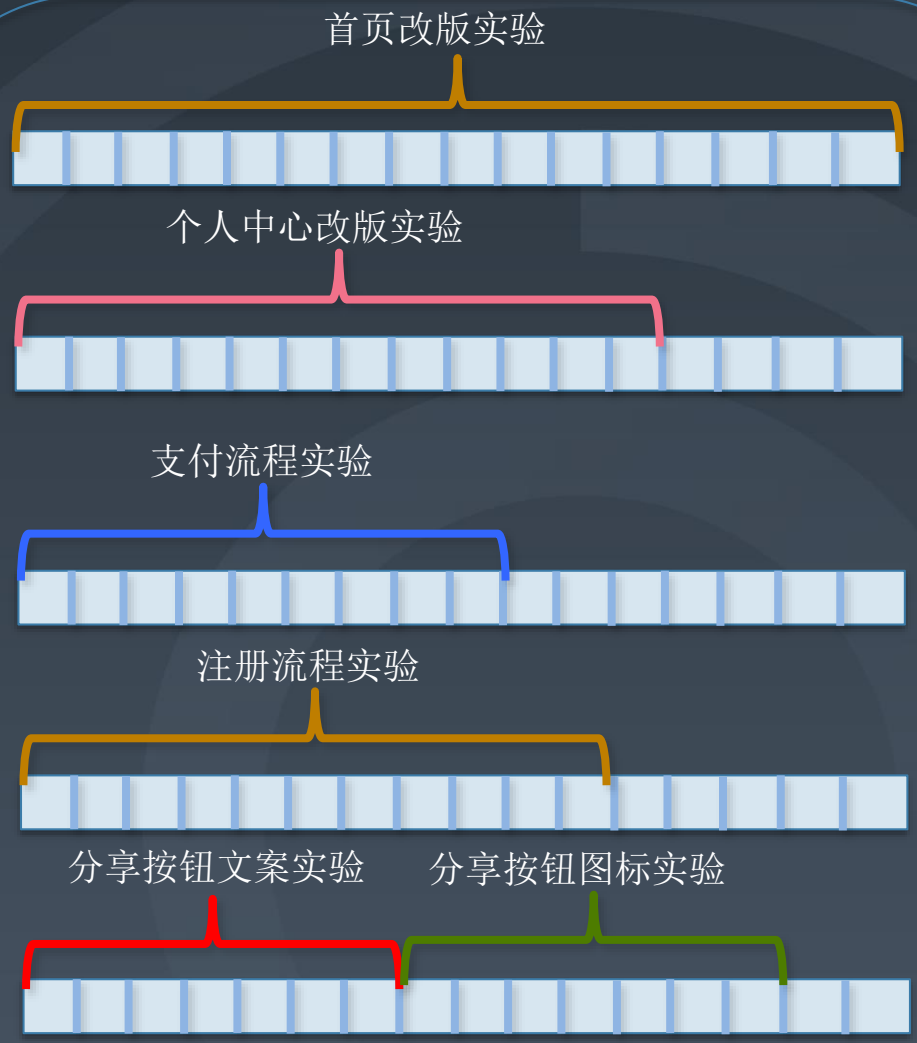


排队等待的实验：

- 支付流程实验
- 注册流程实验
- 分享按钮实验

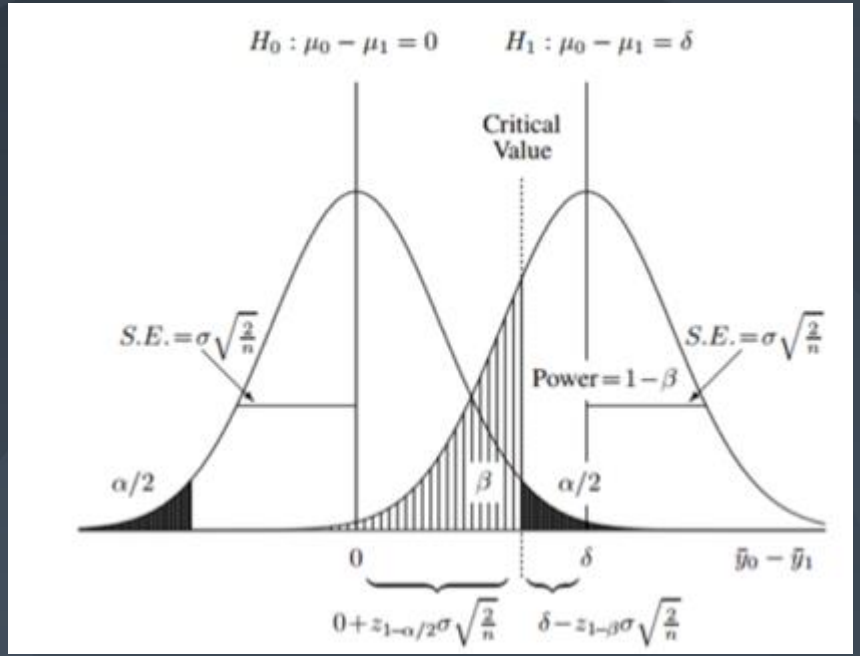
- 没有分层流量机制时存在的限制：
  - 每个用户最多只能参加一个A/B测试实验
  - 多个实验不能同时使用全体用户进行测试，可能因为人群覆盖度不够高导致结果偏差
  - 每个实验的可用实验流量受限于其他正在进行的实验，缺乏灵活的流量分配机制

VS



A/B测试之科学统计算法

- 科学统计
  - » 使用了科学的统计分析方法来对试验数据进行分析，并给出可靠的实验结果
- 区间估计
  - » 使用区间估计，给出了95%置信区间，因此避免了点估计带来的决断风险
- 统计显著性判断
  - » 通过p-value来判断不同实验版本之间差异显著性
- 统计功效判断
  - » 通过Power来判断不同实验版本统计功效是否充足
- 精益分析
  - » 对试验数据进行去噪音处理，去除噪音数据，以提高统计结果的质量



A/B测试系统关键技术点

分流准确性

- 用户OS版本
- 浏览器版本
- 设备型号
- 应用版本
- 屏幕尺寸
- 系统语言
- SDK版本
- 用户自定义标签
- 分层分流

统计结果可信度

- 平均值
- 方差
- 抽样误差
- 正态分布
- 置信区间
- 统计显著性

可视化编辑埋点

- UI控件识别与编辑
- 不同机型兼容性保障
- 标准UI控件
- 自定义控件
- 多层UI控件
- UI控件树

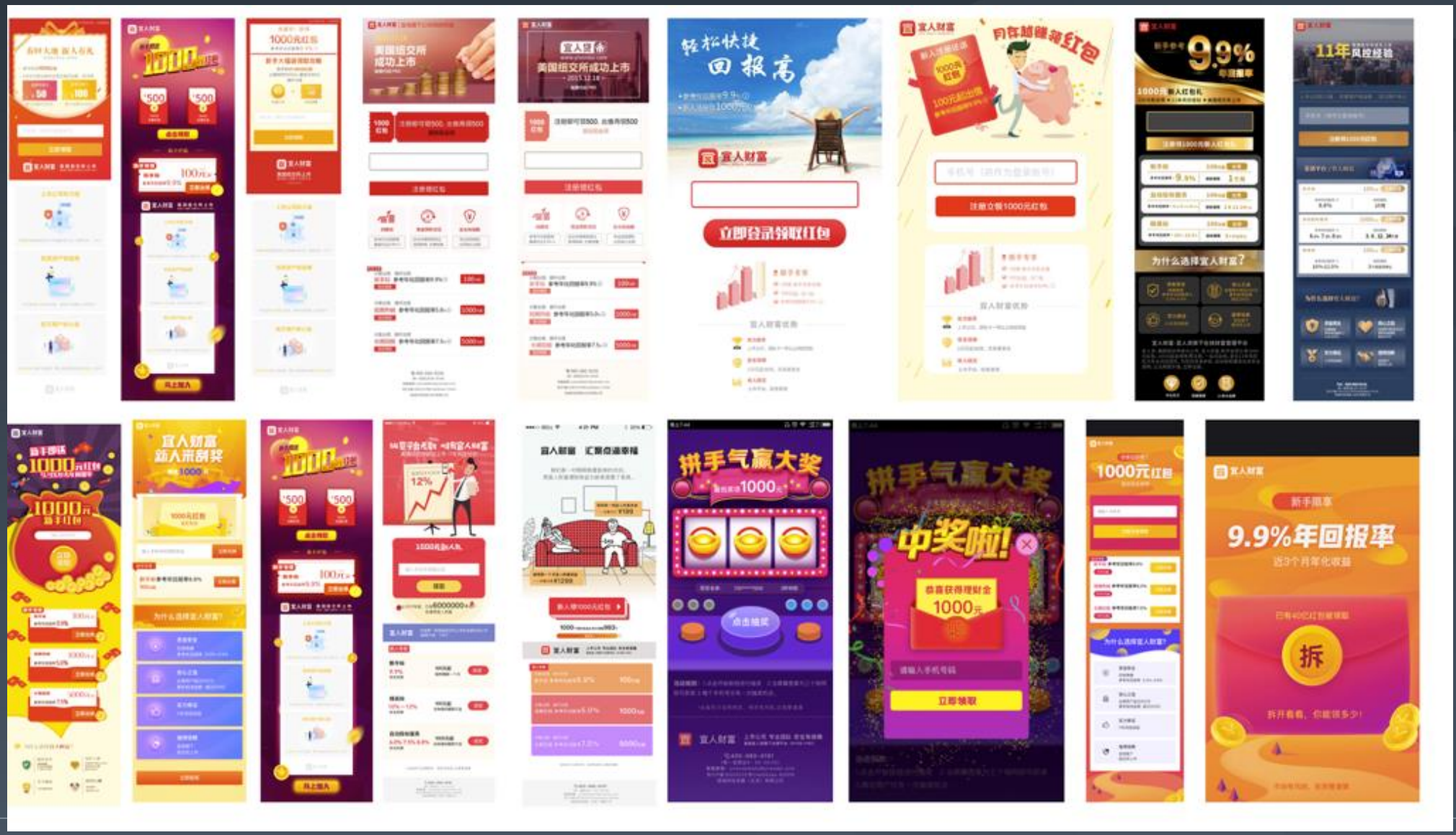
SDK稳定性

- 在SDK发版流程中引入Testin的兼容性测试和功能性测试
- iOS 70款机型兼容性
- Android主流600款机型兼容性
- Web主流浏览器兼容性
- 20KB+ JavaScript SDK
- 350KB+ iOS/Android SDK
- 数据本地聚合减少数据量
- Https通信协议

数据后台稳定性

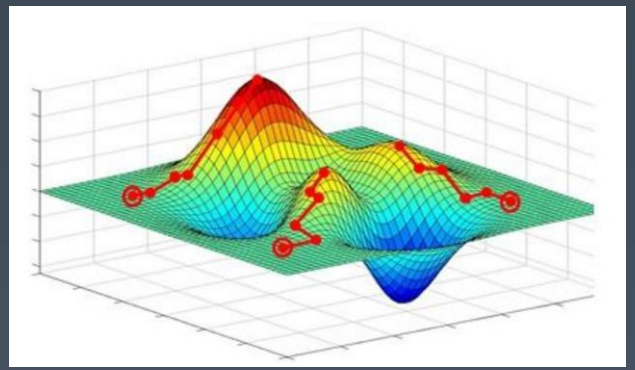
- 数据实时摄入
- 数据实时计算、展示
- 数据准确性
- 数据完整性
- 支持千万级日活
- 可扩展性
- 高可靠性

如何对多渠道多版本做“智能“A/B测试？



基于强化学习的智能优化引擎 （业界首家）

- 基于上下文&多臂赌博机Contextual & Multi-armed Bandit，专利[2017103029700自动生成多页面组合版本，2017102790713自动版本选优和流量调控]等强化学习算法实现自动智能选优
- 不需监督，7*24小时智能版本选优、分流调整，保持运营水准持续提升；
- 行业领先App复杂场景、大用户样本持续训练、迭代进化；
- 历史数据智能分析比较，持续增长；
- 连续决策，挑选动作行为来最大化将来的累计回报，牺牲立即回报来获得更多的长期回报；
- 人工辅助调整接口



极客时间 | 企业服务

这是一封正经“福利”贴

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养 企业转型加快，团队能力又跟不上	VS	从基础到进阶，超100+一线实战 技术专家带你系统化学习成长
团队成员技能水平不一， 难以一“敌”百人需求	VS	解决从小白到资深技术人所遇到 80%的问题
寻求外部培训，奈何价更高且 集中式学习	VS	多样、灵活的学习方式，包括 音频、图文 和视频
学习效果难以统计，产生不良循环	VS	获取员工学习报告，查看学习 进度，形成闭环

课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

 # 极客时间企业账号 # 解决技术人成长路上的学习问题

Geekbang 极客邦科技 | TGO 鲲鹏會

TGO鲲鹏会

汇聚全球科技领导者的高端社群

全球12大城市

850+高端科技领导者

使命 Mission

为社会输送更多优秀的科技领导者

愿景 Vision

构建全球领先的有技术背景优秀人才的学习成长平台

扫描二维码，了解更多内容

THANKS!