

A/B 测试中 20 个必须知道的问题



掘金官方 (/u/5fc9b6410f4f) [+ 关注](#)

2016.06.02 20:21 字数 6996 阅读 1842 评论 0 喜欢 13

(/u/5fc9b6410f4f)

本文来自 Topgeek 投稿。

在网站和移动产品设计和开发中、以及互联网产品运营中，我们经常会面临多个产品设计和运营方案的选择，比如某个按钮是用红色还是用蓝色，是放左边还是放右边。传统的解决方法通常是集体讨论表决，或者由某位专家或领导来拍板，实在决定不了时也有随机选一个上线的。虽然传统解决办法多数情况下也是有效的，但 A/B 测试 (A/B Testing) 可能是解决这类问题的一个更好的方法。

在软件开发中，产品需求通过多种技术手段来实现; A/B 测试实验提供了一个有价值的方式来评估新功能对客户行为的影响。

在运营过程中，AB 测试用得更加普遍，比如发送邮件或者广告，先拿小样本，测试多个版本，数据表明哪一个广告或邮件的转化率高，就用哪一个邮件或广告。

1 什么是 A/B 测试？

A/B 测试是一种流行的网页优化方法，可以用于增加转化率注册率等网页指标。简单来说，就是为同一个目标制定两个方案（比如两个页面），将产品的用户流量分割成 A/B 两组，一组试验组，一组对照组，两组用户特点类似，并且同时运行。试验运行一段时间后分别统计两组用户的表现，再将数据结果进行对比，就可以科学的帮助决策。比如在这个例子里，50% 用户看到 A 版本页面，50% 用户看到 B 版本页面，结果 A 版本用户转化率 23%，高于 B 版本的 11%，在试验流量足够大的情况下，我们就可以判定 A 版本胜出，然后将 A 版本页面推送给所有的用户。



Paste_Image.png

AB测试本质上是个分离式组间实验，以前进行AB测试的技术成本和资源成本相对较高，但现在一系列专业的可视化实验工具的出现，AB测试已越来越成为网站优化常用的方法。

A/B测试其实是一种“先验”的实验体系，属于预测型结论，与“后验”的归纳性结论差别巨大。A/B测试的目的在于通过科学的实验设计、采样样本代表性、流量分割与小流量测试等方式来获得具有代表性的实验结论，并确信该结论在推广到全部流量可信。

2 有天然存在的AB测试吗？

A/B 测试并不是互联网测试新发明的方法，事实上，自然界也存在着类似 A/B 测试的事件，比如下图中的达尔文雀。



Paste_Image.png

达尔文雀主要生活在太平洋东部加拉帕戈斯（Galapagos）的一个名为伊莎贝拉（Isabela）的岛上，一部分生活在岛的西部，另一部分生活在岛的东部，由于生活环境的细微不同它们进化出了不同的喙。这被认为是自然选择学说上的一个重要例证。

同一种鸟，究竟哪一种喙更适合生存呢？自然界给出了她的解决方案，让鸟儿自己变异（多个设计方案），然后优胜劣汰。具体到达尔文雀这个例子上，不同的环境中的也有不同的解决方案。

上面的例子虽然和网站设计无关，但包含了 A/B 测试最核心的思想，即：

- 1、多个方案并行测试；
- 2、每个方案只有一个变量（比如鸟喙）不同；
- 3、以某种规则优胜劣汰。



×

下载简书App
创作你的创作

分享图标

二维码图标

(/apps
/download?utm_source=stc)

需要特别留意的是第 2 点，它暗示了 A/B 测试的应用范围，——必须是单变量。

3 什么情况不适合做 A/B 测试？

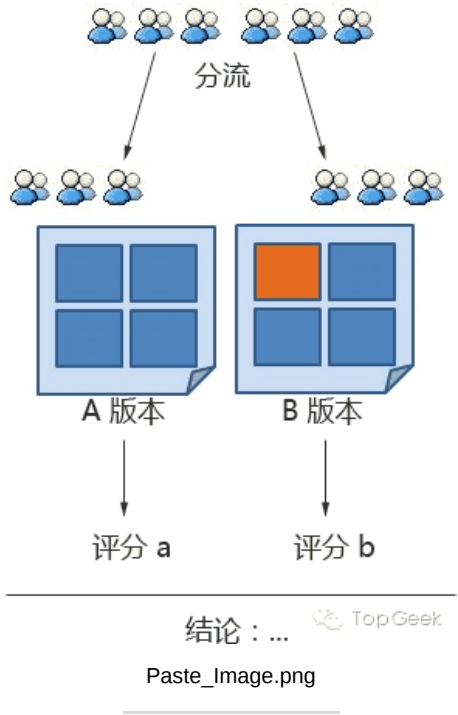
有时多个设计稿可能会有非常大的差异，这样的情况一般不太适合做 A/B 测试，因为它们的变量太多了，变量之间会有较多的干扰，很难通过 A/B 测试的方法来找出各个变量对结果的影响程度。

另外，虽然 A/B 测试名字中只包含 A、B，但并不是说它只能用于比较两个方案的好坏，事实上，你完全可以设计多个方案进行测试，“A/B 测试”这个名字只是一个习惯的叫法。

回到网站设计，一般来说，每个设计方案应该大体上是相同的，只是某一个地方有所不同，比如某处排版、文案、图片、颜色等。然后对不同的用户展示不同的方案。

要注意，不同的用户在他的一次浏览过程中，看到的应该一直是同一个方案。比如他一开始看到的是 A 方案，则在此次会话中应该一直向他展示 A 方案，而不能一会儿让他看 A 方案，一会儿让他看 B 方案。同时，还需要注意控制访问各个版本的人数，大多数情况下我们会希望将访问者平均分配到各个不同的版本上。要做到这些很简单，根据 cookie（比如 cookie 会话ID的最后一位数字）决定展示哪个版本就是一个不错的方法。

下面是 A/B 测试示意图：



可以看到，要实现 A/B 测试，我们需要做以下几个工作：



×

下载简书App
创作你的创作

(/apps/download?utm_source=stc)




- 1、开发两个（或多个）不同的版本并部署；
- 2、收集数据；
- 3、分析数据，得出结果。

4 哪些公司在做AB测试？

A/B测试如同GitHub、Docker、APM一样在美国市场已经被各类企业逐渐采用，比如Google、Airbnb等。

其测试范围也不仅仅局限于网页优化，目前移动端的A/B测试需要同时支持前端（Web/H5、iOS、Android）及后端（Node.js、PHP、Java）。

5 什么阶段的公司适合做AB测试？

AB测试你自己做是要花很大的人力、物力，大公司有很大的用户，做AB测试的话，是可以持续投入的，每个投入的提升增长价值也很大，是公司中最为重要的。

很多中小型的公司具备条件，但不一定有经验或能力执行和分析，不过现在也有些第三方服务公司提供了工具，方便做AB测试，降低了门槛，比如吆喝科技（<http://www.appadhoc.com>）在这方面做得非常好。（<http://www.appadhoc.com%E5%BC%89%E5%9C%A8%E8%BF%99%E6%96%B9%E9%9D%A2%E5%81%9A%E5%BE%97%E9%9D%9E%E5%B8%B8%E5%A5%BD%E3%80%82>）

初创公司，在产品还没验证的时候，或者用户量非常小的时候，不适合做AB测试。



Paste_Image.png

6 如何利用A/B测试做增长？

AB 测试是撬动理性增长的最重要工具之一。AB测试背后的理念是在于用数据来帮助你做决策，来帮助你做更好的决策，很多东西就不再是靠艺术创造、靠想象、靠拍脑袋来做，而是靠数据，像你写代码、做分析的时候那样一种很理性的模式。

7 A/B测试的数据结果出来后，应该怎么样选择？

从数据结果分析客观的效果，但往往也需要根据用户体验和总收益做一个折衷。

《增长黑客》作者范冰讲过一个百姓网的案例。百姓网之前有段时间销售员和产品经理撕逼，销售人员是觉得为了获得更多销售额，所以我们必须是用户给钱越多，我们给他越大的特权。



A/B测试是撬动理性增长的最重要工具之一

Paste_Image.png

他们想像左图这样，用户在我这个平台上发布的小帖子以后，谁给的钱多，给得最多的我给你置顶，同时又给你一个广告位，就是红色标量，其他的给钱没那么多的，在相对置顶比较高的位置。就是你越给钱，我越给你一些标签把你位置提得越高，这是销售人员的思维；产品经理的思维是右边这种，虽然你给了钱，你是我们的金主，你很重要，但是我要重视我们的产品体验，如果说你给钱我就让你上去的话，其实这上面满眼看得都是广告，而且谁给钱谁就上，那就有点像百度了，像现在这个样子他们就提出我们的产品在右边，不管你给了多少钱，我最多就给你个高亮，所以你的位置我不控制，当时为了这个原因，双方激烈的撕逼。撕逼一般是没有结果的，因为公说公有理，婆说婆有理，后来他们想到组织一次 AB 测试，下发了两波用户，看这两波用户看到两种不同的页面，哪波用户最后转化率高，带来的收入高，还有其他一些指标。

结果是怎样的？

说 x

各自呈现两

的作你的创作

下载简书App

(/apps

/download?utm_source=stc)

大家从直觉判断，一定觉得产品经理的决定是对的，最后一定是用了产品经理的方案。

测试结果，右边产品经理方案是好的，他的数据更高，但是最后用了左边的方案。为什么？因为测试结果反馈显示，这两个方案虽然右边更好，但是他这个好的方案只是精确到小数点后面的千分位，只是比前一种方案好了那么一点，虽然是好了那么一点，但是左边的方案更吸金，左边的能吸引到大家更多的往里投钱，更多的花钱，既然是只好了这么一点点，当然要用左边的。于是经过测试以后，他们最后用了左边的方案，这是大家没有想到的结果。

因为 AB 测试固然重要，AB 测试的结果的确右边好，但是有的时候要结合实际，如果说差别不是很大的话，你可能要选一种赚钱更多的方式，这是 AB 测试一个很大的价值，大家不要偏信数据，不要被数据给完全左右，有的时候结合一些你的理性的思考。

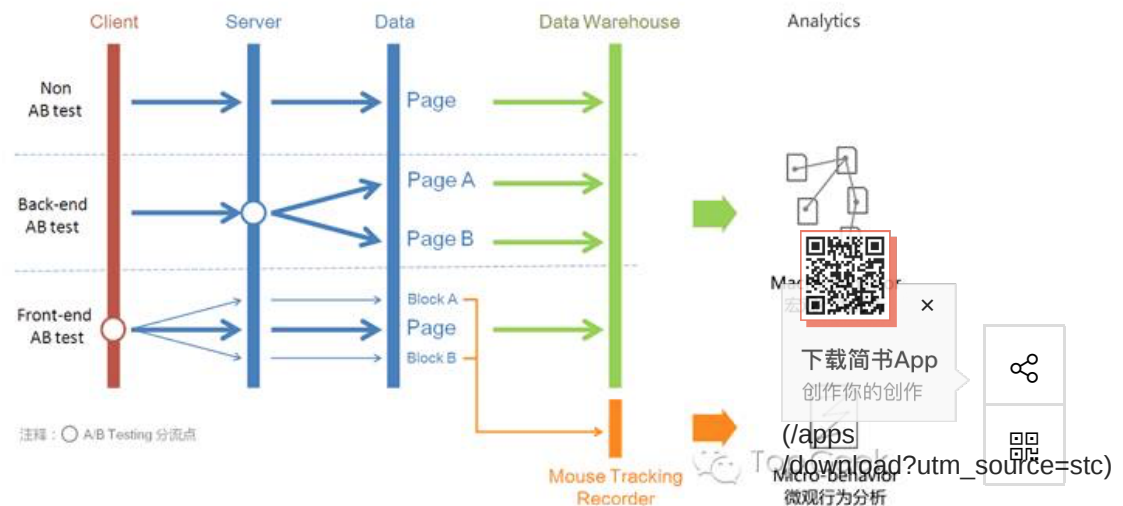
1. AB测试的具体实施流程是什么样的呢？

其实非常简单，可能在座的大家都有一定地印象，比如说你有一个网页，就是你用户流量的页面，你用上面的绿色作为一个代表，如果你现在的页面只有23%的转化率，你希望通过某种黑客方式，把它调一调，改一改，改成蓝色的页面，希望能够提升它的转化率。

那么可以用AB测试怎么做呢？就让来访流量的访客，一半或者一部分的访客看到绿色的老版本，一部分的访客看到蓝色的新版本，这些用户自己是无感知的，他们并不知道自己是分配在实验里面，他们依然按照自己的行为去操作，他们会买东西，会退出或者怎么样，然后你看他的转化率，有没有发生变化，假如我们看到一个很糟糕的现象，这个蓝色的版本，它的转化率反而降低了只有11%了，结果你的老版本还胜出的，就说明你改进的方案不成功，于是你会想其他的方案再去改，总会找到能够提升转化率的方法。

。

A/B testing 部署概念图



Paste_Image.png

9 有哪些AB测试需要注意的经验或规则？

1. 效果惊人，一些很微小的改动，它就可能造成对你KPI巨大的影响。
2. 大多数改动都不会带来大幅度提高KPI，所以你需要耐心。
1. Twyman法则，他是凡是看上去很出人意料的图表，通常都是因为数据统计错了。
 4. 各个产品几乎都不一样，你复制他人的经验，往往都没有什么效果。
 5. 任何能加速用户响应时间的改动，都会带来KPI的正向提升。
 6. 点击率是很容易提高的，但是流失率是很难改进的，千万不要把精力放在优化某个页面点击率上。
 7. 尽量不要做很复杂的大量改动的实验，而是要做很简单的小的迭代。

10 Facebook在增长过程中怎样使用A/B测试？

据前Facebook工程师，现峰瑞资本技术合伙人覃超介绍，facebook做增长的流程为四步：

1 设计关键数据面板 2 关注核心动作 3 发现增长规律和模式 4 灰度发布和AB测试

具体在灰度发布和A/B测试分为以下步骤：

- 1 计划：根据新功能制定改版计划；
 - 2 预期：数据会如何变化；
 - 3 设置多版本：逐步开放给用户；
 - 4 清除：清除老的版本。
- 6个月内所有版本完全线上灰度发布，通过不断进行用户流量分割的方式进行实验，获得无Bug口碑。

11 还有什么领域也用AB测试？

对照实验，也叫随机实验和A/B测试，曾在多个领域产生深远的影响，其中包括医药，农业，制造业和广告。

通过随机化和适当的实验设计，实验构建了科学的因果关系，这就是为什么对照实验（A/B测试）是药物测试的最高标准。

正是考虑到后验方法的局限性，西医（现代医学科学）首先引入了 A/B 测试来验证新药的疗效。新药的验证可能是这样一个流程：100 位患者，被测试医生分为 AB 两组，注意患者自己并不知道自己被分组，注意 AB 两组患者的健康情况是一致的；A 组患者将会得到试验新药，B 组患者将会得到长的和新药几乎一样的安慰剂；如果最终 A 组患者比 B 组的疗效更好，才能证明新药的药效。



实验x
下载简书App
创作你的创作



/download?utm_source=stc)

12 A/B测试的价值是什么？

AB测试的实验能力可以用更科学方法来评估规划过程中不同阶段的想法价值。

A/B测试其实是一种“先验”的实验体系，属于预测型结论，与“后验”的归纳性结论差别巨大。A/B测试的目的在于通过科学的实验设计、采样样本代表性、流量分割与小流量测试等方式来获得具有代表性的实验结论，并确信该结论在推广到全部流量可信。

通过值得信赖的实验来加速创新。通过解决技术和文化的挑战，我们给软件开发人员、项目经理和设计师一副“公正的耳朵”，帮助他们听取客户真实的诉求以及用数据驱动的决策。

13 A/B测试的应用场景有哪些？

A/B测试这种方法论的应用非常广泛，包括在Web产品、移动产品、数字广告优化领域的应用。应用场景由小到大可以可以分为：

元素/控件层面

功能层面

产品层面

公司层面

14 A/B测试中需要用到的基本概念有哪些？

样本空间、样本特征、实验流量

假定这是个电商的APP，产品有100万用户

样本空间：100万用户

样本特征：这100万用户有各式各样的特点（性别、地域、手机品牌与型号、甚至是不是爱点按钮等行为。。）

实验流量：100万用户成为100%的流量；假定将这100万用户根据样本特征与相似性规则分为100组，那每组就是1万人，这1万人就是1%的流量

采样、代表性误差、聚类

相似性采样：在A/B测试的实验中，需要保证小流量的实验具备代表性，也就是说1%的流量做出来的实验结果，可以推广到100%的用户，为了保证这一点，需要保证1%的流量的样本特征与100%流量的样本特征具备相似性。（说个最简单的逻辑：假设小米手机用户均匀的分到这100组中，那第一组的所有小米手机用户的特征与第100组的所有小米手机用户具备相似性）

代表性误差：代表性误差，又称抽样误差。主要是指在用样本数据向总体进行推断时所产生的随机误差。从理论上讲，这种误差是不可避免的，但是它是可以计算并且加以控制的。（继续小米。。尽管把小米用户均匀的分成了100组，但是不能完全保证每个组里



下载简书App
创作你的创作



的小米用户的数量、性别、地域等特征完全一样，这就带来了实验误差风险)

聚类：物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类，也就是在分配小米用户的过程中，需要按照实验目的的不同把特征相似性高的用户认为是一类用户，比如定义100次点击为高频点击，可能在某些情况下也会认为99次点击的用户跟100次点击的用户是一类用户。

置信度与置信区间

在统计学中，一个概率样本的置信区间 (Confidence interval) 是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度。置信区间给出的是被测量参数的测量值的可信程度，即前面所要求的“一定概率”。这个概率被称为置信水平。

置信度：简单来将表示可信程度，一般来说95%的置信度就很好了，一些及其严苛的A/B测试实验才会到99%的置信度。差别在于，越高的置信度得出结论的实验时间越长、流量要求越高。

置信区间：从前面的概念中也讲了，1%的流量尽管具备了代表性，但是跟100%的流量还是有差异的嘛，所以实验结果的评判要有一定的前提的，置信度就是这个前提，置信区间表示在这个置信度的前提下，实验结果很可能会落在一个区间内，比如下图，95%的置信度的前提下，置信区间为 $[-2.3\%, +17.4\%]$ ，可以解读为这个A/B测试的实验既有可能使“点击次数”降低2.3%，又有可能提升17.4%。说明这个实验结果还不稳定，可能是试验时间短或者是流量不够。

15 数据化驱动决策与确定性提升是什么意思？

数据化驱动决策：A/B测试是典型的靠谱数据化驱动决策，先用A/B测试的方式，让比如1%或者5%的用户进行实验，让用户用实际的行为来告诉你哪个好。比如这1%或者5%的用户通过“点击次数”这个指标告诉你，他们不喜欢橙色的设计。这就是数据化驱动决策，不用一屋子人你拍桌子我瞪眼的争辩到底那个设计好，让真实的用户跟数据告诉你到底哪个更好。

确定性提升：这就更好解释了，有了这么个工具，每次只有效果好了才会上线，也就意味着每次优化都能比以前更好，大大提高用户的体验和产品经理的自信心。

16 Airbnb的产品是怎么样做AB测试的？

Airbnb经常用灰度发布和A/B测试对重要页面的修改和流程上的调优，通过1%或者5%的用户，看其实际对用户的数据影响（访问时间增加、留存提高、转化率提高等），决定此修改到底是100%发布还是被砍掉。

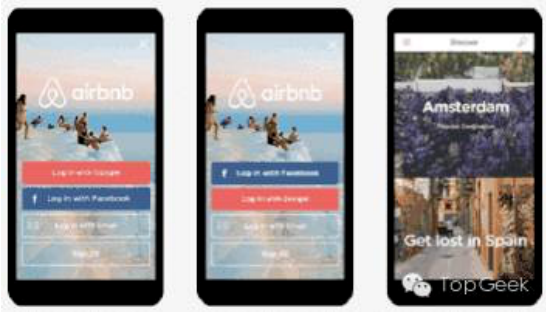


下载简书App

创作你的创作



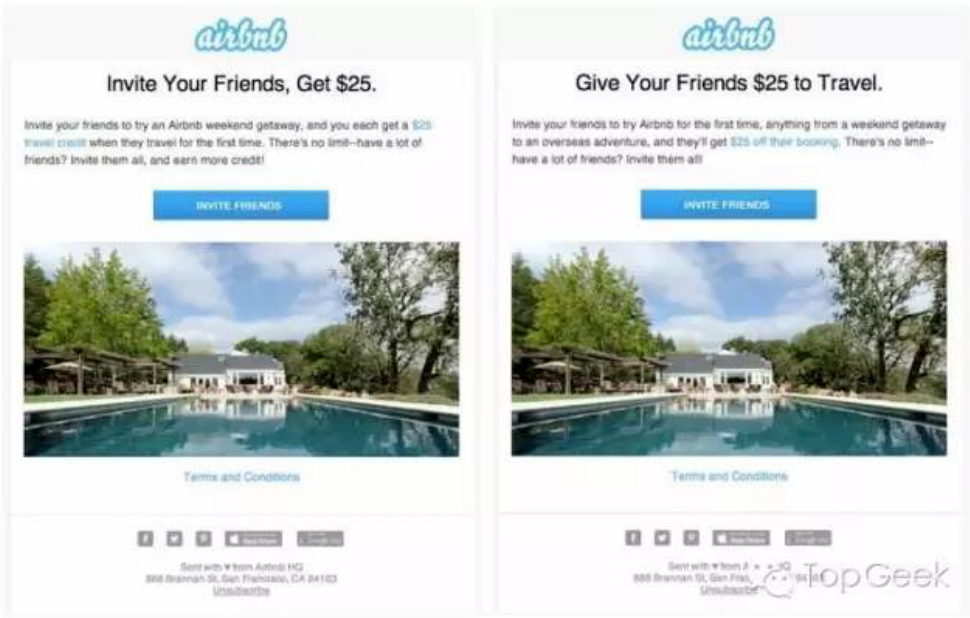
/download?utm_source=stc)



Paste_Image.png

Airbnb 从第一天就开始做 A/B 测试，不仅在自己的体系里做，还用第三方工具做，保证所有的决策，从产品，到运营，乃至到战略，都是经过数据驱动的优化决策。每一个改动，都先用 1% 的流量来试验，然后再推到 5%，再到 10%，到 20%，到 50%，最后再发布给所有用户。

通过A/B测试，他们还有一个关于推介文案的结论：给用户展示“利他”的文案，比“利己”的更容易带来转化。如图所示，告诉用户“邀请好友可以获得25美元”的效果就不如“给你的好友赠送25美元的旅行经费”更打动人。



Paste_Image.png

17 Google是怎么样做AB测试的？

Google每个月从上百个A/B测试中找到十几个有效方案，月营收提升2%左右，有app 10亿美元规模的。很难解释的是广告位左移一个像素带来X%的增收，左移两个像素带来Y%的亏损。



下载简书App
创作你的创作



在Google，任何产品改动需要A/B测试才能上线。

Google X 生命科学分部的负责人 Andy Conrad 在《财富》的一篇文章中曾提到：

对于一个问题 Larry 会尝试用 1、2 种办法去解决，并且在策略上会对两者都同时下注。

Google 几乎所有的产品目录似乎都要进行大型的 A/B 测试。正如 Google 的搜索引擎不断从 Web 上收集数据加以学习和改进一样，Google 公司本身也是这么运作的。它给单个问题提供了多个解决方案，希望能从中决出优胜者。

这种多产品策略对于 Google 的长期健康来说是好的，但它也浪费了许多资源。到处都是重复的工作，但 Google 的 Adsense 和 Adwords 带来了那么多的收入，至少现在 Google 挥霍得起。

Google 往往喜欢针对同一客户群推出多项竞争产品。这样的话，如果一个产品失败了，也许另一个产品能够补上。最极端的例子是 Google 的即时通信解决方案。Android 上一度曾出现过 4 款不同的产品：Google Talk、Google+ Messenger、Messaging（Android 的短信应用）以及 Google Voice。Google Hangouts 最终胜出，把其他的都合进了一个平台。

Google 平时就是这样折腾的。其行动表明，自己并不相信一个问题只有一种解决方案，哪怕这样会让用户的日子好过得多。因为它需要应对外部各个领域的竞争对手，而且 Google 似乎也认为没理由竞争就不能出自内部——让自己的产品自相残杀。

18 在线销售的定价策略能否用AB测试？

伴随着产品迭代、促销等等因素影响，什么时候降价是对自己最有利的策略，完全可以A/B测试来解决。

19 移动端基于A/B测试的灰度发布怎么做？

就目前移动端的产品来说，iOS的应用商店审核期是个大大大坑，任何BUG打补丁还得再来一遍，也就意味着补丁的审核期内用户带着BUG使用，这个太致命了，用户的获取成本高的吓人，因为这个流失太不值得了，基于A/B测试的灰度发布更重要的不是优化，而是保护性发布，先通过小流量的实际用户测试，有BUG或者新版本体验不好，可以立即回滚到老版本，简单有效。

20 为什么很多公司实施A/B测试效果并不好？

大多数的产品或功能上线前都会进行测试，实际上很多的测试行为并不科学。很多定向的用户测试经常会有这个弊端，简单来说，如果新上线的一个功能，工程师都说好，那是不是意味着所有的用户都觉得好？很多情况下是否定的。例子比较简单，实际上很多A/B测试方法并没有考虑到这个问题，以至于最后得出的结论跟实际情况差异巨大。



很 x
所有的研发
下载简书App
创作你的创作
App



/download?utm_source=stc)

要解决这个问题，对采样、聚类、流量分割等要求非常的高，这也是为什么A/B测试工具不能像很多统计工具一样，埋个点看数据，再根据数据反推业务逻辑，而是要充分与业务结合，从一开始就应该考虑业务策略，让用户去选择适合其口味的产品。

通过AB测试来优化产品的方法在国外已经被广泛应用，现在这种代表先进生产力的方法如同GitHub、Docker、APM一样也正在逐渐被国内广大开发团队所接纳。如果自己公司里面缺乏专业能力和经营，可以尝试用www.AppAdhoc.com优化平台来提高产品的设计、研发、运营和营销的效率，降低产品决策风险，同时也能够帮助用户用数据优化移动广告，让流量的变现价值更大。

现在的互联网公司尤其是创业型公司面临着前所未有的竞争压力，好的想法与用户接受的想法有着各种不可逾越的鸿沟。特别是伴随着激烈的竞争，谁能领先一步可能就变成了赢者通吃的局面。

如果你希望能够提升竞争力，快速验证想法，迭代产品，做数据驱动的增长，建议你来参加互联网产品增长大会，听听国内通过低成本预算获得几亿用户的著名公司创始人们怎么说，如饿了么联合创始人汪渊、触宝科技联合创始人兼任 CEO王佳梁，WiFi万能钥匙联合创始人张发有讲讲增长的秘密。

相信你投资的268元门票必将获得100倍甚至10000倍的回报。

以及一些有过成功增长经验的专家，包括陆金所网站产品管理部副总经理唐灏，《增长黑客》作者范冰，GrowingIO CEO (前LinkedIn高级总监) 张溪梦，吆喝科技CEO(前Google工程师) 王晔，360奇酷粉丝运营总监类延昊，Teambition 增长团队负责人钱卓群，触宝科技增长团队负责人杨乘骁，昭合投资合伙人(前Movoto公司中国总经理)陈世欣等。
点击[这里](http://eventm.31huiyi.com/311999130#rd?sukey=3903d1d3b699c208a1bd61219c70fb4656c7907d39099db813783f0c128b8a97fba3ba386ec3c96a7ba8f166436000bb) (<http://eventm.31huiyi.com/311999130#rd?sukey=3903d1d3b699c208a1bd61219c70fb4656c7907d39099db813783f0c128b8a97fba3ba386ec3c96a7ba8f166436000bb>)进行报名。

 掘金日报 (/nb/1900285) 举报文章 © 著作权归作者所有



掘金官方 (/u/5fc9b6410f4f)

写了 2484108 字，被 8847 人关注，获得了 9237 个喜欢 (/u/5fc9b6410f4f)

掘金 —— 高质量的技术社区 <https://juejin.im>

+ 关注



下载简书App
创作你的创作

/apps/download?utm_source=stc

喜欢 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-like-button)

1



更多分享



下载简书 App ▶

随时随地发现和创作内容





nshu.io

33.jpg)

(/apps/download?utm_source=nbc)

被以下专题收入，发现更多相似内容

- 

A/B测试 (/c/99dc28ea22e4?utm_source=desktop&utm_medium=notes-included-collection)
- 

PM&&Design (/c/e69e85ee348c?utm_source=desktop&utm_medium=notes-included-collection)



×

下载简书App

创作你的创作





(/apps
/download?utm_source=stc)