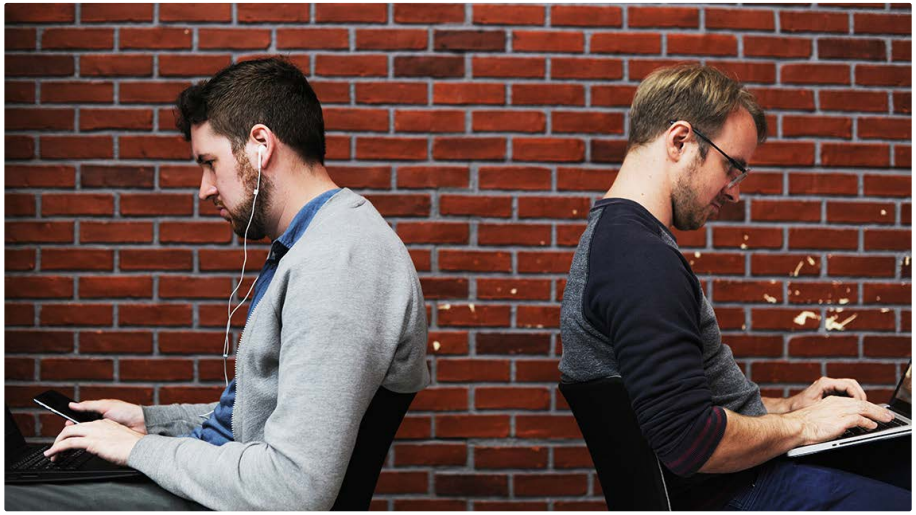
每个工程师都应该了解的：A/B测试

2017-11-17 朱赟



每个工程师都应该了解的：A/B测试

朱赟

00:00 / 00:00

说到 A/B 测试，不论你是工程师、数据科学家、还是产品经理，应该对这个概念都不陌生。

简单来说，A/B 测试是一种数据分析手段，它可以对产品特性、设计、市场、营销等方面进行受控实验。在实验中，数据样本被分到两个“桶”中，分别加以不同的控制和处理，然后对采集回来的信息进行对比分析。

举一个例子。

假如你想修改 UI 上一个模块的交互设计，这个模块的内容是引导用户点击“下一步”按钮，但是你不知道设计改动前后哪一种效果更佳。

于是你通过 A/B 测试，让一部分用户体验新的 UI，另一部分用户继续使用旧的 UI，再对采集回来的数据进行分析，对不同组用户在这个页面上的转化率进行比较，观察在哪一种 UI 下，用户更愿意往下走。有了数据分析，我们就可以判断新的设计是否改进了用户体验。

原理就这么简单。下面我会从自己使用 A/B 测试的经验出发，重点说一说 A/B 测试中需要注意哪些问题，观点会比较侧重于工程师视角，但是对产品经理也会有帮助。

第一点：永远不要过分相信你的直觉。 有时候，我们会觉得一个功能特征的改动是理所当然的，更新后效果肯定更好，做什么 A/B 测试，这显然是画蛇添足。

这就像一个资深的程序员修改线上代码一样：这样改，一定不会出问题。我们当然不否认这样的情况存在，但每当你开始有这样的念头时，我建议你先停下来，仔细地想一想，是不是就不那么确定了呢？

把你的想法和别的工程师、设计师、产品经理深入交流一下，看看他们会不会有不同的意见和建议。不同的角色背景也不同，考虑问题的方式也就不一样。当你不确定哪种方式更好的时候，A/B 测试就是你最好的选择。

第二点：实验样本的数量和分配很重要。 如果你的实验注定没有太多数据，也许就不要去做 A/B 测试了，小样本偏差会很大，帮不了太多的忙，除非你的测试结果出现“一边倒”的情况。

另外，请确保你在 A 组和 B 组随机分配的数据是绝对公平的。也就是说，你的分配算法不会让两个桶的数据产生额外的干扰。

比如，不要按不同时间段把用户分配到不同的组里，因为在不同时间段使用产品的用户本身就会出现一些不同的情况。区域分配也存在同样的问题，这些都可能导致偏差。

第三点：分析的维度尽可能全面。 文章开头举的例子是说，虽然你最在乎的是用户转化率，但是功能改动可能会影响很多指标，这些指标都要尽可能地测量和分析。

比如，虽然 A 组转化率略高于 B 组，但是 A 组点击后会引发 API 调用流程的变化，结果延迟高出很多，或者出错率变高了，那么 A 依然不是更好的设计。

换句话说，A/B 测试不能只关注单一指标，测试目标虽然是转化率，但倘若高转化率的方案会导致其他风险，比如提高了出错率，也应当舍弃。

第四点：其它组的改动对 A/B 测试产生的影响。 当 A/B 测试成为一个广泛使用的工具后，产品很多特性的改动都会用到这个工具。这也就意味着，当你在采集数据做分析的时候，别人也在做同样的事，只不过策略和数据样本不同。

换句话说，你在跑 A/B 测试比较 A 和 B 的优劣，另一个同事在跑 A/B 测试比较 C 和 D 的优劣，结果因为实现细节的原因，A 组中大部分样本同样也是 C 组改动过的样本。这样一来，两个实验可能就会相互影响。因此，你要做足够的分析，确保实验结果考虑到了这种相关性的影响。

第五点：比较值的趋势必须是收敛的，而不是发散的。 要想比较结果有实际的统计意义，一定是每天采集数据的比较结果逐步收敛，最终趋于稳定。如果一周内 A 比较好，后面又开始波动，B 变得更好，这样来回波动的结果是没有太大参考价值的。

另外，即使比较值趋于稳定，还要确保这个稳定数据所处的阶段不在一个特殊时期。如果恰好有促销或者类似的市场活动，那么即便获得了稳定的结果，这个结果也不一定是普适的。

第六点：数据埋点。 数据的埋点和采集是 A/B 测试成功的关键。

怎么样进行埋点呢？总体来说，这其实和每个公司的代码架构有很大的关系。公司使用哪种方式触动事件、记录事件，尽可能地重用。

前端埋点一般可以采集实时数据，后端埋点可以采集实时事件，也可能是一些聚合数据。要视具体情况和应用而定。

第七点：形成一个流程，或者设计一个工具。这一点很重要。A/B 测试作为一个工具，只有在它足够灵活、好用的情况下，才能更广泛地应用到日常的产品迭代和开发中。虽然说这个方法很简单，但是做好一套包括埋点、采集、处理和具备 UI 的工具，会让工程师事半功倍。

第八点：试图给每个结果一个合理的解释。不用过分相信数据，也不要拿到什么分析结果都照单全收。试着去给每个结果一个合理的解释，不要觉得结果比期望值还好，就不用思考为什么结果如此完美。这可能并不是一件好事情，实际情况是：如果解释不了，可能它就是个 Bug。

第九点：必要的时候重新设计实验。很多实验会有不同版本，每个版本都会根据实验结果做一些改动和调整。如果发现实验设计上有漏洞，或是代码实现有问题，那就需要随时调整或者重新设计实验，重新取样、分析。实验的版本控制，会让分析和重新设置的过程更加快捷。

第十点：不同客户端分开进行实验。Web 端、iOS、Android 尽可能分开观察。很多时候你会发现，同样的实验数据对比，在不同的客户端会有完全不同的结果。如果不分开，很可能让数据变得难以解读，或者出现“将只对移动客户端成立的结果扩展到 Web 端”，这样以偏概全的错误。

最后，我们来做一个小结。今天我结合自己的实际工作经验，为你讲述了 A/B 测试中需要注意一些问题。

A/B 测试是一种行之有效的产品验证和功能改进方法，很多互联网公司，如Google、Facebook、Airbnb 等都有自己的 A/B 测试工具，他们会基于工具和数据验证自己的想法，持续进行功能改进、推动产品的发展。

如果你也在做 A/B 测试实验，可以对照我在文本中提到的那些问题来思考，相信你可以做出更好的测试结果。

Jesse

2017-11-17

学习了，不过很多初创类的公司 很少有A B测试。

氮

2017-11-18

提到测试或实验，不得不提**双盲实验**。这个在医疗领域比较常见。一种新疗法，或一款新药是否有效，必须经过**大样本随机双盲实验**。尽可能地排除安慰剂效应或者实验人员的主观臆断。

文中关于设计A/B测试，也可以看做是设计一个“双盲实验”。关于数据的分析，推荐“信号与噪声”这本书，里面许多原则可以通用。

刘剑

2017-11-17

朱老师可否讲一下在移动App上做A / B测试遇到的坑呢？A / B测试需要哪些技术资源配合？比如：客户端、服务器端如何管理测试版本？是否需要跟正式环境隔离？

A/B测试我遇到的情况：

1.IOS的A / B测试就需要有企业开发者账号，但是有些企业是申请不下来的
2.A / B测试点可能是非核心决定要素，可能导致误判
3.A / B测试如果想效果好，有一个前提是有明确的用户画像，后续工作就清晰和明确的多了

---

fenghao

2017-11-17

logging太重要，很多时候看到结果需要解释，发现没有log结果又要再开实验很费时间

Seven _ dong

2017-11-25

说个具体的技术实现，用ELK可以方案的实现出一套A/B Testing 系统

逗逼师父

2017-11-17

产品修改以事实为依托，这样大家都能接受，而且工程师和产品都能看到自己努力的结果，有了即时反馈就更容易产生驱动力。这种方式的确值得学习，受教了。

李红元

2018-05-24

讲的真好。

---

reneaimececi

2018-05-20

请教一个问题：A/B Testing跟另一个常提到的Canary金丝雀发布是同一个机制么？

彭超

2018-04-23

AB测试是和灰度发布一个意思么

吴天

2018-02-28

首先建立一套数据埋点和采集工具

85后小卡

2017-12-27

“请确保你在 A 组和 B 组随机分配的数据是绝对公平的”，有一种方案是：完全随机（不区分用户属性，如性别，年龄段，地域等），还有一种是方案是：根据用户属性来划分AB，例如上一次已经分配A给了一个男性用户，本次用户如果还是男性就分配B，如果是女性还是分配给A，这样能保证A和B组里面的用户属性（比如男女比例）都大致相同。不知大神会建议采用哪种方案？

walt

2017-12-24

A/B测试如何实施呢，具体说埋点、采集、分析等一套UI工具如何搭建

沉思猿

2017-11-21

A/B测试在后端开发中有什么应用场景吗？

英子

2017-11-20

今天刚跟组里的小姑娘聊到尝试站在读者或者用户的角度思考问题。这一点有时候太难了，就像女神说的，我们过分相信自己的直觉，有时候感觉某一事情的发生理所当然，殊不知，这理所当然的范围只是我们个人的思维。以前听过一句话，这世界上根本不存在感同身受，大概是说感情的，不过，也说明完全理解他人，或者说局外人要做出和局内人对某件事一样的反应该有多难。你之于我是局外人，产品之于开发是局外人，用户之于所有产品开发者是局外人，以自己的思维推测别人的反馈应该是出错率很高的一件事儿。

---

smith

2017-11-18

杨澜的声音与你有相同之处

法老

2017-11-17

AB测试需要多大的数据量才能达到效果呢？

---

mengwen

2017-11-17

很想知道research scientist如果参与产品改动会以什么样的方式介入