A/B testing

Emma Good, egood@kth.se

April 2019

1 Introduction

Testing is important. Code usually is (or at least should be) tested before deployment to detect bugs, to make sure that it runs correctly and that the results are satisfactory. While almost every company test their code for their platforms, testing the content on the platform could easily be forgotten. While you might have the best and smartest ideas and back-end of them all, sub-par content might keep people from engaging with the platform. Instead on relying on what marketers ”feel” will give the best response, you should get some real data on what works and what doesn’t. This is where A/B testing comes in.

2 What is A/B testing?

The idea behind A/B testing is to compare two (or more) versions of a feature to each other, and to get feedback on which one gets the best response. What the best response is depends on the context, but a common metric for websites is which version gets the most clicks. One common use case is to compare a new, proposed feature with the feature currently in use to find out if the new version will achieve better results.

The test is performed by dividing the users into different groups where each group receive a different version of the feature to be tested. In the case where we have a new feature being tested against and old feature, we might assign the new feature to half of the users and have the other half as a ”control group” that receives the old version. After running the test for the specified time the results of the test are evaluated, and we will find out if one version performed better, and we will have the data to prove it. [1]

One simple example is illustrated in Figure 1, where we have two ’buy’ buttons of different colors on a product page on an online store. We might want to perform an A/B test to find out if the proposed new variant A would result in more clicks than the old variant B. During a test period of a week, we let 50% of users receive variant A when loading a product page on our store, and the other 50% will receive the original variant B. During the test period, we measure the number of clicks on the button made by both groups, and when the week is over we look at the results and notice that the group that received variant A clicked on the 'buy' button 10% more than the group that got variant B. This shows that updating the button to variant A for all users could be a good idea.

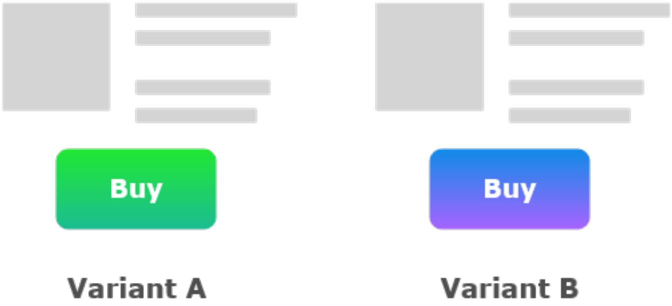

Figure 1: Two different versions of a 'buy' button.

3 Anatomy of an A/B test

To get the most out of an A/B test it is important that it is planned well, as with any experiment. Below are some important steps that should be followed in order to conduct a successful A/B test.

Test one variable at a time

The first part of an A/B test is to figure out what variable (feature) you want to test. You might want to test many different variables, but only one variable should be tested at a time. If multiple variables are tested at the same time, you won't know which variable affected the change.

Hypothesis

When the variable to be tested has been chosen, the next step is to create an hypothesis. The hypothesis states what you will change, what you think the outcome of the test will be, and why. Setting up the hypothesis helps to make sure that you have done your research before starting the test and has thought trough the experiment. [2]

Create the versions

Now that we have chosen the variable and created a hypothesis it's time to create the new alternative. Based on the hypothesis, we create a new "challenger" version that we think will give us better results.

Divide the groups

When dividing the groups it is important to make sure that the test groups are big enough to get a reasonable amount of data to be able to tell if there is a difference in results between the versions. How to divide the test groups depends on the platform we are testing. If we are testing two different versions of a link text in an e-mail, we can easily divide the list of recipients into equally sized test groups. Then we send one version of the e-mail to one test group and the other version to another group, and log which of the groups clicked the link the most. If we are A/B testing a website, we can send some percentage of visitors the new challenger version and some others the original version.

Since dividing the groups and keeping track of the visitors from each group can be a bit tricky, it is a good idea to use an A/B testing tool that will help with this. Some popular tools are Google Analytics and Google Optimize, Optimizely, VWO, and Adobe Target [3].

Decide on the desired significance of the results

Now it's time to think about how significant we want our test results to be in order to justify picking one version over the other when the test has ended and the results are in. We need to choose a confidence level that reflects how sure we want to be that the test results indicate that one version actually performed better than the other, rather than the results being random. We usually want a high confidence level at a minimum of 95%, especially if the test took a lot of time to set up. For a smaller test with less strict requirements, a lower confidence level can be good enough. [4]

Patience

Now it's time to start the test and to deploy the different versions. With a website it can be hard to specify a specific length of the test, since you need to get enough visitors to generate the amount of data needed to make a good decision. Depending on the site this can take from a couple of hours to several weeks. Just let the test take it's time.

Analyze the results

When enough data is gathered it's time to decide if the results are statistically significant. Calculate the confidence level of the results and compare it to the desired confidence level you set earlier.

Take action

Congratulations, you successfully conducted an A/B test! After analyzing the results you will know if one version performed better than the other, and can choose to deploy that version. If neither version was statistically better, you have learned that the chosen variable probably doesn't affect the performance. This failing data can be used to improve a future iteration of the A/B test.

Start planing for the next A/B test

There is always room for improvement. The test hopefully pointed to one variable that could be optimized for better performance, so now it's time to find others that can be optimized. Almost every variable can be A/B tested, so there is a lot of potential improvement that can be done.

4 Case studies

The following section will present some successful A/B tests performed by companies and organizations to showcase how A/B testing has been used in real life and the impact it can have.

4.1 Netflix

Every product change at Netflix is A/B tested before becoming the new default. A Netflix user is usually a part of several A/B tests at any given time, as long as the tests doesn't conflict with each other. [5] One of areas where A/B testing is most important for Netflix is when it comes to selecting which artwork to display for movies and series. The artworks need to be engaging enough to catch the interest of a user fast, otherwise the user might leave Netflix and go do something else. By using A/B testing Netflix has found that just having a good artwork for a movie/series could significantly increase user engagement and the number of views for a title. [6]

Another example of where Netflix has used A/B testing is on their landing page. A Netflix survey had shown that 46% of the respondents wanted to browse the available titles before signing up. After deciding to implement this feature Netflix A/B tested five versions of the new landing page, where visitors could browse titles before signing up, against the original landing page, without title browsing. For the first test round, Version 1 was tested against the original page, and the winner would move on to be tested against Version 2, where the winner would be tested against Version 3. This process was repeated until all five new versions had been tested, and the best version would be the one to be deployed. When all the versions had been tested, the results showed that the original landing page, without title browsing, had outperformed all of the new versions. This proves that users don't always know what they want, and that A/B testing is worth spending time on. In Netflix had chosen to add the browsing feature without A/B testing it first, they might not had known why their new feature performed badly. [7]

4.2 Obama 2008 presidential campaign

In 2007, the Obama 2008 campaign team conducted an A/B test on the Obama campaign landing page that helped them raise $60 million more than they might have without the test. The landing page had two variables that were tested; a 'media' placement with a large picture as on top of the page and a sign-up button below the picture. The campaign team made four versions of the button with different text, and three different pictures and three different videos for the media placement. Every combination of media and button were tested against each other (known as multivariate testing), so a visitor would be randomly assigned one of the 24 possible combinations, and the website tracked how many from each group that signed up.

The results showed that the combination of a picture of the Obama family and the button text "Learn more" lead to a 40.6% increase in sign-up rate. The campaign staff's favourite, one of the videos, was one of the worst performing medias. If the campaign page had not been A/B tested, that video would most likely have been chosen, and the number of people that signed up, and later donated to the campaign might have been a lot lower. [8]

4.3 Huffington Post

The Huffington Post is an American news site, and uses A/B testing to find the best headlines for some of their articles. Visitors are randomly shown one of two headlines for an article. Due to the site's high amount of traffic, it only takes about five minutes until enough data is gathered to calculate which headline gets the most clicks. That headline is then shown for all of the site's visitors. The Huffington Post editors have also noticed that putting the author's name above headline almost always generates more clicks than not having the name.

5 Conclusion

A/B testing is a powerful technique that can help optimize a platform, and should not be neglected. Nowadays we are constantly bombarded with new web pages, e-mails, and apps. In order to be noticed you need your platform to be the best it can be. By rigorously A/B testing every variable you can, you might find that features you thought were great actually gave you worse results than you could have had. There are great tools available that will help with A/B testing, so there is really no reason why people shouldn't do it besides not having heard of A/B testing (yet). Relying on what you feel will perform the best might not always work out as well as it could. If you can get data on what actually will work, you might learn more about your users and could perhaps make even better optimizations in the future. Even if an A/B test might not show that one version performs better than another, you will have learned from the experience. Life is short, optimize.

References

[1] "A/B Testing." [Online]. Available: https://www.optimizely.com/optimization-glossary/ab-testing/

[2] "How to Write a Solid A/B Test Hypothesis," Jan. 2015. [Online]. Available: https://blog.optimizely.com/2015/01/29/why-an-experiment-without-a-hypothesis-is-dead-on-arrival/

[3] "10 Best A/B Testing Tools (That Work in 2019)," Sep. 2018. [Online]. Available: https://www.ventureharbour.com/best-a-b-testing-tools/

[4] L. Kolowich, "How to Do A/B Testing: A Checklist You'll Want to Bookmark." [Online]. Available: https://blog.hubspot.com/marketing/how-to-do-a-b-testing

[5] N. T. Blog, "It's All A/Bout Testing," Apr. 2016. [Online]. Available: https://medium.com/netflix-techblog/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15

[6] ——, "Selecting the best artwork for videos through A/B testing," May 2016. [Online]. Available: https://medium.com/netflix-techblog/selecting-the-best-artwork-for-videos-through-a-b-testing-f6155c4595f6

[7] "The Registration Test Results Netflix Never Expected," Nov. 2015. [Online]. Available: https://apptimize.com/blog/2015/11/netflix-registration-ab-test/

[8] "How Obama Raised $60 Million by Running a Simple Experiment," Nov. 2010. [Online]. Available: https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/