# COMP33812: Software Evolution Coursework 2

Alex-Radu Malan
9770386

# A/B Testing – Successes and Challenges

## Introduction

A/B testing is an experiment used in statistics that are concerned about ideas and how the variants of an idea would get more attention from the public (2). Usually used in web analytics, this type of testing is about different variants of the same website getting experimented on a chosen group of people in order to have a result that will reveal the most accessed variant of the website and the most loved one. A/B testing is used amongst a large range of companies that are using the web environment to promote their business such as Amazon, Google, Twitter, Yahoo, etc. (10). Besides companies, there are a variety of other websites that belong to individuals, charities and other forms of organisation which will test different versions of their website in order to know what variant will be set for final release (9). In the following report, we will review the successes and challenges the A/B testing experiment is going through in order to help the developers create a better user experience, gain profits and provide a better platform for the users based on the data collected from them. One of the first times this experiment has been conducted was in the 1700s when there was a British crew of members with a captain leading them(5). They were having health problems because of the lack of Vitamin C which is essential; the captain divided the crew into two groups: a control group and a treatment group. The control group represents the people that would continue their normal diet, while the treatment group would start to add limes to their diet. After a while, the captain observed the fact that the health of the treatment group was starting to improve while the control group members would have decreased health than before. In conclusion, the whole crew started to eat limes in order to be healthy.

## Successes

The story with the captain adapted for the 21st Century would be testing on the same type of groups - control and treatment - but mainly used in web analytics. Users would use a website/platform such as Amazon, for example, and the user experience would lack certain features or implementation. The development team want to improve the platform in order to increase the sales and user satisfaction, so they will have to make the decisions based on the user data. The team is assigning to a certain group that is part of their user base the same platform, but with a new feature or implementation, while the other group would continue to use the same features, nothing new added. After testing the new platform for a period of time they will get an outcome also known as Overall Evaluation Criterion(OEC) that would provide the development team with the differences found between the two groups.

First of all, the most important part in creating and conducting such experiment is the expecting result, so we would have a certain expectation rather than just

waiting for variables and then think how to interpret them (7). In the outcome prediction equation there are properties such as the confidence level, so that in 5% of the cases there is no difference between the control and treatment group, power property, which mean that the rest of 80 to 95% there will be a difference and another important part would be the standard error which can be reduced by increasing the sample size(the users involved in the experiment). After setting the expectations for the experiment we need to have ways of conducting it. One of the most popular ones is called "Treatment Ramp-Up" which refers to the fact that we will take a proportion of 99% control group, and 1% treatment group in the beginning. Usually, an experiment is conducted on a week-time period so from the start to the end we will keep on increasing the treatment percent while in the meantime decreasing the control group percent. Another way of having the experiment is by using automation. Amazon and other online retailers are using automation for quick response to the user such as recommendations ("People also bought:") which led to increased number of sales. Most of the time the users that are part of an experiment of such a type will either get an ID linked to their account, which will tell if they are if they are involved or not and even know what group they are part of, or by using caching which will remember all the details for a certain user next time he/she will enter the platform, this being also called a randomisation algorithm(11).

Amazon is the biggest online store in the world nowadays and one component that helped it reach this far is the A/B testing used in order to improve the way people are shopping online. One of the ideas for the Amazon development team was to check on what buttons the users are testing and also in what order. Basically, this would help in understanding the path a user is going through when he/she search for an item or buys an item on the Amazon website. The problem that came to them was the fact that in a certain layout or webpage there are multiple components and the number of choices a user has is too big. In order to solve this problem, they reduced the decision space and then implement and conduct this experiment based on the Bayesian model. Bayesian model would determine the path a certain user is following on the website by using probabilities of choosing the same components when purchasing something, for example, if they would use the same path like searching for the item in the search bar, followed by scrolling through the list of results, selecting the result that matches their expectation, adding it to the basket and the process goes so on so forth.

The Amazon team used three types of practices in order to create variants of the webpage for the users (4). To have a better understanding of those practices they used clothes as a comparison. The first practice would be the Bandit Algorithm. When you want to go on a date, for example, the clothes you are wearing does not matter when you have a horse mask on your face. The Bandit Algorithm takes the most unsuccessful component from the page and makes sure that it is not used on any other pages that relate to a certain topic. Basically,

besides Halloween, a horse mask is not used by anyone which will make the usability of it to decrease to less than 1% of all of the time a person is wearing clothes. The second practice used is called the Hill-Climbing Algorithm in which a person would have a certain outfit and then try all the pants available to see which one would best fit with the current outfit. In the webpage, a certain component would be changed in order to match the overall webpage outline and components. The last practice used is called Model Interaction. This refers to creating components for a webpage that would best fit that webpage, also called a fast generalisation. In terms of clothes is like picking an outfit based on the mentality of "I think this might fit". The most important part of these experiments is to conduct them for a long period of time since the probabilities would get better and better based on how much data the algorithm will get from the website users. Another thing that was taken into consideration was the Thompson sampling (3) that is used with the Bayesian Control Rule. The main idea is that a certain user will learn how to use the platform better after using it for a longer period of time rather than taking his first-time use of the platform into consideration that much. Those practices were a small bit of how Amazon reached success.

**Challenges**

Although there are benefits of developing platforms with A/B testing there are challenges as well. For the Amazon team described above everything was great with the clothes/component part but there was one more problem that nobody ever taught about. What if you and your date set a meeting at a certain street corner, you dress for restaurant and she/he dress for tennis playing. Here is a problem that is related to the content of the website and how the components will match that content. Amazon is having different components and options for each type of content. A simple example would be a coke drink and a ruler. You can choose to have a 300ml, 1L coke and you can choose to have a 1m or 2m ruler, but not a 2m coke. That is why it is important to adapt the components on the content as well and match each content and product type rather than generalising the components. ¬

Challenges arise from different parts of the experiment, for example when a migration of a software is made we have to be sure that there is no difference between two groups of users with the same version of the platform, this is also called A/A testing since it is the same platform delivered to two groups. Another problem would be how do we perceive the results of an experiment, from what perspective should we review. When the Amazon recommendation was tested, an executive would not agree on it with the idea that people would be distracted from the checkout if we keep on recommending them other products, which was a bad idea since the recommendations feature increased the sales in order of magnitude. Besides this, the ads that a user is exposed to should be reviewed as well. If a user is not clicking on a certain type of ad, then

we should change the ad instead of keeping it. In the meantime of testing the platform there might be problems of browser compatibility or a certain type of code is not compatible to a certain percentage of the treatment group so that means that the data should be mined so that small problem like this would not interfere with the results or the experience a testing user should receive in the experiment. The speed is also important for the user satisfaction. When testing a new feature which most of the time is not optimised for the platform since we do not know if we will keep it or not, the load time will increase which can lead to unsatisfied users that will also lead to decreasing number of sales. Amazon lost 1% of the sales because the users would need to wait an extra 100msec. Google got a 20% revenue drop when the users had to wait 500msec more. Another factor is the day of the week, which refers to the fact that some users will enter the website more in the weekend for example of in another day of the week which is something to take into consideration. Consistency and the newness effect is also important because when you give something new to the users, they will click randomly in order to understand the new feature which might mess up the metrics for the first week of use. The number of variants of the website should take into consideration the return on investment (ROI). A feature should be able to generate revenue that would cover the amount of money invested into other unsuccessful variants or features added to the website.
After all, one of the most important and big challenges in the A/B testing is the press release problem. One of the most famous press release examples is the Facebook flower like. For a certain amount of users, Facebook released a new feature that would allow the user to like with a flower. The new feature was published in articles and newspapers which reached Facebook users that were in the control group and could not have access to this feature, making them "sad" that they are unable to use the feature.

**Conclusion**

In conclusion, A/B testing and other practices that are into the sphere of software evolution were not existing in the past and their use in the platforms are very welcomed since it is important to listen to what users want and need instead of developing a platform based on what one person is deciding. The most important thing is to solve the challenges and problems that arise, always test the features instead of going with a mentality of "That will not hurt anybody" and also run experiments continuously instead of just once in a while. It is important to always get data from the users in order to improve the platform and its way of satisfying users.

# Biography


1. Amazon Web Services. (2017). *A/B Testing at Scale – Amazon Machine Learning Research | Amazon Web Services*. [online] Available at: https://aws.amazon.com/blogs/machine-learning/ab-testing-at-scale-amazon-machine-learning-research/

2. En.wikipedia.org. (n.d.). *A/B testing*. [online] Available at: https://en.wikipedia.org/wiki/A/B_testing

3. En.wikipedia.org. (n.d.). *Thompson sampling*. [online] Available at: https://en.wikipedia.org/wiki/Thompson_sampling

4. Guy Ernest (2017). *An Eficient Bandit Algorithm for Realtime Multivariate Optimization*. KDD 2017 Applied Data Science Paper. [online] Available at: http://www.kdd.org/kdd2017/papers/view/an-efficient-bandit-algorithm-for-realtime-multivariate-optimization

5. Kohavi, Ron, Randal M. Henne, and Dan Sommerfield. "Practical Guide to Controlled Experiments on the Web: Listen to your Customers not to the HiPPO." *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007. [http://www.robotics.stanford.edu/~ronnyk/2007GuideControlledExperiments.pdf]

6. Lu, Luo, and Chuang Liu. "Separation Strategies for Three Pitfalls in A/B Testing." *Proceedings of the 2nd Workshop on User Engagement Optimization at KDD '14*. ACM. 2014. [http://www.ueo-workshop.com/wpcontent/uploads/2014/04/Separation-strategies-for-three-pitfalls-in-AB-testing_withacknowledgments.pdf]

7. Optimizely.com. (n.d.). *A/B Testing*. [online] Available at: https://www.optimizely.com/optimization-glossary/ab-testing/?redir=uk

8. Sarah Clinch (2018). *Evolving the User Experience with A/B Tests*. [online] Available at: https://online.manchester.ac.uk/bbcswebdav/pid-5977097-dt-content-rid-22334206_1/xid-22334206_1

9. Unbounce. (n.d.). *What is A/B testing?*. [online] Available at: https://unbounce.com/landing-page-articles/what-is-ab-testing/

10. VWO (n.d.). *AB testing - The Complete Guide*. [online] Available at: https://vwo.com/ab-testing/

11. Xu, Ya, et al. "From infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks." *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. [https://content.linkedin.com/content/dam/engineering/siteassets/pdfs/ABTestingSocialNetwork_share.pdf]