World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

GirGut: An AI-powered, developer-First AB testing platform mimicking natural selection for web applications

Leela Gowtham Yanamaddi [1,*] and Balaji Kummari [2]

[1] *CEO and VP of Engineering, scale.jobs 537 Payne Rd, Woodstock, GA, USA 30188.*
[2] *CTO, scale.jobs 1-84, Beside Venugopala Swamy Temple, Rayanapadu, Vijayawada, AP 521241, India.*

World Journal of Advanced Research and Reviews, 2024, 21(02), 2035-2044

Publication history: Received on 16 January 2024; revised on 24 February 2024; accepted on 28 February 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.2.0406

Abstract

A high-quality product experience is becoming more important to consumers as a whole as a result of societal progress. The competition is constantly raising the bar for product specifics in their pursuit of a high profit conversion rate. Rapid, high-quality product iteration is essential for product providers looking to increase user viscosity and activity, which in turn increases the profit conversion rate. By inserting logs and analysing statistical data, A/B testing can determine which iterative strategy is more effective by conducting experiments on target users. This paper introduces GirGut, an innovative, open-source AB testing platform designed specifically for web developers. GirGut leverages artificial intelligence to generate and evolve test variants, mimicking the process of natural selection to optimize user engagement and conversion rates. By combining ease of use with powerful AI capabilities, GirGut aims to revolutionize the way developers approach experimentation and optimization in web applications.

Keywords: AI-Powered; AB Testing; Web Applications; Product

1. Introduction

The success of software companies hinges on iterative software development and time to market. Innovation through the discovery of new software features or the experimentation with software modifications is at the heart of this. Software businesses frequently use A/B testing [1,2] to facilitate such innovation in practice. Two versions of a program are tested in the field to see which one performs better; these versions might have anything from a slightly different user interface layout to brand new functionality. This method is called A/B testing, which is also called online controlled experimentation or continuous experimentation. Specifically, marketing conversion rates, subscription service members' lifetime values (LTVs), and website click rates are used to compare and contrast the two versions [3, 4]. Google, Meta, LinkedIn, and Microsoft are just a few of the well-known internet businesses that use A/B testing widely in practice [5, 6]. We were unable to find a single study that examined the current state of the art (i.e., the state of research as opposed to the state of practice) in A/B testing, despite the fact that it is widely employed in practice. To propel further research in the area of A/B testing, a comprehensive review like this one is essential. Three prior studies [7,8] investigated various facets of A/B testing research, including study subjects, A/B testing trial types, and A/B testing metrics and tooling. Still, there is a lack of a holistic review of the current state of the art in these studies, which would shed light on the various uses of A/B testing, the stakeholders' responsibilities in their design, the tests' execution, and the outcomes' interpretation. These understandings are critical for situating and comprehending A/B testing within the larger context of software engineering. We conducted a comprehensive literature review to address this matter [9]. The purpose of our study is to shed light on the current landscape of A/B testing research so that future studies can build upon our findings. The study could also help practitioners find ways to incorporate A/B testing into their everyday work.

* Corresponding author: Leela Gowtham Yanamaddi

1.1. Why use AI for A/B testing?

Landing pages, user interfaces, and other marketing prototypes can undergo A/B testing to identify the optimal version prior to full launch. You divide your listeners into at least two categories. Both view the original version (A) and the changed version (B), but only one can interact with it. What follows is a process for monitoring interactions, evaluating outcomes, and enhancing content.

Figure 1 Shows How AI and traditional methods stack up against each other

1.2. How to Use AI for A/B Testing

The seven ways AI can revolutionise A/B testing are detailed below.

1.2.1. Real-Time Data Analysis to Enhance Decision-Making

Big data insights can be processed in real-time by A/B testing platforms driven by AI. Their ability to spot intricate patterns, trends, and other variables allows for more accurate testing. Multi-Armed Bandit (MAB) algorithms are one example of a test design that showcases AI real-time analysis. For example, it optimises ad placement and content recommendation in real time, and it distributes visitors to variants that are performing better. MAB prioritises ads that demonstrate higher performance as user data accumulates, allocating ad impressions in real-time. It can also change the suggested material depending on how the viewer has interacted with it recently. One app that utilised nGrow's MAB algorithm to decrease user turnover was Amma, a pregnancy tracker. MAB enhanced retention for both iOS and Android users by 12% with real-time automation and optimisation of push notifications. In addition, the group learnt more about their target demographic. New territories can be better planned for, and user involvement can be optimised.

1.2.2. Predictive Analytics to Boost Accuracy

AI predictions save you from testing inefficient versions and having incorrect notions. The managing director of analytics at Zuko, Alun Lucas, showed me the ropes. In order to find the answers to the following questions, he analysed Zuko's form analytics data using artificial intelligence techniques like ChatGPT:

- Which form fields are giving me the most trouble? How are the numbers different from the last period? In order to improve the user experience and decrease desertion in the problem fields that have been identified, what options can we explore? Before they become big problems, predictive analytics can find problems with your data forms or user processes.

1.2.3. Personalized Testing to Create Tailored Experiences

With the help of AI, you can divide your audience into subsets defined by their demographics, interests, and actions. You can target specific demographics with your A/B testing if you're in the fashion product recommendation business, for instance. Clients, price hunters, and environmentally concerned consumers should be considered. Consultancy head Ellie Hughes of Eclipse Group thought this method was useful for checking prototypes before putting them into production. Algorithms like photo-based recommendations and personalised search rankings were among those she tried. What became of it? It improved her clients' experience and presented a strong argument for investing more in AI. "The value wasn't in the production of an algorithm as an output," Hughes says. It was all about the ingenious design of an experiment to demonstrate the financial worth of incorporating AI into studies.

1.2.4. Multivariate Testing to Reveal Useful Insights

With A/B testing, you may expand the range of options from just A and B to all the way to A to Z. According to Ellie Hughes, "A/B testing can involve multiple variants and more complex experimental designs, such as multivariate testing [...] to optimise various elements simultaneously." This dispels the idea that A/B testing is only used to compare two versions.

2. Literature review

The purpose of A/B testing is to evaluate two software versions, called variants A and B, by observing how users interact with the system and drawing conclusions about which version is better. An experiment, or the real A/B test, is designed to compare the versions after a hypothesis is formed. Alternate hypothesis testing (A/B testing) uses real-world systems instead of simulated ones. Figure 2 depicts the three primary steps of an A/B test [11]. Creating an A/B test is the initial step in the A/B testing process. Many factors are defined in this experimental design, including the hypothesis, the population sample to be tested, the length of the trial, and the A/B metrics to be measured. In order to find out which variation is better during the experiment, the A/B metrics are utilised. The amount of clicks, the click-through rate (CTR), and the number of sessions are all examples of A/B metrics [12]. Running the A/B test in a live software environment is the second step in A/B testing. Both versions are tested in a real-world environment, and the population sample is distributed evenly between them. In order to assess the experiment once it has finished (in the allotted time), the system records pertinent data during execution. Data that is relevant can either directly relate to the A/B metrics that were defined or can indirectly allow for advanced analysis during the review phase to obtain more insights from the A/B tests that were run. Conducting an analysis of the trial is the third step in A/B testing. Following completion of the A/B test, the original hypothesis is then tested, usually using a statistical test like a students test or Welsh's t-test [13, 14]. The designer can then proceed with the feature rollout or the creation of new A/B variations to test in future A/B tests depending on the test's outcome.

2.1. Controlled experiments vs A/B testing

Conventionally, one way to empirically test a theory is using a controlled experiment [15]. Independent and dependent variables are the two main categories of experimental variables in a controlled setting. An example of an independent variable would be a control group that uses the current state of the art to tackle a specific design challenge and a treatment group that uses a newly suggested strategy; both groups are under the control of the experimenters in order to test the hypothesis. Examples of dependent variables are the fault density and productivity measured in a design task, which are used to compare the results of the control and treatment groups in an experiment. As an example, a recently suggested design approach has a much lower fault density than the state-of-the-art technique, but further research is needed to determine the productivity, after which the hypothesis is evaluated and conclusions are taken based on the results. Software engineering is only one of many modern scientific disciplines that makes extensive use of controlled experiments, which are commonplace in psychology, pharmaceutics, education, and many more [16].

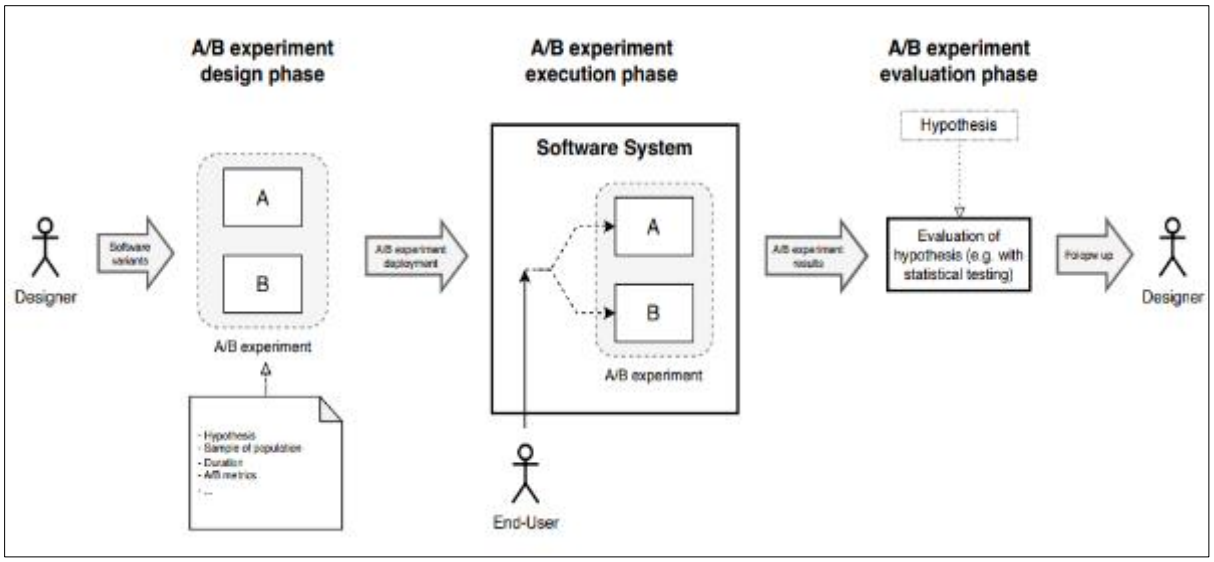

Figure 2 General A/B testing process

Unlike traditional controlled tests, which take place in a controlled environment, A/B testing evaluates software features or variants on end-users of a running system by means of controlled experiments. This is why "online controlled experimentation" is a common name for A/B testing. The goal of A/B testing is to put theories to the test in real-world software systems with actual users serving as the experimental population. When doing A/B testing, it is common practice to test hypotheses about how to enhance the user experience (UX) [17], the design of the user interface (UI) [18], the rate at which users click on links, and the evaluation of non-functional requirements in distributed services [19].

2.2. Social aspects of A/B testing.

Receiving input from users is a crucial social component of A/B testing. User experience optimisation and prioritisation constitute a large part of A/B testing. This social aspect is the subject of two studies that we found. Provide a literature assessment on methods for gathering data and client input within the framework of software R&D in [19]. The writers discuss the current methods for gathering customer feedback and organising data collection in the literature. They also include the software development stages where these methods are utilised, as well as the key obstacles and limits of these methods. An useful tool for getting user input on prototypes is A/B testing, which is one of the methods described by the authors. Discuss the difficulties and consequences of big companies not sharing customer data in [20]. The authors provide a concrete example to illustrate the serious problems that arise when developers are required to either obtain user feedback again or create products without the qualitative data that was collected during the pre-development phase.

2.3. A/B testing case studies

There is a science to the art of product and website design. An effective strategy is to run A/B tests, which involve comparing two versions of a product or website in a controlled environment to see which one yields better results. There is a standard format for A/B testing, which is also called split testing:

- Find a problem
- Create a hypothesis of how you could solve it
- Create a new design or different copy based on your hypothesis
- Test the new version against the old one
- Analyze the results

In spite of this, there are a plethora of options available to you inside this framework in terms of A/B testing tools, data formats, and data collection methods. Looking at instances of successful A/B testing is a great method to learn and get better:

2.3.1. Bannersnack: landing page

Bannersnack, a company offering online ad design tools, knew they wanted to improve the user experience and increase conversions—in this case, sign-ups—on their landing page.

Bannersnack sought to understand user behaviour on the page using Hotjar Heatmaps since they were unsure of where to begin. Thanks to heatmaps, the business could tell exactly where people were clicking and which parts of the site were largely unnoticed. Bannersnack might use this information to make educated guesses about how to enhance the experience, and then they could construct a version to test alongside the original. By iteratively checking heatmaps, Bannersnack got closer and closer to their target outcomes during the testing process. They realised they required a bigger CTA button with a better contrast ratio, and as a result, sign-ups went up by 25%.

2.4. Turum-burum: checkout flow

The goal of the digital UX design agency Turum-burum was to increase sales for the Ukrainian online shoe retailer Intertop. During the user experience study phase, Turum-burum collected data on Intertop's checkout page using Hotjar Surveys, more precisely an exit-intent pop-up. Just before a user could click the "Exit" button, the survey would enquire, "Why would you like to stop placing the order?" Nearly half (444 people) of those who took the survey reported being unable to finish the purchase process. Making and testing theories was the following stage. Some of the adjustments that were examined were dividing the webpage into blocks, lowering the number of form fields, and adding a time-saving autofill function.

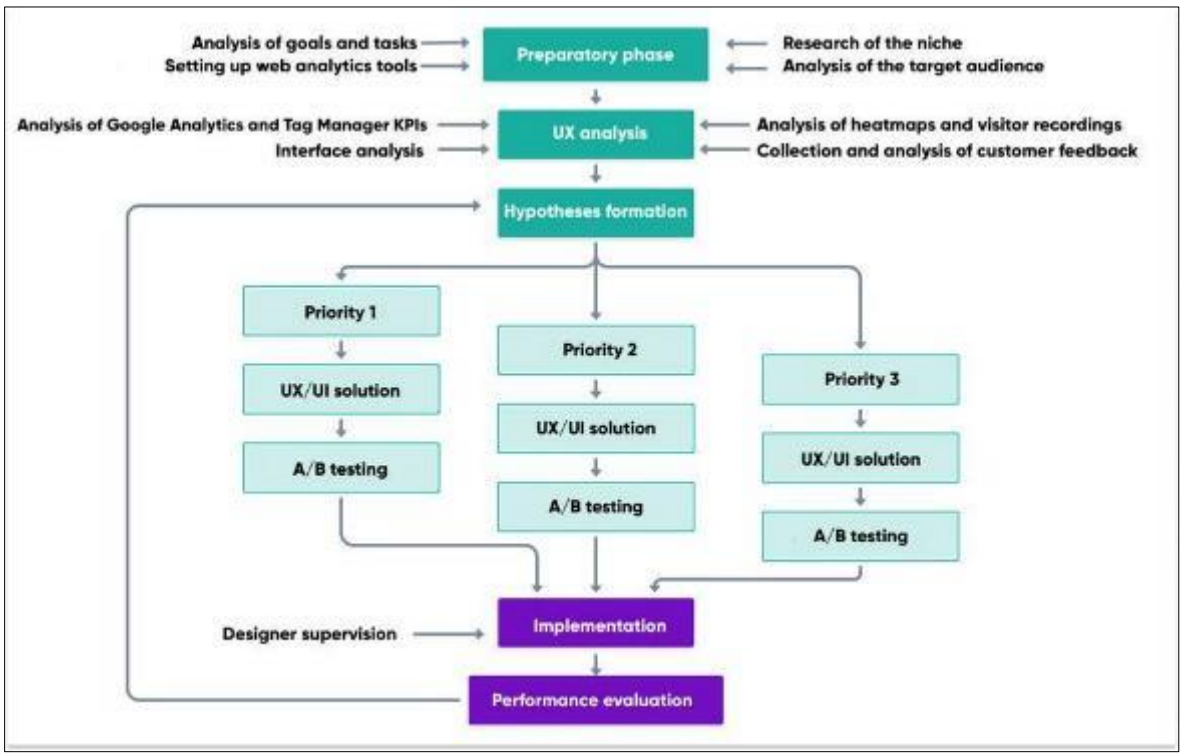

Figure 3 Flow chart

Evolutionary Site Redesign (ESR), Turum-conversion burum's rate optimisation (CRO) model, relies heavily on A/B testing. The business tracked user reactions to every page update using Hotjar Recordings and Heatmaps. Through the use of heatmaps and recordings, the team was able to identify patterns in the click and scroll behaviour of users and identify spots of friction, such as angry clicks, that customers experienced when navigating the checkout process. The end product? The test variant saw a 54.68% rise in Intertop's conversion rate. There was an 11.35% drop in the checkout bounce rate and a 11.46% increase in average revenue per user (ARPU) after the improvements were formally implemented.

2.4.1. Spotahome: new features

There is no need for stuffiness or anxiety during A/B testing. Spotahome, an online booking company, hosts Hotjar Watch Parties to keep things light and entertaining. Video conferencing allows developers to digitally gather and review user interactions with new features.

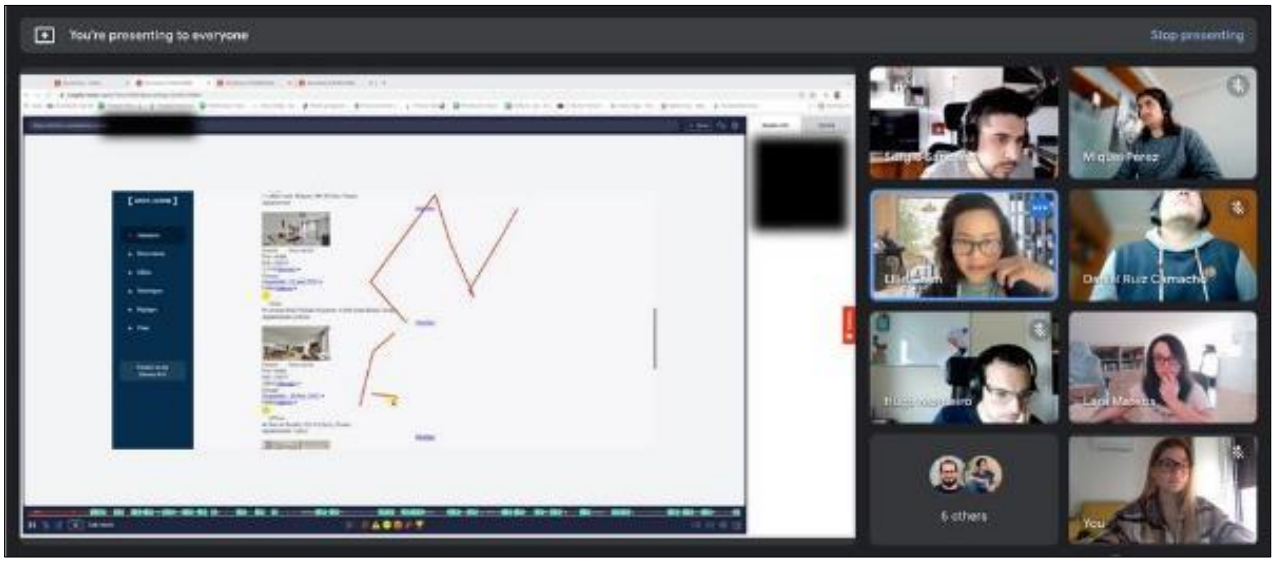

Figure 4 Video conferencing

At pizza parties, the Spotahome team watches Hotjar recordings to gauge the success of new features. For instance, engineers discovered a faulty button while reviewing user experience recordings of their new sign-up flow. They might have scowled and moaned when they noticed it, but that was the exact time they caught an issue that could have cost them conversions.

2.4.2. The Good: mobile homepage

Experts in conversion rate optimisation for online markets Swiss Gear is a retailer of outdoor, travel, and camping products; The Good was tasked with increasing mobile conversion rates for this customer. The Good first used Google Analytics to find out when, where, and why users left the website in order to identify any problems or obstacles. In order to draw attention to visitors' click and scroll behaviours, the company used Hotjar Heatmaps, which are free forever, and this quantitative data served as a starting point. Then, they used the qualitative data gleaned from Hotjar Recordings to speculate on possible enhancements based on user behaviour. The Good put their theories to the test by recording and analysing user behaviour using heatmaps following each test.

2.4.3. Re:member: application form

Re:member, a credit card firm based in Scandinavia, detected an issue with their funnel. According to Google Analytics, a large number of eligible leads came from affiliate sites, however they abandoned the cart before registering for a credit card.
Steffen Quistgaard, a senior marketing specialist at re:member, selectively retrieved session records and click maps from affiliate sites using Hotjar filters. Quistgaard spotted people browsing to the homepage, scrolling up and down, and hovering over the advantages section when studying these sessions. Quistgaard concluded from these patterns of behaviour that leads were reluctant and may benefit from more convincing content on the form.

2.4.4. Every.org: donation flow

An interesting observation was made by Dave Sharp, Senior Product Designer at Every.org, while reviewing session records. He noted a spike in angry clicks, which can be defined as a succession of repeated clicks within a brief period, on their donation form. Dave, after seeing a lot of sessions, came to the conclusion that the two calls to action on the form were making users very confused and angry. The donation flow was redesigned by Every.org to consist of two pages, each with a single call to action button. After that, it was compared to the initial version. The A/B testing method resulted in a staggering 26.5% increase in conversions.

3. AI for A/B testing: how artificial intelligence optimizes your marketing experiments

3.1. Introduction: Experimentation Unleashed with A/B Testing

The huge world of marketing is full of opportunities, and the key to unlocking those opportunities is experimentation. A/B testing is one of the most potent marketing strategies available. This easy-to-implement technique compares two versions of a marketing asset (web page, email, etc.) to see which one the target audience prefers. A/B testing has long been an essential part of successful marketing since it allows for the use of hard facts rather than speculation to direct decisions. On the other hand, AI is a game-changer that is altering A/B testing in this new digital age. Join me on a thrilling adventure as we delve into the ways artificial intelligence is revolutionising A/B testing and boosting marketing campaigns for companies all around the globe.

3.2. A/B Testing: Understanding the Basics

We need to understand A/B testing's foundational principles before we can plunge into the AI-powered revolution. The essence of A/B testing is that it is a scientific approach to advertising. It entails contrasting two iterations of a marketing component, such a landing page, email, or ad, to ascertain which one the intended audience responds to better. Because it is so easy to use, A/B testing is quite powerful. It is essential for successful marketing since it removes uncertainty and gives factual data for making informed decisions.

3.3. The Power of AI in A/B Testing

Every industry has been profoundly affected by the advent of artificial intelligence (AI), and marketing is no different. Adding AI to A/B testing elevates it to a new level. Data patterns and trends can be discovered in real-time via A/B testing platforms driven by AI, which is able to process enormous amounts of data. The capacity to delve farther into complex analysis allows for better forecasting and decision-making.

3.4. The Benefits of AI in A/B Testing

- Improved Speed and Scope: Using AI, it is possible to test numerous variables at once, which speeds up the process without sacrificing accuracy.
- Complex Data Analysis: Deep cause-and-effect correlations between variables can be uncovered by AI systems that discover nuanced patterns in data.
- Predictive Capabilities: Artificial intelligence (AI) can learn from past data to predict future trends, which helps marketers to foresee customer actions and conversion rates.

4. Real-World Examples of AI and A/B Testing

To put the theoretical benefits into practice, let's look at some real-life instances of organisations that have used AI and A/B testing in their marketing efforts.

4.1. Example 1: Amazon's AI-Powered Personalization

Amazon uses artificial intelligence and split testing to tailor consumers' buying experiences. Amazon increases conversion rates by personalising the shopping experience for each consumer by constantly evaluating new product recommendation algorithms.

4.2. Example 2: Netflix's Content Recommendation Engine

By combining AI with A/B testing, Netflix is able to fine-tune its content suggestion engine. To improve consumer retention and performance in the streaming sector, Netflix does testing on several AI-driven recommendation algorithms to make sure the correct material gets the right viewer.

4.3. Example 3: HubSpot's Email Marketing Optimization

To enhance their email marketing strategies, HubSpot employs A/B testing that is powered by AI. If a company wants to increase their open rates and conversions, they can use HubSpot to test different email features like subject lines and calls to action to find the ones that work best.

4.4. Implementing AI in A/B Testing: A Step-by-Step Guide

There is a simple way to incorporate AI into A/B testing, despite how intimidating it sounds. I've laid down all you need to know in this detailed tutorial:

4.4.1. Step 1: Define Your Objectives

Before you get into A/B testing powered by AI, make sure you clearly identify the objectives of your marketing experiment. Whether your goal is more interaction, more purchases, or improved click-through rates, knowing this will help you focus your efforts.

4.4.2. Step 2: Choose the Right AI Tool

It is critical for your company to choose the appropriate AI tool. Think at things like price, integration options, customer service, and how easy it is to use before making a final decision.

4.4.3. Step 3: Create Different Variants

Get your goals and resources in order, and then make some variations for your A/B test. Find out what your audience responds to most effectively by testing one variable at a time.

4.4.4. Step 4: Analyze Your Results

Examine the outcomes of your split testing with the help of your AI tool. You can learn more about your audience's tastes with the help of AI's pattern recognition capabilities.

4.4.5. Step 5: Apply Your Learnings

Incorporate the findings from your AI-powered A/B tests into your advertising campaigns. Because A/B testing is iterative, getting the most out of it requires constant tweaking.

4.5. Best Practices for AI-Powered A/B Testing

To make the most of AI in A/B testing, consider these best practices:

- Focus on One Variable: Perform tests on a single component at a time to guarantee precise results.
- Aim for Statistical Significance: Collect enough information to draw valid conclusions.
- Consider the Human Element: Consider the preferences of your audience when interpreting the data.
- Embrace Data-Driven Decision-Making: Use a lot of data to make AI work for you.
- Keep Testing: The success of A/B testing depends on continuous testing and learning.

4.6. Future-Proof Your Marketing Strategy with AI and A/B Testing

Maintaining a leading edge is critical for companies to thrive in today's fast-paced digital landscape. A/B testing greatly benefits from the incorporation of cutting-edge technology, such as AI. Through the provision of quicker and more precise information, AI enhances marketing experiments, hence transforming the decision-making process for marketers. Nashville and beyond firms may use AI marketing to cement their strategy for the future, boost consumer engagement, and achieve extraordinary growth. We at Digital Resource have a team of specialists that are well-versed in the potential of artificial intelligence marketing in Nashville. For marketing campaigns that achieve results you never imagined possible, let us show you how to harness AI's full potential. Initiate the process right now to take your marketing to the next level.

5. Conclusion

The use of A/B testing allows for feature adoption decisions to be based on data. It finds extensive application in many sectors and by major tech firms like Microsoft, Google, and Meta. Subjects of A/B tests, methods for designing and executing A/B tests, and reported open research topics were all part of this comprehensive literature analysis. The most typical things to test with A/B testing, according to our observations, are algorithms, visual elements, and changes to workflows or processes. The most common application domains for A/B testing are web, search engine, and e-commerce. Traditional split-tests with two versions are the gold standard when it comes to A/B testing, with engagement measures like conversion rate or impression count serving as metrics to measure the potential of the A/B variants. There are a few of main papers that show interest in bootstrapping, and hypothesis tests for equality testing

are commonly used to analyse A/B test findings. We came up with the terms "concept designer," "experiment architect," and "setup technician" to describe the three jobs that stakeholders play when creating A/B tests. Evaluating the performance of A/B tests empirically is the gold standard. In addition to the primary A/B metrics, data about the product or system as well as data focused on users are primarily collected in order to analyse the A/B test results more thoroughly

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Abhishek, Vineet, Mannor, Shie, 2017. A nonparametric sequential test for online randomized experiments. In: Proceedings of the 26th International Conference on World Wide Web Companion. In: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 610–616. http://dx.doi.org/10.1145/3041021.3054196.

[2] Agarwal, Deepak, Long, Bo, Traupman, Jonathan, Xin, Doris, Zhang, Liang, 2014. LASER: A scalable response prediction platform for online advertising. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM '14, Association for Computing Machinery, New York, NY, USA, pp. 173–182. http://dx.doi.org/10.1145/2556195.2556252.

[3] Alfaro-Flores et al., 2021 Alfaro-Flores Rafael, Salas-Bonilla José, Juillard Loic, Esquivel-Rodríguez Juan Experiment-driven improvements in human-in-the-loop machine learning annotation via significance-based A/B testing 2021 XLVII Latin American Computing Conference, CLEI (2021), pp. 1-9, 10.1109/CLEI53233.2021.9639977.

[4] Aharon, Michal, Somekh, Oren, Shahar, Avi, Singer, Assaf, Trayvas, Baruch, Vogel, Hadas, Dobrev, Dobri, 2019b. Carousel ads optimization in yahoo gemini native. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 1993–2001. http://dx.doi.org/10.1145/3292500. 3330740

[5] Bakshy, Eytan, Frachtenberg, Eitan, 2015. Design and analysis of benchmarking experiments for distributed internet services. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 108–118. http://dx.doi.org/10.1145/2736277.2741082.

[6] Budylin, Roman, Drutsa, Alexey, Katsev, Ilya, Tsoy, Valeriya, 2018. Consistent transformation of ratio metrics for efficient online controlled experiments. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18, Association for Computing Machinery, New York, NY, USA, pp. 55–63. http://dx.doi.org/10.1145/3159652.3159699.

[7] Cai, Tianchi, Cheng, Daxi, Liang, Chen, Liu, Ziqi, Gu, Lihong, Xie, Huizhi, Zhang, Zhiqiang, Zeng, Xiaodong, Gu, Jinjie, 2021. LinkLouvain: Link-aware A/B testing and its application on online marketing campaign. In: Jensen, Christian S., Lim, Ee-Peng, Yang, De-Nian, Lee, Wang-Chien, Tseng, Vincent S., Kalogeraki, Vana, Huang, Jen-Wei, Shen, Chih-Ya (Eds.), Database Systems for Advanced Applications. Springer International Publishing, Cham, pp. 499–510.

[8] Cámara, Javier, Kobsa, Alfred, 2009. Facilitating controlled tests of website design changes: A systematic approach. In: Gaedke, Martin, Grossniklaus, Michael, Díaz, Oscar (Eds.), Web Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 370–378.

[9] Deng, Alex, Li, Yicheng, Lu, Jiannan, Ramamurthy, Vivek, 2021. On post-selection inference in A/B testing. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. KDD '21, Association for Computing Machinery, New York, NY, USA, pp. 2743–2752. http://dx.doi.org/10.1145/3447548. 3467129.

[10] Deng, Alex, Lu, Jiannan, Litz, Jonthan, 2017. Trustworthy analysis of online A/B tests: Pitfalls, challenges and solutions. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17, Association for Computing Machinery, New York, NY, USA, pp. 641–649. http://dx.doi.org/10.1145/3018661. 3018677.

[11] Deng, Alex, Xu, Ya, Kohavi, Ron, Walker, Toby, 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. WSDM '13, Association for Computing Machinery, New York, NY, USA, pp. 123–132. http://dx.doi.org/10.1145/2433396.2433413

[12] Esteller-Cucala, Maria, Fernandez, Vicenc, Villuendas, Diego, 2019. Experimentation pitfalls to avoid in a/b testing for online personalization. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. In: UMAP'19 Adjunct, Association for Computing Machinery, New York, NY, USA, pp. 153–159. http://dx.doi.org/10.1145/3314183.3323853.

[13] Fabijan, Aleksander, Arai, Benjamin, Dmitriev, Pavel, Vermeer, Lukas, 2021. It takes a flywheel to fly: Kickstarting and growing the a/b testing momentum at scale. In: 2021 47th Euromicro Conference on Software Engineering and Advanced Applications. SEAA, pp. 109–118. http://dx.doi.org/10.1109/SEAA53835.2021. 00023.

[14] Fabijan, Aleksander, Dmitriev, Pavel, McFarland, Colin, Vermeer, Lukas, Holmström Olsson, Helena, Bosch, Jan, 2018. Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. J. Softw.: Evol. Process 30 (12), e2113. http://dx.doi.org/10.1002/smr.2113, arXiv:https: //onlinelibrary.wiley.com/doi/pdf/10.1002/smr.2113, e2113 JSME-17-0210.R2.

[15] Galster, Matthias, Weyns, Danny, 2016. Empirical research in software architecture: How far have we come? In: 2016 13th Working IEEE/IFIP Conference on Software Architecture. WICSA, IEEE Press, Los Alamitos, CA, USA, pp. 11–20. http://dx.doi. org/10.1109/WICSA.2016.10.

[16] Giaimo, Federico, Andrade, Hugo, Berger, Christian, 2020. Continuous experimentation and the cyber–physical systems challenge: An overview of the literature and the industrial perspective. J. Syst. Softw. 170, 110781. http://dx.doi.org/10.1016/j.jss. 2020.110781.

[17] Gomez-Uribe, Carlos A., Hunt, Neil, 2016. The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6 (4), http://dx.doi. org/10.1145/2843948.

[18] Ju, Nianqiao, Hu, Diane, Henderson, Adam, Hong, Liangjie, 2019. A sequential test for selecting the better variant: Online A/B testing, adaptive allocation, and continuous monitoring. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19, Association for Computing Machinery, New York, NY, USA, pp. 492–500. http://dx.doi.org/10.1145/3289600.3291025.

[19] Kaplan, Jared, McCandlish, Sam, Henighan, Tom, Brown, Tom B., Chess, Benjamin, Child, Rewon, Gray, Scott, Radford, Alec, Wu, Jeffrey, Amodei, Dario, 2020. Scaling laws for neural language models. arXiv:2001.08361.

[20] Keele, Staffs, et al., 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, Ver. 2.3 EBSE Technical Report. EBSE. Kharitonov, Eugene, Drutsa, Alexey, Serdyukov, Pavel, 2017. Learning sensitive combinations of A/B test metrics. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. WSDM '17, Association for Computing Machinery, New York, NY, USA, pp. 651–659. http://dx.doi.org/10.1145/3018661. 3018708.