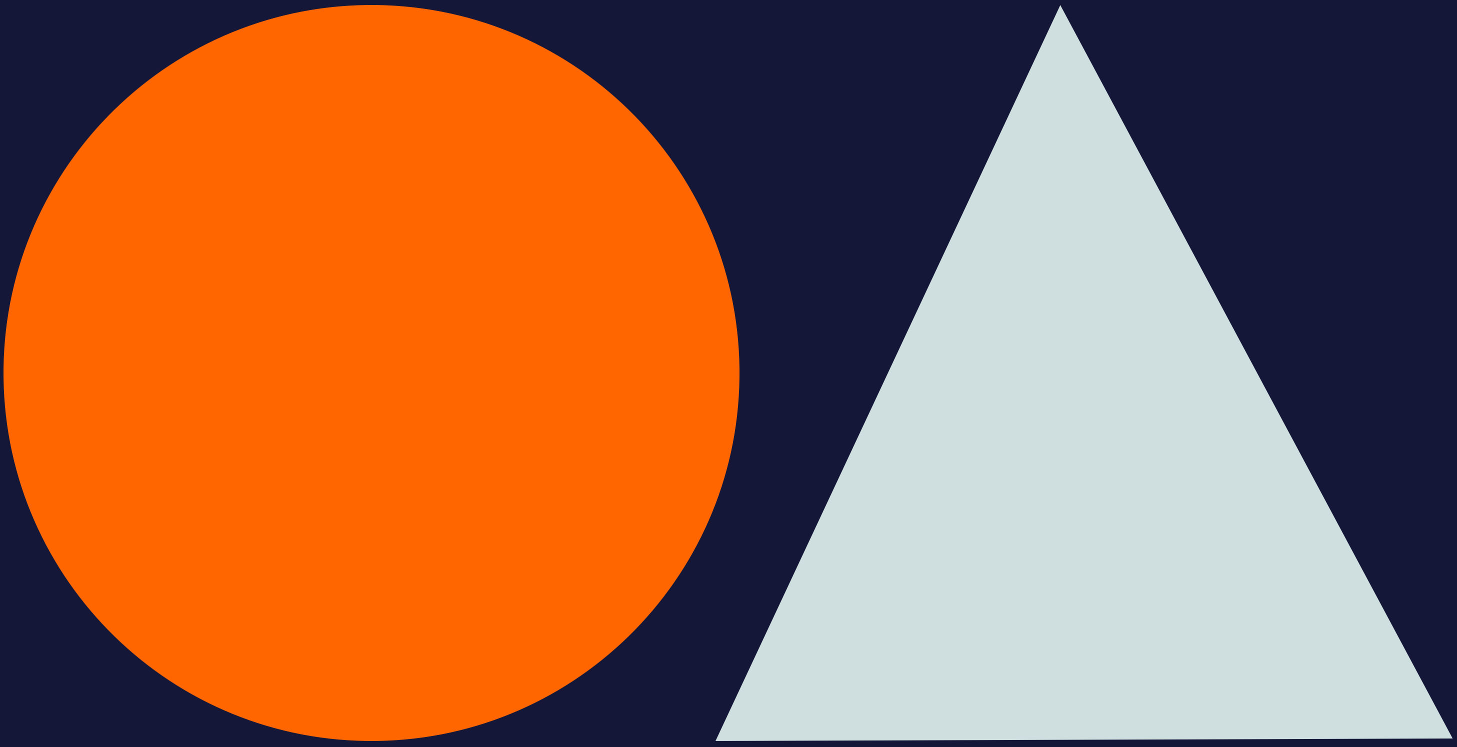
The Data Driven Marketer's Guide to A/B Testing

outlier

CHAPTER I

Introduction

As a data driven marketer, gone are the days when you can throw money at programs and simply hope for a favorable outcome. Today you need a scientific data driven approach for making and validating your decisions. Key to this approach is an experimentation mindset that allows you to not only test your marketing hypothesis, but also generate the data necessary to measure and benchmark your choices so that you can continuously measure and improve results.

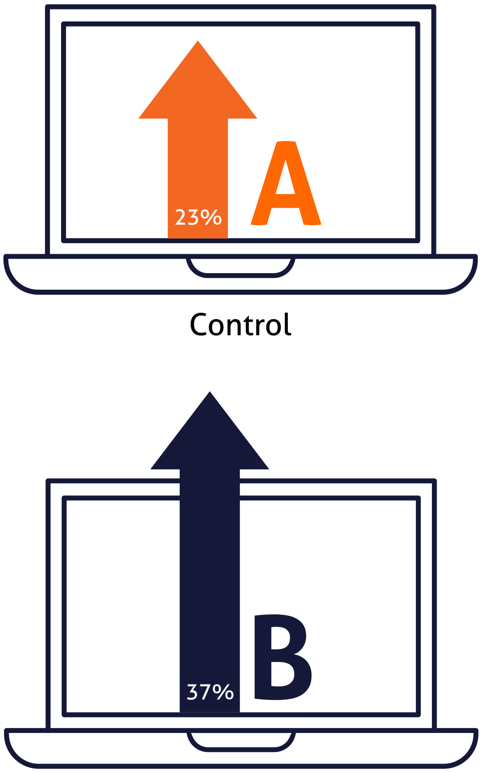


Few tools better encapsulate the data driven experimentation mindset than A/B testing. Simply put A/B testing is one of the most effective ways to both produce and leverage the data necessary to make and measure decisions.

What is A/B Testing?

A/B testing, also known as split testing, is an experiment conducted to test two versions of something (variable A & B) at the same time, allowing you to observe which variable, A or B, delivers the best result. The current version, version A, is referred to as the *control*, and version B, modified in some respect, is the *treatment*.

While A/B testing is most commonly associated with the testing of websites and app's, A/B Testing is and can be used in a variety of industries and applications, everything from product design to marketing.



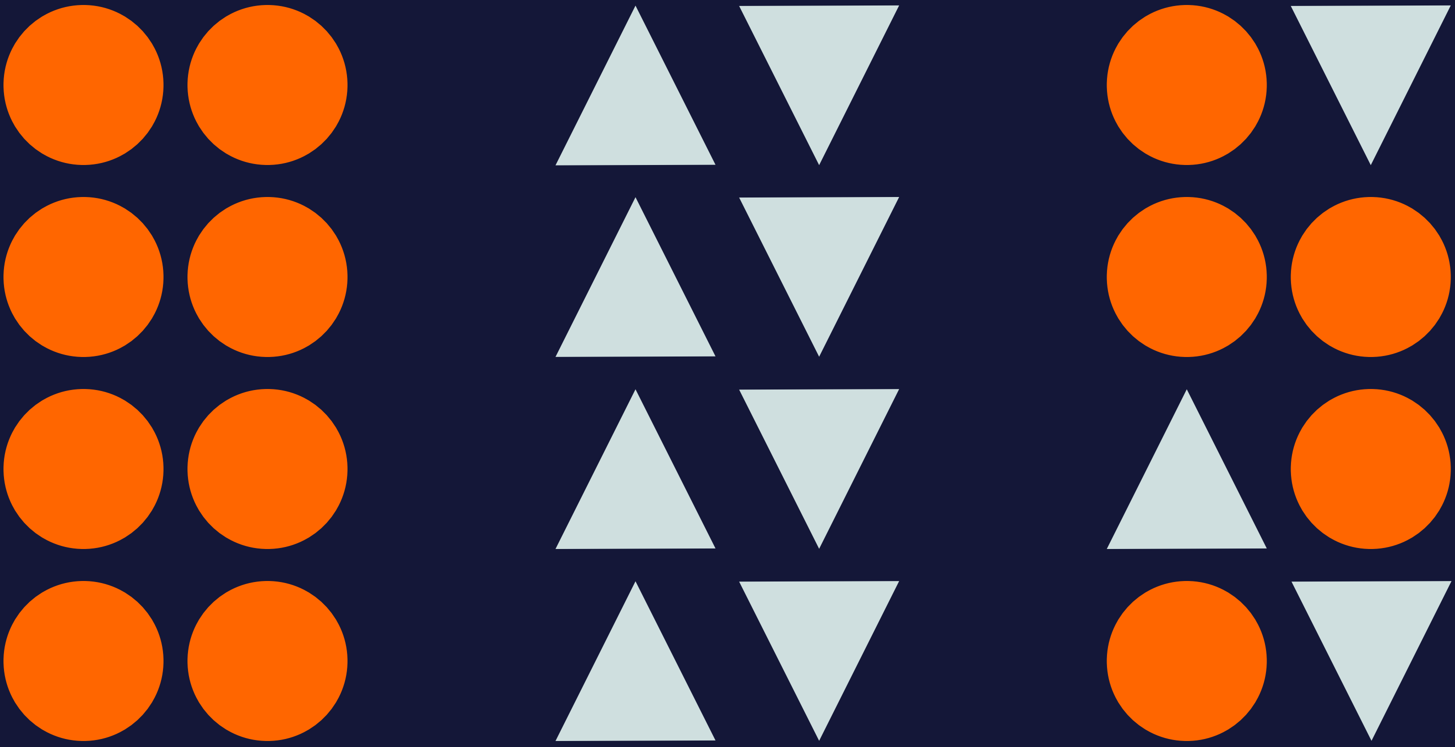
How Does A/B Testing Work?

Conducting an A/B test requires that you first decide what it is you want to test and then how you want to evaluate its performance. For example lets say you are in charge of Marketing for a direct to consumer woman's fashion company. You decide to test changing the location of your click to join button for your customer loyalty program. Your metric is the number of consumers that click the button. To run the test you show two sets of users (assigned at random when they visit the site) the different versions, where the only variable is the location of the 'join the loyalty program' button. In this case performance is measured by which location prompts the most clicks.

CHAPTER II

Testing is Hard

While the example above captures the basic premise of A/B testing, for most organizations today, leveraging A/B testing requires a more complex approach and understanding than the simple example above.



A. Control & Treatment

When discussing A/B testing the terms Control & Treatment are frequently used. The control group represents the group of subjects that are set aside and do not receive the new testing treatment, thus allowing you to measure the impact of the independent variable being tested.

B. Population Sampling

When running an A/B Test, the most challenging question to answer is how many observations do the control and treatment groups each need to be statistically significant. Why? A threshold number of tests must be met to make the test result trustworthy. *Spoiler Alert:* Generally speaking, the answer is a lot more than you think!

Depending on your company, sample size, and the A/B test variables, statistically significant results could take hours, days or weeks -- and in theory you should not restrict the time in which you're gathering results. However in the real world, your world, when does any decision come with unlimited time?

When time is of the essence, Statisticians have developed a numerical technique, called Power Analysis (also known as a sensitivity analysis), to determine the number of observations needed. Power Analysis relies on two inputs, the size of the change you expect to measure and the numerical confidence you want in your results. For example, you expect a 5% change in customer behavior with 95% confidence. We won't go into the mathematics behind Power Analysis (there is a great example here) because there are plenty of off the shelf software products you can use to calculate it for you.

What do you do when you don't have enough observations to make your results statistically significant? You will need to relax one of your constraints, either larger changes (20% instead of 5%) or a lower level of confidence (80% instead of 95%).

View a deep-dive into the mathematics behind Power Analysis here.

c. Multi Variate Testing

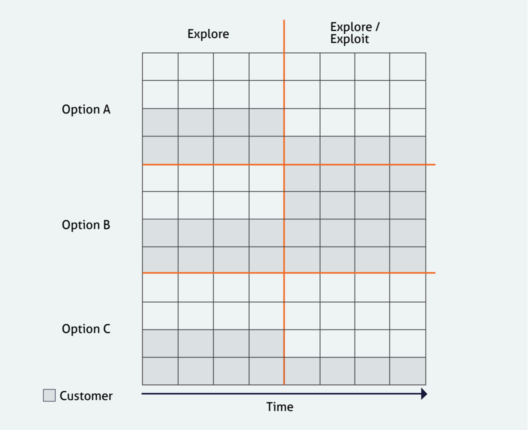
In today's data driven marketing world, rarely are key decisions simply an A/B choice. Instead you probably have tens or hundreds of variables to choose from. When faced with this scenario simple A/B testing, testing each variable against the control, takes too much time. When this is the case Multi-Variate Testing is more appropriate. In a Multi-Variate test (also known as a Multi-armed Bandit Test) you test multiple variables at the same time and quickly move to the most effective option. This is accomplished by dividing the problem into two parts:

Explore: This phase tests the possible options to see which performs best.

Exploit: This phase uses the best option to get the best performance.

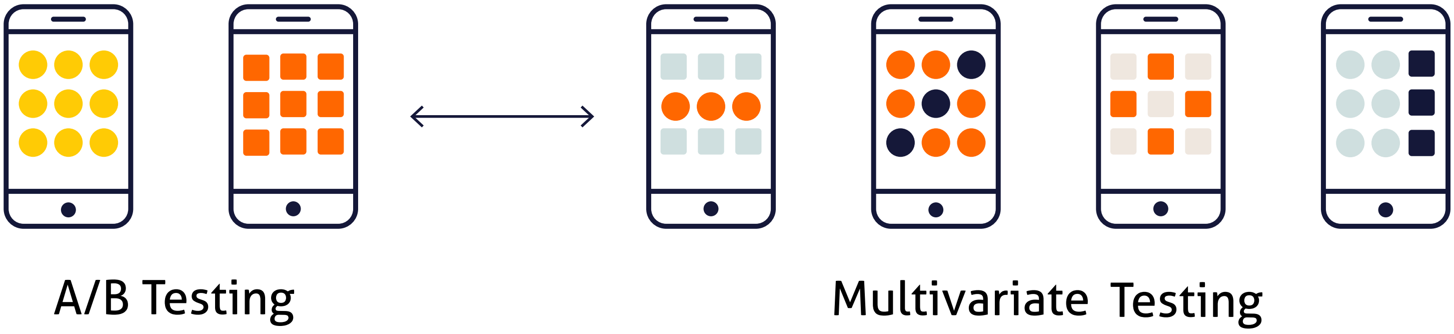
There are a few different ways that you can run Multi-Variate tests, but here I will focus on a common method known as Epsilon-Greedy, which runs both Explore and Exploit at the same time! It does this by dividing your customers into two groups (Explore and Exploit). Typically, your Explore group will be 20% of your activity and in that group all available options will be tested side by side. The Exploit group will be the other 80% of your activity and use whichever option is performing best in the Explore tests.

Of course, at first, you have no best option so all activity will be used for Explore. However, as soon as one option shows progress it will be used for both Explore and Exploit. For this reason, the progression of your Epsilon-greedy test will look like the following:



As you can see, the test initially tests all three options at the same time, but as soon as it's clear that Option B is best it moves more customers to that option.

The advantage of Multi-Variate testing is that you can avoid wasting time since it will identify and use the best options on its own. However, you need to have completely interchangeable options for this testing technique! That means it works well for testing email subject lines and colors on a website, but will be hard to apply to things like pricing and product features.



CHAPTER III

The Responsible Way to Speed

"Our success at Amazon is a function of how many experiments we do per year, per month, per week, per day." – Jeff Bezos

When your marketing activities run on hundreds if not thousands of variables and time is of the essence, the only practical solution is running multiple tests simultaneously. In many cases simultaneous testing presents no additional challenges or issues, save the additional work. For example:

- If the groups of customers exposed to each test are mutually exclusive, so that no customers participating in Test 1 are also participating in Test 2.
- If the overlap between customers in Test 1 and Test 2 is very small (say 1% of all customers) so any error introduced should be minor.
- If the tests are of features so distinct and different that they cannot influence the same customer behavior(s).

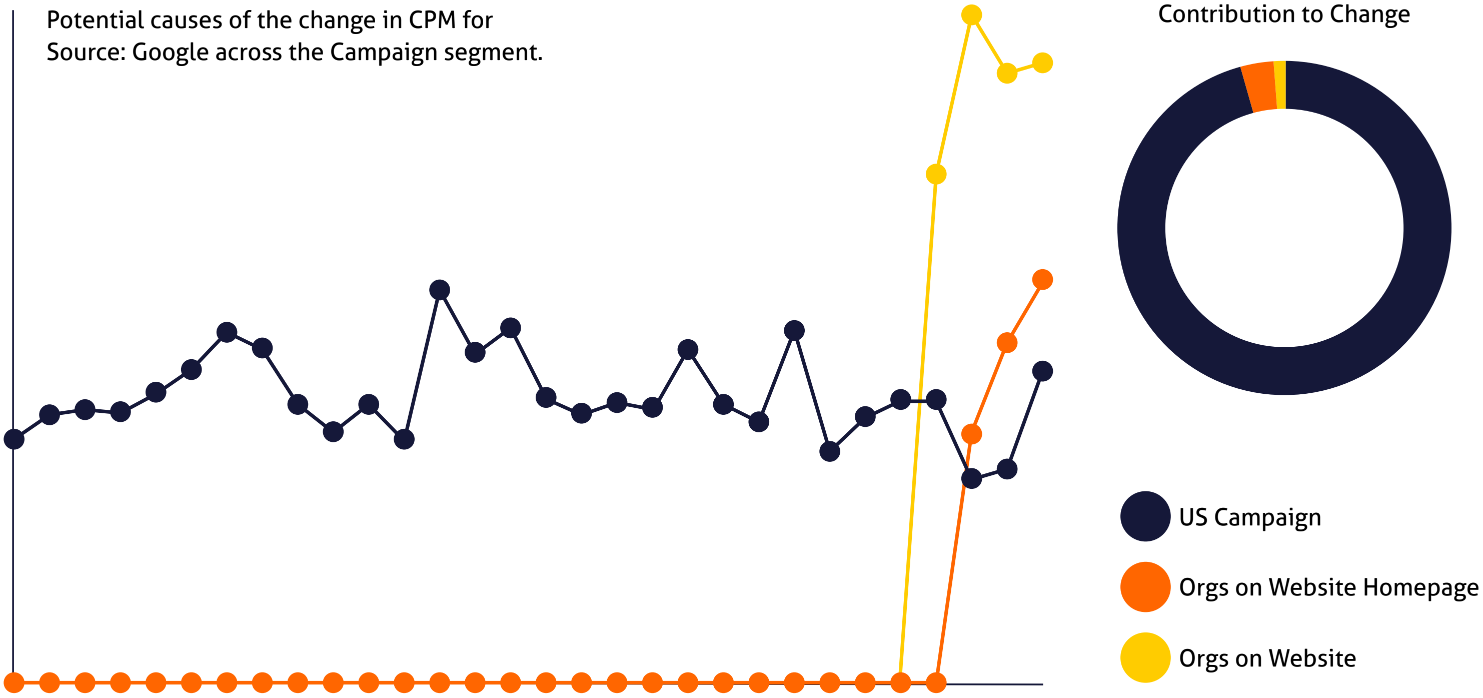
A great use case for simultaneous testing is Lookalike advertising. Once you have segmented your Lookalike audiences' the next step is to A/B test which audience performs better.

A simple way to do this is to leverage the default 'Events' tracking found in both Facebook and Google. For example if you are an e-commerce provider, Facebook's standard e-commerce relevant events include things like Add-To-Cart, Page-View and Purchase. To run a simple A/B test create Lookalike audiences from two versions of the same event, (for example the purchasing customers of two distinct but similar positioning products) then test to determine which Lookalike audience performs better.

Companies who leverage Outlier can track these campaigns in two ways. The first is if you set up two lookalike campaigns, Outlier analyzes the campaigns' performance on a daily basis via our Root Cause Analysis feature. For example you'd see in your daily story that "Lookalike campaign A" is contributing more to your KPIs, like page views. The second way is over time, Outlier will model the behavior expected from the A/B test campaigns and you can decide which one is best to optimize around.

If you need to run multiple tests but cannot meet one of those criteria, you will need to use your judgment. The argument for running tests simultaneously is that the danger in having error and bias in your test results is better than having no data to make the decision at all. Conversely, there is no point in running a test if you cannot clearly rely on the results.

The most dangerous scenario is testing with customers exposed to multiple tests; making it difficult to discern which features affected that customers' behavior! Was it Option A of Test 1 or Option B of Test 2 that caused the change in outcome? The more tests you are running the harder it might be to tell the difference.



Potential causes of the change in CPM for Source: Google across the Campaign segment.

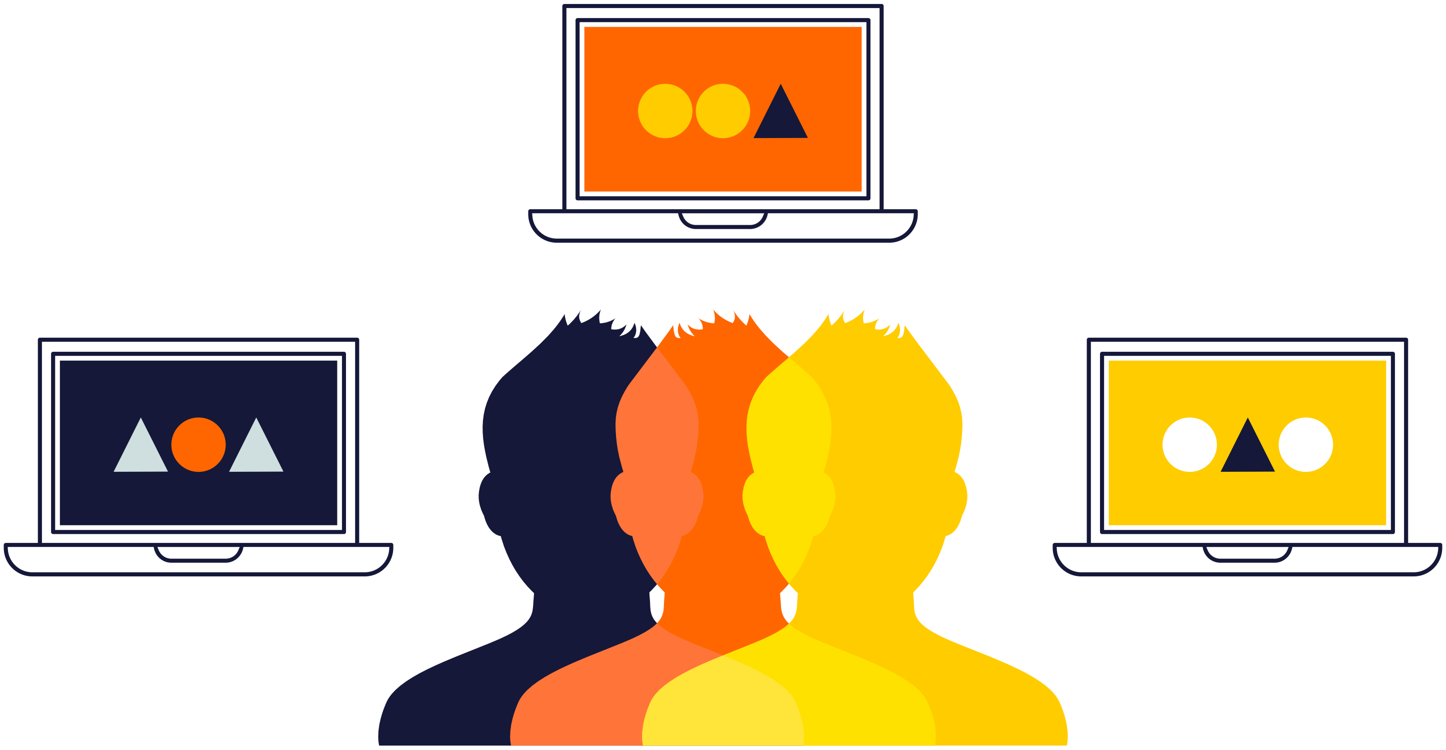
CHAPTER IV

Traps to Avoid

Whether you are running an A/B or Multi-Variate Test there are many common traps you can fall into that will cause your results to be misleading. These are some of the most common traps and how to avoid them:

1. **Small samples.** Yes, you may have 1 million customers but how many of these customers use the feature you are going to test? If you only have 100 customers using that feature you may not have a large enough sample to get reliable results from your A/B Test. Before running a test, be sure to understand the required sample size and that you have enough customer activity to create the observations you need.
2. **Vague Hypotheses.** Your test is designed to test something, but what is that something? If you aren't crystal clear on what you are testing and what the expected results are from the test then you won't be able to trust results. It is not as simple as "I think Option A will generate more revenue". You need to be specific in your test hypothesis or else you can't guarantee that other factors influenced the result. A good test might be "I think that changing our email subject line to X will increase the email option rates by Y%".
3. **Competing Tests.** If you run more than one test at a time with the same group of customers, your tests may be competing with each other. How will you know if the improvement you see in Option A from Test 1 is real or a result of Option B of Test 2? Running multiple tests can cause all sorts of data pollution when tests share the same customers.

That last trap is a doozy! I have no doubt that you want to run more than one test at a time, but if doing so jeopardizes all of your tests what should you do?



Competing tests can influence each other and cause data pollution.

··outlier

About Outlier

In today's data driven world making the right decision requires identifying the unexpected changes in your business so that you can make more informed decisions quickly.

Adding more dashboards to monitor your business simply isn't enough. What's required is the ability for your marketers, data scientist and analyst to perform deep analysis across all your business data. Outlier's Automated Business Analysis platform discovers hidden patterns and relationships that are impossible to find without significant investment in artificial intelligence and machine learning algorithms. Adding any integration only takes minutes, and each integration helps global organizations learn more about different aspects of their business.

Sign up for a custom demo to learn how Outlier can help you discover the data you need to take you're A/B testing to the next level. Find the unexpected with Outlier.

outlier