NAMOGOO

A B TESTING

Best Practices Learned from 100s of Tests on Large eCommerce Websites

INTRODUCTION

A/B testing It's an easy concept to grasp. Test two different versions of the same page, button, feature, or whatever. Monitor the success of version A or B. Wash, rinse, repeat. Pretty simple, in theory.

In practice, A/B testing is one of the most complex practices that an eCommerce operation can run. It's hard to put a time stamp on when and how A/B testing originated but we can comfortably say that it was first introduced by academic statisticians at the turn of the 20th century.

However, **unlike academic A/B testing, the A/B tests performed by online retailers are not performed in a laboratory environment, and are very difficult to control.** Every eCommerce operation is different, which makes it very hard to create an academic experiment whose results would be replicable across various brands and web properties.

At Namogoo, we help online retailers recover lost revenue by preventing unwanted ad injections from redirecting web visitors to other websites and negatively impacting conversion rates and revenue. However, to do that we must first test to see what percentage of a website's traffic is infected by ad injections, and demonstrate its impact on revenue. **As part of this process, my team and I have run hundreds of A/B tests on some of the leading websites in the Alexa top 100.**

Online retailers are like fingerprints, no two are the same. However, our rich and diverse experience has led us to identify the methodologies that have been proven to be successful by the retailers we work with and should be applied to any A/B test, regardless of the circumstances. Keep in mind, proper A/B testing requires awareness of all the little things impacting your tests, which make it difficult to attribute your conversion increase (or decrease) back to a specific test. In order to reduce this uncertainty, and statistical noise, I've compiled our comprehensive guide to A/B testing best practices.

Hope you enjoy the eBook and feel free to reach out adam.segal@namogoo.com

Adam Segal
Director, Customer
Solution Consulting
NAMOGOO

TABLE OF CONTENTS

CHAPTER 1

How to Create Better A/B Tests with A/A Testing

One practice that online retailers should employ to better understand their A/B tests is A/A testing. Despite having a lower profile and less hype, A/A testing can be an important extra step toward making sure that A/B tests are run correctly; and providing useful, actionable insights instead of just leading you astray with statistical noise.

A/A TESTING VS. A/B TESTING: WHAT'S THE DIFFERENCE?

A/B testing is a way to compare two versions of a single variable, and determine if changing that variable will result in a meaningful and significant change in outcomes.

An **A/A test is like an A/B test, except that both versions are identical**—there's no significant difference between the test version and the control version; at least there shouldn't be...

A/A Testing
A/B Testing
VS
BOTH VERSIONS ARE IDENTICAL
TWO VERSIONS OF A SINGLE VARIABLE

WHY RUN AN A/A TEST WITH IDENTICAL VERSIONS?

There are certain factors that can create statistical noise within an A/B test. This can lead you to draw false conclusions, and take actions based on misleading results. If you want to be scientific about your testing and improve accuracy you will require **multiple independent tests to be run, in order to ensure that no single individual test is producing misleading results.** However, that process can be impractical so an A/A test is the next best thing.

WHAT'S WRONG WITH RUNNING AN A/B TEST WITHOUT A/A TESTING FIRST?

One of the main purposes of the A/A test is to help identify if there are issues with the splitting methodology being used to create the two test segments for an A/B test. **Without running**

an A/A test first, we run the risk of having two groups that have significant differences in their characteristics or behaviors.

If the two groups are found to be too heterogeneous (significantly different) and non-representative of their sample, this could result in a "winner" being declared based on a false premise. If the test version with the highest conversion rate was being fed a completely different customer segment than the version that "failed," what did you really learn from the test?

Therefore, the A/A test is most commonly used to QA the technical setup of the A/B test and surface any issues.

WHEN TO RUN AN A/A TEST

When adopting new testing tools, it's important to ensure that those tools are splitting your traffic such that the likelihood of being placed in either the "A" or "B" group is the same for all visitors, regardless of what browser or device they're using, their geolocation, and so on. At Namogoo, the Customer Solutions Consulting (CSC) team, which I head, tests the impact of our solution on "infected users" browsing some of the world's largest eCommerce websites. We do this by implementing various testing methodologies. However, we almost always recommend an A/A test to "clean the pipes" and make sure everything is running smoothly.

> In fact, I recommend running an A/A test at least once a year to verify the integrity of the splitting mechanisms your conversion optimization tools are using.

For example, a short while ago, our team ran an A/B test for a well-known eCommerce subscription service. Due to technical issues, they were required to temporarily revert pages back from the changes we had made, causing the test to function like an A/A test. Surprisingly, both test groups responded differently despite being served identical pages.

We quickly noticed this behavior and were able to find and resolve the issue when we identified a discrepancy within the segmentation of their site traffic. Specifically, **it was related to iOS devices, which resulted in a split that created an unequal distribution of iOS users, placing more of them into the control group instead of the test group.** Given that the majority of visitors were using iOS devices, the outcome of this test did not accurately represent their actual customer population.

A/B TESTING BEST PRACTICES START WITH A/A TESTING

The cornerstone of an A/B test that yields meaningful information is a pool of randomly and equally distributed test subjects that exhibit homogeneous behaviors and characteristics. Relics from previous tools and tests can affect whether you actually get the right distribution of homogeneous groups in your A/B test.

I recommend running an A/A test for any new testing tool you're using, along with annual inspections to ensure that the testing groups identify and behave similarly (within reasonable variances).

By running an A/A test on a large enough sample of the population, you can feel confident that the test outcome is not being distorted by statistical noise, which can be difficult to detect without multiple rounds of testing. When unidentified statistical noise is allowed to influence test results, false conclusions may be drawn about which testing variant had the better conversion rate.

SOLVING IMPLEMENTATION ISSUES WITH A/A TESTING

Another example of the importance of A/A testing was discovered when I ran a test with a client who had recently brought in a dedicated team to work with the new testing tool they had acquired. My team assumed that the tool had been implemented correctly, and was providing accurate reporting to the analytics system.

Due to an unrelated technical issue, the client removed the Namogoo tag from the test variation while the test continued to run and report. With the tag removed from the variation, the A/B test effectively became an A/A test, and started reporting different results between the two variations, which is a red flag.

Upon investigation, the team discovered an implementation issue that caused the A/B testing tool to behave incorrectly. This issue affected every test the client had run, rendering them all invalid.

> Every decision made based on the test results provided through prior A/B testing had been made based on erroneous data. Only by inadvertently running the equivalent of an A/A test were the problems with the initial implementation revealed.

Decisions made as a result of running a test through the incorrectly deployed A/B testing tool were made based on erroneous data, and, upon realization, their impact was palpable.

There's no way to know if those A/B tests correctly identified the "winning" variations or not. When the test started running in de facto A/A mode, both the test group and the control group started or continued behaving differently, even though they were no longer being served different variations.

This showed that the testing tool was not calibrated correctly; if both the Control Group and the Test Group were being served the same variation, then both groups should have behaved in the same way.

CONCLUSION

When running an A/B test, our objective is to set a baseline by using a control group. Then we measure the effect of the changes we want to make, by comparing a version with those changes implemented against the unchanged version served to the control group. If the null hypothesis can be rejected without any changes made, this throws the entire testing mechanism into question.

How can we be sure that our A/B testing tool has established a control group that provides a fair and accurate representation? By performing an A/A test we're ensuring that the control group is representative of the demographics and behavior of your average visitor. If one group in an A/A test outperforms the other by a statistically significant margin—typically 5%—it means that the group populations are different.

For this reason, I recommend running an A/A test at least once a year to ensure that your test results will be valid.

CHAPTER 2

The Secret To Preventing A/B Testing Errors: It's Your Control Group

Faced with two different directions to go with your marketing campaign? Unable to decide between two different redesigns for your home page, both of which you're in love with? You know what these situations call for: A/B testing. Your customer is always right, so set up an A/B test and let them show you which path to take—that always works! Except when it doesn't.

When an A/B test leads you astray, it's usually because the control group wasn't giving you a representative sample to test your new treatment against. You have too many distinctive personalities, in your Control group — people who aren't like your regular visitor traffic, either in their demographics or their behavior.

This is why great minds wrestle with the question of how to put together a Control group that isn't going to throw your test results all out of whack.

A/B TESTS: WHAT COULD GO WRONG?

A/B testing can provide valuable insights into how better to engage with visitors to your website and capture their interest, but there's always a risk. A test that isn't set up properly can generate misleading, erroneous results. Acting on bad data from an A/B test defeats the purpose of A/B testing in the first place. When designing an A/B test, one of the most important goals should be to build a plan that reduces the likelihood of errors.

There are two main error types we want to avoid:

- Type I: Rejecting the null hypothesis when it is actually correct
- Type II: Accepting the null hypothesis when it is actually not correct

You're pregnant!
TYPE I ERROR
You're not pregnant!
TYPE II ERROR

The null hypothesis is the presupposition that there's no relation between the two variables you're testing. In other words, a specific change to the web page would make no difference to user behavior.

A Type I error, also known as a false positive, occurs when one declares the null hypothesis to be false, when it's actually true. A Type II error, a false negative, occurs when the null hypothesis is accepted while in reality, the variation resulted in a distinct, measurable change.

THE CHALLENGE OF AVOIDING ERRORS IN A/B TESTING

Because we're dealing with statistics and limited sample sizes, there's no way to completely eliminate the chance of errors occurring, but there are ways to minimize the risk. In particle physics for example, the threshold for rejecting the null hypothesis is extreme to say the least. The discovery of the Higgs Boson was measured with an alpha value of 0.0000003, a 1 in 3.5 million chance that the data was observed by chance. But we don't have the time, resources, or hadron colliders of CERN, so, what's the best way to avoid these two types of errors?

One idea that sometimes gets floated is using two control groups. You might assume that this would ensure greater accuracy and reduce the likelihood of these errors occurring, but as we'll soon see, **having two control groups can actually increase the chances that these types of errors will manifest.**

> Oftentimes, a company will run an A/A test to verify that their control population is homogeneous in characteristics and in behavior. This would ensure a fair A/B test, and can help determine the sample size necessary to run an accurate one.

In an A/A test, both versions of the web page served up to visitors are identical. In theory, both versions should elicit statistically similar actions from visitors. If you get the expected results from an A/A test, then you should be able to feel confident that your A/B test will produce accurate, meaningful data.

APPROACHES TO USING 2 CONTROL GROUPS

This is where the idea of using two control groups for an A/B test comes in. An A/A/B test, the thinking goes, would allow us to simultaneously run a test and ensure that its outcome is not being impacted by outliers or anomalies in the control group.

There are two different approaches we could take with A/A/B testing.

Approach One: Two A Groups Validating Each Other

Let's say we set up an A/A/B test with the assumption that the second control group will validate the first one. If the Test Group performs differently from both Control Groups, then we would consider the results to be accurate; if not, we would assume there were issues with the composition of one or both of the control groups and disregard the results.

If both Control Groups yield statistically similar results, we would proceed under the assumption that our control groups were properly established and that the test had generated accurate results.

Approach Two: Two A Groups Enter, One A Group Leaves

Another approach would be to look at the second Control Group as a "backup", if we decide that the first Control Group is an unrepresentative population.

That's right, this is the "Thunderdome" approach: you pit two Control Groups against each other and only one of them gets counted. After you run the test, you look at the results of your two Control Groups, and make a determination as to which one better represents your mean visitor traffic.

THE PROBLEMS WITH A/A/B TESTING

Sounds good on paper, right? In reality, there are a few different ways this could play out. Lucille Lu, a data scientist at Twitter, crunched the numbers on A/A/B testing and found that neither method produced more accurate results than simply using a larger pool of visitor traffic in a single Control group.

Out of all possible scenarios, there were only a minority in which A/A/B testing did reduce error rates. In all the rest, either the error rates were skewed in a different direction, or the test was subjected to confirmation bias errors. If a customer's data team views their primary responsibility as exposing vendor claims to be false, then confirmation bias has a decent head start in the race.

Here's what can happen when you try the first approach to A/A/B testing, where we discard the results when the two Control Groups don't agree with each other. In a situation where the null hypothesis is correct, A/A/B testing can reduce Type I errors by more than 4%, but at a cost. The added constraint of measuring against two control groups causes a sharp increase in the rate of Type II errors—as much as 19%, depending on how well the B treatment is actually performing.

Using the second approach, where we make a later determination as to which Control Group is "better", and compare only the results from that Control Group with the Test Group, there's a huge risk of confirmation bias. We would be choosing which Control Group to use after the results of the test have already come in. At this point, it would be tempting to select the group that confirms the results we were hoping to get.

Even if we assume a total absence of human bias, and that some objective analysis can be found to choose the better control group, there are still problems with this approach.

Lu's statistical analysis shows that in cases where there is such a sufficient divergence between the two control groups that a better one can be identified, you would actually get a group that better represents the true mean by pooling the two control groups together rather than choosing one over the other.

In the real world, the difference would be even more stark. For example, when asking children who drank Coke to select a color from an array. The Control Groups would be made up of children who did not drink Coke and requesting them to select a color from a fixed array. If results are different for the two Controls, which is the control you should use?

In strictly controlled laboratory environments the answer would be a simple one. However, in real life, and when people are involved, several parameters that we may not be aware of can influence results like age, education, social-economical background, color blindness, etc.

In theory, we would expect a randomly selected control group to remain consistent, regardless of how you slice it up. In reality, splitting a control group is counterproductive to the objective of minimizing bias and errors when running controlled, randomized experiments.

CONCLUSION

Ultimately, **when subjected to a rigorous statistical analysis, the idea of using two control groups in an A/B test just doesn't hold up as a viable way to reduce error rates.** While there are different ways to approach the implementation of A/A/B testing, they only produce more accurate results under very specific conditions, and can expose you to an even greater risk of false negatives and results tainted by human bias.

Using two control groups may make intuitive sense as a way to impose greater accuracy on an A/B test, but the statistical analysis shows this to be an illusion.

On the bright side, that same analysis shows that eCommerce companies that have sufficient traffic to perform an A/A/B test have an even better, more accurate alternative; which is to simply pool what would have been your two Control Groups into a single, larger control group.

Pooling groups gets you a better estimate of the true mean than selecting the better half of a split group does. This means the results you'll get from an A/B test with a larger Control Group are more likely to be accurate and useful for making marketing decisions. It also relieves you of the requirement to define a method of choosing the "better" group, without letting your subjective biases influence your choice. In the next chapter, we'll examine how to avoid one of the biggest reasons A/A/B tests fail - cross contamination.

CHAPTER 3

How to Avoid Cross-contamination in A/B Testing

Scientists have it pretty good when they need to run rigorous and sterile tests. To make sure test results remain uncontaminated by external factors, they have autoclaves, cleanrooms, coveralls, and all sorts of equipment to make sure the test isn't being skewed by anything from outside—not even the tiniest particle of dirt.

On the other hand, **internet marketers often struggle to carry out basic A/B tests without some sort of cross-contamination throwing off the results.** Unfortunately, when you ask the IT department if there's any way to clean up your visitor traffic by routing it through an autoclave, they look at you like you're crazy.

Cross-contamination, also known as data pollution, is a frequently encountered problem in A/B testing. These tests are based on the premise that your two different groups of traffic, A and B, are going to have nearly identical composition of their constituent members in terms of demographics, origin, and behavior. When they arrive at your website, each group should be presented with the exact same content and experience until they are diverted to the A and B versions of the page you are testing.

> Cross-contamination happens when a visitor clicks on some different content, has an unplanned experience, or makes some other impressions in between arriving at your site and landing on the A or B test pages.

In other words, they're leaving the "flow" to which the test is supposed to be limited, then coming back to the test with some ideas in their head that may influence the behaviors your test is evaluating. This can happen when the test is not deployed to your entire website, or when the visitor needs to access third-party information before they can complete part of the conversion process.

HOW CROSS-CONTAMINATION OCCURS?

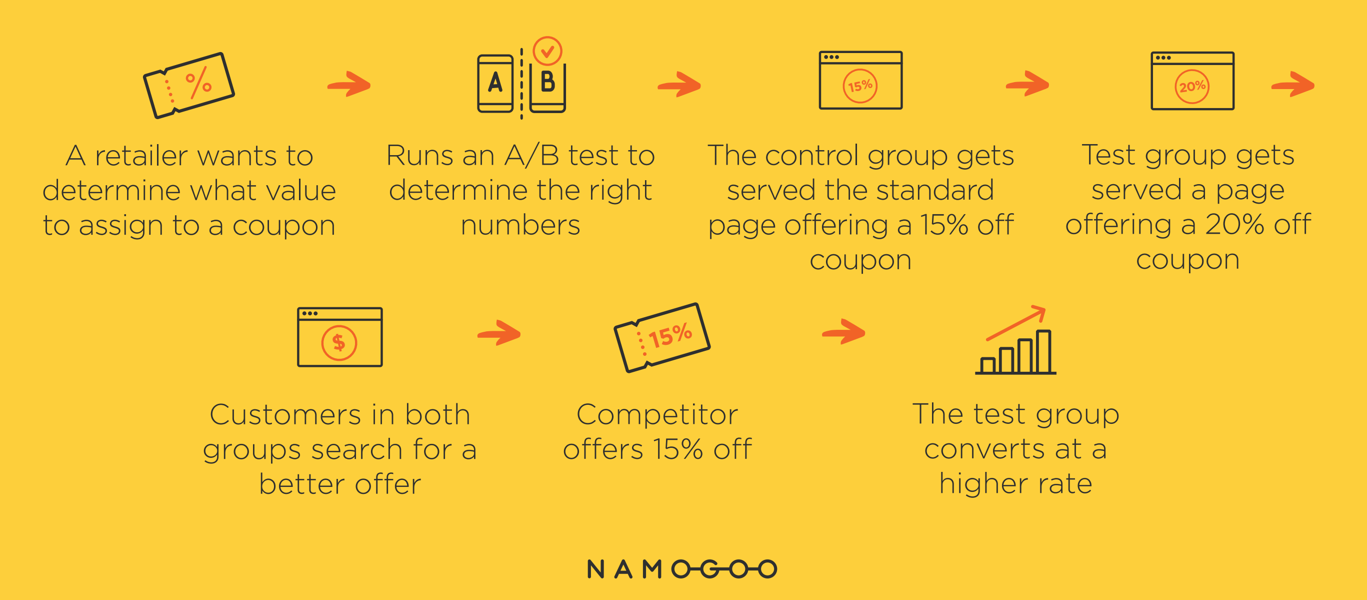
The problem of cross-contaminated results in A/B testing probably occurs more often than expected. It can be very difficult for organizations to identify when, and in what manner, cross-contamination may have occurred. The results won't necessarily look skewed or incorrect after the fact, especially if they skew in the direction that the organization was hoping or expecting to see.

Here's an example of how cross-contamination typically occurs. Let's say a retailer wants to determine what value to assign to a coupon, in order to motivate customers to actually use it. They run an A/B test to determine the right numbers. The control group gets served the standard page offering a 15% off coupon, and the test group gets served a page offering a 20% off coupon.

Both groups get served the page fairly early in the conversion process. Then, some customers in both groups see the coupon offer and start Googling around to see if this is the best deal they're going to get. It just so happens that a competitor is offering their own 15% coupon at the same time.

The customers in your test group see that your 20% off coupon is a better deal, so they place their orders using it. When you get your test results back, you see that the test version was the big winner.

An Example of A/B Testing Cross Contamination:



A retailer wants to determine what value to assign to a coupon →

Runs an A/B test to determine the right numbers →

The control group gets served the standard page offering a 15% off coupon →

Test group gets served a page offering a 20% off coupon →

Customers in both groups search for a better offer →

Competitor offers 15% off →

The test group converts at a higher rate

Here's the problem:

The customers that placed orders in the test group weren't reacting to a 20% coupon in isolation. They were reacting to a coupon that was a better deal than what the competitor was offering. Something outside of the test, and outside of your control, was influencing your customers' behavior, and by extension the validity of your test results. How will the retailer know if the 20% coupon really would have outperformed the 15% coupon if their competitor hadn't been making a similar offer at the same time?

The competitor's sale was what we'd call a confounding variable—a variable that has the potential to influence the results of the test, but is not accounted for within the test itself.

AVOID CONTAMINATION WITH PROPER FLOW

At Namogoo, we experienced this problem recently when we helped a customer run an A/B test. The customer decided that the test would only run within a specific conversion flow. Taken in isolation, this flow was sufficiently closed-off for testing purposes, but users could still navigate out to areas which were not part of the scope of the test.

We were able to determine that the areas outside the testing scope contained information that increased the likelihood of visitors coming into contact with confounding variables which had the potential to influence the outcome of the test.

We were also able to determine that these areas outside the testing scope contained information that, if visitors were exposed to it, would increase the likelihood of confounding variables influencing the outcome of the test.

The solution we developed preserved the integrity of the test by working with the customer to limit the test to only the customer population segment that stayed exclusively within the appropriate conversion funnel. When it could be determined that a user had left the funnel and visited any other page, they were excluded from the test.

THE MANY FORMS OF CONFOUNDING VARIABLES

The purpose of avoiding cross-contamination is to ensure that there are no confounding variables that are going to increase the statistical noise in your test, or diminish the effect of the treatment you're testing.

Confounding variables can show up in many forms. Here's a partial list of events that, if they occur while A/B testing is taking place, can introduce confounding variables that can invalidate your test results:

- Launching or suspending a marketing campaign
- Holidays, especially shopping-related ones like Christmas
- Major news or announcements related to your business
- A new product launch
- A significant website redesign
- Google or other major search engines updating their algorithms
- Internet outages
- Changes to the settings or goals of the A/B test itself

For an A/B test to render up useful information, every customer who comes to the point of making their choice to proceed with either the control or test version needs to have arrived there without other factors having influenced them.

An A/B test becomes unusable, or too compromised to extract meaningful data from, when a significant event occurs in the middle of the test. Say a company is running an A/B test and right in the middle of it, a tabloid runs an explosive and lengthy interview with a disgruntled ex-employee. Every customer who engages with your site is going to have a controversial news story about the company in their thoughts, and it's impossible to say how much that's going to affect their choices and feelings.

In that case, it would be better to postpone the test, then run it again later when things are nice, quiet, and boring. **My team has had to restart many a test when significant changes were made to the website**

during the test, and when rollbacks have accidentally removed the Namogoo tag from the test group. At this point, the test is compromised and must be restarted.

THE CHALLENGE OF SETTING UP A "CLEAN" TEST

To be sure, controlling for potential cross-contamination is one of the most challenging parts of setting up a valid, informative A/B test. It requires scrutinizing the paths of both the control subjects and the test subjects, looking to remove any variables that could skew the results.

For example, if you were looking to determine if offering free shipping would increase your sales, there might be a considerable difference in behavior between customers who arrived at your site organically versus customers who have been lured in by a coupon offered in an email campaign. Unless you can control for that, you might end up incorrectly declaring a winning scenario.

> In some cases, trying to figure out how to eliminate cross contamination can help you reframe your test in a way that will ultimately yield more informative results.

In the above example, recognizing the potential for cross-contamination might force you to make an explicit choice between A/B testing a free shipping announcement on your shopping cart page, or a free shipping coupon given out via email marketing.The initial question—"does offering free shipping increase sales?"—might have been agnostic as to where the customer first becomes aware of the offer. In trying to construct a proper A/B test to measure its effect, it becomes clear that the test cannot really proceed without making an explicit determination as to just what type of "free shipping" offer you're trying out.

It's also important to recognize that even if you've set up your test properly, changes you make to other parts of your site (or really, any notable actions your company takes) can have an influence on the behavior of the users you're testing. If you want reliable A/B test results, then make sure you wait until the test is over before you launch a new ad campaign, introduce a brand-new product, or overhaul the look of your website.

CONCLUSION

A successful A/B test is completed by carefully choosing and setting up the initial conditions of the test, eliminating potential opportunities to influence your test subjects, and scrutinizing for any possible contamination that might be introducing confounding variables and throwing your results into question.

However, keep in mind, that even **an A/B test setup with extreme scrutiny can still encounter complications over time. For that reason, it is imperative to maintain a persistent experience and consistent user groupings for the entire duration of the A/B test.** As part of our scoping process, we help our customers identify any potential for data pollution and work with them to remove those conditions from the test segments.

CHAPTER 4

How to Prevent Bot Traffic from Skewing Your A/B Tests

While things aren't quite so bad that we're sending Terminators back in time to stop evil A.I. computers from taking over, there's no question that bots are a growing problem on the internet. There are bots that infiltrate social media to spread misleading stories; bots that launch hacking and denial-of-service attacks against web servers; and bots that mimic human traffic to invade ecommerce sites and scrape data—or probe for security holes.

Some of these "bad" bots are looking for competitive data like pricing or inventory levels, while others may be actively attempting to commit theft or fraud. Even seemingly innocuous bots can cause trouble for ecommerce merchants trying to make decisions based on accurate estimates of human traffic to their website.

For merchants who want to understand what's going on with their customers—their preferences, their pain points, what drives their purchases, what they're hoping to find more of—bot activity poses a serious hazard to their ability to collect and analyze data that tells the real story.

HUMAN VS. BOT TRAFFIC

Imagine driving down the highway and seeing that more than four out of every ten cars you pass is being controlled by a driverless A.I. That's the current situation in web traffic. According to statistics from 2017, 42.2% of all web traffic is bot-driven.

If your business has an online presence, be assured that it's getting hit up by bots every day. Many of them are malicious-the kind that don't tell web servers that they're bots. According to that same 2017 study, more than half of the bots encountered were the "bad" kind.

Some bots are used for legitimate purposes, to test software or gather data for search engines and other above-board enterprises. These bots announce themselves to your web server, the web server knows they're bots, and their visits can be excluded from reports on human traffic patterns.

Bad bots range in complexity from very simple scripts to complex, sophisticated programs that imitate human web surfing behavior in a variety of ways. They can even use proxy servers and alter their IP addresses to conceal their actual connection point. Web servers have ways to tell bots and humans apart, so bot programmers have to keep coming up with ways to avoid detection so their bots can't be blocked. The Interactive Advertising Bureau (IAB) even publishes a list of known bots so merchants can avoid and manage them.

> Most of these bots are designed for a specific purpose. They will persistently visit sites over and over again to achieve their programmed goals, which could be cracking user accounts, stealing cardholder information, scanning for vulnerabilities, or scraping data. Sometimes bots even join together to launch a concerted attack, forming what's called a "botnet."

Airlines are a great example of an industry that has a huge problem with bots. **Shady online travel agencies use bad bots to scrape airline websites and mobile apps for cheap fares to resell.** Competitors use bots to gather market intelligence about rates and schedules. Criminals attempt account takeovers, so they can perpetrate credit card or reward program fraud. Bad bots give the bad guys an advantage in snatching up protected data.

An airline's website, mobile app, and API are very enticing prizes for those sending out bad bots to do their dirty work. This is because the products they're selling are highly sought after, subject to changing rates, and limited in availability in terms of both time and quantity. What that means is there's always an incentive for competitors and fraudsters to keep cooking up newer, better bots, custom-made to slip past the airlines' cyberdefenses and gain entry.

Even if an airline can prevent actual theft, account takeovers, or fraudulent purchases, the bots are still doing harm. When the airline wants to gather data about website usage to improve the experience for their actual customers, how much of it is going to be skewed and inaccurate due to their high volume of malicious bot traffic?

THE IMPACT OF BOTS ON A/B TESTING

One way that bots can become a big problem for ecommerce sites is during A/B testing. When bot traffic is indistinguishable from human traffic, it can have a dramatic impact on any A/B tests you're running.

There are two ways for bots to mess up an otherwise well-constructed A/B test. First, when attempting to determine the proper composition of the control group, bots can throw off the data you're using to determine what a "typical" visitor to your site looks and behaves like. Bots might take a keen interest in pages or products that your human visitors largely ignore, or vice versa.

The second issue is when the test goes live and bots are being directed to your A and B treatments. Their behavior gets recorded and factored into the final results.

From a marketing perspective, it's not helpful or useful information to know whether a cybercriminal's bot likes the A version of a page better than the B version. Data gleaned from bot activity is just statistical noise, and if a significant portion of the test population turns out to be bots instead of human traffic—which is entirely possible, especially if you're in a vulnerable industry like airlines—then you're going to end up with an A/B test that was a total waste of time and labor.

HOW TO CONTROL FOR BOTS WHEN A/B TESTING

So, now you know that bad bots are a big problem, possibly accounting for close to half the incoming traffic to your website. You know that they're specifically designed to imitate human behavior as much as possible, so that they cannot be detected or blocked. Given this information, it may seem like there's no way to ensure that you can ever get reliable results from A/B testing.

Don't panic! While bots may have permanently ruined the pastime of arguing with strangers on Twitter, your A/B test should still be running at a scale where knowledge and diligent efforts will enable you to suss out the bad bots, and zero in on the identifying characteristics that differentiate them from human visitor website traffic.

Here's an example we worked with. A top online travel website was running an A/B test with Namogoo. We noticed some anomalies in the data, and upon further investigation we found that the reason for the anomalous behavior was a botnet attack devised to scrape data. The customer had to pause all A/B tests until the bot issue could be resolved.

Evasive bots that try to conceal their identity as bots pose a challenge, as they are much more difficult to detect than simple scripted bots, and aren't found on the IAB list of known bots.

However, merchants have an advantage in that they should know their customers well enough to know what their traffic "should" look like, in terms of demographics, geographical location, visit length, shopping patterns, and so on.

The best way to identify bots is to pay close attention to the behavior of visitors to your website. Often, there's not a single telltale sign that clearly flags a visitor as a bot, but various clues that, when taken together, will strongly suggest that the visitor's behavior is not being driven by the interests and priorities you would expect to see in a normal customer.

Here are some questions Namogoo asks when we're trying to clear the bot noise out of an A/B test:

Is there a range of IP addresses having multiple sessions within an unrealistic time period?

Are visitors who are viewing 10+ pages spending less time, on average, on each page than a visitor normally would?

Do your error logs show visitors who are trying to query pages or filenames that don't actually exist on your server?

Do any visitors have a habit of failed login attempts, abandoned carts,or other unusual activity?

Unfortunately, there's no one-size-fits-all bot detection framework you can run all your traffic through for analysis. **Every website is different, and therefore every bot that targets that site will look and behave differently.** There simply isn't a method of defense that works better than really knowing your website, your customers, and your products on a deep level. This lets you see when visitors are interacting with your site in a way that just doesn't feel normal based on your average customer interaction.

CONCLUSION

Bots are crafty, and every day their programmers find new ways to hide and disguise them. However, a keen analytical eye and deep customer knowledge can identify bots posing as humans.

A/B testing is only as good as the groups you're testing. It's of vital importance for anyone running an A/B test to make sure that both their control group and the visitors being diverted to their test treatment are composed of representative populations of real, live, human customers.

Maybe someday the bots will evolve to a level of complexity where they're spending their own money on ecommerce products and services, but until that day comes, you'll want to make a serious effort to keep them out of your tests and analytics.

SUMMARY

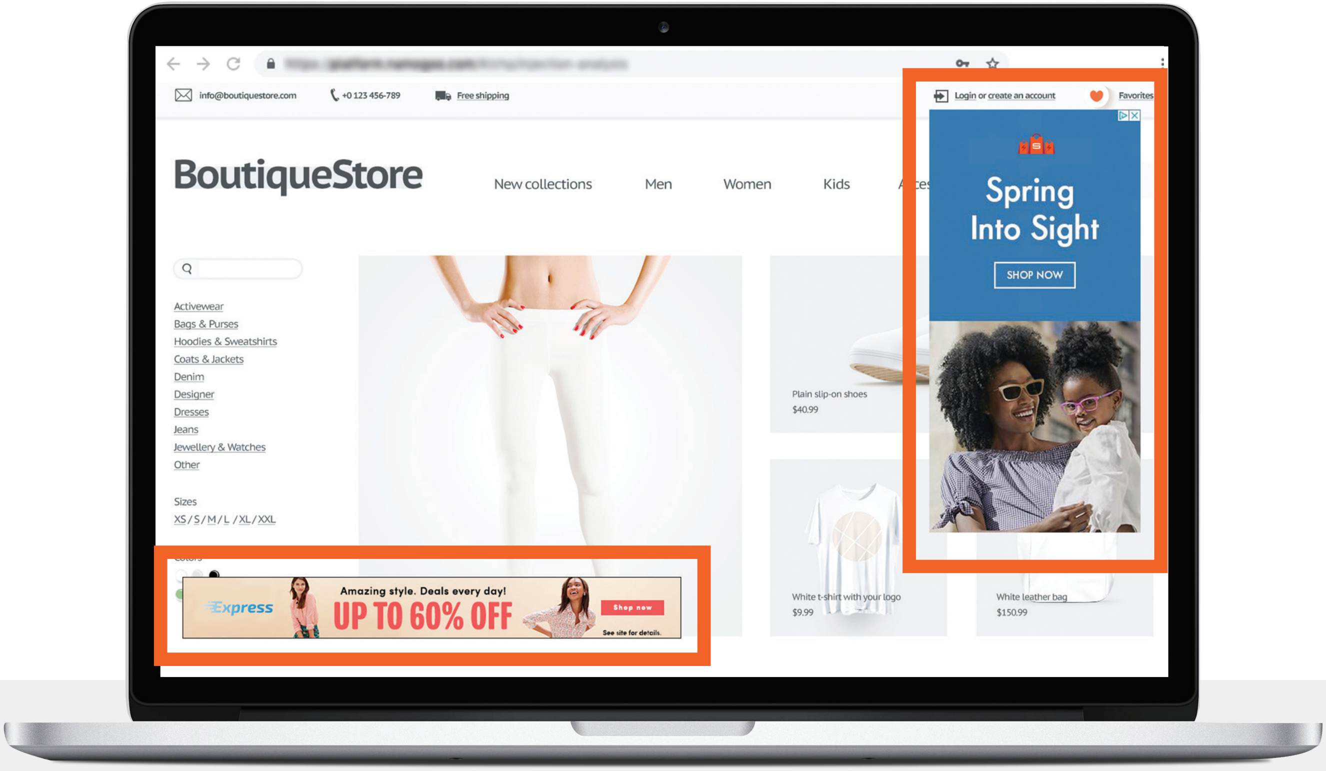
Cost of acquisition, lifetime customer value, conversion rates, average order value, cart abandonment - these are the measurements that determine the success or failure of an eCommerce operation, and A/B testing is the engine that drives these KPIs up or down.

In fact, A/B tests are such an integral part of digital commerce that 71% of companies are running at least 2 A/B tests a month. When implemented correctly, A/B tests can be an extremely powerful weapon in your arsenal. However, like any powerful weapon it needs to be used wisely or it could end up generating more hard than good. I hope you use these best practices of A/B testing to lay a foundation for skyrocketing your KPIs. Happy testing!

15-25% OF YOUR TRAFFIC IS BEING INTERRUPTED BY UNAUTHORIZED ADS

Find out what percentage of your customers are being diverted and to which competitors.

GET A FREE WEBSITE ANALYSIS



Namogoo helps online businesses enhance customer journeys and business results. With over 500 million web sessions analyzed each day, Namogoo's disruptive client-side platform enables online businesses to deliver a distraction-free user experience by blocking unauthorized product ads injected into consumer web sessions, and gaining full visibility and intelligence over all third- and fourth-party services running on their site. The world's largest online brands rely on Namogoo to gain control over their online customer experience and consistently improve business metrics. For more information, visit namogoo.com.