The Fundamentals of Experimentation

TWILIO segment

Growth Masters

Introduction

Netflix. Google. Facebook. These are just a few of the names that are famous for embracing a culture of experimentation. Evidenced by their multi-billion or trillion dollar valuations, the strategy appears to be paying off. No wonder Reforge Founder and CEO Brian Balfour lists “maximize learning” as his first principle of top performing growth teams.

Fortunately, you don’t have to be an industry behemoth to unlock the value of experimentation. Smaller companies can also learn how to make better decisions faster through a culture of experimentation.

In this quick-start guide, you’ll learn the fundamentals of experimentation. We’ll first define key terms and describe different types of experiments. Next, we’ll offer frameworks for prioritizing experiments and tips for creating an experiment roadmap. Finally, we’ll suggest ideas for what to test, highlight popular experimentation tools, and explore a handful of case studies that reveal the kinds of results you can expect from your own experimentation program.

Table of contents

Glossary

A/B testing

A type of randomized experiment in which you divide your traffic in half and send it to two pages where a single element has been changed to discover which variant performs best.

Bounce rate

The percentage of visitors who leave the website after viewing a single page.

Confidence interval

A range of values within which the true value lies.

Conversion

When a visitor takes a desired action on a website.

Conversion rate

The number of visitors who take a desired action on a website divided by the total number of visitors to that website.

Conversion rate optimization

The practice of growing the percentage of users who take a desired action on a website.

Experiment

A scientific method for testing a well-defined hypothesis.

Hypothesis

A prediction based on limited evidence that must be tested to determine whether it's true.

Multipage testing

A type of randomized experiment in which you change a single variable on multiple webpages and compare it with your baseline data to discover which variant performs best.

Multivariate testing

A type of randomized experiment in which you test multiple versions of multiple variables on a webpage simultaneously to discover which combination performs best.

Null hypothesis

The prediction that no statistical significance exists between two variables.

Sample size

A selection of people from a population chosen to participate in a study.

Segmentation

The act of dividing a population into groups based on shared traits.

Split URL testing

A type of randomized experiment in which you divide your traffic in half and send it to two different versions of the same page to discover which variant performs best.

Standard error

The difference between the total population's average and the sample's average.

Statistical significance

An outcome that is likely not due to random chance but is instead due to a specific cause.

Test duration

The amount of time you run an experiment for.

Type 1 error

The act of rejecting a null hypothesis when it's actually true.

Type 2 error

A failure to reject a null hypothesis that is false.

A tried, tested, and true process for experimentation

To reap benefits like faster learning, better decision-making, and rapid growth, it's important to implement a repeatable process for experimentation. Otherwise, your business will never operate at the speed needed to make significant gains over your competitors.

Next we share a 10-step process for conducting experiments, which was developed in partnership with high-performing growth leaders at companies like Zendesk, Lyft, and Shopify. (Download our 2021 Growth Report for more actionable insights.)

1. Identify a gap or opportunity

The first step is to identify a problem to solve. A good place to start is your customer data platform (CDP), which you can use to capture customer behavior in a few clicks and then build unique audiences based on that data for comparison (and testing) purposes. For example, you can easily see who your most and least profitable customers are and how they engage with your marketing or product. Alternatively, you can compare your performance against external benchmarks or internal targets to figure out which metric needs work. These are both good strategies for identifying opportunities for growth.

"Data is both your measuring stick and your dousing rod. It helps you determine where you are and your progression toward where you want to be, but it's also the raw material that helps you understand where to focus your efforts to grow faster."

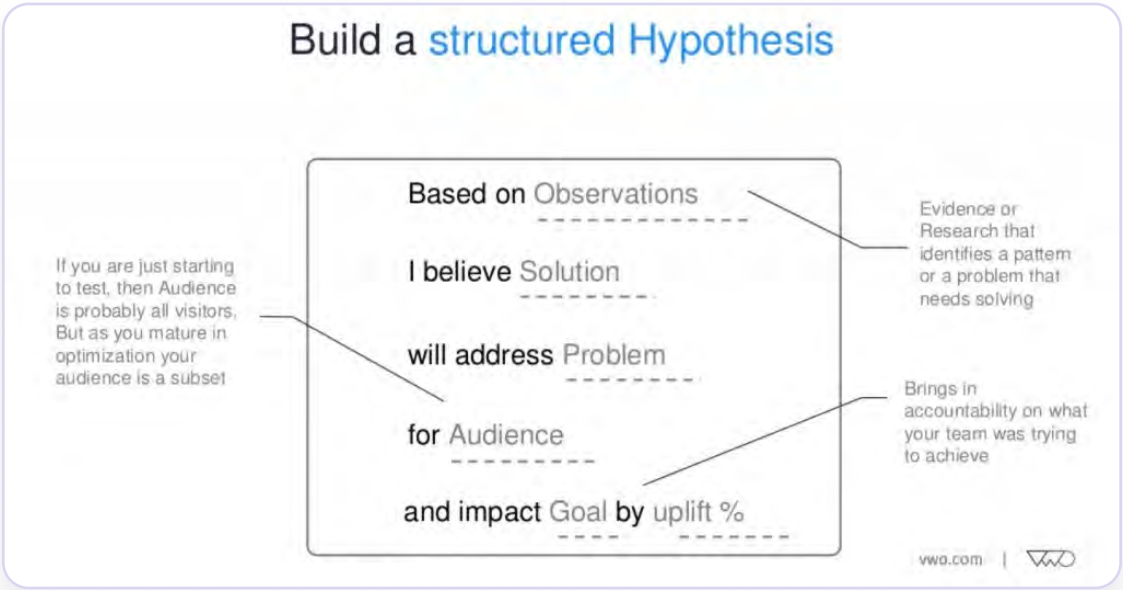
Brian Kotlyar
SVP, Demand Generation, New Relic

2. Decide how to solve the problem

The next step is to brainstorm ways to solve the issue. This is where your gut instinct comes into play. While you wouldn't want to make a critical decision on gut alone, you can and should take inspiration from what your gut tells you. After all, you're about to put it to the test! Try to come up with as many ideas as possible before choosing the one you want to move forward with.

3. State your hypothesis

A hypothesis is a prediction based on limited evidence that must be tested to determine whether it's true. It invites you to document the problem, propose a solution, and anticipate results in just a few sentences. In this step, you'll craft a precise hypothesis that can be proven wrong. For a simple yet effective hypothesis, use this framework: Based on [observations], I believe [solution] will address [problem] for [audience] and impact [metric] by [uplift].



4. Prioritize your experiments

No company can test every idea they have. Instead, they must prioritize their list of experiments, choosing the one likely to have the greatest impact. Later, we'll describe and evaluate three popular frameworks for prioritizing your tests: HiPPO, or highest paid person's opinion; RICE, or reach, impact, confidence, and ease; and PIE, or potential, importance, and ease. Until then, know that it's possible to objectively assess the potential impact of your experiments.

5. Create your test plan

Once you've identified which experiment you want to run, it's time to create your test plan. This document should include the purpose of your test, a copy of your hypothesis, details about the different variants you plan to test, specifics about how and what you will measure, the location where your test will take place, details about your audience and sample size, and your planned test duration.

GitHub

"It's tempting, when you're a growth practitioner, to look at all the opportunities and start a lot of conversations about all the things you can fix. You end up getting a lot of people excited, but you lose momentum because you don't have a ton of resources. Focus is so, so important."

Thibault Imbert
VP, Growth, GitHub

6. Validate your idea

Before launching an experiment, savvy growth leaders first validate their hypotheses. To do so, they typically turn to user research. By putting your variants in front of a handful of real users, you reduce the risk of burning your entire sample with a broken experience. This protects your brand from unnecessary harm.

7. Launch your test

Once you've validated your hypothesis, it's time to release your experiment out into the wild. The faster you do so, the faster you will learn—and the faster you learn, the faster you'll maximize your impact. Your test duration will depend on your target sample size and monthly traffic, but experts agree it should last at least one full business cycle.

8. Evaluate the results

Once your test has run its course, it's time to crunch the numbers. Do they reach statistical significance? Are the results in favor of or against the change? What have you learned? You may also notice unintended consequences, so it's important to develop a clear picture of what actually happened. For example, you may notice an increase in paid signups instead of free trials. While that may not have been your prediction or goal, it still provides you with important lessons.

MURAL

“Insights are the lifeblood of growth. They’re what fuels the machine.”

Lauren Schuman
VP, Product Growth, MURAL

9. Implement changes

Informed by the outcome, you can then determine next steps. For example, if the results were in favor of the change and were statistically significant, you’ll need to deploy the proposed changes to the entire population. If not, you can still document what you’ve learned and ideate future tests inspired by the failure.

10. Share your learnings

In this final step, you’ll share what you’ve learned with the broader team, including leadership. That way, everyone has an opportunity to adjust their strategies in response to your newly uncovered insights. Presenting the outcomes of your experiments also ensures that team members don’t duplicate efforts in the future.

KLAVIYO

“If you want to be a successful growth team, don’t waste your time doing experiments someone else has done. Do everyone a favor and share whatever you have and everyone wins.”

James de Feu
Director, Self-Serve Growth, Klaviyo

4 types of experiments you can run

When many people think of experimentation, they automatically jump to a/b testing. But, while A/B testing is a popular and effective tool for experimentation, it's not the only one out there. In this section, we cover four different kinds of experiments: A/B tests, split URL tests, multivariate tests, and multipage tests.

A/B tests

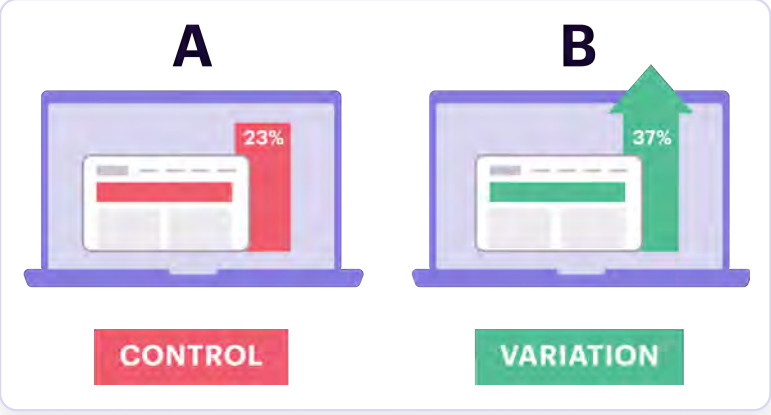
A/B testing is a type of randomized experiment in which you divide your traffic in half and send it to two pages where a single element has been changed to discover which variant performs best. A/B tests are one of the best options for businesses with under 100,000 unique visitors per month.

ADVANTAGES

- A/B tests are relatively simple to design and execute.
- You can resolve debates around whose hypothesis is correct.
- You can get statistically significant results with less traffic.

DISADVANTAGES

- You are limited to two variations of a single element.
- It's impossible to understand interactions between elements.



When to use it

Use an A/B test when you're testing a single element, such as the color or placement of a call-to-action button.

Split URL tests

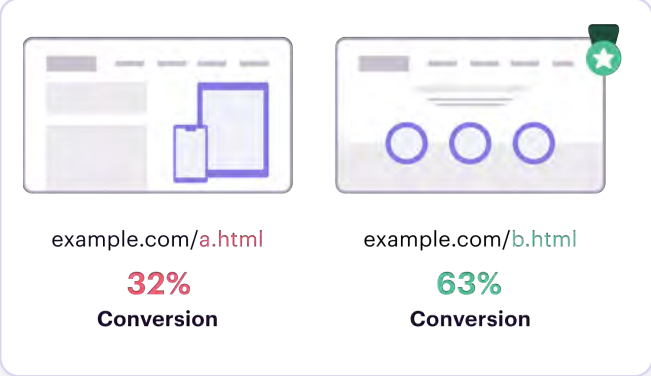
A/B testing is a type of randomized experiment in which you divide your traffic in half and send it to two pages where a single element has been changed to discover which variant performs best. A/B tests are one of the best options for businesses with under 100,000 unique visitors per month.

ADVANTAGES

- High-risk, high-reward: Because split URL tests involve substantial changes, they can deliver substantial results.
- You can get statistically significant results with less traffic.

DISADVANTAGES

- You are unable to detect the impact of specific elements, so you'll gain fewer insights into the problem at hand.
- To achieve substantial changes, you'll need more design and development resources.
- Out of all the experiment types, this is the most complex one to design and execute.



When to use it

Use split URL testing when you are refreshing an entire page. This experiment is especially useful during rebrands.

Multivariate tests

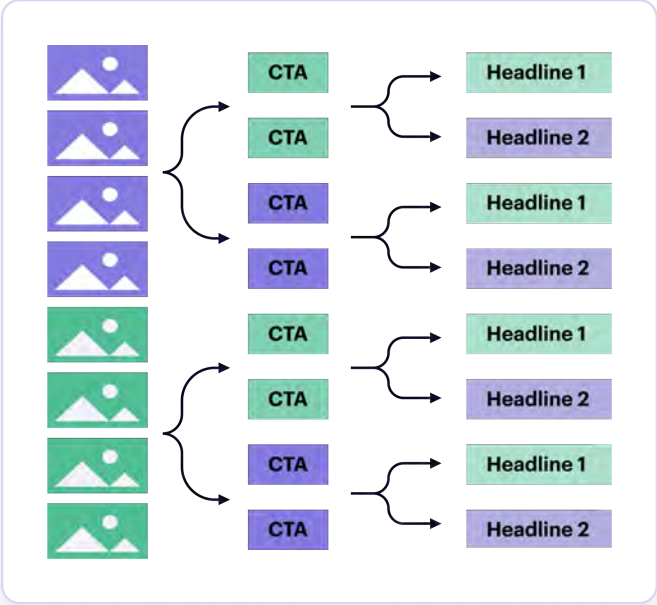
Multivariate testing is a type of randomized experiment in which you test multiple versions of multiple variables on a webpage simultaneously to discover which variant performs best. Multivariate tests are only recommended for businesses with over 100,000 unique visitors per month.

ADVANTAGES

- Multivariate tests can save you time, since there is no need to conduct a series of A/B tests.
- You can also gain more insights: In addition to understanding how individual elements contribute to results, you can see how variables interact with one another.

DISADVANTAGES

- Multivariate tests are moderately complex to design and execute.
- While you save time by avoiding a series of A/B tests, it takes longer to get results from a multivariate test in comparison with an A/B test.
- You can only get statistically significant results with lots of traffic.
- It's difficult to interpret the contributions of individual elements.
- To test multiple variables at once, you'll need more design and development resources.



When to use it

Use multivariate tests when you want to know how multiple elements interact. For example, a multivariate test would be useful for figuring out how call-to-action button placement and color interact.

Multipage tests

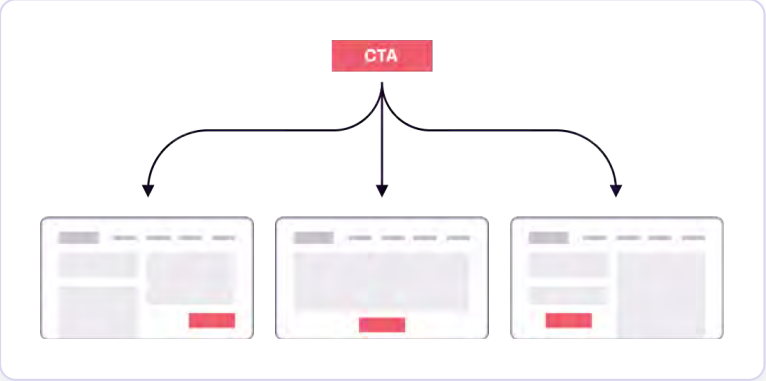
Multipage testing is a type of randomized experiment in which you change a single variable on multiple webpages and compare it with a control to discover which variation performs best. It is best suited to businesses with over 100,000 unique visitors per month, but can also be useful for companies with less traffic if they're willing to embrace a long test duration.

ADVANTAGES

- You can test a single element across the entire website at once.
- It can help you calculate your site-wide conversion rate.
- It helps you maintain a consistent user experience.

DISADVANTAGES

- It takes a longer time to decipher results.
- You can only get statistically significant results with lots of traffic or a prolonged test duration.



When to use it

Use a multipage test when you want to deliver a consistent user experience. For example, you may want to change your call-to-action copy across your entire website, as opposed to confusing visitors with different CTAs on different pages.

8 ideas for what you can test

Copy

Two types of copy exist on your website: headlines and body copy. Headlines are short and should instantly grab the visitor's attention. Body copy is longer and should concisely communicate the benefits of your offer, persuading readers to take action. Both should be geared toward your specific audience. You may also want to incorporate psychology principles like urgency and scarcity into your copy.

Beyond the words themselves, you can experiment with tone of voice, readability, or even formatting. Test whether your audience prefers clear copy or sophisticated prose. Test whether visitors resonate with an authoritative tone or something a little more empathetic. Test whether they prefer large paragraphs presented in a serif font or short, bulleted points sans serif. The possibilities are endless.

CTAs

Calls to action should be clear and actionable. They should also fit the stage the customer is in. Often but not always, you can start the CTA with an action verb that causes the visitor to imagine taking the desired action. Apart from that, there is plenty to test.

You can test variations on the copy, size, placement, and color of CTA buttons. For example, a green button might perform better than a red button in some circumstances. Similarly, a large CTA button might better capture the attention of your audience in comparison with a small or mid-sized button. Perhaps centering your CTA button on the page would result in more clicks than if it was right- or left-aligned. You're only limited by the limits of your imagination.

Depth

When it comes to content depth, the only general rule is keep what's most important above the fold. Outside of that, there are two considerations to keep in mind: The first is whether your offer is a high-commitment offer, such as an annual software contract, or a low-commitment offer, such as a free ebook. Typically, the higher the commitment, the longer the page you'll need to convert your target audience. Shorter pages, on the other hand, often work perfectly well for low-commitment offers.

The other consideration to keep in mind is whether your audience prefers short, skimmable pages or long, detailed pages. According to a study conducted at Brown University, two types of consumers exist: As "explanation fiends" learn more about a product, they become more likely to buy it. In contrast, as "explanation foes" learn more about a product, they become less likely to buy it. Hence why it's important to test what works for your audience.

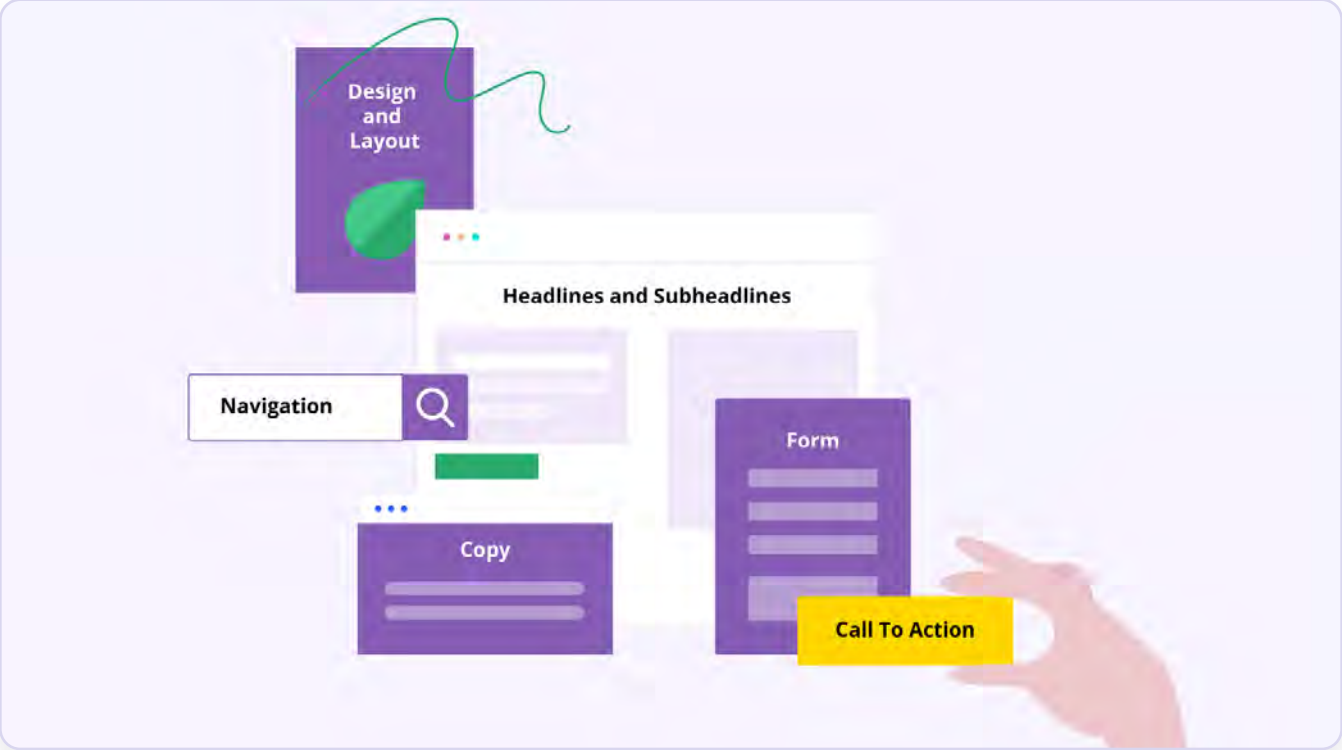
Design and layout

For the most effective user experience, you want the design to include plenty of white space, a contrasting color scheme, minimal distractions, and only the most important copy. But design is highly subjective, meaning there are many ways to design a page. For that reason, it's ripe for testing. To optimize the visitor's experience, try experimenting with different visuals, color schemes, and spacing.

Your layout should help guide the audience from one piece of important information to the next, preferably ending with your call-to-action button. When it comes to layout, you can test things like whether a single- or multiple-column layout performs better. Often, a combination of both is what's needed. You can also conduct a test to see whether an F- or Z-pattern layout works best for your audience.

Forms

To prevent friction, experts suggest you should aim for as few form fields as necessary. However, some companies intentionally add friction to attract fewer but better qualified leads. Either way, you should test what works: a short form that asks for only the essentials, or a long form that provides you with rich details about your target audience? You can also experiment with the order of questions, microcopy, and progress bars.



Navigation

Whether you're testing changes to your homepage or creating a new landing page from scratch, it's important to consider your audience's navigation needs. When browsing a website, for example, visitors typically want to be able to click around to find what they need. In contrast, you actually want to keep visitors on a landing page once they land there, as it improves your conversion rate.

To promote a seamless user experience, place your navigation bar in a predictable place, matching your visitor's expectations. Also, group similar content together in your navigation bar. Beyond that, test away! Try experimenting with how comprehensive your navigation is, what it links out to, and how it changes when you scroll over it.

Offers

Outside of testing the copy, design, and layout, you can test the offer itself. Are future customers more interested in a free assessment, which evaluates their performance in a certain area and provides strategic recommendations, or a free trial, which allows them to get into the product and click around? Are current customers more interested in a discount for switching from a monthly to annual subscriber or a bonus for referring their friends and family? These are the kinds of questions you should be asking as you consider what offers to test.

Trust

Trust indicators include social proof, such as testimonials, product reviews, and expert recommendations; customer logos, especially household names; media mentions; guarantees; and security certifications. As the name suggests, these elements reassure future customers that they can trust you enough to buy from you. Testing can help you discover which types of trust indicators most resonate with your audience, and which specific elements drive the most conversions.

Where to source ideas for new experiments

- Product walk-throughs
- On-page surveys and live chat
- Funnel analysis
- User session replays
- Web and product analysis
- Mouse tracking
- Heat maps
- User testing
- Technical analysis

The top 3 experiment prioritization frameworks

When it comes to prioritizing your experiments, it's a best practice to use a systematic method for evaluating things like their potential impact and the effort required to pull them off. Establishing a clear set of criteria helps you avoid subjective decision-making and increases the likelihood that you will choose to run tests with the highest potential. As a result, you're more likely to achieve the speed required for rapid business growth.

In this section, we highlight three popular frameworks for prioritizing your experiments: HiPPO, RICE, and PIE.

HiPPO

HiPPO, which stands for "highest paid person's opinion," is what you want to avoid. This autocratic form of decision-making is highly subjective and thus subject to error. Many startups follow this strategy without realizing it. To achieve a true culture of experimentation, it's important to embrace a democratic, evidence-based approach to prioritization.

PIE

PIE is a simple prioritization framework that was invented by Widerfunnel Founder Chris Goward. It balances three factors: potential, importance, and ease.

Potential looks at how much growth is possible. Your lowest-performing assets have the most room to grow, so give them a 9 or 10. Your highest-performing assets have little room to grow, so give them a 1 or 2.

Importance refers to how valuable the experiment is. If conducting a test on a webpage with high volume and pricey traffic, give it a 9 or 10. If conducting a test on a webpage with low volume and cheap traffic, give it a 1 or 2.

Ease evaluates how difficult the test will be to implement. Tests that require fewer resources earn a higher score, like a 9 or 10, whereas tests that require many resources earn a lower score, such as a 1 or 2.

LIFT Zone	Potential	Importance	Ease	PIE Score
Homepage	10	10	8	9.3
Checkout	8	10	9	9.0
Product page	10	9	7	8.7

To calculate your PIE score, add up your potential, importance, and ease scores and divide them by 3. For example, if your potential is 5, your importance is 9, and your ease is 4, your pie score will be 6 out of a possible 10 marks.

RICE

RICE was invented by Intercom's team of product managers. It's an approach that weighs four factors: reach, impact, confidence, and ease.

Reach is an estimate of the number of people (e.g., customers, users, visitors) who will be affected by a change within a certain period of time. It helps prevent you from prioritizing smaller tests with less reach.

Impact encourages you to focus on experiments that have the potential to move the needle. Because it's hard to measure precisely, you choose from a scale: If the result of your test could have a massive impact, give it a 3. If it could have a big impact, give it a 2. If it could have a moderate impact, give it a 1. If it would likely have a small impact, give it a 0.5. If it would have a minimal impact, give it a 0.25.

Confidence is your estimation of how sure you are that your hypothesis is correct. If you have data to back up your estimated reach, impact, and ease, give yourself a 100% confidence score. If you have data to back up two out of three, give yourself an 80% confidence score. If you only have data to back up one out of three, give yourself a 50% confidence score.

Effort is your prediction for how long the experiment will take you and all of your teammates to produce. Assign 1 point for every month it will take. Use 0.5 for anything that takes less than a month.

$$\frac{\text{Reach x Impact x Confidence}}{\text{Effort}} = \text{RICE SCORE}$$

To calculate your RICE score, multiply reach by impact by confidence and then divide the total by the amount of effort you expect to put in. For example, if your reach was 500, your impact was 3, your confidence was 100%, and your effort was 2, your score would be 750.

Why you should also build an experiment roadmap

Now that you've prioritized your list of experiments, it's time to plan when, where, and how you will execute them. Unlike an experiment backlog, which is essentially just a queue of tests waiting for their moment in the spotlight, an experiment roadmap helps you develop a regular cadence of tests with defined timelines and stakeholders. Using this approach, you plan and organize much of the work in advance, which ensures you never miss a beat. By building an experiment roadmap, you unlock benefits like better resource planning, improved speed and efficiency, and faster business growth.

When building out your own experiment roadmap, consider these fields:

- Test name
- Hypotesis
- Total variations
- Sample size
- Start date
- Resource investment
- Location
- Stakeholders
- Implementation plan
- Test description
- Priority
- Posible side effects
- Test duration
- End date
- Baseline data
- Outcome
- Why'd it work?
- Future tests

Experimentation tools

Running experiments often requires begging engineering to pull the necessary data, trawling through that data to divide it into unique audiences, and finally, sending those audiences over to a testing platform.

That's where a CDP comes in: Segment will help you collect, validate, segment, and send all the data you need to the testing platform of your choice in just a few clicks. That means you can spend more time actually testing out new strategies and less time wrangling data.

The tools we've highlighted here all integrate with Segment, which means your life will be that much easier when it comes to setting up your own experimentation system. For more information about how each tool integrates with Segment, visit our Integration Catalog.

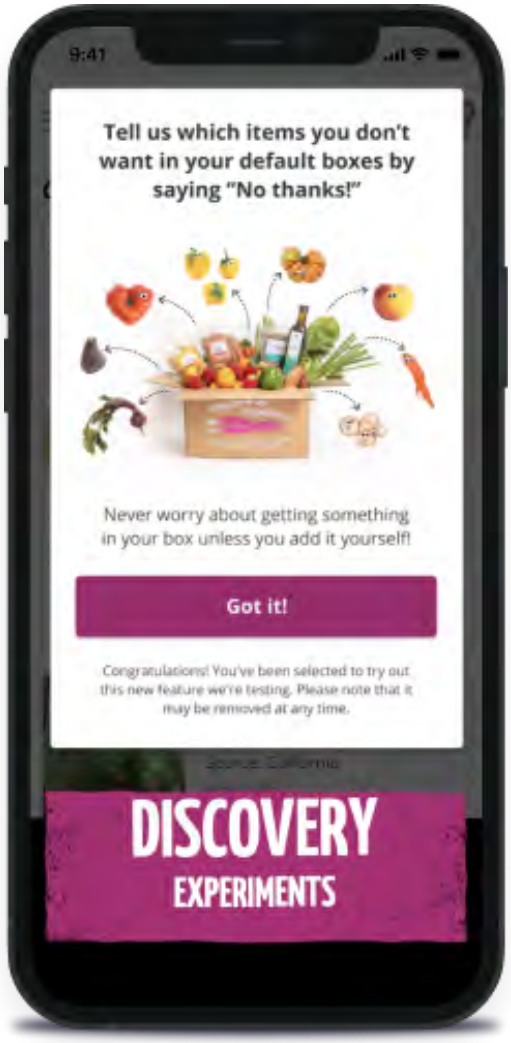
TWILIO segment	Used by over 20,000+ businesses to collect, clean, and activate their customer data, Segment is the leading Customer Data Platform
TAPLYTICS	Taplytics is a modern A/B testing software for product and marketing teams
Kameleoon	Kameloon provides A/B testing and AI-powered personalization solutions to help marketers, product owners, and developers maximize customer engagement and conversion.
split	Split powers your product decisions with a unified solution for feature flagging and experimentation.
Optimizely	Oprimizely is a feature flagging and experimentation platform fo anything with a network connection, inculding webistes, mobile apps, and smart devices.
VWO	VWO is an all-in-one platform that helps you conduct visitor research, build an optimization roadmap, and run continuous experimentation.
APPTIMIZE AN AIRSHIP COMPANY	Apptimize empowers product teams to efficiently run A/B tests, rollout and manage new features, and deliver personalized user experiences.
A/B SMARTLY	A/B Smartly provides an on-premise, full-stack experimentation platform for engineering and product teams that do continuous experimentation embedded into their development process.

Experimentation case studies

How Imperfect Foods grew customer retention by 21% by giving customers the power to choose

Imperfect Foods offers imperfect (yet delicious) produce, affordable pantry items, and quality meat and dairy on a weekly subscription basis via their website. VP Product Patti Chan and her team of product managers, engineers, and designers hypothesized that allowing customers to select foods they didn't want in their monthly box would improve customer satisfaction and loyalty. As a test, they launched this functionality to 50% of their LA subscribers and waited anxiously for the results. To their delight, customers who used the new feature were **21% more likely to be retained** than users who did not have access.

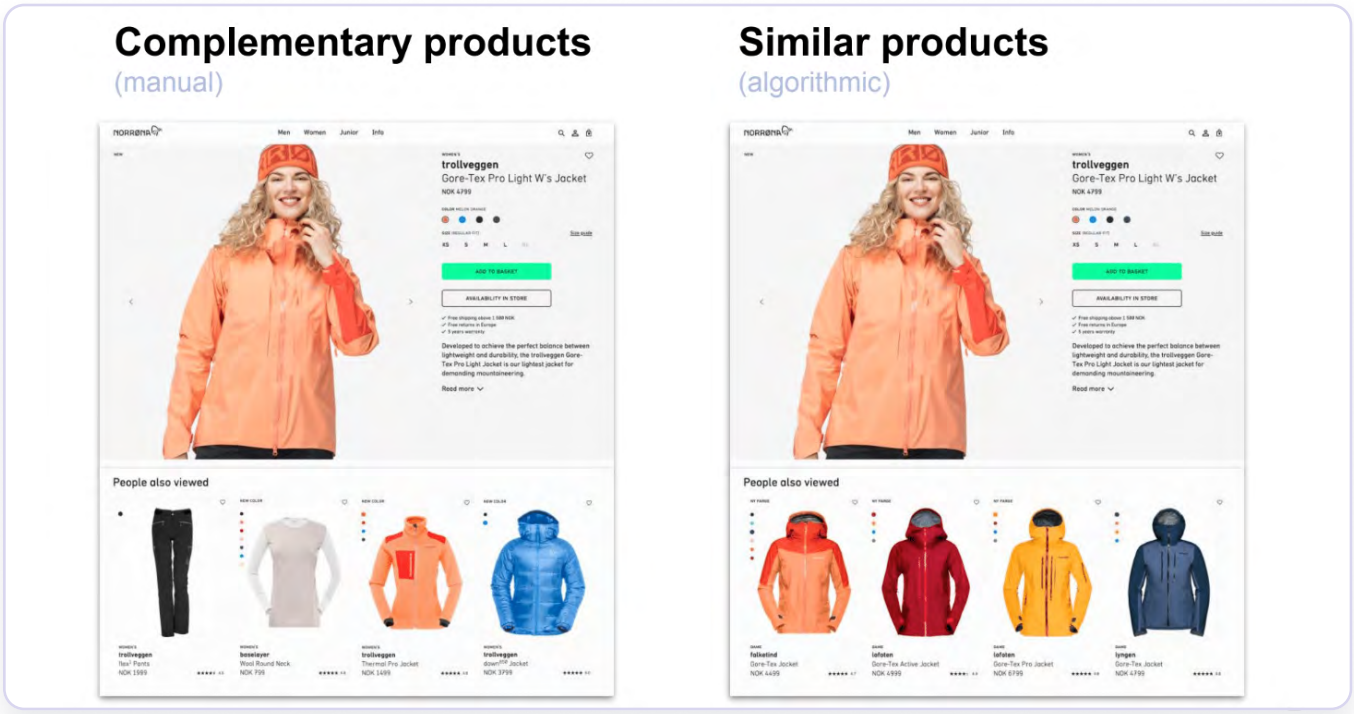
Read the full story >



How Norrøna built a recommendation engine that increased conversions by 50%

Norrøna has been the leading brand of outdoor clothing in Scandinavia for more than 90 years. Product Manager Thomas Gariel hypothesized that algorithmic recommendations would outperform manual recommendations, having a significant impact on revenue while reducing a ton of manual labor. Using Segment and BigQuery, they built a recommendation engine and soon discovered that it offered similar products, as opposed to the complementary products they had manually suggested. After a 16-week A/B test, Thomas Gariel discovered that the algorithm drove **50% more conversions** than their previous manual recommendation system—a significant win

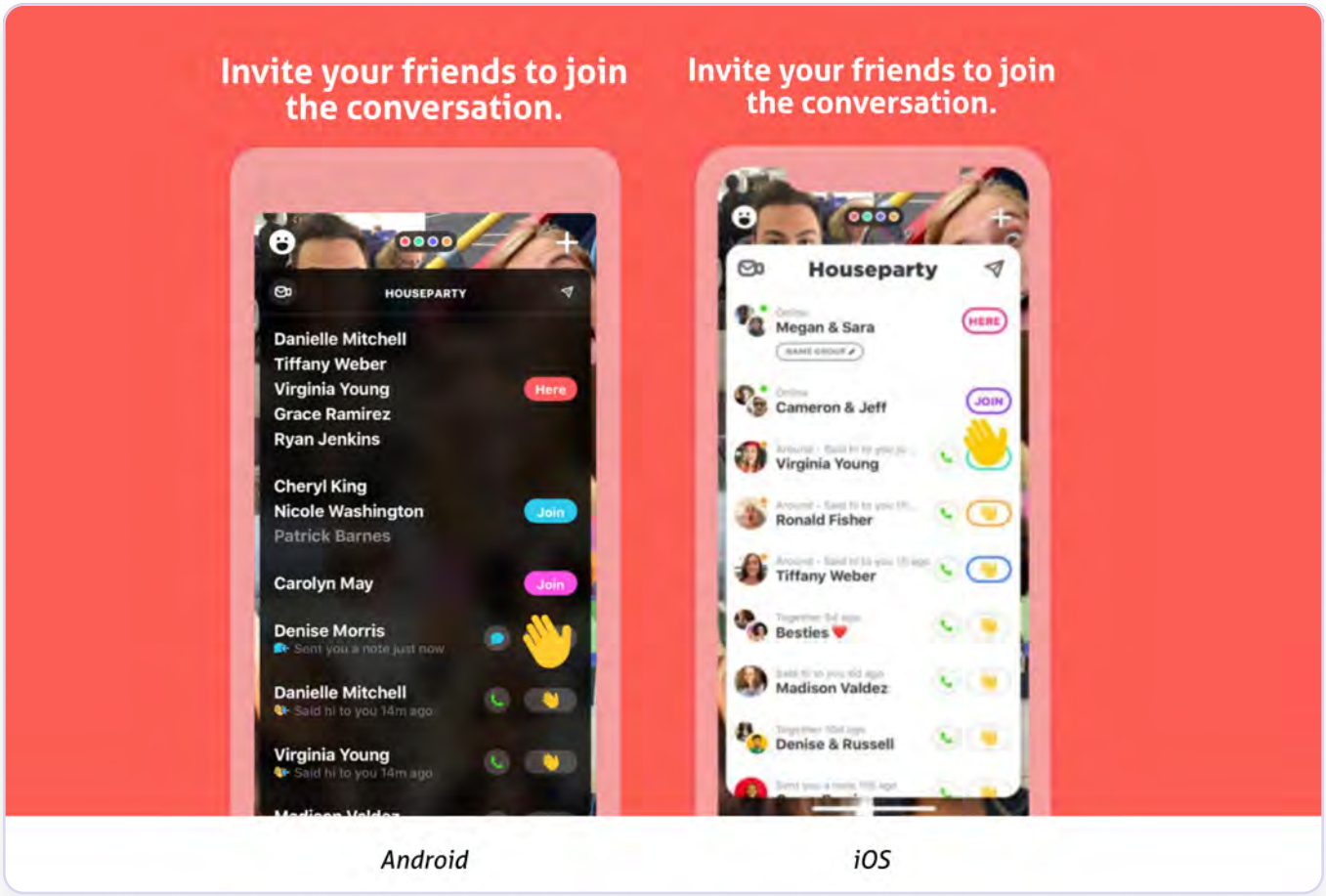
Read the full story >



How Houseparty grew friend invites 64% on iOS and 27% on Android by personalizing by device

Houseparty is an app where tens of millions of users video chat with up to 8 friends at once. Jeff Needles, who ran special projects for Houseparty at the time this story was collected, hypothesized that Android users would behave very differently than their iOS counterparts when inviting their friends to the platform. He was right: After testing 16 different variations of invite text, the highest performing entries differed for Android and iOS users. Whereas the iOS users invited more friends with the text "let's (house)party", the Android users invited more friends with the text "we need to talk". This one experiment **grew friend invites by 64% on iOS** and 27% on Android, a significant jump in one of their biggest drivers of new Houseparty users

Read the full story >



Take the first step toward data-driven decision-making

Now that you've learned the fundamentals of experimentation, you're well on your way to making better business decisions.

By investing in Segment, the industry's #1 customer data platform, you can empower every employee to leverage your customer data to drive rapid business growth—without getting lost in it.

Join the 20,000+ customers who've given Segment first-class ratings across G2, TrustRadius, GetApp and Capterra.

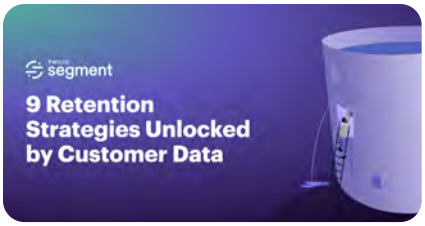
Learn how to supercharge your growth with Twilio Segment
Schedule a demo

Want to learn more?

The Fundamentals of Customer Acquisition

We share a comprehensive overview of the channels, strategies, and best practices that will help you master the art of customer acquisition.

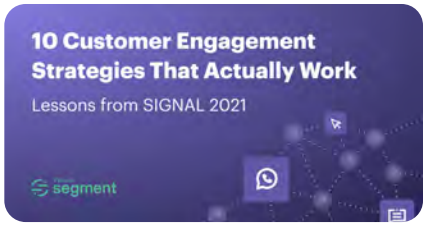
Download guide >



9 Retention Strategies Unlocked by Customer Data

Learn 9 retention strategies that leverage customer data to drive loyalty and lifetime value.

Download report >



10 customer engagement strategies that actually work

We compiled a list of the 10 essential customer engagement strategies shared at SIGNAL conference, from leaders at Peloton, Intuit, and more.

Download guide >