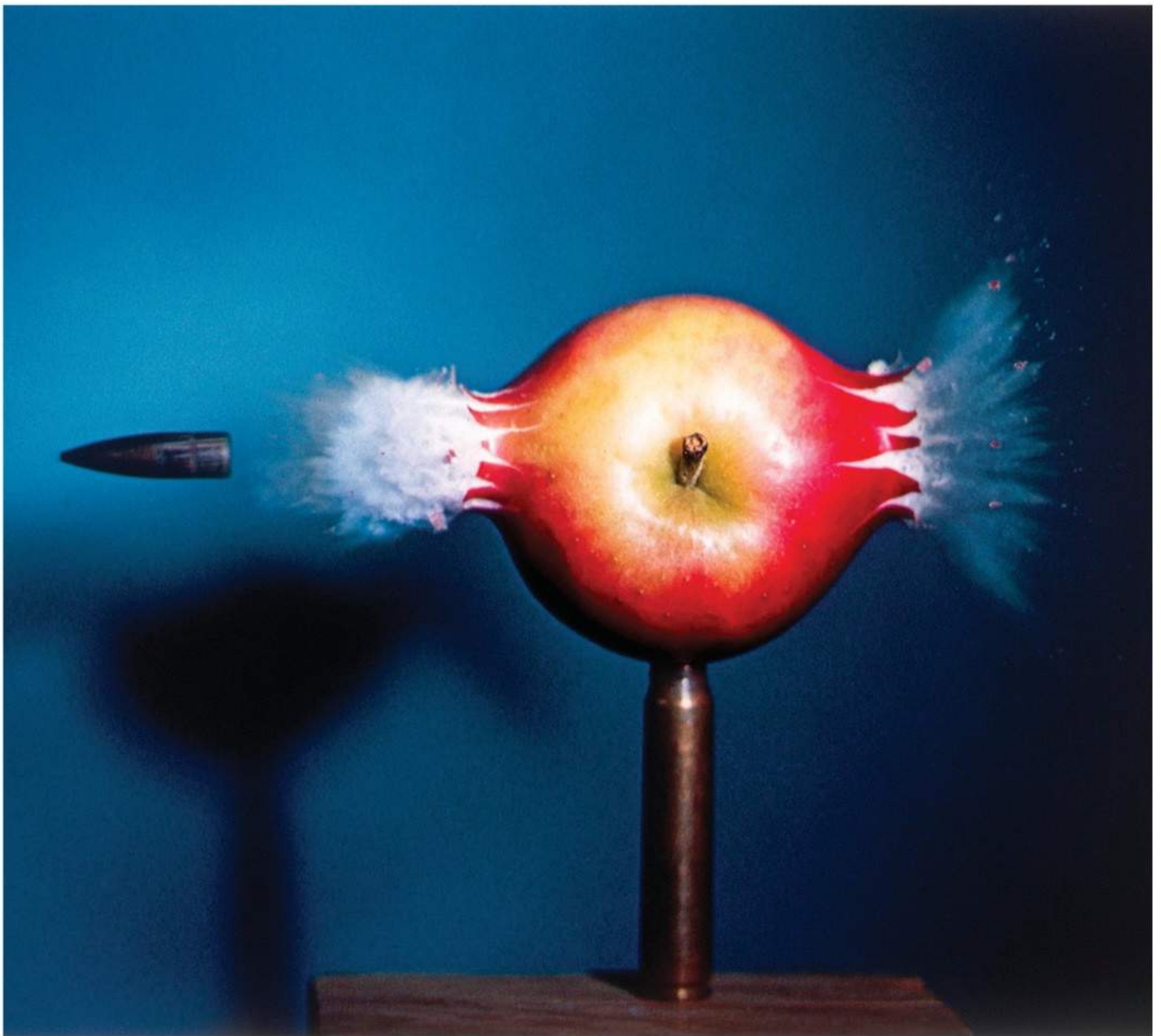

Spotlight

 PRODUCTIVE INNOVATION 



Spotlight



Harold Edgerton was known for his experiments with high-speed photography and used stroboscopic equipment to capture moments in time.

Building a Culture of Experimentation

It takes more than good tools. It takes a complete change of attitude.

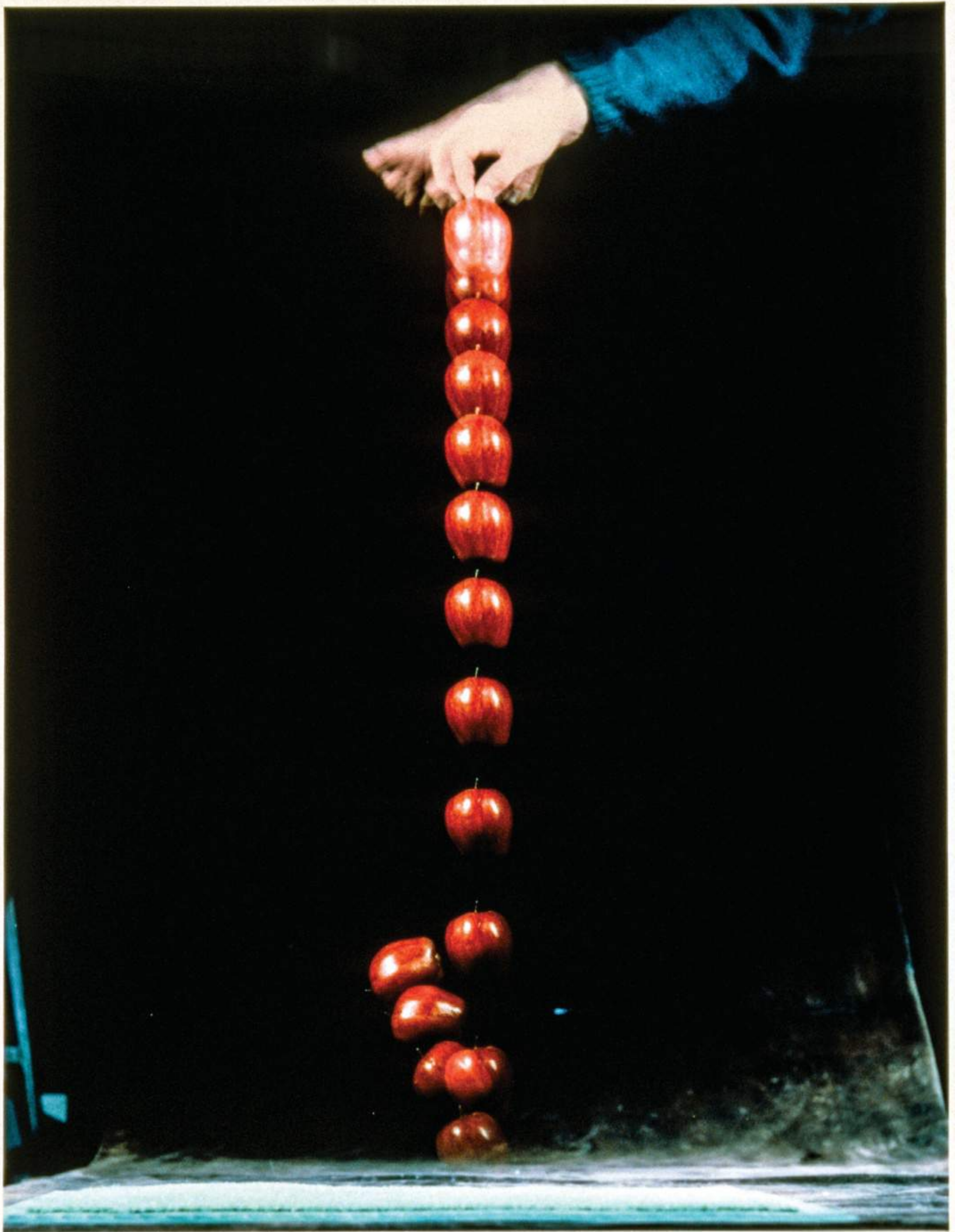


Stefan Thomke
Professor, Harvard
Business School

IN DECEMBER 2017, just before the busy holiday travel season, Booking.com's director of design proposed a radical experiment: testing an entirely new layout for the company's home page. Instead of offering lots of options for hotels, vacation rentals, and travel deals, as the existing home page did, the new one would just feature a small window asking where the customer was going, the dates, and the number of people in the party, and present three simple options: "accommodations," "flights," and "rental cars." All the content and design elements—pictures, text, buttons, and messages—that Booking.com had spent years optimizing would be eliminated.

Gillian Tans, Booking.com's CEO at the time, was skeptical. She worried that the change would cause confusion among the company's loyal customers. Lukas Vermeer, then the head of the firm's core experimentation team, bet a bottle of champagne that the test would "tank"—meaning it would drive down the company's critical performance metric: customer conversion, or how many website visitors made a booking. Given that pessimism, why didn't senior management just veto the trial? Because doing so would have violated one of Booking.com's core tenets: Anyone at the company can test anything—without management's permission.

Harold Edgerton ©2010 MIT. Courtesy of MIT Museum



Spotlight

Booking.com runs more than 1,000 rigorous tests simultaneously and, by my estimates, more than 25,000 tests a year. At any given time, quadrillions (millions of billions) of landing-page permutations are live, meaning two customers in the same location are unlikely to see the same version. All this experimentation has helped transform the company from a small Dutch start-up to the world's largest online accommodation platform in less than two decades.

Booking.com isn't the only firm to discover the power of online experiments. Digital giants such as Amazon, Facebook, Google, and Microsoft have found them to be a game changer when it comes to marketing and innovation. They've helped Microsoft's Bing unit, for instance, make dozens of monthly improvements, which collectively have boosted revenue per search by 10% to 25% a year. (See "The Surprising Power of Online Experiments," HBR, September–October 2017.) Firms without digital roots—including FedEx, State Farm, and H&M—have also embraced online testing, using it to identify the best digital touchpoints, design choices, discounts, and product recommendations.

"In an increasingly digital world, if you don't do large-scale experimentation, in the long term—and in many industries the short term—you're dead," Mark Okerstrom, the CEO of Expedia Group told me. "At any one time we're running hundreds, if not thousands, of concurrent experiments, involving millions of visitors. Because of this, we don't have to guess what customers want; we have the ability to run the most massive 'customer surveys' that exist, again and again, to have them tell us what they want."

But in studying more than a dozen organizations and analyzing anonymized data on experiments from upwards of 1,000, I have seen that Booking.com, Expedia, and their ilk are the exception. Instead of running hundreds or thousands of online tests a year, many firms run no more than a few dozen that have little impact.

If testing is so valuable, why don't companies do it more? After examining this question for several years, I can tell you that the central reason is culture. As companies try to scale up their online experimentation capacity, they often find that the obstacles are not tools and technology but shared behaviors, beliefs, and values. For every experiment

that succeeds, nearly 10 don't—and in the eyes of many organizations that emphasize efficiency, predictability, and "winning," those failures are wasteful.

To successfully innovate, companies need to make experimentation an integral part of everyday life—even when budgets are tight. That means creating an environment where employees' curiosity is nurtured, data trumps opinion, anyone (not just people in R&D) can conduct or commission a test, all experiments are done ethically, and managers embrace a new model of leadership. In this article, I'll look at several companies that have managed to do those things well, focusing in particular on Booking.com, which has one of the strongest cultures of experimentation I have found.

CULTIVATE CURIOSITY

Everyone in the organization, from the leadership on down, needs to value surprises, despite the difficulty of assigning a dollar figure to them and the impossibility of predicting when and how often they'll occur. When firms adopt this mindset, curiosity will prevail and people will see failures not as costly mistakes but as opportunities for learning.

IDEA IN BRIEF

THE OPPORTUNITY

In an increasingly digital world, randomized, controlled A/B experiments are an extremely valuable way to create or improve online experiences.

THE OBSTACLE

Culture—not tools and technology—prevents companies from conducting the hundreds, even thousands, of tests they should be doing annually and then applying the results.

THE REMEDY

Create an environment in which curiosity is nurtured, data trumps opinion, anyone can conduct a test, all experiments are done ethically, and managers embrace a new model of leadership.



It's actually less risky to run a large number of experiments than a small number. If a company does only a handful of experiments a year, it may have only one success—or none. Then failure is a big deal.

A classic example concerns an incident at Amazon involving a revision of *Air Patriots*, a game for mobile devices in which players defend towers from attack with a squadron of planes. When Amazon launched a new version of it, the development team was taken aback by the response: The seven-day user-retention rate dropped by an astonishing 70%, and revenue fell 30%. The team discovered that it had inadvertently increased the game's difficulty by about 10%. Amazon quickly shipped a fix, but the developers wondered if making the game easier could produce large *gains* in retention and revenue. To find out, they ran a test with four new levels of difficulty, in addition to a control, and learned that the easiest variant did the best. After some further refinements, Amazon launched a new version—and this time users played 20% longer and revenue increased by 20%. An accident had led to a surprising insight, which became the starting point for new experiments.

Unfortunately, this kind of reaction is an anomaly. At many companies the risk associated with experiments makes managers reluctant to allocate resources to them. But the gains enjoyed by companies that have made the leap of faith should give others the courage to follow them.

Many organizations are also too conservative about the nature and amount of experimentation. Overemphasizing the importance of successful experiments may encourage employees to focus on familiar solutions or those that they already know will work and avoid testing ideas that they fear might fail. And it's actually less risky to run a large

number of experiments than a small number. At Booking.com, only about 10% of experiments generate positive results—meaning that “B,” a modification that attempts to improve something (sales, repeat usage, click-through rates, or the time users spend on the site, for example), performs better among randomly assigned users than “A,” the control, which is the status quo. (In addition to A/B tests, Booking.com also runs more-complex tests that assess more than one modification at the same time.) But when you conduct a large volume of experiments, a low success rate still translates into a significant number of successes, which, in turn, diminish the financial and emotional costs of the failures. If a company does only a handful of experiments a year, it may have only one success or, if it's unlucky, none. Then failure is a big deal.

At the companies I studied, the success rate for ideas tested early in the development of a brand-new offering is even lower. Early failures, however, allow developers to quickly eliminate unfavorable options and refocus their efforts on more-promising alternatives.

In experimental cultures, employees are undaunted by the possibility of failure. “The people who thrive here are curious, open-minded, eager to learn and figure things out, and OK with being proven wrong,” said Vermeer, who now oversees all testing at Booking.com. The firm's recruiters look for such people, and to make sure they're empowered to follow their instincts, the company puts new hires through a rigorous onboarding process, which includes experimentation training, and then gives them access to all testing tools.

INSIST THAT DATA TRUMP OPINIONS

The empirical results of online experiments must prevail when they clash with strong opinions, no matter whose opinions they are. This is the attitude at Booking.com, but it's rare among most firms for an understandable reason: human nature. We tend to happily accept “good” results that confirm our biases but challenge and thoroughly investigate “bad” results that go against our assumptions.

The remedy is to implement the changes that experiments validate with few exceptions. As one director at Booking.com told me, “If the test tells you that the header of the website should be pink, then it should be pink. You always follow the test.”

Getting executives in the top ranks to abide by this rule isn't easy. (As the American writer Upton Sinclair once quipped, “It is difficult to get a man to understand something, when his salary depends upon his not understanding it!”) But it's vital that they do: Nothing stalls innovation faster than a so-called HiPPO—highest-paid person's opinion.

Note that I'm not saying that all management decisions can or should be based on online experiments. Some things are very hard, if not impossible, to conduct tests on—for example, strategic calls on whether to acquire a company.

But if everything that can be tested online is tested, experiments can become instrumental to management decisions and fuel healthy debates. Sometimes, those discussions might result in a conscious choice to overrule the data. That's what happened with one decision involving a comedy series at



Netflix, which has built a sophisticated infrastructure for large-scale experimentation. According to a *Wall Street Journal* article published in 2018, the company's executives were torn when tests showed that a promotion featuring an image of only Lily Tomlin, one of the stars of *Grace and Frankie*, resulted in more clicks by potential viewers than promotions featuring both Tomlin and her costar, Jane Fonda. The content team worried that excluding Fonda would alienate the actress and possibly violate her contract. After heated debates that pitted empirical evidence against “strategic considerations,” Netflix

chose to use images that included both actresses, even though customer data didn't support the decision. However, the experimental evidence made the trade-offs more transparent.

DEMOCRATIZE EXPERIMENTATION

As I've noted, any employee at Booking.com can launch an experiment on millions of customers without management's permission. About 75% of its 1,800 technology and product staffers actively use the company's experimentation platform. Standard templates allow them to set up tests with minimal effort,

and processes like user recruitment, randomization, the recording of visitors' behavior, and reporting are automated. A core experimentation team and five satellite teams used to provide training and support to the whole organization, but because the firm's needs evolved, that structure was recently changed to four central teams that report to Vermeer and specialists (“ambassadors”) that are placed in product teams.

To get things rolling, individuals or teams fill out an electronic form, which is visible to all and includes the name of the experiment, its purpose, the main beneficiaries (customers or suppliers),

Andreas Feininger/The LIFE Picture Collection via Getty Images



A long-exposure photograph by Andreas Feininger captures the light trail of a helicopter.

related past experiments, and the number of modifications to be tried out in A/B, A/B/C, or A/B/n tests. Once an experiment is up and running, the team watches it closely for the first few hours; if its primary or secondary metrics tank quickly, the team can stop the test. After that initial period, the platform continues to automatically run data-quality checks and sends warning messages if something is odd. To encourage openness, Booking.com maintains a central searchable repository of past experiments, with full descriptions of successes, failures, iterations, and final decisions. And everyone can see the real-time information generated by ongoing experiments.

“Somewhat ironically, the centralizing of our experimentation infrastructure is what makes our organizational decentralization possible,” Vermeer explained to me. “Everyone uses the same tools. This fosters trust in each other’s data and enables discussion and accountability. While some companies, like Microsoft, Facebook, and Google, may be more technically advanced in areas like machine learning, our use of simple A/B tests makes us more successful in getting all people involved; we have democratized testing throughout the organization.”

Democratization, of course, has its challenges. One is the risk that teams or individuals could break something on Booking.com’s high-traffic website, causing it to crash. Another is that each team has to set its own direction and figure out which user problems it wants to solve. That requires extensive training and ongoing discussions among team members about what the right problems are. Debates are encouraged, and people reach out to colleagues if

they see anything that strikes them as questionable. Just as anyone can launch an experiment, anybody can stop one. However, this happens only on the rare occasion when an experiment has gone catastrophically awry—for example, if someone is alone in the office at night and sees that an experiment is causing a key metric like the customer conversion rate to plunge and will cost the company millions of dollars in revenues if it continues.

This system gives teams the autonomy they need to try out new approaches they believe are valuable and allows people throughout the company to monitor the experiments and provide feedback in real time. It truly liberates everyone to test any idea about how to improve Booking.com’s business.

BE ETHICALLY SENSITIVE

When contemplating new experiments, companies must think carefully about whether users would consider the tests to be unethical. While the answer isn’t always clear-cut, organizations that fail to examine this question risk sparking a backlash. Take the weeklong experiment that Facebook ran in 2012 to learn whether emotional states were contagious on its platform. Facebook rejiggered its news feed—an algorithmically curated list of posts, stories, and activities—to see whether viewing fewer positive news stories led people to reduce their number of positive posts. The network also tested whether the reverse happened when people were exposed to fewer negative news stories. The experiment involved nearly 690,000 randomly selected users, about 310,000 of whom were unwittingly exposed to manipulated emotional expressions in their news feeds, while the rest were subjected to control conditions in which a corresponding number of randomly chosen posts were omitted.

When researchers from Facebook and Cornell University published the results in an academic journal, public outrage

erupted. Facebook’s data science team had been running experiments on unsuspecting users for years without controversy, but the emotional manipulation struck a nerve. Critics raised concerns about whether the participants’ consent to Facebook’s general data-use policy sufficed; they felt the company should have made it clearer that users could opt out of testing and that data was collected for research. From a learning perspective, the experiment was a success: It found that emotional contagion existed online, though the effect was very small. But some users felt that Facebook had exploited them in the name of science.

Research suggests that companies that test new ideas first face greater customer scrutiny than competitors that implement new practices without conducting any experiments. In a published analysis of 16 studies in domains such as health care, vehicle design, and global poverty, bioethicist Michelle Meyer and her collaborators concluded that participants considered A/B tests to be more morally questionable than the universal implementation of an untested practice (A or B) on the entire population—even when both treatments were unobjectionable.

Clearly, ethics training and some kind of oversight are necessary. The challenge is conducting the latter in ways that don’t make people overly cautious or tangle them in red tape. For those precise reasons, Booking.com has shied away from imposing rules from on high about what kind of tests can be run. Instead, it encourages employees to ask whether an experiment or proposed practice would help or hurt customers. “I’d rather stay away from policing or

Spotlight

ethical review boards,” David Vismans, Booking.com’s chief product officer, told me. “That’s not a scalable solution. You’d create a bottleneck, and testing police don’t make people feel like they’re empowered.” Instead, the company encourages debates in internal online forums that are open to all employees. The debates can be vigorous and have tackled issues like the use of techniques to persuade customers to complete transactions (for example, messages such as “Please book now or you will lose this reservation” or “Only three rooms left”). “I would rather have a community that is self-correcting,” Vismans explained.

To that end, Booking.com’s onboarding process also includes ethics training. LinkedIn, another company with a large experimentation program, takes a slightly different approach. It has created internal guidelines that state the company won’t run experiments “that are intended to deliver a negative member experience, have a goal of altering members’ moods or emotions, or override existing members’ settings or choices.”

EMBRACE A DIFFERENT LEADERSHIP MODEL

By democratizing experimentation and following test results where they lead, companies can enable employees to make good decisions on their own and accelerate innovation and improvements. But if most decisions are made this way, what’s left for senior leaders to do, beyond developing the company’s strategic direction and tackling big decisions such as which acquisitions to make? There are at least four things:

Set a grand challenge that can be broken into testable hypotheses and key performance metrics. Employees need to see how their experiments support an overall strategic goal. Say Booking.com’s senior leaders challenged employees to design the best online experience in the industry. They might expect that a superior experience would generate more customer traffic, which would attract more suppliers to Booking.com’s platform, helping expand the customer base and activity even more. To discover ways to pursue that goal, employees could devise hypotheses and related metrics—for instance, that underlining important text would increase conversion rates by making critical information easier to find, and that a “one click, no cost” cancellation option would boost user return rates without causing net hotel bookings to drop.

Put in place systems, resources, and organizational designs that allow for large-scale experimentation. Scientifically testing nearly every idea requires infrastructure: instrumentation, data pipelines, and data scientists. Several third-party tools and services make it easy to try experiments, but to scale things up, senior leaders must tightly integrate the testing capability into company processes. Doing so requires striking the right balance between centralization and decentralization.

In centralized groups, dedicated specialists such as developers, user interface designers, and data analysts can run experiments for the entire company and focus on introducing state-of-the-art methods and tools. But if testing is limited to a small group of specialists, it will be hard to scale up experimentation and

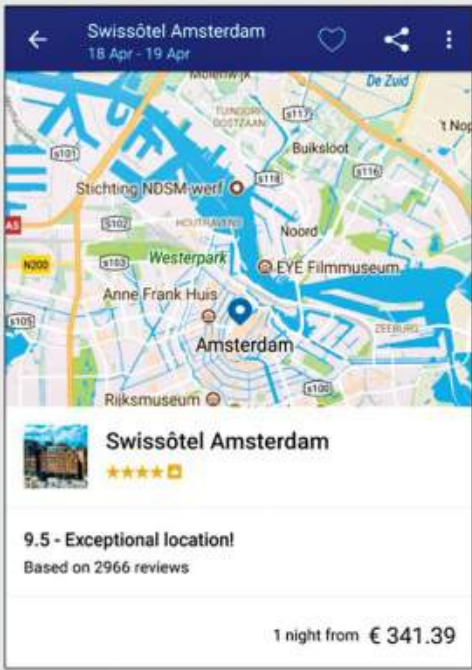
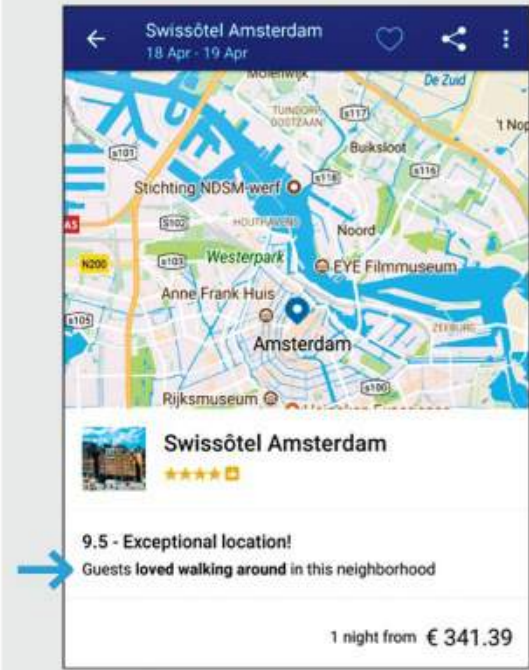
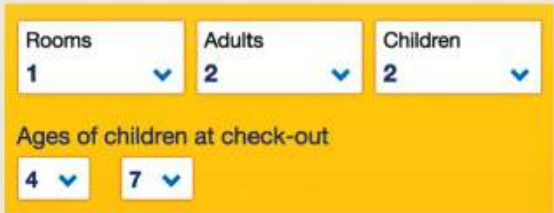

change a company’s culture. In decentralized testing, firms spread specialist teams throughout different business units. While this approach expands experimentation to more parts of the organization, it can hinder knowledge sharing and lead to conflicting goals and poor coordination among specialists. Decentralization may be needed to get the broader organization involved at first, but after that, firms should turn to improving their experimentation capabilities. That’s what Booking.com did. It initially used satellite teams to spread experimentation across the company but found that they were too busy supporting users to focus on building firmwide capabilities. To address that problem and align the teams better, Booking.com recently switched to a center-of-excellence model that supports business units, standardizes the company’s approach to experimentation, and makes sure that best practices are adopted and followed.

Be a role model. Leaders have to live by the same rules as everyone else and subject their own ideas to tests. “You can’t have an ego, thinking that you always know best,” Tans told me. “If I, as the CEO, say to someone, ‘This is what I want you to do because I think it’s good for our business,’ employees would literally look at me and say, ‘OK, that’s fine, we are going to test it and see if you are right.’” Bosses ought to display intellectual humility and be unafraid to admit, “I don’t know.” They should heed the advice of Francis Bacon, the father of the scientific method: “If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties.”

Recognize that words alone won’t change behavior. Ultimately, being a leader in an experiment-driven organization means letting go and empowering employees to perform their own tests—which doesn’t happen by simply telling people that they can do so. It requires a concerted effort like IBM’s.

How Booking.com Experiments with Site Improvements

Every day, employees at the company use A/B tests to try out their ideas for tweaks. Below are two examples.

SCENARIO #1		SCENARIO #2
<p>Hypothesis</p> <p>Highlighting a neighborhood's walkability helps users make better decisions about property location.</p>		<p>Hypothesis</p> <p>Displaying the checkout date when users select the age of children in their party improves their experience.</p>
<p>A</p> <p>The Control</p> <p>Shows the site's current practice</p> 	<p>B</p> <p>The Treatment</p> <p>Adds walkability information</p> 	<p>A</p> <p>The Control</p> <p>Shows the site's current practice</p> 
<p>The Result</p> <p>The treatment had no significant impact on the key metric. The current practice is kept in place.</p>		<p>B</p> <p>The Treatment</p> <p>Adds the checkout date above children's ages</p> 
<p>The Result</p> <p>The treatment had a significant positive impact on the key metric, and the change is implemented.</p>		<p>The Result</p> <p>The treatment had a significant positive impact on the key metric, and the change is implemented.</p>

In 2015 experimentation wasn't a core activity at IBM; the company's IT function offered to run tests, but they were costly, were charged back to business units, and had to follow a rigid process. The testing capacity consisted of just one specialist, who was also the gatekeeper and who rejected many proposed experiments because he felt that they weren't strong-enough candidates. As a result, the company ran only 97 tests that year.

Then, Ari Sheinkin, IBM's head of marketing analytics at the time, took over experimentation and, with the backing of the chief marketing officer, empowered over 5,500 marketers worldwide to conduct their own tests. To induce them to do so, Sheinkin took a number of steps. He installed easy-to-use tools, created a center of excellence to provide support, introduced a framework for conducting disciplined experiments,

offered training for everyone, and made online tests free for all business groups. He also conducted an initial "testing blitz" during which the marketing units had to run a total of 30 online experiments in 30 days. After that, he held quarterly contests for the most innovative or most scalable experiments. He also employed more-forceful tactics: IBM tied part of marketing units' budgets to experimentation plans. These efforts