

ing novels and biographies. But these hold little interest for me as well. I always find myself asking, "Would that happen in other situations? What's the more general principle?" Their stories feel small and unrepresentative.

What I have tried to present in this book is something that, for me, is like nothing else. It is based on data and numbers; it is illustrative and far-reaching. And yet the data is so rich that you can visualize the people underneath it. When we zoom in on every minute of Edmonton's water consumption, I *see* the people getting up from their couch at the end of the period. When we zoom in on people moving from Philadelphia to Miami and starting to cheat on their taxes, I *see* these people talking to their neighbors in their apartment complex and learning about the tax trick. When we zoom in on baseball fans of every age, I *see* my own childhood and my brother's childhood and millions of adult men still crying over a team that won them over when they were eight years old.

At the risk of once again sounding grandiose, I think the economists and data scientists featured in this book are creating not only a new tool but a new genre. What I have tried to present in this chapter, and much of this book, is data so big and so rich, allowing us to zoom in so close that, without limiting ourselves to any particular, unrepresentative human being, we can still tell complex and evocative stories.

6

ALL THE WORLD'S A LAB

February 27, 2000, started as an ordinary day on Google's Mountain View campus. The sun was shining, the bikers were pedaling, the masseuses were massaging, the employees were hydrating with cucumber water. And then, on this ordinary day, a few Google engineers had an idea that unlocked the secret that today drives much of the internet. The engineers found the best way to get you clicking, coming back, and staying on their sites.

Before describing what they did, we need to talk about correlation versus causality, a huge issue in data analysis—and one that we have not yet adequately addressed.

The media bombard us with correlation-based studies seemingly every day. For example, we have been told that those of us who drink a moderate amount of alcohol tend to be in better health. That is a correlation.

Does this mean drinking a moderate amount will improve

one's health—a causation? Perhaps not. It could be that good health causes people to drink a moderate amount. Social scientists call this reverse causation. Or it could be that there is an independent factor that causes both moderate drinking and good health. Perhaps spending a lot of time with friends leads to both moderate alcohol consumption and good health. Social scientists call this omitted-variable bias.

How, then, can we more accurately establish causality? The gold standard is a randomized, controlled experiment. Here's how it works. You randomly divide people into two groups. One, the treatment group, is asked to do or take something. The other, the control group, is not. You then see how each group responds. The difference in the outcomes between the two groups is your causal effect.

For example, to test whether moderate drinking causes good health, you might randomly pick some people to drink one glass of wine per day for a year, randomly choose others to drink no alcohol for a year, and then compare the reported health of both groups. Since people were randomly assigned to the two groups, there is no reason to expect one group would have better initial health or have socialized more. You can trust that the effects of the wine are causal. Randomized, controlled experiments are the most trusted evidence in any field. If a pill can pass a randomized, controlled experiment, it can be dispensed to the general populace. If it cannot pass this test, it won't make it onto pharmacy shelves.

Randomized experiments have increasingly been used in the social sciences as well. Esther Duflo, a French economist at MIT, has led the campaign for greater use of experiments in developmental economics, a field that tries to figure out the best

ways to help the poorest people in the world. Consider Duflo's study, with colleagues, of how to improve education in rural India, where more than half of middle school students cannot read a simple sentence. One potential reason students struggle so much is that teachers don't show up consistently. On a given day in some schools in rural India, more than 40 percent of teachers are absent.

Duflo's test? She and her colleagues randomly divided schools into two groups. In one (the treatment group), in addition to their base pay, teachers were paid a small amount—50 rupees, or about \$1.15—for every day they showed up to work. In the other, no extra payment for attendance was given. The results were remarkable. When teachers were paid, teacher absenteeism dropped in half. Student test performance also improved substantially, with the biggest effects on young girls. By the end of the experiment, girls in schools where teachers were paid to come to class were 7 percentage points more likely to be able to write.

According to a *New Yorker* article, when Bill Gates learned of Duflo's work, he was so impressed he told her, "We need to fund you."

THE ABCS OF A/B TESTING

So randomized experiments are the gold standard for proving causality, and their use has spread through the social sciences. Which brings us back to Google's offices on February 27, 2000. What did Google do on that day that revolutionized the internet?

On that day, a few engineers decided to perform an experiment on Google's site. They randomly divided users into two groups. The treatment group was shown twenty links on the search results pages. The control group was shown the usual ten. The engineers then compared the satisfaction of the two groups based on how frequently they returned to Google.

This is a revolution? It doesn't seem so revolutionary. I already noted that randomized experiments have been used by pharmaceutical companies and social scientists. How can copying them be such a big deal?

The key point—and this was quickly realized by the Google engineers—is that experiments in the digital world have a huge advantage relative to experiments in the offline world. As convincing as offline randomized experiments can be, they are also resource-intensive. For Duflo's study, schools had to be contacted, funding had to be arranged, some teachers had to be paid, and all students had to be tested. Offline experiments can cost thousands or hundreds of thousands of dollars and take months or years to conduct.

In the digital world, randomized experiments can be cheap and fast. You don't need to recruit and pay participants. Instead, you can write a line of code to randomly assign them to a group. You don't need users to fill out surveys. Instead, you can measure mouse movements and clicks. You don't need to hand-code and analyze the responses. You can build a program to automatically do that for you. You don't have to contact anybody. You don't even have to tell users they are part of an experiment.

This is the fourth power of Big Data: it makes randomized experiments, which can find truly causal effects, much, much easier to conduct—anytime, more or less anywhere, as long as you're online. In the era of Big Data all the world's a lab.

This insight quickly spread through Google and then the rest of Silicon Valley, where randomized controlled experiments have been renamed "A/B testing." In 2011, Google engineers ran seven thousand A/B tests. And this number is only rising.

If Google wants to know how to get more people to click on ads on their sites, they may try two shades of blue in ads—one shade for Group A, another for Group B. Google can then compare click rates. Of course, the ease of such testing can lead to overuse. Some employees felt that because testing was so effortless, Google was overexperimenting. In 2009, one frustrated designer quit after Google went through forty-one marginally different shades of blue in A/B testing. But this designer's stand in favor of art over obsessive market research has done little to stop the spread of the methodology.

Facebook now runs a thousand A/B tests per day, which means that a small number of engineers at Facebook start more randomized, controlled experiments in a given day than the entire pharmaceutical industry starts in a year.

A/B testing has spread beyond the biggest tech firms. A former Google employee, Dan Siroker, brought this methodology to Barack Obama's first presidential campaign, which A/B-tested home page designs, email pitches, and donation forms. Then Siroker started a new company, Optimizely, which allows organizations to perform rapid A/B testing. In 2012,

Optimizely was used by Obama as well as his opponent, Mitt Romney, to maximize sign-ups, volunteers, and donations. It's also used by companies as diverse as Netflix, TaskRabbit, and *New York* magazine.

To see how valuable testing can be, consider how Obama used it to get more people engaged with his campaign. Obama's home page initially included a picture of the candidate and a button below the picture that invited people to "Sign Up."



Was this the best way to greet people? With the help of Siroker, Obama's team could test whether a different picture and button might get more people to actually sign up. Would more people click if the home page instead featured a picture of Obama with a more solemn face? Would more people click if the button instead said "Join Now"? Obama's team showed users different combinations of pictures and buttons and measured how many of them clicked the button. See if you can predict the winning picture and winning button.

Pictures Tested



Buttons Tested



The winner was the picture of Obama's family and the button "Learn More." And the victory was huge. By using that combination, Obama's campaign team estimated it got 40 per-

Winning Combination



cent more people to sign up, netting the campaign roughly \$60 million in additional funding.

There is another great benefit to the fact that all this gold-standard testing can be done so cheap and easy: it further frees us from our reliance upon our intuition, which, as noted in Chapter 1, has its limitations. A fundamental reason for A/B testing's importance is that people are unpredictable. Our intuition often fails to predict how they will respond.

Was your intuition correct on Obama's optimal website?

Here are some more tests for your intuition. The *Boston Globe* A/B-tests headlines to figure out which ones get the most people to click on a story. Try to guess the winners from these pairs:

ONE OF THESE HEADLINES WAS WAY BETTER THAN THE OTHER IN GETTING CLICKS.

	HEADLINE A	HEADLINE B
1.	Can the SnotBot drone save the whales?	Can this drone help save the whales?
2.	Of course "deflated balls" is a top search term in Massachusetts	This top Mass. Google search term is pretty embarrassing
3.	Hookup contest at heart of St. Paul rape trial	No charges in prep school sex scandal
4.	Woman makes bank off rare baseball card	Woman makes \$179,000 off rare baseball card
5.	MBTA projects annual operating deficit will double by 2020	Get ready: the MBTA's deficit is about to double
6.	How Massachusetts helped win you the right to birth control access	How Boston University helped end "crimes against chastity"
7.	When the first subway opened in Boston	Cartoons from when the first subway opened in Boston
8.	Victim and family in prep-school rape trial blame toxic culture	Victim and family in prep-school rape trial releases statement
9.	Guy in "Free Brady" hat is only one able to foil Miley Cyrus prank	Pats fan gets an eyeful for recognizing an undercover Miley Cyrus

Finished your guesses? The answers are in bold below.

	HEADLINE A	HEADLINE B	WINNER?
1.	Can the SnotBot drone save the whales?	Can this drone help save the whales?	53% more clicks for A
2.	Of course "deflated balls" is a top search term in Massachusetts	This top Mass. Google search term is pretty embarrassing	986% more clicks for B

(continued on next page)

	HEADLINE A	HEADLINE B	WINNER?
3.	Hookup contest at heart of St. Paul rape trial	No charges in prep school sex scandal	108% more clicks for B
4.	Woman makes bank off rare baseball card	Woman makes \$179,000 off rare baseball card	38% more clicks for A
5.	MBTA projects annual operating deficit will double by 2020	Get ready: the MBTA's deficit is about to double	62% more clicks for B
6.	How Massachusetts helped win you the right to birth control access	How Boston University helped end "crimes against chastity"	188% more clicks for B
7.	When the first subway opened in Boston	Cartoons from when the first subway opened in Boston	33% more clicks for A
8.	Victim and family in prep-school rape trial blame toxic culture	Victim and family in prep-school rape trial releases statement	76% more clicks for B
9.	Guy in "Free Brady" hat is only one able to foil Miley Cyrus prank	Pats fan gets an eyeful for recognizing an undercover Miley Cyrus	67% more clicks for B

I predict you got more than half right, perhaps by considering what you would click on. But you probably did not guess all of these correctly.

Why? What did you miss? What insights into human behavior did you lack? What lessons can you learn from your mistakes?

We usually ask questions such as these after making bad predictions.

But look how difficult it is to draw general conclusions from the *Globe* headlines. In the first headline test, changing a single word, "this" to "SnotBot," led to a big win. This might suggest

more details win. But in the second headline, "deflated balls," the detailed term, loses. In the fourth headline, "makes bank" beats the number \$179,000. This might suggest slang terms win. But the slang term "hookup contest" loses in the third headline.

The lesson of A/B testing, to a large degree, is to be wary of general lessons. Clark Benson is the CEO of ranker.com, a news and entertainment site that relies heavily on A/B testing to choose headlines and site design. "At the end of the day, you can't assume anything," Benson says. "Test literally everything."

Testing fills in gaps in our understanding of human nature. These gaps will always exist. If we knew, based on our life experience, what the answer would be, testing would not be of value. But we don't, so it is.

Another reason A/B testing is so important is that seemingly small changes can have big effects. As Benson puts it, "I'm constantly amazed with minor, minor factors having out-sized value in testing."

In December 2012, Google changed its advertisements. They added a rightward-pointing arrow surrounded by a square.

Hotels
www.example.com
 Special rates until the end of the month. No booking fees, book your room now!
 

Dublin Hotels
www.example.com
 Browse hundreds of hotels in Dublin, sort by price, location and user reviews.
 

Hotels in Ireland
www.example.com
 Compare prices of 1000s of hotels all over Ireland!
 

AdChoices 

Notice how bizarre this arrow is. It points rightward to absolutely nothing. In fact, when these arrows first appeared,

many Google customers were critical. Why were they adding meaningless arrows to the ad, they wondered?

Well, Google is protective of its business secrets, so they don't say exactly how valuable the arrows were. But they did say that these arrows had won in A/B testing. The reason Google added them is that they got a lot more people to click. And this minor, seemingly meaningless change made Google and their ad partners oodles of money.

So how can you find these small tweaks that produce outsize profits? You have to test lots of things, even many that seem trivial. In fact, Google's users have noted numerous times that ads have been changed a tiny bit only to return to their previous form. They have unwittingly become members of treatment groups in A/B tests—but at the cost only of seeing these slight variations.

Centering Experiment (Didn't Work)

Best Selling iPad 2 Case

The ZAGGmate™ - Tough Aluminum Case
with built in Bluetooth Keyboard
www.zagg.com

Green Star Experiment (Didn't Work)

Foster's Hollywood Restaurant Reviews, Madrid, Spain ...
www.tripadvisor.co.uk > ... > Madrid > Madrid Restaurants ~ TripAdvisor ~
★★★★ Rating: 3 - 118 reviews
Foster's Hollywood, Madrid: See 118 unbiased reviews of Foster's Hollywood, rated 3 of 5 on TripAdvisor and ranked #3647 of 6489 restaurants in Madrid

New Font Experiment (Didn't Work)

Live Stock Market News

Free Charts, News and Tips from UTVi Experts. Visit us
Today!
UTVi.com/Stocks

The above variations never made it to the masses. They lost. But they were part of the process of picking winners. The road to a clickable arrow is paved with ugly stars, faulty positionings, and gimmicky fonts.

It may be fun to guess what makes people click. And if you are a Democrat, it might be nice to know that testing got Obama more money. But there is a dark side to A/B testing.

In his excellent book *Irresistible*, Adam Alter writes about the rise of behavioral addictions in contemporary society. Many people are finding aspects of the internet increasingly difficult to turn off.

My favorite dataset, Google searches, can give us some clues as to what people find most addictive. According to Google, most addictions remain the ones people have struggled with for many decades—drugs, sex, and alcohol, for example. But the internet is starting to make its presence felt on the list—with “porn” and “Facebook” now among the top ten reported addictions.

TOP ADDICTIONS REPORTED TO GOOGLE, 2016

Drugs	Alcohol	Gambling
Sex	Sugar	Facebook
Porn	Love	

A/B testing may be playing a role in making the internet so darn addictive.

Tristan Harris, a “design ethicist,” was quoted in *Irresistible*

explaining why people have such a hard time resisting certain sites on the internet: "There are a thousand people on the other side of the screen whose job it is to break down the self-regulation you have."

And these people are using A/B testing.

Through testing, Facebook may figure out that making a particular button a particular color gets people to come back to their site more often. So they change the button to that color. Then they may figure out that a particular font gets people to come back to their site more often. So they change the text to that font. Then they may figure out that emailing people at a certain time gets them coming back to their site more often. So they email people at that time.

Pretty soon, Facebook becomes a site optimized to maximize how much time people spend on Facebook. In other words, find enough winners of A/B tests and you have an addictive site. It is the type of feedback that cigarette companies never had.

A/B testing is increasingly a tool of the gaming industry. As Alter discusses, *World of Warcraft* A/B-tests various versions of its game. One mission might ask you to kill someone. Another might ask you to save something. Game designers can give different samples of players' different missions and then see which ones keep more people playing. They might find, for example, that the mission that asked you to save a person got people to return 30 percent more often. If they test many, many missions, they start finding more and more winners. These 30 percent wins add up, until they have a game that keeps many adult men holed up in their parents' basement.

If you are a little disturbed by this, I am with you. And we

will talk a bit more about the ethical implications of this and other aspects of Big Data near the end of this book. But for better or worse, experimentation is now a crucial tool in the data scientists' tool kit. And there is another form of experimentation sitting in that tool kit. It has been used to ask a variety of questions, including whether TV ads really work.

NATURE'S CRUEL—BUT ENLIGHTENING—EXPERIMENTS

It's January 22, 2012, and the New England Patriots are playing the Baltimore Ravens in the AFC Championship game.

There's a minute left in the game. The Ravens are down, but they've got the ball. The next sixty seconds will determine which team will play in the Super Bowl. The next sixty seconds will help seal players' legacies. And the last minute of this game will do something that, for an economist, is far more profound: the last sixty seconds will help finally tell us, once and for all, Do advertisements work?

The notion that ads improve sales is obviously crucial to our economy. But it is maddeningly hard to prove. In fact, this is a textbook example of exactly how difficult it is to distinguish between correlation and causation.

There's no doubt that products that advertise the most also have the highest sales. Twentieth Century Fox spent \$150 million marketing the movie *Avatar*, which became the highest-grossing film of all time. But how much of the \$2.7 billion in *Avatar* ticket sales was due to the heavy marketing? Part of the reason 20th Century Fox spent so much money on pro-