Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing

3 authors, including:

Some of the authors of this publication are also working on these related projects:

Project   Online Controlled Experiments and A/B Testing View project

Project   Machine Learning C++ - MLC++ View project

Trustworthy Online Controlled Experiments

A Practical Guide to A/B Testing

Getting numbers is easy; getting numbers you can trust is hard. This practical guide by experimentation leaders at Google, LinkedIn, and Microsoft will teach you how to accelerate innovation using trustworthy online controlled experiments, or A/B tests. Based on practical experiences at companies that each runs more than 20,000 controlled experiments a year, the authors share examples, pitfalls, and advice for students and industry professionals getting started with experiments, plus deeper dives into advanced topics for experienced practitioners who want to improve the way they and their organizations make data-driven decisions.

Learn how to:

- Use the scientific method to evaluate hypotheses using controlled experiments
- Define key metrics and ideally an Overall Evaluation Criterion
- Test for trustworthiness of the results and alert experimenters to violated assumptions
- Interpret and iterate quickly based on the results
- Implement guardrails to protect key business goals
- Build a scalable platform that lowers the marginal cost of experiments close to zero
- Avoid pitfalls such as carryover effects, Twyman's law, Simpson's paradox, and network interactions
- Understand how statistical issues play out in practice, including common violations of assumptions

RON KOHAVI is a vice president and technical fellow at Airbnb. This book was written while he was a technical fellow and corporate vice president at Microsoft. He was previously director of data mining and personalization at Amazon. He received his PhD in Computer Science from Stanford University. His papers have more than 40,000 citations and three of them are in the top 1,000 most-cited papers in Computer Science.

DIANE TANG is a Google Fellow, with expertise in large-scale data analysis and infrastructure, online controlled experiments, and ads systems. She has an AB from Harvard and an MS/PhD from Stanford, with patents and publications in mobile networking, information visualization, experiment methodology, data infrastructure, data mining, and large data.

YA XU heads Data Science and Experimentation at LinkedIn. She has published several papers on experimentation and is a frequent speaker at top-tier conferences and universities. She previously worked at Microsoft and received her PhD in Statistics from Stanford University.

"At the core of the Lean Methodology is the scientific method: Creating hypotheses, running experiments, gathering data, extracting insight and validation or modification of the hypothesis. A/B testing is the gold standard of creating verifiable and repeatable experiments, and this book is its definitive text."
– Steve Blank, Adjunct professor at Stanford University, father of modern entrepreneurship, author of *The Startup Owner's Manual* and *The Four Steps to the Epiphany*

"This book is a great resource for executives, leaders, researchers or engineers looking to use online controlled experiments to optimize product features, project efficiency or revenue. I know firsthand the impact that Kohavi's work had on Bing and Microsoft, and I'm excited that these learnings can now reach a wider audience."
– Harry Shum, EVP, Microsoft Artificial Intelligence and Research Group

"A great book that is both rigorous and accessible. Readers will learn how to bring trustworthy controlled experiments, which have revolutionized internet product development, to their organizations"
– Adam D'Angelo, Co-founder and CEO of Quora and former CTO of Facebook

"This book is a great overview of how several companies use online experimentation and A/B testing to improve their products. Kohavi, Tang and Xu have a wealth of experience and excellent advice to convey, so the book has lots of practical real world examples and lessons learned over many years of the application of these techniques at scale."
– Jeff Dean, Google Senior Fellow and SVP Google Research

"Do you want your organization to make consistently better decisions? This is the new bible of how to get from data to decisions in the digital age. Reading this book is like sitting in meetings inside Amazon, Google, LinkedIn, Microsoft. The authors expose for the first time the way the world's most successful companies make decisions. Beyond the admonitions and anecdotes of normal business books, this book shows what to do and how to do it well. It's the how-to manual for decision-making in the digital world, with dedicated sections for business leaders, engineers, and data analysts."
– Scott Cook, Intuit Co-founder & Chairman of the Executive Committee

"Online controlled experiments are powerful tools. Understanding how they work, what their strengths are, and how they can be optimized can illuminate both specialists and a wider audience. This book is the rare combination of technically authoritative, enjoyable to read, and dealing with highly important matters"
– John P.A. Ioannidis, Professor of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics at Stanford University

"Which online option will be better? We frequently need to make such choices, and frequently err. To determine what will actually work better, we need rigorous controlled experiments, aka A/B testing. This excellent and lively book by experts from Microsoft, Google, and LinkedIn presents the theory and best practices of A/B testing. A must read for anyone who does anything online!"
– Gregory Piatetsky-Shapiro, Ph.D., president of KDnuggets, co-founder of SIGKDD, and LinkedIn Top Voice on Data Science & Analytics.

"Ron Kohavi, Diane Tang and Ya Xu are the world's top experts on online experiments. I've been using their work for years and I'm delighted they have now teamed up to write the definitive guide. I recommend this book to all my students and everyone involved in online products and services."
– Erik Brynjolfsson, Professor at MIT and Co-Author of
The Second Machine Age

"A modern software-supported business cannot compete successfully without online controlled experimentation. Written by three of the most experienced leaders in the field, this book presents the fundamental principles, illustrates them with compelling examples, and digs deeper to present a wealth of practical advice. It's a "must read"!
– Foster Provost, Professor at NYU Stern School of Business & co-author of the
best-selling *Data Science for Business*

"In the past two decades the technology industry has learned what scientists have known for centuries: that controlled experiments are among the best tools to understand complex phenomena and to solve very challenging problems. The ability to design controlled experiments, run them at scale, and interpret their results is the foundation of how modern high tech businesses operate. Between them the authors have designed and implemented several of the world's most powerful experimentation platforms. This book is a great opportunity to learn from their experiences about how to use these tools and techniques."
– Kevin Scott, EVP and CTO of Microsoft

"Online experiments have fueled the success of Amazon, Microsoft, LinkedIn and other leading digital companies. This practical book gives the reader rare access to decades of experimentation experience at these companies and should be on the bookshelf of every data scientist, software engineer and product manager."
– Stefan Thomke, William Barclay Harding Professor, Harvard Business School,
Author of *Experimentation Works: The Surprising Power of Business Experiments*

"The secret sauce for a successful online business is experimentation. But it is a secret no longer. Here three masters of the art describe the ABCs of A/B testing so that you too can continuously improve your online services."
– Hal Varian, Chief Economist, Google, and author of
Intermediate Microeconomics: A Modern Approach

"Experiments are the best tool for online products and services. This book is full of practical knowledge derived from years of successful testing at Microsoft Google and LinkedIn. Insights and best practices are explained with real examples and pitfalls, their markers and solutions identified. I strongly recommend this book!"
– Preston McAfee, former Chief Economist and VP of Microsoft

"Experimentation is the future of digital strategy and 'Trustworthy Experiments' will be its Bible. Kohavi, Tang and Xu are three of the most noteworthy experts on experimentation working today and their book delivers a truly practical roadmap for digital experimentation that is useful right out of the box. The revealing case studies they conducted over many decades at Microsoft, Amazon, Google and LinkedIn are organized into easy to understand practical lessens with tremendous depth and clarity. It should be required reading for any manager of a digital business."
– Sinan Aral, David Austin Professor of Management,
MIT and author of *The Hype Machine*

"Indispensable for any serious experimentation practitioner, this book is highly practical and goes in-depth like I've never seen before. It's so useful it feels like you get a superpower. From statistical nuances to evaluating outcomes to measuring long term impact, this book has got you covered. Must-read."

<div align="right">– Peep Laja, top conversion rate expert, Founder and Principal of CXL</div>

"Online experimentation was critical to changing the culture at Microsoft. When Satya talks about "Growth Mindset," experimentation is the best way to try new ideas and learn from them. Learning to quickly iterate controlled experiments drove Bing to profitability, and rapidly spread across Microsoft through Office, Windows, and Azure."

<div align="right">– Eric Boyd, Corporate VP, AI Platform, Microsoft</div>

"As an entrepreneur, scientist, and executive I've learned (the hard way) that an ounce of data is worth a pound of my intuition. But how to get good data? This book compiles decades of experience at Amazon, Google, LinkedIn, and Microsoft into an accessible, well-organized guide. It is the bible of online experiments."

<div align="right">– Oren Etzioni, CEO of Allen Institute of AI and
Professor of Computer Science at University of Washington</div>

"Internet companies have taken experimentation to an unprecedented scale, pace, and sophistication. These authors have played key roles in these developments and readers are fortunate to be able to learn from their combined experiences."

<div align="right">– Dean Eckles, KDD Career Development Professor in Communications and
Technology at MIT and former scientist at Facebook</div>

"A wonderfully rich resource for a critical but under-appreciated area. Real case studies in every chapter show the inner workings and learnings of successful businesses. The focus on developing and optimizing an "Overall Evaluation Criterion" (OEC) is a particularly important lesson."

<div align="right">– Jeremy Howard, Singularity University, founder of fast.ai,
and former president and chief scientist of Kaggle</div>

"There are many guides to A/B Testing, but few with the pedigree of Trustworthy Online Controlled Experiments. I've been following Ronny Kohavi for eighteen years and find his advice to be steeped in practice, honed by experience, and tempered by doing laboratory work in real world environments. When you add Diane Tang, and Ya Xu to the mix, the breadth of comprehension is unparalleled. I challenge you to compare this tome to any other - in a controlled manner, of course."

<div align="right">– Jim Sterne, Founder of Marketing Analytics Summit and
Director Emeritus of the Digital Analytics Association</div>

"An extremely useful how-to book for running online experiments that combines analytical sophistication, clear exposition and the hard-won lessons of practical experience."

<div align="right">– Jim Manzi, Founder of Foundry.ai, Founder and former CEO and
Chairman of Applied Predictive Technologies, and author of *Uncontrolled:
The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*</div>

"Experimental design advances each time it is applied to a new domain: agriculture, chemistry, medicine and now online electronic commerce. This book by three top experts is rich in practical advice and examples covering both how and why to experiment online and not get fooled. Experiments can be expensive; not knowing what works can cost even more."

– Art Owen Professor of Statistics, Stanford University

"This is a must read book for business executives and operating managers. Just as operations, finance, accounting and strategy form the basic building blocks for business, today in the age of AI, understanding and executing online controlled experiments will be a required knowledge set. Kohavi, Tang and Xu have laid out the essentials of this new and important knowledge domain that is practically accessible."

– Karim R. Lakhani, Professor and Director of Laboratory for Innovation Science at Harvard, Board Member, Mozilla Corp.

"Serious 'data-driven' organizations understand that analytics aren't enough; they must commit to experiment. Remarkably accessible and accessibly remarkable, this book is a manual and manifesto for high-impact experimental design. I found its pragmatism inspirational. Most importantly, it clarifies how culture rivals technical competence as a critical success factor."

– Michael Schrage, research fellow at MIT's Initiative on the Digital Economy and author of *The Innovator's Hypothesis: How Cheap Experiments Are Worth More than Good Ideas*

"This important book on experimentation distills the wisdom of three distinguished leaders from some of the world's biggest technology companies. If you are a software engineer, data scientist, or product manager trying to implement a data-driven culture within your organization, this is an excellent and practical book for you."

– Daniel Tunkelang, Chief Scientist at Endeca and former Director of Data Science and Engineering at LinkedIn

"With every industry becoming digitized and data-driven, conducting and benefiting from controlled online experiments becomes a required skill. Kohavi, Tang and Yu provide a complete and well-researched guide that will become necessary reading for data practitioners and executives alike."

– Evangelos Simoudis, Co-founder and Managing Director Synapse Partners; author of *The Big Data Opportunity in Our Driverless Future*

"The authors offer over 10 years of hard-fought lessons in experimentation, in the most strategic book for the discipline yet"

– Colin McFarland, Director Experimentation Platform at Netflix

"The practical guide to A/B testing distills the experiences from three of the top minds in experimentation practice into easy and digestible chunks of valuable and practical concepts. Each chapter walks you through some of the most important considerations when running experiments - from choosing the right metric to the benefits of institutional memory. If you are looking for an experimentation coach that balances science and practicality, then this book is for you."

– Dylan Lewis, Experimentation Leader, Intuit

"The only thing worse than no experiment is a misleading one, because it gives you false confidence! This book details the technical aspects of testing based on insights from some of the world's largest testing programs. If you're involved in online experimentation in any capacity, read it now to avoid mistakes and gain confidence in your results."

- Chris Goward, Author of *You Should Test That!*,
Founder and CEO of Widerfunnel

"This is a phenomenal book. The authors draw on a wealth of experience and have produced a readable reference that is somehow both comprehensive and detailed at the same time. Highly recommended reading for anyone who wants to run serious digital experiments."

- Pete Koomen, Co-founder, Optimizely

"The authors are pioneers of online experimentation. The platforms they've built and the experiments they've enabled have transformed some of the largest internet brands. Their research and talks have inspired teams across the industry to adopt experimentation. This book is the authoritative yet practical text that the industry has been waiting for."

– Adil Aijaz, Co-founder and CEO, Split Software

Trustworthy Online Controlled Experiments

A Practical Guide to A/B Testing

RON KOHAVI
Microsoft

DIANE TANG
Google

YA XU
LinkedIn

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Contents

Preface

How to Read This Book

If we have data, let's look at data.
If all we have are opinions, let's go with mine
— Jim Barksdale, Former CEO of Netscape

Our goal in writing this book is to share practical lessons from decades of experience running online controlled experiments at scale at Amazon and Microsoft (Ron), Google (Diane), and Microsoft and LinkedIn (Ya). While we are writing this book in our capacity as individuals and not as representatives of Google, LinkedIn, or Microsoft, we have distilled key lessons and pitfalls encountered over the years and provide guidance for both software platforms and the corporate cultural aspects of using online controlled experiments to establish a data-driven culture that informs rather than relies on the HiPPO (Highest Paid Person's Opinion) (R. Kohavi, HiPPO FAQ 2019). We believe many of these lessons apply in the online setting, to large or small companies, or even teams and organizations within a company. A concern we share is the need to evaluate the trustworthiness of experiment results. We believe in the skepticism implied by Twyman's Law: *Any figure that looks interesting or different is usually wrong*; we encourage readers to double-check results and run validity tests, especially for breakthrough positive results. Getting numbers is easy; getting numbers you can trust is hard!

Part I is designed to be read by everyone, regardless of background, and consists of four chapters.

- Chapter 1 is an overview of the benefits of running online controlled experiments and introduces experiment terminology.
- Chapter 2 uses an example to run through the process of running an experiment end-to-end.
- Chapter 3 describes common pitfalls and how to build experimentation trustworthiness, and

- Chapter 4 overviews what it takes to build an experiment platform and scale online experimentation.

Parts II through V can be consumed by everyone as needed but are written with a focus on a specific audience. Part II contains five chapters on fundamentals, such as Organizational Metrics. The topics in Part II are recommended for everyone, especially leaders and executives. Part III contains two chapters that introduce techniques to complement online controlled experiments that leaders, data scientists, engineers, analysts, product managers, and others would find useful for guiding resources and time investment. Part IV focuses on building an experimentation platform and is aimed toward engineers. Finally, Part V digs into advanced analysis topics and is geared toward data scientists.

Our website, https://experimentguide.com, is a companion to this book. It contains additional material, errata, and provides an area for open discussion. The authors intend to donate all proceeds from this book to charity.

Acknowledgments

PART I

Introductory Topics for Everyone

1

Introduction and Motivation

> One accurate measurement is worth more than a thousand expert
> opinions
> – *Admiral Grace Hopper*

In 2012, an employee working on Bing, Microsoft's search engine, suggested changing how ad headlines display (Kohavi and Thomke 2017). The idea was to lengthen the title line of ads by combining it with the text from the first line below the title, as shown in Figure 1.1.

Nobody thought this simple change, among the hundreds suggested, would be the best revenue-generating idea in Bing's history!

The feature was prioritized low and languished in the backlog for more than six months until a software developer decided to try the change, given how easy it was to code. He implemented the idea and began evaluating the idea on real users, randomly showing some of them the new title layout and others the old one. User interactions with the website were recorded, including ad clicks and the revenue generated from them. This is an example of an A/B test, the simplest type of controlled experiment that compares two variants: A and B, or a *Control and a Treatment*.

A few hours after starting the test, a revenue-too-high alert triggered, indicating that something was wrong with the experiment. The Treatment, that is, the new title layout, was generating too much money from ads. Such "too good to be true" alerts are very useful, as they usually indicate a serious bug, such as cases where revenue was logged twice (double billing) or where only ads displayed, and the rest of the web page was broken.

For this experiment, however, the revenue increase was valid. Bing's revenue increased by a whopping 12%, which at the time translated to over $100M annually in the US alone, without significantly hurting key user-experience metrics. The experiment was replicated multiple times over a long period.
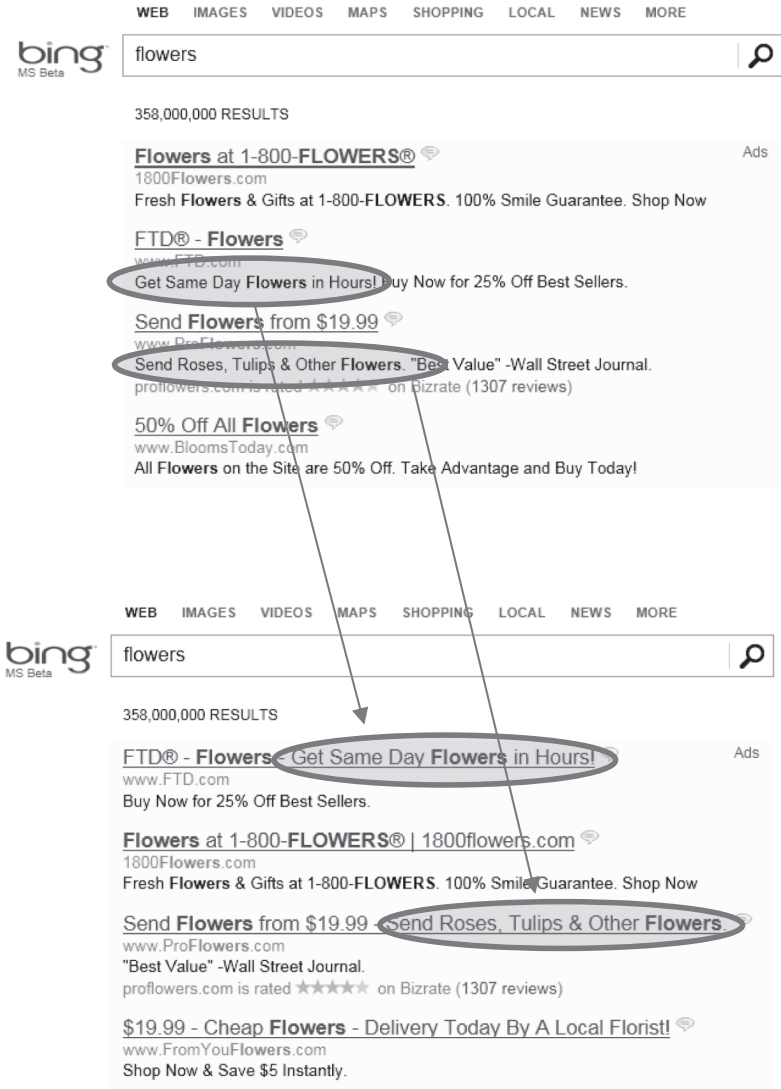
Figure 1.1 An experiment changing the way ads display on Bing

The example typifies several key themes in online controlled experiments:

- It is hard to assess the value of an idea. In this case, a simple change worth over $100M/year was delayed for months.
- Small changes can have a big impact. A $100M/year return-on-investment (ROI) on a few days' work for one engineer is about as extreme as it gets.

- Experiments with big impact are rare. Bing runs over 10,000 experiments a year, but simple features resulting in such a big improvement happen only once every few years.
- The overhead of running an experiment must be small. Bing's engineers had access to ExP, Microsoft's experimentation system, which made it easy to scientifically evaluate the idea.
- The overall evaluation criterion (OEC, described more later in this chapter) must be clear. In this case, revenue was a key component of the OEC, but revenue alone is insufficient as an OEC. It could lead to plastering the web site with ads, which is known to hurt the user experience. Bing uses an OEC that weighs revenue against user-experience metrics, including Sessions per user (are users abandoning or increasing engagement) and several other components. The key point is that user-experience metrics did not significantly degrade even though revenue increased dramatically.

The next section introduces the terminology of controlled experiments.

Online Controlled Experiments Terminology

Controlled experiments have a long and fascinating history, which we share online (Kohavi, Tang and Xu 2019). They are sometimes called A/B tests, A/B/n tests (to emphasize multiple variants), field experiments, randomized controlled experiments, split tests, bucket tests, and flights. In this book, we use the terms *controlled experiments* and *A/B tests* interchangeably, regardless of the number of variants.

Online controlled experiments are used heavily at companies like Airbnb, Amazon, Booking.com, eBay, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yahoo!/Oath, and Yandex (Gupta et al. 2019). These companies run thousands to tens of thousands of experiments every year, sometimes involving millions of users and testing everything, including changes to the user interface (UI), relevance algorithms (search, ads, personalization, recommendations, and so on), latency/performance, content management systems, customer support systems, and more. Experiments are run on multiple channels: websites, desktop applications, mobile applications, and e-mail.

In the most common online controlled experiments, users are randomly split between variants in a persistent manner (a user receives the same variant in multiple visits). In our opening example from Bing, the Control was the original display of ads and the Treatment was the display of ads with longer

**100%
Users**

**50%
Users**

**50%
Users**

Control
Existing
System

Treatment
Existing System
with Feature X

**Users' interactions instrumented,
analyzed & compared**

**Analyze at the end
of the experiment**
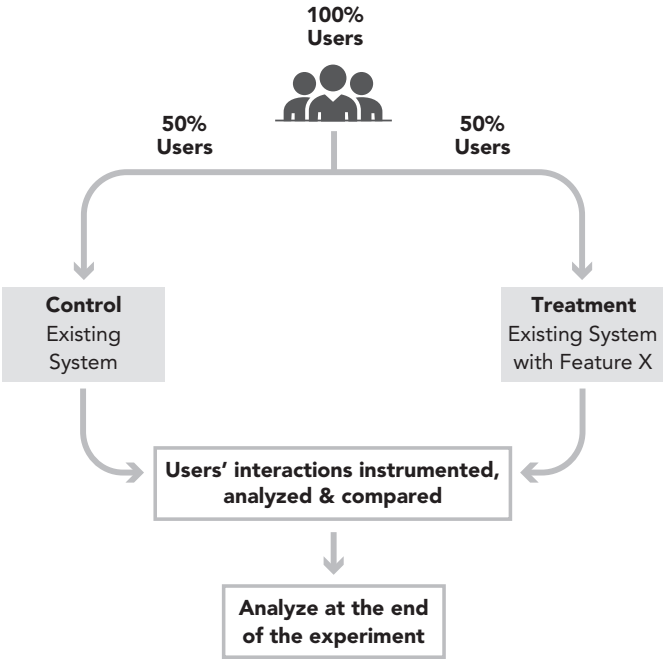
Figure 1.2 A simple controlled experiment: An A/B Test

titles. The users' interactions with the Bing web site were instrumented, that is, monitored and logged. From the logged data, metrics are computed, which allowed us to assess the difference between the variants for each metric.

In the simplest controlled experiments, there are two variants: Control (A) and Treatment (B), as shown in Figure 1.2.

We follow the terminology of Kohavi and Longbottom (2017), and Kohavi, Longbottom et al. (2009) and provide related terms from other fields below. You can find many other resources on experimentation and A/B testing at the end of this chapter under Additional Reading.

Overall Evaluation Criterion (OEC): A quantitative measure of the experiment's objective. For example, your OEC might be active days per user, indicating the number of days during the experiment that users were active (i.e., they visited and took some action). Increasing this OEC implies that users are visiting your site more often, which is a great outcome. The OEC must be measurable in the short term (the duration of an experiment) yet believed to causally drive long-term strategic objectives (see *Strategy, Tactics, and their Relationship to Experiments* later in this chapter and Chapter 7). In the case of a search engine, the OEC can be a combination of usage (e.g., sessions-per-user),

relevance (e.g., successful sessions, time to success), and advertisement revenue (not all search engines use all of these metrics or only these metrics).

In statistics, this is often called the *Response* or *Dependent* variable (Mason, Gunst and Hess 1989, Box, Hunter and Hunter 2005); other synonyms are *Outcome*, *Evaluation* and *Fitness Function* (Quarto-vonTivadar 2006). Experiments can have multiple objectives and analysis can use a balanced scorecard approach (Kaplan and Norton 1996), although selecting a single metric, possibly as a weighted combination of such objectives is highly desired and recommended (Roy 2001, 50, 405−429).

We take a deeper dive into determining the OEC for experiments in Chapter 7.

Parameter: A controllable experimental variable that is thought to influence the OEC or other metrics of interest. Parameters are sometimes called *factors* or *variables*. Parameters are assigned *values*, also called *levels*. In simple A/B tests, there is commonly a single parameter with two values. In the online world, it is common to use univariable designs with multiple values (such as, A/B/C/D). Multivariable tests, also called *Multivariate Tests* (MVTs), evaluate multiple parameters (variables) together, such as font color and font size, allowing experimenters to discover a global optimum when parameters interact (see Chapter 4).

Variant: A user experience being tested, typically by assigning values to parameters. In a simple A/B test, A and B are the two variants, usually called Control and Treatment. In some literature, a variant only means a Treatment; we consider the Control to be a special variant: the existing version on which to run the comparison. For example, in case of a bug discovered in the experiment, you would abort the experiment and ensure that all users are assigned to the Control variant.

Randomization Unit: A pseudo-randomization (e.g., hashing) process is applied to units (e.g., users or pages) to map them to variants. Proper randomization is important to ensure that the populations assigned to the different variants are similar statistically, allowing causal effects to be determined with high probability. You must map units to variants in a persistent and independent manner (i.e., if user is the randomization unit, a user should consistently see the same experience, and the assignment of a user to a variant should not tell you anything about the assignment of a different user to its variant). It is very common, and we highly recommend, to use *users* as a randomization unit when running controlled experiments for online audiences. Some experimental designs choose to randomize by pages, sessions, or user-day (i.e., the experiment remains consistent for the user for each 24-hour window determined by the server). See Chapter 14 for more information.

Proper randomization is critical! If the experimental design assigns an equal percentage of users to each variant, then each user should have an equal chance of being assigned to each variant. Do not take randomization lightly. The examples below demonstrate the challenge and importance of proper randomization.

- The RAND corporation needed random numbers for Monte Carlo methods in the 1940s, so they created a book of a million random digits generated using a pulse machine. However, due to skews in the hardware, the original table was found to have significant biases and the digits had to re-randomized in a new edition of the book (RAND 1955).
- Controlled experiments were initially used in medical domains. The US Veterans Administration (VA) conducted an experiment (drug trial) of streptomycin for tuberculosis, but the trials failed because physicians introduced biases and influenced the selection process (Marks 1997). Similar trials in Great Britain were done with blind protocols and were successful, creating what is now called a watershed moment in controlled trials (Doll 1998).

No factor should be allowed to influence variant assignment. Users (units) cannot be distributed "any old which way" (Weiss 1997). It is important to note that random does not mean "haphazard or unplanned, but a deliberate choice based on probabilities" (Mosteller, Gilbert and McPeek 1983). Senn (2012) discusses some myths of randomization.

Why Experiment? Correlations, Causality, and Trustworthiness

Let's say you're working for a subscription business like Netflix, where X% of users churn (end their subscription) every month. You decide to introduce a new feature and observe that churn rate for users using that feature is X%/2, that is, half. You might be tempted to claim causality; the feature is reducing churn by half. This leads to the conclusion that if we make the feature more discoverable and used more often, subscriptions will soar. Wrong! Given the data, no conclusion can be drawn about whether the feature reduces or increases user churn, and both are possible.

An example demonstrating this fallacy comes from Microsoft Office 365, another subscription business. Office 365 users that see error messages and experience crashes have lower churn rates, but that does not mean that Office 365 should show more error messages or that Microsoft should lower code
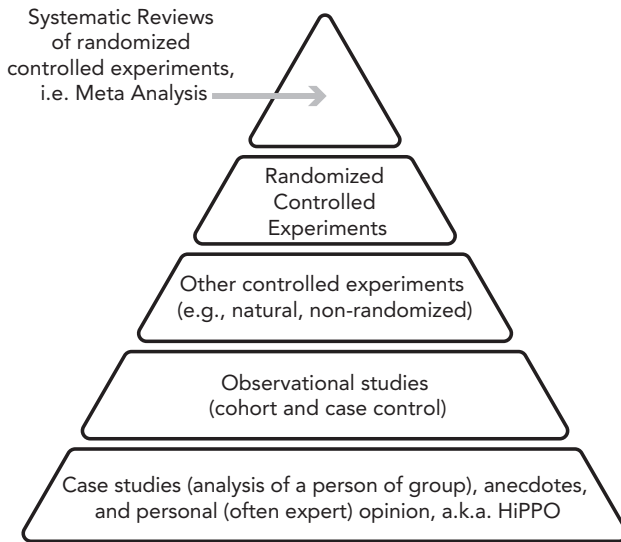
Figure 1.3 A simple hierarchy of evidence for assessing the quality of trial design (Greenhalgh 2014)

quality, causing more crashes. It turns out that all three events are caused by a single factor: usage. Heavy users of the product see more error messages, experience more crashes, and have lower churn rates. Correlation does not imply causality and overly relying on these observations leads to faulty decisions.

In 1995, Guyatt et al. (1995) introduced the hierarchy of evidence as a way to grade recommendations in medical literature, which Greenhalgh expanded on in her discussions on practicing evidence-based medicine (1997, 2014). Figure 1.3 shows a simple hierarchy of evidence, translated to our terminology, based on Bailar (1983, 1). Randomized controlled experiments are the gold standard for establishing causality. Systematic reviews, that is, meta-analysis, of controlled experiments provides more evidence and generalizability.

More complex models, such as the *Levels of Evidence* by the Oxford Centre for Evidence-based Medicine are also available (2009).

The experimentation platforms used by our companies allow experimenters at Google, LinkedIn, and Microsoft to run tens of thousands of online controlled experiments a year with a high degree of trust in the results. We believe online controlled experiments are:

- The best scientific way to establish causality with high probability.
- Able to detect small changes that are harder to detect with other techniques, such as changes over time (sensitivity).

- Able to detect unexpected changes. Often underappreciated, but many experiments uncover surprising impacts on other metrics, be it performance degradation, increased crashes/errors, or cannibalizing clicks from other features.

A key focus of this book is highlighting potential pitfalls in experiments and suggesting methods that improve trust in results. Online controlled experiments provide an unparalleled ability to electronically collect reliable data at scale, randomize well, and avoid or detect pitfalls (see Chapter 11). We recommend using other, less trustworthy, methods, including observational studies, when online controlled experiments are not possible.

Necessary Ingredients for Running Useful Controlled Experiments

Not every decision can be made with the scientific rigor of a controlled experiment. For example, you cannot run a controlled experiment on mergers and acquisitions (M&A), as we cannot have both the merger/acquisition and its counterfactual (no such event) happening concurrently. We now review the necessary technical ingredients for running useful controlled experiments (Kohavi, Crook and Longbotham 2009), followed by organizational tenets. In Chapter 4, we cover the experimentation maturity model.

1. There are experimental units (e.g., users) that can be assigned to different variants with no interference (or little interference); for example, users in Treatment do not impact users in Control (see Chapter 22).
2. There are enough experimental units (e.g., users). For controlled experiments to be useful, we recommend thousands of experimental units: the larger the number, the smaller the effects that can be detected. The good news is that even small software startups typically get enough users quickly and can start to run controlled experiments, initially looking for big effects. As the business grows, it becomes more important to detect smaller changes (e.g., large web sites must be able to detect small changes to key metrics impacting user experience and fractions of a percent change to revenue), and the sensitivity improves with a growing user base.
3. Key metrics, ideally an OEC, are agreed upon and can be practically evaluated. If the goals are too hard to measure, it is important to agree on surrogates (see Chapter 7). Reliable data can be collected, ideally cheaply and broadly. In software, it is usually easy to log system events and user actions (see Chapter 13).

4. Changes are easy to make. Software is typically easier to change than hardware; but even in software, some domains require a certain level of quality assurance. Changes to a recommendation algorithm are easy to make and evaluate; changes to software in airplane flight control systems require a whole different approval process by the Federal Aviation Administration (FAA). Server-side software is much easier to change than client-side (see Chapter 12), which is why calling services from client software is becoming more common, enabling upgrades and changes to the services to be done more quickly and using controlled experiments.

Most non-trivial online services meet, or could meet, the necessary ingredients for running an agile development process based on controlled experiments. Many implementations of software+services could also meet the requirements relatively easily. Thomke wrote that organizations will recognize maximal benefits from experimentation when it is used in conjunction with an "innovation system" (Thomke 2003). Agile software development is such an innovation system.

When controlled experiments are not possible, modeling could be done, and other experimental techniques might be used (see Chapter 10). The key is that if controlled experiments can be run, they provide the most reliable and sensitive mechanism to evaluate changes.

Tenets

There are three key tenets for organizations that wish to run online controlled experiments (Kohavi et al. 2013):

1. The organization wants to make data-driven decisions and has formalized an OEC.
2. The organization is willing to invest in the infrastructure and tests to run controlled experiments and ensure that the results are trustworthy.
3. The organization recognizes that it is poor at assessing the value of ideas.

Tenet 1: The Organization Wants to Make Data-Driven Decisions and Has Formalized an OEC

You will rarely hear someone at the head of an organization say that they don't want to be data-driven (with the notable exception of Apple under Steve Jobs, where Ken Segall claimed that "we didn't test a single ad. Not for print, TV,

billboards, the web, retail, or anything" (Segall 2012, 42). But measuring the incremental benefit to users from new features has cost, and objective measurements typically show that progress is not as rosy as initially envisioned. Many organizations will not spend the resources required to define and measure progress. It is often easier to generate a plan, execute against it, and declare success, with the key metric being: "percent of plan delivered," ignoring whether the feature has any positive impact to key metrics.

To be data-driven, an organization should define an OEC that can be easily measured over relatively short durations (e.g., one to two weeks). Large organizations may have multiple OECs or several key metrics that are shared with refinements for different areas. The hard part is finding metrics measurable in a short period, sensitive enough to show differences, and that are predictive of long-term goals. For example, "Profit" is not a good OEC, as short-term theatrics (e.g., raising prices) can increase short-term profit, but may hurt it in the long run. Customer lifetime value is a strategically powerful OEC (Kohavi, Longbottom et al. 2009). We cannot overemphasize the importance of agreeing on a good OEC that your organization can align behind; see Chapter 6.

The terms "data-informed" or "data-aware" are sometimes used to avoid the implication that a single source of data (e.g., a controlled experiment) "drives" the decisions (King, Churchill and Tan 2017, Knapp et al. 2006). We use data-driven and data-informed as synonyms in this book. Ultimately, a decision should be made with many sources of data, including controlled experiments, surveys, estimates of maintenance costs for the new code, and so on. A data-driven or a data-informed organization gathers relevant data to drive a decision and inform the HiPPO (Highest Paid Person's Opinion) rather than relying on intuition (Kohavi 2019).

Tenet 2: The Organization Is Willing to Invest in the Infrastructure and Tests to Run Controlled Experiments and Ensure That Their Results Are Trustworthy

In the online software domain (websites, mobile, desktop applications, and services) the necessary conditions for controlled experiments can be met through software engineering work (see *Necessary Ingredients for Running Useful Controlled Experiments*): it is possible to reliably randomize users; it is possible to collect telemetry; and it is relatively easy to introduce software changes, such as new features (see Chapter 4). Even relatively small websites have enough users to run the necessary statistical tests (Kohavi, Crook and Longbotham 2009).

Controlled experiments are especially useful in combination with Agile software development (Martin 2008, K. S. Rubin 2012), Customer Development process (Blank 2005), and MVPs (Minimum Viable Products), as popularized by Eric Ries in *The Lean Startup* (Ries 2011).

In other domains, it may be hard or impossible to reliably run controlled experiments. Some interventions required for controlled experiments in medical domains may be unethical or illegal. Hardware devices may have long lead times for manufacturing and modifications are difficult, so controlled experiments with users are rarely run on new hardware devices (e.g., new mobile phones). In these situations, other techniques, such as *Complementary Techniques* (see Chapter 10), may be required when controlled experiments cannot be run.

Assuming you can run controlled experiments, it is important to ensure their trustworthiness. When running online experiments, getting numbers is easy; getting numbers you can trust is hard. Chapter 3 is dedicated to trustworthy results.

Tenet 3: The Organization Recognizes That It Is Poor at Assessing the Value of Ideas

Features are built because teams believe they are useful, yet in many domains most ideas fail to improve key metrics. Only one third of the ideas tested at Microsoft improved the metric(s) they were designed to improve (Kohavi, Crook and Longbotham 2009). Success is even harder to find in well-optimized domains like Bing and Google, whereby some measures' success rate is about 10–20% (Manzi 2012).

Fareed Mosavat, Slack's Director of Product and Lifecycle tweeted that with all of Slack's experience, only about 30% of monetization experiments show positive results; "if you are on an experiment-driven team, get used to, at best, 70% of your work being thrown away. Build your processes accordingly" (Mosavat 2019).

Avinash Kaushik wrote in his Experimentation and Testing primer (Kaushik 2006) that "80% of the time you/we are wrong about what a customer wants." Mike Moran (Moran 2007, 240) wrote that Netflix considers 90% of what they try to be wrong. Regis Hadiaris from Quicken Loans wrote that "in the five years I've been running tests, I'm only about as correct in guessing the results as a major league baseball player is in hitting the ball. That's right – I've been doing this for 5 years, and I can only 'guess' the outcome of a test about 33% of the time!" (Moran 2008). Dan McKinley at Etsy (McKinley 2013) wrote "nearly everything fails" and for features, he wrote "it's been humbling to

realize how rare it is for them to succeed on the first attempt. I strongly suspect that this experience is universal, but it is not universally recognized or acknowledged." Finally, Colin McFarland wrote in the book *Experiment!* (McFarland 2012, 20) "No matter how much you think it's a no-brainer, how much research you've done, or how many competitors are doing it, sometimes, more often than you might think, experiment ideas simply fail."

Not every domain has such poor statistics, but most who have run controlled experiments in customer-facing websites and applications have experienced this humbling reality: *we are poor at assessing the value of ideas*.

Improvements over Time

In practice, improvements to key metrics are achieved by many small changes: 0.1% to 2%. Many experiments only impact a segment of users, so you must dilute the impact of a 5% improvement for 10% of your users, which results in a much smaller impact (e.g., 0.5% if the *triggered* population is similar to the rest of the users); see Chapter 3. As Al Pacino says in the movie *Any Given Sunday*, "…winning is done inch by inch."

Google Ads Example

In 2011, Google launched an improved ad ranking mechanism after over a year of development and incremental experiments (Google 2011). Engineers developed and experimented with new and improved models for measuring the quality score of ads within the existing ad ranking mechanism, as well as with changes to the ad auction itself. They ran hundreds of controlled experiments and multiple iterations; some across all markets, and some long term in specific markets to understand the impact on advertisers in more depth. This large backend change – and running controlled experiments – ultimately validated how planning multiple changes and layering them together improved the user's experience by providing higher quality ads, and improved their advertiser's experience moving towards lower average prices for the higher quality ads.

Bing Relevance Example

The Relevance team at Bing consists of several hundred people tasked with improving a single OEC metric by 2% every year. The 2% is the sum of the Treatment effects (i.e., the delta of the OEC) in all controlled experiments that

shipped to users over the year, assuming they are additive. Because the team runs thousands of experiment Treatments, and some may appear positive by chance (Lee and Shen 2018), credit towards the 2% is assigned based on a replication experiment: once the implementation of an idea is successful, possibly after multiple iterations and refinements, a *certification* experiment is run with a single Treatment. The Treatment effect of this certification experiment determines the credit towards the 2% goal. Recent development suggests shrinking the Treatment effect to improve precision (Coey and Cunningham 2019).

Bing Ads Example

The Ads team at Bing has consistently grown revenue 15−25% per year (eMarketer 2016), but most improvements were done inch-by-inch. Every month a "package" was shipped, the results of many experiments, as shown in Figure 1.4. Most improvements were small, some monthly packages were even known to be negative, as a result of space constraints or legal requirements.
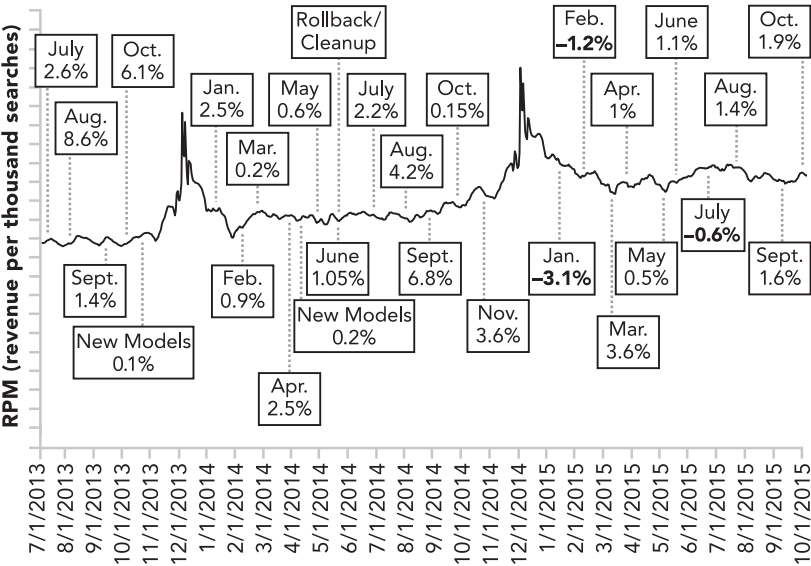
(*) Numbers have been perturbed for obvious reasons

Figure 1.4 Bing Ad Revenue over Time (y-axis represents about 20% growth/ year). The specific numbers are not important

It is informative to see the seasonality spikes around December when purchase intent by users rises dramatically, so ad space is increased, and revenue per thousand searches increases.

Examples of Interesting Online Controlled Experiments

Interesting experiments are ones where the absolute difference between the expected outcome and the actual result is large. If you thought something was going to happen and it happened, then you haven't learned much. If you thought something was going to happen and it didn't, then you've learned something important. And if you thought something minor was going to happen, and the results are a major surprise and lead to a breakthrough, you've learned something highly valuable.

The Bing example at the beginning of this chapter and those in this section are uncommon successes with surprising, highly positive, results. Bing's attempt to integrate with social networks, such as Facebook and Twitter, are an example of expecting a strong result and not seeing it – the effort was abandoned after many experiments showed no value for two years.

While sustained progress is a matter of continued experimentation and many small improvements, as shown in the section *Bing Ads Example*, here are several examples highlighting large surprising effects that stress how poorly we assess the value of ideas.

UI Example: 41 Shades of Blue

Small design decisions can have significant impact, as both Google and Microsoft have consistently shown. Google tested 41 gradations of blue on Google search results pages (Holson 2009), frustrating the visual design lead at the time. However, Google's tweaks to the color scheme ended up being substantially positive on user engagement (note that Google does not report on the results of individual changes) and led to a strong partnership between design and experimentation moving forward. Microsoft's Bing color tweaks similarly showed that users were more successful at completing tasks, their time-to-success improved, and monetization improved to the tune of over $10 M annually in the United States (Kohavi et al. 2014, Kohavi and Thomke 2017).

While these are great examples of tiny changes causing massive impact, given that a wide sweep of colors was done, it is unlikely that playing around with colors in additional experiments will yield more significant improvements.

You could save $30 today with the Amazon Visa® Card:

Your current subtotal: $32.20
Amazon Visa discount: - $30.00   ▶ Find out how
Your new subtotal: $2.20

Save $30 off your first purchase, earn **3% rewards**, get a **0% APR***, and pay **no annual fee**.

Figure 1.5 Amazon's credit card offer with savings on cart total

Making an Offer at the Right Time

In 2004, Amazon placed a credit-card offer on the home page. It was highly profitable but had a very low click-through rate (CTR). The team ran an experiment to move the offer to the shopping cart page that the user sees after adding an item, showing simple math highlighting the savings the user would receive, as shown in Figure 1.5 (Kohavi et al. 2014).

Since users adding an item to the shopping cart have clear purchase intent, this offer displays at the right time. The controlled experiment demonstrated that this simple change increased Amazon's annual profit by tens of millions of dollars.

Personalized Recommendations

Greg Linden at Amazon created a prototype to display personalized recommendations based on items in the user's shopping cart (Linden 2006, Kohavi, Longbottom et al. 2009). When you add an item, recommendations come up; add another item, new recommendations show up. Linden notes that while the prototype looked promising, "a marketing senior vice-president was dead set against it," claiming it would distract people from checking out. Greg was "forbidden to work on this any further." Nonetheless, he ran a controlled experiment, and the "feature won by such a wide margin that not having it live was costing Amazon a noticeable chunk of change. With new urgency, shopping cart recommendations launched." Now multiple sites use cart recommendations.

Speed Matters a LOT

In 2012, an engineer at Microsoft's Bing made a change to the way JavaScript was generated, which shortened the HTML sent to clients significantly, resulting in improved performance. The controlled experiment showed a surprising number of improved metrics. They conducted a follow-on

experiment to estimate the impact on server performance. The result showed that performance improvements also significantly improve key user metrics, such as success rate and time-to-success, and each 10 millisecond performance improvement (1/30th of the speed of an eye blink) pays for the fully loaded annual cost of an engineer (Kohavi et al. 2013).

By 2015, as Bing's performance improved, there were questions about whether there was still value to performance improvements when the server was returning results in under a second at the 95th percentile (i.e., for 95% of the queries). The team at Bing conducted a follow-on study and key user metrics still improve significantly. While the relative impact on revenue was somewhat reduced, Bing's revenue improved so much during the time that each millisecond in improved performance was worth more than in the past; every four milliseconds of improvement funded an engineer for a year! See Chapter 5 for in-depth review of this experiment and the criticality of performance.

Performance experiments were done at multiple companies with results indicating how critical performance is. At Amazon, a 100-millisecond slow-down experiment decreased sales by 1% (Linden 2006b, 10). A joint talk by speakers from Bing and Google (Schurman and Brutlag 2009) showed the significant impact of performance on key metrics, including distinct queries, revenue, clicks, satisfaction, and time-to-click.

Malware Reduction

Ads are a lucrative business and "freeware" installed by users often contains malware that pollutes pages with ads. Figure 1.6 shows what a resulting page from Bing looked like to a user with malware. Note that multiple ads (high-lighted in red) were added to the page (Kohavi et al. 2014).

Not only were Bing ads removed, depriving Microsoft of revenue, but low-quality ads and often irrelevant ads displayed, providing a poor user experience for users who might not have realized why they were seeing so many ads.

Microsoft ran a controlled experiment with 3.8 million users potentially impacted, where basic routines that modify the DOM (Document Object Model) were overridden to allow only limited modifications from trusted sources (Kohavi et al. 2014). The results showed improvements to all of Bing's key metrics, including Sessions per user, indicating that users visited more often or churned less. In addition, users were more successful in their searches, quicker to click on useful links, and annual revenue improved by several million dollars. Also, page-load time, a key performance metric we previously discussed, improved by hundreds of milliseconds for the impacted pages.
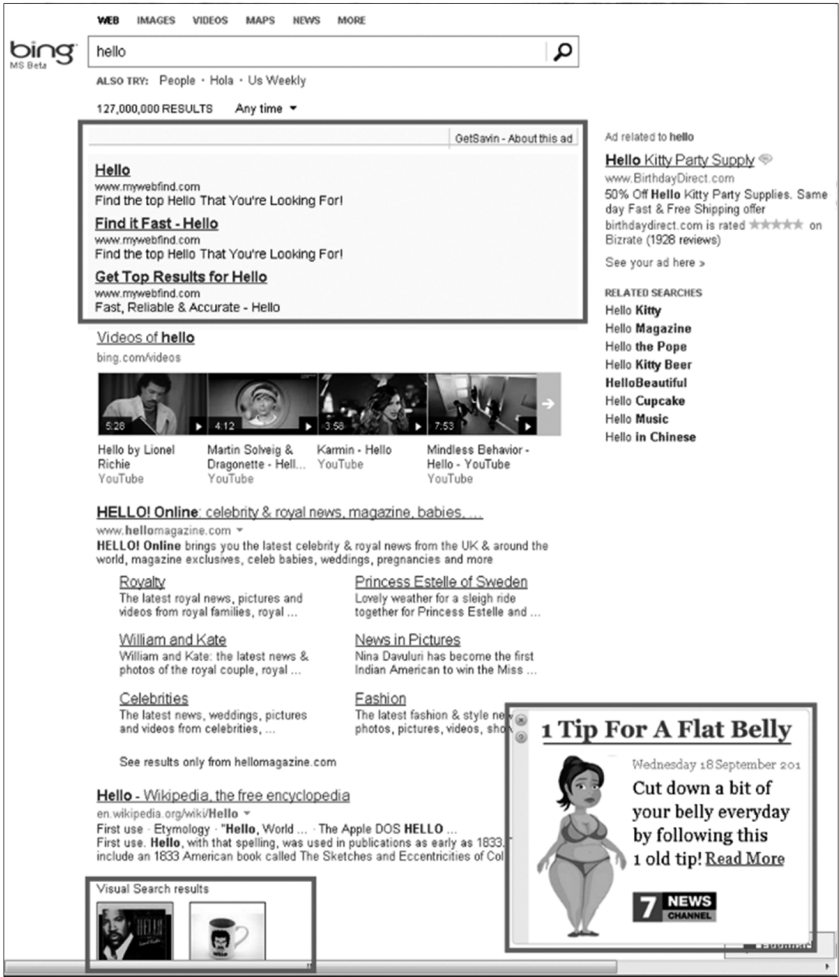
Figure 1.6 Bing page when the user has malware shows multiple ads

Backend Changes

Backend algorithmic changes are often overlooked as an area to use controlled experiments (Kohavi, Longbottom et al. 2009), but it can yield significant results. We can see this both from how teams at Google, LinkedIn, and Microsoft work on many incremental small changes, as we described above, and in this example involving Amazon.

Back in 2004, there already existed a good algorithm for making recommendations based on two sets. The signature feature for Amazon's
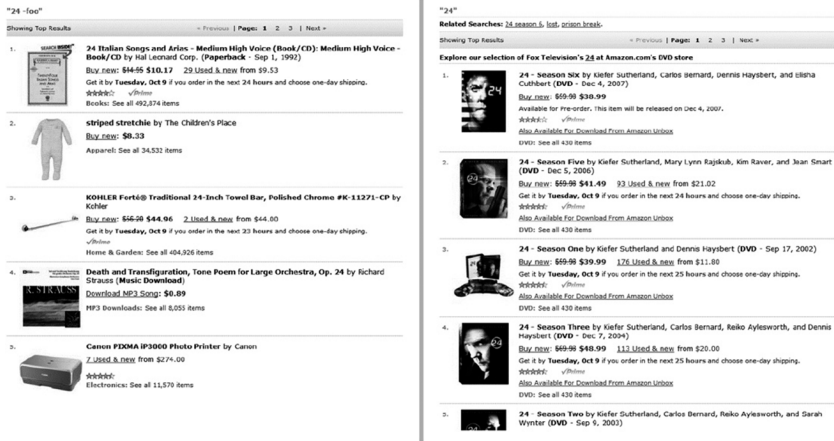
Figure 1.7 Amazon search for "24" with and without BBS

recommendation was "People who bought item X bought item Y," but this was generalized to "People who *viewed* item X bought item Y" and "People who viewed item X *viewed* item Y." A proposal was made to use the same algorithm for "People who *searched* for X bought item Y." Proponents of the algorithm gave examples of underspecified searches, such as "24," which most people associate with the TV show starring Kiefer Sutherland. Amazon's search was returning poor results (left in Figure 1.7), such as CDs with 24 Italian Songs, clothing for 24-month old toddlers, a 24-inch towel bar, and so on. The new algorithm gave top-notch results (right in Figure 1.7), returning DVDs for the show and related books, based on what items people actually purchased after searching for "24." One weakness of the algorithm was that some items surfaced that did not contain the words in the search phrase; however, Amazon ran a controlled experiment, and despite this weakness, this change increased Amazon's overall revenue by 3% – hundreds of millions of dollars.

Strategy, Tactics, and Their Relationship to Experiments

When the necessary ingredients for running online controlled experiments are met, we strongly believe they should be run to inform organizational decisions at all levels from strategy to tactics.

Strategy (Porter 1996, 1998) and controlled experiments are synergistic. David Collis of Lean Strategy wrote that "rather than suppressing

entrepreneurial behavior, effective strategy encourages it – by identifying the bounds within which innovation and experimentation should take place" (Collis 2016). He defines a *lean strategy process*, which guards against the extremes of both rigid planning and unrestrained experimentation.

Well-run experiments with appropriate metrics complement business strategy, product design, and improve operational effectiveness by making the organization more data driven. By encapsulating strategy into an OEC, controlled experiments can provide a great feedback loop for the strategy. Are the ideas evaluated with experiments improving the OEC sufficiently? Alternatively, surprising results from experiments can shine a light on alternative strategic opportunities, leading to pivots in those directions (Ries 2011). Product design decisions are important for coherency and trying multiple design variants provides a useful feedback loop to the designers. Finally, many tactical changes can improve the operational effectiveness, defined by Porter as "performing similar activities better than rivals perform them" (Porter 1996).

We now review two key scenarios.

Scenario 1: You Have a Business Strategy and You Have a Product with Enough Users to Experiment

In this scenario, experiments can help hill-climb to a local optimum based on your current strategy and product:

- Experiments can help identify areas with high ROI: those that improve the OEC the most, relative to the effort. Trying different areas with MVPs can help explore a broader set of areas more quickly, before committing significant resources.
- Experiments can also help with optimizations that may not be obvious to designers but can make a large difference (e.g., color, spacing, performance).
- Experiments can help continuously iterate to better site redesigns, rather than having teams work on complete site redesigns that subject users to primacy effects (users are *primed* in the old feature, i.e., used to the way it works) and commonly fail not only to achieve their goals, but even fail to achieve parity with the old site on key metrics (Goward 2015, slides 22−24, Rawat 2018, Wolf 2018, Laja 2019).
- Experiments can be critical in optimizing backend algorithms and infrastructure, such as recommendation and ranking algorithms.

Having a strategy is critical for running experiments: the strategy is what drives the choice of OEC. Once defined, controlled experiments help

accelerate innovation by empowering teams to optimize and improve the OEC. Where we have seen experiments misused is when the OEC is not properly chosen. The metrics chosen should meet key characteristics and not be gameable (see Chapter 7).

At our companies, not only do we have teams focused on how to run experiments properly, but we also have teams focused on metrics: choosing metrics, validating metrics, and evolving metrics over time. Metric evolution will happen both due to your strategy evolving over time but also as you learn more about the limitations of your existing metrics, such as CTR being too gameable and needing to evolve. Metric teams also work on determining which metrics measurable in the short term drive long-term objectives, since experiments usually run over a shorter time frame. Hauser and Katz (1998) wrote that "the firm must identify metrics that the team can affect today, but which, ultimately, will affect the firm's long-term goals" (see Chapter 7).

Tying the strategy to the OEC also creates *Strategic Integrity* (Sinofsky and Iansiti 2009). The authors point out that "Strategic integrity is not about crafting brilliant strategy or about having the perfect organization: It is about getting the right strategies done by an organization that is aligned and knows how to get them done. It is about matching top-down-directed perspectives with bottom-up tasks." The OEC is the perfect mechanism to make the strategy explicit and to align what features ship with the strategy.

Ultimately, without a good OEC, you are wasting resources – think of experimenting to improve the food or lighting on a sinking cruise ship. The weight of passenger safety term in the OEC for those experiments should be extremely high – in fact, so high that we are not willing to degrade safety. This can be captured either via high weight in the OEC, or, equivalently, using passenger safety as a guardrail metric (see Chapter 21). In software, the analogy to the cruise ship passenger safety is software crashes: if a feature is increasing crashes for the product, the experience is considered so bad, other factors pale in comparison.

Defining guardrail metrics for experiments is important for identifying what the organization is *not* willing to change, since a strategy also "requires you to make tradeoffs in competing – to choose what not to do" (Porter 1996). The ill-fated Eastern Air Lines flight 401 crashed because the crew was focused on a burned-out landing gear indicator light, and failed to notice that the autopilot was accidentally disengaged; altitude, a key guardrail metric, gradually decreased and the plane crashed in the Florida Everglades in 1972, resulting in 101 fatalities (Wikipedia contributors, Eastern Air Lines Flight 401 2019).

Improvements in operational efficiencies can provide long-term differentiated advantage, as Porter noted in a section titled "Japanese Companies Rarely have Strategies" (1996) and Varian noted in his article on Kaizen (2007).

Scenario 2: You Have a Product, You Have a Strategy, but the Results Suggest That You Need to Consider a Pivot

In Scenario 1, controlled experiments are a great tool for hill climbing. If you think of the multi-dimensional space of ideas, with the OEC as the "height" that is being optimized, then you may be making steps towards a peak. But sometimes, either based on internal data about the rate of change or external data about growth rates or other benchmarks, you need to consider a pivot: jumping to a different location in the space, which may be on a bigger hill, or changing the strategy and the OEC (and hence the shape of the terrain).

In general, we recommend always having a portfolio of ideas: most should be investments in attempting to optimize "near" the current location, but a few radical ideas should be tried to see whether those jumps lead to a bigger hill. Our experience is that most big jumps fail (e.g., big site redesigns), yet there is a risk/reward tradeoff: the rare successes may lead to large rewards that compensate for many failures.

When testing radical ideas, how you run and evaluate experiments changes somewhat. Specifically, you need to consider:

- The duration of experiments. For example, when testing a major UI redesign, experimental changes measured in the short term may be influenced by primacy effects or change aversion. The direct comparison of Treatment to Control may not measure the true long-term effect. In a two-sided marketplace, testing a change, unless sufficiently large, may not induce an effect on the marketplace. A good analogy is an ice cube in a very cold room: small increases to room temperature may not be noticeable, but once you go over the melting point (e.g., 32 Fahrenheit), the ice cube melts. Longer and larger experiments, or alternative designs, such as the country-level experiments used in the Google Ads Quality example above, may be necessary in these scenarios (see also Chapter 23).
- The number of ideas tested. You may need many different experiments because each experiment is only testing a specific tactic, which is a component of the overall strategy. A single experiment failing to improve the OEC may be due to the specific tactic being poor, not necessarily indicating that the overall strategy is bad. Experiments, by design, are testing specific hypotheses, while strategies are broader. That said, controlled experiments help refine the strategy, or show its ineffectiveness and encourage a pivot (Ries 2011). If many tactics evaluated through controlled experiments fail, it may be time to think about Winston Churchill's saying: "However beautiful the strategy, you should occasionally look at the results." For about two years, Bing had a strategy of integrating with social media, particularly

Facebook and Twitter, opening a third pane with social search results. After spending over $25 million on the strategy with no significant impact to key metrics, the strategy was abandoned (Kohavi and Thomke 2017). It may be hard to give up on a big bet, but economic theory tells us that failed bets are sunk costs, and we should make a forward-looking decision based on the available data, which is gathered as we run more experiments.

Eric Ries uses the term "achieved failure" for companies that successfully, faithfully, and rigorously execute a plan that turned out to have been utterly flawed (Ries 2011). Instead, he recommends:

> The Lean Startup methodology reconceives a startup's efforts as experiments that test its strategy to see which parts are brilliant and which are crazy. A true experiment follows the scientific method. It begins with a clear hypothesis that makes predictions about what is supposed to happen. It then tests those predictions empirically.

Due to the time and challenge of running experiments to evaluate strategy, some, like Sinofsky and Iansiti (2009) write:

> ... product development process as one fraught with risk and uncertainty. These are two very different concepts ... We cannot reduce the uncertainty – you don't know what you don't know.

We disagree: the ability to run controlled experiments allows you to significantly reduce uncertainty by trying a Minimum Viable Product (Ries 2011), getting data, and iterating. That said, not everyone may have a few years to invest in testing a new strategy, in which case you may need to make decisions in the face of uncertainty.

One useful concept to keep in mind is EVI: Expected Value of Information from Douglas Hubbard (2014), which captures how additional information can help you in decision making. The ability to run controlled experiments allows you to significantly reduce uncertainty by trying a Minimum Viable Product (Ries 2011), gathering data, and iterating.

Additional Reading

There are several books directly related to online experiments and A/B tests (Siroker and Koomen 2013, Goward 2012, Schrage 2014, McFarland 2012, King et al. 2017). Most have great motivational stories but are inaccurate on the statistics. Georgi Georgiev's recent book includes comprehensive statistical explanations (Georgiev 2019).

The literature related to controlled experiments is vast (Mason et al. 1989, Box et al. 2005, Keppel, Saufley and Tokunaga 1992, Rossi, Lipsey and Freeman 2004, Imbens and Rubin 2015, Pearl 2009, Angrist and Pischke 2014, Gerber and Green 2012).

There are several primers on running controlled experiments on the web (Peterson 2004, 76−78, Eisenberg 2005, 283−286, Chatham, Temkin and Amato 2004, Eisenberg 2005, Eisenberg 2004); (Peterson 2005, 248−253, Tyler and Ledford 2006, 213−219, Sterne 2002, 116−119, Kaushik 2006).

A multi-armed bandit is a type of experiment where the experiment traffic allocation can be dynamically updated as the experiment progresses (Li et al. 2010, Scott 2010). For example, we can take a fresh look at the experiment every hour to see how each of the variants has performed, and we can adjust the fraction of traffic that each variant receives. A variant that appears to be doing well gets more traffic, and a variant that is underperforming gets less.

Experiments based on multi-armed bandits are usually more efficient than "classical" A/B experiments, because they gradually move traffic towards winning variants, instead of waiting for the end of an experiment. While there is a broad range of problems they are suitable for tackling (Bakshy, Balandal and Kashin 2019), some major limitations are that the evaluation objective needs to be a single OEC (e.g., tradeoff among multiple metrics can be simply formulated), and that the OEC can be measured reasonably well between re-allocations, for example, click-through rate vs. sessions. There can also be potential bias created by taking users exposed to a bad variant and distributing them unequally to other winning variants.

In December 2018, the three co-authors of this book organized the First Practical Online Controlled Experiments Summit. Thirteen organizations, including Airbnb, Amazon, Booking.com, Facebook, Google, LinkedIn, Lyft, Microsoft, Netflix, Twitter, Uber, Yandex, and Stanford University, sent a total of 34 experts, which presented an overview and challenges from breakout sessions (Gupta et al. 2019). Readers interested in challenges will benefit from reading that paper.