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.

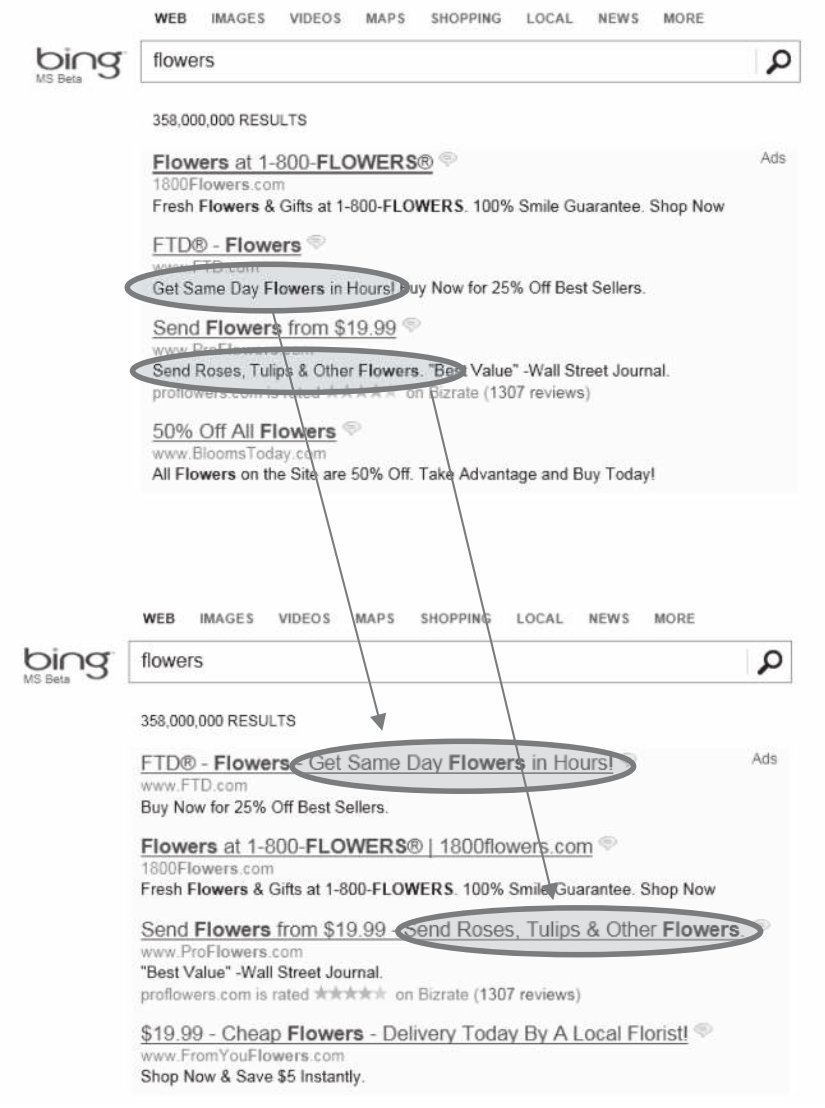

Figure 1.1 An experiment changing the way ads display on Bing

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

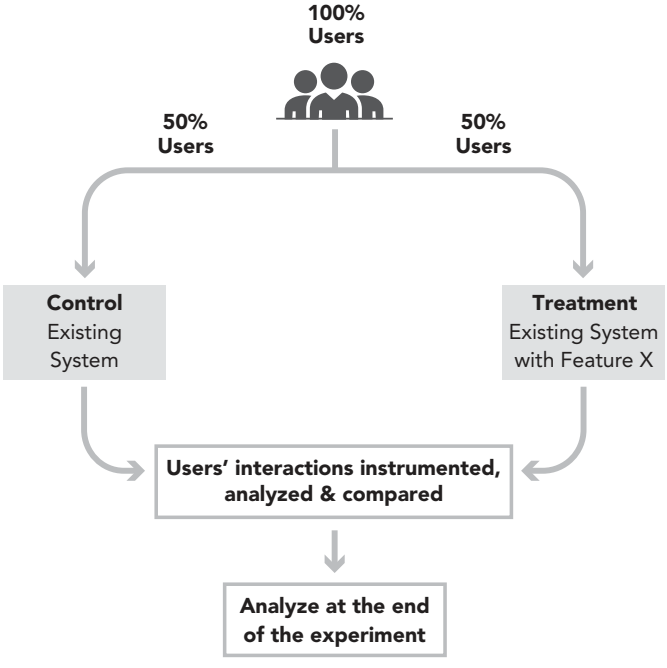

Figure 1.2 A simple controlled experiment: An A/B Test

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

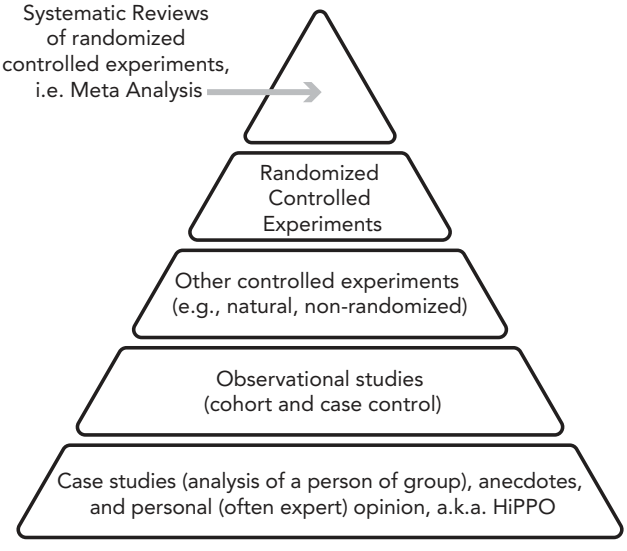

Figure 1.3 A simple hierarchy of evidence for assessing the quality of trial design (Greenhalgh 2014)

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,