O'REILLY®

Analytical Skills for AI & Data Science

Building Skills for an AI-Driven Enterprise

Daniel Vaughan

CHAPTER 2

Intro to Analytical Thinking

In the last chapter, I defined *analytical thinking* as the ability to translate business problems into prescriptive solutions. There is a lot to unpack from this definition, and this will be our task in this chapter.

To really understand the power of prescriptive solutions, I will start by precisely defining each of the three stages present in any analysis of business decisions: these are the descriptive, predictive, and prescriptive steps we have already mentioned in Chapter 1.

Since one crucial skill in our analytical toolbox will be formulating the right business questions from the outset, I will provide an initial glimpse into this topic. Spoiler alert: we only care about business questions that entail business decisions. We will then dissect decisions into levers, consequences, and business results. The link between levers and consequences is intermediated by *causation*, so I will spend quite a bit of time talking about this topic. Finally, I will talk about the role that uncertainty plays in business decisions. Each of these topics is tied to one skill that will be developed throughout the book.

> **What Is a Lever?**
>
> In the context of this book, "levers" are synonymous with "actions" or "decisions," so whenever we say that "we want to pull some lever to obtain a business outcome," this means that we are looking for suitable actions or decisions.

Descriptive, Predictive, and Prescriptive Questions

In Chapter 1, we saw that data maturity models usually depict a nice, smooth road that starts at the descriptive stage, goes through the predictive plateau, and finally ascends to the predictive summit. But why is this the case? Let's start by understanding what these mean, and then we can discuss why commentators and practitioners alike believe that this is the natural ascension of the data evolution.

In a nutshell, *descriptive* relates to how things are, *predictive* to how we believe things will be, and *prescriptive* to how things ought to be. Take Tyrion Lannister's quote in the *Game of Thrones* "The Dance of Dragons" episode: "It's easy to confuse what is with what *ought* to be, especially when *what is* has worked out in your favor" (my emphasis). Tyrion seems to be claiming that we have the tendency to confuse the descriptive and prescriptive when things turn out well, in what may well be a form of confirmation bias. Incidentally, when the outcome is negative, our tendency is to think that this was the worst possible result and attribute our fate to some version of Murphy's Law.

In any case, as this discussion shows, the prescriptive stage is a place where we can rank different options so that words like "best" or "worst" make any sense at all. It follows that the prescriptive layer can never be inferior to the descriptive one, as in the former we can always make the best decision.

But what about prediction? To start, its intermediate ranking is at least problematic, since description relates to the current state and prescription to the *quality of decisions*, and prediction is an input to make decisions, which may or may not be optimal or even good. The implicit assumption in all maturity models is that the quality of decisions can be improved when we have better predictions about the underlying uncertainty in the problem; that good predictions allow us to plan ahead and move proactively, instead of reacting to the past with little or no room to maneuver. That said, this really is an assumption as there's nothing inherent about prediction that makes it improve the outcomes for our businesses.

When Predictive Analysis Is Powerful: The Case of Cancer Detection

Let's take an example where better prediction can make a huge difference: cancer detection. Oncologists usually use some type of visual aid such as X-rays or the more advanced CT scans for early detection of different pathologies. In the case of lung cancer, an X-ray or a CT scan is a description of the patient's current health status. Unfortunately, visual inspection is ineffective unless the disease has already reached a late stage, so description here, by itself, may not provide enough time for a proactive reaction. AI has shown remarkable prowess in predicting the existence of lung cancer from inspecting CT scans, by identifying spots that will eventually turn out to be malignant. But prediction can only take us so far. A doctor should then recommend the right course of action for the patient to fully recover. AI provides the predictive muscle, but humans prescribe the treatment.

Descriptive Analysis: The Case of Customer Churn

Let's run a somewhat typical descriptive analysis of a use case that most companies have dealt with: customer churn or attrition. We will see that without guidance from our business objectives, this type of analysis might take us to a dead end.

What Is Customer Churn?

In case you don't recognize the term, customer churn is the rate at which customers stop using a company's product or service per period of time. For instance, if your company's monthly rate of churn is 5%, this means that 5 out of 100 customers that were purchasing from you at the beginning of the period are no longer doing so when the month ends. As you might imagine, the exact definition varies from industry to industry and depends crucially on the expected frequency of purchases (think about a credit card).

The main reason we care about churn is that our customer acquisition costs are generally substantially larger than the corresponding retention costs, so having a proactive churn control strategy has become an objective in and of itself.

Describing churn

Suppose that your boss wants to get churn under control. As a first step, she may ask you to diagnose the magnitude of the problem. After wrangling with the data, you come up with the following two plots (Figure 2-1). The left plot shows a time series of daily churn rates. Confidently, you state two things: after having a relatively stable beginning of the year, churn is now on the rise. Second, there is a clear seasonal pattern, with weekends having lower than average churn. In the right panel you show that municipalities with higher average incomes also have higher churn rates, which of course is a cause for concern since your most valuable customers may be switching to other companies.

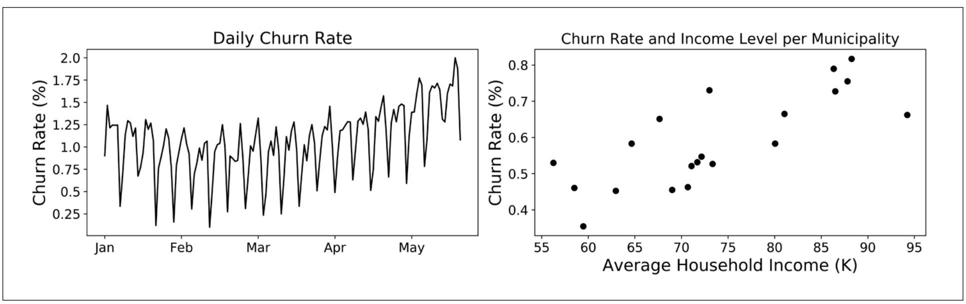

Figure 2-1. Descriptive analysis of our company's churn rate

This is a great example of what can be achieved with descriptive analysis, thanks to our remarkable ability to recognize patterns in the data. Here we quickly identified a change in the trend (churn is accelerating), the existence of strong seasonal effects, and a positive correlation between churn rates and average household income in the scatterplot.

But this also highlights some of its shortcomings. First, as you've probably heard, *correlation does not imply causation*, a topic that will be discussed at length later in this chapter. Related to this, successful root cause analysis requires our ability to create theories about cause and effect. Without these theories we cannot aim at providing alternative courses of action to improve our company's situation. Inspecting data without advancing some plausible explanations is the perfect recipe for making your analytics and data science teams waste valuable time.

The Trap of Finding Actionable Insights

One common catchphrase among consultants and vendors of big data solutions is that once they are given enough data, your data analysts and data scientists will be able to find *actionable insights*.

This is a common trap among business people and novice data practitioners: the idea that given some data, if we inspect it long enough, these actionable insights will emerge, almost magically. I've seen teams spend weeks waiting for the actionable insights to appear, without any luck.

Experienced practitioners reverse engineer the problem: start with the question, formulate hypotheses, and use your descriptive analysis to find evidence against or in favor of these hypotheses. Note the difference: under this approach we actively search for actionable insights by first deciding where to look for them, as opposed to waiting for them to emerge from chaos.

Predicting churn

As a next step, your boss may ask you to *predict* churn in the future. How should you proceed? It really depends on what you want to achieve with this analysis. If you work in finance, for example, and you're interested in forecasting the income statement for the next quarter, you'd be happy to predict aggregate churn rates into the future. If you are in the marketing department, however, you may want to predict which customers are at risk of leaving the company, possibly because you may try using different retention campaigns.

Prescribing courses of action to reduce churn

Finally, suppose that your boss asks you to recommend alternative courses of action to *reduce* the rate of customer churn. This is where the prescriptive toolkit becomes quite handy and where the impact of making good decisions can be most appreciated. You may then pose a cost-benefit analysis for customer retention and come up with a rule that maximizes your customer lifetime value (CLV).

Customer Lifetime Value (CLV)

How should we value our customers? One approach is to assign the current value derived from each one of them. The problem with this short-term view is that companies invest in their customers all the time, from acquisition to retention, marketing, etc., so to value those investments we also need the long-run view from the revenues side.

Several decades ago, people started looking at customers as assets, and under this approach, the right metric is the *stream* of profits derived from them. One difficulty with the stream approach is that at any time our customers may decide to change companies, so we need to incorporate an uncertain time window into the analysis.

The CLV measures the discounted present value of all profits obtained from a relationship with one customer along their expected duration with the company.

For instance, assuming a monthly discount rate of 1%, a new customer who will keep purchasing our goods and services for the next 11 months, leaving a monthly profit of 1 dollar, will have a CLV of $\$1 + \$1/(1.01) + \$1/(1.01)^2 + \cdots + \$1/(1.01)^{10} = 10.4$ dollars. In practice, to compute the CLV we need an estimate of the expected duration of a customer's relationship with us, as well as an estimate of how profits change over time.

We will have the opportunity to go into greater detail on this use case, but let me just single out two characteristics of any prescriptive analysis: as opposed to the two previous analyses, here we actively recommend courses of action that can improve our position, by way of incentivizing a likely-to-leave customer to stay longer with us. Second, prediction is used as an input in the decision-making process, helping us calculate *expected* savings and costs. AI will help us better estimate these quantities, which is necessary for our proposed decision rule. But it is this decision rule that creates value, not prediction itself.

One of the objectives of this book is to prepare us to translate business questions into prescriptive solutions, so don't worry if it's not obvious yet. We will have time to go through many step-by-step examples.

Business Questions and KPIs

One foundational idea in the book is that value is derived from *making decisions*. As such, prediction in the form of machine learning is just an input to create value.

In this book, whenever we talk about business questions, we will always have in mind business decisions. Surely, there are business questions that are purely informative and no actions are involved. But since our aim is to systematically create value, we will only consider actionable questions. As a matter of fact, one byproduct of this book is that we will learn to look for actionable insights in an almost automatic fashion.

It suggests the question, then, of *why* we have to make a decision. Only by answering this question will we be able to know how to measure the appropriateness or not of the choices we make. Decisions that cannot be judged in the face of any relevant evidence are to be discarded. As such, we will have to learn how to select the right metrics to track our performance. Many data science projects and business decisions fail not because of the logic used but because the metrics were not right for the problem.

There is a whole literature on how to select the right key performance indicators (KPIs), and I believe I have little to add on this topic. The two main characteristics I look for are *relevance* and *measurability*. A KPI is relevant when it allows us to clearly assess the results from our decisions *with respect* to the business objective. Notice that this doesn't have to do with how pertinent the business question is, but rather, with whether we are able to evaluate if the decision worked or not, and by how much. It follows that a good KPI should be measurable, and this should be with little or no delay with respect to the time when the decision was made. Not only is there an opportunity cost of delayed measurement, but it may also be harder to identify the root cause.

KPIs to Measure the Success of a Loyalty Program

Let's briefly discuss one example. Suppose that our chief marketing officer asks us to evaluate the creation of a loyalty program for the company. Since the question starts with an action (i.e., to create the loyalty program or not), it immediately registers for us as a business problem. What metrics should we track? To answer this let's start the sequence of *why* questions.

> **The Sequence of Why Questions**
>
> The following example showcases a technique that I call *the sequence of* why *questions*. It is used to identify the business metric that we want to optimize.
>
> It works by starting with what you, your boss, or your colleagues may think you want to achieve and questions the reasons for focusing on this objective. Move one step above and repeat. It terminates when you're satisfied with the answer. Just in passing, recall that to be satisfied you must have a relevant and measurable KPI to quantify the business outcome you will focus on.

Our *why* questions, then, are as follows:

- Create a loyalty program. *Why?*
- Because you want to reward loyal customers. *Why?*
- Because you want to incentivize customers to stay longer with the company. *Why?*
- Because you want to increase your revenues in the longer term. *Why?*

And of course, the list can go on. The important thing is that the final answer to these questions will usually let you clearly identify what KPI is relevant for the problem at hand, and any intermediate metrics that may provide useful; if it's also measurable, then you have found the right metric for your problem.

Consider the second question, for example. Why would anyone want to reward loyal customers? They are already loyal, without the need for any extrinsic motivation, so this strategy may even backfire. But putting aside the underlying reasoning, why is loyalty meaningful and how would you go about measuring the impact of the reward? I argue that loyalty by itself is not meaningful: we prefer loyal customers to not-so-loyal customers because they represent a more stable stream of revenues in the future. If you're not convinced, think about those loyal but *unprofitable* customers. Do you still rank their loyalty as high as before? If loyalty per se is not what you're pursuing, then you should keep going down the sequence of *why* questions.

Just for the sake of the discussion, suppose that you still want to reward loyal customers. How do we measure if the program worked, or put differently, what is a good KPI for this? One commonly used method is to directly ask our customers, as done with the Net Promoter Score (NPS). To calculate the NPS we first ask our customers how likely they are to recommend us as a company on a scale from 0 to 10. We then classify them into *Promoters* (9 to 10), *Detractors* (0 to 6), and *Passive* (7, 8). Individual answers are finally aggregated into the NPS by subtracting the percentage of detractors from the percentage of promoters.

On the bright side, this is a pretty *direct* assessment: we just go and ask our customers if they value the reward. It can't get more straightforward than that. The problem here is that humans act on motivations, so we generally can't tell if the answer is truthful, or if there is some underlying motive and they're trying to game our system. This type of strategic consideration matters when we assess the impact of our decisions.

An alternative is to let the customers indirectly *reveal* their level of satisfaction through their actions, say from the amount or frequency of their recent transactions, or through a lower churn rate for those who receive the reward relative to a well-designed control group.[1] Companies will always have customer surveys, and they should be treated as a potentially rich source of information. But a good practice is to always check if what they *say* is supported by their actions.

An Anatomy of a Decision: A Simple Decomposition

Figure 2-2 shows the general framework we will use to decompose and understand business decisions. Starting from the right, it is useful to repeat one more time that we *always start with the business*. If your objective is unclear or fuzzy, most likely the decision shouldn't be made at all. Companies tend to have a bias for action, so fruitless decisions are sometimes made. This may not only have unintended negative consequences on the business side; it could also take a toll on employees' energy and morale. Moreover, we now take for granted that our business objective can be measured through relevant KPIs. This is not to say that metrics arise naturally: as highlighted in a later example, we must choose our metrics carefully.

It is generally the case that we can't simply manipulate those business objectives ourselves (remember Enron?), so we need to take some actions or pull some levers in order to try to generate results. Actions themselves map to a set of consequences that directly affect our business objective. To be sure: *we* pull the levers, and our business objectives depend on consequences that arise when the "environment" reacts. The environment can be humans or technology, as we will see later.

1 We will talk about designing experiments or A/B tests later in this chapter.

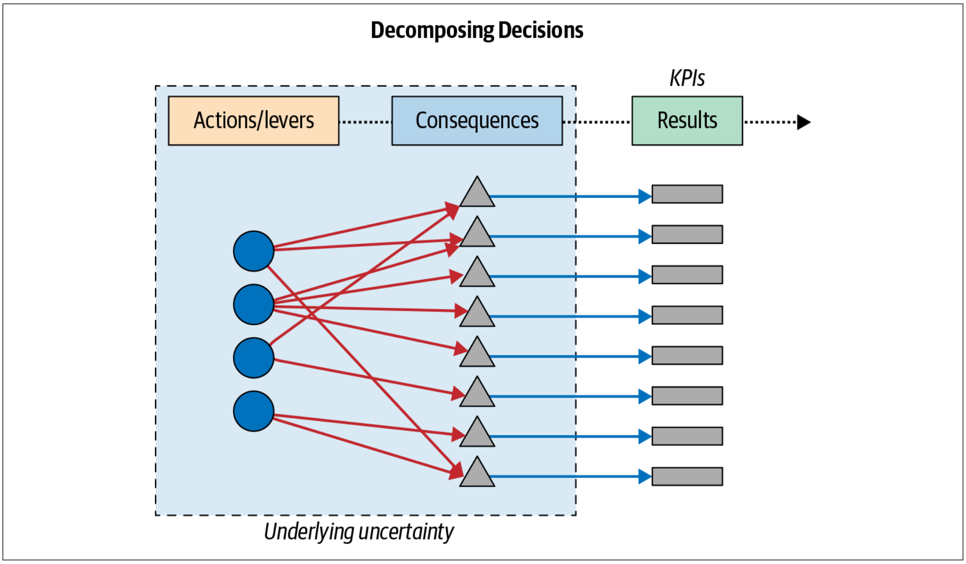

Figure 2-2. Decomposing decisions: actions, consequences, and business results

Even if the mapping is straightforward (most times it isn't), it's still mediated by uncertainty, since at the time of the decision it is impossible to know exactly what the consequences will be. We will use the powers of AI to embrace this underlying uncertainty, allowing us to make better decisions. But make no mistake: *value is derived from the decision, and prediction is an input to make better decisions.*

Difference Between Actions, Consequences, and Results

In case you haven't figured out the role that consequences play in the decomposition, here's an example. Suppose that our objective is to increase our revenues. To do so we decided to pull the pricing lever and offer some discounts to our customers. The consequence from our action is that our customers increase their spend on our brand, which itself generates higher revenues.

- *Action*: offer a discount
- *Consequence*: customers increase their demand for our product
- *Outcome*: revenues increase

To sum up, in our daily lives and in business, we generally pursue well-chosen, measurable objectives. Decision-making is the act of choosing among competing actions to attain these objectives. Data-driven decision-making is acting upon evidence to assess alternative courses of action. Prescriptive decision-making is the science of choosing the action that produces the best results for us; we must therefore be able to rank our choices relative to a measurable and relevant KPI.

An Example: Why Did You Buy This Book?

One example should illustrate how this decomposition works for *every* decision we make (Figure 2-3). Take your choice to purchase this book. This is an action you already made, but, surely, you could have decided otherwise. Since we always start with the business problem, let me imagine what type of problem you were trying to solve.

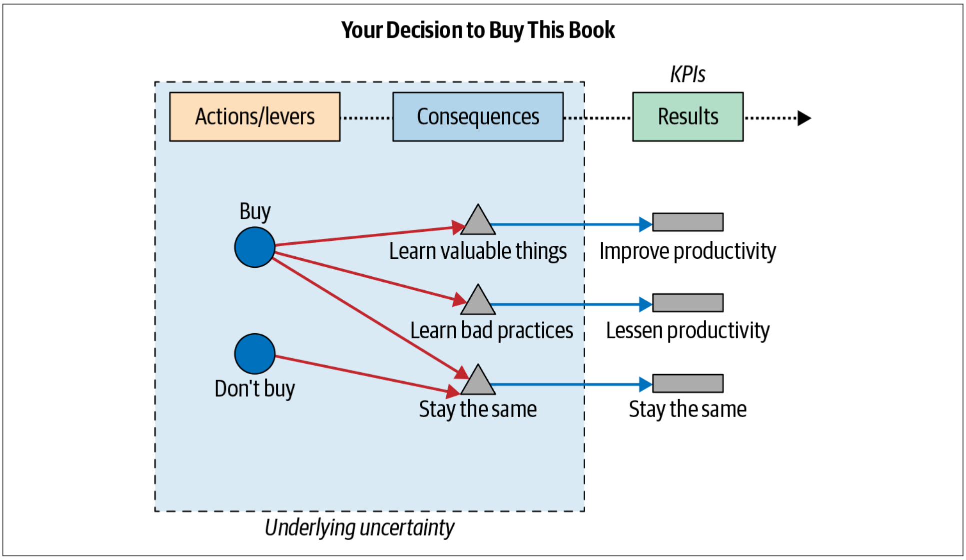

Figure 2-3. Decomposing your decision to buy this book

I don't know what objective you were solving when you decided to make the purchase, but in my case, I would've been interested in advancing my career. I will thus assume that the key metric you want to optimize is your productivity, and following our discussion on KPIs, I will conveniently assume that it is measurable.

Since you're reading the book, I'll also simplify all interesting details and just take two possible actions: buy or not buy. As the figure shows, if you buy the book there are at least three possible consequences: you learn valuable things, you learn bad practices, or you learn nothing. Naturally, each of these consequences impact your productivity.

If you don't buy the book many things can happen. For instance, you may get a sudden burst of inspiration and start understanding the intricacies of your job, thereby improving your productivity. Though plausible, we will appeal to Occam's razor and keep the most likely consequence that your knowledge and productivity stay the same.

Occam's Razor

When there are many plausible explanations for a problem, the principle known as Occam's razor appeals for the simplest one. Similarly, in statistics, when we have many possible models to explain an outcome, if we apply this principle we would attempt to use the most parsimonious one.

Don't worry if this isn't entirely clear now; Chapter 5 will be devoted entirely to improving our simplifying skills.

Finally, the difficulty here is that you don't really know what consequence will follow at the time you make the decision. For instance, contrary to your beliefs, it could be that O'Reilly made a mistake by signing this book or author. Unfortunately you will only know once you read it (so please do). This is the underlying uncertainty of this specific decision.

To sum up, notice how a simple action helped us to clearly and logically find the problem being solved, a set of levers, their consequences, and the underlying uncertainty. You can use this decomposition with any decision you make.

A Primer on Causation

The upcoming chapters will delve into each of the stages in the decomposition, so there will be enough time to understand where these levers come from and how they map to consequences. It is important, though, to stop now and recognize that this mapping is mediated by *causal* forces.

Going back to the saying that "correlation does not imply causation," no matter how many times we've heard about it, it is still very common to get the two terms confused. Our brain evolved to become a powerful pattern-recognizing machine, but we are not so well equipped to distinguish causation from correlation.[2]

Defining Correlation and Causation

Strictly speaking, correlation is the presence or absence of any linear dependencies in two or more variables. Less formally, two variables are correlated if they tend to "move together."

2 To be fair, even after taking into account this apparent impairment, we are by far the most sophisticated causal creatures that we know of, and we are infinitely superior to machines (since at the time of writing, they completely lack the ability, and it is not even clear when this ability may be achieved or if it's achievable at all).

Causality is harder to define, so let us take the shortcut followed by almost everyone: a relation of causality is one of cause and effect. X (partially) causes Y if Y is (partially) an effect of X. The "partial" qualifier is used because rarely is one factor the unique source of a relationship.

One can also define causality in terms of *counterfactuals*: *had X not taken place, is it true that Y had been observed?* If the answer is positive, then it is unlikely that a causal relationship from X to Y exists. Again, the qualifier "unlikely" is important and related to the previous "partial" qualifier: there are causal relations that only occur if the right combination of conditions is present.

Scatterplots like the one in Figure 2-1 are very good at depicting correlations between two variables, but unfortunately can't guide us in our quest to understand causation. To do so, it's quite standard to ask counterfactual questions in both directions and use Occam's razor to select a subset of plausible explanations.

Some Difficulties in Estimating Causal Effects

Estimating the causal impact on outcome Y of pulling a lever $X \implies Y$ is paramount since we are trying to engineer optimal decision-making. The analogy is not an accident: like the engineer who has to understand the laws of physics to build skyscrapers, bridges, cars, or planes, the analytical leaders of today must have some level of understanding of the causal laws mediating our own actions and their consequences to make the best possible decisions. And this is something that humans must do; AI will help us later in the decision-making process, but we must first overcome the causal hurdles.

Problem 1: We can't observe counterfactuals

As discussed in the previous sections, there are several problems that make our identification of causal effects much harder. The first one is that we only observe the facts, so we must imagine alternative *counterfactual* scenarios. It is an understatement that one of the most important skills analytical thinkers must develop is to question the initial interpretation given to empirical results, and to come up with counterfactual alternatives to be tested. Would the consequences be different had we pulled different levers, or the same levers but under different conditions?

Let's stop briefly to discuss what this question entails. Suppose we want to increase lead conversion in our telemarketing campaigns. Tom, a junior analyst who took one class in college on Freudian psychoanalysis, suggests that female call center representatives should have higher conversion rates, so the company decides to have its very capable group of female representatives make all outbound calls for a day. The next day, they meet to review the results: lead conversion went from the normal 5% to an outstanding 8.3%. It appears that Freud was right, or better, that Tom's decision to take the class had finally proven correct. Or does it?

To get the right answer, we need to imagine a customer receiving one call from the female representative in one universe, and the *exact same* call from a male representative in a parallel universe (Figure 2-4). Exact customer, exact timing, exact mood, and exact message; everything is the same in the two scenarios: we only change the tone of voice from that of a male to a female. Needless to say, putting in practice such a counterfactual sounds impossible. Later in this chapter, we will describe how we can simulate these impossible counterfactuals through well-designed randomized experiments or A/B tests.

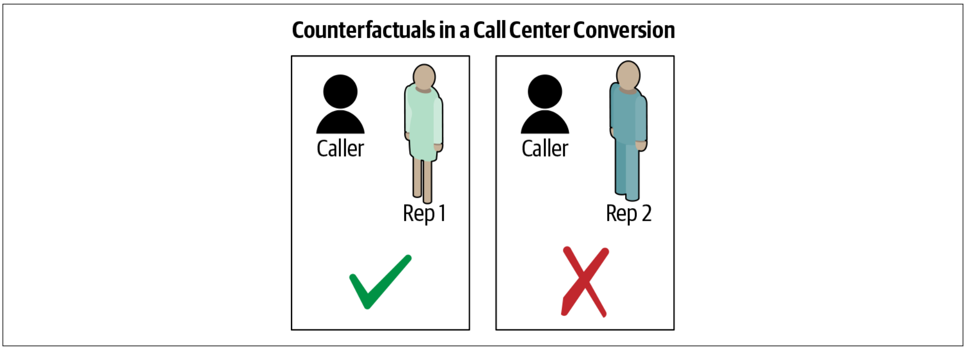

Figure 2-4. Counterfactual analysis of lead conversion rates in a call center

Problem 2: Heterogeneity

A second problem is *heterogeneity*. Humans are intrinsically different, each and every one the product of both our genetic makeup and lifetime experiences, creating unique worldviews and behaviors. Our task is not only to estimate how behavior changes when we choose to pull a specific lever—the causal effect—but we must also take care of the fact that different customers react differently. An influencer recommending our product will have different effects on you and me: I may now be willing to try it, while you may choose to remain loyal to your favorite brand. How do we even measure heterogenous effects?

Figure 2-5 shows the famous bell curve, the normal distribution, the darling of statistical aficionados. I'm using it here to represent the natural variation we may encounter when analyzing our customers' response when our influencer recommends our product. Some of his followers, like me, will accept the cue and react positively—represented as an action right of the vertical dashed line, the average response across all followers, followers' followers, and so on. Some will have no reaction whatsoever, and some may even react negatively—that's the beauty of human behavior; we sometimes get the full spectrum of possible actions and reactions. The shape of the distribution has important implications, and in reality, our responses may not be as symmetric; we may have longer left or right tails and reactions may be skewed toward the positive or the negative. The important thing here is that people react differently, making things even more difficult for us when we try to estimate a causal effect.

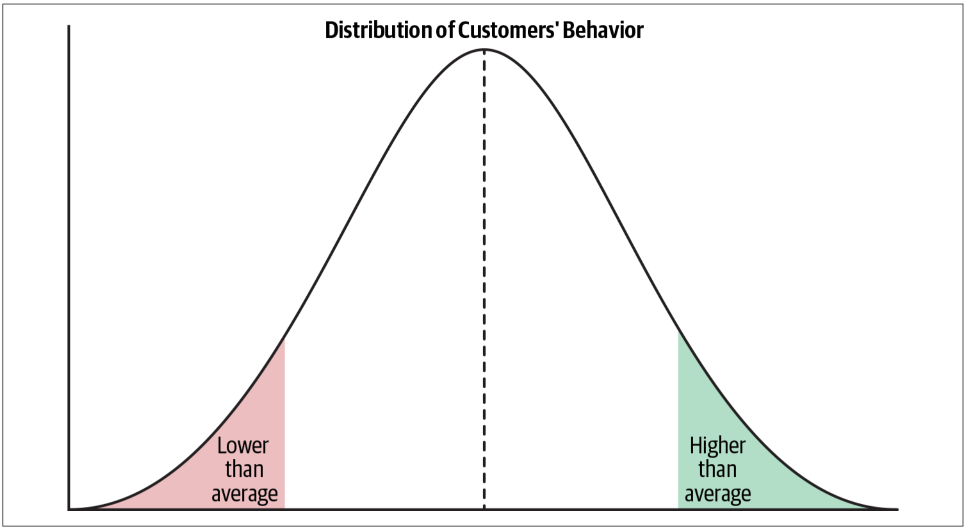

Figure 2-5. A normal distribution as a way to think about customer heterogeneity

The way we usually deal with heterogeneity is by dispensing of it by estimating a unique response, usually given by the average or the mean (the vertical line in Figure 2-5). The mean, however, is overly sensitive to extreme observations, so we may sometimes replace it with the median, which has the property that 50% of responses are lower (to the left) and 50% higher (to the right); with bell-shaped distributions the mean and the median are conveniently the same.

Problem 3: Confounders

When searching for causal relationships it's quite common to start by plotting scatterplots like the one in Figure 2-6 where each marker denotes a pair of (x, y) observations.

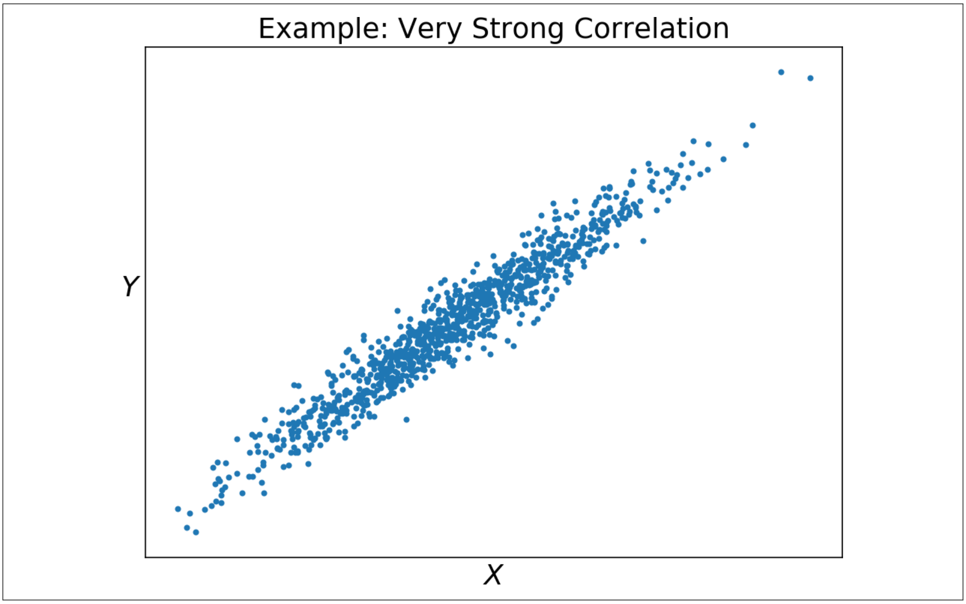

Figure 2-6. A simulation of two highly correlated variables

You may be tempted to assert that in this case there is clear evidence that X causes Y or vice versa—it is common to interpret scatterplots as relationships from the variable in the horizontal axis to outcomes on the vertical axis—but as Example 2-1 shows, this interpretation is faulty:

Example 2-1. Simulating the effect of a third unaccounted variable on the correlation of the other two

```
# fix a seed for our random number generator and number of observations to simulate
np.random.seed(422019)
nobs = 1000
# our third variable will be standard normal
z = np.random.randn(nobs,1)
# let's say that z --> x and z--> y
# Notice that x and y are not related!
x = 0.5 + 0.4*z + 0.1*np.random.randn(nobs,1)
y = 1.5 + 0.2*z + 0.01*np.random.randn(nobs,1)
```

To be sure, a third variable z positively affects both x and y, creating this spurious correlation. If we can control for this third variable (also known as a *confounder*), we may be able to get a better sense of the net relationship between the two variables of interest.

Consider the examples shown in Figure 2-7. The top left panel plots a measure of global CO_2 emissions and per capita real Gross Domestic Product (GDP) in Mexico for the period 1900–2016. The top right panel plots the number of divorces in Wales and England against Mexican GDP for 1900–2014. The bottom panel plots the three time series, indexed so that the 1900 observation is 100.[3]

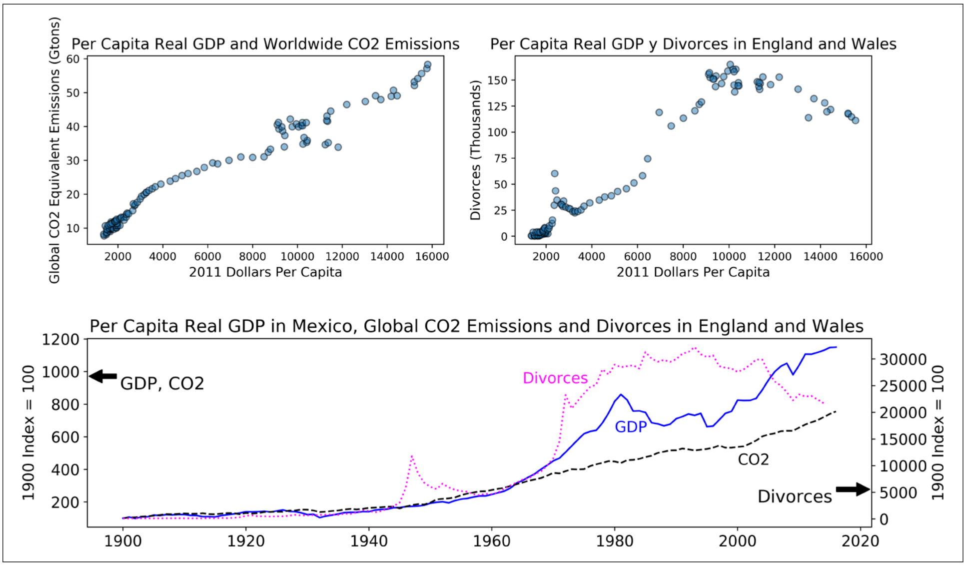

Figure 2-7. The top left panel plots global CO_2 emissions against real per capita Gross Domestic Product (GDP) for Mexico for the period 1900–2016; the top right panel does the same, replacing CO_2 emission with the number of divorces in Wales and England during 1900–2014; the bottom plot shows the time series for each of these variables

If we just inspected the scatterplots, we would be tempted to conclude that global emissions and divorces in the UK are somehow causally related to economic growth in Mexico. In this case, however, a third variable is responsible for such spurious correlation: statisticians and econometricians call a *time trend* the natural growth rate of a variable when plotted against time. The bottom panel shows that indeed these growth rates were very similar across the three variables in specific time periods.

Once we identify a confounder we can just *control* for it in our predictive algorithms (see the Appendix). But the problem of finding confounders is far from straightforward, so this task has to be done by us (and is thus not easily automatable).

3 Sources: GDP data comes from *https://oreil.ly/9J_wb*. CO_2 emissions from *https://oreil.ly/9J3XF*. Divorce rates from *https://oreil.ly/t_1x-*.

Problem 4: Selection effects

One final problem is the prevalence of selection effects. This usually arises because we choose the customer segments we want to act upon, or customers self-select themselves, or both. An important result in causal inference is that if we wish to estimate the causal effect from a treatment by comparing the average outcomes of two groups, we need to find a way to eliminate selection bias.[4]

> **Selection Bias and Causal Effects**
>
> Because of selection bias we may over- or underestimate a causal effect when we just take the difference in average outcomes across treated and control groups. Stated as an equation:
>
> $$\text{Observed Difference in Means} = \text{Causal Effect} + \text{Selection Bias}$$

It is standard practice to plot average outcomes as in the top panel of Figure 2-8. In this case, the outcome for the control is 0.29 units (let's say hundreds of dollars) higher than for those exposed to our action or lever. This number corresponds to the left-hand side of the previous equation. The bottom panel shows the corresponding distributions of outcomes. Using the mean to calculate differences is standard practice, but it is useful to remember that there are a full spectrum of responses, in some cases with a clear overlap between the two groups: the shaded areas show responses from customers in the two groups that are indistinguishable from each other.

4 Hereafter I will use the term "treated" or "those who receive a treatment" to refer to those customers that are exposed to our action or lever. This jargon is common in the statistical analysis of experiments and was originally borrowed from the analysis of medical trials.

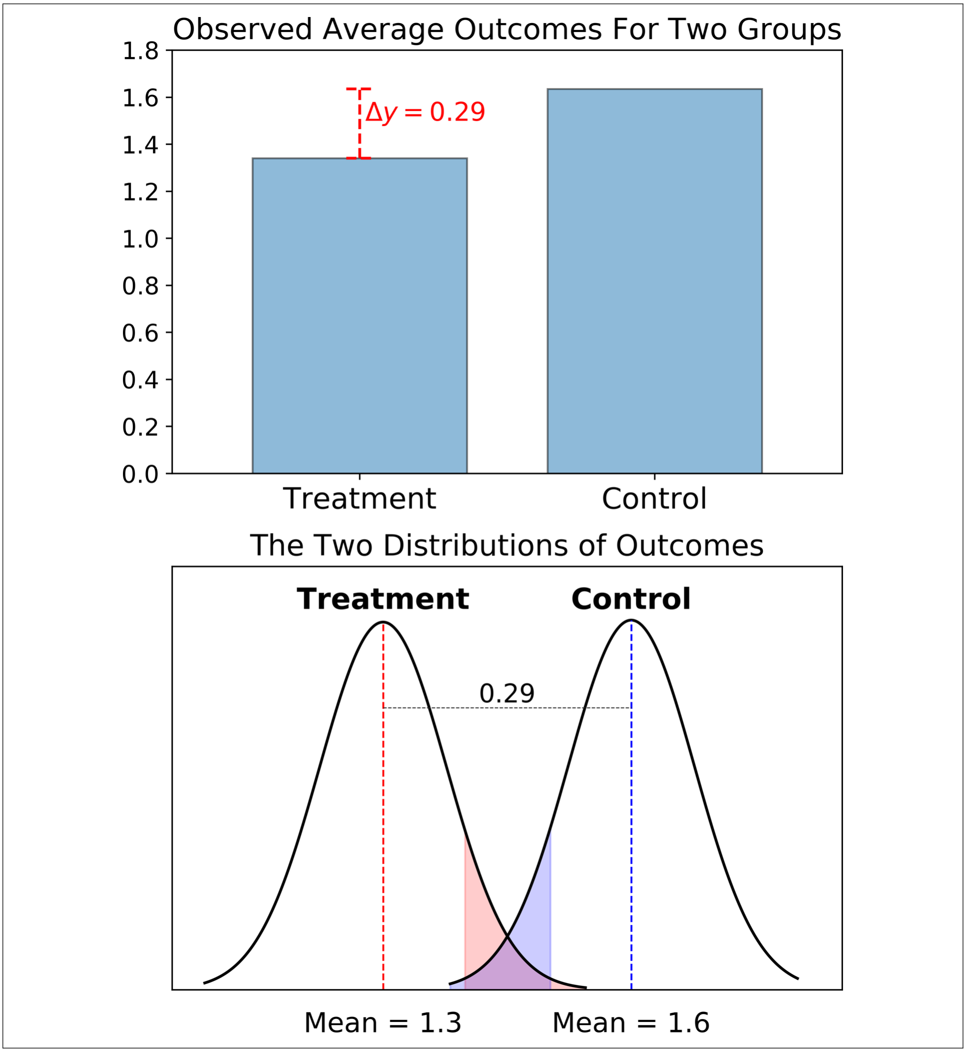

Figure 2-8. The top panel plots the observed differences in average outcomes for treatment and control groups; the bottom panel shows the actual distributions of outcomes

In any case, the difference in observed outcomes (left-hand side) is not enough for us since we already know that it is potentially biased by selection effects; since our interest is in estimating the causal effect, we must therefore devise a method to cancel this pervasive effect.

Statisticians and econometricians, not to mention philosophers and scientists, have been thinking about this problem for centuries. Since it is physically impossible to get an exact copy of each of our customers, is there a way to assign our treatments and circumvent the selection bias? It was Ronald A. Fisher, the famous 20th-century statistician and scientist, who put on firm grounds the method of experimentation, the most prevalent among practitioners when we want to estimate causal effects. The idea is simple enough to describe without making use of technical jargon.

A/B testing

In the industry it's quite standard to eliminate selection effects by running A/B tests, and the most data-driven companies run thousands of such experiments each year to find causal estimates that drive their decision-making.

I will devote several pages to A/B testing in the Appendix, so I'll just give a very superficial description of the technique here. Our objective is to estimate the causal effect of pulling a lever X on some output metric Y. Say that we wish to quantify the impact that a price discount has on our revenues.

We run an A/B test by splitting our customers into two groups: the A group acts as a control and gets the standard price. In contrast, the B group gets the price discount. Crucially, to avoid selection biases we choose our groups randomly, so that when we compare the average profits across groups, we can rest assured that we in fact estimated the causal effect. I left out all of the interesting technical details, so if you're interested, please consult the Appendix.

Uncertainty

We have now talked about each of the stages in the decomposition: starting with the business, we reverse engineer the actions or levers that impact our objective and corresponding KPIs, mediated by some consequences. However, since decisions are made under uncertainty, this mapping from actions to consequences is not known to us at the time of the decision. But by now we already know that uncertainty is not our enemy and that we can embrace it thanks to the advances in predictive power of AI.

But why do we have uncertainty? Let us first discuss what this uncertainty is not, and then we can talk about what it is. Think about flipping a coin. We know that with a balanced coin the chances it falls on heads are 50% and that the final outcome cannot be fully anticipated from the outset. Since we have played heads and tails for most of our lifetimes, this is an example of randomness that is quite close and natural to us.

This is not, however, the type of uncertainty we have when we are making decisions, and that is good news for us. The fact that ours is not pure randomness allows us to use powerful predictive algorithms, combined with our knowledge of the problem, to select input variables—also known as features—to create a prediction. With pure randomness, the best thing we can do is learn or model the distribution of outcomes and derive some theoretical properties that allow us to make smart choices or predictions.[5]

The four main sources of uncertainty *when we make decisions* are our need to simplify, heterogeneity, complex and strategic behavior arising from social interactions, and pure ignorance about the phenomenon, each of which will be described in turn. Note that as analytical thinkers, we should always know where uncertainty comes from, but it is not uncommon that we end up being taken by surprise.

Uncertainty from Simplification

One of my favorite quotes—commonly ascribed to Albert Einstein—is that "everything should be made as simple as possible. But not simpler." In the same vein, statistician George Box famously said that "all models are wrong, but some are useful." Models are simplifications, metaphors that help us understand the workings of the highly complex world we live in.

I cannot emphasize enough the importance that learning to simplify has for the modern analytical thinker. We will have enough time in Chapter 5 to exercise our analytical muscle through some well-known techniques, but we should now discuss the toll that simplification has.

As analytical thinkers and decision-makers we constantly face the trade-off between getting a good-enough answer or devoting more time to develop a more realistic picture of the problem at hand. We must decide how much uncertainty we're comfortable with and how much we are willing to accept in order to get a timely solution. But this calibration takes practice, as Einstein succinctly puts it in the first quote.

One clear example of the powers and dangers of simplification is maps. Figure 2-9 shows a section of the official Transit for London (TfL) tube map on the left and a more realistic version on the right, also by the transportation authority. With the objective of making our transportation decisions fast and easy, a map trades-off realism for ease-of-use. As users of the map, we now face uncertainty about the geography, distances, angles, and even the existence of possible relevant venues such as parks or museums. But to a first approximation we feel comfortable with this choice of granularity since our first objective is being able to get from our origin to a destination. We can take care of the remaining parts of the problem later.

5 In the coin tossing example, for instance, after observing the outcomes we may end up modeling the distribution as Bernoulli trials and predict a theoretically derived expected value (number of trials times the estimated probability of heads, say).

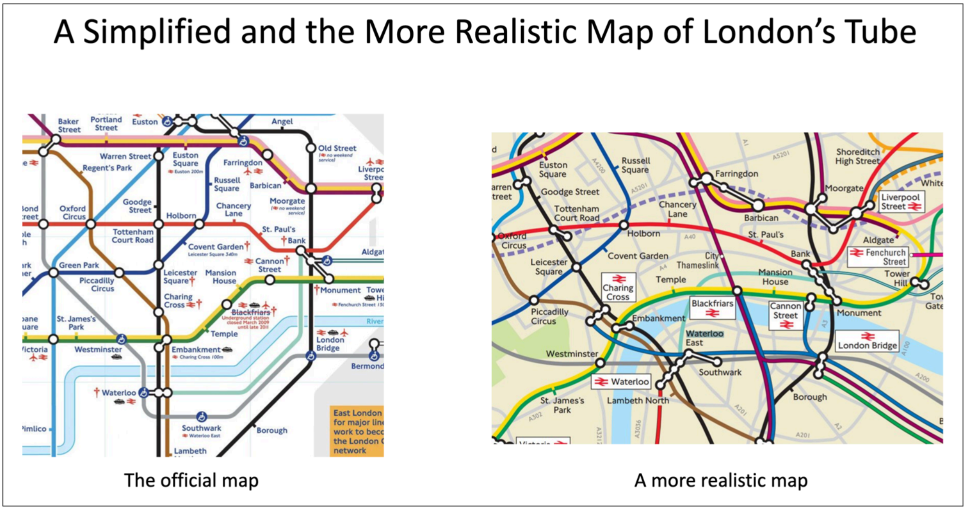

Figure 2-9. Sections of the London underground maps—the left panel corresponds to the official tube map, while the right panel shows a more realistic version of the same section

This last point takes me to another related issue: one common simplification technique is to divide a complex problem into simpler subproblems that can each be tackled independently; computer scientists call this the *divide and conquer* technique. When each of these subproblems gives rise to some uncertainty, nothing guarantees that the resulting uncertainty after aggregation becomes more tractable (unless we impose some simplifying assumptions to start with).

The moral of this story is that we should always remember that simplifying a problem usually brings additional uncertainty to the table. As Box, the statistician, commented, "…the approximate nature of the model must always be borne in mind".

Uncertainty from Heterogeneity

One important source of uncertainty when making business decisions comes from the fact that our customers react in very different ways. This wide variety of behaviors, tastes, and responses can be modeled with the use of distributions since that's how we generally deal with uncertainty (recall Figure 2-5). By doing so we can dispense with the nitty-gritty details of how and why outcomes are so diverse, and just focus on how uncertainty affects our final outcomes. This modeling approach is quite handy and forces us to know some basic properties about distributions.

Take the case of the *uniform distribution*. While it is most commonly assumed for simplification purposes, it can also be used if there's no reason to believe that outcomes will tend to accumulate. To give a concrete example, think about how people waiting for a train during peak hours end up being distributed across the platform. If their goal is to find a seat and enter the train as quickly as possible, it is most natural that they end up distributing uniformly.

We have already encountered the *normal distribution*, which is quite pervasive in the sciences. It is sometimes used for simplification purposes as it has some highly desirable properties (linearity, additivity), but it also arises naturally in many settings. For instance, we may appeal to a version of the central limit theorem, which states that under certain conditions, the distribution of averages or sums of numbers ends up being close enough to a normal.

Other commonly used distributions are power-law (or heavy tailed) distributions, which, contrary to the Gaussian distribution, have longer tails.[6] For instance, when modeling the reach or just the number of followers that your influencer has, we may resort to a power-law distribution, but there are many other examples where these distributions arise most naturally.[7]

Figure 2-10 shows the results of drawing one million observations from uniform, normal, and power-law distributions.

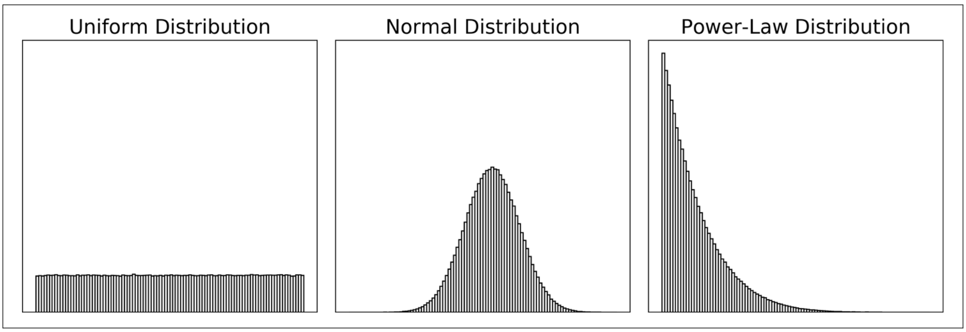

Figure 2-10. Histograms for the results of drawing one million observations from a uniform (left), normal (center), and power-law (right) distribution

Uncertainty from Social Interactions

Another source of uncertainty arises from the simple fact that we are social animals continuously interacting with each other. While this has been taking place for hundreds of thousands of years, the explosion of interactions with modern social networks has made it even more salient and prevalent.

6 The normal distribution accumulates 99% of the possible outcomes within 2.57 standard deviations from the mean and 99.9% within almost 3.3 standard deviations.

7 Other examples and applications of power-law distributions in business can be found in Crawford, Christopher G. et al., "Power law distributions in entrepreneurship: Implications for theory and research" *Journal of Business Venturing* 30, no. 5 (September 2015): 696-713. *https://oreil.ly/pSxTh*.

A first source of uncertainty comes from the strategic nature of our interactions with our customers and workforce, just to give two examples. With customer retention offers, for instance, it is not uncommon that customers understand our workings and motivations and end up gaming our system. Similarly, compensation schemes are quite commonly gamed by our sales executives, giving rise to somewhat unexpected results like delayed sales when goals have been or are unlikely to be reached.

But uncertainty may also arise from nonstrategic and very simple decision rules. One well-studied example is John Conway's Game of Life, which evolves in a two-dimensional grid such as the one depicted in Figure 2-11.[8] At any given time, each colored pixel can only interact with its immediate neighbors, thereby creating three possible outcomes: it lives, dies, or multiplies. There are only three simple rules of interaction, and depending on the initial conditions, you can get completely different outcomes that appear to be random to any observer.

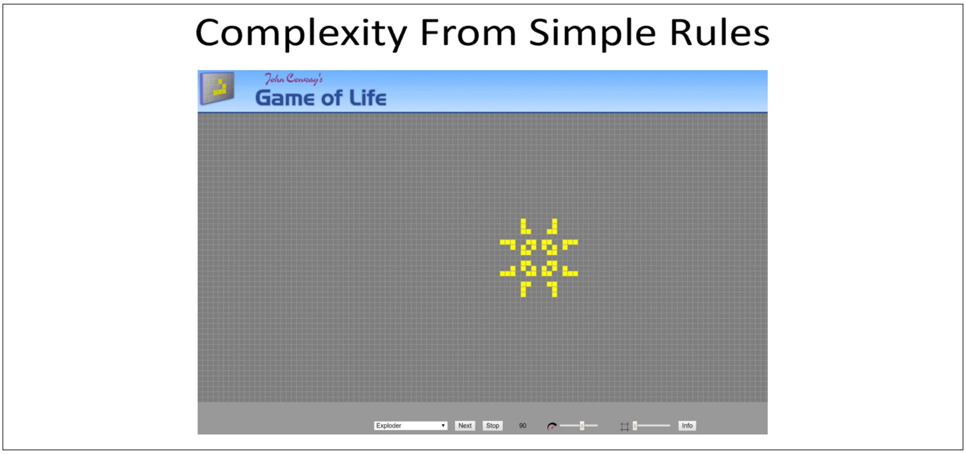

Figure 2-11. John Conway's Game of Life: a plethora of aggregate phenomena arises from three simple rules of how each cell or pixel interact with its neighbors

You may wonder if this is something worth your time and attention, or if it's just an intellectual curiosity. For a start, it should serve as a cautionary tale that even simple rules of behavior can create complex outcomes, so we don't really need sophisticated consumers trying to game our systems. But social scientists have also been using these tools to make sense of human behavior so, at the minimum, they ought to be useful for us when making decisions in our businesses.

8 You can "play" the game yourself at *https://playgameoflife.com* and marvel at the rich diversity of outcomes that can be generated by simple deterministic rules. See also *https://oreil.ly/6ruzw*.

Uncertainty from Ignorance

The last source of uncertainty is pure ignorance, as many times we simply don't know what will happen when a lever is pulled, and we are also unaware of the likely distribution of outcomes. In this case, it is not uncommon to start by assuming that outcomes follow a uniform or a normal distribution, later improving our knowledge by some sort of experimentation.

A company's ability to scale testing at the organizational level can create a rich knowledge base to innovate and create value in the medium-to-long term. But there is always a trade-off: we may need to sacrifice short-term profits for medium-term value and market leadership. That's why we need a new brand of analytical decision-makers in our organizations.

Key Takeaways

- *Analytical thinking* is the ability to identify and translate business questions into prescriptive solutions.
- *Value is created by making decisions*: we create value for our companies by making better decisions. Prediction is only one input necessary in our decision-making process.
- *Stages in the analysis of decisions*: there are generally three stages when we analyze a decision: we first gather, understand, and interpret the facts (descriptive stage). We then may wish to predict the outcomes of interest. Finally, we choose the levers to pull to make the best possible outcome (prescriptive stage).
- *Prescriptive decision-making*: decision-making is the act of choosing among competing actions to attain specific objectives. *Data-driven* decision-making is acting upon evidence to assess alternative courses of action. *Prescriptive* decision-making is the science of choosing the action that produces the best results for us.
- *Anatomy of a decision*: we choose an action that may have one or several consequences that impact our business outcomes. Since generally we don't know which consequence will result, this choice is made under conditions of uncertainty. The link between actions and consequences is mediated by causality.
- *Start with the business*: since our aim is to find the best course of action, we'd better be optimizing for the right question. So start with the business. One side benefit is that we usually enlarge the menu of levers available to us.
- *As important as asking the right question is the selection of the metrics to measure the impact of our decision-making*: many data science projects fail not because of the logic used but because we used the wrong set of metrics to measure the impact for our business question. Good metrics should be relevant and measurable.

- *One important skill for us to develop is the ability to create counterfactuals*: since causation mediates the mapping from actions to consequences, we must strengthen our ability to imagine alternative theories of why our business objectives follow from our actions.
- *Estimating causal effects has several important difficulties*: selection biases abound, so directly estimating the causal effect of a lever is generally not possible. We also need to master the use of counterfactual thinking and dealing with heterogenous effects.

Further Reading

Almost every book on data science or big data describes the distinction between descriptive, predictive, and prescriptive analysis. You may check Thomas Davenport's now classic *Competing on Analytics* or any of its sequels (Harvard Business Press), or Bill Schmarzo's *Big Data: Understanding How Data Powers Big Business*, or any of its prequels and sequels (Wiley).

The anatomy of decisions used here follows that literature and is quite standard. We will come back to this topic in Chapter 6, where I will provide sufficient references.

My favorite treatments of causality can be found in the books by Joshua Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics* (Princeton University Press) and their most recent, *Mastering 'Metrics': The Path from Cause to Effect* (Princeton University Press). If you are interested, you can find there the mathematical derivation of the equality between difference in observed outcomes and causal effects plus selection bias. They also present alternative methods to identify causality from *observational data*, that is, from data that was not obtained through a well-designed test.

A substantially different approach to causal reasoning can be found in Judea Pearl and Dana Mackenzie's *The Book of Why: The New Science of Cause and Effect* (Basic Books). Scott Cunningham's *Causal Inference: The Mixtape* provides a great bridge between the two approaches, focusing mostly on the first literature (econometrics of causal inference) but devoting a chapter and several passages to Pearl's approach using causal graphs and diagrams. At the time of writing, this book is free to download on his website.

I will provide many references on A/B testing in the Appendix. My discussion of uncertainty follows many ideas in Scott E. Page's *The Model Thinker: What You Need to Know to Make Data Work for You* (Basic Books). This is a great place to start thinking about simplification and modeling, and provides many examples of when and where distinct distributions, complex behavior, and network effects appear in real life.