The Book of Why: The New Science of Cause and Effect – Pearl and Mackenzie

Introduction: Mind Over Data

Every science that has thriven has thriven upon its own symbols.

– Augustus de Morgan (1864)

This book tells the story of a science that has changed the way we distinguish facts from fiction, and yet has remained under the radar of the general public. The consequences of the new science are already impacting crucial facets of our lives and have the potential to affect more, from the development of new drugs to the control of economic policies, from education and robotics to gun control and global warming. Remarkably, despite the apparent diversity and incommensurability of these problem areas, the new science embraces them all under a unified methodological framework that was practically non-existent two decades ago.

The new science does not have a fancy name: I call it simply "causal inference," as do many of my colleagues. Nor is it particularly high-tech. The ideal technology that causal inference strives to emulate is in our own mind. Some tens of thousands of years ago, humans began to realize that certain things cause other things, and that tinkering with the former could change the latter. No other species grasps this, certainly not to the extent that we do. From this discovery came organized societies, then towns and cities, and eventually the science-based and technology-based civilization we enjoy today. All because we asked a simple question: "Why?"

Causal inference is all about taking this question seriously. It posits that the human brain is the most advanced tool ever devised for managing causes and effects. Our brains store an incredible amount of causal knowledge which, supplemented by data, could be harnessed to answer some of the most pressing questions of our time. More ambitiously, once we really understand the logic behind causal thinking, we could emulate it on modern computers and create an "artificial scientist.” This would be a smart robot that discovers yet unknown phenomena, finds explanations to pending scientific dilemmas, designs new experiments and continually extracts more causal knowledge from the environment.

But before we can venture to speculate on such futuristic developments, it is important to understand the achievements that causal inference has tallied thus far. We will explore the way that it has transformed the thinking of scientists in almost every data-informed discipline, and how it is about to change our lives.

The new science addresses seemingly straightforward questions like these:

- How effective is a given treatment in preventing a disease?
- Did the new tax law cause our sales to go up, or was it our advertising campaign?
- What is the health-care cost attributable to obesity?
- Can hiring records prove an employer is guilty of sex discrimination?
- I’m about to quit my job, should I?

The common feature of these questions is that they are all concerned with cause-and-effect relationships. We can recognize them through words such as “preventing”, “cause,” “attributable to”, “discrimination” and “should I.” Such words are common in everyday language, and our society constantly demands answers to such questions. Yet until very recently science gave us no means even to articulate them, let alone answer them.

By far the most important contribution of causal inference to mankind has been to turn this scientific neglect into a thing of the past. Causal inference has spawned a simple mathematical language to articulate causal relationships that we know as well as those we wish to find out about. The ability to express this information in mathematical form has unleashed a wealth of powerful and principled methods for combining our knowledge with data and answering causal questions like the five above.

I have been lucky to be part of this scientific development for the past quarter of a century. I have watched its progress take shape in students' cubicles and research laboratories, and I have heard its breakthroughs resonate in somber scientific conferences, far from the limelight of public attention. Now, as we enter the era of strong artificial intelligence, and many tout the endless possibilities of Big Data and deep learning, I find it timely and exciting to present to the reader some of the most adventurous paths that the new science is taking, the way it impacts data science, and the many ways in which it will change our lives in the 21st century.

When you hear me describe these achievements as a "new science," you may be skeptical. You may even ask: "Why hasn't this been done a long time ago?" Say when Virgil first proclaimed, "Lucky is he who has been able to understand the causes of things" (29 BC). Or when the founders of modern statistics, Francis Galton and Karl Pearson, first discovered that population data can shed light on scientific questions. There is a long tale behind their unfortunate failure to embrace causation at this juncture, which we will tell in the historical sections of this book. But the most serious impediment, in my opinion, has been the fundamental gap between the vocabulary in which causal questions are cast and the traditional vocabulary in which scientific theories are communicated.

To appreciate the depth of this gap, imagine the difficulties that a scientist would face in trying to express some obvious causal relationships, say that the barometer reading B tracks the atmospheric pressure P. We can easily write down this relationship in an equation such as $B = kP$, where k is some constant of proportionality. The rules of algebra now permit us to rewrite this same equation in a wild variety of forms, for example, $P =B/k$, $k =B/P$, or $B – kP = 0$. They all mean the same thing—that if we know any two of the three quantities, the third is determined. None of the letters, k, B, P is in any mathematical way privileged over any of the others. How then can we express our strong conviction that it is the pressure that causes the barometer to change and not the other way around? And if we cannot express even this, how can we hope to express the many other causal convictions that do not have mathematical formulas, such as "The rooster's crow does not cause the sun to rise?"

My college professors could not do it and never complained. I would be willing to bet that none of yours ever did either. We now understand why: Never were they shown a mathematical language of causes, nor were they shown its benefits. It is in fact an indictment of science that it has neglected to develop such a language for so many generations. Everyone knows that flipping a switch will cause a light to turn on or off, or that a hot, sultry summer afternoon will cause sales to go up at the local ice cream parlor. Why then have scientists not captured such obvious facts in formulas, as they did with the basic laws of optics, mechanics, or geometry? Why have they allowed these facts to languish in bare intuition, deprived of mathematical tools that have enabled other branches of science to flourish and mature?

Part of the answer is that scientific tools are developed to meet scientific needs. Precisely because we are so good in handling questions about switches, ice cream, and barometers, it was not obvious that we needed special mathematical machinery to handle them. But, as scientific curiosity increased, and causal questions were posed in complex legal, business, medical and policy-making situations, we found ourselves lacking the tools and principles that mature science should provide.

Belated awakenings of this sort are not uncommon in science. For example, until about 400 years ago, people were quite happy with their natural ability to manage the uncertainties in daily life, from crossing a street to risking a fist fight. It was only after gamblers invented intricate games of chance, sometimes carefully designed to trick us into bad choices, that mathematicians like Christiaan Huygens, Blaise Pascal and Pierre de Fermat found it necessary to develop what we call today probability theory. Likewise, it was only when insurance organizations demanded accurate estimates of life annuity that mathematicians like Edmond Halley (1693) and Abraham de Moivre (1725) began looking at mortality tables to calculate life expectancies. Similarly, astronomers' demands for accurate predictions of celestial motion led Jacob Bernoulli, Pierre-Simon de Laplace, and Carl Friedrich Gauss to develop a theory of errors to help us extract signals from noise. These methods were all predecessors of today's statistics.

Ironically, the need for a theory of causation began to surface at the same time that statistics came into being. In fact modern statistics hatched out of the causal questions that Galton and Pearson asked about heredity and out of their ingenious attempts to answer them from cross-generation data. Unfortunately, they failed in this endeavor and, rather than pause to ask "Why?", they declared those questions off limits, and turned to develop a thriving, causality-free enterprise called statistics.

This was a critical moment in the history of science. The opportunity to equip causal questions with a language of their own came very close to being realized, but was squandered. In the following years, these questions were declared unscientific and went underground. Despite heroic efforts by the geneticist Sewall Wright (1889-1988), causal vocabulary was virtually prohibited for more than half a century. And when you prohibit speech, you prohibit thought, and you stifle principles, methods, and tools.

Readers do not have to be scientists to witness this prohibition. In Statistics 101, every student learns to chant: "Correlation is not causation." With good reason! The rooster crow is highly correlated with the sunrise, yet it does not cause the sunrise.

Unfortunately, statistics took this common-sense observation and turned it into a fetish. It tells us that correlation is not causation, but it does not tell us what causation *is*. In vain will you search the index of a statistics textbook for an entry on "cause." Students are never allowed[1] to say that X is the cause of Y—only that X and Y are *related* or *associated*.

Because of this prohibition, mathematical tools to manage causal questions were deemed unnecessary, and statistics focused its attention exclusively on how to summarize data, not on how to interpret data. A shining exception was geneticist Sewall Wright, whose invention of path analysis in the 1920s was a direct ancestor of the methods we will entertain in this book. However, path analysis was badly underappreciated in statistics and its satellite communities, and languished for decades in its embryonic status. What should have been the first step toward causal inference remained the *only* step until the 1980s. The rest of statistics, including the many disciplines that looked to it for guidance, remained in the Prohibition era, falsely believing that the answers to all scientific questions reside in the data, if only we knew how to unveil them through clever data-mining tricks.

Much of this data-centric history is still haunting us today. We live in an era when Big Data is presumed to be the solution to all our problems. Courses in "data science" are proliferating in our universities, and jobs for "data scientists" are plentiful in the companies that participate in the "data economy." But one thing I hope to convince you of in this book is that data are profoundly dumb. Data can tell you that the people who took a medicine recovered faster than people who did not take it, but they can't tell you why. Maybe they took the medicine because they could afford it, and they would have recovered just as fast without it.

Over and over again, in science and in business, we see situations where mere data aren't enough. Most big-data enthusiasts, while somewhat aware of these limitations, continue the chase after data-centric intelligence, as if we were still in the Prohibition era.

[1] With possibly one exception: If we have performed a randomized controlled trial, as discussed in Chapter 4.

As I mentioned earlier, things have changed dramatically in the past three decades. Nowadays, thanks to carefully crafted causal models, contemporary scientists can address problems that would have been considered unsolvable or even beyond the pale of scientific inquiry just a few decades ago. For example, only a hundred years ago, the question of whether cigarette smoking causes a health hazard would have been considered unscientific. The mere mention of the words "cause" or "effect" would create a storm of objections in any reputable statistical journal.

Even two decades ago, asking a statistician a question like "Was it the aspirin that stopped my headache?" would have been like asking if he believed in voodoo. To quote an esteemed colleague of mine, it would be "more of a cocktail conversation topic than a scientific inquiry." But today, such questions are posed routinely and answered with mathematical precision by epidemiologists, social scientists, computer scientists, and at least some enlightened economists and statisticians. To me, this change is nothing short of a revolution. I dare to call it the Causal Revolution, a scientific shakeup that embraces rather than denies our innate cognitive gift of understanding cause and effect.

The Causal Revolution did not happen in a vacuum; it has a mathematical secret behind it which can be best described as a calculus of causation, which answers some of the hardest problems ever asked about cause-effect relationships. I am thrilled to unveil this calculus to readers of this book, not only because the turbulent history of its development is intriguing, but even more because I expect that the full potential of this calculus will be developed one day beyond what I can imagine … perhaps even by a reader of this book.

The calculus of causation consists of two languages: causal diagrams, to express what we know, and a symbolic language, resembling algebra, to express what we want to know. The causal diagrams are simply dot-and-arrow pictures that summarize our existing scientific knowledge. The dots represent variables of interest, and the arrows represent known or suspected causal relationships between those variables, namely, which variable "listens" to which others. These diagrams are extremely easy to draw, comprehend and use, and the reader will find literally dozens of them in the pages of this book. If you can navigate using a map of one-way streets, then you can understand causal diagrams, and you can solve the type of questions we posed at the beginning of this preface.

Though causal diagrams are my tool of choice in this book, as well as the last 35 years of my research, they are not the only kind of causal model possible. Some scientists (e.g., econometricians) prefer to work with mathematical equations, and others (e.g., hard core statisticians) with a list of assumptions that ostensibly summarizes the structure of the diagram. But regardless of language, the model should depict, however qualitatively, the *process that generates the data*: in other words, the cause-effect forces that operate in the environment and shape the data generated.

Side by side with this diagrammatic "language of knowledge," we also have a symbolic "language of queries" to express the questions we want answers to. For example, if we are interested in the effect of a drug (D) on life expectancy (L), then our query might be written symbolically as: P(L|*do*(D)). In other words, what is the probability (P) that a typical patient would survive L years if made to take the drug? This question describes what epidemiologists would call an *intervention* or a *treatment* and corresponds to what we measure in a clinical trial. In many cases we may also wish to compare P(L|*do*(D)) with P(L|*do*(not-D)); the latter describes patients denied treatment, also called the "control" patients. The *do* operator signifies that we are dealing with an intervention rather than a passive observation; there is nothing remotely similar to this operator in classical statistics.

The reason we must invoke an intervention operator *do*(D) is to ensure that the observed change in lifespan L is due to the drug itself and is not *confounded* with other factors that tend to shorten or lengthen life. If, instead of intervening, we let the patient himself decide whether to take the drug, his decision may be influenced by those other factors, and lifespan differences between taking and not taking the drug would no longer be solely due to the drug. For example, suppose those who took the drug chose it because they are terminally ill. Such persons are surely different from those who did not take the drug, and the comparison would reflect differences in the severity of their disease, rather than the effect of the drug. By contrast, forcing patients to take the drug or to refrain from taking it, regardless of preconditions, would wash away preexisting differences and would provide us a valid comparison.

Mathematically, we write the observed frequency of lifespan L among patients who voluntarily take the drug as P(L|D), which is the standard *conditional probability* used in statistical textbooks. This expression stands for the probability (P) of lifespan L conditional on *seeing* the patient take drug D. Note that P(L|D) may be totally different from P(L|*do*(D)). This difference between *seeing* and *doing* is fundamental and explains why we do not regard the falling barometer to be a cause of the coming storm. Seeing the barometer fall increases the probability of the storm, while forcing it to fall does not affect this probability.

This confusion between seeing and doing has resulted in a fountain of paradoxes, some of which we will entertain in this book. A world devoid of P(L|*do*(D)) and governed solely by P(L|D) would be a strange world indeed. For example, patients would avoid going to the doctor to reduce the probability that they are seriously ill; cities would dismiss their firefighters to reduce the number of fires that break out; doctors would recommend a drug to male and female patients, but not to patients with undisclosed gender; and so on. It is hard to believe that less than three decades ago science did operate in such a world—the *do* operator did not exist.

One of the crowning achievements of the Causal Revolution has been to explain how to predict the effects of an intervention without actually enacting it. It would never have been possible if we had not, first of all, defined the *do*-operator so that we can ask the right question and, second, devised a way to emulate it by non-invasive means.

When the scientific question of interest involves retrospective thinking, we call on another type of expression that is unique to causal reasoning, called a *counterfactual*. For example, suppose that Joe took drug D and died a month later; our question of interest is whether the drug might have caused his death. To answer this question we need to imagine a scenario where Joe was about to take the drug but changed his mind. Would he have been alive?

Again, classical statistics deals only with summarization of data, so it does not provide even a language for asking that question. Causal inference provides a notation, and more importantly it offers a solution. Just as in the case of predicting the effect of interventions (mentioned above), in many cases we can emulate human retrospective thinking with an algorithm that takes what we know about the observed world and produces an answer about the counterfactual world. This "algorithmization of counterfactuals" is another gem uncovered by the Causal Revolution.

Counterfactual reasoning, which deals with questions like "What if?" and "What might have been?", might strike some readers as unscientific. Indeed, the answers to such questions can never be confirmed or refuted by empirical observations. Yet our minds make "What if?" and "What might have been?" judgments all the time, very reliably and reproducibly. We all understand, for instance, that “had my rooster been silent this morning, the sun would have risen just as well.” This consensus stems from the fact that counterfactuals are not products of whimsy but reflect the very structure of our world model. Two people who share the same causal model will also share all counterfactual judgments.

Counterfactuals are the building blocks of moral behavior as well as scientific thought. The ability to reflect back on one's past actions and envision alternative scenarios is the basis of free will and social responsibility. The algorithmization of counterfactuals invites thinking machines to benefit from this ability and participate in this (until now) uniquely human way of thinking about the world.

My mention of thinking machines in the last paragraph is intentional. I came to this subject as a computer scientist working in the area of artificial intelligence (AI), which entails two points of departure from most of my colleagues in the causal inference arena. First, in the world of AI, you do not really understand a topic until you can teach it to a stupid robot. That is why you will find me emphasizing and re-emphasizing notation, language, vocabulary and grammar. For example, I obsess over whether or not we can *express* a certain claim in a given language, and whether one claim *follows* from others. It is amazing how much one can learn from just following the grammar of scientific utterances. My emphasis on language also comes from a deep conviction that language shapes our thoughts. You cannot answer a question that you cannot ask, and you cannot ask a question that you have no words for. As a student of philosophy and computer science, my attraction to causal inference has largely been triggered by the excitement of seeing an orphaned scientific language making it from birth to maturity.

My background in machine learning has given me yet another incentive for studying causation. In the late 1980s, I realized that machines' lack of understanding of causal relations was perhaps the biggest roadblock to the achievement of human-level intelligence. In the last chapter of this book I will return to my roots, and together we will explore what implications the Causal Revolution might have for artificial intelligence. I believe that strong AI is an achievable goal, and that it is not to be feared, precisely because causality is part of the solution. A causal reasoning module will give machines the ability to reflect on their mistakes, to pinpoint weaknesses in their software, to function as moral entities, and to converse naturally with humans about their own choices and intentions.

A Blueprint for Causal Inference

In our era, I am sure that many readers have heard terms like “knowledge,” “information,” “intelligence” and “data,” and some may feel confused about the differences between them or how they interact. Now I am proposing to throw another term, “causal model,” into the mix, and the reader may justifiably wonder if this will only add to the confusion.

It will not! In fact, it will anchor the elusive notions of science, knowledge and data in a concrete and meaningful setting, and we will elucidate how the three work together to produce answers to difficult scientific questions. In Figure 1, I have drawn a blueprint for a “causal inference engine” that might handle causal reasoning for a future artificial intelligence. It’s important to realize that this is not only a blueprint for the future; it is also a guide to how causal models work in scientific applications *today* and how they interact with data.

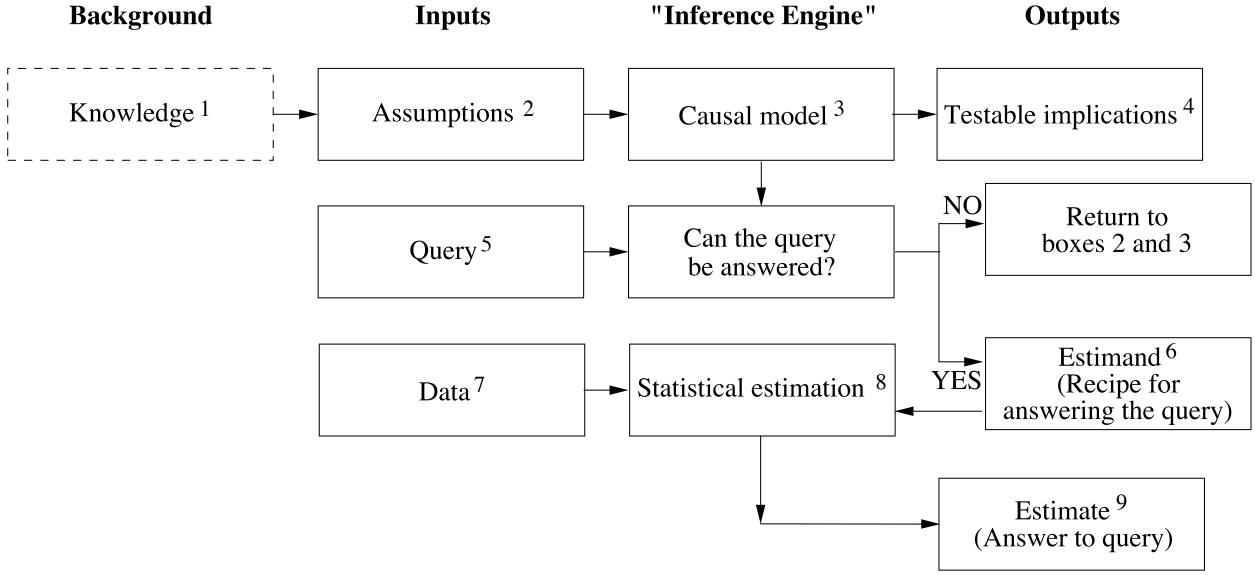

Figure 1. How an "inference engine" combines data with causal knowledge to produce answers to queries of interest. Dashed boxes are not part of the engine but required for building it. Arrows could also be drawn from boxes 4 and 9 to box 1, but we have opted to keep the diagram simple.

The inference engine is a machine that accepts three different kinds of inputs, Assumptions, Queries, and Data, and produces three kinds of outputs. The first of the outputs is a Yes/No decision as to whether the given query can in theory be answered under the existing causal model, assuming perfect and unlimited data. If the answer is Yes, the inference engine next produces an Estimand. This is a mathematical formula that can be thought of as a recipe for generating the answer from any hypothetical data, whenever they are available. Finally, after the inference engine has received the Data input, it will use the recipe to produce an actual Estimate for the answer, along with statistical estimates of the amount of uncertainty in that estimate. This uncertainty reflects the limited size of the data set, as well as possible measurement errors or missing data.

To dig more deeply into the chart, I have labeled the boxes 1 through 9, which I will annotate in the context of the query, "What is the effect of drug D on lifespan L?"

1. "Knowledge" stands for traces of experience the reasoning agent has had in the past, including past observations, past actions, education and hearsay, that are deemed relevant to the query of interest. The dotted box around "Knowledge" indicates that it remains implicit in the mind of the agent, and is not explicated formally in the model.
2. Scientific research always requires simplifying assumptions, that is, statements which the researcher deems worthy of making explicit on the basis of the available Knowledge. While most of the researcher's knowledge remains implicit in his or her brain, only Assumptions see the light of day and are encapsulated in the model. They can in fact be

read from the model, which has led some logicians to conclude that a model is nothing more than a list of assumptions. Computer scientists take exception to this claim, noting that the way assumptions are represented can make a profound difference in one's ability to specify them correctly, draw conclusions from them, and even extend or modify them in light of compelling evidence.

3. Various options exist for causal models: causal diagrams, structural equations, logical statements, etc. I am strongly sold on causal diagrams for nearly all applications, primarily due to their transparency but also due to the explicit answers they provide to many of the questions we wish to ask. For the purpose of constructing the diagram, the definition of "causation" is simple, if a little metaphorical: a variable X is a cause of Y if Y "listens" to X and decides its value in response to what it hears. For example, if we suspect that a patient's life span L "listens" to whether or not drug D was taken, then we call D a cause of L and draw an arrow from D to L in a causal diagram. Naturally, the answer to our query about D and L is likely to depend on other variables as well, which must also be represented in the diagram along with their causes and effects. (Below we will denote them collectively by Z.)
4. The listening pattern prescribed by the paths of the causal model usually results in observable patterns or dependencies in the data. These patterns are called "testable implications" because they can be used for testing the model. These are statements like "there is no path connecting D and L," which translates to a statistical statement, "D and L are independent," i.e., finding D does not change the likelihood of L. If the data contradict this implication, then we need to revise our model. Such revisions require another engine which obtains its inputs from Box 4 and 7 and computes the "degree of

fitness," that is, the degree to which the Data are compatible with the model's assumptions. For simplicity, we did not show this second engine in Figure 1.

5. Queries submitted to the inference engine are the scientific questions that we want to answer. They must be formulated in causal vocabulary. For example, what is $P(L|\mathit{do}(D))$? One of the main accomplishments of the Causal Revolution has been to make this language scientifically transparent as well as mathematically rigorous.
6. "Estimand" comes from Latin, meaning "that which is to be estimated." This is a statistical quantity to be estimated from the data that, once estimated, can legitimately represent the answer to our query. While it is written as a probability formula, for example $P(L|D, Z) \times P(Z)$, it can also be thought of as a recipe for answering the causal query from the type of data we have, *assuming* it can be answered.

   It's very important to realize that, contrary to traditional estimation in statistics, some queries may not be answerable under the current causal model, even after collecting any amount of data. For example, if our model shows that both D and L depend on a third variable Z (say, the stage of a disease) and if we do not have any way to measure Z, then the query $P(L|\mathit{do}(D))$ cannot be answered. In that case it is a waste of time to collect data. Instead we need to go back and refine the model, either by adding new scientific knowledge that might allow us to estimate Z or by making simplifying assumptions (at the risk of being wrong)—for example, that the effect of Z on D is negligible.

7. Data are the ingredients that go into the estimand recipe. It is critical to realize that data are profoundly dumb about causal relationships. They tell us about quantities like $P(L|D)$ or $P(L|D, Z)$. It is the job of the estimand to tell us how to bake these statistical quantities into one expression that, based on the model assumptions, is logically equivalent to the causal query, say $P(L|\mathit{do}(D))$.

Notice that the whole notion of estimands and in fact the whole top part of Figure 1 does not exist in traditional methods of statistical analysis. There, the estimand and the query coincide. For example, if we are interested in the proportion of people among those with lifetime L who took the drug D, we simply write this query as P(D|L). The same quantity would be our estimand. This already specifies what proportions in the data need to be estimated and requires no causal knowledge. For this reason, some statisticians to this day find it extremely hard to understand why some knowledge lies outside the province of statistics and why data alone cannot make up for lack of scientific knowledge.

8. The estimate is what comes out of the oven! However, it is only approximate because of one other real-word fact about data: they are always only a finite sample from a theoretically infinite population. In our running example, the sample consists of the patients we choose to study. Even if we choose them at random, there is always some chance that the proportions measured in the sample are not representative of the proportions in the population at large. Fortunately, the discipline of statistics gives us many, many ways to manage this uncertainty—maximum likelihood estimators, propensity scores, confidence intervals, significance tests, etc.
9. In the end, if our model is correct and our data are sufficient, we get an answer to our causal query, such as: Drug D increases the lifespan L of diabetic patients Z by 30 percent, plus or minus 20 percent. Hooray! The answer will also add to our scientific knowledge (Box 1) and, if things did not go the way we expected, it might suggest some improvements to our causal model (Box 3).

This flowchart may look complicated at first, and you might wonder whether it is really necessary. Indeed, in our ordinary lives, we are somehow able to make causal judgments without consciously going through such a complicated process, and certainly without resorting to the mathematics of probabilities and proportions. Our causal intuition alone is usually sufficient for handling the kind of uncertainty we find in household routines, or even in our professional lives. But if we want to teach a dumb robot to think causally, or if we are pushing the frontiers of scientific knowledge, where we do not have intuition to guide us, then a carefully structured procedure like this is mandatory.

I especially want to highlight the role of Data in the above process. First, notice that we collect data only *after* we posit the causal model, *after* we state the scientific query we wish to answer, and *after* we derive the estimand. This contrasts with the traditional statistical approach, mentioned above, which does not even have a causal model.

But in our present-day scientific world, there is a new challenge to sound reasoning about causes and effects. While awareness of the need for a causal model has grown by leaps and bounds among the sciences, many researchers in artificial intelligence would like to skip the hard step of constructing or acquiring a causal model and rely solely on data for all cognitive tasks. The hope—and at present, it is only a silent hope—is that the data themselves will guide us to the right answers whenever causal questions come up.

I am an outspoken skeptic of this trend, because I know how profoundly dumb data are about causes and effects. For example, information about the effects of actions or interventions is simply not available in raw data, unless it is collected by controlled experimental manipulation. By contrast, if we are in possession of a causal model, we can often predict the result of an intervention from hands-off, intervention-free data.

The case for causal models becomes even more compelling when we seek to answer counterfactual queries such as "What would have happened had we acted differently?" We will discuss counterfactuals in great detail because they are the most challenging queries for any artificial intelligence. They are also at the core of the cognitive advances that made us human, and the imaginative abilities that have made science possible. We will also explain why any query about the *mechanism by which causes transmit their effects*—the most prototypical "Why?" question—is actually a counterfactual question in disguise. Thus, if we ever want robots to answer "Why?" questions or even understand what they mean, we must equip them with a causal model and teach them how to answer counterfactual queries, as in Figure 1.

Another advantage causal models have which data mining and deep learning lack is adaptability. Note that in Figure 1, the estimand is computed on the basis of the causal model alone, prior to examining the specifics of the data. This makes the causal inference engine supremely adaptable, because the estimand computed is good for *any* population that is compatible with the qualitative model, regardless of the numerical relationships among the variables.

To see why this adaptability is important, compare this engine with a learning agent—maybe a human, maybe a deep-learning algorithm, maybe a human using a deep-learning algorithm—who tries to learn solely from the data. By observing the outcome L of many patients given drug D, she is able to predict the probability that a patient with characteristics Z will survive L years. Now she is transferred to a different hospital, in a different part of town, where the population characteristics (diet, hygiene, work habits) are different. Even if these new characteristics merely modify the numerical relationships among the variables recorded, she will still have to retrain herself and learn a new prediction function all over again. That's all that a deep-learning program can do: fit a function to data. On the other hand, if she possessed a model of how the drug operates, and its causal structure remains intact in the new location, then the estimand she obtained in training would remain valid. It could be applied to the new data to generate a new population-specific prediction function.

Many scientific questions look differently "through a causal lens," and I have delighted in playing with this lens, which over the last 25 years has been increasingly empowered by new insights and new tools. I hope and believe that readers of this book will share in my delight. Therefore, I'd like to close this introduction with a preview of some of the coming attractions in this book.

In Chapter 1 we will assemble the three steps of observation, intervention, and counterfactuals into the Ladder of Causation, the central metaphor of this book. It will also expose you to the basics of reasoning with causal diagrams, our main modeling tool. This chapter will set you well on your way to becoming a proficient causal reasoner—and in fact, you will be far ahead of generations of data scientists who attempted to interpret data through a model-blind lens, oblivious to the distinctions that the Ladder of Causation illuminates. In Chapter 2 we will tell the bizarre story of how the discipline of statistics inflicted causal blindness on itself, with far-reaching effects for all sciences that depend on data. We will also tell the story of one of the great heroes of this book, the geneticist Sewall Wright, who in the 1920s drew the first causal diagrams and for many years was one of the few scientists who dared to take causality seriously.

Chapter 3 will relate the equally curious story of how I became a convert to causality, through my work in AI and particularly on Bayesian networks. These were the first tool that allowed computers to think in "shades of gray"—and for a time I believed this was the key to unlocking AI. Towards the end of the 1980s I became convinced that I was wrong, and this chapter tells about my journey from prophet to apostate. Nevertheless, Bayesian networks remain a very important tool for AI and still encapsulate much of the mathematical foundation of causal diagrams. In addition to a gentle, causality-minded introduction to Bayes' Rule and Bayesian methods of reasoning, Chapter 3 will entertain the reader with examples of real-life applications of Bayesian networks.

Chapter 4 will tell about the major contribution of statistics to causal inference: the randomized controlled trial (RCT). From a causal perspective, the RCT is a man-made tool for uncovering the query P(L|*do*(D)), which is a property of nature. Its main purpose is to disassociate variables of interest (say, D and L) from other variables (Z) that would otherwise affect them both. Disarming the distortions produced by such "lurking variables" has been a century-old problem called "confounding." We will walk the reader through surprisingly simple solution to the general confounding problem, which you will be able to acquire in ten minutes of playful tracing of paths in the diagram.

In Chapter 5 we will give an account of a seminal moment in the history of causation and indeed the history of science, when statisticians struggled with the question, "Does smoking cause lung cancer?" Unable to use their favorite tool, the randomized controlled trial, they struggled to agree on an answer, or even on how to make sense of the question. The smoking debate brings the importance of causality into its sharpest focus. Millions of lives were lost or shortened because scientists did not have an adequate language or methodology for answering causal questions.

Chapter 6 will, I hope, be a welcome diversion for the reader after the serious matters of Chapter 5. This is a chapter of paradoxes: the Monty Hall paradox, Simpson's paradox, Berkson's paradox and others. Classical paradoxes like these can be enjoyed as brain-teasers, but they have a serious side too, especially when viewed from a causal perspective. In fact, almost all of them represent clashes with causal intuition, and reveal therefore the anatomy of that intuition. They were a "canary in a coal mine" that should have alerted scientists to the fact that human intuition is grounded in causal, not statistical logic. I believe that the reader will enjoy this new twist on his or her favorite old paradoxes.

In Chapters 7-9 we will, finally, take the readers on a thrilling ascent of the Ladder of Causation. We start in Chapter 7 with questions about intervention, and explain how my students and I went through a twenty-year struggle to automate the answers to *do*-type questions. We succeeded, and this chapter will explain the guts of the "causal inference engine," which produces the Yes/No answer and the estimand in Figure 1. In the course of studying this engine, the reader will be empowered to spot certain patterns in the causal diagram that deliver immediate answers to the causal query. These patterns are called back-door adjustment, front-door adjustment, and instrumental variables, the workhorses of causal inference in practice.

In Chapter 8, we will take you to the top of the ladder by discussing counterfactuals. These have been seen as a fundamental part of causality at least since 1748, when the Scottish philosopher David Hume proposed the following somewhat contorted definition of causation: "We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed." David Lewis, a philosopher at Princeton University who died in 2001, pointed out that Hume really gave two definitions, not one. The first is the "regularity" definition (i.e., the cause is regularly followed by the effect) and the second one is counterfactual ("if the first object had not been…"). While philosophers and scientists had mostly paid attention to the regularity definition, Lewis argued that the counterfactual definition aligns more closely with human intuition: "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."

Readers will be excited to find out that we can now move past the academic debates and compute an actual value (or probability) for any counterfactual query, no matter how convoluted. Of special interest are questions concerning necessary and sufficient causes of observed events. For example, how likely is it that the defendant's action was a necessary cause of the claimant's injury? How likely is it that man-made climate change is a sufficient cause of a heat wave?

Finally, in Chapter 9 we will discuss the topic of mediation. You may have wondered, when we talked about drawing arrows in a causal diagram, whether we should draw an arrow from drug D to lifespan L if the drug affects lifespan only by way of its effect on blood pressure Z (a *mediator*). In other words, is the effect of D on L direct or indirect? And if both, how do we assess their relative importance? Such questions are not only of great scientific interest, but they also have practical ramifications; if we understand the mechanism through which a drug acts, we might be able to develop other drugs with the same effect that are cheaper, or have fewer side effects. The reader will be pleased to discover how this age-old quest for a mediation mechanism has been reduced to an algebraic exercise, and how scientists are using the new tools in our causal toolkit to solve such exercises.

Chapter 10 will bring the book to a close by coming back to the problem that initially led me to causation: the problem of automating human-level intelligence (sometimes called "strong AI"). I believe that causal reasoning is essential for machines to communicate with us in our own language about policies, experiments, explanations, theories, regret, responsibility, free will, and obligations—and, eventually, to make their own moral decisions.

If I could sum up the message of this book in one pithy phrase, it is that *you are smarter than your data*. Data do not understand causes and effects; humans do. My hope is that the new science of causal inference will enable us to better understand how we do it, because there is no better way to understand ourselves than to emulate ourselves. In the age of computers, this new understanding also brings with it the prospect of amplifying our innate abilities, so that we can make better sense of data, be it big or small.