1 A Computational Approach to Causality

Chapter Summary

As discussed in the previous chapter, distinguishing cause from effect is essential in science, in particular for understanding complex and living systems. In this chapter, we will provide a general overview of approaches bridging computation and causality. We begin by surveying concepts and outlining the current state of the field to convey its richness, its multiple difficulties, and its challenges. In particular, we survey how causality has been addressed from different conceptual standpoints. Interestingly, two of these strands reflect the tension that has characterised AI research since the 1956 Dartmouth workshop considered the inception of AI. The first modern take on causality derives from a discrete logical and symbolic mode of thinking. Central in this paradigm is the belief that to construct intelligent reasoning, i.e., causally competent systems, we need symbolic representation and computations thereof. The second framework that we discuss in this chapter essentially displaces causality as represented in terms of discrete symbolic elements in favour of statistical continuous distributions. It has pre-empted progress in the study of causality and replaced it with the study of correlations and associations, plunging the field of causality into a spell of wintry aridity. Correlations in data and observations assumed saliency during this period spanning several decades and still prevail in all areas of science. The third line of thinking flips the problem on its head. Instead of saying ‘let us assume a statistical model and ask how our data could or could not be accounted for by it’, a machine learning approach reformats the problem as a learning task. Given the data, it says, ‘let us discern an implicit model capturing the regularities that characterise it’. There is also a major line of research that merges statistical and learning approaches to finding patterns in data. Bayesian approaches, including network representation, are examples of this. Yet, as we will see, both the statistical and the learning approach often confound causes and effects in data, a shortcoming inherited from traditional statistics. This is not entirely surprising, as many of these approaches were not primarily designed for elucidating causality from data without recourse to probability distributions. The fourth paradigm that we review originates in what could be considered a scientific computing and dynamical systems mode of thinking. In practice, it uses a model which is generative in a mechanistic sense (e.g., differential equations), not a merely

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

statistical model, and it asks whether such a generative model is sufficient to account for the observations. The advantage is that, here, candidate models are made explicit and can be tested against data and predictions. Yet the sufficiency of a model does not imply its necessity, as there could well be other casual models (explanations) for the same set of observations. This chapter sets the stage for the major ways scientists have dealt with causality in the quantitative sciences, exploring their advantages and inherent limitations. It paves the way for critically inquiring into what we need to accomplish in order to come to grips with causality. The chapter ends with a set of multiple choice and discussion questions to establish key takeaway points.

1.1 The Challenge of Causality

Below we will review several concepts – as outlined above – that are relevant to the objective of this book, which is to explain how the theories of computability, dynamical systems, and algorithmic complexity can help address the challenge of causal discovery, specifically by helping to generate mechanistic models underlying data and observations, that is, models that one can run on a computer, for example, in order to make educated guesses about the cause of a system's behaviour, its likely future behaviour, and how it may be manipulated.

Because there are many things we cannot understand or predict, we tend to believe that we are dominated by chance. However, science tells us that much of what we believe to be random is in fact highly determined by events in our world, events that precede and determine other events. Science, or at least classical physics, tells us that everything happens for a reason. What turns out to be difficult is to find or identify these reasons. Later in this book, we will explain that determinism and predictability are not the same thing and do not imply each other. In particular, we will see that determinism does not imply predictability.

Causality is a property of objects and systems that have been produced by a cause, as opposed to having occurred or appeared by chance. Causality is the property that distinguishes science from magic, and is the driver of the scientific method as it attempts to discover the causes of what happens in the natural world. The ultimate aim of science is to understand the world through simplified causal models and thereby to predict and change events. These events can be of any kind, for example, producing a new drug to treat a disease, or designing a new electric plant to generate more energy. Most of the time, the connection between cause and effect is anything but trivial. One of the most popular examples is the so-called butterfly effect, the idea that a small perturbation in one part of the world can cause a cascade of unpredictable effects on the other side of the world.

This type of butterfly phenomenon is referred to as *non-linearity*. Non-linearity means that effects feed back into themselves, becoming further magnified in complex

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

and unpredictable ways. Later on in this book we will explore some of the technical aspects of complex systems.

At the core of the challenge of causality is something more mundane, and that is the fact that we rarely witness a process unfolding in real time. We usually start studying something when it has already happened, whether it is how dinosaurs became extinct, how multi-cellular life emerged on earth, or how a virus spreads. And when we do happen to witness a process of interest unfolding, or perform experiments to make it happen, we may attempt to isolate the system of interest from other causes, but in fact we rarely succeed. In the real world, we must necessarily deal with incomplete data from limited observations, and confront systems interacting with other systems at all scales, forcing us to settle for educated guesses about causation rather than strive for complete certainty. This is why we are forced to use tools such as statistics to study probable causes of natural phenomena. Usually, systems interacting with each other appear to us as *noise* as they cannot be distinguished from the data of interest, thus impeding our understanding of events.

Science has managed to come up with partial solutions to some of these challenges. Scientists perform experiments to force events to happen in real time so as to watch them unfold firsthand. Researchers have also come up with tools that we like to group under different rubrics, such as logic, statistics, probability, and information theory, to mention a few. But all of these have the same purpose even if they approach the problem in different ways, namely, to empower scientists with suitable means to characterise the relationship between observations, causes, and effects. To the action of discovering the direct causes of natural phenomena, we give the name science, and all subfields of science, from data science and logic to astrophysics and biology, are devoted to this activity.

In the next section, we will discuss causality and models of the world, in particular mechanistic models.

1.2 What Is a Mechanistic Model?

Central to causality is the notion of a ‘source’ whence a process arises, a generating mechanism underlying an observation. We often talk about primordial causes as first principles. We also think in terms of what we call a mechanistic model, a model that can be followed from cause to effect step by step, as in an algorithm.

What has come to be known as a Rube Goldberg machine after its creator, a cartoonist, is the perfect illustration of a mechanistic model, because in this rather silly, convoluted mechanism, every process corresponds to an action and a consequence, all linked in a long chain of causes and effects (see Fig. 1.1). In this cartoon, a so-called self-operating napkin is activated by the action of a person using a spoon to eat soup. It is comical because the mechanism for activating the napkin comprises a series of risibly sophisticated tasks, only to achieve the rather mundane end result of moving the napkin.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

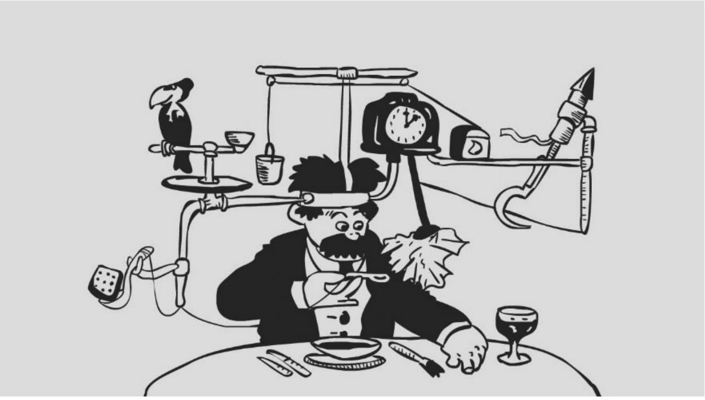


Figure 1.1 Self-operating napkin cartoon by Rube Goldberg illustrating a long and rather silly chain of intermediate cause-effect events with a trivial end result.

A Rube Goldberg machine, or Goldberg machine for short, illustrates another important aspect of the challenge of causality, particularly in relation to the area of modelling, which is that if we were not able to witness the processes leading to the movement of the napkin, we would not be able to rule out a Goldberg machine as the generating mechanism. The funniest thing about a Goldberg machine is that it is very transparent – we see its inner workings. But if we were only looking at the napkin, we wouldn't know whether it is the hand of a person that moves the napkin or a series of comically sophisticated actions. In other words, a Goldberg machine also illustrates that it is generally impossible to know the underlying cause of an event.

While we would like to say that a Rube Goldberg machine is among the least likely explanations for what happens in the world, we truly do not know, and we have to find arguments in favour of or against such explanations. Perhaps what strikes us as funny in this cartoon is not that far from what actually happens inside the human mind, for example, or at a molecular level, even if it does not involve gears, pulleys, and parrots. Clearly, a causal model like this one is far from random because the mechanism constrains the degrees of freedom of movement so that movement can only happen in a constrained space and can only be activated by a previously activated movement. The more control from original causes, the less random the effect when movement is limited in space and controlled by a prior cause.

For many centuries a geocentric model of the universe reigned and was based on what we call epicycles (Fig. 1.2). This model was used to explain the movement of the stars and planets in the sky, with the earth occupying the centre of the universe. Planets traced an additional movement around a small circle called an epicycle, which, in turn, moved along a larger circle called a deferent. Both circles rotated clockwise and were roughly parallel to the plane of the sun's orbit. This explained the apparent retrograde motion of the five planets known at the time, but one can see how complex the epicycles model was.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

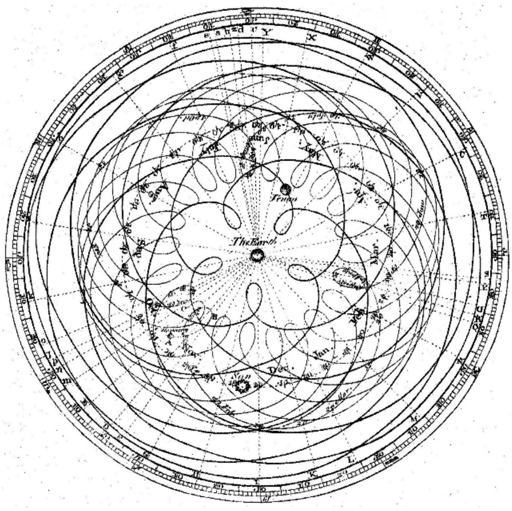


Figure 1.2 Intricate planetary motions as observed in the night sky leading to Ptolemy's epicycles model.

Even though the epicycles model of the universe was, for most purposes, incorrect, it was able to explain and describe, with virtually the same accuracy as heliocentric models, the movements of objects in the sky (see Fig. 1.3).

One of the most difficult movements to explain was that of the interior planets (closer to the sun than the earth), because these planets appeared to move in one direction and then suddenly switched to the opposite direction depending on the relative positions of the planet and the earth in their revolution around the sun. This retrograde movement was difficult to explain without placing the earth outside the centre of the solar system, so the epicycles were needed if the heliocentric model was to be causal and mechanistic.

Moreover, both the epicycles and the later heliocentric models had about the same predictive power for certain features, meaning that it was possible to run the models into the future to simulate the movements of the sun, planets, and stars as they appeared in the sky and make predictions (see Fig. 1.4). For example, they explained changes in the apparent distances of the planets from the earth.

So both the geocentric and heliocentric interpretations involved mechanistic models, with the epicycles being a concomitant of the mechanistic geocentric model and Kepler's laws of planetary motion, for instance, being associated with the mechanistic heliocentric model of Copernicus. Clearly, the effectiveness of both the geocentric and heliocentric models are an indication that generating mechanisms and predictions are not necessarily related. Nevertheless, mechanistic models are very important because they are not passive descriptions of an active process; they can actually be built, run, and followed step by step. This is why they are called mechanistic.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

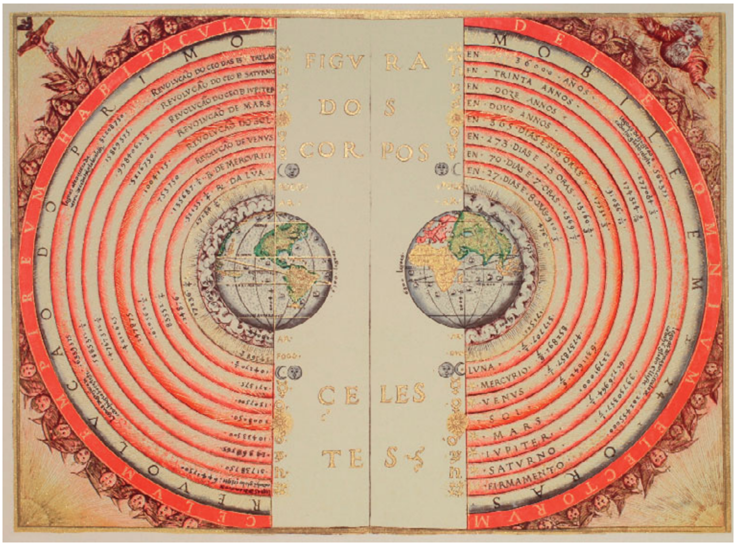


Figure 1.3 Heavenly bodies – An illustration of the Ptolemaic geocentric system by Portuguese cosmographer and cartographer Bartolomeu Velho, 1568 (Bibliothèque Nationale, Paris).

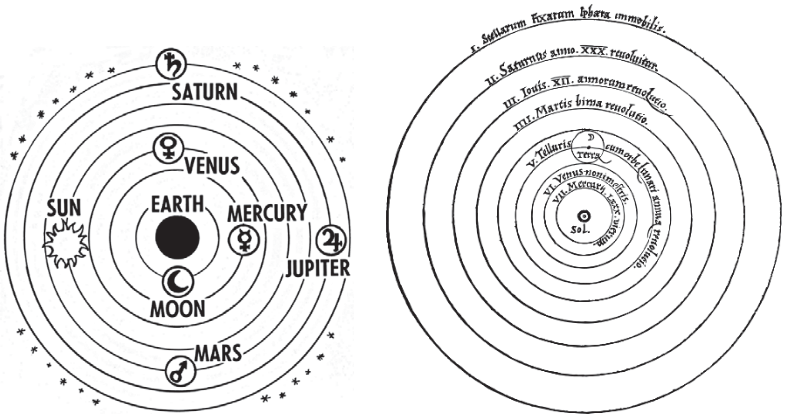

SATURN
VENUS
SUN
EARTH
MERCURY
JUPITER
MOON
MARS
I. Stellarum fixarum ſphæra immobilis.
II. Saturnus anno. XXX. revoluitur.
III. Iovis. XII. annorum reuolutio.
IIII. Martis bima reuolutio.
V. Telluris cum orbe lunari annua reuolutio.
Terra
VI. Venus nono menſe
VII. Mercurij. LXXX. dierum.
Sol.

Figure 1.4 An illustration of the (left) geocentric and (right) heliocentric models. Source: Archives of Pearson Scott Foresman, donated to the Wikimedia Foundation.

A mechanistic model may reveal more information by virtue of being mechanistic as opposed to being, say, merely descriptive. For instance, in the heliocentric model, if the sun's cycle requires 2 gears 20 times larger than the gear for the earth's moon, the mechanical model would suggest that the diameter of the sun is 2 multiplied by 20 times the diameter of the moon and would also shed light on the moon's distance from the sun, especially as observed during solar eclipses, when the moon and sun appear to be about the same size. Geocentric models would have required more convoluted explanations to reveal just the right distances at all times. And this makes mechanistic models more important than descriptive models.

1.3 The First Approach: Logic and Symbolic Reasoning

When humans started thinking about the world, they could not explain most of it and they tended to attribute everything to magic or the will of capricious gods. This is exactly the kind of explanation that is not mechanistic. Non-mechanistic models can come in many forms. Many cultures have attached importance to astronomical events, connecting them to human and terrestrial affairs, and to a certain extent we continue to do so. Every time they face something inexplicable some people are still apt to attribute it to extraterrestrial intelligences or paranormal phenomena. But the first record of progress toward finding tools to separate fact from myth and understand the natural world is to be found in Greek philosophy, in the development of a form of formal logic in arithmetic and geometry that remains the basis of modern science.

Aristotle was the first to deal with the principles of formal logic in a systematic way. The history of logic is the history of valid inference. In fact ‘logos’ means ‘reason’ in Greek, among other things. Logic as a formal way of reasoning was reinvigorated in the mid-nineteenth century, at the beginning of a revolutionary period when the subject developed into a rigorous and formal discipline that took as its paradigm the method of proof used in Greek mathematics.

This copy of Euclid’s Elements (Fig. 1.5) was published in 1573 but it was written around 300 BC, long before the Bible itself. The concept of mathematical proof reached maturity at this point, and was systematically used thereafter.

The early thirteenth century witnessed a recovery of Greek philosophy after its eclipse in the Middle Ages. And to return to mechanistic models, a philosopher named William of Occam was exercised by the problem of finding arguments to rule out

E·VCLIDIS
ELEMENTORVM
LIBRI XV. GRAEcè & Latinè,

Quibus, cùm ad omnem Mathematicæ ſcientiæ partem, tùm ad quamlibet Geometriæ tractationem, facilis comparatur aditus.

Ἐπίγραμμα παλαιόν.
Σχήματα πέντε Πλάτωνος, ἃ Πυθαγόρας σοφὸς εὗρε.
Πυθαγόρας σοφὸς εὗρε, Πλάτων δ' ἀρίδηλ' ἐδίδαξεν,
Εὐκλείδης ἐπὶ τοῖσι κλέος περικαλλὲς ἔτευξεν.

IN ME MORS, IN ME VITA.

PARISIIS,
Apud Hieronymum de Marnef, & Gulielmum Cauellat, ſub Pellicano, monte D. Hilarij.
1573.

Figure 1.5 Euclidis – Elementorum libri XV Paris, Hieronymum de Marnef & Guillaume Cavelat, 1573 (second edition after the 1557 ed.); in-8, 350, (2)pp. THOMAS-STANFORD, Early Editions of Euclid’s Elements, no. 32. Mentioned in T. L. Heath’s translation. Private collection H. Zenil.

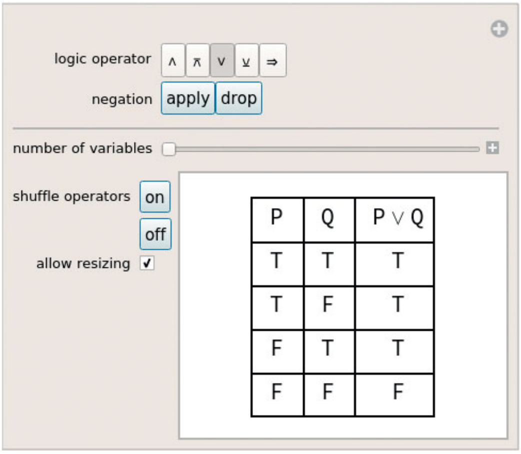
https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

explanations along the lines of a Rube Goldberg machine, that is, possible but overly complicated explanations for even the simplest phenomena.

Today we know his argument as ‘Occam’s razor’, also known as the ‘law of parsimony’. Occam’s razor establishes that when presented with competing models, one should select the one that makes the fewest assumptions. So, in the case of the self-activated napkin drawn by Rube Goldberg, we would simply discard all the mechanisms involved if we had not seen them in operation firsthand, and we would instead favour simpler models as possible explanations. For example, that the person themselves or someone else was manipulating the napkin and not an overly complicated machine.

In science, Occam’s razor is used as a guiding principle in the development and selection of theoretical models, but we will also see that there is good evidence in favour of Occam’s razor. One of my own contributions to the topic is to show that Occam’s razor is not only formalised by a concept central to this book, the concept of algorithmic complexity, but also that numerical approximations to algorithmic complexity provide strong evidence in its favour. We will see this later, formally and in detail; we will see how algorithmic complexity is connected with the kind of simplicity advocated by Occam.

The development of modern ‘symbolic’ and ‘mathematical’ logic is owed to authors such as George Boole, Gottlob Frege, Bertrand Russell, Giuseppe Peano, David Hilbert, and Kurt Gödel. One of the most basic kinds of mathematical logic is that of propositions that can be assigned a truth value, either true or false (Fig. 1.6).



P	Q	P ∨ Q
T	T	T
T	F	T
F	T	T
F	F	F

Figure 1.6 Charles Sanders Peirce appears to be one of the earliest logicians (1893) to devise a truth table matrix. While truth tables are relatively modern, their underlying principles are similar to those of the Aristotelian syllogism (a subset of the current Boolean operators), which Frege, Boole, and others would much later build upon to formalise what is today known as first-order predicate logic. A program designed to illustrate the concept is available at: http://demonstrations.wolfram.com/TruthTables/.

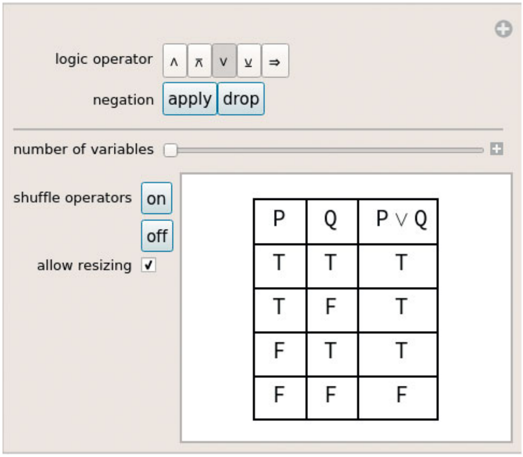
This type of logic is equivalent to a type of algebra known as Boolean logic, after George Boole. Boolean logic turned out to be especially important for computer science because it fits nicely with the binary system, in which each bit has a value of either 1 or 0, comparable to True or False. Boolean logic will be important later on in the book as we will be using a type of system called a Boolean network. Boolean networks are networks with propositional formulae, such as the ones on a truth table, that process information according to the way in which the network is connected.

In the truth table shown in Fig. 1.6, for example, there is a formula meaning P and Q and written P AND Q, where P and Q are propositions that can be False or True, represented by ‘F’ or ‘T’, which can also simply be 0 and 1. The connective ‘AND’ is called a Boolean operator and assigns a value to two propositions, telling us whether they can be true or false together. For example, for the operator AND, only when the two propositions P and Q are True can the final formula be True.

Say the proposition P tells us that ‘the lights are on’ while proposition Q tells us that ‘there are empty seats’. Both statements may be True or False, but if we say that ‘The lights are on AND there are empty seats’ such an assertion can only be True if both P and Q are true at the same time. This is different from the Boolean operator OR, where if we claim that ‘The lights are on OR there are empty seats’ then the proposition denoted by P OR Q can be True if either of the two propositions is True, that is, even if one is False, for example, if the lights are off but there are empty seats (see Fig. 1.7).

These kinds of formulae can be arbitrarily long and so they become less trivial and can implement more complicated structures when combined as, for example, is shown in Fig. 1.8.

It can now be clearly seen how this simple logic can be deeply connected to causality. For example, the formula P OR Q can only be True if either P, Q, or both are True, and thus the observation that P OR Q is True or False tells us something about the truth



logic operator ∧ ⊼ ∨ ⊻ ⇒

negation apply drop

number of variables

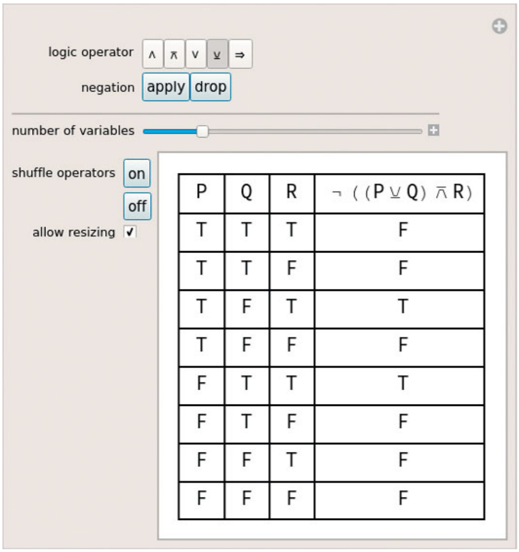
shuffle operators on off

allow resizing ☑

P	Q	P ∨ Q
T	T	T
T	F	T
F	T	T
F	F	F

Figure 1.7 The truth table for the Boolean operator OR. Source: http://demonstrations.wolfram.com/TruthTables/.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press



allow resizing ☑

P	Q	R	$\neg((P \veebar Q) \barwedge R)$
T	T	T	F
T	T	F	F
T	F	T	T
T	F	F	F
F	T	T	T
F	T	F	F
F	F	T	F
F	F	F	F

Figure 1.8 The truth table for a formula with combined logic operators. Source: http://demonstrations.wolfram.com/TruthTables/.

value of the causes, that is, the truth values of P AND Q individually when the truth value of the formula is True or False.

However, in general, mathematics and logic are more interested in truth, which is only indirectly connected to causality, because a proof is not a generating mechanism as it cannot be constructive. That is, it is impossible to write a program to effectively carry out the details of a mathematical proof, as in the case of Peano's famous diagonalisation methods. But in the twentieth century, truth and proof in logic were uncoupled in a very fundamental way, and this somehow broke the direct connection to causality.

Nevertheless, as we will see in this book when covering aspects of the theory of computation, this rupture was essential to understanding some fundamental properties of the ways in which we can achieve truth and discover causation by means of pure reasoning, logic, mechanism, and algorithm. Indeed, that logic and proof were uncoupled did not mean that we regressed to magic and astrology to explain complex phenomena. On the contrary, it led to new methods of calculation and to the way in which we approach science today – through the lens of computation. And even more importantly, it led to the concept of universal computation, which we will explore in detail in the following chapters. But it is important to understand that causality is very difficult to calculate, for the same reasons that led to the rupture of the connection between truth and proof by formal logic.

1.4 The Second Approach: Probability and Statistics

Following what we think of as a first revolution in the formal approach to the problem of causation, the advent of logical inference, there was a second revolution, so to speak,

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

Figure 1.9 The current statistical approach to science assumes that processes are mostly random, so when a deviation, such as a black swan, is observed in a long sequence of other uninteresting events (white swans), it generates great surprise. However, if one assumes that a sequence of events may not be randomly arranged and has been, for example, carefully sorted, then black swan events can be expected, provided the underlying mechanism allows for them or even promotes them.

consisting of an attempt to quantify the idea of chance, of things simply happening without a particular reason, as opposed to being produced or having a cause. There are all sorts of interesting epistemological challenges entailed in such an enterprise.

For example, science, and particularly classical mechanics, establishes that everything has a cause, and that there is no such thing as chance. This may suggest that something appearing to happen by ‘chance’ is precisely that, an appearance.

Italian and French mathematicians such as Cardano, Pascal, Fermat, and Borel were the first to make attempts to characterise chance in a mathematical way, and we owe them many of the concepts and ideas still in use today in probability and statistics. Further progress was made by German-speaking mathematicians such as Jacob Bernoulli and, much later, Richard von Mises, laying the foundations of modern probability theory.

But it was Andrey Kolmogorov who formulated the definitive version of what we know today as probability theory in its current formal mode, called an ‘axiomatisation’ in mathematics. Interestingly, both von Mises and Kolmogorov felt they had arrived at a limited and weak characterisation of randomness by means of probability distributions. However, while some crucial concepts necessary to an effective characterisation of randomness were unavailable to von Mises, Kolmogorov combined ideas from the rising field of computation to arrive at what we know today as the mathematical definition of randomness, which is at the centre of this book and the techniques introduced and studied here.

The aim of classical probability and traditional statistics is to help solve the problem of causal inference by calculating probability distributions. A key criticism of the way statistics approaches causality has to do with its dependence on assumptions and expectations based on probability distributions. Indeed, without knowing the generating cause of the appearance of, for example, five thousand white swans parading in single file, a statistician would approach the problem of guessing the colour of the next swan to appear by simply calculating a distribution based on how many times white swans have been spotted versus swans of other colours (Fig. 1.9).

Viewed through the lens of traditional statistics, the sudden appearance of a black swan in the long parade of white swans would be a great surprise and would be considered an oddity – and may in fact be so. However, if there was a swan owner releasing the swans in a specific order who had decided to release all the white ones first followed by the black ones, we would have made the erroneous assumption that swans were

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

appearing at random and thus that a black swan was an oddity when we saw one for he first time. In other words, using, or attempting to devise a model of a generating mechanism not based on an assumption of randomness has the advantage of leading to better explanations, and affords the means to make more accurate predictions. Of course it is one thing to attempt and another to achieve such an objective, but statistics is not designed to generate candidate models but rather to describe data probabilistically.

Its limitations may include confounding causes and effects, that is, either difficulty isolating a common cause in certain tangled situations, as happens frequently in science and is among the main considerations in designing experiments, or else a lack of clarity as regards what is cause and what effect. For example, a certain study may claim that children who watch a lot of TV are the most violent. Clearly, TV makes children more violent, it says. But this effect could easily be produced by some other cause. Perhaps it could be that violent children watch more TV than less violent ones because they are more likely to be reclusive.

In cases such as this, what one needs is to introduce what is called a 'control experiment' to sort out whether the children were already violent and became more addicted to TV as a result of their violent personalities. Another problem is that one cannot reset the same child as violent and then non-violent or vice versa, so the experiment has to be performed on an already existing population that will necessarily be influenced by other indirect causes that cannot be completely isolated. The purpose of control experiments, however, is to control the most obvious bias or confounding cases, so as to be able to draw more meaningful conclusions. Control experiments are always of the form 'what would have happened if something else had or had not happened, or what would happen if we do or do not apply or remove a certain other influence'.

Pierre-Simon Laplace was the first to use what are called uniform priors when faced with a complete lack of knowledge, that is, a distribution that assumes that all events are equally likely. He introduced a principle known as the 'principle of insufficient reason,' also known as the 'principle of indifference' (Fig. 1.10).

The 'principle of insufficient reason' is similar to Occam's razor in that it is a guiding principle with no strong evidence in favour or against. The principle states that if there are n possible causes indistinguishable except perhaps by their names, then possible causes should each be assigned a probability equal to $1/n$, that is, equal probability, and none should be discarded or ruled out.

While the principle of insufficient reason is a reasonable principle to follow, we will challenge some of its assumptions because while it may be desirable to assign all possible causes non-zero probability, as suggested by another principle, the Principle of Multiple Explanations, which establishes that if several theories are consistent with the observed data we should retain them all, there are strong reasons to assign different rather than equal probabilities to different explanations. In fact, the 'principle of indifference' seems to contradict Occam's razor, which suggests that we not assign equal probability to overly complicated causes unless there is a good reason for doing so.

All these assumptions made in traditional statistics reveal that there is a highly subjective component in the way that classical probability deals with causality, particularly in the absence of data and knowledge about the generating source, which is pretty much

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

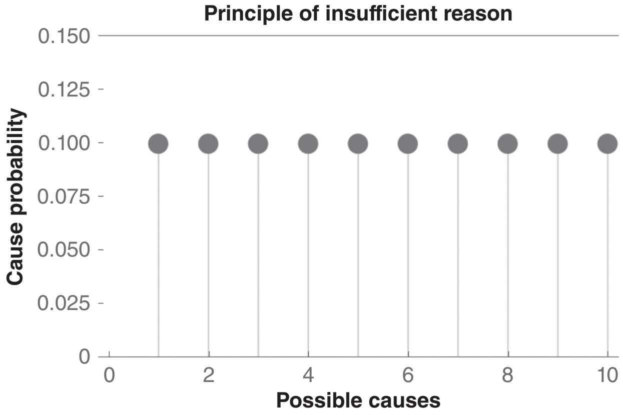


Figure 1.10 Principle of insufficient reason: when no other information is consistent with the data, or the data is not observed, this principle suggests that we assume a distribution where all events are equally possible, so if 10 events are possible causes of an effect, all them can, independently, be responsible for the said effect with probability 1/10.

the general case. Nonetheless, all these methods are widely accepted and used in the field of machine learning, for instance.

We will challenge the use of uniform priors, that is, the use of uniform distributions as a first assumption when undertaking the study of causal systems, as opposed to random systems, also called stochastic systems. We have shown that challenging the common use of these uniform priors affords interesting insights – such as the acceleration in the convergence of biological evolution – in comparison to assuming uniform random mutations, something that we will cover in the last chapter, which will return to all this in greater detail.

The idea that 'probability' should be interpreted as the 'subjective degree of belief in a proposition' was proposed by John Maynard Keynes in the early 1920s, but even today, methods such as Shannon entropy are taken far more seriously than they should be. Entropy, in the sense in which Claude Shannon introduced it, is usually presented as a measure of surprise in the context of communication.

As we will see, Shannon entropy is a measure of degree of uncertainty, which reflects one's own lack of knowledge rather than any objective indeterminacy in phenomena. So, contrary to what's generally claimed, we will demonstrate that Shannon entropy is not a syntactic measure at all, but a highly semantic one – though this does not mean that it is necessarily better or worse. Shannon entropy is interesting because it introduced logic and computation as descriptions and operations of information. We will explain in greater detail later why some of these ways to characterise randomness are often, if not always, very fragile. We will illustrate this with examples, but it is nevertheless important that you have a grasp of these concepts.

A fundamental concept in statistics is that of correlation, not a minor concept but actually the heart of statistics in a sense. Statistics is all about finding statistical patterns in the form of regularity. As we will see, anything more sophisticated than that will be missed by statistical approaches to inferring causal mechanisms in data.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

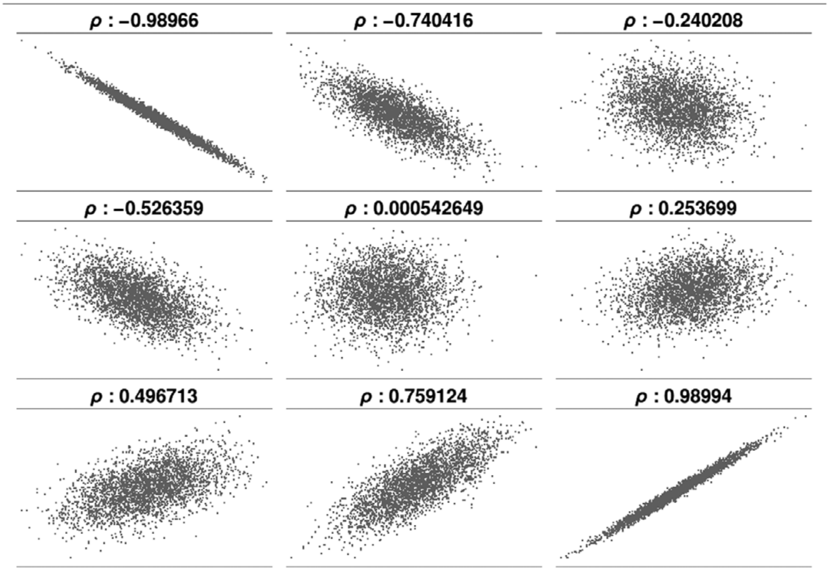


Figure 1.11 Correlation values for different scatter plots. Source: Justin Matejka and George Fitzmaurice, ACM CHI 2017. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. (www.autodesk.com/research/publications/same-stats-different-graphs).

A statistical regularity can be, for example, the tendency of some data points to lie on a plane, or of a time series to display periods. These are typical plots of positive and negative correlation. Think of two processes from which we obtain data, wishing to ascertain whether they correlate and/or are causally connected by correlation, say, a time series. A time series is a collection of data points sorted by time. Think of the x-axis as recording the values of one time series and the y-axis those of the other one. Then one can test whether the data points get aligned, meaning that they distribute in a similar fashion.

Correlation test values used to measure the strength of association between variables, usually denoted by the Greek letter ρ (rho), are typically given between -1 and 1; when ρ is close to 1 or -1 the data is positively or negatively correlated, respectively (Fig. 1.11).

The plots in Fig. 1.11 are called scatter plots. There are several ways to measure correlation, but they are all very similar and consist of measuring distances among data points. One of the most popular measures is called Pearson's correlation, which gauges the correlation among data point values. Another popular measure is Kendall's or Spearman's correlation. These measures rank correlation when only the order matters.

Because of the limitations mentioned above, traditional statistics often leads, with high probability, to spurious models from false negatives and false positives. A false positive or false negative is a regularity in the data that appears real but is only an artefact giving the wrong impression regarding its cause or effect. Figure 1.12 shows examples of false negatives, meaning that the correlation test, quantified by rho, suggests that there is no correlation between axis x and y. But if we look at the plots themselves, we

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

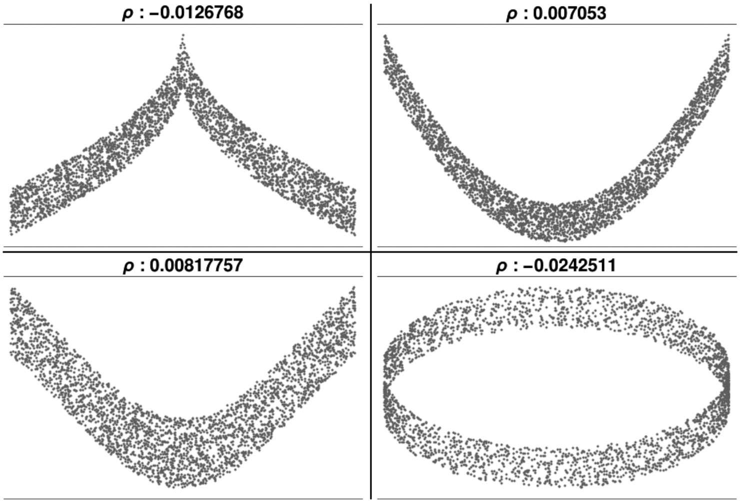


Figure 1.12 False negatives. Source: Justin Matejka and George Fitzmaurice, ACM CHI 2017. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. (www.autodesk.com/research/publications/same-stats-different-graphs).

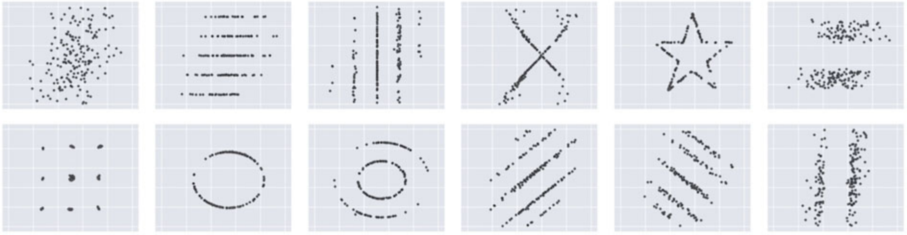


Figure 1.13 Scatter plots with equal correlation values but different structures. Source: Justin Matejka and George Fitzmaurice, ACM CHI 2017. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. (www.autodesk.com/research/publications/same-stats-different-graphs).

see structure immediately, suggesting that something interesting is happening in the way the data points are distributed along the axes. However, rho is almost 0 in these four plots, suggesting that, on the contrary, nothing interesting is happening.

These are other examples where not only is the correlation 0 but all plots have exactly the same rho value, despite there clearly being lots of different structures (see Fig. 1.13).

The Stanford Encyclopedia of Philosophy refers to attempts to analyse causation in terms of statistical patterns as ‘regularity theories’ of causation [2]. Statistical regularities are only a subset of the possible properties that a phenomenon may display. A statistical approach offers an explanation for the distribution of data but leaves to the scientist the arduous task of formulating an interpretation in order to come up with a model underlying the data. Traditionally, what a scientist does is to fit a curve. Then the equation of the curve is taken as the generating model, both to explain the distribution of data and to make predictions. In the typical case of positive correlation, for example,

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

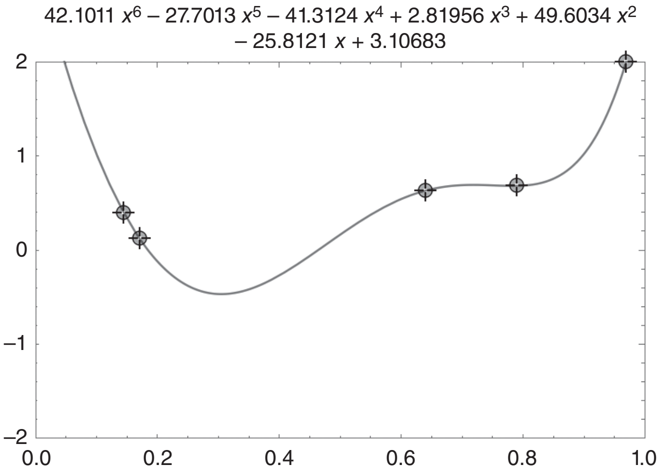


Figure 1.14 Polynomial fit. One can always find a polynomial of degree equal to or greater than the number of data points that fits the data points perfectly, but most likely this is only necessary when such data points have no apparent common cause and are algorithmically random. When at least one data point can be explained by a combination of the others, then the polynomial equal to or greater than the number of original data points is most certainly over-fitting the data.

it is not difficult to fit a line. This is called a linear fit because the function is linear as fitted by a line.

However, one can always force a curve to pass through any number of data points using a polynomial of degree proportional to the number of points. One can see how, by increasing the degree of the polynomial, one can make the curve go through or very close to the data points (see Fig. 1.14).

All these limitations of traditional statistical approaches to causality can be summarised in what is one of the most common adages in the field: ‘association is not causation’, or ‘correlation does not imply causation’, though lack of correlation does not mean lack of causation. In other words, that you can fit a curve to a set of data points does not necessarily mean that the curve actually has anything to do with said data points.

In [3], the degree to which regression and correlation can be misleading was made clear. All the scatter plots in Fig. 1.15 have identical values, that is, mean, standard deviation, and Pearson’s correlation to 2 decimal places (x mean = 54.26, y mean = 47.83, x SD = 16.76, y SD = 26.93, Pearson’s $R = -0.06$), yet clearly they were generated in different ways. They started from some points scattered randomly with little displacement then using an annealing technique they pushed every dot to a target image (a dinosaur), keeping all summary statistics the same. Clearly, once the data points reach the dinosaur a lot of computation has been invested and the result is a long chain of highly directed cause and effect, as opposed to more random configurations. Yet all these forms are very different and represent shapes that do not imply

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

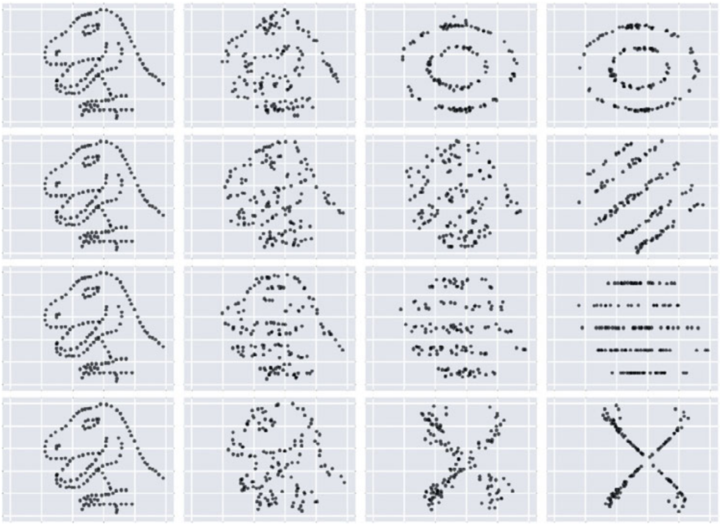


Figure 1.15 Deceptive stats, graphs with the same statistical correlation values forming all sorts of shapes, from a dinosaur to various other shapes that tools like correlation are blind to. Credit: Justin Matejka and George Fitzmaurice, ACM CHI 2017, Source: www.autodeskresearch.com/publications/samestats.

randomness. Some versions of traditional statistics, including Shannon entropy, may distinguish certain trivial cases, but will fail in many ways, as illustrated in these plots. What is missing is a method other than statistics that may be used to infer the underlying algorithmic probability and linked chains of cause and effect. In the chapters to follow we will propose that some of this can be tackled using computation and algorithmic probability.

Another example that can serve to illustrate the limitations of statistical analysis can be found in Fig. 1.16. According to the statistical description of the phenomenon, car arrivals at any point on an open road follow what is known as a Poisson distribution. This is because distributions can have characteristic shapes when they are plotted. However, what the statistical approach describes is the effect of the mechanistic cause and not the generating mechanism. It may provide clues to the causes that are then left to the scientist to interpret, but the statistics in and of themselves do not provide a model.

The reason why a Poisson distribution is produced is that slower drivers cause faster drivers to cluster in their wake, producing knots of cars running together on the highways and arriving at gas stations at about the same time. But it is not the Poisson distribution that causes the cars to cluster, nor does it suggest how or why this happens. In contrast, a mechanistic approach attempts to provide a causal model that may help design ways to manipulate the effects, as it points out the exact submechanisms that can be changed to achieve a different end. Probability and statistics have led a revolution in the study of causality, but they have, in some fundamental way, exhausted their potential

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

Figure 1.16 Times between cars in common traffic on a road can be described with statistics and an associated probability distribution (Poisson), but the cause can only be suggested or advanced externally. Taking a different approach, one can look for causes directly. In this case, the cause is that slower cars lead to clustering. Source: JArrevillaga/Wikimedia, licensed under the CC BY-SA 3.0 licence.

to engender further progress in modern science. This book is all about trying to provide an alternative and complement to traditional statistics and classical probability. We will see how what we call algorithmic information dynamics (AID) provides interesting tools to help reveal and deal with mechanistic causes.

It is a common observation that some events are less predictable than others. When are we justified in regarding a given phenomenon as random?

The most common approach in classical probability theory is to frame events within a set-theoretic framework. Events have an outcome that belongs to a set of possible outcomes. Traditionally, events are assumed to be repeatable, and such a repetition is called a trial. Outcomes are direct observables. Let us denote the probability P of an event A_n as $\mathrm{P}(A_n)$. For example, in the roll of a die, events $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$ are elements of a space of possible outcomes of the experiment. Two events A and B are considered mutually exclusive, or disjoint, if the occurrence of A rules out the occurrence of B. In the dice example, if a roll turns up ‘three dots’, this event rules out the event ‘six dots’.

1.5 The Third Approach: Perturbation Analysis and Machine Learning

In the previous sections we saw how much progress had been made by formalising the concepts of causality and chance with mathematics, in particular with logic,

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

probability, and statistics. Another revolution brings these ideas together and takes probability and statistics to their limits.

How informative an experiment may be in determining the strength of a cause depends on a number of factors, such as whether there are control experiments to discard or rule out certain causes, as we have seen before, but also how good the measurements are, how well placed the observer that makes them, both at the input and output levels of a system, and how distinguishable individual events are.

We have seen how correlation is central to statistics, but also how limited it can be. One way to make the most of statistics is by performing systematic perturbations according to a probabilistic calculus introduced by Judea Pearl and colleagues, with a view to finding possible causes that can easily be ruled out upon further inspection. These ideas are at the forefront of the practice of probabilistic causal discovery. Perturbation analysis allows us to update beliefs under conditions of uncertainty based on previous knowledge derived from, e.g., performing perturbations or observing a stream of data. It also accommodates the concept of a control experiment. For example, if skin cancer is related to sun exposure and failure to wear sunscreen, then, using Bayes' theorem, a person's exposure to sunlight and failure to wear sunscreen can be used to assess the probability that they will develop skin cancer more accurately than would be possible if one knew nothing about their degree of exposure.

Central to this kind of calculation is the work of Thomas Bayes (circa 1701–1761). In its simplest form, Bayes' theorem establishes a relationship between the probabilities of two events A and B, $\mathrm{P}(A)$ and $\mathrm{P}(B) \neq 0$, and the conditional probabilities $\mathrm{P}(A|B)$ and $\mathrm{P}(B|A)$. Bayes' theorem establishes that:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}. \tag{1.1}$$

Bayes' theorem as a rule provides some kind of 'backwards probability' where $\mathrm{P}(A)$ is traditionally called the *prior* (or initial degree of belief in A) and $\mathrm{P}(A|B)$ is the constructed *posterior*.

As in the case of the experiments with time series mentioned in previous sections, we can ask whether two time series resulting from different observations are causally connected to each other (see Fig. 1.17). Now, traditional statistics would typically suggest that the behaviour of two time series, let us call them X and Z, are causally connected because they are correlated, but there are several other cases that are not decidable following a simple correlation test.

A first possibility is that the time series simply show similar behaviour without being at all causally connected, despite the possible positive result of a correlation test. Another possibility is that they are causally connected, but correlation does not tell us whether it is a case of X affecting Z, or vice versa. Yet another possibility is that the two have some common cause upstream in their causal paths and that X and Z are therefore not directly causally connected but connected through a third cause Y that is concealed from the observer. So how are we to test all or some of these hypotheses?

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

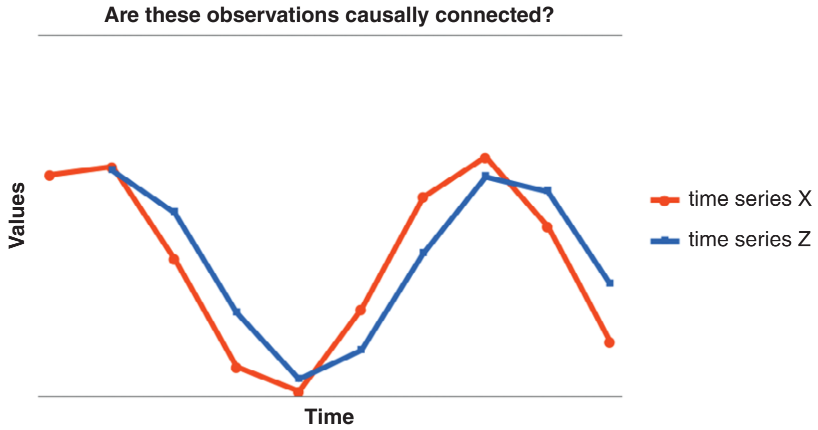


Figure 1.17 Time series as an example of a sequence of observations to be tested for a causal relationship.

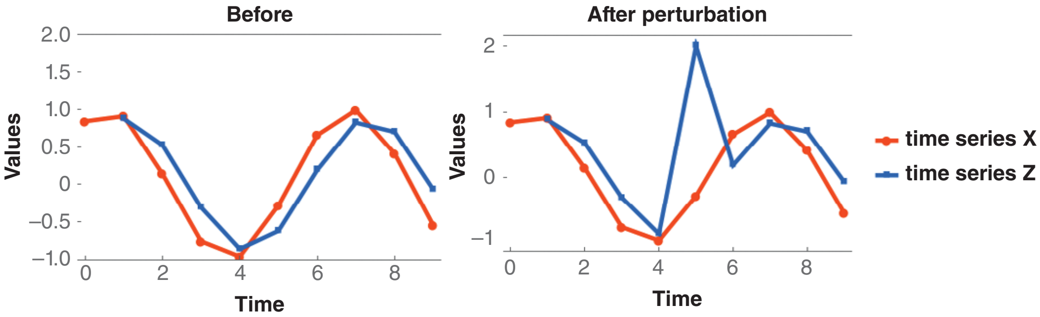


Figure 1.18 Time series X and Z before and after perturbation.

One way to do so is to perform a perturbation on one time series and see how the perturbation spreads to the other time series. Let us perturb a data point in the time series Z; let us say that we multiply by -2 the data point in position 5. We can see that nothing has happened to the values of time series X; it looks exactly the same as before (see Fig. 1.18).

So if we perturb the values in the time series Z, at least for this data point we can see that X remains the same. This suggests that there is no causal influence of Z on X.

However, if the perturbation is applied to a value of X, Z changes and follows the direction of the new value, suggesting that the perturbation of X has a causal influence on Z (see Fig. 1.19). From behind the scenes, we can reveal that Z is the moving average of X, which means that each value of Z takes two values of X to calculate, and so is a function of X. The results of these perturbations produce evidence in favour of a causal relationship between these processes if we did not know that they were related by the function we just described.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

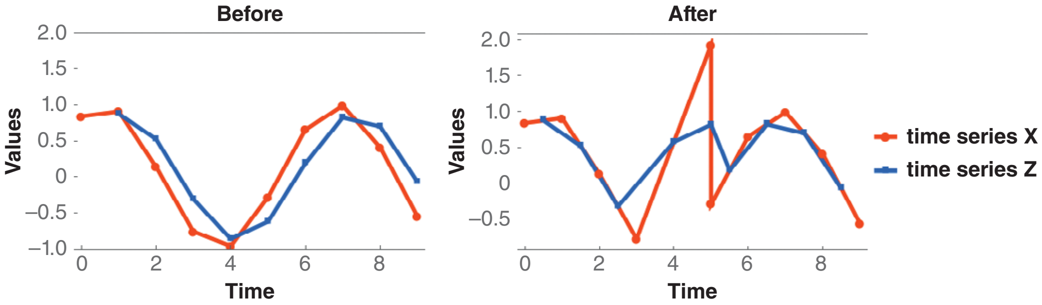


Figure 1.19 Time series X and Z before and after another perturbation.

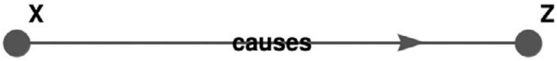


Figure 1.20 Causal relationship between series X and Z.

This suggests that it is X which causally precedes Z. So we can say that this single perturbation suggests the causal relationship shown in Fig. 1.20.

It is important to consider that performing interventions is not always simple or possible in reality, but their simulation or emulation can offer great insights. This kind of simulation of ‘impossible’ situations is what Judea Pearl identifies a ‘counterfactual’, i.e., the opposite of a possible fact. Simulation and emulation will be at the heart of our AID approach. Think of drugs as a way to perturb a biological system. Some drugs may have adverse consequences, so interventions in the way of prescriptions are highly regulated and one cannot easily do experiments relating to diseases on humans. But assuming that one may perform such interventions, in cases where an intervention does not lead to a change, it produces evidence against a causal connection between the events.

Moreover, the underlying argument is that after an intervention, correlation may or may not occur, and thus one still relies on classical statistics to make the final calls if an inference engine that uses probability theory, regression, and correlation is not replaced with something that does not. So while perturbation analysis does help to rule out some cases, it still inherits the pitfalls of statistics and correlation analysis, for the simple reason that we have not yet changed the tools; we have merely performed more experiments on the data. And this is what AID is about, taking probability and statistics out of the core of candidate models, even if these models must still rely on some statistics or on a probability distribution – which we will never be completely rid of.

As mentioned above, there are three possible types of causal relationships that can be represented in what is known as a directed acyclic graph, that is, a graph that has arrows implying a cause and effect relationship but has no loops, because a loop would make a cause into the cause of itself, or an effect into its own cause, something that is not allowed because it would be incommensurate with causality (see

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

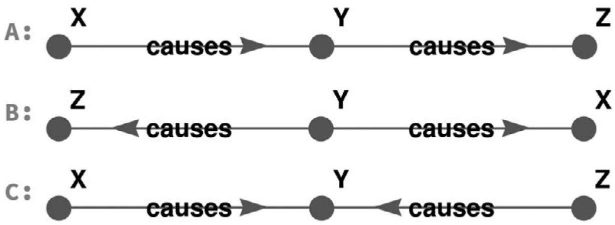


Figure 1.21 Direct acyclic graphs of all possible causal relationships among three unlabelled variables with no self-loops, i.e., no effect is the cause of itself.

Fig. 1.21). We will cover all this graph jargon in the next chapter. In these graphs, nodes are events and events are linked to each other if there is a direct cause-and-effect relationship.

In the first case, labelled A in orange, the event X is the cause of event Y, and Y is the cause of event Z, and so X is said to be an indirect cause of Z. In general, we are of course always more interested in direct causes, because almost anything can be an indirect cause of anything else. In the second case, B, an event Y is a direct cause of both Z and X. Finally, in case C, the event Y has 2 causes, X and Z. With an interventionist calculus such as the one performed on the time series above, one may rule out some cases but not all of them, but more importantly, the perturbation analysis offers the means to start constructing a model explaining the system and data rather than merely describing it in terms of simpler correlations.

In our approach to causality, we incorporate the ideas of an interventionist calculus and perturbation analysis in AID, and we replace traditional probability theory and classical statistics and correlation by a fully model-driven approach that is fed by data but is not merely data-driven. The idea is to systematically produce the most likely generating models that explain the observed behaviour. In the case of our two time series experiments, the time series X is produced by the mathematical function $f(x) = \sin(x)$, and thus $\sin(x)$ is the generating mechanism of time series X. On the other hand, the generating mechanism of Z is MovAvg($f(x)$), and clearly MovAvg($f(x)$) depends on $f(x)$ which is $\sin(x)$, but $\sin(X)$ does not depend on MovAvg($f(x)$), so how could we have guessed these two functions that statistics alone cannot systematically find? In other words, we need some sort of method to infer the function or algorithm behind the data. Having found MovAvg($f(x)$) and then $f(x) = \sin(x)$, we could not only establish the actual causal relationship, we could also produce the observation and any number of other data points in the future with perfect accuracy.

Some of this perturbation analysis and the graphical approach to causality was based on what are called Bayesian networks, involving the calculation of conditional probability distributions with the additional feature of asking ‘what happens with some variable representing a possible event if this other variable representing another event is observed, perturbed, or simulated’. These ideas on an interventionist calculus are fundamental to causality and are incorporated into AID.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

1.5.1 Perturbation Analysis and Interventionist Calculus

Judea Pearl's main contribution was to come up with the idea of perturbations, counterfactuals, a calculus, and graphical representation of cause and effect. We built upon Pearl's idea of perturbing systems. But as we have established, Pearl could not move away completely from traditional statistics and probability theory because he had to resort to the very tools he was trying to leave behind to compare the original and the perturbed or simulated system. This means that anyone using Pearl's do-calculus would be forced to resort to regression or correlation in the absence of better tools. In AID we take the next step toward removing probability from candidate causal models. Once a system is perturbed or simulated, one uses AID to compare the result to the original system, that is, one examines the differences in the candidate models explaining the original and the new model. AID incorporates perturbations and counterfactuals naturally. In AID, we apply all possible perturbations to a system, including those that can be seen as counterfactuals, and then we look at how the sets of candidate models change. In contrast, perturbing a system or estimating a counterfactual effect in Pearl's calculus in practice involves comparing distributions, which means that we fall into the same regression/correlation trap.

Candidate models in AID are computable models that explain both the original data and the new (perturbed or simulated) data. Then one can see how much the underlying model has changed as a result of the perturbation and decide which changes impact the dynamics of the original system, something that Pearl's calculus is unable to do without an inference engine that does not offer the do-calculus, or at least not without having to resort to probability distributions, which require traditional statistics.

There is a relationship between causal graphs, entropy, and algorithmic complexity that we explore throughout the book. Algorithmic complexity is a generalisation of entropy and has absolutely no problem dealing with Pearl's causal graphs; they are simply not necessary because we are no longer dealing with a language based on probability distributions (which are usually difficult if not impossible to access in the best case).

Solomonoff was aware that the theory of algorithmic probability, on which AID is heavily reliant, was the ultimate optimal theory of inference. The AI community also agreed that Solomonoff's theory was the answer to causal inference, but because it was uncomputable and there was not enough computational power back in the 60s – and indeed until recently – it was difficult to really explore the field numerically, and most researchers simply turned their backs on it. It is easy to see how Marvin Minsky could say that the way forward, possibly the only or the most important way forward, was algorithmic probability. He was speaking of AI at a panel discussion on The Limits of Understanding at the World Science Festival, NYC, on 14 December 2014, just one year before he passed away. Marvin Minsky is widely considered to be the founding father of artificial intelligence. His astonishing claim describes what turns out to be precisely the objective of our research programme and the main purpose of this book. To quote his closing statement to the panel:

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

> It seems to me that the most important discovery since Gödel was the discovery by Chaitin, Solomonoff, and Kolmogorov of the concept called Algorithmic Probability, which is a fundamental new theory of how to make predictions given a collection of experiences and this is a beautiful theory, everybody should learn it, but it's got one problem, that is, that you cannot actually calculate what this theory predicts because it is too hard, it requires an infinite amount of work. However, it should be possible to make practical approximations to the Chaitin, Kolmogorov, Solomonoff theory that would make better predictions than anything we have today. Everybody should learn all about that and spend the rest of their lives working on it.

1.6 The Fourth Approach: Model-Driven Inference, Dynamical Systems, and Computation

1.6.1 Introducing Computation in Causal Analysis

Causality and computation have been linked since the inception by way of the concept of calculation. Figure 1.22 shows an ancient Greek calculator used to predict celestial events such as eclipses. It is known as the Antikythera Mechanism after the island on which it was found. It is the first known 'analogue computer', believed to have been made circa 87 BC, that is, more than 2,000 years ago. Possibly built by Archimedes, this device is very similar to the mechanical artifact shown above that simulate the planetary movements of the solar system. These cogwheels would turn in a precise fashion to make a prediction based on the epicyles model that we encountered above. Figure 1.23 shows a computer program designed to showcase the way in which the mechanism worked and what it was able to predict.

Not everything is known about this ancient device. For example, it is not known what additional components were needed to simulate and display planetary motions. Turning the handle causes a sequence of interlocking gears to rotate, moving the Sun

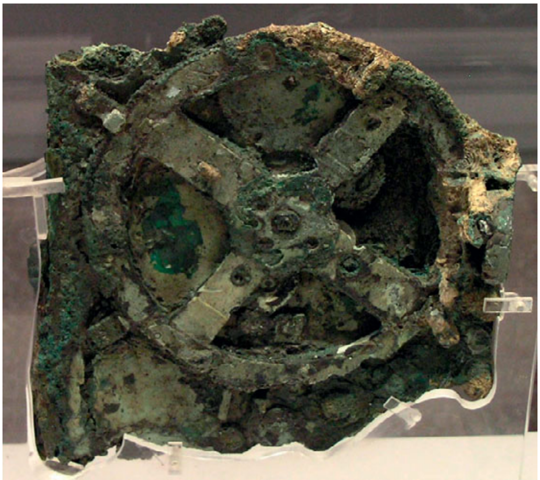
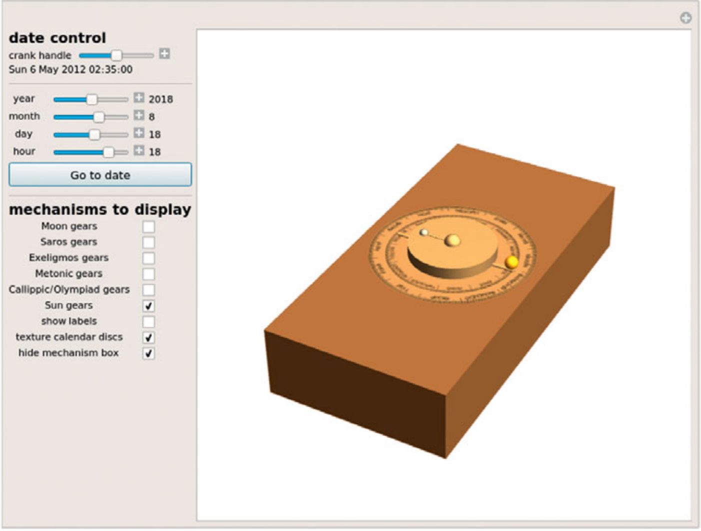


Figure 1.22 The Antikythera mechanism. Courtesy: National Archaeological Museum, Athens.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press



date control

crank handle

Sun 6 May 2012 02:35:00

year 2018

month 8

day 18

hour 18

mechanisms to display

Moon gears □

Saros gears □

Exeligmos gears □

Metonic gears □

Callippic/Olympiad gears □

Sun gears ☑

show labels □

texture calendar discs ☑

hide mechanism box ☑

Figure 1.23 A computer simulation showing the presumed inner workings of the Antikythera mechanism. Source: http://demonstrations.wolfram.com/AntikytheraMechanism/. Contributed by Adam P. Goucher.

and Moon markers around a calendar disc. The reason that researchers have cracked this device with partial success, inhabiting the mind and knowledge of the people who designed it 2,000 years ago, is that it is a mechanical device disclosing the model that it mechanistically simulates.

As we have seen before, mechanistic models such as this one does not necessarily represent the actual mechanisms of the movements of the celestial bodies that they are attempting to simulate, but they do capture some of their regularities. What is most interesting is that while the mechanism was not the cause, the mechanism itself was the cause of the numerical calculations indicating the future position of the celestial bodies. This is literally a computer simulation performed more than 2,000 years ago, illustrating how seriously the concept of causation was approached even back then – with the use of a mechanical calculator or (analogue) computer.

What we will see in this textbook is that modern computer simulations are not very different from this one. Simulations can be used to make optimal predictions and even to try to figure out the most likely mechanistic models of natural processes. This simple artefact actually illustrates the kinds of concepts that bring together computation and simulation to serve the purpose of finding possible causes for natural or astronomical phenomena, just as they set out to do on the Greek island of Antikythera.

Figure 1.24 shows various landmark approaches to the problem of causal discovery. One can see how until very recently most approaches were based largely on data, with almost no model production. This was the statistics-led approach of the last few decades that is still most widely used today in the practice of research and science. In mathematics, however, we have dealt with models all the time. These are mostly

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

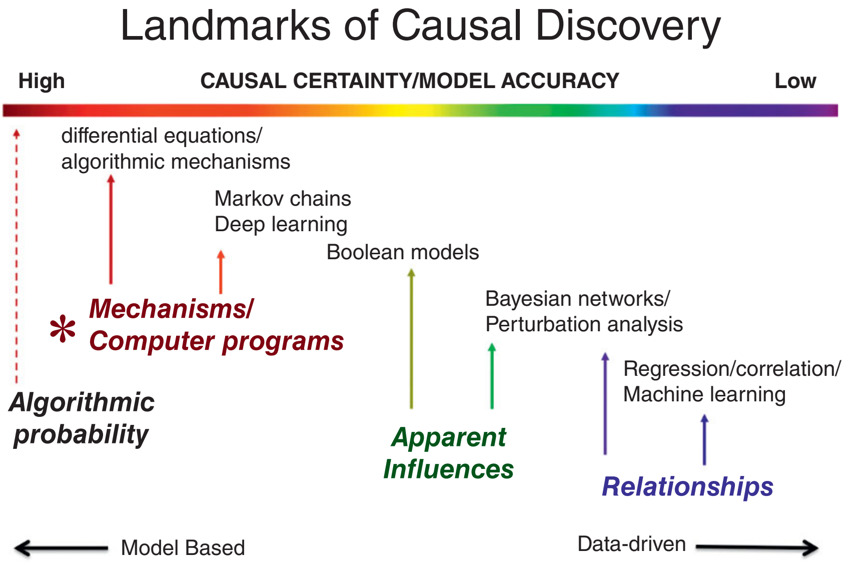


Figure 1.24 Landmarks of causal discovery indicating the place of model-driven approaches, with algorithmic information dynamics based on algorithmic probability being the ultimate theory of optimal model inference.

theoretical models in the area of dynamical systems, which we will cover in detail in a future chapter using differential equations, one of the cornerstones of the field, used in producing numerical and computer simulations. One can also see how perturbation analysis and Bayesian networks move things in the direction of models.

Statistical machine learning (which we will refer to as ML) cannot find or generate mechanistic models from data. Sometimes the work of scientists is facilitated by AI and ML approaches, but traditionally neither AI nor ML can produce models by themselves unless they are empowered by an inference engine of a symbolic type. Thus, if the fundamental distinction between the statistical analysis of patterns and classification versus mechanistic generative models is not taken into consideration, confusion may occur. One example of this comes from biomarker discovery research in the medical domain. Since a number of biomolecules exhibit correlated behaviour they will appear as clusters in data analysis. There are therefore several distinct subsets of biomarkers that are equally effective as predictors of a given phenotype, such as disease. Machine learning and data-driven analytical techniques are excellent tools for finding such correlated patterns in large data sets. Yet, such sets of biomarkers are not equivalent to the mechanisms of disease. The generative model of a disease is not necessarily the same as a set of features correlated with the disease. The lack of appreciation of this (academic) distinction has caused a significant amount of confusion, for example, through neglect of the lack of repeatability in biomarkers.

Trends and methods in these areas, including deep learning and deep neural networks, are black-box approaches to data classification – and even prediction – that work

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

amazingly well but provide little to no understanding of generating mechanisms. As a consequence, they also fail to be scalable to domains for which they are not trained. Supervised learning typically requires that tons of data be trained before anything interesting can be done, and training is needed every time such methods must tackle (even slightly) different data.

It seems reasonable to expect ML and AI to get deeper into model-driven approaches, leaving traditional statistics behind by incorporating algorithmic principles. This means promoting fundamental science rather than throwing more computational resources at data-related problems, as is characteristic of current ML approaches. This is exactly what we are doing with AID, and we will encourage you to test it on your own data for your own purposes, and tell you how to do so in the last chapter involving real-world applications.

Our aim is to go all the way and start from the opposite end of the spectrum in the landmarks of causal discovery, at the most model-driven end, and from there move to the data-driven side when necessary, in a feedback loop, so that a model is generated and improved upon based on the data observed, and once a candidate model is produced or chosen it is tested against new data until we are left with only a handful of likely mechanistic models that we can study and exploit. At this extreme end of the spectrum of model production, generation, and selection, is a theory called algorithmic probability, which is going to be at the centre of our methods and applications in this book.

This book is all about establishing a strong bridge between data- and model-based approaches, between observations and perturbations, and it offers a causal calculus based on the theory of algorithmic probability as the optimal and ultimate theory of induction. To this end, we connect different areas of science such as perturbation analysis, complex networks, information theory, dynamical systems, algorithmic complexity, and even machine learning, where we think our approaches will help better understand fundamental concepts such as deep learning, and will contribute to refining tools and making them more powerful in the quest to uncover causes.

1.6.2 Computational Mechanics

Our approach can be viewed as complementing the area of computational mechanics (traditionally based on, for example, Markov processes and Bayesian inference) with an inference engine powered by algorithmic probability and an empirical estimation of the so-called universal distribution (the distribution associated with algorithmic probability). Similar to program synthesis in inductive inference, albeit mostly of a theoretical nature, there are approaches (such as AIXI) that combine algorithmic probability with decision theory, replacing the prior with the universal distribution, something that AID also suggests. In actual deployments, however, many of these approaches circumvent uncomputability and intractability by relying heavily on popular lossless compression algorithms such as Lempel–Ziv–Welch (LZW), minimum description length approaches, Monte Carlo search, and Markov processes, thereby effectively adopting weaker models of computation. Another set of approaches are based upon

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

Levin's search and other variations, based on a dovetailing algorithm interleaving computer programs one step at a time, from shortest to longest, with each program assigned a fraction of time proportional to its probability during each iteration.

Some methods used by AID can be linked to resource-bounded algorithmic (Kolmogorov–Chaitin) complexity, which imposes an upper bound on program length, which in turn effectively requires the adoption of a linear finite automaton computational model that can be continuously improved upon by throwing more computational resources at it, which we circumvent by allowing improvements while increasing the running time (though in practice each calculation is restricted to a resource-bounded calculation).

1.7 Modelling from Observation

1.7.1 Systems as Black or Grey Boxes

The real world always behaves in more complicated ways than do theoretical models, and often when studying an object we have to treat it as a black box because we cannot see behind or inside it, whether it is the brain or a computer. Take as an example an aeroplane, a man-made artefact for every part of which we have accurate blueprints. We can understand the basics of how it flies but hardly anyone understands in detail the way in which millions of components work together in an aeroplane such as an Airbus A380 or a Boeing 747, with their roughly six million parts (see Fig. 1.25). Instead, manufacturers specialise in different parts and different aspects of their assembly.

Thus even engineers at Boeing and Airbus are compelled to see aeroplanes as black boxes. And this with something that humans have designed and manufactured. The situation is even more complicated with things that have evolved naturally and with which we may have less of a shared history and no involvement at all in their design, such as biological organisms. And even natural phenomena such as the weather or fluid dynamics turn out to be extremely difficult to model and predict. This is because we always have a very partial point of view, we are overwhelmed by the number of

Figure 1.25 Pictures taken by H. Zenil at the main Boeing factory in Everett, WA.

interactions involved, we cannot easily isolate them from one another and at all possible scales. This is why we have to always deal with apparently noisy and incomplete data and why we need to learn to model complex systems with our best tools.

One way to do so is to try to understand a system by simulating it. In this process of modelling, the first step is to understand the value and also the limitations of performing an observation. Let me show you an extremely oversimplified case consisting of an unknown system behind a black box. What typically happens is that the observer is at some point of the system that can be identified as its output.

One can also see how the input can actually be an induced input, an experiment, or a perturbation of the system. It is like throwing a stone at something to see how it reacts. What one can throw at black-box functions are inputs in a certain sequence.

However, behind a black box there could be a function that may appear to be the function we identify, but that in fact produces the outputs in a much more convoluted way, say by adding a random number and subtracting the same random number and thus behaving like the identify function while not being the simplest identify function, thus dissembling the actual operating mechanism.

So it is relevant to ask what made us think that the function behind a black box is actually the most simple version of the identify function, that is, that there was not instead a Rube Goldberg machine pretending to calculate the identify function but doing so in some incredibly, risibly, sophisticated way. Indeed we cannot ever be completely certain that a function is the function that appears to be without actually opening the black box, or that it is a mathematical function at all. It only appears to be so in the range given and for the type of perturbations or stones thrown. This is similar to the black swan problem in statistics, where one can only see the output but not where the swans are coming from, that is, the generating mechanism, and may be misleading at thinking that black swans are impossible to see until we do.

It is also worth noting that though we can follow an order in the input sequence, observations could have been carried out at random without loss of generality as long as the sample is congruent with the sought confidence level of the function behind. However, things are not always that straightforward, and as soon as we get into slightly more complicated cases, it may be more difficult to establish a relationship between the sequence of perturbations. One can call the sequence of observations a sample of the behaviour of the function.

Obviously all this becomes much more difficult in the real world because we usually have no idea of the magnitudes of the possible inputs, nor is there necessarily a privileged order. But we can see what the input for a biological organism may be. For example, to eat or drink can be an input, just as to learn or to read may be considered an input for the mind.

1.7.2 Noise versus Sampling

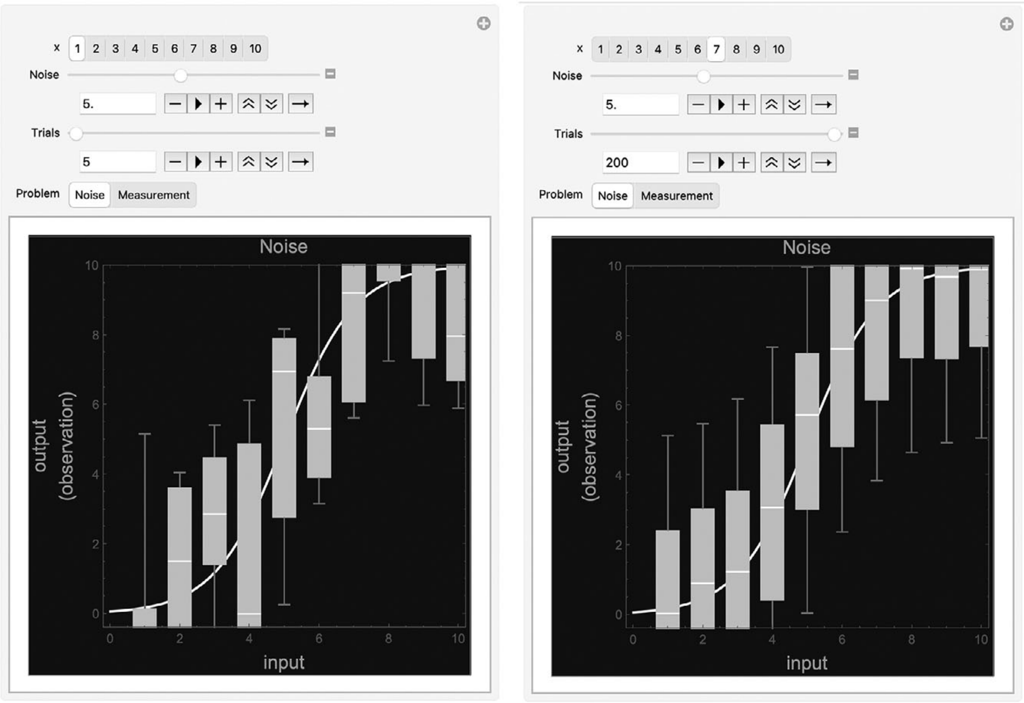
How informative can a single observation or a collection of observations be when it comes to producing a reasonable model with some degree of confidence? In other words, what type and how many observations should we perform to decide whether

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

the function behind a black box is the one we have hypothesised? How many input and output pairs is it sufficient to gather in order to infer an underlying function? Is this a property of the observer or the observed? We will see that the quantity and quality of an experiment depends both on the conditions of the experiment and the capabilities of the observer.

Let us look at this sigmoid function that is usually a representation of how a neuron works, because there is a short interval, called a threshold, where after it is reached, the output behaviour of the system radically changes. There are two main factors allowing or preventing us from gathering enough information about the system to correctly infer the function. One is how much noise there is in the environment, and the other is how precise our measurements are. Notice that these two factors may not be independent and one may condition the other. For example, poor measurement capabilities may look like noise, and noise may look like inaccuracies of measurement. This is why, traditionally, tools are calibrated in a controlled experiment to assess how good they are before being used in more complicated cases.

In Fig. 1.26 we see some large boxes going up and down the sigmoid function represented by a white line. These boxes represent how off a measurement can be from the actual function value for an input on the X axis in case of noise or measurement



x 1 2 3 4 5 6 7 8 9 10
Noise 5.
Trials 5
Problem Noise Measurement

x 1 2 3 4 5 6 7 8 9 10
Noise 5.
Trials 200
Problem Noise Measurement

Figure 1.26 The problem of measurement from sampling as an observer in the face of noise (at source or as a result of measuring limitations). Increasing the number of observations (right) reduces uncertainty (box size) and increases accuracy (median).

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

inaccuracies, and what the measurements or observations on the Y axis would look like. If these boxes are too large, values will start overlapping, preventing us from making educated guesses. The smaller the error, the better and faster we can guess at the generating function.

Notice that when the error bars in yellow are too large, the mean indicated by a small horizontal white line inside the yellow bars will converge to the true value of the function, illustrating how the larger number of observations and measurements increases our chances of finding the generating mechanism. Even though individual measurements indicated by the red dots on the Y axis are highly misleading, this shows that increasing the number of samples increases the accuracy of the prediction.

This is related to what is known as the ‘law of large numbers’, a principle of probability according to which the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer as more trials are performed. A trial consists of a repetition of the same input value several times, so that even if the underlying system is completely deterministic, and perhaps even if the error bars are fixed, the result may be variant output values, but their average will converge to the true value. A trial is also often called a replicate, because the idea is to replicate the same experiment.

Another interesting observation is that the error bars may be of different lengths and even depend on the place they occupy along the function. In this case the error bars are very similar and do not depend on the function. The kind of noise they introduce is then called additive and linear, but more complicated cases exist as well.

The relationship between noise and the number of samples is thus proportional; the greater the noise the larger the number of samples needed, and the smaller in magnitude the noise, the fewer the samples needed. We will see how the field of information theory can help us make all these choices.

What is most interesting in these examples is that no matter how oversimplified, they illustrate how everything may be studied in a similar fashion. There is always an input and an output in a system of interest, even in areas such as biology and cognition, as we will see in the last chapter.

For example, an input can be a drug and the output the development of a disease, an input can be heating a protein and the output the way in which it folds, an input can be the accumulation of clouds and the output is whether it rains or not. Notice also that inputs are usually outputs from other systems, and outputs are usually inputs for other systems. So every aspect of these examples is related to causality.

In the next section we will see how we may study these systems by introducing computation into the study of causation.

1.8 Causality as Computation

In the previous section we saw how functions could be concealed behind or inside black boxes, and how we could infer them by performing experiments such as perturbing them and then recording the corresponding observations. Functions can also be seen

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

as computer programs. As we have said before, the identify function, for example, can have an infinite number of implementations. For example, the simplest version $f(x) = x$ will look exactly the same as $f(x) = x + a - a$ from the perspective of an observer feeding in the input and observing the output, and versions of the same function can be very complicated. If one only cares about the output, then functions may suffice as the object of study, but if we care about mechanisms, then other types of time-dependent equations and computer programs are better representations. Representations that have a time variable are traditionally the object of study of an area called dynamical systems. A cellular automaton is a type of discrete dynamical system that we will be using throughout the book.

An intuitive way to explain what a cellular automaton is is to look at the way they work, using what we call their space-time diagrams. In a space-time diagram of one of the simplest instances of a cellular automaton, called an elementary cellular automaton (ECA) as introduced by Stephen Wolfram, time runs from top to bottom, starting from an initial tape placed on the top. Cellular automata run on a discrete space that looks like a Sudoku grid. One may start the system with a black cell and then apply a set of local rules from top to bottom, row by row. The set of rules is called the rule icon. The rule icon of a cellular automaton dictates how each cell changes over time. You can see that the rule that corresponds to the initial black cell surrounded by white cells is a black cell in the second row, and so on.

This particular example illustrating how a cellular automaton works is slightly misleading because each row should be updated in parallel. In contrast, in this example, the cells highlighted in red show the local rule that was applied at each step sequentially.

A cellular automaton has the advantage of being a highly visual computer program whose evolution can be inspected and followed step-by-step in real time. In other words, it is a sort of transparent computer program that allows us to understand every single one of its components visually, tracking it down at the lowest scale. It is a deterministic system and hence causal, so it will play an important role in our computational approach to the exploration of causality.

Notice that time can be seen as flowing from top to bottom while space runs sideways. This is why the diagram is called a space-time diagram, and it provides a 2-dimensional view of the evolution of a 1-dimensional discrete dynamical system. Notice that one can also start from a less simple initial configuration rather than from the black cell, for example, a random initial configuration. Yet the rules are applied in exactly the same fashion. A minor technicality is what happens at the borders of the system. Each rule is defined every three cells, but at the beginning and end on both right and left extremes one cannot evaluate the last two cells without a third neighbour. In these cases, convention dictates that we take the missing cells from the opposite side, as seen in the example in Fig. 1.27.

Cellular automata can be seen as part of a larger ensemble of means to simulate aspects of causal reality using what is known as an agent-based approach, given that individual cells can be seen as agents interacting with other agents according to specific rules. Agent-based models and cellular automata are bottom-up approaches for constructing models of causal phenomena. Cellular automata can be studied as black,

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

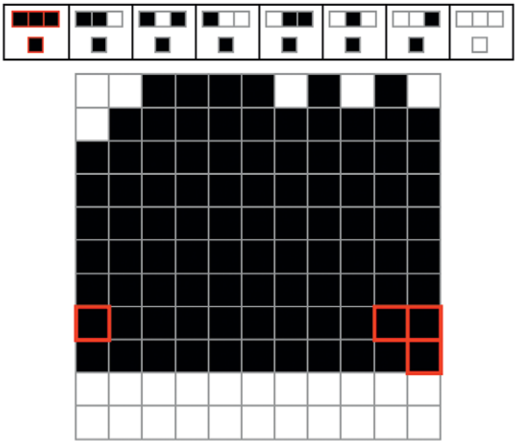


Figure 1.27 Illustration of the application of local rules in the evolution of a cellular automaton. Cells at the boundary take cells on the opposite extreme to continue applying rules. This is called a cylinder configuration. Adapted from http://demonstrations.wolfram.com/CellularAutomataEvaluation/. Based on code by Paul-Jean Letourneau.

grey, and transparent boxes for the purpose of testing old and designing new tools. Traditionally, cellular automata are completely transparent boxes, as already noted, as their evolution is very visual, but as we have also seen, this is not always the case in the real world.

So if you are an ideal observer you would probably witness the whole space-time diagram running, and if sufficient observations were possible the rule icon could be cracked, that is, the computer code and generating mechanism for a particular cellular automaton. Thus a cellular automaton is an ideal simplification on which ideas related to causality can be studied and tested by, for example, concealing the cellular automaton's evolution to see what an observer looking at the last output can infer about it (see Fig. 1.28).

We can perform some sort of organised attack on this particular cellular automaton's black box to see if we can crack the code. The strategy consists in giving it an ordered set of initial conditions based on a binary enumeration and then seeing what cellular automaton it may be by looking at the output, just as we did with mathematical functions.

It looks as if for any input, as marked in red in the example, the output is always blank, so the rule icon must be something that takes any input and then writes a white cell if this is an elementary cellular automaton (Fig. 1.29).

This rule is called Rule 0 because the output is all white cells and, taken as if they were zeros in binary, they would represent the zero in decimal. However, not all rules are so simple. Figure 1.30 demonstrates the so-called ECA rule 30, also in binary, according to Wolfram's enumeration scheme for ECA. And its evolution, depicted here from 1 to 70 steps, looks quite random, and has even been used as a random number generator in the past. Trying to crack the code by only observing the last row at each

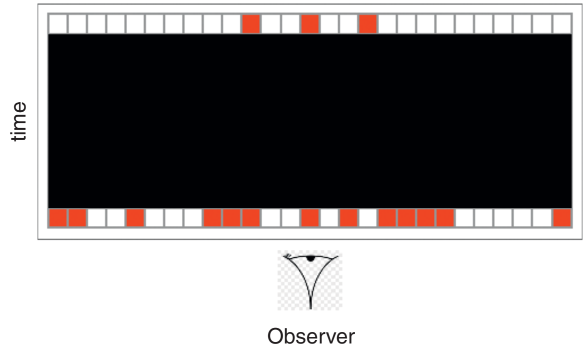


Figure 1.28 A typical example in the real world is an observer looking at the output of a system, with such a system typically being a black box about which little, if anything, can be known or extracted first hand, because either the process that is the output is already complete or it is the result of multi-scale top-down and bottom-up causation to which we have no easy access. We have to find the best tools to infer what the black box may be doing just by looking at the output for different natural or induced inputs.

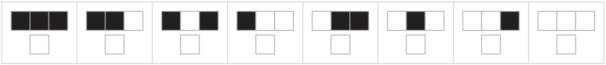


Figure 1.29 All local rules write a white cell for any input.

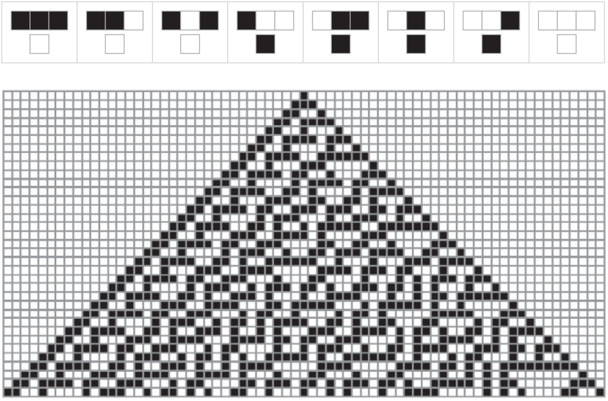


Figure 1.30 Evolution diagram of the ECA rule 30. The icon on top is the rule determining how each cell will be coloured according to the previous three cells evolving from top to bottom and starting from a single black cell.

time is much more difficult than in the previous examples, and one has to perform the observations in the right place to avoid being misled, given that the rule actually presents some regularities in certain places, such as on the left hand side.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

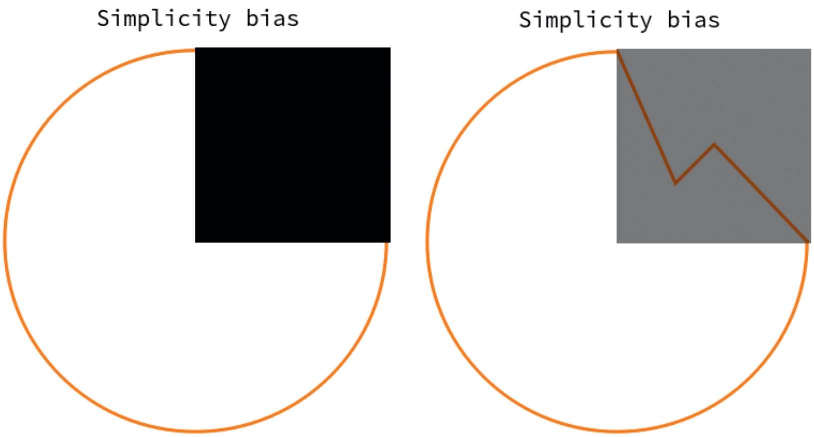


Figure 1.31 Simplicity bias: faced with the challenge of guessing what is behind the circle, the mind will tend to choose the shape that completes the picture in such a way as to minimise surprise, uncertainty, entropy, and algorithmic complexity. Most likely, you'd think that what is behind the black box (left figure) is the complete circle and you would react with surprise if what the black box conceals is a more complicated shape like the one on the right.

A simulation is, by definition, a series of causally connected events, a deterministic evolving system that we call a dynamical system. We will see in the following chapters how we can model and simulate evolving systems, and not only crack their generating codes from partial observations but also reconstruct these systems from disordered observations, and then even reprogram them to perform different computations and behave in different ways.

1.9 Inference and Complexity

In this section, we will briefly explain how causality and inferring functions and computer programs are related to complexity. Imagine you see an object with some part of it blocked by another object, in this case a black square (see Fig. 1.31). Typically, if you ask people what this object could be, they will complete the picture and suspect it to be a full circle. In some sense it seems that our minds are hardwired to complete a picture using the simplest possible shape. We seem deeply biased towards simple forms. If behind the black square there is something else, we would certainly be a little surprised.

Before getting into the technical details of entropy, as we will do later in the book, remember that we had said that entropy as defined by Shannon is traditionally taken as a measure of surprise. Now, something like classical information theory may be able to describe but not explain this bias towards simple forms by establishing that we tend to favour configurations that surprise us less, that is, that have low entropy. But why is this so? We have suggested that such hardwiring for simple things comes from living

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

in a world that is not random, our minds having thus evolved with a high degree of algorithmic structure. We will discuss this in more detail, but this example is meant to show how inference is related to complexity, or rather to simplicity as opposed to randomness, and even to some form of subjectivity that may be cognitive or perhaps even more fundamental. And we will see that the type of randomness we are talking about is not statistical in nature but algorithmic.

Perhaps a useful way to explain how something is complex as opposed to simple is to evoke the way we classify certain human diseases, especially since in the last chapter we will be applying all these ideas, concepts, and the tools based on them to molecular biology and genetics, fields that are deeply relevant to human disease and syndromes. Most diseases are complex, with scientists dealing with challenges related to the observer, the quality and quantity of measurements, apparent noise, and highly interactive systems with multiple and intertwined causes.

Some diseases, such as multiple sclerosis, Alzheimer's, Parkinson's, and most cancers are very complex, in that they can be produced by multiple factors, not just one. They depend on many variables, both genetic and environmental, and are highly unpredictable. In contrast, simple diseases have single or easily identifiable and isolable causes. They may arise from punctual genetic mutations, as is the case with a certain type of breast cancer. The outcome of simple diseases is much easier to predict in that they have well-defined effects. Examples of simple diseases and conditions under this specific definition include cystic fibrosis, Down syndrome, and Huntington's disease. Table 9.1 illustrates what a complex disease looks like as opposed to a simple disease.

How may a classification of this type, contrasting simple and complex diseases, help us to, for example, treat these diseases and conditions? A basic implication of this classification is that one-for-all drugs can only work well for simple diseases, because having multiple causes and factors, complex diseases will appear for different reasons in different people. Hence the current pharmaceutical approach to and business model in medicine is inadequate, and a new paradigm is necessary. That paradigm is known as personalised medicine, and the idea is that one should be able to manufacture a specific drug for a specific person. So what scientists aim for, or should aim for, is an understanding of causes, so as to have a better chance of producing new drugs to direct the way in which a disease develops instead of merely controlling its effects.

This type of simple versus random behaviour can also be modelled mathematically with a view to studying ways of better understanding it, even if it will often be oversimplified. We will explore this issue in further detail later, but for the present we can again resort to cellular automata as an example. If you examine the response of the elementary cellular automaton with rule number 10 (see Fig. 1.32), you will find that no matter what the input is, the system separates the input signals into clear stripes that do not interact with each other, and always produces the same qualitative behaviour. So in a very simplified way it is modelling an aspect of a simple disease. In all four cases, the same rule is applied and the same behaviour obtained for very different initial conditions.

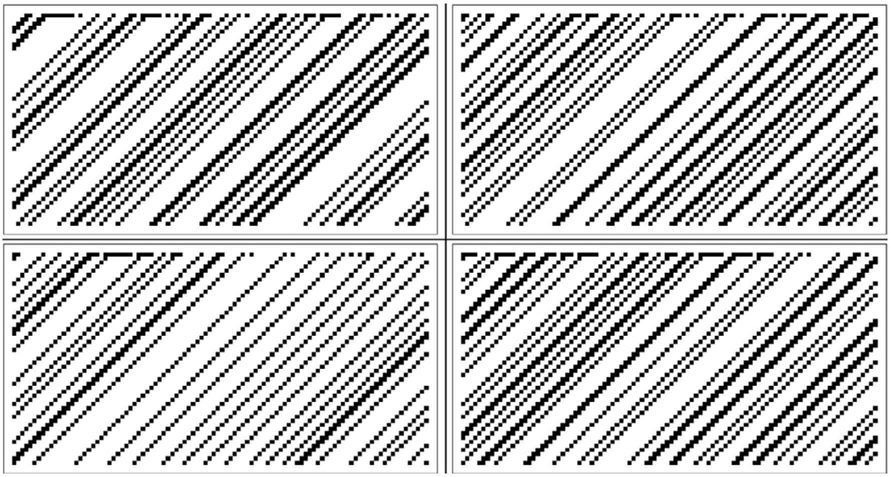


Figure 1.32 Behaviour of the ECA rule 10 for different initial conditions.

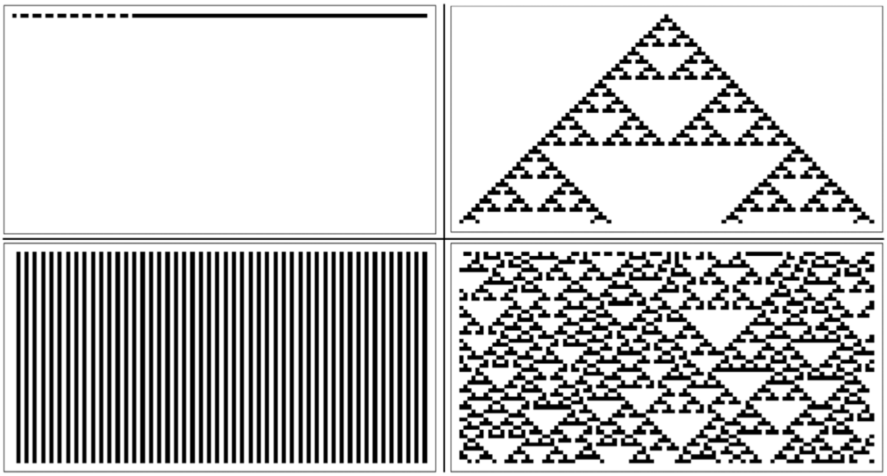


Figure 1.33 Behaviour of the ECA rule 22 for different initial conditions.

But if you take the behaviour of more complicated rules, such as rule 22 depicted in Fig. 1.33, you will find that even small changes in the initial conditions or small perturbations along the way will have a significant impact on the final outcome. This behaviour is more similar, for example, to the way in which cancer tumours evolve, often growing and spreading in unpredictable ways. In all four cases, the same rule, rule 22, was applied, but for different initial conditions. The result is highly unpredictable.

One of the topics that this book will cover is how to characterise these behavioural differences, especially when the rule is unknown but we want to understand the system from a hacker's perspective, that is, when we wish to know the underlying generating mechanism where we have no access to a source code – in artificial examples like this

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

one, but also in real-world systems such as biological organisms like cells. The goal will be not only to attempt to crack the code behind natural phenomena but to try to manipulate and reprogram the way in which these systems evolve and behave, and to that end we will go back and forth between artificial and natural systems in applications to the biological and cognitive sciences.

1.10 The Future of Causality

In all fairness, dealing with causality is extremely difficult, and we used to do well with tools such as traditional statistics and classical probability theory. A major development in causality has been information theory, which is in some sense a clever extension of traditional statistics, one that has sometimes been dangerous because it may give the impression of dealing with causality in different ways than traditional statistics and probability theory, while in fact it does not. The concept of entropy was developed in the context of statistical mechanics in the early twentieth century, as an attempt to understand the average behaviour of small objects such as particles, atoms and molecules. In the late 1940s, Claude Shannon formulated a version of entropy similar to the one adopted in statistical mechanics, a branch of quantum physics, intended to help evaluate the capacity of communication channels. Shannon was working at the time at Bell Labs, a company concerned with communication lines and a precursor of today's AT&T. But it was Andrey Kolmogorov, standing on the shoulders of such giants as Richard von Mises and Alan Turing, who soon found out that probability theory was very limited in dealing with fundamental aspects of randomness and hence causality.

Yet, we have not made much use of what we know about mathematical randomness, as opposed to statistical randomness, either in data science or science in general, as we tackle the challenge of causality. To characterise randomness is key in the area of causation because it is exactly what one needs to rule out when using observations and models to come up with a candidate mechanism to explain the data at hand. One does not want to attribute chance where it does not obtain, and vice versa.

The founders of the theory of algorithmic complexity, namely Andrey Kolmogorov, Gregory Chaitin, Ray Solomonoff, and Leonid Levin, came up with a mathematical solution to the challenges of randomness, causality discovery, and inference.

Indeed, Solomonoff and Levin showed that the concept of algorithmic probability was the ultimate solution to the problem of causal inference, suggesting that intelligence was likely just computation, in a fundamental sense. The two concepts, algorithmic complexity and algorithmic probability, are faces of the same coin, deeply related to each other. However, while some applied scientists may have heard of algorithmic randomness, and a smaller number about algorithmic probability, very few actually use them in addressing the challenge of causation.

The advantages of algorithmic complexity and algorithmic probability in data science have been largely underused if not ignored, because unlike other, simpler, measures such as those derived from statistics, including Shannon entropy, measures of algorithmic complexity are much more difficult to estimate. One thing to bear in mind

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

is that there will always be some statistics involved, because the only way to attain full certainty about causation is not only to witness a process of interest firsthand but also to isolate the system and have access to all the relevant processes at all scales, something that is difficult if not necessarily impossible in the real world.

In this book we will introduce what we think is an exciting – hopefully for you too – new approach to tackling the problem of causality based on algorithmic complexity by way of algorithmic probability. We will call this area Algorithmic Information Dynamics or Algorithmic Dynamics for short.

It is pertinent to return to the claim about algorithmic probability made by Marvin Minsky, a founding father of the field of artificial intelligence, just a few months before he passed away, and quoted above. Minsky recognised that the way forward was through AID and the work of Solomonoff and Chaitin.

Both Minsky and Solomonoff attended the Dartmouth College conference/workshop, the conference sometimes seen in the US as the inception of the new field of AI. Clearly, algorithmic probability is not just another pet topic in complexity science but a fundamental pillar of science, as we will see throughout this book.

1.11 Conclusion

In this textbook we offer an approach to causal discovery that is model driven and based on the theory of algorithmic probability in order to help discover and produce candidate generative models for observed data. We have made some progress in applying these new tools to data to deal with causation, and they themselves build on the progress made throughout history, from logic to statistics, probability, dynamical systems, and computation, including perturbation analysis, but they displace traditional statistics and the need for estimations of probability distributions from their central position.

We think that AID may be part of a shift away from the dominance that statistics and even equations have long exerted over science, and towards computation and machine learning, but a different kind of machine learning, rooted in algorithmic approaches rather than traditional statistics, as is the case with machine learning today.

We will see how AID can help predict and even reconstruct the phase space and space-time behaviour of complex non-linear dynamical systems with limited information. But first, in the next chapters we will cover many more topics needed to understand AID, now that we have offered a brief overview that highlights the significance and some of the many aspects and challenges of causality.

1.12 Practice Questions

1. The study of how things and events happen or come into being and are causally connected with each other as opposed to occurring by chance is called:
   (a) Non-linearity
   (b) Causality

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

   (c) Probability
   (d) Entropy
2. Scientists perform experiments mainly to:
   (a) Justify the costs of lab equipment
   (b) Discover causal knowledge
   (c) Find correlations
3. The interaction or evolution of non-trivial systems may often seem:
   (a) Trivial and easy
   (b) Complex and noisy
   (c) Quantum mechanical
4. Determinism implies that future events are easily predictable.
   (a) True
   (b) False
5. A mechanistic model means that a model:
   (a) Can be followed from cause to effect, like an algorithm, step by step
   (b) Cannot incorporate noise and is never complex
   (c) Is always the right explanation
   (d) Cannot run on a computer
6. A Rube Goldberg machine illustrates that:
   (a) Causality is always easy to identify no matter how laughable
   (b) Non-linearity is complex and interesting
   (c) Computation is universal and ubiquitous
   (d) Trivial outcomes can be produced by needlessly sophisticated but causally connected tasks
7. Both the geocentric and heliocentric models of the solar system are mechanistic models.
   (a) True
   (b) False
8. The geocentric model has a lot of predictive power even if considered wrong.
   (a) True
   (b) False
9. Occam's razor is also known as:
   (a) The law of big numbers
   (b) The law of logic
   (c) The law of parsimony
10. Occam's razor states that when presented with competing models one should favour:
    (a) The one that satisfactorily explains the observed phenomena with fewest assumptions
    (b) The one that satisfactorily explains the observed phenomena with more and more complex assumptions
11. Occam's razor is a criterion for:
    (a) Simplicity
    (b) Complexity

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

12. Occam's razor generally rules out Goldberg machine-like explanations for a given outcome.
    (a) True
    (b) False
13. If P = True and Q = False, P AND Q is:
    (a) True
    (b) False
14. If P = True and Q = False, P OR Q is:
    (a) True
    (b) False
15. The concept of algorithmic (Kolmogorov) complexity is colloquially related to Occam's razor.
    (a) True
    (b) False
16. In science, Occam's razor is used as:
    (a) The ultimate test for truth
    (b) A substitute for the experimental validation of theoretical models
    (c) A guiding principle for the production of scientific models
17. Causal inference draws conclusions based on:
    (a) Chains of effects
    (b) The detection of noise
    (c) Differential equations
    (d) The use of computers
18. Classical probability and traditional statistics usually mean the study of science:
    (a) Through the use of universal computation
    (b) Through assumptions of probability distributions
    (c) Through analysis of possible generating mechanistic models
19. A fundamental challenge in the study of causality is:
    (a) The difficulty of describing data probabilistically
    (b) The increasing entropy of the universe
    (c) Disentangling causes from effects and determining mechanistic models
20. A uniform prior distribution assumes that:
    (a) Some events are more likely than others
    (b) Rare events have probability equal to zero, no matter how many events are possible
    (c) All events are equally likely
21. Laplace's 'principle of insufficient reason' states that if there are n indistinguishable causes for an effect, then:
    (a) Each possible cause has probability equal to n
    (b) Each possible cause has probability equal to 1
    (c) Each possible cause has probability equal to 0
    (d) Each possible cause has probability equal to $1/n$
22. Shannon's entropy is a measure of:
    (a) Causality between observed variables

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

(b) Lack of knowledge about the underlying probability distribution
(c) Cost of building a universal computer

23. Correlation implies causation.
(a) True
(b) False

24. Which of the following is one of the latest practices in probabilistic causal discovery?
(a) Applying the Wiener filter
(b) Comparing time series to test correlation
(c) Analysis of the response to perturbations and counterfactuals
(d) Estimation of fractal dimensions

25. One way to test a causal relation between two time series is to:
(a) Perform a perturbation on one and see if it has an effect on the other
(b) Reorder the data
(c) Merge the time series
(d) Calculate the Spearman correlation coefficient

26. If we encounter a case where perturbations do not lead to a change in the effects, it is evidence of:
(a) Strong causal relationship between the events
(b) Lack of a causal relationship between the events

27. In a directed acyclic graph that represents causal relationships:
(a) Nodes are events and events are linked with each other if there's a direct cause and effect between them
(b) Nodes are causes and causes are linked with each other if there's an event between them

28. In a causal directed acyclic graph there may be no loops because:
(a) It would become chaotic
(b) It would make an effect redundant
(c) It would make a cause into the cause of itself

29. Computer simulations in the context of causality are typically used to:
(a) Find likely mechanisms for natural or artificial processes
(b) Classify data

30. The Antikythera mechanism is an example of a _____ model used to predict celestial events.
(a) Mechanistic
(b) Statistical

31. The study of dynamical systems is an example of a:
(a) Statistics-driven approach
(b) Model-driven approach

32. Conventional machine learning (regression, support vector machines, etc.) are examples of a:
(a) Statistics-driven approach
(b) Model-driven approach

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

33. We often have to treat an object of study like a ‘black box’ because:
    (a) The real world is complicated and we rarely or never witness phenomena unfolding in real time at all scales
    (b) Concealing the underlying mechanism of a system allows us to use tools like statistics
34. A ‘black box’:
    (a) Exposes its underlying mechanisms
    (b) Conceals its underlying operating mechanism
35. ‘The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed’ is an enunciation of:
    (a) The law of parsimony
    (b) The law of large numbers
    (c) Bayes’s theorem
    (d) The 2nd law of thermodynamics
36. When sampling, an increased amount of noise means that to provide a reliable estimate we need a _____ number of samples.
    (a) Small
    (b) Large
37. A cellular automaton is a _____ dynamical system.
    (a) Continuous
    (b) Discrete
38. Agent-based models and cellular automata are _____ approaches for constructing models of causal phenomena.
    (a) Top-down
    (b) Bottom-up
39. During the execution of a cellular automaton, each of the rows in its space-time diagram is updated:
    (a) Sequentially, from left to right
    (b) Sequentially, from right to left
    (c) In parallel
40. In a typical space-time evolution of an elementary cellular automaton, time flows:
    (a) From top to bottom
    (b) From bottom to top
    (c) From left to right
    (d) From right to left
41. The space-time diagram of an elementary cellular automaton shows a _____ view of the evolution of a _____ system.
    (a) One dimensional, two dimensional
    (b) Two dimensional, one dimensional
    (c) Three dimensional, two dimensional
    (d) Three dimensional, n-dimensional

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

42. Shannon's entropy is typically interpreted as a measure of:
    (a) Change over time
    (b) Meaning
    (c) Surprise
    (d) Complexity
43. What is one plausible explanation for why humans prefer simpler structures?
    (a) Human emotions block our ability to generate complex models of the world
    (b) We have adapted to environments that are not random
    (c) Neuron synapses fire less rapidly than thoughts
44. Diseases with well-defined causes and predictable effects are, in general, cases of
    (a) Simple diseases
    (b) More complex diseases
45. If a cellular automaton produces the same qualitative behaviour no matter what the input is, the cellular automaton's rule is:
    (a) Simple
    (b) Complex
46. If scientists want to develop personalised medicine, they should:
    (a) Focus on changing the 'source code' that causes diseases in the first place
    (b) Treat the conditions that these diseases cause by using generalised drugs
    (c) Study massive amounts of social data
47. The concept of entropy was first motivated by:
    (a) Developing mechanistic tools to understand the motion of celestial bodies
    (b) Computational methods to understand the behaviour of large social systems
    (c) The study and development of tools to understand the behaviour of small particles
48. Algorithmic probability is the optimal solution to:
    (a) Neural networks
    (b) Causal inference and characterising randomness
    (c) Statistics
49. Currently, data science does not usually use algorithmic complexity/probability because:
    (a) It is very difficult to calculate
    (b) Statistics already gives us an understanding of causality
    (c) Information theory has been unsuccessful
50. Algorithmic probability can be used practically because:
    (a) It is statistical in nature
    (b) It is possible to make approximations of what it is able to predict
    (c) It is widely used already

1.13 Discussion Questions

The purpose of the discussion points below is to facilitate your understanding of the material in the chapter by interrogating the concepts discussed in it. In particular, what

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press

are the limits of the concepts? Are there any hidden presuppositions in our reasoning when we think about causality, for example?

1. Bertrand Russell made the argument in a classic essay (1911) that causality was a relic of a bygone age: ‘The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.’ What did he mean? Why did he make this statement? Do you think it is true or false?
2. In physics, at the most fundamental level, there are essentially fluctuating quantum fields. Here the idea or notion of causation appears to dissolve into nothingness. Do you agree with this description, and if not why not? Discuss its pros and cons. If this characterisation of the state of affairs has some validity, then why is it that we devote scientific effort to understanding causality in the specialised sciences? Might not our continuing to do so reflect a poor understanding of the world?
3. Gödel found closed, time-like solutions to Einstein’s equations. Explain what he found. Discuss the implications such solutions, which allow time-travel into the past, have or do not have for the notion of causation and the scientific study of causality.
4. Discuss and enumerate sufficient and necessary assumptions about the world that would, in concert, imply that there is no causation in the world, and that the scientific study of causality is fundamentally flawed. Could this be a possibility? If not, why not?
5. Formulate some additional discussion questions in the spirit of the above that would, in effect, open new avenues for investigation beyond our current way of thinking and analysing causality.

https://doi.org/10.1017/9781108596619.004 Published online by Cambridge University Press