Models We Can Understand

We have not always insisted on understanding our predictions. For example, some of the Founding Fathers of the United States made daily records of the weather and the factors they thought were related to it: when plants start blooming, the first frost, and so forth. They hoped this aggregated data would reveal reliable correlations, such as the daffodils' blooming early signifying that there's a good chance it will be a wet summer. Until the early 1900s, that sort of weather forecasting worked better than not predicting at all.

As Nate Silver explains in *The Signal and the Noise*, this is statistical forecasting: we gather data and use it to make an informed guess about what will happen, based on the assumption that the data is expressing a regularity.[10] Silver says that is how hurricanes were predicted until about thirty years ago. It works pretty well, at least as long as the natural system is fairly consistent.

Statistical forecasting doesn't need a model of the sort proposed in 1900 by Vilhelm Bjerknes, which we looked at in chapter 1. Bjerknes's model explained the dynamics of global weather using seven variables and Newtonian physics: relevant factors connected by rules governing their interactions.[11] But there was a problem: even using only seven variables, the computations were so complex that in 1922 a mathematician named Lewis Fry Richardson spent six full weeks doing the work required to predict the weather on a day years earlier, based on data gathered on the days before it. He wasn't even close. After all that grueling work, Richardson's calculated air pressure was 150 times too high.[12]

These days we track hundreds of variables to forecast the weather, as well as to predict the longer-term changes in our climate. We do so with computers that chortle at the 1940s computer—the ENIAC (Electronic Numerical Integrator and Computer)—that took twenty-four hours to predict the next day's weather.[13] Nevertheless, until machine learning, we relied on the model-based technique that harks back to Pierre-Simon Laplace's demon: if we know the rules governing the behavior of the seven factors that determine the weather,

and if we have the data about them for any one moment in the life of the Earth, we should be able to predict what the next moment's weather will be.

The problem is that so very many factors can affect the weather. In fact, Silver says "the entire discipline of chaos theory developed out of what were essentially frustrated attempts to make weather forecasts."[14] Literally everything on the surface of the planet affects the weather to one degree or another. It is not a coincidence that the example forever associated with Chaos Theory involves a butterfly that creates a catastrophic weather event thousands of miles away.

So if we want to make a prediction about a system like the weather—a third level of predictive complexity, in the terms discussed in the previous chapter—we seem to be left with bad choices. We can rely on statistics and hope that we've been gathering the relevant ones, and that the future will repeat the patterns of the past as surely as the Nile overflows after the Dog Star returns. Or we can figure out the laws governing change and hope that the system is as simple as the model we're using . . . and that it is not disrupted by, say, the Krakatoa volcano that erupted in 1883, spewing forth enough ash to cool the seasons for a year and to chill the oceans for a full century afterward.[15]

Bjerknes's seven-factor weather model has the advantage of providing a working model that at least crudely reflects its conceptual model. But we don't always insist on that. The following four examples show different ways successful working models may or may not coincide with our conceptual models. They'll also let us see how deeply machine learning models break with our traditional practices and age-old assumptions about how things happen . . . and our assumptions about how suited we humans are to understanding what happens.

Spreadsheets

Although computerized spreadsheets date back to the early 1960s,[16] they only became widely used after 1978 when Dan Bricklin, a student working on his MBA at Harvard Business School, was annoyed

by a class assignment that required calculating the financial implications of a merger. Having to recalculate all the dependent numbers when any single variable changed was more than irksome.[17]

So, in the spring of 1978, Bricklin prototyped a spreadsheet on an Apple II personal computer, using a gaming controller in place of a mouse.[18] With the rise of PCs and with the decision by Bricklin and his partner, Bob Frankston, not to patent the software,[19] spreadsheets became a crucial way businesses understood themselves and made decisions: a company's conceptual model of itself now could be expressed in a working model that let the business see the effects of the forces affecting it and of decisions the company was contemplating.

In a remarkably prescient article in 1984, Stephen Levy wrote, "It is not far-fetched to imagine that the introduction of the electronic spreadsheet will have an effect like that brought about by the development during the Renaissance of double-entry bookkeeping."[20] He was right. "The spreadsheet is a tool, and it is also a world view—reality by the numbers," Levy wrote.

A spreadsheet is what a business looks like to a traditional computer: quantitative information connected by rules. The rules—formulas—and some of the data, such as fixed costs, are relatively stable. But some of the data changes frequently or even constantly: sales, expenses, headcount, and so on. Personal computers running spreadsheets made keeping the working model up to date so easy and fast that a new decision-making process was made feasible: a spreadsheet is a temptation to fiddle, to try out new futures by plugging in different numbers or by tweaking a relationship. This makes them very different from most traditional models, which focus on representing unchanging relationships, whether they're Newtonian laws or the effect that raising taxes has on savings. Spreadsheets are models that encourage play: you "run the numbers," but then you poke at them to try out "what if this" or "what if that." This was a model meant to be played with.

A spreadsheet thus is a simple example of a working model based on a fully understandable conceptual model. It lets you plug in data or play with the rules to see what the future might or could look like.

Of course, they are inexact, they can't capture all of the relationships among all of the pieces, and the predictions made from their models may be thrown off by events that no one predicted. Because spreadsheets are tools and not perfect encapsulations of every possible eventuality, we accept some distance between the working model and the conceptual model, and between the conceptual model and the real world. We continue to use them because, as George E. P. Box said, "[a]ll models are wrong but some are useful."[21]

Armillary

In the Galileo Museum in Florence sits a beautiful set of nested geared rings, 6.5 feet tall.[22] If we today had to guess the point of this intricate mechanism just by looking at it, we might suppose that it's some type of clock. If we were contemporaries of it, we'd be far more likely to recognize that it shows the positions of the major heavenly bodies in Earth's skies for any night.

Antonio Santucci finished this object, called an armillary, in 1593, after five years of work. Although forty-six years earlier Nicolaus Copernicus had shown that the Earth revolves around the sun, Santucci still put the Earth at the center, circled by seven rings that display the positions of the seven known planets. An eighth ring has the fixed stars on it, as well as the markings of the zodiac. Adjust the rings on this wood-and-metal machine, and the planets and fixed stars will align relative to one another and to the Earth. Now gild it and paint in the four winds, the shield of your patron's Medici-related family, and an image of God Himself, and you have a beautiful room-size model of the universe.[23]

According to no less an authority than Galileo, even Santucci eventually came around to Copernicus's idea.[24] But the armillary's model of the universe is odd beyond its Earth-centric view. It simulates the movement of the heavenly bodies using only circles as components of the mechanism because, from the early Greeks on, it was commonly assumed that because the heavens were the realm of perfection, and circles were the perfect shape, the heavenly bodies must move in perfect circles. That makes the planets a problem, for they

wander through Earth's sky in distinctly noncircular ways; *planet* comes from the Greek word for *wanderer*. Therefore, if the armillary were to be truthful to its conceptual model, not only did it have to get the planets in the right places relative to Earth, it also had to do it the way the universe does: by using circles. So Santucci set smaller gears turning as they revolved around larger gears that were themselves turning, adding in as many as necessary to model the paths of the planets accurately.[25]

The result is a successful working model that uses a convoluted mechanism dictated by a conceptual model that has been shown to be wildly wrong.

The error in its conceptual model also happens to make the working model quite beautiful.

Tides

"Unlike the human brain, this one cannot make a mistake."[26]

That's how a 1914 article in *Scientific American* described a tide-predicting machine made of brass and wood that made mistakes all the time. And its creators knew it.

Newton had shown that the gravitational pull of the sun and moon accounted for the rise and fall of the tides around Earth. But his formulas only worked approximately, for, as the *Scientific American* article pointed out,

> *the earth is not a perfect sphere, it isn't covered with water to a uniform depth, it has many continents and islands and sea passages of peculiar shapes and depths, the earth does not travel about the Sun in a circular path, and Earth, Sun and Moon are not always in line. The result is that two tides are rarely the same for the same place twice running, and that tides differ from each other enormously in both times and in amplitude.*[27]

In his book *Tides: The Science and Spirit of the Ocean*, Jonathan White notes, "There are hundreds of these eccentricities, each call-

ing out to the oceans—some loudly, some faintly, some repeating every four hours and others every twenty thousand years." Newton knew he was ignoring these complications, but they were too complicated to account for. (It's quite possible that he never saw an ocean himself.)[28]

It was Laplace who again got Newton righter than Newton did, creating formulas that included the moon's eight-year cycle of distances from the Earth, its varying distance north and south of the equator, the effect of the shape and depth of the ocean's basin, the texture of the ocean floor, the water's fluctuating temperatures, and other conditions.[29]

This added nuance to Newton's model, but a vast number of additional factors also affect the tides. It took about another hundred years for Lord Kelvin, in 1867, to come up with a way of predicting tides that takes all the factors into account without having to know what all of them are.[30]

As the 1914 *Scientific American* article explains it, imagine a pencil floating up and down in an ocean, creating a curve as it draws on a piece of paper scrolling past it. Imagine lots of pencils placed at uniform distances from one another. Now imagine the ocean lying still, without any bodies exerting gravitational forces on it. Finally, imagine a series of fictitious suns and moons above Earth in exactly the right spots for their gravity to pull that pencil to create exactly those curves. Wherever you have a curve that needs explaining, add another imaginary sun or moon in the right position to get the expected result. Lord Kelvin ended up with a "very respectable number" of imaginary suns and moons circling the Earth, as the article puts it. If adding sea serpents would have helped, presumably Lord Kelvin would have added them as well.[31]

With the assistance of George Darwin—brother of Charles—Lord Kelvin computed formulas that expressed the pull of these imaginary bodies, then designed a machine that used chains and pulleys to add up all of those forces and to draw the tidal curves. By 1914, this had evolved into the beast feted in the *Scientific American* article: fifteen thousand parts that, combined, could draw a line showing the tides at any hour.

Lord Kelvin was in fact not the first to imagine a science-fiction Earth circled by multiple suns and moons that create the wrinkled swells and ebbs of tides caused by the vagaries of the Earth's geography, topology, weather, and hundreds of other factors. Laplace himself "imagined a stationary Earth with these tide components circling as satellites."[32] Lord Kelvin's machine and its iterations took this to further levels of detail, while accepting that the actual tides are subject to still more factors that simply could not be captured in the machine's model—the influx of melted snow from a particularly long winter, the effect of storms, and all the other influences Earth is heir to. The *Scientific American* article could claim the machine never makes a mistake because Kelvin's machine was as accurate as the tools and data of the time allowed, so it became the accuracy we counted as acceptable . . . all while relying on a fictitious model.

It set this level of accuracy by building a working model that is knowingly, even wildly, divorced from its conceptual model.

The River

In 1943, the US Army Corps of Engineers set Italian and German prisoners of war to work building the largest scale model in history: two hundred acres representing the 41 percent of the United States that drains into the Mississippi River. By 1949 the model was being used to run simulations to determine what would happen to cities and towns along the way if water flooded in. It's credited with preventing $65 million in damage from a flood in Omaha in 1952.[33] In fact, some claim its simulations are more accurate than the existing digital models.[34]

Water was at the heart of another type of physical model: the MONIAC (Monetary National Income Analogue Computer) economic simulator built in 1949 by the New Zealand economist Alban William Housego Phillips.[35] The MONIAC used colored water in transparent pipes to simulate the effects of Keynesian economic policies. Tanks of water represented "households, business, government, exporting

and importing sectors of the economy," measuring income, spending, and GDP.[36]

It worked, given its limitations. The number of variables it could include was constrained by the number of valves, tubes, and tanks that could fit in a device about the size of a refrigerator.[37] But because it only took account of a relative handful of the variables that influence the state of a national economy, it was far less accurate than the Mississippi River simulator. Yet the flow of water through a river the size of the Mississippi is also affected by more variables than humans can list. So how could the Mississippi model get predictions so right?

The Mississippi had the advantage of not requiring its creators to have a complete conceptual model of how a river works. For example, if you want to predict what will happen if you place a boulder in a rapids, you don't have to have a complete model of fluid dynamics; you can just build a working scale model that puts a small rock into a small flow. So long as scale doesn't matter, your model will give you your answer. As Stanford Gibson, a senior hydraulic engineer in the Army Corps of Engineers, said about the Mississippi basin project, "The physical model will simulate the processes on its own."[38]

So this working model can deal with more complexity *because* it doesn't have a conceptual model: it puts the actual forces to use in a controlled and adjustable way. Because the model is not merely a symbolic one—real water is rolling past a real, scaled-down boulder—the results aren't limited by what we know to factor in. That's the problem with the MONIAC: it sticks with factors that we know about. It's like reducing weather to seven known factors.

Still, the Mississippi River basin model may seem to make no assumptions about what affects floods, but of course it does. It assumes that what happens at full scale also happens at 1/2000 scale, which is not completely accurate for the dynamics of water; for example, the creators of a model of San Francisco Bay purposefully distorted the horizontal and vertical scales by a factor of ten in order to get the right flow over the tidal flats.[39] Likewise, the Mississippi model does

not simulate the gravitational pull of the sun and the moon. Nor does it grow miniature crops in the fields. The model assumes those factors are not relevant to the predictions it was designed to enable. Using the Mississippi model to simulate the effects of climate change or the effect of paddle wheelers on algae growth probably wouldn't give reliable results, for those phenomena are affected by factors not in the model and are sensitive to scale.

The Mississippi model wasn't constructed based on an explicit conceptual model of the Mississippi River basin, and it works without yielding one. Indeed, it works because it doesn't require us to understand the Mississippi River: it lets the physics of the simulation do its job without imposing the limitations of human reason on it. The result is a model that is more accurate than one like the MONIAC that was constructed based on human theory and understanding. So the advent of machine learning is not the first time we have been presented with working models for which we have no conceptual model.

But, as we'll see, machine learning is making clear a problem with the very idea of conceptual models. Suppose our concepts and the world they model aren't nearly as alike as we've thought? After all, when it comes to the Mighty Mississippi, the most accurate working model lets physical water flow deeper than our conceptual understanding.

Despite the important differences among all these models—from spreadsheets to the Mississippi—it's the similarities that tell us the most about how we have made our way in a wildly unpredictable world.

In all these cases, models *stand in* for the real thing: the armillary is not the heavenly domain, the spreadsheet is not the business, the tubes filled with colored water are not the economy. They do so by *simplifying* the real-world version. A complete tidal model would have to include a complete weather model, which would have to include a complete model of industrial effects on the climate, until the entire world and heavens have been included. Models simplify systems until they yield acceptably accurate predictions.

Models thereby assume that we humans can identify the elements that are relevant to the thing we are modeling: the factors, rules, and principles that determine how it behaves. Even the model of the Mississippi, which does not need to understand the physics of fluid dynamics, assumes that floods are affected by the curves and depths of the river and not by whether the blue vervain growing along the sides of the river are in flower. This also implies that models assume some degree of *regularity*. The armillary assumes that the heavenly bodies will continue to move across the skies in their accustomed paths; the tidal machine assumes the gravitational mass of the sun and moon will remain constant; the spreadsheet assumes that sales are always going to be added to revenues.

Because the simplification process is *done by human beings*, models reflect our strengths and our weaknesses. The strengths include our ability to see the order beneath the apparent flux of change. But we are also inevitably prone to using unexamined assumptions, have limited memories and inherent biases, and are willing to simplify our world to the point where we can understand it.

Despite models' inescapable weaknesses due to our own flawed natures, they have been essential to how we understand and control our world. They have become the *stable frameworks that enable us to predict and explain the ever-changing and overwhelming world in process all around us*.

Beyond Explanation

We are transitioning to a new type of working model, one that does not require knowing how a system works and that does not require simplifying it, at least not to the degree we have in the past. This makes the rise of machine learning one of the most significant disruptions in our history.[40]

In the introduction, we talked about Deep Patient, a machine learning system that researchers at Mount Sinai Hospital in New York fed hundreds of pieces of medical data about seven hundred thousand patients. As a result, it was able to predict the onset of

diseases that have defied human diagnostic abilities. Likewise, a Google research project analyzed the hospital health records of 216,221 adults. From the forty-six billion data points, it was able to predict the length of a patient's stay in the hospital, the probability that the patient would exit alive, and more.[41]

These systems work: they produce probabilistically accurate outcomes. But why?

Both of these examples use *deep learning*, a type of machine learning that looks for relationships among the data points without being instructed what to look for. The system connects the nodes into a web of probabilistic dependencies, and then uses that web—an "artificial neural network"—to refine the relationships again and again. The result is a network of data nodes, each with a "weight" that is used to determine whether the nodes it is connected to will activate; in this way, artificial neural networks are like the brain's very real neural network.

These networks can be insanely complicated. For example, Deep Patient looked at five hundred factors for each of the hundreds of thousands patients whose records it analyzed, creating a final data set of two hundred million pieces of data. To check on a particular patient's health, you run her data through that network and get back probabilistic predictions about the medical risks she faces. For example, Deep Patient is unusually good at telling which patients are at risk of developing schizophrenia, a condition that is extremely hard for human doctors to predict.[42]

But the clues the system uses to make these predictions are not necessarily like the signs doctors typically use, the way tingling and numbness can be an early sign of multiple sclerosis, or sudden thirst sometimes indicates diabetes. In fact, if you asked Deep Patient how it came to classify people as likely to develop schizophrenia, there could be so many variables arranged in such a complex constellation that we humans might not be able see the patterns in the data even if they were pointed out to us. Some factor might increase the probability of a patient becoming schizophrenic but only in conjunction with other factors, and the set of relevant factors may itself vary widely, just as your spouse dressing more formally might

mean nothing alone but, in conjunction with one set of "tells," might be a sign that she is feeling more confident about herself and, with other sets, might mean that she is aiming for a promotion at work or is cheating on you. The number and complexity of contextual variables mean that Deep Patient simply cannot always explain its diagnoses as a conceptual model that its human keepers can understand.

Getting explanations from a machine learning system is much easier when humans have programmed in the features the system should be looking for. For example, the Irvine, California–based company Bitvore analyzes news feeds and public filings to provide real-time notifications to clients about developments relevant to them. To do this, its dozens of algorithms have been trained to look for over three hundred different sorts of events, including CEO resignations, bankruptcies, lawsuits, and criminal behavior, all of which might have financial impacts. Jeff Curie, Bitvore's president, says that it's like having several hundred subject experts each scouring a vast stream of data.[43] When one of these robotic experts finds something relevant to its area of expertise, it flags it, tags it, and passes it on to the rest, who add what they know and connect it to other events. This provides clients—including intelligence agencies and financial houses—not just an early warning system that sounds an alarm but also contextualized information about the alarm.

Bitvore's system is designed so that its conclusions will always be explicable to clients. The company's chief technology officer, Greg Bolcer, told me about a time when the system flagged news about cash reserves as relevant to its municipal government clients. It seemed off, so Bolcer investigated. The system reported that the event concerned not cash reserves but a vineyard's "special reserve" wines and was of no relevance to Bitvore's clients. To avoid that sort of machine-based confusion, Bitvore's system is architected so that humans can always demand an explanation.[44]

Bitvore is far from the only system that keeps its results explicable. Andrew Jennings, the senior vice president of scores and analytics at FICO, the credit-scoring company, told me, "There are a number of long standing rules and regulations around credit scoring in the

US and elsewhere as a result of legislation that require[s] people who build credit scores to manage the tradeoff between things that are predictively useful and legally permitted."[45] Machine learning algorithms might discover—to use a made-up example—that the Amish generally are good credit risks but, say, Episcopalians are not. Even if this example were true, that knowledge could not be used in computing a credit score because US law prevents discrimination on the basis of religion or other protected classes. Credit score companies are also prohibited from using data that is a surrogate for these attributes, such as an applicant's subscribing to *Amish Week* magazine or, possibly, the size of someone's monthly electricity bills.

There are additional constraints on the model that credit score companies can use to calculate credit risk. If a lender declines a loan application, the lender has to provide the reasons why the applicant's score was not higher. Those reasons have to be addressable by the consumer. For example, Jennings explained, an applicant might be told, "Your score was low because you've been late paying off your credit cards eight times in the past year," a factor that the applicant can improve in the future.

But suppose FICO's manually created models turn out to be less predictive of credit risk than a machine learning system would be? Jennings says that they have tested this and found the differences between the manual and machine learning models to be insignificant. But the promise of machine learning is that there are times when the machine's inscrutable models may be more accurately predictive than manually constructed, human-intelligible ones.

As such systems become more common, the demand for keeping their results understandable is growing. It's easy to imagine a patient wanting to know why some future version of Deep Patient has recommended that she stop eating high-fat foods, or that she preemptively get a hysterectomy. Or a job applicant might want to know whether her race had anything to do with her being ruled out of the pool of people to interview. Or a property owner might want to know why a network of autonomous automobiles sent one of its cars through her fence as part of what that network thought was the op-

timal response to a power line falling onto a highway. Sometimes these systems will be able to report on what factors weighed the heaviest in a decision, but sometimes the answer will consist of the weightings of thousands of factors, with no one factor being dominant. These systems are likely to become more inexplicable as the models become more complex and as the models incorporate outputs from other machine learning systems.

But it's controversial. As it stands, in most fields developers generally implement these systems aiming at predictive accuracy, free of the requirement to keep them explicable. While there is a strong contingent of computer scientists who think that we will always be able to wring explanations out of machine learning systems, what counts as an explanation, and what counts as understanding, is itself debatable.[46] For example, the *counterfactual* approach proposed by Sandra Wachter, Brent Mittelstadt, and Chris Russell at Oxford could discover whether, say, race was involved in why someone was put into the "do not insure" bin by a machine learning application: in the simplest case, resubmit the same application with only the race changed, and if the outcome changes, then you've shown race affected the outcome.[47] It does not at all take away from the usefulness of the counterfactual approach to point out that it produces a very focused and minimal sense of "explanation," and even less so of "understanding."

In any case, in many instances, we'll accept the suggestions of these systems if their performance records are good, just as we'll accept our physician's advice if she can back it up with a study we can't understand that shows that a treatment is effective in a high percentage of cases—and just as many of us already accept navigation advice from the machine learning–based apps on our phones without knowing how those apps come up with their routes. The riskier or more inconvenient the medical treatment, the higher the probability of success we'll demand, but the justification will be roughly the same: a good percentage of people who follow this advice do well. That's why we took aspirin—initially in the form of willow bark—for thousands of years before we understood why it works.

As machine learning surpasses the predictive accuracy of old-style models, and especially as we butt our heads against the wall of inexplicability, we are coming to accept a new model of models, one that reflects a new sense of how things happen.

Four New Ways of Happening

Suppose someday in the near future your physician tells you to cut down on your potassium intake; no more banana smoothies for you. When you ask why, she replies that Deep Asclepius—a deep learning system I've made up—says you fit the profile of people who are 40 percent more likely to develop Parkinson's disease at some point in their lives if they take in too much potassium (which I'm also making up).

"What's that profile?" you may ask.

Your physician explains: "Deep Asclepius looks at over one thousand pieces of data for each person, and Parkinson's is a complex disease. We just don't know why those variables combine to suggest that you are at risk."

Perhaps you'll accept your physician's advice without asking about her reasons, just as you tend to accept it when your physician cites studies you're never going to look up and couldn't understand if you did. In fact, Deep Asclepius's marketers will probably forestall the previous conversation by turning the inexplicability of its results into a positive point: "Medical treatment that's as unique as you are . . . and just as surprising!"

Casual interactions such as these will challenge the basic assumptions of our past few thousand years of creating models.

First, we used to assume that we humans made the models: in many cases (but not all, as we've seen) we came up with the simplified conceptual model first, and then we made a working model. But *deep learning's models are not created by humans*, at least not directly.[48] Humans choose the data and feed it in, humans head the system toward a goal, and humans can intercede to tune the weights and the outcomes. But humans do not necessarily tell the machine

what features to look for. For example, Google fed photos that included dumbbells into a machine learning system to see if it could pick out the dumbbells from everything else in the scene. The researchers didn't give the system any characteristics of dumbbells to look for, such as two disks connected by a rod. Yet without being told, the system correctly abstracted an image of two disks connected by a bar. On the other hand, the image also included a muscular arm holding the dumbbell, reflecting the content of the photos in the training set.[49] (We'll talk in the final chapter about whether that was actually a mistake.)

Because the models deep learning may come up with are not based on the models we have constructed for ourselves, they can be opaque to us. This does not mean, however, that deep learning systems escape human biases. As has become well known, they can reflect and even amplify the biases in the data itself. If women are not getting hired for jobs in tech, a deep learning system trained on existing data is likely to "learn" that women are not good at tech. If black men in America are receiving stiffer jail sentences than white men in similar circumstances, the training based on that data is very likely to perpetuate that bias.[50]

This is not a small problem easily solved. Crucially, it is now the subject of much attention, research, and development.

The second assumption about models now being challenged comes from the fact that our conceptual models cover more than one case; that's what makes them models. We have therefore tended to construct them out of general principles or rules: Newton's laws determine the paths of comets, lowering prices tends to increase sales, and all heavenly bodies move in circles, at least according to the ancient Greeks. Principles find simpler regularities that explain more complex particulars. But *deep learning models are not generated premised on simplified principles, and there's no reason to think they are always going to produce them*, just as A/B testing may not come up with any generalizable rules for how to make ads effective.

Sometimes a principle or at least a rule of thumb does emerge from a deep learning system. For example, in a famous go match between Lee Sedol, a world-class master, and Google's AlphaGo, the computer

initially played aggressively. But once AlphaGo had taken over the left side of the board, it started to play far more cautiously. This turned out to be part of a pattern: when AlphaGo is 70 percent confident it's going to win, it plays less aggressively. Perhaps this is a generalizable heuristic for human go players as well.[51] Indeed, in 2017, Google launched a program that brings together human players and AlphaGo so that the humans can learn from the machine.[52]

A later version of AlphaGo took the next step. Rather than training AlphaGo on human games of go, the programmers fed in nothing but the rules of the game and then had the machine play itself. After just three days, the system so mastered the game that it was able to beat the prior version of AlphaGo a hundred games out of a hundred.[53] When experts studied the machine-vs.-machine games that Google published, some referred to the style of play as "alien."[54]

Isn't that the literal truth?

If so, it's because of the third difference: *deep learning systems do not have to simplify the world to what humans can understand.*

When we humans build models for ourselves, we like to find the general principles that govern the domain we're modeling. Then we can plug in the specifics of some instance and read out the date and time of an eclipse or whether the patient has type 2 diabetes. Deep learning systems typically put their data through artificial neural networks to identify the factors (or "dimensions") that matter and to discern their interrelationships. They typically do this several times, sometimes making the relationships among the pieces understandable only by understanding the prior pass, which may have surpassed our understanding on its own.

The same holds for the data we input in order to get, say, a diagnosis from my hypothetical Deep Asclepius system. Deep Asclepius doesn't have to confine itself to the handful of factors a patient is typically asked to list on a three-page form while sitting in the waiting room. It can run the patient's lifetime medical record against its model, eventually even pulling in, perhaps, environmental data, travel history, and education records, noting relationships that might otherwise have been missed (and assuming privacy has been

waived). Simplification is no longer required to create a useful working model.

The success of deep learning suggests to us that the world does not separate into neatly divided events that can be predicted by consulting a relative handful of eternal laws. The comet crossing paths with Jupiter, Saturn, and the sun is not a three-body or four-body problem but rather an all-body problem, for, as Newton well knew, every gravitational mass exerts some pull on every other. Calculating a comet's path by computing the gravitational effect of three massive bodies is a convenient approximation that hides the alien complexity of the truth.

As we gasp at what our machines can now do, we are also gasping at the clear proof of what we have long known but often suppressed: our old, oversimplified models were nothing more than the rough guess of a couple of pounds of brains trying to understand a realm in which everything is connected to, and influenced by, everything.

Fourth, where we used to assume that our conceptual models were stable if not immutable, everything being connected to everything means that *machine learning's model can constantly change*. Because most of our old models were based on stable principles or laws, they were slower to change. The classic paradigm for this was put forward by Thomas Kuhn in his 1962 book *The Structure of Scientific Revolutions*. Historically, Kuhn says, a science's overarching model (which he calls a paradigm) maintains itself as data piles up that doesn't fit very well.[55] At some point—it's a nonlinear system—a new paradigm emerges that fits the anomalous data, as when germ theory replaced the long-held idea that diseases such as malaria were caused by bad air. But changes in machine learning models can occur simply by retraining them on new data. Indeed, some systems learn continuously. For example, our car navigation systems base our routes on real-time information about traffic and can learn from that data that Route 128 tends to get backed up around four o'clock in the afternoon. This can create a feedback loop as the navigation system directs people away from Route 128 at that time, perhaps

reducing the backups. These feedback loops let the model constantly adjust itself to changing conditions and optimize itself further.

As we'll see, this reveals a weakness in our traditional basic strategy for managing what will happen, for the elements of a machine learning model may not have the sort of one-to-one relationship that we envision when we search for the right "lever" to pull. When everything affects everything else, and when some of those relationships are complex and nonlinear—that is, tiny changes can dramatically change the course of events—butterflies can be as important as levers.

Overall, these changes mean that while models have been the stable frameworks that enable explanation, *now we often explain something by trying to figure out the model our machines have created*.

The only real continuity between our old types of models and our new ones is that both are representations of the world. But one is a representation that we have created based on our understanding, a process that works by reducing the complexity of what it encounters. The other is generated by a machine we have created, and into which we have streamed oceans of data about everything we have thought might possibly be worth noticing. The source, content, structure, and scale of these two types of representations are vastly, disconcertingly different.

Explanation Games

"JAL 123 was twelve minutes into its flight when a bang was heard on the flight deck."

On August 12, 1985, thirty-two minutes after that, the pilots lost their struggle to keep the plane aloft as its right wingtip clipped a mountain. The Boeing 747 came down with such force that three thousand trees in its path were destroyed. Of its 509 passengers, 505 were killed. It is to this day the plane crash that claimed the most victims.[56]

The task facing the investigators who arrived from multiple organizations and countries was made more difficult by the impend-