3

Prediction Machine Magic

What do Harry Potter, Snow White, and Macbeth have in common? These characters are all motivated by a prophecy, a prediction. Even in *The Matrix*, a film seemingly about intelligent machines, the human characters' belief in predictions drives the plot. From religion to fairy tales, knowledge of the future is consequential. Predictions affect behavior. They influence decisions.

The ancient Greeks revered their many oracles for an apparent ability to predict, sometimes in riddles that fooled the questioners. For example, King Croesus of Lydia was considering a risky assault on the Persian Empire. The king did not trust any particular oracle, so he decided to test each before asking for advice about attacking Persia. He sent messengers to each oracle. On the hundredth day, the messengers were to ask the various oracles what Croesus was doing *at that moment*. The oracle at Delphi predicted most accurately, so the king asked for and trusted its prophecy.[1]

As in Croesus's case, predictions can be about the *present*. We predict whether a current credit card transaction is legitimate or

fraudulent, whether a tumor in a medical image is malignant or benign, whether the person looking into the iPhone camera is the owner or not. Despite its Latin root verb (*praedicere*, meaning to make known beforehand), our cultural understanding of prediction emphasizes the ability to see otherwise hidden information, whether in the past, present, or future. The crystal ball is perhaps the most familiar symbol of magical prediction. While we may associate crystal balls with fortune-tellers predicting someone's future wealth or love life, in *The Wizard of Oz*, the crystal ball allowed Dorothy to see Auntie Em in the present. This brings us to our definition of prediction:

PREDICTION is the process of filling in missing information. Prediction takes information you have, often called "data," and uses it to generate information you don't have.

The Magic of Prediction

Several years ago, Avi (one of the authors) noticed a large, unusual transaction in a Las Vegas casino on his credit card. He hadn't been in Las Vegas. He had only been there once a long time before; the losing bet of gambling doesn't appeal to his economist way of seeing the world. After an extensive conversation, his card provider reversed the transaction and replaced the card.

Recently, a similar problem occurred. Someone had used Avi's credit card for a purchase. This time Avi didn't see it on his statement and didn't have to deal with the painstaking process of explaining it to a polite but firm customer service representative. Instead, he received a proactive call that his card had been compromised and that a new card was already in the mail.

The credit card provider had accurately inferred, based on Avi's spending habits and a myriad of other available data, that the transaction was fraudulent. The credit card company was so confident that they did not even block his card for a few days while they carried out

an investigation. Instead, like magic, the company sent a replacement without his having to do anything. Of course, the credit card provider did not have a crystal ball. It had data and a good predictive model: a prediction machine. Better prediction allowed it to reduce fraud while, as Ajay Bhalla, Mastercard's president of enterprise risk and security, put it, "solving a major consumer pain point of being falsely declined."[2]

Business applications are well aligned with our definition of prediction as the process of filling in missing information. Credit card networks find it is useful to know whether a recent credit card transaction is fraudulent. The card network uses information about past fraudulent (and nonfraudulent) transactions to predict whether a particular recent transaction is fraudulent. If so, then the credit card provider can prevent future transactions on that card and, if the prediction is made quickly enough, then, perhaps even the current one.

This notion—taking information of one kind and turning it into information of another kind—is at the heart of one of AI's recent main achievements: language translation, a goal that has been around for all of human civilization, even enshrined in the millennia-old story of the Tower of Babel. Historically, the approach to automatic language translation was to hire a linguist—an expert on the rules of language—to exposit rules and translate them into a way they could be programmed.[3] This is how, for instance, you might take a Spanish phrase and, beyond simply substituting word for word, understand that you need to swap the order of nouns and adjectives to make it a readable English sentence.

The recent advances in AI, however, have enabled us to recast translation as a prediction problem. We can see the seemingly magical nature of the use of prediction for translation in the sudden change in the quality of Google's translation service. Ernest Hemingway's *The Snows of Kilimanjaro* begins beautifully:

> Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa.

One day in November 2016, in translating a Japanese version of Hemingway's classic short story into English via Google, Professor Jun Rekimoto, a computer scientist at the University of Tokyo, read:

> Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa.

The next day, the Google translation read:

> Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa.

The difference was stark. Overnight, the translation had gone from clearly automated and clunky to a coherent sentence, from someone struggling with a dictionary to seemingly fluent in both languages.

Admittedly, it wasn't quite at the Hemingway level, but the improvement was extraordinary. Babel appeared to have returned. And this change was no accident or quirk. Google had revamped the engine underlying its translation product to take advantage of the recent advances in AI that are our focus here. Specifically, Google's translation service now relied on deep learning to supercharge prediction.

Language translation from English to Japanese is about predicting the Japanese words and phrases that match the English. The missing information to be predicted is the set of Japanese words and the order in which they belong. Take data from a foreign language and predict the correct set of words in the right order in a language you know, and then you can understand another language. Do it really well, and you might not realize translation is taking place at all.

Companies have wasted no time in putting this magical technology to commercial use. For example, over 500 million people in China already use a deep learning–powered service developed by iFlytek to translate, transcribe, and communicate using natural language. Landlords use it to communicate with tenants in other languages, hospital patients use it to communicate with robots for directions, doctors use it to dictate a patient's medical details, and

drivers use it to communicate with their vehicles.[4] The more the AI is used, the more data it collects, the more it learns, and the better it becomes. With so many users, the AI is improving rapidly.

How Much Better Is Prediction Than It Used to Be?

The changes in Google Translate illustrate how machine learning (of which deep learning is a subfield) has dramatically reduced the costs of quality-adjusted prediction. For the same cost in terms of computational capacity, Google can now provide higher-quality translations. The cost of producing the same quality of prediction has dropped significantly.

Innovations in prediction technology are having an impact on areas traditionally associated with forecasting, such as fraud detection. Credit card fraud detection has improved so much that credit card companies detect and address fraud before we notice anything amiss. Still, this improvement seems incremental. In the late 1990s, the leading methods caught about 80 percent of fraudulent transactions.[5] These rates improved to 90–95 percent in 2000 and to 98–99.9 percent today.[6] That last jump is a result of machine learning; the change from 98 percent to 99.9 percent has been transformational.

The change from 98 percent to 99.9 percent might *seem* incremental, but small changes are meaningful if mistakes are costly. An improvement from 85 percent to 90 percent accuracy means that mistakes fall by one-third. An improvement from 98 percent to 99.9 percent means mistakes fall by a factor of twenty. An improvement of twenty no longer seems incremental.

The drop in the cost of prediction is transforming many human activities. Just as the first applications of computing applied to familiar arithmetic problems like census tabulations and ballistics tables, many of the first applications of inexpensive prediction from machine learning applied to classic prediction problems. In addition to fraud detection, these included creditworthiness, health insurance, and inventory management. Creditworthiness involved predicting the

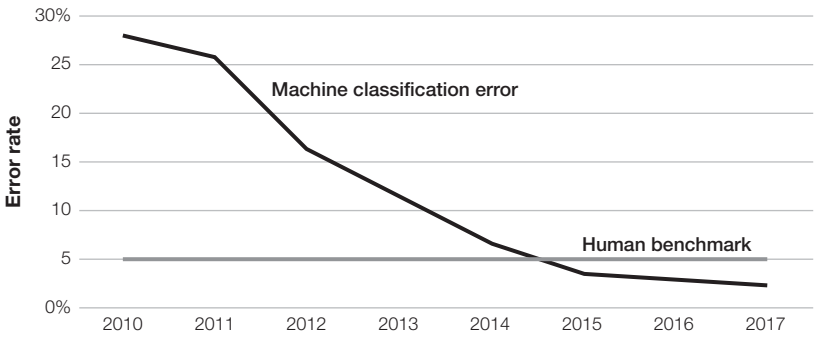
likelihood that someone would pay back a loan. Health insurance involved predicting how much an individual would spend on medical care. Inventory management involved predicting how many items would be in a warehouse on a given day.

More recently, entirely new classes of prediction problems emerged. Many were nearly impossible before the recent advances in machine intelligence technology, including object identification, language translation, and drug discovery. For example, the ImageNet Challenge is a high-profile annual contest to predict the name of an object in an image. Predicting the object in an image can be a difficult task, even for humans. The ImageNet data contains a thousand categories of objects, including many breeds of dog and other similar images. It can be difficult to tell the difference between a Tibetan mastiff and a Bernese mountain dog, or between a safe and a combination lock. Humans make mistakes around 5 percent of the time.[7]

Between the first year of the competition in 2010 to the final contest in 2017, prediction got much better. Figure 3-1 shows the accuracy of the contest winners by year. The vertical axis measures the error rate, so lower is better. In 2010, the best machine predictions made mistakes in 28 percent of images. In 2012, the contestants used deep learning for the first time and the error rate plunged to 16 percent. As Princeton professor and computer scientist Olga Russakovsky notes,

FIGURE 3-1

Image classification error over time



"2012 was really the year when there was a massive breakthrough in accuracy, but it was also a proof of concept for deep learning models, which had been around for decades."[8] Rapid improvements in the algorithms continued, and a team beat the human benchmark in the competition for the first time in 2015. By 2017, the vast majority of the thirty-eight teams did better than the human benchmark, and the best team had fewer than half as many mistakes. Machines could identify these types of images better than people.[9]

The Consequences of Cheap Prediction

The current generation of AI is a long way from the intelligent machines of science fiction. Prediction does not get us HAL from *2001: A Space Odyssey*, Skynet from *The Terminator*, or C3PO from *Star Wars*. If modern AI is just prediction, then why is there so much fuss? The reason is because prediction is such a foundational input. You might not realize it, but predictions are everywhere. Our businesses and our personal lives are riddled with predictions. Often our predictions are hidden as inputs into decision making. Better prediction means better information, which means better decision making.

Prediction is "intelligence" in the espionage sense of "obtaining of useful information."[10] Machine prediction is artificially generated useful information. Intelligence matters. Better predictions lead to better outcomes, as we illustrated with the fraud-detection example. As the cost of prediction continues to fall, we are discovering its usefulness for a remarkably broad range of *additional* activities and, in the process, enabling all sorts of things, like machine language translation, that were previously unimaginable.

KEY POINTS

- Prediction is the process of filling in missing information. Prediction takes information you have, often called "data," and uses it to generate information you don't have. In addition to

generating information about the future, prediction can generate information about the present and the past. This happens when prediction classifies credit card transactions as fraudulent, a tumor in an image as malignant, or whether a person holding an iPhone is the owner.

- The impact of small improvements in prediction accuracy can be deceptive. For example, an improvement from 85 percent to 90 percent accuracy seems more than twice as large as from 98 percent to 99.9 percent (an increase of 5 percentage points compared to 2). However, the former improvement means that mistakes fall by one-third, whereas the latter means mistakes fall by a factor of twenty. In some settings, mistakes falling by a factor of twenty is transformational.
- The seemingly mundane process of filling in missing information can make prediction machines seem magical. This has already happened as machines see (object recognition), navigate (driverless cars), and translate.