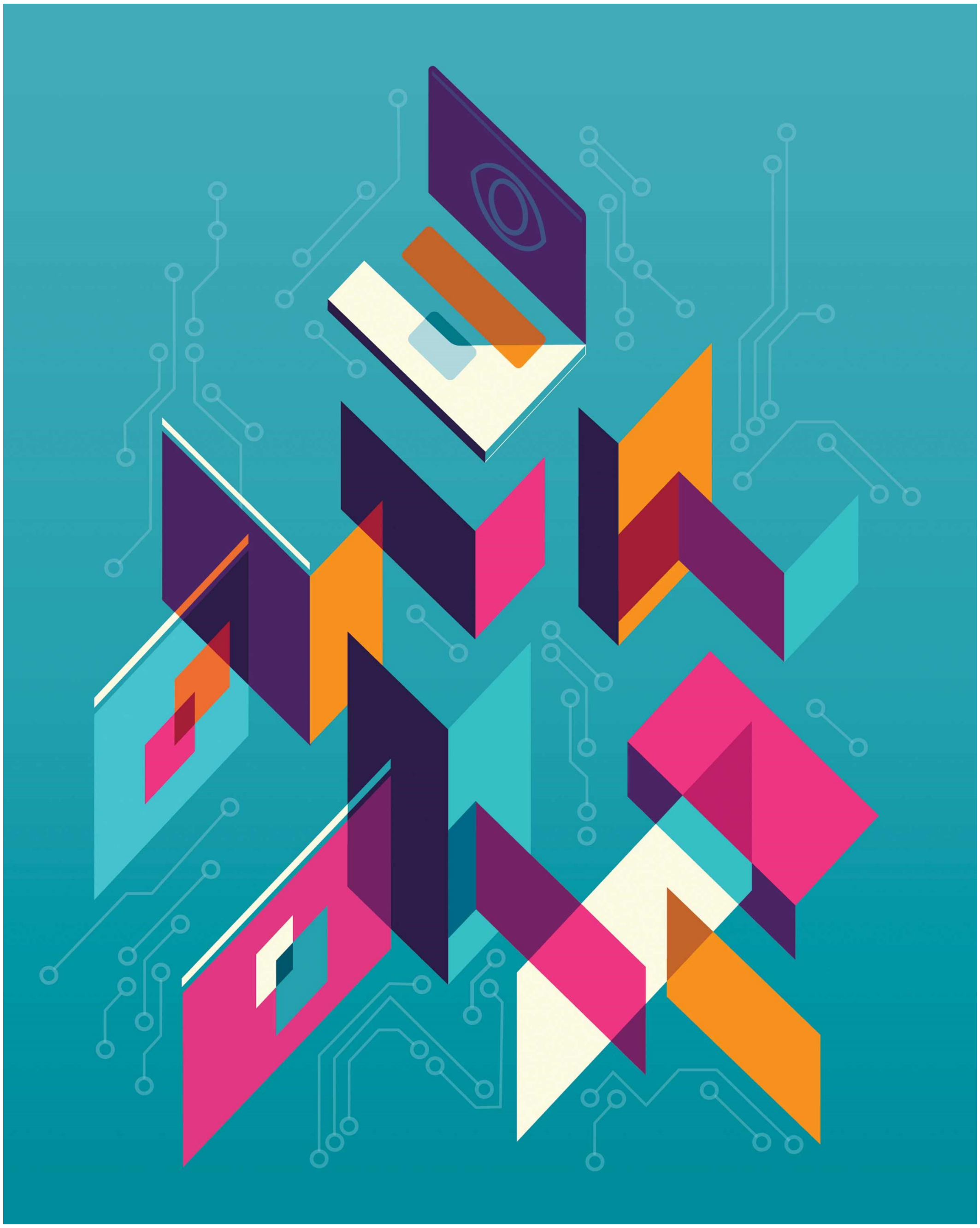

How to Win with Machine Learning

hbr.org

Executive Summary

Many companies can dramatically improve their products and services by using machine learning—an application of artificial intelligence that involves generating predictions from data inputs. Amazon, Google, and other tech giants are already experts at taking advantage of this technology. Smaller enterprises and late entrants, however, may be unsure how to do likewise to gain market share for themselves.

This article suggests that early movers will be successful if they have enough training data to make accurate predictions and if they can improve their algorithms by quickly incorporating feedback derived from customers' behavior. Latecomers will need a different approach to be competitive: The secret for them is to find untapped sources of training or feedback data, or to differentiate themselves by tailoring predictions to a special niche.

Idea in Brief

The Challenge

As more companies deploy machine learning for AI-enabled products and services, they face the challenge of carving out a defensible market position, especially if they are latecomers.

How to Get Ahead

The most successful AI users capture a good pool of training data early and then exploit feedback data to open up a value gap—in terms of prediction quality—between themselves and later movers.

How to Catch Up

Latecomers can still secure a foothold if they can find sources of superior training data or feedback data, or if they tailor their predictions to a specific niche.

Leer en español

The past decade has brought tremendous advances in an exciting dimension of artificial intelligence—machine learning. This technique for taking data inputs and turning them into predictions has enabled tech giants such as Amazon, Apple, Facebook, and Google to dramatically improve their products. It has also spurred start-ups to launch new products and platforms, sometimes even in competition with Big Tech.

Consider BenchSci, a Toronto-based company that seeks to speed the drug development process. It aims to make it easier for scientists to find needles in haystacks—to zero in on the most crucial information embedded in pharma companies' internal databases and in the vast wealth of published scientific research. To get a new drug candidate into clinical trials, scientists must run costly and time-consuming experiments. BenchSci realized that scientists could conduct fewer of these—and achieve greater success—if they applied better insights from the huge number of experiments that had already been run.

Indeed, BenchSci found that if scientists took advantage of machine learning that read, classified, and then presented insights from scientific research, they could halve the number of experiments normally required to advance a drug to clinical trials. More specifically, they could use the technology to find the right biological reagents—essential substances for influencing and measuring protein expression. Identifying those by combing through the published literature rather than rediscovering them from scratch helps significantly cut the time it takes to produce new drug candidates. That adds up to potential savings of over $17 billion annually, which, in an industry where the returns to R&D have become razor-thin, could transform the market. In addition, many lives could be saved by bringing new drugs to market more quickly.

What is remarkable here is that BenchSci, in its specialized domain, is doing something akin to what Google has been doing for the whole of the internet: using machine learning to lead in search. Just as Google can help you figure out how to fix your dishwasher and save you a long trip to the library or a costly repair service, BenchSci helps scientists identify a suitable reagent without incurring the trouble or expense of excessive research and experimentation. Previously, scientists would often use Google or PubMed to search the literature (a process that took days), then read the literature (again spending days), and then order and test three to six reagents before choosing one (over a period of weeks). Now they search BenchSci in minutes and then order and test one to three reagents before choosing one (conducting fewer tests over fewer weeks).

Many companies are already working with AI and are aware of the practical steps for integrating it into their operations and leveraging its power. But as that proficiency grows, companies will need to consider a broader issue: How do you take advantage of machine learning to create a defensible moat around the business—to create something that competitors can't easily imitate? In BenchSci's case, for instance, will its initial success attract competition from Google—and if so, how does BenchSci retain its lead?

In the following pages, we explain how companies entering industries with an AI-enabled product or service can build a sustainable competitive advantage and raise entry barriers against latecomers. We note that moving early can often be a big plus, but it's not the whole story. As we discuss, late adopters of the new technology can still advance—or at least recover some lost ground—by finding a niche.

Making Predictions with AI

Businesses use machine learning to recognize patterns and then make predictions—about what will appeal to customers, improve operations, or help make a product better. Before you can build a strategy around such predictions, however, you must understand the inputs necessary for the prediction process, the challenges involved in getting those inputs, and the role of feedback in enabling an algorithm to make better predictions over time.

A prediction, in the context of machine learning, is an information output that comes from entering some data and running an algorithm. For example, when your mobile navigation app serves up a prediction about the best route between two points, it uses input data on traffic conditions, speed limits, road size, and other factors. An algorithm is then employed to predict the fastest way to go and the time that will take.

The key challenge with any prediction process is that training data—the inputs you need in order to start getting reasonable outcomes—has to be either created (by, say, hiring experts to classify things) or procured from existing sources (say, health records). Some kinds of data are easy to acquire from public sources (think of weather and map information). Consumers may also willingly supply personal data if they perceive a benefit from doing so. Fitbit and Apple Watch users, for example, allow the companies to gather metrics about their exercise level, calorie intake, and so forth through devices that users wear to manage their health and fitness.

Moving early can often be a big plus, but it's not the whole story.

Obtaining training data to enable predictions can be difficult, however, if it requires the cooperation of a large number of individuals who do not directly benefit from providing it. For instance, a navigation app can collect data about traffic conditions by tracking users and getting reports from them. This allows the app to identify likely locations for traffic jams and to alert other drivers who are heading toward them. But drivers already caught in the snarls get little direct payoff from participating, and they may be troubled by the idea that the app knows where they are at any moment (and is potentially recording their movements). If people in traffic jams decline to share their data or actually switch off their geolocators, the app's ability to warn users of traffic problems will be compromised.

Another challenge may be the need to periodically update training data. This isn't always an issue; it won't apply if the basic context in which the prediction was made stays constant. Radiology, for example, analyzes human physiology, which is generally consistent from person to person and over time. Thus, after a certain point, the marginal value of an extra record in the training database is almost zero. However, in other cases algorithms may need to be frequently updated with completely new data reflecting changes in the underlying environment. With navigational apps, for instance, new roads or traffic circles, renamed streets, and similar changes will render the app's predictions less accurate over time unless the maps that form part of the initial training data are updated.

In many situations, algorithms can be continuously improved through the use of feedback data, which is obtained by mapping actual outcomes to the input data that generated predictions of those outcomes. This tool is particularly helpful in situations where there can be considerable variation within clearly defined boundaries. For instance, when your phone uses an image of you for security, you will have initially trained the phone to recognize you. But your face can change significantly. You may or may not be wearing glasses. You may have gotten a new hairstyle, put on makeup, or gained or lost weight. Thus the prediction that you are you may become less reliable if the phone relies solely on the initial training data. But what actually happens is that the phone updates its algorithm using all the images you provide each time you unlock it.

Creating these kinds of feedback loops is far from straightforward in dynamic contexts and where feedback cannot be easily categorized and sourced. Feedback data for the smartphone face-recognition app, for example, creates better predictions only if the sole person inputting facial data is the phone's owner. If other people look similar enough to get into the phone and continue using it, the phone's prediction that the user is the owner becomes unreliable.

It can also be dangerously easy to introduce biases into machine learning, especially if multiple factors are in play. Suppose a lender uses an AI-enabled process to assess the credit risk of loan applicants, considering their income level, employment history, demographic characteristics, and so forth. If the training data for the algorithm discriminates against a certain group—say, people of color—the feedback loop will perpetuate or even accentuate that bias, making it increasingly likely that applicants of color are rejected. Feedback is almost impossible to incorporate safely into an algorithm without carefully defined parameters and reliable, unbiased sources.

Building Competitive Advantage in Prediction

In many ways, building a sustainable business in machine learning is much like building a sustainable business in any industry. You have to come in with a sellable product, carve out a defensible early position, and make it harder for anyone to come in behind you. Whether you can do that depends on your answers to three questions:

1. Do you have enough training data?

At the get-go, a prediction machine needs to generate predictions that are good enough to be commercially viable. The definition of "good enough" might be set by regulation (for example, an AI for making medical diagnoses must meet government standards), usability (a chatbot has to work smoothly enough for callers to respond to the machine rather than wait to speak to a human in the call center), or competition (a company seeking to enter the internet search market needs a certain level of predictive accuracy to compete with Google). One barrier to entry, therefore, is the amount of time and effort involved in creating or accessing sufficient training data to make good-enough predictions.

This barrier can be high. Take the case of radiology, where a prediction machine needs to be measurably better than highly skilled humans in order to be trusted with people's lives. That suggests that the first company to build a generally applicable AI for radiology (one that can read any scanned image) will have little competition at first because so much data is needed for success. But the initial advantage may be short-lived if the market is growing rapidly, because in a fast-growing market the payoff from having access to the training data will probably be large enough to attract multiple big companies with deep pockets.

This, of course, means that training-data entry requirements are subject to the economics of scale, like so much else. High-growth markets attract investments, and over time this raises the threshold for the next new entrant (and forces everyone already in the sector to spend more on developing or marketing their products). Thus the more data you can train your machines on, the bigger the hurdle for anyone coming after you, which brings us to the second question.

2. How fast are your feedback loops?

Prediction machines exploit what has traditionally been the human advantage—they learn. If they can incorporate feedback data, then they can learn from outcomes and improve the quality of the next prediction.

The extent of this advantage, however, depends on the time it takes to get feedback. With a radiology scan, if an autopsy is required to assess whether a machine-learning algorithm correctly predicted cancer, then feedback will be slow, and although a company may have an early lead in collecting and reading scans, it will be limited in its ability to learn and thus sustain its lead. By contrast, if feedback data can be generated quickly after obtaining the prediction, then an early lead will translate into a sustained competitive advantage, because the minimum efficient scale will soon be out of the reach of even the biggest companies.

When Microsoft launched the Bing search engine in 2009, it had the company's full backing. Microsoft invested billions of dollars in it. Yet more than a decade later, Bing's market share remains far below Google's, in both search volume and search advertising revenue. One reason Bing found it hard to catch up was the feedback loop. In search, the time between the prediction (offering up a page with several suggested links in response to a query) and the feedback (the user's clicking on one of the links) is short—usually seconds. In other words, the feedback loop is fast and powerful.

Peter Greenwood

By the time Bing entered the market, Google had already been operating an AI-based search engine for a decade or more, helping millions of users and performing billions of searches daily. Every time a user made a query, Google provided its prediction of the most relevant links, and then the user selected the best of those links, enabling Google to update its prediction model. That allowed for constant learning in light of a constantly expanding search space. With so much training data based on so many users, Google could identify new events and new trends more quickly than Bing could. In the end, the fast feedback loop, combined with other factors—Google's continued investment in massive data-processing facilities, and the real or perceived costs to customers of switching to another engine—meant that Bing always lagged. Other search engines that tried to compete with Google and Bing never even got started.

3. How good are your predictions?

The success of any product ultimately depends on what you get for what you pay. If consumers are offered two similar products at the same price, they will generally choose the one they perceive to be of higher quality.

Prediction quality, as we've already noted, is often easy to assess. In radiology, search, advertising, and many other contexts, companies can design AIs with a clear, single metric for quality: accuracy. As in other industries, the highest-quality products benefit from higher demand. AI-based products are different from others, however, because for most other products, better quality costs more, and sellers of inferior goods survive by using cheaper materials or less-expensive manufacturing processes and then charging lower prices. This strategy isn't as feasible in the context of AI. Because AI is software-based, a low-quality prediction is as expensive to produce as a high-quality one, making discount pricing unrealistic. And if the better prediction is priced the same as the worse one, there is no reason to purchase the lower-quality one.

For Google, this is another factor explaining why its lead in search may be unassailable. Competitors' predictions often look pretty similar to Google's. Enter the word "weather" into Google or Bing, and the results will be much the same—forecasts will pop up first. But if you enter a less common term, differences may emerge. If you type in, say, "disruption," Bing's first page will usually show dictionary definitions, while Google provides both definitions and links to research papers on the topic of disruptive innovation. Although Bing can perform as well as Google for some text queries, for others it's less accurate in predicting what consumers are looking for. And there are few if any other search categories where Bing is widely seen as superior.

Catching Up

The bottom line is that in AI, an early mover can build a scale-based competitive advantage if feedback loops are fast and performance quality is clear. So what does this mean for late movers? Buried in the three questions are clues to two ways in which a late entrant can carve out its own space in the market. Would-be contenders needn't choose between these approaches; they can try both.

Identify and secure alternative data sources.

In some markets for prediction tools, there may be reservoirs of potential training data that incumbents have not already captured. Going back to the example of radiology, tens of thousands of doctors are each reading thousands of scans a year, meaning that hundreds of millions (or even billions) of new data points are available.

Early entrants will have training data from a few hundred radiologists. Of course, once their software is running in the field, the number of scans and the amount of feedback in their database will increase substantially, but the billions of scans previously analyzed and verified represent an opportunity for laggards to catch up, assuming they are able to pool the scans and analyze them in the aggregate. If that's the case, they might be able to develop an AI that makes good-enough predictions to go to market, after which they too can benefit from feedback.

Latecomers could also consider training an AI using pathology or autopsy data rather than human diagnoses. That strategy would enable them to reach the quality threshold sooner (because biopsies and autopsies are more definitive than body scans), though the subsequent feedback loop would be slower.

Alternatively, instead of trying to find untapped sources of training data, latecomers could look for new sources of feedback data that enable faster learning than what incumbents are using. (BenchSci is an example of a company that has succeeded in doing this.) By being first with a novel supply of faster feedback data, the newcomer can then learn from the actions and choices of its users to make its product better. But in markets where feedback loops are already fairly rapid and where incumbents are operating at scale, the opportunities for pulling off this approach will be relatively limited. And significantly faster feedback would likely trigger a disruption of current practices, meaning that the new entrants would not really be competing with established companies but instead displacing them.

Differentiate the prediction.

Another tactic that can help late entrants become competitive is to redefine what makes a prediction "better," even if only for some customers. In radiology, for example, such a strategy could be possible if there is market demand for different types of predictions. Early entrants most likely trained their algorithms with data from one hospital system, one type of hardware, or one country. By using training data (and then feedback data) from another system or another country, the newcomer could customize its AI for that user segment if it is sufficiently distinct. If, say, urban Americans and people in rural China tend to experience different health conditions, then a prediction machine built to diagnose one of those groups might not be as accurate for diagnosing patients in the other group.

Latecomers could look for new sources of feedback data that enable faster learning.

Creating predictions that rely on data coming from a particular type of hardware could also provide a market opportunity, if that business model results in lower costs or increases accessibility for customers. Many of today's AIs for radiology draw upon data from the most widely used X-ray machines, scanners, and ultrasound devices made by GE, Siemens, and other established manufacturers. However, if the algorithms are applied to data from other machines, the resulting predictions may be less accurate. Thus a late entrant could find a niche by offering a product tailored to that other equipment—which might be attractive for medical facilities to use if it is cheaper to purchase or operate or is specialized to meet the needs of particular customers.

CONCLUSION

The potential of prediction machines is immense, and there is no doubt that the tech giants have a head start. But it's worth remembering that predictions are like precisely engineered products, highly adapted for specific purposes and contexts. If you can differentiate the purposes and contexts even a little, you can create a defensible space for your own product. Although the devil is in the details of how you collect and use data, your salvation rests there as well.

Nonetheless, the real key to competing successfully with Big Tech in industries powered by intelligent machines lies in a question that only a human can answer: What is it that you want to predict? Of course, figuring out the answer is not easy. Doing so necessitates a deep understanding of market dynamics and thoughtful analysis of the potential worth of specific predictions and the products and services in which they are embedded. It is therefore perhaps not surprising that the lead investor in BenchSci's Series A2 financing was not one of the many local Canadian tech investors but rather an AI-focused venture capital firm called Gradient Ventures—owned by Google.

A version of this article appeared in the September–October 2020 issue of *Harvard Business Review*.

■