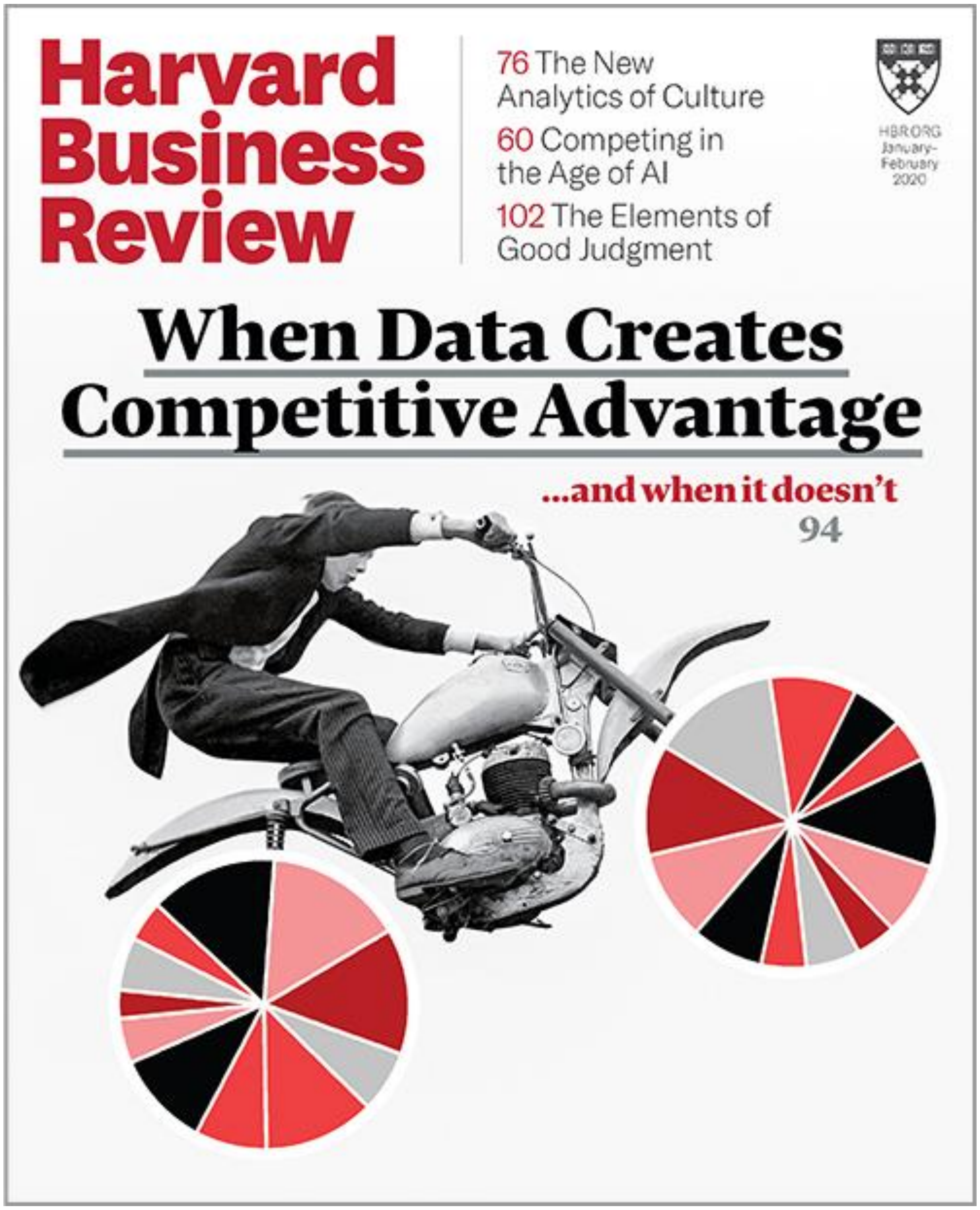
STRATEGY

When Data Creates Competitive Advantage

by Andrei Hagiu and Julian Wright

From the January–February 2020 Issue
January–February 2020 Issue


Harvard Business Review
76 The New Analytics of Culture
60 Competing in the Age of AI
102 The Elements of Good Judgment
HBR.ORG January–February 2020
When Data Creates Competitive Advantage
...and when it doesn't
94

Many executives and investors assume that it’s possible to use customer-data capabilities to gain an unbeatable competitive edge. The more customers you have, the more data you can gather, and that data, when analyzed with machine-learning tools, allows you to offer a better product that attracts more customers. You can then collect even more data and eventually marginalize your competitors in the same way that businesses with sizable network effects do. Or so the thinking goes. More often than not, this assumption is wrong. In most instances people grossly overestimate the advantage that data confers.

The virtuous cycles generated by data-enabled learning may look similar to those of regular network effects, wherein an offering—like a social media platform—becomes more valuable as more people use it and ultimately garners a critical mass of users that shuts out competitors. But in practice regular network effects last longer and tend to be more powerful. To establish the strongest competitive position, you need them *and* data-enabled learning. However, few companies are able to develop both. Nevertheless under the right conditions customer-generated data can help you build competitive defenses, even if network effects aren’t present. In this article we’ll walk you through what those conditions are and explain how to evaluate whether they apply to your business.

What Has Changed?

Companies built on data have been around for a long time. Take credit bureaus and the information aggregators LexisNexis, Thomson Reuters, and Bloomberg, just to name a few. Those companies are protected by significant barriers to entry because of the economies of scale involved in acquiring and structuring huge amounts of data, but their business models don’t involve gleaning data from customers and mining it to understand how to improve offerings.

Gathering customer information and using it to make better products and services is an age-old strategy, but the process used to be slow, limited in scope, and difficult to scale up. For automakers, consumer-packaged-goods companies, and many other traditional manufacturers, it required crunching sales data, conducting customer surveys, and holding focus groups. But the sales data often wasn’t linked to individual customers, and since surveys and focus groups were expensive and time-consuming, only data from a relatively small number of customers was collected.

That changed dramatically with the advent of the cloud and new technologies that allow firms to quickly process and make sense of vast amounts of data. Internet-connected products and services can now directly collect information on customers, including their personal details, search behavior, choices of content, communications, social media posts, GPS location, and usage patterns. After machine-learning algorithms analyze this “digital exhaust,” a company’s offerings can be automatically adjusted to reflect the findings and even tailored to individuals.

These developments make data-enabled learning much more powerful than the customer insights companies produced in the past. They do not, however, guarantee defensible barriers.

Building Moats with Data-Enabled Learning

To determine to what degree a competitive advantage provided by data-enabled learning is sustainable, companies should answer seven questions:

1. How much value is added by customer data relative to the stand-alone value of the offering?

The higher the value added, the greater the chance that it will create a lasting edge. Let’s look at a business where the value of customer data is very high: Mobileye, the leading provider of advanced driver-assistance systems (ADAS), which include collision-prevention and lane-departure warnings for vehicles. Mobileye sells its systems mainly to car manufacturers, which test them extensively before incorporating them into their products. It’s crucial for the systems to be fail-safe, and the testing data is essential to improving their accuracy. By gathering it from dozens of its customers, Mobileye has been able to raise the accuracy of its ADAS to 99.99%.

While insights from data are powerful, they don’t guarantee defensible barriers.

Conversely, the value of learning from customers is relatively low for makers of smart televisions. Some now include software that can provide personalized recommendations for shows or movies based on an individual’s viewing habits as well as what’s popular with other users. So far, consumers don’t care much about this feature (which is also offered by streaming service providers such as Amazon and Netflix). They largely consider TV size, picture quality, ease of use, and durability when making purchasing decisions. If learning from customers was a bigger factor, perhaps the smart TV business would be less competitive.

2. How quickly does the marginal value of data-enabled learning drop off?

In other words, how soon does the company reach a point where additional customer data no longer enhances the value of an offering? The more slowly the marginal value decreases, the stronger the barrier is. Note that when answering this question, you should judge the value of the learning by customers’ willingness to pay and not by some other application-specific measure, such as the percentage of chat-bot queries that could be answered correctly or the fraction of times a movie recommendation was clicked on.

Let’s say you graphed the accuracy of Mobileye’s ADAS as a function of customer usage (total miles driven by car manufacturers testing it) and found that a few manufacturers and a moderate level of testing would be sufficient to achieve, say, 90% accuracy—but that a lot more testing with a bigger set of car manufacturers would be needed to get to 99%, let alone

99.99%. Interpreting that to mean that the customer data's marginal value was rapidly decreasing would, of course, be incorrect: The value of the additional 9-percentage-point (or even a 0.99-point) improvement in accuracy remains extremely high, given the life-or-death implications. It would be difficult for any individual car manufacturer—even the largest one—to generate the necessary amount of data on its own or for any potential Mobileye competitors to replicate the data. That's why Mobileye was able to carve out a dominant position in the ADAS market, making it a highly attractive acquisition for Intel, which bought it for $15 billion in 2017.

When the marginal value of learning from customer data remains high even after a very large customer base has been acquired, products and services tend to have significant competitive advantages. You can see this with systems designed to predict rare diseases (such as those offered by RDMD) and online search engines such as Baidu and Google. Although Microsoft has invested many years and billions of dollars in Bing, it has been unable to shake Google's dominance in search. Search engines and disease-prediction systems all need huge amounts of user data to provide consistently reliable results.

A counterexample of a business where the marginal value of user data drops off quickly is smart thermostats. These products need only a few days to learn users' temperature preferences throughout the day. In this context data-enabled learning can't provide much competitive advantage. Although it launched the first smart thermostats that learn from customer behavior in 2011, Nest (acquired by Google in 2014) now faces significant competition from players such as Ecobee and Honeywell.

3. How fast does the relevance of the user data depreciate?

If the data becomes obsolete quickly, then all other things being equal, it will be easier for a rival to enter the market, because it doesn't need to match the incumbent's years of learning from data.

All the data Mobileye has accumulated over the years from car manufacturers remains valuable in the current versions of its products. So does the data on search-engine users that Google has collected over decades. Although searches for some terms may become rare over time while searches for new ones might start appearing more frequently, having years of historical search data is of undeniable value in serving today's users. Their data's low depreciation rate helps explain why both Mobileye and Google Search have proved to be very resilient businesses.

With casual social games for computers and mobile devices, however, the value of learning from user data tends to decrease quickly. In 2009 this market took off when Zynga introduced its highly successful FarmVille game. While the company was famous for relying heavily on user-data analytics to make design decisions, it turned out that the insights learned from one game did not transfer very well to the next: Casual social games are subject to fads, and user preferences shift quickly over time, making it difficult to build sustainable data-driven competitive advantages. After a few more successes, including FarmVille 2 and

CityVille, Zynga stopped producing new hits, and in 2013 it lost nearly half its user base. It was superseded by game makers like Supercell (Clash of Clans) and Epic Games (Fortnite). After reaching a peak of $10.4 billion in 2012, Zynga's market value languished below $4 billion for most of the next six years.

4. Is the data proprietary—meaning it can't be purchased from other sources, easily copied, or reverse-engineered?

Having unique customer data with few or no substitutes is critical to creating a defensible barrier. Consider Adaviv, a Boston-area start-up we've invested in, which offers a crop-management system that allows growers (now primarily of cannabis) to continuously monitor individual plants. The system relies on AI, computer-vision software, and a proprietary data-annotation technique to track plant biometrics not visible to the human eye, such as early signs of disease or lack of adequate nutrients. It then translates the data into insights that growers can use to prevent disease outbreaks and improve yields. The more growers Adaviv serves, the broader the range of variants, agricultural conditions, and other factors it can learn about, and the greater the accuracy of its predictions for new and existing customers. Contrast its situation with that of spam-filter providers, which can acquire user data relatively cheaply. That helps explain the existence of dozens of such providers.

It's important to keep in mind that technological progress can undermine a position based on unique or proprietary data. A case in point is speech-recognition software. Historically, users needed to train the software to understand their individual voices and speech patterns, and the more a person used it, the more accurate it became. This market was dominated by Nuance's Dragon solutions for many years. However, the past decade has seen rapid improvements in speaker-independent speech-recognition systems, which can be trained on publicly available sets of speech data and take minimal or no time to learn to understand a new speaker's voice. These advances have allowed many companies to provide new speech-recognition applications (automated customer service over the phone, automated meeting transcript services, virtual assistants), and they're putting increasing pressure on Nuance in its core markets.

5. How hard is it to imitate product improvements that are based on customer data?

Even when the data is unique or proprietary and produces valuable insights, it's difficult to build a durable competitive advantage if the resulting enhancements can be copied by competitors without similar data.

A couple of factors affect companies' ability to overcome this challenge. One is whether the improvements are hidden or deeply embedded in a complex production process, making them hard to replicate. Pandora, the music-streaming service, benefits from this barrier. Its offering leveraged the firm's proprietary Music Genome Project, which categorized millions of songs on the basis of some 450 attributes, allowing Pandora to customize radio stations to individual users' preferences. The more a user listens to his or her stations and rates songs up or down, the better Pandora can tailor musical selections to that user. Such customization cannot be easily imitated by any rival because it is deeply tied to the Music Genome Project. In contrast, the design improvements based on learning from the customer use of many office-productivity software products—such as Calendly for coordinating calendars and Doodle for polling people about meeting times—can be easily observed and copied. That's why dozens of companies offer similar software.

The second factor is how quickly the insights from customer data change. The more rapidly they do so, the harder they are for others to imitate. For example, many design features of the Google Maps interface can be easily copied (and they have been, by Apple Maps, among others). But a key part of Google Maps' value is its ability to predict traffic and recommend optimal routes, which is much harder to copy because it leverages real-time user data that becomes obsolete within minutes. Only companies with similarly large user bases (such as Apple in the United States) can hope to replicate that feature. Apple Maps is closing the gap with Google Maps in the United States, but not in countries where Apple has a relatively small user base.

6. Does the data from one user help improve the product for the same user or for others?

Ideally, it will do both, but the difference between the two is important. When data from one user improves the product for that person, the firm can individually customize it, creating switching costs. When data from one user improves the product for other users, this can—but may not—create network effects. Both kinds of enhancements help provide a barrier to entry, but the former makes *existing* customers very sticky, whereas the latter provides a key advantage in competing for *new* customers.

For example, Pandora was the first big player in digital music streaming but then fell behind Spotify and Apple Music, which are still growing. As we noted, Pandora's main selling point is that it can tailor stations to each user's tastes. But learning across users is very limited: An individual user's up-or-down votes allow Pandora to identify music attributes that the user likes and then serve that person songs sharing those attributes. In contrast, Spotify focused a lot more on providing users with sharing and discovery features, such as the ability to search and listen to other people's stations, thereby creating direct network effects and luring additional customers. Pandora's service remains available only in the United States (where it has a base of loyal users), while Spotify and Apple Music have become global players. And though Pandora was acquired by Sirius XM for $3.5 billion in February 2019, Spotify became a public company in April 2018 and as of early November 2019 was worth $26 billion. Clearly, customization based on learning from an individual user's data helps keep existing customers locked in, but it doesn't lead to the type of exponential growth that network effects produce.

7. How fast can the insights from user data be incorporated into products?

Rapid learning cycles make it hard for competitors to catch up, especially if multiple product-improvement cycles occur during the average customer's contract. But when it takes years or successive product generations to make enhancements based on the data, competitors have more of a chance to innovate in the interim and start collecting their own user data. So the competitive advantage from customer data is stronger when the learning from *today's* customers translates into more-frequent improvements of the product for those same customers rather than just for *future* customers of the product or service. Several of the product examples we've discussed already—maps, search engines, and AI-based crop-management systems—can be quickly updated to incorporate the learning from current customers.

A counterexample is offered by direct online lenders, such as LendUp and LendingPoint, which learn how to make better loan decisions by examining users' repayment history and how it correlates with various aspects of users' profiles and behavior. Here, the only learning that is relevant to *current* borrowers is that from *previous* borrowers, which is already reflected in the contracts and rates that current borrowers are offered. There's no reason for borrowers to care about any future learning that the lender may benefit from, since their existing contracts won't be affected. For that reason, customers don't worry about how many other borrowers will sign up when deciding whether to take a loan from a particular lender. Existing borrowers might prefer to stick with their current lenders, which know them better than other lenders do, but the market for new borrowers remains very competitive.

Does Data Confer Network Effects?

The answers to questions 6 and 7 will tell you whether data-enabled learning will create true network effects. When learning from one customer translates into a better experience for other customers *and* when that learning can be incorporated into a product fast enough to benefit its current users, customers will care about how many other people are adopting the product. The mechanism at work here is very similar to the one underlying network effects with online platforms. The difference is that platform users prefer to join bigger networks because they want more people to interact with, not because more users generate more insights that improve products.

Let's look at Google Maps again. Drivers use it in part because they expect many others to employ it too, and the more traffic data the software gathers from them, the better its predictions on road conditions and travel times. Google Search and Adaviv's AI-based crop-management system also enjoy data-enabled network effects.

Often companies can level the playing field by buying data from alternative sources.

Like regular network effects, data-enabled ones can create barriers to entry. Both types of effects present a huge cold-start, or chicken-or-egg, challenge: Businesses aiming to build regular network effects need to attract some minimum number of users to get the effects started, and those aiming to achieve data-enabled network effects need some initial amount of data to start the virtuous cycle of learning.

Despite these similarities, regular network effects and data-enabled network effects have key differences, and they tend to make advantages based on the regular ones stronger. First, the cold-start problem is usually less severe with data-enabled network effects, because buying data is easier than buying customers. Often, alternative sources of data, even if not perfect, can significantly level the playing field by removing the need for a big customer base.

Second, to produce lasting data-enabled network effects, the firm has to work constantly to learn from customer data. In contrast, as Intuit cofounder Scott Cook has often said, "products that benefit from [regular] network effects get better while I sleep." With regular network effects, interactions between customers (and possibly with third-party providers of complementary offerings) create value even if the platform stops innovating. Even if a new social network offered users objectively better features than Facebook does (for instance, better privacy protection), it would still have to contend with Facebook's powerful network effects—users want to be on the same social platform as most other users.

Third, in many cases nearly all the benefits of learning from customer data can be achieved with relatively low numbers of customers. And in some applications (like speech recognition), dramatic improvements in AI will reduce the need for customer data to the point where the value of data-enabled learning might disappear completely. Regular network effects, on the other hand, extend further and are more resilient: An additional customer still typically enhances value for existing customers (who can interact or transact with him or her), even when the number of existing customers is already very large.

CONCLUSION

As even mundane consumer products become smart and connected—new kinds of clothing, for instance, can now react to weather conditions and track mileage and vital signs—data-enabled learning will be used to enhance and personalize more and more offerings. However, their providers won't build strong competitive positions unless the value added by customer data is high and lasting, the data is proprietary and leads to product improvements that are hard to copy, or the data-enabled learning creates network effects.

In the decades ahead, improving offerings with customer data will be a prerequisite for staying in the game, and it may give incumbents an edge over new entrants. But in most cases it will not generate winner-take-all dynamics. Instead, the most valuable and powerful businesses for the foreseeable future will be those that are both built on regular network effects and enhanced by data-enabled learning, like Alibaba's and Amazon's marketplaces, Apple's App Store, and Facebook's social networks.

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