AGENDA

Advancing economics in business

oxera

January 2021

Is knowledge power? Data-enabled learning and competitive advantage

Contact
Matthew Johnson
Partner

New technologies make data-enabled learning much more powerful than the customer insights produced by such techniques in the past. They do not, however, guarantee long-lasting barriers that prevent entry by rivals. In this article, Andrei Hagiu, Associate Professor of Information Systems at Boston University, and Julian Wright, Oxera Associate and Professor of Economics at the National University of Singapore, discuss seven factors that determine whether data-enabled learning creates a sustainable competitive advantage

In recent years, much attention has been focused on the role that data can play in providing firms with a competitive advantage. Companies with more data can use that data and machine learning algorithms to produce a superior product, allowing them to attract more customers, from whom they gather more data, and so on. We see this virtuous cycle (which we call 'data-enabled learning') playing out in an ever-increasing array of digital and cloud-based products and services.

Examples include:

- Google Maps and Waze (traffic predictions improve with more drivers using them);
- Netflix, StitchFix, Spotify, Tinder, TikTok, and TrueFit (recommendations improve with more users/usage);
- speech-recognition software, virtual assistants, and chatbots (accuracy improves with more individual usage and/or more users);
- smart devices like the Nest thermostat or the Eight Sleep bed (customisation improves with more individual usage);
- autonomous vehicle systems like those being developed by Cruise, Mobileye and Waymo (accuracy improves with more usage and testing).

Indeed, by now, we've seen data-enabled learning examples in practically every sector, cutting across fields as diverse as farming (Prospera), healthcare (Notable Labs), law (Luminance) and security (VAAK). It is little wonder why 'AI' and 'big data' have become buzzwords used by corporate executives and entrepreneurs to attract investor interest.

In this article, we address what determines the extent of competitive advantage created by data-enabled learning, drawing on our recent 2020 academic working paper 'Data-enabled learning, network effects and competitive advantage',[1] and our guide for executives, 'When Data Creates Competitive Advantage',[2] published in the January/February 2020 edition of the *Harvard Business Review*. In a subsequent article, we will look at the implications for data network effects and public policies related to data-enabled learning.

Before proceeding, it is worth noting that data-enabled learning and the virtuous cycle it generates are not entirely new phenomena. Companies used to survey their customers or use focus groups, incorporating the resultant insights into the next versions of their products. However, this process was slow—it took months or even years. Today, due to the rise of cloud-based products and services, as well as the ability to efficiently store and process vast amounts of data, data-enabled learning has become much faster and more consequential. Products and services can often be improved in real time, while consumers are still using them, which was not possible previously. Moreover, because the learnings may be tied to individual customer data, these real-time product and service improvements can now be customised.

Does having a lot more data give firms a strong competitive advantage?

Not necessarily.

There are several reasons why the competitive advantage obtained from data may be overstated.

First, customer data is often just not that important for creating value relative to the many other things that a firm can improve upon. A smart TV might be able to collect voluminous amounts of data on customers' TV-watching habits, but if people do not see a lot of value in their TV recommending what shows they should watch, then data-enabled learning will not afford incumbent TV brands much of an advantage. When buying TVs, consumers place a lot more weight on superior picture quality and larger screen size than on recommendations about what to watch. This is in contrast to the data and recommendation engine that drives TikTok, which is arguably the primary factor behind its success.

A **second** related point is that the value of learning generated from *additional* usage may diminish after a modest amount of data has been collected. This makes it easier for rival companies to close the gap by generating the modest level of usage required to achieve most of the value from learning. This is likely the case for smart thermostats: such products quickly learn a user's temperature preferences throughout the day, and so data-enabled learning cannot provide much of a competitive advantage. This helps explain why Nest (acquired by Google in 2014), the first producer of smart thermostats, now faces significant competition from the likes of ecobee and Honeywell.

We can capture the above two points in Figure 2, which shows how the value of learning from customer data may increase with usage. For data-enabled learning to create a strong competitive advantage, one would need the curve to increase significantly above the stand-alone value of the product *and* continue to strongly increase even as usage increases substantially (as is the case for the linear curve in Figure 2 overleaf).

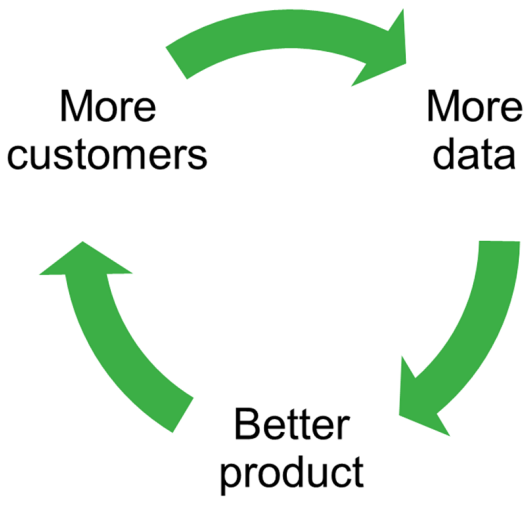

Figure 1 Virtuous cycle of learning from customer data

Source: Oxera.

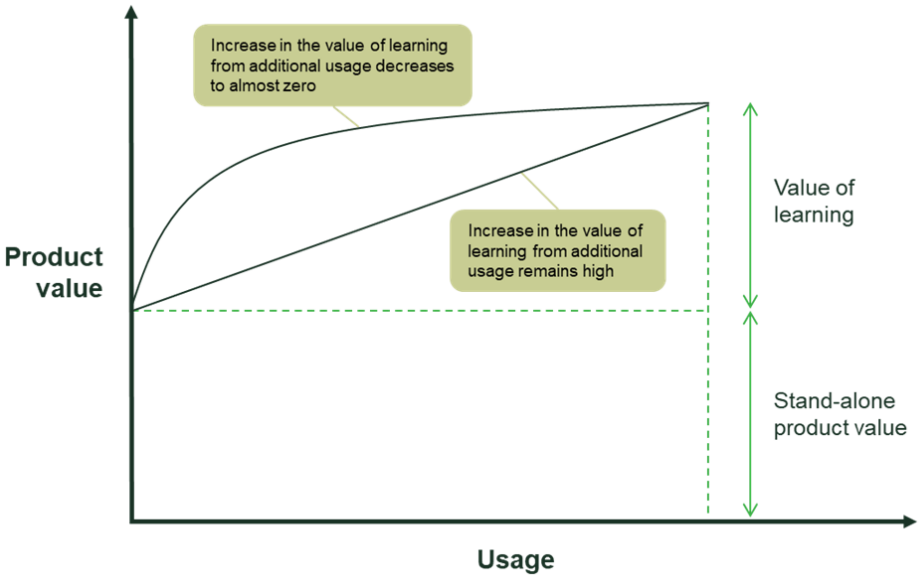

Figure 2 Illustration of how the value of learning from customer data increases with usage

Source: Oxera.

The second point is more subtle. Even though the total value of learning is the same for both curves, the linear curve, in which the value of learning from additional usage remains high throughout, may provide a stronger competitive advantage for the firm that is ahead. This is because for the other, highly concave curve, one can obtain most of the value of learning with a much lower level of usage, and so it is easier for a new entrant to get close enough to compete effectively with the market leader.

Third, for the effective value of learning from data to remain high, the relevance of data should not depreciate too quickly relative to the speed at which the company learns from new data generated from its existing customers. If data depreciates quickly, the firm that is ahead cannot use data to gain a lasting advantage.

A common reason why the value of data-enabled learning may not depreciate quickly is the importance of 'edge cases'. Edge cases are scenarios that happen infrequently, such as a dust storm in the case of autonomous vehicles, or an unusual search query in the case of online search engines. The ability to handle edge cases may not be very important for some applications (e.g. when recommending what movie to watch) but may be critical for others (e.g. in the case of autonomous vehicles or advances in rare diseases).

The role of edge cases highlights another point—that the value of learning should be judged by how much it adds to the **value derived by users**, and not by some technical measure of accuracy. When edge cases matter, little economic value may be created by data-enabled learning until the accuracy level is sufficiently close to 100%, after which further small improvements in accuracy may increase the economic value substantially (since it makes the technology safe enough to be widely deployed without further training).

To illustrate, consider Advanced Driver Assistance Systems (ADAS) such as that offered by Mobileye, which provides collision prevention and lane-departure warnings for automobiles. If we were to graph the level of accuracy as a function of usage (total test miles driven by all Mobileye's customers), the result would be something like the higher concave curve in Figure 2, where it would take a moderate level of usage to achieve 90% accuracy but a lot more usage to get to 99%, let alone 99.99%. However, that would be misleading, because the economic value of that additional 9.99%, or even 0.99%, is of course extremely high given the life-or-death implications.

Fourth, our research highlights that in determining which firm has the competitive advantage, it is not only the current stock of data that a firm may hold that matters, but also the scope to learn from future data. A firm that has less data today but more scope to learn from future data can afford to offer today's consumers greater subsidies (e.g. a high-quality product at a very low price, or even for free) because it obtains more incremental value from attracting them. Accordingly, an entrant that has a better algorithm may still be able to win against an incumbent who initially has more data.

Fifth, a very different reason why having a lot of customer data may not create a lasting competitive advantage is that a rival may be able to copy or imitate the resulting product improvements. If data-enabled learning leads to product improvements that are publicly observable, a rival can provide the same features without needing the data. This is true for a variety of software products, where the design features based on learning from customer usage can be easily observed and copied. Contrast this with product improvements that are hidden or deeply embedded in a complex production process, which makes them hard to replicate by rivals. For example, when a firm obtains detailed feedback on its call centre staff and calls from its customers, it can improve its call centre performance (e.g. assigning more experienced staff to more difficult calls, targeting both feedback and training to underperforming staff), and there is no way for these improvements to be copied by a rival firm that is entering the market.

Sixth, there may exist alternative sources of data that are relatively easy or inexpensive to obtain, which new entrants can use to train their algorithms, thereby improving their products to the point at which they can start attracting customers organically. For instance, spam filter providers can acquire user data relatively cheaply, which helps to explain the existence of dozens of such providers—and the same goes for firms that offer the service of producing captions and subtitles for videos. Even in less obvious cases, there may often be reasonable substitutes for the required data that a rival can acquire to start competing.

Consider VAAK, a Tokyo-based start-up that provides retail stores with AI-powered software enabling them to spot signs of shoplifting behaviour. It acquired 100,000 hours of shop surveillance data to train its algorithm. The problem is that similar shop surveillance footage can be relatively easily accessed or acquired by many start-ups, so VAAK's hopes of building sustainable competitive advantage really depend on how much *new* learning it can derive from each additional retail store it serves, which may be quite modest. On the other hand, a recommender system like the one powering TikTok, which takes advantage of the unique nature of its users' preference data, is much harder to compete with because there is no good substitute data set.

Relatedly, there is a growing number of publicly available data sets that firms can use to get started, and which researchers have used to develop improved algorithms for standard tasks. Consider speech-recognition software. Historically, users needed to train this type of software (the most popular of which was Nuance's Dragon voice recognition) to their individual voices and speech patterns. Such speaker-dependent software would get more accurate the more it was used by the same speaker. However, over the last decade there have been rapid improvements in speaker-independent speech recognition

systems, which require minimal or no training to understand a particular speaker's voice after having been pre-trained on existing databases. This has allowed a multitude of companies to provide speech-recognition applications such as automated customer service over the phone and automated meeting transcript services.

Extrapolating from this example, we expect to see other domains in which researchers advance algorithms in fundamental ways (possibly trained on publicly available data) to the point that they become widely available and used by hundreds or thousands of different firms. This suggests that in such domains, data is unlikely to provide a long-lasting competitive advantage.

While large existing data sets can sometimes be valuable, at other times they are not because of subtle (but important) nuances in the way data is used across different applications. For this reason, there can also be a tendency to exaggerate the ability of companies with large existing data sets to leverage that data to new applications. Consider the company x.ai,[3] which provides an AI agent that helps to set up meetings by communicating with human contacts via email. One might think that Google's massive amount of data from Google Calendar, Gmail and Search would give it an overwhelming advantage in this space. However, in a discussion of whether Google's vast data set poses a threat to x.ai, founder Dennis Mortensen remarked:[4]

> They actually have no data set, as in zero. Nothing. Because there is no agent human negotiation […] that exists in that data set. And you can't label it in the past, as in, what if the agent said this, what would the human then say. So you have to then go out and say now the agent says this to a human, what is the response—that becomes part of your training data set.

A **seventh** and final reason why data-enabled learning may only create limited competitive advantage is that data from one user may improve the product for that user but not for others. In our research, we call this 'within-user learning' to emphasise that the firm's learning from each user's history is only relevant to that user. For example, smart devices (e.g. thermostats) rely mostly on within-user learning. Such within-user learning is good from a firm's perspective because customisation creates a switching cost for *existing* customers, making it less likely they will switch to a competitor after spending time using a product. But this does not provide the firm with an advantage in competing for *new* customers, which it would only get in the presence of across-user learning (i.e. the learning from one customer helping to make the product more valuable for other customers too). Instead, it is the combination of both across-user and within-user learning that provides the most defensibility for firms.

When will data-enabled learning enhance strong competitive positions?

In conclusion, as even the most mundane consumer products become smart and connected (e.g. clothing and yoga mats),[5] data-enabled learning will be used to enhance and personalise more and more offerings. However, their providers will not build strong competitive positions unless the value added by customer data is high, lasting, proprietary, admits few substitutes, and leads to product improvements that are hard to copy.

Professor Andrei Hagiu

Professor Julian Wright

Contact

matthew.johnson@oxera.com

Matthew Johnson

wright.economics@gmail.com

Professor Julian Wright

[1] Hagiu, A. and Wright, J. (2020), 'Data-enabled learning, network effects and competitive advantage', June, https://bit.ly/3iQqbXZ.

[2] Hagiu, A. and Wright, J. (2020), 'When Data Creates Competitive Advantage', *Harvard Business Review*, January–February, https://bit.ly/3iLCuVi.

[3] See x.ai.

[4] Interview with Harry Stebbings on *The Twenty Minute VC*, 24 June 2015, at 8.17, https://bit.ly/2Y8EYUx.

[5] See Larkin, M. (2018), 'This Smart Clothing Could Transform Wearables – And Your Wardrobe', *Investor's Business Daily*, 5 May, https://bit.ly/2Y8Uks0; and smartmat.com, https://bit.ly/3a5y2x3.