

Data Centers Drive the Digital Economy, Everyday Experiences, and Innovation

HBR
Analytic
Services
BRIEFING
PAPER

Sponsored by



Data Centers and How They Are Shaping the Future of Digital Infrastructure

Data centers have been enabling critical services, commerce, education, and other aspects of daily life for years without much fanfare. In 2022, the widespread availability of generative artificial intelligence increased awareness of data center capabilities and overall impact of data centers on the digital economy, communities, and the future digital infrastructure.

Business and public sector organizations recognize the opportunity of using other modern, computer-intensive applications—as well as artificial intelligence (AI)—to improve processes and services, but they also realize they need reliable infrastructure to support advanced workloads. Data center providers such as CoreSite are rethinking facilities and operations to optimize the performance of AI and other workloads for their customers while also innovating to address concerns around energy consumption, sustainability, and physical security.

In this Harvard Business Review
Analytic Services briefing paper, unbiased analysis of trends and interviews with industry experts provides insights on topics such as evolving power solutions and cooling technologies, data center types and architectures, cloud and digital ecosystem interconnection, and the ongoing innovation to support future needs. Discussion also includes how colocation providers can help customers enable AI, focus on competitive differentiators, reduce operational costs, and evaluate data center providers for partnership.

Data Centers Drive the Digital Economy, Everyday Experiences, and Innovation

Many business executives do not realize the central role data centers play in driving the global economy and the everyday digital experiences that impact everyone. Data centers support complex enterprise and government agency use cases such as product development, platforms for artificial intelligence (AI) and generative AI (gen AI), financial forecasting and analysis, the 9-1-1 system, medical imaging, and complex health care and scientific applications, as well as many other digital use cases. Data centers are also the technology backbone for functions like cell phone use, GPS mapping, and content streaming.

THE DATA CENTER MARKET, not surprisingly, makes a significant contribution to the global economy, which the International Monetary Fund projected to be approximately \$110.06 trillion in 2024. The global data center market was valued at \$242.72 billion in 2024, with North America a dominant segment at 38.8% of the market, according to the Fortune Business Insights "Data Center Market Size" report, published in May 2025.¹ That same report predicts the global data center market will grow to \$269.79 billion by the year 2032.

And the market is evolving as well as growing. "This industry is constantly changing," says Josh Levi,

president of the Leesburg, Va.-based Data Center Coalition (DCC). "From 2012 onward, you saw tremendous demand for cloud services and migration from on-premises data centers to workloads in the cloud. That hit a fever pitch during the pandemic, and in many ways, it has accelerated. A large part of that is certainly attributable to new workloads, including gen AI, support for autonomous vehicles, and a lot of the more data-intensive uses like medical imagery. Those trends continue to play out."

That's a familiar perspective.
"Right now is an inflection point,"
says Roger Strukhoff, chief research
officer of the Rockville, Md.-based



New architectural designs are being built into new data centers, and advanced power generation and cooling technologies are coming into play.

Liquid cooling is rapidly becoming the primary approach for heat management.

Other changes to data center evolution are being driven by customer demand when they are evaluating potential data center providers, which has led to data center design and operational technologies like data center interconnect and direct-to-cloud connections.

1

Data centers have expanded the capacity, range, and flexibility of the services they provide, and they continue to do so, enabling increasingly sophisticated and interconnected client networks as well as servers capable of supporting myriad workloads.

International Data Center Authority (IDCA). "People are building these [advanced data centers] over the last few years [using servers with] more highly efficient chips to get more power out of a smaller space."

Data centers have expanded the capacity, range, and flexibility of the services they provide, and they continue to do so, enabling increasingly sophisticated and interconnected client networks as well as servers capable of supporting myriad workloads. New architectural designs are being built into new data centers, and advanced power generation and cooling technologies are coming into play, all with the goal of enabling organizations large and small, public and private, to have the IT infrastructure and processing capacity to operate more efficiently and with greater resilience.

This report will help inform and educate technology and business executives on how colocation data centers continue to support the digital services that society and enterprises rely on every day, and on how to devise an effective and efficient strategy for ensuring they have sufficient data center capacity and services. This strategy will allow companies to focus on their core business and better position themselves for the future.

Data Centers Evolve to Meet Demand

The only constant among modern data centers may be change. The role of data centers has evolved to meet customer needs and more advanced use cases, requiring a shift in the architectural/operational design. "We have seen new builds where folks have paused and changed configurations with their data center build to accommodate new workloads, so it's a very dynamic environment," says DCC's Levi. "We see traditional data centers and AI data centers and a whole lot of things in between, so there is clearly the need for flexibility."

Another architectural evolution Levi is seeing is the development of campus-style configurations for new data center locations. "You have four, six, or 12 data centers located on contiguous parcels, operated by the same developer," says

Levi. This type of configuration conveys a variety of benefits, he says, including economies of scale, the ability to combine compute capacity across facilities, and a shared workforce operating the data centers within the campus.

Power-generation and cooling technologies for data centers have also evolved considerably. "You're seeing the intensity increase in terms of the amount of power you need to run the graphics processing units—GPUs—and to eject the heat associated with those more intense workloads," says Levi. "Power continues to be used for two main things—one is compute, and the other is removing heat."

Those two things go hand in hand. "Looking at new compute workloads [like generative AI], how do you cool those servers? Some data centers may have their racks filled with rows of servers. You may also see areas within the data center or server racks where you have a lot of space between GPUs to help manage heat," says Levi.

But data centers are always striving to improve efficiency, which is measured by a power usage effectiveness (PUE) ratio. "Data centers are always trying to get to a more efficient [lower] PUE by measuring how much power consumption is for the actual data processing and how much is for AC and lights," says IDCA's Strukhoff. "Outside of core processing and networking activities, most of the power data centers use is for cooling right now. Electronics produce heat. That doesn't change."

Adds Levi, "The question for new builds is how do you maximize that density [of GPUs] where you get compute capacity without overwhelming the ability to remove that heat."

Without significantly improved efficiency, the industry's use of power is likely to be unbridled. At current growth rates around the world, particularly in the U.S., demand for electricity to power this expanding data center landscape is projected to double or triple by 2028, according to the "2024 Report on U.S. Data Center Energy Use" prepared by the Berkeley, Calif.-based Lawrence Berkeley National Laboratory for the U.S. Department of Energy in December 2024.² Data center power demand could reach between 325 and 580

terawatt hours (TWh) by 2028, representing 6.7% to 12% of the total electricity consumption in the U.S. By comparison, data center power demand in 2018 was only 60 TWh, or 1.9% of U.S. energy consumption. FIGURE 1

The prevailing effort among data center operators to lower PUE involves the use of ways in addition to air cooling to reduce server temperature. "Liquid cooling is accomplished today primarily with water. But as more and more heat is generated, companies are looking at more expensive solutions involving chemical mixtures, applied direct-to-chip, as the industry calls it, or through deep immersion," says Strukhoff.

Liquid cooling is rapidly becoming the primary approach for heat management. "Liquid cooling is a big one, particularly for organizations that have their data center located near natural resources to cool the water or cool the coolant instead of having to use energy to do the heat transfer," says John O'Neill Sr., chief innovation officer at Jefferson, Ohiobased technology consultancy Azure Innovators.

Liquid cooling technologies convey several benefits, he says, adding, "That saves a lot of money and helps the equipment to run cooler, which allows [the equipment] to last longer and not consume as much energy." Not all data centers have liquid cooling capabilities at this point, however, which could become a key consideration for companies evaluating a data center provider.

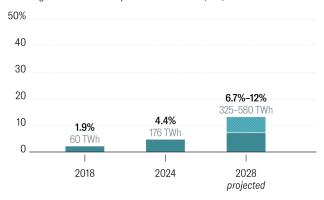
Levi sees other advanced cooling technologies now emerging. "Some data centers are piloting liquid submerged cooling to remove the heat and keep the servers at temperature," he says. "We're also seeing direct-to-chip cooling now, where you have tubes that bring the liquid cooling agent directly to the chips themselves. And, of course, there is still ambient cooling technology, by designing racks [to use] airflow to remove the heat."

Data centers are always looking to manage power costs, as well. "Energy remains one of the largest operating expenses, so there continues to be tremendous focus on optimizing use of electricity. At the same time, our [DCC] members remain focused on sustainability goals around water and power," says Levi.

FIGURE 1

Data Center Energy Demand Expected to Surge Data centers will demand increasing amounts of electricity in the coming years

Percentage of total U.S. consumption in terawatt hours (TWh)



Source: Lawrence Berkeley National Laboratory, December 2024

Azure Innovators' O'Neill agrees that exploring more environmentally sustainable ways to generate power is becoming a significant focus. "Larger organizations have done a lot to enhance their efficiency and their footprint," he says. "You see a lot of data centers being powered at least partially by renewable energy, with solar panels on the roof or a wind turbine on the property. That can help lower operational expenses associated with modern data centers. Also, with a modern hardware stack, power consumption and heat generation are much lower."

Cloud Connections

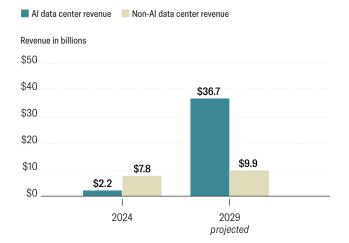
Other changes to data center evolution are being driven by customer demand when they are evaluating potential data center providers, which has led to data center design and operational technologies like data center interconnect (DCI) and direct-to-cloud connection. DCI uses networking technologies to connect multiple data centers, and thus helps facilitate sharing resources and balancing workloads, especially in a campus environment. The resulting business benefits include the ability to reduce latency, provide high service availability, simplify compliance, enhance cybersecurity, rapidly scale up or down, and streamline disaster recovery.

Direct-to-cloud connection is just that—a direct connection from the data center to a cloud service provider. "There are many advantages to this," says O'Neill. "The first that come to mind are disaster recovery, business continuity, and

FIGURE 2

Artificial Intelligence Workloads to Drive Revenue

Data center revenue from artificial intelligence workloads alone is expected to increase dramatically



Source: Cushman & Wakefield, May 2025

incident response. For conventional disaster recovery, you'd have a hot, [a] warm, and [a] cold site at a physical location nearby. Larger data centers would have a leased line or fiber connection between those locations so they could just shift operations."

Establishing a direct-to-cloud connection reduces the expense of establishing and maintaining both a primary and a backup data center. "Now cloud providers can do away with most of that expense. Cloud resources can be spun up so fast, so you can pay a small amount to have it semiprovisioned, a site that has enough storage and compute resources carved out. You're not paying the full price until you have to go in there and flip the on switch," says O'Neill.

A direct-to-cloud connection creates significant advantages in more-efficient and less-expensive disaster recovery. "Even smaller companies can figure it out," says O'Neill. "After 9/11, some companies had their backup data centers a couple of blocks from their primary and they were both wiped out, so they took months to get back."

Having a direct-to-cloud connection can also help a company more effectively respond to cybersecurity events. "If you have a breach incident, you need to have a clean network. You don't know how far the [breach has] moved laterally. There are layers like an onion," says O'Neill. "The fastest way to business continuity is to have a clean network where you can spin up virtual machines you know are not infected to get operations going."

Besides improving disaster recovery and cybersecurity breach response, direct-to-cloud connections provide reduced latency on a regular basis. "Most organizations have moved core services to cloud—their FinOps, their DevOps, whatever it is. When they move those to the cloud, the bigger the bandwidth connection, obviously the higher the performance," says O'Neill.

Supporting the Surge in Artificial Intelligence

Continuing to improve data center performance and capacity is becoming increasingly critical to support compute-intensive workloads like AI, medical imaging, and autonomous vehicles. "There are a lot of different deployments right now trying to determine the best way to maximize the benefits and meet these challenges," says Levi. "The largest parts [of the challenge] are the power required and the heat created by the density of those workloads. You have different flavors of AI workloads with different requirements."

The use of AI, gen AI, and other multicloud services continues to expand, leading to increased demand for data center services and therefore increased data center revenue, especially for those supporting AI workloads. In fact, data center revenue growth from AI workloads alone is forecast to eclipse non-AI workload revenue by 2029, according to the "Global Data Center Market Comparison" report prepared by Chicago-based commercial real estate developer Cushman & Wakefield.³ This finding demonstrates how significant AI usage will be for modern enterprises and the data centers that support them. FIGURE 2

Levi sees organizations not only using their data centers to support AI workloads but also employing AI within the data center to make operations more efficient. "Think not only of supporting AI workloads, but also data centers starting to utilize AI within their data center to move workloads between different servers or even between different data centers. [AI also has the] ability to look at real-time cooling

Most organizations have moved core services to cloud—their FinOps, their DevOps, whatever it is. When they move those to the cloud, the bigger the bandwidth connection, obviously the higher the performance.

John O'Neill Sr., chief innovation officer, Azure Innovators

and heat maps within a facility and adjust accordingly to maximize efficiency," he says. "[The data center industry] is not only the digital infrastructure that supports AI but also [is] an industry leading the way [to] utilizing AI to increase efficiencies and benefits."

The Benefits of Colocation

While the data center market continues to expand across the globe, colocation data centers captured the largest share of the market in 2024 at 34.8%, according to the Fortune Business Insights "Data Center Market" report. Colocation facilities provide flexibility and scalability, which help companies relying on colocation services—referred to as tenants—quickly and easily scale their data center capacity to accommodate surges in demand. Colocation data centers also run at a lower PUE, reducing the amount of energy required.

"In the old days, everyone had their own data center and IT department," says Strukhoff. "Then cloud computing came along, and suddenly a single facility with a lot of iron in it could serve lots of different customers. [However], some companies don't like losing control of their data for certain legal requirements, like a medical company with HIPAA regulations. So along comes colocation. 'We'll build the building, we'll get the power, we'll keep the lights on. Put your stuff in these racks and rent the space from us."

Colocation data centers allow tenant companies to maintain control of their equipment and their data. "You don't lose control, but they're doing the grunt work for you. They're also a lot more flexible. If you have an anticipated spike, you can rent a few more racks a lot quicker than building them. All the headaches involved in building these things go to them," he says.

Levi agrees that this type of data center serves increasingly important requirements. "There continues to be a strong need for [multitenant] facilities. We have data centers that have 25, 100, or 200 tenants located within [them].

That's how the multitenant models work. They are clearly an important part of [the data center] ecosystem."

He also sees flexibility within the multitenant model. "Think about the opportunities [for companies] to do partnerships," he explains. "Let's say you have a big-box retailer and a credit card company doing a deal. There's an opportunity to put their servers side by side within a multitenant facility so those two workloads can collaborate. There's a lot of appeal to that approach."

Economies of scale are also a significant aspect of multitenant data centers. "Not everybody needs the scale you get in leasing an entire data center," says Levi. "This part of the ecosystem has seen a lot of migration as some operators have moved toward the build-to-suit multitenant model to meet demand from businesses requiring data space and ability to scale up quickly. [Using a multitenant data center] frees up tenants to run their own business. It's an important part of the ecosystem and one with a lot of runway left."

Being able to shift the burden of data center operations is the most significant benefit for companies working with a colocation data center provider—often called a "colo"—according to Strukhoff. "All I have to do is send my systems there, and they hook them up. I don't have to worry about power, don't have to worry about cooling. It's almost like a bank. I put my money in a bank, and they're just holding it and providing all these services," he says.

As companies across all industries face an increasingly competitive environment, outsourcing data center operations becomes more important. "Companies want to focus on competitive advantage. They want to focus on making money," he says. "They don't see managing all this hardware as a core skill. The colos say, 'We'll do that.' Colos have the expertise."

Locating data center services closer to a customer base is also an advantage. "Because of scale, colos have lots of locations, so you can have a data center in as many regions as you want, depending on how your company is set up and where your customers are located," says Strukhoff.

Data Centers by Design

There are several major categories of data centers: enterprise, cloud, hyperscale, managed, colocation/multitenant, and edge. There are also several specialized types of data centers, such as carrier hotels, modular, and micro data centers.

Enterprise: Enterprise data centers are private, self-managed facilities that support data operations for a single large organization. These data centers may also include multiple locations around the globe to support operations closer to different customer bases. Enterprise data centers are typically architected and operated as multiple sub-data centers, each supporting critical functions. These sub-data centers include an internet data center to support web servers, an extranet data center to support business-to-business transactions, and an intranet data center to support internal business applications and operations.

Cloud: Cloud data centers are owned, operated, and managed by the large cloud providers Google, AWS, Microsoft Azure, Oracle, IBM, and others. These large cloud service providers manage all the applications, platforms, and data stored in their data centers. Most of these data centers are for the public cloud, but there are also data centers run by private cloud providers. Cloud data centers convey multiple benefits, particularly cost savings (especially when there is a direct cloud connection), convenience, and security.

Hyperscale: Hyperscale data centers are configured to handle the rising demands for data processing capacity. These massive data centers often run thousands of servers and can rapidly scale up to meet surges in demand. They serve one or two large enterprises with massive data storage and management requirements.

Managed: In a managed data center, the provider is responsible for the monitoring, maintenance, and management of network, compute, and storage resources. Managed data centers typically offer varying management levels, from fully to partially managed. Full management covers all technical and administrative details, while partial management involves a shared level of responsibility and allows the customer a higher level of control over and responsibility for its own data.

Colocation/Multitenant: Colocation or multitenant data centers—colos—are data centers that host and provide remote hands services for servers and networking equipment on behalf of multiple customers or tenants. Businesses not interested in owning and operating their own data centers can engage with a colocation data center to rent server space and request technical support services.

Companies can also rent rack space and have the colocation provider house and operate the company's own servers.

Colocation data centers are typically multitenant, meaning they host servers for multiple business customers, often as many as hundreds of tenants.

The colocation data center provider supplies the physical facility, electrical power, cooling capacity, system updates, and physical security. Colocation data centers also help businesses realize economies of scale by sharing space and power with small, medium, and large businesses and also allow businesses to connect directly with each other. Colocation data centers can also be established in a campus type of setting, where multiple colocation data centers are positioned nearby to ensure ample capacity and rapid scalability with minimal latency.

Edge: Edge data centers are the most recent development in data center architecture. These centers are architected and built as smaller data centers located closer to the business's customer base and to where the data is generated, processed, and acted on. Edge data centers are built on a concept similar to edge computing, which brings compute power closer to the systems such as cell towers using that power to expedite operations.

While still in the development phase, edge data centers will be smaller and built closer to customer operations. A company with a strong customer base in New York, Hartford, and Boston, for example, might have edge data centers built in or near those cities. This strategy allows data access with minimal latency.

It's attractive to be able to farm that out knowing you can get online right away, but if you build it yourself, it could take 18 months. 37

Roger Strukhoff, chief research officer, International Data Center Authority

Location is critical when it comes to ensuring low latency. With traditional use cases such as content streaming, low latency creates higher-quality user experiences. And as AI platforms move to inference from model training, having the AI workload closer to users is critical for real-time data applications. Locating data centers closer to the customer base can also impact cloud connectivity and cost predictability.

Partnering With a Provider

Any company interested in outsourcing its data center operations will need to make several key strategic decisions, such as location, the range of services provided, the expertise of the potential provider, and—naturally—the expense. When initially engaging with a partner, a company will also need to evaluate its latency requirements. "How quickly do they need their data, and for what purposes?" asks Levi.

Besides latency, Levi lists physical security, proximity, scalability, and cost as critical considerations. "They'll want to evaluate the levels of security they actually require, because data centers have variability in terms of security," he says. "They'll also want to consider proximity in some cases. By locating close to the users, if you have an issue with a server at 2 a.m., you can jump in your car and show up there quickly. That can be a benefit, as well."

Scalability and flexibility are also significant factors. "If you're a company that anticipates growing, how quickly are you going to grow? [You will need] the ability to reserve additional capacity and be able to scale up," says Levi. While some companies will remain relatively static, others may require the ability to scale up rapidly. "Think about a social media app where suddenly every teenager wants to access the latest thing, [so] they have to scale up really quickly."

While the basics of cost and location are clearly vital, Strukhoff stresses the importance of contract parameters such as length and flexibility, as well as the range of services the data center provider will offer. "It's business basics," he says. "How large are the instances? What [central or graphics processing unit] are [the customers] using? How much memory or storage will you have?"

While the colocation data center tenant maintains ownership of the equipment, the colocation data center operates that equipment. The data center also provides services such as direct cloud connections, an interconnection platform, deployment monitoring, intermarket data center connections, and remote hands services.

Engaging a colocation data center provider also eases capacity planning. "It's attractive to be able to farm that out knowing you can get online right away, but if you build it yourself, it could take 18 months," he says.

When an organization is evaluating the services of a potential data center provider, it should consider the following critical capabilities that provider must bring to bear.

Advanced power generation/cooling. Data center providers continue to explore advanced power-generation and cooling technologies, including renewable energy, liquid cooling, and direct-to-chip cooling. "Water is the best cooling agent out there," says Levi, citing data centers that are now submerging servers in liquid to remove heat.

Campus architecture. Designing the physical layout of data centers in a campus-type environment can facilitate interconnection, reducing latency and expediting shared workloads. "There are economies of scale in being able to centralize load latency from facility to facility," says Levi.

Capacity. Capacity is clearly a primary factor. The more server storage and processing power a data center has, the greater its capacity. "You have to calculate [capacity requirements] based on data you have," says O'Neill.

Direct cloud connection. Direct cloud connections also greatly reduce latency, reduce data egress costs, and help ensure cost predictability and business continuity. These connections have created a "huge advantage" for things like disaster recovery, says O'Neill.

Interconnection. Interconnections to other data centers can also ensure scalability, as well as greatly increasing

[Data center providers] are acquiring land now for the ability to build the future. It's an industry that's organizing its supply chain now with the expectation of having certainty in 2027 and 2028 around the equipment they will need. 3

Josh Levi, president, Data Center Coalition

efficiency by sharing workloads. Interconnections are part of data center topology, says Strukhoff. "With high-speed connections, why not?"

Scalability. Having the additional server space and actual physical space can help ensure the data center can rapidly scale up as needed. "Companies will need to secure the ability to scale up as quickly as possible with [their] service provider," says Levi.

Security and compliance. Data centers also continue to advance not only data and workload security but also redundant, diverse, and layered cybersecurity, as well as physical security of the facilities. "The big risk is cybersecurity," says Strukhoff. "The colos [can also often] provide a higher level of [physical] security."

Data Centers of the Future

The data center industry will certainly expand its capacity and range of services. "Part of the DNA within the data center industry is to not only anticipate the future but [also] to invent the future," says Levi. "This is not a backward-looking industry, not a today-looking industry. [Data center providers] are acquiring land now for the ability to build the future. It's an industry that's organizing its supply chain now with the expectation of having certainty in 2027 and 2028 around the equipment they will need." The industry isn't just working to help invent the future of IT infrastructure but is also working alongside other technology leaders to accommodate advancements in cloud computing, software-as-a-service applications, and enterprise resource planning systems.

Companies that rely on data center operators can expect increasing capacity and flexibility, says Strukhoff. "[A data center provider] gives you flexibility you need to plan," he says. "You can write your own scenario and future-proof yourself that way."

Strukhoff hopes to see more development around the globe. "There's a huge imbalance. Forty percent of data

centers are in the U.S.," he says. "There's a screaming need for more development throughout the world."

The server, power, and cooling technologies continue to advance, as well, according to Levi. "When I went through my first data center in 2003, I needed a long-sleeve shirt because it was cold," he says. "They were trying to keep everything super cool to maintain the operating environment."

He points out that hardware heat tolerance and durability have improved substantially. "Those environments can operate more warmly now, and that creates all kinds of efficiencies with energy and water and cooling. I expect that would continue," says Levi. "I am also seeing a lot of emphasis by chip developers on trying to maximize compute while also maximizing energy efficiency, the ability to run these workloads in a more-efficient way that generates less heat."

Levi also expects to see continued innovation in terms of new data center configurations, architecture, and cooling technologies. "You're going to continue to see those kinds of innovations as a business imperative," he says. And data centers will continue to play a pivotal role in driving the digital experiences that affect everyone and the entire spectrum of modern business.

Endnotes

- 1 Fortune Business Insights, "Data Center Market Report," May 2025. https://www.fortunebusinessinsights.com/data-center-market-109851.
- 2 Lawrence Berkeley National Laboratory, "2024 Report on U.S. Data Center Energy Use," prepared for the U.S. Department of Energy, December 2024. https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers.
- 3 Cushman & Wakefield, "Global Data Center Market Comparison," May 2025. https://cushwake.cld.bz/globaldatacentermarketcomparison-05-2025-global-central-en-content.



VISIT US ONLINE

hbr.org/hbr-analytic-services

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject-matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.