McKinsey & Company

Electric Power & Natural Gas Practice

# Powering AI: How real is the risk of overbuilding?

The power infrastructure being built for data centers in the United States is unlikely to become stranded—even if compute demand falls short of current projections.

*by Humayun Tai, Jesse Noffsinger, and Sam DeFabrizio*
*with Raissa Dantas*

July 2026

**'Are we building too much power?'** As the growing demand for electricity to supply data centers has come to dominate conversations in the US power sector, and forecasts for data center demand continue to increase, we are hearing that question everywhere. The question has at least two dimensions: whether players in the data center value chain are building too much data center capacity to support compute, and whether players in the power value chain are at risk of overinvesting based on that build.

While some signals are mixed, our analysis suggests that the answer for both dimensions is probably "No." In fact, for the power sector, we find that a short-term *shortfall* is likely. For power sector players, the greater near-term risk might be underbuilding rather than overbuilding.

In this article, we assess the outlook for US data center power demand and supply through 2035, drawing on data center evolution scenarios and a state-by-state analysis of grid capacity. We evaluate how a potential shortfall may be filled and examine implications and opportunities for various US power-sector players.

## Data center overbuild is unlikely

Approximately 75 percent of the anticipated US power demand growth over the next decade is expected to come from data centers, which are being built at a rate requiring the equivalent of almost 30 gigawatts (GW) of power[1] per year. Whether that growth materializes depends on whether demand for AI compute will last. Several indicators suggest that it will.

According to McKinsey's "The state of AI in 2025" survey, roughly one-third of respondents report scaling AI, 39 percent report measurable EBIT impact, and 64 percent cite innovation benefits.[2] Large language models are continuing to get better, expanding their potential applications—with no sign of an imminent technological plateau. Investments are increasingly tied to contracted capacity utilization and concentrated among scaled, investment-grade data center operators—factors that point to durable, committed growth rather than speculation. Finally, demand is anchored by sophisticated, well-capitalized hyperscalers, whose capital spending has grown approximately 50 percent annually from 2022 to 2026 while maintaining strong margins and cash flow.

While on balance these signs point to sustained growth, the timing and extent of companies' AI adoption remain somewhat uncertain. While companies' intent to use AI is high, execution is uneven. Many cite talent and cultural barriers to adoption, and 71 percent of organizations report negative implementation outcomes. At the same time, parts of the AI ecosystem exhibit bubble-like characteristics: More than half of the new unicorns in 2025 were AI-focused, including several multibillion-dollar firms without commercialized products. Outside the hyperscaler core, some AI-native players and infrastructure intermediaries are pursuing aggressive expansion with high leverage or negative cash flow. Broader macro and geopolitical factors—such as tighter capital markets, elevated power prices, or supply chain disruptions—could also dampen demand

[1] Represents approximately 20 gigawatts of incremental IT demand plus the additional electricity required for cooling systems, power distribution equipment, and reliability buffers.
[2] Respondents included 1,993 individuals across 105 nations representing the full range of regions, industries, company sizes, functional specialties, and tenures.

and slow the pace of data center build-out. These uncertainties could temper growth in compute demand.

## How might power demand play out in the next decade?

Given the uncertainty around future AI demand, players in the power sector are evaluating the risk of overbuilding and stranded assets. The concern is reasonable: even in a low data center demand scenario, the US power sector could add more than 150 GW of power capacity to supply the IT load demand by 2030 (see sidebar, “Counting gigawatts”). But unlike the fiber-optic overbuild of the early 2000s, which could serve only one purpose (delivering high-speed data), electricity infrastructure serves a broad and growing base of demand beyond data centers (including refreshing aging infrastructure), reducing the risk that new power assets become stranded.

## Counting gigawatts

**On the demand side,** data center *IT load* represents the maximum power consumption that the site could see from servers, storage, and networking equipment. It is the metric that data center operators typically use to describe their capacity, although the power sector needs multiple other metrics to understand what will be required to serve that load. Data center interconnection capacity represents the peak power a site might draw at any one time, including IT capacity, peak cooling needs, and any engineering buffers to ensure that the critical equipment is not starved for power when it is most under stress. This figure might be 50–80 percent higher than the IT capacity requested, and it’s important to understand that IT capacity (what hyperscalers report) and data center interconnection requests are not apples-to-apples metrics, although both are measured in gigawatts.

On the supply side, *generation nameplate capacity* refers to the maximum theoretical output of a power plant under ideal conditions, while *derated capacity/power output* adjusts this figure to reflect the dependable capacity that can realistically be delivered after accounting for technology-specific availability, maintenance, outages, fuel constraints, and weather conditions. For example, for a data center connected to the grid that is reported as 500 megawatts (MW), this IT load may require 900 MW of interconnection capacity. This could be served by 1,000 MW of gas plants, which have an assumed outage ratio of 10 percent. It is critical to understand which gigawatts are being discussed, to avoid misunderstanding the power implications both for the data center and the grid.

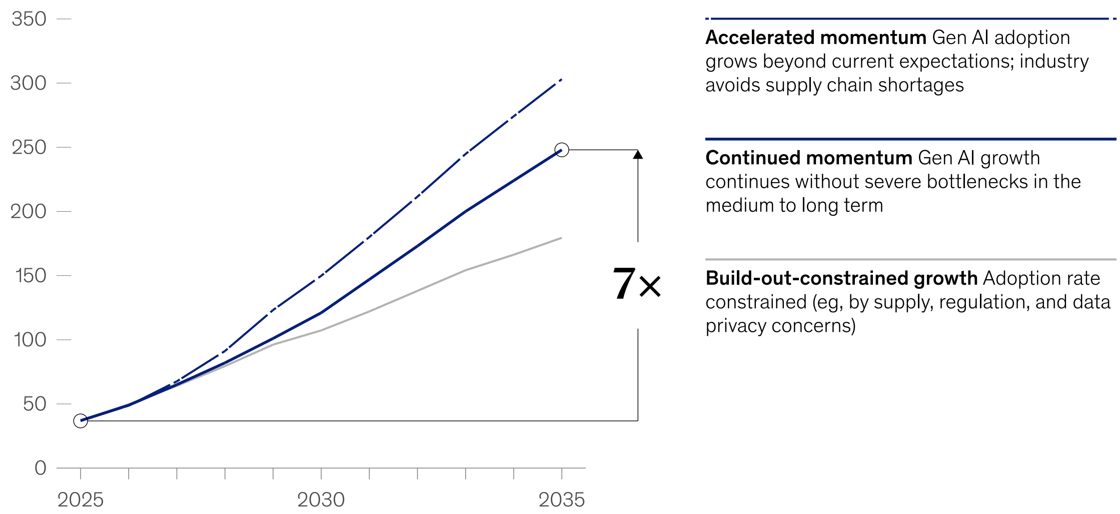
To better understand the range of possible futures, we modeled three scenarios for data center demand through 2035. An accelerated-momentum scenario assumes adoption grows beyond current expectations as chip and algorithm efficiency improve, and that the industry avoids supply chain shortages. The continued-momentum scenario assumes trends continue roughly on the same trajectory as today, with moderate bottlenecks in supply chains and infrastructure. And the constrained-growth scenario assumes slower adoption because of supply limitations, regulatory pressure, and public concerns around both data privacy and community impacts of data centers. Community pushback is emerging as a force restricting data center development; four gigawatts of projects were postponed in the first quarter of 2026 as a result of community action.

The power demand trajectory is relatively certain in all scenarios through 2030, given announced projects, secured interconnection queues, and publicly disclosed capital plans from hyperscalers and AI infrastructure providers. Our model suggests data center power demand could grow approximately 27 percent annually through 2030, reaching 121 GW of data center IT demand (Exhibit 1).

Exhibit 1

## Data center power demand could grow sevenfold by 2035.

**US data center IT load demand scenarios,** gigawatts




Source: Data Center Demand Model

McKinsey & Company

Beyond 2030, uncertainty increases and the difference across scenarios widens, but all point to slower growth in capacity from 2030–35. This slowdown reflects a potential saturation in demand for chips and slower AI adoption, as well as a shift from building new data centers to replacing today's components with more energy-efficient, next-generation chips that add compute per unit of power. For the continued-momentum scenario, we expect a sevenfold increase in data center IT demand by 2035.

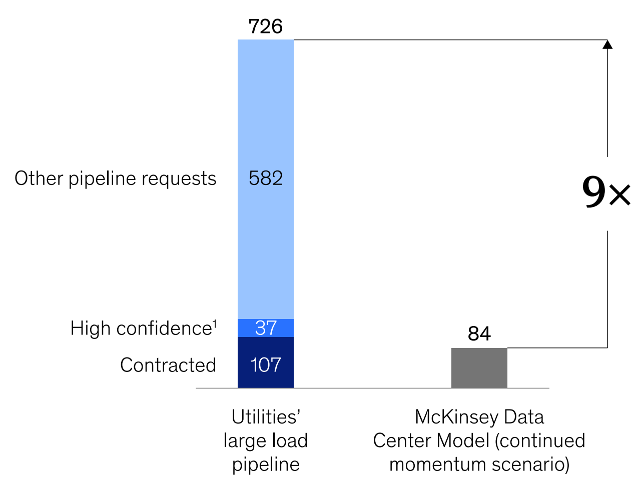
### Swollen interconnection backlogs paint a distorted picture

The massive backlog in load interconnection pipeline requests could be read as evidence of a demand bubble. Yet many of these requests are speculative and duplicative: Developers and hyperscalers are submitting multiple requests to see which secures power first, and early-stage developers are trying to secure interconnection rights before selling a project to a well-capitalized party that will do the building and install the IT hardware. Demand estimates based on such interconnection requests are likely highly distorted. Indeed, our analysis shows that large-load interconnection pipelines are nine times as large as the 2030 demand in our continued-momentum data center scenario (approximately 84 GW of new IT load, or approximately 120 GW of new total power demand), not including off-grid projects (Exhibit 2). Considering only contracted and high-confidence projects (those expected to move forward within the next 12 months) still amounts to a twofold overshoot of our projected demand.

Exhibit 2

## Pipeline inflation is likely significant, with large load capacity requests nine times as large as credible builds.

**Large load pipeline vs expected additional US data center demand, 2030,** gigawatts




[1]Projects expected to begin within 12 months.

McKinsey & Company

Distinguishing between credible, financeable projects and speculative or duplicative proposals is difficult, especially given how quickly power projects are being built today and the fact that most projects are good-faith development efforts, even without a future tenant secured. As developers rush to secure land and power, interconnection requests often outpace chip availability, contributing to concerns about power overbuild.

If some of these projects do not pan out, inflated requests for power will still cause problems across the value chain. They can interfere with electrical equipment ordering and lead times; strain engineering, procurement, and construction (EPCs) issuance capacity; complicate utility planning assumptions; increase stranded cost risk; and affect how investors, hyperscalers, colocators, and neocloud providers allocate capital across their portfolios. As a result, utilities have begun implementing large-load tariffs, larger down payments, and other measures to curb queue gaming, incentivize applications that are most likely to be built, and protect the customer base from bearing costs for projects that do not materialize. After implementing a large-load tariff, American Electric Power Ohio saw an approximately 81 percent reduction in its large-load pipeline, dropping from 30.0 GW to 5.64 GW of requested capacity.[3]

**Power demand extends beyond AI**

In the event that data center compute growth demand slows, generation and transmission built to support AI loads are unlikely to be stranded. These assets can serve other power demands and will likely strengthen system reliability and resilience. Coal plants still provide 15 percent[4] of US electricity generation, and many of these are likely to exit the fleet by 2030 as they pass their 65-year lifetime (Exhibit 3). In total, coal and aging thermal plant retirements could remove 50–75 GW of capacity from the system by 2030.

Even if data center demand slows, components of a power system that were expected to supply AI power could be repurposed or redeployed to serve broader system needs, including demand from large commercial and industrial businesses.

## Power pressures point to a tight US grid in the coming years

Our scenarios point to a range of US power demand outcomes in the coming decade. We expect that supplying the datacenter demand will require a variety of power solutions, some of which come at a higher cost and incur additional emissions to meet the rapid demand growth than would have been necessary at a more measured pace. Notably, those outcomes will vary widely by region. To better understand how these pressures may play out in the US grid, we took a more granular look at the power generation system.

We developed a power generation tightness metric that combines current state-level generation “headroom” (generation capacity above the steady load), planned generation additions (net of retirements), and projected demand growth through 2030 (Exhibit 4).

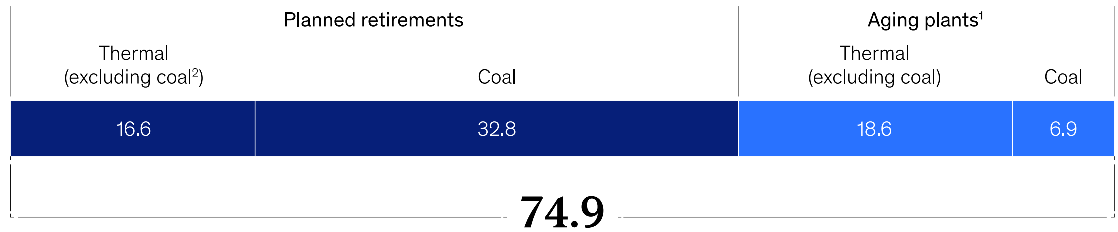
Aggregated across all states, the analysis reveals supply bottlenecks at the national level. The

[3] To ensure that transmission upgrades that exclusively serve a large load site are not paid by the broader customer base, many utilities are instituting tariffs.
[4] “U.S. energy facts explained,” US Energy Information Administration, April 24, 2025.

Exhibit 3

## Up to 75 gigawatts of coal and other thermal generation capacity is expected to retire by 2030 as aging plants surpass 65 years of operation.

**Cumulative generation capacity of plants with retirement dates by 2030 and aging steam gas plants older than 65 years, 2025–30,** gigawatts




Note: Excludes nuclear plants.
[1]Plants more than 65 years old by 2030 without an announced retirement date.
[2]Steam turbines represent 11.3 gigawatts (GW) of the total 15.9 GW of planned retiring thermal capacity and 17.4 GW of the total 22.6 GW of expected retiring thermal capacity.
Source: Preliminary Monthly Electric Generator Inventory, April 2026, US Energy Information Administration

McKinsey & Company

United States has roughly 40 GW of dispatchable headroom to accommodate near-term load growth and roughly 100 GW of capacity committed to be built. But 50–75 GW of coal and steam-gas capacity are expected to retire, and projected demand growth is approximately 120 GW,[5] pointing to a nationwide capacity need for 30–55 GW by 2030.

This shortfall will not be evenly distributed. Load growth will likely concentrate in existing and emerging data center hubs such as Northern Virginia, Phoenix, Louisiana, and parts of Texas—regions favored for their land availability, straightforward permitting, supportive tax policies, and robust digital ecosystems, but where generation and transmission capacity are often already constrained. Mid-Atlantic, Gulf, and certain Midwestern states, where projects are beginning to materialize, are already nearly out of existing headroom.

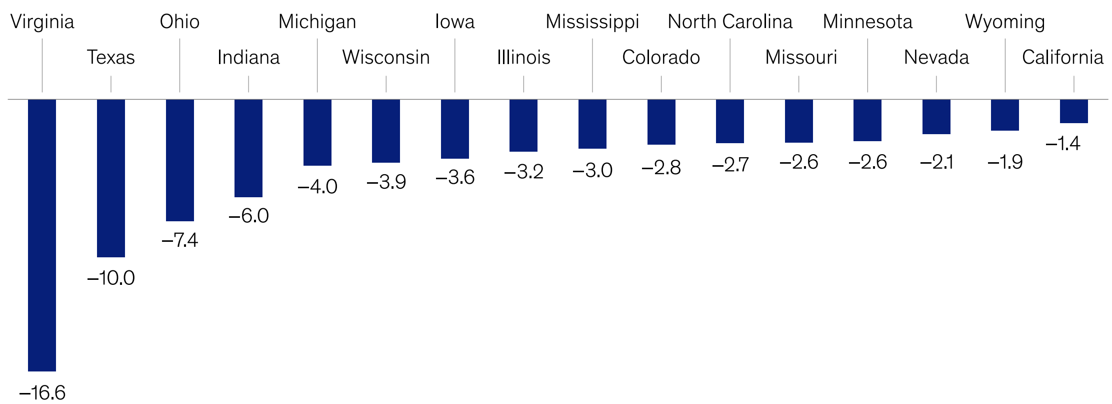
In the United States' largest power market, regional transmission organization PJM, which encompasses much of the mid-Atlantic, the grid has been stable. But PJM's very tight capacity market suggests it may be approaching saturation. In Texas (managed predominantly by the Electric Reliability Council of Texas, or ERCOT), transmission expansion may be more feasible, but equipment and supply chain constraints are delaying new grid infrastructure and firm generation. In other parts of the country, expanding the transmission network with new, high-voltage lines can often take more than a decade, meaning that both power generation and transmission infrastructure can limit data center expansion. At the current pace of growth,

[5] Incremental IT demand by 2030 (that is, 121–137 GW) converted into nameplate generation capacity.

Exhibit 4

## The estimated 2030 power gap varies widely by state.

**Top states by estimated 2030 power gap,[1]** gigawatts




[1]Forecast of incremental demand growth less incremental anticipated supply (new generation, retirements, and current headroom) by 2030; excludes transmission across states.
Source: US Energy Information Administration; McKinsey Global Energy Perspective

McKinsey & Company

misalignment between where stakeholders want to build data centers and where generation, energy demand, and grid infrastructure can quickly scale risks turning into energy supply problems in parts of the United States.

To address these looming gaps and ensure grid reliability, utilities are bringing retired plants back online and extending the lifetimes of coal and gas plants. These measures provide important temporary system relief but are likely insufficient, and, on balance, they delay the energy transition while the biggest buyers of new power, the hyperscalers, maintain their pledge to serve load with clean energy.

Demand-side flexibility—shifting electricity use away from peak times—could head off grid shortages, but data centers have shown little willingness to curtail AI workloads, in part because the value to companies of AI computing far exceeds the cost of power. Other approaches, such as on-site battery storage or contracted virtual power plants[6], may provide more flexibility to reach the same result, but these approaches have reliability tradeoffs. The important caveat here is that if these non-firm sources allow sites to connect sooner than waiting for firm power, they may become widely adopted as companies choose “some power” over “no power”. Ultimately, though, new energy-producing resources will need to be brought online.

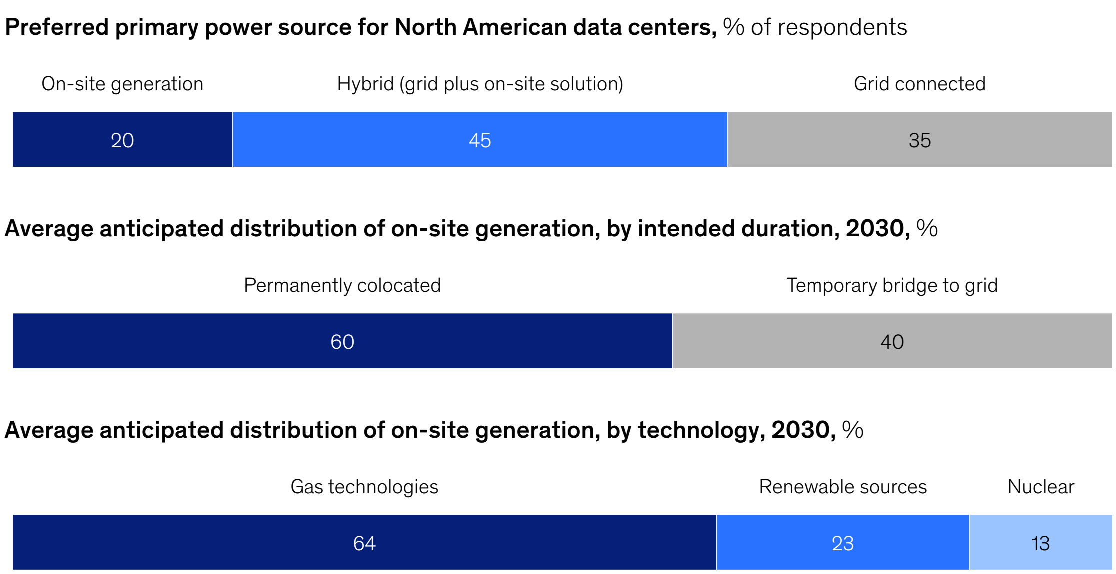
[6]A virtual power plant is a network of power sources, such as home solar panels, smart thermostats, or batteries that are coordinated by software to feed electricity to the grid or remove demand when needed.

**Data center players are turning to on-site power**
In the near term, data center players are turning to on-site power and management solutions such as gas engines and turbines, as well as fuel cells and battery storage. The McKinsey 2025 North America Electric Power and Natural Gas Market Intelligence Survey of power industry leaders revealed that 65 percent of respondents expect to deploy some form of on-site power because of system uptime requirements, resilience concerns, and a need for speed to market (Exhibit 5).[7] Nearly 60 percent of respondents expect to have permanent on-site generation by 2030, even after the grid becomes available—an unprecedented development for the modern US power system. Of those planning to have on-site power, 64 percent expect to rely on natural gas. On-site generation is shifting from a contingency to a core part of the power stack.

While natural gas is often viewed as the fastest path to adding firm power capacity, competition for pipeline infrastructure could cause delays in building gas-powered generation—both in front of and behind the meter—particularly where redundant pipeline access is needed to meet reliability requirements. These constraints are expected to influence where data centers can secure power and scale most rapidly. Siting closer to gas basins reduces the infrastructure constraints and burden of constructing new pipelines, where rights of way can be difficult to obtain within the short timeframe most data centers are targeting.

Exhibit 5

## Data center players report a likely permanent shift to on-site power generation, with gas expected to be the primary source.




Source: McKinsey 2025 North America Electric Power and Natural Gas Market Intelligence Survey (n = 46 data center respondents)

McKinsey & Company

[7] In the fourth quarter of 2025, we conducted a double-blind survey of 145 power sector leaders to understand how industry leaders are navigating power procurement and evaluating power solutions and providers. The survey explored four areas: power procurement processes, preferences across power sources, willingness to pay for select attributes, and perspectives on power value chain providers. Of the respondents, 46 were US-based experts with deep data center knowledge, representing hyperscalers, co-locators, and utilities across the power value chain. More than half were senior directors or above.

### How will power needs be met after 2030?

Beyond 2030, our analyses point to a wide range of outcomes, but several structural dynamics are likely to hold. In addition to the growing role of gas, we expect solar paired with battery storage to sustain strong momentum, given its cost competitiveness, modularity, and speed to deploy, though this is contingent to some degree on stable policy support, interconnection reform, and continued declining costs of storage. Even with a boost from storage, the inherent intermittency of solar means this combination will be part of the broader grid energy system rather than the sole on-site source of large-scale power.

And even as load growth to 2035 and beyond remains somewhat uncertain, the solutions emerging to cover short-term power needs, particularly extensions of older thermal plants and approaches that curtail peak demand, mean that new infrastructure will remain valuable.

In the United States, any material growth in the role of nuclear power over the next decade is likely to come from extensions and upgrades to existing plants rather than from new plants. Emerging technologies such as small modular nuclear reactors and next-generation geothermal energy are attracting increasing capital and offtaker interest, leading to a scaling opportunity. Despite the promise of these solutions to add low-carbon firm power, they are unlikely to materially impact system capacity before the mid-2030s, even as individual projects gain momentum and point to future opportunity. The challenge for leaders is about sequencing multiple solutions to balance speed, reliability, and cost rather than choosing a single pathway. In practice, fast-to-deploy options such as solar and batteries are being paired with firmer capacity such as gas, while hyperscalers continue to explore lower-carbon firm solutions—including nuclear, carbon capture and storage, and geothermal—to stay aligned with long-term decarbonization commitments.

---

The central question for power sector leaders is how to build for a future in which both demand and the grid's architecture continue to evolve. In that future, access to power, not compute, may prove to be the scarcest resource, and the most valuable investments will be those that remain useful across a wide range of demand outcomes. The ideal environment would lead players to maximize the value of both power infrastructure and the benefits of load growth for the broader power system. In that context, grid connections will become even more important, as the benefits of load growth—including the potential to alleviate affordability concerns across the power sector—can only be realized when that load is integrated into the system.

The leaders across the value chain will likely be those who combine early access to scarce power resources with thoughtful risk management. Faced with a market defined by structural imbalance, the players best positioned to succeed are those who navigate uncertainty with discipline: derisking projects under evolving market rules, deploying capital aligned to strategic signposts, managing a portfolio of solutions, and developing commercial solutions that create durable value for all parties.

**Humayun Tai** is a senior partner in McKinsey's New York office, **Jesse Noffsinger** is a partner in the Seattle office, **Sam DeFabrizio** is a partner in the New Jersey office, and **Raissa Dantas** is a consultant in the Bay Area office.

The authors wish to thank Arjita Bhan, Emma Clifford, Jess He, Sandrine Veillette, and Shu Chern Lim for their contributions to this article.

This article was edited by Jessica Marshall, an executive editor in the Seattle office.