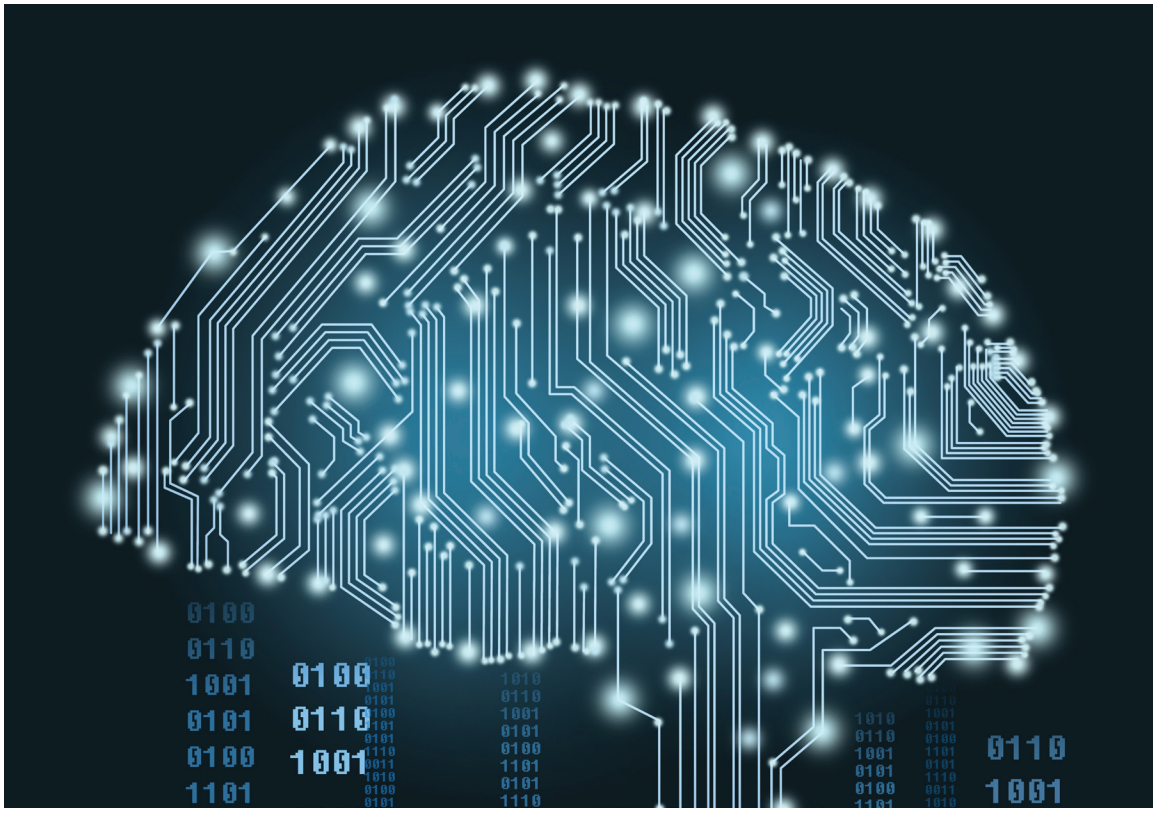
Artificial intelligence: The time to act is now

Advanced Electronics December 2017



Gaurav Batra
Andrea Queirolo
Nick Santhanam

Artificial intelligence: The time to act is now

Artificial intelligence will soon change how we conduct our daily lives. Are companies prepared to capture value from the oncoming wave of innovation?

Pity the radiology department at your local hospital. Yes, they have a fine MRI machine and powerful software to generate the images. But that's where the machines bog down. The radiologist has to find and read the patient's file, examine the images, and make a determination. What if artificial intelligence (AI) could jump-start that process by enabling real-time and more accurate diagnoses or guidance, beyond what human eyes can see?

Thanks to technological advances over the past few years, manufacturers are close to offering such leading-edge MRI solutions. In fact, they're exploring new AI applications that span virtually every major industry, from industrials to the public sector. With better algorithms and increased stores of data, the error rate for computer calculations is now often similar to or better than those of human beings for image recognition and several other cognitive functions. Hardware performance has also improved drastically, allowing machines to process this unprecedented amount of data. That has been a major driver of the improvement in the accuracy of AI models.

Within AI, deep learning (DL) represents the area of greatest untapped potential. (For more information on AI categories, see sidebar, "The evolution of AI"). This technology relies on complex neural networks that process information using various architectures, comprised of layers and nodes, that approximate the functions of neurons in a brain. Each set of nodes in the network performs a different pattern analysis, allowing DL to deliver far more sophisticated insights than earlier AI tools. With this increased sophistication comes greater needs for leading-edge hardware and software.

Well aware of AI's massive potential, leading high-tech companies have taken early steps to win in this market. But the industry is still nascent and a clear recipe for success hasn't emerged. So how can companies capture value and see a return on their huge AI investments?

Our research, as well as interactions with end customers of AI, suggests that six tenets will ring true once the dust settles. First off, value capture will initially be limited in the consumer space, and companies will achieve most value by focusing on enterprise "microverticals"—specific use cases within select industries. Our analysis of the technology stack also suggests that opportunities will vary by layer and that the most successful companies will pursue end-to-end solutions, often through partnerships or acquisitions. For certain hardware players, AI might represent a reversal of fortune, after years of waning interest from investors who gravitated toward software, combined with heavy commoditization that depressed margins. We believe that the advent of AI opens significant opportunities, with solutions in both the cloud and the edge generating strong end-customer demand. But our most important takeaway is that companies need to act quickly. Those that make big bets now and overhaul their traditional strategies will emerge as the winners.

The nuts and bolts of the AI market

Despite the hype about AI, the market can intimidate even the most fearless analysts and investors. No standard definition of the technology stack has emerged in the industry, making it difficult to understand the crowded competitive field. Of the hundreds of companies jockeying for market share, who is offering what?

The evolution of AI

Artificial intelligence (AI) was born in the 1950s, when the English polymath Alan Turing created a test to determine a machine's ability to mimic human cognitive functions, including perception, reasoning, learning, and problem solving. AI grew with the rise of machine learning (ML)—wherein systems absorb and "learn" from data. They then use this knowledge base to make better predictions and decisions over time. In 2010, the advent of deep neural networks ushered in the deep learning (DL) era.

All ML and DL solutions require two steps: training and inference. Take the software in autonomous cars. To help systems detect obstacles in the road, developers present images to the neural net—for instance, those of dogs or pedestrians—and perform recognition tests. Network parameters are then refined until the neural net displays high accuracy in visual detection. After the network has viewed millions of images and is fully trained, it enables recognition of dogs and pedestrians during the inference phase.

Training now accounts for about 95 percent of AI-related workloads in the public cloud because most AI applications are still relatively immature and require huge amounts of data to refine them. As AI models mature, inference will gain more share in the cloud. In fact, DL inference could account for 30 to 40 percent of public-cloud workloads over the next three to five years, with training dropping to 60 to 70 percent. Inference will also gain share with the rise of edge computing (which takes place within devices), as innovation enables low-power, high-performance inference chips.

To bring some clarity to the seemingly chaotic supply landscape, we divided the machine-learning (ML) and DL technology stack into nine layers, across services, training, platform, interface, and hardware (Exhibit 1). Some companies are competing in multiple layers, while others are concentrating on only one or two. As we'll discuss later, companies that limit their focus to specific layers may find themselves at a disadvantage.

Edge and cloud solutions

Traditionally, most AI applications have resided in the cloud—a network of remote servers—for both training and inference. However, inference at the edge will become increasingly common for applications where latency in the order of microseconds is mission critical. With self-driving cars, for instance, the decision of braking or accelerating must occur with near-zero latency, making inference on the edge the optimal option.

Edge computing will also emerge as the favored choice for applications where privacy issues and data bandwidth are paramount, such as AI-enabled CT-scan diagnostics. The growth of edge computing will create new opportunities for all players along the technology stack, particularly for hardware developers.

Our core beliefs about the future of AI

AI is positioned to disrupt our world. McKinsey Global Institute estimates that rapid advances in automation and artificial intelligence will have a significant impact on the way we work and our productivity. To capture value in this growing market, companies are experimenting with different strategies, technologies, and opportunities, all of which require large investments. While much uncertainty still persists, companies that heed the following points will be better positioned to win.

Exhibit 1 **The machine-learning and deep-learning technology stack has nine discrete layers.**

Technology stack			Definition	Example(s)
Services	Solution and use case	9	Solution to problems using trained deep-learning model	Autonomous vehicles (visual recognition)
Training	Data types	8	Data presented to artificial-intelligence (AI) system based on a specific application	Labeled versus unlabeled
	Methods	7	Techniques for optimizing the model weights for the specific application given data	Unsupervised, supervised, reinforcement
Platform	Architecture	6	Structured approach to extract features from data given the specific problem	Convolutional neural network, recurrent neural network
	Algorithm	5	A set of rules that gradually modifies the weights of the neural network to achieve optimal inference, as defined by the training method	Back propagation, evolutionary, contrasted divergence
	Framework	4	Software packages to define architectures and invoke algorithms on the hardware through the interface	Caffe, Torch, Theano
Interface		3	Classes within framework that determine and facilitate communication between software and underlying hardware	Compute unified device architecture, Open Computing Language
Hard-ware	Head node	2	Hardware unit that orchestrates and coordinates computations among accelerators	Central processing units (CPUs)
	Accelerator	1	Silicon chip designed to perform highly parallel operations required by AI	Training: graphic processing units (GPUs), field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs) Inference: CPUs, GPUs, ASICS, FPGAs

Source: Expert interviews; literature search

1. Value capture will initially be limited in the consumer sector

The first consumer AI offerings share a common trait: they enhance products but don't directly contribute to the bottom line. Most of these come from large and well-known tech players, including some online translation and photo-tagging services, or digital voice assistants on mobile phones. Such product enhancements definitely appeal to consumers—they may, for instance, increase the amount of time someone spends on a web site—but they don't produce a direct uptick in sales or revenue. If smaller companies create similar offerings, they often find that sales are limited or nonexistent because consumers gravitate to free solutions. Large players also have access to a significantly larger pool of consumer data—the lifeblood of AI—which allows them to develop more accurate and insightful AI solutions for consumers. With the free products from large players winning most market share, AI value capture will be limited in the consumer sector, in the immediate term.

This may not be the case in the future, however, since newer, fee-based offerings are entering the market, including in-home assistants. The next wave of consumer AI will see even more innovation as automakers and others introduce new products. Take autonomous cars. Some consumers may be content with vehicles in which AI enables

autonomous braking, but others will want more features, such as complete self-driving capabilities, even if they must pay a premium.

2. Enterprise winners will focus on microverticals in promising industries

Our early analysis of data from McKinsey Global Institute, combined with expert interviews and research, revealed nearly 600 discrete uses for AI across major industries. Of these, about 400 require some level of ML and 300 require DL capabilities. Many of the most interesting AI applications are still in the pilot stage and haven't been deployed at scale yet. Here are a few AI applications that could see high demand over the next few years because of their strong visual-perception and processing capabilities:

- Governments can use AI to scan video and identify suspicious activity in public places, or apply AI algorithms to detect potential cyberattacks. Many military applications, including drones, also rely on AI. Beyond security, AI is finding a role in traffic control, including sensors and cameras that allow light signals to change their timing and sequence based on the number of cars on the road.

- As with the public sector, banks are beginning to use AI to detect suspicious behavior, such as patterns suggestive of money laundering. AI algorithms can also help process transactions and make decisions, often with greater accuracy than human employees. For instance, AI algorithms might reveal that certain overlooked characteristics increase the odds that a particular transaction is fraudulent.

- Within retail, AI is already helping with theft detection and it could bring further enhancements to automated checkouts. Several retailers, are piloting systems that use cameras and sensors to detect when shoppers take or return items from the store. After the customers leave the store, their accounts are charged for the total. Other retailers use in-store video to optimize sales associates' coverage. If cameras detect a shopper lingering before a display, the system notifies an associate to provide assistance. In the future, we could see even more enhancements in this area, including AI systems that identify customers with high purchase potential by looking at various characteristics—facial expression (as a signifier of mood), clothing, and number of companions. They could then alert associates about the location of these shoppers within the store.

Companies face a difficult task when deciding which opportunities to pursue, among the hundreds available, but they can narrow their options through a structured approach. The first step involves picking an industry focus. It's true that a company's expertise and capabilities will influence this decision, but players should also consider industry characteristics, including the sector's size. Also important is the potential for disruption within an industry, which we estimated by looking at the number of AI use cases, start-up equity funding, and the total economic impact of AI, defined as the extent to which solutions reduced costs, increased productivity, or otherwise benefited the bottom line in a retrospective analysis of various applications. The greater the economic benefit, the more likely that customers will pay for an AI solution. Exhibit 2 shows the data that we compiled for 17 industries for AI-related metrics.

Just as AI value varies by industry, so does maturity. For instance, the industrial sector could gain big from AI, but member companies are not as ready to embrace these solutions as their counterparts in the automotive industry. For producers of AI products and services, this means that value capture will be staggered, with some industries initially producing higher returns than others.

Exhibit 2 **In each industry, artificial-intelligence demand will depend on market size, pain points, and willingness to pay.**

	Market size	Pain points		Willingness to pay
	Global industry size, $ trillion	Artificial intelligence (AI) use cases, #	Start-up equity raised,[1] $ billion	Average AI economic impact,[2] %
Public & social sector	25+	50+	1.0+	5–10
Retail	10–15	50+	0.5–1.0	5–10
Healthcare	5–10	50+	1.0+	15–20
Banking	15–25	50+	1.0+	<5
Industrials	5–10	50+	0.5–1.0	10–15
Basic materials	5–10	10–30	<0.5	15–20
Consumer packaged goods	15–25	10–30	0.5–1.0	5–10
Automotive & assembly	5–10	10–30	0.5–1.0	10–15
Telecom	<5	30–50	<0.5	20+
Oil & gas	5–10	30–50	<0.5	<5
Chemicals & agriculture	5–10	10–30	<0.5	5–10
Pharmaceuticals & medical products	<5	10–30	<0.5	20+
Transport & logistics	5–10	30–50	<0.5	5–10
Insurance	<5	30–50	<0.5	15–20
Media & entertainment	<5	10–30	<0.5	15–20
Travel	<5	10–30	<0.5	5–10
Technology	<5	10–30	<0.5	10–15

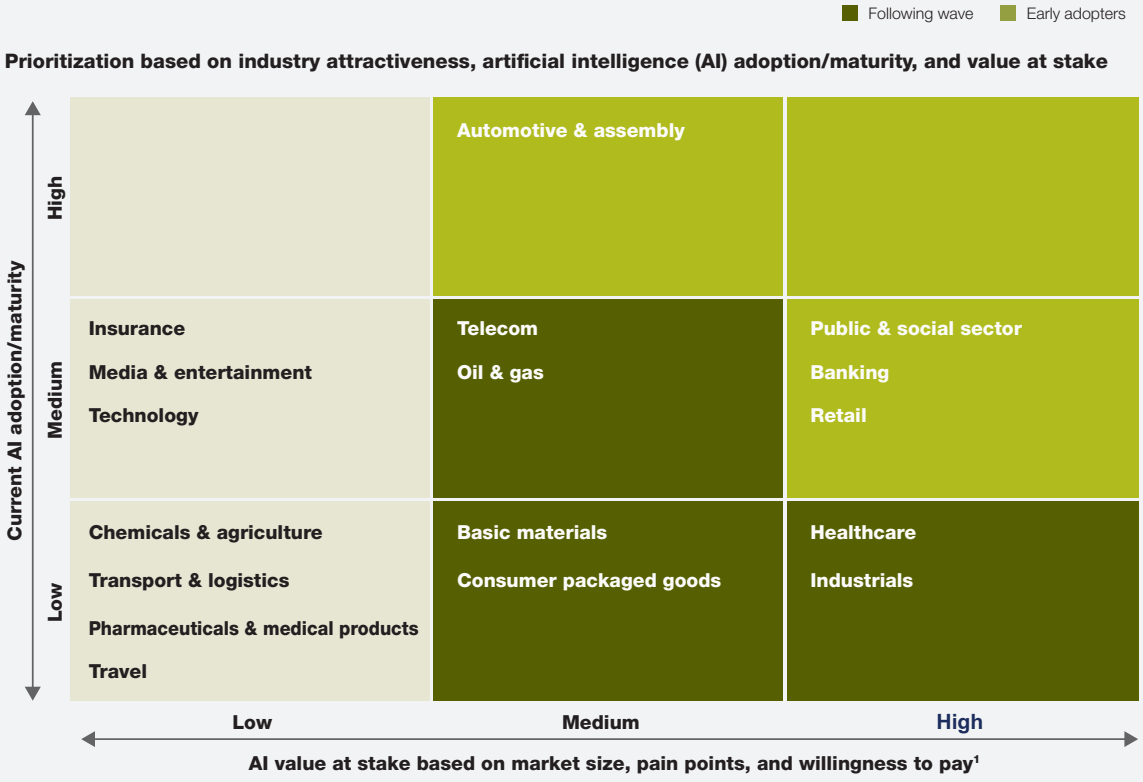
[1] For cross-industry start-ups, equity amount was assumed to be distributed based on global industry size.
[2] Economic impact is the sum of value related to all use cases divided by global industry size.

Source: Crunchbase; expert interviews; IDC; IHS; McKinsey Global Institute analysis

When we considered value at stake in combination with maturity, it became clear that several industries now offer the strongest opportunities for AI: public sector, banking, retail, and automotive (Exhibit 3). While the public sector's prominence may seem surprising in an age where governments are cutting budgets, many officials see the value of AI in improving efficiency and efficacy, and they are willing to provide funding. As they plan their AI strategies, suppliers should focus their investments on potential consumers of AI solutions who are willing to be the first domino.

Microverticals. Once companies have chosen one industry, or a few, as their focus, they must go a step further by selecting particular use cases—which we call microverticals—where they will concentrate. Buyers aren't interested in AI just because it's an exciting new technology—instead, they want AI to generate a solid return on investment (ROI) by solving specific problems, saving them money, or increasing sales. For instance, a manufacturing plant that wants to reduce machine downtime won't simply look for an AI provider that's well known in the industrial space; it will instead seek a company with proven predictive-maintenance

Exhibit 3 **The value at stake from artificial intelligence varies across industries—and so does the readiness for adoption.**



[1] Pain points were identified based on number of use cases and start-up equity. Willingness to pay was based on the total economic value of AI to an industry.

Source: Crunchbase; expert interviews; IDC; IHS; McKinsey Global Institute analysis

expertise and solutions. If an AI provider tried to offer a horizontal solution—one that customers could apply across a variety of unrelated use cases—the value proposition would not be as compelling. End customers would question whether the solution's ROI could justify its greater expense, especially if it applied to several use cases that they considered unimportant or irrelevant.

3. Companies must have end-to-end solutions to win in AI

To win in AI, companies must offer, or orchestrate, end-to-end solutions across all nine layers of the technology stack because many enterprise customers struggle to implement piecemeal solutions. A hospital, for instance, would prefer to purchase a system that included both an MRI machine and AI software that makes a diagnosis, rather than getting these components separately and then trying to make them work together. In addition to increasing sales, suppliers with end-to-end solutions can capture a strategic foothold with customers and accelerate adoption. Nvidia, for instance, offers its Drive PX platform as a module, not just a chip, to provide an end-to-end solution for autonomous driving. The

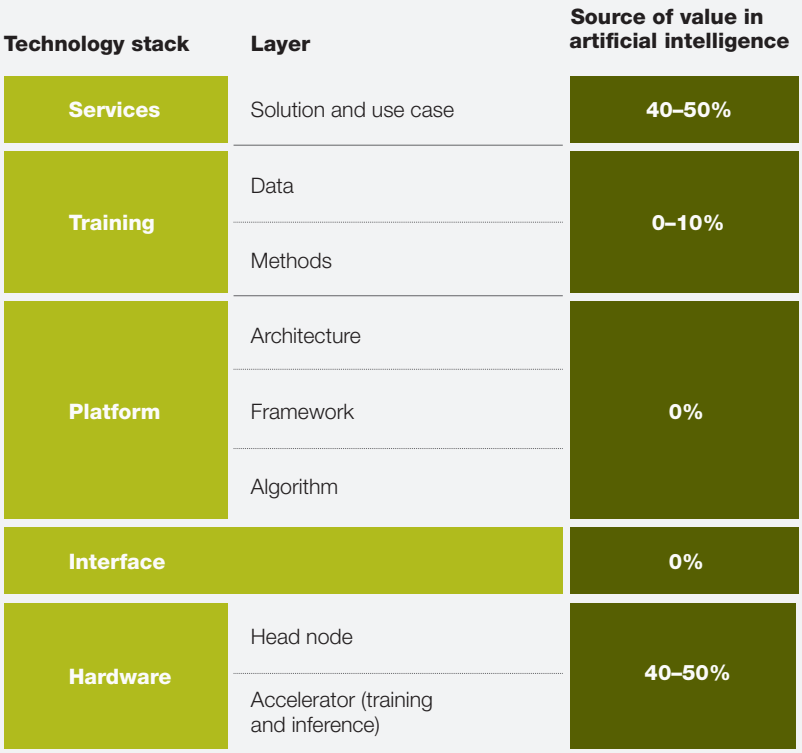
platform combines processors, software, cameras, sensors, and other components to provide real-time images of the environment surrounding a car. It can also identify its location on a map and plan a safe path forward for vehicles.

Large hardware and software players often expand their AI portfolio across the stack by acquiring other companies. While deal making is common across industries, it's more prevalent within AI because of the need for end-to-end solutions. There have been over 250 acquisitions involving private companies with AI expertise since 2012, with 37 of these occurring in the first quarter of 2017.[1] To compete with these giants, many start-ups are undertaking partnerships to position themselves as system integrators for AI solutions.

4. In the AI technology stack, most value will come from solutions or hardware

Within the AI technology stack, our analysis of future trends suggests that each layer will directly generate a different amount of profit, or value. Most value will be concentrated in two areas (Exhibit 4). First—and somewhat surprisingly, given industry trends—many of the best opportunities will come from hardware (head nodes, inference accelerators, and training accelerators). Together, we estimate that these components will account for 40 to 50 percent of total value to AI vendors.

Exhibit 4 **Most opportunities for monetization will come from the hardware and use case/solution layers of the artificial intelligence stack.**



Source: Expert interviews; McKinsey analysis.

While hardware has become commoditized in many other sectors, this trend won't reach AI any time soon because hardware optimized to solve each microvertical's problems will provide higher performance, when total cost of ownership is considered, than commodity hardware, such as general-purpose central processing units (CPUs). For instance, accelerators optimized for convolutional neural networks are best for image recognition and thus would be chosen by medical-device manufacturers. But accelerators optimized for long short-term memory networks are better suited to speech recognition and language translation and thus would appeal to makers of sophisticated virtual home assistants. With every use case having slightly different requirements, each one will need partially customized hardware.

In another pattern that departs from the norm, software (defined as the platform and interface layers) is unlikely to be the sole long-term differentiator in AI. As seen with the advent of DL accelerators, hardware alone or in combination with software will likely enable significant performance improvements, such as decreased latency or power consumption. In this environment, players will need to be selective about hardware choices.

Another 40 to 50 percent of the value from AI solutions will come from services, which includes solutions and use cases. System integrators, who often have direct access to customers, will capture most of these gains by bringing solutions together across all layers of the stack.

For the immediate future, other areas of the AI stack won't generate much profit, even though they may generate indirect value that will drive growth in the DL ecosystem. For instance, data and methods, both elements of training, now deliver only up to 10 percent of a typical AI supplier's value. This pattern occurs because most data comes from end users of AI solutions, rather than third-party providers. A market for data may eventually emerge in the consumer and enterprise world, however, making this layer of the stack relatively more attractive in the future.

5. Specific hardware architectures will be critical differentiators for both cloud and edge computing

With the growth of AI, hardware is fashionable again, after years in which software drew the most corporate and investor interest. Our discussions with end users suggest that interest will be strong for both cloud and edge solutions, depending on the use case. Cloud will continue to be the favored option for many applications, given its scale advantage. Within cloud hardware, customers and suppliers vary in their preference for application-specific integrated circuit (ASIC) technology over graphics processing units (GPUs), and the market is likely to remain fragmented.

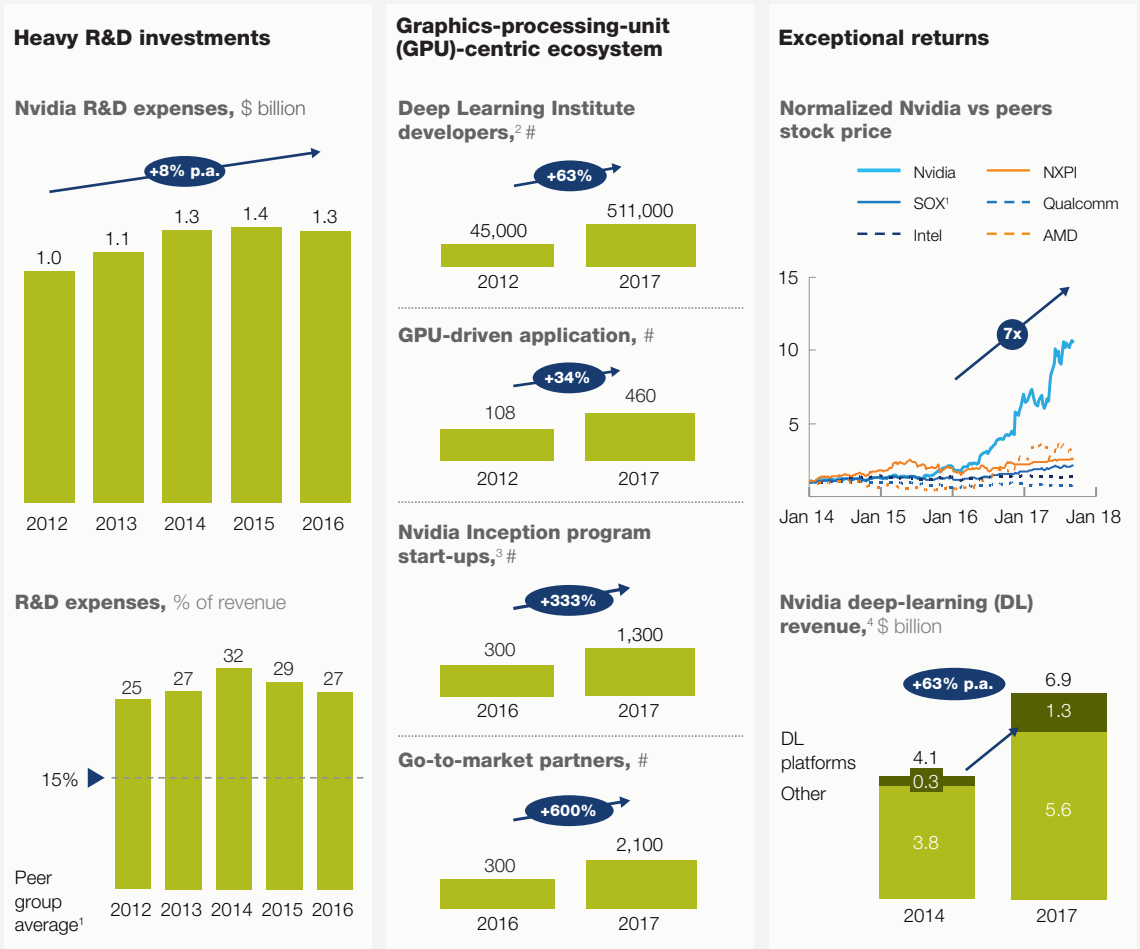
That said, we also see an important and growing role for inference at the edge, where low latency or privacy concerns are critical, or when connectivity is problematic. At the edge, ASICs will win in the consumer space because they provide a more optimized user experience, including lower power consumption and higher processing, for many applications. Enterprise edge will see healthy competition among field programmable gate arrays, GPUs, and ASIC technology. However, ASICs may have an advantage because of their superior performance per watt, which is critical on the edge. We believe that they could dominate specific enterprise applications when demand levels are strong enough to justify their high development costs.

6. The market is taking off already—companies need to act now and reevaluate their existing strategies

Although technology companies may not know exactly how AI demand is evolving, they recognize the enormous opportunity within DL and want to capture it. With the technology still evolving, and with multiple players implementing wildly different strategies, the recipe for success is still uncertain.

The big players are already making their moves, with leading businesses going in directions that defy current wisdom. To consider just one example, Nvidia has increased its R&D expenditures for AI by about 8 percent annually from 2012 to 2016, when they reached $1.3 billion (Exhibit 5). Those costs represent about 27 percent of Nvidia's total revenue—much higher than the peer group average of 15 percent—and they show that Nvidia is willing to take a different path than many semiconductor companies, which

Exhibit 5 **Nvidia is investing heavily in artificial intelligence, gaining market share and exceptional returns.**



[1] Peer group includes various semiconductor companies in the PHLX Semiconductor Sector Index (SOX).
[2] GPU developers trained through Nvidia programs.
[3] Start-ups that are part of Nvidia Inception Program.
[4] Deep-learning revenue for Nvidia is defined as Datacenter (HPC, artificial intelligence) and auto.

Source: Company financials; Nvidia; press search; S&P; McKinsey analysis

are aggressively cutting R&D expenditures. Nvidia has also taken massive steps to create an end-to-end product ecosystem focused on its GPUs. The company is aggressively training developers on the skills needed to make use of GPUs for DL, funding start-ups that proliferate the use of its GPUs for DL, forming partnerships to create end-to-end solutions that incorporate its products, and increasing the number of GPU-driven applications. Other companies that follow such unconventional strategies could also be rewarded with exceptional returns.

Nvidia's success shows that tech companies won't win in AI by maintaining the status quo. They need to revise their strategy now and make the big bets needed to develop solid AI offerings. With so much at stake, companies cannot afford to have a nebulous or tentative plan for capturing value. So what are their main considerations as they forge ahead? Our investigation suggests the following emerging ideas on the classic questions of business strategy:

- *Where to compete.* When deciding where to compete companies have to look at both industries and microverticals. They should select the use cases that suit their capabilities, give them a competitive advantage, and address an industry's most pressing needs, such as fraud detection for credit-card transactions.

- *How to compete.* Companies should be searching now for partners or acquisitions to build ecosystems around their products. Hardware providers should go up the stack, while software players should move down to build turnkey solutions. It's also time to take a new look at monetization models. Customers expect AI providers to assume some of the risk during a purchase, and that could result in some creative pricing options. For instance, a company might charge the usual price for an MRI machine that also has AI capabilities and only require additional payment for any images processed using DL.

- *When to compete.* High-tech companies are rewarded for sophisticated, leading-edge solutions, but a focus on perfection may be detrimental in AI. Early entrants can improve and rapidly gain scale to become the standard. Companies should focus on strong solutions that allow them to establish a presence now, rather than striving for perfection. With an early success under their belt, they can then expand to more speculative opportunities.

If companies wait two to three years to establish an AI strategy and place their bets, we believe they are not likely to regain momentum in this rapidly evolving market. Most businesses know the value at stake and are willing to forge ahead, but they lack a strong strategy. The six core beliefs that we've outlined here can point them in the right direction and get them off to a solid start. The key question is which players will take this direction before the window of opportunity closes.

[1] "The race for AI: Google, Baidu, Intel, Apple in a rush to grab artificial intelligence startups," CB Insights, June 21, 2017, cbinsights.com.

Gaurav Batra is a partner in McKinsey's Washington, DC, office, **Andrea Queirolo** is an associate partner in the New York office, and **Nick Santhanam** is a senior partner in the Silicon Valley office.

The authors wish to thank Eylul Harputlugil, Sam Morton, Harish Soundararajan, and Ben Byungchol Yoon for their contributions to this article.

Contact for distribution: Gaurav Batra
Phone: +1 650 842 5612
Email: Gaurav_Batra@McKinsey.com

December 2017
Designed by Sydney Design Studio