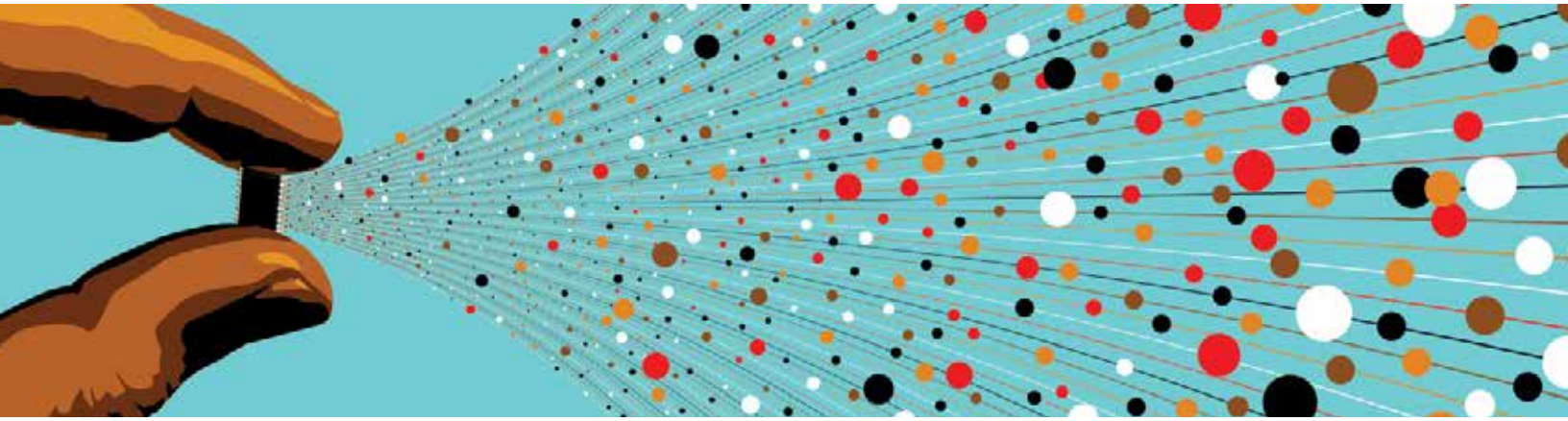


Big data and the opportunities it creates for semiconductor players

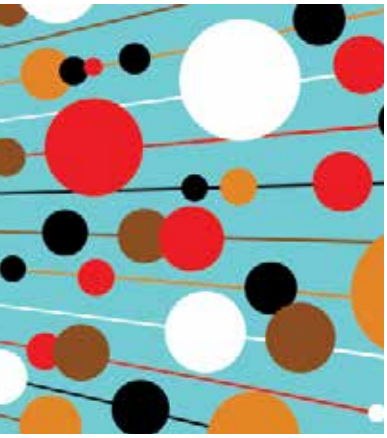
The wave of big data is likely to reshape not only how business gets done but also the pockets of opportunity for semiconductor players. In this article, we explain the nuts and bolts of big data and present a semiconductor-centric view on segments likely to grow most rapidly.

Harald Bauer, Pratap Ranade, and Sid Tandon

The era of big data is upon us. A deluge of business data flows into corporate data centers each day, faster, it seems, than anyone can sort through it. At the same time, consumers going about their day—communicating, browsing, buying, sharing, searching—create their own enormous trails of data. And the volume of data generated by a wide range of sensors, such as those in pipelines, throughout power plants, and on machinery around the factory floor, as well as in smartphones, GPS systems, and connected consumer-electronics devices, presents an entirely new category of "machine data"—generated without explicit human intervention.

The question, then, is what this phenomenon means. Is the proliferation of data simply evidence of an increasingly intrusive world? Or can big data play a useful economic role? While most research into big data thus far has focused on the question of its volume, McKinsey's detailed study of the topic makes the case that the business and economic possibilities of big data and its wider implications are important issues that business leaders and policy makers must tackle.

Before digging into those issues, we define what we consider to be big data and discuss its



growth rates. We then offer an analysis of the challenges and opportunities that big-data and advanced-analytics technologies will present for semiconductor companies, and we conclude with a look at the strategic impact of big data for these companies.

What is big data?

Big data is the confluence of Internet data, business data, and sensor data that together requires a new generation of technical architectures and analytics to process. Such data, if analyzed properly, will help companies large and small unlock new sources of value.

Of course, businesses have tracked performance data for ages, so we must develop a more crisp definition of big data in order to illustrate the difference between it and garden-variety data. Big data has five defining characteristics; any data set that embodies at least the first three characteristics could be considered big data.

First, the scale is significantly larger than traditional data sets. Big data is normally massive. It is usually measured in petabytes, and the databases that house it are designed to scale, ingesting additional classes of information over time. Second, big data is characterized by high dimensionality (thousands or even millions of dimensions for each data element), creating unique challenges for analysis. A third characteristic is the sheer diversity of data. Big data is usually semistructured or even unstructured (for example, tweets from users around the world), and it is often amalgamated across various sources. Frequently, it is a blend of several types of data, which would gum up traditional analytical tools. Big data also flows at a rapid rate, forcing analytical engines to be able to capture, process, and analyze a rushing river of information to enable real-time decision making. The final characteristic of big data is that companies typically use adaptive or machine learning–based analytics that generate better results as the size of the total data set increases.

How big is the wave?

Sizing big data presents a challenge, as the sheer volume of information becomes difficult for humans to interpret. We estimate, for example, that the amount of data stored in enterprise systems, on a global basis, exceeded seven exabytes in 2010. New data stored by consumers that year added another six exabytes to the total. To put these very large numbers in context, the data that companies and individuals are producing and storing is equivalent to filling more than 60,000 US Libraries of Congress. If all words spoken by humans were digitized as text, they would total about five exabytes—less than the new data stored by consumers in a year.

How fast is this data pile growing? Various estimates put the rate of growth at between 23 and 40 percent a year. Recently, Martin Hilbert and Priscila López published a paper in *Science* that analyzed total global storage and computing capacity from 1986 to 2007.[1] Their analysis showed that global storage capacity grew at an annual rate of 23 percent over that period (to more than 290 exabytes in 2007 for all analog and digital media) and that general-purpose computing capacity grew at a much higher annual rate, 58 percent.

IDC estimates that the total amount of data created and replicated in 2009 was 800 exabytes—

[1]Martin Hilbert and Priscila López, "The world's technological capacity to store, communicate, and compute information," *Science*, April 2011, Volume 332, Number 6025, pp. 60–5.

enough to fill two stacks of DVDs that would reach all the way to the moon. The research firm projected that this volume would grow by 44 times to 2020, an implied annual growth rate of 40 percent.

All sources agree that the growth trend in the generation of data has been accelerating at a healthy clip. That will only complicate the challenge of extracting insight from the building mountain of data.

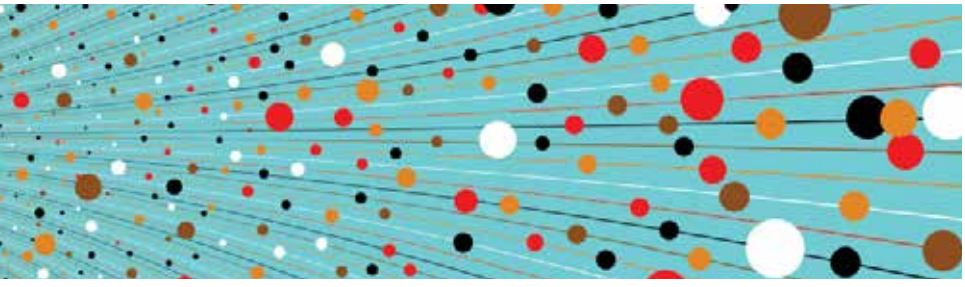
Big data's impact on semiconductor companies

To assess the challenges and opportunities that this trend poses for the semiconductor industry, it makes sense to look at the sorts of demands that data and advanced analytics will put on computer hardware. We will look at four elements involved in creating value from big data: collecting distributed data, extracting meaningful patterns from noisy data, refreshing insights in real time from flowing data, and determining the architecture for chips of the future.

I. Collecting distributed data

To start, you must understand what type of data needs to be collected and how you can go about collecting it. The type of data collected is critical to the impact that the data can have. For example, a retailer used to need to know the name, address, and zip code of each customer to do business. Now, that barely captures the outermost layer of the information available about each customer. Tweets, Facebook posts, user-generated video, GPS modules, and accelerometers in mobile phones are just a few examples of sources for such information. In addition, there is a wide range of structured and unstructured data available that could help modern retailers deliver targeted sales pitches through the most relevant medium at the perfect moment. As data are collected for individuals across other dimensions, more interesting patterns may emerge. Nike+ FuelBands, for example, track a user's motion throughout the day, and iPhone applications such as SleepCycle monitor the quality of their sleep at night. Even more powerful are the sensors for industrial applications. They are being embedded to track information from all aspects of a complex operation such as an oil refinery, a power station, or a mine.

To understand how data are collected from the viewpoint of semiconductor companies, it helps to break the overarching task into three elements: field nodes, network infrastructure, and back-end analytics. Field nodes are the millions of networked sensors that are being embedded in the world around us, whether in mobile phones, smart energy meters, automobiles, or industrial machines. They are engineered to communicate with other electronics, making up the "Internet of Things," also referred to as machine-to-machine communications. Sales of these types of devices—comprising an analog front end, an embedded microcontroller unit, and a radio-frequency chip—are currently increasing at a compound annual growth rate of 35 percent. In all, we expect there will be a global field-node installed base of



Analysis is the key to turning the multitude of data into useful insights.

200 million units by 2015, representing a market worth between $3 billion and $5 billion. These field nodes are, by design, low consumers of power, and as such, they are popular as a tool to drive efficiencies in industries ranging from retail to health care, from manufacturing environments to oil and gas processing.

"Digital oil fields," for example, are emerging as a key technology to optimize production costs for oil fields. (Typically, the production phase accounts for about 42 percent of the cost of producing a barrel of oil.) Digital fields aggregate data from arrays of field nodes, including seismic sensors, flow monitors, and oil-rig and tanker GPS and telemetry. The data are aggregated and managed in real time at operations and decision centers, relying heavily on automated pattern recognition and decision making, significantly lowering the cost of production.

In health care, remote health-monitoring nodes allow physicians to monitor patients' vital signs via low-power wireless signals. This data stream enables preventive medicine and therefore reduces medical costs, since it is frequently cheaper to treat a patient before the condition deteriorates and becomes an emergency. In the public sector, this technology is being deployed to reduce traffic congestion by coordinating data from sensors embedded in the road surface with smart parking meters and even water-supply-management systems. Policing is another application. Over the last 15 years, New York City experienced a 60 percent drop in crime by adopting predictive policing efforts that integrate data from closed-circuit TV cameras, real-time news feeds, and mining of CompStat data to assess the real-time likelihood of crimes by type and by location.

Once data are collected, information can be transmitted over wireless or wireline data networks. Telecommunications carriers such as AT&T and Sprint have launched services aimed at facilitating the transport of data collected from field nodes. These services are being tailored for enterprises in different industries, including health care, transportation, and energy generation and distribution. The focus of the network is to ensure that the data collected by field nodes are transported back to central clusters of computers for analysis in a fast, reliable, and secure manner.

Analysis is the key to turning the multitude of data into useful insights. With the right tools in place, businesses can uncover patterns and connections within the data that would not be obvious to human analysts. Combining data from the Web, field nodes, and other sources in an effort to capture multiple attributes of a target (which could be a customer, a location, or a product) is the first step. As you sift through billions of data points across hundreds or even millions of dimensions for patterns, you encounter what is known as the "curse of dimensionality"—data analysis gets exponentially harder as the dimensionality of the information increases. Big data sets have very high

dimensionality. A general rule of thumb, used by several start-ups dedicated to machine learning (Exhibit 1), is that once you have more than 15 dimensions in a group of data, you start to see significant benefits from applying machine-learning techniques instead of classical methods to analyze the data. This brings us to the second task for semiconductors in the age of big data: extracting meaning from the mountain of data through advanced analytics.

II. Extracting signal from noise

IT professionals tend to think of big data primarily as a database-management challenge. After all, it does involve large numbers of data points. But handling the high dimensionality in the data is as much, if not more, of a challenge. If Facebook wanted to track an individual user, for instance, it would know right off the bat where the user lived, what his or her e-mail address was, who that person's friends are, how often they communicate with each other, and perhaps where they bank or shop. Each of those aspects is a dimension. In all, the count of dimensions for any one customer or user could easily reach 1,000. Now, if Facebook wanted to use this data to segment its user base, the task would quickly become too much for any human mind to easily comprehend.

One of the techniques used by machine learning practitioners to crunch this geyser of data is to find the "natural dimensionality" of the data they wish to analyze (Exhibit 2). A support-vector machine (SVM) is a well-established, powerful

Exhibit 1

Machine learning seeks to automate understanding.

Machine learning is the science (or art) of building algorithms that can recognize patterns in data and improve as they learn

It uses a bottom-up approach:
- Learn model structure from the data
- Separate training and testing

It includes two types of learning:
- Supervised (classification, regression)
- Unsupervised (clustering, structuring, detection)

It has broad applications:
- Speech and gesture recognition (Kinect)
- Natural language processing (Siri)
- Vision (iPhoto facial recognition)
- Recommendation (Netflix, Amazon)
- Time-series prediction (Rebellion)
- Medicine (Equinox Pharma)

Over the past few years,[1] machine learning has exploded, but it is still at the knee of an S-curve

1957	1980–1990s	2004	2008	2009	2010	2011	2012
Perceptrons invented	Novel algorithms SVMs, HMMs[2] created	Google MapReduce published	Netflix prize offered[3]	Start-ups emerge (eg, Cloudera, Kaggle, Tresata)	Watson won Jeopardy; Siri released for iPhone 4	First major industry conference, Strata 2011, held	Companies build in-house machine-learning capability (eg, Facebook, LinkedIn, Twitter)

[1]Dates are approximate.
[2]SVMs: support-vector machines; HMMs: hidden Markov models.
[3]Progress prize awarded in 2008.

Exhibit 2

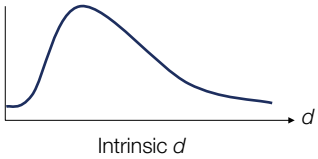
Machine learning seeks to discover intrinsic structures embedded in high-dimensional observations.

EXAMPLE: UNSUPERVISED LEARNING

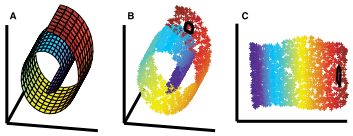
Linear methods are insensitive to higher-order nonlinear structures

The "curse of dimensionality" motivates the need to reduce the complexity of data

Classification performance



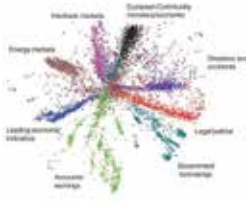
Linear methods cannot unroll the Swiss roll



More sophisticated methods are necessary to uncover patterns in high-*d* data

Linear method applied to ~800,000 news stories

Autoencoder network applied to same data



Source: G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, July 2006, Volume 313, Number 5786, pp. 504–7; Sam T. Roweis and Lawrence K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, December 2000, Volume 290, Number 5500, pp. 2323–6

machine-learning algorithm often used for classification of large data sets. It transforms data into a higher-dimensional feature space where the data are sorted and separated by a "hyperplane." Feature-space transformations are leveraged not only by SVMs but also by many leading machine-learning algorithms. Such types of linear-algebra transformations require matrix operations in high-dimensional spaces, which resemble the transformations that graphics processors use to quickly render beautiful images for video games. This is a class of operation known in Flynn's taxonomy (a classification of computer architectures) as SIMD or MIMD—that is, single instruction or multiple instructions on multiple data. The phrase "multiple data" in this case refers to the multiple elements of one vector data point. SISD, on the other hand, represents single instruction on single data—a more traditional workload.

These machine-learning algorithms, such as SVMs, tend to outperform traditional statistical methods for classifying complex data sets. Furthermore, the task of combing through large data sets is actually quite similar to the types of operations found in modern graphics processing. After all, a computer monitor or high-definition TV screen is essentially a matrix, and an individual vector is similar to the number of pixels in one row on a screen.

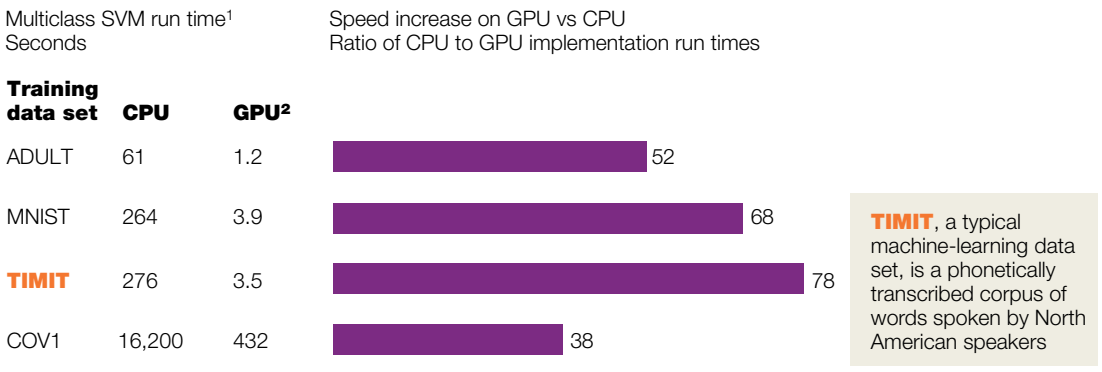
Graphics processing units (GPUs) have been optimized to do the massively parallel linear algebra and matrix math that is behind the sorts of high-powered animations found in home game systems such as Microsoft's Xbox LIVE or Sony's PlayStation 3. A new class of GPUs sold for general-purpose computing, known as GPGPUs, is already catching on. In fact, as workload shifts to the cloud, GPGPU clusters could present an important opportunity area for semiconductor players in the server space. Amazon.com was the first major cloud player to launch a GPGPU instance of its popular Elastic Compute Cloud (EC2) offering; it did so in November 2010.

Optimizing code for the GPGPU, however, remains one of the primary barriers to adoption. The upper bound with regard to how fast code can run in a massively parallelized environment is described by Amdahl's law, which states that the degree of speed increase is inversely proportional to the share of sequential code, measured by run time. That said, the observed speed increases for a range of machine-learning algorithms have varied from 43 to 800 times the normal speed when run on GPGPUs rather than CPUs. Researchers at the Toyota Technological Institute (a joint effort with the University of Chicago), for example, demonstrated speed increases of 40 to 80 times on GPUs in 2011 for the multiclass SVM—a core machine-learning algorithm (Exhibit 3).

However, parallelizing code to enable programs to run on GPGPUs presents significant challenges. First, having multiple threads operating at the same time, with a few shared variables across

Exhibit 3

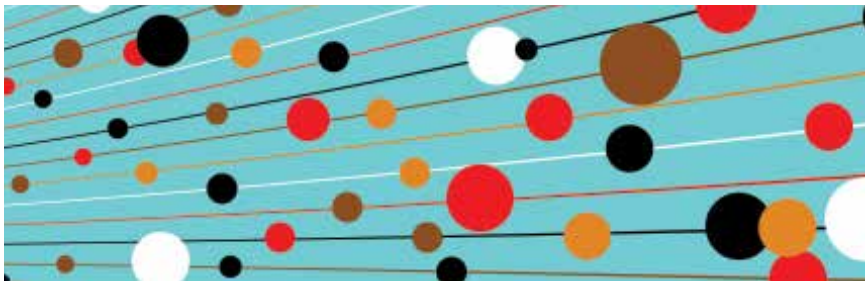
Binary and multiclass kernel-based SVMs, a core set of machine-learning algorithms, can operate 40 to 80 times faster on GPUs.



Testing system specs:
- Intel Core i7 920 CPU; 12G memory
- 2x NVIDIA Tesla C1060 graphics cards (4G memory each): only 1 card used for GPU implementation

[1]SVM: support-vector machine; run times do not include time spent during initialization (or clustering).
[2]Graphics processing unit.

Source: A. Cotter, J. Keshet, and N. Srebro, "Proceedings of the 17th ACM SIGKDD," an international conference on knowledge discovery and data mining, 2011; K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines." *Journal of Machine Learning Research*, March 2002, pp. 2265–92



them, can create an entirely new class of software bugs. The most common example would be "race conditions," which arise from errors in "locking" shared variables while a particular thread is operating on the program. While that is bad enough, race conditions can also create a condition known as "parallel slowdown," where the entire program actually functions slower when parallelized than when run in a linear fashion.

A second challenge comes from the human mind. Academics and IT professionals with experience in parallel programming have highlighted repeatedly that the human brain thinks sequentially, not in parallel, which makes parallel programming conceptually less intuitive and more challenging. Last, parallelization requires programming in a new language. The two most popular are CUDA and OpenCL; these are not considered easy to learn, nor are they easy to use.

Now that we understand a bit more about the nature of the new analytical workloads and their implications for processing horsepower, we shift to the third important task that chips will encounter in their efforts to make big data powerful: generating insights in real time.

III. Making decisions in real time

Putting together the pieces of insight from big-data harvests overnight is good. But putting them together to inform business decisions in real time is even better. Here we encounter the third task that semiconductor companies must accomplish in order to participate in the big-data revolution: real-time analytics.

Depending on the business and its specific context, a given company may refresh its customer data several times a day, or even several times an hour. And if it wants to track customers in real time, the company will need to know when, for example, you begin to shut down your PC at work and head for the subway or commuter-rail service in the evening. To target you with an advertisement that appears on the screen of your mobile phone in time for you to walk by a branch location, that company will need a system optimized to make marketing decisions in real time. The GPGPUs discussed in the previous section help reduce the raw computation times needed to run machine-learning pattern-recognition algorithms, but the biggest bottleneck for real-time analytics is the speed of memory access. The roadblock is the time it takes a CPU or GPGPU to read and write information from cache, random-access memory (RAM), and the hard-disk drives or flash memory where the data are stored.

Data flows from storage, such as hard disks, to RAM and then to cache memory, getting physically close to the processor at each step,

With the rise of big data and the Internet of Things, the trend toward integration of more functions onto a single piece of silicon is likely to continue.

enabling increasingly fast access. More memory, closer to the processor, is essential for speed. More and more companies are therefore looking to shift toward in-memory computing. The elements that make this possible are larger caches, above 512 megabytes, and faster interconnects. These pieces allow even a robust business to store in memory, say, the last minute of customer data across a large retail network. That gives the cache one minute to refresh its data, and because the processor is not involved in that refresh, it can concentrate on the decision-making end of the process, thereby speeding up the decision-making engine.

IV. Moving toward the elegant, all-in-one smart chip of the future

The microchip is evolving at a brisk clip. More and more functions that used to reside on discrete chips are being integrated into a single chip. With the rise of big data and the Internet of Things, the trend toward integration of more functions onto a single piece of silicon is likely to continue.

According to James Jian-Qiang Lu and Ken Rose of Rensselaer Polytechnic Institute and Susan Vitkavage of Sematech, the next generation of devices will likely be 3-D integrated circuits (ICs) with an array of sensors (GPS, accelerometers, microelectromechanical systems components, and biosensors, to name a few) deposited on one layer and large, ultra-high-density cache memory on another layer. These two layers will be integrated both vertically and horizontally, forming a single circuit. Those multilayered circuits could usher in an era of smart, integrated devices, constantly collecting and transmitting data about the world around us.

A bright future is just around the corner

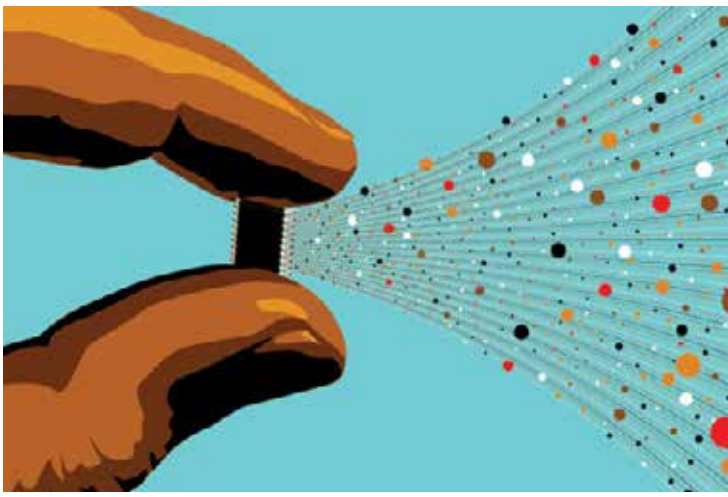
The types of innovations discussed in the last section are at the forefront of laboratory research today, but given a few years' time, these sorts of innovations will be appearing in fabs and foundries around the globe. This presents a range of opportunities for semiconductor companies, whether in the development of sensors or in the field of integration, which takes both engineering prowess and manufacturing skill to flourish.

All semiconductor companies must adapt to the era of big data and the Internet of Things.

Fortunately, there are many, many ways semiconductors—whether they are CPUs, GPGPUs, chip-based sensors, integrated field nodes, 3-D ICs, or optical chip interconnects—will power the core of the machines that drive the next S-curve in productivity and insight generation. Big data presents a huge opportunity for semiconductor companies to power this next phase of growth.

• • •

Given the context of the broad big-data revolution, it will be important for chip players to acknowledge that software analytics will be as important as their precious semiconductors are. To tackle the challenges of deploying advanced analytics for the big-data world, semiconductor companies will benefit from alliances with or even acquisitions of the software and middleware players also working on their pieces of the big-data puzzle. Additional collaborations with systems original equipment manufacturers, specifically those that are working to design solutions for big-data analytics, could also prove helpful for semiconductor companies. These sorts of collaborations will benefit both parties as they work to uncover the right approach to real-time, large-scale data processing. New ideas for hardware and software elements will occur during testing, and that leads to new market opportunities for both partners. While the technical challenges are significant, the opportunity for semiconductor players in the age of big data is substantial.



Harald Bauer (Harald_H_Bauer@McKinsey.com) is a principal in McKinsey's Frankfurt office, **Pratap Ranade** (Pratap_Ranade@McKinsey.com) is a consultant in the New York office, and **Sid Tandon** (Sid_Tandon@McKinsey.com) is a consultant in the Silicon Valley office. The authors would like to acknowledge the valuable contributions of Christian Bienia, Sri Kaza, Kai Shen, and Michael Vidne.