DOI:10.1145/3231589

**Advances in neurotechnologies are reigniting opportunities to bring neural computation insights into broader computing applications.**

BY JAMES B. AIMONE

# Neural Algorithms and Computing Beyond Moore's Law

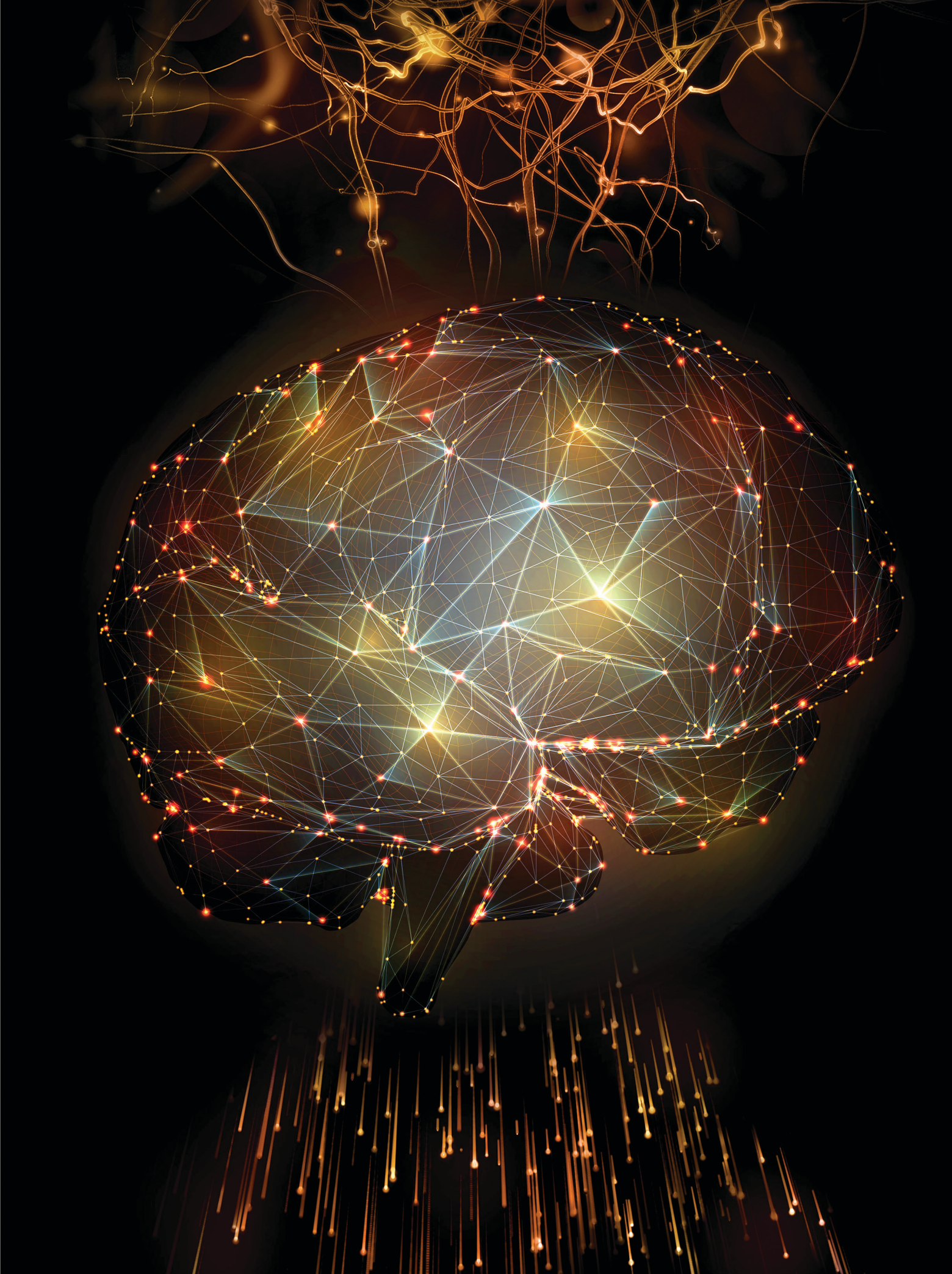
THE IMPENDING DEMISE of Moore's Law has begun to broadly impact the computing research community.[38] Moore's Law has driven the computing industry for many decades, with nearly every aspect of society benefiting from the advance of improved computing processors, sensors, and controllers. Behind these products has been a considerable research industry, with billions of dollars invested in fields ranging from computer science to electrical engineering. Fundamentally, however, the exponential growth in computing described by Moore's Law was driven by advances in materials science.[30,37] From the start, the power of the computer has been limited by the density of transistors. Progressive advances in how to manipulate silicon through advancing lithography methods and new design tools have kept advancing computing in spite of perceived limitations of the dominant fabrication processes of the time.[37]

There is strong evidence that this time is indeed different, and Moore's Law is soon to be over for good.[3,38] Already, Dennard scaling, Moore's Law's lesser known but equally important parallel, appears to have ended.[11] Dennard's scaling refers to the property that the reduction of transistor size came with an equivalent reduction of required power.[8] This has real consequences—even though Moore's Law has continued over the last decade, with feature sizes going from ~65nm to ~10nm; the ability to speed up processors for a constant power cost has stopped. Today's common CPUs are limited to about 4GHz due to heat generation, which is roughly the same as they were 10 years ago. While Moore's Law enables more CPU cores on a chip (and has enabled high power systems such as GPUs to continue advancing), there is increasing appreciation that feature sizes cannot fall much further, with perhaps two or three further generations remaining prior to ending.

Multiple solutions have been presented for technological extension of Moore's Law,[3,33,38,39] but there are two main challenges that must be addressed. For the first time, it is not immediately evident that future materials

## » key insights

- **While Moore's Law is slowing down, neuroscience is experiencing a revolution, with technology enabling scientists to have more insights into the brain's behavior than ever before and thus positioning the neuroscience field to provide a long-term source of inspiration for novel computing solutions.**
- **Extending the reach of brain-inspiration into computing will not only make current AI methods better, but looking beyond the brain's sensory systems can also expand the reach of AI into new applications.**
- **Realizing the full potential of brain-inspired computing requires increased collaborations and sharing of knowledge between the neuroscience, computer science, and neuromorphic hardware communities.**



will be capable of providing a long-term scaling future. While non-silicon approaches such as carbon nanotubes or superconductivity may yield some benefits, these approaches also face theoretical limits that are only slightly better than the limits CMOS is facing.[31] Somewhat more controversial, however, is the observation that requirements for computing are changing.[33,39] In some respects, the current limits facing computing lie beyond what the typical consumer outside of the high-performance computing community will ever require for floating point math. Data-centric computations such as graph analytics, machine learning, and searching large databases are increasingly pushing the bounds of our systems and are more relevant for a computing industry built around mobile devices and the Internet. As a result, it is reasonable to consider the ideal computer is not one that is better at more FLOPS, but rather one that is capable of providing low-power computation more appropriate for a world flush with "big data." While speed remains an important driver, other considerations—such as algorithmic capabilities— are increasingly critical.

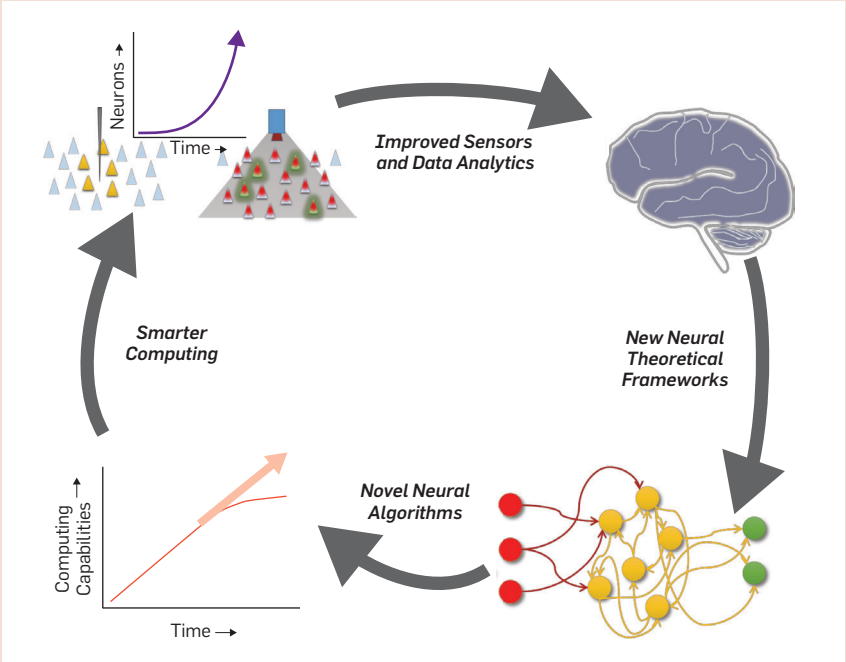
For these reasons, neural computing has begun to gain increased attention as a post-Moore's Law technology. In many respects, neural computing is an unusual candidate to help extend Moore's Law. Neural computing is effectively an algorithmic and architectural change from classic numerical algorithms on von Neumann architectures, as opposed to exploiting a novel material to supplant silicon. Further, unlike quantum computation, which leverages different physics to perform computation, neural computing likely falls within the bounds of classic computing theoretical frameworks. Whereas quantum computation can point to exponential benefits on certain tasks such as Shor's quantum algorithm for factoring numbers;[34] neural computing architecture's most likely path to impact is through polynomial trade-offs between energy, space, and time. Such benefits can be explicitly formalized and are potentially quite impactful for certain applications.[1] However, there is limited evidence that neural architectures can be more powerful on generic applications than the general purpose architectures used today.[33]

The identification of neuromorphic technologies as a potential driver beyond Moore's Law[39] forces the question of whether neural computing is truly a paradigm that will permit exponential scaling going forward, or rather would it represent a "one-off" gain of efficiency in some dimension, such as power efficiency? While potentially impactful, such a value proposition would not represent a long-lasting scaling capability. While to some the distinction between these two futures may appear semantic, there is a considerable difference. If neural architectures indeed represent only a one-time gain to accelerate a handful of algorithms, then it perhaps merits some consideration by specialized communities. However, if neural computation were to actually represent a scalable technology, it would justify a significant research investment from the trillion-dollar computing industry.

This article posits that new computational paradigms that leverage emerging neuroscience knowledge represent a distinctly new foundation for scaling computing technology going forward. Instead of relying on continual advances in miniaturization of devices; neural computing is positioned to benefit from long-lasting intellectual advances due to our parallel gain of knowledge of the brain's function (Figure 1). In effect, because the materials science and chemistry of devices has been extensively optimized, we may achieve greater impact by looking to the brain for neural inspiration and hopefully achieve a continual advancement of our neural computing capabilities through algorithmic and architectural advances. Arguably, the recent successes of deep artificial neural networks (ANNs) on artificial intelligence applications is a compelling first step of this process, but the perspective offered here will contend that more extensive incorporation of insights from brain will only continue to improve our computational capabilities. This influence of neu-

**Figure 1. Moore's Law has helped initiate a potential positive feedback loop between neural data collection and improved computation.**

Moore's Law has enabled the miniaturization of sensors and improved the analytics necessary to improve neural data collection. This increased neural data has the potential to dramatically improve our ability to extract knowledge from the brain and incorporate deeper brain-derived capabilities into new algorithms and architectures; in turn furthering the advances of computing technology.



roscience can then more effectively transition to novel computational architectures and more efficient use of silicon's capabilities.

### Knowledge of the Brain Is Undergoing Its Own Dramatic Scaling

Several critical efforts are underway that make such a revolutionary perspective possible. Major government funded efforts in neuroscience, such as the BRAIN Initiative in the U.S., the Human Brain Project in the European Union, and the China Brain Project, are focused on studying the brain at a systems level that they argue will maximize the computational understanding of neural circuits. Several major non-profit efforts, most notably the Allen Institute for Brain Sciences and the Howard Hughes Medical Institute Janelia Research Campus, similarly have developed large programs to systematically study neural circuits. As a specific example, the BRAIN Initiative has a guiding goal of recording a million neurons simultaneously in awake, behaving animals.[4] Such a goal would have been unfathomable only a few years ago; however, today it increasingly appears within neuroscientists' grasp. Somewhat ironically, the advances in neuroscience sensors which allow neuroscientists to measure the activity of thousands of neurons at a time have been fueled in large part by the miniaturization of devices described by Moore's Law. It has been noted the increase in numbers of neurons recorded within a single experiment has itself undergone an exponential scaling over recent decades.[36]

Similarly, large-scale efforts seeking to reconstruct the "connectome" of the brain are becoming more common.[19] In contrast to ANNs, neural circuits are highly complex and vary considerably across brain regions and across organisms. This connectome effectively represents the graph on which biological neural computation occurs, and many neuroscientists argue that knowing this connectivity is critical for understanding the wide range of neural computations performed by the brain. While the technology to image these large-scale connectomes is increasingly available, there is a growing appreciation that challenges surrounding data analysis and storage are likely to become the limiting factor of neurotechnology as opposed to simply achieving higher resolution sensor technologies.[5,12]

**New computational paradigms that leverage emerging neuroscience knowledge represent a distinctly new foundation for scaling computing technology going forward.**

This rise of large-scale neuroscience efforts focused on high-throughput characterization of the brain rests on many decades of substantial progress in understanding biological neural circuits, but it is notable that neuroscience's influence on computation has been relatively minor. While neural networks and related methods have been experiencing a renaissance in recent years, the advances that led to deep learning did not derive from novel insights about neurobiological processing, but rather from a few key algorithmic advances and the availability of large-volumes of training data high-performance computing platforms such as GPUs.[23]

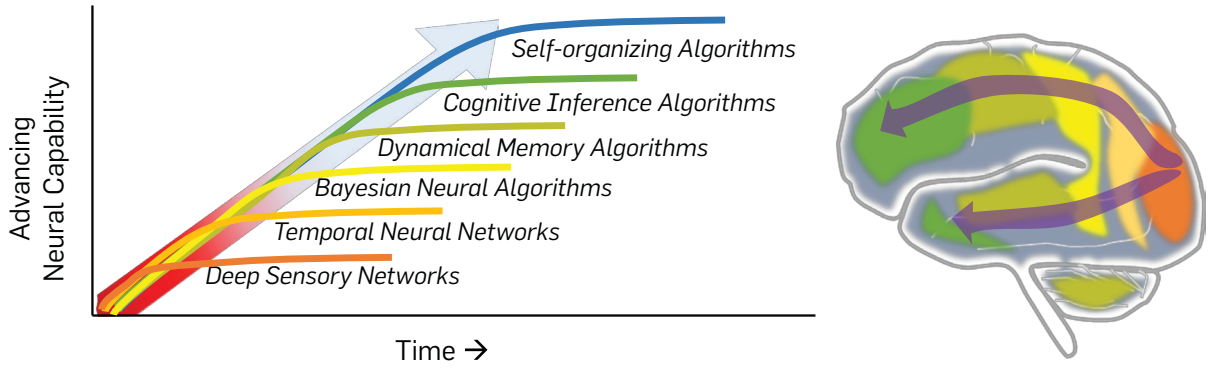
While advances in neuroscience are not responsible for the recent successes in machine learning, there are reasons that it will be more important to look to the brain going forward. For example, the brain may offer novel computational mechanisms that enable the machine learning field to impact domains that still require human intervention, in the same sense that Moore's Law benefited from disruptive shifts in materials science. Two such areas are the requirements of current deep learning techniques for high-quality training data and the capabilities targeted by machine learning applications. Of most immediate concern is the data requirement of machine learning methods. While large volumes of data are increasingly common in many applications, obtaining high-quality data—defined by both well calibrated sensors and effective annotations—is often incredibly expensive and time consuming. As a result, the ability for deep learning-related methods to impact domains with inappropriately structured data has been limited, even in domains where this is relatively straightforward for human operators.

More efficient use of data is an area of intensive machine learning research today, and has seen some recent improvements with regularization techniques such as "dropout"[35] and generative adversarial networks, or GANs,

**Figure 2. The continued scaling of neural computing need not rely on improved materials, but rather can be achieved by looking elsewhere within the brain.**

Today, we are exploiting advances of conventional ANNs at large scale, but there are already trends toward more temporal based neural networks such as long short-term memory. We are poised to benefit from a series of these technological advances, brining neural algorithms closer to the more sophisticated computational potential of the brain.

| Algorithm Class | Current Algorithms | Inspiration | Application |
|---|---|---|---|
| Deep Vision Processing | Deep Convolutional Networks (VGG, AlexNet, GoogleNet), HMax, Neocognitron | Hierarchy of sensory nuclei and early sensory cortices | Static feature extraction (e.g., images) and pattern classification |
| Temporal Neural Networks | Deep Recurrent Networks (e.g., long short-term memory), Hopfield Networks | Local recurrence of most biological neural circuits, especially higher sensory cortices | Dynamic feature extraction (e.g., videos, audio) and classification |
| Bayesian Neural Algorithms | Predictive Coding, Hierarchical Temporal Memory, Recursive Cortical Networks | Substantial reciprocal feedback between "higher" and "lower" sensory cortices | Inference across spatial and temporal scales |
| Dynamical Memory and Control Algorithms | Liquid State Machines, Echo State Networks, Neural Engineering Framework | Continual dynamics of hippocampus, cerebellum, and prefrontal and motor cortices | Online learning content-addressable memory and adaptive motor control |
| Cognitive Inference Algorithms | Reinforcement learning (e.g., Deep Q-learning) Neural Turing Machines | Integration of multiple modalities and memory into prefrontal cortex, which provides top-down influence on sensory processing | Context and experience dependent information processing and decision making |
| Self-organizing Algorithms | Neurogenesis Deep Learning | Initial development and continuous refinement of neural circuits to specific input and outputs | Automated neural algorithm development for unknown input and output transformations |




that help utilize poorly labeled data,[14] but there are many reasons to believe that the brain's approach to maximizing the utility of observed data in both developmental and adult learning is a notable area where brain-inspiration can dramatically improve computing. More broadly, it is useful to consider neuroscience's impact on computing capabilities. In general, machine learning has focused primarily on tasks most associated with sensory processing in the brain. The increased knowledge of neural circuits of non-sensory regions, such as the hippocampus, pre-frontal cortex, and striatum, may well provide opportunities for radically different approaches to algorithms that provide computational intelligence.

## A Potential Timeline of Brain-Inspired Capabilities in Computation

Here, I describe one outlook for neural algorithm "scaling," wherein the community benefits from the development of progressively more advanced brain-like capabilities in algorithms. This work comes from the perspective that the rapid increase in available experimental data is a trend that is unlikely to end soon. While the BRAIN Initiative goal of simultaneously recording one million neurons may appear impressive, that number of neurons is only a tiny fraction of a rodent cortex; and the diversity of neural regions and complex behaviors suggests a plethora of algorithms wait to be defined. A major indicator of this progress will be potential developments in theoretical neuroscience. Neuroscience has long been a field where the ability to collect data has constrained the development of robust neural theories, and it is a growing hope that the ability to measure large populations of neurons simultaneously will inspire the development of more advanced neural theories that that previously would have been dismissed as

non-falsifiable due to our inability to perform the requisite experiments in the brain.[7,40]

Figure 2 illustrates one potential path by which more sophisticated neural algorithms could emerge going forward. Because of advances in neuroscience, it is reasonable to expect the current deep learning revolution could be followed by complementary revolutions in other cognitive domains. Each of these novel capabilities will build on its predecessors—it is unlikely that any of these algorithms will ever go away, but it will continue to be used as an input substrate for more sophisticated cognitive functions. Indeed, in most of the following examples, there is currently a deep learning approach to achieving that capability (noted in the second column). Importantly, this description avoids an explicit judgment between the value of neural-inspired and neural-plausible representations; the neuroscience community has long seen value in representing cognitive function at different levels of neural fidelity. Of course, for computing applications, due to distinct goals from biological brains, it is likely that some level of abstraction will often outperform algorithms relying on mimicry; however, for theoretical development, it will likely be more effective for researchers to represent neural computation in a more biologically plausible fashion.

**1. Feed-forward sensory processing.** The use of neural networks to computationally solve tasks associated with human vision and other sensory systems is not a new technology. While there have been a few key theoretical advances, at their core deep learning networks, such as deep convolutional networks, are not fundamentally different from ideas being developed in the 1980s and 1990s. As mentioned earlier, the success of deep networks in the past decade has been driven in large part due to the availability of sufficiently rich datasets as well as the recognition that modern computing technology, such as GPUs, are effective at training at large scale. In many ways, the advances that have enabled deep learning have simply allowed ANNs to realize the potential that the connectionist cognitive science community has been predicting for several decades.

From a neuroscience perspective, deep learning's success is both promising and limited. The pattern classification function that deep networks excel at is only a very narrow example of cognitive functionality, albeit one that is quite important. The inspiration it takes from the brain is quite restricted as well. Deep networks are arguably inspired by neuroscience that dates to the 1950s and 1960s, with the recognition by neuroscientists like Vernon Mountcastle, David Hubel, and Torsten Wiesel that early sensory cortex is modular, hierarchical, and has representations that start simple in early layers and become progressively more complex. While these are critical findings that have also helped frame cortical research for decades, the neuroscience community has built on these findings in many ways that have yet to be integrated into machine learning. One such example, described here, is the importance of time.

**2. Temporal neural networks.** We appear to be at a transition point in the neural algorithm community. Today, much of the research around neural algorithms is focused on extending methods derived from deep learning to operate with temporal components. These methods, including techniques such as long short-term memory, are quickly beginning to surpass state of the art on more time-dependent tasks such as audio processing.[23] Similar to more conventional deep networks, many of these time-based methods leverage relatively old ideas in the ANN community around using network recurrence and local feedback to represent time.

While this use of time arising from local feedback is already proving powerful, it is a limited implementation of the temporal complexity within the brain. Local circuits are incredibly complex; often numerically dominating inputs and outputs to a region and consisting of many distinct neuron types.[17] The value of this local complexity likely goes far beyond the current recurrent ANN goals of maintaining a local state for some period of time. In addition to the richness of local biological complexity, there is the consideration of what spike based information processing means with regard to contributing information about time. While there is significant discussion around spike-based neural algorithms from the perspective of energy efficiency; less frequently noted is the ability of spiking neurons to incorporate information in the time domain. Neuroscience researchers are very familiar with aspects of neural processing for which "when" a spike occurs can be as important as whether a spike occurs at all, however this form of information representation is uncommon in other domains.

Extracting more computational capabilities from spiking and the local circuit complexity seen in cortex and other regions has the potential to enable temporal neural networks to continue to become more powerful and effective in the coming years. However, it is likely the full potential of temporal neural networks will not be fully realized until they are fully integrated into systems that also include the complexity of regional communication in the brain, such as networks configured to perform both top-down and bottom-up processing simultaneously, such as neural-inspired Bayesian inference networks.

**3. Bayesian neural algorithms.** Even perhaps more than the time, the most common critique from neuroscientists about the neural plausibility of deep learning networks is the general lack of "top-down" projections within these algorithms. Aside from the optic nerve projection from retina to the LGN area of the thalamus, the classic visual processing circuit of the brain includes as much, and often more, top-down connectivity between regions (for example, V2→V1) as it contains bottom-up (V1→V2).

Not surprisingly, the observation that higher-level information can influence how lower-level regions process information has strong ties to well-established motifs of data processing based around Bayesian inference. Loosely speaking, these models allow data to be interpreted not simply by low-level information assembling into higher features unidirectionally, but also by what is expected—either acutely based on context or historically based on past experiences. In effect, these high-level "priors" can bias low-level processing toward more accurate interpretations of what the input means in a broader sense.

While the extent to which the brain is perfectly explained by this Bayesian perspective is continually debated, it is quite clear the brain does use higher-level information, whether from memory, context, or across sensory modalities, to guide perception of any sensory modality. If you expect to see a cloud shaped like a dog, you are more likely to see one. The application of these concepts to machine learning has been more limited, however. There are cases of non-neural computer vision algorithms based on Bayesian inference principles,[22] though it has been challenging to develop such models that can be trained as easily as deep learning networks. Alternatively, other algorithms, such as Recursive Cortical Networks (RCNs),[13] Hierarchical Temporal Memory (HTM),[2] and predictive networks (PredNet)[24] have been developed that also leverage these top-down inputs to drive network function. These approaches are not necessarily explicitly Bayesian in all aspects, but do indicate that advances in this area are occurring.

Ultimately, however, this area will be enabled by increased knowledge about how different brain areas interact with one another. This has long been a challenge to neuroscientists, as most experimental physiology work was relatively local and anatomical tracing of connectivity has historically been sparse. This is changing as more sophisticated physiology and connectomics techniques are developed. For example, the recently proposed technique to "bar-code" neurons uniquely could enable the acquisition of more complete, global graphs of the brain.[20]

Of course, the concept of Bayesian information processing of sensory inputs, like the previous two algorithmic frameworks described previously, is skewed heavily toward conventional machine learning tasks like classification. However, as our knowledge of the brain becomes more extensive, we can begin to take algorithmic inspiration from beyond just sensory systems. Most notable will be dynamics and memory.

**4. Dynamical memory and control algorithms.** Biological neural circuits have both greater temporal and architectural complexity than classic ANNs. Beyond just being based on spikes and having feedback, it is important to consider that biological neurons are not easily modeled as discrete objects like transistors, rather they are fully dynamical systems exhibiting complex behavior over many state variables. While considering biological neural circuits as complex assemblies of many dynamical neurons whose interactions themselves exhibit complex dynamics seems intractable as an inspiration for computing, it is worth noting that there is increasing evidence that it is possible to extract computational primitives from such neural frameworks, particularly when anatomy constraints are considered. Increasingly, algorithms like liquid state machines (LSMs)[25] have been introduced that abstractly emulate cortical dynamics loosely by balancing activity in neural circuits that exhibit chaotic (or near chaotic) activity. Alternatively, by appreciating neural circuits as programmable dynamical systems, approaches like the neural engineering framework (NEF) have shown that complex dynamical algorithms can be programmed to perform complex functions.[10]

While these algorithms have shown that dynamics can have a place in neural computation, the real impact from the brain has yet to be appreciated. Neuroscientists increasingly see regions like the motor cortex, cerebellum, and hippocampus as being fundamentally dynamical in nature: it is less important what any particular neuron's average firing rate is, and more important what the trajectory of the population's activity is.

The hippocampus makes a particularly interesting case to consider here. Early models of the hippocampus were similar to Hopfield networks—memories were represented as auto-associative attractors that could reconstruct memories from a partial input. These ideas were consistent with early place cell studies, wherein hippocampal neurons would fire in specific locations and nowhere else. While a simple idea to describe, it is notable how for roughly forty years this idea has failed to inspire any computational capabilities. However, it is increasingly appreciated that the hippocampus is best considered from a dynamical view: place cell behavior has long been known to be temporally modulated and increasing characterization of "time cells" is indicative that a more dynamical view of hippocampal memory is likely a better description of hippocampal function and potentially more amenable to inspiring new algorithms.

Of course, developing neural-inspired dynamical memory and control algorithms has the potential to greatly advance these existing techniques, but the real long-lasting benefit from neural computing will likely arise when neuroscience provides the capability to achieve higher-level cognition in algorithms.

**5. The unknown future: Cognitive inference algorithms, self-organizing algorithms and beyond.** Not coincidentally, the description of these algorithms has been progressing from the back of the brain toward the front, with an initial emphasis on early sensory cortices and eventually progressing to higher level regions like motor cortex and the hippocampus. While neural machine learning is taking this back-to-front trajectory, most of these areas have all received reasonably strong levels of neuroscience attention historically—the hippocampus arguably is as well studied as any cortical region. The "front" of the brain, in contrast, has continually been a significant challenge to neuroscientists. Areas such as the prefrontal cortex and its affiliated subcortical structures like the striatum have remained a significant challenge from a systems neuroscience level, in large part due to their distance from the sensory periphery. As a result, behavioral studies of cognitive functions such as decision making are typically highly controlled to eliminate any early cortical considerations. Much of what we know from these regions originates from clinical neuroscience studies, particularly with insights from patients with localized lesions and neurological disorders, such as Huntington's and Parkinson's diseases.

As a result, it is difficult to envision what algorithms inspired by prefrontal cortex will look like. One potential direction are recent algorithms based on deep reinforcement learning, such as AlphaGo's deep Q-learning,

which has been successful in winning against humans in games of presumably complex decision making.[29] These deep reinforcement algorithms today are very conventional, only touching loosely on the complexity of neuromodulatory systems such as dopamine that are involved in reinforcement learning, yet they are already disrupting many of the long-held challenges in artificial intelligence. Still, the reliance of modern reinforcement learning techniques on training methods from deep learning limits their utility in domains where training data is sparse; whereas biological neural circuits can rapidly learn new tasks on comparably very few trials.

While learning more about the frontal and subcortical parts of the brain offer the potential to achieve dramatic capabilities associated with human cognition, there is an additional benefit that we may achieve from considering the longer time-scales of neural function, particularly around learning. Most current neural algorithms learn through processes based on synaptic plasticity. Instead, we should consider that the brain learns at pretty much all relevant time-scales – over years even in the case of hippocampal neurogenesis and developmental plasticity. Truly understanding how this plasticity relates to computation is critical for fully realizing the true potential of neural computation.

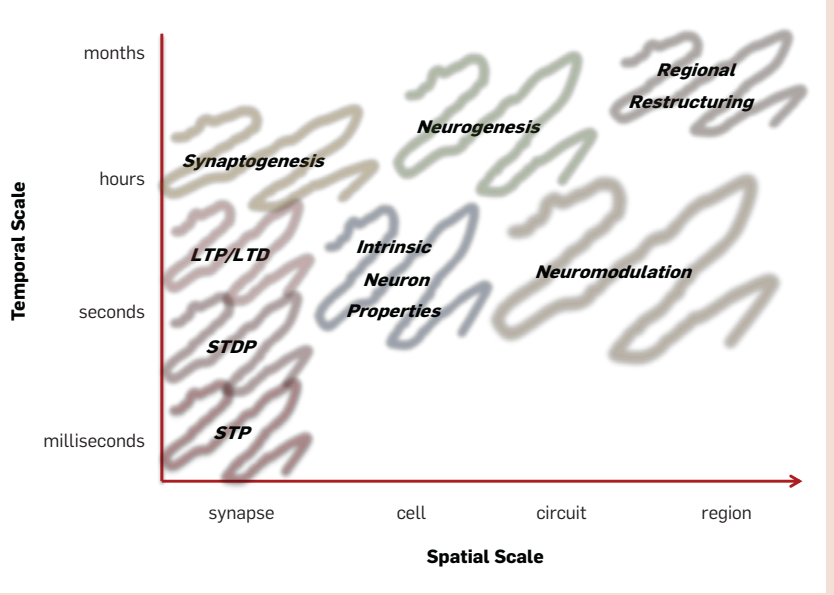
We and others have begun to scrape the surface of what lifelong learning in algorithms could offer with early signs of success. For instance, adding neurogenesis to deep learning enables these algorithms to function even if the underlying data statistics change dramatically over time,[9] although the relative simplicity of deep learning is not suitable to leverage much of the potential seen in the biological process. The recently announced Lifelong Learning in Machines program by DARPA promises to explore this area deeper as well. Ultimately, the implications of these differing learning mechanisms in the brain are highly dependent on the diversity of neural circuit architectures; thus, as algorithms begin to better leverage neural circuits in their design, the opportunities to extend algorithm function through learning will likely appear.

**Progress in neural algorithms will build on itself**. This progression of neural functionality very well could make combined systems considerably more powerful than the individual components. The neural Turing machine concept took inspiration from the brain's higher-level working memory capabilities by combining neural networks with conventional computing memory resources.[15] It is likely this process can be performed entirely within a neural context—if a hippocampal-inspired one-shot learning algorithm could enable the continuous adaptation of a deep learning network, it could considerably increase the long-term utility of that algorithm. This consideration of amortized cost of an algorithm is of course a very different approach to evaluating the costs and benefits of computing, but the prospect of continuously learning neural systems will require some sort of long-term evaluation.[9] Further, just as the brain uses the hippocampus to provide a short-term memory function to complement the long-term memory of several sensory cortices, it is likely that future neural systems could be constructed in a modular manner whereby different combinations of neural components can amplify the performance on different functions.

While this discussion has focused primarily on the long-term benefits of modular neural algorithms, this predicted succession of algorithmic capabilities would be well positioned to be amplified by corresponding advances in computing architectures and materials.[16] Today, deep learning has already begun to substantially influence the design of computer architectures such as GPUs and specialized deep learning architectures such as Google's TPU[18] and implicitly underlying devices and materials. While materials have often been researched with respect to the ubiquitous binary transistor function common to von Neumann architectures, new architectures inspired by novel neural algorithm classes may introduce entirely new desirable characteristics for devices. For example,

**Figure 3. Biological neural circuits exhibit learning at many spatial and temporal scales.**

At synaptic scales, short-term plasticity (STP) changes synapse strengths at very rapid (spike-to-spike) intervals, whereas spike-timing dependent plasticity (STDP) and long-term potentiation / depression (LTP/LTD) affect synaptic strengths over longer time scales, and are more analogous to neural network learning. Learning is not restricted to existing synapses, with neurons also changing their dynamics in response to inputs, and at longer timescales adding new synapses (synaptogenesis) and neurons (neurogenesis) does occur in select brain regions. Finally, learning occurs at macroscopic scales as well; with neuromodulators affecting neuronal dynamics over large brain regions; and potentially even restructuring of brain regions seen at long timescales in response to injury.



dynamical neural algorithms inspired by prefrontal and motor cortex may be best implemented on more dynamics-friendly devices capable of smooth state transitions as opposed to the very stiff and reliable operational characteristics of transistors today. One particular area where neural architectures could begin to have dramatic impact would be intrinsic capabilities for learning and self-organization. While we are still a long-way away from understanding neural development from a computational theory perspective, the availability of such functionality at an architectural level will likely be very disruptive, particularly as algorithms leveraging more brain-like plasticity mechanisms are introduced.

### Can Neuroscience Really Drive Computing Long-Term?

While an argument has been made for why neural computing could provide the computing industry with a future beyond Moore's Law, by no means is this future assured. Aside from the clear technical challenges that lie ahead related to implementing the intellectual trajectory laid out here; there are considerable social challenges that must be addressed as well.

Arguably, the greatest urgency is to inspire the broader neuroscience community to pursue developing theories that can impact neural computing. While there is considerable reason to believe that our knowledge of the brain will continue to accelerate through improved neurotechnologies, the path by which that knowledge can be leveraged into a real impact on computing is not well established. In particular, it is notable that much of the deep learning revolution was driven by computer scientists and cognitive scientists basing algorithms primarily on concepts well established in neuroscience in the 1940s and 1950s. There are several examples to have optimism, however. The IARPA MICrONS program, which is part of the U.S. BRAIN Initiative, aims directly at the challenge of leveraging high-throughput neuroscience data in novel algorithm development.[6] Google's DeepMind—a company started by cognitive neuroscientists—is at the forefront of successfully integrating neural concepts such as reinforcement learning into machine learning algorithms.[29] The EU Human Brain Project has been successful at renewing interest in neuromorphic technologies in the computer science and electrical engineering communities.

**The greatest urgency is to inspire the broader neuroscience community to pursue developing theories that can impact neural computing.**

Nevertheless, there must be a more robust investment by neuroscientists if computing is to benefit from the revolutions underway in experimental neuroscience. This is particularly important if neural influence is to move beyond computer vision—a community that has long had ties to neuroscience vision researchers. For example, despite a historic level of attention and understanding that is roughly comparable to that of visual cortex,[26] the hippocampus has had arguably very little influence on computing technologies, with only limited exploration of hippocampal-inspired spatial processing in simultaneous localization and mapping (SLAM) applications[28] and almost no influence on computer memory research.

A renewed focus by neuroscientists on bringing true brain-inspiration to computation would be consistent with the field's broader goals in addressing the considerable mental health and neurological disorders facing society today.[4] Many of the clinical conditions that drive neuroscience research today can be viewed as impairments in the brain's internal computations, and it is not unreasonable to argue that taking a computing-centric perspective to understanding neurologically critical brain regions such as the striatum and hippocampus could facilitate new perspectives for more clinically focused research.

A second, related challenge is the willingness of the computing communities to incorporate inspiration from a new source. Computing advances have been driven by materials for decades, with reduced emphasis on addressing the underlying von Neumann architecture. Given the perceived plateauing of this classic path, there is now considerable investment in neural architectures; efforts such as IBM TrueNorth[27] and the SpiNNaker[21] and BrainScales[32] systems out of the EU HBP have focused on powerful architectural alternatives in anticipation of neural algorithms. Other more-device

driven efforts are focused on using technologies such as memristors to emulate synapses. To some extent, these approaches are seeking to create general purpose neural systems in anticipation of eventual algorithm use; but these approaches have had mixed receptions due to their lack of clear applications and the current success of GPUs and analytics-specific accelerators like the TPU. It is reasonable to expect that new generations of neural algorithms can drive neuromorphic architectures going forward, but the parallel development of new strategies for neural algorithms with new architecture paradigms is a continual challenge. Similarly, the acceptance of more modern neuroscience concepts by the broader machine learning community will likely only occur when brain-derived approaches demonstrate an advantage that appeared insurmountable using conventional approaches (perhaps once the implications of Moore's Law ending reach that community); however, once such an opportunity is realized, the deep learning community is well-positioned to take advantage of it.

One implication of the general disconnect between these very different fields is that few researchers are sufficiently well versed across all of these critical disciplines to avoid the sometimes-detrimental misinterpretation of knowledge and uncertainty from one field to another. Questions such as "Are spikes necessary?" have quite different meanings to a theoretical neuroscientist and a deep learning developer. Similarly, few neuroscientists consider the energy implications of complex ionic Hodgkin-Huxley dynamics of action potentials, however many neuromorphic computing studies have leveraged them in their pursuit of energy efficient computing. Ultimately, these mismatches demand that new strategies for bringing 21st century neuroscience expertise into computing be explored. New generations of scientists trained in interdisciplinary programs such as machine learning and computational neuroscience may offer a long-term solution; but in the interim, it is critical that researchers on all sides are open to the considerable progress made these complex, well-established domains in which they are not trained.

**Acknowledgments**

The author thanks Kris Carlson, Erik Debenedictis, Felix Wang, and Cookie Santamaria for critical comments and discussions regarding the manuscript. The authors acknowledge financial support from the DOE Advanced Simulation and Computing program and Sandia National Laboratories' Laboratory Directed Research and Development Program. Sandia National Laboratories is a multiprogram laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

This article describes objective technical results and analysis. Any subjective views or opinions that might be expressed do not necessarily represent the views of the U.S. Department of Energy or the U.S. Government. C

**References**

1. Agarwal, S. et al. Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. *Frontiers in Neuroscience 9*.
2. Ahmad, S. and Hawkins, J. Properties of sparse distributed representations and their application to hierarchical temporal memory. arXiv:1503.07469.
3. Association, S.I. and Corporation, S.R. Rebooting the IT Revolution: A Call to Action, 2015.
4. Bargmann, C. et al. BRAIN 2025: A scientific vision. *Brain Research Through Advancing Innovative Neurotechnologies Working Group Report to the Advisory Committee to the Director, NIH*. U.S. National Institutes of Health, 2014; http://www.nih.gov/science/brain/2025/.
5. Bouchard, K.E. et al. High-performance computing in neuroscience for data-driven discovery, integration, and dissemination. *Neuron 92*, 3, 628–631.
6. Cepelewicz, J. The U.S. Government launches a $100-million 'Apollo project of the brain.' *Scientific American*.
7. Churchland, A.K. and Abbott, L. Conceptual and technical advances define a key moment for theoretical neuroscience. *Nature Neuroscience 19*, 3, 348–349.
8. Dennard, R.H., Gaensslen, F.H., Rideout, V.L., Bassous, E. and LeBlanc, A.R. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid-State Circuit 9*, 5, 256¬–268.
9. Draelos, T.J. et al. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *Proceedings of the 2017 International Joint Conference on Neural Networks*. IEEE, 526–533.
10. Eliasmith, C. and Anderson, C.H. *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, Cambridge, MA, 2004.
11. Esmaeilzadeh, H., Blem, E., Amant, R.S., Sankaralingam, K. and Burger, D. Dark silicon and the end of multicore scaling. In *Proceedings of the 2011 38th Annual International Symposium on Computer Architecture*. IEEE, 365–376.
12. Gao, P. and Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology 32*, 148–155.
13. George, D. et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science 358*, 6368, eaag2612.
14. Goodfellow, I. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014, 2672–2680.
15. Graves, A., Wayne, G. and Danihelka, I. Neural Turing machines; arXiv:1410.5401.
16. Indiveri, G., Linares-Barranco, B., Hamilton, T.J., van Schaik, A., Etienne-Cummings, R., Delbruck, T., Liu, S.-C., Dudek, P., Häfliger, P. and Renaud, S. Neuromorphic Silicon Neuron Circuits. Frontiers in Neuroscience, 5. 73.
17. Jiang, X. et al. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science 350*, 6264, aac9462.
18. Jouppi, N.P. et al. Datacenter performance analysis of a tensor-processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*. ACM, 2017, 1–12.
19. Kasthuri, N. et al. Saturated reconstruction of a volume of neocortex. *Cell 162*, 3, 648–661.
20. Kebschull, J.M., da Silva, P.G., Reid, A.P., Peikon, I.D., Albeanu, D.F. and Zador, A.M. High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron 91*, 5, 975–987.
21. Khan, M.M. et al. SpiNNaker: Mapping neural networks onto a massively-parallel chip multiprocessor. In *Proceedings of the IEEE 2008 International Joint Conference on Neural Networks*. IEEE, 2849–2856.
22. Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science 350*, 6266, 1332–1338.
23. LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *Nature 521*, 7553, 436–444.
24. Lotter, W., Kreiman, G. and Cox, D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605.08104.
25. Maass, W., Natschläger, T. and Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation 14*, 11, 2531–2560.
26. Marr, D. Simple memory: A theory for archicortex. *Philosophical Trans. Royal Society of London. Series B, Biological Sciences*. 23-81.
27. Merolla, P.A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science 345*, 6197, 668–673.
28. Milford, M.J., Wyeth, G.F. and Prasser, D., RatSLAM: A hippocampal model for simultaneous localization and mapping. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*. IEEE, 403–408.
29. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature 518*, 7540, 529–533.
30. Moore, G.E., Progress in digital integrated electronics. *Electron Devices Meeting*, (1975), 11–13.
31. Nikonov, D.E. and Young, I.A. Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. In *Proceedings of the IEEE 101*, 12, 2498–2533.
32. Schemmel, J., Briiderle, D., Griibl, A., Hock, M., Meier, K. and Millner, S. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 1947–1950.
33. Shalf, J.M. and Leland, R. Computing beyond Moore's Law. *Computer 48*, 12, 14–23.
34. Shor, P.W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Review 41*, 2, 303–332.
35. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Machine Learning Research 15*, 1, 1929–1958.
36. Stevenson, I.H. and Kording, K.P. How advances in neural recording affect data analysis. *Nature Neuroscience 14*, 2, 139–142.
37. Thompson, S.E. and Parthasarathy, S. Moore's Law: The future of Si microelectronics. *Materials Today 9*, 6, 20–25.
38. Waldrop, M.M. The chips are down for Moore's Law. *Nature News 530*, 7589, 144.
39. Williams, R.S. and DeBenedictis, E.P. *OSTP Nanotechnology-Inspired Grand Challenge: Sensible Machines* (ext. ver. 2.5).
40. Yuste, R. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience 16*, 8, 487–497.

**James B. Aimone** (jbaimon@sandia.gov) is Principal Member of Technical Staff at the Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA.