Cognitive Computing and Big Data Analytics
By Judith Hurwitz, Marcia Kaufman and Adrian Bowles

CHAPTER
14

Future Applications for Cognitive Computing

The development of cognitive computing is at the early stages; however, the building blocks to create this new generation of systems are in place. Over the coming decade there will be many advances in both hardware and software that will impact the future of this important technology. So, the future of cognitive computing will be a combination of evolution and revolution. The evolutionary aspects of cognitive computing are foundational technologies such as security, data visualization, machine learning, natural language processing, data cleaning, management, and governance. There will be revolutions in the capability of systems to improve human-to-machine interactions. In addition, some of the biggest revolutions will come in the areas of hardware innovation.

For decades, advances in chip technology were based on increasing levels of component density and systems integration. Although conventional architectures will continue to improve along these lines, fundamentally different architectures are emerging that will have a bigger impact on cognitive computing performance. Neuromorphic architectures, which are "brain inspired" and use processing elements modeled after neurons, will have a profound impact on speed and portability. In particular, neuromorphic hardware will bring a new level of performance for scale up and will allow data to be processed closer to the source, including direct processing on mobile devices. Quantum computing architectures, based on properties of quantum mechanics, offer great promise for fast processing of large data sets that are

often found in cognitive computing applications. This new generation of chips and systems will enable demand for context-aware computing to be met. This chapter looks forward to the coming decade and what is coming and what will be possible.

Requirements for the Next Generation

The need to share knowledge has always been a top requirement for large and small organizations. Myriad attempts have been made over the decades to try to create learning systems that could codify knowledge in a way that does not require years of coding and software development. Emerging technologies that speed the capability to manage and interpret data to gain insights are emerging. A number of important innovations will change the way organizations can translate data into knowledge that is dynamic, sharable, and predictable.

Leveraging Cognitive Computing to Improve Predictability

Advanced analytics is going to be integrated with cognitive solutions. As cognitive computing matures, companies will find more automated methods of capturing and ingesting massive amounts of data to create solutions. As the corpora of data expand with more experience, it will be possible to incorporate advanced analytics algorithms to a corpus or subset of available data for analysis to determine next best actions or to correlate data to find hidden patterns. This will require a set of tools that can also automate the process of vetting data sources to ensure that data quality is at the level it needs to be. After analysis has been completed, the results can be moved into the cognitive system to update the machine learning models. This will be part of the process of ensuring that a cognitive system can take advantage of the wealth of knowledge and expertise to make better decisions.

The New Life Cycle for Knowledge Management

In a sense, there will be a new life cycle of knowledge management. You begin by creating a hypothesis for the problem you want to solve; you then ingest all the data that is relevant to that problem area; and then vet the data sources, cleanse them, and verify those sources. You train the data, apply natural language processing (NLP) and visualization, and refine the corpus. After the system is put into use, the data is continuously analyzed with predictive analytic algorithms to understand what is changing. Then the process starts all over again. This life cycle from hypothesis through Big Data analytics creates a sophisticated and dynamic learning environment (see Figure 14-1).

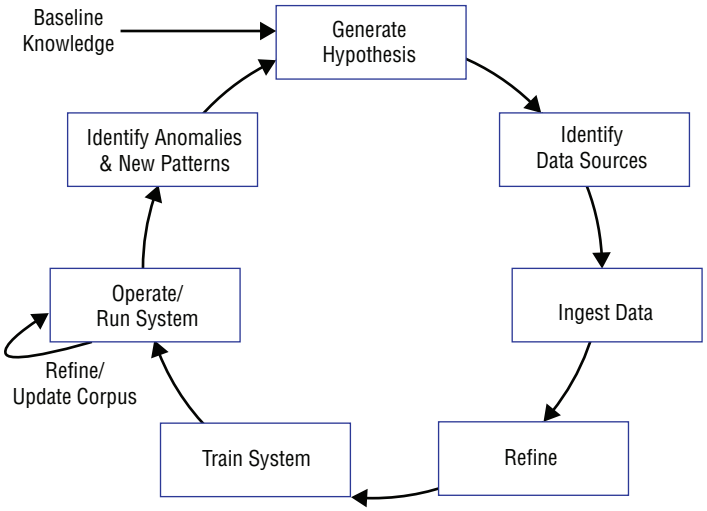

Figure 14-1: The life cycle of knowledge management

Creating Intuitive Human-to-Machine Interfaces

The most sophisticated applications in the first generation of cognitive systems rely heavily on a natural language interface. NLP will continue to be the foundation of how we interact with cognitive systems. However, there will be additional interfaces available for use depending on the nature of the task. For example, there are times when the interface needs to provide visualization so that the researcher can determine where a pattern exists that requires additional exploration. If a biotech researcher is trying to determine the affinity between a disease molecule and a potential therapy, visually detecting patterns will speed the development of a potentially powerful new drug. Other interfaces are also beginning to emerge. For example, improvements in voice recognition technology that can detect emotions such as fear through detection of hesitance could be useful in guiding a user and system through a complex process. When the voice indicates that the instructions are unclear, the system will react with a new explanation. Over time, the system could begin to create new sets of directions that are clearly for a majority of users. A voice recognition system could be helpful in working with the elderly. If the system can detect panic or evidence of a stroke through slurred speech and other cues, it could send help to the elderly individual living at home.

One of the most intriguing experiments with visual interfaces is from an experiment called BabyX developed at the University of Auckland at the Laboratory for Animate Technologies. It is creating "live computational

models of the face and brain by combining Bioengineering, Computational and Theoretical Neuroscience, Artificial Intelligence and Interactive Computer Graphics Research," according to the University's website (`http://www.abi.auckland.ac.nz/en/about/our-research/animate-technologies.html`). The University explains the BabyX project:

> *BabyX is an interactive animated virtual infant prototype. BabyX is a computer generated psychobiological simulation under development in the Laboratory of Animate Technologies and is an experimental vehicle incorporating computational models of basic neural systems involved in interactive behaviour and learning.*
>
> *These models are embodied through advanced 3D computer graphics models of the face and upper body of an infant. The system can analyse video and audio inputs in real time to react to the caregiver's or peer's behaviour using behavioural models.*
>
> *BabyX embodies many of the technologies we work on in the Laboratory and is under continuous development, in its neural models, sensing systems and also the realism of its real time computer graphics.*

The laboratory researchers developed a visual modeling technique that enabled programmers to build, visually model, and animate neural systems. The language being developed is called Brain Language (BL). Armed with this language, researchers can interactively work with the simulations and model new behavior. To gain some insights into the potential for this type of interface, you may want to view some of the videos of BabyX (`http://vimeo.com/97186687`).

Requirements to Increase the Packaging of Best Practices

Most cognitive computing applications are based on custom projects in collaboration with subject matter experts. As with any emerging technology sector, pioneers often blaze their own trail. Over time as there are more and more implementations, it will be possible for these results to be codified into patterns that can be used with other projects looking to solve similar problems. Initially, there will be a set of foundational services that developers can use. However, over time there will be a set of packaged services that has been proven through multiple uses by organizations in similar industries. In a sense there is a corollary with what is now thought of as a packaged application. The difference is that a traditional packaged application is a black box. The user can change data, add rules and business process, but the application itself is sealed from the user.

In a packaged cognitive system, there is a level of transparency. First, it will be critical to understand the assumptions and hypotheses that are built into models as well as the source of the data in the package. In this way, a user could use a subset of the package if the use case is different. There will also be packages that are ubiquitous best practices that will become industry standards. This will

have many varied uses for these packaged cognitive applications from training new professionals in a complex field to creating new cognitive applications within a few months rather than a year or more.

Technical Advancements That Will Change the Future of Cognitive Computing

We have made it clear through this book that we are at the early stages of the evolution and maturation of cognitive computing. Many of the foundational technologies are already in place. However, we still require the evolution of other technologies to get to the predictability and repeatability needed to make systems easy to create and manage. Speed of learning is perhaps the area most in need of innovation. Real-time processing is at the heart of fast learning. On the software side, data has to be analyzed in real time especially to process information in data-rich environments such as video, images, voice, and signals from sensors. These systems will require better clarity, and faster identification of the meaning of these signals. The key to success is improving the time to meaning, not just data acquisition. For example, getting to the point at which the system recognizes and understands the actions of a specific individual within a video fast enough to respond in a threat situation enables more meaningful outcomes. Identifying and then processing the relationships between data in real time can help establish context.

Future innovations in software and hardware will transform what today is complicated and time-consuming for data analytics. Today, to gain this level of expertise requires a lot of manual effort. In the future, machine learning will become more abstracted into the fabric of the development environment. It will be possible to interact with a system in real time as a pattern or connection is detected from the data. This evolution is required as we move from data to information to knowledge. The faster we can process knowledge and understand patterns and context, the sooner we can begin to make discoveries that change the pace of innovation and discoveries across markets and industries.

What the Future Will Look Like

What will a cognitive system look like in the future? The changes in technology that are needed will not happen all at once. Rather, there are two time horizons to consider: the first five years and then the long term, moving out to the next decade. There are three facets that will define the future of cognitive computing: software innovation, hardware transformation, and availability of refined and trusted data sources. All these are predicated on the development of standards. Before discussing the type of technology that will be at play in the future, take a quick look at what might be expected in five years and then in the more distant future.

The Next Five Years

There will be considerable change over the next five years. One of the most significant changes will be in the number of well-defined foundational and industry-specific components that are based on foundational and industry-specific elements. For example, there will be a service that can automatically build ontologies based on deep analysis of text in natural language within a domain. Today, the process requires a lot of manual intervention and consensus building. Although people are still expected to have the last word for the foreseeable future, their participation in the process will diminish as ontology building software learns from experience.

Within the travel market, there will be services that might automate the process of building correlations between destinations, predicted weather patterns, and social media data. As interfaces become standardized, it will be possible to automatically link these services together. These functional services will likely be packaged together into workloads using emerging container standards to enable cost-effective cloud deployment.

There will be a series of well-defined services for everything from ingesting specific types of data to analyzing that data in real time and providing visual interfaces that indicate where patterns exist and what they mean. Natural language interfaces will enable users to select the type of interfaces that are most appropriate for the type of analysis being done. One of the newer approaches gaining traction is to move from simply reporting or displaying data to delivering a "story" that explains the data using a narrative interface. Telling a story about how elements of data are related to each other based on wanted outcomes will bring the best clarity in certain situations. Today, many applications rely on graphs and charts to tell a story about the meaning of data.

In other situations when a customer asks a question on a retail selling site, they will be shown a set of products that best matches a keyword. However, what if that engine has a better understanding of the context of consumer's intent? The consumer looking for a sleeping bag might be ready to purchase sleeping bags for the whole family. The successful site would help build a story just for you, based on your needs, your aspirations, and even your financial constraints. Now you have moved from being shown a single item to being shown a story of your future engagement. The world of camping has many nuances, and there could be other products and services that a smart retailer could offer to the consumer. This new type of system will be as cognitive as you allow it to be. Trust and the ability to grant permission for one engagement or over the life of a relationship between a consumer and vendor will be key to the future of engagement.

Imagine a scenario in which the traveler is equipped with a cognitive trip system. The system knows your destination, your preferences for the way you drive, the gas stations along the way, the health of your car, your preferences in food, and the type of hotels you like to stay at. With the right level of input

and the right level of security, that system could make your reservations, alert you to alternative routes well in advance, and indicate that you should stop to get your car repaired—although a really smart system might advise you not to leave home on a trip if it can't be completed without predictable auto repairs or maintenance. The system could alert a store that has an item that you need and could even negotiate a price and alert you to the pickup time.

You can apply the same approach to how you deal with your insurance company. You may negotiate a deal with that company based on your habits that will be tracked by a device you wear (assuming that you have granted permission). The information you provide to your insurance company will be aggregated with hundreds of thousands of other insured customers to gain an understanding the level of risk. This could either drive costs down as insurance companies better understand actual risks or create a sharing economy pool of people with similar profiles. It could also lead to new government policies as cognitive systems for human capital management intervene to prevent catastrophic loss for the uninsurable, which could lead to unrest because those things can't be hidden in a transparent, connected society with access to communications and cognitive tools.

Looking at the Long Term

As the individual technologies that have been discussed continue to mature, we will see them built into the fabric of cognitive systems or platforms rather than assembled from discrete components. Learning will happen in real time and increasingly be influenced by gestures, facial expressions, and seemingly off-hand comments. These systems will, therefore, automatically understand context from events and data from yesterday or from 5 years ago. These systems will store and continuously analyze all social media history in a deep way. Armed with this level of analysis, the cognitive system will anticipate what you might do next and understand why.

Keep in mind, even in ten years permission-based interactions will be the rule. However, there will be more automated techniques that assume your permission level and then ask for confirmation. The system, in fact, is built by analyzing patterns across millions or perhaps billions of interactions. The level of permission the consumer allows will determine the interaction and level of security in this environment. The optimal system will act in the background, making suggestions or recommended actions when necessary but remaining silent most of the time. In essence, you will be dealing with an advanced automated agent that will allow you to create a persona for yourself that is comfortable for you. The agent software gets to know your preferences and personality over time based on the data that you provide directly and the learning system behind it that makes assumptions based on accumulated information. The system will be designed with a set of rules of etiquette based on how humans are comfortable interacting with machines.

In essence, this is the personal digital assistant for the new cognitive era. Rather than the physical device, it may take the form of a representation or personal agent in the cloud with the multiple types of context-appropriate interfaces available at the time of engagement. It could be you and your personal interactions; it could also be the interface to your washing machine. We have already begun to enter an era where the Internet is ubiquitous, but in the coming era, you will always be connected—unless you choose to disconnect. Depending on the situation, the interface is a natural language, a gesture, or a physical action. The cognitive system captures the nuances of your interactions and changes its interaction based on your changing needs and conditions. The system is constantly learning from your behavior and activities behind the scenes. It modifies its actions based on the learning over time. This technique will be widely applied to everything from traffic patterns in a city to security infrastructure.

As more devices with embedded sensors become ubiquitous, the level of data and actions will explode. Professional athletes will be equipped with sensors that know if they suffer a concussion even before a physician examines them. That same type of sensor-based device could warn a construction worker of an obstacle that he should avoid.

These types of cognitive systems will have potential to break down barriers for humans. A sensor-based device with a sophisticated interface can provide a different level of interaction with people who have trouble interacting in social situations. The system is nonjudgmental. Individuals on the autism spectrum could be helped by a system that learns the best ways to interact and has the potential to open lines of communications that have been blocked. The cognitive system adapts to the communication style most effective for different individuals with different disorders. It could be helpful for elderly suffering from Alzheimer's disease.

The most significant change in the coming decade is that cognitive computing will become part of the fabric of computing. Therefore, it will have a profound impact on many industries and many of the tasks that humans do. Machine learning and advanced analytics will be built into every application. Natural language interfaces will continue to be the foundation of how we interact with systems. Eventually, natural language processing will become a utility service rather than a separate market.

Emerging Innovations

What will it take to get from individual handcrafted systems to the state in which these technologies are deeply embedded in everything you use?

A number of existing technologies that are instrumental in cognitive computing are going to evolve over the next 5 years. That will improve the capability

of systems to be created faster with greater capability to solve complex issues. This section discusses the key technologies.

Deep QA and Hypothesis Generation

Today, Deep QA—which may require a system to generate a series of probing questions for a human to answer for the system to navigate multiple levels of meaning—is rare in practice. In IBM's Watson it is used interactively in a conversational mode with experts to refine their quest for possible answers in complex domains. For example, a doctor may describe a set of symptoms relevant to a patient, and Watson may ask questions that help it to narrow the range of possible answers or increase confidence in one or more diagnoses. It may ask if a particular test has been ordered or ask for more details about a family history. Deep QA requires the system to keep track of all the information that has been provided in previous answers for a session, and only ask further questions when the human answer can help it improve its own performance. It will evaluate the possible answers it may give and assign a confidence level in each, but look at what additional evidence could change that confidence to decide whether to ask for additional information.

If the learning experiences of a lot of systems that answer related questions are shared, that body of knowledge about the process could become a reusable pattern across a domain. In healthcare, for example, there may be enough deep QA analysis to discover the optimal treatment for a specific type of skin cancer because enough data exists— when aggregated—and enough analysis has been done on that data that has been vetted by the best experts in the world. Over time, some hypotheses will have been proven and accepted, so the same query asked at a later date may require less analysis and fewer generated hypotheses as the corpus matures. We may never run out of problems to solve, but for the most part we will see the process of problem solving in complex domains begin to coalesce around cognitive computing. Much like the scientific method guides discovery in the natural sciences, discovery through deep QA and hypothesis generation and testing is likely to become the default approach for many professional disciplines.

NLP

Advances in NLP have been dramatic in recent years as evidenced by the capability of IBM's Watson to derive meaning from unstructured text under conditions of intentional difficulty. (The QA format of Jeopardy! presents “answers” that may be ambiguous or require context or familiarity with idiomatic speech, and contestants must determine the meaning of the answer before identifying the most appropriate question as a response.) This format is challenging for many humans, but Watson had little difficulty finding the relevant meaning,

or alternatively, recognizing when it had low confidence in its answer. The Watson team prepared for the event by studying the way Jeopardy! writers used speech in the past. Those lessons will be valuable as IBM and others extend NLP technology to handle more general cases of slang, colloquialisms, regional dialogues, industry-specific jargon, and the like. A lot of the training is involved with understanding the context of language. NLP systems or services must understand state and conditions that may have been set previously.

Automating translation between natural languages that capture deep meaning remains a difficult problem for NLP. Vocabularies may be mapped from one language to another with reasonable precision (English to French, for example), but natural language communication involves strings or sentences built in to paragraphs and stories that may have explicit and implicit references to meaning expressed in other strings, paragraphs, or even historical references. A key NLP innovation—assuming some common constructs among languages that map to the same underlying deep structures—would be the identification and emulation of the manual process used by expert human translators to discover rules or heuristics they may be applying unconsciously. Analyzing different well-respected translations of books, for example, to identify commonalities and different interpretations, will provide insights into these rules. Today, even some shallow language analysis is so processor-intensive that mobile systems have to send the sentence or string from the device to a cloud-based service before responding. Enabling deep translation on the fly for more than simple statements on mobile devices will require these breakthroughs, or alternatively more powerful NLP chips on the devices themselves.

Cognitive Training Tools

It is tedious and time-consuming to build a corpus today by training a system based on ingested knowledge. A lot of trial and error and human judgment are involved for every new corpus. Much of the training work that is human-intensive today will become automated as we use current generation cognitive computing systems to examine the process to help build better tools. Similar to the way every generation of high-precision manufacturing tools were built with the previous generation of less sophisticated tools, cognitive computing technology will be used iteratively to discover ways to improve the process of building cognitive computing solutions.

Bias in training is one of the most important issues that will have to be addressed. With a lot of unstructured data and no standards to understand that data, experts make judgments based on their own experiences, which are biased because most have never seen the entire universe of possible interpretations. (Even in narrow medical specialties, for example, the most experienced practitioner has rarely seen every possible set of symptoms or treatment outcomes.) However, they aren't even aware of the bias they are bringing to the situation.

In the future, as cognitive tools become more powerful and apply more cognitive learning, it will be easier to determine the source of a bias and point that out to the expert.

Data Integration and Representation

Today, connectors, adapters, encapsulation, and interfaces are used to deal with complex data integration. Although this is sufficient if you have a good understanding of the data sources and they are well vetted, it is a different matter when you begin to bring thousands of data sources together. Data integration needs to be automated with a cognitive process so that the system begins to look for patterns across data sources and detect anomalies to see if they represent new, important relationships that were unknown before or problems with a data source being inconsistent.

You saw that ontologies can codify common understanding of complex relationships within a domain, but implementing an ontology is actually a crutch. In a perfect world, a cognitive computing system would not need an ontology because it could dynamically build its own model of the universe by understanding the relationships and context—but that works only if there is enough data and experience and it can process and understand fast enough. Today, we create ontologies so that performance is acceptable with current system constraints. If you could do that processing on the fly, you wouldn't have to predetermine what the ontology would be; you could discover an ontology rather than building one. With sufficient processing power, an ontology would actually be a system state during execution. It would be generated only on demand if it were required for auditing purposes, perhaps to understand why a decision or recommendation was made.

Emerging Hardware Architectures

Hardware innovations in both the short and long term will have a dramatic impact on the evolution of cognitive computing. Today, it is primarily traditional hardware systems that are used to build cognitive systems. Although parallel structures are used, these systems are still general purpose von Neumann architecture computers, in which all the actual processing takes place in registers within central processing units (CPUs) (or in adjunct processors such as graphical processing units [GPUs]). The real breakthroughs that are on the horizon over the next several years include major changes in chip architectures and programming models.

Complementary to the efforts in software and data architectures, we are seeing two different approaches to hardware architectures evolve. One is based on modeling neurosynaptic behavior (the relationship between neurons and synapses in the brain) directly in hardware. These neuromorphic chips feature many small processing elements that are most tightly interconnected to near

neighbors to communicate much like human brain neurons pass signals via chemical or electrical synapses.

The second promising approach is quantum computing, which is based on quantum mechanics (quantum physics), a branch of physics that explores physical properties at nano-scale. Unlike conventional computers whose fundamental unit of storage and processing is the bit (binary digit) which must be a 1 or 0 at any given time, quantum computers use the qubit (quantum bit), which may be in more than one state at any given time. The next two sections explore the prospects for these competing architectural approaches.

Neurosynaptic Architectures

Why should you look at this new generation of hardware architectures? Simply put, the complexity of identifying and managing relationships between data elements at the scale required for cognitive computing—Big Data—requires enormous computing resources with conventional architectures. Fundamentally, the challenge today is to partition the data effectively to funnel it into an architecture that processes 64 bits of data at a time.

The current basic Intel microarchitecture, for example, used in the Core i7 processor (found in many laptops) and the Xeon family of processors (used in Tianhe-2, currently the world's fastest supercomputer) processes data in increments of 64 bits. Over the past decades, computer scientists have developed elaborate workarounds to compensate for the limitations of hardware. For example, it is relatively easy to add processors to a cluster or system. The individual processors in Tianhe-2 are no faster than those in a modern laptop, but it links together 260,000 of them to harness 3,120,000 cores operating in parallel. The difficult part is to effectively distribute the workload across those similarly architected processors. Some cognitive computing techniques such as hypothesis generation are inherently parallel. Based on the data, it may be desirable to generate hundreds of hypotheses and then process them independently on different processors, cores, or threads.

Another task that would be valuable in a cognitive computing application is real-time image processing in a manner similar to human vision. That also requires mapping millions of bytes of information to look for patterns, which humans do in parallel rather than by breaking up the problem into sequential tasks. For still images, this can take thousands of processors. (The Google experiment mentioned in Chapter 2, "Cognitive Computing Defined," used 16,000 processors just to identify cats.) For video, the problem is much more difficult. A high-definition camcorder typically generates approximately 5 gigabytes of data per minute recording at 30 frames per second. If you want to analyze all the images, you need to analyze each frame and compare it to prior and subsequent frames to find patterns. For example, when evaluating video of a crime scene, detectives look for people whose behavior is not like the rest of the crowd. A

human can do that relatively easily with a single video stream, but when multiple streams are involved, it becomes a daunting task that could be automated with sufficient processing power. For most applications today, it is impractical to do large-scale hypothesis generation and evaluation or real-time video analysis.

Now contrast this first to the neurosynaptic hardware approach. The current large-scale leader in this field is IBM's TrueNorth (developed with funding from DARPA), a neurosynaptic chip with 1 million neuron-inspired processing units and 256 million synapses (connections between the processing units, similar to a computer bus but a lot more powerful and faster). Instead of improving performance by adding additional 64-bit register-limited machines, scaling up with a neuromorphic chip builds in the parallelism because while each neural processing unit executes a single function, it communicates with many others. Like neurons in the brain, they are so physically close and connected that they communicate virtually instantaneously. Test systems have already been constructed with multiple TrueNorth chips yielding a system with 16 M neurons and 4 B synapses.

The underlying principle that is modeled in neurosynaptic chips is Hebb's Rule, commonly simplified as "cells that fire together, wire together," —meaning that neurons in close proximity that fire together (actually in rapid sequence) reinforce learning. This was postulated in Donald O. Hebb's 1949 book, *The Organization of Behavior*, which formed the basis for much of the current understanding of associative learning and the development of parallelized pattern matching algorithms. Mapping the behavior of these human brain elements to fundamental constructs in the hardware architecture provides a natural bridge between the way we look at a problem and the way we solve it, which gives neuromorphic computing great appeal (as "brain-inspired" hardware). In the near future you can expect to see billions of processing units per neurosynaptic chip with trillions of synapses. When these chips are assembled into systems, the result will be a new standard for scalable parallelism that has practical applications for pattern matching and learning in cognitive computing systems.

Commercialization of this architecture will require a new programming model, a sophisticated software development environment and an ecosystem of professionals and companies to create a new industry around this model. Efforts to develop these tools and skills are already underway, but in the immediate future you may see hybrid solutions in which neuromorphic approaches will be combined with conventional computers. Similar to the way the average computer today often incorporates special processors for graphics and sound, neuromorphic chips integrated with a conventional system will enable you to take advantage of conventional programming models for much of the required preprocessing.

Why is this architectural approach so important? The emerging architecture enables you to populate each of the millions of neurons in parallel, rather than artificially constraining you to a 64-bit bandwidth for actual processing. When the data is loaded in these neurons, the chip or system can search for patterns in real time. Applications that now are impractical for conventional systems—for

example, massively parallel hypothesis processing in medicine and scientific exploration or human-like vision processing become feasible. Parallelism without partitioning is a huge advantage for neuromorphic architectures. The acts of partitioning and reassembling results take time and add complexity. Although there are multiple research efforts to build large-scale neurosynaptic chips, the same approach to mimicking neurosynaptic processing is already being commercialized in smaller scale special purpose chip sets for mobile devices. Qualcomm has a production chip set called Zeroth that is intended to capture patterns of human behavior based on the usage of the mobile device to provide context-aware services. This is planned to be put into production by 2015.

The architectures operate in parallel efficiently so that the total power consumption for a unit of work is lower than that of a register-based architecture. This makes these architectures appealing for mobile devices and at scale will reduce the power and space requirements for data centers. Scalability (up and down) and a simple architectural model will make the adoption of neuromorphic chips inevitable for some cognitive computing applications.

Quantum Architectures

The fundamental concept behind a quantum computer is to go beyond a binary, two-state (on/off; that is, 1s and 0s) atomic processing unit to a multistate unit called the qubit. A qubit can have multiple states as defined by the physics of quantum mechanics, including being in multiple states simultaneously (superposition). Conceptually, this will be extremely difficult to popularize because it is beyond the mathematical and scientific knowledge and experience of most of the world's population, but it is the most natural way to process quantum algorithms for learning and discovery. Quantum computers can be simulated using conventional computers by mapping each of the possible states to binary states, but, of course, the performance overhead is significant. For example, in a single conventional 64-bit register, you could represent 2^{64} values (ranging from a string of all 64 0s to 64 1s, or $1.8 * 10^{19}$). In a qubit with three possible values (0,1, or both) 64 qubits could represent 3^{64} values or $3.4 * 10^{30}$, that is, 200 billion times bigger than the binary solutions and impossible to process on a conventional system in anything approaching real time. And in theory, quantum computers can scale without the artificial register restriction, which makes them attractive for massively parallel computations and processing existing quantum algorithms. Like neuromorphic computing, quantum computing will require entirely different programming models, skills, and tools.

Perhaps the most significant barrier to quantum computing is that it requires physical materials to actually be in these superposition states, which requires the processing units to operate at a temperature near absolute zero. That precludes any mobile applications and modestly sized system installations, at least for the time being. Still, the performance potential is too great to ignore. Today, we

are seeing significant research and investment in quantum computing by IBM, Google, and DWave (which focuses exclusively on quantum computing). Google has set up a new effort to build its own quantum computer for AI research with academic researchers in the University of California system while continuing to support the independent efforts of DWave.

The energy, space, cooling, and mathematical skills requirements will keep quantum computing from becoming mainstream in the next decade. Although neuromorphic architectures are expected to grow in popularity quickly and be more pervasive at all levels than quantum computing, quantum architectures will continue to attract research funding because it is well understood that a few breakthroughs could lead to fundamentally faster supercomputers.

Alternative Models for Natural Cognitive Models

Although neuromorphic and quantum computing architectures are based on approaches to established science—neuroscience and quantum mechanics, respectively—that have active research communities in place, they are being challenged by a new approach pioneered by Jeff Hawkins. Hawkins, who changed the way we think about mobile devices when he introduced the Palm Pilot, has an alternative view of human learning. He founded the Redwood Center for Theoretical Neuroscience in 2002 to support research into a layered model of learning based on the functioning of the neocortex. His company Numenta is building applications and an infrastructure for cognitive computing based on his theory of the way the brain stores, processes, and retrieves information about events. His approach is based on the role of the neocortex in human memory as the central organizing principle for computer architecture rather than neurons and synapses. Although it is too early to evaluate the potential for this approach, its Grok for Analytics machine learning anomaly detection product has already demonstrated that it may be useful even if the theory behind it isn't ultimately adopted by the scientific community at large.

Summary

In the future, cognitive systems will be defined as an integrated environment, which means that software and hardware will work as though they are a single integrated system. This new architecture will scale up and down depending on the use case. For applications such as smarter cities and smarter healthcare, the high-end architectures will enable machine learning in near real time. With personal devices and sensor-based assistants, hardware embedded at the end points will provide processing at the source. This convergence between hardware, software, and connectivity will provide the platform for a huge flood of new use cases and applications for cognitive technologies.