1

A NEW ERA OF COMPUTING

IBM's Watson computer created a sensation when it bested two past grand champions on the TV quiz show *Jeopardy!* Tens of millions of people suddenly understood how "smart" a computer could be. This was no mere parlor trick; the scientists who designed Watson built upon decades of research in the fields of artificial intelligence and natural-language processing and produced a series of breakthroughs. Their ingenuity made it possible for a system to excel at a game that requires both encyclopedic knowledge and lightning-quick recall. In preparation for the match, the machine ingested millions of pages of information. On the TV show, first broadcast in February 2011, the system was able to search that vast storehouse in response to questions, size up its confidence level, and, when sufficiently confident, beat the humans to the buzzer. After more than five years of intense research and development, a core team of about twenty scientists had made a very public breakthrough. They demonstrated that a computing system—using traditional strengths

and overcoming assumed limitations—could beat expert humans in a complex question-and-answer competition using natural language.

Now IBM scientists and software engineers are busy improving the Watson technology so it can take on much bigger and more useful tasks. The *Jeopardy!* challenge was relatively limited in scope. It was bound by the rules of the game and the fact that all the information Watson required could be expressed in words on a page. In the future, Watson will take on more open-ended problems. It will ultimately be able to interpret images, numbers, voices, and sensory information. It will participate in dialogue with human beings aimed at navigating vast quantities of information to solve extremely complicated yet common problems. The goal is to transform the way humans get things done, from health care and education to financial services and government.

One of the next challenges for Watson is to help doctors diagnose diseases and assess the best treatments for individual patients. IBM is working with physicians at Cleveland Clinic and Memorial Sloan-Kettering Cancer Center in New York to train Watson for this new role. The idea is not to prove that Watson could do the work of a doctor but to make Watson a useful aid to a physician. The *Jeopardy!* challenge pitted man *against* machine; with Watson and medicine, man *and* machine are taking on a challenge together—and going beyond what either could do on its own. It's impossible for even the most accomplished doctors to keep up with the explosion of new knowledge in their fields. Watson can keep up to date,

though, and provide doctors with the information they need. Diseases can be freakishly complicated, and they express themselves differently in each individual. Within the human genome, there are billions of combinations of variables that can figure in the course of a disease. So it's no wonder that an estimated 15 to 20 percent of medical diagnoses are inaccurate or incomplete.[1] Doctors know a lot about diseases and the practice of medicine. What they need help with is using evidence-based medicine to better evaluate and treat individuals.

Dr. Larry Norton, a world-renowned oncologist at Memorial Sloan-Kettering Cancer Center who is helping to train Watson, believes the computer will be able to synthesize encyclopedic medical and patient information to help physicians more quickly and easily identify treatment options for complex health conditions. "This is more than a machine," Larry says. "Computer science is going to evolve rapidly and medicine will evolve with it. This is coevolution. We'll help each other."[2]

THE COMING ERA OF COGNITIVE COMPUTING

Watson's potential to help with health care is just one of the possibilities opening up for next-generation technologies. Scientists at IBM and elsewhere are pushing the boundaries of science and technology fields ranging from nanotechnology to artificial intelligence with the goal of creating machines that do much more than calculate and organize and find patterns in data—they sense,

learn, reason and interact naturally with people in powerful new ways. Watson's exploits on TV were one of the first steps into a new phase in the evolution of information technology—the era of cognitive computing.

During this era, humans and machines will become more interconnected. Thomas Malone, director of the MIT Center for Collective Intelligence, says a big question for researchers as the era of cognitive computing unfolds is: How can people and computers be connected so that collectively they act more intelligently than any person, group, or computer has ever done before?[3] This avenue of thought stretches back to the computing pioneer J. C. R. Licklider, who led the U.S. government project that evolved into the Internet. In 1960 he authored a paper, "Man-Computer Symbiosis," where he predicted that "in not too many years, human brains and computing machines will be coupled together very tightly and the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."[4] That time is fast approaching.

The new era of computing is not just an opportunity for society; it's also a necessity. Only with the help of smart machines will we be able to deal adequately with the exploding complexity of today's world and successfully address interlocking problems like disease and poverty and stress on natural systems. Computers today are brilliant idiots. They have tremendous capacities for storing information and performing numerical calculations—far superior to those of any human. Yet when it comes to

another class of skills, the capacities for understanding, learning, adapting, and interacting, computers are woefully inferior to humans; there are many situations where computers can't do a lot to help us.

Up until now, that hasn't mattered much. Over the past sixty-plus years, computers have transformed the world by automating defined tasks and processes that can be codified in software programs in series of procedural "if A, then B" statements—expressing logic or mathematical equations. Faced with more complex tasks or changes in tasks, software programmers add to or modify the steps in the operations they want the machine to perform. This model of computing—in which every step and scenario is determined in advance by a person—can't keep up with the world's evolving social and business dynamics or deliver on its potential. The emergence of social networking, sensor networks, and huge storehouses of business, scientific, and government records creates an abundance of information tech-industry insiders call "big data." Think of it as a parallel universe to the world of people, places, things, and their interrelationships. This digital universe is growing at about 60 percent each year.[5]

The volume of data creates the potential for people to understand the environment around us with a depth and clarity that was simply not possible before. Governments and businesses struggle to come to grips with complex situations, such as the inner workings of a city or the behavior of global financial markets. In the cognitive era, using the new tools of decision science, we will be able to apply new kinds of computing power to huge amounts of data

and achieve deeper insight into how things really work. Armed with those insights, we can develop strategies and design systems for achieving the best outcomes—taking into account the effects of the variable and the unknowable. Think of big data as a natural resource waiting to be mined. And in order to tap this vast resource, we need computers that "think" and interact more like we do.

The human brain evolved over millions of years to become a remarkable instrument of cognition. We are capable of sorting through multitudes of sensory impressions in the blink of an eye. For instance, faced with the chaotic scene of a busy intersection, we're able to instantly identify people, vehicles, buildings, streets, and sidewalks and see how they relate to one another. We can recognize and greet a friend we haven't seen for ten years even while sensing and prioritizing the need to avoid stepping in front of a moving bus. Today's computers can't do that.

With the exception of robots, tomorrow's computers won't need to navigate in the world the way humans do. But to help us think better they will need the underlying humanlike characteristics—learning, adapting, interacting, and some form of understanding—that make human navigation possible. New cognitive systems will extract insights from data sources that are almost totally opaque today, such as population-wide health-care records, or from new sources of information, such as sensors monitoring pollution in delicate marine environments. Such systems will still sometimes be programmed by people using "if A, then B" logic, but programmers won't have to anticipate every procedure and every rule. Instead,

computers will be equipped with interpretive capabilities that will let them learn from the data and adapt over time as they gain new knowledge or as the demands on them change.

The goal isn't to replicate human brains, though. This isn't about replacing human thinking with machine thinking. Rather, in the era of cognitive systems, humans and machines will collaborate to produce better results, each bringing their own superior skills to the partnership. The machines will be more rational and analytic—and, of course, possess encyclopedic memories and tremendous computational abilities. People will provide expertise, judgment, intuition, empathy, a moral compass, and human creativity.

To understand what's different about this new era, it helps to compare it to the two previous eras in the evolution of information technology. The tabulating era began in the nineteenth century and continued into the 1940s. Mechanical tabulating machines automated the process of recording numbers and making calculations. They were essentially elaborate mechanical abacuses. People used them to organize data and make calculations that were helpful in everything from conducting a national population census to tracking the performance of a company's sales force. The programmable computing era—today's technologies—emerged in the 1940s. Programmable machines are still based on a design laid out by the Hungarian American mathematician John von Neumann. Electronic devices governed by software programs perform calculations, execute logical sequences of steps, and

store information using millions of zeros and ones. Scientists built the first such computers for use in decrypting encoded messages in wartime. Successive generations of computing technology have enabled everything from space exploration to global manufacturing-supply chains to the Internet.

Tomorrow's cognitive systems will be fundamentally different from the machines that preceded them. While traditional computers must be programmed by humans to perform specific tasks, cognitive systems will learn from their interactions with data and humans and be able to, in a sense, program themselves to perform new tasks. Traditional computers are designed to calculate rapidly; cognitive systems will be designed to draw inferences from data and pursue the objectives they were given. Traditional computers have only rudimentary sensing capabilities, such as license-plate-reading systems on toll roads. Cognitive systems will augment our hearing, sight, taste, smell, and touch. In the programmable-computing era, people have to adapt to the way computers work. In the cognitive era, computers will adapt to people. They'll interact with us in ways that are natural to us.

Von Neumann's architecture has persisted for such a long time because it provides a powerful means of performing many computing tasks. His scheme called for the processing of data via calculations and the application of logic in a central processing unit. Today, the CPU is a microprocessor, a stamp-sized sliver of silicon and metal that's the brains of everything from smartphones and laptops to the largest mainframe computers. Other major

components of the von Neumann design are the memory, where data are stored in the computer while waiting to be processed, and the technologies that bring data into the system or push it out. These components are connected to the central processing unit via a "bus"—essentially a highway for data. Most of the software programs written for today's computers are based on this architecture.

But the design has a flaw that makes it inefficient: the von Neumann bottleneck. Each element of the process requires multiple steps where data and instructions are moved back and forth between memory and the CPU. That requires a tremendous amount of data movement and processing. It also means that discrete processing tasks have to be completed linearly, one at a time. While we have introduced some parallelism, it's not enough. For decades, computer scientists have been able to rapidly increase the capabilities of CPUs by making them smaller and faster. But we're reaching the limits of our ability to make those gains at a time when we need even more computing power to deal with complexity and big data. And that's putting unbearable demands on today's computing technologies—mainly because today's computers require so much energy to perform their work.

What's needed is a new architecture for computing, one that takes more inspiration from the human brain. Data processing should be distributed throughout the computing system rather than concentrated in a CPU. The processing and the memory should be closely integrated to reduce the shuttling of data and instructions back and forth. And discrete processing tasks should be executed

simultaneously rather than serially. A cognitive computer employing these systems will respond to inquiries more quickly than today's computers; less data movement will be required and less energy will be used.

Today's von Neumann–style computing won't go away when cognitive systems come online. New chip and computing technologies will extend its life far into the future. In many cases, the cognitive architecture and the von Neumann architecture will be employed side by side in hybrid systems. Traditional computing will become ever more capable while cognitive technologies will do things that were not possible before. Already, cloud, social networking, mobile, and new ways to interact with computing from tablets to glasses are fueling the desire for cognitive systems that will, for example, both harvest insights from social networks and enhance our experiences within them.

Should we fear the cognitive machines? MIT professors Erik Brynjolfsson and Andrew McAfee warn in their book, *Race Against the Machine*, that one of the side effects of this generation of advances in computing is they are coming at the expense of existing jobs. We believe, though, that the most important effect of these technologies will be in assisting people to do what they are unable to do today, vastly expanding the problems we can solve and creating new spheres of innovation for every industry. And like previous eras of computing, this will take a tremendous amount of innovation over decades. "These new capabilities will affect everything. It will be like the discovery of DNA," predicts Ralph Gomory, a pioneer of

applied mathematics who was director of IBM Research in the 1970s and 1980s and later head of the Alfred P. Sloan Foundation.[6]

HOW COGNITIVE SYSTEMS WILL HELP US BE SMARTER

As smart as human beings are, there are many things that we can't do or simply can't process in time to affect the outcome of a situation. Cognitive systems in many cases help us overcome our limitations.

COMPLEXITY

We have difficulty rapidly processing large amounts of information. We also have problems understanding the interactions among elements of large systems, such as the interplay of chemical compounds in the human body or the dynamics of financial markets. With cognitive computing, we will be able to harvest insights from huge quantities of data to handle complex situations, make more accurate predictions about the future, and better anticipate the unintended consequences of actions.

City mayors, for instance, already can begin to make sense of the interrelationships among urban subsystems—everything from electrical grids to weather to subways to demographic trends to issues reported or expressed by citizens. One example is monitoring social media during a major storm to spot patterns of words and images that

indicate critical problems in particular neighborhoods. Much of this information will come from sensors—video cameras, instruments that detect motion, and devices that spot anomalies. Mobile phones will also be used as anonymized sensors that help city planners understand the movements of people and accurately predict the effects and financial impact of various actions.

EXPERTISE

With the help of cognitive systems, we will be able to see the big picture and make better decisions. This is especially important when experience in an area is limited or we're trying to address problems that cut across professional or practical domains.

For instance, police are beginning to gather crime statistics and combine them with information about demographics, events, building blueprints, and weather to produce better analysis and safer cities. Armed with abundant data, police chiefs can set strategies and deploy resources more effectively—even predicting where and when crimes are likely to happen. Patrol officers benefit by gaining a wealth of information about unfamiliar locations; situational intelligence will be extremely useful when they knock on someone's door. The ability to achieve such comprehensive understanding of situations at every level will be an essential tool for managing a city and will become one of the most important factors in the economic growth and competitiveness of cities.

OBJECTIVITY

We all possess biases based on personal experience, professional background, and intuition about what works and what doesn't, as well as the influence of group dynamics. Cognitive systems can make it possible for us to be more objective in our decision making.

Corporations will become socially networked businesses made up of people and systems in collaboration. Sophisticated analytic engines will understand how an organization works, its competitive environment, capabilities, resources, and ecosystem of partners. Professionals will experience data in new ways and will be prompted to consider a number of hypotheses based on inferences and scenarios constructed by a system using relevant contextual data. The system will make hard-to-spot connections and help guide individuals in achieving business goals.

IMAGINATION

Professional background, training, and experience make it difficult to envision dramatically different settings and ranges of choice. Cognitive systems will help us discover and explore new and contrarian ideas.

Today, it takes ten to fifteen years and up to $1 billion to bring a single new drug to market. The research-and-development teams of pharmaceutical companies are beginning to use cognitive systems to explore existing biological information in new ways and to discover new

treatments utilizing compounds that have already been proven safe in other applications. Furthermore, cognitive systems can spot hidden patterns in the text of thousands of published articles to help identify new opportunities for drug treatments. Using computer modeling and simulation, researchers can run experiments to validate these findings "in silico" much faster than can be done in a life sciences "wet" lab. With the aid of cognitive machines, researchers and engineers will be able to explore millions of combinations in ways that could shave years and hundreds of millions of dollars from the development process.

SENSES

We can only take in and make sense of so much raw, physical information. With cognitive systems, computer sensors combined with analytics software will vastly extend our ability to gather and process such information.

Imagine a world where individuals carry their own personal cognitive system in a handheld device. These personal cognitive assistants would carry on conversations with us and acquire knowledge about us, in part, from observing what we see, say, touch, and type so they can better anticipate our wishes. In addition, the assistant might be able to use sophisticated sensing to monitor threats to a person's well-being. If there's carbon monoxide in a room, for example, the device might be able to alert its user.

Over time, humans have evolved to be more successful as a species. We continually adapt to overcome our

limitations. This partnership with computers is simply the latest step in a long process of adaptation. Just as the personal computer, the Internet, mobile communications, and social networking have given rise to tens of thousands of software applications, Web services, and smartphone apps, the cognitive era will produce a similar explosion of creativity.

Think of the coming technologies as cognitive apps. At the enterprise level, corporations might use these apps for handling mergers and acquisitions, crisis management, competitive analysis, and product design. The technologies already exist to assist a team within a company that's in charge of sizing up acquisition candidates using a cognitive M&A app. The team can augment its knowledge of potential targets, many of which are private companies, by gaining a deeper understanding of data, including transactions and partnerships in the public domain, intellectual property filings, and insights from Web sharing and other social channels. The cognitive software maps the links and the nature of the interactions. The M&A team will also track the performance of previously acquired companies. Those insights, constantly updated and integrated in a learning system, will help the team identify risks and synergies and become smarter with each acquisition.

Moreover, such apps delivered from the cloud will help individuals tackle highly complex though common decisions, for example, selecting a college, making investments, choosing among insurance options, and purchasing a car or home. Say you're starting a family and you and your spouse want to move out of your tiny

apartment into a house in a neighborhood or town with good public schools. The cognitive house-hunting app will start by understanding your needs and sizing up places to live, school systems, commuting factors, and costs of living. The app will even tap into social networks to discover hidden truths, such as the fact that the property a house was built on was once part of landfill. You will be able to conduct an ongoing conversation with the app as you learn more about what's available and the app learns more about you and the things you care about. Then, when you begin shopping for houses and mortgages, the app will guide you through the thicket of information and considerations that will inform your decisions. And, finally, when the two of you are standing in the foyer of your dream house wondering if you really can afford it, and you're under pressure to make a bid immediately, you can consult your trusted app for advice on one of the most critical decisions you will ever make.

TECHNOLOGY BREAKTHROUGHS: OPPORTUNITIES AND NECESSITIES

Much of the progress in science and technology comes in small increments. Scientists and engineers build on top of the innovations that came before. Consider the tablet computer. The first such devices appeared on the scene back in the 1980s. They had large, touch-sensitive screens but weighed nearly five pounds and were an inch and a half thick. They were more like bricks than books, and about all

you could do with them was scrawl brief memos and fill out forms. After thirty years of gradual improvements, we have slim, light, powerful tablets that combine the features of a telephone, a personal computer, a television, and more.

There's nothing wrong with incremental innovation. It's absolutely necessary, and, sometimes, its results are both delightful and transformational. A prime example is the iPhone. With its superior navigation and abundance of easy-to-use applications, this breakthrough product spawned a burst of smartphone innovation, which combined with the social-networking phenomenon to produce a revolutionary shift in global human behavior. Yet, technologically, the iPhone was built on top of many smartphone advances that preceded it.

New waves of progress, however, require majorly disruptive innovations—things like the transistor, the microchip, and the first programmable computers. These are the advances that fundamentally change our world.

Today, many of the core technologies that provide the basic functions for traditional computers are mature; they have been in use for decades. In some cases, each wave of improvements is less profound than the wave that preceded it. We're reaching the point of diminishing returns. Yet the demands on computing technology are growing exponentially.

Soon incremental innovation will no longer be sufficient. People who demand the most from computers are already running into the limits of today's circuitry. Michel McCoy, director of the Simulation and Computing Program at the U.S. Lawrence Livermore National Laboratory,

is among those calling for a nationwide initiative involving national laboratories and businesses to come up with radical new approaches to microprocessor and computer-system design and software programming. "In a sense, everything we've done up until this point has been easy," he says. "Now we have reached a physics-dominated threshold in the design of microprocessors and computing systems which, unless we do something about it, is essentially going to stagnate progress."[7]

We need more radical innovations. In the years ahead, a number of fundamental advances in science and technology will be required to make progress. In order to appreciate the scope and breadth of these innovations, we can conceptualize a cognitive system as one of those colorful Russian wooden dolls where each of the smaller dolls nests inside a slightly larger one.

The top layer is the way we interact with computers and get them to do what we want. The big innovation at this outer layer is "learning systems," which we will explore more deeply in chapter 2. The goal is to create machines that do not require as much programming by humans. Instead they'll be "taught" by people, who will set objectives for them. As they learn, the machines will work out the details of how to meet those goals.

The next layer represents how we organize and interpret data, which we'll discuss in chapter 3. Today's databases do an excellent job of organizing information in columns and rows. Tomorrow's are being designed to manage huge volumes of different kinds of data, place information in context, and crunch data in real time.

The next layer represents the architecture or design of systems—how we fit together all the physical components that make up a computer. The challenge here, which we address in chapter 5, is creating data-centric computers. The designers of computing systems have long treated logic and memory as separate elements. Now, they will meld the components together, first, on circuit boards and, later, on single microchips. Also, they'll move the processing to the data, rather than visa versa.

Finally, in the innermost layer is nanotechnology employed in building core components of a cognitive system, where we manipulate matter at the molecular and atomic scale. In chapter 6, we'll explore what it will take to invent a new physics of computing. To overcome the limits of today's microchip technology, scientists must shift to new nanomaterials and new approaches to switching from one digital state to another. Possibilities include harnessing quantum mechanics or chips driven by "synapses and neurons" for data processing.

A NEW CULTURE OF INNOVATION

We're still in the early stages of the emergence of this new era of computing. Progress will require a willingness to make big bets, take a long-term view, and engage in open collaboration. We'll explore the elements of the culture of innovation in each of the subsequent chapters in what we call the *journeys of discovery*. An absolutely critical aspect of the culture of innovation will be the ambition and

capabilities of the inventors themselves. For rapid progress to be made in the new era of computing, young people must be inspired to become scientists, and they must be educated by teachers using superior tools and techniques. They have to be rewarded and given opportunities to challenge everything we think we know about how the world works. It requires dedication and investment by all of society's institutions, including families, local communities, governments, universities, and businesses.

When we ask scientists at IBM Research what motivates them, the answer is often that they want to change the world—not in minuscule increments but in great leaps forward. Mark Ritter, a senior manager in IBM Research's Physical Sciences Department, leads an effort, inspired by the human brain, to rethink the entire architecture of computing for the era of cognitive systems. As a child, Mark, whose father was a plumber, had an intense curiosity about how things work on a fundamental level. It was his good fortune that his grandparents, who lived near his family in Grinnell, Iowa, had two neighbors who were physics professors at Grinnell College. One of the physicists, whom Mark pestered with science questions while the neighbor repaired his VW in the driveway, lent Mark a book on particle physics when he was about twelve years old. As a teenager, Mark bicycled over to the campus to attend physics lectures. He built a simple gas laser in the basement of his home. It was the beginning of decades of inquiry into how things work and how they can work better. A few years ago, after more than twenty years in IBM Research, Mark and his colleagues recognized that

the computing model designed for mid-twentieth-century demands was running out of gas. "This is the most exciting time in my career," Mark says. "The old ways of doing things aren't going to solve efficiently the big, real-world problems we face."

For his part, Dr. Larry Norton of Memorial Sloan-Kettering Cancer Center is driven to transform the way medicine is practiced. Ever since he can remember, he was motivated by the desire to do something with his life that would improve the world. Born in 1947, he grew up at a time when people saw science as a powerful means of solving humanity's problems. He recalls a thought-crystallizing experience when he was an undergraduate at the University of Rochester. He lived in a dorm where students often gathered in the mornings for freewheeling discussions of politics, values, and ethics. He was already contemplating a career in medicine, and the topic that day was, if you were a doctor and had done everything medical science could offer to save a patient but she died anyway, how would you feel? The students were split. "I realized I would feel terrible about it," he says. "Offering everything available isn't enough. I should have done better. And since, because of limitations in the world's knowledge, I couldn't do better, I should be involved in moving things forward."

During his forty-year career, Larry has been an innovator in cancer treatment. Among his contributions is the central role he played in developing the Norton-Simon hypothesis, a once-revolutionary but now widely used approach to chemotherapy. In Larry's years as a clinician,

he has saved the lives of many patients, but, of course, some of his patients have died. Those deaths haunt him. He believes that he owes it to those people and to their children to improve the treatment of cancer and, ultimately, to help eradicate the disease. He sees his work with Watson as another way to contribute. Through a machine, he can share his knowledge and expertise with other physicians. He can help save cancer victims from a distance—people he has never met.

You will meet a host of innovators in this book. There's Dharmendra Modha, the IBM Research manager who is leading a team of IBM and university scientists in a quest to mimic the human brain. And there's Murali Ramanathan, a professor of pharmaceutical sciences and neurology at State University of New York, Buffalo, who is studying the role of genetic and environmental factors in multiple sclerosis. These innovators are remarkable people. We depend on them to produce the surprising advances that knock the world off kilter and, ultimately, have the potential to make it a better place. We will need many of them to make the transition to the era of cognitive systems. In the end, this era is not about machines but about the people who design and use them.