3

HANDLING BIG DATA

The global financial crisis of 2008 had many causes, but, boiled down to the essentials, it comes to this: Institutions and individuals didn't fully appreciate the risks they were taking. Lenders issued high-risk home loans and securitized them without performing enough due diligence. Investment banks packaged and resold them in ways that obscured the underlying risks. Credit-rating agencies gave the securities their seal of approval. Investors around the globe snapped them up. Government regulators failed to spot flaws in the system.[1]

The crisis was a failure of information. Though institutions and individuals had access to immense quantities of data about both the financial system as a whole and their own parts in it, they were not capable of seeing what was going on well enough to make prudent decisions. Major changes are needed to prevent this kind of thing from happening again, including more effective regulatory monitoring and more responsible behavior by participants in the system. But there's an additional element that could

help prevent such a calamity: the emergence of technology designed to help people understand the workings of the entire system more deeply so they can better assess the risks they face. Think of it as an early-warning system. The financial crisis is a dramatic example of where technologists have the ability to harness data to help solve a complex and critical problem.

Remember the metaphor of the nested Russian dolls that we used earlier to illustrate the layers of cognitive computing? In chapter 2, we described the outermost layer—the transformation of computers into powerful learning systems. Chapter 3 explores the doll that nests inside that one—the new approach that these computers will need for managing and analyzing data.

Today, we are witnessing the emergence of a new force in society and business: big data. Organizations and individuals are faced with a torrent of data, everything from structured information such as transaction records to a wide variety of unstructured information—still images, video, audio, and sensor data. The biggest new source of data is the so-called Internet of things, data produced by sensors and harvested via the Internet. The sensors involved range from the RFID tags that retailers use to track merchandise to video cameras that capture the flow of traffic. Every day, as a group, human beings generate about 3 exabytes of computer data—a prodigious output that is expected to produce a data universe of 40 zettabytes of digital stuff by 2020.[2] A zettabyte is a decidedly *big* number: a 1 followed by 21 zeros. One zettabyte of storage would hold 250 billion two-hour HD movies. This flood of information should be

tremendously valuable. When you get down to it, big data is nothing less than the digital expression of life in the raw, in all of its richness and ambiguity. As we mentioned in chapter 1, it's useful to think of all of these data as a new natural resource, potentially even more valuable than resources like oil, coal, or natural gas because data are an essential ingredient in everything we do.

But a resource isn't worth much unless you can take full advantage of it. Today, less than 1 percent of the digital data that have been collected is actually analyzed.[3] This profusion of data is difficult to capture, make sense of, and move around. And, unfortunately, today's computing systems aren't up to the task of handling all of this raw information in an efficient and affordable way. That's why we need to develop new systems for managing big data. We need a new generation of data storage, management, and analytics tools that will improve our ability to gather, meld, and make sense of huge amounts of data and, in some cases, to perform a complex ballet of tasks in real time.

The new generation of tools must be designed to handle the four Vs of big data: *volume*, *variety*, *velocity*, and *veracity.*

- *Volume:* The sheer amount of data we're gathering is a challenge in itself. While the cost of transporting and storing a unit of data declines steadily, the amount we collect is increasing at a much faster rate, so we have to find new ways of moving and storing it.
- *Variety:* Until fairly recently, most business and government information was stored in simple spreadsheets or

relational databases made up of columns and rows—easily accessed using standard queries. But a host of newer data types, including video, geospatial information, webpages, and speech, require new ways of managing data.

- *Velocity:* In some scenarios, dealing with the velocity of data will be essential. So rather than first gathering and storing data in databases, we must develop effective ways of analyzing data in motion, drawing actionable insights, and keeping only the bits that are essential.
- *Veracity:* Much of the new data we have access to now are not precise or certain—or clearly organized. A lot of it is just so much noise. The challenge is to make sure that we're getting the right information, that it's accurate, and that we're able to draw the right conclusions from it.

HOW TO SOLVE THE BIG DATA PROBLEM

A tremendous amount of progress has been made in the way we handle data since the dawn of the programmable computing era. The first commercial disk drive, introduced by IBM in 1956, was the size of a small car and stored just five megabytes of data. Today, a standard laptop computer comes with a two-and-a-half-inch disk drive capable of storing 1 terabyte of data, one million megabytes. Early database-management systems stored information in lists organized by categories and subcategories. The computer user had to manually navigate through a series of lists to find a particular nugget of information. In contrast, today's relational databases store information in columns

and rows that can be searched with standard query languages. Early computers calculated probabilities of particular outcomes based on statistical analysis of data related to relatively narrow problems. Today's advanced analytics software programs find patterns in large sets of data and extract meaning from them.

In order to handle the four Vs of big data, though, major disruptive advances will be required in the storage, management and analysis of data.

VOLUME

One of the most significant breakthroughs we see coming for storage of huge amounts of data is the ability to use cognitive techniques to more efficiently manage the flow of data within computing systems. Today, data are stored using a variety of technologies, depending on how often access is needed and the cost of the storage medium. Magnetic tape is the cheapest form of storage, but data retrieval is slow. Disk drives are more expensive but faster to access. Data storage on memory chips is lightning quick but even more expensive. Typically, database administrators move data from one technology to another based on rigid rules and schedules.

But that approach isn't satisfactory in the era of big data. So scientists at the IBM Research lab in Zurich, Switzerland, are developing a sophisticated system for optimizing the storage of data for particular uses by considering how much it costs to store the data and how quickly the computer system will need to fetch specific

pieces of information. The system will learn by studying the application it's being asked to manage. Then it will make adjustments as it monitors the application in action—putting bits of data in the right places at the right times.[4]

For instance, an application for monitoring security on a city street will store most of the video and audio it gathers on tape or disk drives. But when it picks up anomalies that might indicate security threats, it will store data about them on memory chips and revisit that information frequently as additional sensor data are gathered. Only the most relevant data are stored on the most expensive media, and only for the time they are needed.

VARIETY

Over the past decade or so, computer scientists and mathematicians have become quite proficient at handling specific types of data by using specialized tools that do one thing very well. For example, a sales manager can employ data analytics software to reveal past sales trends and forecast future sales, using a specific tool for a specific data type. But that approach doesn't work for complex operational challenges such as managing cities, global supply chains, or power grids, where many interdependencies exist and many different kinds of data have to be taken into consideration.

What's required is a holistic approach to data management and analysis. You have to be able to mash together different kinds of data and act on them with different

kinds of tools. An early example of this approach is a project that IBM researchers in New York and Brazil completed on behalf of the city of Rio de Janeiro. Rio has recurring flooding and landslide problems in many hilly neighborhoods. The researchers used data describing the physics of the atmosphere to create a mathematical model of how storms are likely to unfold in Rio. With it, they can predict up to forty hours ahead of time how much rain will fall in a particular location with 90 percent accuracy. They also used topographical data and historical flood information to create a model of how the water would flow in the streets once it hit the ground. After combining the two models, they were able to predict flooding and landslide conditions. They also consider factors such as traffic patterns on city streets. In this way, they're making it possible for the city to anticipate flooding and landslides and put resources in place ahead of time to deal with them—including deciding where to deploy emergency crews and where to open shelters. By combining the different data types and techniques in this way, they created a system that gathers a wide variety of information, evaluates it systematically, and reaches conclusions the way humans do.

VELOCITY

Most of today's computing tasks involve data that have been gathered and stored in databases. The data make a stationary target. But, increasingly, vitally important insights can be gained from analyzing information that's on the move. This technology can be used for everything

from real-time stock trading to drilling for oil a half-mile below the surface of the ocean to monitoring patients' health in hospitals. Consider the benefits of real-time monitoring of the vital signs of preemies in a neonatal ward. Babies don't show signs of infection the way adults to. The danger is that they will become very sick before the problem is detected and die before an effective treatment can be administered. But by monitoring their vital signs and detecting subtle patterns that indicate an infection has set in, physicians can spot problems and begin treatments twelve to twenty-four hours earlier—potentially saving lives.

This approach is called streams analytics. Rather than placing the data in a database first, the computer analyzes it as it comes in from a variety of sources, continually refining its understanding of the data as conditions change. This is the way humans process information. Streams analytics programs are extremely sophisticated. Each kind of data must be monitored on its own, translated so it is compatible with other data types, integrated with other data, and analyzed again in the context of the overall flow of information. The translation part is particularly challenging. Spoken words, for instance, have to be converted to machine-interpretable form.

In the coming years, streams programs will become more cognitive and adaptive, explains IBM researcher Nagui Halim, head of the streams technology project. He foresees a time when managers will provide a basic description of what they want to a system to do and the system will follow those specifications to design an

application made up of preprogrammed components and then launch the application to get the job done. As the applications gain real-world experience with the data, the software will continuously sharpen its understanding of what people are looking for and refine the answers it provides.[5]

VERACITY

Organizations face huge challenges as they attempt to get their arms around the complex interactions between natural and human-made systems. The enemy is uncertainty. In the past, since computing systems didn't handle uncertainty well, the tendency was to pretend that it didn't exist. Today, it's clear that that approach won't work anymore. So rather than trying to eliminate uncertainty, people have to embrace it.

One strategy is to build analytics software that makes it possible to identify an optimal way of getting things done based on mathematical probabilities while taking uncertainty into account. The technique is called stochastic optimization. It has been used for some time but is being further developed to handle the problems of big data. The optimization discipline has its roots in the early days of computing. One of the pioneers in the field was Ralph Gomory, who became the head of IBM Research. After he got his Ph.D. in mathematics from Princeton in 1954 and served in the U.S. Navy, the navy hired him as a consultant and asked him to use mathematical techniques to help design the optimal naval taskforce—how many

aircraft carriers, battleships, and so on. He invented a new method, called integer programming, to help solve the problem. Initially, he did this work with pencil and paper, then, with a mechanical calculating machine. Finally, he was invited by the Rand Corporation, a think tank in California, to use their large-scale computer, one of only a few that existed at the time.[6]

Ralph created a program that could compute an optimal design given a specific set of choices. But today we're faced with immensely complicated problems that contain so many possibilities that it would take a long time, even with a powerful computer, to explore all of them using Ralph's deterministic techniques. In fact, solving some of the more complex problems would require more computing power than exists on earth. That's why in the era of big data we need to use stochastic optimization to get our answers.

The term *stochastic* is derived from a Greek word for "aim." It refers to the fact that when an archer shoots a series of arrows at a target, most will miss the bull's-eye and will create a cluster around it. Where each arrow falls depends on a combination of factors, including the skill and visual acuity of the archer, his emotions, the quality of the bow and arrow, weather conditions, and the distance from the target. Stochastic optimization takes into account such variables.

To perform stochastic optimization, computer scientists use probability theory, the branch of mathematics concerned with the analysis of random phenomena, to analyze complex situations and figure out how to achieve

the best outcome while allowing for the effects of variability. The first step is building a mathematical model—translating a real-world situation into a formula. They use probability theory to identify the mostly likely effects of the variables then perform a series of analyses or simulations, based on a subset of the possibilities, to determine the most likely outcomes. Previously, using less sophisticated techniques, it might take days of computation to arrive at a useful result; now it can be done in seconds. In addition, as models are tested against real-world outcomes, they learn and get better over time.

To see how stochastic optimization works, consider an electricity grid. Today, many electrical-distribution systems are outfitted with smart meters that make it possible for consumers and operators of the system to know how much energy is being used in real time. Based on that information, consumers can make informed choices about their consumption levels, and operators can predict with a reasonable certainty what demand will be at a specific time. But what happens if the weather changes suddenly or there's a failure in the system and power is in short supply? Typically, the operator would go to the spot energy market and pay a premium for additional power. In a worst-case scenario, the utility might not be able to meet demand, resulting in service brownouts.

Using stochastic optimization, mathematicians at IBM Research have built a computing system that can predict what incentives would be required for consumers to reduce demand in response to a sudden change in supply. Rather than having to decide what to do on a case-by-case

basis, consumers can preprogram their smart energy-management systems to respond instantly and to dial back energy use under certain circumstances. Over time, cognitive computer programs will be designed that learn from the ongoing interactions and continuously optimize the incentive systems.[7] The consumer's devices will be in touch with the operator's devices, and, as a result of their automated negotiations, the operators will be able to avoid paying a premium for energy on the spot market and, in some cases, to prevent service outages.

Stochastic optimization addresses some of the challenges posed by the uncertainty inherent in big data, but not all of them. Machines, on their own, don't necessarily recognize discrete pieces of information that are related to one another but are scattered among multiple data storehouses. How do they sort through all the evidence and draw conclusions? They must learn to understand context. Today, the most advanced approach to achieving that aim is an emerging class of technology called entity analytics. The software collects diverse data about real-world entities (people, places, and things) and makes it possible to identify who or what they are and to figure out the relationships among them.

One of the pioneers of the field of contextual analytics is Jeff Jonas, the chief scientist of IBM Entity Analytics. Jeff took an unusual path to get to IBM. A serial entrepreneur for two decades, he developed the entity analytics technology to help casinos quickly detect the presence of known cheaters and card counters. His big breakthrough came after a casino asked him to find a way to ensure it was not

doing business with people who had been barred by the gaming regulators. At the time, this casino had eighteen different kinds of watch lists and information available from other internal sources, ranging from job applications and hotel reservations to enrollment in the loyalty club. Jeff's technology, known as NORA, for Non-Obvious Relationship Awareness, helped casinos make connections among related pieces of information, which, among other things, helped them spot the notorious MIT card-counting team that was featured in the best-selling book *Bringing Down the House* and the movie *21*. IBM bought Jeff's company, Systems Research & Development, and since then he has added new features, including the ability to analyze streams of information in real time.

Jeff's software essentially builds and maintains a giant jigsaw puzzle on the fly, making connections between newly acquired data and older information. With each additional piece of information, more context accumulates, and the software's insights and understanding improve. The technology might conclude, for example, that the Jonathan W. Smith who is applying for a credit card in New York City is the Jon Smith who was convicted of credit card fraud in Jersey City five years earlier. The system draws a conclusion based on data matches, such as date of birth, previous addresses, phone numbers, and other bits of data.

The system is quite nimble. New data can be ingested in real time, and new insights can be drawn from them immediately. It can even change its mind about a conclusion it has reached based on new evidence. And the system

can even spot connections that are meaningful and alert people to them before they have to ask. Says Jeff: "Because there are not enough humans to dream up and ask every question every day, the data must find the data and the things that are relevant must find you."[8]

We're still in the early days of being able to gather and understand context in effective ways. The IBM Research scientist Sam Adams foresees the emergence of computing systems that will gather huge amounts of information from a large number of sources, find patterns and correlations and linkages in it, and then apply specialized information-management software that Sam calls *context engines*. The process is like finding many needles in a field full of haystacks, combining those needles, then searching the smaller but still substantial collection for the needles made of gold—the truly valuable slivers of insight.[9]

All of this may sound frightfully inefficient, but it's the way we humans make sense of the world around us. We have experiences, gather sense- and language-based information, see patterns, learn from successes and mistakes, draw conclusions, and act on them. Increasingly, machines will operate in a similar way. Computing systems will need different kinds of data-organizing structures to accomplish this. A technology called the graph database could become essential. This allows people to envision a web of relationships that lies hidden in collections of unstructured data. For instance, consider how the technology could be used in mining insights from social networks. Jim Smith, who lives in New York City, has friendships on social-networking sites with a number of people who, like

him, list Bruce Springsteen among their favorite performers. Several of the others live in Buffalo and went to the same university as Jim, the State University of New York at Buffalo. Some of the others have indicated on their social networks that they "like" the Chipotle restaurant chain. Springsteen is performing in Buffalo next month. Armed with this web of information, an Internet travel site might consider offering Jim an airline and entertainment package at a discount combining airfare to Buffalo, a night in a hotel, two tickets to the Springsteen concert, and a coupon for dinner at Chipotle. The travel site's computer system will use the graph database in much the same way that a human being makes connections among his or her relationships and interests.

THE EMERGENCE OF PERVASIVE ANALYTICS

We're at an inflection point in the evolution of data analytics. The combination of massive amounts of information and the tools to deal with it creates a new opportunity that has huge implications for business, society, and individuals. Today, computers are everywhere, thanks, in large part, to the revolution in mobile communications that has brought us all manner of smartphones and digital tablets. Now we're on the front edge of a second wave of everywhere-ness as data analytics also becomes pervasive. Thanks to mobile communications, we no longer need to be tethered to the monitor and keyboard to get new insights because analytics allows individuals

and organizations to have the information they need when they need it. The insights from analytics will soon no longer be reserved for an elite few. In the future, these insights will be available to just about anybody.

Pervasive analytics will enrich our private lives, as well. People love the apps on their smartphones that allow them to ask a simple question, such as, "Where's the nearest pizzeria?" and get the answer via a friendly robotic voice in a matter of seconds. But what if you could ask more difficult questions, such as, "Can I afford to pay for the car?" with the implicit assumption that the app already knows you need to help out with your kid's college tuition next year, or, "Is it worth fixing my car?" knowing that the app is tuned in to the mileage, maintenance record, and book value of your car. Suddenly, you can get not just handy information but useful insights that help you conduct your life more successfully.

SCENARIO: THE COGNITIVE ENTERPRISE

In the coming era of cognitive systems, organizations will acquire powerful new capabilities for using big data to make better decisions. Dario Gil, an executive at IBM Research who develops technology solutions for the energy and natural resources industries, believes that for organizations to maximize gains from cognitive systems, they will need to transform the way they operate to become more dynamic, transparent, and collaborative.

And he believes the spaces in which they operate will influence whether they can achieve the transformation.

That's why he designed the Cognitive Enterprise Lab, a special workspace at the IBM Research in Yorktown Heights, N.Y. As Dario sees it, the way companies do business is expressed in the way they organize their work environments. And most companies organize things in the same ways—with private offices for executives, cubicles for everybody else, conference rooms for group meetings, the boardroom, and, in some cases, special rooms for teleconferencing. Over time, the ways that people behave and interact in these rooms become codified. Traditional power hierarchies often determine who sits where and who says what. The status quo is reinforced. Think of it this way: ergonomics is destiny. The Cognitive Enterprise Lab is a designed to feel different from other work environments, signaling that the old rules of interaction have been suspended. The room is not even rectangular; instead, one of the four walls slants at an angle. The chairs and tables can easily be reconfigured depending on how the space is to be used.

But the technology is what truly sets the space apart. One entire wall is covered with a large high-definition visualization screen upon which presentations and graphics can be displayed and manipulated using hand gestures. Other screens designed for aiding group interactions can be positioned around the room by moving them via tracks suspended from the ceiling. Cameras capture speakers for video conferencing no matter where

they sit or stand in the room—and also record facial expressions and body language. (The video might be analyzed later by the computer system, and insights drawn from it can be given back to the speakers so they see how people reacted to specific moments of a presentation.) Acoustic arrays help with speech recognition and processing. The idea is to create the richest possible environment for people's interactions with data and one another in order to enhance the group's collective intelligence—a term used by MIT's Thomas Malone. Dario says, "The goal is to have the system partner with the humans inside the cognitive environments to transform the decision-making activities of teams that are dealing with very complex problems."[10]

Imagine how a large oil company might use cognitive capabilities in such a space a few years from now. A critical decision-making juncture for the oil-exploration business is whether to bid to secure leases on new oil fields. Today, typically, the analysis that leads to such decisions is performed in relative isolation by a handful of scientists, financial analysts, and executives who rely on information from seismic studies and publications from scholarly and trade journals. The information is then combined with their own experience-driven hunches about where rich and accessible oil deposits might be located. Geologists, geophysicists, petroleum engineers, and financial analysts each have their own sources of data, expressed in the vocabularies of their domains. The decision maker in charge of the bid speaks to each group of specialists separately to gather information and advice.

But what if all of the parties to the decision were united in an environment such as the Cognitive Enterprise Lab? There, the group could carry on its conversations in partnership with a learning system that has access to millions of pages of scientific and industry documents, data about the fields being considered, and information about the company's financial resources and about current and future market scenarios. The system could integrate the data, analytics, and simulations from different sources so the group can perform cross-domain analysis of the options. When considering how much oil will be available in a particular location and how much it will cost to extract, members can call up visualizations of the scientific data on the shared digital wall or personal screens, where the data are expressed in an individually tailored way. Because everyone is talking together and viewing and interacting with the data at the same time, it's less likely that there will be misunderstandings or that anybody will make recommendations or decisions that are not fully supported by the evidence.

This environment could be used in a wide variety of situations. A group of doctors could gather to evaluate a particularly complex case. Public health officials could manage a pandemic. Venture capitalists could size up a new category of investments. Consumer-electronics companies could plan the launch of new gizmos. Government bodies could develop new legislation. City managers could respond to natural disasters. In the Cognitive Enterprise Lab, it is more likely that the best ideas—backed by evidence—will win.

Dario's lab illustrates a key point about the era of cognitive systems. No longer will the deep analysis of complex situations belong only to data specialists. Instead, problem solvers of many different talents will be able to interact with one another and with data in ways that were not possible before.

JOURNEY OF DISCOVERY: THE ULTIMATE BIG DATA CHALLENGE

The Square Kilometre Array is one of the most ambitious scientific projects ever undertaken. Its organizers plan on setting up a massive radio telescope made up of more than half a million antennas spread out across vast swaths of Australia and southern Africa. When it's completed in 2024 or so, astronomers will be able to analyze the data to better comprehend the history of the universe and the nature of matter. The SKA is a scientific sibling to the Large Hadron Collider in Switzerland. There, scientists are studying the tiniest elemental particles for answers to some of the fundamental riddles of existence. In contrast, "Our laboratory is the whole universe," says Marco de Vos, managing director of ASTRON, the Netherlands Institute for Radio Astronomy, which, along with IBM, is proposing an information-technology system to manage the SKA's data.[11]

The SKA is an iconic example of the need in the coming era for what we call radical collaboration. It's such a complex project, requiring so many skills and so much investment, that many participants are required. Ten nations

are backing the project, and thousands of scientists and hundreds of companies are ultimately expected to play roles in designing, building, and operating it. IBM and ASTRON are melding their skills and resources in a true partnership, and they have invited other organizations to participate. Already, Square Kilometre Array South Africa, an agency of the South African government, has joined. In addition, ASTRON and IBM plan on tapping a global network of hundreds of subject experts from industry and academia to help design the system.

The SKA is the ultimate big data challenge. The telescope will collect a veritable deluge of radio signals from outer space—amounting to fourteen exabytes of digital data per day. (The data collected by the SKA in a single day would take nearly two million years to playback on an iPod.)[12] Because the telescope is to be made from so many individual antennas, the antennas are to be so widely scattered, and such a large volume of data is being gathered, a novel computing system must be developed to manage the process of gathering, storing, and analyzing data from end to end.

ASTRON and IBM have combined forces on a five-year initiative called DOME to create such a system. Their work together began years before, when they collaborated on the information-technology system for a radio telescope called LOFAR. LOFAR is radically different from conventional radio telescopes. Unlike the traditional large steel-mesh dishes pointed at the heavens, the more than 10,000 LOFAR antennas, which act like a single virtual telescope, are low profile and low cost. The antennas are set up in

clusters on vast fields in the Netherlands and elsewhere in northern Europe. The antennas don't move. The signals gathered by them are combined in a clever way to target objects of interest in the sky. ASTRON hopes LOFAR, which is now up and running, will be used as a model for the SKA telescope. It will be the test for the DOME project and one of the pilot projects testing the SKA antenna technology.

During a recent visit to one of the LOFAR fields, ASTRON's Albert-Jan Boonstra, a codirector of the DOME project, explained how the equipment works. One antenna type is made up of five-foot-tall posts that are held in place by four wires, which are also key elements of the antenna systems. A small disk at the top of the post contains electronics that amplify the radio waves and then transmit the data along underground cables to a nearby metal box the size of a small car, where more processing takes place. The other antennas are spear-shaped metal pieces arrayed in clusters on electronic circuit boards and covered with Styrofoam sheets and plastic tarps. Most of the data is sent by fiber-optic cables to an IBM Blue Gene supercomputer at nearby University of Groningen, where it is filtered and correlated.[13]

Albert-Jan grew up in a small village in the Netherlands and has been obsessed with space since watching the Apollo moon landing on TV as a ten-year-old. Inspired, he and a friend made their own rocket out of tin cans and wires, with gasoline for fuel—only to see it explode and burn on the launch pad. "I'm fascinated with the idea that

we're tiny humans in an endless, nearly empty universe. I wanted to see how far we can look into space," he says.

But Albert-Jan knows he won't be able to penetrate deep into space without some significant advances in computer science. In fact, the DOME project will require more than a half-dozen major breakthroughs because of the huge amount of data and the fact that the antennas are to be scattered over such a wide area. The signals from outer space are a combination of valuable data and meaningless noise and have to be processed to sort out the useful stuff. To control costs, designers of the computing system have to figure out how to minimize the amount of energy used for processing data. At the same time, since so much of the energy in computing is required to move data around, they have to discover ways to move the data as little as possible. Faced with these challenges, Albert-Jan and Marco engaged IBM Research–Zurich in a sort of show-and-tell for emergent technologies. In a series of meetings held at the lab, IBMers presented twenty proposals, of which ASTRON ultimately accepted seven, each intended to overcome a major information-technology hurdle posed by the SKA.

While these technology projects are focused on the SKA, they could help solve problems across a wide array of computer science fields, from data management and analytics to computer system design and chip design. The SKA "is an extreme situation," says Ton Engbersen of IBM Research–Zurich, the codirector of the DOME project. "When we solve it, we'll have a good handle on solving the other big data challenges."[14]

Several of the DOME subprojects feature cognitive computing technologies. The most audacious of them, called Algorithms & Machines, is focused on creating an ultrasophisticated software program that will help the team design the entire DOME system in a holistic manner—so the technology can handle the extreme data-processing demands of the SKA without breaking the international organization's budget. It's one of most demanding system-optimization challenges ever—the great-grandchild of Ralph Gomery's U.S. Navy fleet-optimization project in the 1950s.

Algorithms & Machines is the brainchild of Ronald Luijten, a Dutchman who has worked at IBM Research–Zurich for twenty-eight years. He's a student of computer history and was impressed when the supercomputing pioneer Seymour Cray recounted in a speech years ago that he designed the architecture for the Cray 1 supercomputer in the early 1970s by essentially locking five engineers in an office with him and a white board for two weeks. Such a feat would be impossible today. Ronald figures that it would require upward of 250 experts in a wide array of computer science domains to design the architecture of the DOME system. Hiring all those people and bringing them together isn't practical. So instead with Algorithms & Machines, he and his team are gathering all of the pertinent knowledge in a repository, setting the parameters for the entire computing system, and creating optimization algorithms. The system will learn what they want and prepare a recommendation on how to fulfill their needs. Ronald sees this effort as nothing less than an attempt to

map out the contours of the new era of computing. "This design-exploration tool will help us figure out how to design everything from the chip of the future to the data-center of the future," he says.

Big data creates gigantic challenges, but it also offers the potential of tremendous advances in the ways human beings use information. In these huge data-intensive projects—the SKA, Watson—global communities of scientists and domain specialists are gathering large data sets and building sophisticated models for understanding natural and human phenomena. In a sense, with this data and with the tools to understand it, we're creating a collective intelligence that can be shared and used by many for the betterment of humankind. That's the ultimate promise of big data.