The Big Idea: The Next Scientific Revolution

by Tony Hey

From the Magazine (November 2010)

A visitor walking the halls of Microsoft Research's campus in Redmond, Washington, today is likely to overhear discussions not only about computer science but about a surprising variety of other subjects, from which way a galaxy rotates, to a new AIDS vaccine, to strategies for managing the planet's precious supply of fresh water.

What could these issues possibly have in common? And why would Microsoft—ostensibly a software company—be involved with them? The simple answer is data—vast amounts of data. So vast that when we run the programs that analyze some of the databases, the temperature of the building that houses 10,000 microprocessors shoots up several degrees. Today our computer scientists find themselves in partnership with leading scientists in a wide array of disciplines—astronomy, biology, chemistry, hydrology, oceanography, physics, and zoology, just to name a few—working on efforts such as drug development, alternative energy, and health care cost containment. And, yes, even commercial software projects. We believe that a new generation of powerful software tools, which support collaboration and data exploration on an unprecedented scale, are about to enable revolutionary discoveries in these fields.

For decades computer scientists have tried to teach computers to think like human experts by embedding in them complex rules of linguistics and reasoning. Up to now, most of those efforts have failed to come close to generating the creative insights and solutions that come naturally to the best scientists, physicians, engineers, and marketers. The most talented experts not only have a deep understanding of data but also are able to see the possibilities "between the columns"; they can find the nonobvious connections within or between disciplines that make all the difference.

We have reached a point, however, where even the experts are drowning in data. Digital information is streaming in from all sorts of sensors, instruments, and simulations, overwhelming our capacity to organize, analyze, and store it. Moore's Law has for decades accurately predicted that the number of transistors that could be placed on an integrated circuit would double every two years, and until recently, this decrease in transistor size was accompanied by increased microprocessor performance. To increase performance today, we must program multiple processors on multicore chips and exploit parallelism. The multicore revolution has arrived just as we face an exponential increase in data. That increase is not a challenge we can address with patches and upgrades; we must rethink our whole approach to data-intensive science. Which is why, several years ago, our colleague and Turing Award winner, the late Jim Gray, proposed what he called "the fourth paradigm" for scientific exploration. Jim's vision of powerful new tools to analyze, visualize, mine, and manipulate scientific data may represent the only systematic hope we have for solving some of our thorniest global challenges.

The Four Paradigms of Science

Experimentation

Beginning in ancient Greece and China, people tried to explain their observations through natural laws instead of supernatural causes.

Theory

By the 17th century, scientists like Isaac Newton tried to make predictions for new phenomena and

would verify hypotheses by conducting experiments.

Computation and Simulation
The advent of high-performance computers in the latter half of the 20th century allowed scientists to explore regimes inaccessible to experiment and theory, such as climate modeling or galaxy formation, by numerically solving systems of equations on a large scale and in fine detail.

Data Mining
Using more-powerful computers, scientists begin with the data and direct programs to mine enormous databases for relationships. In essence, they use computers to discover the rules by studying the data.

The first two paradigms for scientific exploration and discovery, experiment and theory, have a long history. The experimental method can be traced back to ancient Greece and China, when people tried to explain their observations through natural rather than supernatural causes. Modern theoretical science originated with Isaac Newton in the 17th century. After high-performance computers were developed in the latter half of the 20th century, Nobel Prize winner Ken Wilson identified computation and simulation as a third paradigm for scientific exploration. Detailed computer simulations capable of solving equations on a massive scale allowed scientists to explore fields of inquiry that were inaccessible to experiment and theory, such as climate modeling or galaxy formation.

By the Numbers
400
The estimated number of agencies involved in managing California's water supply. Coordinating and analyzing the data they generate can make water management more effective.

$10 Million
The amount one health care provider discovered in underpayments within the first six months of using deep-data-mining tools

$500 Versus $15,000
The cost of monitoring a patient that data-driven techniques predict will be rehospitalized, versus the potential cost of treating a readmitted patient

The fourth paradigm also involves powerful computers. But instead of developing programs based on known rules, scientists begin with the data. They direct programs to mine enormous databases looking for relationships and correlations, in essence using the programs to discover the rules. We consider big data part of the solution, not the problem. The fourth paradigm isn't trying to replace scientists or the other three methodologies, but it does require a different set of skills. Without the ability to harness sophisticated computer tools that manipulate data, even the most highly trained expert would never manage to unearth the insights that are now starting to come into focus.

Saving Lives with "Machine Learning"
Let's start with an example of the kind of thinking that drives this type of research. In the 1980s my colleague Eric Horvitz, while training at a Veterans Administration hospital as part of his medical education, observed a disturbing phenomenon. During the holiday season, the hospital experienced a surge in admissions for congestive heart failure. Each year, some patients who had otherwise successfully managed their health despite a weakened heart would reach a tipping point after a salty

holiday meal. That extra salt caused their bodies to retain additional fluids, which would lead to lung congestion and labored breathing—and often to a visit to the emergency room.

Those post-turkey collapses were expensive in every sense of the word. They could be fatal for some patients—sometimes quite rapidly, sometimes by causing a downward spiral of failing physiological systems that took days to weeks. Other, luckier patients were effectively stabilized, but most still required a stay of a week or more that would typically cost the VA system $10,000 to $15,000 a patient. (Today those bills would be far higher.)
More than two decades later, Eric and his colleagues at Microsoft Research have developed analyses that can predict with impressive accuracy whether a patient with congestive heart failure who is released from the hospital will be readmitted within 30 days. This feat is not based on programming a computer to run through the queries a given diagnostician would ask or on an overall estimate of how many patients return. Rather, this insight comes from what we call "machine learning," a process by which computer scientists direct a program to pore through a huge database—in this instance, hundreds of thousands of data points involving hundreds of evidential variables of some 300,000 patients. The machine is able to "learn" the profiles of those patients most likely to be readmitted by analyzing the differences between cases for which it knows the outcome. Using the program, doctors can then plug in a new patient's data profile to determine the probability of his or her "bouncing back" to the hospital.

In one sense we owe this project to a human expert spotting a nonobvious connection: Eric not only earned his MD but also has a PhD in computer science, and he realized that machine-learning techniques similar to the ones he and his team had used to analyze Seattle traffic patterns could work for this important health care challenge. In 2003 they had developed methods of predicting traffic jams by analyzing massive quantities of data, which included information on the flow of traffic over highways, weather reports, accidents, local events, and other variables that had been gathered over several years. The team's new program compared data about patients who were and were not readmitted, and unearthed relationships among subtle evidence in a patient's clinical history, diagnostic tests, and even socioeconomic factors, such as whether the patient lived alone. This integration was not trivial: Information on a patient's living situation, for example, may reside in a social worker's report, not on a medical chart. It is unlikely that a single clinician involved in a patient's care could ever process the volume of variables sufficient to make a prediction like this.

The economic impact of this prediction tool could be huge. If physicians or hospitals understand a patient's likelihood of being readmitted, they can take the right preventive steps. As Eric explains: "For chronic conditions like congestive heart disease, we can design patient-specific discharge programs that provide an effective mix of education and monitoring, aimed at keeping the patients in stable, safe regimes. Such programs can include visits or calls from a nurse, or special scales that indicate dangerous changes in a patient's fluid balance and communicate them to the doctor. If we can spend even $500 or $1,000 on postdischarge programs for patients who have the highest likelihood of being rehospitalized, we can minimize readmissions and actually save money while enhancing health outcomes."

It's no wonder that health insurers and hospital chains are lining up to talk about this. And it doesn't take much imagination to list other types of businesses that could benefit from this kind of data intensive discovery as well.

On Wall Street, massive data-mining programs are already tracking "sympathetic movements," or related trading patterns among different investment vehicles. Hedge funds and large money managers are placing millions of dollars in bets every day based on these data-discovered relationships.

On the operational side of business, the possibilities are endless. Companies will be able to do massive analyses of customers and business opportunities using programs that unearth patterns in price, buying habits, geographic region, household income, or myriad other data points. The large quantities of available data on advertising effectiveness, customer retention, employee retention, customer satisfaction, and supply chain management will allow firms to make meaningful predictions about the behavior of any given customer or employee and the likelihood of gaps in service or supply. And more and more, we find companies using data techniques to spot irregularities in payments and receivables. These programs can predict, for example, the revenues that should be collected for a given list of delivered services. One health care provider we have worked with in New Mexico discovered $10 million in underpayments within the first six months of using such data-mining tools.

The relevance of the old joke "only half of all advertising dollars are successful—we just don't know which half" will be imperiled by the new analytical tools. An electronic entertainment company in the Philippines is using Microsoft data-mining technology to customize its sales pitches to individual customers, based on extensive analysis of such factors as past buying patterns, age, gender, financial profile, and location. Almost immediately after implementing this technique, the company saw its response rate for offers for ringtones and other products double.

With all those business opportunities, some ask why Microsoft Research is working on so many global health and environmental projects. After all, aren't those projects that the Bill & Melinda Gates Foundation might fund? Yes, but the reason Microsoft Research has several dozen computer scientists working on them is that they involve some of the most enormous data stores imaginable and constitute an invaluable testing ground. We need to expand our own thinking and the capabilities of our tools by working on the biggest problems out there, which happen to be of immense importance to humanity. Tackling these problems also opens more opportunities for collaboration and experiments. When there is a compelling incentive for experts in different disciplines to work together and share data in a transparent environment, we're likely to make the fastest progress. As Jim Gray used to say, astronomy data are valuable precisely because they have no commercial value.

Plug-and-Play Ocean Research

One such ambitious environmental project involves ocean science and is now under construction beneath the cool Pacific waters west of Washington State and British Columbia. It's impossible to overstate the importance of the oceans, which cover 70% of the Earth's surface and make up the planet's largest ecosystem. The oceans drive weather systems; are the source of powerful, still largely unpredictable hazards such as tsunamis and hurricanes; store much more carbon than the atmosphere, vegetation, and soil; and are a critical food source.

And yet, in many ways we understand more about the surfaces of Mars and Venus than about the seafloors. Water is opaque to the electromagnetic radiation that allows us to explore the heavens; that's why the mainstays of our oceanographic research have been submarines, ships, and satellites. That is about to change. On a patch of the Pacific's floor, oceanographers involved with the U.S. National Science Foundation's $600 million Ocean Observatories Initiative (OOI) have mapped out a network of nodes that is designed to offer what my colleague Roger Barga wryly calls "USB for the ocean." OOI will lay 1,500 miles of cable to and around the patch, providing power, internet access, and the ability to record and time-stamp data on phenomena scientists will study with all sorts of devices, ranging from simple temperature sensors to remote-controlled robots to state-of-the-art gene sequencers.

The project aims to involve scientists from all over the world. The ability to measure and analyze natural processes—such as silt buildup or changes in the density of microscopic organisms—is unprecedented. But the amount of information OOI will generate could swamp the effort if the data

aren't cleverly organized and stored. That's why Roger and his team are using work-flow technology to manage the data collected and are figuring out how to store data in the shared computing cloud, so they don't overwhelm any one facility and so scientists, students, and interested citizens everywhere can access them. The team is working out the data standards that will allow analysis programs to combine findings from different experiments into one larger analysis. That's called "interoperability," and it's crucial to making these scientific mashups work, because researchers will want to combine and compare data generated by predictive models in laboratories, as well as data from other sources, with data from the OOI network on the seafloor.

"This new era draws on the emergence, and convergence, of many rapidly evolving new technologies," Roger observes. The exploration will be focused on finding correlations across ocean events that will enhance our understanding of—and perhaps our ability to predict—land, ocean, and atmospheric interactions. Scientists will be able to measure such previously inaccessible underwater phenomena as erupting volcanoes, major migration patterns of sea life, earthquakes, and giant storms. Real-time video and new data visualization tools will allow students, educators, and the public at large to watch these events unfold and, in some cases, even conduct their own experiments. "The internet will emerge as the most powerful oceanographic tool on the planet," Roger predicts.

New video and data tools will allow everyday citizens to watch undersea events unfold and even conduct their own experiments.

OOI is unleashing the creativity of oceanographers worldwide, who are developing new kinds of instruments to plug into this undersea lab. One is a washing-machine-size DNA sequencer designed to operate unmanned and underwater. It will filter in local creatures, capture and sample their DNA, and then send the results to scientists on shore. That ability alone is impressive. Layer on the ability to merge the DNA information gathered with data about pollution levels, acidity, ocean temperatures, or the presence of migratory species that may affect the food chain—all of which are collected by other researchers—and we have the birth of a new era of oceanographic science.

Is there a business dimension to all of this? Well, for starters, imagine what might happen if a chemist at an energy company who was developing spill amelioration technology could consult a database on these organisms' DNA. He or she would be able to instantly call up genetic profiles of the microorganisms in the waters surrounding a spill and predict how they were likely to interact with the chemicals or solutions under consideration. Today's scientists grappling with the aftereffects of the massive deepwater oil spill in the Gulf of Mexico do not have comprehensive baseline measures of the ocean's health and are relying instead on "downstream" indicators, such as the health of fish. Other interoperability tools refined for OOI could offer more prosaic, but no less important, insights. For example, a retail marketing executive sitting at a desk might receive a daily report generated by a program that combs the data streaming in from point-of-sale terminals throughout the world in real time, flagging anomalous patterns of sales and returns, and make connections that most retailers would never think to look for.

Solutions for Disease and Droughts

One way the fourth paradigm achieves faster breakthroughs is by allowing the general population to interact with databases and contribute knowledge that will advance discoveries. In the Seattle traffic effort, for example, volunteers with GPS devices in their cars helped gather critical data about local traffic routes simply by driving them. These methods were later extended to the task of predicting flows on all streets in greater metropolitan areas and now enable traffic-sensitive routing for 72 cities in North America, available today in Bing Maps. (See the sidebar "Crowdsourcing in the Heavens" for a description of another effort that's taking place in astronomy.) Soon all sorts of citizen-scientists in different fields will likely use devices as simple as cell phones or laptops to

collect specialized information and analyze it.

Crowdsourcing in the Heavens

There's already one field in which citizen-scientists play a key role in guiding discovery: astronomy.

Most astronomy data today are gathered by charge-coupled devices, or CCDs—through robotic systems that collect far more information than the world's roughly 10,000 professional astronomers could ever evaluate in their lifetimes. However, there are at least one million amateur astronomers, who now have a way to get in on the action and make real contributions.

In 2007 a group of astronomers wrote a web-based application called Galaxy Zoo, which created a clever, gamelike user interface for a database of astronomical information collected by the Sloan Digital Sky Survey. It turns out that people can do certain kinds of galaxy classifications visually that computers are not yet very good at. So the project made it fun for the public to participate in the classifications, which also helped the astronomers test a theory that spiral galaxies tended to rotate clockwise. Galaxy Zoo was launched with a data set made up of a million galaxies imaged with a robotic telescope. Participants looked at the images and classified the galaxies as "right-handed" (meaning they rotated clockwise) or "left-handed" (rotating counterclockwise). With so many galaxies, the team thought that it might take at least two years for the site's visitors to work through them all. Within 24 hours of launch, however, the site was receiving 70,000 classifications an hour, and more than 50 million classifications were received by the project during its first year, from almost 150,000 people. The effort refuted the idea that most spiral galaxies were right-handed. It turns out that only half of them were. Even more amazing, a Dutch schoolteacher participating in the project found a strange galaxy that so baffled astronomers it ended up getting the attention of the Hubble telescope.

In 2008, Microsoft introduced the WorldWide Telescope and gave astronomers and the general public access to interactive 3-D images of the sky, planets, and galaxies. Visitors can view the images through a standard Explorer browser and visualize the same data that professional astronomers use. WWT incorporates the Galaxy Zoo classifications and allows every stargazer to call up the actual coordinates of remote galaxies, streaking comets, and spidery nebulae. Visualization tools such as WWT can actually transform scientists' ability to gain insights from data, sometimes with the help of ordinary citizens. —T.H.

My research team has a project in India, for instance, that allows nonmedical personnel in remote areas to diagnose certain illnesses with the help of cell phones. Using them, people dial into an enormous database of medical information, fill in answers to a set of questions, and receive valuable diagnoses on the spot. This system could someday be used to track and study the spread of diseases, particularly infectious ones. With large numbers of people performing quick diagnostics that feed into a database, public officials and health care workers can see where outbreaks are occurring, how fast they're moving, and what kind of symptoms are appearing. Machine learning can enter the loop in real time, constantly comparing every new case with every other case of this and other infectious outbreaks—and looking for patterns that might aid prevention efforts.

The stress this kind of ambitious project puts on every aspect of current technology—processing power; demand for parallel programmers; and data storage, curation, and publishing—is enormous. Unless curation of the data is actually built into a project's design, for example, the scientists involved usually try to figure it out ad hoc, which tends to lead to brittle, local solutions that don't scale up. Scientists and policy makers, however, do not have the luxury of waiting until everything

is figured out before taking action on urgent problems such as climate change or water shortages or planning for hurricanes or tsunamis.

Consider the plight of California, where the population is projected to increase from about 38 million today to more than 50 million by 2040. Says Jeff Dozier, a professor in the School of Environmental Science and Management at the University of California, Santa Barbara: “The availability of water drives California’s economy. Historically, we’ve tried to manage the supply of water to meet demand. We may not be able to do that anymore. Everyone would love a reliable, uniform supply, but that’s not what nature gives us. We will need much better technology to predict the amount of water we will have in a given year.”

Predicting water stores from snowpack is a much more difficult problem than it might appear, Dozier explains. Satellites collect huge volumes of data on snowpack, but they are still insufficient because they mainly reveal the snow’s surface characteristics. To manage runoff, we need to know the “water equivalent,” or the amount of water that would result from snowmelt. We can estimate the water equivalent from the weight of the snow, but that is difficult to measure across large stretches of variable terrain. The challenge: How do scientists combine data from satellites and surface measurements with information on economics and governance to better estimate, calibrate, and manage water supplies? In California alone, there are at least 400 different agencies that manage water. Microsoft is working with scientists at the University of California, Berkeley, and the Lawrence Berkeley National Laboratory to acquire and curate historic hydrologic data so that they can be used more effectively with data from new sensor networks to create better prediction models.

Through data analysis, scientists are zeroing in on a way to stop HIV in its tracks.

In another urgent arena, Microsoft’s David Heckerman, another MD with a PhD in computer science, is using data-intensive scientific discovery in the fight against the human immunodeficiency virus. “In several years in a single patient, HIV mutates about as much as the influenza virus has mutated in its known history,” he explains. That’s why developing a vaccine to thwart it has been so difficult. Moreover, the mutations seen in one individual are quite different from those seen in another, thanks to variability in human immune systems. David and his team are analyzing data about individual viral mutations in thousands of subjects, trying to zero in on the elements of the virus that are vulnerable to attack by the immune system. By creating a vaccine that can trigger a person’s own immune system to attack those elements, they hope to stop the virus in its tracks. He and his Harvard collaborator Bruce Walker expect to begin testing the first vaccine based on this work soon.

Shifting Gears—and Standards

Endeavors like vaccine development or fields like human genomics involve a limited number of disciplines but absolutely enormous amounts of data unique to each individual. In efforts to better characterize an environmental phenomenon like ocean processes or climate change, it’s not only the volume of data about any one factor but the number of disciplines and data sources that is daunting. Comprehensive calculations of warming trends might require factoring in measurements of radiant heat reflected from polar ice sheets, wasting of floating ice shelves caused by small increases in ocean temperature, the health of mangrove forests in tropical climates, global insect-hatching trends, climate changes captured in tree rings, CO_2 levels preserved in stored ice cores—and more. Creating standards for collecting, storing, and mashing together such data will become increasingly important as scientists deploy more and more sensors.

Critically, too, most of us believe scientific publishing will change dramatically in the future. We foresee the end product today—papers that discuss an experiment and its findings and just refer to

data sets—morphing into a wrapper for the data themselves, which other researchers will be able to access directly over the internet, probe with their own questions, or even mash into their own data sets in creative ways that yield insights the first researcher might never have dreamed of. The goal, as Jim Gray put it so well, is "a world in which all of the science literature is online, all of the science data are online and they interoperate with each other. Lots of new tools are needed to make this happen."

Beyond Microsoft: How Other Tech Companies Help Advance Science by: Daniel McGinn
Computer scientists are powering breakthroughs in health care, climate change, and other disciplines.

In early 2009 the Centers for Disease Control and Prevention began hearing reports of a severe flu outbreak in Mexico. For help investigating them, it turned to Google.org, the tech company's philanthropic arm. Since November 2008, Google.org has run a project called Flu Trends, which uses a sophisticated algorithm to track aggregated flu-related internet searches and estimate how the illness is spreading through a population. The Googlers didn't have a flu-tracking program for Mexico, but their engineers quickly built an experimental model. The results suggested this new illness—which became the swine flu pandemic—was widespread in Mexico but wasn't nearly as lethal as CDC researchers had originally feared. "In a pandemic, you want as many reliable sources of information as possible," says Jacquelline Fuller, a top executive at Google.org. "Flu Trends can help health officials draw a picture of what's actually happening on the ground." Google now operates Flu Trends in 28 countries.

It's an example of how companies besides Microsoft are applying data technology to scientific and social problems. At Google.org, the early work focused mostly on grant making. But now, Fuller says, "we've realized that what we can bring to these global challenges is our expertise in technology and our strength in providing information and data." Among the organization's projects: Earth Engine, which uses satellite imagery and analytics to track deforestation, a key cause of climate change. The group also uses technology to assist in crises; after the earthquakes in Haiti and Chile, for instance, Google worked with the U.S. State Department to create an online Person Finder database that relatives could use to search for or post information about missing loved ones.

While most of Google.org's work is altruistic, other companies see revenue opportunities in this space. In 2008 IBM launched its Smarter Planet initiative, which CEO Sam Palmisano has identified as a primary driver of IBM's future growth. Smarter Planet takes IBM's expertise in analytics and integration and applies it to problems like traffic management in Stockholm, water management in Malta, and health care in China. (In one application IBM created technology to help doctors in Guang Dong Hospital test the efficacy of traditional Chinese treatments.) "This is a nice intersection of some very tough societal issues that need the strength and expertise of a company like IBM, but has a nice commercial aspect to it," says Michael Valocchi, an IBM vice president for Global Business Services. And as more organizations become adept at mixing technology and science in new ways, the benefits will go far beyond profits.

While the realization of this goal would mean positive changes for society and the planet, the fourth paradigm also will inevitably create great business opportunities. For example, David Heckerman's genomic analysis of HIV is just one small piece of the much bigger agenda of personalized medicine. The pharmaceutical industry is betting that finding out which drugs are most effective for someone with a particular genetic profile will bring a whole new dimension to drug design.

Microsoft's Health Solutions Group is integrating medical records and images as a first step in providing a set of smart tools to help the pharmaceutical industry fulfill this vision.

All scientific disciplines, including computer science, need to collaborate to realize the power of the fourth paradigm and solve important problems for humanity. The answers are hiding amid vast mountains of numbers—and it's within our reach to find them.