The Machine That Would Predict the Future

If you dropped all the world's data into a black box, could it become a crystal ball that would let you see the future—even test what would happen if you chose A over B? One researcher thinks so, and he could soon get a billion euros to build it

By David Weinberger

In the summer and fall of last year, the Greek financial crisis tore at the seams of the global economy. Having run up a debt that it would never be able to repay, the country faced a number of potential outcomes, all unpleasant. Efforts to slash spending spurred riots in the streets of Athens, while threats of default rattled global financial markets. Many economists argued that Greece should leave the euro zone and devalue its currency, a move that would in theory help the economy grow. "Make no mistake: an orderly euro exit will be hard," wrote New York University economist Nouriel Roubini in the *Financial Times*. "But watching the slow disorderly implosion of the Greek economy and society will be much worse."

No one was sure exactly how the scenario would play out, though. Fear spread that if Greece were to abandon the euro, Spain and Italy might do the same, weakening the central bond of the European Union. Yet the *Economist* opined that the crisis would "bring more fiscal-policy control from Brussels, turning the euro zone into a more politically integrated club." From these consequences would come yet further-flung effects. Migrants heading into the European Union might shift their travel patterns into a newly affordable Greece. A drop in tourism could limit the spread of infectious disease. Altered trade routes could disrupt native ecosystems. The question itself is simple—Should Greece drop the euro?—but the potential fallout is so far-reaching and complex that even the world's sharpest minds found themselves unable to grasp all the permutations.

Questions such as this one are exactly what led Dirk Helbing, a physicist and the chair of sociology at the Swiss Federal Institute of Technology Zurich, to propose a €1-billion computing system that would effectively serve as the world's crystal ball.

Photograph by Dan Saelinger

SIGNS
POINT TO
YES

David Weinberger is a senior researcher at Harvard University's Berkman Center for Internet and Society and co-director of the Harvard Library Innovation Laboratory at Harvard Law School. His latest book is *Too Big to Know,* which is being published in January 2012.

Helbing's system would simulate not just one area of finance or policy or the environment. Rather it would simulate everything all at once—a world within the world—spitting out answers to the toughest questions policy makers face. The centerpiece of this project, the Living Earth Simulator, would attempt to model global-scale systems—economies, governments, cultural trends, epidemics, agriculture, technological developments, and more—using torrential data streams, sophisticated algorithms, and as much hardware as it takes. The European Commission was so moved by Helbing's pitch that it chose his project as the top-ranked of six finalists in a competition to receive €1 billion.

The system is the most ambitious expression of the rise of ”big data," a trend that is striking many scientists as being on a par with the invention of the telescope and microscope. The exponential growth of digitized information is bringing together computer science, social science and biology in ways that let us address questions we just otherwise could not have posed, says Nicholas Christakis, a social scientist and professor of medicine at Harvard University. As an example, he points to the ubiquity of mobile phones that create oceans of information about where individuals are going, what they are buying, and even traces of what they are thinking. Combine that with other kinds of data—genomics, economics, politics, and more—and many experts believe we are on the cusp of opening up new worlds of inquiry.

"Scientific advance is often driven by instrumentation," says David Lazer, an associate professor in the College of Computer and Information Science at Northeastern University and a supporter of Helbing's project. Tools attract the tasks, or as Lazer puts it: "Science is like the drunk looking for his keys under the lamppost because the light is better there." For Helbing's supporters, the ranks of which include dozens of respected scientists all over the world, €1 billion can buy a pretty bright light.

Many scientists are not convinced of the need to gather the world's data in a centralized collection, though. Better, they argue, to form data clouds on the Internet, connected by links to make them useful to all. A shared data format will give more people the opportunity to poke around through the data, find hidden connections and create a marketplace of competing ideas.

NEXT TOP MODEL

FINDING CORRELATIONS in sets of data is nothing out of the ordinary for modern science, even if those sets are now gigantic and the correlations span astronomical distances. For example, researchers have amassed so much anonymized data about human behavior that they have begun to unravel the complex behavioral and environmental factors that trigger "diseases of behavior" such as type 2 diabetes, says Alex Pentland, director of the Massachusetts Institute of Technology's Human Dynamics Laboratory. He says that mining big data this way makes the seminal Framingham study of cardiovascular disease—which, starting in 1948 followed 5,209 people—look like a focus group study.

Yet Helbing's FuturICT Knowledge Accelerator and Crisis-Relief System, as it's formally known, goes beyond data mining. It will include global Crisis Observatories that will search for nascent problems such as food shortages or emerging epidemics, as well as a Planetary Nervous System that aggregates data from sensor systems spread around the globe. But the heart of the FuturICT project is the Living Earth Simulator, an effort to model the myriad social, biological, political and physical forces at work in the world and use them to gain insight into the future.

Models have been with us for generations. In 1949 Bill Phillips, an engineer and economist from New Zealand, unveiled a model of how the U.K. economy worked that he had constructed out of plumbing supplies and a cannibalized windshield-wiper motor. Colored water simulated the flow of income based on "what if" adjustments in consumer spending, taxes and other economic activities. Although it is of course primitive by today's standards, it expresses the basics of modeling: stipulate a set of relations among factors, feed in data, watch the outcome. If the predictions are off, that itself becomes valuable information that can be used to refine the model.

Our society could no more function without models than without computers. But can you add enough pipes and pumps to model not only, say, the effect of volcano eruptions on short-term economic growth but also the effect of that change on all the realms of human behavior it touches, from education to the distribution of vaccines? Helbing thinks so. His confidence comes in part from his success modeling another complex system: highway traffic. By simulating the flow of vehicles on a computer, he and his colleagues came up with a model that showed (again, on a computer) that you could end stop-and-go delays by reducing the distance between moving vehicles. (Unfortunately, the distance is so small that it would require cars driven by robots.) Likewise, Helbing describes a project he consulted on that modeled the movement of pedestrians during the hajj in Mecca, resulting in a billion dollars of street and bridge rejiggering to prevent deaths from trampling. Helbing sees his FuturICT system as, in essence, a scaled-up elaboration of these traffic models.

Yet this type of agent-based modeling works only in a very narrow set of circumstances, according to Gary King, director of the Institute for Quantitative Social Science at Harvard. In the case of a highway or the hajj, everyone is heading in the same direction, with a shared desire to get where they are going as quickly and safely as possible. Helbing's FuturICT system, in contrast, aims to model systems in which people are acting for the widest

IN BRIEF

Researchers plan to build a computing system that would model the entire world to predict the future.

The project would be powered by the enormous data streams now available to researchers.

Yet models are not perfect; many researchers think they will never be able to capture the world's complexities.

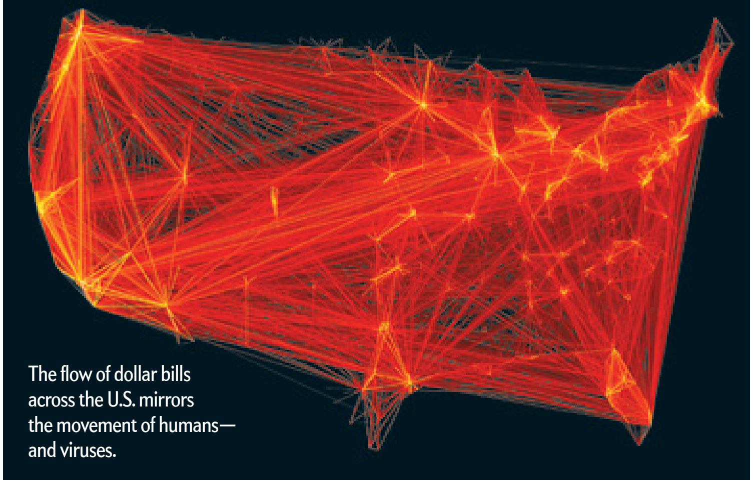
A better knowledge machine may arise out of Web-like principles such as interconnection and argument.

PROP STYLING BY LAURIE RAAB, FOR HALLEY RESOURCES (*preceding pages*)

PUBLIC HEALTH APPLICATIONS

Disease Follows the Money

Imagine a novel in which a deadly flu virus emerges. Where will it spread? Physicists and epidemiologists have begun to tap enormous data streams to make predictions about how a pandemic might play out—and what can be done to stop it. Scientists took data from the Where's George project, which tracks the location of millions of dollar bills as they move across the U.S., to model how 2009's H1N1 flu virus would likely spread. Other researchers used air and land traffic patterns in the same way. The studies demonstrated both the promise and problems of big data: they accurately predicted *where* the flu would spread, but they severely undercounted the number of people who would end up infected.



The flow of dollar bills across the U.S. mirrors the movement of humans—and viruses.

FROM "THE STRUCTURE OF BORDERS IN A SMALL WORLD," BY C. THIEMANN, F. THEIS, D. GRADY, R. BRUNE AND D. BROCKMANN, IN *PLOS ONE*, VOL. 5, NO. 11; 2010

variety of reasons (from selfish to altruistic); where their incentives may vary widely (getting rich, getting married, staying out of the papers); where contingencies may erupt (the death of a world leader, the arrival of UFOs); where there are complex feedback loops (an expert's financial model brings her to bet against an industry, which then panics the market); and where there are inputs, outputs and feedback loops from related models. The economic model of a city, for example, depends on models of traffic patterns, agricultural yields, demographics, climate and epidemiology, to name a few.

Beyond the problem of sheer complexity, scientists raise a number of interrelated challenges that such a comprehensive system would have to overcome. To begin with, we don't have a good theory of social behavior from which to start. King explains that when we have a solid idea of how things work—in physical systems, for example—we can build a model that successfully predicts outcomes. But whatever theories of social behavior we do have fall far short of the laws of physics in predictive power.

Nevertheless, King points to another possibility: if we have enough data, we can build models based on some hints about what creates regularities, even if we don't know what the laws are. For example, if we were to record the temperature and humidity at each point over the globe for a year, we could develop fairly accurate weather forecasts without any understanding of fluid dynamics or solar radiation.

We have already begun to use data to tease out some of these regularities in human systems, says Albert-László Barabási, director of the Center for Complex Network Research at Northeastern University and an adviser to the project. For example, Barabási and his colleagues recently unveiled a model that predicts with 90 percent accuracy where people will be at 5 P.M. tomorrow based purely on their past travel patterns. This knowledge does not assume anything about psychology, or technology, or the economy. It just looks at past data and extrapolates from there.

Yet sometimes the volume of data needed to make these approaches work dwarfs our capabilities. To get the same accuracy in a problem that requires you to consider 100 different interacting factors as you would in a two-dimensional problem, the number of data points required goes up into the number-of-stars-in-the-universe range, according to Cosma Shalizi, a statistician at Carnegie Mellon University. He concludes that unless you resign yourself to using simple models that fail to capture the full complexity of social behavior, "getting good models from data alone is hopeless."

FuturICT will not just rely on one model, however complex. Helbing says it will combine "computer science, complexity science, systems theory, social sciences (including economics and political sciences), cognitive science" and other fields. Yet combining models also creates problems of exploding complexity. "Let's say weather and traffic each have 10 outcomes," King says. "And now you want to know about both. So how many things do we need to know? It's not 20, it's 100. That doesn't make it hopeless. It just means the data requirements go up very quickly."

To further add to the challenge, news of a model's conclusions can alter the situation it is modeling. "This is the big scientific question," says Alessandro Vespignani, director of the Center for Complex Networks and Systems Research at Indiana University and the project's lead data planner. "How can we develop models that include feedback loops or real-time data monitors that let us continuously update our algorithms and get new predictions" even as the predictions affect their own conditions?

The models also have to be incredibly intricate and particular. For example, if you ask an economic model if your city should reclaim some land and if the model does not take account how that decision affects the food chain, it can generate a result that might be good economics but disastrous for the environment. With 10 million species, simply learning which one eats what is a daunting task. Further, relevant variances in food do not stop at the species level. Jesse Ausubel, an environmental scientist at the Rockefeller University, points out that by analyzing the DNA of the contents of the stomachs of bats, we can know for sure exactly what bats eat. But the food source of bats in a specific cave might be different from the food source of bats of the same species a few miles away. Without crawling through the guano-coated particularities cave by cave, experts relying on interrelated models may encounter unreliable and cascading effects.

So while in theory we might be able to construct models of complicated phenomena even when we do not have any underlying laws on which to build them, the practical difficulties quickly turn exponential. There is always another layer of detail, always another factor that may prove critical in the final accounting; without a prior understanding of how humans operate, we cannot know when our accounting is final.

It is not at all clear that human brains will be capable of understanding why the supercomputers have come up with the answers they have.

Big data have given rise to many successes in genomics and astrophysics, but success in one field may not be evidence that we can succeed when fields are interdependent in highly complex ways. Perhaps we can make stepwise progress. Or there may be a natural limit to the power of models for systems as complex as those that involve human activity. Human systems, after all, are subject to the two hallmarks of unpredictability: black swans and chaos theory.

KNOWLEDGE WITHOUT UNDERSTANDING

ON DECEMBER 17, 2010, Mohamed Bouazizi, a street vendor in the small Tunisian town of Sidi Bouzid, set himself on fire in a protest against the local culture of corruption. That singular act set into motion a popular revolution that burned across the Arab world, leading to uprisings that overthrew decades of dictatorial rule in Egypt, Libya and beyond, upending forever the balance of power in the world's most oil-rich region.

What model would have been able to foresee this? Or the attacks of September 11, 2001, and the extent of their effects? Or that the Internet would go from an obscure network for researchers to a maker and breaker of entire industries? This is the black swan problem popularized by Nassim Nicholas Taleb in his 2007 best seller of the same name. "The world is always more complex than models," Ausubel says. "It's always something."

What's worse, the social, political and economic systems that Helbing wants to understand are not merely complex. They are chaotic. Each depends on hundreds of unique factors, all intricately interrelated and profoundly affected by the state from which they started. Everything happens for a reason in a chaotic system, or, more exactly, everything happens for so many reasons that events are unpredictable except in the broadest of strokes. For example, Jagadish Shukla, a climatologist at George Mason University and president of the Institute of Global Environment and Society, told me that while we can now forecast the weather five days ahead, "we may not be able to get beyond day 15. [No] matter how many sensors you put in place, there will still be some errors in the initial conditions, and the models we use are not perfect." He adds, "The limitations are not technological. They are the predictability of the system."

Shukla is careful to distinguish weather from climate. We may not be able to predict whether it will rain in the afternoon exactly 100 years from now, but we can with some degree of reliability predict what the average ocean temperature will be. "Even though climate is a chaotic system, it still does have predictability," Shukla says. And so it would be for Helbing's models.

"Detailed financial-market moves are probably much less predictable than weather," Helbing wrote in an e-mail, "but the fact that a financial meltdown would happen sooner or later could be derived from certain macroeconomic data (for example, that consumption in the U.S. grew bigger than incomes over many years)." But we don't need a set of supercomputers, galaxies of data and €1 billion to know that.

If the aim is to provide scientifically based advice to policy makers, as Helbing emphasizes when justifying the expense, some practical issues arise. For one thing, it is not at all clear that human brains will be capable of understanding why the supercomputers have come up with the answers that they have. When the model is simple enough—say, a hydraulic model of the British economy—we can backtrack through a model run and realize that the drawdown of personal savings accounts was an unexpected effect of raising taxes too quickly. But sophisticated models derived computationally from big data—and consequently tuned by feeding results back in—might produce reliable results from processes too complex for the human brain. We would have knowledge but no understanding.

When I asked Helbing about this limitation, he paused before saying he thought it likely that human-understandable general principles and equations would probably emerge because they did when he studied traffic. Still, the intersection of financial systems, social behaviors, political movements, meteorology and geology is orders of magnitude more complex than three lanes of traffic moving in the same direction. So humans may not be able to understand why the model predicts disaster if Greece goes off the euro.

Without understanding why a particular course of action is the best one, a president or prime minister would never be able to act on it—especially if the action seems ridiculous. Victoria Stodden, a statistician at Columbia University, imagines a policy maker who reads results from the Living Earth Simulator and announces, "To pull the world out of our economic crisis, we must set fire to all the world's oil wells." That will not be actionable advice if the policy maker cannot explain why it is right. After all, even with scientists virtually universally aligned about the danger of climate change, policy makers refuse to prepare for the future predicted by every serious environmental model.

NERDS ARGUING WITH NERDS

THESE AND OTHER practical problems arise because FuturICT as Helbing currently describes it assumes that such a large, complex effort requires a central organization to take charge. Helbing would oversee a global project that would assemble the hardware, collect data and return results.

It's not what John Wilbanks, vice president of science at Creative Commons, would do. Wilbanks shares Helbing's enthusiasm for big data. But his instincts hew to the Internet, not the institution. He is a leading figure in an ongoing project to organize various "data commons" that anyone can make use of. The aim is to let the world's scientists engage in an open market of ideas, models and results. It is the opposite approach to planning out a formalized institution with organized inputs and high-value outputs.

The two approaches focus on different values. A data commons might not have the benefits of up-front, perfect curation that a closed system has, but Wilbanks believes it more than makes up for that in "generativity," a term from Jonathan Zittrain's 2008 *The Future of the Internet:* "a system's capacity to produce unanticipated change through unfiltered contributions from broad and varied audiences." The Web, for example, allows everyone to participate, which is why it is such a powerful creative engine. In Wilbanks's view, science will advance most rapidly if scientists have access to as much data as possible, if that information is open to all, is easy to work with, and can be pulled together across disciplines, institutions and models.

Over the past few years a new "language" for data has emerged that makes Wilbanks's dream far more plausible. It grows out of principles enunciated in 2006 by Tim Berners-Lee, inventor of the World Wide Web. In this "linked data" format, information comes in the form of simple assertions: *X* is related to *Y* in some specified way; the relation can be whatever the person releasing the data wants. For example, if Creative Commons wanted to release its staffing information as linked data, it would make it available in a series of "triples": [John Wilbanks] [leads] [Science at Creative Commons], [John Wilbanks][has an e-mail address of] [johnsemail@creativecommons.org], and so forth.

Further, because many John Wilbanks live in the world and because "leads" has many meanings, each element of these triples would include a Web link that points at an authoritative or clarifying source. For example, the "John Wilbanks" link might point to his home page, to the page about him at CreativeCommons.org or to his Wikipedia entry. "Leads" might point to a standard vocabulary that defines the type of leadership he provides.

This linked structure enables researchers to connect data from multiple sources without having first to agree on a single abstract model that explains the relations among all the pieces. This lowers the cost of preparing the data for release. It also increases the value of the data after they have been released.

A linked-data approach increases the number of eyeballs that could in theory pay attention to any particular data set, thus increasing the likelihood that someone will stumble across an interesting signal. More hypotheses will be tested, more models tried. "Your nerds and my nerds need to have arguments," Wilbanks says. "They need to argue about whether the variables and the math in the models are right and whether the assumptions are right." The world is so chaotic that our best chance to make sense of it—to catch a financial meltdown in time—is to get as many nerds poking at it as we can. For Wilbanks and his tribe, making the data open and interoperable is the first step—the transformative step. Among the groups entering the fray certainly will be institutions that have assembled great minds and built sophisticated models. But the first and primary condition for the emergence of the truth is the fray itself. Nerds arguing with nerds.

Wilbanks and Helbing both see big data as transformative, and both are hopeful that far more social behavior can be understood scientifically than we thought just a few years ago. When Helbing is not trying to persuade patrons by painting a picture of how the Living Earth Simulator will avert national bankruptcies and global pandemics—as Barabási observes, "If you want to convince politicians, you have to talk about the outcomes"—he acknowledges that FuturICT will support multiple models that compete with one another. Further, he is keen on gathering the biggest collection of big data in history and making almost all of it public. (Some will have to stay private because it comes under license from commercial providers or because it contains personal information.)

Nevertheless, the differences are real. Helbing and his data architect, Vespignani, do not stop with the acknowledgment that the FuturICT institution will support multiple models. "Even weather forecasts are made with multiple models," Vespignani says. Then he adds, "You combine them and get a statistical inference of what the probabilistic outcome will be." For Helbing and him, the value is in this convergence toward a single answer.

The commons view also aims at convergence toward truth, of course. But as a networked infrastructure, it acknowledges and even facilitates fruitful disagreement. Scientists can have different models, different taxonomies, different nomenclatures, but they can still talk with one another because they can follow their shared data's links back to some known anchor on the Internet or in the real world. They can, that is, operate on their own and yet still communicate and even collaborate. The differences won't resolve into a single way of talking about the world because—Wilbanks argues—there may be differences of culture, starting point, even temperament. The data-commons approach recognizes, acknowledges and even embraces the persistence of difference.

WHAT KNOWLEDGE IS

THE OBVIOUS QUESTION is the practical one: Which approach is going to work better, where "working better" means advancing the state of the science and producing meaningful (and accurate) answers to hard questions about the future?

The answer may come down to a disagreement about the nature of knowledge itself. We have for a couple of millennia in the West thought of knowledge as a system of settled, consistent truths. Perhaps that exhibits the limitations of knowledge's medium more than of knowledge itself: when knowledge is communicated and preserved by writing it in permanent ink on paper, it becomes that which makes it through institutional filters and that which does not change. Yet knowledge's new medium is not a publishing system so much as a networked public. We may get lots of knowledge out of our data commons, but the knowledge is more likely to be a continuous argument as it is tugged this way and that. Indeed, that is the face of knowledge in the age of the Net: never fully settled, never fully written, never entirely done.

The FuturICT platform hopes to build a representation of the world sufficiently complete that we can ask it questions and rely on its answers. Linked data, on the other hand, arose (in part) in contrast to the idea that we can definitively represent the world in logical models of all the many domains of life. Knowledge may come out of the commons, even if that commons is not itself a perfect representation of the world.

Unless, of course, the messy contention of ideas—nerds arguing with nerds—is a more fully true representation of the world.

MORE TO EXPLORE

The Semantic Web. Tim Berners-Lee, James Hendler and Ora Lassila in *Scientific American,* Vol. 284, No. 5, pages 34–43; May 2001.

Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room. David Weinberger. Basic Books (in press). **www.hyperorg.com/blogger**

The FuturICT Project: **www.futurict.eu**

SCIENTIFIC AMERICAN ONLINE
What big data means for knowledge: **ScientificAmerican.com/dec2011/big-data**