Science | DOI:10.1145/1839676.1839682 Gregory Goth

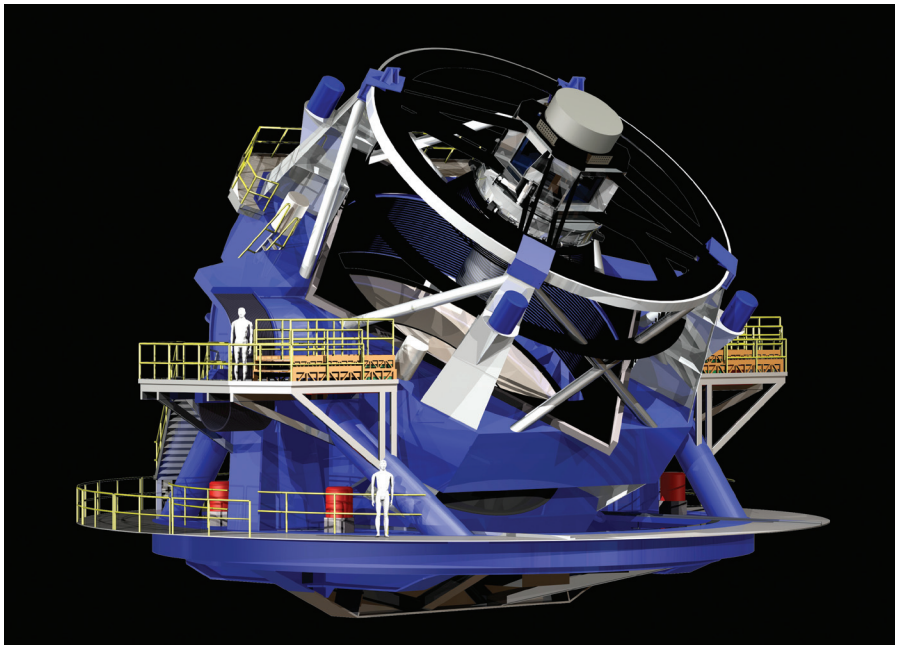
Turning Data Into Knowledge

Today's data deluge is leading to new approaches to visualize, analyze, and catalog enormous datasets.

THE AMOUNT OF data available to scientists of nearly every discipline has almost become a "Can you top this?" exercise in numbers.

The Sloan Digital Sky Survey (SDSS), for example, is often cited as a prime example. Since the survey's 2.5-meter telescope first went online in 1998, more than 2,000 refereed publications have been produced, but they use just 10% of the survey's available imaging data, according to a recent U.S. National Science Foundation workshop on data-enabled science in the mathematical and physical sciences. Once the next-generation, state-of-the-art Large Synoptic Survey Telescope (LSST) goes online in 2016, however, it is estimated to be capable of producing a SDSS-equivalent dataset every night for the next 10 years. Another often-cited example is the Large Hadron Collider. It will generate two SDSS's worth of data each day.

On the surface, then, the scientific community's mandate seems clear: create better computational tools to visualize, analyze, and catalog these enormous datasets. And to some extent, there is wide agreement these tasks must be pursued.



The Large Synoptic Survey Telescope will have the ability to survey the entire sky in only three nights.

IMAGE COURTESY OF LSST CORPORATION

Some leading computational research scientists believe, however, that progress in utilizing the vast expansion of data will best be attacked on a project-by-project basis rather than by a pan-disciplinary computational blueprint.

"In theory, you might think we should all be working together, and the reality might be that each of the people working on their own discipline are achieving the results they need to scientifically," says Dan Masys, M.D., chairman of biomedical informatics at Vanderbilt University. "There's a cost of communication that reaches an irreducible minimum when you work

across disciplinary boundaries, and sometimes it's worth it.

"But the grander potential for synergy that's often spoken of at the level of federal funding agencies probably doesn't happen as much as people think would be best for science," Masys continues. "You can't push that rope all that well, because it depends on the art of the possible with respect to technologies and the vision of the scientists doing the work."

Tom Mitchell, chairman of the machine learning department at Carnegie Mellon University (CMU), concurs with Masys' assessment. "I think it starts from the bottom up and at some point you'll see commonalities across domains," he says. As an example, he cites time series algorithms being developed by CMU colleague Eric Xing that may also be useful for brain imaging work Mitchell is undertaking.

"There's an example I think is probably pretty representative of how it's going to go," Mitchell says. "People encounter problems and have to design algorithms to address them, but time series analysis is a pretty generic problem. So I think bottom up it will grow and then they will start connecting across [different disciplines]."

Vanderbilt's Masys is about to begin a collaboration with computational biologists from Oak Ridge National Laboratory. Masys says the Oak Ridge scientists' optimization of Vanderbilt's fundamental algorithms and the lab's teraflop-capable architecture will likely speed processing of problems involving multiplying "several million genome data points by several thousand people" from five days to three hours—a prime example of focused intradisciplinary collaboration and leading-edge hardware.

New Perspectives on Data

Both Mitchell and Randal Bryant, dean of the school of computer science at CMU, cite the influence of commercial companies for helping to expand the concept of what kind of data, and what kind of data storage and computational architectures, can produce useful scientific results.

"The commercial world, Google and its peers, have been the drivers on the data side, much more than the traditional sciences or universities," says Bryant, who cites the example of a Google cluster running a billion-word index that outperformed the Big Iron architecture of the "usual suspects" in a 2005 language-translation contest sponsored by the U.S. National Institute of Standards and Technology.

The availability of such large datasets can lead to serendipitous discoveries such as one made by Mitchell and his colleagues, using a trillion-word index Google had originally provided for machine translation projects. "We found we could build a computational model that predicts the neural activity that will show up in your brain when you think about an arbitrary noun," Mitchell says. "It starts by using a trillion-word collection of text provided to us by Google, and looks up the statistical properties of that word in the text; that is, if you give it the word 'telephone', it will look up how often 'telephone' occurs with words from a long list of verbs—for example, how often does it occur with 'hug', or 'eat', and so on.

"To Google's credit they put this out on the Web for anybody to use, but they were thinking it would be used by researchers working on translation—and it turned out to be useful for something else."

Meanwhile, the LSST project is planning multiple vectors by which its huge dataset—all of which will be publicly available in near-real time—will aid research by professional astronomers; programs at museums, secondary schools, and other institutions; and citizen scientists. The project's goal, say the organizers, is "open source, open data."

"We will develop methods for engaging the public so anyone with a Web browser can effectively explore aspects of the LSST sky that interest and impact the public," according to the LSST organizers. "We will work with the IT industry on enhanced visualization involving dynamic graphics overlays from metadata and provide tools for public query of the LSST database."

The LSST organization's hope, then, is that the distributed nature of allowing any researcher at any level to access the data will result in a plethora of projects—a kind of "given enough eye-

Ubiquitous Computing

Intel's Friendly, Smart Machines

Context-aware computing, in which devices understand what a user is doing and anticipate his or her needs without being asked, are the next step in the evolution of smart machines, says Justin Rattner, Intel vice president and chief technology officer.

In his keynote address at IDF2010, the recent Intel Developer Forum in San Francisco, Rattner laid out a vision in which computers use a variety of sensors—microphones, accelerometers, and global positioning systems (GPSs)—combined with "soft sensors" such as calendars and social networks, to track people's activity and figure out how the devices can help. For instance, a device might locate someone at her office, hear the sound of human voices, crosscheck her calendar, and conclude she's in a business meeting, then suggest to the husband trying to call her that this wouldn't be a good time to interrupt.

A television remote control, using unsupervised learning in which it continuously collects data and makes inferences about what's going on around it, could learn to recognize which person is holding it—based on how the user moves, what angle he holds it at, and how fast he presses the buttons—then make personalized recommendations for shows, based on past preferences. A prototype Personal Vacation Assistant, developed with Fodor's Travel, uses GPS location, time of day, and past behavior to recommend restaurants and tourist sites. Data, collected over time and shared among devices, is run through an inference algorithm that examines the input and generates confidence scores to determine what is likely going on.

Collecting this data will require giving users control over what gets shared, and allow them to turn off sensors, Rattner says. He gives no timeline for introducing such applications, but says, "We believe that context-aware computing is poised to fundamentally change the way we interact and relate to the devices that we use today."

—Neil Savage

balls" approach to massive datasets.

However, even massive datasets are sometimes not complete enough to deliver definitive results. Recent discoveries in biomedical research have revealed that even a complete index of the human genome's three billion pairs of chemical bases has not greatly accelerated breakthroughs in health care, because other crucial medical data is missing. A study of 19,000 women, led by researchers at Brigham and Women's Hospital in Boston, used data constructed from the National Human Genome Research Institute's catalog of genome-wide association study results published between 2005 and June 2009—only to find that the single biggest predictor of heart disease among the study's cohort is self-reported family history. Correlating such personal data with genetic indexes on a wide demographic scale today is nearly impossible as an estimated 80% of U.S.-based primary-care physicians do not record patient data in electronic medical records (EMRs). Recent government financial incentives are meant to spur EMR adoption, but for the immediate future, crucial data in biomedical research will not exist in digital form.

Another issue in biomedical research is the reluctance of traditionally trained scientists to accept datasets that were not created under the strict parameters required by, for example, epidemiologists and pharmaceutical companies.

CMU's Mitchell says this arena of public health research could be in the vanguard of what may be the true crux of the new data flood—the idea that the provenance of a given dataset should matter less than the provenance of a given hypothesis.

"The right question is, Do I have a scientific question and a method for answering it that is scientific, no matter what the dataset is?" Mitchell asks. Increasingly, he says, computational scientists will need to frame their questions and provide data for an audience that extends far beyond their traditional peers.

"We're at the beginning of the curve of a decades-long trend of increasingly evidence-based decision-making across society, that's been noticed by people in all walks of life," he says. "For example, the people at the public policy school at CMU came to the machine learning department and said,

"The right question is, Do I have a scientific question and a method for answering it that is scientific, no matter what the dataset is?" asks Tom Mitchell.

'We want to start a joint Ph.D. program in public policy and machine learning, because we think the future of policy analysis will be increasingly evidence-based. And we want to train people who understand the algorithms for analyzing and collecting that evidence as well as they understand the policy side.'" As a result, the joint Ph.D. program was created at CMU. C

Further Reading

Duda, S.N. Cushman, C., and Masys, D.R.
An XML model of an enhanced dictionary to facilitate the exchange of pre-existing clinical research data in international studies, *Proceedings of the 12th World Congress on Health Informatics*, Brisbane, Australia, August 20–24, 2007.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., and Just, M.A.
Predicting human brain activity associated with the meanings of nouns, *Science 320*, 5880, May 30, 2008.

Murray-Rust, P.
Data-driven science—a scientist's view, NSF/JISC Repositories Workshop position paper, April 10, 2007.

Newman, H.B.
Data intensive grids and networks for high energy and nuclear physics: drivers of the formation of an information society, World Summit on the Information Society, Pan-European Ministerial Meeting, Bucharest, Romania, November 7–9, 2002.

Thakar, A.R.
The Sloan Digital Sky Survey: drinking from the fire hose, *Computing in Science and Engineering 10*, 1, Jan./Feb. 2008.

Gregory Goth is an Oakville, CT-based writer who specializes in science and technology.

Research & Development

Paper Chase

Due to enormous governmental investments in research and development, scientists in many Asian countries are steadily increasing their number of papers published in scientific journals.

The Asia-Pacific region increased its total of published science articles from 13% in the early 1980s to slightly more than 30% in 2009, according to the Thomson Reuters National Science Indicators, an annual database of the number of articles published in about 12,000 internationally recognized journals. China leads the pack with 11% in 2009, up from 0.4% in the early 1980s, followed by Japan with 6.7% and India with 3.4%. In contrast, the ratio of articles from scientists in the U.S. decreased to 28% in 2009, down from 40% in the early 1980s.

In all, 25 nations have increased their research, but none more so than Singapore. With a population of just five million, the nation published 8,500 articles in 2009, compared with only 200 in 1981. Singapore now allocates 3% of its gross domestic product to research and development, a figure expected to rise to 3.5% by 2015.

The increase in scientific publications, especially in East Asian countries, reflects a "phenomenal" increase in funding, Simon Marginson, a professor of higher education at the University of Melbourne, told *The New York Times*. Marginson attributed the increase in research output to governments' commitment to establishing knowledge-intensive economies. "It's very much not simply about knowledge itself—it's about its usefulness throughout the economy. I think that that economic vision is really the principal driver," Marginson said.

Another reason for increased publications is that many Asian universities now receive additional funding to have their papers translated into English, the language used by the majority of academic journals.

—Phil Scott