The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

The Fourth Paradigm

The Fourth Paradigm

Data-Intensive Scientific Discovery

EDITED BY
**TONY HEY, STEWART TANSLEY,
AND KRISTIN TOLLE**

MICROSOFT RESEARCH
REDMOND, WASHINGTON

Second printing, version 1.1, October 2009.

ISBN 978-0-9825442-0-4

Printed in the United States of America.

The information, findings, views, and opinions contained in this publication are those of the authors and do not necessarily reflect the views of Microsoft Corporation or Microsoft Research. Microsoft Corporation does not guarantee the accuracy of any information provided herein.

Microsoft Research
http://research.microsoft.com

For Jim

CONTENTS

3. SCIENTIFIC INFRASTRUCTURE

4. SCHOLARLY COMMUNICATION

FINAL THOUGHTS

Foreword

GORDON BELL | Microsoft Research

This book is about a new, fourth paradigm for science based on data-intensive computing. In such scientific research, we are at a stage of development that is analogous to when the printing press was invented. Printing took a thousand years to develop and evolve into the many forms it takes today. Using computers to gain understanding from data created and stored in our electronic data stores will likely take decades—or less. The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives.

In many instances, science is lagging behind the commercial world in the ability to infer meaning from data and take action based on that meaning. However, commerce is comparatively simple: things that can be described by a few numbers or a name are manufactured and then bought and sold. Scientific disciplines cannot easily be encapsulated in a few understandable numbers and names, and most scientific data does not have a high enough economic value to fuel more rapid development of scientific discovery.

It was Tycho Brahe's assistant Johannes Kepler who took Brahe's catalog of systematic astronomical observations and discovered the laws of planetary motion. This established the division between the mining and analysis of captured and carefully archived experimental data and the creation of theories. This division is one aspect of the Fourth Paradigm.

In the 20th century, the data on which scientific theories were based was often buried in individual scientific notebooks or, for some aspects of "big science," stored on magnetic media that eventually become unreadable. Such data, especially from individuals or small labs, is largely inaccessible. It is likely to be thrown out when a scientist retires, or at best it will be held in an institutional library until it is discarded. Long-term data provenance as well as community access to distributed data are just some of the challenges.

Fortunately, some "data places," such as the National Center for Atmospheric Research[1] (NCAR), have been willing to host Earth scientists who conduct experiments by analyzing the curated data collected from measurements and computational models. Thus, at one institution we have the capture, curation, and analysis chain for a whole discipline.

In the 21st century, much of the vast volume of scientific data captured by new instruments on a 24/7 basis, along with information generated in the artificial worlds of computer models, is likely to reside forever in a live, substantially publicly accessible, curated state for the purposes of continued analysis. This analysis will result in the development of many new theories! I believe that we will soon see a time when data will live forever as archival media—just like paper-based storage—and be publicly accessible in the "cloud" to humans and machines. Only recently have we dared to consider such permanence for data, in the same way we think of "stuff" held in our national libraries and museums! Such permanence still seems far-fetched until you realize that capturing data provenance, including individual researchers' records and sometimes everything about the researchers themselves, is what libraries insist on and have always tried to do. The "cloud" of magnetic polarizations encoding data and documents in the digital library will become the modern equivalent of the miles of library shelves holding paper and embedded ink particles.

In 2005, the National Science Board of the National Science Foundation published "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," which began a dialogue about the importance of data preservation and introduced the issue of the care and feeding of an emerging group they identified as "data scientists":

> The interests of data scientists—the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection—lie in having their creativity and intellectual contributions fully recognized." [1]

[1] www.ncar.ucar.edu

THE FOURTH PARADIGM: A FOCUS ON DATA-INTENSIVE SYSTEMS AND SCIENTIFIC COMMUNICATION

In Jim Gray's last talk to the Computer Science and Telecommunications Board on January 11, 2007 [2], he described his vision of the fourth paradigm of scientific research. He outlined a two-part plea for the funding of tools for data capture, curation, and analysis, and for a communication and publication infrastructure. He argued for the establishment of modern stores for data and documents that are on par with traditional libraries. The edited version of Jim's talk that appears in this book, which was produced from the transcript and Jim's slides, sets the scene for the articles that follow.

Data-intensive science consists of three basic activities: capture, curation, and analysis. Data comes in all scales and shapes, covering large international experiments; cross-laboratory, single-laboratory, and individual observations; and *potentially individuals' lives.*[2] The discipline and scale of individual experiments and especially their data rates make the issue of tools a formidable problem. The Australian Square Kilometre Array of radio telescopes project,[3] CERN's Large Hadron Collider,[4] and astronomy's Pan-STARRS[5] array of celestial telescopes are capable of generating several petabytes (PB) of data per day, but present plans limit them to more manageable data collection rates. Gene sequencing machines are currently more modest in their output due to the expense, so only certain coding regions of the genome are sequenced (25 KB for a few hundred thousand base pairs) for each individual. But this situation is temporary at best, until the US$10 million X PRIZE for Genomics[6] is won—100 people fully sequenced, in 10 days, for under US$10,000 each, at 3 billion base pairs for each human genome.

Funding is needed to create a generic set of tools that covers the full range of activities—from capture and data validation through curation, analysis, and ultimately permanent archiving. Curation covers a wide range of activities, starting with finding the right data structures to map into various stores. It includes the schema and the necessary metadata for longevity and for integration across instruments, experiments, and laboratories. Without such explicit schema and metadata, the interpretation is only implicit and depends strongly on the particular programs used to analyze it. Ultimately, such uncurated data is guaranteed to be lost. We

[2] http://research.microsoft.com/en-us/projects/mylifebits
[3] www.ska.gov.au
[4] http://public.web.cern.ch/public/en/LHC/LHC-en.html
[5] http://pan-starrs.ifa.hawaii.edu/public
[6] http://genomics.xprize.org

must think carefully about which data should be able to live forever and what additional metadata should be captured to make this feasible.

Data analysis covers a whole range of activities throughout the workflow pipeline, including the use of databases (versus a collection of flat files that a database can access), analysis and modeling, and then data visualization. Jim Gray's recipe for designing a database for a given discipline is that it must be able to answer the key 20 questions that the scientist wants to ask of it. Much of science now uses databases only to hold various aspects of the data rather than as the location of the data itself. This is because the time needed to scan all the data makes analysis infeasible. A decade ago, rereading the data was just barely feasible. In 2010, disks are 1,000 times larger, yet disc record access time has improved by only a factor of two.

DIGITAL LIBRARIES FOR DATA AND DOCUMENTS: JUST LIKE MODERN DOCUMENT LIBRARIES

Scientific communication, including peer review, is also undergoing fundamental changes. Public digital libraries are taking over the role of holding publications from conventional libraries—because of the expense, the need for timeliness, and the need to keep experimental data and documents about the data together.

At the time of writing, digital data libraries are still in a formative stage, with various sizes, shapes, and charters. Of course, NCAR is one of the oldest sites for the modeling, collection, and curation of Earth science data. The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, which is normally associated with supplying computational power to the scientific community, was one of the earliest organizations to recognize the need to add data to its mission. SDSC established its Data Central site,[7] which holds 27 PB of data in more than 100 specific databases (e.g., for bioinformatics and water resources). In 2009, it set aside 400 terabytes (TB) of disk space for both public and private databases and data collections that serve a wide range of scientific institutions, including laboratories, libraries, and museums.

The Australian National Data Service[8] (ANDS) has begun offering services starting with the Register My Data service, a "card catalog" that registers the identity, structure, name, and location (IP address) of all the various databases, including those coming from individuals. The mere act of registering goes a long way toward organizing long-term storage. The purpose of ANDS is to influence national policy on data management and to inform best practices for the curation

[7] http://datacentral.sdsc.edu/index.html
[8] www.ands.org.au

of data, thereby transforming the disparate collections of research data into a cohesive collection of research resources. In the UK, the Joint Information Systems Committee (JISC) has funded the establishment of a Digital Curation Centre[9] to explore these issues. Over time, one might expect that many such datacenters will emerge. The National Science Foundation's Directorate for Computer and Information Science and Engineering recently issued a call for proposals for long-term grants to researchers in data-intensive computing and long-term archiving.

In the articles in this book, the reader is invited to consider the many opportunities and challenges for data-intensive science, including interdisciplinary cooperation and training, interorganizational data sharing for "scientific data mashups," the establishment of new processes and pipelines, and a research agenda to exploit the opportunities as well as stay ahead of the data deluge. These challenges will require major capital and operational expenditure. The dream of establishing a "sensors everywhere" data infrastructure to support new modes of scientific research will require massive cooperation among funding agencies, scientists, and engineers. This dream must be actively encouraged and funded.

REFERENCES

[1] National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," Technical Report NSB-05-40, National Science Foundation, September 2005, www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

[2] Talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007, http://research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm. (Edited transcript also in this volume.)

[9] www.dcc.ac.uk

Jim Gray on eScience: A Transformed Scientific Method

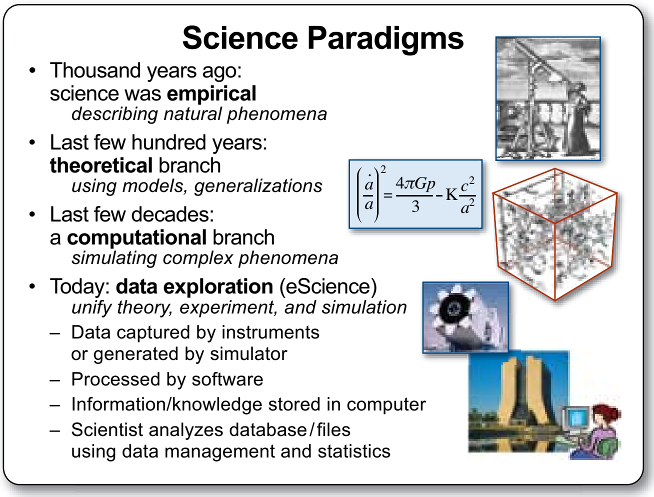
Based on the transcript of a talk given by Jim Gray to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]

EDITED BY **TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE** | Microsoft Research

We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization. Today, the tools for capturing data both at the mega-scale and at the milli-scale are just dreadful. After you have captured the data, you need to curate it before you can start doing any kind of data analysis, and we lack good tools for both data curation and data analysis. Then comes the publication of the results of your research, and the published literature is just the tip of the data iceberg. By this I mean that people collect a lot of data and then reduce this down to some number of column inches in *Science* or *Nature*—or 10 pages if it is a computer science person writing. So what I mean by data iceberg is that there is a lot of data that is collected but not curated or published in any systematic way. There are some exceptions, and I think that these cases are a good place for us to look for best practices. I will talk about how the whole process of peer review has got to change and the way in which I think it is changing and what CSTB can do to help all of us get access to our research.

[1] National Research Council, http://sites.nationalacademies.org/NRC/index.htm; Computer Science and Telecommunications Board, http://sites.nationalacademies.org/cstb/index.htm.

[2] This presentation is, poignantly, the last one posted to Jim's Web page at Microsoft Research before he went missing at sea on January 28, 2007—http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt.



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files
  using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

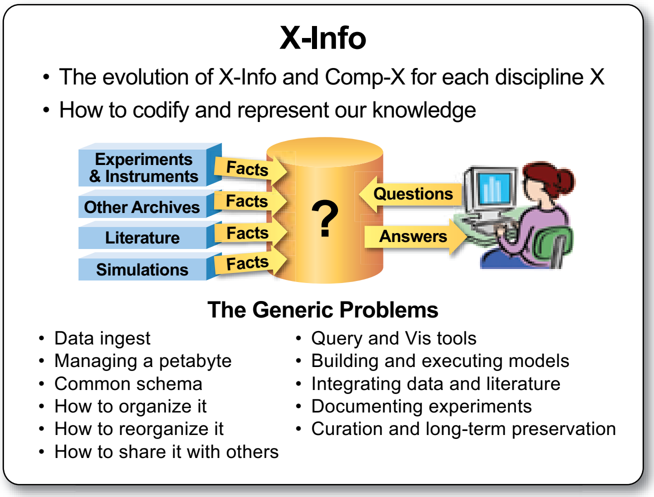
FIGURE 1

eSCIENCE: WHAT IS IT?

eScience is where “IT meets scientists.” Researchers are using many different methods to collect or generate data—from sensors and CCDs to supercomputers and particle colliders. When the data finally shows up in your computer, what do you do with all this information that is now in your digital shoebox? People are continually seeking me out and saying, “Help! I’ve got all this data. What am I supposed to do with it? My Excel spreadsheets are getting out of hand!” So what comes next? What happens when you have 10,000 Excel spreadsheets, each with 50 workbooks in them? Okay, so I have been systematically naming them, but now what do I do?

SCIENCE PARADIGMS

I show this slide [Figure 1] every time I talk. I think it is fair to say that this insight dawned on me in a CSTB study of computing futures. We said, “Look, computational science is a third leg.” Originally, there was just experimental science, and then there was theoretical science, with Kepler’s Laws, Newton’s Laws of Motion, Maxwell’s equations, and so on. Then, for many problems, the theoretical models grew too complicated to solve analytically, and people had to start simulating. These simulations have carried us through much of the last half of the last millennium. At this point, these simulations are generating a whole lot of data, along with



X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge

Experiments & Instruments
Facts
Other Archives
Facts
Literature
Facts
Simulations
Facts
?
Questions
Answers

The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share it with others
- Query and Vis tools
- Building and executing models
- Integrating data and literature
- Documenting experiments
- Curation and long-term preservation

FIGURE 2

a huge increase in data from the experimental sciences. People now do not actually look through telescopes. Instead, they are "looking" through large-scale, complex instruments which relay data to datacenters, and only then do they look at the information on their computers.

The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, *fourth paradigm* for scientific exploration [1].

X-INFO AND COMP-X

We are seeing the evolution of two branches of every discipline, as shown in the next slide [Figure 2]. If you look at ecology, there is now both computational ecology, which is to do with simulating ecologies, and eco-informatics, which is to do with collecting and analyzing ecological information. Similarly, there is bioinformatics, which collects and analyzes information from many different experiments, and there is computational biology, which simulates how biological systems work and the metabolic pathways or the behavior of a cell or the way a protein is built.

This is similar to Jeannette Wing's idea of "computational thinking," in which computer science techniques and technologies are applied to different disciplines [2].

The goal for many scientists is to codify their information so that they can exchange it with other scientists. Why do they need to codify their information? Because if I put some information in my computer, the only way you are going to be able to understand that information is if your program can understand the information. This means that the information has to be represented in an algorithmic way. In order to do this, you need a standard representation for what a gene is or what a galaxy is or what a temperature measurement is.

EXPERIMENTAL BUDGETS ARE ¼ TO ½ SOFTWARE

I have been hanging out with astronomers for about the last 10 years, and I get to go to some of their base stations. One of the stunning things for me is that I look at their telescopes and it is just incredible. It is basically 15 to 20 million dollars worth of capital equipment, with about 20 to 50 people operating the instrument. But then you get to appreciate that there are literally thousands of people writing code to deal with the information generated by this instrument and that millions of lines of code are needed to analyze all this information. In fact, the software cost dominates the capital expenditure! This is true at the Sloan Digital Sky Survey (SDSS), and it is going to continue to be true for larger-scale sky surveys, and in fact for many large-scale experiments. I am not sure that this dominant software cost is true for the particle physics community and their Large Hadron Collider (LHC) machine, but it is certainly true for the LHC experiments.

Even in the "small data" sciences, you see people collecting information and then having to put a lot more energy into the analysis of the information than they have done in getting the information in the first place. The software is typically very idiosyncratic since there are very few generic tools that the bench scientist has for collecting and analyzing and processing the data. This is something that we computer scientists could help fix by building generic tools for the scientists.

I have a list of items for policymakers like CSTB. The first one is basically to foster both building tools and supporting them. NSF now has a cyberinfrastructure organization, and I do not want to say anything bad about them, but there needs to be more than just support for the TeraGrid and high-performance computing. We now know how to build Beowulf clusters for cheap high-performance computing. But we do not know how to build a true data grid or to build data stores made out of cheap "data bricks" to be a place for you to put all your data and then analyze the information. We have actually made fair progress on simulation tools, but not very much on data analysis tools.

PROJECT PYRAMIDS AND PYRAMID FUNDING

This section is just an observation about the way most science projects seem to work. There are a few international projects, then there are more multi-campus projects, and then there are lots and lots of single-lab projects. So we basically have this Tier 1, Tier 2, Tier 3 facility pyramid, which you see over and over again in many different fields. The Tier 1 and Tier 2 projects are generally fairly systematically organized and managed, but there are only relatively few such projects. These large projects can afford to have both a software and hardware budget, and they allocate teams of scientists to write custom software for the experiment. As an example, I have been watching the U.S.-Canadian ocean observatory—Project Neptune—allocate some 30 percent of its budget for cyberinfrastructure [3]. In round numbers, that's 30 percent of 350 million dollars or something like 100 million dollars! Similarly, the LHC experiments have a very large software budget, and this trend towards large software budgets is also evident from the earlier BaBar experiment [4, 5]. But if you are a bench scientist at the bottom of the pyramid, what are you going to do for a software budget? You are basically going to buy MATLAB[3] and Excel[4] or some similar software and make do with such off-the-shelf tools. There is not much else you can do.

So the giga- and mega-projects are largely driven by the need for some large-scale resources like supercomputers, telescopes, or other large-scale experimental facilities. These facilities are typically used by a significant community of scientists and need to be fully funded by agencies such as the National Science Foundation or the Department of Energy. Smaller-scale projects can typically get funding from a more diverse set of sources, with funding agency support often matched by some other organization—which could be the university itself. In the paper that Gordon Bell, Alex Szalay, and I wrote for *IEEE Computer* [6], we observed that Tier 1 facilities like the LHC get funded by an international consortium of agencies but the Tier 2 LHC experiments and Tier 3 facilities get funded by researchers who bring with them their own sources of funding. So funding agencies need to fully fund the Tier 1 giga-projects but then allocate the other half of their funding for cyberinfrastructure for smaller projects.

[3] www.mathworks.com

[4] http://office.microsoft.com/en-us/excel/default.aspx

LABORATORY INFORMATION MANAGEMENT SYSTEMS

To summarize what I have been saying about software, what we need are effectively "Laboratory Information Management Systems." Such software systems provide a pipeline from the instrument or simulation data into a data archive, and we are close to achieving this in a number of example cases I have been working on. Basically, we get data from a bunch of instruments into a pipeline which calibrates and "cleans" the data, including filling in gaps as necessary. Then we "re-grid"[5] the information and eventually put it into a database, which you would like to "publish" on the Internet to let people access your information.

The whole business of going from an instrument to a Web browser involves a vast number of skills. Yet what's going on is actually very simple. We ought to be able to create a Beowulf-like package and some templates that would allow people who are doing wet-lab experiments to be able to just collect their data, put it into a database, and publish it. This could be done by building a few prototypes and documenting them. It will take several years to do this, but it will have a big impact on the way science is done.

As I have said, such software pipelines are called Laboratory Information Management Systems, or LIMS. Parenthetically, commercial systems exist, and you can buy a LIMS system off the shelf. The problem is that they are really geared towards people who are fairly rich and are in an industrial setting. They are often also fairly specific to one or another task for a particular community—such as taking data from a sequencing machine or mass spectrometer, running it through the system, and getting results out the other side.

INFORMATION MANAGEMENT AND DATA ANALYSIS

So here is a typical situation. People are collecting data either from instruments or sensors, or from running simulations. Pretty soon they end up with millions of files, and there is no easy way to manage or analyze their data. I have been going door to door and watching what the scientists are doing. Generally, they are doing one of two things—they are either looking for needles in haystacks or looking for the haystacks themselves. The needle-in-the-haystack queries are actually very easy—you are looking for specific anomalies in the data, and you usually have some idea of what type of signal you are looking for. The particle physicists are looking

[5] This means to "regularize" the organization of the data to one data variable per row, analogous to relational database normalization.

for the Higgs particle at the LHC, and they have a good idea of how the decay of such a heavy particle will look like in their detectors. Grids of shared clusters of computers are great for such needle-in-a-haystack queries, but such grid computers are lousy at trend analysis, statistical clustering, and discovering global patterns in the data.

We actually need much better algorithms for clustering and for what is essentially data mining. Unfortunately, clustering algorithms are not order N or N log N but are typically cubic in N, so that when N grows too large, this method does not work. So we are being forced to invent new algorithms, and you have to live with only approximate answers. For example, using the approximate median turns out to be amazingly good. And who would have guessed? Not me!

Much of the statistical analysis deals with creating uniform samples, performing some data filtering, incorporating or comparing some Monte Carlo simulations, and so on, which all generates a large bunch of files. And the situation with these files is that each file just contains a bundle of bytes. If I give you this file, you have to work hard to figure out what the data in this file means. It is therefore really important that the files be self-describing. When people use the word *database,* fundamentally what they are saying is that the data should be self-describing and it should have a schema. That's really all the word *database* means. So if I give you a particular collection of information, you can look at this information and say, "I want all the genes that have this property" or "I want all of the stars that have this property" or "I want all of the galaxies that have this property." But if I give you just a bunch of files, you can't even use the concept of a galaxy and you have to hunt around and figure out for yourself what is the effective schema for the data in that file. If you have a schema for things, you can index the data, you can aggregate the data, you can use parallel search on the data, you can have ad hoc queries on the data, and it is much easier to build some generic visualization tools.

In fairness, I should say that the science community has invented a bunch of formats that qualify in my mind as database formats. HDF[6] (Hierarchical Data Format) is one such format, and NetCDF[7] (Network Common Data Form) is another. These formats are used for data interchange and carry the data schema with them as they go. But the whole discipline of science needs much better tools than HDF and NetCDF for making data self-defining.

[6] www.hdfgroup.org

[7] www.unidata.ucar.edu/software/netcdf

DATA DELIVERY: HITTING A WALL

The other key issue is that as the datasets get larger, it is no longer possible to just FTP or grep them. A petabyte of data is very hard to FTP! So at some point, you need indices and you need parallel data access, and this is where databases can help you. For data analysis, one possibility is to move the data to you, but the other possibility is to move your query to the data. You can either move your questions or the data. Often it turns out to be more efficient to move the questions than to move the data.

THE NEED FOR DATA TOOLS: LET 100 FLOWERS BLOOM

The suggestion that I have been making is that we now have terrible data management tools for most of the science disciplines. Commercial organizations like Walmart can afford to build their own data management software, but in science we do not have that luxury. At present, we have hardly any data visualization and analysis tools. Some research communities use MATLAB, for example, but the funding agencies in the U.S. and elsewhere need to do a lot more to foster the building of tools to make scientists more productive. When you go and look at what scientists are doing, day in and day out, in terms of data analysis, it is truly dreadful. And I suspect that many of you are in the same state that I am in where essentially the only tools I have at my disposal are MATLAB and Excel!

We do have some nice tools like Beowulf[8] clusters, which allow us to get cost-effective high-performance computing by combining lots of inexpensive computers. We have some software called Condor[9] that allows you to harvest processing cycles from departmental machines. Similarly, we have the BOINC[10] (Berkeley Open Infrastructure for Network Computing) software that enables the harvesting of PC cycles as in the SETI@Home project. And we have a few commercial products like MATLAB. All these tools grew out of the research community, and I cannot figure out why these particular tools were successful.

We also have Linux and FreeBSD Unix. FreeBSD predated Linux, but somehow Linux took off and FreeBSD did not. I think that these things have a lot to do with the community, the personalities, and the timing. So my suggestion is that we should just have lots of things. We have commercial tools like LabVIEW,[11]

[8] www.beowulf.org
[9] www.cs.wisc.edu/condor
[10] http://boinc.berkeley.edu
[11] www.ni.com/labview

for example, but we should create several other such systems. And we just need to hope that some of these take off. It should not be very expensive to seed a large number of projects.

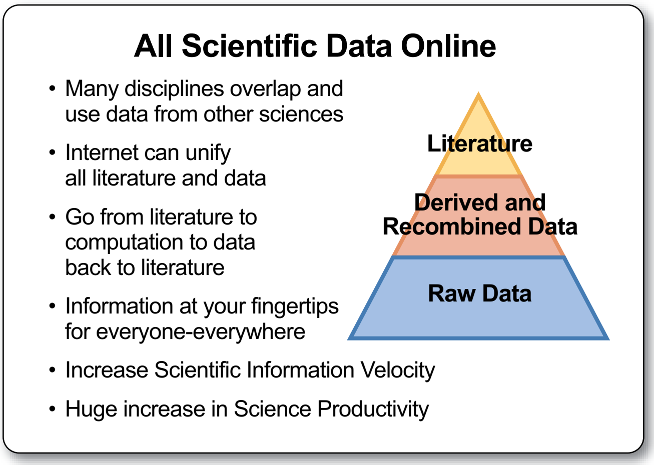
THE COMING REVOLUTION IN SCHOLARLY COMMUNICATION

I have reached the end of the first part of my talk: it was about the need for tools to help scientists capture their data, curate it, analyze it, and then visualize it. The second part of the talk is about scholarly communication. About three years ago, Congress passed a law that recommended that if you take NIH (National Institutes of Health) funding for your research, you should deposit your research reports with the National Library of Medicine (NLM) so that the full text of your papers should be in the public domain. Voluntary compliance with this law has been only 3 percent, so things are about to change. We are now likely to see all of the publicly funded science literature forced online by the funding agencies. There is currently a bill sponsored by Senators Cornyn and Lieberman that will make it compulsory for NIH grant recipients to put their research papers into the NLM PubMed Central repository.[12] In the UK, the Wellcome Trust has implemented a similar mandate for recipients of its research funding and has created a mirror of the NLM PubMed Central repository.

But the Internet can do more than just make available the full text of research papers. In principle, it can unify all the scientific data with all the literature to create a world in which the data and the literature interoperate with each other [Figure 3 on the next page]. You can be reading a paper by someone and then go off and look at their original data. You can even redo their analysis. Or you can be looking at some data and then go off and find out all the literature about this data. Such a capability will increase the "information velocity" of the sciences and will improve the scientific productivity of researchers. And I believe that this would be a very good development!

Take the example of somebody who is working for the National Institutes of Health—which is the case being discussed here—who produces a report. Suppose he discovers something about disease X. You go to your doctor and you say, "Doc, I'm not feeling very well." And he says, "Andy, we're going to give you a bunch of tests." And they give you a bunch of tests. He calls you the next day and says,

[12] See Peter Suber's Open Access newsletter for a summary of the current situation: www.earlham.edu/~peters/fos/newsletter/01-02-08.htm.



All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity

FIGURE 3

"There's nothing wrong with you. Take two aspirins, and take some vacation." You go back a year later and do the same thing. Three years later, he calls you up and says, "Andy, you have X! We figured it out!" You say, "What's X?" He says, "I have no idea, it's a rare disease, but there's this guy in New York who knows all about it." So you go to Google[13] and type in all your symptoms. Page 1 of the results, up comes X. You click on it and it takes you to PubMed Central and to the abstract "All About X." You click on that, and it takes you to the *New England Journal of Medicine*, which says, "Please give us $100 and we'll let you read about X." You look at it and see that the guy works for the National Institutes of Health. Your tax dollars at work. So Lieberman[14] and others have said, "This sucks. Scientific information is now peer reviewed and put into the public domain—but only in the sense that anybody can read it if they'll pay. What's that about? We've already paid for it."

The scholarly publishers offer a service of organizing the peer review, printing the journal, and distributing the information to libraries. But the Internet is our distributor now and is more or less free. This is all linked to the thought process that society is going through about where intellectual property begins and ends. The scientific literature, and peer reviewed literature in particular, is probably one of the places where it ends. If you want to find out about X, you will probably be

[13] Or, as Jim might have suggested today, Bing.
[14] The Federal Research Public Access Act of 2006 (Cornyn-Lieberman).

able to find out that peach pits are a great treatment for X. But this is not from the peer reviewed literature and is there just because there's a guy out there who wants to sell peach pits to you to cure X. So the people who have been pioneering this movement towards open access are primarily the folks in healthcare because the good healthcare information is locked up and the bad healthcare information is on the Internet.

THE NEW DIGITAL LIBRARY

How does the new library work? Well, it's free because it's pretty easy to put a page or an article on the Internet. Each of you could afford to publish in PubMed Central. It would just cost you a few thousand dollars for the computer—but how much traffic you would have I don't know! But curation is not cheap. Getting the stuff into the computer, getting it cross-indexed, all that sort of stuff, is costing the National Library of Medicine about $100 to curate each article that shows up. If it takes in a million articles a year, which is approximately what it expects to get, it's going to be $100 million a year just to curate the stuff. This is why we need to automate the whole curation process.

What is now going on is that PubMed Central, which is the digital part of the National Library of Medicine, has made itself portable. There are versions of PubMed Central running in the UK, in Italy, in South Africa, in Japan, and in China. The one in the UK just came online last week. I guess you can appreciate, for example, that the French don't want their National Library of Medicine to be in Bethesda, Maryland, or in English. And the English don't want the text to be in American, so the UK version will probably use UK spellings for things in its Web interface. But fundamentally, you can stick a document in any of these archives and it will get replicated to all the other archives. It's fairly cheap to run one of these archives, but the big challenges are how you do curation and peer review.

OVERLAY JOURNALS

Here's how I think it might work. This is based on the concept of overlay journals. The idea is that you have data archives and you have literature archives. The articles get deposited in the literature archives, and the data goes into the data archives. Then there is a journal management system that somebody builds that allows us, as a group, to form a journal on X. We let people submit articles to our journal by depositing them in the archive. We do peer review on them and for the ones we like, we make a title page and say, "These are the articles we like" and put it into the archive as well. Now, a search engine comes along and cranks up the page rank on all of those articles as being good because they are now referenced by this very significant front page. These articles, of course, can also point back to the data. Then there will be a collaboration system that comes along that allows people to annotate and comment on the journal articles. The comments are not stored in the peer reviewed archive but on the side because they have not been peer reviewed—though they might be moderated.

The National Library of Medicine is going to do all this for the biomedical community, but it's not happening in other scientific communities. For you as members of the CSTB, the CS community could help make this happen by providing appropriate tools for the other scientific disciplines.

There is some software we have created at Microsoft Research called Conference Management Tool (CMT). We have run about 300 conferences with this, and the CMT service makes it trivial for you to create a conference. The tool supports the whole workflow of forming a program committee, publishing a Web site, accepting manuscripts, declaring conflicts of interest and recusing yourself, doing the reviews, deciding which papers to accept, forming the conference program, notifying the authors, doing the revisions, and so on. We are now working on providing a button to deposit the articles into arXiv.org or PubMed Central and pushing in the title page as well. This now allows us to capture workshops and conferences very easily. But it will also allow you to run an online journal. This mechanism would make it very easy to create overlay journals.

Somebody asked earlier if this would be hard on scholarly publishers. And the answer is yes. But isn't this also going to be hard for the IEEE and the ACM? The answer is that the professional societies are terrified that if they don't have any paper to send you, you won't join them. I think that they are going to have to deal with this somehow because I think open access is going to happen. Looking around the room, I see that most of us are old and not Generation Xers. Most of us join these organizations because we just think it's part of being a professional in that field. The trouble is that Generation Xers don't join organizations.

WHAT HAPPENS TO PEER REVIEW?

This is not a question that has concerned you, but many people say, "Why do we need peer review at all? Why don't we just have a wiki?" And I think the answer is that peer review is different. It's very structured, it's moderated, and there is a degree of confidentiality about what people say. The wiki is much more egalitarian.

I think wikis make good sense for collecting comments about the literature after the paper has been published. One needs some structure like CMT provides for the peer review process.

PUBLISHING DATA

I had better move on and go very quickly through publishing data. I've talked about publishing literature, but if the answer is 42, what are the units? You put some data in a file up on the Internet, but this brings us back to the problem of files. The important record to show your work in context is called the data provenance. How did you get the number 42?

Here is a thought experiment. You've done some science, and you want to publish it. How do you publish it so that others can read it and reproduce your results in a hundred years' time? Mendel did this, and Darwin did this, but barely. We are now further behind than Mendel and Darwin in terms of techniques to do this. It's a mess, and we've got to work on this problem.

DATA, INFORMATION, AND KNOWLEDGE: ONTOLOGIES AND SEMANTICS

We are trying to objectify knowledge. We can help with basic things like units, and what is a measurement, who took the measurement, and when the measurement was taken. These are generic things and apply to all fields. Here [at Microsoft Research] we do computer science. What do we mean by planet, star, and galaxy? That's astronomy. What's the gene? That's biology. So what are the objects, what are the attributes, and what are the methods in the object-oriented sense on these objects? And note, parenthetically, that the Internet is really turning into an object-oriented system where people fetch objects. In the business world, they're objectifying what a customer is, what an invoice is, and so on. In the sciences, for example, we need similarly to objectify what a gene is—which is what GenBank[15] does.

And here we need a warning that to go further, you are going to bump into the O word for "ontology," the S word for "schema," and "controlled vocabularies." That is to say, in going down this path, you're going to start talking about semantics, which is to say, "What do things mean?" And of course everybody has a different opinion of what things mean, so the conversations can be endless.

The best example of all of this is Entrez,[16] the Life Sciences Search Engine,

[15] www.ncbi.nlm.nih.gov/Genbank
[16] www.ncbi.nlm.nih.gov/Entrez

created by the National Center for Biotechnology Information for the NLM. Entrez allows searches across PubMed Central, which is the literature, but they also have phylogeny data, they have nucleotide sequences, they have protein sequences and their 3-D structures, and then they have GenBank. It is really a very impressive system. They have also built the PubChem database and a lot of other things. This is all an example of the data and the literature interoperating. You can be looking at an article, go to the gene data, follow the gene to the disease, go back to the literature, and so on. It is really quite stunning!

So in this world, we have traditionally had authors, publishers, curators, and consumers. In the new world, individual scientists now work in collaborations, and journals are turning into Web sites for data and other details of the experiments. Curators now look after large digital archives, and about the only thing the same is the individual scientist. It is really a pretty fundamental change in the way we do science.

One problem is that all projects end at a certain point and it is not clear what then happens to the data. There is data at all scales. There are anthropologists out collecting information and putting it into their notebooks. And then there are the particle physicists at the LHC. Most of the bytes are at the high end, but most of the datasets are at the low end. We are now beginning to see mashups where people take datasets from various places and glue them together to make a third dataset. So in the same sense that we need archives for journal publications, we need archives for the data.

So this is my last recommendation to the CSTB: foster digital data libraries. Frankly, the NSF Digital Library effort was all about metadata for libraries and not about actual digital libraries. We should build actual digital libraries both for data and for the literature.

SUMMARY

I wanted to point out that almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, “data-intensive” science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen.

EDITORS' NOTE

The full transcript and PowerPoint slides from Jim's talk may be found at the Fourth Paradigm Web site.[17] The questions and answers during the talk have been extracted from this text and are available on the Web site. (Note that the questioners have not been identified by name.) The text presented here includes minor edits to improve readability, as well as our added footnotes and references, but we believe that it remains faithful to Jim's presentation.

REFERENCES

[1] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009, doi: 10.1126/science.1170411.

[2] J. Wing, "Computational Thinking," *Comm. ACM*, vol. 49, no. 3, Mar. 2006, doi: 10.1145/1118178.1118215.

[3] NSF Regional Scale Nodes, http://rsn.apl.washington.edu.

[4] Large Hadron Collider (LHC) experiments, http://public.web.cern.ch/Public/en/LHC/LHCExperiments-en.html.

[5] BaBar, www.slac.stanford.edu/BFROOT.

[6] G. Bell, J. Gray, and A. Szalay, "Petascale Computational Systems," *IEEE Computer*, pp. 110–112, vol. 39, 2006, doi: 10.1109/MC.2006.29.

[17] www.fourthparadigm.org

1. EARTH AND ENVIRONMENT

Introduction

DAN FAY | Microsoft Research

CHANGE IS INEVITABLE—the Universe expands, nature adapts and evolves, and so must the scientific tools and technologies that we employ to feed our unrelenting quest for greater knowledge in space, Earth, and environmental sciences. The opportunities and challenges are many. New computing technologies such as cloud computing and multicore processors cannot provide the entire solution in their generic forms. But effective and timely application of such technologies can help us significantly advance our understanding of our world, including its environmental challenges and how we might address them.

With science moving toward being computational and data based, key technology challenges include the need to better capture, analyze, model, and visualize scientific information. The ultimate goal is to aid scientists, researchers, policymakers, and the general public in making informed decisions. As society demands action and responsiveness to growing environmental issues, new types of applications grounded in scientific research will need to move from raw discovery and eliciting basic data that leads to knowledge to informing practical decisions. Active issues such as climate change will not wait until scientists have all the data to fill their knowledge gaps.

As evidenced by the articles in this part of the book, scientists are indeed actively pursuing scientific understanding through the use of new computing technologies. Szalay and Blakeley describe Jim Gray's informal rules for data-centric development and how they serve as a blueprint for making large-scale datasets available through the use of databases, leveraging the built-in data management as well as the parallel processing inherent in SQL servers.

In order to facilitate informed decisions based on reliable scientific evidence, Dozier and Gail explore how the applied use of technology and current scientific knowledge is key to providing tools to policy and decision makers. Hunt, Baldocchi, and van Ingen describe the changes under way in ecological science in moving from "science in the small" to large collaborations based on synthesis of data. These aggregated datasets expose the need for collaborative tools in the cloud as well as easy-to-use visualization and analysis tools. Delaney and Barga then provide compelling insights into the need for real-time monitoring of the complex dynamics in the sea by creating an interactive ocean laboratory. This novel cyberinfrastructure will enable new discoveries and insights through improved ocean models.

The need for novel scientific browsing technologies is highlighted by Goodman and Wong. To advance the linkage across existing resources, astronomers can use a new class of visualization tools, such as the WorldWide Telescope (WWT). This new class of tool offers access to data and information not only to professional scientists but also the general public, both for education and possibly to enable new discoveries by anyone with access to the Internet. Finally, Lehning et al. provide details about the use of densely deployed real-time sensors combined with visualization for increased understanding of environmental dynamics—like a virtual telescope looking back at the Earth. These applications illustrate how scientists and technologists have the opportunity to embrace and involve citizen scientists in their efforts.

In Part 1 and throughout the book, we see new sensors and infrastructures enabling real-time access to potentially enormous quantities of data, but with experimental repeatability through the use of workflows. Service-oriented architectures are helping to mitigate the transition to new underlying technologies and enable the linkage of data and resources. This rapidly evolving process is the only mechanism we have to deal with the data deluge arising from our instruments.

The question before us is how the world's intellectual and technological resources can be best orchestrated to authoritatively guide our responses to current and future societal challenges. The articles that follow provide some great answers.

Gray's Laws: Database-centric Computing in Science

ALEXANDER S. SZALAY
The Johns Hopkins University

JOSÉ A. BLAKELEY
Microsoft

The explosion in scientific data has created a major challenge for cutting-edge scientific projects. With datasets growing beyond a few tens of terabytes, scientists have no off-the-shelf solutions that they can readily use to manage and analyze the data [1]. Successful projects to date have deployed various combinations of flat files and databases [2]. However, most of these solutions have been tailored to specific projects and would not be easy to generalize or scale to the next generation of experiments. Also, today's computer architectures are increasingly imbalanced; the latency gap between multi-core CPUs and mechanical hard disks is growing every year, making the challenges of data-intensive computing harder to overcome [3]. What is needed is a systematic and general approach to these problems with an architecture that can scale into the future.

GRAY'S LAWS

Jim Gray formulated several informal rules—or laws—that codify how to approach data engineering challenges related to large-scale scientific datasets. The laws are as follows:

1. Scientific computing is becoming increasingly data intensive.
2. The solution is in a "scale-out" architecture.
3. Bring computations to the data, rather than data to the computations.

4. Start the design with the “20 queries.”
5. Go from “working to working.”

It is important to realize that the analysis of observational datasets is severely limited by the relatively low I/O performance of most of today’s computing platforms. High-performance numerical simulations are also increasingly feeling the “I/O bottleneck.” Once datasets exceed the random access memory (RAM) capacity of the system, locality in a multi-tiered cache no longer helps [4]. Yet very few high-end platforms provide a fast enough I/O subsystem.

High-performance, scalable numerical computation also presents an algorithmic challenge. Traditional numerical analysis packages have been designed to operate on datasets that fit in RAM. To tackle analyses that are orders of magnitude larger, these packages must be redesigned to work in a multi-phase, divide-and-conquer manner while maintaining their numerical accuracy. This suggests an approach in which a large-scale problem is decomposed into smaller pieces that can be solved in RAM, whereas the rest of the dataset resides on disk. This approach is analogous to the way in which database algorithms such as sorts or joins work on datasets larger than RAM. These challenges are reaching a critical stage.

Buying larger network storage systems and attaching them to clusters of compute nodes will not solve the problem because network/interconnect speeds are not growing fast enough to cope with the yearly doubling of the necessary storage. Scale-out solutions advocate simple building blocks in which the data is partitioned among nodes with locally attached storage [5]. The smaller and simpler these blocks are, the better the balance between CPUs, disks, and networking can become. Gray envisaged simple “CyberBricks” where each disk drive has its own CPU and networking [6]. While the number of nodes on such a system would be much larger than in a traditional “scale-up” architecture, the simplicity and lower cost of each node and the aggregate performance would more than make up for the added complexity. With the emergence of solid-state disks and low-power motherboards, we are on the verge of being able to build such systems [7].

DATABASE-CENTRIC COMPUTING

Most scientific data analyses are performed in hierarchical steps. During the first pass, a subset of the data is extracted by either filtering on certain attributes (e.g., removing erroneous data) or extracting a vertical subset of the columns. In the next step, data are usually transformed or aggregated in some way. Of course, in more complex datasets, these patterns are often accompanied by complex joins among multiple datasets, such as external calibrations or extracting and analyzing different parts of a gene sequence [8]. As datasets grow ever larger, the most efficient way to perform most of these computations is clearly to move the analysis functions as close to the data as possible. It also turns out that most of these patterns are easily expressed by a set-oriented, declarative language whose execution can benefit enormously from cost-based query optimization, automatic parallelism, and indexes.

Gray and his collaborators have shown on several projects that existing relational database technologies can be successfully applied in this context [9]. There are also seamless ways to integrate complex class libraries written in procedural languages as an extension of the underlying database engine [10, 11].

MapReduce has become a popular distributed data analysis and computing paradigm in recent years [12]. The principles behind this paradigm resemble the distributed grouping and aggregation capabilities that have existed in parallel relational database systems for some time. New-generation parallel database systems such as Teradata, Aster Data, and Vertica have rebranded these capabilities as "MapReduce in the database." New benchmarks comparing the merits of each approach have been developed [13].

CONNECTING TO THE SCIENTISTS

One of the most challenging problems in designing scientific databases is to establish effective communication between the builder of the database and the domain scientists interested in the analysis. Most projects make the mistake of trying to be "everything for everyone." It is clear that that some features are more important than others and that various design trade-offs are necessary, resulting in performance trade-offs.

Jim Gray came up with the heuristic rule of "20 queries." On each project he was involved with, he asked for the 20 most important questions the researchers wanted the data system to answer. He said that five questions are not enough to see a broader pattern, and a hundred questions would result in a shortage of focus. Since most selections involving human choices follow a "long tail," or so-called 1/f distribution, it is clear that the relative information in the queries ranked by importance is logarithmic, so the gain realized by going from approximately 20 ($2^{4.5}$) to 100 ($2^{6.5}$) is quite modest [14].

The "20 queries" rule is a moniker for a design step that engages the domain scientist and the database engineer in a conversation that helps bridge the semantic gap between nouns and verbs used in the scientific domain and the entities and relationships stored in the database. Queries define the precise set of questions in terms of entities and relationships that domain scientists expect to pose to the database. At the end of a full iteration of this exercise, the domain scientist and the database speak a common language.

This approach has been very successful in keeping the design process focused on the most important features the system must support, while at the same time helping the domain scientists understand the database system trade-offs, thereby limiting "feature creep."

Another design law is to move from working version to working version. Gray was very much aware of how quickly data-driven computing architecture changes, especially if it involves distributed data. New distributed computing paradigms come and go every other year, making it extremely difficult to engage in a multi-year top-down design and implementation cycle. By the time such a project is completed, the starting premises have become obsolete. If we build a system that starts working only if every one of its components functions correctly, we will never finish.

The only way to survive and make progress in such a world is to build modular systems in which individual components can be replaced as the underlying technologies evolve. Today's service-oriented architectures are good examples of this. Web services have already gone through several major evolutionary stages, and the end is nowhere in sight.

FROM TERASCALE TO PETASCALE SCIENTIFIC DATABASES

By using Microsoft SQL Server, we have successfully tackled several projects on a scale from a few terabytes (TB) to tens of terabytes [15-17]. Implementing databases that will soon exceed 100 TB also looks rather straightforward [18], but it is not entirely clear how science will cross the petascale barrier. As databases become larger and larger, they will inevitably start using an increasingly scaled-out architecture. Data will be heavily partitioned, making distributed, non-local queries and distributed joins increasingly difficult.

For most of the petascale problems today, a simple data-crawling strategy over massively scaled-out, share-nothing data partitions has been adequate (MapReduce, Hadoop, etc.). But it is also clear that this layout is very suboptimal when a good index might provide better performance by orders of magnitude. Joins between tables of very different cardinalities have been notoriously difficult to use with these crawlers.

Databases have many things to offer in terms of more efficient plans. We also need to rethink the utility of expecting a monolithic result set. One can imagine crawlers over heavily partitioned databases implementing a construct that can provide results one bucket at a time, resulting in easier checkpointing and recovery in the middle of an extensive query. This approach is also useful for aggregate functions with a clause that would stop when the result is estimated to be within, for example, 99% accuracy. These simple enhancements would go a long way toward sidestepping huge monolithic queries—breaking them up into smaller, more manageable ones.

Cloud computing is another recently emerging paradigm. It offers obvious advantages, such as co-locating data with computations and an economy of scale in hosting the services. While these platforms obviously perform very well for their current intended use in search engines or elastic hosting of commercial Web sites, their role in scientific computing is yet to be clarified. In some scientific analysis scenarios, the data needs to be close to the experiment. In other cases, the nodes need to be tightly integrated with a very low latency. In yet other cases, very high I/O bandwidth is required. Each of these analysis strategies would be suboptimal in current virtualization environments. Certainly, more specialized data clouds are bound to emerge soon. In the next few years, we will see if scientific computing moves from universities to commercial service providers or whether it is necessary for the largest scientific data stores to be aggregated into one.

CONCLUSIONS

Experimental science is generating vast volumes of data. The Pan-STARRS project will capture 2.5 petabytes (PB) of data each year when in production [18]. The Large Hadron Collider will generate 50 to 100 PB of data each year, with about 20 PB of that data stored and processed on a worldwide federation of national grids linking 100,000 CPUs [19]. Yet generic data-centric solutions to cope with this volume of data and corresponding analyses are not readily available [20].

Scientists and scientific institutions need a template and collection of best practices that lead to balanced hardware architectures and corresponding software to deal with these volumes of data. This would reduce the need to reinvent the wheel. Database features such as declarative, set-oriented languages and automatic parallelism, which have been successful in building large-scale scientific applications, are clearly needed.

We believe that the current wave of databases can manage at least another order of magnitude in scale. So for the time being, we can continue to work. However, it is time to start thinking about the next wave. Scientific databases are an early predictor of requirements that will be needed by conventional corporate applications; therefore, investments in these applications will lead to technologies that will be broadly applicable in a few years. Today's science challenges are good representatives of the data management challenges for the 21st century. Gray's Laws represent an excellent set of guiding principles for designing the data-intensive systems of the future.

REFERENCES

[1] A. S. Szalay and J. Gray, "Science in an Exponential World," *Nature*, vol. 440, pp. 23–24, 2006, doi: 10.1038/440413a.

[2] J. Becla and D. Wang, "Lessons Learned from Managing a Petabyte," CIDR 2005 Conference, Asilomar, 2005, doi: 10.2172/839755.

[3] G. Bell, J. Gray, and A. Szalay, "Petascale Computational Systems: Balanced Cyber-Infrastructure in a Data-Centric World," *IEEE Computer,* vol. 39, pp. 110–112, 2006, doi: 10.1109/MC.2006.29.

[4] W. W. Hsu and A. J. Smith, "Characteristics of I/O traffic in personal computer and server workloads," *IBM Sys. J.*, vol. 42, pp. 347–358, 2003, doi: 10.1147/sj.422.0347.

[5] A. Szalay, G. Bell, et al., "GrayWulf: Scalable Clustered Architecture for Data Intensive Computing," Proc. HICSS-42 Conference, Hawaii, 2009, doi: 10.1109/HICSS.2009.750.

[6] J. Gray, Cyberbricks Talk at DEC/NT Wizards Conference, 2004; T. Barclay, W. Chong, and J. Gray, "TerraServer Bricks – A High Availability Cluster Alternative," Microsoft Technical Report, MSR-TR-2004-107, http://research.microsoft.com/en-us/um/people/gray/talks/DEC_Cyberbrick.ppt.

[7] A. S. Szalay, G. Bell, A. Terzis, A. S. White, and J. Vandenberg, "Low Power Amdahl Blades for Data-Intensive Computing," http://perspectives.mvdirona.com/content/binary/AmdahlBladesV3.pdf.

[8] U. Roehm and J. A. Blakeley, "Data Management for High-Throughput Genomics," *Proc. CIDR*, 2009.

[9] J. Gray, D. T. Liu, M. A. Nieto-Santisteban, A. S. Szalay, G. Heber, and D. DeWitt, "Scientific Data Management in the Coming Decade," *ACM SIGMOD Record*, vol. 34, no. 4, pp. 35–41, 2005; also MSR-TR-2005-10, doi: 10.1145/1107499.1107503.

[10] A. Acheson et al., "Hosting the .NET Runtime in Microsoft SQL Server," ACM SIGMOD Conf., 2004, doi: 10.1145/1007568.1007669.

[11] J. A. Blakeley, M. Henaire, C. Kleinerman, I. Kunen, A. Prout, B. Richards, and V. Rao, ".NET Database Programmability and Extensibility in Microsoft SQL Server," ACM SIGMOD Conf., 2008, doi: 10.1145/1376616.1376725.

[12] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," OSDI, 2004, doi: 10.1145/1327452.1327492.

[13] A. Pavlo et al., "A Comparison of Approaches to Large-Scale Data Analysis," ACM SIGMOD Conf., 2009, doi: 10.1145/1559845.1559865.

[14] C. Anderson. *The Long Tail.* New York: Random House, 2007.

[15] A. R. Thakar, A. S. Szalay, P. Z. Kunszt, and J. Gray, "The Sloan Digital Sky Survey Science Archive: Migrating a Multi-Terabyte Astronomical Archive from Object to Relational DBMS," *Comp. Sci. and Eng.*, vol. 5, no. 5, pp. 16–29, Sept. 2003.

[16] A. Terzis, R. Musaloiu-E., J. Cogan, K. Szlavecz, A. Szalay, J. Gray, S. Ozer, M. Liang, J. Gupchup, and R. Burns, "Wireless Sensor Networks for Soil Science," *Int. J. Sensor Networks,* to be published 2009.

[17] Y. Li, E. Perlman, M. Wan, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink, "A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence," *J. Turbul.*, vol. 9, no. 31, pp. 1–29, 2008, doi: 10.1080/14685240802376389.

[18] Pan-STARRS: Panoramic Survey Telescope and Rapid Response System, http://pan-starrs.ifa.hawaii.edu.

[19] A. M. Parker, "Understanding the Universe," in *Towards 2020 Science,* Microsoft Corporation, 2006, http://research.microsoft.com/towards2020science/background_overview.htm.

[20] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science,* vol. 323, no. 5919, pp. 1297–1298, 2009, doi: 10.1126/science.1170411.

The Emerging Science of Environmental Applications

JEFF DOZIER
University of California, Santa Barbara

WILLIAM B. GAIL
Microsoft

The science of Earth and environment has matured through two major phases and is entering a third. In the first phase, which ended two decades ago, Earth and environmental science was largely discipline oriented and focused on developing knowledge in geology, atmospheric chemistry, ecosystems, and other aspects of the Earth system. In the 1980s, the scientific community recognized the close coupling of these disciplines and began to study them as interacting elements of a single system. During this second phase, the paradigm of Earth system science emerged. With it came the ability to understand complex, system-oriented phenomena such as climate change, which links concepts from atmospheric sciences, biology, and human behavior. Essential to the study of Earth's interacting systems was the ability to acquire, manage, and make available data from satellite observations; in parallel, new models were developed to express our growing understanding of the complex processes in the dynamic Earth system [1].

In the emerging third phase, knowledge developed primarily for the purpose of scientific understanding is being complemented by knowledge created to target practical decisions and action. This new knowledge endeavor can be referred to as the *science of environmental applications.* Climate change provides the most prominent example of the importance of this shift. Until now, the climate science community has focused on critical questions involving basic knowledge, from measuring the amount of change to determining the causes. With the basic understanding now well established, the demand for climate applications knowledge is emerging. How do we quantify and monitor total forest biomass so that carbon markets can characterize supply? What are the implications of regional shifts in water resources for demographic trends, agricultural output, and energy production? To what extent will seawalls and other adaptations to rising sea level impact coasts?

These questions are informed by basic science, but they raise additional issues that can be addressed only by a new science discipline focused specifically on applications—a discipline that integrates physical, biogeochemical, engineering, and human processes. Its principal questions reflect a fundamental curiosity about the nature of the world we live in, tempered by the awareness that a question's importance scales with its relevance to a societal imperative. As Nobel laureate and U.S. Secretary of Energy Steven Chu has remarked, "We seek solutions. We don't seek—dare I say this?—just scientific papers anymore" [2].

To illustrate the relationships between basic science and applications, consider the role of snowmelt runoff in water supplies. Worldwide, 1 billion people depend on snow or glacier melt for their water resources [3]. Design and operations of water systems have traditionally relied on historical measurements in a stationary climate, along with empirical relationships and models. As climates and land use change, populations grow and relocate, and our built systems age and decay, these empirical methods of managing our water become inaccurate—a conundrum characterized as "stationarity is dead" [4]. Snowmelt commonly provides water for competing uses: urban and agricultural supply, hydropower, recreation, and ecosystems. In many areas, both rainfall and snowfall occur, raising the concern that a future warmer climate will lead to a greater fraction of precipitation as rain, with the water arriving months before agricultural demand peaks and with more rapid runoff leading to more floods. In these mixed rain and snow systems, the societal need is: How do we sustain flood control and the benefits that water provides to humans and ecosystems when changes in the timing and magnitude of runoff are likely to render existing infrastructure inadequate?

The solution to the societal need requires a more fundamental, process-based understanding of the water cycle. Currently, historical data drive practices and decisions for flood control and water supply systems. Flood operations and reservoir flood capacity are predetermined by regulatory orders that are static, regardless of the type of water year, current state of the snowpack, or risk of flood. In many years, early snowmelt is not stored because statistically based projections anticipate floods that better information might suggest cannot materialize because of the absence of snow. The more we experience warming, the more frequently this occurrence will impact the water supply [5]. The related science challenges are: (1) The statistical methods in use do not try to estimate the basin's water balance, and with the current measurement networks even in the U.S., we lack adequate knowledge of the amount of snow in the basins; (2) We are unable to partition the input between rain and snow, or to partition that rain or snow between evapotranspiration and runoff; (3) We lack the knowledge to manage the relationship between snow cover, forests, and carbon stocks; (4) Runoff forecasts that are not based on physical principles relating to snowmelt are often inaccurate; and (5) We do not know what incentives and institutional arrangements would lead to better management of the watershed for ecosystem services.

Generally, models do not consider these kinds of interactions; hence the need for a *science of environmental applications*. Its core characteristics differentiate it from the basic science of Earth and environment:

- **Need driven versus curiosity driven.** Basic science is question driven; in contrast, the new applications science is guided more by societal needs than scientific curiosity. Rather than seeking answers to questions, it focuses on creating the ability to seek courses of action and determine their consequences.

- **Externally constrained.** External circumstances often dictate when and how applications knowledge is needed. The creation of carbon trading markets will not wait until we fully quantify forest carbon content. It will happen on a schedule dictated by policy and economics. Construction and repair of the urban water infrastructure will not wait for an understanding of evolving rainfall patterns. Applications science must be prepared to inform actions subject to these external drivers, not according to academic schedules based on when and how the best knowledge can be obtained.

- **Consequential and recursive.** Actions arising from our knowledge of the Earth often change the Earth, creating the need for new knowledge about what we have changed. For example, the more we knew in the past about locations of fish populations, the more the populations were overfished; our original knowledge about them became rapidly outdated through our own actions. Applications science seeks to understand not just those aspects of the Earth addressed by a particular use scenario, but also the consequences and externalities that result from that use scenario. A recent example is the shift of agricultural land to corn-for-ethanol production—an effort to reduce climate change that we now recognize as significantly stressing scarce water resources.

- **Useful even when incomplete.** As the snowpack example illustrates, actions are often needed despite incomplete data or partial knowledge. The difficulty of establishing confidence in the quality of our knowledge is particularly disconcerting given the loss of stationarity associated with climate change. New means of making effective use of partial knowledge must be developed, including robust inference engines and statistical interpretation.

- **Scalable.** Basic science knowledge does not always scale to support applications needs. The example of carbon trading presents an excellent illustration. Basic science tells us how to relate carbon content to measurements of vegetation type and density, but it does not give us the tools that scale this to a global inventory. New knowledge tools must be built to accurately create and update this inventory through cost-effective remote sensing or other means.

- **Robust.** The decision makers who apply applications knowledge typically have limited comprehension of how the knowledge was developed and in what situations it is applicable. To avoid misuse, the knowledge must be characterized in highly robust terms. It must be stable over time and insensitive to individual interpretations, changing context, and special conditions.

- **Data intensive.** Basic science is data intensive in its own right, but data sources that support basic science are often insufficient to support applications. Localized impacts with global extent, such as intrusion of invasive species, are often difficult for centralized projects with small numbers of researchers to ascertain. New applications-appropriate sources must be identified, and new ways of observing (including the use of communities as data gatherers) must be developed.

Each of these characteristics implies development of *new knowledge types* and *new tools for acquiring that knowledge.* The snowpack example illustrates what this requirement means for a specific application area. Four elements have recently come together that make deployment of a measurement and information system that can support decisions at a scale of a large river basin feasible: (1) accurate, sustained satellite estimates of snow-covered area across an entire mountain range; (2) reliable, low-cost sensors and telemetry systems for snow and soil moisture; (3) social science data that complement natural and engineered systems data to enable analysis of human decision making; and (4) cyberinfrastructure advances to integrate data and deliver them in near real time.

For snow-dominated drainage basins, the highest-priority scientific challenge is to estimate the spatial distribution and heterogeneity of the *snow water equivalent*—i.e., the amount of water that would result if the snow were to melt. Because of wind redistribution of snow after it falls, snow on the ground is far more heterogeneous than rainfall, with several meters of differences within a 10 to 100 m distance. Heterogeneity in snow depth smoothes the daily runoff because of the variability of the duration of meltwater in the snowpack [6]; seasonally, it produces quasi-riparian zones of increased soil moisture well into the summer. The approach to estimating the snow water equivalent involves several tasks using improved data: (1) extensive validation of the satellite estimates of snow cover and its reflectivity, as Figure 1 on the next page shows; (2) using results from an energy balance reconstruction of snow cover to improve interpolation from more extensive ground measurements and satellite data [7]; (3) development of innovative ways to characterize heterogeneity [8]; and (4) testing the interpolated estimates with a spatially distributed runoff model [9]. The measurements would also help clarify the accuracy in precipitation estimates from regional climate models.

This third phase of Earth and environmental science will evolve over the next decade as the scientific community begins to pursue it. Weather science has already built substantial capability in applications science; the larger field of Earth science will need to learn from and extend those efforts. The need for basic science and further discovery will not diminish, but instead will be augmented and extended by this new phase. The questions to address are both practically important and intellectually captivating. Will our hydrologic forecasting skill decline as changes in precipitation diminish the value of statistics obtained from historic patterns? Where will the next big climate change issue arise, and what policy actions taken today could allow us to anticipate it?

Equally important is improving how we apply this knowledge in our daily lives. The Internet and mobile telephones, with their global reach, provide new ways to disseminate information rapidly and widely. Information was available to avoid much of the devastation from the Asian tsunami and Hurricane Katrina, but we

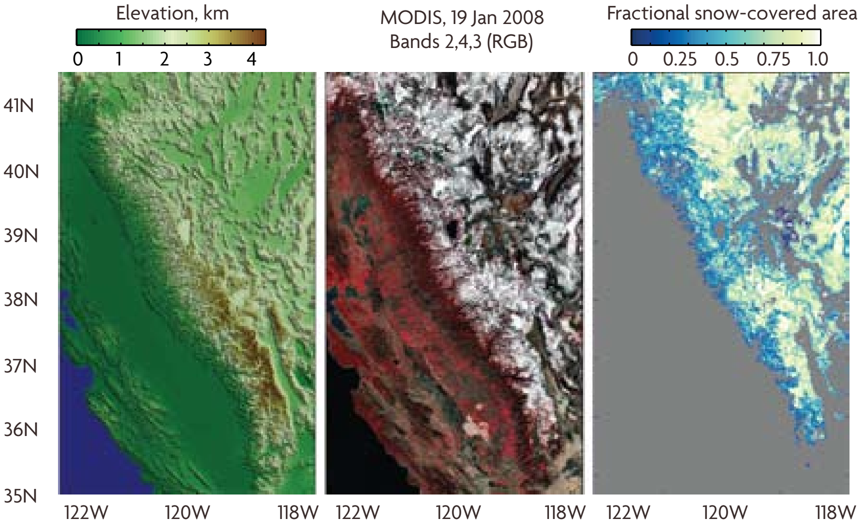


FIGURE 1.
An illustration of the type of data that are useful in analyzing the snow cover. The left panel shows elevations of the Sierra Nevada and Central Valley of California, along with a portion of northwestern Nevada. The middle panel shows the raw satellite data in three spectral bands (0.841–0.876, 0.545–0.565, and 0.459–0.479μm) from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS), which provides daily global data at 250 to 1000 m resolution in 36 spectral bands. From seven "land" bands at 500 m resolution, we derive the fractional snow-covered area—i.e., the fraction of each 500 m grid cell covered by snow, shown in the right panel [10].

lacked the tools for rapid decision making and communication of needed actions. Applications science is therefore integrative; it couples understanding of physical phenomena and research into the ways that people and organizations can use better knowledge to make decisions. The public as a whole can also become an important contributor to localized Earth observation, augmenting our limited satellite and sensor networks through devices as simple as mobile phone cameras. The ability to leverage this emerging data-gathering capability will be an important challenge for the new phase of environmental science.

The security and prosperity of nearly 7 billion people depend increasingly on our ability to gather and apply information about the world around us. Basic environmental science has established an excellent starting point. We must now develop this into a robust science of environmental applications.

REFERENCES

[1] National Research Council, *Earth Observations from Space: The First 50 Years of Scientific Achievement.* Washington, D.C.: National Academies Press, 2007.

[2] R. DelVecchio, "UC Berkeley: Panel looks at control of emissions," *S.F. Chronicle,* March 22, 2007.

[3] T. P. Barnett, J. C. Adam, and D. P. Lettenmaier, "Potential impacts of a warming climate on water availability in snow-dominated regions," *Nature,* vol. 438, pp. 303–309, 2005, doi: 10.1038/nature04141.

[4] P. C. D. Milly, J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, "Stationarity is dead: whither water management?" *Science,* vol. 319, pp. 573–574, 2008, doi: 10.1126/science.1151915.

[5] R. C. Bales, N. P. Molotch, T. H. Painter, M. D. Dettinger, R. Rice, and J. Dozier, "Mountain hydrology of the western United States," *Water Resour. Res.,* vol. 42, W08432, 2006, doi: 10.1029/2005WR004387.

[6] J. D. Lundquist and M. D. Dettinger, "How snowpack heterogeneity affects diurnal streamflow timing," *Water Resour. Res.,* vol. 41, W05007, 2005, doi: 10.1029/2004WR003649.

[7] D. W. Cline, R. C. Bales, and J. Dozier, "Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling," *Water Resour. Res.,* vol. 34, pp. 1275–1285, 1998, doi: 10.1029/97WR03755.

[8] N. P. Molotch and R. C. Bales, "Scaling snow observations from the point to the grid element: implications for observation network design," *Water Resour. Res.,* vol. 41, W11421, 2005, doi: 10.1029/2005WR004229.

[9] C. L. Tague and L. E. Band, "RHESSys: regional hydro-ecologic simulation system—an object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling," *Earth Int.,* vol. 19, pp. 1–42, 2004.

[10] T. H. Painter, K. Rittger, C. McKenzie, R. E. Davis, and J. Dozier, "Retrieval of subpixel snow-covered area, grain size, and albedo from MODIS," *Remote Sens. Environ.,* vol. 113, pp. 868–879, 2009, doi: 10.1016/j.rse.2009.01.001.

Redefining Ecological Science Using Data

JAMES R. HUNT
University of California, Berkeley, and the Berkeley Water Center

DENNIS D. BALDOCCHI
University of California, Berkeley

CATHARINE VAN INGEN
Microsoft Research

Ecology is the study of life and its interactions with the physical environment. Because climate change requires rapid adaptation, new data analysis tools are essential to quantify those changes in the midst of high natural variability. Ecology is a science in which studies have been performed primarily by small groups of individuals, with data recorded and stored in notebooks. But large synthesis studies are now being attempted by collaborations involving hundreds of scientists. These larger efforts are essential because of two developments: one in how science is done and the other in the resource management questions being asked. While collaboration synthesis studies are still nascent, their ever-increasing importance is clear. Computational support is integral to these collaborations and key to the scientific process.

HOW GLOBAL CHANGES ARE CHANGING ECOLOGICAL SCIENCE

The global climate and the Earth's landscape are changing, and scientists must quantify significant linkages between atmospheric, oceanic, and terrestrial processes to properly study the phenomena. For example, scientists are now asking how climate fluctuations in temperature, precipitation, solar radiation, length of growing season, and extreme weather events such as droughts affect the net carbon exchange between vegetation and the atmosphere. This question spans many Earth science disciplines with their respective data, models, and assumptions.

These changes require a new approach to resolving resource management questions. In the short run of the next few decades, ecosystems cannot be restored to their former status. For example, with a warming climate on the West Coast of the United States, can historical data from coastal watersheds in southern California be used to predict the fish habitats of northern California coastal watersheds? Similarly, what can remote sensing tell us about deforestation? Addressing these challenges requires a synthesis of data and models that spans length scales from the very local (river pools) to the global (oceanic circulations) and spans time scales from a few tens of milliseconds to centuries.

AN EXAMPLE OF ECOLOGICAL SYNTHESIS

Figure 1 shows a simple "science mashup" example of a synthesis study. The graph compares annual runoff from relatively small watersheds in the foothills of the Sierra Nevada in California to local annual precipitation over multiple years. Annual runoff values were obtained from the U.S. Geological Survey (USGS) for three of the gauging stations along Dry Creek and the Schubert University of California experimental field site.[1] Long-term precipitation records from nearby rain gauges were obtained from the National Climatic Data Center.[2] The precipitation that does not run off undergoes evapotranspiration (ET) that is largely dominated by watershed vegetation. In these watersheds, a single value of 400 mm is observed over all years of data. A similar value of annual ET was obtained by independent

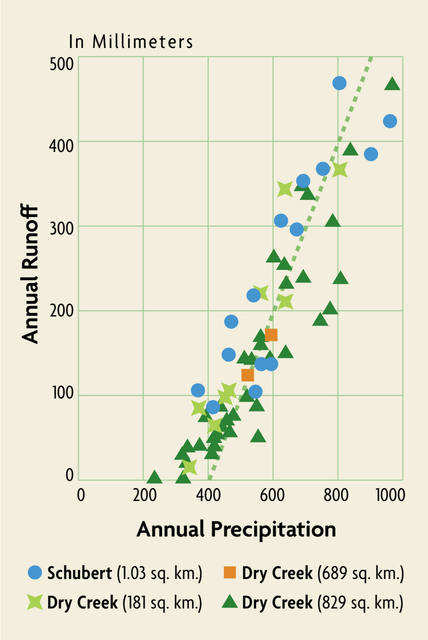


FIGURE 1.

Simple annual water balance to estimate evapotranspiration in Sierra Nevada foothill watersheds. The dashed line represents an annual ET of 400 mm.

[1] http://waterdata.usgs.gov/nwis
[2] www.ncdc.noaa.gov

measurement from atmospheric sensors deployed over an oak savannah ecosystem at the AmeriFlux Tonzi Ranch tower.[3] This synthesis of historical data defines a watershed model appropriate for historical conditions and provides a reference frame for addressing climate change effects in a highly variable system.

THE COMING FLOOD OF ECOLOGICAL DATA

These new synthesis studies are enabled by the confluence of low-cost sensors, remote sensing, Internet connectivity, and commodity computing. Sensor deployments by research groups are shifting from short campaigns to long-term monitoring with finer-scale and more diverse instruments. Satellites give global coverage particularly to remote or harsh regions where field research is hampered by physical and political logistics. Internet connectivity is enabling data sharing across organizations and disciplines. The result of these first three factors is a data flood. Commodity computing provides part of the solution, by allowing for the flood to be paired with models that incorporate different physical and biological processes and allowing for different models to be linked to span the length and time scales of interest.

The flood of ecological data and ecological science synthesis presents unique computing infrastructure challenges and new opportunities. Unlike sciences such as physics or astronomy, in which detectors are shared, in ecological science data are generated by a wide variety of groups using a wide variety of sampling or simulation methodologies and data standards. As shown earlier in Figure 1, the use of published data from two different sources was essential to obtain evapotranspiration. This synthesis required digital access to long records, separate processing of those datasets to arrive at ET, and finally verification with independent flux tower measurements. Other synthetic activities will require access to evolving resources from government organizations such as NASA or USGS, science collaborations such as the National Ecological Observatory Network and the WATERS Network,[4] individual university science research groups such as Life Under Your Feet,[5] and even citizen scientist groups such as the Community Collaborative Rain, Hail and Snow Network[6] and the USA National Phenology Network.[7]

While the bulk of the data start out as digital, originating from the field sensor,

[3] www.fluxdata.org:8080/SitePages/siteInfo.aspx?US-Ton
[4] www.watersnet.org
[5] www.lifeunderyourfeet.org
[6] www.cocorahs.org
[7] www.usanpn.org

radar, or satellite, the historic data and field data, which are critical for the science, are being digitized. The latter data are not always evenly spaced time series; they can include the date of leaf budding, or aerial imagery at different wavelengths and resolutions to assess quantities throughout the watershed such as soil moisture, vegetation, and land use. Deriving science variables from remote sensing remains an active area of research; as such, hard-won field measurements often form the ground truth necessary to develop conversion algorithms. Citizen science field observations such as plant species, plant growth (budding dates or tree ring growth, for example), and fish and bird counts are becoming increasingly important. Integrating such diverse information is an ever-increasing challenge to science analysis.

NAVIGATING THE ECOLOGICAL DATA FLOOD

The first step in any ecological science analysis is data discovery and harmonization. Larger datasets are discoverable today; smaller and historic datasets are often found by word of mouth. Because of the diversity of data publishers, no single reporting protocol exists. Unit conversions, geospatial reprojections, and time/length scale regularizations are a way of life. Science data catalog portals such as SciScope[8] and Web services with common data models such as those from the Open Geospatial Consortium[9] are evolving.

Integral to these science data search portals is knowledge of geospatial features and variable namespace mediation. The first enables searches across study watersheds or geological regions as well as simple polygon bounding boxes. The second enables searches to include multiple search terms—such as "rainfall," "precipitation," and "precip"—when searching across data repositories with different naming conventions. A new generation of metadata registries that use semantic Web technologies will enable richer searches as well as automated name and unit conversions. The combination of both developments will enable science data searches such as "Find me the daily river flow and suspended sediment discharge data from all watersheds in Washington State with more than 30 inches of annual rainfall."

MOVING ECOLOGICAL SYNTHESIS INTO THE CLOUD

Large synthesis datasets are also leading to a migration from the desktop to cloud computing. Most ecological science datasets have been collections of files. An example is the Fluxnet LaThuile synthesis dataset, containing 966 site-years of sensor

[8] www.sciscope.org
[9] www.opengeospatial.org

data from 253 sites around the world. The data for each site-year is published as a simple comma-separated or MATLAB-ready file of either daily aggregates or half-hourly aggregates. Most of the scientists download some or all of the files and then perform analyses locally. Other scientists are using an alternative cloud service that links MATLAB on the desktop to a SQL Server Analysis Services data cube in the cloud. The data appears local, but the scientists need not be bothered with the individual file handling. Local download and manipulation of the remote sensing data that would complement that sensor data are not practical for many scientists. A cloud analysis now in progress using both to compute changes in evapotranspiration across the United States over the last 10 years will download 3 terabytes of imagery and use 4,000 CPU hours of processing to generate less than 100 MB of results. Doing the analysis off the desktop leverages the higher bandwidth, large temporary storage capacity, and compute farm available in the cloud.

Synthesis studies also create a need for collaborative tools in the cloud. Science data has value for data-owner scientists in the form of publications, grants, reputation, and students. Sharing data with others should increase rather than decrease that value. Determining the appropriate citations, acknowledgment, and/or co-authorship policies for synthesis papers remains an open area of discussion in larger collaborations such as Fluxnet[10] and the North American Carbon Program.[11] Journal space and authorship limitations are an important concern in these discussions. Addressing the ethical question of what it means to be a co-author is essential: Is contributing data sufficient when that contribution is based on significant intellectual and physical effort? Once such policies are agreed upon, simple collaborative tools in the cloud can greatly reduce the logistics required to publish a paper, provide a location for the discovery of collaboration authors, and enable researchers to track how their data are used.

HOW CYBERINFRASTRUCTURE IS CHANGING ECOLOGICAL SCIENCE

The flood of ecological data will break down scientific silos and enable a new generation of scientific research. The goal of understanding the impacts of climate change is driving research that spans disciplines such as plant physiology, soil science, meteorology, oceanography, hydrology, and fluvial geomorphology. Bridging the diverse length and time scales involved will require a collection of cooperating models. Synthesizing the field observations with those model results at key length

[10] www.fluxdata.org
[11] www.nacarbon.org/nacp

and time scales is crucial to the development and validation of such models.

The diversity of ecological dataset size, dataset semantics, and dataset publisher concerns poses a cyberinfrastructure challenge that will be addressed over the next several years. Synthesis science drives not only direct conversations but also virtual ones between scientists of different backgrounds. Advances in metadata representation can break down the semantic and syntactic barriers to those conversations. Data visualizations that range from our simple mashup to more complex virtual worlds are also key elements in those conversations. Cloud access to discoverable, distributed datasets and, perhaps even more important, enabling cloud data analyses near the more massive datasets will enable a new generation of cross-discipline science.

A 2020 Vision for Ocean Science

JOHN R. DELANEY
University of Washington
ROGER S. BARGA
Microsoft Research

The global ocean is the last physical frontier on Earth. Covering 70 percent of the planetary surface, it is the largest, most complex biome we know. The ocean is a huge, mobile reservoir of heat and chemical mass. As such, it is the "engine" that drives weather-climate systems across the ocean basins and the continents, directly affecting food production, drought, and flooding on land. Water is effectively opaque to electromagnetic radiation, so the seafloor has not been as well mapped as the surfaces of Mars and Venus, and although the spatial relationships within the ocean basins are well understood to a first order, the long- and short-term temporal variations and the complexities of ocean dynamics are poorly understood.

The ultimate repository of human waste, the ocean has absorbed nearly half of the fossil carbon released since 1800. The ocean basins are a source of hazards: earthquakes, tsunamis, and giant storms. These events are episodic, powerful, often highly mobile, and frequently unpredictable. Because the ocean basins are a vast, but finite, repository of living and non-living resources, we turn to them for food, energy, and the many minerals necessary to sustain a broad range of human lifestyles. Many scientists believe that underwater volcanoes were the crucible in which early life began on Earth and perhaps on other planets. The oceans connect all continents; they are owned by no one, yet they belong to all of us by virtue of their mobile nature. The oceans may be viewed as the common heritage of humankind, the responsibility and life support of us all.

OCEAN COMPLEXITY

Our challenge is to optimize the benefits and mitigate the risks of living on a planet dominated by two major energy sources: sunlight driving the atmosphere and much of the upper ocean, and internal heat driving plate tectonics and portions of the lower ocean. For more than 4 billion years, the global ocean has responded to and integrated the impacts of these two powerful driving forces as the Earth, the oceans, the atmosphere, and life have co-evolved. As a consequence, our oceans have had a long, complicated history, producing today's immensely complex system in which thousands of physical, chemical, and biological processes continually interact over many scales of time and space as the oceans maintain our planetary-scale ecological "comfort zone."

Figure 1 captures a small fraction of this complexity, which is constantly driven by energy from above and below. Deeper understanding of this "global life-support system" requires entirely novel research approaches that will allow broad spectrum, interactive ocean processes to be studied simultaneously and interactively by many scientists—approaches that enable continuous *in situ* examination of linkages among many processes in a coherent time and space framework. Implementing these powerful new approaches is both the challenge and the vision of next-generation ocean science.

HISTORICAL PERSPECTIVE

For thousands of years, humans have gone to sea in ships to escape, to conquer, to trade, and to explore. Between October 1957 and January 1960, we launched the first Earth-orbiting satellite and dove to the deepest part of the ocean. Ships, satellites, and submarines have been the mainstays of spatially focused oceanographic research and exploration for the past 50 years. We are now poised on the next threshold of technological breakthrough that will advance oceanic discovery; this time, exploration will be focused on the time domain and interacting processes. This new era will draw deeply on the emergence, and convergence, of many rapidly evolving new technologies. These changes are setting the scene for what Marcel Proust called "[t]he real voyage of discovery, [which] lies not in seeking new landscapes, but in having new eyes."

In many ways, this "vision" of next-generation oceanographic research and

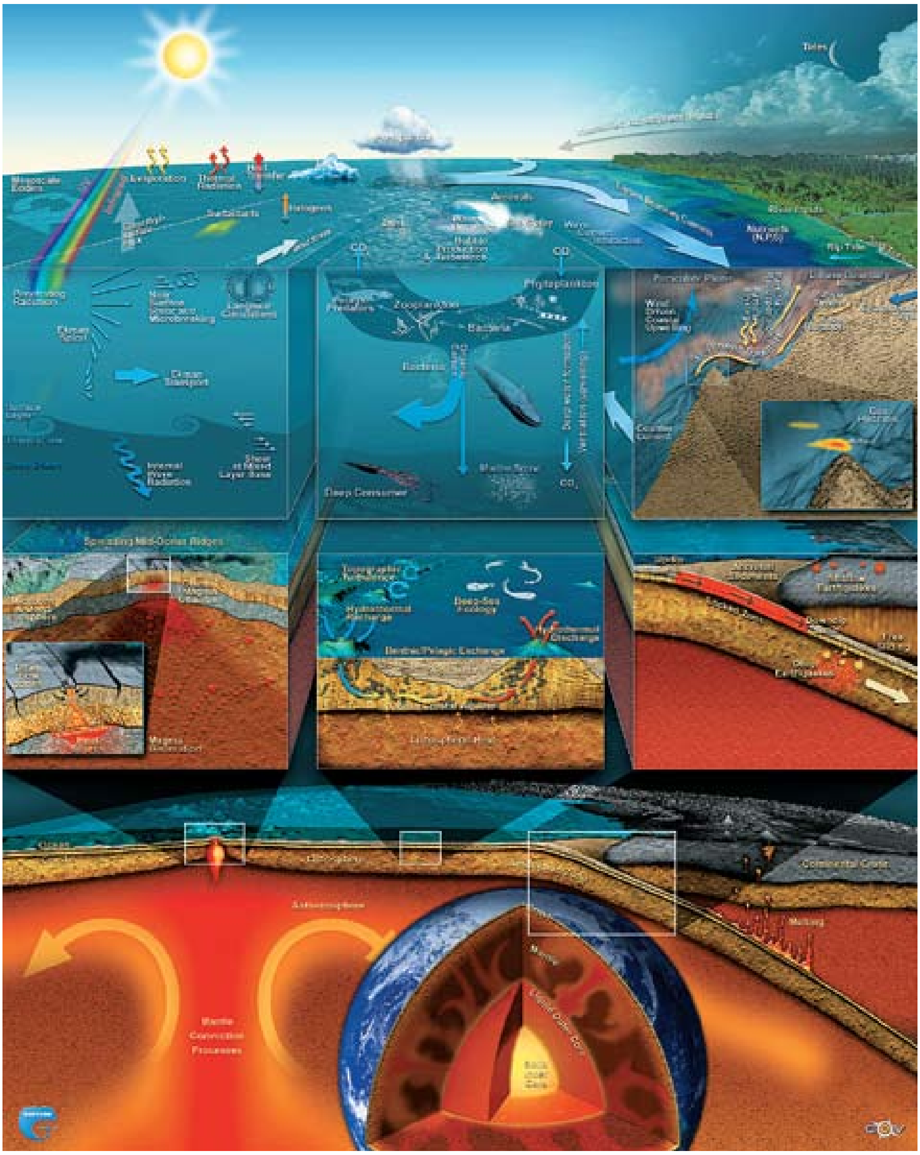


FIGURE 1.

Two primary energy sources powerfully influence the ocean basins: sunlight and its radiant energy, and internal heat with its convective and conductive input. Understanding the complexity of the oceans requires documenting and quantifying—in a well-defined time-space framework over decades—myriad processes that are constantly changing and interacting with one another.

Illustration designed by John Delaney and Mark Stoermer;
created by the Center for Environmental Visualization (CEV) for the NEPTUNE Program.

education involves utilizing a wide range of innovative technologies to simultaneously and continuously "see," or sense, many different processes operating throughout entire volumes of the ocean *from a perspective within the ocean.* Some of these same capabilities will enable remote *in situ* detection of critical changes taking place within selected ocean volumes. Rapid reconfiguration of key sensor arrays linked to the Internet via submarine electro-optical cables will allow us to capture, image, document, and measure energetic and previously inaccessible phenomena such as erupting volcanoes, major migration patterns, large submarine slumps, big earthquakes, giant storms, and a host of other complex phenomena that have been largely inaccessible to scientific study.

THE FOURTH PARADIGM

The ocean has been chronically under-sampled for as long as humans have been trying to characterize its innate complexity. In a very real sense, the current suite of computationally intensive numerical/theoretical models of ocean behavior has outstripped the requisite level of actual data necessary to ground those models in reality. As a consequence, we have been unable to even come close to useful predictive models of the real behavior of the oceans. Only by quantifying powerful episodic events, like giant storms and erupting volcanoes, within the context of longer-term decadal changes can we begin to approach dependable predictive models of ocean behavior. Over time, as the adaptive models are progressively refined by continual comparison with actual data flowing from real systems, we slowly gain the ability to predict the future behavior of these immensely complex natural systems. To achieve that goal, we must take steps to fundamentally change the way we approach oceanography.

This path has several crucial steps. We must be able to document conditions and measure fluxes *within the volume of the ocean, simultaneously and in real time,* over many scales of time and space, regardless of the depth, energy, mobility, or complexity of the processes involved. These measurements must be made using co-located arrays of many sensor types, operated by many investigators over periods of decades to centuries. And the data must be collected, archived, visualized, and compared immediately to model simulations that are explicitly configured to address complexity at scales comparable in time and space to the actual measurements.

This approach offers three major advantages: (1) The models must progressively emulate the measured reality through constant comparison with data to capture the real behavior of the oceans in "model space" to move toward more predictive simulations; (2) When the models and the data disagree, assuming the data are valid, we must immediately adapt at-sea sensor-robot systems to fully characterize the events that are unfolding because they obviously offer new insights into the complexities we seek to capture in the failed models; (3) By making and archiving all observations and measurements in coherently indexed time and space frameworks, we can allow many investigators (even those not involved in the data collection) to examine correlations among any number of selected phenomena during, or long after, the time that the events or processes occur. If the archived data are immediately and widely available via the Internet, the potential for discovery rises substantially because of the growing number of potential investigators who can explore a rapidly expanding spectrum of "parameter space." For scientists operating in this data-intensive environment, there will be a need for development of a new suite of scientific workflow products that can facilitate the archiving, assimilation, visualization, modeling, and interpretation of the information about all scientific systems of interest. Several workshop reports that offer examples of these "workflow products" are available in the open literature [1, 2].

EMERGENCE AND CONVERGENCE

Ocean science is becoming the beneficiary of a host of powerful *emergent* technologies driven by many communities that are entirely external to the world of ocean research—they include, but are not limited to, nanotechnology, biotechnology, information technology, computational modeling, imaging technologies, and robotics. More powerful yet will be the progressive *convergence* of these enabling capabilities as they are adapted to conduct sophisticated remote marine operations in novel ways by combining innovative technologies into appropriate investigative or experimental systems.

For example, computer-enabled support activities must include massive data storage systems, cloud computing, scientific workflow, advanced visualization displays, and handheld supercomputing. Instead of batteries and satellites being used to operate remote installations, electrical power and the vast bandwidth of optical fiber will be used to transform the kinds of scientific and educational activities that can be conducted within the ocean. Adaptation of industry-standard electro-optical cables for use in oceanographic research can fundamentally change the nature of human telepresence throughout the full volume of the oceans by introducing unprecedented but routinely available power and bandwidth into "ocean space." High-resolution optical and acoustic sensing will be part of the broader technology of "ocean imaging systems." These approaches will include routine use of high-definition video, in stereo if needed, as well as high-resolution sonar, acoustic lenses, laser imaging, and volumetric sampling. Advanced sensor technologies will include chemical sensing using remote, and mobile, mass spectrometers and gas chromatographs, eco-genomic analysis, and adaptive sampling techniques.

AN INTEGRATED APPROACH

After decades of planning [3, 4], the U.S. National Science Foundation (NSF) is on the verge of investing more than US$600 million over 6 years in the construction and early operation of an innovative infrastructure known as the Ocean Observatories Initiative (OOI) [4]. The design life of the program is 25 years. In addition to making much-needed high-latitude and coastal measurements supported by relatively low-bandwidth satellite communications systems, this initiative will include a transformative undertaking to implement electro-optically cabled observing systems in the northeast Pacific Ocean [5-7] off the coasts of Washington, Oregon, and British Columbia, as illustrated in Figure 2.[1]

These interactive, distributed sensor networks in the U.S. and Canada will create a large-aperture "natural laboratory" for conducting a wide range of long-term innovative experiments within the ocean volume using real-time control over the entire "laboratory" system. Extending unprecedented power and bandwidth to a wide range of interactive sensors, instruments, and robots distributed throughout the ocean water, at the air-sea interface, on the seafloor, and below the seafloor within drill holes will empower next-generation creativity and exploration of the time domain among a broad spectrum of investigators. The University of Washington leads the cabled component of the NSF initiative, known as the Regional Scale Nodes (formerly known, and funded, as NEPTUNE); the University of Victoria leads the effort in Canada, known as NEPTUNE Canada. The two approaches were conceived jointly in 2000 as a collaborative U.S.-Canadian effort. The Consortium for Ocean Leadership in Washington, D.C., is managing and integrating the entire OOI system for NSF. Woods Hole Oceanographic Institution and the University of California, San Diego, are responsible for overseeing the Coastal-Global and Cyber-Infrastructure portions of the program, respectively. Oregon State University and Scripps Institution of Oceanography are participants in the Coastal-Global portion of the OOI.

[1] www.interactiveoceans.ocean.washington.edu

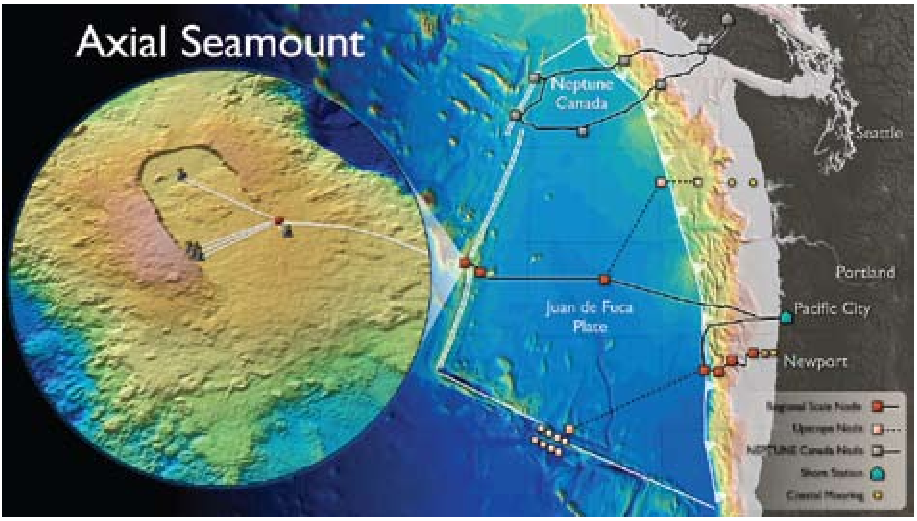


FIGURE 2.
A portion of the OOI focuses on the dynamic behavior of the Juan de Fuca Plate and the energetic processes operating in the overlying ocean and atmosphere. Recent modifications in the Regional Scale Nodes (RSN) have focused on delivery of the elements shown in red, and the pink components are future expansion. The inset shows the crest of Axial Seamount along the active Juan de Fuca Ridge. Each square block site will provide unprecedented electrical power and bandwidth available for research and education. Many of the processes shown in Figure 1 can be examined at the sites here.

Image created by CEV for OOI-RSN.

The cabled ocean observatory approach will revolutionize ocean science by providing interactive access to ocean data and instruments 24/7/365 over two to three decades. More than 1,200 kilometers of electro-optical submarine cable will deliver many tens of kilowatts of power to seafloor nodes, where instruments that might spread over a 50 km radius for each node will be plugged in directly or via secondary extension cables. The primary cable will provide between 2.5 and 10 gigabit/sec bandwidth connectivity between land and a growing number of fixed sensor packages and mobile sensor platforms. We expect that a host of novel approaches to oceanography will evolve based on the availability of *in situ* power and bandwidth. A major benefit will be the real-time data return and command-control of fleets of remotely operated vehicles (ROVs) and autonomous underwater vehicles

FIGURE 3.
Next-generation scientists or citizens. This virtual picture shows a deep ocean octopus, known as Grimpoteuthis, *and a portion of a submarine hydrothermal system on the Juan de Fuca Ridge. Such real-time displays of 3-D HD video will be routine within 5 years.*

Graphic designed by Mark Stoermer and created by CEV for NEPTUNE in 2005.

(AUVs). The infrastructure will be adaptable, expandable, and exportable to interested users. Data policy for the OOI calls for all information to be made available to all interested users via the Internet (with the exception of information bearing on national security).

Hardwired to the Internet, the cabled observatories will provide scientists, students, educators, and the public with virtual access to remarkable parts of our planet that are rarely visited by humans. In effect, the Internet will be extended to the seafloor, with the ability to interact with a host of instruments, including HD video live from the many environments within the oceans, as illustrated in Figure 3. The cabled observatory systems will be able to capture processes at the scale of the tectonic plate, mesoscale oceanic eddies, or even smaller scales. Research into representative activities responsible for climate change, major biological productivity at the base of the food chain, or encroaching ocean acidification (to name a few) will be readily conducted with this new infrastructure. Novel studies of mid-ocean spreading centers, transform faults, and especially processes in the subduction zone at the base of the continental slope, which may trigger massive earthquakes in the Pacific Northwest, will also be addressable using the same investment in the same cabled infrastructure.

This interactive ocean laboratory will be enabled by a common cyberinfrastructure that integrates multiple observatories, thousands of instruments, tens of thousands of users, and petabytes of data. The goals of the cabled ocean observatory can be achieved only if the at-sea portion is complemented by state-of-the-art information technology infrastructure resulting from a strong collaborative effort between computer scientists and ocean scientists. Such collaboration will allow scientists to interact with the ocean through real-time command and control of sensors; provide models with a continuous data feed; automate data quality control and calibration; and support novel approaches to data management, analysis, and visualization.

WHAT IS POSSIBLE?

Figure 4 on the next page depicts some of the potentially transformative capabilities that could emerge in ocean science by 2020. In the long term, a key element of the introduction of unprecedented power and bandwidth for use within the ocean basins will be the potential for bold and integrative designs and developments that enhance our understanding of, and perhaps our ability to predict, the behavior of Earth, ocean, and atmosphere interactions and their bearing on a sustainable planetary habitat.

CONCLUSION

The cabled ocean observatory merges dramatic technological advancements in sensor technologies, robotic systems, high-speed communication, eco-genomics, and nanotechnology with ocean observatory infrastructure in ways that will substantially transform the approaches that scientists, educators, technologists, and policymakers take in interacting with the dynamic global ocean. Over the coming decades, most nations will implement systems of this type in the offshore extensions of their territorial seas. As these systems become more sophisticated and data become routinely available via the Internet, the Internet will emerge as the most powerful oceanographic research tool on the planet. In this fashion, the legacy of Jim Gray will continue to grow as we learn to discover truths and insights within the data we already have "in the can."

While the cabled observatory will have profound ramifications for the manner

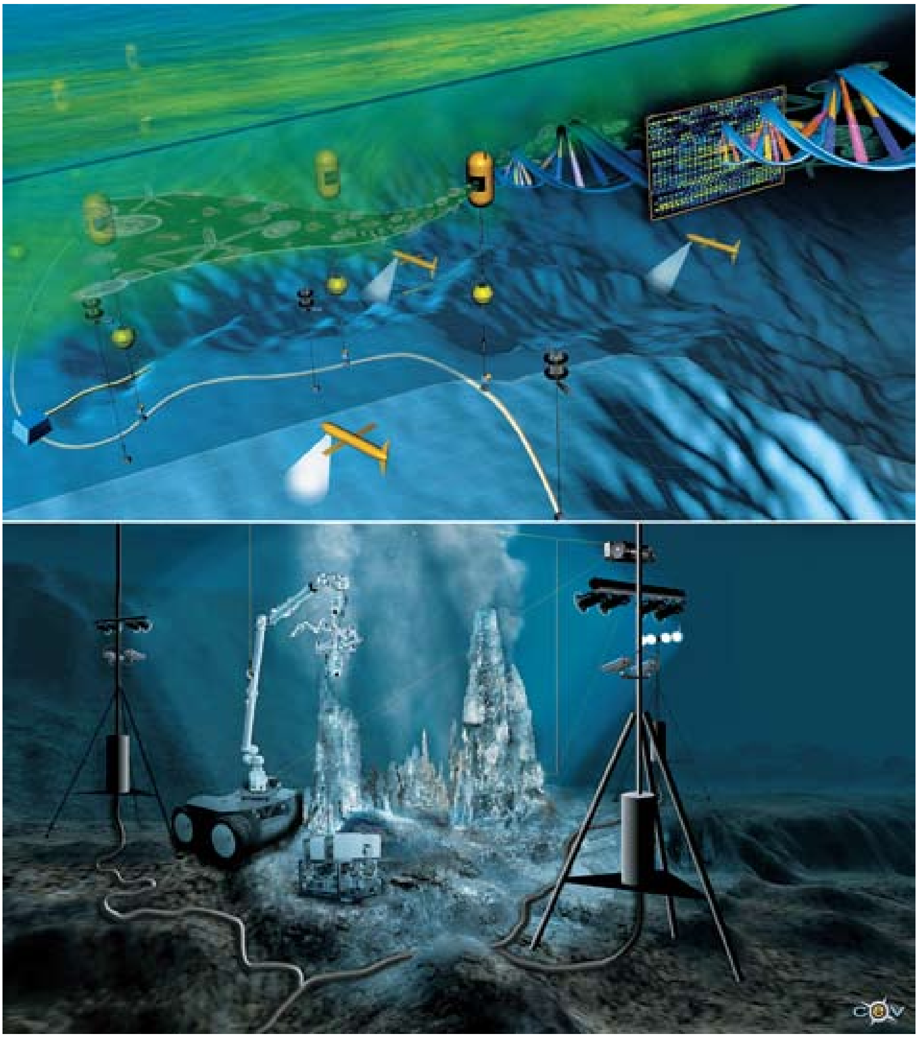


FIGURE 4.

Some of the transformative developments that could become routine within 5 years with the added power of a cabled support system. The top image shows miniaturized genomic analysis systems adapted from land laboratories to the ocean to allow scientists, with the flip of a switch in their lab hundreds of miles away, to sample ambient flow remotely and run in situ *gene sequencing operations within the ocean. The data can be made available on the Internet within minutes of the decision to sample microbes in an erupting submarine volcanic plume or a seasonally driven phytoplankton bloom. The lower part shows a conceptual illustration of an entire remote analytical-biological laboratory on the seafloor that allows a variety of key measurements or dissections to be made* in situ *using stereo high-definition video to guide high-precision remote manipulations.*

Scientific concepts by Ginger Armbrust and John Delaney; graphic design by Mark Stoermer for CEV.

in which scientists, engineers, and educators conduct their professional activities, the most far-reaching effects may be a significant shift in public attitudes toward the oceans as well as toward the scientific process. The real-time data and high-speed communications inherent in cabled remote observing systems will also open entirely new avenues for the public to interact with the natural world.

In the final analysis, having predictive models of how the ocean functions based on decades of refining sophisticated computer simulations against high-quality observations from distributed sensor networks will form the basis for learning to manage, or at least adapt to, the most powerful climate modulating system on the planet—the global ocean.

ACKNOWLEDGMENTS

We gratefully acknowledge the significant influence of Jim Gray, who unflinchingly stated that this cabled ocean observing approach using high-bandwidth and real-time data flow would be integral to human progress and understanding of the world we live in. We are also pleased to acknowledge the support of the University of Washington, the National Science Foundation, the Consortium for Ocean Leadership, and the Microsoft External Research group for technical collaboration and financial support. NSF and the National Oceanographic Partnership Program were particularly supportive of the early development of the NEPTUNE concept from 1998 to 2005, through grants to J. R. Delaney. Deborah Kelley, Nancy Penrose, and Mark Stoermer contributed significantly to the preparation of this manuscript and to conversations bearing on the content.

REFERENCES

[1] "Project Trident: A Scientific Workflow Workbench Brings Clarity to Data," http://research.microsoft.com/en-us/collaboration/focus/e3/workflowtool.aspx.

[2] Two URLs for the NSF Workshop on Challenges of Scientific Workflows:
http://grids.ucs.indiana.edu/ptliupages/publications/IEEEComputer-gil.pdf
http://vtcpc.isi.edu/wiki/index.php/Main_Page.

[3] National Research Council of the National Academies, *Enabling Ocean Research in the 21st Century: Implementation of a Network of Ocean Observatories.* Washington, D.C.: National Academies Press, 2003, p. 220.

[4] "Ocean Observatories Initiative (OOI) Scientific Objectives and Network Design: A Closer Look," 2007, http://ooi.ocean.washington.edu/cruise/cruiseFile/show/40. Ocean Leadership Web site for the Ocean Observatories Initiative: www.oceanleadership.org/programs-and-partnerships/ocean-observing/ooi.

[5] J. R. Delaney, F. N. Spiess, S. C. Solomon, R. Hessler, J. L. Karsten, J. A. Baross, R. T. Holcomb, D. Norton, R. E. McDuff, F. L. Sayles, J. Whitehead, D. Abbott, and L. Olson, "Scientific rationale for establishing long-term ocean bottom observatory/laboratory systems," in *Marine Minerals:*

Resource Assessment Strategies, P. G. Teleki, M. R. Dobson, J. R. Moor, and U. von Stackelberg, Eds., 1987, pp. 389–411.

[6] J. R. Delaney, G. R. Heath, A. D. Chave, B. M. Howe, and H. Kirkham, “NEPTUNE: Real-time ocean and earth sciences at the scale of a tectonic plate,” *Oceanography,* vol. 13, pp. 71–83, 2000, doi: 10.1109/OCEANS.2001.968033.

[7] A. D. Chave, B. St. Arnaud, M. Abbott, J. R. Delaney, R. Johnson, E. Lazowska, A. R. Maffei, J. A. Orcutt, and L. Smarr, “A management concept for ocean observatories based on web services,” *Proc. Oceans'04/Techno-Ocean'04,* Kobe, Japan, Nov. 2004, p. 7, doi: 10.1109/OCEANS.2004.1406486.

Bringing the Night Sky Closer: Discoveries in the Data Deluge

ALYSSA A. GOODMAN
Harvard University

CURTIS G. WONG
Microsoft Research

THROUGHOUT HISTORY, ASTRONOMERS have been accustomed to data falling from the sky. But our relatively newfound ability to store the sky's data in "clouds" offers us fascinating new ways to access, distribute, use, and analyze data, both in research and in education. Here we consider three interrelated questions: (1) What trends have we seen, and will soon see, in the growth of image and data collection from telescopes? (2) How might we address the growing challenge of finding the proverbial needle in the haystack of this data to facilitate scientific discovery? (3) What visualization and analytic opportunities does the future hold?

TRENDS IN DATA GROWTH

Astronomy has a history of data collection stretching back at least to Stonehenge more than three millennia ago. Over time, the format of the information recorded by astronomers has changed, from carvings in stone to written records and hand-drawn illustrations to photographs to digital media.

While the telescope (c. 1600) and the opening up of the electromagnetic spectrum beyond wavelengths visible to the human eye (c. 1940) led to qualitative changes in the nature of astronomical investigations, they did not increase the volume of collected data nearly as much as did the advent of the Digital Age.

Charge-coupled devices (CCDs), which came into widespread use by the 1980s, and equivalent detectors at non-optical wavelengths became much more efficient than traditional analog media (such as photographic plates). The resulting rise in the rate of photon collection caused the ongoing (and potentially perpetually accelerating) increase in data available to astronomers. The increasing capabilities and plummeting price of the digital devices used in signal processing, data analysis, and data storage, combined with the expansion of the World Wide Web, transformed astronomy from an observational science into a digital and computational science.

For example, the Large Synoptic Survey Telescope (LSST), coming within the decade, will produce more data in its first year of operation—1.28 petabytes—than any other telescope in history by a significant margin. The LSST will accomplish this feat by using very sensitive CCDs with huge numbers of pixels on a relatively large telescope with very fast optics (f/1.234) and a wide field of view (9.6 square degrees), and by taking a series of many shorter exposures (rather than the traditional longer exposures) that can be used to study the temporal behavior of astronomical sources. And while the LSST, Pan-STARRS, and other coming astronomical megaprojects—many at non-optical wavelengths—will produce huge datasets covering the whole sky, other groups and individuals will continue to add their own smaller, potentially more targeted, datasets.

For the remainder of this article, we will assume that the challenge of managing this explosive growth in data will be solved (likely through the clever use of "cloud" storage and novel data structures), and we will focus instead on how to offer better tools and novel technical and social analytics that will let us learn more about our universe.

A number of emerging trends can help us find the "needles in haystacks" of data available over the Internet, including crowdsourcing, democratization of access via new browsing technologies, and growing computational power.

CROWDSOURCING

The Sloan Digital Sky Survey was undertaken to image, and measure spectra for, millions of galaxies. Most of the galaxy images had never been viewed by a human because they were automatically extracted from wide-field images reduced in an automated pipeline. To test a claim that more galaxies rotate in an anticlockwise direction than clockwise, the Sloan team used custom code to create a Web page that served up pictures of galaxies to members of the public willing to play the online Galaxy Zoo game, which consists primarily of classifying the handedness of the galaxies. Clever algorithms within the "Zoo" serve the same galaxy to multiple users as a reference benchmark and to check up on players to see how accurate they are.

The results from the first year's aggregated classification of galaxies by the public proved to be just as accurate as that done by astronomers. More than 50 million classifications of a million galaxies were done by the public in the first year, and the claim about right/left handed preference was ultimately refuted. Meanwhile, Hanny Van Arkel, a schoolteacher in Holland, found a galaxy that is now the bluest known galaxy in the universe. It has come under intense scrutiny by major telescopes, including the Very Large Array (VLA) radio telescope, and will soon be scrutinized by the Hubble Space Telescope.

DEMOCRATIZING ACCESS VIA NEW BROWSING TECHNOLOGIES

The time needed to acquire data from any astronomical object increases at least as quickly as the square of the distance to that object, so any service that can accumulate custom ensembles of already captured images and data effectively brings the night sky closer. The use of archived online data stored in a "data cloud" is facilitated by new software tools, such as Microsoft's WorldWide Telescope (WWT), which provide intuitive access to images of the night sky that have taken astronomers thousands and thousands of hours of telescope time to acquire.

Using WWT (shown in Figure 1 on the next page), anyone can pan and zoom around the sky, at wavelengths from X-ray through radio, and anyone can navigate through a three-dimensional model of the Universe constructed from real observations, just to see what's there. Anyone can notice an unusual correspondence between features at multiple wavelengths at some position in the sky and click right through to all the published journal articles that discuss that position. Anyone can hook up a telescope to the computer running WWT and overlay live, new images on top of online images of the same piece of sky at virtually any wavelength. Anyone can be guided in their explorations via narrated "tours" produced by WWT users. As more and more tours are produced, WWT will become a true "sky browser," with the sky as the substrate for conversations about the universe. Explorers will navigate along paths that intersect at objects of common interest, linking ideas and individuals. Hopping from tour to tour will be like surfing from Web page to Web page now.

But the power of WWT goes far beyond its standalone ability. It is, and will continue to be, part of an ecosystem of online astronomy that will speed the progress of both "citizen" and "professional" science in the coming years.

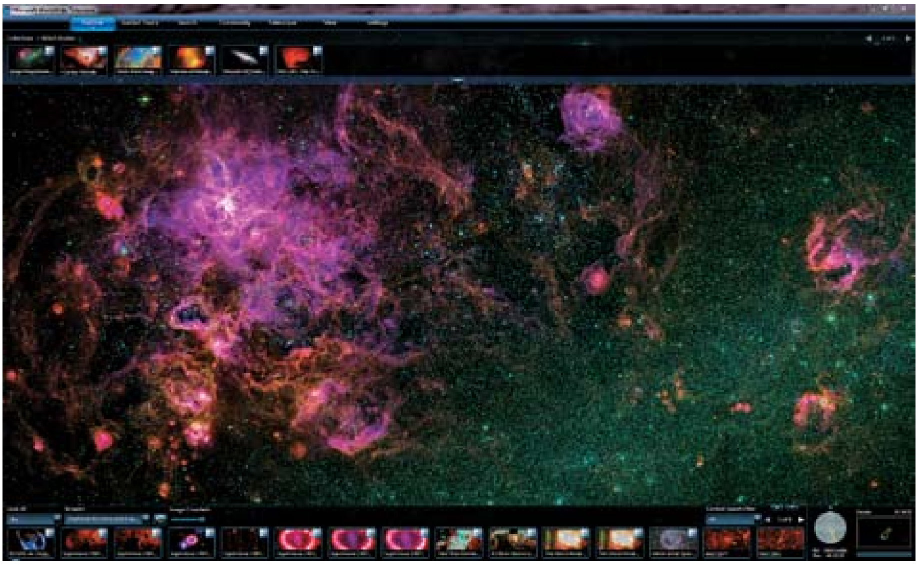


FIGURE 1.
WorldWide Telescope view of the 30 Doradus region near the Large Magellanic Cloud.

Image courtesy of the National Optical Astronomy Observatory/National Science Foundation.

Microsoft, through WWT, and Google, through Google Sky, have both created API (application programming interface) environments that allow the sky-browsing software to function inside a Web page. These APIs facilitate the creation of everything from educational environments for children to "citizen science" sites and data distribution sites for professional astronomical surveys.

Tools such as Galaxy Zoo are now easy to implement, thanks to APIs. So it now falls to the astronomical and educational communities to capitalize on the public's willingness to help navigate the increasing influx of data. High-school students can now use satellite data that no one has yet analyzed to make real discoveries about the Universe, rather than just sliding blocks down inclined planes in their physics class. Amateur astronomers can gather data on demand to fill in missing information that students, professionals, and other astronomers ask for online. The collaborative and educational possibilities are truly limitless.

The role of WWT and tools like it in the professional astronomy community will also continue to expand. WWT in particular has already become a better way to access all-sky surveys than any extant professional tool. WWT, as part of international "virtual observatory" efforts, is being seamlessly linked to quantitative and research tools that astronomers are accustomed to, in order to provide a beautiful contextual viewer for information that is usually served only piecemeal. And it has already begun to restore the kinds of holistic views of data that astronomers were used to before the Digital Age chopped up the sky into so many small pieces and incompatible formats.

GROWING COMPUTATIONAL POWER

In 10 years, multi-core processors will enhance commodity computing power two to three orders of magnitude beyond today's computers. How will all this computing power help to address the data deluge? Faster computers and increased storage and bandwidth will of course enable our contemporary approaches to scale to larger datasets. In addition, fully new ways of handling and analyzing data will be enabled. For example, computer vision techniques are already surfacing in consumer digital cameras with face detection and recognition as common features.

More computational power will allow us to triage and potentially identify unique objects, events, and data outliers as soon as they are detected and route them to citizen-scientist networks for confirmation. Engagement of citizen scientists in the alerting network for this "last leg" of detection can be optimized through better-designed interfaces that can transform work into play. Interfaces could potentially connect human confirmation of objects with global networks of games and simulations where real-time data is broadly distributed and integrated into real-time massive multiplayer games that seamlessly integrate the correct identification of the objects into the games' success metrics. Such games could give kids the opportunity to raise their social stature among game-playing peers while making a meaningful contribution to science.

VISUALIZATION AND ANALYSIS FOR THE FUTURE

WWT offers a glimpse of the future. As the diversity and scale of collected data expand, software will have to become more sophisticated in terms of how it accesses data, while simultaneously growing more intuitive, customizable, and compatible.

The way to improve tools like WWT will likely be linked to the larger challenge of how to improve the way visualization and data analysis tools can be used together in all fields—not just in astronomy.

Visualization and analysis challenges are more common across scientific fields than they are different. Imagine, for example, an astronomer and a climate scientist working in parallel. Both want to study the properties of physical systems as observed within a spherical coordinate system. Both want to move seamlessly back and forth between, for example, spectral line observations of some sources at some specific positions on a sphere (e.g., to study the composition of a stellar atmosphere or the CO_2 in the Earth's atmosphere), the context for those positions on the sphere, and journal articles and online discussions about these phenomena.

Today, even within a discipline, scientists are often faced with many choices of how to accomplish the same subtask in analysis, but no package does all the subtasks the way they would prefer. What the future holds is the potential for scientists, or data specialists working with scientists, to design their own software by linking componentized, modular applications on demand. So, for example, the astronomer and the climate scientist could both use some generalized version of WWT as part of a separate, customized system that would link to their favorite discipline- or scientist-specific packages for tasks such as spectral-line analysis.

CONCLUSION

The question linking the three topics we have discussed here is, "How can we design new tools to enhance discovery in the data deluge to come in astronomy?" The answer seems to revolve around improved *linkage* between and among existing *resources*—including citizen scientists willing to help analyze data; accessible image browsers such as WWT; and more customized visualization tools that are mashed up from common components. This approach, which seeks to more seamlessly connect (and reuse) diverse components, will likely be common to many fields of science—not just astronomy—in the coming decade.

Instrumenting the Earth: Next-Generation Sensor Networks and Environmental Science

MICHAEL LEHNING
NICHOLAS DAWES
MATHIAS BAVAY
WSL Institute for
Snow and Avalanche
Research SLF

MARC PARLANGE
École Polytechnique
Fédérale de Lausanne

SUMAN NATH
FENG ZHAO
Microsoft Research

Increasing environmental challenges worldwide and a growing awareness of global climate change indicate an urgent need for environmental scientists to conduct science in a new and better way. Existing large-scale environmental monitoring systems, with their coarse spatiotemporal resolution, are not only expensive, but they are incapable of revealing the complex interactions between atmospheric and land surface components with enough precision to generate accurate environmental system models.

This is especially the case in mountainous regions with highly complex surfaces—the source of much of the world's fresh water and weather patterns. The amount of data required to understand and model these interactions is so massive (terabytes, and increasing) that no off-the-shelf solution allows scientists to easily manage and analyze it. This has led to rapidly growing global collaboration among environmental scientists and computer scientists to approach these problems systematically and to develop sensing and database solutions that will enable environmental scientists to conduct their next-generation experiments.

NEXT-GENERATION ENVIRONMENTAL SCIENCE

The next generation of environmental science, as shown in Figure 1, is motivated by the following observations by the atmospheric science community: First, the most prominent challenge in weather and climate prediction is represented by land-atmosphere interaction processes. Second, the average effect of a patchy surface on the atmosphere can be very different from an effect that is calculated by averaging a particular surface property such as temperature or moisture [1-3]—particularly in the mountains, where surface variability is typically very high.

Figure 2 shows an example of this—a highly complex mountain surface with bare rocks, debris-covered permafrost, patchy snow cover, sparse trees, and shallow and deep soils with varying vegetation. All of these surface features can occur within a single kilometer—a resolution that is typically not reached by weather forecast models of even the latest generation. Existing models of weather prediction and climate change still operate using a grid resolution, which is far too coarse (multiple kilometers) to explicitly and correctly map the surface heterogeneity in the mountains (and elsewhere). This can lead to severe errors in understanding and prediction.

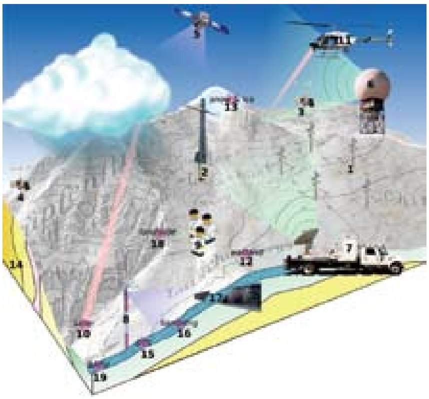


FIGURE 1.
A typical data source context for next-generation environmental science, with a heterogeneous sensor deployment that includes (1) mobile stations, (2) high-resolution conventional weather stations, (3) full-size snow/weather stations, (4) external weather stations, (5) satellite imagery, (6) weather radar, (7) mobile weather radar, (8) stream observations, (9) citizen-supplied observations, (10) ground LIDAR, (11) aerial LIDAR, (12) nitrogen/methane measures, (13) snow hydrology and avalanche probes, (14) seismic probes, (15) distributed optical fiber temperature sensing, (16) water quality sampling, (17) stream gauging stations, (18) rapid mass movements research, (19) runoff stations, and (20) soil research.

In next-generation environmental science, data resolution will be addressed using densely deployed (typically wireless) sensor networks. Recent developments in wireless sensing have made it possible to instrument and sense the physical world with high resolution and fidelity over an extended period of time. Wireless connections enable reliable collection of data from remote sensors to send to laboratories for processing, analyzing, and archiving. Such high-resolution sensing enables scientists to understand more precisely the variability and dynamics of environmental parameters. Wireless sensing also provides scientists with safe and convenient visibility of *in situ* sensor deploy-

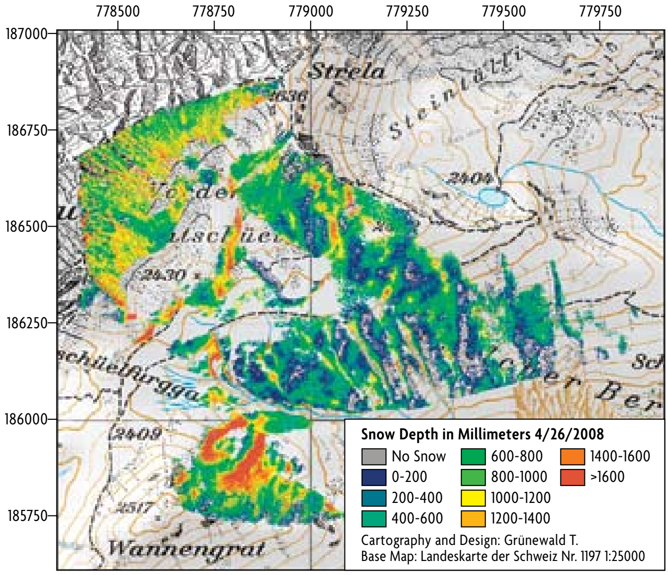


FIGURE 2.
Terrestrial laser scan for snow distribution in the Swiss Alps showing typical patchy snow cover.

ments and allows them to enable, debug, and test the deployments from the laboratory. This helps minimize site visits, which can be costly, time consuming, and even dangerous.

However, dense sensor deployments in harsh, remote environments remain challenging for several reasons. First, the whole process of sensing, computation, and communication must be extremely energy efficient so that sensors can remain operational for an extended period of time using small batteries, solar panels, or other environmental energy. Second, sensors and their communication links must be fairly robust to ensure reliable data acquisition in harsh outdoor environments. Third, invalid sensor data due to system failures or environmental impacts must be identified and treated accordingly (e.g., flagged or even filtered from the dataset). Although recent research (including the Swiss Experiment and Life Under Your Feet) partially addresses these issues, further research is needed to address them in many production systems.

MANAGING AND EXPLORING MASSIVE VOLUMES OF SENSOR DATA

High-resolution environmental sensing introduces severe data management challenges for scientists. These include reliably archiving large volumes (many terabytes) of data, sharing such data with users within access control policies, and maintaining sufficient context and provenance of sensor data using correct metadata [4].

Environmental scientists can use commercial database tools to address many of the data management and exploratory challenges associated with such a massive influx of data. For example, Microsoft's SenseWeb project [5] provides an infrastructure, including an underlying Microsoft SQL Server database, for archiving massive amounts of sensor data that might be compressed and distributed over multiple computers. SenseWeb also maintains suitable data indexes and enables efficient query processing to help users quickly explore the dataset to find features for detailed analysis [5-7]. But even with these capabilities, SenseWeb hits just the tip of the iceberg of the challenging data management tasks facing environmental scientists. Additional tools are necessary to efficiently integrate sensor data with relevant context and provide data provenance. Querying such data in a unified framework remains challenging. More research is also needed to deal with uncertain data that comes from noisy sensors and to handle the constant data flow from distributed locations.

To better understand environmental phenomena, scientists need to derive and apply various models to transform sensor data into scientific and other practical results. Database technology can help scientists to easily integrate observational data from diverse sources, possibly distributed over the Internet, with model assessments and forecasts—a procedure known as *data assimilation.* Sophisticated data mining techniques can allow scientists to easily explore spatiotemporal patterns of data (both interactively as well as in batch on archived data). Modeling techniques can provide correct and timely prediction of phenomena such as flooding events, landslides, or avalanche cycles, which can be highly useful for intervention and damage prevention, even with just a few hours of lead time. This very short-term forecasting is called *nowcasting* in meteorology.

Scientists in the Swiss Experiment project[1] have made progress in useful data assimilation and nowcasting. One case study in this project applies advanced sensors and models to forecasting alpine natural hazards [8]. A refined nowcast relies on the operational weather forecast to define the target area of a potential storm that

[1] www.swiss-experiment.ch

would affect a small-scale region (a few square kilometers) in the mountains. The operational weather forecast should allow sufficient time to install local mobile stations (such as SensorScope stations[2]) and remote sensing devices at the target area and to set up high-resolution hazard models. In the long term, specialized weather forecast models will be developed to allow much more precise local simulation.

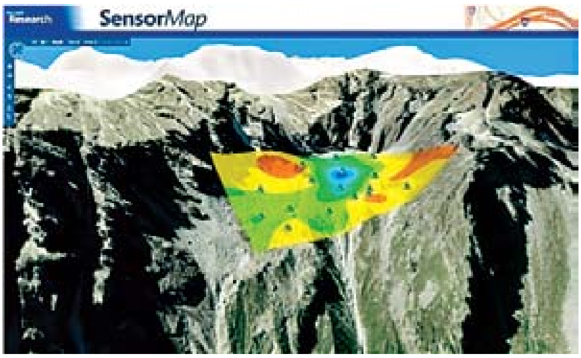


FIGURE 3.
SensorMap showing temperature distribution overlaid on 3-D mountain terrain.

To increase the public's environmental awareness and to support decision and policy makers, useful findings from scientific experiments must be presented and disseminated in a practical fashion. For example, SenseWeb provides a Web-based front end called SensorMap[3] that presents real-time and historical environmental factors in an easy-to-understand visual interface. It overlays spatial visualizations (such as icons showing current air pollution at a location or images showing distribution of snowfalls) over a browsable geographic map, plays the visualizations of selected environmental datasets as a movie on top of a geographic map, and shows important trends in historic environmental data as well as useful summaries of real-time environmental data. (See Figure 3.) At present, such platforms support only a limited set of visualizations, and many challenges remain to be solved to support the more advanced visualizations required by diverse audiences.

WORLDWIDE ENVIRONMENTAL MONITORING

We have described the next-generation environmental monitoring system as isolated—focused on a particular region of interest such as a mountain range, ice field, or forest. This is how such environmental systems are starting to be deployed. However, we foresee far more extensive monitoring systems that can allow scientists to share data with one another and combine and correlate data from millions of

[2] www.swiss-experiment.ch/index.php/SensorScope:Home
[3] www.sensormap.org

sensors all over the world to gain an even better understanding of global environmental patterns.

Such a global-scale sensor deployment would introduce unprecedented benefits and challenges. As sensor datasets grow larger, traditional data management techniques (such as loading data into a SQL database and then querying it) will clearly prove inadequate. To avoid moving massive amounts of data around, computations will need to be distributed and pushed as close to data sources as possible [7]. To reduce the storage and communication footprint, datasets will have to be compressed without loss of fidelity. To support data analysis with reasonable latencies, computation should preferably be done over compressed data [9]. Scientific analysis will also most likely require additional metadata, such as sensor specifications, experiment setups, data provenance, and other contextual information. Data from heterogeneous sources will have to be integrated in a unified data management and exploration framework [10].

Obviously, computer science tools can enable this next-generation environmental science only if they are actually used by domain scientists. To expedite adoption by domain scientists, such tools must be intuitive, easy to use, and robust. Moreover, they cannot be "one-size-fits-all" tools for all domains; rather, they should be domain-specific custom tools—or at least custom variants of generic tools. Developing these tools will involve identifying the important problems that domain scientists are trying to answer, analyzing the design trade-offs, and focusing on important features. While such application engineering approaches are common for non-science applications, they tend not to be a priority in science applications. This must change.

CONCLUSION

The close collaboration between environmental science and computer science is providing a new and better way to conduct scientific research through high-resolution and high-fidelity data acquisition, simplified large-scale data management, powerful data modeling and mining, and effective data sharing and visualization. In this paper, we have outlined several challenges to realizing the vision of next-generation environmental science. Some significant progress has been made in this context—such as in the Swiss Experiment and SenseWeb, in which an advanced, integrated environmental data infrastructure is being used by a variety of large environmental research projects, for environmental education, and by individual scientists. Meanwhile, dramatic progress is being made in complementary fields such as basic sensor technology. Our expectation is that all of these advances in instrumenting the Earth will help us realize the dreams of next-generation environmental science—allowing scientists, government, and the public to better understand and live safely in their environment.

REFERENCES

[1] M. Bavay, M. Lehning, T. Jonas, and H. Löwe, "Simulations of future snow cover and discharge in Alpine headwater catchments," *Hydrol. Processes,* vol. 22, pp. 95–108, 2009, doi: 10.1002/hyp.7195.

[2] M. Lehning, H. Löwe, M. Ryser, and N. Raderschall, "Inhomogeneous precipitation distribution and snow transport in steep terrain," *Water Resour. Res.*, vol. 44, 2008, doi: 10.1029/2007WR006545.

[3] N. Raderschall, M. Lehning, and C. Schär, "Fine scale modelling of the boundary layer wind field over steep topography," *Water Resour. Res.*, vol. 44, 2008, doi: 10.1029/2007WR006544.

[4] N. Dawes, A. K. Kumar, S. Michel, K. Aberer, and M. Lehning, "Sensor Metadata Management and Its Application in Collaborative Environmental Research," presented at the 4th IEEE Int. Conf. e-Science, 2008.

[5] A. Kansal, S. Nath, J. Liu, and F. Zhao, "SenseWeb: An Infrastructure for Shared Sensing," *IEEE MultiMedia,* vol. 14, no. 4, pp. 8–13, Oct. 2007, doi: 10.1109/MMUL.2007.82.

[6] Y. Ahmad and S. Nath, "COLR-Tree: Communication Efficient Spatio-Temporal Index for a Sensor Data Web Portal," presented at the Int. Conf. Data Engineering, 2008, doi: 10.1.1.65.6941.

[7] A. Deshpande, S. Nath, P. B. Gibbons, and S. Seshan, "Cache-and-Query for Wide Area Sensor Databases," Proc. 22nd ACM SIGMOD Int. Conf. Management of Data Principles of Database Systems, 2003, doi: 10.1145/872757.872818.

[8] M. Lehning and C. Wilhelm, "Integral Risk Management and Physical Modelling for Mountainous Natural Hazards," in *Extreme Events in Nature and Society,* S. Albeverio, V. Jentsch, and H. Kantz, Eds. Springer, 2005.

[9] G. Reeves, J. Liu, S. Nath, and F. Zhao, "Managing Massive Time Series Streams with MultiScale Compressed Trickles," *Proc. 35th Int. Conf. Very Large Data Bases,* 2009.

[10] S. Nath, J. Liu, and F. Zhao, "Challenges in Building a Portal for Sensors World-Wide," presented at the First Workshop on World-Sensor-Web, 2006, doi: 10.1109/MPRV.2007.27.

2. HEALTH AND WELLBEING

Introduction

SIMON MERCER | Microsoft Research

PART 2 OF THIS BOOK EXPLORES the remarkable progress and challenges we are seeing in the most intimate and personal of our sciences, the one with the most immediate impact on all of us across the planet: the science of health and medicine.

The first article sets the scene. Gillam et al. describe the progress of medical science over human history and make a strong case for a convergence of technologies that will change the face of healthcare within our lifetime. The remaining articles shed light on the convergent strands that make up this larger picture, by focusing on particular medical science challenges and the technologies being developed to overcome them.

Any assertion that the coming healthcare revolution will be universal is credible only if we can demonstrate how it can cross the economic and social divides of the modern world. Robertson et al. show that a combination of globally pervasive cell phone technology and the computational technique of Bayesian networks can enable collection of computerized healthcare records in regions where medical care is sparse and can also provide automated, accurate diagnoses.

An understanding of the human brain is one of the grand challenges of medicine, and Lichtman et al. describe their approach to the generation of the vast datasets needed to understand this most complex of structures. Even imaging the human brain at the subcellular level, with its estimated 160 trillion synaptic connections, is a challenge that will test the bounds of data storage, and that is merely the first step in deducing function from form.

An approach to the next stage of understanding how we think is presented by Horvitz and Kristan, who describe techniques for recording sequences of neuronal activity and correlating them with behavior in the simplest of organisms. This work will lead to a new generation of software tools, bringing techniques of machine learning/artificial intelligence to generate new insights into medical data.

While the sets of data that make up a personal medical record are orders of magnitude smaller than those describing the architecture of the brain, current trends toward universal electronic healthcare records mean that a large proportion of the global population will soon have records of their health available in a digital form. This will constitute in aggregate a dataset of a size and complexity rivaling those of neuroscience. Here we find parallel challenges and opportunities. Buchan, Winn, and Bishop apply novel machine learning techniques to this vast body of healthcare data to automate the selection of therapies that have the most desirable outcome. Technologies such as these will be needed if we are to realize the world of the "Healthcare Singularity," in which the collective experience of human healthcare is used to inform clinical best practice at the speed of computation.

While the coming era of computerized health records promises more accessible and more detailed medical data, the usability of this information will require the adoption of standard forms of encoding so that inferences can be made across datasets. Cardelli and Priami look toward a future in which medical data can be overlaid onto executable models that encode the underlying logic of biological systems—to not only depict the behavior of an organism but also predict its future condition or reaction to a stimulus. In the case of neuroscience, such models may help us understand how we think; in the case of medical records, they may help us understand the mechanisms of disease and treatment. Although the computational modeling of biological phenomena is in its infancy, it provides perhaps the most intriguing insights into the emerging complementary and synergistic relationship between computational and living systems.

The Healthcare Singularity and the Age of Semantic Medicine

MICHAEL GILLAM
CRAIG FEIED
JONATHAN HANDLER
ELIZA MOODY
Microsoft

BEN SHNEIDERMAN
CATHERINE PLAISANT
University of Maryland

MARK SMITH
MedStar Health Institutes for Innovation

JOHN DICKASON
Private practice

In 1499, when Portuguese explorer Vasco da Gama returned home after completing the first-ever sea voyage from Europe to India, he had less than half of his original crew with him—scurvy had claimed the lives of 100 of the 160 men. Throughout the Age of Discovery,[1] scurvy was the leading cause of death among sailors. Ship captains typically planned for the death of as many as half of their crew during long voyages. A dietary cause for scurvy was suspected, but no one had proved it. More than a century later, on a voyage from England to India in 1601, Captain James Lancaster placed the crew of one of his four ships on a regimen of three teaspoons of lemon juice a day. By the halfway point of the trip, almost 40% of the men (110 of 278) on three of the ships had died, while on the lemon-supplied ship, every man survived [1]. The British navy responded to this discovery by repeating the experiment—*146 years later.*

In 1747, a British navy physician named James Lind treated sailors suffering from scurvy using six randomized approaches and demonstrated that citrus reversed the symptoms. The British navy responded, 48 years later, by enacting new dietary guidelines requiring citrus, which virtually eradicated scurvy from the British fleet overnight. The British Board of Trade adopted similar dietary

[1] 15th to 17th centuries.

practices for the merchant fleet in 1865, *an additional 70 years later.* The total time from Lancaster's definitive demonstration of how to prevent scurvy to adoption across the British Empire was 264 years [2].

The translation of medical discovery to practice has thankfully improved substantially. But a 2003 report from the Institute of Medicine found that the lag between significant discovery and adoption into routine patient care still averages 17 years [3, 4]. This delayed translation of knowledge to clinical care has negative effects on both the cost and the quality of patient care. A nationwide review of 439 quality indicators found that only half of adults receive the care recommended by U.S. national standards [5].

THE IMPACT OF THE INFORMATION EXPLOSION IN MEDICINE

Despite the adoption rate of medical knowledge significantly improving, we face a new challenge due to the exponential increase in the rate of medical knowledge discovery. More than 18 million articles are currently catalogued in the biomedical literature, including over 800,000 added in 2008. The accession rate has doubled every 20 years, and the number of articles per year is expected to surpass 1 million in 2012, as shown in Figure 1.

Translating all of this emerging medical knowledge into practice is a staggering challenge. Five hundred years ago, Leonardo da Vinci could be a painter, engineer, musician, and scientist. One hundred years ago, it is said that a physician might have reasonably expected to know everything in the field of medicine.[2] Today, a typical primary care doctor must stay abreast of approximately 10,000 diseases and syndromes, 3,000 medications, and 1,100 laboratory tests [6]. Research librarians estimate that a physician in just one specialty, epidemiology, needs 21 hours of study per day just to stay current [7]. Faced with this flood of medical information, clinicians routinely fall behind, despite specialization and sub-specialization [8].

The sense of information overload in medicine has been present for surprisingly many years. An 1865 speech by Dr. Henry Noyes to the American Ophthalmologic Society is revealing. He said that "medical men strive manfully to keep up their knowledge of how the world of medicine moves on; but too often they are the first to accuse themselves of being unable to meet the duties of their daily calling…." He went on to say, "The preparatory work in the study of medicine is so great, if adequately done, that but few can spare time for its thorough performance…." [9]

[2] www.medinfo.cam.ac.uk/miu/papers/Hanka/THIM/default.htm

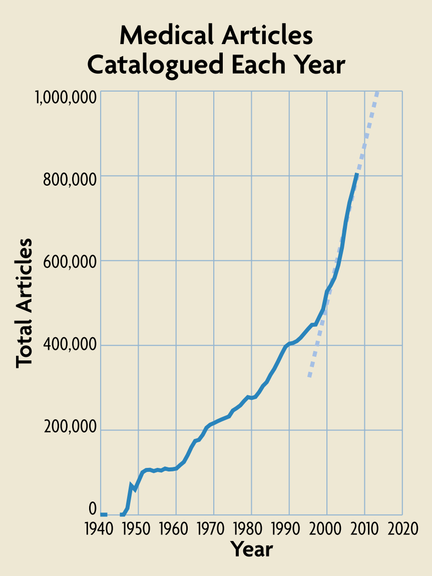


FIGURE 1.
The number of biomedical articles catalogued each year is increasing precipitously and is expected to surpass 1 million in 2012.

COULD KNOWLEDGE ADOPTION IN HEALTHCARE BECOME NEARLY INSTANTANEOUS?

The speed at which definitive medical discoveries have broadly reached medical practice over the last two millennia has progressively increased, as shown in Figure 2 on the next page.

Focusing on the last 150 years, in which the effects of industrialization and the information explosion have been most acute, the trajectory flattens slightly but remains largely linear, as the figure shows. (An asymptotic fit yields an r^2 of 0.73, whereas the linear fit is 0.83.)

Given that even the speed of light is finite, this trend will inevitably be asymptotic to the horizontal axis. Yet, if the linearity can be sufficiently maintained for a while, the next 20 years could emerge as a special time for healthcare *as the translation from medical knowledge discovery to widespread medical practice becomes nearly instantaneous.*

The proximity of this trajectory to the axis occurs around the year 2025. In response to the dramatic computational progress observed with Moore's Law and the growth in parallel and distributed computing architectures, Ray Kurzweil, in *The Singularity Is Near*, predicts that 2045 will be the year of the Singularity, when computers meet or exceed human computational ability and when their ability to recursively improve themselves can lead to an "intelligence explosion" that ultimately affects all aspects of human culture and technology [10]. Mathematics defines a "singularity" as a point at which an object changes its nature so as to attain properties that are no longer the expected norms for that class of object. Today, the dissemination path for medical information is complex and multi-faceted, involving commercials, lectures, brochures, colleagues, and journals. In a world with nearly instantaneous knowledge translation, dissemination paths would become almost entirely digital and direct.

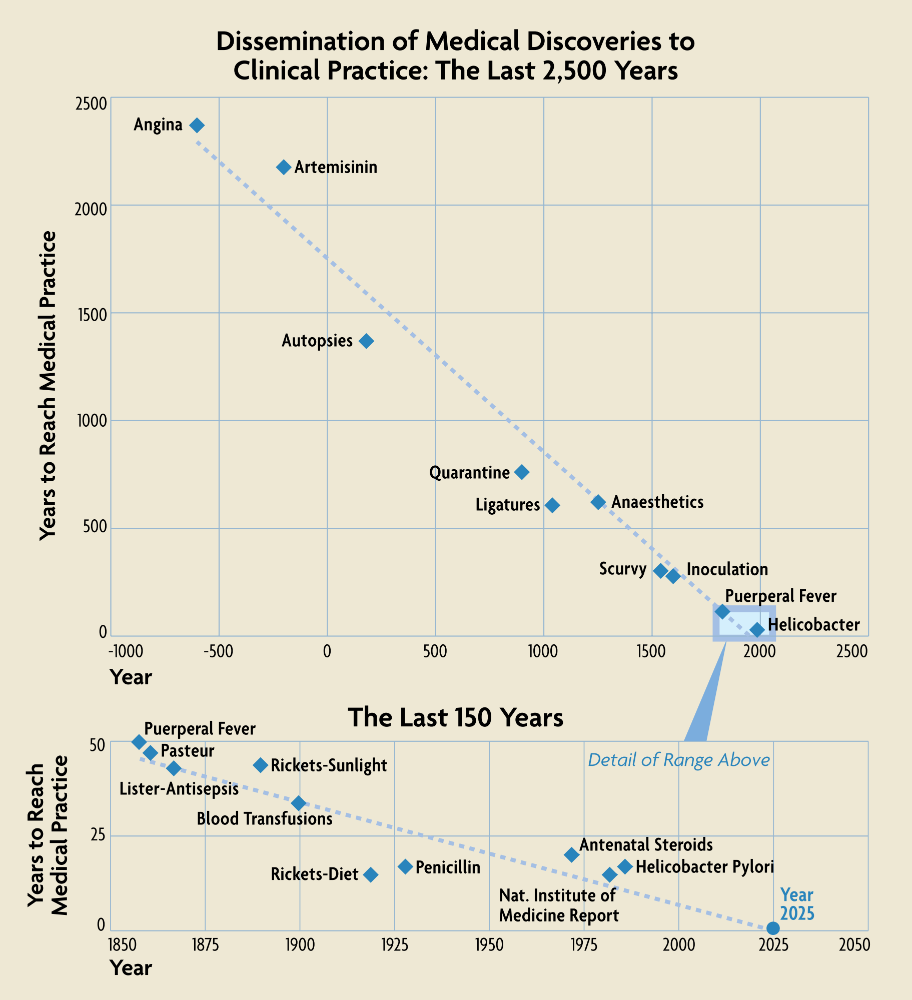


FIGURE 2.
While it took 2,300 years after the first report of angina for the condition to be commonly taught in medical curricula, modern discoveries are being disseminated at an increasingly rapid pace. Focusing on the last 150 years, the trend still appears to be linear, approaching the axis around 2025.

While the ideas around a technological singularity remain controversial,[3] the authors refer to this threshold moment, when medical knowledge becomes "liquid" and its flow from research to practice ("bench to bedside") becomes frictionless and immediate, as the "Healthcare Singularity."

THE PROMISES OF A POST–HEALTHCARE SINGULARITY WORLD

Rofecoxib (Vioxx) was approved as safe and effective by the U.S. Food and Drug Administration (FDA) on May 20, 1999. On September 30, 2004, Merck withdrew it from the market because of concerns about the drug's potential cardiovascular side effects. The FDA estimates that in the 5 years that the drug was on the market, rofecoxib contributed to more than 27,000 heart attacks or sudden cardiac deaths and as many as 140,000 cases of heart disease [11]. Rofecoxib was one of the most widely used medications ever withdrawn; over 80 million people had taken the drug, which was generating US$2.5 billion a year in sales.[4]

Today, it is reasonable to expect that after an FDA announcement of a drug's withdrawal from the market, patients will be informed and clinicians will immediately prescribe alternatives. But current channels of dissemination delay that response. In a post–Healthcare Singularity world, that expectation will be met. To enable instantaneous translation, journal articles will consist of not only words, but also bits. Text will commingle with code, and articles will be considered complete only if they include algorithms.

With this knowledge automation, every new medication will flow through a cascade of post-market studies that are independently created and studied by leading academics across the oceans (effectively "crowdsourcing" quality assurance). Suspicious observations will be flagged in real time, and when certainty is reached, unsafe medications will disappear from clinical prescription systems in a rippling wave across enterprises and clinics. The biomedical information explosion will at last be contained and harnessed.

Other scenarios of knowledge dissemination will be frictionless as well: medical residents can abandon the handbooks they have traditionally carried that list drugs of choice for diseases, opting instead for clinical systems that personalize healthcare and geographically regionalize treatments based on drug sensitivities that are drawn in real time from the local hospital microbiology lab and correlated with the patient's genomic profile.

[3] http://en.wikipedia.org/wiki/Technological_singularity
[4] http://en.wikipedia.org/wiki/Rofecoxib

Knowledge discovery will also be enhanced. Practitioners will have access to high-performance, highly accurate databases of patient records to promote preventive medical care, discover successful treatment patterns [12, 13], and reduce medical errors. Clinicians will be able to generate cause-effect hypotheses, run virtual clinical trials to deliver personalized treatment plans, and simulate interventions that can prevent pandemics.

Looking farther ahead, the instantaneous flow of knowledge from research centers to the front lines of clinical care will speed the treatment and prevention of newly emerging diseases. The moment that research labs have identified the epitopes to target for a new disease outbreak, protein/DNA/RNA/lipid synthesizers placed in every big hospital around the world will receive instructions, remotely transmitted from a central authority, directing the on-site synthesis of vaccines or even directed antibody therapies for rapid administration to patients.

PROGRESS TOWARD THE HEALTHCARE SINGULARITY

Companies such as Microsoft and Google are building new technologies to enable data and knowledge liquidity. Microsoft HealthVault and Google Health are Internet based, secure, and private "consumer data clouds" into which clinical patient data can be pushed from devices and other information systems. Importantly, once the data are in these "patient clouds," they are owned by the patient. Patients themselves determine what data can be redistributed and to whom the data may be released.

A February 2009 study by KLAS reviewed a new class of emerging data aggregation solutions for healthcare. These enterprise data aggregation solutions ("enterprise data clouds") unify data from hundreds or thousands of disparate systems (such as MEDSEEK, Carefx, dbMotion, Medicity, and Microsoft Amalga).[5] These platforms are beginning to serve as conduits for data to fill patient data clouds. A recent example is a link between New York-Presbyterian's hospital-based Amalga aggregation system and its patients' HealthVault service.[6] Through these links, data can flow almost instantaneously from hospitals to patients.

The emergence of consumer data clouds creates new paths by which new medical knowledge can reach patients directly. On April 21, 2009, Mayo Clinic announced the launch of the Mayo Clinic Health Advisory, a privacy- and security-enhanced

[5] www.klasresearch.com/Klas/Site/News/PressReleases/2009/Aggregation.aspx

[6] http://chilmarkresearch.com/2009/04/06/healthvault-ny-presbyterian-closing-the-loop-on-care

online application that offers individualized health guidance and recommendations built with the clinical expertise of Mayo Clinic and using secure and private patient health data from Microsoft HealthVault.[7] Importantly, new medical knowledge and recommendations can be computationally instantiated into the advisory and applied virtually instantaneously to patients worldwide.

New technology is bridging research labs and clinical practice. On April 28, 2009, Microsoft announced the release of Amalga Life Sciences, an extension to the data-aggregation class of products for use by scientists and researchers. Through this release, Microsoft is offering scalable "data aggregation and liquidity" solutions that link three audiences: patients, providers, and researchers. Companies such as Microsoft are building the "pipeline" to allow data and knowledge to flow through a *semantically interoperable* network of patients, providers, and researchers. These types of connectivity efforts hold the promise of effectively instantaneous dissemination of medical knowledge throughout the healthcare system. The Healthcare Singularity could be the gateway event to a new Age of Semantic Medicine.

Instantaneous knowledge translation in medicine is not only immensely important, highly desirable, valuable, and achievable in our lifetimes, but perhaps even inevitable.

REFERENCES

[1] F. Mosteller, "Innovation and evaluation," *Science*, vol. 211, pp. 881–886, 1981, doi: 10.1126/science.6781066.

[2] J. Lind, *A Treatise of the Scurvy* (1753). Edinburgh: University Press, reprinted 1953.

[3] E. A. Balas, "Information Systems Can Prevent Errors and Improve Quality," *J. Am. Med. Inform. Assoc.*, vol. 8, no. 4, pp. 398–399, 2001, PMID: 11418547.

[4] A. C. Greiner and Elisa Knebel, Eds., *Health Professions Education: A Bridge to Quality.* Washington, D.C.: National Academies Press, 2003.

[5] E. A. McGlynn, S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, et al., "The quality of health care delivered to adults in the United States," *N. Engl. J. Med.*, vol. 348, pp. 2635–2645, 2003, PMID: 12826639.

[6] T. H. Davenport and J. Glaser, "Just-in-time delivery comes to knowledge management," *Harv. Bus. Rev.*, vol. 80, no. 7, pp. 107–111, 126, July 2002, doi: 10.1225/R0207H.

[7] B. S. Alper, J. A. Hand, S. G. Elliott, S. Kinkade, M. J. Hauan, D. K. Onion, and B. M. Sklar, "How much effort is needed to keep up with the literature relevant for primary care?" *J. Med. Libr. Assoc.*, vol. 92, no. 4, pp. 429–437, Oct. 2004.

[8] C. Lenfant, "Clinical Research to Clinical Practice — Lost in Translation?" *N. Engl. J. Med.*, vol. 349, pp. 868–874, 2003, PMID: 12944573.

[9] H. D. Noyes, *Specialties in Medicine,* June 1865.

[7] www.microsoft.com/presspass/press/2009/apr09/04-21MSMayoConsumerSolutionPR.mspx

[10] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin Group, 2005, p. 136.

[11] D. J. Graham, D. Campen, R. Hui, M. Spence, C. Cheetham, G. Levy, S. Shoor, and W. A. Ray, "Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study," *Lancet,* vol. 365, no. 9458, pp. 475–481, Feb. 5–11, 2005.

[12] C. Plaisant, S. Lam, B. Shneiderman, M. S. Smith, D. Roseman, G. Marchand, M. Gillam, C. Feied, J. Handler, and H. Rappaport, "Searching Electronic Health Records for temporal patterns in patient histories: A case study with Microsoft Amalga," *Proc. Am. Med. Inform. Assoc.,* Washington, D.C., Nov. 2008.

[13] T. Wang, C. Plaisant, A. Quinn, R. Stanchak, B. Shneiderman, and S. Murphy, "Aligning temporal data by sentinel events: Discovering patterns in electronic health records," *Proc. ACM CHI2008 Human Factors in Computing Systems Conference,* ACM, New York, Apr. 2008, pp. 457–466, doi: 10.1145/1357054.1357129.

HEALTH AND WELLBEING

Healthcare Delivery in Developing Countries: Challenges and Potential Solutions

JOEL ROBERTSON
DEL DEHART
Robertson Research Institute

KRISTIN TOLLE
DAVID HECKERMAN
Microsoft Research

Bringing intelligent healthcare informatics to bear on the dual problems of reducing healthcare costs and improving quality and outcomes is a challenge even in countries with a reasonably developed technology infrastructure. Much of medical knowledge and information remains in paper form, and even where it is digitized, it often resides in disparate datasets and repositories and in diverse formats. Data sharing is uncommon and frequently hampered by the lack of foolproof de-identification for patient privacy. All of these issues impede opportunities for data mining and analysis that would enable better predictive and preventive medicine.

Developing countries face these same issues, along with the compounding effects of economic and geopolitical constraints, transportation and geographic barriers, a much more limited clinical workforce, and infrastructural challenges to delivery. Simple, high-impact deliverable interventions such as universal childhood immunization and maternal childcare are hampered by poor monitoring and reporting systems. A recent *Lancet* article by Christopher Murray's group concluded that "immunization coverage has improved more gradually and not to the level suggested by countries' official reports of WHO and UNICEF estimates. There is an urgent need for independent and contestable monitoring of health indicators in an era of global initiatives that are target-

The NxOpinion health platform being used by Indian health extension workers.

oriented and disburse funds based on performance." [1]

Additionally, the most recent report on the United Nations Millennium Development Goals notes that "pneumonia kills more children than any other disease, yet in developing countries, the proportion of children under five with suspected pneumonia who are taken to appropriate health-care providers remains low." [2] Providing reliable data gathering and diagnostic decision support at the point of need by the best-trained individual available for care is the goal of public health efforts, but tools to accomplish this have been expensive, unsupportable, and inaccessible.

Below, we elaborate on the challenges facing healthcare delivery in developing countries and describe computer- and cell phone–based technology we have created to help address these challenges. At the core of this technology is the NxOpinion Knowledge Manager[1] (NxKM), which has been under development at the Robertson Research Institute since 2002. This health platform includes a medical knowledge base assembled from the expertise of a large team of experts in the U.S. and developing countries, a diagnostic engine based on Bayesian networks, and cell phones for end-user interaction.

SCALE UP, SCALE OUT, AND SCALE IN

One of the biggest barriers to deployment of a decision support or electronic health record system is the ability to scale. The term "scale up" refers to a system's ability to support a large user base—typically hundreds of thousands or millions. Most systems are evaluated within a narrower scope of users. "Scale out" refers to a system's ability to work in multiple countries and regions as well as the ability to work across disease types. Many systems work only for one particular disease and are not easily regionalized—for example, for local languages, regulations, and processes. "Scale in" refers to the ability of a system to capture and benchmark against a single

[1] www.nxopinion.com/product/knowledgemng

individual. Most systems assume a generic patient and fail to capture unique characteristics that can be effective in individualized treatment.

With respect to scaling up, NxKM has been tested in India, Congo, Dominican Republic, Ghana, and Iraq. It has also been tested in an under-served inner-city community in the United States. In consultation with experts in database scaling, the architecture has been designed to combine multiple individual databases with a central de-identified database, thus allowing, in principle, unlimited scaling options.

As for scaling out to work across many disease types and scaling in to provide accurate individual diagnoses, the amount of knowledge required is huge. For example, INTERNIST-1, an expert system for diagnosis in internal medicine, contains approximately 250,000 relationships among roughly 600 diseases and 4,000 findings [3]. Building on the earlier work of one of us (Heckerman), who developed efficient methods for assessing and representing expert medical knowledge via a Bayesian network [4], we have brought together medical literature, textbook information, and expert panel recommendations to construct a growing knowledge base for NxKM, currently including over 1,000 diseases and over 6,000 discrete findings. The system also scales in by allowing very fine-grained data capture. Each finding within an individual health record or diagnostic case can be tracked and monitored. This level of granularity allows for tremendous flexibility in determining factors relating to outcome and diagnostic accuracy.

With regard to scaling out across a region, a challenge common to developing countries is the exceptionally diverse and region-specific nature of medical conditions. For example, a disease that is common in one country or region might be rare in another. Whereas rule-based expert systems must be completely reengineered in each region, the modular nature of the NxKM knowledge base, which is based on probabilistic similarity networks [4], allows for rapid customization to each region. The current incarnation of NxKM uses region-specific prevalence from expert estimates. It can also update prevalence in each region as it is used in the field. NxKM also incorporates a modular system that facilitates customization to terms, treatments, and language specific to each region. When region-specific information is unknown or unavailable, a default module is used until such data can be collected or identified.

DIAGNOSTIC ACCURACY AND EFFICIENCY

Studies indicate that even highly trained physicians overestimate their diagnostic accuracy. The Institute of Medicine recently estimated that 44,000 to 98,000 preventable deaths occur each year due to medical error, many due to misdiagnosis [5]. In developing countries, the combined challenges of misdiagnoses and missing data not only reduce the quality of medical care for individuals but lead to missed outbreak recognition and flawed population health assessment and planning.

Again, building on the diagnostic methodology from probabilistic similarity networks [4], NxKM employs a Bayesian reasoning engine that yields accurate diagnoses. An important component of the system that leads to improved accuracy is the ability to ask the user additional questions that are likely to narrow the range of possible diagnoses. NxKM has the ability to ask the user for additional findings based on value-of-information computations (such as a cost function) [4]. Also important for clinical use is the ability to identify the confidence in the diagnosis (i.e., the probability of the most likely diagnosis). This determination is especially useful for less-expert users of the system, which is important for improving and supervising the care delivered by health extension workers (HEWs) in developing regions where deep medical knowledge is rare.

GETTING HEALTHCARE TO WHERE IT IS NEEDED: THE LAST MILE

Another key challenge is getting diagnostics to where they are most needed. Because of their prevalence in developing countries, cell phones are a natural choice for a delivery vehicle. Indeed, it is believed that, in many such areas, access to cell phones is better than access to clean water. For example, according to the market database Wireless Intelligence,[2] 80 percent of the world's population was within range of a cellular network in 2008. And figures from the International Telecommunication Union[3] show that by the end of 2006, 68 percent of the world's mobile subscriptions were in developing countries. More recent data from the International Telecommunications Union shows that between 2002 and 2007, cellular subscription was the most rapid growth area for telecommunication in the world, and that the per capita increase was greatest in the developing world.[4]

Consequently, we have developed a system wherein cell phones are used to access a centrally placed NxKM knowledge base and diagnostic engine implemented on a PC. We are now testing the use of this system with HEWs in rural India. In addition to providing recommendations for medical care to the HEWs, the phone/

[2] www.wirelessintelligence.com

[3] www.itu.int

[4] www.itu.int/ITU-D/ict/papers/2009/7.1%20teltscher_IDI%20India%202009.pdf

central-PC solution can be used to create portable personal health records. One of our partner organizations, School Health Annual Report Programme (SHARP), will use it to screen more than 10 million Indian schoolchildren in 2009, creating a unique virtual personal health record for each child.

Another advantage of this approach is that the data collected by this system can be used to improve the NxKM knowledge base. For example, as mentioned above, information about region-specific disease prevalence is important for accurate medical diagnosis. Especially important is time-critical information about the outbreak of a disease in a particular location. As the clinical application is used, validated disease cases, including those corresponding to a new outbreak, are immediately available to NxKM. In addition, individual diagnoses can be monitored centrally. If the uploaded findings of an individual patient are found to yield a low-confidence diagnosis, the patient can be identified for follow-up.

THE USER INTERFACE

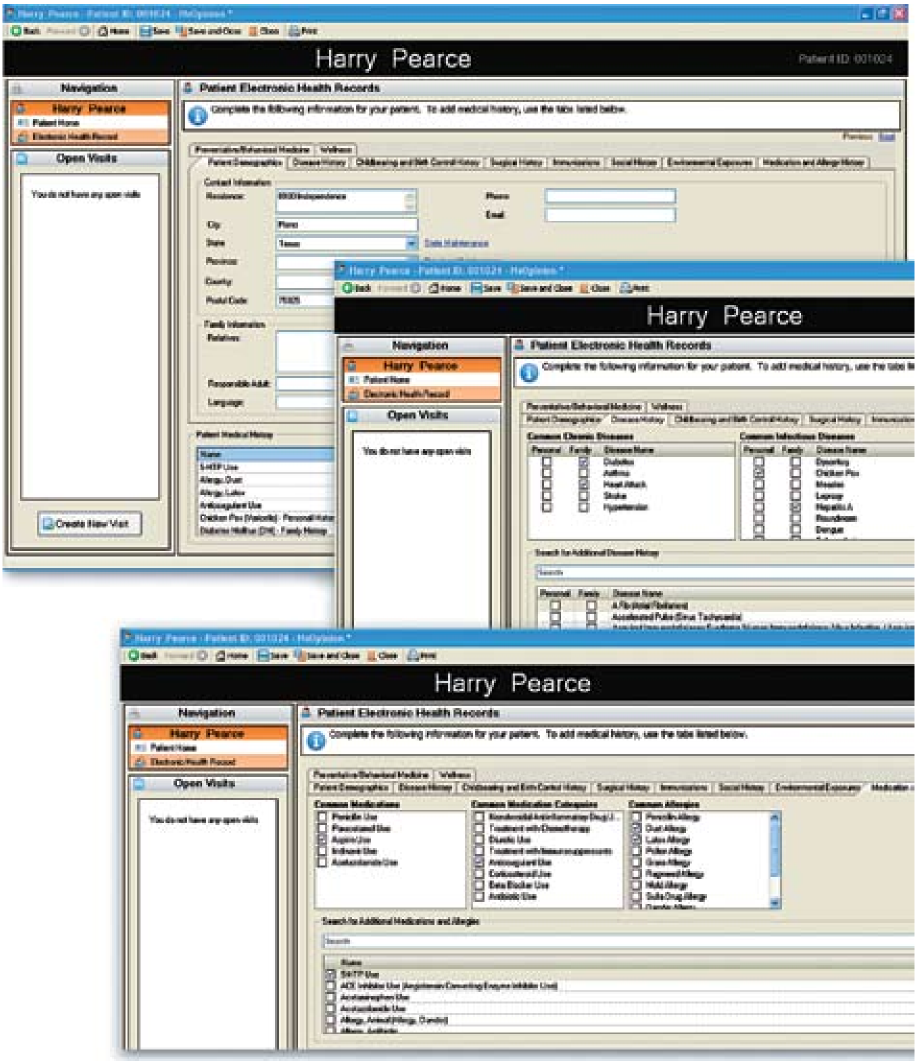
A challenge with cellular technology is the highly constrained user interface and the difficulty of entering data using a relatively small screen and keypad. Our system simplifies the process in a number of ways. First, findings that are common for a single location (e.g., facts about a given village) are prepopulated into the system. Also, as mentioned above, the system is capable of generating questions—specifically, simple multiple-choice questions—after only basic information such as the chief complaint has been entered. In addition, questions can be tailored to the organization, location, or skill level of the HEW user.

It is also important that the user interface be independent of the specific device hardware because users often switch between phones of different designs. Our interface application sits on top of a middle-layer platform that we have implemented for multiple devices.

In addition to simple input, the interface allows easy access to important bits of information. For example, it provides a daily summary of patients needing care, including their diagnosis, village location, and previous caregivers.

DATA-SHARING SOLUTIONS

Even beyond traditional legacy data silos (such as EPIC and CERNER) [5], barriers to sharing critical public health data still exist—including concerns about privacy and sovereignty. Data availability can also be limited regionally (e.g., in India and South Africa), by organizations (e.g., the World Health Organization,



Harry Pearce - Patient ID: 001024 - NxOpinion *

Back Forward Home Save Save and Close Close Print

Harry Pearce

Patient ID: 001024

Navigation

Harry Pearce

Patient Home

Electronic Health Record

Open Visits

You do not have any open visits

Create New Visit

Patient Electronic Health Records

Complete the following information for your patient. To add medical history, use the tabs listed below.

Preventative/Behavioral Medicine | Wellness

Patient Demographics | Disease History | Childbearing and Birth Control History | Surgical History | Immunizations | Social History | Environmental Exposures | Medication and Allergy History

Contact Information

Residence: 8900 Independence

City: Plano

State: Texas

Province:

County:

Postal Code: 75025

Phone

Email

Family Information

Relatives:

Responsible Adult:

Language:

Patient Medical History

Name

5-HTP Use

Allergy, Dust

Allergy, Latex

Anticoagulant Use

Chicken Pox (Varicella) - Personal History

Diabetes Mellitus (DM) - Family History

Common Chronic Diseases

Personal	Family	Disease Name
□	☑	Diabetes
□	□	Asthma
□	☑	Heart Attack
□	□	Stroke
□	□	Hypertension

Common Infectious Diseases

Personal	Family	Disease Name
□	□	Dysentery
☑	□	Chicken Pox
□	□	Measles
□	□	Leprosy
□	☑	Hepatitis A
□	□	Roundworm
□	□	Dengue

Search for Additional Disease History

Search

Personal	Family	Disease Name
□	□	A-Fib (Atrial Fibrillation)
□	□	Accelerated Pulse (Sinus Tachycardia)

Common Medications

□ Penicillin Use
□ Paracetamol Use
☑ Aspirin Use
□ Indinavir Use
□ Acetazolamide Use

Common Medication Categories

□ Nonsteroidal Anti-Inflammatory Drug (...
□ Treatment with Chemotherapy
□ Diuretic Use
□ Treatment with Immunosuppressants
☑ Anticoagulant Use
□ Corticosteroid Use
□ Beta Blocker Use
□ Antibiotic Use

Common Allergies

□ Penicillin Allergy
☑ Dust Allergy
☑ Latex Allergy
□ Pollen Allergy
□ Grass Allergy
□ Ragweed Allergy
□ Mold Allergy
□ Sulfa Drug Allergy

Search for Additional Medications and Allergies

Search

Name

☑ 5-HTP Use
□ ACE Inhibitor Use (Angiotensin-Converting Enzyme Inhibitor Use)
□ Acetaminophen Use
□ Acetazolamide Use
□ Allergy, Animal (Allergy, Dander)

NxOpinion's innovative approach, which shows data when you want it, how you want it, and where you want it, using artificial intelligence.

World Vision, or pharmaceutical companies), or by providers (e.g., insurance companies and medical provider groups). Significant public health value resides in each of these datasets, and efforts should be made to overcome the barriers to gathering data into shared, de-identified global databases. Such public datasets, while useful on their own, also add significant value to proprietary datasets, providing valuable generic context to proprietary information.

NxKM imports, manages, and exports data via *publish sets*. These processes allow various interest groups (governments, public health organizations, primary care providers, small hospitals, laboratory and specialty services, and insurance providers) to share the same interactive de-identified (privacy-preserving) global database while maintaining control of proprietary and protected data.

LOOKING FORWARD

Several challenges remain. While better educated HEWs are able to use these data collection and diagnostic decision support tools readily, other HEWs, such as Accredited Social Health Activists (ASHAs) and other front-line village workers, are often illiterate or speak only a local dialect. We are exploring two potential solutions—one that uses voice recognition technology and another that allows a user to answer multiple-choice questions via the cell phone's numeric keypad. Voice recognition technology provides added flexibility in input, but—at least so far—it requires the voice recognizer to be trained by each user.

Another challenge is unique and reproducible patient identification—verification that the subject receiving treatment is actually the correct patient—when there is no standard identification system for most under-served populations. Voice recognition combined with face recognition and newer methods of biometrics, along with a corroborating GPS location, can help ensure that the patient who needs the care is the one actually receiving treatment.

Another barrier is data integrity. For example, most rural individuals will report diagnoses that have not been substantiated by qualified medical personnel and could be erroneous. We have attempted to mitigate this issue by using an inference engine that allows for down-weighting of unsubstantiated evidence.

Deploying systems that work anywhere in the world can lead to the creation of a massive amount of patient information. Storing, reconciling, and then accessing that information in the field, all while maintaining appropriate privacy and security, are exceptionally challenging when patient numbers are in the millions (instead of tens of thousands, as with most current electronic health record systems). Further, feeding verified data on this scale back into the system to improve its predictive capability while maintaining the ability to analyze and retrieve specific segments (data mine) remains difficult.

A final, and perhaps the greatest, obstacle is that of cooperation. If organizations, governments, and companies are willing to share a de-identified global database while protecting and owning their own database, medical science and healthcare can benefit tremendously. A unified database that allows integration across many monitoring and evaluation systems and databases should help in quickly and efficiently identifying drug resistance or outbreaks of disease and in monitoring the effectiveness of treatments and healthcare interventions. The global database should support data queries that guard against the identification of individuals and yet provide sufficient information for statistical analyses and validation. Such technology is beginning to emerge (e.g., [6]), but the daunting challenge of finding a system of rewards that encourages such cooperation remains.

SUMMARY

We have developed and are beginning to deploy a system for the acquisition, analysis, and transmission of medical knowledge and data in developing countries. The system includes a centralized component based on PC technology that houses medical knowledge and data and has real-time diagnostic capabilities, complemented by a cell phone–based interface for medical workers in the field. We believe that such a system will lead to improved medical care in developing countries through improved diagnoses, the collection of more accurate and timely data across more individuals, and the improved dissemination of accurate and timely medical knowledge and information.

When we stop and think about how a world of connected personal health records can be used to improve medicine, we can see that the potential impact is staggering. By knowing virtually every individual who exists, the diseases affecting that person, and where he or she is located; by improving data integrity; and by collecting the data in a central location, we can revolutionize medicine and perhaps even eradicate more diseases. This global system can monitor the effects of various humanitarian efforts and thereby justify and tailor efforts, medications, and resources to specific areas. It is our hope that a system that can offer high-quality diagnoses as well as collect and rapidly disseminate valid data will save millions of lives. Alerts and responses can become virtually instantaneous and can thus lead to the identification of drug resistance, outbreaks, and effective treatments in a fraction of the time it takes now. The potential for empowering caregivers in developing countries though a global diagnostic and database system is enormous.

REFERENCES

[1] S. S. Lim, D. B. Stein, A. Charrow, and C. J. L. Murray, "Tracking progress towards universal childhood immunisation and the impact of global initiatives: a systematic analysis of three-dose diphtheria, tetanus, and pertussis immunisation coverage," *Lancet,* vol. 372, pp. 2031–2046, 2008, doi: 10.1016/S0140-6736(08)61869-3.

[2] *The Millennium Development Goals Report.* United Nations, 2008.

[3] R. A. Miller, M. A. McNeil, S. M. Challinor, F. E. Masarie, Jr., and J. D. Myers, "The Internist-1/Quick Medical Reference Project—Status Report," *West. J. Med.* vol. 145, pp. 816–822, 1986.

[4] D. Heckerman. *Probabilistic Similarity Networks.* Cambridge, MA: MIT Press, 1991.

[5] L. Kohn, J. Corrigan, and M. Donaldson, Eds. *To Err Is Human: Building a Safer Health System.* Washington, D.C.: National Academies Press, 2000.

[6] C. Dwork and K. Nissim, "Privacy-Preserving Datamining on Vertically Partitioned Databases," *Proc. CRYPTO,* 2004, doi: 10.1.1.86.8559.

Discovering the Wiring Diagram of the Brain

JEFF W. LICHTMAN
R. CLAY REID
HANSPETER PFISTER
Harvard University

MICHAEL F. COHEN
Microsoft Research

THE BRAIN, THE SEAT OF OUR COGNITIVE ABILITIES, is perhaps the most complex puzzle in all of biology. Every second in the human brain, billions of cortical nerve cells transmit billions of messages and perform extraordinarily complex computations. How the brain works—how its function follows from its structure—remains a mystery.

The brain's vast numbers of nerve cells are interconnected at synapses in circuits of unimaginable complexity. It is largely assumed that the specificity of these interconnections underlies our ability to perceive and classify objects, our behaviors both learned (such as playing the piano) and intrinsic (such as walking), and our memories—not to mention controlling lower-level functions such as maintaining posture and even breathing. At the highest level, our emotions, our sense of self, our very consciousness are entirely the result of activities in the nervous system.

At a macro level, human brains have been mapped into regions that can be roughly associated with specific types of activities. However, even this building-block approach is fraught with complexity because often many parts of the brain participate in completing a task. This complexity arises especially because most behaviors begin with sensory input and are followed by analysis, decision making, and finally a motor output or action.

At the microscopic level, the brain comprises billions of neurons, each connected to other neurons by up to several thousand synaptic connections. Although the existence of these synaptic circuits has been appreciated for over a century, we have no detailed circuit diagrams of the brains of humans or any other mammals. Indeed, neural circuit mapping has been attempted only once, and that was two decades ago on a small worm with only 300 nerve cells. The central stumbling block is the enormous technical difficulty associated with such mapping. Recent technological breakthroughs in imaging, computer science, and molecular biology, however, allow a reconsideration of this problem. But even if we had a wiring diagram, we would need to know what messages the neurons in the circuit are passing—not unlike listening to the signals on a computer chip. This represents the second impediment to understanding: traditional physiological methods let us listen to only a tiny fraction of the nerves in the circuit.

To get a sense of the scale of the problem, consider the cerebral cortex of the human brain, which contains more than 160 trillion synaptic connections. These connections originate from billions of neurons. Each neuron receives synaptic connections from hundreds or even thousands of different neurons, and each sends information via synapses to a similar number of target neurons. This enormous fan-in and fan-out can occur because each neuron is geometrically complicated, possessing many receptive processes (dendrites) and one highly branched outflow process (an axon) that can extend over relatively long distances.

One might hope to be able to reverse engineer the circuits in the brain. In other words, if we could only tease apart the individual neurons and see which one is connected to which and with what strength, we might at least begin to have the tools to decode the functioning of a particular circuit. The staggering numbers and complex cellular shapes are not the only daunting aspects of the problem. The circuits that connect nerve cells are nanoscopic in scale. The density of synapses in the cerebral cortex is approximately 300 million per cubic millimeter.

Functional magnetic resonance imaging (fMRI) has provided glimpses into the macroscopic 3-D workings of the brain. However, the finest resolution of fMRI is approximately 1 cubic millimeter per voxel—the same cubic millimeter that can contain 300 million synapses. Thus there is a huge amount of circuitry in even the most finely resolved functional images of the human brain. Moreover, the size of these synapses falls below the diffraction-limited resolution of traditional optical imaging technologies.

Circuit mapping could potentially be amenable to analysis based on color coding of neuronal processes [1] and/or the use of techniques that break through the diffraction limit [2]. Presently, the gold standard for analyzing synaptic connections is to use electron microscopy (EM), whose nanometer (nm) resolution is more than sufficient to ascertain the finest details of neural connections. But to map circuits, one must overcome a technical hurdle: EM typically images very thin sections (tens of nanometers in thickness), so reconstructing a volume requires a "serial reconstruction" whereby the image information from contiguous slices of the same volume is recomposed into a volumetric dataset. There are several ways to generate such volumetric data (see, for example, [3-5]), but all of these have the potential to generate astonishingly large digital image data libraries, as described next.

SOME NUMBERS

If one were to reconstruct by EM all the synaptic circuitry in 1 cubic mm of brain (roughly what might fit on the head of a pin), one would need a set of serial images spanning a millimeter in depth. Unambiguously resolving all the axonal and dendritic branches would require sectioning at probably no more than 30 nm. Thus the 1 mm depth would require 33,000 images. Each image should have at least 10 nm lateral resolution to discern all the vesicles (the source of the neurotransmitters) and synapse types. A square-millimeter image at 5 nm resolution is an image that has ~4 x10^{10} pixels, or 10 to 20 gigapixels. So the image data in 1 cubic mm will be in the range of 1 petabyte (2^{50} ~ 1,000,000,000,000,000 bytes). The human brain contains nearly 1 million cubic mm of neural tissue.

SOME SUCCESSES TO DATE

Given this daunting task, one is tempted to give up and find a simpler problem. However, new technologies and techniques provide glimmers of hope. We are pursuing these with the ultimate goal of creating a "connectome"—a complete circuit diagram of the brain. This goal will require intensive and large-scale collaborations among biologists, engineers, and computer scientists.

Three years ago, the Reid and Lichtman labs began working on ways to automate and accelerate large-scale serial-section EM. Focusing specifically on large cortical volumes at high resolution, the Reid group has concentrated on very high throughput as well as highly automated processes. So far, their work has been published only in abstract form [3], but they are confident about soon having the first 10 terabytes of volumetric data on fine-scale brain anatomy. Physiological experiments can now show the function of virtually every neuron in a 300 μm cube. The new EM data has the resolution to show virtually every axon, dendrite, and synapse—the physical connections that underlie neuronal function.

The problem of separating and tracking the individual neurons within the volume remains. However, some successes have already been achieved using exotic means. Lichtman's lab found a way to express various combinations of red, green, and blue fluorescent proteins in genetically engineered mice. These random combinations presently provide about 90 colors or combinations of colors [1]. With this approach, it is possible to track individual neurons as they branch to their eventual synaptic connections to other neurons or to the end-organs in muscle. The multicolor labeled nerves (dubbed "brainbow"), shown in Figure 1, are reminiscent of the rainbow cables in computers and serve the same purpose: to disambiguate wires traveling over long distances.

Because these colored labels are present in the living mouse, it is possible to track synaptic wiring changes by observing the same sites multiple times over minutes, days, or even months.

Reid's lab has been able to stain neurons of rat and cat visual cortices such that they "light up" when activated. By stimulating the cat with lines of different orientations, they have literally been able to see which neurons are firing, depending on the specific visual stimulus. By comparing the organization of the rat's visual cortex to that of the cat, they have found that while a rat's neurons appear to be randomly organized based on the orientation of the visual stimulus, a cat's neurons exhibit remarkable structure. (See Figure 2.)

Achieving the finest resolution using EM requires imaging very thin slices of neural tissue. One method begins with a block of tissue; after each imaging pass, a

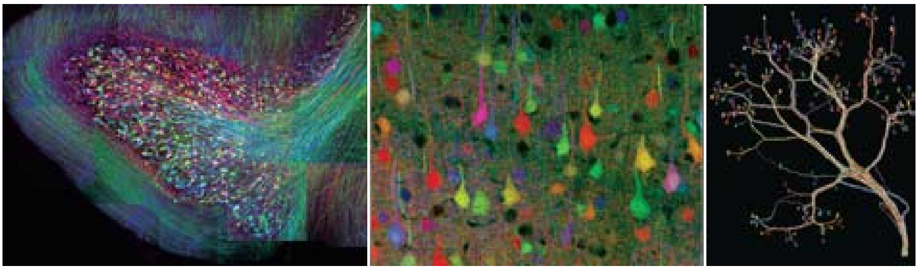


FIGURE 1.
Brainbow images showing individual neurons fluorescing in different colors. By tracking the neurons through stacks of slices, we can follow each neuron's complex branching structure to create the treelike structures in the image on the right.

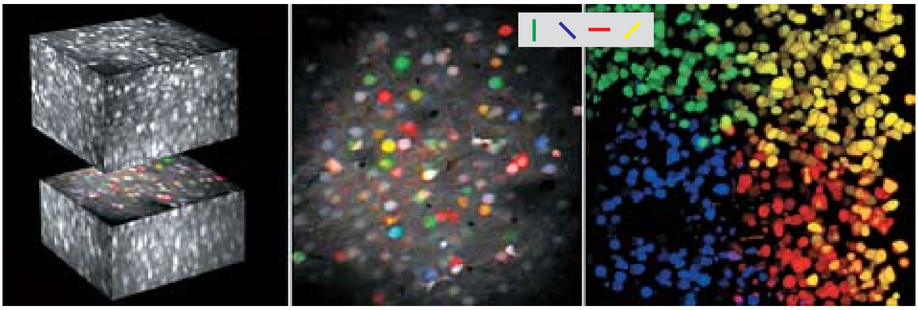


FIGURE 2.
Neurons in a visual cortex stained in vivo *with a calcium-sensitive dye. Left: A 3-D reconstruction of thousands of neurons in a rat visual cortex, obtained from a stack of images (300 µm on a side). The neurons are color coded according to the orientation of the visual stimulus that most excited them. Center: A 2-D image of the plane of section from the left panel. Neurons that responded to different stimulus orientations (different colors) are arranged seemingly randomly in the cortex. Inset: Color coding of stimulus orientations. Right: By comparison, the cat visual cortex is extremely ordered. Neurons that responded preferentially to different stimulus orientations are segregated with extraordinary precision. This image represents a complete 3-D functional map of over 1,000 neurons in a 300x300x200 µm volume in the visual cortex [6, 7].*

thin slice is removed (and destroyed) from the block, and then the process is repeated. Researchers in the Lichtman group at Harvard have developed a new device—a sort of high-tech lathe that they are calling an Automatic Tape-Collecting Lathe Ultramicrotome (ATLUM)—that can allow efficient nanoscale imaging over large tissue volumes. (See Figure 3 on the next page.)

The ATLUM [3] automatically sections an embedded block of brain tissue into thousands of ultrathin sections and collects these on a long carbon-coated tape for later staining and imaging in a scanning electron microscope (SEM). Because the process is fully automated, volumes as large as tens of cubic millimeters—large enough to span entire multi-region neuronal circuits—can be quickly and reliably reduced to a tape of ultrathin sections. SEM images of these ATLUM-collected sections can attain lateral resolutions of 5 nm or better—sufficient to image individual synaptic vesicles and to identify and trace all circuit connectivity.

The thin slices are images of one small region at a time. Once a series of individual images is obtained, these images must be stitched together into very large images

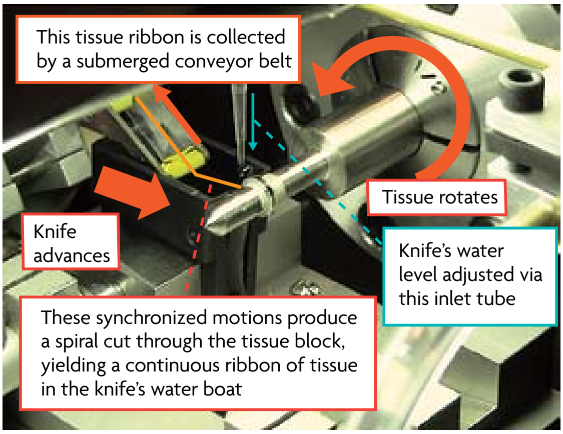

This tissue ribbon is collected by a submerged conveyor belt
Tissue rotates
Knife advances
Knife's water level adjusted via this inlet tube
These synchronized motions produce a spiral cut through the tissue block, yielding a continuous ribbon of tissue in the knife's water boat

FIGURE 3.
The Automatic Tape-Collecting Lathe Ultramicrotome (ATLUM), which can allow efficient nanoscale imaging over large tissue volumes.

and possibly stacked into volumes. At Microsoft Research, work has proceeded to stitch together and then interactively view images containing billions of pixels.[1] Once these gigapixel-size images are organized into a hierarchical pyramid, the HD View application can stream requested imagery over the Web for viewing.[2] This allows exploration of both large-scale and very fine-scale features. Figure 4 shows a walkthrough of the result.

Once the images are captured and stitched, multiple slices of a sample must be stacked to assemble them into a coherent volume. Perhaps the most difficult task at that point is extracting the individual strands of neurons. Work is under way at Harvard to provide interactive tools to aid in outlining individual "processes" and then tracking them slice to slice to pull out each dendritic and axonal fiber [8, 9]. (See Figure 5.) Synaptic interfaces are perhaps even harder to find automatically; however, advances in both user interfaces and computer vision give hope that the whole process can be made tractable.

Decoding the complete connectome of the human brain is one of the great challenges of the 21st century. Advances at both the biological level and technical level are certain to lead to new successes and discoveries, and they will hopefully help answer fundamental questions about how our brain performs the miracle of thought.

[1] http://research.microsoft.com/en-us/um/redmond/groups/ivm/ICE
[2] http://research.microsoft.com/en-us/um/redmond/groups/ivm/HDView

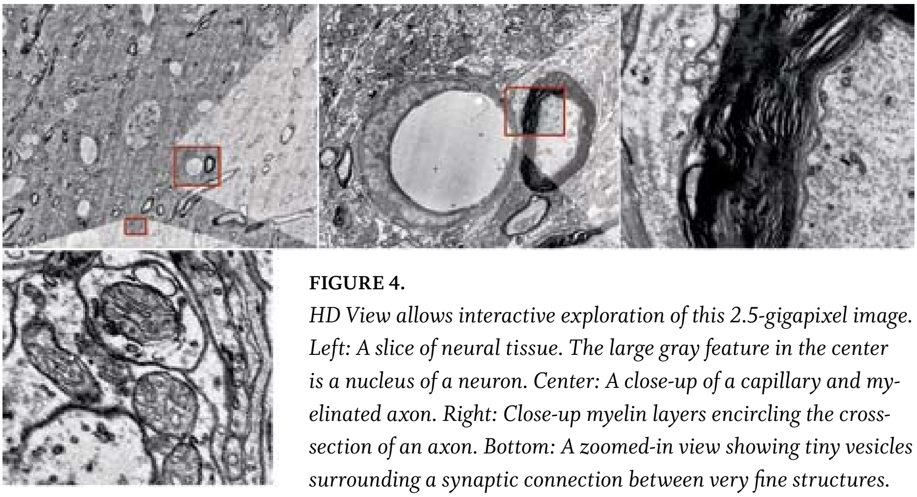


FIGURE 4.
HD View allows interactive exploration of this 2.5-gigapixel image. Left: A slice of neural tissue. The large gray feature in the center is a nucleus of a neuron. Center: A close-up of a capillary and myelinated axon. Right: Close-up myelin layers encircling the cross-section of an axon. Bottom: A zoomed-in view showing tiny vesicles surrounding a synaptic connection between very fine structures.

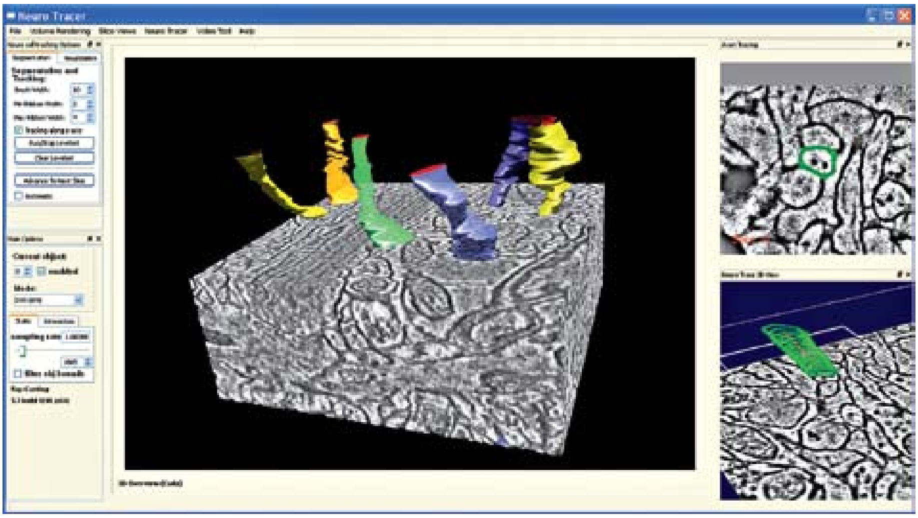


FIGURE 5.
NeuroTrace allows neuroscientists to interactively explore and segment neural processes in high-resolution EM data.

REFERENCES

[1] J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman, “Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system,” *Nature*, vol. 450, pp. 56–62, 2007, doi: 10.1038/nature06293.

[2] S. Hell, “Microscopy and its focal switch,” *Nature Methods*, vol. 6, pp. 24–32, 2009, doi: 10.1038/NMeth.1291.

[3] D. Bock, W. C. Lee, A. Kerlin, M. L. Andermann, E. Soucy, S. Yurgenson, and R. C. Reid, “High-throughput serial section electron microscopy in mouse primary visual cortex following in vivo two-photon calcium imaging,” *Soc. Neurosci. Abstr.*, vol. 769, no. 12, 2008.

[4] W. Denk and H. Horstmann, “Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure,” *PLoS Biol.*, vol. 2, p. e329, 2004, doi: 10.1017/S1431927606066268.

[5] K. J. Hayworth, N. Kasthuri, R. Schalek, and J. W. Lichtman, “Automating the Collection of Ultrathin Serial Sections for Large Volume TEM Reconstructions,” *Microsc. Microanal.*, vol. 12, pp. 86–87, 2006.

[6] K. Ohki, S. Chung, Y. H. Ch’ng, P. Kara, and R. C. Reid, “Functional imaging with cellular resolution reveals precise microarchitecture in visual cortex,” *Nature*, vol. 433, pp. 597–603, 2005, doi:10.1038/nature03274.

[7] K. Ohki, S. Chung, P. Kara, M. Hübener, T. Bonhoeffer, and R. C. Reid, “Highly ordered arrangement of single neurons in orientation pinwheels,” *Nature*, vol. 442, pp. 925–928, 2006, doi:10.1038/nature05019.

[8] W. Jeong, J. Beyer, M. Hadwiger, A. Vazquez, H. Pfister, and R. Whitaker, “Scalable and Interactive Segmentation and Visualization of Neural Processes in EM Datasets,” *IEEE Trans. Visual. Comput. Graphics*, Oct. 2009.

[9] A. Vazquez, E. Miller, and H. Pfister, “Multiphase Geometric Couplings for the Segmentation of Neural Processes,” *Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, June 2009.

Toward a Computational Microscope for Neurobiology

ERIC HORVITZ
Microsoft Research

WILLIAM KRISTAN
University of California,
San Diego

Although great strides have been made in neurobiology, we do not yet understand how the symphony of communication among neurons leads to rich, competent behaviors in animals. How do local interactions among neurons coalesce into the behavioral dynamics of nervous systems, giving animals their impressive abilities to sense, learn, decide, and act in the world? Many details remain cloaked in mystery. We are excited about the promise of gaining new insights by applying computational methods, in particular machine learning and inference procedures, to generate explanatory models from data about the activities of populations of neurons.

NEW TOOLS FOR NEUROBIOLOGISTS

For most of the history of electrophysiology, neurobiologists have monitored the membrane properties of neurons of vertebrates and invertebrates by using glass micropipettes filled with a conducting solution. Mastering techniques that would impress the most expert of watchmakers, neuroscientists have fabricated glass electrodes with tips that are often less than a micron in diameter, and they have employed special machinery to punch the tips into the cell bodies of single neurons—with the hope that the neurons will function as they normally do within larger assemblies. Such an approach has provided data about the membrane voltages and action potentials of a single cell or just a handful of cells.

However, the relationship between neurobiologists and data about nervous systems is changing. New recording machinery is making data available on the activity of large populations of neurons. Such data makes computational procedures increasingly critical as experimental tools for unlocking new understanding about the connections, architecture, and overall machinery of nervous systems.

New opportunities for experimentation and modeling on a wider scale have become available with the advent of fast optical imaging methods. With this approach, dyes and photomultipliers are used to track calcium levels and membrane potentials of neurons, with high spatial and temporal resolution. These high-fidelity optical recordings allow neurobiologists to examine the simultaneous activity of populations of tens to thousands of neurons. In a relatively short time, data available about the activity of neurons has grown from a trickle of information gleaned via sampling of small numbers of neurons to large-scale observations of neuronal activity.

Spatiotemporal datasets on the behaviors of populations of neurons pose tantalizing inferential challenges and opportunities. The next wave of insights about the neurophysiological basis for cognition will likely come via the application of new kinds of computational lenses that direct an information-theoretic "optics" onto streams of spatiotemporal population data.

We foresee that neurobiologists studying populations of neurons will one day rely on tools that serve as *computational microscopes*—systems that harness machine learning, reasoning, and visualization to help neuroscientists formulate and test hypotheses from data. Inferences derived from the spatiotemporal data streaming from a preparation might even be overlaid on top of traditional optical views during experiments, augmenting those views with annotations that can help with the direction of the investigation.

Intensive computational analyses will serve as the basis for modeling and visualization of the intrinsically high-dimensional population data, where multiple neuronal units interact and contribute to the activity of other neurons and assemblies, and where interactions are potentially context sensitive—circuits and flows might exist dynamically, transiently, and even simultaneously on the same neuronal substrate.

COMPUTATION AND COMPLEXITY

We see numerous opportunities ahead for harnessing fast-paced computations to assist neurobiologists with the science of making inferences from neuron population data. Statistical analyses have already been harnessed in studies of populations of neurons. For example, statistical methods have been used to identify and characterize neuronal activity as trajectories in large dynamical state spaces [1]. We are excited about employing richer machine learning and reasoning to induce explanatory models from case libraries of neuron population data. Computational procedures for induction can assist scientists with teasing insights from raw data on neuronal activity by searching over large sets of alternatives and weighing the plausibility of different explanatory models. The computational methods can be tasked with working at multiple levels of detail, extending upward from circuit-centric exploration of local connectivity and functionality of neurons to potentially valuable higher-level abstractions of neuronal populations—abstractions that may provide us with simplifying representations of the workings of nervous systems.

Beyond generating explanations from observations, inferential models can be harnessed to compute the *expected value of information*, helping neuroscientists to identify the best next test to perform or information to gather, in light of current goals and uncertainties. Computing the value of information can help to direct interventional studies, such as guidance on stimulating specific units, clamping the voltage of particular cells, or performing selective modification of cellular activity via agonist and antagonist pharmacological agents.

We believe that there is promise in both automated and interactive systems, including systems that are used in real-time settings as bench tools. Computational tools might one day even provide real-time guidance for probes and interventions via visualizations and recommendations that are dynamically generated during imaging studies.

Moving beyond the study of specific animal systems, computational tools for analyzing neuron population data will likely be valuable in studies of the construction of nervous systems during embryogenesis, as well as in comparing nervous systems of different species of animals. Such studies can reveal the changes in circuitry and function during development and via the pressures of evolutionary adaptation.

SPECTRUM OF SOPHISTICATION

Neurobiologists study nervous systems of invertebrates and vertebrates across a spectrum of complexity. Human brains are composed of about 100 billion neurons that interact with one another via an estimated 100 trillion synapses. In contrast, the brain of the nematode, *Caenorhabditis elegans (C. elegans)*, has just 302 neurons. Such invertebrate nervous systems offer us an opportunity to learn about the principles of neuronal systems, which can be generalized to more complex systems, including our own. For example, *C. elegans* has been a model system for research on the structure of neuronal circuits; great progress has been achieved in mapping the precise connections among its neurons.

Many neurobiologists choose to study simpler nervous systems even if they are motivated by questions about the neurobiological nature of human intelligence. Nervous systems are derived from a family tree of refinements and modifications, so it is likely that key aspects of neuronal information processing have been conserved across brains of a range of complexities. While new abstractions, layers, and interactions may have evolved in more complex nervous systems, brains of different complexities likely rely on a similar neuronal fabric—and there is much that we do not know about that fabric.

In work with our colleagues Ashish Kapoor, Erick Chastain, Johnson Apacible, Daniel Wagenaar, and Paxon Frady, we have been pursuing the use of machine learning, reasoning, and visualization to understand the machinery underlying decision making in *Hirudo*, the European medicinal leech. We have been applying computational analyses to make inferences from optical data about the activity of populations of neurons within the segmental ganglia of *Hirudo*. The ganglia are composed of about 400 neurons, and optical imaging reveals the activity of approximately 200 neurons at a time—all the neurons on one side of the ganglion. Several frames of the optical imaging of *Hirudo* are displayed in Figure 1. The brightness

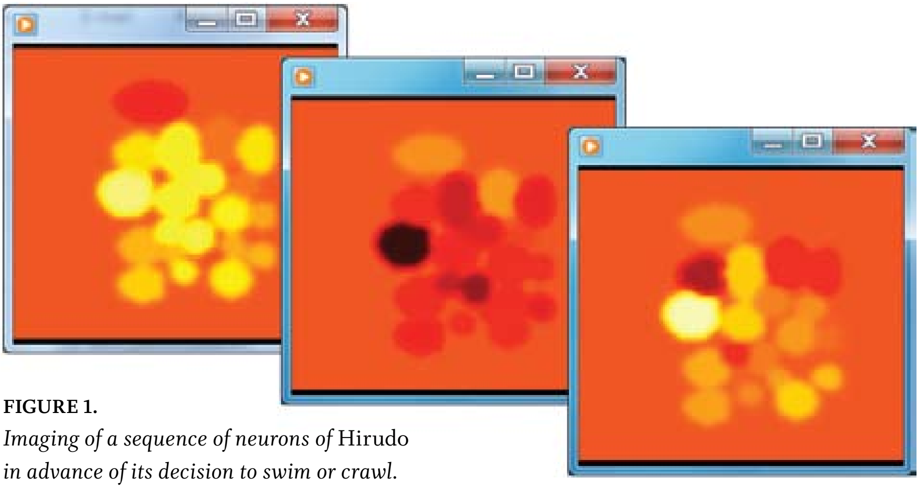


FIGURE 1.
Imaging of a sequence of neurons of Hirudo *in advance of its decision to swim or crawl.*

of each of the imaged neurons represents the level of depolarization of the cells, which underlies the production of action potentials.

We are developing analyses and assembling tools in pursuit of our vision of developing computational microscopes for understanding the activity of neuronal populations and their relationship to behavior. In one approach, we generate graphical probabilistic temporal models that can predict the forthcoming behavior of *Hirudo* from a short window of analysis of population data. The models are generated by searching over large spaces of feasible models in which neurons, and abstractions of neurons, serve as random variables and in which temporal and atemporal dependencies are inferred among the variables. The methods can reveal modules of neurons that appear to operate together and that can appear dynamically over the course of activity leading up to decisions by the animal. In complementary work, we are considering the role of neuronal states in defining trajectories through state spaces of a dynamical system.

EMERGENCE OF A COMPUTATIONAL MICROSCOPE

We have started to build interactive viewers and tools that allow scientists to manipulate inferential assumptions and parameters and to inspect implications visually. For example, sliders allow for smooth changes in thresholds for admitting connections among neurons and for probing strengths of relationships and membership in modules. We would love to see a world in which such tools are shared broadly among neuroscientists and are extended with learning, inference, and visualization components developed by the neuroscience community.

Figure 2 on the next page shows a screenshot of a prototype tool we call the MSR Computational Microscope, which was developed by Ashish Kapoor, Erick Chastain, and Eric Horvitz at Microsoft Research as part of a broader collaboration with William Kristan at the University of California, San Diego, and Daniel Wagenaar at California Institute of Technology. The tool allows users to visualize neuronal activity over a period of time and then explore inferences about relationships among neurons in an interactive manner. Users can select from a variety of inferential methods and specify modeling assumptions. They can also mark particular neurons and neuronal subsets as focal points of analyses. The view in Figure 2 shows an analysis of the activity of neurons in the segmental ganglia of *Hirudo*. Inferred informational relationships among cells are displayed via highlighting of neurons and through the generation of arcs among neurons. Such inferences can help to guide exploration and confirmation of physical connections among neurons.

FIGURE 2.
Possible connections and clusters inferred from population data during imaging of Hirudo.

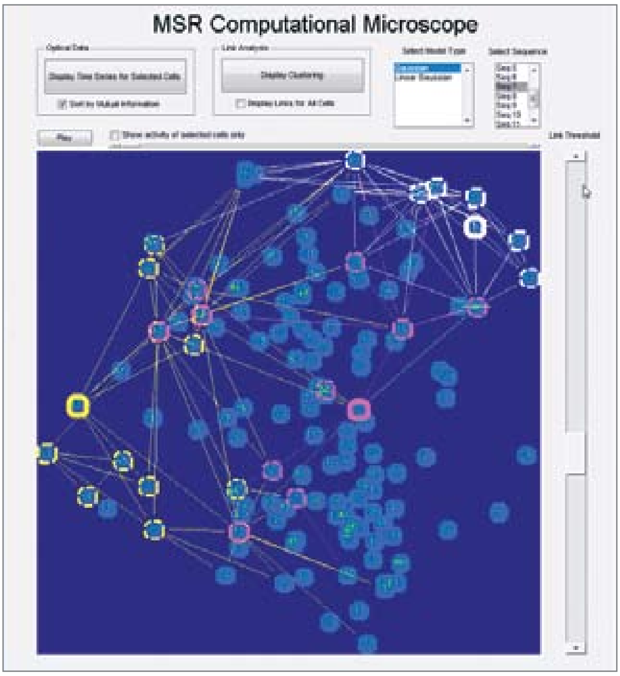

MSR Computational Microscope
Optical Data
Display Time Series for Selected Cells
☑ Sort by Mutual Information
Link Analysis
Display Clustering
□ Display Links for All Cells
Select Model Type
Gaussian
Linear Gaussian
Select Sequence
Seq 6
Seq 8
Seq 7
Seq 8
Seq 9
Seq 10
Seq 11
Play
□ Show activity of selected cells only
Link Threshold

FIGURE 3.
Inferred informational relationships among neurons in a segmental ganglion of Hirudo. *Measures of similarity of the dynamics of neuronal activity over time are displayed via arcs and clusters.*

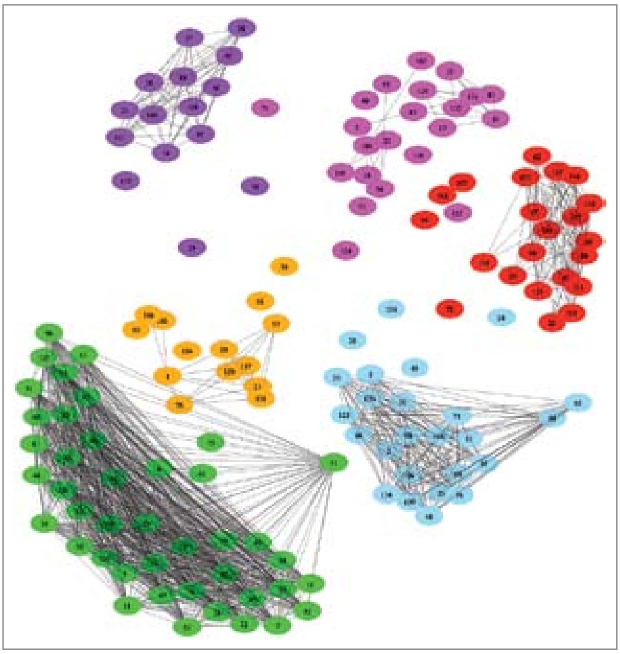


Figure 3 shows another informational analysis that spatially clusters cells that behave in a similar manner in the ganglia of *Hirudo* over a set of trials. The analysis provides an early vision of how information-theoretic analyses might one day help neurobiologists to discover and probe interactions within and between neuronal subsystems.

We are only at the start of this promising research direction, but we expect to see a blossoming of analyses, tools, and a broader sub-discipline that focuses on the neuroinformatics of populations of neurons. We believe that computational methods will lead us to effective representations and languages for understanding neuronal systems and that they will become essential tools for neurobiologists to gain insight into the myriad mysteries of sensing, learning, and decision making by nervous systems.

REFERENCES

[1] K. L. Briggman, H. D. I. Abarbanel, and W. B. Kristan, Jr., “Optical imaging of neuronal populations during decision-making,” *Science*, vol. 307, pp. 896–901, 2005, doi: 10.1126/science.110.

A Unified Modeling Approach to Data-Intensive Healthcare

IAIN BUCHAN
University of Manchester

JOHN WINN
CHRIS BISHOP
Microsoft Research

THE QUANTITY OF AVAILABLE HEALTHCARE DATA is rising rapidly, far exceeding the capacity to deliver personal or public health benefits from analyzing this data [1]. Three key elements of the rise are electronic health records (EHRs), biotechnologies, and scientific outputs. We discuss these in turn below, leading to our proposal for a unified modeling approach that can take full advantage of a data-intensive environment.

ELECTRONIC HEALTH RECORDS

Healthcare organizations around the world, in both low- and high-resource settings, are deploying EHRs. At the community level, EHRs can be used to manage healthcare services, monitor the public's health, and support research. Furthermore, the social benefits of EHRs may be greater from such population-level uses than from individual care uses.

The use of standard terms and ontologies in EHRs is increasing the structure of healthcare data, but clinical coding behavior introduces new potential biases. For example, the introduction of incentives for primary care professionals to tackle particular conditions may lead to fluctuations in the amount of coding of new cases of those conditions [2]. On the other hand, the falling cost of devices for remote monitoring and near-patient testing is leading to more capture of objective measures in EHRs, which can provide less biased signals but may create the illusion of an increase in disease prevalence simply due to more data becoming available.

Some patients are beginning to access and supplement their own records or edit a parallel health record online [3]. The stewardship of future health records may indeed be more with individuals (patients/citizens/consumers) and communities (families/local populations etc.) than with healthcare organizations. In summary, the use of EHRs is producing more data-intensive healthcare environments in which substantially more data are captured and transferred digitally. Computational thinking and models of healthcare to apply to this wealth of data, however, have scarcely been developed.

BIOTECHNOLOGIES

Biotechnologies have fueled a boom in molecular medical research. Some techniques, such as genome-wide analysis, produce large volumes of data without the sampling bias that a purposive selection of study factors might produce. Such datasets are thus more wide ranging and unselected than conventional experimental measurements. Important biases can still arise from artifacts in the biotechnical processing of samples and data, but these are likely to decrease as the technologies improve. A greater concern is the systematic error that lies outside the data landscape—for example, in a metabolomic analysis that is confounded by not considering the time of day or the elapsed time from the most recent meal to when the sample was taken. The integration of different scales of data, from molecular-level to population-level variables, and different levels of directness of measurement of factors is a grand challenge for data-intensive health science. When realistically complex multi-scale models are available, the next challenge will be to make them accessible to clinicians and patients, who together can evaluate the competing risks of different options for personalizing treatment.

SCIENTIFIC OUTPUTS

The outputs of health science have been growing exponentially [4]. In 2009, a new paper is indexed in PubMed, the health science bibliographic system, on average every 2 minutes. The literature-review approach to managing health knowledge is therefore potentially overloaded. Furthermore, the translation of new knowledge into practice innovation is slow and inconsistent [5]. This adversely affects not only clinicians and patients who are making care decisions but also researchers who are reasoning about patterns and mechanisms. There is a need to combine the mining

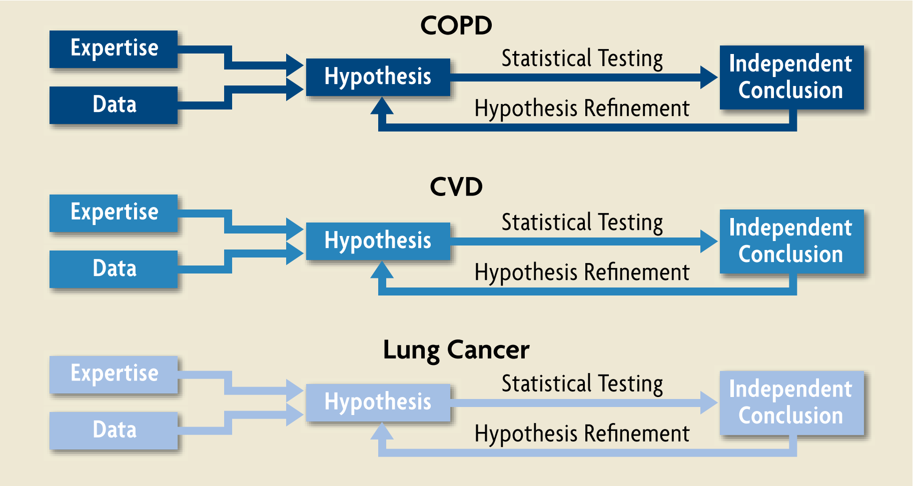


FIGURE 1.
Conventional approaches based on statistical hypothesis testing artificially decompose the healthcare domain into numerous sub-problems. They thereby miss a significant opportunity for statistical "borrowing of strength." Chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), and lung cancer can be considered together as a "big three" [6].

of evidence bases with computational models for exploring the burgeoning data from healthcare and research.

Hypothesis-driven research and reductionist approaches to causality have served health science well in identifying the major independent determinants of health and the outcomes of individual healthcare interventions. (See Figure 1.) But they do not reflect the complexity of health. For example, clinical trials exclude as many as 80 percent of the situations in which a drug might be prescribed—for example, when a patient has multiple diseases and takes multiple medications [7]. Consider a newly licensed drug released for general prescription. Clinician X might prescribe the drug while clinician Y does not, which could give rise to natural experiments. In a fully developed data-intensive healthcare system in which the data from those experiments are captured in EHRs, clinical researchers could explore the outcomes of patients on the new drug compared with natural controls, and they could potentially adjust for confounding and modifying factors. However, such adjustments might be extremely complex and beyond the capability of conventional models.

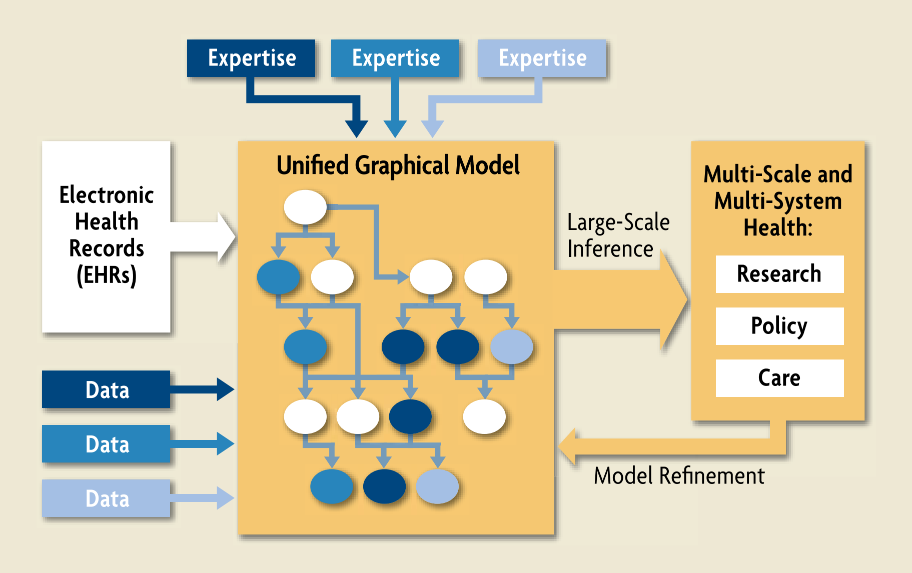


FIGURE 2.
We propose a unified approach to healthcare modeling that exploits the growing statistical resources of electronic health records in addition to the data collected for specific studies.

A UNIFIED APPROACH

We propose a unified modeling approach that can take full advantage of a data-intensive environment without losing the realistic complexity of health. (See Figure 2.) Our approach relies on developments within the machine learning field over the past 10 years, which provide powerful new tools that are well suited to this challenge. Knowledge of outcomes, interventions, and confounding or modifying factors can all be captured and represented through the framework of probabilistic graphical models in which the relevant variables, including observed data, are expressed as a graph [8]. Inferences on this graph can then be performed automatically using a variety of algorithms based on local message passing, such as [9]. Compared with classical approaches to machine learning, this new framework offers a deeper integration of domain knowledge, taken directly from experts or from the literature, with statistical learning. Furthermore, these automatic inference algorithms can scale to datasets of hundreds of millions of records, and new tools such as Infer.NET allow rapid development of solutions within this framework [10]. We illustrate the application of this approach with two scenarios.

In scenario 1, an epidemiologist is investigating the genetic and environmental factors that predispose some children to develop asthma. He runs a cohort study of 1,000 children who have been followed for 10 years, with detailed environmental and physiological measures as well as data on over half a million of the 3 million genetic factors that might vary between individuals. The conventional epidemiology approach might test predefined hypotheses using selected groups of genetic and other factors. A genome-wide scanning approach might also be taken to look for associations between individual genetic factors and simple definitions of health status (e.g., current wheeze vs. no current wheeze at age 5 years). Both of these approaches use relatively simple statistical models. An alternative machine learning approach might start with the epidemiologist constructing a graphical model of the problem space, consulting literature and colleagues to build a graph around the organizing principle—say, "peripheral airways obstruction." This model better reflects the realistic complexity of asthma with a variety of classes of wheeze and other signs and symptoms, and it relates them to known mechanisms. Unsupervised clustering methods are then used to explore how genetic, environmental, and other study factors influence the clustering into different groups of allergic sensitization with respect to skin and blood test results and reports of wheezing. The epidemiologist can relate these patterns to biological pathways, thereby shaping hypotheses to be explored further.

In scenario 2, a clinical team is auditing the care outcomes for patients with chronic angina. Subtly different treatment plans of care are common, such as different levels of investigation and treatment in primary care before referral to specialist care. A typical clinical audit approach might debate the treatment plan, consult literature, examine simple summary statistics, generate some hypotheses, and perhaps test the hypotheses using simple regression models. An alternative machine learning approach might construct a graphical model of the assumed treatment plan, via debate and reference to the literature, and compare this with discovered network topologies in datasets reflecting patient outcomes. Plausible networks might then be used to simulate the potential effects of changes to clinical practice by running scenarios that change edge weights in the underlying graphs. Thus the families of associations in locally relevant data can be combined with evidence from the literature in a scenario-planning activity that involves clinical reasoning and machine learning.

THE FOURTH PARADIGM: HEALTH AVATARS

Unified models clearly have the potential to influence personal health choices, clinical practice, and public health. So is this a paradigm for the future?

The first paradigm of healthcare information might be considered to be the case history plus expert physician, formalized by Hippocrates more than 2,000 years ago and still an important part of clinical practice. In the second paradigm, a medical record is shared among a set of complementary clinicians, each focusing their specialized knowledge on the patient's condition in turn. The third paradigm is evidence-based healthcare that links a network of health professionals with knowledge and patient records in a timely manner. This third paradigm is still in the process of being realized, particularly in regard to capturing the complexities of clinical practice in a digital record and making some aspects of healthcare computable.

We anticipate a fourth paradigm of healthcare information, mirroring that of other disciplines, whereby an individual's health data are aggregated from multiple sources and attached to a unified model of that person's health. The sources can range from body area network sensors to clinical expert oversight and interpretation, with the individual playing a much greater part than at present in building and acting on his or her health information. Incorporating all of this data, the unified model will take on the role of a "health avatar"—the electronic representation of an individual's health as directly measured or inferred by statistical models or clinicians. Clinicians interacting with a patient's avatar can achieve a more integrated view of different specialist treatment plans than they do with care records alone.

The avatar is not only a statistical tool to support diagnosis and treatment, but it is also a communication tool that links the patient and the patient's elected network of clinicians and other trusted caregivers—for what-if treatment discussions, for example. While initially acting as a fairly simple multi-system model, the health avatar could grow in depth and complexity to narrow the gap between avatar and reality. Such an avatar would not involve a molecular-level simulation of a human being (which we view as implausible) but would instead involve a unified statistical model that captures current clinical understanding as it applies to an individual patient.

This paradigm can be extended to communities, where multiple individual avatars interact with a community avatar to provide a unified model of the community's health. Such a community avatar could provide relevant and timely information for use in protecting and improving the health of those in the community. Scarce community resources could be matched more accurately to lifetime healthcare needs, particularly in prevention and early intervention, to reduce the severity and/or duration of illness and to better serve the community as a whole. Clinical, consumer, and public health services could interact more effectively, providing both social benefit and new opportunities for healthcare innovation and enterprise.

CONCLUSION

Data alone cannot lead to data-intensive healthcare. A substantial overhaul of methodology is required to address the real complexity of health, ultimately leading to dramatically improved global public healthcare standards. We believe that machine learning, coupled with a general increase in computational thinking about health, can be instrumental. There is arguably a societal duty to develop computational frameworks for seeking signals in collections of health data if the potential benefit to humanity greatly outweighs the risk. We believe it does.

REFERENCES

[1] J. Powell and I. Buchan, "Electronic health records should support clinical research," *J. Med. Internet Res.*, vol. 7, no. 1, p. e4, Mar. 14, 2005, doi: 10.2196/jmir.7.1.e4.

[2] S. de Lusignan, N. Hague, J. van Vlymen, and P. Kumarapeli, "Routinely-collected general practice data are complex, but with systematic processing can be used for quality improvement and research," *Prim. Care. Inform.*, vol. 14, no. 1, pp. 59–66, 2006.

[3] L. Bos and B. Blobel, Eds., *Medical and Care Compunetics 4*, vol. 127 in Studies in Health Technology and Informatics series. Amsterdam: IOS Press, pp. 311–315, 2007.

[4] B. G. Druss and S. C. Marcus, "Growth and decentralization of the medical literature: implications for evidence-based medicine," *J. Med. Libr. Assoc.*, vol. 93, no. 4, pp. 499–501, Oct. 2005, PMID: PMC1250328.

[5] A. Mina, R. Ramlogan, G. Tampubolon, and J. Metcalfe, "Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge," *Res. Policy*, vol. 36, no. 5, pp. 789–806, 2007, doi: 10.1016/j.respol.2006.12.007.

[6] M. Gerhardsson de Verdier, "The Big Three Concept - A Way to Tackle the Health Care Crisis?" *Proc. Am. Thorac. Soc.*, vol. 5, pp. 800–805, 2008.

[7] M. Fortin, J. Dionne, G. Pinho, J. Gignac, J. Almirall, and L. Lapointe, "Randomized controlled trials: do they have external validity for patients with multiple comorbidities?" *Ann. Fam. Med.*, vol. 4, no. 2, pp. 104–108, Mar.–Apr. 2006, doi: 10.1370/afm.516.

[8] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[9] J. Winn and C. Bishop, "Variational Message Passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, 2005.

[10] T. Minka, J. Winn, J. Guiver, and A. Kannan, Infer.NET, Microsoft Research Cambridge, http://research.microsoft.com/infernet.

Visualization in Process Algebra Models of Biological Systems

LUCA CARDELLI
Microsoft Research

CORRADO PRIAMI
Microsoft Research -
University of Trento
Centre for Computational
and Systems Biology and
University of Trento

IN A RECENT PAPER, NOBEL LAUREATE PAUL NURSE calls for a better understanding of living organisms through "both the development of the appropriate languages to describe information processing in biological systems and the generation of more effective methods to translate biochemical descriptions into the functioning of the logic circuits that underpin biological phenomena." [1]

The language that Nurse wishes to see is a formal language that can be automatically translated into machine executable code and that enables simulation and analysis techniques for proving properties of biological systems. Although there are many approaches to the formal modeling of living systems, only a few provide executable descriptions that highlight the mechanistic steps that make a system move from one state to another [2]. Almost all the techniques related to mathematical modeling abstract from these individual steps to produce global behavior, usually averaged over time.

Computer science provides the key elements to describe mechanistic steps: algorithms and programming languages [3]. Following the metaphor of molecules as processes introduced in [4], process calculi have been identified as a promising tool to model biological systems that are inherently complex, concurrent, and driven by the interactions of their subsystems.

Causality is a key difference between language-based modeling approaches and other techniques. In fact, causality in concurrent languages is strictly related to the notion of concurrency or independence of events, which makes causality substantially different from temporal ordering. An activity A causes an activity B if A is a necessary condition for B to happen and A influences the activity of B—i.e., there is a flow of information from A to B. The second part of the condition defining causality makes clear the distinction between precedence (related only to temporal ordering) and causality (a subset of the temporal ordering in which the flow of information is also considered) [5]. As a consequence, the list of the reactions performed by a system does not provide causal information but only temporal information. It is therefore mandatory to devise new modeling and analysis tools to address causality.

Causality is a key issue in the analysis of complex interacting systems because it helps in dissecting independent components and simplifying models while also allowing us to clearly identify cross-talks between different signaling cascades. Once the experimentalist observes an interesting event in a simulation, it is possible to compact the previous history of the system, exposing only the preceding events that caused the interesting one. This can give precise hints about the causes of a disease, the interaction of a drug with a living system (identifying its efficacy and its side effects), and the regulatory mechanisms of oscillating behaviors.

Causality is a relationship between events, and as such it is most naturally studied within discrete models, which are in turn described via algorithmic modeling languages. Although many modeling languages have been defined in computer science to model concurrent systems, many challenges remain to building algorithmic models for the system-level understanding of biological processes. These challenges include the relationship between low-level local interactions and emergent high-level global behavior; the incomplete knowledge of the systems under investigation; the multi-level and multi-scale representations in time, space, and size; and the causal relations between interactions and the context awareness of the inner components. Therefore, the modeling formalisms that are candidates to propel algorithmic systems biology should be complementary to and interoperable with mathematical modeling. They should address parallelism and complexity, be algorithmic and quantitative, express causality, and be interaction driven, composable, scalable, and modular.

LANGUAGE VISUALIZATION

A fundamental issue in the adoption of formal languages in biology is their usability. A modeling language must be understandable by biologists so they can relate it to their own informal models and to experiments.

One attempt by biologists to connect formal languages and informal descriptions of systems involved the use of a constrained natural language organized in the form of tables that collect all the information related to the structure and dynamic of a system. This narrative representation is informative and structured enough to be compiled into formal description that is amenable to simulation and analysis [6, 7]. Although the narrative modeling style is not yet visual, it is certainly more readable and corresponds better to the intuition of biologists than a formal (programming) language.

The best way to make a language understandable to scientists while also helping to manage complexity is to visualize the language. This is harder than visualizing data or visualizing the results of simulations because a language implicitly describes the full kinetics of a system, including the dynamic relationships between events. Therefore, language visualization must be dynamic, and possibly reactive [8], which means that a scientist should be able to detect and insert events in a running simulation by direct intervention. This requires a one-to-one correspondence between the internal execution of a formal language and its visualization so that the kinetics of the language can be fully reflected in the kinetics of the visualization and vice versa.

This ability to fully match the kinetics of a general (Turing-complete) modeling language to visual representations has been demonstrated, for example, for pi-calculus [9], but many practical challenges remain to adapting such general methods to specific visualization requirements. (See Figure 1 on the next page.) One such requirement, for example, is the visualization and tracking of molecular complexes; to this end, the BlenX language [10] and its support tools permit explicit representation of complexes of biological elements and examination of their evolution in time [11]. (See Figure 2 on page 103.) The graphical representation of complexes is also useful in studying morphogenesis processes to unravel the mechanistic steps of pattern formation. (See Figure 3 on page 104.)

ANALYSIS

Model construction is one step in the scientific cycle, and appropriate modeling languages (along with their execution and visualization capabilities) are important, particularly for modeling complex systems. Ultimately, however, one will want to analyze the model using a large number of techniques. Some of these techniques may be centered on the underlying mathematical framework, such as the analysis of

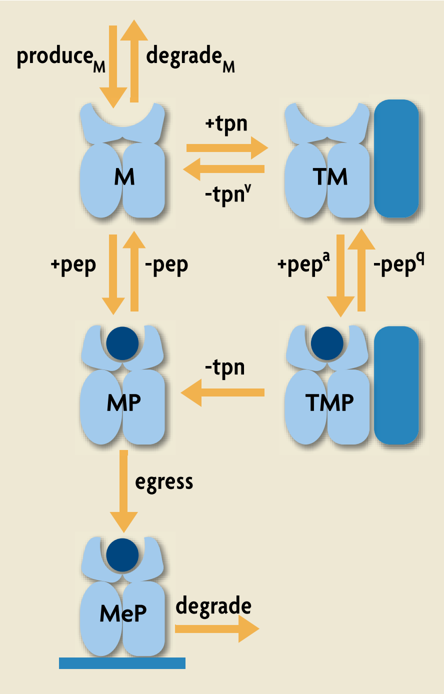


FIGURE 1.

This diagram can be placed in 1:1 correspondence with formal stochastic pi-calculus models [9, 12, 13] so that one can edit either the diagrams or the models. The nodes represent molecular states (the node icons are just for illustration), and the labeled arcs represent interactions with other molecules in the environment. The models use a biochemical variant of pi-calculus with rate weight as superscripts and with +/- for binding and unbinding.

differential equations, Markov chains, or Petri nets generated from the model. Other techniques may be centered on the model description (the language in which the model is written). For example, we may want to know whether two different model descriptions actually represent the same behavior, by some measure of behavior equivalence. This kind of model correspondence can arise, for example, from apparently different biological systems that work by the same fundamental principles. A similar question is whether we can simplify (abstract) a model description and still preserve its behavior, again by some measure of behavior equivalence that may mask some unimportant detail.

Behavioral equivalences are in fact a primary tool in computer science for verifying computing systems. For instance, we can use equivalences to ensure that an implementation is in agreement with a specification, abstracting as much as possible from syntactic descriptions and instead focusing on the semantics (dynamic) of specifications and implementations. So far, biology has focused on syntactic relationships between genes, genomes, and proteins. An entirely new avenue of research is the investigation of the semantic equivalences of biological entities populating complex networks of interactions. This approach could lead to new visions of systems and reinforce the need for computer science to enhance systems biology.

Biology is a data-intensive science. Biological systems are huge collections of in-

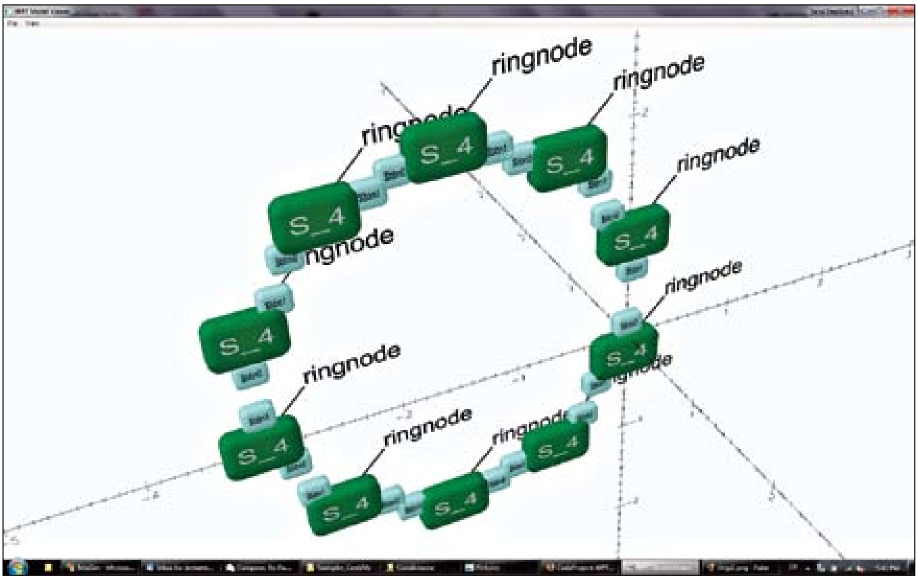


FIGURE 2.
The green S boxes in the diagram represent entities populating the biological system under consideration. The light blue rectangles attached to the green boxes represent the active interfaces/ domains available for complexation and decomplexation. The diagram shows how the simulation of the BlenX specification formed a ring complex and provides the position and the connections between boxes for inspection.

teracting components. The last decade of research has contributed to identifying and classifying those components, especially at the molecular level (gene, metabolites, proteins). To make sense of the large amount of data available, we need to implicitly represent them in compact and executable models so that executions can recover the available data as needed. This approach would merge syntax and semantics in unifying representations and would create the need for different ways of storing, retrieving, and comparing data. A model repository that represents the dynamics of biological processes in a compact and mechanistic manner would therefore be extremely valuable and could heighten the understanding of biological data and the basic biological principles governing life. This would facilitate predictions and the optimal design of further experiments to move from data collection to knowledge production.

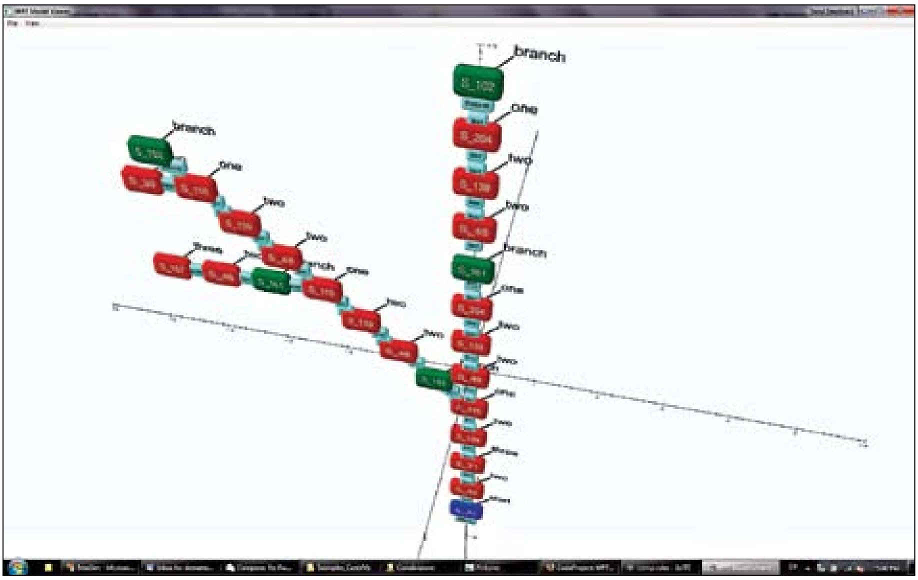


FIGURE 3.
The green, red, and blue S boxes in the diagram represent different species populating the biological system under consideration. The light blue rectangles attached to the boxes represent the active interfaces/domains available for complexation and decomplexation. The diagram elucidates how patterns are formed in morphogenesis processes simulated by BlenX specifications.

ANALYSIS VISUALIZATION

Executable models need visualization to make their execution interactive (to dynamically focus on specific features) and reactive (to influence their execution on the fly). Execution is one form of analysis; other analysis methods will need visualization as well. For complex systems, the normal method of "batch" analysis, consisting of running a complex analysis on the model and then mining the output for clues, needs to be replaced with a more interactive, explorative approach.

Model abstraction is an important tool for managing complexity, and we can envision performing this activity interactively—for example, by lumping components together or by hiding components. The notion of lumping will then need an appropriate visualization and an appropriate way of relating the behavior of the original components to the behavior of the lumped components. This doesn't mean visualizing the modeling language, but rather visualizing an abstraction function between models. We therefore suggest visualizing the execution of programs/models in such a way that the output is linked to the source code/model specification and the graphical abstraction performed by the end user is transformed into a formal program/model transformation. The supporting tool would then check which properties the transformation is preserving or not preserving and warn the user accordingly.

All the above reinforces the need for a formal and executable language to model biology as the core feature of an *in silico* laboratory for biologists that could be the next-generation high-throughput tool for biology.

ACKNOWLEDGMENTS

The authors thank Andrew Phillips and Lorenzo Dematté for preparing the figures.

REFERENCES

[1] P. Nurse, "Life, Logic and Information," *Nature*, vol. 454, pp. 424–426, 2008, doi: 10.1038/454424a.

[2] J. Fisher and T. Henzinger, "Executable Cell Biology," *Nature Biotechnology*, vol. 25, pp. 1239–1249, 2007, doi: 10.1038/nbt1356.

[3] C. Priami, "Algorithmic Systems Biology: An opportunity for computer science," *Commun. ACM*, June 2009, doi: 10.1145/1506409.1506427.

[4] A. Regev and E. Shapiro, "Cells as computation," *Nature*, vol. 419, p. 343, 2002, doi: 10.1038/419343a.

[5] P. Degano and C. Priami, "Non-interleaving semantics of mobile processes," *Theor. Comp. Sci.* vol. 216, no. 1–2, pp. 237–270, 1999.

[6] M. L. Guerriero, J. Heath, and C. Priami, "An automated translation from a narrative language for biological modelling into process algebra," *Proc. of CMSB 2007*, LNBI 4695, 2007, pp. 136–151, doi: 10.1007/978-3-540-75140-3_10.

[7] M. L. Guerriero, A. Dudka, N. Underhill-Day, J. Heath, and C. Priami, "Narrative-based computational modelling of the Gp130/JAK/STAT signalling pathway," *BMC Syst. Biol.*, vol. 3, no. 1, p. 40, 2009, doi: 10.1186/1752-0509-3-40.

[8] S. Efroni, D. Harel, and I. R. Cohen, "Reactive Animation: Realistic Modeling of Complex Dynamic Systems," *Computer*, vol. 38, no. 1, pp. 38–47, Jan. 2005, doi: 10.1109/MC.2005.31.

[9] A. Phillips, L. Cardelli, and G. Castagna, "A Graphical Representation for Biological Processes in the Stochastic Pi-calculus," *Trans. Comput. Syst. Biol., VII* - LNCS 4230, 2006, pp. 123–152, doi: 10.1007/11905455_7.

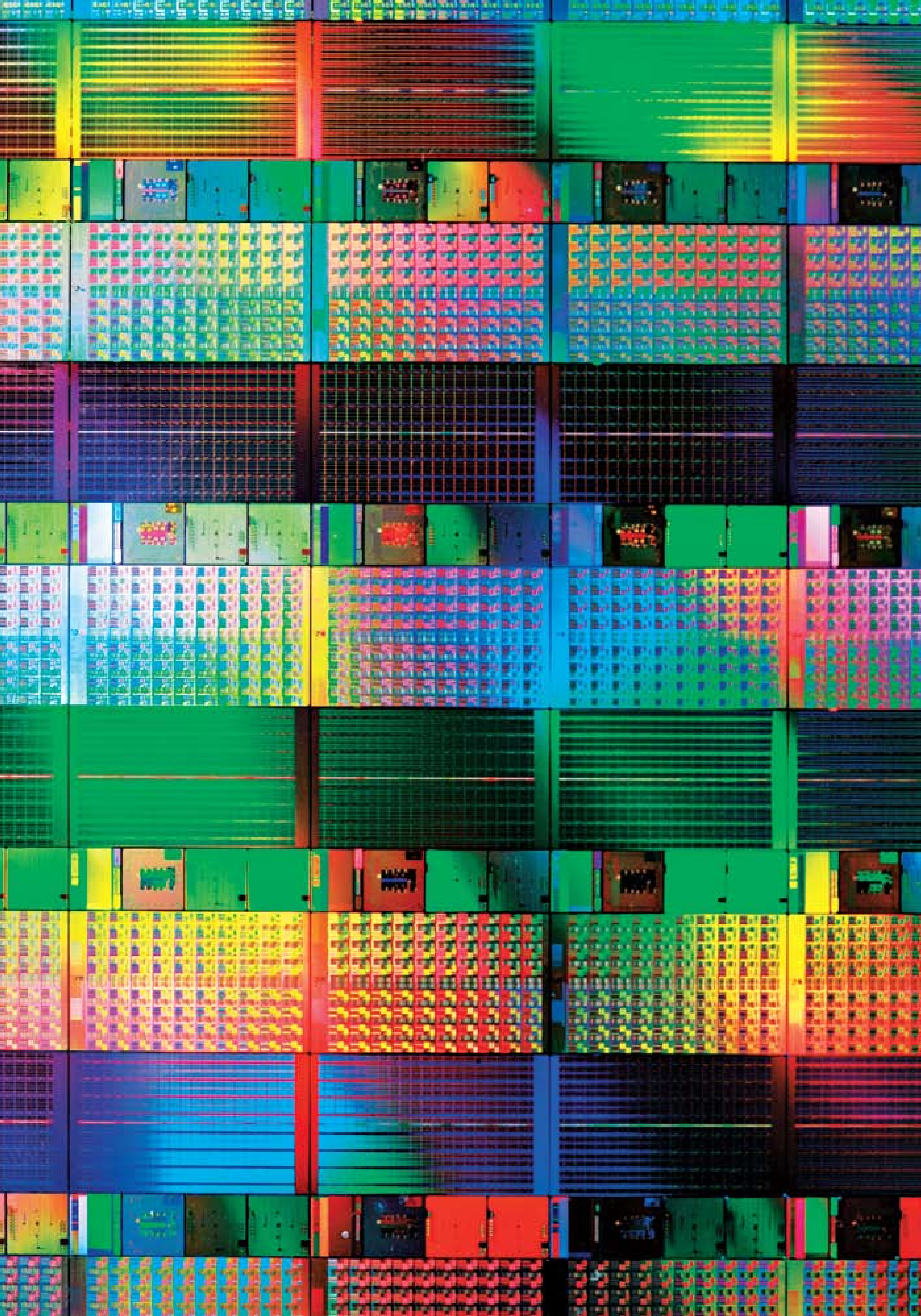
[10] L. Dematté, C. Priami, and A. Romanel, "The BlenX Language: a tutorial," *Formal Meth. Comput. Syst. Biol.*, LNCS 5016, 2008, pp. 313–365, doi: 10.1145/1506409.1506427.

[11] L. Dematté, C. Priami, and A. Romanel, "The Beta Workbench: a computational tool to study the dynamics of biological systems," *Brief Bioinform*, vol. 9, no. 5, pp. 437–449, 2008, doi: 10.1093/bib/bbn023.

[12] C. Priami, "Stochastic pi-calculus," *Comp. J.*, vol. 38, no. 6, pp. 578–589, 1995, doi: 10.1093/comjnl/38.7.578.

[13] A. Phillips and L. Cardelli, "Efficient, Correct Simulation of Biological Processes in Stochastic Pi-calculus," *Proc. Comput. Meth. Syst. Biol.*, Edinburgh, 2007, pp. 184–199, doi: 10.1007/978-3-540-75140-3_13.

3. SCIENTIFIC INFRASTRUCTURE



Introduction

DARON GREEN | Microsoft Research

WARNING! The articles in Part 3 of this book use a range of dramatic metaphors, such as "explosion," "tsunami," and even the "big bang," to strikingly illustrate how scientific research will be transformed by the ongoing creation and availability of high volumes of scientific data. Although the imagery may vary, these authors share a common intent by addressing how we must adjust our approach to computational science to handle this new proliferation of data. Their choice of words is motivated by the opportunity for research breakthroughs afforded by these large and rich datasets, but it also implies the magnitude of our culture's loss if our research infrastructure is not up to the task.

Abbott's perspective across all of scientific research challenges us with a fundamental question: whether, in light of the proliferation of data and its increasing availability, the need for sharing and collaboration, and the changing role of computational science, there should be a "new path for science." He takes a pragmatic view of how the scientific community will evolve, and he is skeptical about just how eager researchers will be to embrace techniques such as ontologies and other semantic technologies. While avoiding dire portents, Abbott is nonetheless vivid in characterizing a disconnect between the supply of scientific knowledge and the demands of the private and government sectors.

To bring the issues into focus, Southan and Cameron explore the "tsunami" of data growing in the EMBL-Bank database—a nucleotide sequencing information service. Throughout Part 3 of this book, the field of genetic sequencing serves as a reasonable proxy for a number of scientific domains in which the rate of data production is brisk (in this case, a 200% increase per annum), leading to major challenges in data aggregation, workflow, backup, archiving, quality, and retention, to name just a few areas.

Larus and Gannon inject optimism by noting that the data volumes are tractable through the application of multicore technologies—provided, of course, that we can devise the programming models and abstractions to make this technical innovation effective in general-purpose scientific research applications.

Next, we revisit the metaphor of a calamity induced by a data tidal wave as Gannon and Reed discuss how parallelism and the cloud can help with scalability issues for certain classes of computational problems.

From there, we move to the role of computational workflow tools in helping to orchestrate key tasks in managing the data deluge. Goble and De Roure identify the benefits and issues associated with applying computational workflow to scientific research and collaboration. Ultimately, they argue that workflows illustrate primacy of method as a crucial technology in data-centric research.

Fox and Hendler see "semantic eScience" as vital in helping to interpret interrelationships of complex concepts, terms, and data. After explaining the potential benefits of semantic tools in data-centric research, they explore some of the challenges to their smooth adoption. They note the inadequate participation of the scientific community in developing requirements as well as a lack of coherent discussion about the applicability of Web-based semantic technologies to the scientific process.

Next, Hansen et al. provide a lucid description of the hurdles to visualizing large and complex datasets. They wrestle with the familiar topics of workflow, scalability, application performance, provenance, and user interactions, but from a visualization standpoint. They highlight that current analysis and visualization methods lag far behind our ability to create data, and they conclude that multidisciplinary skills are needed to handle diverse issues such as automatic data interpretation, uncertainty, summary visualizations, verification, and validation.

Completing our journey through these perils and opportunities, Parastatidis considers how we can realize a comprehensive knowledge-based research infrastructure for science. He envisions this happening through a confluence of traditional scientific computing tools, Web-based tools, and select semantic methods.

A New Path for Science?

MARK R. ABBOTT
Oregon State University

The scientific challenges of the 21st century will strain the partnerships between government, industry, and academia that have developed and matured over the last century or so. For example, in the United States, beginning with the establishment of the National Science Foundation in 1950, the nation's research university system has blossomed and now dominates the basic research segment. (The applied research segment, which is far larger, is primarily funded and implemented within the private sector.)

One cannot overstate the successes of this system, but it has come to be largely organized around individual science disciplines and rewards individual scientists' efforts through publications and the promotion and tenure process. Moreover, the eternal "restlessness" of the system means that researchers are constantly seeking new ideas and new funding [1, 2]. An unexpected outcome of this system is the growing disconnect between the supply of scientific knowledge and the demand for that knowledge from the private and government sectors [3, 4]. The internal reward structure at universities, as well as the peer review system, favors research projects that are of inherent interest to the scientific community but not necessarily to those outside the academic community.

NEW DRIVERS

It is time to reexamine the basic structures underlying our research enterprise. For example, given the emerging and urgent need for new approaches to climate and energy research in the broad context of sustainability, fundamental research on the global climate system will continue to be necessary, but businesses and policymakers are asking questions that are far more interdisciplinary than in the past. This new approach is more akin to scenario development in support of risk assessment and management than traditional problem solving and the pursuit of knowledge for its own sake.

In climate science, the demand side is focused on feedback between climate change and socioeconomic processes, rare (but high-impact) events, and the development of adaptive policies and management protocols. The science supply side favors studies of the physical and biological aspects of the climate system on a continental or global scale and reducing uncertainties (e.g., [5]). This misalignment between supply and demand hampers society's ability to respond effectively and in a timely manner to the changing climate.

RECENT HISTORY

The information technology (IT) infrastructure of 25 years ago was well suited to the science culture of that era. Data volumes were relatively small, and therefore each data element was precious. IT systems were relatively expensive and were accessible only to experts. The fundamental workflow relied on a data collection system (e.g., a laboratory or a field sensor), transfer into a data storage system, data processing and analysis, visualization, and publication.

Figure 1 shows the architecture of NASA's Earth Observing System Data and Information System (EOSDIS) from the late 1980s. Although many thought that EOSDIS was too ambitious (it planned for 1 terabyte per day of data), the primary argument against it was that it was too centralized for a system that needed to be science driven. EOSDIS was perceived to be a data factory, operating under a set of rigorous requirements with little opportunity for knowledge or technology infusion. Ultimately, the argument was not about centralized versus decentralized but rather who would control the requirements: the science community or the NASA contractor. The underlying architecture, with its well-defined (and relatively modest-sized) data flows and mix of centralized and distributed components, has remained undisturbed, even as the World Wide Web, the Internet, and the volume of online data have grown exponentially.

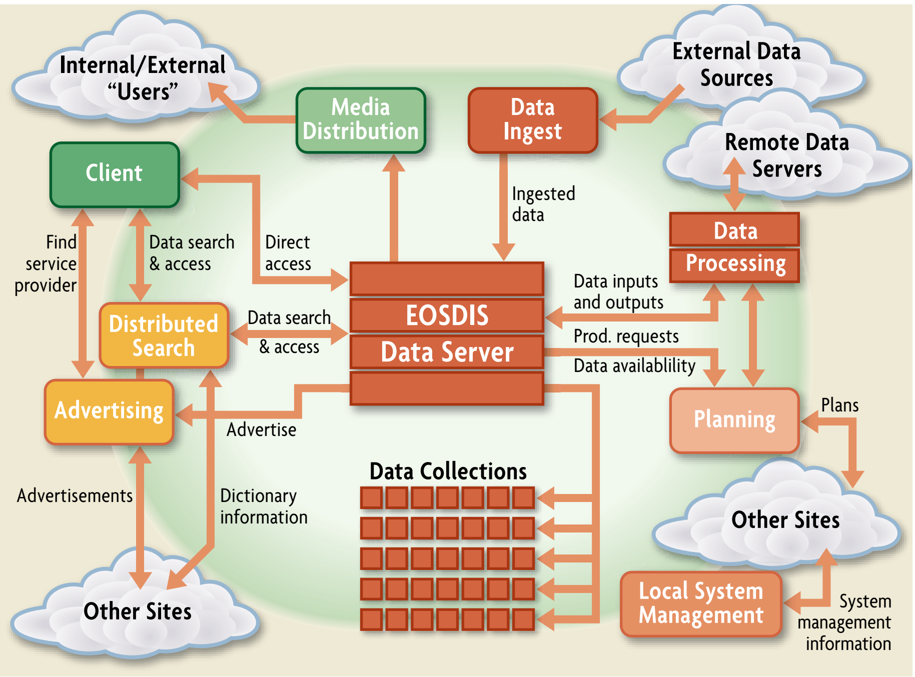


FIGURE 1.
NASA's Earth Observing System Data and Information System (EOSDIS) as planned in 1989.

THE PRESENT DAY

Today, the suite of national supercomputer centers as well as the notion of "cloud computing" looks much the same as the architecture shown in Figure 1. It doesn't matter whether the network connection is an RS-232 asynchronous connection, a dial-up modem, or a gigabit network, or whether the device on the scientist's desktop is a VT100 graphics terminal or a high-end multicore workstation. Virtualized (but distributed) repositories of data storage and computing capabilities are accessed via network by relatively low-capability devices.

Moore's Law has had 25 years to play out since the design of EOSDIS. Although we generally focus on the increases in capacity and the precipitous decline in the price/performance ratio, the pace of rapid technological innovation has placed enormous pressure on the traditional modes of scientific research. The vast amounts of data have greatly reduced the value of an individual data element, and we are no longer data-limited but insight-limited. "Data-intensive" should not refer just to the centralized repositories but also to the far greater volumes of data that are network-accessible in offices, labs, and homes and by sensors and portable devices. Thus, data-intensive computing should be considered more than just the ability to store and move larger amounts of data. The complexity of these new datasets as well as the increasing diversity of the data flows is rendering the traditional compute/data-center model obsolete for modern scientific research.

IMPLICATIONS FOR SCIENCE

IT has affected the science community in two ways. First, it has led to the *commoditization* of generic storage and computing. For science tasks that can be accomplished through commodity services, such services are a reasonable option. It will always be more cost effective to use low-profit-margin, high-volume services through centralized mechanisms such as cloud computing. Thus more universities are relying on such services for data backup, e-mail, office productivity applications, and so on.

The second way that IT has affected the science community is through radical *personalization*. With personal access to teraflops of computing and terabytes of storage, scientists can create their own compute clouds. Innovation and new science services will come from the edges of the networks, not the commodity-driven datacenters. Moreover, not just scientists but the vastly larger number of sensors and laboratory instruments will soon be connected to the Internet with their own local computation and storage services. The challenge is to harness the power of this new network of massively distributed knowledge services.

Today, scientific discovery is not accomplished solely through the well-defined, rigorous process of hypothesis testing. The vast volumes of data, the complex and hard-to-discover relationships, the intense and shifting types of collaboration between disciplines, and new types of near-real-time publishing are adding pattern and rule discovery to the scientific method [6]. Especially in the area of climate science and policy, we could see a convergence of this new type of data-intensive research and the new generation of IT capabilities.

The alignment of science supply and demand in the context of continuing scientific uncertainty will depend on seeking out new relationships, overcoming language and cultural barriers to enable collaboration, and merging models and data to evaluate scenarios. This process has far more in common with network gaming than with the traditional scientific method. Capturing the important elements of data preservation, collaboration, provenance, and accountability will require new approaches in the highly distributed, data-intensive research community.

Instead of well-defined data networks and factories coupled with an individually based publishing system that relies on peer review and tenure, this new research enterprise will be more unruly and less predictable, resembling an ecosystem in its approach to knowledge discovery. That is, it will include loose networks of potential services, rapid innovation at the edges, and a much closer partnership between those who create knowledge and those who use it. As with every ecosystem, emergent (and sometimes unpredictable) behavior will be a dominant feature.

Our existing institutions—including federal agencies and research universities—will be challenged by these new structures. Access to data and computation as well as new collaborators will not require the physical structure of a university or millions of dollars in federal grants. Moreover, the rigors of tenure and its strong emphasis on individual achievement in a single scientific discipline may work against these new approaches. We need an organization that integrates natural science with socioeconomic science, balances science with technology, focuses on systems thinking, supports flexible and interdisciplinary approaches to long-term problem solving, integrates knowledge creation and knowledge use, and balances individual and group achievement.

Such a new organization could pioneer integrated approaches to a sustainable future, approaches that are aimed at understanding the variety of possible futures. It would focus on global-scale processes that are manifested on a regional scale with pronounced socioeconomic consequences. Rather than a traditional academic organization with its relatively static set of tenure-track professors, a new organization could take more risks, build and develop new partnerships, and bring in people with the talent needed for particular tasks. Much like in the U.S. television series *Mission Impossible*, we will bring together people from around the world to address specific problems—in this case, climate change issues.

MAKING IT HAPPEN

How can today's IT enable this type of new organization and this new type of science? In the EOSDIS era, it was thought that relational databases would provide the essential services needed to manage the vast volumes of data coming from the EOS satellites. Although database technology provided the baseline services needed for the standard EOS data products, it did not capture the innovation at the edges of the system where science was in control. Today, semantic webs and ontologies are being proposed as a means to enable knowledge discovery and collaboration. However, as with databases, it is likely that the science community will be reluctant to use these inherently complex tools except for the most mundane tasks.

Ultimately, digital technology can provide only relatively sparse descriptions of the richness and complexity of the real world. Moreover, seeking the unusual and unexpected requires creativity and insight—processes that are difficult to represent in a rigid digital framework. On the other hand, simply relying on PageRank[1]-like statistical correlations based on usage will not necessarily lead to detection of the rare and the unexpected. However, new IT tools for the data-intensive world can provide the ability to "filter" these data volumes down to a manageable level as well as provide visualization and presentation services to make it easier to gain creative insights and build collaborations.

The architecture for data-intensive computing should be based on storage, computing, and presentation services at every node of an interconnected network. Providing standard, extensible frameworks that accommodate innovation at the network edges should enable these knowledge "ecosystems" to form and evolve as the needs of climate science and policy change.

REFERENCES

[1] D. S. Greenberg, *Science, Money, and Politics: Political Triumph and Ethical Erosion.* Chicago: University of Chicago Press, 2001.

[2] National Research Council, *Assessing the Impacts of Changes in the Information Technology R&D Ecosystem: Retaining Leadership in an Increasingly Global Environment.* Washington, D.C.: National Academies Press, 2009.

[3] D. Sarewitz and R. A. Pielke, Jr., "The neglected heart of science policy: reconciling supply of and demand for science," *Environ. Sci. Policy,* vol. 10, pp. 5–16, 2007, doi: 10.1016/j.envsci.2006.10.001.

[4] L. Dilling, "Towards science in support of decision making: characterizing the supply of carbon cycle science," *Environ. Sci. Policy,* vol. 10, pp. 48–61, 2007, doi: 10.1016/j.envsci.2006.10.008.

[5] Intergovernmental Panel on Climate Change, *Climate Change 2007: The Physical Science Basis.* New York: Cambridge University Press, 2007.

[6] C. Anderson, "The End of Theory," *Wired*, vol. 16, no. 7, pp. 108–109, 2008.

[1] The algorithm at the heart of Google's search engine.