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

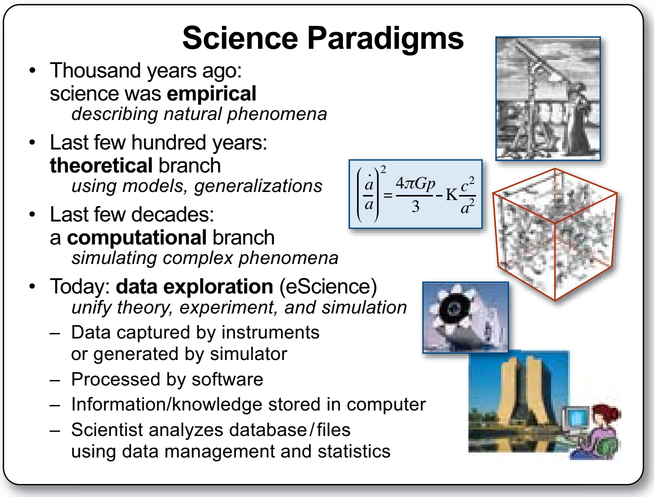
Based on the transcript of a talk given by Jim Gray to the NRC-CSTB[1] in Mountain View, CA, on January 11, 2007[2]

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

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

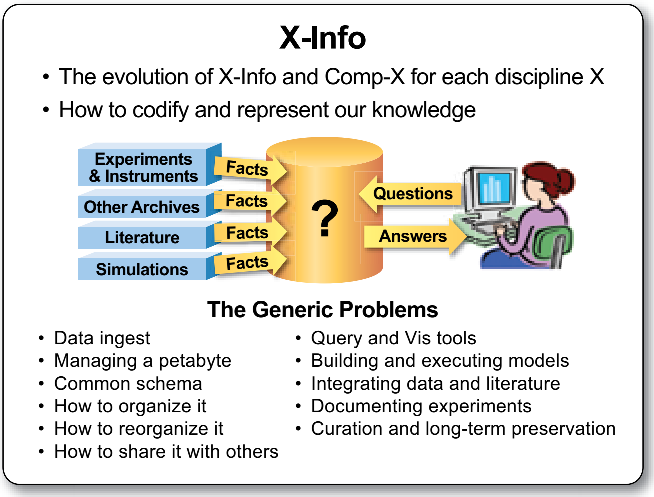
FIGURE 1

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

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.

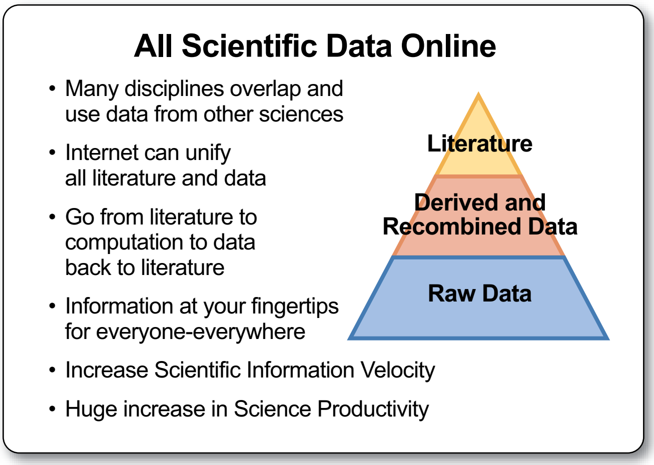
THE COMING REVOLUTION IN SCHOLARLY COMMUNICATION

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