



The Future of Scientific Knowledge Discovery in Open Networked Environments: Summary of a Workshop

ISBN
978-0-309-26791-5

200 pages
8 1/2 x 11
PAPERBACK (2012)

Paul F. Uhler, Rapporteur; Board on Research Data and Information; Policy and Global Affairs; National Research Council

 Add book to cart

 Find similar titles

 Share this PDF



Visit the National Academies Press online and register for...

- ✓ Instant access to free PDF downloads of titles from the
 - NATIONAL ACADEMY OF SCIENCES
 - NATIONAL ACADEMY OF ENGINEERING
 - INSTITUTE OF MEDICINE
 - NATIONAL RESEARCH COUNCIL
- ✓ 10% off print titles
- ✓ Custom notification of new releases in your field of interest
- ✓ Special offers and discounts

Distribution, posting, or copying of this PDF is strictly prohibited without written permission of the National Academies Press. Unless otherwise indicated, all materials in this PDF are copyrighted by the National Academy of Sciences. Request reprint permission for this book

THE FUTURE OF SCIENTIFIC KNOWLEDGE DISCOVERY IN OPEN NETWORKED ENVIRONMENTS

Summary of a Workshop

Paul F. Uhler, Rapporteur

Board on Research Data and Information
Policy and Global Affairs

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by the National Science Foundation under Grant No. 1042078. This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-26791-5

International Standard Book Number-10: 978-0-309-26791-9

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Room 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; <http://www.nap.edu>.

Copyright 2012 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

THE FUTURE OF SCIENTIFIC KNOWLEDGE DISCOVERY IN OPEN NETWORKED ENVIRONMENTS: A National Workshop

Steering Committee

John Leslie King, Chair

William Warner Bishop Collegiate Professor of Information
University of Michigan

Hal Abelson

Professor, Massachusetts Institute of Technology

Francine Berman

Vice President of Research, Rensselaer Polytechnic Institute

Bonnie Carroll

President, Information International Associates

Michael Carroll

Professor, American University, Washington College of Law

Alyssa Goodman

Professor, Harvard University

Sara Graves

Director, Information Technology and Systems Center
University Professor of Computer Science
University of Alabama in Huntsville

Michael Lesk

Professor, Rutgers University

Gilbert Omenn

Professor, University of Michigan

Project Staff

Paul F. Uhler

Board Director
The National Academies

Daniel Cohen

Program Officer
The National Academies
[on detail from Library of Congress]

Cheryl Levey

Senior Program Associate
The National Academies

Board on Research Data and Information – Membership

Michael Lesk, *Chair* (until 11/2011)
Rutgers University

Roberta Balstad, *Vice Chair* (until 11/2011)
Columbia University

Francine Berman, *Co-Chair*
Rensselaer Polytechnic Institute

Clifford Lynch, *Co-Chair*
Coalition for Networked Information

Maureen Baginski (until 11/2011)
Serco

Laura Bartolo
Kent State University

R. Steven Berry (until 11/2011)
University of Chicago

Christine Borgman (until 11/2011)
University of California, Los Angeles

Philip Bourne
University of California, San Diego

Norman Bradburn (until 11/2011)
University of Chicago

Henry Brady
University of California, Berkeley

Mark Brender
GeoEye Foundation

Bonnie Carroll
Information International Associates

Michael Carroll
American University, Washington College of
Law

Sayed Choudhury
The Johns Hopkins University

Keith Clarke
University of California, Santa Barbara

Paul A. David
Stanford University

Kelvin Droegemeier
University of Oklahoma

Clifford Duke
Ecological Society of America

Barbara Entwisle
University of North Carolina

Stephen Friend
Sage Bionetworks

Michael Goodchild (until 11/2011)
University of California, Santa Barbara

Alyssa Goodman (until 11/2011)
Harvard University

Margaret Hedstrom
University of Michigan

Michael Keller (until 11/2011)
Stanford University

Alexa T. McCray
Harvard Medical School

Michael R. Nelson (until 11/2011)
Georgetown University

Daniel Reed (until 11/2011)
Microsoft Research, Microsoft Inc.

Alan M. Title
Lockheed Martin Advanced Technology
Center

Ann J. Wolpert
Massachusetts Institute of Technology

Cathy H. Wu (until 11/2011)
University of Delaware and
Georgetown University Medical Center

Board on Research Data and Information Staff

Paul F. Uhler
Board Director

Daniel Cohen
Program Officer
[on detail from Library of Congress]

Subhash Kuvelker
Senior Program Officer

Cheryl Levey
Senior Program Associate

Preface and Acknowledgments

Digital technologies and networks are now part of everyday work in the sciences, and have enhanced access to and use of scientific data, information, and literature significantly. They offer the promise of accelerating the discovery and communication of knowledge, both within the scientific community and in the broader society, as scientific data and information are made openly available online.

The phrase “scientific knowledge discovery in open networked environments” is subject to many definitions. For purposes of this project, the focus was on computer-mediated or computational scientific knowledge discovery, taken broadly as any research processes enabled by digital computing technologies. Such technologies may include data mining, information retrieval and extraction, artificial intelligence, distributed grid computing, and others. These technological capabilities support computer-mediated knowledge discovery, which some believe is a new paradigm in the conduct of research.

The emphasis was primarily on digitally networked data, rather than on the scientific, technical, and medical literature. The meeting also focused mostly on the advantages of knowledge discovery in open networked environments, although some of the disadvantages were raised as well.

The workshop brought together a set of stakeholders in this area for intensive and structured discussions. The purpose was not to make a final declaration about the directions that should be taken, but to further the examination of trends in computational knowledge discovery in the open networked environments, based on the following questions and tasks:

1. Opportunities and Benefits: What are the opportunities over the next 5 to 10 years associated with the use of computer-mediated scientific knowledge discovery across disciplines in the open online environment? What are the potential benefits to science and society of such techniques?

2. Techniques and Methods for Development and Study of Computer-mediated Scientific Knowledge Discovery: What are the techniques and methods used in government, academia, and industry to study and understand these processes, the validity and reliability of their results, and their impact inside and outside science?

3. Barriers: What are the major scientific, technological, institutional, sociological, and policy barriers to computer-mediated scientific knowledge discovery in the open online environment within the scientific community? What needs to be known and studied about each of these barriers to help achieve the opportunities for interdisciplinary science and complex problem solving?

4. Range of Options: Based on the results obtained in response to items 1–3, define a range of options that can be used by the sponsors of the project, as well as other similar organizations, to obtain and promote a better understanding of the computer-mediated scientific knowledge discovery processes and mechanisms for openly available data and information online across the scientific domains. The objective of defining these options is to improve the activities of the sponsors (and other similar organizations) and the activities of researchers that they fund externally in this emerging research area.

The first day of the 2-day meeting consisted primarily of invited expert speakers, who addressed tasks 1–3. This was followed immediately by a workshop on the second day to leverage the expertise of the invitees to address task 4, based on the discussions of tasks 1–3 on the first day. The slides presented by the speakers at the meeting are posted on the National Academy of Sciences’ Board on Research Data and Information Web site and the entire meeting was webcast.¹

This report has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The committee’s role was limited to planning and convening the workshop. The views contained in the report are those of the individual workshop participants and do not necessarily represent the views of all workshop participants, the steering committee, or the National Academies.

It can be argued that too much time has passed since the meeting took place, and that the results of this effort are not timely enough to provide insight. In fact, the elapsed time between the events reported here and this report has provided time to assess the issues with more care.

We are grateful to the National Science Foundation (NSF) for support of this project under NSF Grant Number 1042078. This volume has been reviewed in draft form by individuals chosen for their technical expertise, in accordance with procedures approved by the National Research Council’s Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for quality. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We wish to thank the following individuals for their review of this report: Sayeed Choudhury, Johns Hopkins University; Stephen Hilgartner, Cornell University; Michael Kurtz, Harvard University; Robert McDonald, Indiana University; Mark Parsons, University of Colorado; Jack Stankovic, University of Virginia; and Katherine Strandburg, New York University.

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the content of the individual papers. Responsibility for the final content of the papers rests with the individual authors.

¹Available at http://sites.nationalacademies.org/PGA/brdi/PGA_060424.

We would especially like to recognize the contributions of Daniel Cohen, on assignment to the National Academies from the U.S. Library of Congress, who assisted with the editing and the production of the manuscript Cheryl Levey of the board staff also helped with the review process and the preparation of this volume. Finally, we would like to thank Raed Sharif for his editorial support in completing this manuscript.

John Leslie King
Steering Committee Chair

Paul F. Uhler
Project Director

Contents

| | |
|--|-----------|
| 1. Opening Session | 1 |
| <i>Introduction, 1</i> | |
| John Leslie King | |
| <i>Opening Remarks by Project Sponsors, 3</i> | |
| Alan Blatecky | |
| Sylvia Spengler | |
| <i>Keynote Address: An Overview of the State of the Art, 7</i> | |
| Tony Hey | |
| Discussion, 17 | |
| 2. Experiences with Developing Open Scientific Knowledge Discovery in Research and Applications | 19 |
| <i>Case Studies</i> | |
| International Online Astronomy Research, 19 | |
| Alberto Conti | |
| Integrative Genomic Analysis, 25 | |
| Stephen Friend | |
| Geoinformatics: Linked Environments for Atmospheric Discovery, 31 | |
| Sara Graves | |
| <i>Implications of the Three Scientific Knowledge Discovery Case Studies – The User Perspective</i> | |
| International Online Astronomy Research, c. 2011, 37 | |
| Alyssa Goodman | |
| Integrative Genomic Analysis, 45 | |
| Joel Dudley | |
| Geoinformatics, 55 | |
| Mohan Ramamurthy | |
| Discussion, 61 | |
| 3. How Might Open Online Knowledge Discovery Advance the Progress of Science? 69 | |
| Technological Factors | |
| Session Chair: Hal Abelson | |
| Interoperability, Standards, and Linked Data, 71 | |
| James Hendler | |
| National Technological Needs and Issues, 77 | |
| Deborah Crawford | |
| Discussion, 81 | |
| <i>Sociocultural, Institutional, and Organizational Factors, 83</i> | |
| Session Chair: Michael Lesk | |
| Sociocultural Dimensions, 85 | |
| Clifford Lynch | |
| Institutional Factors, 89 | |
| Paul Edwards | |
| Discussion, 96 | |
| <i>Policy and Legal Factors, 99</i> | |
| Session Chair: Michael Carroll | |
| Legal Aspects, 101 | |
| Michael Madison | |
| Knowledge Discovery in Open Networked Environments: Some Policy Issues, 107 | |
| Gregory A. Jackson | |
| Discussion, 114 | |
| <i>How Can We Tell? What Needs to Be Known and Studied to Improve Potential for Success?, 117</i> | |
| Session Chair: Francine Berman | |
| Introduction, 119 | |

| | |
|--|------------|
| Francine Berman An Academic Perspective, 121 Victoria Stodden A Government Perspective, 127 Walter L. Warnick Discussion, 130 | |
| 4. Summary of Workshop Results from Day One and Discussion of Additional Issues | 133 |
| <i>Introduction, 135</i> Bonnie Carroll <i>Opportunities and Benefits for Automated Scientific Knowledge Discovery in Open Networked Environments, 137</i> Puneet Kishor <i>Techniques and Methods for Development and Study of Automated Scientific Knowledge Discovery, 147</i> Alberto Pepe <i>Barriers to Automated Scientific Knowledge Discovery in Open Networked Environments, 155</i> Alberto Pepe <i>Range of Options for Further Research, 163</i> Puneet Kishor | |
| 5. Appendix: Workshop Agenda | 179 |

1. Opening Session

Introduction

John Leslie King
-University of Michigan-

As the chair of the committee that is organizing this two-day event on “The Future of Scientific Knowledge Discovery in Open Networked Environments,” I would like to thank primarily the National Science Foundation, but also the Library of Congress and the Department of Energy, for sponsoring this project. The event was organized by two boards of the National Academies, the Board on Research Data and Information in collaboration with the Computer Sciences Telecommunications Board. In my opening remarks, I would like to address two points: one on the motivation for this event and the other one on the means.

The motivation lies in the fact that we are undergoing a relatively rapid change in capability, because of technology and improved knowledge. What is possible today was not even thinkable some years ago for data generation, storage, management, and sharing. Developments in these areas are having a rather profound impact on the nature of science.

This is not new historically, however. When the technologies of optics were developed and the microscope and telescope were created, these technologies changed what we *could* do, but they did not immediately change what we *did* do. Over time, there was an evolution of social conventions, of data record keeping, and so forth. Take for example the classical story of Tycho Brahe, who spent years looking through telescopes and taking detailed measurements, and later his data were instrumental in Johannes Kepler’s proof that the orbits of the planets were elliptical rather than circular.

That kind of practice has been repeated many times in the history of science, and it is being repeated now. Some of my colleagues at the University of Michigan are working on the ATLAS Project at the Large Hadron Collider in CERN, the European Center for Nuclear Research. ATLAS is an interesting operation. It has nearly 3,000 principal investigators, and many of the papers produced through this project have more pages dedicated to the listing of authors than to the paper itself. Obviously they have had to develop new conventions to deal with this situation. If we ask them about how they know who did what in a project or a paper, they will say that they just know. That is probably more tacit knowledge than it is explicit knowledge.

The ATLAS detector is capable of generating a petabyte of data per second; most of these data are thrown away immediately as not relevant for the work. The amount of data left is still huge, however, and CERN distributes large volumes of data globally through different networks. Such big and important operations have to be planned. We cannot just expect that information technology people will take care of that simply because nobody has ever done it before. It took years to plan the data distribution plan for CERN. This, of course, did not involve the details of how the data get used, how credit gets assigned, and so forth. The point I am trying to make is that this is an emerging and very challenging frontier, and that we are still

at the infant stage of it. The means are to address four issues: (1) the opportunity–benefit space; (2) the techniques that we know now and those that we think are going to emerge; (3) the barriers we have to overcome; and (4) the options that we can use to move forward.

Opening Remarks by Project Sponsors

Alan Blatecky
-National Science Foundation-

Thinking about the cyberinfrastructure and the impact of data, we need to understand that it is an ecosystem with multiple components that are dependent upon each other. Figure 1-1 presents a very complex and interconnected world of cyberinfrastructure. Although data is a key component, there are other components that impact data dramatically; this includes networking, software, computational resources, expertise, and so on.

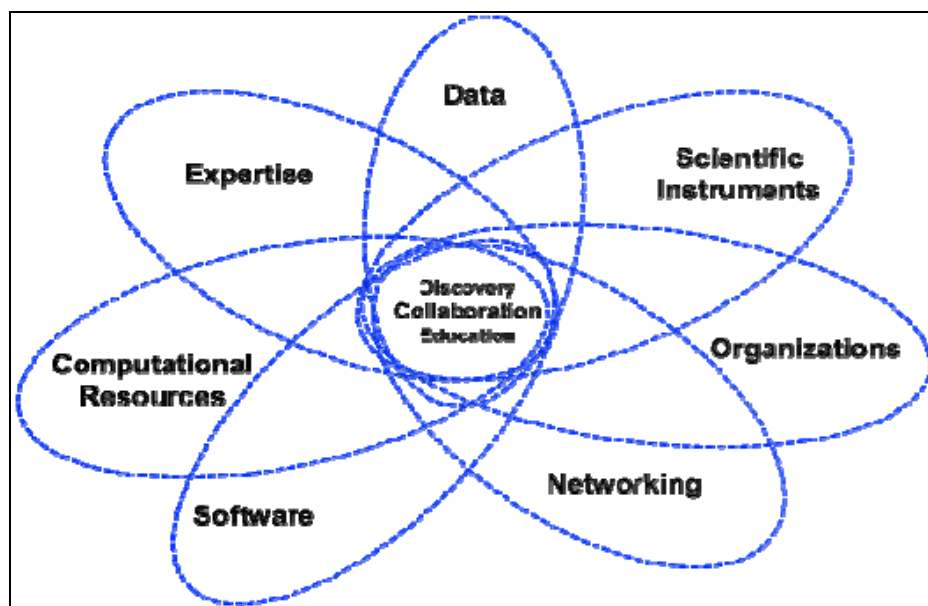


FIGURE 1-1 Cyberinfrastructure.
SOURCE: Alan Blatecky

Furthermore, Figure 1-1 makes two points. First, data cannot be considered as a separate entity. Second, as data capabilities grow, this has an impact on networking requirements, on organizations, on expertise, and so forth—and also vice versa. Advances in one component affect other areas. As mentioned with the ATLAS Project example in the opening presentation, as advanced scientific instruments are developed, they enable tremendous new capabilities. The challenge is how to accommodate and take advantage of these new capabilities, and how does this impact the cyberinfrastructure environment?

Figure 1-2 addresses four data challenges.

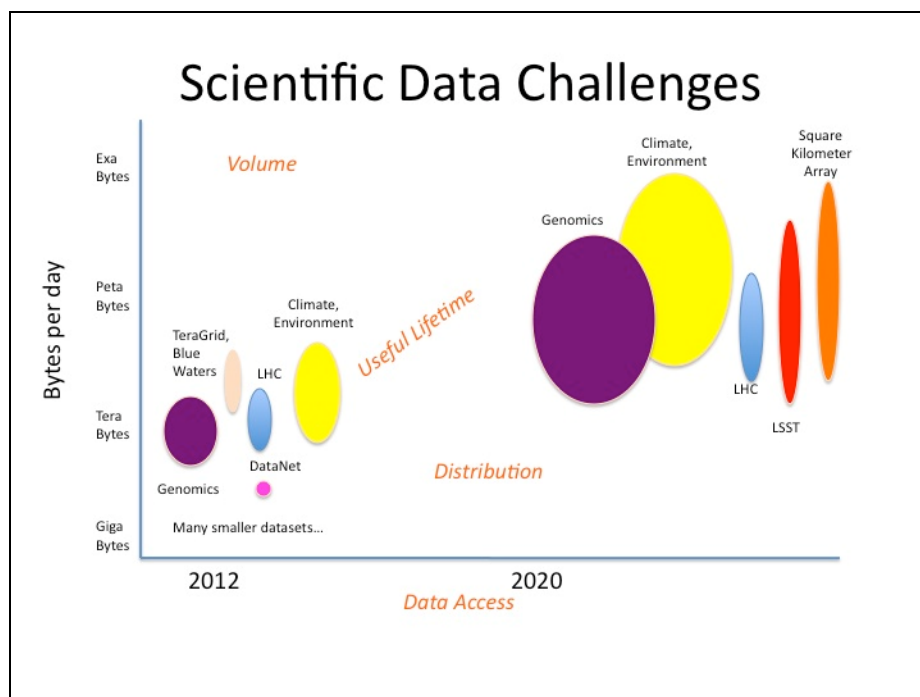


FIGURE 1-2 Scientific data challenges.
SOURCE: Alan Blatecky

The first point is the growth and useful lifetime in data from today to what is expected in 2020. While terabytes and petabytes of data and information are currently being generated, in less than 10 years we are going to have volumes in petabytes and exabytes. In short, there is going to be a tremendous growth in the amount of scientific data being generated, and there can be a long lifetime associated with the data. Therefore, issues of curation, sustainability, preservation, and metadata approaches need to be addressed.

The second challenge is the volume of data by discipline, instrument, or observatory. In Figure 1-2, the vertical dimension of the ellipses is intended to illustrate the volume of data being generated by that area. As Prof. King pointed out, the Large Hadron Collider (LHC) will be generating a gigabyte per second—multiple terabytes a day, in essence—and the figure illustrates how much volume that is.

The third challenge is distribution. In Figure 1-2, the horizontal dimension of the ellipse illustrates the amount of distribution for that data. For example, the data from the LHC or Large Synoptic Survey Telescope will be stored in a few locations; hence, the ellipse is fairly narrow. Genomics data or climate data, on the other hand, will be located at literally hundreds, if not thousands, of locations. The point here is that distribution is a very different challenge from volume or lifetime. Genomic data are being generated at incredible rates at many universities and medical centers, and managing that data across hundreds of sites is a significant challenge.

The fourth challenge deals with data access. How will the data be accessed and what policies are required for access (e.g., privacy, security, integrity)? These are some of the challenges we need to address.

The point of all of this is exactly what Prof. King suggested in his opening presentation—that science, technology, and society will be transformed by data. Modern science will depend on it. We need to be focusing on multidisciplinary activities and collaborations. In this new Age of Observation, there is a whole new world opening up for everyone; this is similar to what happened with the development of telescopes, only now data will be the “instrument” that will change the way we understand the world.

The last point I will make is that the National Science Foundation (NSF) Advisory Committee on Cyberinfrastructure established six task forces that worked for almost 2 years to develop a series of recommendations on cyberinfrastructure. The data task force addressed these issues in depth, and as the task force concluded, there are infrastructure components to deal with, as well as culture changes and economic sustainability. It is necessary to determine how to manage data and deal with such diverse issues as ethics, intellectual property rights, and economic sustainability. In this regard, the NSF should address the following issues:

- ❖ *Infrastructure*: Recognize data infrastructure and services (including visualization) as essential research assets fundamental to today’s science and as long-term investments in national prosperity.
- ❖ *Culture Change*: Reinforce expectations for data sharing; support the establishment of new citation models in which data and software tool providers and developers are credited with their contributions.
- ❖ *Economic Sustainability*: Develop and publish realistic cost models to underpin institutional and national business plans for research repositories and data services.
- ❖ *Data Management Guidelines*: Identify and share best practices for the critical areas of data management.
- ❖ *Ethics and Internet Protocol*: Train researchers in privacy-preserving data access.

Sylvia Spengler
National Science Foundation

I am pleased to have this meeting organized, because of the opportunity for the Board on Research Data and Information and the Computer Sciences Telecommunications Board to come together and begin to explore where the policy, legal, social, and computer science challenges intersect in this area. I am also pleased that we were able to bring together individuals who represent different communities and fields. This is a good opportunity to sustain conversations that have been scattered around the globe for the past 15 years.

From a personal perspective, I am gratified to see that some of the issues that I have long worried about in the fields of data and analytics have finally come to the attention of the people who are going to use the data. We have at the table not only scientists from a single scientific domain, but scientists from many different kinds of domains. Science is inherently multidisciplinary and will require different kinds of expertise, different communities, and new kinds of social environments within science. This is an opportunity to begin to think about that.

Keynote Address: An Overview of the State of the Art

Tony Hey
-Microsoft Research-

I will provide a summary of where I think we are in the data-intensive knowledge discovery area and offer a personal view of some of the issues that we are facing. As mentioned earlier, we are in the midst of a revolution, but we have been in this revolution before. I like the example of Kepler, because Tycho Brahe's measurements were sufficiently accurate that Kepler could not fit the planetary orbits to circles and was forced to consider ellipses, which led eventually to Newton's Laws. However, it was actually the data that Tycho Brahe took without a telescope that led Kepler to his conclusion.

Some of my slides are based on Dr. Jim Gray's last talk in 2007. Dr. Gray was a computer scientist at Microsoft until he disappeared at sea in January 2007 while on his sailboat. Dr. Gray made the argument that we are moving into a new area in which people will need different skills.

Obviously the skills for experimental and theoretical science are not going to disappear—we still need and use them—but computational science requires people to know about, for example, numerical methods, algorithms, computer architectures, and parallel programming. This is a set of specialized skills for doing work in areas such as climate prediction and modeling galaxy formation. Today data-intensive science essentially requires a new set of skills for handling and manipulating large amounts of data, visualizing them, and so on. When I ran the U.K.'s *e*-Science Program, the agenda was about tools and infrastructure to support data-intensive science. It was distributed, it was data intensive, and it involved a new way of thinking about what we are doing. That is how I distinguish between *e*-science and data-intensive science.

Figure 1-3 is from Jim Gray's talk.

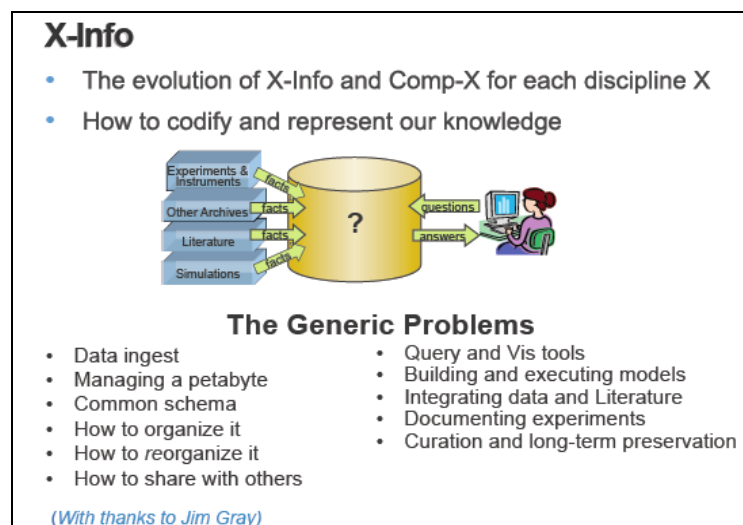


FIGURE 1-3 X-Info.

SOURCE: Microsoft Research

Figure 1-3 shows field X evolving into two separate subdisciplines. There is computational biology, and there is bioinformatics; computational chemistry, and then cheminformatics. Almost every field (e.g., astroinformatics, archeoinformatics) today is being pervaded by informatics and the need to deal with data. It is no longer just science.

Dr. Gray saw that in responding to the problems of getting and manipulating the data there is a need to teach people new skills. Managing a petabyte of data is not easy, because, for example, it is difficult to send a petabyte easily over the Internet. We need to think about organizing and reorganizing the data in case a wrong assumption was made. Other issues to think about include sharing the data, query and visualization, integrating data and literature, and so on. As we can see, there are many new challenges.

Dr. Gray had this vision of moving from literature to computation and then back to literature through a process of linking and combining different sources of data to produce new data and knowledge. He believed that this was the future global digital library for science and that it could lead to a huge increase in scientific productivity, because the research process is fairly inefficient at present, as described in Figure 1-4.

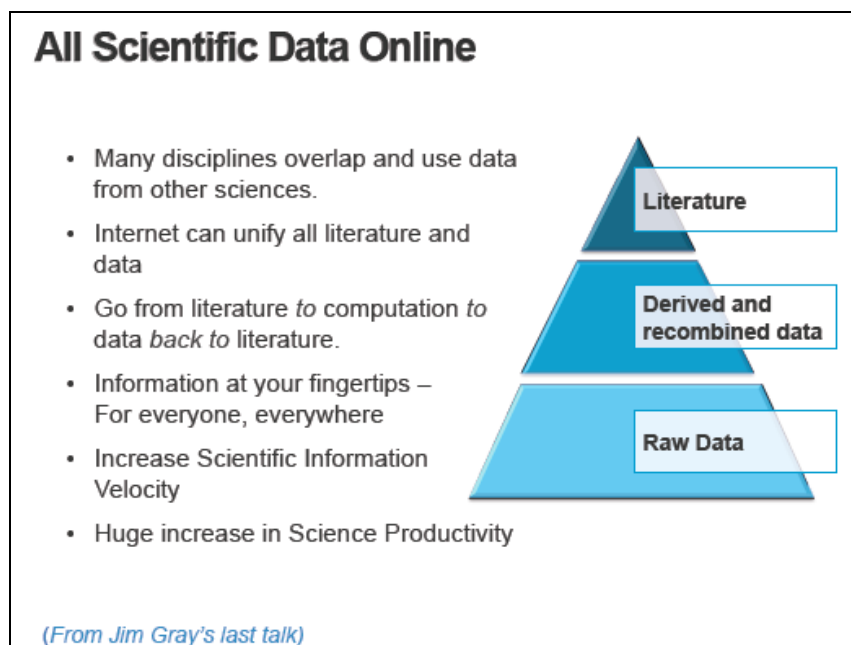


FIGURE 1-4 All scientific data online.
SOURCE: Microsoft Research

The other issue that I would like to emphasize is the reproducibility of science. If a scientist is interested in a paper, accessing the data on which that paper is based is not easy, and it is probably only in the minority of cases that a scientist can do it. So this was Dr. Gray's vision—a big, global-distributed digital library—and that was why I accepted to cochair a National Science Foundation (NSF) Office of Cyberinfrastructure Taskforce on Data and Visualization. We tried to produce a report quickly, because the data agenda is a very important area for all the agencies, not just NSF².

² The final version of the report can be found on the NSF Web site <http://www.nsf.gov/od/oci/taskforces/>.

The report covers areas such as infrastructure, delivery, culture and sociological change, roles and responsibilities, ethics and privacy, economic value and sustainability, and data management guidelines. Also, NSF is now requiring that its grantees have a data management plan.

Another point I would like to make is related to funding. To illustrate this point, I will give an example from the Coastal Ocean Observation Lab at Rutgers University. After a boating or aircraft accident at sea, the U.S. Coast Guard (USCG) historically has relied on current charts and wind gauges to determine where to hunt for survivors. But thanks to data originally collected by Rutgers University oceanographers³ to answer scientific questions about earth-ocean-atmosphere interactions, the USCG has a new resource that promises to literally save lives. This is a powerful example that large datasets can drive myriad new and unexpected opportunities, and it is an argument for funding and building robust systems to manage and store the data.

However, one of the group's frustrations today, unfortunately, is the lack of funding to design and support long-term preservation of data. A large fraction of the data the Rutgers team collects has to be thrown out, because there is no room to store it and no support within existing research projects to better curate and manage the data. "I can get funding to put equipment into the ocean, but not to analyze that data on the back end," says Schofield.

Given the political climate that we are in, there is a need to convince the taxpayers who fund this kind research that it is worthwhile. Thus, we need engagement with the citizens.

Let me provide another example. The Galaxy Zoo uses the Sloan Digital Sky Survey data, which contains around 300 million galaxies. Since there are only about 10,000 professional astronomers, several Oxford University astronomers enlisted public support to help classify them in their Galaxy Zoo project. They have had more than 200,000 individuals participating who have helped classify galaxies and have found various interesting discoveries, such as a blue galaxy called Hanny's Voorwerp. Hanny is a Dutch school teacher. Another example is a new classification of galaxies called "green pea galaxies." Hence, engaging the public is a good strategy.

Then there is the issue of return on investment. Why do scientists want to save the data? To address this issue, I will use this interesting collaboration called the Supernova Factory. Because data curation and management were considered a priority in this project, today the Supernova Factory is a shining example of a significant return on investment, both in financial resources and in scientific productivity. It reduced labor, it reduced false supernova identification by 40 percent, it improved scanning and vetting times by 70 percent, and it reduced labor for search and scanning from six to eight people for 4 hours to one person for 1 hour. Not only did the system pay for itself operationally within one-and-a-half years, but it enabled new science discovery. The important factor in this case is that data curation and management were considered a priority from the very beginning, so the developers built a system that embodied that.

³ At Rutgers University's Coastal Ocean Observation Lab, scientists have been collecting high-frequency radar data that can remotely measure ocean surface waves and currents. The data are generated from antennae located along the eastern seaboard from Massachusetts to Chesapeake Bay.

In his talk, Dr. Gray also listed some calls for action. One of them is related to funding and support of software tools. NSF and other funding agencies are working on developments in support of software tools, but this is always going to be a problem. We can buy some tools off the shelf, but there are always going to be gaps that the commercial software industry does not cover.

I personally came to the conclusion that building complete cyberinfrastructure using open-source software is very difficult. If scientists could buy different pieces of the infrastructure from various vendors and use any relevant open-source software available for them, then they could focus their funding and software development efforts into building the pieces that are not available in the commercial market.

Dr. Gray pointed out that while the biggest projects have a software budget, the small projects generally do not. He felt there was a range of actions that could be done, and that is what he meant by funding generic laboratory information management systems—support going from data capture to publication. His advice was to “let a thousand flowers bloom”—scientific data management research, data analysis, data visualization, and new algorithms and tools.

Figure 1-5 illustrates the data life cycle. It begins with data acquisition and modeling. Because of collaboration, the data are usually distributed. Then the question is how to get the data into a usable form. This would include data analysis, data mining, and visualization. Next is the stage of writing up the data and sharing them in various ways. The final stage in the life cycle is data archiving and preservation, which, for example, can be important for environmental studies where scientists cannot repeat the same measurement. There is no hope of preserving all of the data, though, because this may be too expensive.

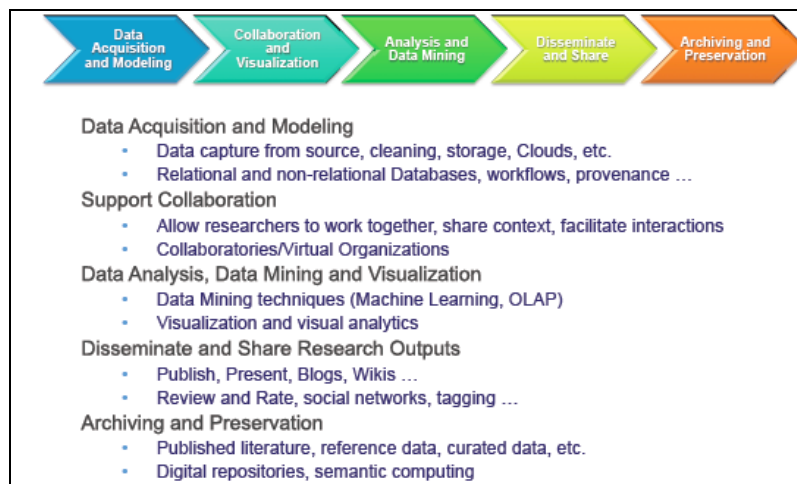


FIGURE 1-5 Data life cycle.
SOURCE: Microsoft Research

I will now talk about a tool called ChronoZoom that we have been collaborating on with Prof. Walter Alvarez at the University of California, Berkeley. What Prof. Alvarez is trying to do is to develop a new “Big History” perspective that begins with the origins of the universe, but makes it possible to look at history on various scales. It will help scientists to put not only biological evolution but also geological evolution and historical events in the system,

and then compare them and see what influenced what in these areas, by looking at it from these different scales. The zoom metaphor is what appealed to Prof. Alvarez, so we built him a tool that enables detailed examination and exploration of huge differences in time scales. This example illustrates that working with such data is not just for the sciences but also for the arts and the humanities.

When I arrived at Microsoft in 2005, Bill Gates was giving a talk at the Supercomputing Conference. Figure 1-6 is one of the slides that he used, which illustrates a new era of research reporting.

Envisioning a New Era of Research Reporting

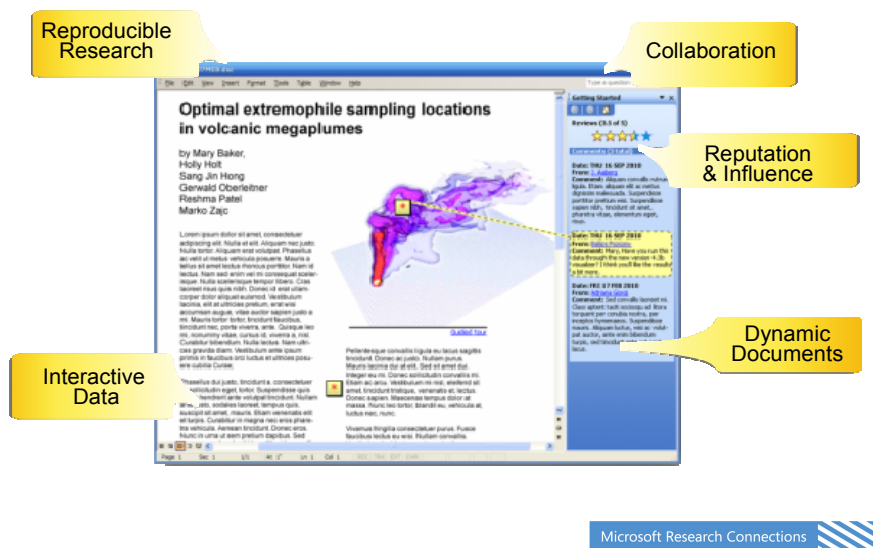


FIGURE 1-6 Envisioning a new era of research reporting.
SOURCE: Microsoft Research

Mr. Gates envisioned that the documents would be dynamic, so that someone could get the data, interact with the data and images, change the parameters, and run it again. Furthermore, besides it being peer reviewed, there could also be a tool like an Amazon rating, or it could be similar to a social network that someone follows. For example, a scientist might always pay attention to what a certain expert from the Massachusetts Institute of Technology (MIT) recommends to read and follow that expert rather than following others in the field. I thought that was a very far-forward-looking vision for Mr. Gates’s talk in 2005.

Another important issue is being able to identify datasets. DataCite⁴ is promoting the use of Digital Object Identifiers for datasets. The scientific community needs to have easier access to research data and to be able to make the data citable. The Open Researcher and Contributor ID (ORCID)⁵ collaboration is trying to solve the issue of contributor name

⁴ DataCite is an international consortium to establish easier access to scientific research data on the Internet, and increase acceptance of research data as legitimate, citable contributions to the scientific record, and to support data archiving that will permit results to be verified and repurposed for future study. (See <http://datacite.org/>)

⁵ ORCID aims to solve the author-contributor name ambiguity problem in scholarly communications by creating a central registry of unique identifiers for individual researchers and an open and transparent linking mechanism

ambiguity—to determine how to know, for example, that Anthony J. G. Hey and Tony Hey are the same person. It is a particular problem in Asia. Thus, ORCID and DataCite are important trends.

There is also a revolution being forced upon us in scholarly communication. Dr. Gray called for establishing digital libraries that support other sciences in the same way that the National Library of Medicine does for medicine. He also called for funding the development of new authoring tools and publication models, and for exploring the development of digital data libraries that contain scientific data and not just the metadata. It does not necessarily have to be done within databases, although that is what Dr. Gray was thinking, but we should certainly support integration with the published literature.

I would like to address the journal subscription issue now, and I will use data from the University of Michigan as an example. The University of Michigan libraries are canceling some journal subscriptions because of budget cuts and the increasing costs of the subscriptions (in some cases, 70 percent of the library collection budget). University Librarian Paul Courant said that about 2,500 were canceled in the 2007 fiscal year. The University Library budget has gone up by an average of 3.1 percent per year since 2004, and according to *Library Journal* magazine, the average subscription price of national arts and humanities journals has increased 6.8 percent per year since 2003. National social science journals increased 9.2 percent and national science journals increased by 8.3 percent. I was trying to get more up-to-date figures, so I contacted a librarian at the university who informed me that there are currently some plans at the university to close some libraries. We should not be closing libraries. We should not spend so much money on subscriptions, especially since the research reported in these journals has already been paid for in some sense.

When I was dean of engineering at the University of Southampton in the United Kingdom, I was supposed to look at the research output of 200 faculty and 500 postdoctoral and graduate students. At the same time, the university library would send me a form every year asking which journals to cancel. If the library cannot afford to subscribe to all the journals in which my staff publish, then how could I start a new field—order some new journals in, for example, nanobioengineering? This is a broken system. Therefore, it seemed absolutely essential to me that a university keep a digital copy of all its research output. As a result, we set up a repository that has now become a university repository.

The university now keeps full-text digital copies of everything that its staff and students produce. It is not just the papers that will eventually appear in *Nature* or other journals, but it is also the research reports, conference papers, theses, and even software and images. No one is asked to violate the publisher's copyright, since these are not necessarily the versions that appear in print. About 70 or 80 percent of publishers allow keeping a nonfinal version on the Web. What is important is to get the scholarly products captured at the time of production with the relevant metadata so that it becomes easy for the researchers to access these materials.

between ORCID and other current author identification schemes. These identifiers, and the relationships among them, can be linked to the researcher's output to enhance the scientific discovery process and to improve the efficiency of research funding and collaboration within the research community.

A fellow dean at Virginia Tech told me that he had required electronic theses to be online in 1997. Virginia Tech had 200,000 requests for PDFs that year. Less than 10 years later they had 20 million requests. These are extraordinary figures. It shows that making your data available is important.

I do not advocate Webometrics as a way of ranking departments or universities, but looking at just the Google Scholar ranking, which looks at papers and citations, we get the list shown in Figure 1-7.

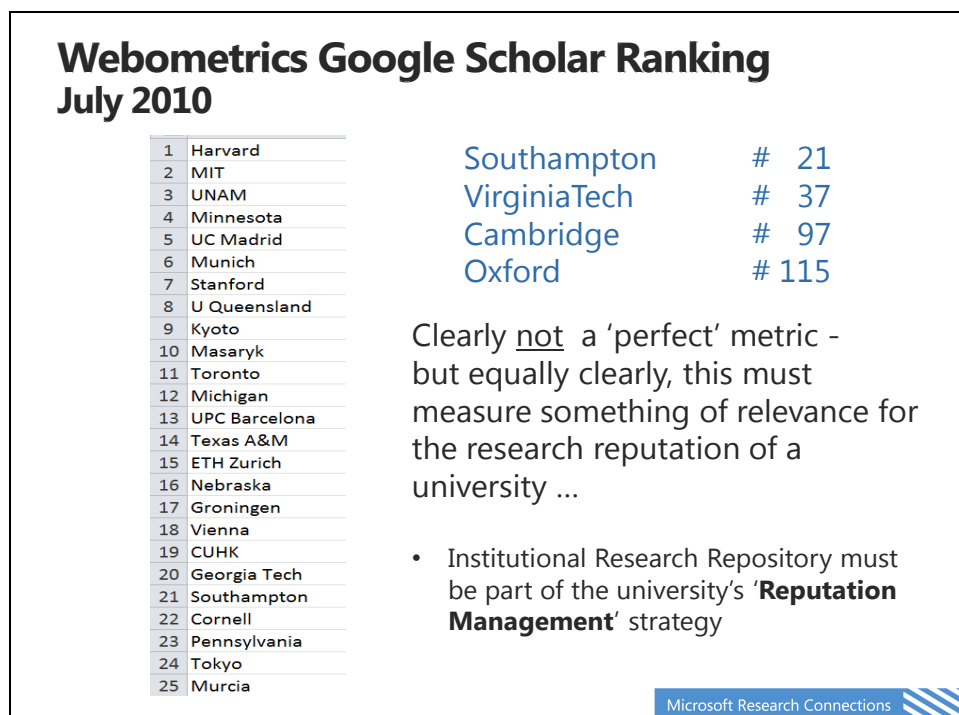


FIGURE 1-7 Webometrics Google Scholar ranking.

SOURCE: Microsoft Research

Harvard University and MIT are at the top. But there is also a large amount of apparent randomness in this ranking. According to citations in Google Scholar, the top university in the United Kingdom is Southampton, at number 21, well above Cambridge and Oxford. This might appear hard to believe, but it is saying something serious about research. If a university is regarded to be a serious research university, it should be showing up on such ranking, because people like such tables. It is a question of reputation management, and I believe that the university research repository is an important piece of that.

As to the future of research repositories, I think they will contain not only full text and the other kinds of scholarly products that I have talked about but also data, images, and software as part of the intellectual output. Some of those items may have an identifier attached that can be used, for example, in a court case. In my view, the university library should be doing that.

Let me offer two examples of how work in this area can be done. There is a large amount of data in databases at the National Library of Medicine so that researchers can have a "walled garden." Dr. David Lipman and colleagues at the National Institutes of Health have

created a system that allows anyone not only to have access to the text provided through PubMed Central but also to have a range of databases and cross-database searching. That is a centralized repository model.

Most fields are not going to be fortunate enough to have such a centralized model, because it is expensive; therefore, it is important to develop a more distributed model. One example is *WorldWideScience.org*. Dr. Walter Warnick and his colleagues at the U.S. Department of Energy are building a consortium whose membership is now up to 65 nations. It offers many publications and databases online, and there is a translation service that allows people to type a query in English, have it translated into Chinese or Russian, and after the search is done, translate the results back into English. The major advantage of this search system over a Google search is that this system can get up to 96 percent unique results that cannot be captured by a simple Google search. This is an example of a distributed system that is getting close to Dr. Gray's vision.

We all know about Vannevar Bush's article "As We May Think"⁵. In a similar vein, the physicist Paul Ginsparg wrote a good article called "As We May Read"⁶. Prof. Ginsparg started arXiv at Los Alamos National Laboratory for use by high-energy physicists. His idea was to reproduce the way that physicists used to circulate preprints back when papers were typed. They used to make photocopies of papers they were submitting to, for example, *Physical Review* and send them around to a hundred institutions. Since he had access to a server, he suggested to other physicists that they send digital copies of their papers to him and he would circulate them. That is essentially how the arXiv started. The project is now located at Cornell University.

In the conclusion of his paper, Prof. Ginsparg wrote, "On a one-decade timescale it is likely that more research communities will join some form of global unified archive system without the current partitioning and access restrictions familiar with the paper medium for the simple reason that it is the best way to communicate knowledge and hence to create new knowledge. Ironically, it is also possible that the technology of the 21st century will allow the traditional players, namely the professional societies and institutional libraries, to return to their dominant role in support of the research enterprise."⁷

Next, to talk about future research cyberinfrastructure, I thought I would go back in history and look at a slide from 2004. These are the six key elements for a global *cyberinfrastructure*, as opposed to *e-Infrastructure*, which was the term used in Europe.

⁵ Atlantic Magazine, July 1945.

⁶ The Journal of Neuroscience, 20 September 2006, 26(38): 9606-9608; doi: 10.1523/JNEUROSCI.3161-06.2006

⁷ Ibid.

Six Key Elements for a Global e-Infrastructure for e-Science (2004)

1. High bandwidth Research Networks
2. Internationally agreed AAA Infrastructure
3. Development Centers for Open Software
4. **Technologies and standards for Data Provenance, Curation and Preservation**
5. **Open access to Data and Publications via Interoperable Repositories**
6. Discovery Services and Collaborative Tools

Microsoft Research Connections

FIGURE I-8 Six Key Elements for a Global e-Infrastructure for e-Science.
SOURCE: Microsoft Research

I highlighted the points on technologies and standards for data provenance, curation and preservation, and open access to data and publications by interoperable repositories. I am not sure we have made much progress since then, but it was interesting that we were thinking along those lines 7 years ago. We funded a number of projects. We set up a digital curation center in 2004. It took the people there some time to find their way, but I think they are doing useful work now. They advise people about data management plans, for example. We also set up a software engineering organization called Open Middleware Infrastructure Institute. Its job was to take some of the software we produced in our e-science projects that got the most use and then do testing, documentation, writing specifications, and finally make the software available. The institute worked to make software usable to scientists who had not written it. It brought the level of the open-source software up to a reasonable engineering level. The NSF should be helping its grantees do similar work.

Another example is the National Center for Text Mining (NCTM). My colleagues in the computer sciences area claim that the technology is now mature. I do not think that is the case. I do not think we have proven that intelligent text mining is used by the community yet. The NCTM can be considered as an experiment that is only halfway successful in this area.

I should also mention JISC, which is an NSF-like funding agency in the United Kingdom. JISC funded many pioneering activities in digital libraries and repositories. I was the chair of the JISC Committee for Support and Research. One project that JISC funded was the TARDIS project, which led to the repository at Southampton. Another JISC project was the MALIBU project, which was about hybrid libraries and the invisible Web—a lot of information is locked behind Z39.50 and not accessible to Web browsers, so how could anyone find it? The final example was the CLADDIER project with the British Atmospheric Data Center, which was concerned with linking data to publications, and, again, it was very interesting.

Let me now focus on where we are going in the future. At present, about two papers are being deposited in the National Library of Medicine every minute. No one can keep up with this rate, so how do we deal with that deluge of publications? That does not even include the data generated for those papers. I think that semantic computing will help. Currently computers are great tools for storing, computing, managing, and indexing huge amounts of data. In the future, we will need computers to help with automatic acquisition, aggregation, correlation, interpretation, discovery, organization, analysis, and inference from all of the world's data. If the system knows that a scientist is looking for information about a star, for example, it should be able to identify which papers contain data about that star and get those papers for the searcher without being asked. It should "understand" what we are looking for. The computer should be computing on your behalf and understanding your intent. These are the important things that we need to do.

I do think that we are moving to a world where all the data are linked, and that at some point a scientist will be informed that paper X is about star Y without even searching for this paper. This is what semantic computing will be able to do. This is coming and is an exciting challenge for the information technology industry and for computer science to see if they can build systems that are useful.

I also think that, in the future, researchers will keep some data and tools locally, but they will also use some services from the cloud. We already know about blogs and social networking in the cloud as well as identity services. Amazon and Microsoft offer storage and computing services. Thus, depending on the area or application, scientists may want to use some kind of cloud service. For example, a scientist may use the cloud to replicate the data to a safe place to store them or to share them for collaboration purposes. Semantic computing is also coming, and I think that we are going to realize Dr. Gray's big digital global library that contains data and publications.

Discussion

DISCUSSANT: There is a scientist at Rutgers University named Christopher Rose who is known for, among other things, his paper about why the search for extraterrestrial intelligence is not worthwhile. His point is that the energy required to fill the universe with so much energy that somebody in another galaxy can pick it up is absolutely unbelievable. He also said that the only thing that makes any sense is that if someone throws an object toward a place outside his or her galaxy, it means that this person has to produce half the energy, and the recipient has to produce the other half in the form of a radar system that detects it. When I hear someone saying that we are going to fill the world with knowledge in the hope that you pick up some of it as it passes your head, I wonder if there is not something to be said for the traditional system of trying to put the person who is producing a bit of chemical knowledge in a room with the people who might use chemical knowledge and see if they recognize the intersection.

DR. HEY: It is a vision, and visions can be wrong. I think that we can do much better. We spend a lot of money in the biological sciences, for example, with everybody repeating similar work.

Also, there was a medical study recently released that found that almost 400 medical research projects did not cite obviously relevant papers on the same subject. I think we can do better. I do not think we will fix all the problems, but I do think we are all fortunate enough to have a National Library of Medicine.

DISCUSSANT: Dr. Hey, if you look at where the scientific community is going with publications and data linkage, arguably this is already being done in the real world. If you go to the *New York Times* Web site, for example, or other publications' sites, you see that the private sector has provided the engine to have the kinds of searchable ranked articles that we would like to have in the scientific community. What do you think it will take to close the gap? And what is the next level of developments that we should be aiming at today?

DR. HEY: If there are commercial systems that work, then we should use them, and we should not try to reproduce tools and open-source software with our limited research funds. It is absolutely important to have standards and interoperability, but there are experiences that we can learn from the commercial world.

Where we are going? Supercomputer centers produce huge amounts of data, and it will be difficult to manage and transfer the petabytes of data that they produce, so I think that we will have supercomputer centers with data centers. I think there will probably be some regional data centers. There are currently attempts to put some of this infrastructure in place, and I think there is an opportunity for the United States to be leading in this exercise. I am pleased to see that NSF is galvanized to do something in this area.

DISCUSSANT: Much of what was shown in the presentation, including the slide from Bill Gates's talk in 2005, now exists. Tools like Mendeley for sharing publications and fancy XML documents with objects embedded inside of documents with different formats do exist. But in astronomy, maybe 1 percent of the astronomers use them and maybe 5 percent know that they exist. What did we learn about this culture shift that is needed to let us use tools and technologies more efficiently?

DR. HEY: I absolutely agree. We used to produce useful tools and give them to the users, but we found that they were not using them 6 months later. Even though we would show them how useful they were and they had all agreed, they did not use them. There is a real problem in engaging scientists. We have to produce tools that are as close as possible to the way that scientists are working now, but that at the same time give them an improvement. Advocates are needed, and it will be important to have some people showing how scientists can do great research with these tools.

2. Experiences with Developing Open Scientific Knowledge Discovery in Research and Applications

Case Studies

International Online Astronomy Research

Alberto Conti

-Space Telescope Science Institute-

I am the archive scientist at the Optical UV Archive at the Space Telescope Science Institute, which is the place that operates the Hubble Space Telescope for the National Aeronautics and Space Administration (NASA) and which is going to operate the successor of Hubble, the James Webb Space Telescope. As archive scientists, we have to do a lot of work to bring data to the users, but we focus mainly on three areas. We like to optimize the science for the community by not just storing the data that we get from users but also trying to deliver value-added data. We also try to develop and nurture innovation in space astronomy, particularly for Hubble and its successor. We are good at that because we have been doing it for 20 years. And we like to collaborate with the next generation of space telescopes as well as with ground-based telescopes, because that field is quite busy.

Figure 2-1 shows some key astronomical missions, which provides an idea of the kind of astronomy and astrophysics data that exist and are being developed.

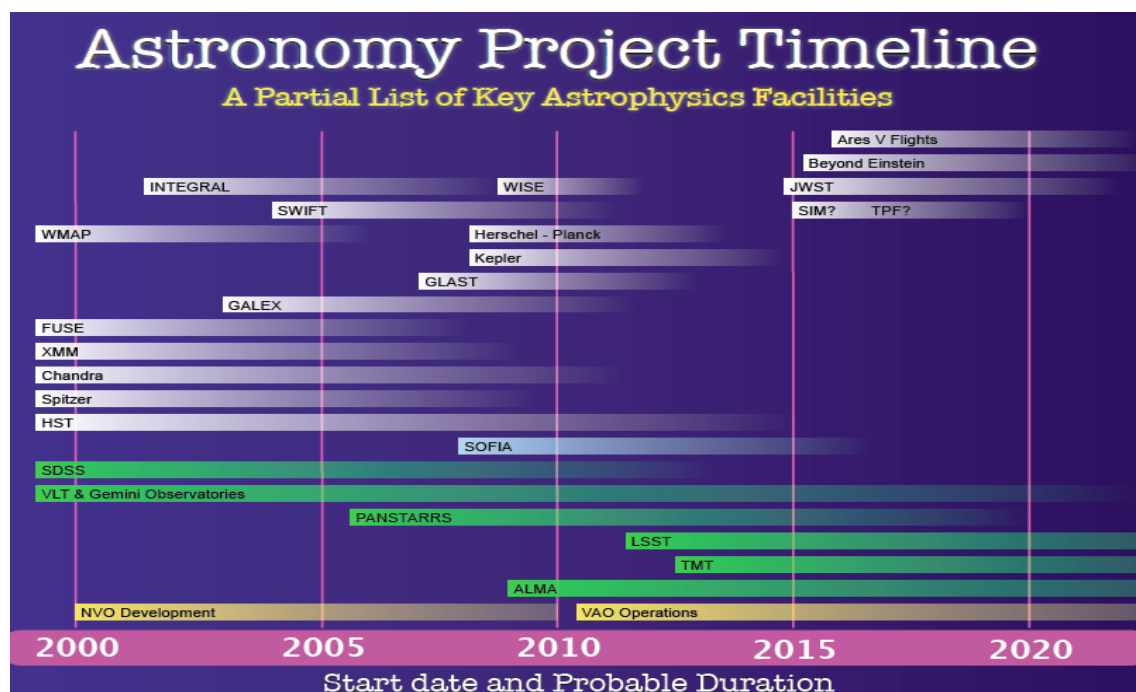


FIGURE 2-1 Astronomy project timeline.
SOURCE: Space Telescope Science Institute

The Kepler mission, for example, is a planet-finding mission. The ones on top are space based, and the bottom ones are ground based, some of which are going to produce very large amounts of data. At my institution, we collaborate directly or indirectly with most of these missions and deal in some way with the data they produce. How are we going to manage the large amounts of data from these missions?

Over the past 25 years, astronomy has been changing quite radically. Astronomers have been very good at building large telescopes to collect more data. We have been much better at building larger mirrors. But we have been a hundred times better than that at building detectors that allow us to collect extremely large amounts of data. Since the detectors roughly follow Moore's Law, every year or so our data collection doubles, and this raises many important issues.

We realize that we are not alone in this area. We know that fields from earth science and biology to economics are dealing with massive amounts of data that must be handled. I am not particularly fond of using the word “*e*-science” to describe this field; I prefer to speak of “data-intensive scientific discovery,” because I think that is exactly the field that we should be moving to, because we are going to be driven by data.

However, while astronomy is similar to other fields in managing big volumes of data, the astronomy field is somehow special—not because the data are intrinsically different and special in their own right, but because they have no commercial value. They belong to every one of us, so they are an ideal test bed for trying out complex algorithms that are based on very large dimensions. Currently there are missions that have in excess of 300 dimensions, which can be very useful if a scientist wants a dataset with many dimensions to analyze.

We also have to be aware that things have been changing, for example, in the geographic information system world, where our perception of our planet has been changed by such tools as TerraServer, Microsoft's Virtual Earth, Google Earth, and Google Maps. The way that we interact with our planet is now different than it was, and as we use all of these tools, we have no concept of what is really going on underneath, because we are focusing on the discoveries. Therefore, some of us are trying to do the same thing for data in other areas.

What is this new size paradigm? Why is so much data a problem?

In Figure 2-2, the red curve represents how much data we collect in our systems as a function of time. The big spike in the middle corresponds to the installation of new instruments for Hubble, for example. This is a trend, but it is not a problem. The real problem is the curve above it, which is the amount of data that we serve to the community. This is a large multiple of the data that we collect, so this potentially can be a problem.

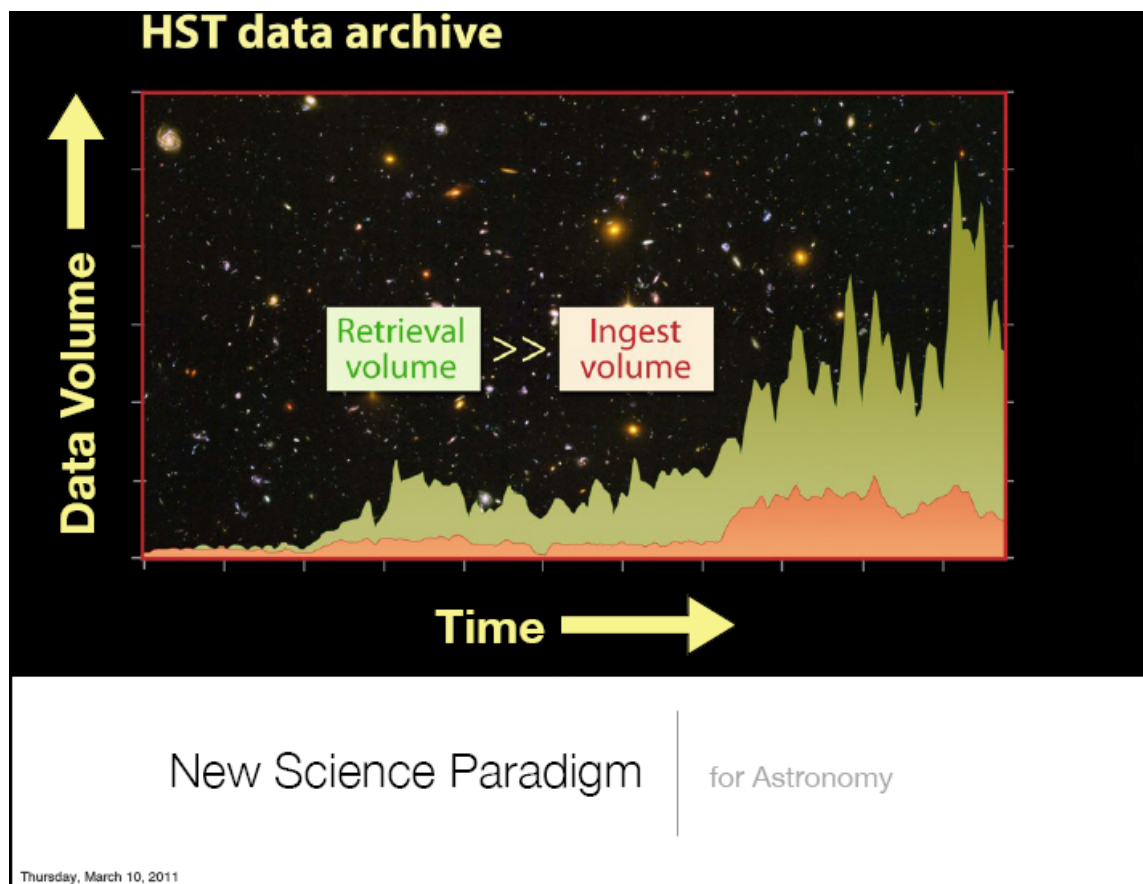


FIGURE 2-2 New science paradigm for astronomy.
SOURCE: Space Telescope Science Institute

This was a concern, at least until a few years ago, because of how we used to work with the data. As a researcher with the Space Telescope Science Institute, I can use a few protocols to access the very different data centers, but every time I interface with a different data center, I have to learn the jargon—the language of that data center. I have to learn all the idiosyncrasies of the data in this particular data center. After I learn all this, my only option is to download the data to my local machine and spend a long time filtering and analyzing them. In astronomy this limits us to small observations of very narrowly selected samples.

In 2001, scientists realized that this was not a good model, so the National Science Foundation (NSF) funded the National Virtual Observatory for Astronomy. The goal was to lower the barrier for access to all the different data centers so that a scientist did not have to worry about the complexities behind the data centers. They just had to worry about how to get data and do their science. The National Virtual Observatory was established in 2001, and for several years it was in the development phase. It moved into the operational phase with funding from NSF and NASA in 2010. By this time it was called the Virtual Astronomical Observatory.

At the same time these developments were taking place in the United States, people across the planet were building their own standards, their own protocols, and their own data centers. At some point it became evident that we could not ignore what was happening around

the world, because many of the ideas that were being proposed to deal with this deluge of data were very worthwhile and could be integrated into the thought process that goes on inside of any astronomical archive. Hence, in the middle of developing the virtual observatory, a collaboration called the International Virtual Observatory Alliance (IVOA) was begun with the goal of lowering the barrier for access at all of the data centers across the world.

Today the IVOA allows scientists to access all these different missions and to become more knowledgeable about what each mission does and how they are different from each other. It also has helped scientists identify where data are stored and what they are called. This is a major leap, but it has come at a cost, and the cost is that we have developed very ad hoc data standards for storing and sharing the data. After a long development process, we adopted protocols that turned out to not be very effective at certain times. Furthermore, after that many years of development, we do not have very effective tools, because doing the data mining is very difficult. We are still in the mode where most of the data have to be downloaded and filtered locally, which is not what we would like to do. Although we have standards and protocols, the IVOA has come under some pressure lately because it is not changing perhaps as fast as some would like.

To some people's credit—particularly Dr. Alex Szalay of Johns Hopkins University, who is one of the founding fathers of the virtual observatory idea—it was predicted that this transition would be chaotic, and it has been chaotic. It is very hard to integrate very large data centers across the planet into a seamless process of discovery, but some people are trying to do that. The chaotic part has come to pass. Unfortunately we do not yet have a uniformly federated and distributed regime of sources across the planet that we can access in a seamless manner, so this is part of our existing problem.

In the future, I think the IVOA would still be at the core of developing such a distributed system. Some people, like me, believe that the IVOA should have a much smaller and limited role in defining the requirements and standards for this distributed system that allows people to access and process on a cloud-like system. The IVOA should try to build standards that are not ad hoc, but rather are based on industry standards. Metadata is another important frontier that we must explore so that we can characterize these data and mine them appropriately. The idea is that we want to enable new science and new discoveries.

Here is what I see as the challenges we are facing. We capture data very efficiently, but we need to reduce obstacles to organizing and analyzing the data. We are also not very good at summarizing, visualizing, and curating these large amounts of data. Another issue is that most of the time the algorithms that produce the papers that we see are never associated with the data. We should consider this as a problem. A scientist should be able to reproduce exactly the research result once he or she finds the data.

Semantic technology is key to being able to deal with many of these obstacles. We have to be prepared, because this will introduce a fundamental change in the way astronomical research is conducted. Another issue is that we do not have available to us an intensive data-mining infrastructure, and if we are to build one ourselves, that will be very expensive. Many of these efforts in astronomy are not funded at a level that would allow us to build our own. A solution for handling very large datasets is emerging. Such solutions are not pervasive in astronomy, which is also a problem. As for hosting, if we host our large datasets on the cloud

for redundancy, for performance, and so on, that will come at a cost, and some archives might not be able to sustain that cost even if it does not seem to be very large.

Finally, I think we have the opportunity to do a few things. We have to be a little humble and accept that we have to ask for help from computer science, statistics, applied mathematics, and from others who can use our high-dimensional data for their purposes, and to apply their useful algorithms. We should leverage partnerships with those who are already doing this work on the Web scale. There are companies—Microsoft, Google, Amazon, and many others—that are handling and trying to solve these problems, and we should definitely partner with them. They may not have *the* solution for us, but they will have a big part of some solution.

Integrative Genomic Analysis

Stephen Friend
-Sage Bionetworks-

There are three points that I would like to focus on today. One concerns the current paradigm shift in medicine, and I am going to make a case for why we need better maps of diseases. The second—and probably most important—point is that it is not about *what* we do, but *how* we do it: It is all about behavior. Third, I want to discuss some aspects of building a precompetitive commons.

When we are working with commercially valuable information—unlike the case with astronomy, there is serious money involved with information in the medical area—a number of issues arise. The cost of drug discovery, which is at the intersection between the knowledge and being able to impact people, is around \$1 to 2 billion per molecule. There is about \$60 billion being spent every year to make drugs, but there are fewer than 30 new drugs that are being approved. People are beginning to realize this lack of efficiency, however, and the good news is that there is information coming along that makes discovery easier. To put it in perspective, even after drugs are approved their rate of effectiveness is low. This is true for cancer drugs and also for the statins used to treat high cholesterol levels. Only 40 percent of patients who take statins have their cholesterol lowered, while more than half do not, and of those who do, less than half see any benefit in decreased occurrence of heart attacks, yet we still give these drugs and say everybody has got to be on a statin.

The problem is that when trying to figure out what is going on inside cells and to understand diseases in terms of pathways, we are using maps that are akin to pre-Kepler models of biology. People get very excited about understanding pathways and believe that if they could just understand the pathway of a disease, they could make sense of it. Almost every single protein has a drug-discovery effort associated with it, and all of these researchers are dreaming that they will be able to make a drug against this or that disease without understanding the context and the circuitry that underlies it.

We have gone through these transitions before. We got beyond the idea that God throws lightning bolts, and we got beyond the concept of Earth being at the center of the universe. In the next 10 to 15 years, we are going to have a change that is similarly dramatic and will be remembered for centuries. It is going to be driven by a recognition that the classical way of thinking about pathways is a narrative approach that we favor, because our minds are wired for the narrative approach.

The reason this change will happen is that technologies are developing—analogue to telescopes—that will allow us to look inside of cells. In the last 10 or 15 years, we have gone from the first sequencing of a whole genome to the point where it will soon be cheaper to get your genome sequenced than it is to get a CT scan or a blood test. Think of that in terms of the amount of information associated with it. It will take a few hours to get your whole genome sequenced, and it will be under \$1,000. The cost of CT scans is going up, while the cost of sequencing genomes is going down.

I know of a project in China that is planning to do the whole genome sequence of everyone in a village with a population of 1 million people. They believe they will get whole genome sequences on the village population in the next 8 years.

The problem we have is that most people are still looking at altered component lists, and astronomers have already gone through this. Altered component lists do not equal knowledge of what is happening. While people are excited about the sequencing of genomes and understanding the specifics, the reality is that we are stuck in a world where biological scientists are still generating one-dimensional slices of information and not understanding that those altered one-dimensional slices of information will not add up to understanding altered function, which is ultimately what you have to do. The reason that this is important is that the circuitry in the cell has, over a billion years, been hardwired to buffer any type of changes, and the complexity that the system has in it to tolerate a modification is hardwired together within one cell, within multiple cells within organs, and so on.

A project started in 2003 at Rosetta in Seattle asked how much it would cost to build a predictive model of disease that used integrated layers of information. By integrated, I do not mean raw sequence leads leading through to a genome. I am talking about integrating different layers of information. An experiment done by Eric Schadt represented a fundamental seismic shift in how we think of this science when it is asked: “Why not use every DNA variation in each of us as a perturbation experiment?”

Imagine that we are each living experiments and that if we know all of those variations, we could take a top-down approach similar to the blue key that is on your computer that feeds any error through to Microsoft or other places. Imagine a way in which we could build knowledge not from looking at publications. This would be very important. I will argue that publications are a derivative, and while they are helpful, that is not where we want to be focused. We should be looking at the raw data and building those up.

Imagine that we could look at putative causal genes, that we could look at disease traits, and that we could be zooming across thousands of individuals and asking what happens at intermediate layers and what happens with causal traits. The central dogma in biology, which is that DNA produces RNA, which in turn produces proteins, allows us to build causal maps that enable us to look collectively at changes in DNA, changes in RNA, and effects on behavior and then begin to ask, “What was responsible for doing what to what?”

There are many mathematical models, from causal inferences to coexpression networks, but it does not matter which statistical tools we use to do top-down systems biology. The result is that we do such tasks as taking brain samples and expression, building correlation matrixes and connection matrixes, then clustering them and building network models that use a key driver of a particular effect, not because it was in the literature, but because it was experimented with in different databases.

A key paper that used this approach was published in *Nature Genetics* in 2005. It was looking for the key drivers of diabetes. The idea behind the paper was to rank the order of the 10 genes most likely to be responsible for going from a normal state to diabetes and from diabetes to normal. Over the next 4 years, that work was validated by the community, which found that 90 percent of the genes on that rank-ordered list were correct. That is unheard of. Those trivial early maps did not have nearly as much power compared to what is coming, but they were good enough to make that type of analysis.

With such an approach, we could look at drugs that are on the market and drugs that are in development and try to identify those that are going to be responsible for certain subtypes of effects on behavior or on cardiovascular function, or we can look for toxicity. This is the type of shift that is emerging in the medical world for understanding compounds. There are already about 80 or 90 papers that have been published in different disease areas. The laboratory of Atul Butte at Stanford is one of the 10 labs that have made an art out of exploring human diseases by going into public datasets and building these types of maps.

The first take-home message is that we have been living in a world where we think of narrative descriptions of diseases—this disease is caused by this one agent. At Sage Bionetworks, we now use maps that are not hardwired narratives, but that look at statistical probabilities and the real complexity. For example, imagine that some components in the cells talk to each other and they need to listen to each other to make a decision on what physiologic module is coming out of that component of the cell that then talks to other parts of the cell. What we are finding is that key proteins aggregate into key regulatory drivers that are able to say, “This is what I need to do.” This is the command language. We are not looking at the genetic language, but rather at a molecular language that determines what is going on within a cell, that says, “I have enough of this,” and interacts with cells in different parts of the body.

The complexity of this approach is outstripping our ability to visualize it and to work with it, but the data are there. Those key ingredients are coming. Maps of diseases will be very accurate for who should get what drugs, and what drugs are able to be developed. This is going to totally revolutionize the way we think of disease. Disease today is still stuck in a symptom-based approach. If you are depressed, you have depression, and you try to find a drug as if all depression were the same. The same is true for diabetes or cancer. The level of the molecular analysis that we will soon be able to do is going to be important, but substantial progress will require significant resources and decades to complete.

Those maps are almost like thin peel-off slices of what is happening, and we realized that the group doing this work could not be one institution or one company. For that reason, a year and a half ago I left the job I had running the oncology division at Merck and started a not-for-profit foundation to do a proof of concept. This foundation, which has about 30 people and a budget of about \$30 million, is asking how to do the same work for genomics that has been done for astronomy. We are now trying in genomics to do the virtual equivalents to what has been done in astronomy. The people who are driving it are Lee Hartwell; Hans Wigzell from the Karolinska Institute; Wang Jun, the visionary who is sequencing the million genomes in China; and Jeff Hammerbacher, who was the chief architect of Facebook. We brought Hammerbacher in because we think that this is not strictly about science. We think we need to ask why a young investigator would want to share his or her data. Why would that investigator allow someone else to work on them? We think the social aspects of this are very important.

I will offer four quick examples. In breast cancer research, we are working on ways to identify key drivers by using publicly available datasets and running coexpression and Bayesian networks. We are going to be publishing that work soon. The second concerns cancer drugs. It turns out that virtually all of the cancer drugs that are the standard of care, the first line of drugs that a patient gets, are usually not very effective. We are looking at Cisplatin doublet therapies for ovarian cancer, Avastin for breast cancer, and Sutent for renal cell carcinoma. We realized that we should identify who is not responding to those drugs so that these patients can be given something else. It has been very difficult to get the National

Institutes of Health and the Food and Drug Administration interested, so I had to go to China for funding.

The third example involves getting data from pharmaceutical companies. The richest trove of data is inside these companies, because they often collect genomic and clinical information when assembling information for a drug filing; but they lock all these data up, because they do not want to share what it was that gave them their advantage. Accordingly we took a different approach. We asked to use the data from the comparative arm rather than the data for the investigational drug. This helped, and all of these companies have said they will give us these data.

That brings up the point that Dr. Conti was making earlier—that when we have all these data, we must worry about the annotation and curation. That is what we are doing. We are saying, “You give us the data, and we will annotate, curate, host, and make them available to anyone.” There is a very big difference between accessible and usable data. We tell the pharmaceutical companies that if they make the data accessible, we will make them usable. I am sure we are going to do it in a bad way that has to be redone, but at least we will get it started.

The fourth example is a project in which we took five labs that are supercompetitive with each other and linked them together so that they could share their data, models, and tools. This type of linking opens up new opportunities. We found that we have broken the siloed principal investigator mentality of science. This is part of that fourth dimension. Science is presently practiced like a feudal system: “I have the money, those are my postdocs, and this is my data.” But imagine a world in which scientists could go anywhere they wanted to get anyone’s data before it is published in the same way that astronomers work. We have found that it is the scientists under 35 years old who know what to do. We have been making openly available global datasets, models, and tools.

I want to describe the behavior we have run into. First, clinical genomic data are accessible, but minimally usable, because there is no incentive to annotate and curate the data. Second, people who get the data think they own the data. If someone pays them to generate the data, somehow they think they own them. We are living in a hunter-gatherer community, which is preventing data sharing, and, for the most part, the principal investigators have no interest in changing that situation. This needs to fundamentally change.

Just as Eisenhower in the 1960s talked about the military-industrial complex, I think we now have a medical-industrial complex in which the goals of individual researchers to get tenure, the goals of institutions to improve their reputation, and the goals of the pharmaceutical industry to make money create a situation in which the patients are not benefiting. There is a time for intellectual property, there is a time for innovation, but the current situation is crippling our ability to share the data. The system is not set up to make it easy to reproduce results, as in astronomy. You do not hold people accountable when they publish a paper to explain what they did so that it can be reproduced. We have to figure all of that out.

Finally, I want to ask why we do not use what the software industry has already figured out for tracking workflows and versioning. We have to change the biologic world so that the evolution of a software project is clear, with a code repository and the branches and releases. What we are doing now at Sage Bionetworks is determining the details on the repositories, collaboration, workflows, and cloud computing. We find that the world is saturated with early

efforts, so it would be absurd to think that we ourselves need to do this. We are interested in such efforts as Taverna, Amalga, and work at Google. We are identifying who has what so that we can stitch this together. The hardest thing is that we do not have the support of patients in this area yet. We have to bring the patients and citizens in. Otherwise the data-sharing, privacy, and access issues are going to tear this approach apart.

Geoinformatics: Linked Environments for Atmospheric Discovery

Sara Graves

-University of Alabama in Huntsville-

I will be giving this presentation on behalf of Dr. Kelvin Droegemeier, who had an emergency to which he has to attend. Dr. Droegemeier's area is geosciences. He is from the University of Oklahoma and serves on the National Science Board. He was planning to talk about a project called Linked Environments for Atmospheric Discovery (LEAD). I worked on LEAD as one of the coinvestigators, as did Mohan Ramamurthy. LEAD was a multi-university and multidiscipline research program that was created by the National Science Foundation (NSF).

I heard on the weather report this morning that 100 million people will be affected by a severe weather situation in the Northeast and other parts of the country, so I thought that it was quite fitting to be talking about LEAD. The annual economic losses due to weather are greater than \$13 billion per year. Every time we try to get people ready for some weather event, there is a cost to that preparation, and if the weather event does not occur as predicted, there are additional costs.

We have a few weather technologies today that we are familiar with, and we use the output from many of them. According to some meteorologists, however, we still have many technological improvements to make. Radars do not adaptively scan, so one of the keys that we were looking into with LEAD was how to get more adaptation and more dynamic processing of information. We wanted to take into account the complexity of all of the factors and not just approach it linearly.

Operational models run largely on fixed schedules and fixed domains, and the cyberinfrastructure that we often use is quite static. We teach our students about current weather data, but we do not allow them to interact with it. The previous two talks discussed how to get more people involved, not just in an academic setting but also in citizen science. That is certainly something we would like to do in this domain as well. Weather is local, high impact, heterogeneous, and rapidly evolving, yet our technologies and our thinking have not evolved that quickly. One of the fundamental research questions we looked into with this project was whether we can better understand the atmosphere, educate others more effectively about it, and forecast more accurately if we adapt our technologies. Even animals adapt and respond, so why should this not be true for the resources that we have?

This was a multi-university project, and there was an associated project called Collaborative Adaptive Sensing of the Atmosphere (CASA), which I will describe shortly, that interacted with LEAD. The goal of LEAD was to revolutionize the abilities of scientists, students, and operational practitioners. The operational practitioners were a key part of the vision, because we were trying to work not just with the typical student scientist but also with the National Oceanic and Atmospheric Administration and some other practitioners to observe, analyze, predict, understand, and respond to intense local weather by interacting with it dynamically and adaptively in real time. Oklahoma and Alabama have many tornadoes. We did not have this project just for that reason, but severe weather does have a local impact.

What does adaptation mean? Let us take the example of a tornadic storm that occurred in Fort Worth in March 2000. A local TV station radar system and a National Weather Service system with 12-hour computer forecast that was valid at 6 p.m. central time did not explicitly show any evidence of precipitation in north Texas. The reality was quite different, because there was in fact a tornado.

The approach that the project took was to have streaming observations. We wanted to have constant observations and not just frequent static pictures. We did a lot of mining of both the streaming data and some of the archived data to feed into the Weather Research and Forecasting (WRF) Model. We needed on-demand computing, so we were using the NSF TeraGrid computational capabilities. The cloud could be used in that respect too. What comes out of the computation models is fed back into the mining engine, which creates a feedback loop.

What does it take to make all of this possible? Of course, you need adaptive weather tools such as adaptive sensors, adaptive cyberinfrastructure, and a user-centered framework in which these tools can mutually interact. We developed and continued working on these tools with LEAD. What LEAD was doing is creating a Web service. We had our linked environments, our processing capabilities scattered around distributed areas, and Web services to produce some of this information much faster. You also need to link together the services and workflows.

One of the lessons that was learned from this project was the recognition that we can create very elaborate workflows and save them in a special database so that it is possible to go back and recreate some of those problems, both for academic purposes and for practitioner and operational capabilities. We did significant work in creating the workflows and some of the tools. Since all of this work is openly available to others, many of these tools and processes are being used today by various entities.

To develop more accurate forecasts, we needed to come up with better tools. We did a lot of ensemble modeling, and we have continued to develop this capability with WRF.

Figure 2-3 shows the current operational radar system in the United States.

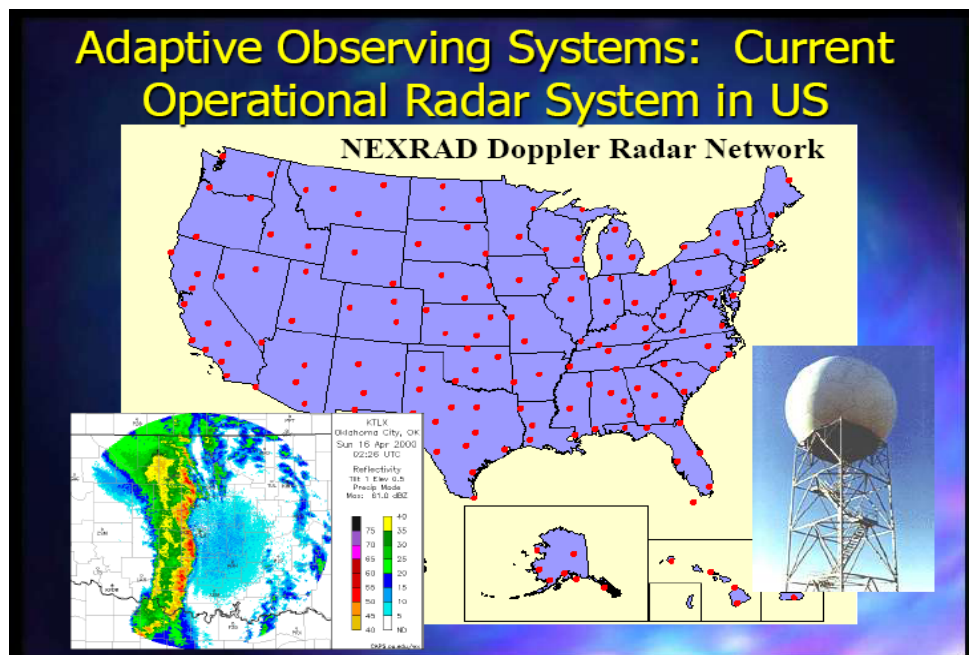


FIGURE 2-3 Current operational radar system in the United States.
SOURCE: Kelvin Droegemeier

There are limitations to NEXRAD (Next-Generation Radar), as it is not set up for some of the issues I have discussed. It is fairly independent of the prevailing weather conditions because of the way it operates, and it is very independent from the models and algorithms that use the data, since the Earth's curvature prevents 72 percent of the atmosphere below 1 kilometer from being observed. There is also a next-generation capability being developed now.

The other project that we worked on was mainly headquartered at Colorado State University, and Dr. Droegemeier was also involved with it. It was focused on collaborative adaptive sensing of the atmosphere. We would mine and find an event that is occurring and that we could see with the radar data. Then, we would try to steer some of these radars to detect the event through dynamic adaptation. Figure 2-4 depicts the Oklahoma test bed where they had the various capabilities.

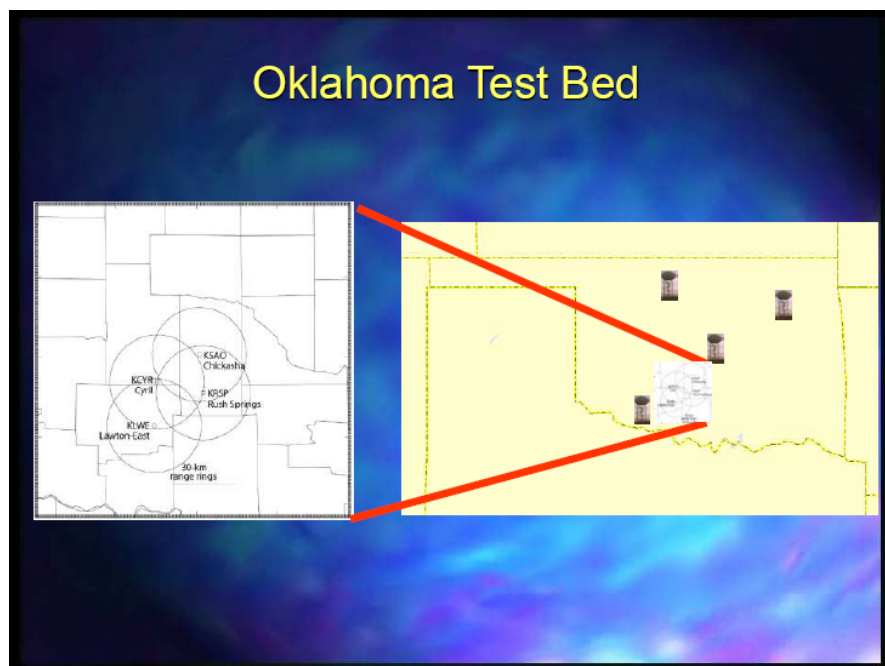


FIGURE 2-4 Oklahoma test bed.
SOURCE: Kelvin Droegemeier

Figure 2-5 is an example of adaptive sampling.

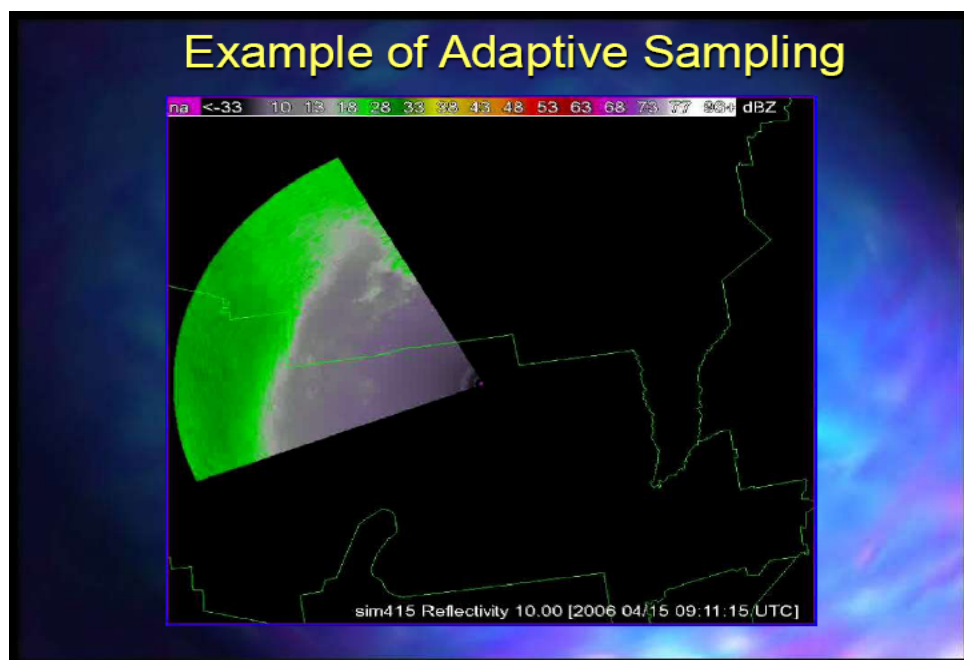


FIGURE 2-5 Example of adaptive sampling.
SOURCE: Kelvin Droegemeier

And finally, Figure 2-6 shows the CASA experiment looking at what they were seeing from the radars and the way they could steer them versus what you could get with NEXRAD.

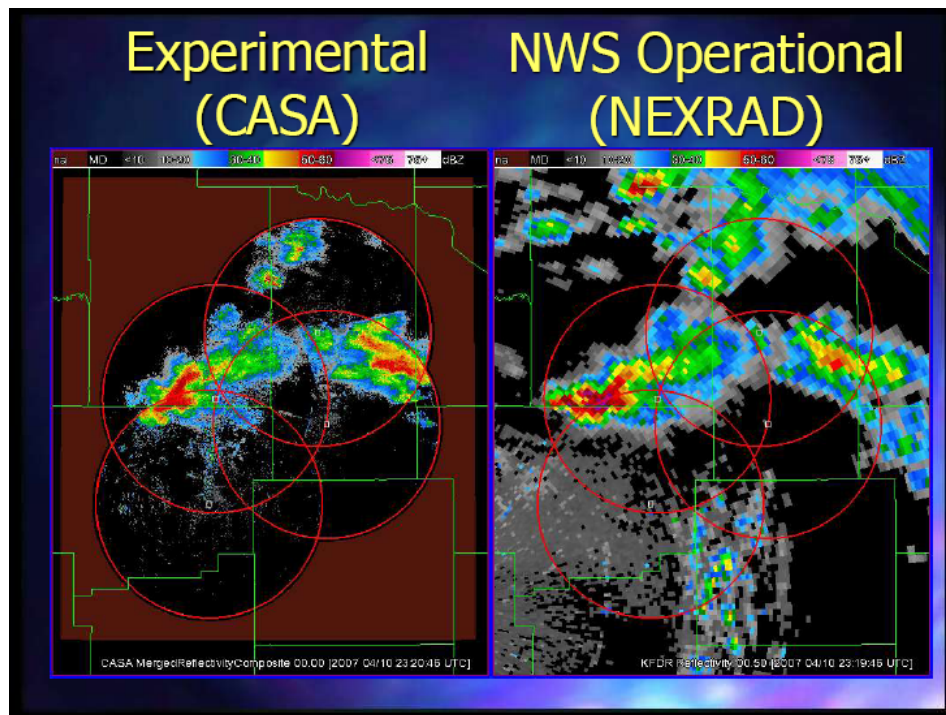


FIGURE 2-6 Comparing CASA with NEXRAD.
SOURCE: Kelvin Droegemeier

The potential, of course, lies with the fact that we are trying to constantly transform and improve our teaching, education, and workforce development.

Implications of the Three Scientific Knowledge Discovery Case Studies – The User Perspective

International Online Astronomy Research, c. 2011

Alyssa Goodman
-Harvard University-

Dr. Conti provided a great overview of the challenges that face our field of astronomy. Instead of focusing on only one example, I will draw a map of the way that things work in our field, and then give a few brief examples.

Figure 2-7 shows water droplets on a spiderweb.



FIGURE 2-7 Image of water droplets on a spider web.
SOURCE: Alyssa Goodman

This is the way that I like to think of “the cloud” with the International Virtual Observatory Alliance. Basically each one of those little water droplets is some data repository, service, or person that provides something in this ecosystem of online astronomy research. The way that those kinds of services are connected on the World Wide Web is how I think that online astronomy research should be done—in a slightly less orchestrated fashion that people might have thought of a decade or more ago. In my work, we call this “Seamless Astronomy.”

There is a researcher in the middle of the image in Figure 2-8. Imagine that this researcher is drawing on two kinds of primary sources of information: data on the left, and literature on the right.

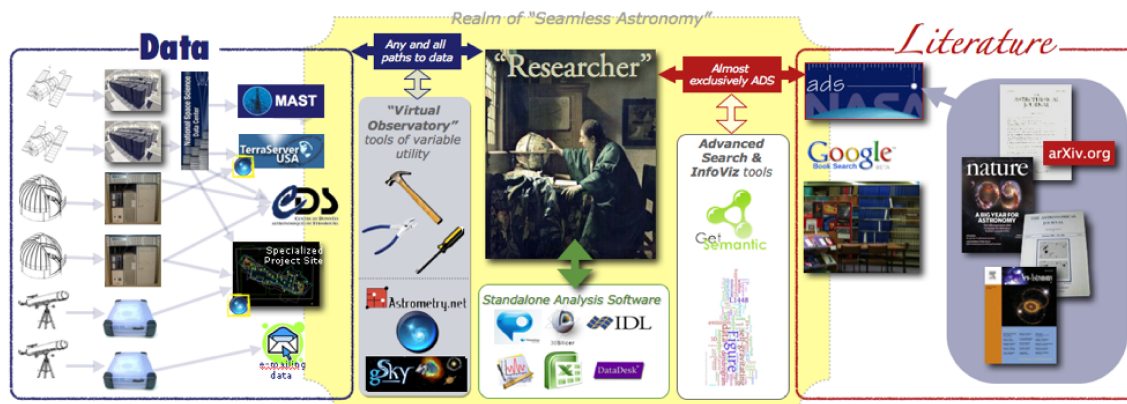


FIGURE 2-8 Realm of Seamless Astronomy.
SOURCE: Alyssa Goodman

In astronomy, our literature is organized very well by a service called the Astrophysics Data System (ADS). Much of what I am going to discuss is how the ADS team is working on connecting the data to the literature through the ADS Labs at the Center for Astrophysics in Cambridge.

I am going to take this Web-like approach and try to explain a few of the interconnections between the two sides of the picture: literature on the right and data on the left. I have reduced this model to just a few services on which I will focus, but as the disclaimer at the bottom says, this slide shows key excerpts from within the astronomy community and excludes more general software that is used, such as papers, graphing and statistic packages, data-handling software, search engines, and so forth. As I mentioned earlier, the literature is very organized, so I have put it all in one accessible repository, and it is accessed through the ADS system. As for the data, there is a set of archives that I will cover shortly.

Figure 2-9 is a view of the international online astronomy ecosystem that I will use to frame the rest of this talk.

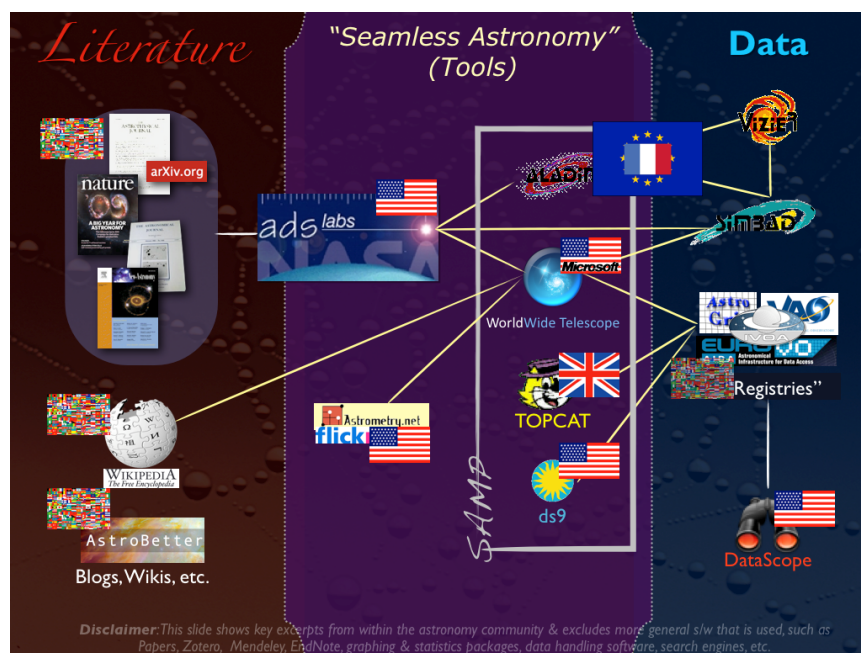


FIGURE 2-9 International online astronomy ecosystem.
SOURCE: Alyssa Goodman

In the middle, I have depicted some tools in a regime called Seamless Astronomy, which offers links between data and literature that are still partly missing, but are now forming and expected to lead to the increase in scientific productivity. The “ADS Labs” is a project in which the ADS team is taking what is already a successful system for accessing the research literature in astronomy and trying to draw that system closer to the data by using the articles themselves as a retrieval mechanism for data. Since ADS already stands for “Astrophysics Data System,” why not develop a way to retrieve articles using the literature as a filter for the data?

There are also linkages between many services that do different work. What you should notice, however, is that there are many yellow lines in the figure, which connect all of these services in a variety of ways. It is important to note that this is not a linear process. There are many different ways that the services can access each other without the user even knowing about it.

Four of these applications—the four highlighted in the “SAMP” box in the figure (SAMP is a message-passing protocol)—are actually connected to each other in a way that allows them to interoperate as a set of services and as a research environment.

I would like to point out that the flags in the figure represent the various countries involved. There is also an international flag that indicates this activity is not associated with any one particular country. Some of the services I am going to talk about are from the United States, but others are not, such as TOPCAT from the U.K. *e-Science* Program.

Figure 2-10 shows a screenshot of what happens when we run many of these services together as well as the SAMP hub that allows the services to talk to each other.

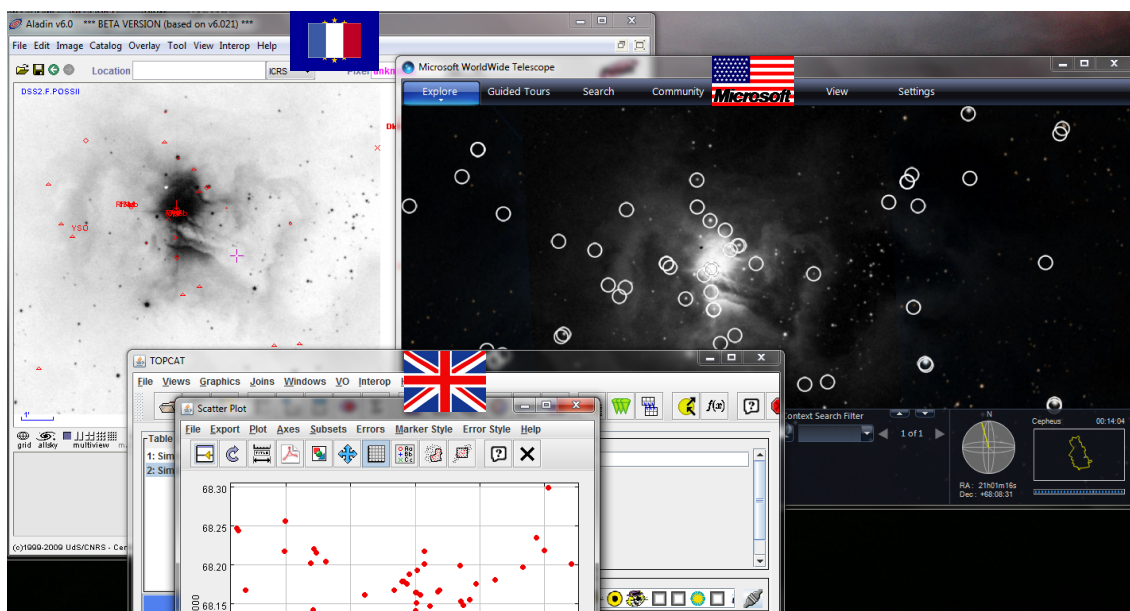


FIGURE 2-10 SAMP hub.
SOURCE: Alyssa Goodman

Suppose that I was looking at some particular region in the sky in one application. I can open another application that looks at the exact same region. Worldwide Telescope can superimpose catalogs in ways that are different from how the first application superimposes catalogs. And TOPCAT, which allows us to statistically manipulate those catalogs, will show the results through live links. It is important to note that these are from three different sets of countries, and the applications are working together through this open-source message-passing protocol. At the moment, my view is that probably only about 5 percent of astronomers know how to do this correctly.

Now let me show you the ADS Labs itself. I should point out that ADS is funded through the U.S. Virtual Observatory funding, provided by the National Aeronautics and Space Administration (NASA). It also has a mirror site in France run by *Centre de Données Astronomiques de Strasbourg* (CDS). ADS Labs was released in January 2011, at the American Astronomical Society meeting in Seattle. The idea was to take ADS and, instead of just doing a keyword search, do more intelligent searching. There are some interesting semantic aspects of this kind of search. Let me give an example of how the ADS can be used.

I was working on an observing schedule recently to observe an obscure quantum mechanical effect called the Zeeman effect in the molecule CH from a particular telescope. I wanted to make sure that no one has ever done this before, so I checked for related papers. If I use the most recent work function of ADS and type “Zeeman effect,” I get something that looks like the image in Figure 2-11. There is a filter list down the side, and I can choose particular authors, key words, missions, objects, and so on. The list of papers that I get will change depending on what is chosen.

The screenshot shows the ADS Labs Beta Search interface. At the top, there are logos for 'ads labs' and 'NASA', and navigation links: Home, Labs Home, ADS Classic, Help, and Sign on. The search results are for 'zeeman effect ch - Most recent'. The search bar shows 'Object: Other object [X] OR Nebula [X]'. On the left, there are filter sections: Authors (Uitenbroek, H (4), Amano, T (2), Angel, J (2), Asensio Ramos, A (2), Balasubramaniam, K (2)), Keywords, Archives, Missions, SIMBAD Objects (Other object (3), Star (3), Nebula (1)), Vizier Tables, Refereed status, and Dates. A dropdown menu for 'NGC 7027 (1)' is highlighted with a green box, showing options: ADS Publications, SIMBAD Info, World Wide Telescope, and Aladin applet. A yellow arrow points to 'World Wide Telescope'.

FIGURE 2-11 ADS Labs Beta Search.
SOURCE: NASA

The green box in the figure highlights some linkages between the data and the literature. If I click on one of the particular objects that seems to be mentioned very frequently in these papers, I get four different choices: (1) I can look up all the publications about that object; (2) I can look up data on that object using the SIMBAD data service from CDS; (3) I can see that object in context using Worldwide Telescope; or (4) I can do a similar thing, although it is a narrower context, in another service called Aladin.

If I clicked on Worldwide Telescope, I would get a view like the one in Figure 2-12. The object turned out to be a planetary nebula that I did not know. Since this is not what I was interested in, I concluded that no one has observed what I was looking at and that this object was irrelevant. But perhaps I became curious about this object, and so I could click on the button at the bottom that says “Research.” By clicking there, I can get another set of choices, linking to yet other services.

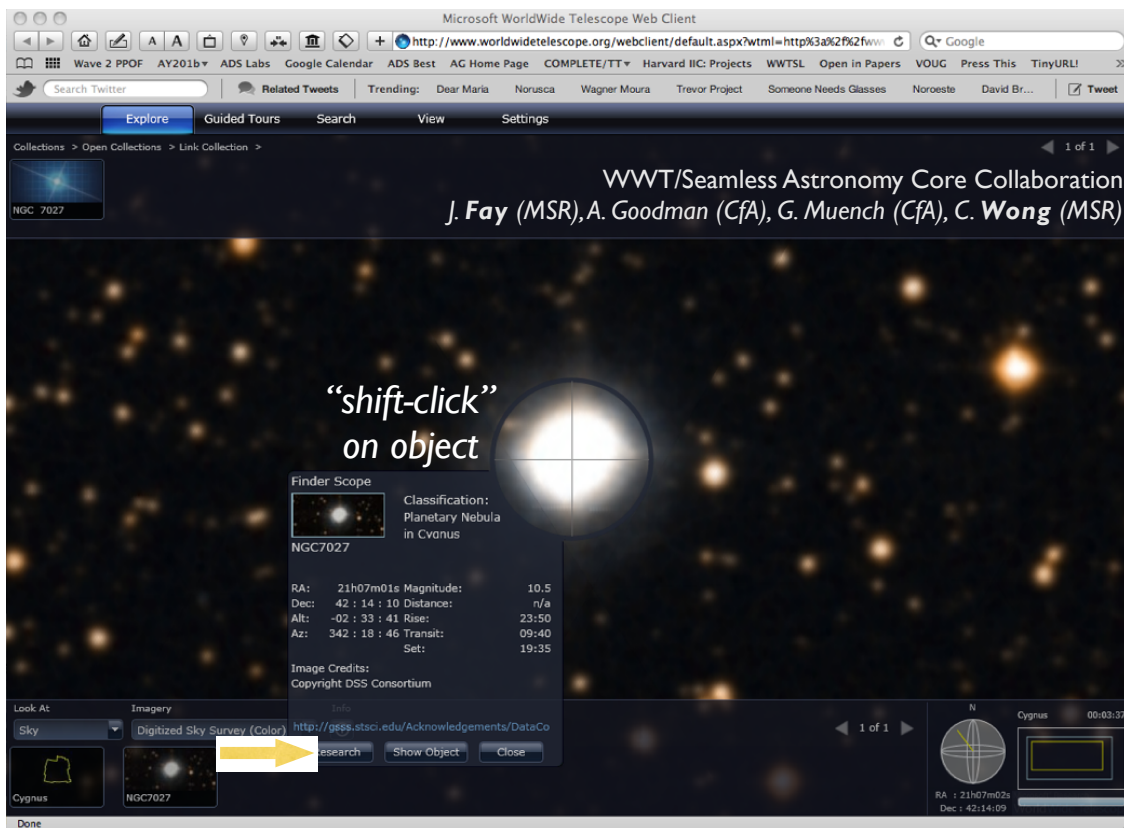


FIGURE 2-12 Worldwide Telescope.
SOURCE: Alyssa Goodman

Basically what we are being offered when clicking on the different options under “Research” is either more data or more literature. The links are such things as “look up all the publications about that particular object,” “look up that object on Wikipedia,” or “look up that object in astronomy databases.” This system offers a lot of flexibility. However, even though the system is supposed to list the most relevant papers, it is employing a very low level of semantics to do the search. For example, because I know this field, I can tell if the list of papers is incomplete. It is partially right, but it is not giving the full picture to a scientist who has never worked in this field, and that can be very misleading.

At this point, I should recognize the people who are involved in this project. Worldwide Telescope was envisioned by Curtis Wong and implemented by Jonathan Fay, who is a software engineer at Microsoft Research. At the Harvard-Smithsonian Center for Astrophysics, Gus Muench and I have been working closely with them since the beginning, but Curtis and Jonathan get all the credit.

Let me give another example. Assume that I find an astronomical image somewhere on the Web or that I go outside on a clear night with a camera and I take a picture of the sky. With this service called *astrometry.net*, I can submit a picture, and the service will take the image and miraculously calculate exactly where it goes in the sky and, if the picture is good enough, even figure out when it was taken.

There was a service that some researchers were hosting to do this work, but they did not have enough money or servers. So they made a group called “astrometry” on Flickr. As illustrated in figure 2-13, when you upload a picture to Flickr, not only does it host the image but it also offers some services that, for example, tell you what is in that picture. As soon as a picture finishes loading, if you add it to the astrometry group, you will soon see all sorts of little boxes pop up on the picture, and it will tell you famous objects in the picture that it recognizes. The reason Flickr can do this is because the *astrometry.net* program went in and blindly solved for the coordinates of the image. On Flickr, after an image is solved by *astrometry.net*, you will see a button at the bottom of the page that says “View in Worldwide Telescope.” If I click this button, then I will see that object overlaid on the sky where it goes, and I can compare it to 40 different layers of multiwavelength data. I can also use all the catalog-searching tools that I showed before and interact with all the statistical software, plotting software, et cetera, live.

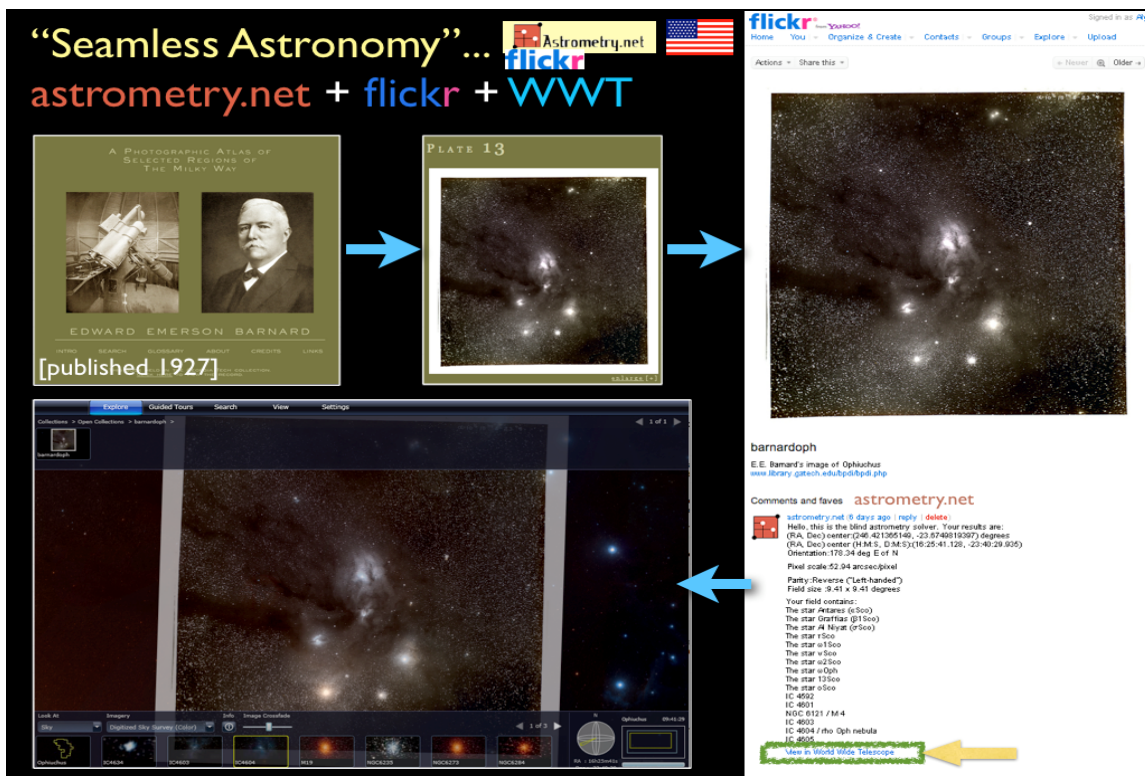


FIGURE 2-13 Seamless Astronomy.
SOURCE: Alyssa Goodman

There are two more projects in which Alberto Pepe is involved. In one of them, we are taking all of the historical images that are online, including all of the literature, and making a historical image map of the sky. With the second project, we can tell where particular papers on the sky are and why those papers were written, so you can get heat maps of articles on the sky.

I will conclude by highlighting a couple of examples where one field of study has borrowed from another. We published a paper (Goodman et al. 2009, *Nature*) 2 years ago⁹ that has embedded in it an interactive three-dimensional (3D) diagram; the idea is that you go there, click it, and you can put the data in the literature (Figure 2-14).

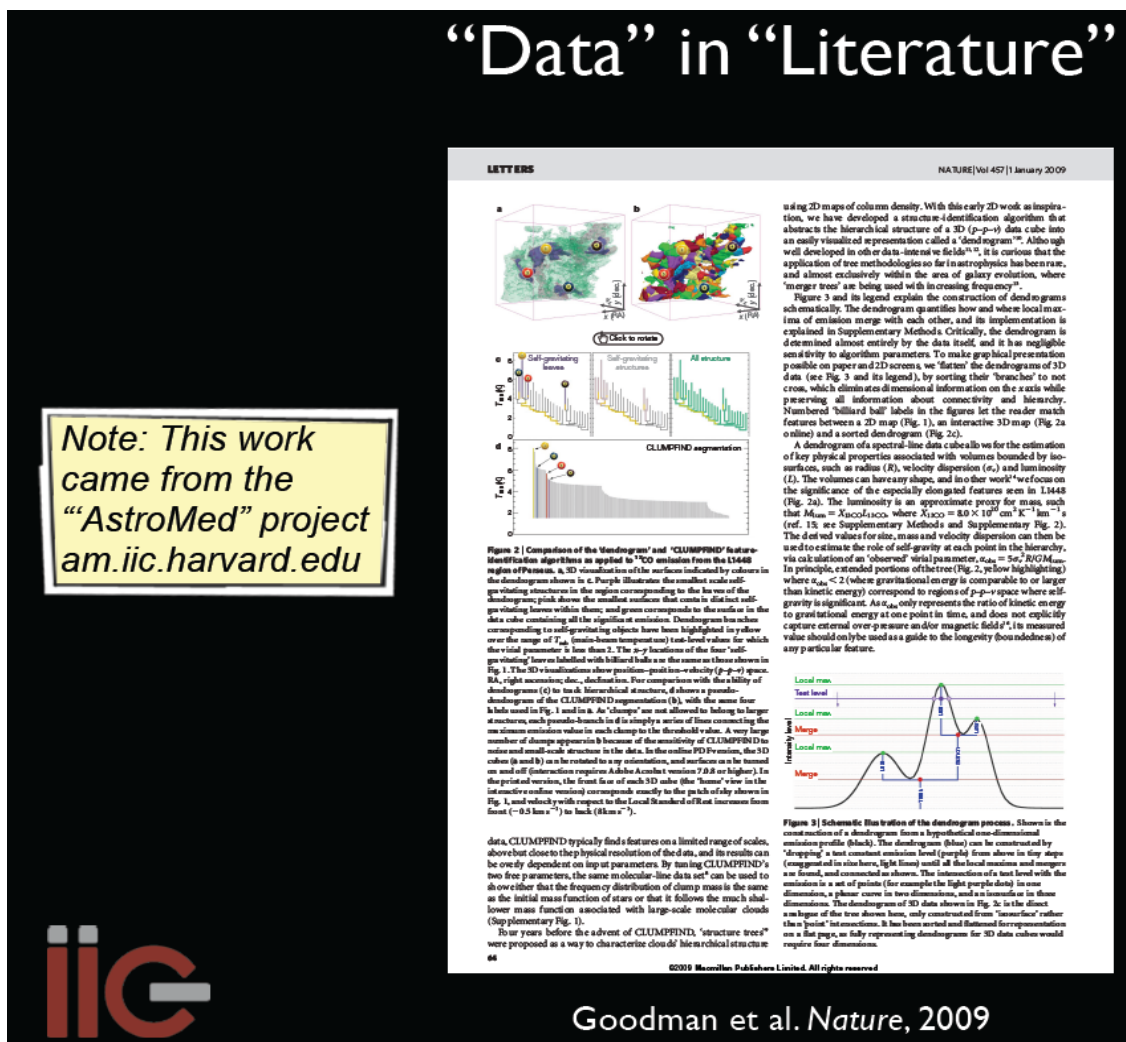


FIGURE 2-14 Data in literature.
SOURCE: Alyssa Goodman

The techniques used to create and study the 3D data were borrowed from medical imaging and used in astronomy. In the new Astronomy Dataverse (theastrodata.org), we have borrowed tools from the social science world as well. The Astronomy Dataverse tools allow individuals or groups to publish their data online in ways that can be retrieved in the future and that are permanent and linked to literature. At present, we are experimenting with the Astronomy Dataverse at the Center for Astrophysics, to let researchers who have small datasets and want to share them with everybody have a way to publish them that is linked to the literature and is persistent.

⁹ *Nature*, Volume 457, January 1, 2009.

Integrative Genomic Analysis

Joel Dudley
-Stanford University-

In this talk, I want to offer some tangible examples showing why it is important to integrate data in a certain way. It is not like a seed that you bury in the ground, and then in a few years it is going to bear fruit. Rather, it is a process that produces immediate discoveries. I will present what we have done in just a few years by dealing with data in this way, essentially on a shoestring budget, and hopefully this will motivate larger investments in this area.

We published a paper¹⁰ a few years ago describing how we took a gold standard set of Type 2 diabetes and obesity genes and wanted to see how well we could recapture them. This is a classification problem. We looked at different modalities individually, for example, genetic experiments that are measuring Type 2 diabetes, or proteomics, another modality for measuring diabetes

In Figure 2-15, the diagonal is essentially a random chance classifier, and anything worthwhile needs to be above it.

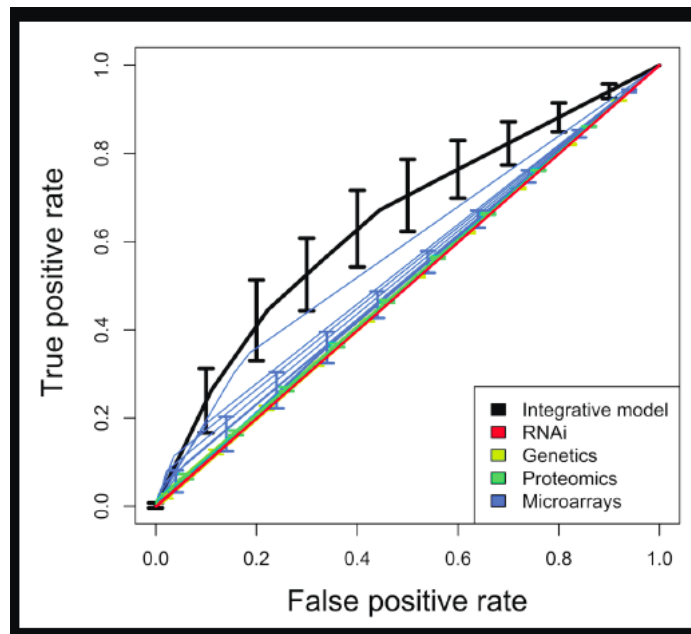


FIGURE 2-15 Rates of Detection for Type 2 Diabetes.
SOURCE: Joel Dudley

We put all the modalities together, which is represented by the black line, and tried to recover a known set of Type 2 diabetes genes. This is a very simple classifier, but what it

¹⁰ S. English and A. J. Butte, "Evaluation and Integration of 49 Genome-wide Experiments and the Prediction of Previously Unknown Obesity-related Genes," *Bioinformatics* (2007) vol. 23 (21), p. 291.

shows is that by putting all the data together, we are able to beat all of the individual modalities. Compared with all the single investigators working very hard on their own single modalities, we were able to surpass each one of them by putting all their data together from the public domain.

One of our main sources of data is the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information. There is also a European counterpart called Array Express. The GEO is a public repository. If a scientist publishes a paper that incorporates gene expression data, the journals require that this data be submitted into GEO. GEO is growing exponentially and has all kinds of useful data. People deposit their data in there reluctantly, but they are doing it. However, as Dr. Friend pointed out, there is a difference between accessibility and usability. GEO is freely available and it is very accessible (e.g., high school students can access and download all the data they want), but it is not very usable.

There are hundreds of thousands of samples. One sample is essentially data derived from microarrays. This is roughly a 10-year-old technology. The very small chips that measure how much a gene is turned on or off in a biological state by measuring the abundance of its expressed transcripts are roughly a 10-year-old technology. There will be certain levels of mRNA expressed off the DNA, and the chip has probes for each of the genes in the genome. By measuring the responses with a scanning laser, you get an analog readout of how much each gene is turned on or off.

To put these data into GEO in a structured format, they used a standard called MIAME (Minimum Information About a Microarray Experiment), but there were no semantic requirements put on the data. As a result, we have all the data in GEO, but we do not know what is being measured. This means that we could not use GEO to find, for example, all the experiments that measure a disease. It is just not set up that way, so we came up with a solution. Since all these experiments are linked to a publication, we could find the publication in PubMed Central and download the Medical Subject Headings, and then map them onto the Unified Medical Language System, which is a huge medical thesaurus created by the National Library of Medicine. In other words, since a paper linked to an experiment must be related to the experiment in some way, we could use the paper annotations to figure out what the experiment is about. We can indeed do that, and it has worked fairly well.

There are other sorts of annotations related to which state is being measured in an experiment. For example, if we are going to do an experiment in which we measure healthy and sick people, some of the chips will be measuring the healthy people, and some chips will be measuring the sick people. There are some annotations that tell us which chip is which. This is very important information to have if there are remote users of the data. Unfortunately it is free text, so that it is not done to a uniform standard. To refer to a person with diabetes, some cases might say “diabetic,” some might say “Type 2 diabetes,” others might say “insulin resistance,” and so on. I saw an experiment where the annotations read “dead” and “not dead.” To figure out which chips belonged to which subset, we did some text mining.

The next problem we encountered was how to determine which genes were being measured on these chips. There are thousands of different platforms. Some of them are custom made by universities. Also, there are different commercial vendors with different technologies and types of probes. We had to determine which genes were on these chips. This problem had not arisen before, because scientists were not trying to map across these different types of

technologies; they got their chip for their particular study and stayed within their own technology. We also found out that these annotations were changing by 10 percent or so every year. We thus had to build an entire system that would constantly update and remap all these probes to tell us which genes were being measured on the different chips.

We then had to determine the quality of the data. To do that we looked at how well data from different labs matched up. For example, we would look at data from one lab using a particular technology that measured Type 2 diabetes in muscle, and compared that with data from another lab using another technology that measured Type 2 diabetes in muscle in completely different patients. It turned out that they match up fairly well—quite a bit better than random. So even though there are all these possible confounders, the quality of the public data is quite good. This was very encouraging.

Once you have all this data annotated in this way, it can be visualized using a matrix. In such a matrix, you can have 300 or so diseases in the rows and then 20,000 genes in columns, which produces a big grid of all diseases and all genes in a very clean, standardized way, showing how the genes change with the different diseases, as illustrated in figure 2-16.

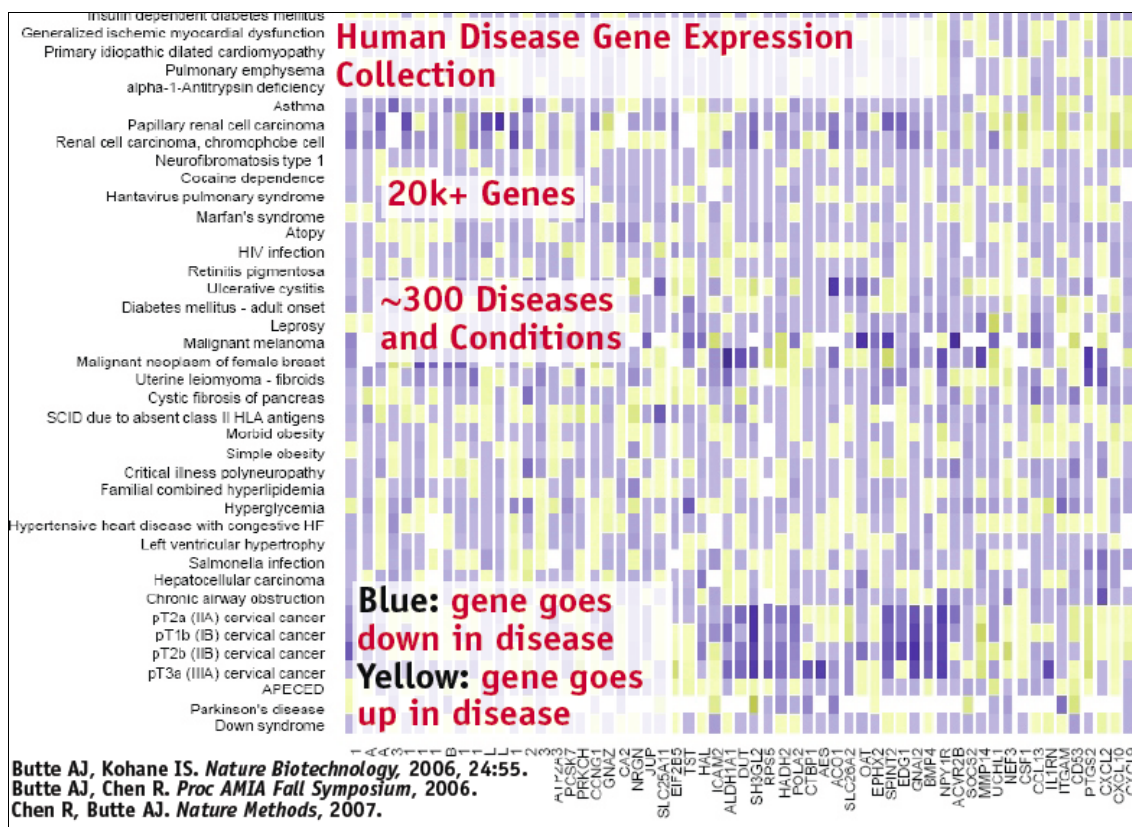


FIGURE 2-16 Human disease gene expression collection.
SOURCE: Joel Dudley

At this point we can start thinking of interesting questions to ask. You could simply cluster that grid and see where the diseases fall with regard to the various molecular bases. This will provide the molecular classification of disease, and then you can see some patterns. Brain

cancers are clustering together, which makes sense, but then you see that pulmonary fibrosis is clustering with cardiomyopathy, which seems strange. It is a lung disease clustering with a heart disease. Typically these are very different diseases, but maybe at their molecular level they have some similarities.

Then you start to wonder what the implications might be for drug treatment. We have all kinds of drugs for cardiomyopathy and hypertrophy and similar cases, but we do not have any drugs for pulmonary fibrosis. Therefore if the molecular basis is so similar, we might be able to borrow drugs from one disease to deal with another.

We do not quite use this representation to figure that out. Instead we have discovered a new way to mine the public data for drug repositioning, as it is called when we borrow a drug from another disease. We have a compendium of clean disease data that is semantically annotated, and we did the same thing for drugs. See Figure 2-17. This gives us both the disease and the drug data. Now we can look at treated versus untreated diseases and get a signature of the disease state versus the healthy state, and the signature in this case is how the genes are changing between the two conditions.

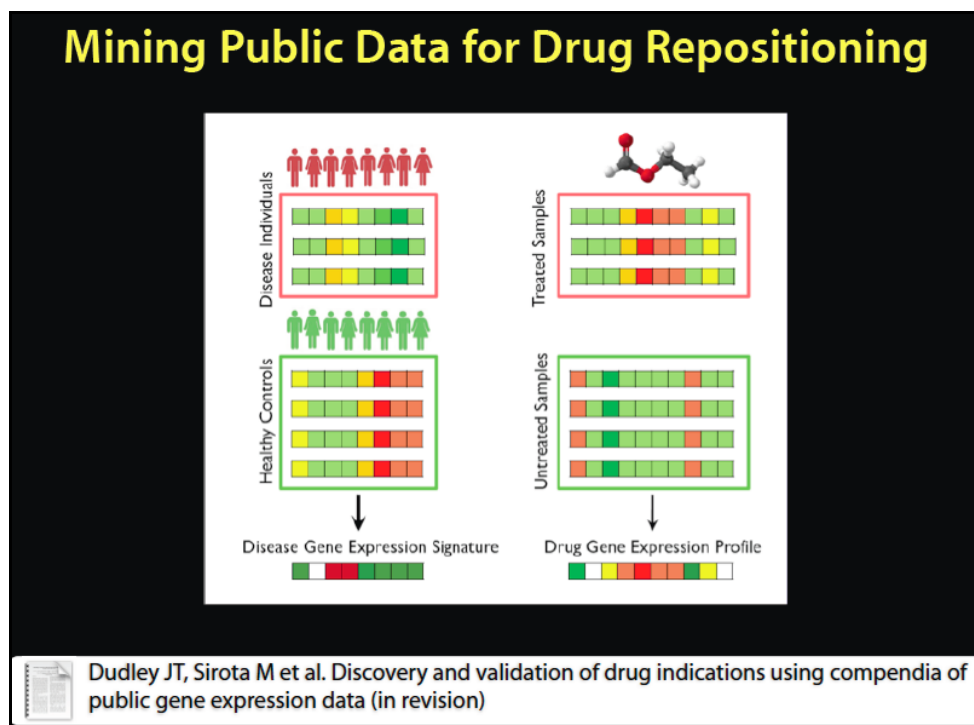


FIGURE 2-17 Mining public data for drug repositioning.
SOURCE: Joel Dudley

Similarly a scientist can look at how the genes are changing in response to different drugs and do pattern matching to match them up statistically. This is exactly what we did, and it produced a representation like the one in Figure 2-18.

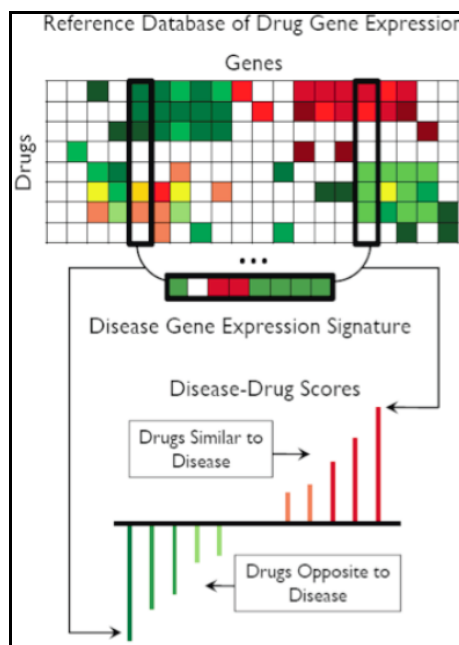


FIGURE 2-18 Reference database of drug gene expression.
SOURCE: Joel Dudley

If a scientist has thousands of drugs and hundreds of diseases, and gives a therapeutic score to each pair, Figure 2-19 shows how the heat map would look.

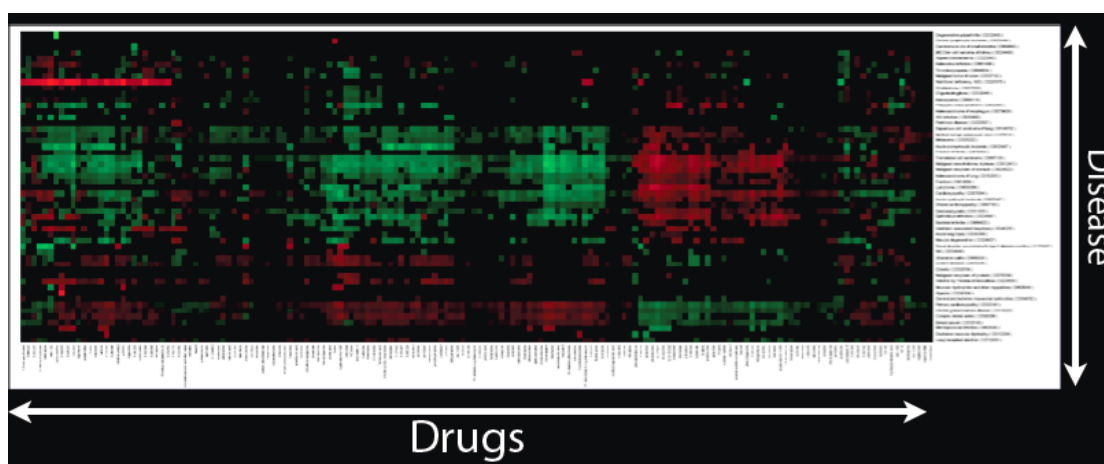


FIGURE 2-19 Heat map of drugs and diseases.
SOURCE: Joel Dudley

After we have done that, we start to see interesting clusters. We might ask why all these drugs and diseases appear in a particular cluster, and check to see if the drugs in the cluster are known to treat the diseases in this same cluster. We might also ask if the drugs that have not been identified as treating those diseases might actually be effective against them.

The first computational prediction that we tested was whether cimetidine, an anti-ulcer drug that anyone can buy inexpensively at any pharmacy, would treat non-small-cell lung cancer. It seemed a very strange prediction, but we thought we would try it (Figure 2-20).

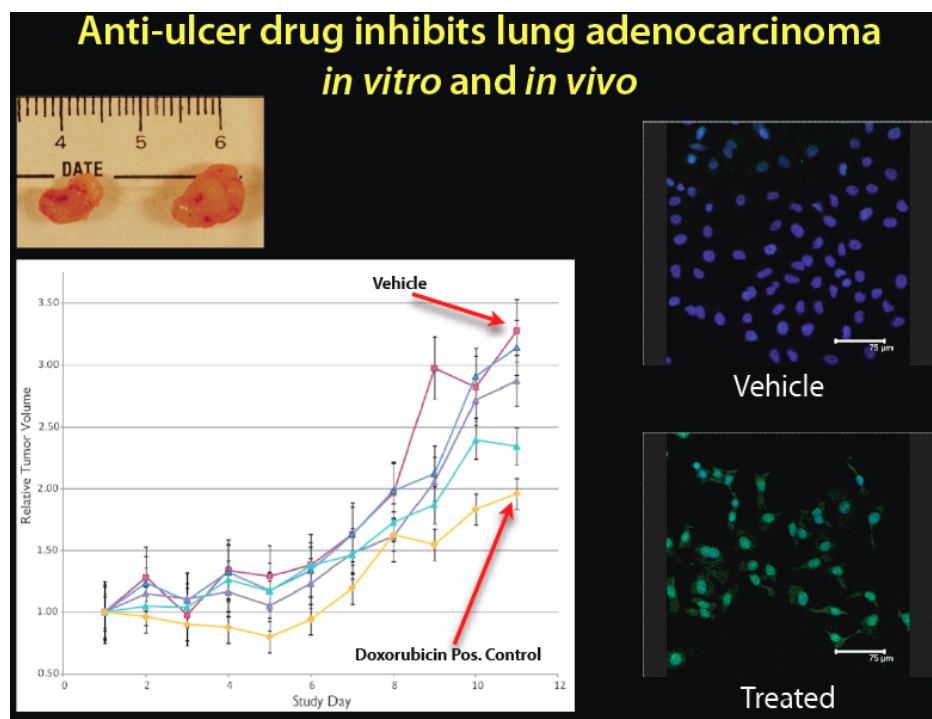


FIGURE 2-20 Anti-ulcer drug inhibits lung adenocarcinoma in vitro and in vivo.
SOURCE: Joel Dudley

When we put human tumors in mice and tested the drug, we got a dose-dependent response. It shrank the tumors, and it was even getting close to a chemotherapeutic drug, doxorubicin. We can see on the top left image the cimetidine-treated tumors on the left and the control on the right. We can also see on the right side the fluorescence microscopy of the actual cancer cells. The green lighting in the bottom right box indicates that when the cells were treated with cimetidine, there was a lot of apoptosis, or cell death, in the cancer cells. We predicted that this over-the-counter anti-ulcer drug would be antineoplastic, or anticancer, and it seemed to be efficacious in this animal and human cell line study.

Another prediction we made was that an antiepileptic drug would treat inflammatory bowel disease, which was another strange connection. We got the rats that have an induced phenotype of inflammatory bowel disease, the TNBS model. A very caustic compound is put in the rat's colon to induce the inflammatory response. When we induced the disease and treated the rats with the drug, it ameliorated the disease. This is a very harsh phenotype, but the drug appeared to reduce the inflammation. This is an anti-epileptic neurological drug working on the colon. At the bottom of the image in Figure 2-21, you can see the pink colons of the rats that were treated, while in the untreated rats on the top you can see all the necrosis.

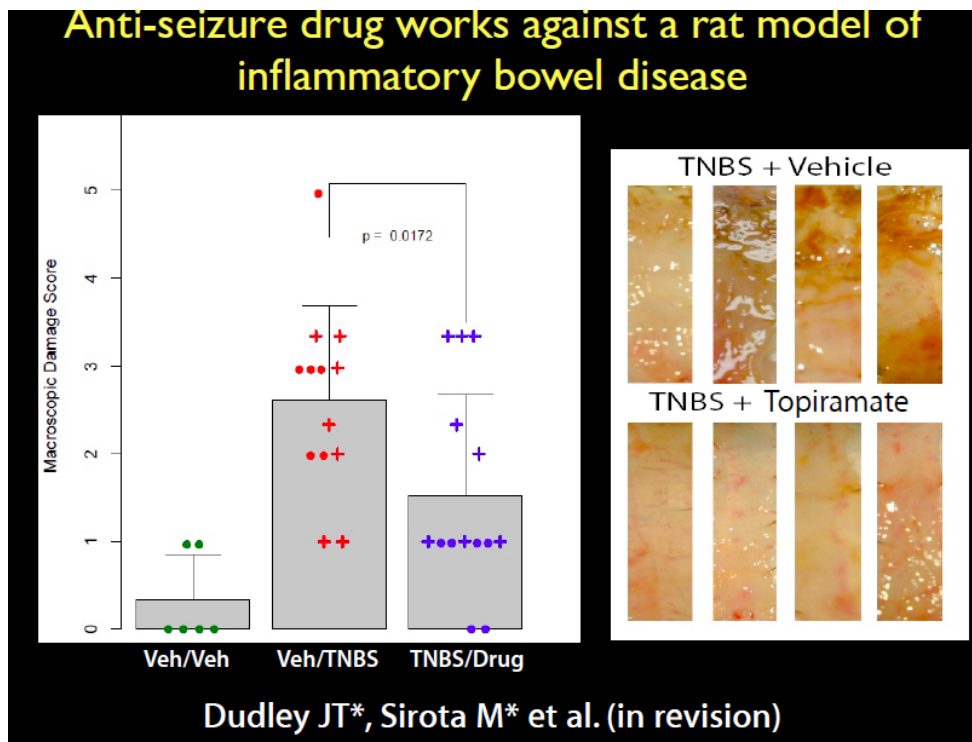


FIGURE 2-21 Antiseizure drug works against a rat model of inflammatory bowel disease.
SOURCE: Joel Dudley

Figure 2-22 provides another recent example. We predicted that a cardiotoxic drug, which was researched quite heavily in the 1980s, but never approved for any disease, would also work for inflammatory disorders. We put it in rats that had an inflammatory phenotype. What can be seen in the table are the blood level measurements of a cytokine called TNF-alpha, which is involved in inflammation. On the right you can see that this heart medication substantially reduced TNF-alpha levels and plasma in rats.

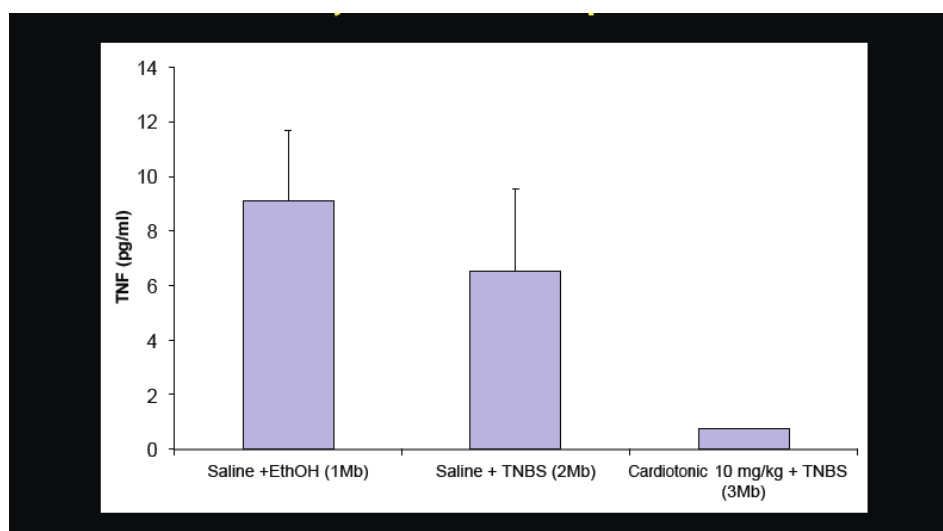


FIGURE 2-22 Cardiotoxic drug inhibits or ameliorates inflammatory cytokine TNF-alfa.
SOURCE: Joel Dudley

Also, we took several experiments related to transplant rejection from the public data, intersected them in a certain way, joined them with a number of serum measurements of proteins, and found a biomarker for transplant rejection. This is a particularly important biomarker, because when someone gets a new kidney, for example, medical personnel stick a big needle through the side of the patient's body, suck out a little tissue around the kidney, look at it under a microscope, and, if this person is lucky, they get the part that is being rejected, and they will try to manage it that way. Just by using the public data and integrating them with public datasets, we computationally predicted a new blood-based biomarker. We are testing it, and it looks like it is also able to be found in the urine. Therefore, instead of putting a big needle into the side of the patient's body, it should be possible to take a urine sample and tell from that whether the kidney will be rejected. Again, this is a repurposing of public data. We did not generate any new data from the patients at Stanford University to do this, although that is where it was tested in the clinic.

There are many more examples from our lab. We just found a new disease risk variant for Type 2 diabetes, which has been validated in a Japanese population of thousands of cases and thousands of controls. We found a new drug target for Type 2 diabetes and validated it. Indeed, these are all validated experimentally. We found biomarkers for medulloblastoma, pancreatic cancer, lung cancer, and atherosclerosis—all validated—and this is just from our lab, but it should not be just from our lab.

The reason we can do this kind of work is that we have significant computational capabilities. Everyone who works in our lab has professional or graduate-level training in molecular biology and computer science. They can draw cell signaling pathways from memory and also program a compiler. It is unfortunate, however, because much of this technology is an expensive commodity, and the people who actually have the questions cannot do anything with their data.

Another example involves the single-nucleotide polymorphism (SNP) chips, which measure genetic variation on about a million different regions of the genome. We can combine them with the gene expression microarrays, and for relatively little expense, a commissioned researcher can then assemble a group of patients and measure both their DNA variation and their gene expression variation. From that, the researcher can figure out such things as which genetic variants are really behind a disease. The SNP chip contains a million probes, however, and the gene expression microarray measures 50,000 probes, so the researcher is looking at 50 billion array comparisons to determine which are related, which would take several years to do on a desktop computer.

We therefore did a case study that showed it would be possible to do the research in the cloud and pull out 100 nodes with a minimal cost (Figure 2-23). It required a lot of programming that the clinical researcher typically would not be able to do, and there are no tools available that allow the researchers who generate the data and have good questions to answer the questions with their own data.

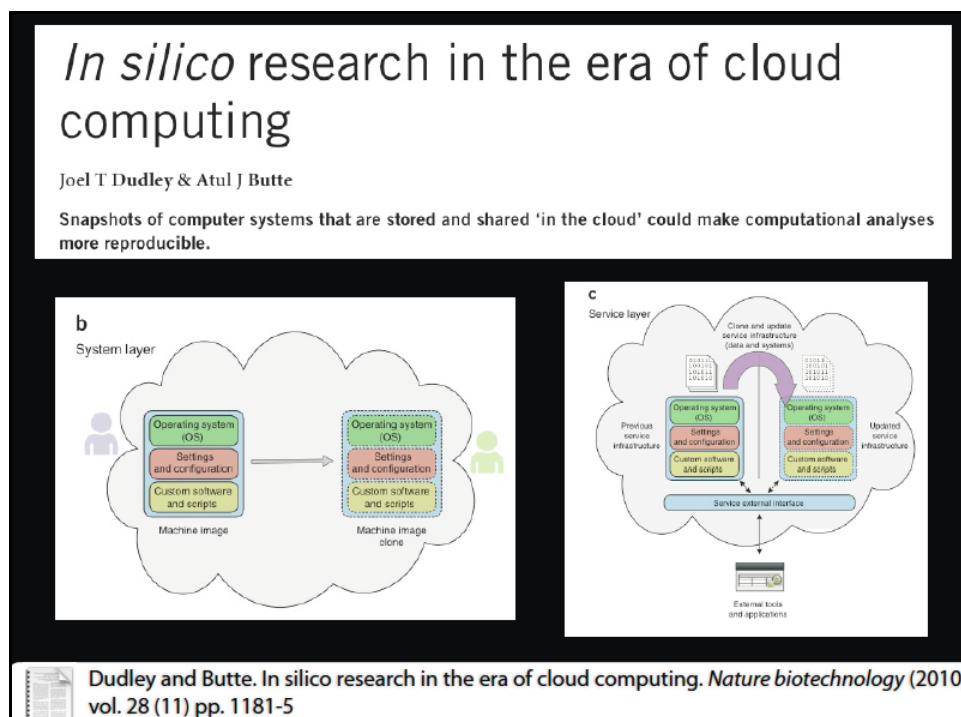


FIGURE 2-23 *In silico* research in the era of cloud computing.
SOURCE: Joel Dudley

I think, however, there is a pragmatic first step that can be taken with cloud computing. Many people will disagree with me on this, but there are many people working on perfect workflow systems, and although people in biology do not seem to like to use them, I think those tools will become the standard. One problem is that people like to keep everything in-house—and probably always will—because they are at the forefront of these research areas, and the standard tools do not necessarily allow them to do what they need to do. One valuable aspect about the cloud is that the entire infrastructure is virtualized, so, for instance, I was able to set up a pipeline for doing drug prediction. It is simply a number of Linux servers and our code. What I can do is to freeze that in place, and at the very least I can hand it over to someone in the cloud who can clone it. I am a big believer in simple first steps and rapid iteration versus locking yourself in a room for years to come up with a perfect system that solves all the problems. I think there is some promise there, and we are starting to see it happen. There are some good tools that are being created to facilitate reproducible computing in the cloud.

Let me now focus on some of the lessons learned from the work in our lab. So far, sticks have worked better than carrots. Biologists hate sharing their data. The data are voluminous and although the researchers are supposed to publish them in GEO, there are many microarray experiments that do not get published. When we find that a paper that is funded by the National Institutes of Health and it has microarray data that are not in GEO, we call the Microarray Gene Expression Data Society, and the people there follow up with the researcher and ask that person to post the relevant data there. We constantly have to do that, so we hope that carrots will work better at some point.

In ontology, we have been able to accomplish much by doing very lightweight and simple annotations on gene identifications. We did not need to model biology perfectly in a higher ontology to do what we needed to do.

Computation is a major bottleneck. Right now there is a privileged computational elite. There are people who cannot even answer questions from their own data on their own patients, and, on the other hand, there are the computational elite who have 500 CPU clusters in their labs and the power to do this kind of work. Obviously this is not fair.

Questions first, data second. What I mean by this is that we did not really know the type of system we needed until we knew the questions we wanted to ask. There are many people in informatics who are pattern hunting and trying to find hypotheses, but it was not until we knew the questions we wanted to ask from the data that it became apparent what type of system we needed, and it was very different than what we had. Finally, to sum up, I would like to emphasize that the advancement of new biology and medicine is possible through public data.

Geoinformatics

Mohan Ramamurthy

-Unidata-

I have been with the Unidata Program for the last 8 years, and before that I was on the faculty at the University of Illinois in the Atmospheric Science Department for 16-plus years. My talk, therefore, will cover the perspectives of the different hats I have worn in the past 25 years or so.

Unidata is a facility funded by the National Science Foundation (NSF) for advancing geoscience education and research. We are part of the University Corporation for Atmospheric Research (UCAR) and a sister program to the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. Part of our mission is to provide real-time meteorological data to users throughout the world. It started with academic universities getting the data, and that data distribution has expanded to include other organizations. We do not give just data to our users; we also give them a suite of tools for analyzing, visualizing, and processing them. One of the software products for which we are best known is NetCDF, because it has become a de facto standard in the geosciences. The data we provide get integrated into research and education in universities, as well as in many high school classrooms. We have more than 30,000 registered users. There are many more who do not register, but get the software. The users come from 1,500 academic institutions and 7,000 organizations in 150 countries around the world, representing all sectors of the geosciences—not just academia, but government, nongovernmental organizations, and the private sector. They get the data and the software from us, because everything we do is available at no cost.

I want to go back almost 25 years to the work that my former colleague, Dr. Robert Wilhelmson, did at the University of Illinois and at the National Center for Supercomputing Applications (NCSA), which ushered in a new era of geoinformatics for my field. That work was the result of a confluence of three factors. First was the availability of high-performance computing. Because the NCSA had been established a few years earlier, he had access to cutting-edge, high-performance computing systems. Second, he had software for visualizing and creating fantastic graphics, which won first prize at the London Computer Graphics Film Festival in 1989 and was also submitted for an Academy Award nomination. Third, there were massive quantities of data coming from state-of-the-art cloud computer models. From those factors, geoinformatics was born, and it revolutionized our understanding of supercell thunderstorms, convective storms that spawn tornadoes, and mesoscale connective systems, such as the one Dr. Graves described earlier.

In February 2011, *Science* magazine had an entire issue devoted to data, and one of the articles focused on climate data and how they are revolutionizing our understanding of climate change—a point that will be woven into my talk throughout this presentation. We know that data are the lifeblood of our science, but we need to move from creating data to creating knowledge and discovering new frontiers.

In 2008, *Wired* magazine published a provocative article with the title, “The End of Theory,”¹¹ which argued that the data deluge makes the traditional scientific method obsolete. I am not sure I agree with that assessment, but it was provocative, because it was essentially modeled after what Google has done in the search business, where all search results and page ranks are driven by data rather than an a priori model of how things should be organized and ranked. This was the original Yahoo model, but that became obsolete as a result of Google’s work.

I want to focus now on the multidisciplinary aspects of the geosciences. Whether we are talking about ozone depletion over the South Pole, mud slides in China, forest fires or deforestation of the Amazon, climate change in general, or the impact of El Niño or La Niña—to make advances in any of these areas—scientists have to utilize data from many geoscience disciplines, related social sciences information, and other ecological information. Francis Bretherton, one of the former directors of NCAR, created the famous earth systems science schema, which scientists now refer to as the “Bretherton Diagram.”

Dr. Bretherton got scientists to realize that they should not be looking at just atmospheric systems, hydrologic systems, or ocean systems. All of these different systems are interconnected, whether it is the cryosphere or the hydrosphere, or even ecology and chemistry, and looking at questions from a systems perspective is the key. However, to do systems science and research and get meaningful results, we need to integrate information across all these different systems as well as to synthesize them across the domains. One of the challenges for geoinformatics is providing the right type of data in the right format to the right application in the right place. One of the key aspects of that process is interoperability of data across all the different disciplines that are involved.

The founding president of UCAR and the first director of NCAR, Dr. Walter Orr Roberts, had a famous mantra: science in service of society. We should not stop at just doing science, but we should keep in mind that science has a meaningful impact on society, which means that we should be thinking about geoinformatics and data services that are end-to-end, that are chained through workflows. For example, scientists should not stop at the prediction part, which was the focus of the work that Dr. Graves presented, but they should go beyond that and couple the predictions with the flooding models, ecosystem models, and so on. If scientists are looking at the societal impact of flooding from tropical cyclones and hurricanes, they should integrate data from atmospheric sciences, oceanography, hydrology, geology and geography, and social sciences, and then use the results with decision support systems. In short, we need geoinformatics studies that encompass these end-to-end aspects.

The issues associated with the deluge of data came up earlier. I like the “sea of data” metaphor for reasons explained in a talk I heard by Tim Killeen, the current Assistant Director for Geosciences at the NSF. When we look at it as a sea of data as opposed to a deluge, we have two options: to drown, or to set sail to new land and make new discoveries. That is very appropriate, because data bring us to new discoveries in the world.

Speaking of a sea of data, here is some information about the growth of data that is expected in the future. Today we have less than 15 petabytes total of atmospheric science data, but that is predicted to grow to 350 petabytes by 2030, with the data coming largely from

¹¹ Available at http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

models, particularly those with high resolution, in ensemble mode, and integrated in time scales from days to 100-plus years. There will also be data coming in from the new generation of remote-sensing systems, including radars and satellites. Dr. Graves mentioned the traditional radar systems, but in Norman, Oklahoma, they are developing a new kind of radar that was previously deployed by the military 20–30 years ago. It is called phased array radar, and it can scan an entire thunderstorm in a matter of seconds, as opposed to the 8 to 10 minutes that it takes today to scan a thunderstorm with traditional radar. Furthermore, the new generation of geostationary weather satellites will have 1,600 channels in the vertical for sounding capability. Current-generation satellite gives 20 channels; so we are talking about an 80-fold increase in the number of channels that give us detailed information about the atmosphere.

The polar orbiting meteorological satellites will also provide large volumes of data once the Joint Polar Satellite System Program is finished. It has been delayed many years now because of various difficulties, but once it is deployed, 3 terabytes of data are expected to be produced every day.

With all these data being produced, how do we extract knowledge and make new discoveries? The field of knowledge extraction and data mining addresses the question—going from data to information to knowledge to wisdom—but the field is still very much in its infancy. Some great work is being done at the University of Alabama in Huntsville under Dr. Graves’s leadership, but our field is still mired in these simple-looking plots that produce spaghetti diagrams of ensemble model output or postage stamp–type maps. These may be of some value when there are only 10 ensemble members, but when there are 1,000 ensemble members, there is no meaningful way to understand which data are together, which are clustered, or which have similar patterns and relationships, so an automated capability to extract that knowledge and information is very much needed.

Every 5 years or so, the Intergovernmental Panel on Climate Change (IPCC) assessment reports are written. The group’s Fourth Assessment Report resulted in the Nobel Peace Prize being awarded to the IPCC, along with Al Gore, for work in bringing awareness about climate change issues. One of the simple actions that happened in the geoinformatics world was that the IPCC Fourth Assessment climate modelers were mandated to submit all of their model output using the NetCDF climate forecast convention to the Program for Climate Model Diagnostics and Intercomparison (PCMDI)–Earth System Grid (ESG) Data Portal. The result was approximately 600 publications within 6 months after the IPCC Fourth Assessment Report activity was completed.

All the model runs were submitted, and they were made available by the PCMDI-ESG Data Portal, so everyone could compare their models immediately because they all used exactly the same structure format standard and convention for their data. That mandate had a huge impact on producing new discoveries in a very short period of time.

The next assessment report is now being prepared and the model runs are getting started. The IPCC is looking at the total assimilation of 90,000 years of climate change model results, consisting of many petabytes of output from 60 experiments and 20 modeling centers around the world. They have been mandated to produce the data and output and to have all of the necessary metadata available for the model runs, which was not required for the last assessment. For the next assessment, all of the models need to document, in a consistent way, the metadata associated with how the model was configured, how it was run, and so forth, and

to make the data available through the Earth System Grid Portal, using the NetCDF-CF format. This again raises the question of how we are going to process all of the data

The other type of ensemble work that I did in my previous work was to make ensemble predictions for hurricanes and land-falling tropical cyclones. In this work, a scientist can create ensemble numbers by varying the initial conditions, boundary conditions, physics of the model, sea surface temperatures, and other criteria a scientist has for a prediction system from the external forcing standpoint. This particular system was generating 150 to 200 runs, considering the different permutations and combinations, and we needed an end-to-end efficient workflow system, as well as an informatics system for providing seamless access to all of the data and information that was created. Therefore, we need to be thinking about geoinformatics systems that are capable of facilitating these things and delivering information.

The hurricane research community has proposed a framework for improving hurricane predictions, and, as part of its metrics, the group is attempting to reduce both average track error and intensity error by 50 percent for a 1-to-5-day forecast. Given that the intensity prediction for tropical cyclones and hurricanes has not improved by even 10 percent in the last 20 years, this would be a gigantic leap, and they are expecting to do very high-resolution predictions as well as ensemble-mode predictions to make that possible. All of the data from this endeavor will be openly available. That was mandated by the authors of the 2008 “Proposed Framework for Addressing the National Hurricane Research and Forecast Improvement Initiatives,” because the group figured this is the only way it is going to realize these results and the metrics they are aiming for.

I also want to talk about data visualization as a way of enabling new insights and discoveries. I will focus on one of the software programs that Unidata has developed, the Integrated Data Viewer (IDV). I am not a geophysicist, but I know that the IDV was used by a group called Unavco, in the GEON (GEOscience Network) project, which was another large NSF Information Technology Research project under the Linked Environments for Atmospheric Discovery project (LEAD), described in the previous talk by Dr. Graves, and which got started about 8 years ago. Unavco adopted the Integrated Data Viewer as the framework for visualizing data from the solid earth community.

Figure 2-24 shows the IDV integrating geophysical data. We are looking at seismic plate activity along the west coast.



FIGURE 2-24 Integrated data viewer (IDV).
SOURCE: Unavco

At the bottom left are stress vectors and fluids, the mantle convection assimilation, and Mount St. Helen's seismic activity underneath. The people who were doing this work told us that they could never see these aspects before this three-dimensional IDV visualization tool became available. It opened their eyes to how the Earth works underneath the surface through plate movements and mantle convection.

I also want to mention geographic information systems (GIS) integration, because I work in a field that is focused on geo-specialties. We need geo-specialty-enabled geoinformatics cyberinfrastructure so that we can integrate location-based information, because many of these processes and events are very much related to the geography. We should not be thinking of GIS as an afterthought. The historical way of thinking about GIS is that atmospheric scientists would do their work and then eventually somebody else would take over from there and put the information into some kind of a GIS framework. Unidata is working to enable NetCDF data models to interface directly with GIS capabilities through the Open Geospatial Consortium standards for Web Coverage Service, Web Features Service, and Web Map Service. As a result, when we have a NetCDF data cube, that data cube becomes available as a Web GIS service, and we can produce maps like the one to the right in Figure 2-24 from that particular capability.

I cannot conclude this talk without talking about data citation, because it has already come up in almost every presentation so far. To me this is the next frontier, I want to give an

anecdotal story to demonstrate that this is not a technical challenge, but a cultural and social challenge. This particular topic came up at the American Meteorological Society. The publication commissioner, who is in charge of all scientific publications of the American Meteorological Society, took a proposal modeled after a Royal Meteorological Society proposal, which was intended to encourage data transparency. He went into the meeting and said that the condition for publication should be that authors be required to agree to make data and software techniques available promptly to readers upon request and that the assets must be available to editors and pre-reviewers, if required, at the time of review. He was basically told that his proposal is a nonstarter, because we will never get the cooperation of the authors. This shows that we still have a lot of work to do.

I also want to talk about globally networked science. The IPCC activity is the gold standard for that, but I also want to mention one other activity that NCAR and Unidata have been involved with, which is the Observing System Research and Predictability Experiment (THORPEX), an interactive and global ensemble project. It is generating 500 gigabytes a day, and the archive is now up to almost a half a petabyte. It brings in data from 10 different modeling centers twice a day. We need to be able to create geoinformatics for these kinds of databases. These are some of the challenges. I want to come back to the point that was made earlier about interfacing social networking systems. As a geoscientist, I have seen GPS, coupled with mobile sensors, revolutionize the geosciences in a major way. I think there are some incredible opportunities, not just for citizens and science, but also for providing workforce development, geospatial awareness, and education for our students at all levels. This is very important, and how we use social networking tools like Facebook for getting the scientific community to provide commentaries, share information and data, and work in a collaborative way is something that is a real opportunity and definitely a challenge.

In closing, I would like to emphasize that we live in an exciting era in which advances in computing and communication technologies, coupled with a new generation of geoinformatics, are accelerating scientific research, creating new knowledge, and leading to new discoveries at an unprecedented rate.

Discussion

DISCUSSANT: Mr. Dudley, in your chart of supporters, I see some National Institutes of Health (NIH) institutes, but none from the National Science Foundation (NSF). However, you were working at the intersection of what NIH and NSF medicine would fund. I find that curious. When I was at NSF 10 years ago, I discussed with NIH colleagues the cultural differences, such as how NSF will fund soft money research and how NSF likes proposals to be more exploratory, while NIH likes them to be more incremental. We did not come to an agreement, though. There was a division director at NSF who liked to encourage joint work between computer science and medicine. Now that you are doing this kind of work, what organizational steps would you take to help in that area?

MR. DUDLEY: Our funding strategy has always been to involve the computer scientists in whatever we are doing. One way we are trying to solve it at Stanford University is by trying to mix the computer scientists and the biologists more, letting the computer science experts figure out what the important problems are in medicine and how they can engage in those problems, especially for the algorithmic issues. We have been struggling with this problem for a long time, and we have just been pragmatic about it and have gone where we can get the money most easily.

DISCUSSANT: If I am collaborating with the doctor, he writes the NIH proposals and hires my students. Who writes the NSF proposals?

MR. DUDLEY: At Stanford, we now have a lot of rebranding. For example, computer science people are rebranding themselves as bioinformatics experts, because they need funding. Therefore, instead of trying to deal with the problem, they rebranded themselves and went to NIH.

DISCUSSANT: As a lawyer, I sometimes think of myself as a social engineer. Most speakers have described a series of social engineering problems, such as institutional culture affecting funding and institutional culture within the disciplines, so it is a big problem. Imagine this scenario: I am at a midtier institution, and what the presentations showed me is that I am the “sharecropper.” You have got both the brain power and the computing power to utilize all the data, process them, and do different kinds of work with them that I will never be able to do. Therefore, my only competitive advantage against you for funding is to hide the data. How do you respond to that? If you are going to break the logjam, what are the incentives for researchers to share their data with you?

MR. DUDLEY: You are exactly right about that, and that was the point I was trying to make when I said there are computational elites. It should not be that way, however. I think that if we can get some good tools and the cloud, we can level the playing field for costs and accessibility for many researchers at midtier institutions. In fact, we wrote a grant for the NIH to build such a system, which was rejected.

DR. FRIEND: I think the point you made is important. I want to separate researchers into data gatherers and data analyzers. Right now the way it works is that the person who collects the data expects to have the right to take it all the way through to the insight, and then the next person should go back and do it again. That system is wrong. It is just inefficient. I think that having core data zones where people who did not generate the data can mine the data and work with them is important enough that funding institutions and journal publishers need

to determine a way to make it happen. A second issue is the matter of humility. It is very hard to get the “Titans on Olympus” to share their data even among themselves.

DR. HEY: We work with some members of the HIV committee. The moment they send us their data, we send them the results with some tools that we developed. There is no real reason we could not make those tools available in the cloud, and we are planning to do that. However, I do think the cloud can empower scientists in certain cases. We are also working with the Ameriflux community. They have about 150 sensor towers and fields all over America. Typically they are owned by a principal investigator, and they take data and publish it with their graduate and postdoctoral students, so it is generally one professor and two students. By putting the data into a central data server, like the SkyServer, the community is gradually realizing that there is much more value to be gained by publishing not just one tower’s data but several. That is a sociocultural change, because instead of one professor and two students we now can have 10 professors and 20 students, which makes it a 30-author paper. For that particular community, this is a big change in culture, and it takes time.

DR. GOODMAN: In astronomy, the situation is very different, and those of us who like the astroinformatics approach are lucky if we have tenure, because people still do not appreciate that kind of approach. Getting the last three photons out of a Keck telescope is what gets someone the “Olympian” status these days. That is changing slowly, but it is a big battle.

MR. UHLIR: I would like to follow up with a comment and a question. The comment has to do with the issue of incentives. Many people talk about the importance of citation of data and the role of that in giving people incentives to share data, through recognition and rewards, and changing the sociology of how people receive their data and share them. The Board on Research Data and Information is about to start a 2-year project in collaboration with some other groups, in particular, the international Committee on Data for Science and Technology (CODATA), on data attributions and citations. We will start with a symposium in Berkeley, California, in August 2011.

The other issue, which most of you have largely avoided, is the human infrastructure element. Obviously this symposium is about automated knowledge discoveries, so you have appropriately focused on the technical infrastructure and automated knowledge tools, but there seem to be two schools of thought about how this is going to be managed and promoted in the future. One is that the technology will disintermediate many professionals in libraries of science or in information management, because everything is going to be automated and work perfectly, and we will live in a wonderful automated information age. The other school of thought is that the deluge of data will require not only more people to manage it, but also a retraining of the information managers and reorientation of the traditional library and information schools. So my question is, which school of thought is correct? Or perhaps they both are. Also, what do the speakers feel needs to be done to develop this kind of new expertise, because a lot of it seems to be done by people mostly in the middle of their careers? Is there a new approach to education—higher education and training at the outset of careers—that is being implemented, and if not, what would you suggest?

DR. GRAVES: There are new programs starting in data science and informatics. We are initiating one at the University of Alabama in Huntsville. I know Rensselaer Polytechnic Institute (RPI) is also starting some programs, as well as several other universities, and they are attempting to be interdisciplinary.

DR. RAMAMURTHY: The creation of Unidata was intended to address the concern that you just expressed. Atmospheric sciences has been at the forefront of data sharing and of bringing data into the classroom and giving students—at all levels, whether it is an introductory atmospheric science class or master’s and Ph.D. students—access to the same kinds of data and tools that are used by professionals in their operational forecasting or in research. Having that facility that has been sustained for 26 years has made the issue of access to atmospheric science information and data systems almost a nonissue from a workforce development, because literally hundreds and thousands of students go through their labs and classrooms and use state-of-the-art software and data technologies in their daily work. The key is having that kind of a sustained infrastructure that will provide the services that are needed for training the next generation of students and for creating a workforce that will be very skilled in dealing with geoinformatics of the future.

DR. CONTI: There is a need for a paradigm shift in astronomy. The vision that the younger generations bring to this area is very refreshing. They suggest that we can solve this problem from a completely different point of view. At the same time, many of them bring very little knowledge of astronomy combined with an extremely high knowledge of computer science, and those have the ability to solve problems much more quickly and therefore to ask many more questions. Therefore, in our programs, we try to find a good balance between people being trained astronomers, but also having a focus on computer science as well, because the future is going to depend on how they are going to manage data and how they can ask new questions that were not asked before.

The other issue is that once we have all these data, our ability to mine them is extremely important in the sense that we need to be able to understand how the data can be manipulated. Therefore, it is important to change the curriculum to reflect the fact that computer science has to be part of our daily life. We find that younger students are used to managing and producing astronomy and looking at astronomy through the Worldwide Telescope. They do not know the inner working or what it takes to produce plenty of data, but they know what they want to do. They know how to discover new pathways through the data. There is a lot of resistance among senior researchers, but perhaps we just need to learn to say, “I do not really understand how you get to this, but show me how, because we can open a new frontier.”

DR. FRIEND: I want to make three points. The first concerns what we have found to be necessary to bridge the gap between the data mining and the understanding of biology. It is very rare that a single individual does both, and so we split it into two parts. We have people that we call network biologists, who may be mathematical physicists, coming toward the models from one direction, and then we have systems biologists, who understand enough of the math and understand the biology. The people who can do that sort of polymath linkage are rare, and so the education and career paths are split in a way that is similar to what happened previously when someone chose to become either a molecular biologist or an enzymologist. It is important to not expect someone to bridge the entire gap.

The second point is that Dr. Eric Schadt, who drove much of this work, and several others have just started a new journal. The journal will have a section that is just for publishing models that are not validated. A validation often comes later, and the point is to make the model available so that people can work on it.

The third issue relates to the discussion about structuring rewards. In medicine, we are convinced that the patients and citizens themselves must break the logjam between the scientists who do not want to provide the data and those who are trying to get access to large amounts of data. If there are electronically generated data from a patient, the patient can get them back. This is going to be a very fundamental change—shifts in how data are going to flow. The development of the law and the role of the citizen are going to be very important in getting the data flowing, because academic institutions are not necessarily going to do it on their own.

DR. HEY: I have three brief comments. First, I do think that research libraries are in danger of being disintermediated. Librarians are caught in a difficult trap. Subscription costs are clearly a crisis, because the budgets are not increasing at the same rate. There is an interesting move toward iSchools, but the question remains: What should they train the next generation of librarians to do?

Second, training will be important in changing the culture of science. We should work with graduate students in each science, because it is clear we are not going to convince the faculty. I think we should train people to have at least depth in a couple of disciplines. Universities change very slowly, however, so my worry is that we will produce bright graduate students with nowhere to go. I think that is a real issue, although some universities are doing things better.

Third, there are three communities: the scientists doing the real research, the computer scientists who know the technologies, but we also need information technology experts, who can turn a research prototype into a tool that people can actually use.

DISUSSANT: I see the success in bioinformatics and geoinformatics as disciplines, and I see communities of scientists working in multidisciplinary ways with other communities, such as statisticians and computer scientists, but we do not see that much in astronomy. One difference is that there was a tipping point reached—certainly this was true in the medical community—that so many papers are being published that it is impossible for anybody to keep up with what is happening. You therefore need some kind of a summarization or aggregation and an informatics approach to give you the high-level picture. I like the historical example at the University of California, Berkeley, where you can move from a higher-level view down to the exact specific event. Imagine doing the same thing with our databases.

Maybe in astronomy we have not quite reached that tipping point. We have learned with the virtual observatory that if it is built, users will not necessarily come, and so many astronomers are not using this infrastructure. Some are using it, but not as many as we would like.

Then, as I was listening to this education discussion, I thought that perhaps that tipping point could be completely different from the one I was thinking. It would be more the education system that is the tipping point, with the younger people who are familiar with social networking and computer-based tools, whether we call it Web 2.0 or Science 2.0. I would like to call it citizen science, that is, getting people in touch with the science and them being inspired to do science, knowing that they approach it with either the skill sets in using these tools or at least an interest in learning to work with them, because that is the way their lives are, and now they discover they can do science with it.

My university, George Mason University, is one of those places, like RPI and a few others, that are offering this sort of informatics space education. We have found that it is hard to draw the students into the program, especially the undergraduates. Our graduate students already understand it, so the graduate program has a good number of students, but our undergraduates still do not. We talk to students and ask what are they interested in, and they say, “We like biology, but I want to do something with computers.” When we finally get these students into our classes, their eyes open up and they say, “I did not know I could do science with computers.” So, in a sense, the younger generation is reaching that tipping point. They are discovering that they can do what they enjoy doing, such as social networking and online experiences, and do science at the same time.

What I would like to see as a goal is not only doing this kind of work at the graduate level, but moving it to the undergraduate level, as we are doing now, and then moving it even further down to the high school students. I would like to teach data mining to school children. It would not be so much the algorithms of data mining, but it would be evidence-based reasoning and evidence-based discovery—basically detective stories—and in this way we could inspire a generation of new scientists just by using a data-oriented approach to the way we teach these topics.

DR. GOODMAN: Perhaps the core or at least part of the problem is the scientific paper as a unit of publication. I will give an example. Seven first-year graduate students in our department started on their own project called Astro Bites. They go to AstroPH, which is the archive server for astronomy, select the articles that they like, and review them. This is now being read all over the world. It is searchable by any search engine, and they can discover useful materials there. This is kind of the junior version of a blog called AstroBetter, which is produced worldwide by astronomers who are of the mindset that the last discussant was talking about—a kind of astroinformatics community that is evolving in a mostly 30-something world. Those are the same people who would like to take their model or dataset and make it available for people to experiment with. Right now there is an interesting concept of the data paper in astronomy. I did not even know what this was, but fortunately when we finished a very big survey about 5 or 6 years ago, I had a postdoctoral student who said, “We have to publish a data paper now,” and I said, “What is a data paper?” It is basically a vacuous paper that says, “Here are my data. Cite me.” She has gotten hundreds or thousands of citations because of that data paper. This should not be this way.

Going back to Mr. Uhlir’s point about people and training, there is a whole new system that is emerging. I do not know exactly what it looks like, but it does not involve 12-page papers as the currency of everything, and it does not involve even looking at those papers to find the data. Such papers are just one of many resources. How we develop the resources by all these other new people, I do not know.

DISCUSSANT: Part of what we are seeing is the realization that a citation count has become an outmoded way of thinking, and yet it is still the H index that gets people into the academy. Take, for example, Danah Boyd. She has 50,000 Twitter followers. She is not too worried about her citation counts—she does fine in her papers. But, again, what does it mean and why does that work?

DISCUSSANT: The right solution is for those of us who are thinking about this area to engage in that conversation, to look for ways to balance the traditional measures with some

new kinds of metrics. We should look for impact metrics, as well as for sustaining papers. People should still write papers, but we need to get a better mix that looks across a much broader range.

DR. GOODMAN: Where does the data scientist, who maybe does not typically get included in a publication, come in?

DR. CONTI: Most of the time we are penalized for being a data scientist. Perhaps one scientist does not have as many citations as some others, and does not care to publish that many data papers; as a result, the tenure-track path is harder.

DISCUSSANT: One key is that those of us who are involved in other people's tenure decisions have to rethink what we examine, and not get the junior faculty to rethink what they are publishing.

DR. FRIEND: Some people who have been looking at this have noted that the music industry is fairly advanced in figuring out microattribution and that there are ways to borrow from that experience, by looking at impacts without using citations. We should think about using tools that others have developed to determine impact factor.

DISCUSSANT: I work for the National Ecological Observatory Network, or NEON. It is funded by the NSF to build a network of 60 sites across the United States to collect data on the effects of climate change and land use on the ecosystem. We will be measuring something like 500-plus variables, and we will be delivering these measurements freely to the community. The measurements range from precipitation and soil pH all the way to things like productivity. We have talked about data-intensive science, of bringing a new era of how scientists can work, but what about data-intensive science in forming our policy and in forming how we design our cities, for example? From NEON's perspective, we will be one of the many data providers, along with the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and other institutions that will be providing similar data, but somebody else has to do the aggregation of those data into knowledge. Those people will need tools, so who is going to build the tools? And will these tools be used by federal agencies, resource management makers, and policy makers to craft how they will plan the nation's national resources?

DR. HEY: We have a couple of projects that show there is some hope. One is a project called Healthy Waterways in Brisbane. They had a big drought, now they have too much water, but it is relevant in either case. They have a system for doing ecology reports on their water supplies, and it took several people several days to get a report written and distributed, which is released every week. By putting in some serious information technology, but nothing research grade, one person can now do it in 1 hour or so, and that person can get much more flow of information, so ecologists have a much better evidence base with which to do emergency measures, for example. The other project is an urbanization project in China where they are building 10-million-people cities. They will need to worry about all the environmental impact issues, as well as designing the city. I think there is hope, but I agree it is a challenge. I think that more case studies are needed.

MR. DUDLEY: Arizona State University has a building called Decision Theater. I do not know how successful it has been, but this building has Liquid Crystal Displays around the wall, 360 degrees, and it was built specifically for policy makers to develop evidence-based

views of policy making. I heard rumors that the state legislature wants to defund the entire Arizona State University, so maybe it was not too successful, but it is something to look into.

DR. RAMAMURTHY: As atmospheric scientists, geoscientists, and geoinformatics people, almost all of the data we produce has societal relevance. It is important not to stop with scientific informatics, but to try to determine a way to interface it with decision-support systems and other kinds of policy tools. In the climate change arena, everyone is focusing on mitigation and adaptation, because it is not just whether the globe is going to be two or four degrees warmer in the next 50 or 100 years, but what it means for the watershed, for agriculture, for urbanization, for coastal communities, and so forth. It is absolutely critical that the scientific information systems and geoinformatics systems can interface with the other tools and other systems. We do not build those systems in my program, but we do think about interfacing with geographic information systems, such as what needs to be extracted from climate-change models to help people talk about what the agricultural picture is going to look like 50 years from now, whether for growing wheat or barley or something else.

DR. GOODMAN: I have a question for the people from the funding agencies. Dr. Hey said “some serious information technology, but nothing research grade.” This is a huge problem. Here is a true story. At a meeting I had yesterday we had an undergraduate researcher who works with a postdoctoral student on a project, who said, “I do not know where we are going to put these data, because we need a couple of terabytes, and the computation facility has a lot of security systems, so they will not let us run the right software,” and so on. I knew that my assistant had some extra money from the division and that she bought a stack of 40 boxes of terabyte drives, so I gave a terabyte drive to each of the researchers, and I solved the problem. The point is that there was an obstacle and that there was no person whose job it was to facilitate that and to make the data usable and streamlined for the future. So my question is, How are we going to pay those people in the future? Where are we going to get those people?

DR. BERMAN: If we think about the impact of digital information—what we need to store, how long we need to store it, how we need to curate it, and how we need to respond to data management policies—the university libraries are reconceptualizing themselves and can step into that role, but they need some help. So the question is, How can we jumpstart the system, so that when universities are thinking about infrastructure costs, they consider the data bill along with the water bill, the liability insurance, and so on?

3. How Might Open Online Knowledge Discovery Advance the Progress of Science?

Technological Factors

Session Chair: Hal Abelson

-MIT-

Interoperability, Standards, and Linked Data

James Hendler

-Rensselaer Polytechnic Institute-

I will focus on the technological challenge of online knowledge discovery. I will first address interoperability issues and then talk about some promising developments.

Scientists can learn a lot about data sharing just from observing what is happening in the outside world. There are many data initiatives and developments outside academia to which we should be paying more attention.

Expressive ontologies are not a scalable solution. They are a necessary solution in certain domains, but they do not solve the interoperability problem widely. They allow a scientist to build his or her silo better, and sometimes they even let the silo get a little wider, but they are not good at breaking down the silos unless the scientist can start linking ontologies, in which case he or she has to deal with the ontology interoperability issue, as well as the costs and difficulty of building them and their brittleness.

I am known for the slogan “A little semantics goes a long way.” I have said this so often that about 10 years ago people started calling it the Hendler hypothesis at the Defense Advanced Research Projects Agency (DARPA).

We are used to thinking that a major science problem is searching for information in papers, and we have forgotten that we also have to find and use the data underlying the science. A traditional scientific view of the world might be, “I get my data, I do my work, and then I want to share it.” But the sharing should be part of how we do science, and other issues such as visualization, archiving, and integration should be in that life cycle too. I will talk about these issues, and then I will discuss the kinds of technologies that are starting to mature in this area. These technologies have not yet solved the problems of science and, in fact, largely have not been applied to the problems of science.

Scientists do use extremely large databases, and many of these data are crucial to society. On the other hand, we scientists tend to be fairly pejorative about something like Facebook, because Facebook is not being used to solve the problems of the world. I wish science could get the number of minutes per day that Facebook gets, which is roughly equivalent to the entire workforce of a very large company for a year. We are also not used to thinking about Facebook as confronting a data challenge per se, but it collects, according to the published figures, 25 terabytes of logged data per day. That sounds like the kind of numbers for large science data collections. Facebook’s valuation is estimated to be well over \$33 billion, which is the size of the entire National Institutes of Health budget. Not surprisingly, they are able to deal with some of these data issues that are discussed here. We need to look at what they are doing and determine if we could take advantage of some of their approaches.

I do not have similar numbers for Google, because they have not been published. The last good estimate I could find was in 2008, which was 20 petabytes per day, but that was 3 years ago. That also does not include the exchange of or the storage of YouTube data, which I cannot find good numbers about either. Google’s valuation in 2011 is about \$190 billion,

which is roughly a third of the Department of Defense budget. Not the research budget—the entire budget.

Therefore, if we think that this kind of work is expensive, yes, it is, but there are other people doing it, and they are investing large amounts of money. It is not surprising that they have been focusing on big data problems in a way different from scientists, and they have been able to explore some areas that are very difficult for us to explore. If a researcher wanted to buy 10,000 computers for data mining purposes, it would be difficult to do because of the lack of resources.

Several speakers talked about semantics in the context of annotation and finding related work in the research literature. It is an important problem, but I do not think it is the key unaddressed problem in science. In fact, I would contend that we have spent a huge amount of money on that problem, much of it on trying to reinvent things that already exist in a better form from open sources in the real world.

Most companies today that want to work with natural language processing start with open-source software. For example, there is a company that is taking everything from the Twitter stream, running it through a number of natural language tools, and doing some categorization, visualization, and other similar work. They did not build any of their language processing software. I am also told that Watson, the IBM Jeopardy computer, had a large team of programmers, but, in fact, that the basic language technology used was mostly off the shelf, and that it was mostly statistical.

Semantic MedLine is a project that the Department of Health and Human Services has invested in. It does a fairly good job but does not understand the literature sufficiently to find exactly what we want. But we are not able to do that in any kind of literature, and Google is working on that problem as well. Hence, I do not see the point of yet another program to do that for yet another subfield or against yet another domain. Instead, we need to start thinking about how to put these kinds of technologies together.

The Web is a hard technology to monitor and track because it moves very fast. It has been moving very fast for a long time, however, and, as scientists, we need to start taking advantage of it much more than reinventing it.

There are a few different tools and models on the Web that are worth thinking about. One is to move away from purely relational models; from assuming that the only way to share data is to have a single common model of the data. In other words, to put data in a database or to create a data warehouse, we need to have a model, and that is done for a particular kind of query efficiency.

Sometimes, however, that efficiency is not the most important factor. Google had to move away from traditional databases to deal with the 20 petabytes of data generated every day. Nobody has a data warehouse that does that, so Google has been inventing many useful techniques. “Big Data” was one of their names for their file system. We cannot easily get the details of how Google does it, but we can get published papers from Google people that will be useful in learning how to do similar work. The NoSQL community is a fairly large and growing movement of people who are saying that when you are dealing with large volumes of data there is a need for something different from the traditional data warehouse.

I had a discussion with some server experts who said, “Just give us the schemas, and we can do this work.” I pointed out that we cannot always get the schemas, that some of these datasets were not even built using database technologies with schemas, and in some cases someone else owns the schema and will not share it. We can still obtain the data, however, either through an Application Programming Interface (API) or through a data-sharing Web site.

Although Facebook’s and Google’s big data solutions are proprietary, the economics of applications, cell phones, and other new technologies are pushing toward much more interoperability and exchangeability for the data, which is mandating some sort of a semantics approach. Consider, for example, the “like” buttons on Facebook. Facebook basically wants their tools and content to be everywhere, not just in Facebook, which means it has to start thinking more about how to get those data, who will get that data, whether it wants data to be shared or not, and what formats and technologies to use. As a result, there are many technologies behind that “like” button. Similarly, there are a number of approaches that Google employs to find better meanings for their advertising technologies. What happens is that Google recognizes that at one point it will not be able to do the work by itself, because there is a long-tailed distribution on the Web. This means that it has to move the work to where Webmasters and other people will be able to develop the semantic representations for their domains.

That is happening with all the search engines, and the big issue now in that area is simple metadata and lightweight semantics. The notion of the complex ontology is getting replaced by the notion of fairly simple descriptive terms—that is, I can probably describe the data in my dataset with 10 or 12 different terms that will be enough for me to put in a federated catalog for people to decide whether they want to read my data dictionary.

The idea here is that if the data are going to be in a 200-page, carefully constructed metadata format with the required field-specific types, and they have to be compliant with the standards of the Internet Engineering Task Force and ISO (International Standards Organization), the vocabularies get harder to work with. It is an engineering problem that is very similar to the integration problem, and what many people are realizing is that you can arrange the data hierarchically and get good results for small investments.

Here is an example. Many governments are putting raw data on the Web now. They are not just putting visualizations of the data online; they are putting the datasets themselves. From a political point of view, the two biggest motivations are enhanced transparency and the chance to inspire innovation. Promoting innovation can proceed in two different ways. First, the governments are hoping that some people will figure out how to make useful and innovative tools using the data. More important, especially for local governments, is that they pay many people to build Web sites and interactive applications that they cannot afford to build anymore.

For instance, a government agency may be able to pay one contractor to build a big system for a problem that is of high priority, but it may have 57 more priorities that it cannot afford to fund. The governments need someone to develop those applications, so if it makes the data available, at least some of those applications are done by other people. The development of such applications is out of the agency’s control, but if the work is getting done at no cost or for some small amount of money, it can start planning strategically for other priorities.

The United States and the United Kingdom are the two leading providers in terms of organizing their data. The United States has about 300,000 datasets, most of them geodata. If

we take out the geodata, there are probably 20,000 to 30,000 datasets that have meaningful raw data. The United Kingdom has less, but it probably has the best in the sense that, for example, you can get the street crime data for the entire country for a number of years. They are releasing very high-quality data down to a local level.

Many countries, including ours, are releasing scientifically interesting data, but you would have to work to find them. To use these data, combining them with other data, can be more important than just looking at them. Those entities releasing data include countries, nongovernmental organizations, cities, and other groups. For example, one of the Indian tribes in New York State has released much of its casino data.

There are groups all over the world working with these data. My group is one, and the Massachusetts Institute of Technology has a group working jointly with Southampton University, mostly on U.K. data. Many of these groups are in academia, but there are many nonacademic groups doing this kind of work. One of my suggestions is to think about data applications. You may have a large database, and if parts of it can be made available through an application, an interface, or through an API, people would be using the data in a sharable—and often an attributable—way.

Getting back to science issues such as attribution and credit, we have one of our government applications in the iTunes store. I know exactly how many people downloaded it yesterday, and how many people are still using an old version.

When some students and I were in China, we discovered that China was releasing much of its economic data, so we took China's gross domestic product (GDP) data and the U.S. GDP data to do a comparison. To do that comparison, we needed to find the exchange rate. Luckily, there is a dataset from the Federal Reserve that has all of the exchange rates for the U.S. dollar weekly for the past 50 years or so. We got those data, we mashed them together, and we got a chart that looks like Figure 3-1:

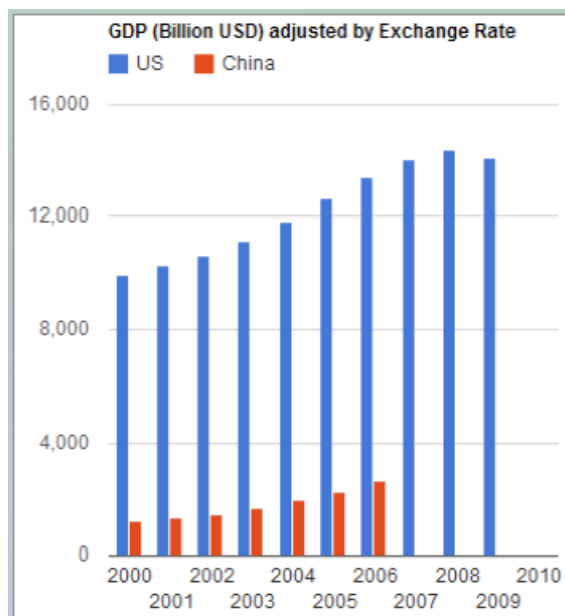


FIGURE 3-1 GDP Chart.
SOURCE: James Hendler

We also divided the GDP by the population (data we found in Wikipedia), so that we could click a button to go between total GDP and per capita GDP. The model was built in less than 8 hours, including the conversion of the data, the Web interface, and the visualization. That is a game-changer. It would completely change the way we work if we could get this down to a few minutes. When my group started working with this kind of technology several years ago, it would take us weeks to do this kind of work. Part of the improvement is because some of the technology was immature, part of it is because the tools are now available commercially, and another part is because there is now visualization technology freely available on the Web. Building visualizations is labor intensive, so by moving to simple visualizations, we can use visualizations much earlier. We are also able to link government data to social media.

One interesting question that has not really been part of the discussion that we have had in the scientific community is how to find data. For example, we noticed that most of the U.S. government data were about the states, and the metadata would represent the data as being about the 50 states, but very few of the sets actually covered all the states. Some states were missing. Some databases included American Samoa, Guam, Puerto Rico, et cetera, and the District of Columbia (which is not officially a state). In this case, there is a very loose definition of a state as opposed to a territory, and no one has much problem with that. But if a researcher wants to find datasets about Alaska, it can be erroneous to just assume that all of the datasets that say they cover the states will actually have Alaska data.

The other problem is that we cannot search for the keyword “Alaska” within the datasets. If there is a column that represents the states, it may be called Alaska, it may be called AK, it may be called state two, or it may be called S17:14B/X. The government has terabytes of data, so how do we find the data that are for Alaska, for example? We cannot just call for building a data warehouse and rationalizing the process, because these datasets are being released by different people in different agencies in different ways.

Thus, metadata becomes very important. Simple and easy-to-collect metadata can allow building faceted browsers and similar tools. It is a real research challenge, however, to determine the kind of metadata for real scientific data that are powerful enough to allow useful searches. With the government data, one problem is that all of the foreign databases are in their own languages, and it is hard for those of us who are English language speakers to figure out what is in the Chinese dataset, unless you hire a Chinese student. You cannot just use Google Translate and expect the result to be academically useful.

People are starting to consider integrating text search and data search. This is an application that one of my staff did, joined with Elsevier’s new SciVerse, which was featured on the U.S. Data.gov site (Figure 3-2).

What this application does is that when we are doing a keyword search for scientific data, it is also looking for datasets that might be related to that same term. We are using very lightweight ontologies that mostly just use the keyword. We are working on making it better. We think it would be good when someone is looking for papers, the data in or about that paper were available, but also finding what is in the world’s data repositories that might be useful.

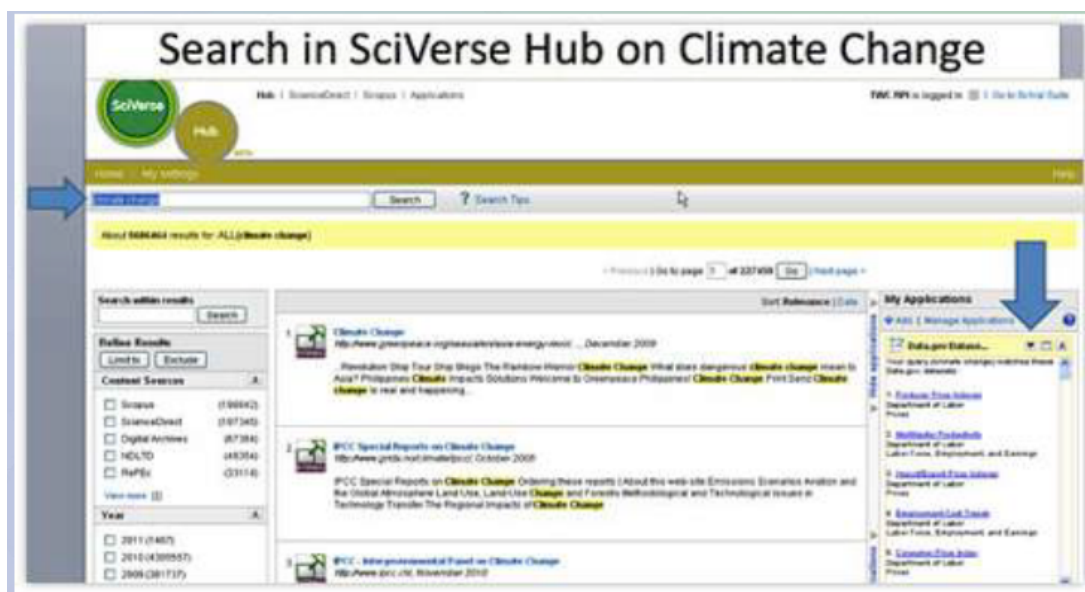


FIGURE 3-2 Search in SciVerse Hub on Climate Change.
Source: James Hendler

When we start thinking about data integration, using data, or searching for data, one thing that happens in the discovery space is that the data become part of the hypothesis formation. That means that looking at, visualizing, and exchanging the data cannot be an expensive add-on to a project. It has got to become a very key part of the standard scientific workflow, with appropriate tools to reduce the costs.

What is promising in this area? We have been looking at linked open data issues outside of science, and some of its promise has been explored (mostly within the ontology area). Genomics and astronomy are two fields where we have actually seen semantic Web technologies deployed, but many other science communities are still mostly thinking about their own data holdings, not about being part of a much larger data universe. It is interesting that when we talk to the Environmental Protection Agency, or to the National Oceanic and Atmospheric Administration, or similar agencies, they say that they are providing data to many communities, so they cannot easily use the ontologies of a particular one. Hence, we need to learn how to map between these approaches.

How do I know what is in a large data store? How do I know what is in a virtual observatory? I need services, metadata, and APIs. I also need to know the rules for using the data.

Other issues we need to think about are related to policies. If I take someone's data from a paper, mix them with someone else's data, and republish them, those people would probably like to get credit for the data being reused. Or if I do some work on your data that you do not like, you might want to rectify it. How do we do that?

National Technological Needs and Issues

Deborah Crawford
-Drexel University-

I am currently the vice provost for research at Drexel University, but I used to work at the National Science Foundation (NSF). I spent almost 20 years at NSF and was fairly significantly involved in the agency's cyberinfrastructure activities. My experience within NSF and now at Drexel has provided me with an interesting perspective on what data-intensive science means in a research university that aspires to be more research intensive. The main message of my talk is that there have been many advances in data-intensive science in some fields, but we have massive amounts of work still to do if data-intensive science is to realize its full potential across all of the disciplines.

There have been many reports issued over the past decade that address the importance of data-intensive science and the role of information technology in advancing science. An important one is the Atkins report that Dr. Hey referred to earlier. In that report, Dan Atkins of the University of Michigan and his committee speak of revolutionizing the conduct of science and engineering research and education through cyberinfrastructure, and they examine democratizing science and engineering. Those were tremendously powerful statements in 2003, and they stimulated a great deal of excitement within the scientific community.

Since then, there have been a number of reports that have specifically addressed data-intensive science, several of which were released in the past couple of years. This is a quote from a joint workshop between the NSF and the U.K. Joint Information Systems Committee that was held in 2007: "The widespread availability of digital content creates opportunities for new kinds of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data."

What we have heard so far in this workshop was from those who I would describe as the visionaries and the trailblazers in science and engineering, representing those communities that were very motivated to take advantage of information technology in their work in order to advance their field of science.

I want to focus now on the long tail of science—the researchers in those fields where the immediate advantages of information technology and collaboration are not so readily apparent. There is tremendous opportunity in those communities, but we do not quite know how to realize those opportunities.

I would like to provide a snapshot of some surveys of researchers working in different communities. The main message is that computer-mediated knowledge discovery practices vary widely among scientific communities and among colleges and universities. In the United States and in the United Kingdom, for example, this is certainly true. There are some colleges and universities that know how to take advantage of their digital technologies and their digital capabilities, while there are others that simply are way behind the curve.

There are three fundamental issues that communities or universities must address: (1) What kinds of data and information are made open, at what stage in the research process, and

how? (2) To which groups of people are the data and information made available and on what terms or conditions? (3) Who develops and who has access to the tools and training to leverage the power of this discovery modality? These are fundamental questions and are key to the realization of data-intensive science.

I am going to talk about two case studies, one conducted in the United Kingdom by the Digital Curation Center, and one conducted within Yale University. Both point to some key features we need to address.

The study done by the Digital Curation Center in the United Kingdom was called “Open to All.” Its purpose was to understand how the principles of digitally enabled openness are translated into practice in a range of disciplines. These are the kinds of questions we need to be asking ourselves to determine the actions that we need to take to make sure that this modality of science is accessible and advantageous to everyone.

In this study, the authors characterized a research life-cycle model and asked different communities how they were using digital technologies within the context of that life cycle to further their science. They surveyed groups among six communities: chemistry, astronomy, image bioinformatics, clinical neuroimaging, language technology, and epidemiology. The range of responses within those different communities was fascinating to see. Surprisingly chemistry was the trailblazing community, at least among the individuals surveyed for this study. The chemists were using social networking tools, Open Notebook Science, wikis, and all the modalities of digital technologies to collaborate, and to collect, analyze, and publish their data. Everything was quite seamless from a community that I had not anticipated to be one of the trailblazing communities.

It was interesting to hear from the clinical neuroimaging group. They were so skeptical of the value of data-intensive science and open data sharing that in this study they refused to even disclose their identities as individuals. We therefore went from one extreme to the other, and we saw the range of practices and values within different scientific communities.

From their conversations with these communities, the group that conducted these case studies in the United Kingdom came up with a list of the perceived benefits of open data-intensive science. It included improving the efficiency of research and not duplicating previous efforts, sharing findings, sharing best practices, and increasing the quality of research and scholarly rigor. This last one was especially true for the members of the chemistry community who were surveyed. They saw a great opportunity in making available in blogs the day-to-day data that they collected—not just raw data, but derived data. They found tremendous benefits in making that open to the community to provide more scholarly rigor.

Among the other perceived benefits were enhancing the visibility of access to science, enabling researchers to ask new questions, and enhancing collaboration in community-building, which all of the groups surveyed agreed was a benefit. There was also a perceived benefit of increasing the economic and social impact of research about which the report was ambiguous. Although economic and social impacts each are treated separately as a perceived benefit, the sense was that the real value could not actually be measured. So, there was a question: Can we create economic value by making our data—essentially our intellectual property—much more openly accessible?

One of the perceived impediments was the lack of evidence of benefits. Researchers who felt this way were not motivated to make their data openly available. Other impediments were the lack of clear incentives to change and the values in academe not being consistent with open data sharing. For tenure and promotion and the drive to publish, the perceptions were that the only way to publish is in the open literature and that it is not in a researcher's interest to share data, because someone else may use the data to advance farther than the one who shared the data. Competitiveness is a big issue.

The conflict with the culture of independence and competition is absolutely related to the lack of clear incentives to change. Other impediments were inadequate skills, a lack of time, and insufficient access to resources. Another big concern was how to train both the scientists who are practicing today and the scientists and engineers of the future. Researchers were also worried about how long it took them to prepare their data for open access, about quality, and about ethical, legal, and other restrictions to access. Those were big issues, especially in the life sciences community.

The report's recommendations called for policies and practices for data managing and sharing. Communities are desperate for guidance on these issues. What have the trailblazers learned that can be shared and applied more broadly? Contributions to the research infrastructure should be valued more—that is something we have heard often. Training and professional development should be provided, and there should be an increased awareness of the potential of open business models. That is related to the attitude among researchers that their data are their intellectual property and they want to derive some value from that; thus, they feel that if they make their data openly available, they are giving that value away. Assessment and quality assurance mechanisms should be developed.

The study done by the Yale University Research Data Taskforce was conducted by an organization within Yale called the Office of Digital Assets and Infrastructure, which is an organization established to accrue the value over time of the digital assets that result from research and scholarly activity within the institution. The office conducted this study to determine the requirements and components of a coherent technical infrastructure, to provide service definitions, and to recommend comprehensive policies to support the life-cycle management of research data at Yale.

Given the discussion earlier about research libraries, this is interesting. Here is Yale University doing a survey that includes both the librarians and the information technology enterprise staff at the university to determine what their faculty base most needs for managing digital data. Very much like the other survey, this one found that data-sharing practices vary widely among the disciplines they surveyed. The researcher has the most at stake when determining what the data-sharing practices are. Yale is going to create an institutional repository to secure, store, and share large volumes of data.

There are certainly some institutional pioneers in this area, such as the Massachusetts Institute of Technology and Indiana University, and much of what I characterize as “slow followers.” A major concern is how an institution can afford to create and maintain infrastructure like this.

Yale University understands that it needs to develop and deliver research data curation services and tools to all of its interested parties, not only in science and engineering, but also in the humanities and in the arts. Recognizing the importance of persistent access and

preservation, Yale is trying to establish a digital-preservation program to determine what they need to preserve data in the long term and what value will be accrued from it.

Data ownership was a big issue in the report. Who actually owns data? Most faculty believed they owned their data. It will be important to develop clear policies regarding data ownership, especially in those communities where data sharing is not a standard practice or is not valued.

My main message, then, is that changes are in play at multiple scales: international, national, institutional, individual, and community. Most of the effort at this time has been focused on the community level, because communities have, for the most part, driven the transition to open data sharing in communities where clear value accrues from that sharing. Culture does not change because we desire it to change; it changes when our practices change. The bottom line is that all of this costs money. We will need to determine the appropriate balance between the cost of improved access to scientific opportunity and the benefits that accrue from it, and those are not easy questions to answer.

Discussion

DISCUSSANT: Is it going to get to the point where IBM's Watson knows the scientific literature and we do not ever need to do any work ourselves on the literature databases?

DR. HENDLER: No. For Jeopardy, Watson took many fairly high-end processors to beat Ken Jennings. For the medical version, they are trying to get that down to a more manageable number—maybe 20–50. They have already announced that the first issue they are pursuing with Watson is differential diagnosis, helping doctors access the medical literature. They designed the computer to be repurposable. This was the plan until 2009, but once they had the agreement that they were going to play Ken Jennings; they stopped doing anything except Jeopardy-related work. Now that they have won the game, they are back to the earlier focus. My guess is it will be 5 or 10 years until that becomes widely deployable.

DR. HEY: I am interested in the financial implications of sharing literature and data. Just as the music industry has to deal with what happens when someone makes a digital copy and distributes it for free, openly available data changes the business model. Similarly, with publications, we do not need printing presses anymore; we have the Web, and we can make digital copies and disseminate them. The big budgets that go to the publishers with their present business models perpetuate the system, but the system is being broken and it is up to universities and university libraries to respond. They should not be doing such deals with the publishers. They should be thinking about their role in the future. They should be thinking about how libraries can assist the research process of the institutions. I think the answer can be found in places like the iSchools. It is up to places like Drexel University and Rensselaer Polytechnic Institute to move forward, but how do we get that to take place? How do we catalyze the major research libraries and major research institutions to act together?

DR. CRAWFORD: There is nothing like a crisis to stimulate action. I know that at Drexel we are having conversations about the future of our libraries. We have an iSchool as well, so there is an opportunity there, because having shared interests in reinventing the model is essentially what it will take. There are trailblazer institutions, and it is a question of whether we can unite and define what the library of the future is in this digital world. But I absolutely agree that it is necessary. It is the future of the knowledge enterprise. I had a meeting with Drexel's president last week in which we talked about the iSchool being the centerpiece of the twenty-first-century university and what that means in the context of research and scholarly activity.

DR. HENDLER: I would also mention projects like VIVO¹², which is funded by the National Science Foundation. It has gotten a few schools involved, and it is growing. They have established the sharing of data and integrating that for the purpose of tracking researchers who agree to become a part of that system. Each library can maintain its own holdings. If I have a list of all my papers that I can get to my university in a way that they can curate, share, and archive, then it does not necessarily mean anymore that it must be in a particular book, in a particular building. We are seeing a lot of pressure in that direction, and it is promising that several of these big libraries are helping to lead that area.

¹² See: <http://vivo.ufl.edu>

I think the idea of a “catalog of catalogs,” the federated catalogs for not just the library holdings but also the people resources at a university—the individual papers and publications of the people, including who, what, when, where aspects—is a very powerful and interesting idea. Exactly how it will be funded remains to be seen, but once we get 100,000 or a million people accessing a page, the money can be found to maintain it. The problem is that until we reach that number, we cannot figure out how to monetize it. This is a situation in which scalable solutions that are experimented with and evolved may be better than trying to design a solution all in one step.

DR. ABELSON: When scientists say, “My data is my intellectual property,” I am not sure under what legal regime someone’s data is their intellectual property. Perhaps it is their trade secret or something similar. But there is tremendous confusion, and I think it is good to have some healthy criticism when people use that language. I would like to resist having people saying their data are their intellectual property.

DR. CRAWFORD: It is a misguided perception, but there is a lot of that.

DR. HENDLER: Dr. Abelson has been working on how we can make these policies more explicit through common-use licenses and the data more sharable, to actually tie the policies in a computable way onto the data and assets.

Sociocultural, Institutional, and Organizational Factors

Session Chair: Michael Lesk
-Rutgers University-

Sociocultural Dimensions

Clifford Lynch

-Coalition for Networked Information-

Sociocultural factors in this area cover a very wide range of topics. In preparing these remarks, I deliberately wanted to avoid some of the very common themes that I expected to hear in this workshop and to focus on a few additional important, but less widely discussed, issues.

I will not thoroughly discuss, for example, the issue of incentives, which is obviously a major problem. The one point I will make in this regard, however, is that we have not talked much about the incentive to receive research funding in the first place, which is a fairly compelling one. If certain data-sharing behaviors are a condition of receiving research funds, and adherence to these behaviors is enforced as a condition of continuing eligibility for funding, this would be a real incentive.

It is interesting to see what is happening in some of the communities with foundations dedicated to the cure of specific, often rare, diseases. These foundations tend to be very focused on what they are doing, which is searching for treatments and cures. It is now common for them to insist that the scientists they fund should make their data openly available at least to other researchers within that community, on an immediate basis. This is the phenomenon of a funder who is focused on solving a problem and is not very interested in accommodating much long-term career-building by the participants. We will see more of this focus.

Let me move to the main issues I want to address. First, I want to talk about the scholarly literature and the realities of mining that literature, both as a research project in and of itself (i.e., to study what is going on in science) and as a pragmatic means of, for example, finding relevant earlier work. There is a longstanding history of researchers trying to extract hypotheses, insights, or facts by computational analysis of large bodies of scholarly literature, sometimes in conjunction with ancillary datasets of various kinds and sometimes not. What has happened with the way people compute on the Web is that we are coming to think about the scholarly record as an object of computation and to see that this computation can facilitate the discovery of new knowledge. It is important to understand that while the potential of this approach is great, the pragmatics of realizing it are phenomenally challenging.

It is clear that very large amounts of the scholarly literature are in one or another electronic format, some formats being far more hospitable to computation than others. They are often available in multiple formats, with the most computationally friendly ones only available internally on the publishers' sites, since they are not designed to be directly read by human readers. But the licenses that universities typically enter into with publishers to use collections of online journals specifically forbid mass copying of the material from those journal collections, as someone would need to do to build up a corpus to compute upon. In fact, if a researcher tries to download a large number of journal articles from many of these databases, the librarian at that university will get a call from one of the publishers, because they have got mass-download detectors that notice a big spike of downloads. The publisher will demand that

the university librarian do something about this or it will shut off the university until it fixes this issue, at which point the researcher will get a phone call acquainting him or her with the terms and conditions under which the university has access.

Fixing this issue is not simply a matter of renegotiating one contract. A large university will have hundreds of such contracts, and thus far, publishers have not been very accommodating of this, with a few exceptions, mostly open-access publishers that do not require license agreements anyway. Of course, there are legitimate publisher concerns that need to be recognized in the renegotiations, but they are not simple to address. In short, there is a substantial pragmatic barrier to getting literature to mine. Note also that any comprehensive subject corpus, most of the time, needs to be drawn from a range of publishers, not just one.

Even if someone can get past the license issues, a second very real challenge to literature mining is the need to normalize data from a wide range of sources into a consistent format. One of the previous presenters talked about what it took to make sense out of one gene array dataset and the cleanup that was needed for that process. Imagine trying to deal with 50 or 60 different data sources, some of which vary depending on whether the journal issue is from 2003 or 1992, or is something from the 1960s that had to be converted into digital form. This is a very significant amount of work, particularly if it has to be done over and over again by different literature-mining groups.

A final subissue on barriers to mining the current scholarly literature base is related to copyright issues. Consider, for example, what happens if someone is performing a computational process on tens of thousands of copyrighted objects. Is the output a derivative work? If so, producing a derivative work from thousands of copyrights raises all sorts of questions. For example, if we did the same computation, but omitted 10 articles, does the fact that we included the 10 articles in our original computation make the computation a derivative of those 10 as well as of the other 9,990, even though they made no difference in the final product? There is not much legal certainty about anything in this area, as far as I know.

Let me address some practical barriers to mining the existing literature base. Today there is little consensus about the role of publishers in facilitating literature mining. We could imagine scenarios in which publishers, or some third party working with publishers, offer literature mining as a service for a database created just for that purpose. This might reduce some barriers, but it also raises concerns about the economics of text mining and the flexibility of the mining technology that it would be possible to apply to the literature base.

Focusing on the idea of computing on a corpus of scholarly literature and a body of experimental results, we also need to think about how to change the character and composition of this corpus to make it as valuable as possible for this kind of computational analysis. For example, it has often been observed that the published record, especially when it was read solely by human beings, has a strong bias toward positive results. People rarely get articles published in high-visibility venues saying, “I tried these six approaches, and none of them worked,” or “This compound seems to be good for nothing, it does not react with anything.” These are valuable contributions in a computational environment, and we need to think carefully, especially in a world full of robotic experimental equipment and large-scale screening systems, about how we get more of this negative information out there so that we can compute on it.

A closely related issue is the tremendous amount of data that industry produces and represents negative results or at least “prepositive results.” We should be able to disseminate and share these results. We need to sort through the notion of value in data, particularly (but far from exclusively) as it affects industry.

There is a school of thought that a database has value even if we invested in it and it proved to be a worthless investment, because if we keep it secret, we can force your competitors to waste time and money recreating that database themselves. Thus, it will consume resources that they could have spent elsewhere, and that will help us in competition. We need to reject that thinking, especially in environments such as the pharmaceutical arena, where we are struggling to contain costs and improve productive research. There is an enormous social cost to secrecy for negative advantage rather than for positive advantage. I do not know how we fix it other than by talking openly about it, but it does seem that we should look for ways to make as much negative data available as possible, in addition to positive data, to facilitate computation on the scholarly record.

One of the other challenges we face is how to make it easy for scientists and scholars to share data. We heard in some of the earlier presentations how difficult that can be. One key is simply giving scientists places to put materials that can be openly available to the world or that can be shared on a rather controlled basis. Anyone who has tried to do inter-institutional data sharing without making them public to the world—to share data just with groups of collaborators—knows how horrible that can be, because you do not have a common federated identity-management system across those institutions, so you have to resort to all kinds of clumsy ad hoc solutions.

One of the contributions that supporting organizations, such as information technology and libraries, can make is to help supply infrastructure that supports simple sharing. I see institutional repositories as key—but far from exclusive—players in that role. We see this from the discipline-related repositories that exist for specific genres of data. Those make sharing very easy, but we need to look for more mechanisms to facilitate that.

The last area I want to talk about is privacy. We face a big burden dealing with personally identifiable data of various kinds, the need in university settings to get involved with institutional review boards (IRBs), and the specific challenges of sharing and reusing data, which seems philosophically opposed to what most IRBs want to achieve. For personally identifiable data, they want to be as constraining and specific as possible. We need to recognize that we are getting into situations where we need to be able to reuse data more fluidly, and we are going to need some serious conversations about how this interacts with policies about privacy and informed consent.

There are several possible approaches in this regard. In many cases we want to anonymize the information, and we need to learn how to do a better job of that technically and to look for easy ways to share anonymized data with an agreement that the recipient will not attempt to de-anonymize it as a second line of defense. In other cases, personally identifiable information may be needed, particularly as we look at the interfaces between personal histories of various kinds and genomics or other kinds of population studies.

I wonder whether there are not some new ways to think about this issue. There is a tremendous interest in citizen science, as was mentioned earlier. There are many people who are interested in documenting their own behaviors, their own physical health. I was at a

meeting recently looking at developments in personal archiving, and it is clear there is a growing interest in collecting and sharing this kind of information among a sector of the population. We also have some very interesting developments in personal genotyping, for example, where communities of people are sharing the results that they get back from companies like 23andMe that do this sort of work.

I wonder if there is some way to connect these kinds of social changes into the process, or at least to think about whether we can do something better than the very constrained kinds of informed consent approaches that are in use today. We are significantly limited right now in doing the kinds of knowledge discovery that cross different kinds of databases, ranging from biological databases to databases about people and about behaviors and populations. I would suggest that is a major cultural and social issue that we need to be exploring in detail.

Institutional Factors

Paul Edwards

-University of Michigan-

The focus of my talk is on institutions and open data. I want to start with a success story. The meteorological data infrastructure is one of the oldest global infrastructures. It goes back to the 1850s. In the interwar period, a rudimentary, global data-sharing system emerged that used shortwave radio, telegraph, and many other technologies. At that time, meteorologists threw away almost all the data they received, because they had no way to use it. Once they got computers, they became able to process that data. Since the 1960s, the combination of a global observing system with many kinds of instruments linked to a global telecommunications system has been used. The system pours data from all over the world into computers that make forecasts and data products and then delivers those to national meteorological services, and it works well.

However, we do have a problem. Figure 3-3 is a graph of the famous global temperature curve as collected by different investigators since the nineteenth century.

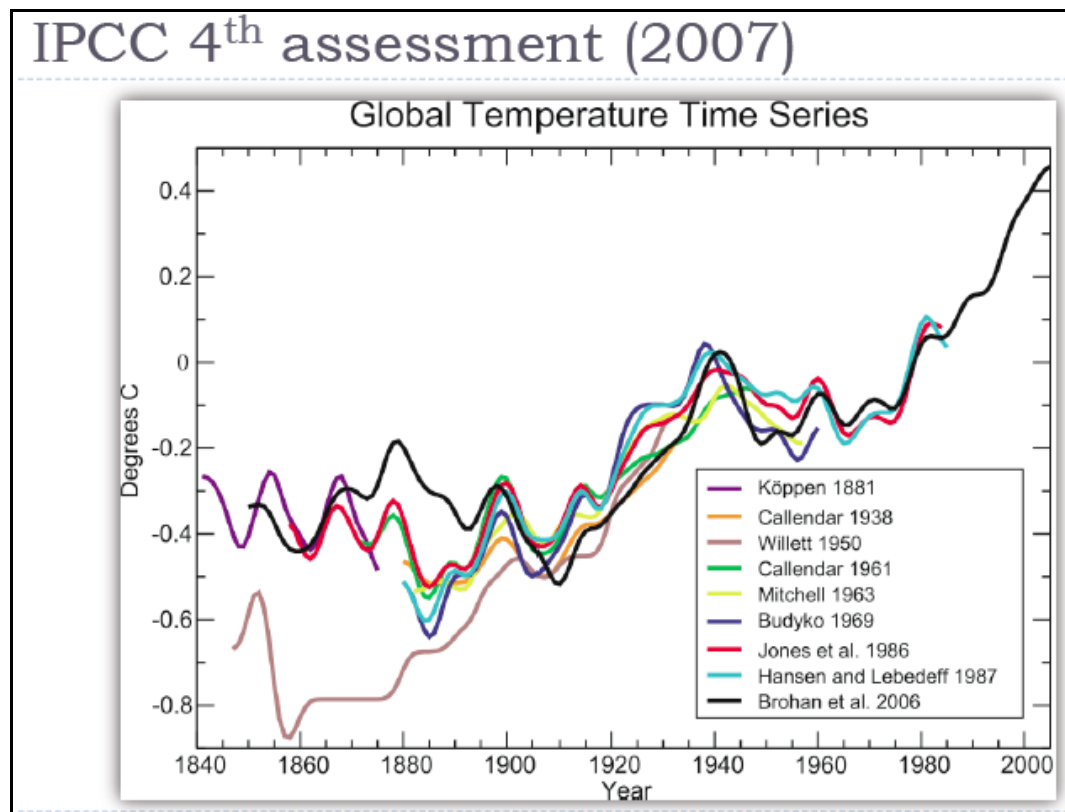


FIGURE 3-3 Global temperature time series.
SOURCE: Intergovernmental Panel on Climate Change

There are enormous differences in the temperatures recorded by the datasets from the nineteenth century and the early part of the twentieth century, especially before 1940. This is not much data by today's standards, but the ability to crunch these numbers was not there before computers. Before 1963, almost all of these datasets included fewer than 450 stations. They did this because these were the stations known to be highly reliable. They were the stations that had been there over the entire period, and which had not made many changes to their locations, instrumentation, and so on. So their data were considered reliable.

The later temperature curves, put together in 1986, 2006, and more recently, use computer models to add in data from stations that were not trustworthy enough earlier—either because their records were incomplete or because they had gone through some changes, such as having been moved or engulfed by an urban environment, which made their record suspect. A recent project at the University of California, Berkeley is trawling the Web to find station records and add them into a global dataset containing more than 30,000 stations. The quality of that dataset is open to question, but it is one of these Web-sourcing ideas that we have heard a lot about today in other presentations.

This story offers an example of an extremely open data system that has existed for more than 150 years. One of the reasons the weather data system works so well is that in the early days of weather data reporting, data were freely exchanged, because they had no economic value. We could not do anything with them, because it was difficult to make much use of them in many ways, so the international telegraph system began to carry weather data at no cost. Since sharing these data was advantageous to everyone, they did it freely.

In the 1990s, that started to change. Some countries are charging for weather data products in an effort to recoup the costs of their observing systems. Figure 3-4 shows datasets that the European Center for Medium Range Weather Forecasts is trying to recuperate now, in order to add them into the global climate record and improve the quality of that record. There are many of these datasets.

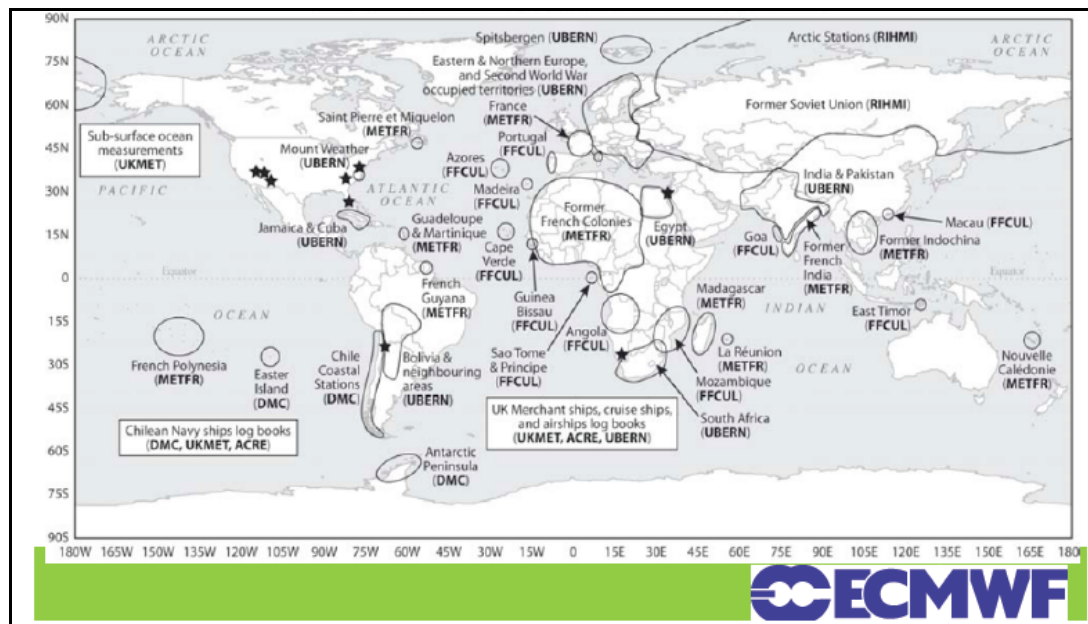


FIGURE 3-4 Focus on pre-1957 meteorological data in sensitive regions.

SOURCE: European Centre for Medium-Range Weather Forecasts

The point of this story is partly that the effort to build this digital climate data record has already been going on for almost 30 years, and it is still under way. This is difficult work, even though the relative size of these datasets is small compared to the size of some of the databases we have now. The results of the global collection effort are good, however. Figure 3-5 presents four independent analyses of the global temperature record. They converge much better—even in the nineteenth century—than the versions created earlier.

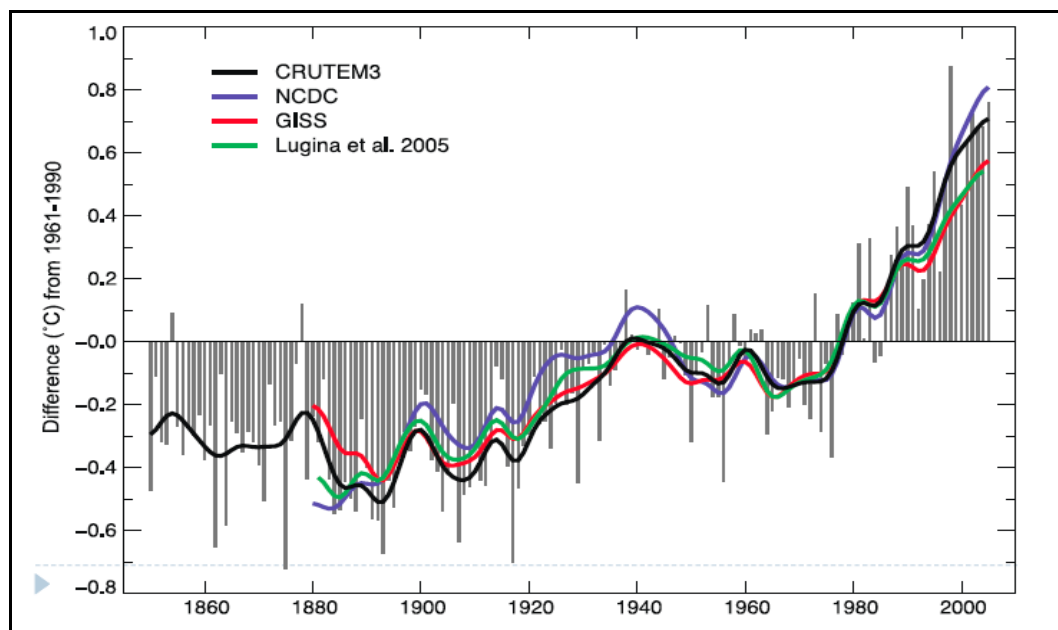


FIGURE 3-5 IPCC AR4 (2007).

SOURCE: Intergovernmental Panel on Climate Change

I want to focus now on data problems. Some of these are institutional and some of them are not. Most of us still work at a local workplace, and almost all of us work for some kind of hierarchical organization that pays our salary. This means that the issues that concern us locally about the institution that we work for will almost always trump concerns about things that are remote. This is true for various psychological and institutional reasons. Studies of face-to-face interaction show, for instance, that it generally works better and faster than interaction at a distance. There is also a trust aspect. The people you know are the people you are more likely to share with, and whose data you are more likely to trust.

The story about meteorology and climate science has some temporal issues that have not come up much in this discussion, but are nonetheless very important. It is easy to think that the data we are trying to federate and collect will be there forever, but of course they will not. The audience may be familiar with the story when Mars data were lost, because the software with which data tapes were made had been lost. Digital information is quite fragile, because the strings of bits cannot be interpreted without the code that created them—and that code is dependent on other code, such as operating systems. Losing data by losing the software and hardware context is the kind of problem that could happen again and that already does happen quite often.

Metadata is another issue that appears in the climate and weather data story. It is not obvious to us what people in the future will need to know about the data we collect now—and

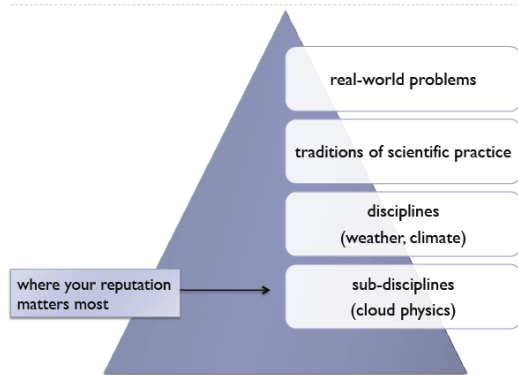
we cannot know that with any certainty. We may not collect what they will need at all, or what we collect may not be sufficient.

Finally, there is a figurative dimension to distance, which is the distance between disciplines. The great promise of open data is that we can work on problems that have never been studied before, because they required collaborations among different fields. That is also a potential failing, however, because if an ecologist wants to work with data from, for example, an atmospheric physicist, the ecologist might be willing to accept that person's data at face value. After all, what does the ecologist know about atmospheric physics? Yet those data may be defective in various ways. We may need to make an extra effort to know more about those data and the problems they may have.

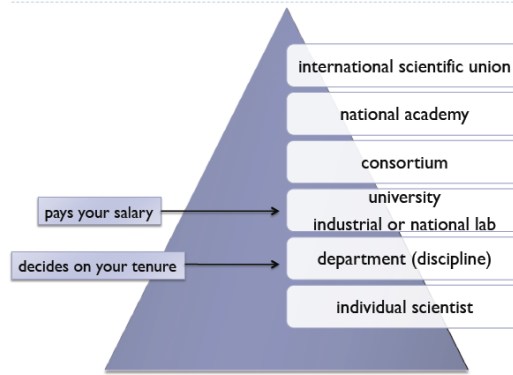
This raises the problem of trust. The further away that some user is from the source of the data, the more possible it is for the members of that user community to find it completely unpersuasive. Hence, sharing data will not automatically facilitate communication and understanding.

The world we think we are in, and want to be in, is now thoroughly networked, with easy sharing of data and information among all scientists and disciplines. But the world that we are actually in is often much more hierarchical and stove-piped. In science, the hierarchy begins with the subdisciplines, rises to the disciplines, and rises again to larger traditions of scientific practice (Figures 3-6 and 3-7). Studying real-world problems, such as climate change, involves many disciplines, but the place where your scientific reputation is made is at the much lower level of your subdiscipline.

Hierarchies and incentives



Hierarchies and incentives



FIGURES 3-6 and 3-7 Hierarchies and incentives.
SOURCE: Paul Edwards

The same observation is true for our institutions. Academic scientists are individual, they work in a department, and the department is part of a university. There might be levels above that all the way up to international professional societies—but the place that pays the salary is the university, and the place where the scientists' reputation is evaluated is their department. Therefore, these relatively local entities can often trump larger collaborations, networks, virtual organizations, and so on.

One of the phenomena we see in the formation of many infrastructures is a phase in which we build gateways—systems that allow us to move information, for example, from one place to another without having to worry about the particular features of every infrastructure involved in the transfer. The meteorological information infrastructure, the World Weather Watch of the U.N. World Meteorological Organization, is a great example. Its bottom level consists of more than 150 national meteorological services. Each one contributes its observational data to the global observing system. The global communication system sends them to central processing centers, where data from the entire planet are processed and then returned to the national weather services.

This is great for understanding and forecasting the weather, but once we start confronting larger problems, such as climate (which involves many Earth systems besides the atmosphere), the World Weather Watch is no longer enough. Therefore, we get another, higher level of institution. For example, the Global Earth Observing System is a collection of systems, and each of those systems is, in fact, a collection of other systems. What I see when I look at the world of scientific institutions is a constant need to climb up a level to try to get an understanding of the general area that is being studied.

I have done some work with people in the Global Organization of Earth System Science Portals (GO-ESSP). These are people who build tools like the Earth System Grid. They are concerned about adequate support and coherence. The problem they face is that their activities for GO-ESSP are, in general, unfunded. They feel strongly that it is necessary, because they want to do things that link the various portals and make them into a more coherent, interoperating system—but nobody is paying them to do that. Thus, again, the institutional level of concern and the one at which they actually want to work are different and conflicting.

My late colleague, Susan Leigh Star, coined this phrase: “Metadata are nobody’s job.” Everybody wants good metadata, but nobody knows who is supposed to create them. We would like to push it off onto the scientists, but they do not want to do it. They are moving on to the next project after they have finished their datasets, and they do not want to go back and document what they have done. Perhaps there is a class of data managers who can handle it, or perhaps it should be the “crowd.” Maybe somehow younger scientists will do it. What we often see in labs is that this work is pushed down onto the graduate students. Maybe social scientists have the magic answer. I am a social scientist, and, as far as I know, we do not.

My research group has been doing some ethnographic work with environmental scientists, and we have observed that there is an obsession with metadata as a *product*. In other words, metadata are conceived as a kind of library catalog that will describe all the data that someone has. The reality of data sharing, however, is quite often that metadata exist mainly as an ephemeral *process*. Many people say that the most important issue when dealing with data collected by someone else is to talk to that person. “I need to know more about what you did, how you collected this data, what it means, what the formats are”—there is an endless list of questions, and it is impossible that all of those will ever be answered by a catalog-style metadata product. One lesson, therefore, is that when someone is conceiving metadata, creating channels for communication among researchers is at least as important as creating the ultimate catalog that we would all like to have.

Finally, I am going to talk about data and software. Data cannot be read without software. An even more important point is that many things that we call data now are actually the product of software. Climate model output is an example, as are instrument data after they have been collected and processed through some sort of data model to be put into the final dataset that is used in a general publication. Which of those things are we talking about? If it is the latter, we need to know exactly how it was processed.

For the last 15 years or so, the journal *Money, Credit, and Banking* has required authors to submit both their code and their data so that their results, in principle, could be replicated. In 2006, researchers found that of 150 articles only 58 actually had some data or a code, and of those, for only 15 could the results be replicated with the code and data the authors provided. The journal editors concluded that there needs to be a stricter system for people who are submitting their code and data. But my question is, how much is it worth? What they did not do was attempt to communicate with the authors to find the missing information.

My last points are related to the problem of the academic reputation system. This is the way the system works now: Scientists have some data sources, they use them to do research, they publish the results, get citations, and that builds their reputations. Somewhere in there are services, and that is where data sharing and software, as well as the writing of scientific software, both sit.

My colleague, James Howison at Carnegie Mellon University, has been thinking about how we might design a reputation system that would work for scientific software writing. Many scientists write little bits of code, or sometimes much more elaborate code. That is part of their job, but it is not their main job. Yet once they have written the software, they get involved in the problem that if their code is going to stay active, they have to maintain it. There is an ecology of software, and as the operating systems and associated software libraries change, this piece of scientific software will eventually stop working. Therefore, the scientist has to keep updating it, to keep modernizing it, if people are going to keep using it. Sometimes this turns into a sideline, and the scientist actually becomes a software developer. More often, however, what happens is that the code eventually just dies, and then there is the problem of replication. How can the research that used it be replicated?

I collect a lot of data, and then what happens to it? Sharing the data requires work from me. Writing out all the metadata is work. One of the best examples we have of a well-functioning collection of large datasets is the collection for CMIP-5, the Coupled Model Intercomparison Project, that my colleague from Unidata spoke about earlier. It is working well because there is a major incentive, which is that if someone wants to have data in the Intergovernmental Panel on Climate Change (IPCC) report, that is the only way to do it. The scientist must fill out the metadata questionnaire. How long does filling out the metadata questionnaire take? Up to 30 days for each model, because there are many runs and the questionnaire has to be filled out separately for each run.

The point is that data sharing is work, and it always will be work. Thus, one of our big issues is to find ways to pay people for that and then also to include it in reputation systems so that they can get credit for it.

There is a conflict between the mandates of institutions like the IPCC, the National Science Foundation, and the National Institutes of Health that want us to publish and share data, but there are few career incentives for sharing data. The career incentives are for the

results that come from the data. That is true of building software. It is also true of sharing data. Moreover, the problems that software developers face are also there for people who want to share data, though perhaps to a lesser degree. Data have to be maintained, which is why the issue of curation is a critical one. Who will maintain the data?

There are many possible solutions, however, and some of them will work for some areas. It is important to realize that there is no one thing called data, and that the differences among scientific fields are enormous. What works in one area may not work in another. Maybe we are talking about a virtual organization, or maybe it is a crowd-sourcing project. Among the people I have been working with, we have often heard Wikipedia, Linux, Mozilla, and Apache held up as models for the future: “This is what we will do. Open-source software works. Open-access models work.” I do not think so. For many fields, there is going to be a problem, because there are not enough people with the right skills.

Metadata standards may help, but someone needs carrots or sticks to get that process to happen. I have heard many people say that the young people will do it. They assume the young people will know how to do it, but if we do not know how to do it, why is it that they will know? They need to be trained. Studies of young people using even simple things like Google Search show that most of them do not know how to use Google Search in an effective and efficient way, and that is equally true of young scientists. Some of them will be very good, but most of them will not, and it might be that we actually need new paradigms of education.

I love the idea of science informatics. Perhaps it will come from an iSchool, but the iSchools we have are not set up for this yet. Perhaps it will come from science departments, but most of the science departments I know are not set up to do that either. Science informatics is not computer science, and most computer science departments are not interested in this problem at all except for their own kinds of models and data. Therefore, it will probably be some combination of computer science, iSchools, and domain science. And the combination may need to be different for different domains and problems.

Discussion

DISCUSSANT: There was a discussion about how difficult it is to get large amounts of text for literature analysis, but people have been trying to do that since the 1970s, long before copyrights became a big issue, and there are large bodies of text that are not available. Do you think the technology works even if someone could get the text, because I do not see that?

DR. LYNCH: Parts of the technology work, and in fact there are some successes to which we can point. The notion that someone has a computer that reads the literature on a certain subject and understands it in a deep way is fairly farfetched. One of the things that Google has suggested, however, is that when someone gets enough data, he or she can do some interesting work through statistical correlations.

DISCUSSANT: There is a general problem about where the credit lies, where the value lies, and who is responsible. But the main question is: What is the business model? Who is responsible among universities for ensuring that there is a cyberinfrastructure and that appropriate credit is given?

DR. LYNCH: I do not think there are simple answers. I do suspect that even in a period where resources are very tight, if we look carefully at most of the cyberinfrastructure investment, it is largely cost effective. This investment should be making the overall research enterprise more effective and providing some real leverage. I think those are the parts that will probably move ahead.

One of the issues in the area of data that we need to be very mindful of is that data preservation per se is almost a pure overhead activity, especially when we get past the first couple of years when people might want to reproduce new results. The activity with the payoff is reuse. That, not preservation, leads to new discovery, so it is important to be focused on facilitating reuse and not just looking for the longest possible list of data that we need to preserve. There are clearly materials we should preserve even if there are not obvious reuses for them right now, but we need to be judicious in that category and compete for the resources.

DISCUSSANT: It used to be that the reason an institution would have a repository was because there were physical objects that made sense to keep locally at an institution. Individuals did not have a chance to use an institutional repository unless they were at an institution.

Consider, for example, a person's personal photography collection. Perhaps this person is in a bizarre hybrid situation, where he or she tries different services and ends up with a few hard drives, where many copies of this material are saved. There should be a different solution, however, where that person could take advantage of a big institution. Long-term trust is the main problem. He or she does not want those photographs to be lost. What if those companies go out of business? That person would still want everything on a personal hard drive.

If we move from digital photographs to scholarly output, the issues are the same. Why is Harvard University, for example, going to have a repository for everybody's information, when it could just have some enterprise solution that gave everything to Microsoft? What is the balance between an institution and the individual?

DR. LYNCH: I am not sure that I parse the landscape the same way you do. I would say that the very critical issue about an institutional repository is that an institution is taking

responsibility for long-term preservation. Institutions and corporations, unlike people, may live forever, or at least indefinitely, and have processes that allow them to operate on much longer timescales than individuals.

DISCUSSANT: If a university is paying money to another institution, is that still an institutional repository?

DR. LYNCH: Yes. The way I view this kind of arrangement is that if the institution identifies the data and takes responsibility for their preservation, then it is making a much longer commitment than an individual could do, perhaps setting up a trust fund or some other mechanism that outlives the individual. Operationally, we may see many universities over time contracting the preservation and curation of specific kinds of data to other universities, or to other for-profit or not-for-profit archives. We may also see the emergence of more consortia, as well as straight-out contracting. The locus of ultimate responsibility is the key thing here.

DISCUSSANT: Does that mean that the universities can act like the individual researcher?

DR. LYNCH: Except that presumably the universities have thought through issues like multisourcing and what happens when a supplier goes broke, how to back up the data, how to audit contractors and move data from one contractor to the next, and other similar issues. The other point I want to make is that we often confuse distribution systems with preservation systems. Flickr is a great example. People put their pictures up on Flickr, and I think of that as a means of distributing them, making them available, and building some social interaction around them. However, there are people who believe this is a preservation venue, and I would urge them not to think that. If you really want your pictures, by all means put them up on Flickr or anywhere else, but keep a set for yourself offsite and make lots of copies.

DR. LESK: In terms of how we get somebody to do the preservation and the data curation, assuming that scientific data is indeed worth having—I wish we had more data on that point—the data are either a public good or a private good. If they are a private good, it means that the owner gets credit for having them, presumably by getting tenure. However, that means waiting for universities to give tenure for collecting data instead of for research results or for the papers interpreting the data. Coming from a university where departments complain that the administration is too slow, I am not optimistic about that.

The other option is that the data are a public good. Dr. Berman made a comment earlier about unfunded mandates. I have the feeling that if the National Science Foundation (NSF) imposed a data management mandate—we are not going to get any more pages in the proposal, but we are going to get 5 percent more money as the upper limit of what we can ask for—I think the world would be completely different. Obviously there would have to be a requirement that the 5 percent be spent on the data curation, but it would completely change the attitude about how the research community does this.

DR. SPENGLER: It is important to point out that the NSF asks for a data management plan. It does not ask a scientist to preserve the data. It does not ask that there always be copies.

Even more important, if we read all the details, we do say that we are willing to have the scientist put requests in the proposed budget for curation purposes. We say that we are willing to support it.

DISCUSSANT: If scientists decide that their data are worth preserving, is there some sort of mechanism for them to appeal that?

DR. SPENGLER: Some of my colleagues and some of the directorates had planned to include in the NSF data management policy that a grantee has to keep the data for 5 years beyond the award period. That got taken out.

DISCUSSANT: Is there a current plan for that?

DR. SPENGLER: No. Some disciplines have mechanisms set up for that, however. For example, if we look at some of the programs within the NSF, we will see that they do not normally say that the NSF wants data management plans. However, researchers will get in trouble if they do not have the data deposited in the designated database when it is time for their annual report or project renewal, because it will not be approved. Hence, some places have thought long term about it, but it requires a very large commitment by a given domain. The social sciences have stepped up in their resources, as has astronomy. Many disciplines and organizations have made more of a commitment, but questions about the long-term cost are still open.

DR. LYNCH: There are some interesting developments at the institutional level as well. For example, earlier this year Princeton University offered a storage service for their researchers that allows them to pay in advance for storage at some rate per terabyte, allegedly for eternity, or as long as Princeton will exist. The university made some projections about the cost of technology, the refresh rate, and the return on capital for the prefunding. We can argue about whether the numbers are right, but it is interesting to see someone step forward with a model like that, and I think we will probably see some more of it.

Policy and Legal Factors

Session Chair: Michael Carroll

-Washington College of Law, American University-

There was a comment made earlier about how researchers think of themselves as hunter-gatherers. This goes to the philosopher John Locke's labor theory: I did it, and therefore I own it, and I have a property right in it.

Lawyers think of property in different ways, but I want to highlight the fact that there is a "there" there, even though the words "intellectual property" are being imprecisely used in this context. The way that an intellectual property lawyer might think about ownership and rights over the data is in terms of the rights over the first copy of the zeroes and ones. In a way, this is almost like tangible property in the sense that when there is only one copy, these zeroes and ones might be owned insofar as I have the medium on which these zeroes and ones sit, and you cannot have it unless I say so. This is a claim of ownership over a form of property.

Intellectual property rights are a layer above the zeroes and ones. They are entirely intangible, and these are rights that tell us what uses are acceptable. May we access the zeroes and ones? May we copy them? May we change the order of the zeroes and ones? These are use regulations.

Think about intellectual property law as access and use regulations that apply to the zeroes and ones rather than being embedded directly in the zeroes and ones. Once we make this conceptual separation, when we think about the policy issues, we are thinking about the regulations over data acquisition, data reuse, and the like.

Legal Aspects

Michael Madison

-University of Pittsburgh School of Law-

I want to highlight some central legal questions in this area. In particular, I am going to focus on intellectual property (IP) law questions. There are some interesting and important privacy law issues as well, but I am not going to discuss those.

Figure 3-8 highlights in a very simplified, stylized way how I organize my thoughts on this topic.

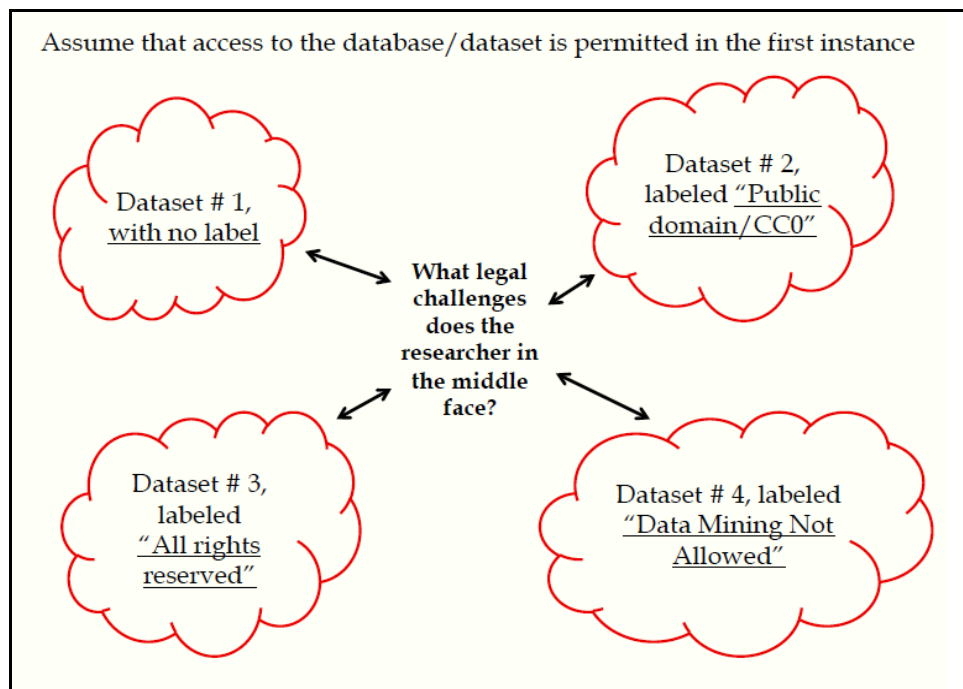


FIGURE 3-8 Intellectual property challenges to data access.
SOURCE: Michael Madison

There are four clouds of different datasets, and a researcher in the middle who is extracting—or recombining or using—data from each of these sources. There is a different kind of a label—or, in one case at least, no label whatsoever—attached to each of those datasets, specifying terms and conditions under which data from that source might be used, recombined, shared, and so forth. The question is, What does the person in the middle do, and how does the legal system frame that set of questions?

In the upper left corner, there is no label whatsoever, just a pot of data. The lower left is “all rights reserved,” or the standard copyright-style notice that would be attached to any kind of copyrighted work and that is often attached to materials that are not actually copyrighted, often out of ignorance or sometimes in an attempt to deliberately claim rights in things that cannot be claimed. It is a default copyright label.

On the lower right side is a context-specific restriction that is phrased “data mining not allowed.” This is to illustrate the idea that people who collect data or create copyrighted works increasingly have the power to custom-design labels or notices that they attach to their works before they send their works out into the world, and they expect both the world to respect that label and the legal system to enforce it somehow.

In the upper right corner is a label of a type that is increasingly common, used to specify or declare that the material inside the pot is in the public domain or, if it was not in the public domain in the first place but in fact was copyrighted to begin with, has now been dedicated to the public domain by some kind of affirmative notice or label. It is the CC0 label, which is a form of Creative Commons license that basically declares that the material covered by the license was copyrighted, but the author of the material, or the owner of the copyrighted material, is affirmatively dedicating that content to the public domain. One of the questions I will address is the extent to which that kind of labeling or notice device is actually effective.

There are three general categories I want to cover, with questions in each. One is the default legal status of data in these environments, both with respect to the underlying data and also with respect to the adjacent information, formats, coding, metadata, and literature that arise from these things. The second one pertains to contracts, licenses, notices, restrictive notices, and enabling notices. The third category is what the legal system has to say about the problem of managing a commons or managing a community of shared resources.

I will begin with some default copyright rules within the intellectual property system—just what the rules are. First, data are treated in copyright law as facts and therefore cannot be copyrighted. They are by design in the public domain. This is one of the few black-letter legal rules that is easy to state. This would include experimental results, observations, sensor readings, and the like. So for people who assert that they own their data, if we were to apply a formal legal model to this, they cannot own the individual items in the dataset. They might, however, be able to claim a copyright in the compilation of the data or the database as a whole.

Regarding collections of data and databases, copyright law gives us this phrase “selection, coordination, or arrangement.” If the selection, coordination, or arrangement of the data, or if the selection, coordination, or arrangement of practically anything, is original—that is, it has been touched by human judgment or creativity in some modest way—then there is at least the spark of a copyright that attaches to that collection. Not to the individual items, but to the collection.

If it is a collection of facts or a dataset that has been compiled in a way that has been organized by human judgment, then copyright may well apply, although the scope of the copyright may be relatively narrow. Copyright comes in different widths. There is narrower copyright for fact-oriented materials and broader copyright for classically entertainment-oriented things such as novels, films, and popular music.

When I say a “narrow” copyright, or what copyright lawyers would call a “thin” copyright, I mean that infringement of that kind of copyright will generally consist of verbatim or wholesale copying of the contents of the work. Copyright law also has this idea of substantially similar copying that can be infringement, but it generally would not apply to a thin copyright in a compilation of facts.

Codes, formats, metadata, data structures, or anything that involves any kind of interpretation or characterization of the data up to and including actual scholarly papers or scholarly output is going to be eligible for copyright protection. The more human judgment, human creativity, or human originality that goes into the production of that material, the stronger the copyright can be. Thus, a scientist's scholarly papers, output, and the literature itself will be copyright-protected in a much more robust way than the underlying datasets. In short, there is a spectrum of copyright protection that starts at a relatively early stage, immediately after the creation or identification of the individual data point.

I will note two other points about default rules. First, if we are talking about a source in the European Union (E.U.), the European Union has the 1996 Database Directive, which operates in Europe in parallel with copyright rights. The Database Directive has some overlap with copyright and some distinct character. The most important feature of the Database Directive is that it prohibits extraction and reuse of a substantial part of the database. Therefore, picking out bits and pieces of an E.U. database may not violate the Database Directive, but if the Directive applies, then using the database wholesale may violate the Database Directive (in the manner that the Directive has been incorporated into national law) as well as local European copyright law, and perhaps other laws.

Copyright law in the United States is subject to fair use. Fair use refers to the use of particular portions of copyrighted works for critical purposes: teaching, research, scholarship, study, and so forth. Fair use in this area is not an especially reliable or useful tool, although it is potentially quite important in several other respects.

One possible application of fair use is to defend the copying of a copyrighted work at an intermediate stage on the way to producing something else. For example, U.S. case law in one instance focuses on intermediate copying for reverse engineering an underlying software product. If someone wants to build an interoperable or complementary software product—the case law generally deals with video games—then it may be permissible to make an intermediate copy of a copyrighted work to extract underlying public domain ideas and then use those public domain ideas to build a complementary work.

A related way in which fair use might help would be related to building and using tools in the database area. If someone wants to be able to connect to an Application Programming Interface to make use of a particular tool, fair use might be a way, using this case law, to exploit that particular possibility.

Second, turning to licenses and contracts, it is important to note that all the default rules, with some exceptions, can be changed by voluntary agreement, by a license or a contract, or a license bundled into a contract. There is some controversy about the extent to which fair use can be negotiated away in a license or a contract. The better view is that fair use is a mandatory legal requirement, but there is some dialogue about that in the world of law practice.

The most important point about contracts and licenses is that there is a distinction between a unilateral notice, or a license as a concept of property law, and a contract, which is something that you would voluntarily assent to. In practice, both in the software world generally and in this space of datasets, there is much ambiguity and confusion about the legal status of particular terms and conditions that come bundled with these collections of data.

A contract is something that typically requires that someone agrees to it, for example, by clicking on the “I Agree” or “I Assent” or “I Acknowledge” button. A license is a property term that grants a person permission to do things and that might require him or her to do some things with the copyrighted material, but a license does not necessarily require that the person agree to something in return. That is the first area that makes licenses tricky.

The second area that makes licenses unclear, besides the fact that a person is not necessarily required to agree to them, is that someone might not necessarily see them. They might come in machine-readable form, but a researcher might not have a human-readable equivalent. Or a researcher might have a human-readable version of a license, but might not have a machine-readable or automated discovery-readable version of the license. Therefore, it might not be apparent enough for a researcher actually to understand that there are some terms and conditions to deal with there.

A third complication is that particular terms and conditions, as illustrated on Figure 3-8, might be different, incompatible, or at least not perfectly aligned from one dataset to another. Thus, if a scientist is doing a large-scale project that entails extracting data or using data from many different sources, that scientist might end up with multiple legal obligations and rights for different chunks of the data. That has been a big problem.

Turning to open-source software, much of the discussion about such software focuses on the General Public License as the standard license for open-source software products. There are many licenses, however, that are open source in their style and their content, and those open-source licenses may not all say precisely the same thing. Therefore, we must be careful not only to read the particular open-source license that we are dealing with but also to be aware that if the source code comes from different sources, with different open-source licenses, we may end up with multiple obligations bundled in a single product.

The next issue about licenses is a bit more optimistic. In the IP world, people tend to think of licenses as a way to prevent people from doing certain things: I will share my content with you, so long as you do not steal it, cheat on it, share it, forward it, and the like. The underlying concept of license-as-permission can be enabling, however. Increasingly over the last decade, there has been innovation in the law in designing licenses that authorize people to do things that you want them to do. Creative Commons licenses are the paradigmatic example of this. Open-source licenses are often used in this way as well. The fact that lawyers have made innovations like Creative Commons and open-source licenses demonstrates the possibility that others could create all kinds of additional new mechanisms for enabling what people can do, alongside ways for restricting what people can do.

Both for the restrictive licenses and enabling licenses, the law is not fully clear regarding which, if any, of this is actually enforceable. The Creative Commons and open-source licenses are put out there and characterized and conceptualized within their respective communities as licenses rather than contracts. The idea behind them is that they are enforceable because the license is attached to the underlying copyrighted object and travels with the underlying copyrighted object. Anybody who encounters the underlying copyrighted object requires access to it along with the accompanying obligation, which is, I think, an interesting and useful concept.

That concept has not genuinely been tested on its merits in the courts, and the problem with it is that there is a dark underside that some people have identified. That is, it is one thing

to attach a happy public domain style or “please use this and share it” kind of style notice, but what is the difference, legally speaking, between attaching a “please go and share this” notice and attaching a “please do not do anything I do not like with this” notice? Suppose I were to share that dataset with you and attach a notice that said you may not do anything: You may not publish your results, you may not criticize my method, and you may not combine or collaborate with anybody else. Legally speaking, it is difficult to distinguish the restrictive notice from the enabling notice, and the legal system has not really come to terms with how to justify one and not the other.

The last point in this area has to do with the character of the enforcement of the license. So long as these licenses are designed by private parties and are attached to copyrighted works or databases by private parties, then enforcement lies with the copyright owner or with the holder of the database or whoever attached the notice. That creates obvious problems in inconsistency and conflicts. One solution has to do with the identity of the enforcer. Having either a government agent or some other third-party neutral in the middle, as custodian of the legal rights for purposes of monitoring and enforcement, is something to consider when designing an enforceable scheme that would accomplish the actual goals of the enterprise.

I am now going to move to the last section, which concerns management of data collected from a variety of sources. Assume that a researcher has identified and overcome various licensing hurdles and IP-rights hurdles associated with the underlying data. The researcher has a project that has data from various sources. Management of this database is often understood primarily by policy choices, but there is also an important role here for law. My research at the University of Pittsburgh focuses on commons communities, with commons understood as a structured or designed pool of resources or a sharing environment. Here is the question: If we have a choice between maintaining public domain status of the underlying content or understanding the mechanics of a particular license or contractual regime, are there potential upsides to a license or contractual regime that might make that more attractive than a public domain alternative?

Consider the questions in Box 3-1:

Managing the results of data collection

Forward-looking issues related to governing a **data commons**:

1. Data/dataset **integrity**.
2. Translation and interoperability of data from different sources, in different formats: designing and enforcing **standards**; managing and maintaining data **consistency**. (**Ensuring PD status of data may be inadequate to deal with this challenge.**)
3. **Who has access** to the new collection of data, and for what purposes?
4. What are participants' **duties and rights** regarding standardization and data consistency, and re-sharing, re-combining, re-using data?
5. How is compliance **monitored and enforced**?
6. When do those duties and rights pass to **downstream parties** who did not obtain the data in the first place?
7. Compare the costs and benefits of **government-sponsored enforcement** with those of private enforcement **via licenses and contracts**, and with those of **informal/community enforcement**.

BOX 3-1

SOURCE: Michael Madison

Box 3-1 identifies governance topics that relate to a community of researchers or to a particular researcher, using data from various resources that might be bundled in a license instrument and made enforceable as part of that license. It is important to identify both the “what”—that is, where things are coming from in terms of the underlying data—and the “who.” We want to decide who can participate in this particular enterprise and which researchers are eligible. We also want to know what their obligations will be, what their duties will be relative to each other and relative to the data, and also what the enforcement mechanism will be. Who has authority to enforce breaches of the underlying protocols? What are the carrots and sticks associated with ensuring compliance, monitoring performance, and making sure that the data is preserved in its integrity over time and combined in a way that is suitable to the underlying license obligations?

This last point is particularly important, and the license instrument can be especially useful for managing downstream obligations. When the questions are how these duties will pass to the next generation of researchers and who you may or may not know about in the beginning, then having a well-designed license instrument might be able to help you specify who those people are and what their obligations will be. That is the mechanics of an open-source software license, but the concept can be ported into this area as well.

In closing, I will quickly respond to the claim that the legal system might provide clarity and a deterministic foundation on which to build policy analysis or other analysis. That is not the case. There is no one-size-fits-all solution that the legal system can offer to these problems. What the Creative Commons example and the open-source software community have shown with its licensing experience is that the set of tools available to manage these collections is much more diverse and complex than it was 10 or 15 years ago.

Knowledge Discovery in Open Networked Environments: Some Policy Issues

Gregory A. Jackson
-EDUCAUSE-

The Coleman Report, released in 1966, was one of the first big studies of American elementary and secondary education, and especially of equality therein. Some years after the initial report, I was working in a group that was reanalyzing data from that report. We had one of what we were told were two remaining copies of a dataset that had been collected as part of that study but had never been used. It was a 12-inch reel of half-inch magnetic tape containing the Coleman “Principals” data, which derived from a survey of principals in high schools and elementary schools.

The first challenge was to decipher the tape itself, which meant trying every possible combination of labeling protocol, track count, and parity bit until one of them yielded data rather than gibberish. Once we did that—the tape turned out to be seven track, even parity, unlabeled—the next challenge was to make sure the codebooks (which gave the data layout—the schema, in modern terms—but not the original questions or choice sets) matched what was on the tape. By the time we did all that, we had decided that the Principals data were not that relevant to our work, and so we put the analysis and the tape aside, and eventually the tape was thrown away.

Unfortunately, apparently what I had had was actually the *last* remaining copy of the Principals data. The other so-called original copy had been discarded on the assumption that our research group would keep the tape. That illustrates what can happen with the LOCKSS strategy (lots of copies keeps stuff safe) for data preservation: If everybody thinks somebody else is keeping a copy, then LOCKSS does not work very well.

Many of us who work in the social sciences, particularly in the policy area, never gather our own data. The research we do is almost always based on data that were collected by someone else, but typically not analyzed by them. This notion that data collectors should be separate from data analysts is actually very well established and routine in my field of work.

The Coleman work came early in my doctoral studies. Most of my work on research projects back then (at the Huron Institute, the Center for the Study of Public Policy, and Harvard’s Administration, Planning, and Social Policy programs in education) involved secondary analysis. Later, when it came time to do my own research, I used a large secondary dataset from the National Longitudinal Study of the High School Class of 1972 (NLS72). This study went on for years and is now a huge longitudinal array of data.

My research question, based on the first NLS72 follow-up, was whether financial aid made any difference in students’ decisions whether to enter college. The answer is yes, but the more complex question is whether that effect is big enough to matter. NLS72 taught me how important the relationship was between those who gather data and those who use it, and so it is good to be here today to reflect on what I have learned since then.

My current employer, EDUCAUSE, is an association of most of the higher-education institutions in the United States and some elsewhere. Among other things, we collect data from our members on different questions: How much do you spend on personal computers? How many helpdesk staff do you have? To whom does the chief information officer report? We gather all of these data, and then our members and many other people use the Core Data Service for all sorts of purposes (Figure 3-9).

Welcome to the EDUCAUSE Core Data Service

EDUCAUSE offers the Core Data Service to our members to address your ongoing need for comparison data about campus information technology environments and practices. You can use the service to help benchmark, plan for, and make decisions about IT on your campus.

*Coming Soon —
The New Annual Survey!*

- Read about the Core Data Service (CDS) Redesign on Dan Updegrove's [blog](#).
- [Subscribe](#) to our e-mail list to receive CDS announcements.




FIGURE 3-9 EDUCAUSE Core Data Service.
SOURCE: Gregory A. Jackson

One of the issues that struck us over time is that we get very few questions from people about what a data item actually means. Users apparently make their own assumptions about what each item of data means, and so although they are all producing research based ostensibly on the same data from the same source, because of this interpretation problem they sometimes get very different results, even if they proceed identically from a statistical point of view.

If issues like this go unexamined, then research based on secondary, “discovered” sources can be very misleading. It is critical, in doing such analysis, to be clear about some important attributes of data that are discovered rather than collected directly. I want to touch quickly on five attributes of discovered data that warrant attention: quality, location, format, access, and support.

Quality

The classic quality problem for secondary analysis is that people use data for a given purpose without understanding that the data collection may have been inappropriate for that purpose. There are two general issues here. One has to do with very traditional measures of data quality: whether the questions were valid and reliable, what the methodology was, and other such attributes. Since that dimension of quality is well understood, I will not discuss it further.

The other is something most people do not usually consider a quality issue, but any archivist would say it is absolutely critical: We have to think about where data came from and why they were gathered—their *provenance*—because why people do things makes a difference in how they do them, and how people do things makes a difference in whether data are reusable.

We hear arguments that this is not true in the hard sciences and is completely true in the social sciences, but the reverse is equally true much of the time. The question of why someone gathered data, therefore, is very important.

One key element of provenance I call *primacy*, which is whether we are getting data from the people who gathered them or there have been intermediaries along the way. People often do not consider that. They say, “I’ve found some relevant data,” and that is the end of it.

I was once assigned, as part of a huge Huron Institute review of federal programs for young children, to determine what we knew about latchkey kids. These are children who come home after school when their parents are not home, and let themselves in with a key (in the popular imagery of the time, with the key kept around their necks on a string). The question was, How many latchkey kids are there?

This was pre-Google, so I did a lot of library research and discovered that there were many studies attempting to answer this question. Curiously, though, all of them estimated about the same number of latchkey kids. I was intrigued by that, because I had done enough data work by then to find such consistency improbable.

I looked more deeply into the studies, determining where each researcher had gotten his or her data. The result was that every one of these studies traced to a single study, and that single study had been done in one town by someone who was arguing for a particular public policy and therefore was interested in showing that the number was relatively high. The original purpose of the data had been lost as they were reused by other researchers, and by the time I reviewed the finding, people thought that the latchkey-kid phenomenon was well and robustly understood based on multiple studies.

The same thing can happen with data mining. We can see multiple studies and think everyone has got separate data, but in reality everyone is using the same data. Therefore, provenance and primacy become very important issues.

Location

In many cases communicating data becomes a financial issue: How do I get data from there to here? If the amount of data is small, the problem can be solved without trade-offs. However, for enormous collections of data—x-ray data from a satellite, for example, or even financial transaction data from supermarkets—how data get from there to here and where and how they are stored become policy issues, because sometimes the only way to get data from source to user and to store them are by summarizing, filtering, or otherwise cleaning or compressing the data. Large datasets gathered elsewhere frequently are subject to such preprocessing, especially when they involve images or substantial detection noise, and this is important for the secondary analyst to know.

Constraints on data located elsewhere arise too. There may be copyright constraints: Someone can use the data, but cannot keep them, or can use them and keep them, but cannot publish anything unless the data collector gets to see—or, worse, approve—the results. All of these things typically have to do with where the data are located, because the conditions accompany the data from where they were. Unlike the original data collector, the secondary analyst has no ability to change the conditions.

Format

There are fewer and fewer libraries that actually have working lantern slide projectors such as the one shown in Figure 3-10. Yet there are many lantern slides that are the only records of certain events or images. In most cases, nobody has the funds to digitize or otherwise preserve those slides. As is the case for lantern slides, whether data can be used depends on their format, and so format affects what data are available. There are three separate kinds of issues related to format: *robustness*, *degradation*, and *description*.



FIGURE 3-10 Projector for obsolete format photographic slides.
SOURCE: Gregory A. Jackson

Robustness has to do with the chain of custody, and especially with accidental or intentional changes. Part of the reason the seven-track Coleman had even parity was so that I could check each 6 bits of data to make sure the ones and zeroes added up to an even number. If they did not, something had happened to those 6 bits in transit, and the associated byte of data could not be trusted. Hence, one question about secondary data is: Is the format robust? Does it resist change, or at least indicate it? That is, does data format stand up to the vagaries of time and of technology?

Degradation is about losing data over time, which happens to all datasets regardless of format. Error-correction mechanisms can sometimes restore data that have been lost, especially if multiple copies of the data exist, but none of those mechanisms is perfect. It is important to know how data might have degraded, and especially what measures have been employed to combat or reverse degradation.

Finally, most data are useless without a *description* of the items in this dataset: not just how the data are recorded on the medium or the database schema, but also how they were

measured, what the different values are, what was done when there were missing data, whether any data correction was done, and so on.

Therefore, as a matter of policy, the “codebook” becomes an important piece of format. Sometimes codebooks come with data, but quite often we get to the codebook by a path that is different from the one leading to the data, or we have to infer what the codebook was by reading someone’s prior research. Both are dangerous. That is what we had to do with the Coleman Principals data, for example, because all we had were summary tables. We had to deduce what the questions were and which questions had which values. It is probably just as well we never used the data for analysis.

Access

Two policy issues arise. The first access issue is *promotion*: Researchers trying to market their data. The risks in that should be obvious, and it is important that secondary analysts seek out the best data sources rather than vice versa.

As an example, as I was preparing this presentation, a furor erupted over a public-radio fundraiser who had been recorded apparently offering to tilt coverage in return for sponsorship. That is not the data issue—rather, it was the flood of “experts” offering themselves as commentators as the media frenzy erupted. Experts were available, apparently and ironically, to document almost any perspective on the relationship between media funding and reporting. The experts that the *Chronicle of Philanthropy* quoted were very different from the ones *Forbes* quoted.

The second issue is *restriction*. Some data are sensitive, and people may not want them to be seen. There are regulations and standard practices to handle this, but sometimes people go further and attempt to censor specific data values rather than restrict access. The most frequent problem is the desire on the part of data collectors to control analysis or publications based on their data.

Most cases, of course, lie somewhere between promotion and censorship. The key policy point is, all data flow through a process in which there may be some degree of promotion or censorship, and secondary analysts ignore that at their peril.

Support

Support has become a big issue for institutions. Suppose a researcher on Campus A uses data that originated with a researcher at Campus B. A whole set of issues arises. Some of them are technical issues. Some of them are coding issues. Many of these I have already mentioned under Location and Format, above.

A wonderful *New Yorker* cartoon¹³ captures the issue perfectly: A car is driving in circles in front of a garage, and one mechanic says to another, “At what point does this become our problem?”

Whatever the issues are, whom does the researcher at Campus A approach for help? For substantive questions, the answer is often doctoral students at A, but a better answer might come from the researcher at B. For technical things, A’s central Information Technology organization might be the better source of help, but some technical questions can only be solved with guidance from the originator at B. Is support for secondary analysis an institutional role, or the researcher’s responsibility? That is, do all costs of research flow to the principal investigator, or are they part of central research infrastructure? In either case, does the responsibility for support lie with the originator—B—or with the secondary researcher? These questions often get answered financially rather than substantively, to the detriment of data quality.

When data collection carries a requirement that access to the data be preserved for some period beyond the research project, a second support question arises. I spoke with someone earlier at this meeting, for example, about the problem of faculty moving from institution to institution. Suppose that a faculty member comes to Campus C and gets a National Science Foundation (NSF) grant. The grant is to the institution. The researcher gathers some data, does his or her own analysis, publishes, and becomes famous. Fame has its rewards: Campus D makes an offer too good to refuse, off the faculty member goes, and now Campus C is responsible for providing access to the data to another researcher, at Campus E. The original principal investigator is gone, and NSF probably has redirected the funds to Campus D, so C is now paying the costs of serving E’s researcher out of C’s own funds. There is no good answer to this issue, and most of the regulations that cause the problem pretend it does not exist.

A Caution about Cautions

Let me conclude by citing a favorite Frazz comic, which I do not have permission to reproduce here. Frazz is a great strip, a worthy successor to the legendary Calvin and Hobbes. Frazz is a renaissance man, an avid runner who works as a school janitor. In one strip, he starts reading directions: “Do not use in the bathtub.” Caulfield (a student, Frazz’s protégé, quite possibly Calvin grown up a bit) reads on: “Nor while operating a motor vehicle.” They continue reading: “And not to be used near a fire extinguisher, not recommended for unlogging plumbing, and you do not stick it in your ear and turn it on.”

Finally, Caulfield says, “Okay, I think we are good,” and he puts on his helmet. Then the principal, watching the kid, who is wearing skates and about to try rocketing himself down an iced-over sidewalk by pointing a leaf blower backwards and turning it on, says, “This

¹³Available at <http://gregj.us/pFj6EN>.

cannot be a good idea.” To which Frazz replies, “When the warnings are that comprehensive, the implication is that they are complete.”

If there is a warning about policy advice, it is that the list I just gave cannot possibly be complete. The kinds of things we have talked about today require constant thought!

Discussion

DISCUSSANT: Is there any foreign country where you think the law works well? There are a variety of regimes around the world, ranging from Afghanistan, which has no intellectual property laws at all, up to us. Is there some place where you think it works well?

MR. CARROLL: When the United States invaded Iraq, shortly after we occupied that country we sent Hilary Rosen, the former head of the recording industry, to Iraq to help them write their copyright law. Generally speaking, the United States has been fairly aggressive over the past 12 or so years in negotiating bilateral trade agreements with a host of modest countries around the world that are explicitly designed to export U.S.-style intellectual property rights, including most all of copyright.

DISCUSSANT: At the moment, the countries that are not part of the Berne Convention include Afghanistan and Somalia and other places, such as East Timor.

MR. CARROLL: The U.S. weather data policies were among the first examples of an open approach to public information. Many of these countries, particularly in Europe, had the cost recovery model, but now Europe is starting to see that open data might be important to promote the innovation that can be built on open platforms and that the cost recovery model for sharing weather data might be inhibiting the pace of innovation. Therefore, one of the arguments in favor of open—rather than closed—systems is that with an open system we get innovation that cannot otherwise be predicted.

DISCUSSANT: I want to make sure that I do not come across as being overly negative or critical of U.S.-style intellectual property rights. One of the great strengths of the U.S.-style copyright system is that it has supported a very lively innovation space over the last decade or two. There is a fair amount of potential for using the existing copyright levers—including the public domain lever, enabling licensing, and so forth, all of which exist as part of the default landscape of American intellectual property law—to build legal institutions that enable the kind of collaboration that we have been talking about today.

MR. UHLIR: I would like to elaborate on the E.U. Database Directive. It is a block to the kinds of automated knowledge discovery and open networking environments that we are discussing here. One of the negative aspects is that an insubstantial part of the database gives rise to the right. So if we take it either qualitatively or quantitatively, as defined by the owner of the database, this leaves the user completely unaware of how much extraction of the data is really a cause of action or an infringement. Second, there is no economic harm needed to trigger the statute. Third, it applies to government data as well as to private-sector data. Fourth, it is potentially a perpetual right, because it is triggered for another 15 years every time the database is updated substantially. Fifth, it is an unprecedented exclusive property right in that it protects investment rather than intellectual creativity, which is a fundamental aspect of copyright law. And sixth, there are no mandatory exclusions or limitations. For example, even though this was supposed to harmonize the law in the European Union, there are at least three E. U. countries that have no exclusions whatsoever, no exceptions for scholarly use or education or any other purpose. This applies to factual information rather than creative content. It is thus highly pernicious in the sense that the controls over factual information and the reintegration of information absolutely forbid the type of reuse of information that scientists typically make.

MR. MADISON: In reaction to that, I want to note that none of this legal landscape is static. It is all moving at different rates of speed. The E.U. Database Directive was a major panic button in the intellectual property policy space for a long time. As Mr. Carroll said, to a certain degree it seems that the panic mode seems to have receded, although on the U.S. side of the Atlantic there seems to be renewed interest in the misappropriation doctrine relative to information from databases as a way to avoid some of the limitations that he described. There needs to be ongoing vigilance.

The other consideration in this area is that, as is often the case, the best defense is a good offense when trying to mitigate the harmful effects of some barriers like this. One way to offset some of the attractiveness of either the E.U.-style regulation or renewed interest in tort law, appropriations-style activity would be for universities, consortia, funders, and others to work proactively to use some of these legal levers, particularly in the area of commons, to build these commons proactively and then to have a proof of concept to show how they can actually succeed in specific settings. There are opportunities to build some of these legal questions into the design of the research enterprises from the beginning. Suppose, for example, that we take the National Science Foundation (NSF) data management plan and leverage off of that, from a legal standpoint, to see if that cannot be made to succeed in some other settings as a proof of concept.

DISCUSSANT: Could you address the problems that face chief information officers (CIOs) of universities, given that there are often residual complaints on these problems?

DR. JACKSON: CIOs are the people typically in charge of the boring part of the use of information technology on a campus. They make sure the training is run, for example. Scientific data discovery poses three major challenges for CIOs.

The first is, particularly for high-volume or time-sensitive scientific data, the question of whether the network will be there to move data in the right way at the right time and whether the storage is there to receive it. There is a set of infrastructure issues, all of which are episodic, meaning that most of the time the datasets sit there unused, then occasionally something happens that results in a peak use that demands all the capacity, and then back to the state of being relatively unused. This is a characteristic of most high-performance networks and computing. It is typically not true for storage. It is a very difficult argument for a CIO to make that the university should invest a lot of money in infrastructure that, on average, is not used. The argument that somebody needs it occasionally is a very difficult argument to make.

The second issue that CIOs face is the support question. This happens, for example, when a faculty member moves on, or a postdoc gets a faculty appointment elsewhere, or a doctoral student finishes a degree. Suddenly someone is looking around and saying, “Who will I rely on for support?” Of course, they will mention the overhead rate and say that the institution should provide the support.

The third issue, which is growing very fast, is the question of how we store data that the institution is required to hold and make available. This is an enormous technical challenge, not because it is hard technically, but because the technology is expensive and should be shared. The institutions are not good at sharing these data or putting them into the Google cloud, however. It sounds attractive, but it is hard to write the right contract. This means that someone at the university must take care of these problems, make sure that the data do not degrade, that error correction gets done, and so on.

One of my favorite questions to ask is, You have a body of data that you need to put in one place and have it available 10 years from now—what mechanism do you choose to do that? There is one good answer to that question, which is that you give it to a relatively expensive service to do it. And every so often—typically a matter of months, maybe every year—this service, which has two copies of the data, is going to read the two copies, compare them to one another, and do a certain amount of error correction and restoration. That is the only way to be sure that the data will be there 10 years from now.

Putting the data on CDs or DVDs does not work. It actually does work if you can manage to print it out on archival quality paper and then copy it, but a lot of data cannot be rendered that way.

MR. CARROLL: Think about the law as providing the default terms of use that apply to data. It was pointed out that those default terms of use are fairly uncertain for data, because it can be difficult to tell exactly where the line is between what is copyrighted and what is not, especially for certain types of data, such as image data.

If we do not do anything, we are still making a policy choice, we are choosing the default terms. Let the uncertainty be the rule.

To defend the NSF's unfunded data management mandate, I think part of what NSF is saying is that researchers have to confront what the policy choices are, what the data management choices are. The researchers can no longer leave those unstated or underdetermined. This is a good idea, as we think about a network environment where people have to work together and we think about longitudinal issues, the demand for the intellectual resources, and the institutional resources, that we should think about such things as what happens to the data tomorrow or 5 or 10 years from now and who has rights to that data. These are the kinds of decisions that are worth making now, because as the flood of data continues, these questions will only become more pressing.

When we lawyers are doing our job right, we are problem solvers. And when we are doing our job really right, we are identifying problems before they have emerged and become serious, and are helping develop a solution. All of these data management issues are only going to become more pressing, and if we leave all these areas undetermined and not thought about, we will end up in a much worse situation in the future. We will find ourselves dependent on somebody else's data, and suddenly they have some ability to claim rights and hold us up, because we did not think about the use and access issues in advance.

Therefore, one important point from this discussion is that these are issues that deserve attention. Attention is expensive, but if we do not attend to it now, it will only be more expensive later.

3. How Can We Tell? What Needs to Be Known and Studied to Improve Potential for Success?

Session Chair: Francine Berman
-Rensselaer Polytechnic Institute-

Introduction

Francine Berman
-Rensselaer Polytechnic Institute-

To set the stage for this session, I want to start by talking about how research itself is evolving. Modern researchers approach their work in multiple ways. There were two approaches that were mentioned in our discussions earlier.

One approach is analogous to looking for a needle in a haystack. The needle is a well-defined question, for example, Does $P=NP$?, which is a classic question in computer science. We know what to do to see if our proposed solution addresses the question, and often the question itself will suggest various approaches to a solution.

Another approach is analogous to identifying where the haystacks are. We may have huge amounts of data and be seeking patterns in the data that are likely to provide useful insights and information. For example, the original Seti@Home application combed through massive amounts of data to identify patterns that may indicate extraterrestrial intelligence of some sort.

To support modern research, today's digital data are being analyzed, used in models and simulations, used to create additional derived data (e.g., visualizations and movies from computer simulations), distributed among a wide variety of participants for filtering, and so on. There is almost no area of modern research that is not being transformed by the availability and ubiquity of digital data.

Another emerging differentiator between today's modes of research and the past is the broad spectrum of individuals involved. In traditional settings, research was done by professional experts—professors, individuals whose profession is research and development, and other experts. Research was their “day job,” and they spent a substantial amount of time doing it. Today we are increasingly seeing some research being done by the “crowd,” many individuals spending varying amounts of time contributing, who are not necessarily professional experts. The millions who have contributed to Seti@Home or Galaxy Zoo have created new information in aggregate. The number of real discoveries and innovations coming from these crowd applications is increasing, and crowd discovery or citizen science is emerging as an innovative new model for research. This also means that the infrastructure for data-enabled research must support these new crowd applications.

During the discussion today, we heard about enabling digital research infrastructure, and the importance of interoperability between its systems and components. We heard about semantic webs and the need for diverse relationships among the data. We focused on the need for human computer interfaces, portals, ways of accessing the data, and ways of using the data. We discussed the different characteristics of the data, about privacy and the sharing of data. We also heard about gaps—gaps between academia and industry, and gaps between earlier generations and the millennial generation who are using data in a very different way.

All of these issues are critical to consider in developing a viable infrastructure to support data-enabled research, and they will inform what we will discuss in this session. Our panelists will examine what works, how we might assess the effectiveness and success of our

infrastructure, and what we are doing. We will also discuss what needs to be done now, and what needs to be studied so that we can continue to accelerate research-driven discovery and innovation in the future.

An Academic Perspective

Victoria Stodden
-Columbia University-

My work focuses on the intersection of computational science and legal and policy issues.

Aspects related to scientific communication are critical to this discussion. As scientists, one of our first duties is to be deeply skeptical. Our first assumption when we see other people's scientific research or the presentation of a scientific fact is that it is not correct. Then we think more and say, "Well, maybe this is right," or we have managed to address the errors in this area or in that area. We move the science forward in that way, and it is a core defining principle of how we should address these issues as technology changes our scientific communication, our modes of research, and the different modalities that we use. At the core, it is about getting at error control.

While it is true that the data deluge is hitting various places and giving us new questions and new solutions, this is not the only way science is being affected. It is also true that fields themselves are becoming transformed.

In my field of statistics, the *Journal of the American Statistical Association* is one of the flagship publications. I took a look at the proportion of computational articles in the journal beginning in 1996. That year a little less than half of the articles, 9 of 20, were computational, and none of them talked about where to get the codes so that I could actually verify those results.

Ten years later, in 2006, 33 of 35 articles published in that journal were computational. Nine percent of those gave me a link to get the code or pointed me to a package. By 2009, all of the articles were computational. Of those, only 16 percent tell me how I can actually get into the code. Therefore, if someone is publishing in this journal as a statistician, he or she is almost certainly publishing computational work. When computational research is done, however, what is communicated in the paper is typically not enough to allow us to understand how the results were generated. In almost all cases, we need to get into the code and the data.

Without access to the code or the data, we are engendering a credibility crisis in computational science. We have enormous amounts of data being collected. What is typical is that the intellectual and scientific contributions are encoded in the software, but the software is not shared. Such things as data filtering, how the data were prepared for analysis, and how the analysis was done, which can all be very deep intellectual contributions, are in the software and are not easily captured in the paper.

Both the data and the code are important. Much of the discussion here is about the data, although code did sometimes come up. It is equally important that they both be shared and be open, which they are not. This is at the root of our credibility crisis in computational science.

How should we proceed? As a scientist, the first thought is to fall back on the scientific method, and how the problems of openness and communication of discoveries have both been

addressed in the other two branches of the scientific method, the deductive and the empirical branches.

In the deductive branch in mathematics and logic, people have worked out what it means to have a proof, to really communicate the thinking behind the conclusions that are being published. Similarly, in the empirical sciences, there are standards that have been developed, such as controlled experiments and the machinery of hypothesis testing, and how they are communicated. In a methods section, there is a very clear way that these results are to be written for publication, designed so that other people can reproduce the thinking and the results themselves.

In computational science, we are now struggling with this issue: how to communicate the innovations that are happening in computational science in such a way that they will meet the standards that have been established in the deductive and empirical branches of science.

My approach is to understand these issues in terms of the reproducibility of computational science results, and this gives me the imperative to share the code and the data. We have seen many interesting examples of how reuse can be facilitated and what happens when someone actually shares open data. This gives rise to a host of issues about ontologies, standards, and metadata. This is framed within the context of reproducibility. The reason that we are putting the data and the code online is to make sure that the results can be verified and reproduced.

Here is an example. In 2007, a series of clinical trials were started at Duke University that have since been terminated, but it took a few years to terminate them. They were based on computational science results in personalized medicine that had been published in prestigious journals, such as *Nature Medicine*. Researchers at the M.D. Anderson Cancer Center tried to replicate the computational work that had gone into the underlying science, and uncovered serious flaws undetected by peer review. The study was plagued with a myriad of issues, such as flipped drug labels in the dataset and errors of alignment between observations in treatment and control groups—errors that are simple to make. The clinical trials were canceled in late 2010, after patients had been assigned into treatment and control groups and had been given drugs. One of the principal investigators resigned from Duke at the end 2010. The point is that we have to assume that errors are everywhere in science, and our focus should be on how we address and root out those errors.

There was a discussion earlier of how the data deluge is a larger manifestation of issues that have been seen before. Also, Dr. Hey gave the example of Brahe and Kepler and how what must have been a data deluge in their context ended up engendering significant scientific discoveries. In that sense, there is nothing fundamentally new here. We are doing the same thing in a methodological sense as we have always done, but we are doing it on a much larger scale. The scope of the questions that we are addressing has changed. In that sense, the nature of the research has changed.

Dr. Hey told us that we need more skills to address this concern. Dr. Friend then said we need verifiability, a point that I have also attempted to make in this talk. This means that the infrastructure and incentives need to adapt to the research reality even though the process of science is not changing in any fundamental way. That in turn means that it will be important to develop tools for reproducibility and collaboration. For example, some presenters also talked about provenance- and workflow-tracking tools and openness in the publication of discoveries.

In short, there are many different efforts that are needed. The solutions in this area will not be something that comes down as an executive order and then all scientists are suddenly open with their data and code. The problems are much too granular, and so the solutions must emerge from within the communities and within the different research and subject areas.

Vis Trails is a scientific workflow management system. It was developed by a team from the University of Utah that is now moving to New York University. It tracks the order in which scientists call their functions and the parameter settings that they have when generating the results. These workflows can then be shared as part of the publication, and they can be regenerated as necessary. Vis Trails also promotes collaboration.

Some recent work by David Donoho and Matan Gavish was presented for the first time in a symposium on reproducibility and computational science, held at the American Association for the Advancement of Science. They have developed a system for automatically storing results in the cloud in a verifiable way as they are generated, and creating an identifier that is associated with each of the published results. For example, if a paper contains a figure, then we would be able to click on it and see how it was done and reproduce the results, as the means to do this are automatically in the cloud.

Another useful tool is colwiz, a name derived from “collective wisdom.” Its purpose is to manage the research process for scientists.

One of the major problems with reproducibility is that, unless a scientist is using these specialized tools, there is no automatic mechanism for researchers to save their steps as they advance. After they have finished an experiment and written the paper, they may find that going back and reproducing the experiment is even more painful than going through it the first time. Thus, tools like colwiz could help both with communicating scientific discovery and with reproducibility.

These issues are all related to the production of scientific data and results. There are also some aspects related to publication and peer review. It is a lot of work to request reviewers, who are already overworked, to review code or data and incorporate them into the review process. Maybe we will get there one day for computational work, but certainly not now.

The journals *Molecular Systems Biology* and the *European Molecular Biology Organization* are publishing the correspondence between the reviewer and the authors, anonymously but openly, along with the actual published results. This is one approach that is being tried to be more open and transparent.

One of the reasons for this practice is that there is a great inequality in the power of the reviewers and the authors. In particular, reviewers can ask for additional experiments and more exploration of data from the person who is trying to get the paper published. Particularly for prestigious journals, reviewers have a lot of power. These journals are trying to balance this power by allowing readers to see the dialogue that took place between editors and authors before a publication.

Furthermore, many journals now have supplemental materials sections in which datasets and code can be housed and made available for experimentation. The sections are not reviewed and so far have had varying amounts of success.

In the February 11, 2011, issue of *Science*, there was an editorial emphasizing that, in addition to data, *Science* is now requiring that code be available for published computational results. That is extremely forward thinking on the part of *Science*. *Science* is folding this into its existing policy, which is that if someone contacts an author after publication and asks for the data—and now the code—the author must make it available.

An approach that other journals have taken is to employ an associate editor for reproducibility. The *Journal of Biostatistics* and the *Biometrical Journal* do that. The associate editor for reproducibility will regenerate and verify the results, and if the editor can produce the same results that are in the manuscript, the journal will Kitemark the published article with a “D” for data or “C” for code. In this case the journal can advertise that readers can have confidence in the results, because they have been independently verified.

There are also new journals that are trying to address the lack of credit authors get for releasing and putting effort into code or data or for attaching metadata. They are trying to address the issue of incentives, and their focus is on open code and open data. *Open Research Computation* offers data notes, for example, and *BioMed Central* has research notes.

PubMed Central and open access are older concepts embedded within the infrastructure of the National Institutes of Health (NIH). But why does it stop with NIH? Could we have a Pub Central for all areas and allow people to deposit their publications when they publish, similar to the NIH policy?

There has been much discussion about the peer-review data management plan at the National Science Foundation (NSF). This is a very important step even though it has been called an unfunded mandate. It is an important experiment in that it creates the possibility of gathering information about how much it will cost and how data should be managed. It is, in a way, a survey of researchers on how they are conceiving of these issues. Maybe the costs are less than NSF worries about, or maybe they are more, but at least we will be able to get a sense of this.

One report that I was involved with along with John King was for the NSF Office of Cyberinfrastructure on virtual communities. We advocated reproducibility as part of the way forward for the collaborative, very high-powered computing problems that we are addressing.

As part of the fallout from the problem I mentioned with the Duke University clinical trials, the Institute of Medicine convened a committee to review omics-based tests for predicting patient outcomes in clinical trials. “Omics” refers to genomics, proteomics, and so on. The committee is chaired by Gil Omenn, and part of its mandate concerns issues of replicability and repeatability and how the articles published in *Nature Medicine* that led to the clinical trials could have gotten through with what were, in hindsight, such glaring errors.

There seems to be a hesitation on the part of some funding agencies to fund the software development or infrastructure necessary to address reproducibility and many of the other issues that we have discussed so far. Let me give an example of an e-mail that was sent to a group e-mail. The author of the e-mail was talking about how his group develops open-source software for research. He is a prominent researcher, who is very well known and very influential. His group develops open-source tools, and it is very difficult for him to get funding even when applying to NIH programs that are targeted at promoting software development and maintenance. In particular, he said, “My group developed an open-source tool written in Java.

We started with microarrays and extended the tool to other data. There were 25,000 downloads of this tool last year. So we submitted a grant proposal. Two reviewers loved it. The third one did not because he or she felt it was not innovative enough. We proposed three releases per year, mapped out the methods we would add, included user surveys, user support, and instructional workshops. We had 100 letters of support.”

This quote is from the negative review: “This is not science. This is software development. This should be done by a company.” We can see that there seems to be a bifurcation in understanding the role that software plays in the development of science.

One idea for the funders of research might be: Why not fund a few projects to be fully reproducible to see what barriers they run into? Is the problem that they do not have repositories where they can deposit their data? Is the problem that they encounter issues of maintaining the software? Where are the problems? Let us do a few experiments to see the stumbling blocks that they encounter.

On the subject of citation and contributions, as we incorporate databases and code, we need to think about how to reward these contributions. Many contributions to databases now are very small, and there are databases where 25,000 people have contributed annotations. Hence, there are questions about how to cite and how to reward this work. What is the relative worth between, for example, a typical article with a scientific idea versus software development versus maintaining the databases? Typically the last two have not been well rewarded, and our discussion here calls that practice into question.

I will end with a figure from a survey I did of the machine-learning community (Figure 3-11). These are the barriers that people said that they faced most dramatically when they were sharing code and sharing data.

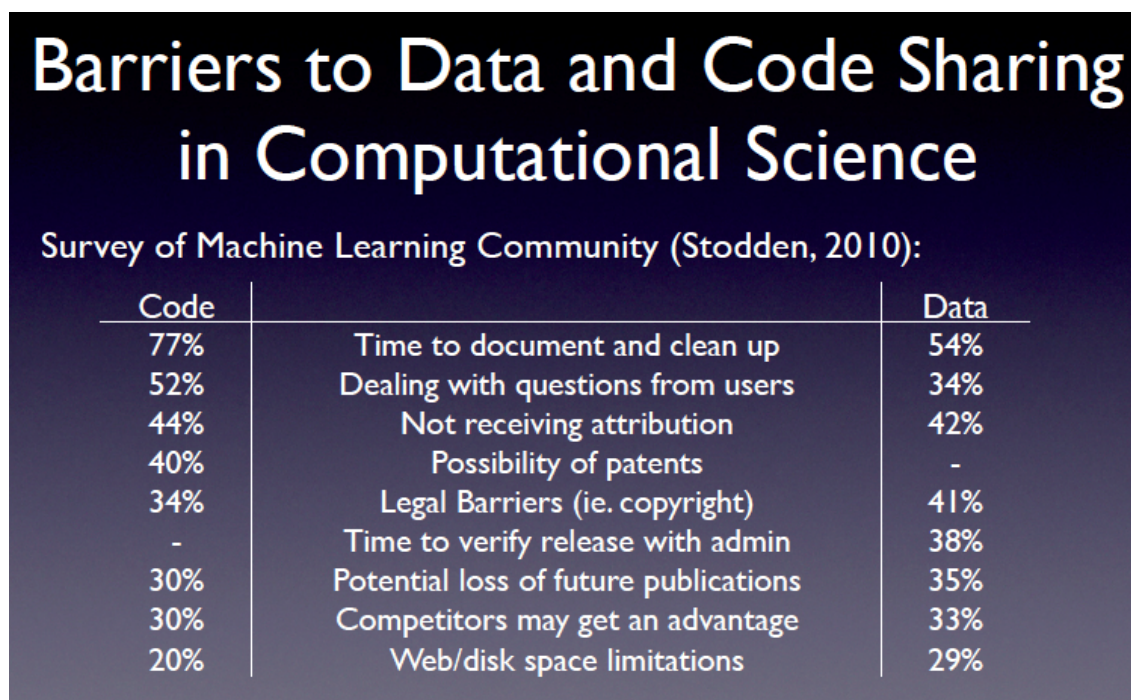


FIGURE 3-11 Barriers to data and code sharing in computational science.
SOURCE: Victoria Stodden

A Government Perspective

Walter L. Warnick
-Department of Energy-

I am the director of the Office of Scientific and Technical Information, which manages many of the scientific and technical information operations of the Department of Energy (DOE). Our goal is to maximize the value of the research and development (R&D) results coming from the department.

To put this into perspective, each government agency has an organization that manages information. Those organizations have gotten together and formed an interagency group called CENDI. Bonnie Carroll is the executive secretary of CENDI. The National Science Foundation is represented in CENDI by Phil Bogden, and the National Institutes of Health (NIH) is represented by Jerry Sheehan. I represent the DOE, and all the other agencies have representatives too, including the Library of Congress, Government Printing Office, Department of Agriculture, Department of the Interior, and practically every other organization that has a big R&D program. Ninety-eight percent of the government's R&D is represented, and several organizations that do not have R&D programs are also in CENDI.

The results of the U.S. government R&D investment, which amounts to about \$140 billion a year, are shared via different formats. There is the text format, which includes journals, e-prints, technical reports, conference proceedings, and more. There is nontext data, which includes numeric datasets, visualizations such as geographic information systems, nonnumeric data such as genome sequences, and much more. And there are other formats, including video and audio.

Each format is in a state of change and presents its own set of challenges. With journal articles, for example, the big issue within the government is public open access versus proprietary access. I think we all agree that the gold standard for text-based scientific technical information is the peer-reviewed journal article. Many highly respected journals are available only by proprietary subscription access. NIH has pioneered a transition to make journal literature publicly accessible. The effort has attracted a lot of attention, and it is still getting a lot of attention within the government. Principal investigators are asked to submit journal-accepted manuscripts for publication in the NIH public-access tool, PubMed Central.

What is significant now is that the America COMPETES Reauthorization Act, which became law in December 2010, calls upon the Office of Science and Technology Policy (OSTP) to initiate a process to determine if public access to journal literature sponsored by government should be expanded. For now, NIH is the only agency that makes a requirement of public access to journal articles. The DOE and other agencies are already empowered by law to adopt that requirement, but we do not have to adopt it, and as a matter of practice we do not. I think that OSTP will soon formulate a committee, which the COMPETES Act calls for, to get input from stakeholders, consider the issues, and develop some recommendations.

Beyond the journal literature, there are gray literature issues, and integrating them with journal literature is important. Gray literature includes technical reports, conference proceedings, and e-prints. It is typically open access, but not all of it is. All of the DOE's

classified research is reported in gray literature, and of course that is closely held, but I am talking here about the open access part of DOE's offerings. Gray literature is often organized into single-purpose databases. For example, in the DOE we have something called the Information Bridge, which has 300,000 technical reports produced by DOE researchers from 1991 to the present. It is all full-text searchable. The average report is 60 pages long, so they are fairly detailed. Many other agencies have similar resources. Other databases handle e-prints, conference proceedings, patents, and more.

DOE pioneered novel and inexpensive ways to make multiple databases act as if they are an integrated whole, one example of which is *Science.gov*. *Science.gov* posts the publications of all the agencies that are in CENDI, so it is a very large virtual collection of databases. It is all searchable, and a single query brings back results ranked by relevance. It has won awards for being easy to use and is an example of transparency in government.

The DOE's largest virtual collection that integrates gray literature, and some journal literature, is *WorldwideScience.org*, which is a computer-mediated scientific knowledge discovery tool. *WorldwideScience.org* makes the knowledge published by, or on behalf of, the governments of 74 countries, including the United States, all searchable by a single query. The amount of content is huge, about 400 million pages. A user can enter a query in any one of nine languages, and the system will search all the databases in the language of the database and then bring back the list of hits in the language requested. It is new, and it is growing very rapidly under the supervision of the international WorldWideScience Alliance.

We also manage nontext sources—the numeric data, genome sequences, and so forth. The main questions are to what extent should such sources be made accessible and for how long. Some agencies are grappling with the issue by formulating data policies. Some agencies require principal investigators to propose data-management plans. The America COMPETES Reauthorization Act calls upon OSTP to initiate a process to encourage agencies to consider these issues, and it is the same part of the act that I mentioned previously that addresses journal literature. Hence, committees stemming from the act are handling both text items and nontext items.

Everything we do entails cost. Whether it is just sharing information or doing analysis of the information, there is always a cost. Here is a way that I talk to my funding sources about cost. Imagine a graph whose vertical axis is the pace of scientific discovery and whose horizontal axis is the percentage of funding for sharing of scientific knowledge (see Figure 3-12). I think everybody agrees that science advances only if knowledge is shared. Therefore, let us postulate an imaginary situation in which no one shared any knowledge. That would take us to the origin of this graph, because there would be no funds expended for the sharing, but there would be no real advance in science either. At the other end of the x-axis, at the 100 percent mark, if we spent all our money sharing and none of it doing bench research, soon your pace of scientific discovery would draw down close to zero too.

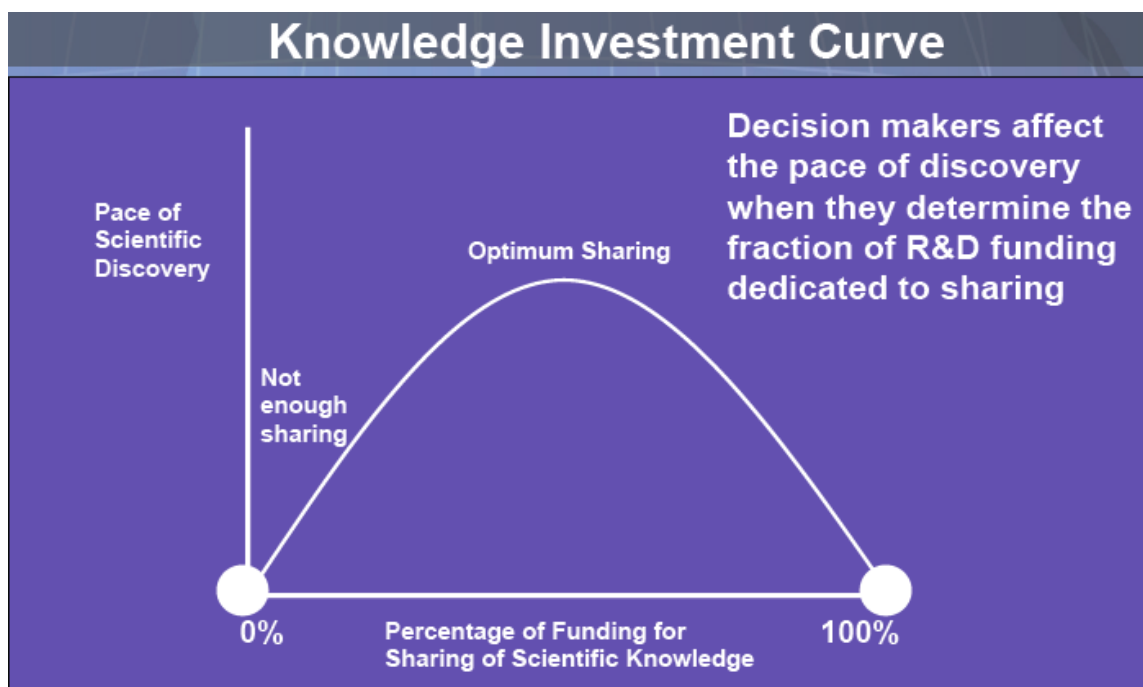


FIGURE 3-12 Knowledge investment curve.
SOURCE: Walter Warnick

We have two data points on this graph, both lying on the x-axis. In between the two points, there is a curve, which we call the Knowledge Investment Curve. We do not know the shape of the curve, but it is likely to have a maximum. The point is that decision makers affect the pace of discovery when they determine the fraction of R&D funding dedicated to sharing. That is the argument I make to my funders.

The point of the Knowledge Investment Curve is to make funders realize that while they can dedicate funds to buy computers, hire more researchers, or build a new facility, they should also weigh that investment against the benefits of getting information out better, sharing it with more people, making searches more precise, and doing the kinds of analyses and data mining we have talked about here today.

It would require a significant research program to calculate what the shape of this curve is and where that optimum is, but we know that such an optimum exists somewhere. Furthermore, the optimum is not the minimum, which is another message I give to the funding sources. If we think that the purpose of an information organization is to be a repository where information goes in, seldom comes back out, and seldom sees the light of day, that is not the optimal expenditure for sharing.

Discussion

DISCUSSANT: My question to Dr. Warnick and Dr. Stodden is related to the knowledge investment curve presented by Dr. Warnick. In a sense, the peak of that graph is the amount of funding spent on infrastructure that enables research versus the amount of funding spent on research. Where do you think that should fall for any example you choose?

DR. WARNICK: We would probably reach a consensus that there is not enough money being spent on, for example, sharing of knowledge and analysis, and the development of some of the tools that were discussed earlier. As to how much below the optimum it is, let me give an example.

Consider the National Library of Medicine as an excellent example of an information management organization. The funding for that organization exceeds the funding for all the other information organizations combined. I am not suggesting that the National Library of Medicine is overfunded, but I will say that the others are underfunded.

DR. STODDEN: I absolutely agree. I think it became much harder than it has been traditionally to share our work. As our science becomes more data intensive and involves code, those two areas add extra expenses involved with sharing that are not wholly taken care of in funder budgets. The science itself, through technology, has changed, and it is making ripple effects through our funding agencies, which have not quite caught up yet, I think.

DISUSSANT: Dr. Hey talked about the new data-intensive work as a new paradigm, yet so much of the discussion has been about things like reproducibility in the traditional sense, but with code and data added, or sharing in the traditional sense, but with code and data added. So where does thinking about new paradigms or new ways of doing things come in? Where do you see that falling in the spectrum of who is responsible and how that affects this whole question?

DR. WARNICK: Even the sharing part is being subjected to new paradigms. Just to make a point, the infrastructure behind *Science.gov* and *WorldwideScience.org* is something that we see very rarely in everyday experience on the Web, and it was developed and matured as a result of some government investment.

To take your point directly about the other kinds of analysis that were discussed earlier, however, since the government is providing \$140 billion of funding for research, then the analysis that gets more mileage out of that research ought to be funded by the government too. Of course, the government always welcomes the idea that the private sector can take and utilize these results, but it must be the funding agencies that do the initial work. I think that the reason why people have not heard the Department of Energy mentioned in this discussion before now is because we were doing very little in this regard compared to the National Institutes of Health (NIH) and the National Science Foundation, and that ought to change.

DR. STODDEN: My perspective within academia is that processes are changing for hiring, promotion, and work committees. The scientists who are clued into these issues of reproducibility, open code, and open data seem to be a little more interested than people who are carrying on in a different paradigm. Academia is conservative in the sense that things change slowly. Therefore, it takes time for these practices to percolate through.

DR. BERMAN: The issue of the gap really intrigues and concerns me, because in our real life, if we want to find a restaurant, we can go to Yelp. We can find which restaurants are near us, what is available, who likes them, and so on. There is an application for that. We can get this information on an iPhone.

Consider taking all of our scientific products and putting them in that world. Is there a place where we can find scientific databases and see who liked them? Can we see who added metadata to them in a very user-friendly way? Can we access them easily?

We are starting to see crossover between the academic world and the world of commercial applications. Phil Bourne has a project called SciVee, where we can show how to do different kinds of experiments or give a talk on a data YouTube-style. We can imagine using many of these commercial types of applications and technologies in academia. Some interesting questions arise: What does it mean if we have a data collection and many people like it? Does that mean it is a good data collection? Does that mean it is an economically sustainable data collection? We should not utilize one set of tools for our academic work and another set of tools for applications in the real world without bridging the gap.

MR. UHLIR: There is a proposed act in Congress, the Federal Research Public Access Act, that broadens the NIH PubMed Central grantee deposit policy to include other agencies with annual research and development budgets of \$100 million or more, although I do not know if it is going to actually become law.

Also, in the list of peer reviews presented earlier, there is one other model that was missing: postpublication review. It is not a traditional peer review. It is an open peer review, it is moderated, and it is ongoing. There are two kinds of this model. One is just commentary, and one is papers generated in response to a big paper. The model I am thinking of is the European Geosciences Union's *Journal of Atmospheric Chemistry and Physics* in Munich. I do not know how many other journals do that, but it is yet another model for a review. Even if the code and the data are not available, people can ask very pointed questions that can be answered by the researchers. That is a potentially valuable way of reviewing results.

In response to Dr. Hendler's comment, there have been several people who have made some intermediate suggestions, such as Dr. Friend's journal of models being published. But the fundamental problem is that we have moved all the print journals wholesale onto the Web without hardly changing the model at all. To obtain good models, one would deconstruct the scholarly communication process used in the print paradigm and reconstruct it in a way that makes sense on the Web. Thus, we are repeating everything we have done before and not really thinking about what the Web allows us to change in order to achieve greater efficiency. I think the print journal system itself is an outmoded way of communicating. I am sure there are all kinds of new paradigms, but I will leave it at that.

4. Summary of Workshop Results from Day One and Discussion of Additional Issues

Introduction

Bonnie Carroll

-Information International Associates-

Today, we will take the results of the discussion from the first day of the workshop and build on them. We would like to hear about some options for what could be done over the next few years and what the value proposition is. Closely coupled with that is the issue of how we will know when we have succeeded.

We had some excellent concepts and ideas during the first day of the workshop that highlight much about the nature of the changes that are ongoing. I heard that we should not stay static and that the community might benefit from a less linear and more dynamic approach. We talked about trailblazers versus the long tail of data activities, and whether we are getting into a credibility crisis, because many of the things we are doing now are not as shareable as they were when they were in print. Many codes are not available, for instance.

The title of the workshop is “The Future of Scientific Knowledge Discovery in Open Networked Environments.” The title recognizes that we are adding a fourth paradigm to the research process. We already have three paradigms—the theoretical, the experimental and observational—but now we have added the computational or the data intensive. The presentations we heard during the first day of the workshop focused mainly on the latter paradigm.

How do these approaches build on each other and work together? Why did we immediately jump to the data-intensive sciences when the title referred to a networked environment? The answers are obvious. First, it is because the technology is so enabling, and second is the data deluge.

In the text that follows, we discuss the four sets of topics identified in the statement of task. Each section begins with an Introductory Summary of the Issues Identified in Day One of the workshop. In the first and fourth sections, Puneet Kishor reviews the sets of issues from the first and fourth items in the statement of task, namely, the Opportunities and the Benefits and the Range of Options. In the second and third sections, Alberto Pepe addresses the Techniques and Methods, and then the Barriers, respectively. Each of these Introductory Summaries of the Issues Identified on Day One of the meeting by these two rapporteurs is then followed by a Summary of the Discussion of each topical area by all the workshop discussants on Day Two of the meeting.

4. Opportunities and Benefits for Automated Scientific Knowledge Discovery in Open Networked Environments

Introductory Summary of the Issues Identified on Day One

Puneet Kishor
-University of Wisconsin-

Among the issues we heard about during the first day of the workshop was that certain problems in science, no matter what domain, are really the same. A scientist performing observational research, for instance, generates data and then manipulates, organizes, analyzes, visualizes, and reports on them, or maybe preserves them, and then the cycle repeats. The digital medium is one of the best ways to communicate knowledge and is probably also the best way to create knowledge. The concept of a knowledge ecosystem, the whole system of producing, storing, and retrieving knowledge, depends on standards, metadata, discovery, association, and dissemination.

One presentation focused on creating a precompetitive commons, using examples from drug discovery, where fierce competitors can cooperate to produce evolving models of diseases. Cooperation can improve the translation of publicly available molecular data into biomarkers, and increase the opportunities for using drugs meant for one disease to treat another disease.

The “crowd” plus the “cloud” concepts are important to the future of the network. There was an assertion that science can learn about data sharing from the World Wide Web, which could be argued either way. Discussants also mentioned exploring linked open data for interdisciplinary uses. Access to many data is very restricted.

Also, there was much focus on the long tail of science, which is defined as those fields where the benefits of information technologies are not immediately apparent. There were conversations about quality, format, access, financial support, and policy issues.

Other topics that were discussed related to opportunities for innovation that we cannot even imagine right now, serendipitous innovation, still others focused on government policies that encourage or even require scientists to share their scientific data and information, such as the America COMPETES Act.

Summary of the Discussion of Opportunities and Benefits

This session examines the opportunities and benefits in a 5- to 10-year time horizon, so it is important that we understand the value proposition. Do we really know what the benefits and the opportunities are?

Building and Sustaining Knowledge Networks and Infrastructure

Can we say that one opportunity is to provide more and better infrastructure? If a common infrastructure is not there, is it possible to have such a common infrastructure so that anybody can use it?

When we hear a term such as *knowledge ecosystem*, it raises the concept of a knowledge network. The discussion on the first day of the workshop about a nonlinear way of thinking or a linear way of doing science reminds us of a neural network. That is, there are many different synapses in the brain that connect different facts and different memories and experiences that lead to an understanding of a situation. The same kind of process happens in science.

We saw the example of the pharmacological studies that found linkages between drugs that were used to treat one disease and that could be used to treat another disease because there was a common gene sequence. There was a connection in the network that showed that the same thing that is being used somewhere else actually does link back. This suggests moving toward not just linked data or a knowledge ecosystem but a linked knowledge system. That is, we could merge those two ideas. We could benefit across all disciplines and truly enable long-tail science—the science that is so removed from our experience that we never would have thought of doing it.

We are likely in a period of persistent restricted resources, however. Hopefully it is not a zero sum game, but assuming it is such an era, what is the best way to deal with that situation? Do we need to think of ways of developing systems that could be used by many different kinds of scientists? Are there general characteristics of such systems? Again, we do not need something only for computer science; we may also want something for the agencies funding research.

Demonstrating the Value of Data Sharing

Change is always going to be before us, so how do we work together in our public-private partnerships to adapt to change? We need to show the value of sharing data to our principal investigators. Science is going to be done differently than it was by scientists 20 or 30 years ago, so we have an opportunity to redefine the science as well.

We can imagine a scenario in which a scientist is speaking to a congressional staffer who has heard that scientists are just “welfare queens” in white coats and wants to know why they are being supported. We need some convincing examples where data have led to important discoveries. Earth sciences data and medical data could provide some useful anecdotes, for instance.

We could survey the number of hours people spend analyzing the databases as opposed to the number of hours they spend reading traditional journals. For example, Princeton University spends about \$10,000 per faculty member per year subscribing to journals. If we could quantify the relative costs, we could put a value on the database. The bottom line is that

when talking to somebody who is not a scientist, such people do not necessarily start with the assumption that science is worth something, so we need more descriptive stories. For example, the digital library program is justified by pointing to Google. We need more compelling stories that appeal to outsiders.

Also, a protein chemist cannot work without the protein data bank. There are many scientists who do not work the way it was done before there were digital databases. They are data-mining researchers. If you take away the databases, their type of research would stop. You will get different answers in different kinds of research, however. For instance, there is much more to online astronomical research than the official National Virtual Observatory or Virtual Astronomical Observatory (VAO) projects. Although the VAO has led to many useful results, so far it has focused only on infrastructure, without building end-user tools. In astronomy or in other sciences, there are databases that are even larger than the protein data bank. Researchers in these fields use them every day, but the results in some seminal papers may not be cited.

Astronomers use data collection facilities such as the Sloan Sky Survey, the Two Micron All Sky Survey (2MASS), and the upcoming Large Synoptic Survey Telescope. Scientists see these huge databases as something on which they have come to rely. What they may not understand well, however, is what would happen if the medium- and small-scale datasets that are attached to literature would also be openly contributed. Incentives alone might not work. Scientists may not be convinced that their datasets are going to be worth depositing, unless the funding agencies make that a condition of their grants.

Big science examples may be the wrong approach to convince scientists, however, because they are too common. In the physics community, researchers would rather work with the Large Hadron Collider (LHC) than their own small data collections, because the LHC will have a huge amount of data. There are other fields in which the generation of the data may not be very expensive, especially when amortized over a longer time period, but may nonetheless be costly to maintain, because many people are involved or for some other reason.

There also may be a potential problem in asserting that it is more efficient to share data, because if this were true, why are scientists asking for more money to do that? If researchers are going to be more efficient, they should be able to perform better science with fewer dollars. If this is not the case, then there is a question about the efficiency of the system. Also, when we say that it is going to be more efficient, another issue is: Where will the savings go? The savings could go into the analysis of existing data, rather than into collecting and organizing new data. Put differently, the budgets may not go down, but the analytical capabilities could increase.

The research funders could identify the different areas where data sharing is extremely important, and they could then establish a reasonable time span for making the data available. Some guidelines to the scientific community could be useful in this regard and constitute an opportunity. For example, we might eliminate the 2-year gap between obtaining the research results to the sharing of those results through publication. This could allow the timelier sharing of the methodology, the data, and even the codes developed for the study. It could lead to real benefits.

Too much specificity concerning what should be shared, however, can be a problem as well. If science were segmented into different categories according to a level of what should be shared, it could be counterproductive. A specific time frame for sharing data absent

countervailing considerations, such as the Health Insurance Portability and Accountability Act (HIPAA) or similar requirements, could be very useful and would signal to scientists that there is recognition that sharing is very important. It would encourage more scientists to start thinking about the next steps after they produce the data, what infrastructure they will need, how they will have to reorganize their own budgets and funding to do that, and so on.

A rigorous deadline can be a problem. Voltaire said, “The best is the enemy of the good.” Everybody is quick to complain that data curation is too much work. What is the cost, however, to the nation’s scientific effort to peer review weak papers? In some fields, preprint distributions and conference presentations have essentially taken over the publication function, and the journals exist so that people can get tenure. A system relying on something like the Cornell arXiv might be better for scholarly communication, if there were something equivalent to page ranking that would help people decide what they should be reading. If the publication system were less strongly tied to tenure, it might be easier for the universities to determine that they could use other metrics. This could save a lot of money if the universities no longer relied on the ranking of publications to determine who would get tenure.

On the one hand, the promise of data sharing is becoming real. On the other hand, there is a powerful reinforcement of Max Planck’s observation that science proceeds funeral by funeral. In the health sciences, the progress of science may be slower than it has been historically. Some scientific processes may even retard the pace of scientific progress.

The most difficult organizations to consult for are the ones that are doing well, because if they need to change the things that they did to make them successful or rich, their view, of course, is that they are fine and they are not receptive to good ideas. The difficulty is to get the existing order to change their approach, when it is needed.

This is a collective action problem. Individual scientists who would like to fulfill their ideals as scientists often find that they cannot. Framing the problem in terms of reproducibility of scientific results can be very appealing to scientists, however. This can be a way that gives them guidance about when to share data and what types of data to share. The same argument can be helpful in convincing people of the importance of openness in this networked environment, because this is a core principle of science. Nobody argues that reproducibility is not important in the sciences, so this can be useful as a guiding framework for what steps are important.

The climate data controversy at the University of East Anglia in the United Kingdom and elsewhere has indicated that transparency is also key. The average person is not going to reproduce those climate simulations, but citizens want some assurance that there is access to how the model simulation was made and what data came out of it. Therefore, transparency and reproducibility are both important.

We also may separate the different axes along which we can argue that there are benefits. We talked about the speed of communicating research results, so presumably the discovery of new results is one aspect. We have talked about value, and a couple of discussants reminded us that the value of discovery, at least when we are talking to the public, is predicated on the listener’s view of the value of the underlying science. If someone believes that many discoveries in the underlying science are not particularly valuable in the first place, being able to make more of them faster does not seem to accomplish very much.

A strong case for value can be made in the biomedical sciences, particularly as research is translated into health care. For example, it might be possible to determine the real dollar value in repurposing pharmaceuticals. It might be worth picking an area in which the economic value is fairly noncontroversial, as opposed to something like basic mathematics or astronomy, and then start quantifying the value and take a look specifically at data reuse in those settings.

One thing that has not been mentioned in this discussion, but was talked about yesterday, is the idea of data as an unexploited resource, because data have not been part of the traditional journal-based way of communicating science. Negative results in early clinical trials were traditionally not reported, for instance. Because doing that work costs money, there might be some quantitative way to get at the value of what is now being withheld.

A couple of the discussants at this meeting were recently at a geosciences data workshop, and the question arose about how to demonstrate the value of initiatives that integrate data and make those data easily available. For example, congressional staff members are interested in problems such as energy, food, and water. If we are asking for additional resources to enable data sharing, it can be useful to tie that to solving some of the issues important to them—to medical benefits, for instance, or better ways of managing natural resources. That also bears on the discussion of incentives and mandates, because when scientists feel that their work is directly related to producing a social benefit, there is additional incentive for them to share the data.

Speeding Up the Pace of Science

The repurposing of data can be a real opportunity in speeding up research. How do we learn from the World Wide Web? What are the inherent opportunities? Companies are investing a lot of money and making advances in communication that people are using every day. Scientists may have a problem in seizing the opportunities, yet everyone is using the commercial media to communicate, to make decisions, and do many other things. Is there something we can learn from the commercial community to apply to science in this regard? Can we encapsulate very concisely some of the opportunities or the benefits?

In March 2011, none of the U.S. newspapers had any headlines about the earthquake and tsunami in Japan, because, of course, it happened early in the morning, and they had already gone to press. In the old days of newspaper-based journalism it would be in the evening edition before most of us knew what was happening. Then when radio and TV journalism came along, they became the media that provided news during the day. Now we can watch videos of the tsunami almost instantaneously on the Internet.

Think about science at that pace. An astronomy course 20 years ago had films that were not very compelling. It was like playing chess. Not only did you have to be smart but also very patient while the other player moved. To keep up with the modern world, the pace of science has to change. It is clear that data science—not just the big data science but the type of science where we can find the appropriate data, pull it together, and quickly get it into a workflow—is what will change the face of science. We need to keep up with that kind of speed.

Much of the reason that we may use the internet as an analogy for how science can be done is because of the speed of change. Just a few years ago, Twitter did not exist. Many people who use it now cannot live without it. We are being asked to solve problems that are changing at the pace of the internet.

The geosciences are well positioned to react to this real-time aspect that was just alluded to, whether it is in the study of earthquakes, tsunamis, flash floods, or tornadoes. Scientists have the capability today to bring that information to be used in real time within seconds of the event happening. This also bears on the point that was made earlier about tying the research to societal benefits that resonate with people in Congress. What does this mean for the protection of lives and property? The rudimentary informatics are in place that can be greatly expanded in the future with the mobility that comes with the smartphones and other mobile devices, so that you do not have to be at your computer anymore to get real-time information.

The Web and semantic data, and linked open data, have had major effects, but not all science may be done at the speed of Twitter. It is one thing to report some results quickly, which is what Twitter and similar tools do; it is another thing to trust science that has been rigorously tested. For example, there is a 10-year study on body fat as a predictor for heart diseases. A 10-year study cannot be compressed into 24 hours, however. Some things just cannot be hastened to completion. This comparison with the speed of Twitter and Facebook, therefore, can be misleading, because it is comparing apples to oranges. The problem with putting these sorts of ideas in a recommendation is that a policy maker might latch onto this and say we want science to be done overnight, which could be even more of a problem rather than a benefit.

Preserving the rigor in science is a very important message, but in the current system there is a major delay from obtaining the research results to publishing them. We should not say: “I am sorry you cannot see my data, because that is how I am going to win my Nobel Prize 30 years from now.” These practices and attitudes are barriers to the goal of solving people’s real needs with science and to better understand the phenomena around them. The 10-year study is great when you need a 10-year study, but not if it is the primary way of doing science. For the next 9 years, you may have people dying who could have been treated by something that might have been ready to work. Verifying, validating, and doing rigorous science, of course, are important. It is the sharing of information that can naturally speed up some of these processes, however.

People in unrelated fields right now may not easily discover prominent work until it has appeared in the archival journal. This means it has been written up and the visualization of data is prepared, sent to the journal, reviewed, and gone through the publication deadline. There are faster approaches that tend to be very different from field to field, but the main issue is the delay of disclosure, the long time it takes to go through this life cycle, workflow, or ecosystem.

We also ought to keep in mind that one size does not fit all for these processes. When we examine the management of data, we may be talking about long-term experiments or about datasets that are developed to help in an emergency. Just those two instances present very different kinds of data types and uses.

Supporting Interdisciplinary Research

Most of the presentations on the challenges from the first day of the workshop were very interdisciplinary, and most of the presentations on the solutions that described what people are building were inside disciplinary boundaries. That actually has to do more with how the funding mechanisms work than with what people would work on in a natural setting, but that interdisciplinarity of the key challenges is important. What happens on the Web is that a

small number of people who can work between areas can transfer huge amounts of information back and forth between those areas. That kind of work tends to be viewed as an ancillary process, not as a major part of doing science. Further, most scientists who have done this kind of work did so at the peril of their careers and may have been fortunate.

Another point is the importance of informal communication. Scientists are used to formal communication. Some of the systems that were mentioned yesterday and some of what the astronomers are doing now have, in addition to the primary scientific channel, an extra social and communicative back channel. For example, informal communication is how scientists may get information from conferences they have not attended personally. Journal papers are the end product that gets put in the library for historical reasons. That is why they are called archival journals and not always the primary reporting of the scientific activity. Recognizing the need for much more of that informal interchange is also important. Those are two key aspects of research—an interdisciplinary approach and informal social communication.

Informal communication can be structured like the World Wide Web: as a network of information and knowledge that you can search and also find the linkages between them. We can take advantage of that type of structure, which does not eliminate the journal system, but improves the communication among different sciences, and not just within a single science.

If we were going to try to find these commonalities that allow us to cross boundaries in data sharing, those are problems that can be addressed. There are high-dimensional image formats, there are complex text formats, there are big table formats, and it does not matter which discipline community produces them. The tools for analyzing, visualizing, and integrating them depend on the structure of the data, not on the contents of the data. Thus, we can look at what those different kinds of datasets are and what the solutions might be.

We have talked about knowledge networks and collaborative environments as being able to communicate and get results out more quickly. One caution is that there are many models for those kinds of practices. As new ways of communicating and collaborating are developed, it will be important to think about what actually is happening (in knowledge and experience), how we collaborate and communicate, and how we capture that so that it can be reused and mined.

One of the issues of citizen science and informal education is the quality of the data, how good the data are for “real scientific research,” and how to describe the quality of the data. That remains an ongoing question.

The National Ecological Observatory Network (NEON) is sponsoring another project, called Budburst. It is a project in which people with cell phones go around and take photos when the first leaves come out from deciduous trees and when flowers first appear, because that timing is expected to change with climate change. It is a very potent opportunity to provide a bit of informal education to citizen scientists. They try to figure out what a plant is, they learn something about how climate and water affect the plants, and when they upload their data, the Web site offers some educational opportunities.

Effective Incentive Systems

There is a 2003 report from the National Academies Press called *Sharing Publication-Related Data and Materials: Responsibilities and Authorship in the Life Sciences*¹⁴. It has many kinds of recommendations that focus on how it is good to share research inputs and what it means to promote the progress of science. It probably has not had a tremendous effect, however. Given that we have been there before, a case can be made for the greater use of deterrents.

The stick-versus-carrot approach has other dimensions. Recently, there was a thought-provoking book called *Switch* by Chip and Dan Heath¹⁵, which talked about the elephant, the writer, and the path. It made the argument that you are never going to get the elephant to move in the desired direction with a stick. It has to be persuaded to go in that direction, and you need the right incentives, the right carrots. More important, however, is that you need the path to be paved. Therefore, we should not focus just on the carrot-versus-stick approach. How can we lower the barriers for the scientists and investigators to share their data, information, and products with the broader world? Not all scientists are opposed to sharing their data because they want to hold onto them for the Nobel Prize. They just do not have the time, resources, or incentives to be able to share them. Agencies can focus on enabling and motivating the investigators to share their data, and on giving them the right tools for effective data sharing and data citation.

It was pointed out that in one discussant's organization, it is the bench scientists who are demanding national data systems to be built, but this is not happening broadly. The use of one of the U.S. climate change data systems has been adopted by 20 countries, and they are supporting it as well. It is the bench scientists who support this, because they want to do better science and they understand it. In that organization, they are the ones who are demanding, so a stick is not necessary. They would love to do it faster and have a greater impact, but they do not have the funding to expand their scope. The money may serve as a carrot or a stick, but if researchers had a little more funding they might expand their impact quite significantly.

In that agency the bench scientists are demanding greater sharing of data, and they are the ones who are realizing that their work is going to be affected. They are changing the way they are doing science. They need some direction from the general science community so that they are not going to be wasting their limited resources. For them, the incentives can be additional funding and a clear direction on how they ought to be documenting and preserving their data.

The separation of science into gathering and analyzing data would mean that people could be scientists with a smaller skill set than is now possible. Such an approach could open up the field to wider participation by people who at the moment cannot easily do so, because they have not gotten all of the training to do both sides of the job.

The Heritage Health Prize is a \$3 million Netflix-like challenge to work on hospital data. It has motivated a huge amount of interdisciplinary effort at Stanford University among the students and faculty in business, computer science, and biology, as well as medical doctors.

¹⁴ Available at http://www.nap.edu/catalog.php?record_id=10613

¹⁵ Chip and Dan Heath, *Switch: How to Change Things When Change is Hard*, Broadway Books, 2010.

People have formed teams to work on this prize, because there is no barrier to form such teams. Many of these prizes have no barriers to participating if the person has the required skills.

The America COMPETES Act explicitly authorizes the National Science Foundation (NSF) to put forth scientific challenges and to have reward money for solving them. It is being “legitimized” with the federal agencies. The question thus arises whether in an open network environment well-documented crowd-sourcing can be an opportunity for advancing knowledge discovery.

Policies of Research Agencies

There was a very compelling diagram presented during the first day of the workshop about the optimum curve that depicted how much of the scientific community’s research investment should go toward data sharing. We could pose that question to the National Institutes of Health (NIH) or to the NSF and ask them to develop a theory based on that and then to manage their portfolios accordingly.

Federal science agency managers may generally understand these areas related to data sharing. Every agency has, in its own way, made a commitment to funding different kinds of resources. For example, NSF has been supporting the National Center for Atmospheric Research along with other research centers for a long time, which has put those institutions in a very good position. Also, NIH, the Department of Energy, and the Department of Agriculture have provided a great deal of genomic funding, which has helped the funded projects to be able to harvest the resulting data and use them for other useful and interesting ends.

A one-size-fits-all solution, a monolithic approach that everybody has to use, is probably not appropriate. There are small and simple solutions, but there are many of them and they are interconnected, they are designed for specific purposes, and they are run at economies of scale. The NSF and other research-funding organizations have their own mechanisms—ones that are good for their scientists. They can offer solutions to their scientists in the way that the GenBank or the Protein Data Bank were offered, and the scientists can choose to use them or, in some cases, have to use them. For many scientists, however, that opportunity for a solution is missing.

Techniques and Methods for Development and Study of Automated Scientific Knowledge Discovery

Introductory Summary of the Techniques and Methods Discussed on Day One

Alberto Pepe

-Harvard—Smithsonian Center for Astrophysics-

Fewer techniques were identified from the discussion on the first day than barriers. One technique that we already have been using for discovery of scientific knowledge is to develop and install open-access repositories that function not only for full text but also for images, data, and software.

The need for alternatives to traditional ranking mechanisms for scholarship was also mentioned. For example, the University of Southampton could be ranked higher than Oxford or Cambridge universities in the United Kingdom, if we look just at the impact of the university on the Web. Therefore, when looking at the number of citations that a university receives, we could use not only traditional scholarship but also the peer-reviewed scholarship that exists online.

Another technique that was mentioned as a need by many presenters is semantic computing, or the development of systems that take semantic meaning into account. As one of the discussants noted, computers are great at storing, managing, indexing, and accessing information, but in the future, they will also need to make sense of all the information that they store, manage, index, and access.

Many presenters also spoke about linked data and the idea of a faceted search. For example, we have seen the new Astrophysics Data System (ADS) Lab interface that allows us to provide an interface to the information as we search for it and to somehow filter down the information for which we are looking .

Yet another technique for scientific knowledge discovery is providing direct access to the data. The example of the ADS Labs and the faceted interface is relevant here as well. Also, one of the presenters referred to using data as a filter to literature. In order to allow the greatest possible access and to enable discovery of scientific knowledge, we could use the data as a filter to the literature and not just use the publications as a filter to the data.

Versioning was another technique that was mentioned throughout the discussion yesterday. A presenter noted that we could probably use softer versioning systems to capture scientific data, practices, and the artifacts that are produced across the scientific life cycle. Two types of integration were raised. One of the techniques required for the enabling of scientific knowledge discovery is the integration of existing scientific tools with Web 2.0 tools, or simply with existing tools on the Web. An example is the astrometry.net site. It is a research tool that is primarily intended for the scientific community. It determines the position of

images uploaded by users on Flickr. This is an excellent example of a tool that seamlessly integrates with an existing application on the Web.

The second type of integration is with social networking sites. Scientists can leverage the user base that already exists in these services to accomplish several objectives. The first is the scientific discovery process. How can a researcher use data and content that are already published by users on social networking sites to enable scientific discovery? Another objective is to enable citizen science. Computer-supported technologies and social media can be used by nonscientists to take part in different scientific enterprises. Finally, how do we use social networking sites to allow people to collect data, either citizen scientists or actual scientists who want to collect and publish data on the Web?

Another issue identified in the discussion was applying large-scale statistical methods and computational models of public scientific data to both predict and track scientific knowledge discoveries and constructs. One example was the heat map used in the biomedical research area that was described on the first day of this meeting. It examined the use of drugs against diseases. This is a kind of research that is made possible by the use of computational science—data science—on publicly available scientific data. It also could be done in parallel with computational social sciences, which attempt to do exactly the same thing using public social data.

Some people suggested that we need discovery tools for scientific knowledge that are data driven—that is, they are bottom-up rather than top-down. Google uses distributed data to organize, rank, and provide access to information rather than using a prior classification scheme. Researchers might do something similar and use the features of the data themselves to discover scientific knowledge.

This is probably the most efficient and popular technique for scientific knowledge discovery today: data visualization. Statistical visualization and other visualization techniques enable knowledge discovery and make possible the observation of scientific phenomena that data alone may not reveal. We heard about different examples of using visualization in astronomy, as well as in geospatial and atmospheric research. We may not be able to see some features of the data just by doing data analysis, so there are some phenomena that visualization allows us to understand and observe.

Finally, returning to the issue of semantics, the notion of using simple metadata and lightweight semantics to annotate the data was brought up. This was described as micro-level semantics. Some presenters suggested that in some cases the use of lightweight semantics—very light annotation schemes—is enough to describe scientific data and provide metadata for the data that someone is trying to discover and explore.

Summary of the Discussion of Techniques and Methods

The title of this workshop is “The Future of Scientific Knowledge Discovery in the Open Networked Environment.” Although much scientific research is taking place in this open networked environment, it is not proceeding as quickly as it could, because of the various barriers that have been identified. It is important to understand these barriers as well as take advantage of the opportunities that we have cited to help us move more quickly into the open networked environment. It will be a mix of both bottom-up and top-down approaches, however. Also, both young and senior scientists work together and there will be success stories in both directions.

Computational Models and Semantic Computing

It should be noted at the outset that the discussion focused on data, rather than the scientific, technical, and medical literature. The sheer size of the datasets that are being managed is a potential barrier, but there is also a potential solution or a technique to deal with that issue. We can ship the algorithm to the data rather than ship the data to the algorithm. Through Web services or some other approach, we could allow a scientist to process the data wherever they are rather than bringing the data to the scientist. It is still difficult to store a petabyte of data on a desktop. This may be needed, however, not because it is hard to move a petabyte dataset, but because it is the way by which the data owner can control what other people are allowed to do.

The idea of shipping the computational programs to the data makes good sense, because the software is much smaller than the data in size. The problem is that scientists often do not work with one dataset. They may be analyzing 10 different datasets. What will they do in those cases? Ship software to 10 different sites and have researchers perform part of the work there, then ship it somewhere else, perform part of the work, etc.? It does not work that way. For example, geographers may do a simple analysis and have perhaps 16 geographic layers for analysis. Therefore, the issue still comes down to access to data.

Researchers often need to develop an intimate knowledge of the data they are using. Sometimes it is a great idea just to ship the algorithm to the data, but if scientists do not actually see the data, if they do not have the data on their computer, and if they are not running some tests and some examples and establish that sort of relationship with the data, then they also cannot run an algorithm. Consequently, sending the algorithm to the data can be a great idea, but in some cases there is a barrier that has to do with how we manage scientific data.

The following anecdote is relevant in this regard. Years ago an astronomer started working with computer scientists. The astronomer’s interest was in distributed data mining, because she had data that the virtual astronomy observatory was making available. For the first few months of their collaboration, the parties were not communicating well. Not only were they using different terminology, but they were not even talking about the same paradigm of research. They finally realized that the astronomer’s interest was to mine distributed data, while the computer scientists’ research interest was the distributed mining of data. They reached a compromise to use both approaches. It was relevant, because for it to be considered research in the computer scientists’ domain, it had to be computer science research, not astronomy research, so just downloading data to the astronomer’s laptop and enabling her to run an algorithm was not computer science research.

To perform distributed principal component analysis (PCA), we need to do an iterative solution on the PCA by shipping parts of it to distributed data sites. The data are never brought to one central location. It is possible to get very good accuracy on the PCA vectors in only a few iterations, however. That was an interesting result for both parties. They had to have enough network bandwidth to get the data to the astronomers, but also, for cases where that was not possible, to have the computer science researchers ship the code to the data.

Offering many collections of data for downloading is a fairly simple service that can be provided either by an individual researcher without much sophisticated work or on an institutional basis with a one-size-fits-all approach through a repository, for example. When you query to database, however, you open up to a Pandora's box regarding computational security. You need to make sure that queries do not run astray in various ways. These can be managed, but require a greater degree of operational sophistication. The last thing we would need as researchers that are encouraged to share more data is media coverage of a high-profile, security breach.

One discussant noted that he has been lucky enough to work with some of the top people in the Web community. He got his credibility when he gave a talk and had the word "the" with a kill ring over it. He said, "On the Web there is no *the*." However, he is still hearing: "*the* way to do this," "*the* way to think about this," "*the community I work with could only do it this way.*"

The open network is not bounded narrowly, which means that some things will work for one group and some things will work for others. It is important not to try to propose the same architecture for all scientific data sharing. We can examine a litany of such approaches going back approximately 40 years. Remember collaboratories and the access grid, and how that was going to change science? Part of the reason it did not was because such concepts were built on a single model that worked great for some people, but not for others. We therefore should try to explore many different ideas, and not have a closed boundary: explore different ways of doing things.

There will be some small solutions that are the only solutions that support certain communities, and we should not resist those solutions because they do not work for the rest of us. At the same time, there will be some very large communities or some aspects of our work that cut across the sciences, and those techniques may benefit greatly from the sort of soft infrastructure that was discussed on the first day of this meeting. It is important to remember that we are talking about a very large-scale, multiple-science, and interdisciplinary set of issues. There is not going to be a single architecture that solves all of that, including the Web itself.

One discussant wanted to underscore a question that was raised earlier in the workshop about how effective some of these techniques are, particularly the text-mining techniques. It could be very useful to find some specific successes for those kinds of methods of knowledge discovery. This is an area of technology that has had a lot of investment. It would be possible to make a lot more investment, but it would be helpful to understand the efficacy of the work before significant investment is made.

Another discussant wanted to know to whom the techniques or the methods or the improvements are specifically directed. There are many stakeholders involved. If the techniques try to address everybody, they address nobody. In the case of an open-access

repository for the medical community to improve pharmaceuticals, who are those stakeholders? There could be equity fund managers, program managers, Congress, the public, and others. They are all different, and they are going to look at it differently. If we do not associate the techniques, methods, and barriers with those people, we may miss the audience in showing them and explaining the impact. As these topics are discussed, the primary stakeholder at issue ought to be identified.

When talking about open-access data repositories, we should look at models too. Much effort has been devoted to open-access repositories for scientific literature and data, but various models could be added to the analysis, because understanding different ones can be key to the future of data-intensive science.

Other issues include semantics, data visualization, and data mining. The purpose of semantics is to help the computer understand. Data visualization is important once there is something to see, while data mining is the computer discoveries that you have in a database. Those are very different things, each of which comes from a different sub-discipline, but they all need to be dealt with each on their own terms. Paul Peters, who died in 1996, said something along the following lines, “In the Paleolithic Age, there was a saying that information was food for thought, but in the Future Age, information will be the predator and the human will be the prey.” If you think about that, this is the ultimate insight in the presaging of data mining and what is really going to happen. Information is going to be finding us, so how do we end up there?

Another issue is that if data are going to be a filter to the literature, then we need to be able to cite data and their connections to the literature. There are many issues in data citation; thus, it becomes much more important that we know how to define what we are talking about and how to name it.

Two more issues were raised in this content. The first is that reproducibility is really hard. Simply having the code and the data does not give you scientific reproducibility today, let alone in 10 or 20 years. There is much more to making sure that things work together and that we can use that code effectively and intelligently to get the right results and be able to interpret these results. It is not at all clear that if we just had the data and the codes, reproduction of the results could be accomplished; they most likely would not.

Also, when discussing code reproducibility or reusability, brand is important. For example, there is Google labs and all their products. There are millions of lines of code in Source Forge, however, that never get reused by anybody, simply because they are not recognized as being a reliable brand of quality software that can be easily obtained and plugged into different tools.

Data Visualization

Visualization is one of the key tools that can enable knowledge discovery. Visualization is important, because it is needed not just to understand the data after they have been analyzed but to understand them while they are being analyzed. For example, sometimes we cannot figure out what is going wrong with a thousand-by-thousand matrix, and it would be useful to just be able to see it and have something that would allow a scientist to do that. It would be good to focus on visualization tools for that reason. However, visualization is not sufficient

alone, because there is an infrastructure that needs to be in place before we can take advantage of visualization.

Data mining and data visualization were described as separate areas. There was a provocative conversation at one of the workshops on data and visualization that Dan Atkins and Tony Hey convened for their advisory group to the National Science Foundation's (NSF) Office for Cyberinfrastructure. Their topical area had covered both data management and visualization, and visualization seemed to be a very siloed kind of technique. We can take a more holistic view of data analytics, however, that comprises mining, visualization, and perhaps some other intermediate or allied techniques. It could be useful to think about these issues, because there is a tremendous amount of siloed work going on in various places, and as the data science scales up, a more holistic approach would be desirable.

One discussant noted that the word “analytics” is seen frequently in the literature these days, but “informatics” better describes what we are talking about. Informatics includes visualization, mining, statistics, semantics, knowledge organization, data management, data organization, and data structures. That is a holistic viewpoint.

What was indicated during the first day of the workshop is that people are willing to use generic visualization tools that are repurposed—which they can get from companies such as Google— as a substitute for what they might have done with higher-end software. We can call this “modular craftsmanship,” which involves people who are very good at making tools and making them more accessible to the point that anyone can plug them into a system. Those tools will most likely continue to get better.

Such tools also could have useful connections to highly specialized data formats that might be specific to a field. The scientific community could use applications that are being developed commercially but are available free of charge. For example, a YouTube for visualizing data, a “VizTube,” could be useful because it represents a class of data functions. There really is a huge amount of work done in information visualization that goes way beyond what scientists themselves have ever done.

Most scientists do not make the connection between amazing graphics in *The New York Times* or the *Guardian* that were produced with some easy, reusable tool and something that they could do for their own data. There is thus an educational aspect here, informing scientists that there are easy-to-use, reusable tools. There is something like this in astronomy, but it is not as easy and as well-behaved technologically as, for example, what newspapers such as the *New York Times* or the *Guardian* do. The *Guardian* is actually an incredible model for graphics and visualization.

When someone takes these tools from the analytics or informatics communities that are generic and freely available, and then tailors them to a scientific use, there can be a hidden cost sometimes to make it work well. It may not be a major cost, but it does need to be covered somehow. That can be a problem, because it is not research in the eyes of the funding entities, even though it can be valuable in enabling much more research.

One of the visualization issues is related to standards. Scientists use tools to create the visualizations and other formats that can be put online. Many of the formats and tools are proprietary, and the vendors who own them are not supporting many of the open visualization

frameworks. The scientific community could benefit from being more aware of the standards for these visualization tools and who is determining them.

A similar situation exists for workflow software. The problem with workflow standards is that there are too many of them, and yet the tools of some groups do not support any of these standards. The tools are just a black box to the user. Therefore, pushing the tool vendors and pushing the scientific community to be smart purchasers can be very important.

Integration of Scientific Discovery Tools with Web 2.0

Another topic concerns the social network sites. How do we get scientists to participate in them? That is a big question that has come up in the Search for Extraterrestrial Intelligence (SETI) community. The issue concerning the encouragement of more participation and how we compete for mindshare or people's time then leads to the "app store" package. The app store can be a real opportunity to get more participation. Apple put a lot of effort into building their App Store brand to make it useful and productive for people. It may be worth thinking about it from a science perspective, especially how to build our professional brands—an NSF brand or some other trusted brand—to indicate that a tool is high quality and should be generally useful to a large swath of scientists. Also, a greater focus on public-private partnerships could be beneficial. Academics then might be able to exploit the tools developed in the private sector more efficiently with the public's money.

One discussant from a company that works a lot with the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) noted that he has been thinking about democratizing access to data and analysis for more than 15 years. The notions of citizen science and public data collection are still seen as fringe activities and not in the center of scientific research. The open-source community likes to say, "If you put enough eyeballs on a problem, all bugs become shallow." The same situation is potentially true of scientific discovery.

For example, a few years ago, there was a well-publicized case involving ordinary citizens who discovered some unique geological formations just by browsing the satellite imagery available through Google Maps or Google Earth. These formations had been there in plain sight, but no one had found them, even though relatively few of them exist, so it was big news.

On a much larger scale, NASA is about to launch the NPOESS Preparatory Project (NPP) satellite.¹⁶ It was never intended to be an operational satellite for applications such as weather forecasting, so a near-real-time, data-access stream is still being discussed. There is an opportunity for direct broadcast systems to share those data for near-real-time use.

Evaluation Techniques

Another methodological issue is related to how do people know whether what they are doing is successful or an improvement? The Web metrics that Dr. Hey mentioned was just one example of a different kind of metric that now exists. Dr. Stodden had a slide in the discussion yesterday about peer review and how we evaluate computational research results that are published and the different modes of doing that.

¹⁶ National Polar-orbiting Operational Environmental Satellite System (NPOESS).

There are techniques for evaluating the literature that are outdated and do not take into consideration the evolving Web paradigm. They are based on the old print paradigm. That is one example of evaluation, and virtually everyone in the academic sector works toward those evaluation criteria, such things as the ISI Citation Index. If they are not updated, those types of evaluation can retard the progress of science and not advance it.

New tools, criteria, metrics, and indicators of progress of what is valuable and of what these new scientific technologies and processes enable could be useful. In particular, there are very few useful indicators for data management and use. Tenure is not based on data work; it is based on outdated publication indicators. It might be useful to discuss the kinds of research that could help change that culture, and the metrics and the evaluation tools, to get a better understanding of what is actually happening rather than what happened a few decades ago.

Barriers to Automated Scientific Knowledge Discovery in Open Networked Environments

Introductory Summary and the Barriers Discussed on Day One

Alberto Pepe

-Harvard—Smithsonian Center for Astrophysics-

This is a summary of the barriers that were repeatedly mentioned in the presentations yesterday. One of the first issues that arose from the discussion was that scientific knowledge discovery efforts exist at the national and international levels, such as the virtual observatory in astronomy, but without any proper coordination mechanisms. Also, ad hoc data standards and protocols are sometimes developed and adopted at the national or organizational levels.

The second barrier that was mentioned yesterday was the issue of interoperability. Infrastructure solutions to knowledge discovery often are not interoperable. Some people referred to the need for a Google-like search and data management service for science, which is something we do not have. Some people mentioned different efforts in the virtual observatory that failed to interoperate in a similar way. The different systems that exist interoperate to a certain point, but usually not across different disciplines.

Funding was another issue that was repeatedly mentioned throughout the presentations. Such things as data hosting, data management, data mining, and cloud computing can be expensive. What are publishable artifacts? What should end up in journals? Some people referred to the importance of considering data, software, algorithms, and other supplemental materials as publishable items.

Regarding proprietary data issues, discussants addressed locked and unlocked data. Some people referred to the rich genomic data that are held by pharmaceutical companies as an example of locked data. Unlocked data include much of the data in astronomy. It was suggested that even with locked data, the owners of such data might want to allow data annotation, because it can add value to the data and, in most cases, makes them reusable. If data are not annotated, they may not be usable at all.

Another topic that was discussed referred to the multiplicity of efforts, solutions, and tools that have been developed for scientific knowledge discovery and data management. We have seen and discussed many of these efforts, and here again the interpretability issue arises. In some cases the various tools are connected to one another, but in other cases they are not connected in a way that makes sense and that allows the tools to communicate.

One topic that was repeatedly highlighted in the presentations was the need for a standardized and widely adopted mechanism to cite and reference scientific data. Although there are some efforts to standardize data citation techniques and mechanisms, there is not a widely adopted standard. For example, many scientists cite data online or just present links to data in the footnotes.

There was also a discussion of sociocultural barriers. Specifically, one of the questions raised was related to whether researchers or social scientists who are interested in scientific knowledge discovery fit in the academic or the industry job market. Another question was related to how they get recognition for their work.

This topic is connected to the next point that was raised, which is the issue of rewards. On many occasions, discussants talked about the changes related to the shape and role of the academic paper and that the reward and recognition systems, including such things as citation and authorship, are also changing. These functions are likely to change even more in the next generation. Today there are not enough incentives to change to the direction of open scholarship and open data. In a way, as mentioned yesterday, such openness conflicts with the culture of independence and competition that we traditionally have seen in scholarship.

A barrier that was mentioned throughout the discussion was related to the difficulty to integrate computer science topics or ideas into the curriculum of the various research disciplines. The refocusing of scientific disciplines and their curricula around computer science and data issues enables not only the solving of problems much more quickly and in an automated fashion but also the asking of novel questions.

Many presentations pointed out that big datasets already exist on the Web and that science is not the only enterprise that produces large datasets. Enterprises such as Facebook and Twitter deal with large amounts of data and other types of large-scale user-generated content, and they are very well funded.

Another barrier that was raised is the simple fact that data searches are hard to do. Indexing, searching, and retrieving scientific data can be difficult. Similarly, extracting knowledge from scientific data in publications can be problematic. For example, extracting hypotheses and results from scholarly articles is very hard to do. It can be accomplished using annotation systems, but usually not in an automated fashion.

The theme of diversity recurred throughout the presentations. Some presenters noted that computer-mediated scientific knowledge discovery has significant differences among scientific communities, disciplines, and institutions.

The issue of disciplinary versus institutional repositories arose several times during our discussions as well. Should there be data and literature repositories at the disciplinary level or at the institutional level? Whatever we choose to do, interoperability of the two platforms is important.

Physical barriers and the fact that distance matters were also discussed. Many large cyberinfrastructure initiatives assume that collaboration can just happen using computer-supported technologies and that scientists can perform work remotely. What we have seen is that most of the time scientific output and trust improves with face-to-face interaction.

The issue of temporal barriers, the idea that software and metadata requirements change with time, was also highlighted. What we collect today may not be what you want tomorrow.

Towards the end of the day, we learned from the lawyers that the law does not always help. The laws that deal with data are not a clearly defined issue. It was also mentioned that with the current legal system, it is difficult to achieve simplicity. The system is designed for securing property, not for sharing knowledge.

Finally, the last point that was discussed was related to the idea that scientific enterprises become progressively more computational, and therefore data and code sharing at the time of publication becomes crucial to the reproducibility of scientific results.

Summary of the Discussion

Certain problems may be common to all fields. We ingest data, manage large amounts of data, organize, analyze, visualize, and preserve them, and therefore the data life cycle is something that might be examined more closely. What many have said is that for us to be able to manage all this data, we ought to lower the barriers to be able to ingest, manage, and visualize the data.

Technical Infrastructure

A point about the need for Google-like search in management service for scientific data was raised during the first day of the workshop. There are a number of such mechanisms. One of them is for astronomy, as in the case of the Astrophysics Data System (ADS), which was discussed during the first day of the workshop and which has complete user adoption. All astronomers use electronic publications, but they do not use the internet to access data. There are many interesting techniques, but we could benefit from having more of them and they could be more broadly available.

Some of these mechanisms allow us to find where data are, but they do not allow us to get the data. Many of them are not at the point of allowing us to have the tools to use the data, so there are plenty of improvements that can be made.

The worldwidescience.org and science.gov portals also were mentioned yesterday. These are tools for deep Web searching that can be better than searching Google for science. They have fairly good use statistics, although not like Google, of course.

Another problem is that as more researchers can generate a lot of data, they cannot visualize them properly or even analyze them. It is becoming extremely difficult. As Mr. Dudley said earlier, he has a group of 20 people that have 500 computers in a room, but that is not typical. That can be a big problem, because scientists can generate huge amounts of data, but the data may just sit there or are lost.

Institutional Factors

The greatest barriers can be based on the social or “soft” infrastructure, not about the technical aspects. The technology moves very rapidly, and the solutions, even if they are not straightforward, are worked on constantly. Institutional structures, models for communication from a human infrastructure standpoint, the legal aspects, and the funding mechanisms are the things that tend to change very slowly. Also, they do not adapt very quickly to new technology changes and always lag behind technological progress.

We are always catching up from an institutional standpoint. The point was made yesterday that we are still mostly stuck in the print paradigm on scholarly communication, even though we have moved wholesale onto the Web. It would be useful to rethink how to organize scholarly communication and how to be more responsive and adaptive to the technological opportunities in a way that will promote greater productivity of the technology that is available.

In this particular activity, we should look to the generation that is called the digital natives. If you know any staffers in Congress, you know that many of them are young, and

senior members of Congress tend to trust these younger people. One of the barriers that could be overcome is the standard structural barrier related to older people making decisions when younger people should be included more in the conversation.

Also, one of the most common explanations given when you ask somebody about barriers is the difficulty of doing the job. Therefore, one issue is getting tools that will make a given task easier. On the soft barrier side, people who have the computing skills to do some of the data work find it difficult to be rewarded in the scientific discipline in which they are working, but also find it difficult to get credit in computer science if they are perceived as working as an assistant to somebody in a scientific area. The same problem exists in statistics. Therefore, another issue is the unwillingness of universities to reward people for activities that seem to be helping someone in a competing department rather than in their own department.

Another institutional barrier is related to legal issues. A large midwestern university, which shall go unnamed, decided to review its copyright policies with respect to faculty. The Statute of Anne, which many people believe to be the first copyright act, was enacted more than 300 years ago, and we still have not figured out exactly what copyright is. At that university, the faculty were upset with a decision by the Technology Transfer Office (TTO), which had been given plenary jurisdiction by a Regental bylaw over all issues related to intellectual property, including copyright. The TTO's position, in effect, held that university researchers may obtain copyright to their works, because historically they have received the copyright, but that the university administration would decide after the fact whether the researchers used extraordinary university resources. If the researchers had done so, then they had to share the proceeds with the university. The university faculty thought this rule was unfair and lobbied to get it changed. This evolved into a standoff between the TTO and the faculty that lasted for 2 years before it was resolved in favor of traditional copyright privileges of faculty.

This is an example of a routine step-by-step process that occurs in some universities regarding arcane problems that we would think have been settled by now. When a university's budget is being cut by tens of millions of dollars a year, this does not sound like a very interesting topic to discuss at the regent level. A separate topic is the complex set of issues concerning the sharing of human subjects data, which has not been looked at nearly enough.

One of the barriers is with the geographic information systems (GIS) mapping. If the data are unavailable for the research, scientists can be disadvantaged in the GIS market, which is one of the markets most in demand besides Twitter and Facebook. A key question is whether the GIS being used is proprietary or open source. If we cannot get to the data, then we cannot do anything. We cannot create any maps without the data. There are no standards to release the data. If the data have not been released, the scientific results suffer as well. For example, data.gov only tells us what is there, but it does not make the data interoperable.

Data Interoperability

In the ideal world, it would be wonderful if everybody could reach everybody else's data. Simply being able to find the data—even if it only gets users to a human-readable page that tells them who to contact to get the actual data—would still be a major breakthrough. Unfortunately we are not there yet. Data.gov is a great example. Every dataset is available if it is in one of the catalogues, but there are other features available to find out more about the data.

GIS is a very important area, but the standards do not exist to interoperate. For example, the guidelines for putting boundaries on the map are not the same as the ones for putting items on the map, which, in turn, are not the same for reporting where things are. Solving that is a complex problem in which the scientific community should participate, but does not have to lead. Finding out if someone has done a study and has the ground truth pixel map for some region or that someone has the scientific data about the weather patterns can be difficult between different discipline communities.

The geosciences community is advanced in the processes required for integrating and interoperating. It is important to separate those two problems and to recognize that the first barrier is finding the data. A second barrier is obtaining them. Then there is the technical integration of the data. An important issue now being addressed is the soft infrastructure for finding it, for knowing who to talk to and what is in a database before investing the money rather than discovering afterward that it does not have what you want.

The Hubble Space Telescope Archive is fairly large, but people do not have a problem finding Hubble data. They know where to go and it is relatively easy. The problem is to serve the data, because the difference between the data that we have and the manner that we serve is gigantic. We serve “N” times the data we have in our archives, where N is a very large number. That number is not sustainable when we get to petascale data, and there is no infrastructure that we can develop that can support these kinds of requests. Thus, the help that we need is in figuring out how to stop this from happening. We do not want to stop people from doing the research, but we cannot serve petabytes of data, because then we are going to become a service that just gives out data and never do anything else, because that would take much of our resources. This is a very fundamental problem: After scientists collect the data, how do they allow others to apply data mining to the data? How do they allow people to apply algorithms to the data without moving petabytes?

We have separated the public outreach portion into a completely separate network from the research one. We know those numbers exactly, and they are very large. The issue is also that new instruments are going to come online and produce a huge amount of data. Even if a scientist wants data for a small portion of the sky, the data are still considerably large. Many people want the same area of the sky, however. Hubble is a flagship mission, and it is possible to do incredible science with it, but how the Hubble archive can sustain the data that scientists want served directly to them is not clear.

Interoperability also is a very important issue for soft and technical infrastructure. Can we imagine a startup today, such as Twitter, trying to build its own cloud without being interoperable, meaning not having an application programming interface (API) with which data can be mined? That cannot happen if someone wants to be successful as a company, but it happens every day in science, because the barrier is so high. How do scientists publish the data? They do not have access to some data programmatically, so that is a very high barrier that really hinders interoperability.

Interoperability is a deeper issue. It is not just about discovery and search. It is the ability to take somebody else’s data and mix them with your own data, which is what many scientists want to do, rather than work with someone else’s data alone. The barriers to that are technological, semantic, and legal. If licenses do not permit scientists to interoperate and mix

the data, they cannot do that. Thus, there are many different barriers that go along with interoperability. Interoperability is being conflated with discovery.

Reward Structures

The issues of a reward structure, workforce development, and the computer science curriculum can be linked. The university department heads that are enlightened and have figured out that they need to add particular courses to their curricula may still forget the data science component. There are university departments that include the data science component, but not that many. Informatics courses include things like data mining, visualization, statistics, and data management. This leads into the issue of workforce development. If we are training the next generation to take over the data-intensive science that we are creating today, then there ought to be courses and curricula that reflect that.

This in turn goes to the issue of the reward structure. For example, one discussant spent 20 years in the National Aeronautics and Space Administration (NASA) system and then went to teach at a university. He reached tenure in his late fifties. It turned out to be harder than he thought, despite having had a successful career and having published. There was a lot of resistance in his college of science—although not in his own department—to astro-informatics, or even just to data science, as a research discipline. Why is that? One reason is that the university administrators do not understand it. A second reason is that most of the places where he can publish are not the typical astronomy journals. They are peer-reviewed conference proceedings. In the physical sciences, unlike in computer sciences, peer-reviewed conference proceedings have a much lower acceptance than traditional science journals. The impact numbers for astronomy journals are in the high double digits, whereas peer-reviewed conference proceedings are in the very low double digits. He had success because he spoke up. He is a senior scientist now, so he can argue with deans and tenured full professors. Younger people do not, so this is definitely a barrier.

One possible solution would be to have some kind of mentorship or advocacy for young scientists who are doing this kind of work and who cannot themselves counter the arguments made by senior professors and deans who say that what they are doing is not research. Younger scientists, unless they are very secure, will not be able to react this way. The younger generations need advocates and mentors to help in this process as they are trained in data-intensive science—or data-oriented science. Statisticians may say that the data they are working with are not large data; they are complex data. It is not so much the size of the datasets, but the orientation of the research, that is different.

Focus of Funding Mechanisms and Science Policy

Another barrier is that the funding mechanisms for science in the United States are set up primarily along disciplinary areas. Even within the larger funding agencies like the National Science Foundation (NSF), it is separated out. Crosscuts for infrastructure and for interdisciplinary work are mainly done by each of the agencies trying to do it themselves. There is very little crosscut infrastructure work between the agencies, although there are notable exceptions. Generally it is difficult to get funding for the crosscuts. The National Academies have written a number of recommendations about the funding of interdisciplinary

work, but these have been almost completely ignored. Both the technical infrastructure and especially the soft infrastructure are absolutely crucial for data work.

Over the past few years there has been a lot of emphasis within the federal government on scientific data policies and plans. For example, the new NSF-issued policy guidance addresses issues related to data management plans. Many research agencies are moving aggressively forward on data policies. The lack of an agency-wide data policy was considered a barrier that set the stage for managing data as an agency resource. Not every agency is doing this, because they are all somewhat different, but among the leaders are NASA and the National Oceanic and Atmospheric Administration (). Also, the Environmental Protection Agency is moving quickly now to develop a complete agency data policy.

In this regard, the report *Harnessing the power of digital data for science and society*¹⁷ discussed data policies and data plans. The Office of Science and Technology Policy's Interagency Working Group on Digital Data is continuing the work on this. One thing that did not become a final recommendation in that report, because it was not acceptable across all agencies, was to have a chief data officer in federal agencies. Data are an asset that needs to be managed, and thus, somebody who understands science and data may be needed at a senior level. There was a follow-up workshop that reemphasized all of these concepts, including the need for a chief data officer. Consequently that is another challenge within the federal government. Is it also a barrier in universities?

To follow up on this point, the incentives of technology transfer offices for data and code dissemination are somewhat orthogonal to those of most research scientists. The primary interest of TTOs is in licensing the technology or working with it in a commercial sense. It is worth recognizing that part of the question that is raised about legal barriers to sharing, at least in the context of this meeting, is a distortion of the traditional incentives for university scientists. That can be a very real barrier at the institutional level to data dissemination.

Following up on the comment about TTOs, this certainly has been a problem for a long time. A counterforce has been emerging in the universities through the interests in institutional repositories, in open scholarly communication, and in other similar initiatives. We need to be careful where to ask questions in the university, depending on the answer we want to get.

Demonstrating Work Impact and Value

One of the biggest barriers within the scientific documentation community is how to prove that you have had an impact. How do you prove that the resources that you use to run an office of scientific and technical information or a defense technical information center really have an impact? There is a lack of metrics, so we are forced to revert to anecdotes, to cases that prove the point or at least incite the imagination. We tried to do that in the *Harnessing* report that was mentioned earlier, and we got nice vignettes, but they are all philosophical: "This will change science" was the assertion, but there was none that actually showed how it did change science. That is one of the biggest barriers, just proving the point.

The anecdote problem is very real. It has been surprisingly hard to get more than a small handful of now overly retold good stories. It seems as if there are two models. One is a

¹⁷ Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Available at http://www.nitrd.gov/about/harnessing_power_web.pdf.

model of data contribution toward a collective resource that everybody uses, like GenBank or the Protein Data Bank, with many people pooling into a common resource to create peer-to-peer data sharing and reuse. That seems to be much more amenable to measuring levels of reuse and impact and maybe even to measuring contribution. For example, if someone says, “I have sequenced four organisms that are in four species that are in GenBank,” it makes a statement that is much more quantifiable than saying, “I collected a lot of ecological data and shared it with three colleagues who used it in their studies.”

That is the barrier to the reward structure. There is the reward structure, and then there is the business case, the value and the impact. Regarding evaluation, there was a meeting in January 2011 organized by the Federation of Earth Science Information Partners (ESIP). The people involved in this organization talk a lot about interoperability and discoverability of data. At this meeting, one of the guest speakers was Ann Doucette, the director of the Evaluators’ Institute at George Washington University. One of the takeaway points that she raised is that if we want to develop metrics that prove the value of an initiative, we need to invest in that up-front. She said we should not wait until the end of the performance period before we ask the institute to get involved. She said if they get involved very early in the research initiative, her group can help do some baseline assessments, so at the end of an initiative that lasts 5 or 6 years, they will be able to develop specific metrics that people can use to prove how the investments have been beneficial.

Range of Options for Further Research

Introductory Summary of the Range of Options Discussed on Day One

Puneet Kishor
-University of Wisconsin-

This is a summary of the options for future research that were discussed yesterday. The options may also be seen as opportunities.

Optimizing science for the community was one option. This is a meta-level issue, so it would be useful to distill it to specifics. Other issues included analyzing knowledge as an ecosystem, rich authoring (integrating data with the literature), and semantic data storage. Building a precompetitive digital commons seems possible.

Then there was comprehensive monitoring of many biomedical traits at the population level. Breaking the siloed principal investigator mentality is important. That is an institutional issue; the funding agencies can enable that.

Another option was to have some sort of code versioning for data. Other options included tools for adaptive modeling, lightweight integration instead of full-scale ontology, and incentives for those who collect data. This last issue came up repeatedly. How do you reward people for collecting data?

Discussants also mentioned the development of decision-support tools, simple metadata and better semantics for data, and making visualization a first-class citizen of science. Quick and easy visualization would be very useful, and it does not have to be complicated. For example, there are many fun things that could be in a YouTube for science.

Simple metadata that can be searched and localized in different languages could be helpful. One discussant noted that we can change culture by changing practices. This is important, because the culture of different scientific disciplines cannot change by itself.

Incentives to produce results versus a mandate to share data were mentioned as future options: There now are incentives to produce results, but few incentives to share data. There are some mandates, however, to share data.

Finally, there were four licensing options for academic data: (1) Do not put any label on your data, in which case, the default rule protects the copyrightable portions of a database; (2) put all of the data in the public domain with a waiver of rights; (3) use an orderly set of well-recognized licenses, such as the Creative Commons family of licenses; or (4) have a complete free-for-all, a babble of licenses. These are the four conceptual buckets for making data available with or without a license.

Summary of the Discussion of the Range of Options for Further Research

Fostering Innovative Approaches

Based on this discussion, there seems to be an important feature of the data-intensive phenomenon. This area is moving relatively quickly and in very interesting ways, and much of that is being done outside of the traditional channels of academia. For example, people are taking data and information and putting them on Flickr or similar tools, even if they are misconstruing Flickr as an archive rather than as a dissemination mechanism. Therefore, there is much that falls outside of the normal boundaries of science and scholarship in the university context.

The younger generations of researchers will set the standards. For example, Facebook was a major global activity in only 5 or 6 years. It may be a young researcher who understands why all this matters and who will start something that is going to spread through the scientific community. In this budget environment, the top-down approach is not effective.

One example is the SuperHappyDevHouse located in the San Francisco Bay Area. It is a huge mansion in Cupertino, California, in which many hackers live. They open it to the public every month or so, and whoever wants to can go and hack all night. At least one very successful company started from SuperHappyDevHouse. There is a similar idea in Mountain View, California, called the Hacker Dojo. For a relatively small amount of money people have rented a huge abandoned warehouse, and the Hacker Dojo is open to anyone who wants to come in and start a company, for example. There are biohackers in there now starting a lab with some old polymerase chain reaction machines they bought on eBay. That is one interesting bottom-up approach.

We are in a new networked environment and in the middle of a change that we cannot completely understand. We can look at other parts of the networked environment to see what has happened there, and this creates an opening to examine the bottom-up approach, which has resulted in many of the successful developments identified here and elsewhere.

A scientist who identifies a problem looks for a solution. Young scientists have new approaches. What do they do with their images, for instance? They deposit their images into Flickr, so Flickr suddenly becomes a scientific database. Some hackers are also biologists. They identify data and they hack on them, and suddenly they are advancing drug discovery in ways that never would have been expected. Many scientists might now recognize themselves in these examples. Scientists have always collaborated in some way, but with the new digital, networked capabilities combining with the social reality, science is becoming even more collaborative.

In the first day of this workshop, there was an interesting discussion about the sociological dimensions to these issues. There might be a new vector of development that happens through the research libraries community. There might be another vector that occurs through the vice presidents for research at universities. There may be other channels in which people respond with more immediate action. It would be interesting to examine what those channels might be.

Identifying, recognizing, rewarding, and supporting the entrepreneurs in this area is another potential approach. It could be useful to compare some of the entrepreneurial efforts—

wherever they fall on the generational spectrum, whether it is people coming through the research pipeline at the beginning of their career or people who are more senior and advanced in their careers—to better established models of research science so that the efforts are not characterized as so dramatically new and transgressive. This would both serve a rhetorical purpose, so that these developments do not come across as threatening to the various audiences, and show that what is going on in these instances does echo some very traditional understandings of collaboration in the sciences. Many scientists have valued the sharing of their efforts and knowledge. Making those values explicit would be a way of being authentic to what the progression is, as well as being politically sensitive to the audiences.

Encouraging Good Network Uses

One interesting thing about the modeling of how the networked entrepreneurial versions tie back to more traditional understandings is that there is a lack of well-defined research on how these collaborative or commons mechanisms work in the networked environment. That is, there is a great deal of descriptive, qualitative, or anecdotal research on the character of science, but there is not much well-modeled, empirical research about the role of the research infrastructure and networked platforms, and how sociological, economic, and legal factors are related.

It is not necessarily an either-or issue, however. There could be multiple messages going out to different communities and mobilizing those communities in positive ways. There are some arguments that are essentially inevitability arguments—that something is going to happen regardless of what the community wants. Because if it is inevitable, we can argue that progress should be made quickly, but if it is inevitable, what is the rush? Thus, if this change is happening irrespective of what people think or prefer, then one option is to get ready to take advantage of it.

On the issue of the bottom-up and top-down approaches, it is worth comparing the United States and the Chinese management methods. There is a real difference in how the two countries approach research and policy formation. Most innovators here—and also largely in Europe—work from the bottom up. The system supports risk-taking and failure, especially in the private sector, and so there is a tremendous amount of creativity and initiative that happens at the working level. Then finally, the policy-making community catches on, and ideally it institutionalizes approaches that are productive. The Chinese model is very top down, and the potential innovators at the bottom generally will not take initiatives that are not approved.

Although the bottom-up approach has produced a great deal of creativity, it has many inefficiencies because there are many dead ends, it is uncoordinated, and it does not always work. Each system has its good aspects, but neither performs optimally. The bottom-up issue is important for this discussion, because virtually all the examples and everything that has been discussed has happened at the initiative of individuals and communities.

From the government top-down perspective, it is worth noting that significant change is occurring. Consider the tenor of the reports, or look at the programs that now include scientific data. The focus used to be on the grid until that first Atkins report came out.¹⁸

¹⁸ *Revolutionizing Science and Engineering Through Cyberinfrastructure*: Report of the National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure. January 2003. Available at <http://www.nsf.gov/od/oci/reports/atkins.pdf>.

One approach to network citizenship from the top down is to emphasize the “network” and articulate the values that are implicit in the idea of the network as the hub of that conversation. A related, but distinct, approach is to put more emphasis on the idea of values and citizenship as part of the scientific research community. That is, how do the traditional values in that domain get expressed in this new technological environment?

The theme of the good network citizen is very interesting. One discussant noted that he works with a project called VIVO (see <http://vivo.ufl.edu>), which uses semantic faculty profiles for “enabling national networking of scientists.” VIVO has been trying to surpass that 80 percent completion rate that is typical for LinkedIn members who get their profiles up to date. VIVO is aiming to get an automatic 80 percent from administrative data and from bibliographic resources, but getting beyond that 80 percent rate has been difficult.

One of the things VIVO has done is to use information in grants. They have been able to do that with some of their funding because the National Institutes of Health (NIH) and the NIH National Center for Research Resources were amenable to supporting that. VIVO has six projects that will have some interesting results with short turnaround times.

The 1970s film *Network* has the classic line “I’m mad as hell, and I’m not going to take this anymore” by a broadcast journalist working in the television networks. That was superseded fairly soon by the current network, which is entirely different. The discussion here has made it clear that the emerging network is an incredibly enabling, individually empowering, and self-organizing system. It is the kind of technology that supports bottom-up research and business, but also enables a Google or a Facebook to arise and become a multibillion-dollar enterprise. The new network’s versatility is potentially available for all the science examples described in the past day. The characteristics of the internet are important, because this environment can completely change the way research and data sharing are done. It is also important to identify the type of research that is needed to get a better understanding of how to make better use of the emerging network for science and applications.

Issues in Reproducing or Reusing Data

Facebook and Twitter open up the science of social networks in a tremendous way. For example, consider data mining capabilities: When you have a huge portion of the planet connected like that, there is a big opportunity for research, just to learn from data mining.

Existing social network firms certainly lend themselves to statistical analysis and research problems involving large datasets, but there is a lot of informatics work that could be done by simple tools that still needs to be funded and developed. As an example, one discussant worked for a confocal microscopy lab. The images from a confocal microscope are not flat, two-dimensional images that you can put onto Flickr. They are three-dimensional. They have multiple channels for different wavelengths of light. In short, they are multidimensional images that a scientist needs to be able to visualize in three dimensions, turn around in a computer, and isolate the different channels of light. Flickr will not do that, so special tools are necessary to make those images useful.

Furthermore, there are about 150-plus different file formats for confocal microscopy, and it is a huge problem, because most of those come bundled with the microscope as part of some proprietary software. The data become unusable after a couple of years, because nobody is using that software or that microscope any longer. That is a good example of a failure of the

bottom-up approach, because the demand for a better file format has not come from the scientists who use those microscopes. They might wish that it would happen, but when they purchase the equipment, they are purchasing a lens first, and a file format is not high on the list. The researchers have not yet come together to demand that the vendors provide a better file format.

Another discussant had a different experience, however. There was a PDF from high-dimensional datasets that she coauthored and published in *Nature*. Part of the point of doing that was to be able to take the file formats, of which there are many proprietary ones and have a format in which people could publish the data, not necessarily in their original richness, but in a way that somebody could reuse them. There are many people in the vanguard of this community who think the PDF format is an obsolescent format. However, this discussant and her coauthors were able to do that because she was at one of the hacker places in Silicon Valley on a trip, and learned that it was possible to easily make three-dimensional (3-D) PDFs. Peter Jackson employed so many 3-D engineers in New Zealand that some formed a company called Right Hemisphere and then licensed its technology to Adobe. Adobe had put a 3-D feature in its free Reader that few people knew about. The authors made a deal with *Nature* that if they got their paper accepted by that publication, the paper would have 3-D content that people would be able to use.

The way that Right Hemisphere normally makes money is with car companies and other companies that use computer-aided design in their pipeline. They set up million-dollar customized pipelines for a community that has 40 or 50 different file formats and wants something in a standard format—a job, a tool, a PDF, a PowerPoint—to come out the other end in such a way that their engineers and their marketing people can just press a button. *Nature* thought it would do this on a trial basis and then get a million dollars or so to buy one of these tools for the magazine. The only problem was that they never got the million dollars to set this up. The discussant and her colleagues are now about to publish the first one in the *Astrophysical Journal*. It may take another 6 years until many more researchers do it.

The idea that people do not care about the file formats they are using or that they are not interested in having their data reproduced or reused varies by discipline and application. The kind of investment someone would make would be to find some smart people who could develop a streamlined way to deploy this million-dollar system on the Web so that others could instantly turn those 40 different confocal microscopy formats into a PDF that they could just send to *Nature*. That is the kind of thing that would enable science—not the actual creation of the content of the article, but the pipeline for making that a whole new kind of article.

There may be one group that could be a very receptive community of end users. There are many resource managers—for example, people in state governments in various areas—who tend to be very appreciative when someone from the scientific community talks to them about what their needs are. It is so important to get the research results to the end user. It is not just putting the data out into the world; it is the data usability that is so important. In many cases, simply having access to the data does them no good at all. They do not know what to do with the data until the scientists and technicians provide the tools. It is important to think not just in terms of publishing research papers and data but also how those results can be used.

Reviewing Computational Research Results

One point Dr. Stodden made on the first day was that the thing people hate to do more than anything else is to review someone else's software. It therefore is unclear how many people would review software that someone else has written. There may be some ways to do that, however. If a piece of software has 500 or 1,000 users, that is a fairly good indication that it is a usable and useful program, and the review could then be much less difficult.

This issue comes up frequently in the discussions about reproducibility and other aspects of working science. A code review as part of the peer-review process before publication is difficult, because it is an enormous amount of work that people do not feel equipped to do. There are middle-ground approaches that might be able to be implemented, however.

Nature published two articles in October 2010 on software in science that discussed how it is used and often broken, that it is generally not reviewed, that it should be open, and so on. Mark Gerstein, a bioinformatics professor at Yale, and Victoria Stodden both wrote a letter to *Nature* commenting on this and saying that the scientific community needed to move toward code review. They suggested a broader adoption of what some journals have done, which is to have an associate editor for reproducibility who will look at the code and try to reproduce it. Then it would not be such a burden that is imposed on reviewers. That is another possible way forward, having code submitted, made open, and then incorporating this aspect of review. *Nature* did not want to publish that letter, but it is on Dr. Stodden's blog, and the idea is being discussed further.

Another discussant noted that she has reviewed code plus data for journal publication and an example of the kind of problems that come up is "the compiler did not run because they changed something in the most recent version of Unix." Every now and then you find real code errors. There are a number of domains that do this as a regular task. A list of approved software is not likely, but different audiences and different kinds of follow-up are definitely worthwhile.

Incentives versus Mandates

Of the 10 stories that *Wired* magazine listed as the biggest science breakthroughs of last year, 5 seemed to be based on the use of databases—2 concerning GenBank and 3 concerning astronomy. The year before, there were three GenBank successes identified, but no astronomy and no earth science breakthroughs.

There are two fundamental human motivations: greed and fear. Greed is working, but very slowly. This carrot–stick dichotomy indicates that it is still something to be sold to researchers rather than something that is a so-called "killer app." Should we be appealing to fear, telling researchers that the real problem with not having data is that somebody in India or China will be running rings around them? Perhaps the most effective approach is to pick one message, whether it be enforce the data management mandates, support tool building, or support sociological analysis of what scientists do with data, and describe why just one of these messages should be used and focus on it, because 15 seconds of attention is all that we are likely to get from the research community and the decision makers.

Much of the successful work happens through bottom-up approaches, and this is a lesson we should learn. Identifying and rewarding the people who are already doing this is a

strong option. Focusing on what can be done from the bottom-up to acknowledge good work and to help keep researchers doing that work can get them more visibility across the sciences and other benefits. The Web can be a very forgiving medium in the sense that if something wrong is fixed, it is okay as long as it is fixed fast. The scientific community generally does not have the kind of culture that gets things moving quickly, however. We do a lot of design and architecture. There is nothing wrong with that, but as this area evolves from the top down, the bottom-up efforts can be rewarded as well.

Journals also can help enforce reporting guidelines if there are standard metadata on which you can get a community to agree. In this case, journals can tell authors that anybody who submits a certain kind of experiment has to include certain information about their experiment. They are effective in applying pressure to authors, because publishing is one of their motivating factors.

Regarding prizes, a couple of years ago something like Kiva¹⁹ was proposed for science—a type of micro-loan that researchers could apply for, say, \$50,000 from the National Science Foundation (NSF), to try something new. Perhaps it would be \$100,000 or \$200,000, depending on the scope. To determine how scientists should meet the data management plan requirements, we could have a contest for a couple of years where people could try different approaches.

If a technology has the right branding, then other people will discover it more easily. Referring to the “app store” example, the app store has applications that get recommended by Apple, and suddenly millions of people download them. Another option is for NSF or some other entity to have a competition with small grants to find approaches that seem to work as widely applicable solutions. The prize would be that they get a brand that says something like “NSF Approved.”

Another similar suggestion focused on some sort of sandbox, where people can experiment with different approaches. The people who would take on these kinds of projects would not take on all the ones that were listed earlier in this summary, but they would have various missions. Some individual investigators might accept a 5 percent tax so that technologies for data management systems could be developed.

There is already a group of people who are reusing data and making good discoveries. We have seen some policies started, such as the call for data management plans, which are helping to get people who produce data to think about managing and sharing them. One of the best incremental steps on the production side right now, other than letting some of those mandates play out and trying to extend them to other funding agencies, will be to continue to encourage people to contribute to collective databases. There are many mechanisms for doing that, such as journals requiring registry numbers. We actually know a good deal about how to get collective databases implemented successfully.

Another option that can be pursued is to enable innovators who are doing productive work. This can be done in different ways. One option would be to highlight the researchers

¹⁹ Kiva is a nonprofit organization with a mission to connect people through lending to alleviate poverty. Leveraging the Internet and a worldwide network of microfinance institutions, Kiva lets individuals lend as little as \$25 to help create opportunity around the world. See <http://www.kiva.org/about>

who are responsible for breakthroughs, as suggested earlier. Another option would be to initiate a prize for the most innovative piece of data reuse for the year—not for the people who supplied the data, but for the person who had the bright idea and went after the data and did something extraordinary with them. Another option is to work with people who are producing results by reusing other people’s data to get them more data. If the theory is that “more data wins,” let us try in some very focused areas to further enable the people who are most active.

The NSF data management plan can be seen as the first iteration of the network plan. That is, from the top-down perspective, the way to break through the discipline silos is to require people within each silo to say how they are going to use their funding to be good network citizens. In what way are they going to be a responsible steward of the research funds with the network in mind? Tell us how you are thinking about the network in your plan to use these resources. Will you make your database available so that it becomes a networked resource, and will you annotate it? The point is that if government agencies start requiring people to plan for using the networks, that could lead to a powerful shift in the way people think about it themselves.

If a checklist of the properties of a good network scientist citizen were compiled, not every scientist would check off all the boxes, because science is heterogeneous. If scientists checked a certain number of those boxes, however, then you would tend to move toward those goals. That would be a documentable approach. When a scientist submits a proposal for funding, he or she can point to a previous project and say, “I was a good network citizen based on this.” For example, one of the boxes could be about reproducibility. If a researcher’s field is not amenable to reproducibility because of the size of its datasets, it might not apply, but it would apply to someone else. There would be some basket of these properties that would determine how good a network citizen you are. That would be something measurable and doable.

The Department of Defense (DOD) has gone completely net-centric. It talks about everything in terms of being net-centric. It might be interesting to look at what the DOD is doing and see how civilian science agencies might think about net-centricity. Keep in mind, however, that the DOD is very command-and-control oriented, so it can do these things top down.

A study could be conducted to look at the impact of the NIH access policy and ask whether it should be extended to any other federal research agencies. Other agencies also could follow NIH’s lead in requiring that those who submit grant proposals include in their bibliographies not only lists of their papers but also lists of the places where that information is publicly available. When a scientist reports on results of prior research, the scientist could list where those data are publicly available. The National Library of Medicine (NLM) has developed many process metrics. These include, for example, determining the percentage of the documents produced with NIH grants that are getting deposited in PubMed Central. That number has substantially increased.

At this juncture, rather than investing heavily in making data available, maybe time is ripe to put some resources into the people who are actively trying to get the data to do some valuable scientific work. Maybe that would be a good strategy for a year or two, and maybe that would be a good thematic response.

Developing the Supporting Infrastructure

The NIH's National Center for Research Resources (subsequently abolished) and the Chinese have a strategic plan for research that includes infrastructure development and funding. Since there is a lot of talk about funding the enabling tools for research, one option could be a similar kind of mechanism in the U.S. government for general science, not just for biomedical science.

This discussion has identified a problem or barrier to the development of infrastructure that crosscuts across science domains. This could be further investigated at a higher level. In fact, some previous reports have noted that, and the President's Council of Advisors on Science and Technology (PCAST) report on the Federal Networking and Information Technology Research and Development (NITRD)²⁰ Program also looked at this.

This infrastructure is hard to develop, however, because it depends in part on a large research group that can support it. The community would have to determine how best to promote and cite that to help solve some of these problems. A top-down management approach is one option, but it could also be good to think about doing it through bottom-up scientific innovation. Whatever approach may be taken requires an advocate to make that happen, and it cannot be done very easily.

Technology Transfer Mechanisms

The role of the university technology transfer office in all of these processes remains a question. The typical instantiation is toward promotion of commercial interests and licensing. What is the relationship between the technology transfer office at the institutional level and the funders themselves? Could that relationship be a bridge toward addressing some of these unfunded areas?

If there is a way to follow up on this idea of a public arm of the technology transfer office, that could be a way to disseminate more broadly the technology, information, and datasets that have commercial potential, but also to do other things, such as resolve ownership issues. If academic datasets were shared more broadly, we would see certain patterns evolving, and those could become templates for how those issues—in a legal and a citation sense—could be sorted out. The technology transfer offices could be one venue in which to build partnerships to encourage this, at least at the institutional level.

Would there be a possibility of organizing some projects? Among the universities and other entities, could a few demonstration projects be organized to show how we see it being done? They would experience various difficulties, but in the end they could see how it can be done.

If a project cannot get funding because it is too “applied,” could the NSF have a technology transfer operation? The idea would be not to fund things that are possible to be commercialized, but rather to fund things that would be used to advance scientific research. If there were a GenBank-like division at NSF, for example, it could encourage the application of the results of research that it funds, although not necessarily all the way through to commercialization. In some cases, such as drug discovery mechanisms, it could lead to commercialization, but in other cases, such as discovering stars in the sky, probably not.

²⁰ Available at <http://www.nitrd.gov/>.

If NSF did have a technology transfer concept, how would it be different from a university's technology transfer function? The NSF does have the interdisciplinary Office of Cyberinfrastructure, which is supporting the development of new software systems, but it tends to be a much higher-level activity at the moment. One of the challenges is determining if this could be done at a specific domain level, or whether we would want to do it more generally.

In addition, a 5 percent tax on research budgets is an option that could be explored further. It is not clear how many people, particularly the program officers at funding agencies, would endorse a 5 percent tax of the funding, but they might. Maybe the agreement could be that if one agreed to pay the tax, it could come with certain benefits, such as access to some repository. That is, researchers could have the carrot and the stick at the same time.

Another model within government science agencies could be to collect technologies that might be broadly used from various directorates or projects and have them available in one place, or at least have pointers to all of them. This is already being done to some extent, but informally. Technologies are created using public research funding, and if they are of value, then they may be put on the Web to let others know about them. There is nothing very systematic, however, that gives NSF or some other agencies credit for having funded something that resulted in a technology or a tool that can be broadly used. So one question would be: Is there a mechanism within the Office of Cyberinfrastructure that could be used to make these tools or technologies available?

There are other government entities that are not explicitly funding agencies—NITRD is a good example—that have the mission of supporting information technology (IT) research and development and have the use of that IT for national priorities, including science. NITRD could be a good ally in this area. The NITRD director is very interested in data issues and data-intensive science, but the organization has no money. All it has is the right to convene groups of experts, mostly in IT hardware, but it could be beneficial for NITRD to be involved.

The recent PCAST report that was discussed above makes the point that the issue of infrastructure needs to be moved up to a level where it is not individual agencies competing with the individual scientists' budgets, but rather at a level where somebody is looking at this infrastructure as a national priority for innovation and science. There is now much more discussion at the top levels and the research agency managers are beginning to see the value of it.

It is harder to do than it sounds, however. At the University of California, for example, anything that costs more than \$500 has been considered capital equipment for many years. This is not rational, because the university has enormous amounts of space. It has junkyards of technologies that cost more than \$500, because if they were more than \$500, they have to be on the capital equipment inventory, and it costs more to dispose of them than to store them. The reason for such a ridiculously low limit is that anything that is considered to be capital equipment does not get taxed for overhead on grants. The principal investigators have wanted to keep the capital equipment threshold as low as possible, because it benefits them in the grants. It sounds preposterous, but California still has a \$500 threshold for capital equipment, and it has been this way for some 15 years.

Publicly funded research organizations are doing projects that are overlapping or similar to each other, whether in medicine or climate change or other areas. It would seem as though they could repackage those projects to demonstrate the multi-disciplinary needs and

benefits of such work. Such projects sometimes also have international partners, such as Brazil, Australia, and Europe, for example. That would allow them to demonstrate something similar on an international scale.

The Virtual Acquisition Office science advisory committee made the mistake of asking the members of the committee to come up with test projects. Some people were very opposed to limiting something like that to a small number of investigators. A regular (small) grant competition could specify that it is closer to infrastructure than research. This would have to be made extremely clear to the reviewers, because otherwise they will not understand. Grant competitions have become a grandiose vision with millions of dollars to do research. That was not what was originally intended.

Cultural Change

Another important feature of this data-intensive phenomenon is that there is a general sense that we ought to change practices and change culture so that the new and better ways are accommodated along with the existing ways of doing things.

How do we change culture by changing practices? How do you change culture? How do you change practices? It is an easy thing to say, but it is a very hard thing to do. It takes a lot of time and effort, and often a lot of pressure, because people's interests at the local level are not served by changing practices to which they have become accustomed.

You can also turn it around to make it “change practices by changing culture.” The idea that you need voluntary kinds of encouragement to stimulate open access has been found to not work very well. That is why we have had to resort to mandates, whether through the journals or legislation or funding mechanisms, to change both culture and practice.

At the same time, so much has been added to the list of mandates that accompany proposals. We have to promote minorities, help K–12 education, and so on. Of course, those who submit proposals claim that they are totally behind these mandates and support them, but frequently they do not do anything to advance these goals.

One of the things that was proposed to the NSF leadership was letting grantees pick what they are going to affect and then incorporate that in the proposal. These interest groups, however, have worked for a long time to get themselves into that mandate list. The fact that not much happens from them being in a mandate list is not as important to them as being in the mandate list itself. There is a kind of culture of incumbency just around being listed. It does not necessarily have anything to do with meaningful change.

We can push hard on this changing-practice, changing-culture approach, because it is hard to do. What can the agencies do to help change practices and cultures? Probably no amount of studies will result in changing practices. Adding something to the mandate list may not change practice either, because the mandate list is already so long that it is just an exercise.

The steps involved in influencing culture are not interchangeable, however. The way to change culture is by changing practice, and this example of NSF requiring data management plans is a good first step, although it is a very tentative step, and there is still a long way to go. Everyone now is holding workshops on how to write the correct data management plan. Maybe in 3 years there will be many more people thinking that this is a good thing to do, because they would have been made to do it and they would have started seeing the benefits of doing so. Thus, there could be some value in recommending specific things for changing practices.

One answer to this question is the old adage “you get what you measure.” If we are not measuring the impact of the data plan, we are not tracking it, we are not enforcing it, and then nobody cares. Suppose, for example, researchers were asked how many minorities they supported on their last NSF grant? If it was one-third of what was promised, then the researchers could get one-third of the requested money on the new grant. You would suddenly have many more people matching those mandates.

It is easy to talk about measurement and enforcement. It is harder to make that work. We may be coming to a point where we need to start thinking about what is going to work.

The NSF seems to be trapped in this situation where it is supposed to be getting ideas from the community about where to go with the science. Although there may be many people interested in this issue, the population of people doing science and being funded by NSF is much larger, so the view being expressed by those sufficiently interested in any particular issue is a minority view within that sourcing process. How do you change this? Where are the pressure points and the leverage that will work, given the limited resources available?

Another discussant made a comment about the socio-cultural attitudes that we may wish to change. Reflecting on the tradition of science, it is important to celebrate that tradition and use it as a basis to build upon. Understanding traditions is very effective in introducing change to different cultures. In working with different cultures you discover that you need to take different approaches. If, for example, we are talking about the transmission of HIV in a culture that has polygamy, it is irrelevant to talk about being faithful. Therefore, if you know the important traditions to select, you can build upon and influence that change more rapidly.

Role of Libraries

In setting up a national information-organizing center that oversees and manages standards and vocabularies, it is important to remember that when data and information are created, they may have to be maintained over long periods of time and adapted as things change. Therefore, there is a certain amount of infrastructure that has to be supported from the top down. It is not possible to get universities to do this in a distributed way.

Libraries are giving much of the advice for data management plans right now. Research libraries have seized upon this as an opportunity to do outreach to the science departments at their universities and to help them figure out how to do the data management plans that the researchers are now being required to do. Hence, research librarians are allies on the ground at universities, and they reach out to the scientists to help disseminate cultural ideas and strategies that funders and policy makers would like also to see implemented.

One of the big questions circulating in the library community currently is what to do with print collections now that we are moving much more toward digital materials. There are a number of issues that have never been encountered before. We have been building up print collections for the past several hundred years, so we know a lot about that. We do not know much about reducing the collections, however. It is hard. It is more of an ecological approach than telling others what they should do.

A problem is that it is very easy for scientists to say that issues or concerns expressed by one community have no corresponding value in their community. There are many ingrained practices, attitudes, and traditions that are candidates for change, but can we deal with them

sensibly? Research librarians may be able to provide assistance to scientists that would not otherwise be available, particularly at the college and university level.

One thing that may be helpful is delineating the roles of different kinds of experts who are involved in the data management process. The funding agencies are like the data investors. The data producers are interested in their data, but not necessarily in how the data will be reused. A problem in some cases is that research funders want data producers to let other people reuse their data, but it is extra work for the producers to deposit their data someplace and to annotate the data in a way that makes them useful. Then you have the data reusers or analyzers as well as those who are the intermediaries between that data producer and the data reuser, acting as the data translator, manager, or marketer. This latter type of person will help the data producers—who really do not want to do this kind of work or think they do not have the tools available to do it—become aware of the practices or tools that exist for them to do this kind of work, and also manage the data and train them in how to do that more efficiently. Research librarians do that kind of work in genomics, for example.

In addition to the NIH Center for Research Resources that was mentioned earlier, there is another institute at NIH that deals entirely with information science and information resources: the National Library of Medicine. There is not an equivalent library for the basic sciences; although NLM has done some work expanding its focus into other related areas, it does not cover geosciences and other disciplines.

In the Agricultural Research Service (ARS) there is the National Agricultural Library. The ARS includes librarians in research project planning and data management. Without trust among the parties, however, we cannot open up a legitimate dialogue, innovative direction, and mutual support. How to build trust really depends on those who are involved in a research project, but without trust they may fight against each other.

Finally, one of the options that was suggested for changing and improving data management practices is to reduce the costs. Better leadership can direct people to what those practices need to be. One research agency has a saying concerning data management: “Should we, could we, and will we?” There is no argument about whether data should be managed, but the scientists want to know how to do that. Clear answers can foster effective leadership in different disciplines for preserving data. Some sort of basic minimum guidelines would be helpful—if not necessarily just tools or methods. Technologies have changed what we can do, but they have not changed what we do. Appropriate leadership can help us define what we do.

Communicating and Influencing Understanding of Scientific Knowledge Discovery in Open Networked Environments

It appears that there are at least four communities that are listening and thinking about these issues. One is government leaders and elites, who believe that there is something big here and are looking for guidance on how to think about it.

Then there are the people in the institutions—librarians, research leaders, and others—who are thinking about these issues from a different perspective. For example, many state universities are now focusing on what to do with the flagship institutions of higher education among the others. The flagship schools are ostensibly the research institutions, but the distinction that some schools are for research and some are for teaching is somewhat dated. One big question here is what the flagships should do for the others. Should they do anything?

This will have an effect on access. A lot of science and scholarship can be done by secondary sources. As those secondary sources get better, the quality of scientific scholarship will improve, and it can be ever more distributed. That is a productivity issue as well as a participation issue. Therefore, this is a second audience, the combination of locally based research entities and the elites of those locales.

A third audience is researchers who are thinking seriously about the approaching change and wondering how to respond. This is important. It is not like it was 40 years ago when the more senior scientists were trained. This is different. What do we do about this? Perhaps that is a channel to approach some of the younger people, because they are coming up through the apprenticeship structure.

The final, fourth, group that is usually not addressed are the creators and innovators—the Facebook and Google types and the people who are doing research in these hacking centers. In short, good things can happen in this self-motivating way. Moreover, they do not have to ask for federal money or congressional approval to do it.

There is a big difference between just communicating versus the process of scientific knowledge discovery. Some of what is being discussed is the need for better tools and applications to do science over Facebook and other new approaches. According to some studies, data practices are driven very much by the kinds of tools they use. By talking to people, we can know immediately the kinds of data practices and the kinds of research that they are doing, just by knowing the kind of tools that they are using.

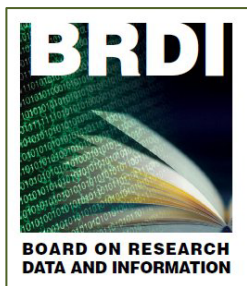
One discussant noted that he just got on a big grant with people at the Mayo Clinic that he has never met in person. He only met them through Twitter, but he is now on their grant. His best paper recommendations come through people he follows on Twitter. Many of his own paper citations can be attributed to tweeting about his work. He also gets invited to conferences through people who follow him on Twitter. This is an example of the value of scientific networking.

Another discussant noted that the topics of creating scientific discoveries and communicating knowledge are being conflated to some extent in this discussion. He uses software and writes code, but does it for a science that can never be practiced on a social network. For example, there are foresters who work with trees that take 50 or 60 years to grow. There is much science that cannot be done socially. At the same time, the communication of science or even enabling the process of science can be done with these tools. He is in user groups, and when he asks for help, people respond to that. He is able to read up quickly and does not have to go through a hierarchy to acquire that knowledge. He can utilize tools that are changeable, because they are all open-source. Therefore, separating the functions of creating knowledge from communicating knowledge is important. There may be some overlap, but they are not the same things at all.

It is helpful to discriminate between the two and to ask what we get from doing that. It is also useful to note that things that we think of as nailed down, such as mathematical proofs, are not settled until the mathematicians say so. Mathematics is essentially a social process. There are some things, such as the proof of Fermat's last theorem, that not all mathematicians agree on at this point. A preponderance of mathematicians think it has been proved, but there are some holdouts who are very strongly opposed to that position.

Scholarship is communication. We should not assume that geology is going to be done on the Web. Geology is probably going to be done with rocks and other things gathered onsite. But the phenomenon of geology—what people agree about or come to hold as geological fact—is going to be socially constructed over time, and that is a communication process. The idea that science and communication are two different things is not valid.

5. Appendix: Workshop Agenda



**The Future of Scientific Knowledge Discovery
in Open Networked Environments:
A National Workshop**
Washington, DC, March 10-11, 2011

Board on Research Data and Information
in collaboration with
Computer Science and Telecommunications Board
National Academy of Sciences

**MARCH 10 AGENDA
Day 1: Workshop**

Venable, LLP, 8 West Conference Center
Capitol Room
575 7th Street, NW, Washington, DC

I. Opening Session

Session Chair: John King, University of Michigan

- | | | |
|------|---|--|
| 8:30 | I.a Chair's Welcoming Remarks | John King, University of Michigan |
| 8:40 | I.b Opening Remarks by the Project Sponsors | Sylvia Spengler and Alan Blatecky, Nat'l Science Foundation |
| 9:00 | I.c Keynote: An Overview of the State of the Art | Tony Hey, Microsoft Research |

**II. Experiences with Developing Open Scientific
Knowledge Discovery Environments and Using Them—
Case Studies and Lessons Learned**

Session Chair: Sara Graves, University of Alabama
at Huntsville

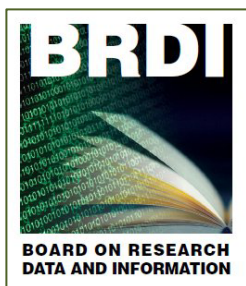
- | | | |
|------|--|--|
| 9:30 | II.a Experiences with developing open scientific knowledge discovery in research and applications | |
|------|--|--|

Case studies:

- | | | |
|---|--|--|
| - | International online astronomy research | Alberto Conti, Space Telescope Sci. Institute |
|---|--|--|

| | | |
|-------|---|---|
| | - Integrative genomic analysis: Sage Bionetworks, Seattle, WA (sagebase.org) | Stephen Friend, Sage Bionetworks |
| | - Geoinformatics: Linked Environments for Atmospheric Discovery (https://portal.leadproject.org/gridsphere/gridsphere) | Kelvin Droegemeier, University of Oklahoma |
| 10:30 | Break | |
| 10:50 | II.b Implications of the three scientific knowledge discovery case studies—the user perspective | |
| | - International online astronomy research | Alyssa Goodman, Harvard University |
| | - Integrative genomic analysis | Joel Dudley, Stanford University |
| | - Geoinformatics | Mohan Ramamurthy, Unidata, UCAR |
| 11:50 | II.c Benefits and drawbacks of the three case studies--Panel Discussion of open scientific knowledge discovery system developers and users Panelists: same 6 speakers as in II.a and II.b, plus session chair | |
| 12:30 | Lunch in Venable cafeteria, 8th floor | |
| | III. How Might Open Online Knowledge Discovery Advance the Progress of Science? | |
| 13:30 | III.a Technological factors Session Chair: Hal Abelson, MIT | |
| | - Interoperability, standards, and linked data | James Hendler, RPI |
| | - National technological needs and issues | Deborah Crawford, Drexel University |
| 14:30 | III.b Socio-cultural, institutional, and organizational factors Session Chair: Michael Lesk, Rutgers University | |
| | - Socio-cultural dimensions | Clifford Lynch, |

| | | |
|-------|---|---|
| | | Coalition for Networked Information |
| | - Institutional factors | Paul Edwards, University of Michigan |
| 15:30 | Break | |
| 15:50 | III.c Policy and legal factors Session Chair: Michael Carroll, Washington College of Law | |
| | - Legal aspects | Michael Madison, University of Pittsburgh School of Law |
| | - Policy Issues | Gregory Jackson, EDUCAUSE |
| 16:50 | III.d How can we tell? What needs to be known and studied to improve potential for success? Session Chair: Fran Berman, RPI | |
| | - Government perspective | Walter Warnick, Office of S&T Information, Department of Energy |
| | - Academic perspective | Victoria Stodden, Columbia University Law School |
| 17:50 | Concluding Remarks | John King, University of Michigan |
| 17:55 | Adjourn | |
| | Closed Session | |
| 18:15 | Reception and working dinner for committee and speakers | |



**The Future of Scientific Knowledge Discovery
in Open Networked Environments:
A National Workshop**
Washington, DC, March 10-11, 2011

Board on Research Data and Information
in collaboration with
Computer Science and Telecommunications Board
National Academy of Sciences

**MARCH 11 AGENDA
Day 2: Workshop**

**National Academy of Sciences Keck Center
Room 100
500 Fifth Street NW, Washington, DC**

I. Recap of Symposium Results and Discussion of Future Issues

Session Chair: Bonnie Carroll, Information International Associates
Rapporteurs: Puneet Kishor, University of Wisconsin, and Alberto Pepe, Harvard

8:45 **1.a Summary of Opportunities to Automated Scientific Knowledge Discovery in Open Networked Environments**

- Discussion to validate results from the symposium and identify issues in next 5-10 years

9:45 **1.b Summary of the Barriers**

- Discussion to validate results from the symposium and identify issues in next 5-10 years

10:45 **Coffee Break**

11:00 **1.c Summary of Techniques and Methods for Development and Study of Automated Scientific Knowledge Discovery**

- Discussion to validate results from the symposium and identify issues in next 5-10 years

12:00 **Lunch in the Keck cafeteria**

13:00 **II. Range of Options for Further Research**

Session Chair: John King, University of Michigan

Based on the results obtained in response to the preceding discussions, define a range of options that can be used by the sponsors of the project, as well as other similar organizations, to obtain and promote a better understanding of the computer-mediated scientific knowledge discovery processes and mechanisms for openly available data and information online across the scientific domains. The objective of defining these options is to improve the activities of the sponsors (and other similar

organizations) and the activities of researchers that they fund externally in this emerging research area.

- Discussion

15:30

End of Meeting

