Knowledge Management and Visualization Tools IN SUPPORT OF DISCOVERY

NSF Workshop Report

Editors:

Katy Börner, *Indiana University*
Luís M. A. Bettencourt, *Los Alamos National Laboratory*
Mark Gerstein, *Yale University*
Stephen M. Uzzo, *New York Hall of Science*

Workshop Webpage: http://vw.slis.indiana.edu/cdi2008

December 2009

Acknowledgements

The organizers would like to thank Maria Zemankova, Directorate for Computer & Information Science & Engineering (CISE) at the National Science Foundation for inspiring and supporting this workshop series and report. Furthermore, we owe thanks to all workshop participants. Many of them revealed their most innovative research and development projects and demonstrated systems under development to their competitors in an attempt to identify the most promising synergies. Several financed their own travel to enable participation by those which needed financial support to attend.

Weixia (Bonnie) Huang, Elisha F. Hardy, Mark A. Price, Marla Fry, Qizheng Bao, Jennifer Coffey, Tracey Theriault at Indiana University provided expertise and time during the organization of the Workshops and the compilation of the Report.

This effort was supported by the National Science Foundation under grant IIS-0750993, the Cyberinfrastructure for Network Science Center at Indiana University (http://cns.slis.indiana.edu), and the New York Hall of Science, which provided a most stimulating venue for the second workshop. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Cover Imagery

How can we gain a more holistic understanding of the structure and dynamics of the information universe that surrounds us? What tools might be most helpful to understand patterns, trends, and outliers so that we can more effectively navigate, manage, and utilize what we collectively know? So far, we use rather primitive tools that give us rather limited views—the metaphor of peeking through a keyhole is intended to capture this.

IMAGE CREDITS (TOP TO BOTTOM)

Map of Science by Kevin Boyack, Dick Klavans and W. Bradford Paley. Appeared in Nature, Vol 444, Dec. 2006.

Society for Neuroscience, 2006 Visual Browser by Bruce W. Herr II (IU and ChalkLab), Gully Burns (USC), David Newman (UCI)

Brain image purchased from http://www.istockphoto.com

INDIANA UNIVERSITY
SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Table of Contents

Executive Summary

This report summarizes the results of two National Science Foundation (NSF) workshops on "Knowledge Management and Visualization Tools in Support of Discovery" that were inspired by NSF's Cyber-Enabled Discovery and Innovation (CDI) initiative. The two workshops took place at the National Science Foundation, Arlington, VA on March 10-11, 2008 and the New York Hall of Science, Queens, NY on April 7-8, 2008.

Major findings of both workshops are captured here. Specifically, the report intends to:

- Capture knowledge management and visualization needs with specific examples for biomedical/ecological research and SoS research.
- Present major challenges and opportunities for the design of more effective tools and CIs in support of scholarly discovery, including a timeline of anticipated science and technology development that will impact tool development.
- Provide recommendations on how current lines of work in academia, government, and industry and promising avenues of research and development can be utilized to design more effective knowledge management, visualization tools and cyberinfrastructures that advance discovery and innovation in 21st century science.

PRINCIPAL FINDINGS

There are several general, confirmative findings:

- **Science is interdisciplinary and global.** Researchers and practitioners need easy access to expertise, publications, software, and other resources across scientific and national boundaries.
- **Science is data driven.** Access to large amounts of high-quality and high-coverage data is mandatory. The "long tail" of data producers/users is larger than the existing major databases and their users.
- **Science is computational.** The design of modular, standardized and easy to use cyberinfrastructures is key for addressing major challenges, such as global warming, or a deeper understanding of how science and technology evolves. Ideally, the "million minds" can share, combine, and improve expertise, data, and tools. It is advantageous for scientists to adapt industry standards, defacto or not, than to have to create their own tools.
- **Science uses many platforms.** Some sciences thrive on Web services and portals, others prefer desktop tools, while some require virtual reality environments, or mobile (handheld) devices.
- **Science is collaborative.** A deeper understanding of how teams "form, storm, norm and perform" will improve our ability to compose (interdisciplinary/international) teams that collaborate effectively.

There were also a number of findings specific to the workshop topic "Knowledge Management and Visualization Tools in Support of Discovery":

- **Formulas and visual imagery** help communicate results across scientific boundaries with different cultures and languages.
- **Users become contributors.** Researchers need more efficient means to share their datasets, software tools, and other resources with each other—just like they share images and videos via Flickr and YouTube today. Overly "top-down" approaches should be discouraged.
- **Science Facebook.** Scholars need more effective ways to find collaborators, monitor research by colleagues and competitors, disseminate their results to a broader audience.
- **Advanced data analyses combined with visualizations** are used to identify patterns, trends, clusters, gaps, outliers and anomalies in massive amounts of complex data. Network science approaches seemed particularly useful in the selected biomedical/ecological and SoS domains.
- **Scientific domains have different affordances.** For example, intuition and approaches developed in the analysis of scholarly data, which is much more readily available and easier to understand than biomedical/ecological data, could be used to study biomedical/ecological data (which requires a large amount of highly specialized background knowledge).

PRINCIPAL RECOMMENDATIONS

Two sets of recommendations resulted. The first set addresses the needs of the biomedical and ecological research community:

- Improve collaboration amongst biological researchers and develop **"Distributed Biomedical Data Repositories"** with common Web-based **"Interactive Visualizations"** to

allow large-scale data to be easily browsed by researchers. (see page 65 in Appendix D)

- A **"Sequenomics Initiative"** is needed to allow mapping nucleic acid and protein sequence possibility spaces over time to create predictive, rather than descriptive, approaches to understanding molecular evolution. (see page 67 in Appendix D)

- Common data repositories and scalable data visualization interfaces of **"Environmental Sensor Data"** are needed to allow long-term trends to be charted, as well as provide opportunities for mash-ups with other data types, such as remote sensing, social/political, and other data sets. (see page 69 in Appendix D)

- An infrastructure/data repository to aid tracing the spatial and temporal **"Migration of Genetic Material in Metagenomic Data"** is needed to improve the understanding of diversity in microbial ecosystems and an evolutionary understanding of the "provenance" of gene structures in microbial biota. (see page 71 in Appendix D)

- Use semantic, social, and visual networks to build a **"cyberinfrastructure for enabling discovery and innovation in genome sciences and neurosciences"** to allow data and tools to be easily published, shared, discovered and linked for enabling scientific collaboration, knowledge discovery and technological innovation. (see page 60 and 80 in Appendix D)

- Using the **"Semantic Web to Integrate Ecological Data and Software"** would allow cataloging and integration of existing and evolving databases and software and create semantic web ontologies for accessing them through software "wrappers" that turn them into easily used and accessed Web services. This could also make ecological data available to the public in real-time and increase understanding and awareness of the biodiversity impacts of global change for policymakers, the public and formal and informal teaching and learning. (see page 66 in Appendix D)

- Developing intuitions and formalisms from "**social network analysis to understand the coupling of activity patterns between tissues in a diseased organ"** would make possible explorations of how genetic, epigenetic and environmental factors contribute to the disregulation of organs through a systems approach. (see page 63 in Appendix D)

The second set proposes solutions for advancing SoS research specifically, and the scholarly research infrastructure in general, including:

- A decentralized, free **"Scholarly Database"** is needed to keep track, interlink, understand and improve the quality and coverage of Science and Technology (S&T) relevant data. (see also page 76 and 77 in Appendix D)

- Science would benefit from a **"Science Marketplace"** that supports the sharing of expertise and resources and is fueled by the currency of science: scholarly reputation. (see page 74 in Appendix D) This marketplace might also be used by educators and the learning community to help bring science to the general public and out of the "ivory tower". (see page 89 in Appendix D)

- A **"Science Observatory"** should be established that analyzes different datasets in real-time to assess the current state of S&T and to provide an outlook for their evolution under several (actionable) scenarios. (see page 72 in Appendix D)

- **"Validate Science Maps"** to understand and utilize their value for communicating science studies and models across scientific boundaries, but also to study and communicate the longitudinal (1980-today) impact of funding on the science system. (see page 81 in Appendix D)

- Design an easy to use, yet versatile, **"Science Telescope"** to communicate the structure and evolution of science to researchers, educators, industry, policy makers, and the general public at large. (see page 87 in Appendix D) The effect of this (and other science portals) on education and science perception needs to be studied in carefully controlled experiments. (see page 88 in Appendix D)

- **"Science of (Team) Science"** studies are necessary to increase our understanding and support the formation of effective research and development teams. (see page 78 and 82 in Appendix D).

- **"Success Criteria"** need to be developed that support a scientific calculation of S&T benefits for society. (see also page 88 in Appendix D)

- A **"Science Life"** (an analog to Second Life) should be created to put the scientist's face on their science. Portals to this parallel world would be installed in universities, libraries and science museums. (see page 80 in Appendix D) The portals would be "fathered and mothered" by domain, as well as learning experts. Their effect on education and science perception should be rigorously evaluated in carefully controlled experiments and improved from a learning science standpoint. (see page 91 in Appendix D)

Section 1

Workshop Organization and Report Structure

Two National Science Foundation (NSF) funded workshops on "Knowledge Management and Visualization Tools in Support of Discovery" were held in 2008. They were inspired by NSF's Cyber-Enabled Discovery and Innovation (CDI) initiative and took place at the National Science Foundation, Arlington, VA on March 10-11, 2008 and the New York Hall of Science, Queens, NY on April 7-8, 2008.

Workshop (I) brought together 17 leading experts in biomedical/ecological research and science of science research, see group photo in Figure 1.1. The participants came from the U.S., Japan, and Europe.

In addition, 20 agency staff members joined different sessions. Appendix A lists the names and biographies of the attendees. The workshop produced a set of challenges faced by biomedical/ecological and SoS researchers and practitioners that might be solvable in the next 10 years. These challenges were documented and usage scenarios were detailed.

Results of Workshop (I) were communicated to participants of Workshop (II) (consisting of about 20 leading academic, government and industry experts in database research and digital libraries, cyberinfrastructure/e-Science design, and

Figure 1.1: *Workshop (I) Participants at the National Science Foundation in Arlington, VA on March 10-11, 2008*

social science from the U.S. and Europe), see group photo in Figure 1.2. A large percentage of participants came from industry–small companies, but also Google and Microsoft attended–adding a spirit of fierce competition, and later collaboration, to the mix. Half a dozen informal science learning experts from the New York Hall of Science joined diverse sessions. Appendix A lists the names and biographies of the attendees. The goal of the second workshop was the identification of socio-technical opportunities, existing and evolving technologies, and incentive structures that might help create, serve, and sustain novel tools and cyberinfrastructures (CIs) that meet the challenges identified in Workshop (I).

The biomedical/ecological and SoS research areas were exemplarily selected based on the expertise of the organizers, see section 2 for an introduction of these domains. Both domains face serious data challenges and need to integrate, mine, and communicate these data in order to generate solutions to important challenges. Plus, they are sufficiently different – solutions that work here might transfer to other domains.

The remainder of this report is organized as follows: section 2 gives a general overview of data access, data navigation, and data management challenges faced by all sciences; section 3 details challenges faced and data analysis, modeling and visualization approaches and tools used in biomedicine/ecology research, SoS research and practice; section 4 presents raw and aggregated workshop results, together with a discussion of concrete research projects identified as most promising by workshop organizers and participants; section 5 concludes this report with a prospective timeline of incremental and transformative socio-technical advances that will most likely make possible qualitatively new tools in biomedical, ecological, science policy and other areas of research and practice.

Appendix A provides biographies of workshop organizers, participants, attending agency staff, and facilitators. Appendix B features the agendas for both workshops. Appendix C has a listing of relevant readings, datasets and tools. Last but not least, Appendix D showcases promising research projects identified by workshop organizers and participants.

Figure 1.2: *Workshop (II) Participants of the New York Hall of Science, Queens, NY on April 7-8, 2008*

Section 2

General Data Management and Visualization Challenges

Massive amounts of (streaming) scientific data and the data intensive and collaborative nature of most scholarly endeavors today (Hey et al., 2009) make knowledge management and visualization tools in support of discovery a necessity. The number of currently active researchers exceeds the number of researchers ever alive. Researchers publish or perish. Some areas of science produce more than 40,000 papers a month. Computer users world-wide generate enough digital data every 15 minutes to fill the U.S. Library of Congress and more technical data have been collected in the past year alone than in all previous years since science began (Parashar, 2009).

We are expected to know more than one can possibly read in a lifetime. Many of us receive many more emails per day than can be processed in 24 hours. We are supposed to be intimately familiar with datasets, tools, and techniques that are continuously changing and increasing in number and complexity. All this, while being reachable twenty-four hours per day, seven days per week.

Not only library buildings and storage facilities, but also databases are filling up more quickly than they can be built. In addition, there are scientific datasets, algorithms, and tools that need to be mastered in order to advance science. No single person or machine can process the entirety of this enormous stream of data, information, knowledge, and expertise.

Over the last several hundred years, our collective knowledge has been preserved and communicated via scholarly works, such as papers or books. Works might report a novel algorithm or approach, report experimental results, or review one area of research among others. Some report several small results, e.g., novel gene-disease linkages, others one large result like the Human Genome. Some confirm results, while others disprove results. Some are filled with formulas, while others feature mostly artistic imagery. Different areas of science have very different publishing formats and quality standards. The description of one and the same result, e.g., a novel algorithm, will look very different if published in a computer science, biology, or physics journal.

Acknowledgments provide linkages to experts, datasets, equipment, and funding information. Since about 1850, citation references have been used. Citation networks of papers show who consumes, elaborates on, or cites whose work. Networks of co-authors can be extracted by counting how often two authors wrote a paper together. The quality of a paper and the reputation of its author(s) are commonly estimated via citation and download counts.

Experts are challenged when asked to identify all claims a scholarly work makes. In some cases, it is simply not known how important today's discovery is for tomorrow's science and society. It is an even harder task to determine how these claims are interlinked with the complex network of prior (supporting and contradicting) results manifested in the shape of papers, books, patents, datasets, software, tools, etc.

Many call this enormous amount of scientific data, information, knowledge, and expertise an "Information Big Bang" or even a "Data Crisis" and propose possible solutions (Association of Research Libraries, 2006; National Science Foundation, 2005; National Science Foundation Cyberinfrastructure Council, 2007; Peterson et al., 2007; President's Council of Advisors on Science and Technology, 2007). A recent *Nature* article titles it "Big Data" (Nature Publishing Group, 2008) and argues that "Researchers need to be obliged to document and manage their data with as much professionalism as they devote to their experiments. And they should receive greater support in this endeavor than they are afforded at present."

Factually, our main and most efficient method of accessing everything we collectively know is search engines. However, the usage of search engines resembles charging a needle with a search query and sticking it into a haystack of unknown size and consistency. Upon pulling the needle out, one checks the linearly sorted items that got stuck on it. This seems to work well for fact-finding. However, it keeps us at the bottom of confirmed and unconfirmed records in the haystack of our collective knowledge. Of course, we can explore local neighborhoods of retrieved records via Web links or citation links. However, there is no 'up' button that provides us with

a more global view of what we collectively know and how everything is interlinked. Novel tools are needed to support identifying patterns, trends, outliers, or the context in which a piece of knowledge was created or can be used. Without context, intelligent data selection, prioritization, and quality judgments become extremely difficult.

Recent work in Visual Analytics (Thomas & Cook, 2005) applies the power of advanced data mining and visualization techniques to support the navigation, sense-making, and management of large-scale, complex, and dynamically changing datasets. These tools have also been called *macroscopes*, a term originally proposed by Joël de Rosnay (de Rosnay & Edwards (Trans.), 1979) and derived from *macro*, great, and *skopein*, to observe. Just as the microscope empowered our naked eyes to see cells, microbes, and viruses, thereby advancing the progress of biology and medicine, or the telescope opened our minds to the immensity of the cosmos and has prepared mankind for the conquest of space, macroscopes promise to help us cope with another infinite: the infinitely complex. Macroscopes give us a 'vision of the whole' and help us 'synthesize'. They let us detect patterns, trends, outliers, and access details in the landscape of science. Instead of making things larger or smaller, macroscopes let us observe what is at once too great, too slow, or too complex for our eyes.

Distinguishing Features of Science

According to Wilson (1998), the distinguishing features of science, as opposed to pseudoscience, are repeatability, economy, mensuration, heuristics, and consilience. **Repeatability** refers to the fact that any science analysis, model, or map can successfully be rerun or regenerated by a scholar other than the author. This requires that datasets be accessible and documented in a way that they can be recompiled. Software must also be made available or documented in sufficient detail so that it can be reimplemented and run with identical parameter settings to produce the exact same results. **Economy** demands that results be presented in the simplest and most informative manner, for both the expert and the general audience. **Mensuration** refers to proper measurement, using universally accepted scales in an unambiguous fashion. **Heuristics** refers to the principle that the best science stimulates further discovery. **Consilience** suggests that consistent explanations and results are most likely to survive.

Macroscopes use visualizations to communicate analysis results. If well designed, these visualizations provide the ability to: comprehend huge amounts of data, reduce search time, reveal relations otherwise not being noticed, facilitate hypothesis formulation, and provide effective sources of communication (Tufte, 1983).

Macroscopes need to be scalable and many will resemble cyberinfrastructures (Atkins et al., 2003) that exploit distributed hardware and software components. They need to be modular and flexible so that they can be quickly adapted and customized for different application domains and user groups that are becoming increasingly larger and more diverse. They need to be reliable and they have to effectively exploit existing hardware that was not necessarily developed for science (National Science Foundation, Office of Cyberinfrastructure, 2009).

Macroscopes for scientists have to support science as, e.g., defined by Edward O. Wilson (1998), see text box. For example, they have to make it easy to re-run and replicate analyses, to compare one analysis results with others, or to analyze a new dataset using the very same set of algorithms and parameter values.

Incentive structures need to be designed such that researchers share expertise and resources instead of keeping them for their personal usage only. Ideally, reputation also goes to those which share and care in support of the global common scientific endeavor. Successful—not exclusively scholarly—examples that show how this might work are Wikipedia (http://www.wikipedia.org), Swivel (http://www.swivel.com), or IBM's Many Eyes (http://manyeyes.alphaworks.ibm.com/manyeyes). The latter two allow anybody to upload, visualize, and interact with data in novel ways. Both sites are highly interactive and socially oriented. Users can form communities, meet like-minded people, and browse the space of their interests. They are driven, populated, and defined by those who congregate there.

Section 3

Data Analysis, Modeling, and Visualization Opportunities in Biomedicine/Ecology and SoS Research

This section presents research needs from the biomedical/ecological and Science of Science (SoS) domains, together with current lines of research as relevant for the workshop topic.

Biomedical/Ecological Research

BIOCHEMISTRY MODELING AND VISUALIZATION

Visualizing complex macromolecular structures has become an important and commonly used tool for studying living processes. A classic example is Watson & Crick's synthesis of Franklin and Wilkins' diffraction data into an easy to comprehend 3-D model of deoxyribonucleic acid (DNA) and then going on to show how this could explain many of the fundamental processes of genetics, see Figure 3.1. The power of this visualization, of course, stems from the fact that the 3-D structure represented a real chemical entity. It also connected two disciplines – chemistry and genetics – with a central visual metaphor that proves useful to this day.

Representing molecular structures and mechanics has quickly evolved into the study of structural genomics and computational biophysics. A number of tools exist that allow researchers to manipulate and visually compare macromolecular structures, including the ability to computationally dock protein structures to understand their interactions. Examples are Chime (Symex Solutions, 2009), RasMol (Sayle, 1999), PyMOL (Schrödinger, 2010), O (Jones, 2010), and MolMovDB (Gerstein and Krebs, 2000). Resources such as Dockground (http://dockground.bioinformatics.ku.edu) represent growing repositories of protein interface data (Douguet et al., 2006).

Many molecular functions can be thought of in this way and the ability to manipulate and visualize 3-D models has become important to understanding the function of macromolecules and their interactions, as well as understanding new approaches to treat disease.

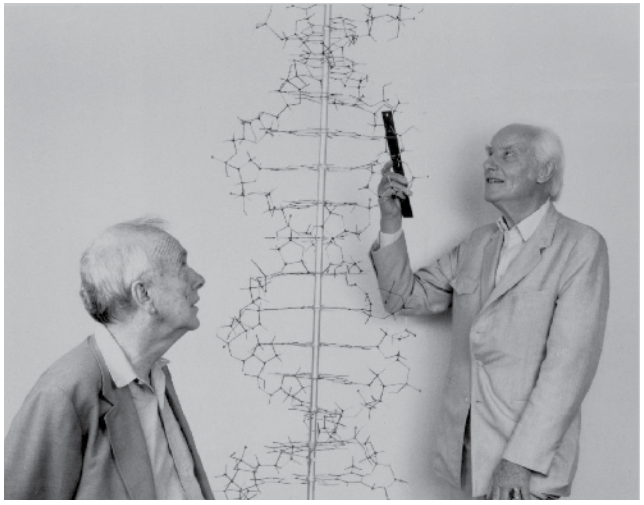


Figure 3.1: *James Watson's and Francis Crick's discovery in 1953 that DNA forms a double helix with complimentary nucleobases interacting via hydrogen bonds allowed understanding of the molecular basis for heredity.*

Credit: Cold Spring Harbor Laboratory Digital Archive, James Watson Collection.

Solid modeling, articulate modeling, and augmented reality take this a step further. The ability for both the interactions and structure of macromolecules to be printed as a tangible, manipulatable solid allows discovery and research into their function that is otherwise difficult, if not impossible, see Figure 3.2.

Articulated models include the ability for the model to be physically reconfigured and positioned to understand folding and docking with other proteins, as well as other important functions, such as the mechanics of viruses, see Figure 3.3.

Augmented reality allows computer-generated models to be associated with tangible models by tracking the physical model to the 3-D representation on the screen, see Figure 3.4.

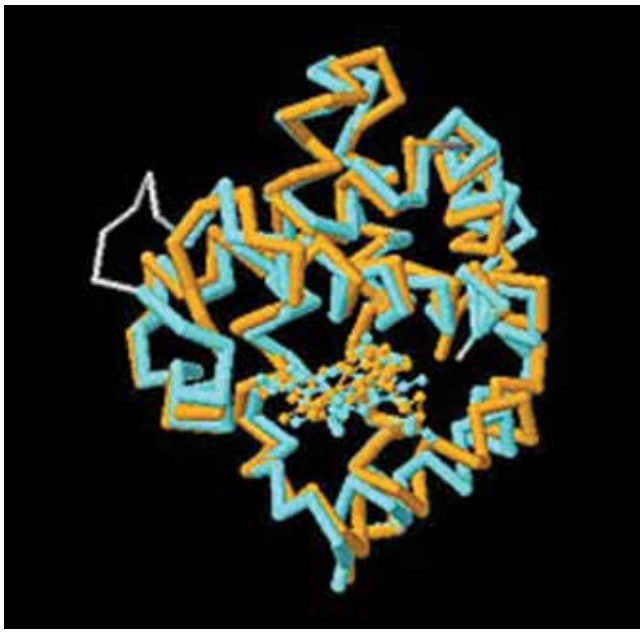


Figure 3.2: *Structural alignment of protein chains 4HHB A and B Using JFatCat paiwise alignment algorithm (rigid) from the Protein Data Bank (http://www.rcsb.org/pdb). This is a snapshot from a 3-D manipulable model generated in Java. Many other alignment and comparison methods are possible.*

Credit: Original structural data generated by Fermi & Perutz (1984).

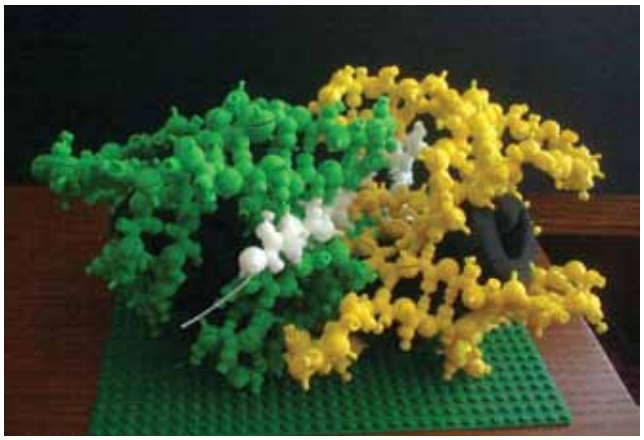


Figure 3.3: *This articulated model of the HIV protease backbone, which is an important drug target in treating AIDS patients, is composed of two identical protein chains (depicted in yellow and green). It functions by cutting other peptide chains (shown in white), in a step that is critical to the maturation of the virus. The folding is accomplished by the interactions of magnets representing hydrogen bond donors and acceptors in the peptide units. The completed model enables manipulation of the two "flaps" that must open to allow the peptide substrate into the active site of the enzyme. Image/Caption is from Molecular Graphics Laboratory Models (and augmentation).*

Credit: Developed in the Olson Laboratory, the Scripps Research Institute. Photo by Arthur Olson, Ph.D. http://mgl.scripps.edu/projects/tangible_models/articulatedmodels.

Figure 3.4: *Computer augmentation of two subunits of the superoxide dismutase (SOD) dimer, which are tracked and manipulated independently. The electrostatic field is shown with small arrows that point along the local field vectors (they appear as small dots in this picture), and the potential is shown with volume rendered clouds, with positive in blue and negative in red.*

Credit: Models (and augmentation) developed in the Olson Laboratory, the Scripps Research Institute. Photo by Arthur Olson, Ph.D. http://mgl.scripps.edu/projects/tangible_models/augmentedreality.

NETWORKS AS A USEFUL PARADIGM IN BIOLOGY AND ECOLOGY

Beyond the scale of simple pairwise interactions, multiple interactions quickly become too complex to model using the structural and molecular biophysical approach. As Carl Woese (2004) and others have pointed out, traditional computational techniques in molecular biology cannot yield an adequate understanding of organisms, cells and other large-scale biological systems.

Systems biology puts an emphasis on new approaches to data gathering, creating databases, data mining and analysis. Using a systems paradigm, systems biology analyzes functional interactions within biological systems over time (Xia et al., 2004). In Proteomics, network analysis is used to link protein interactions for such objectives as understanding the coordinated expression of genes connected by protein interaction and regulatory networks (Han et al., 2004; Luscombe et al., 2004). Networks also provide a way to integrate diverse types of data (Harnad, 2003; Jansen et al., 2003; Lee et al., 2004). The implications of systems biology for medicine involve studying the interactions and combinations of genes that manifest in disease and how cells work as a system, see Figure 3.6.

Molecular biology and network science have also provided useful tools for characterizing whole organisms and ecosystems using a "Computational Ecology" approach. While there are many large scale and localized geospatial mapping approaches to ecological data (such as the USGS National Map Viewer and Global Bioinformation Facility) that help understand species distribution and earth science data, network mapping techniques and other computational approaches help researchers to understand interactions and identify characteristics and indicators for ecosystem health that spatial mapping has not adequately addressed,

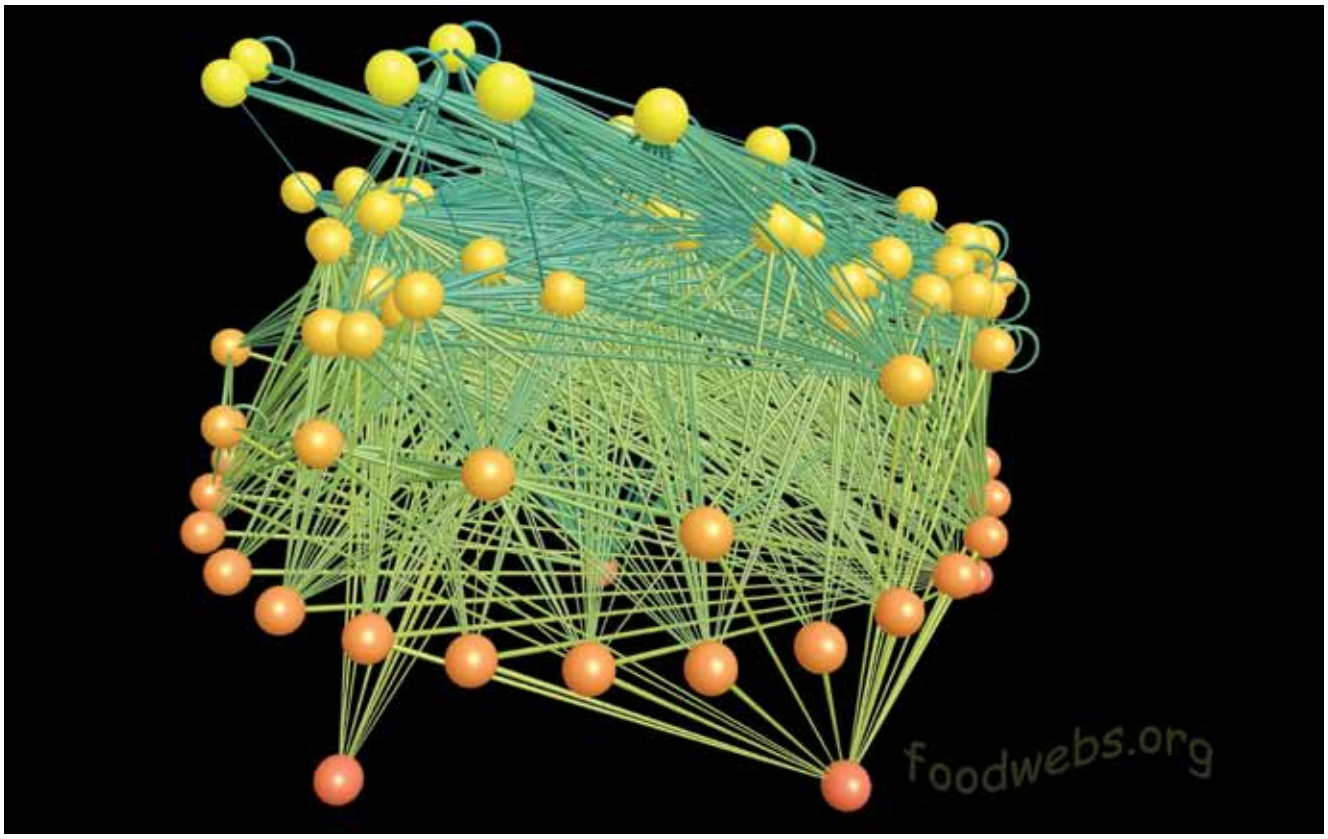


Figure 3.5: *Network map of a Caribbean coral reef in the Puerto Rico Virgin Islands shelf complex. Fish represent 84% of the taxa in this trophic species web. The web structure is organized vertically, with node color representing trophic level. Red nodes represent basal species, orange nodes represent intermediate species, and yellow nodes represent primary predators. Links characterize the interaction between two nodes, and the width of the link attenuates down the trophic cascade.*

Credit: Image produced with FoodWeb3D, written by R. J. Williams and provided by the Pacific Ecoinformatics and Computational Ecology Lab (http://www.foodwebs.org) (Yoon et al., 2004).

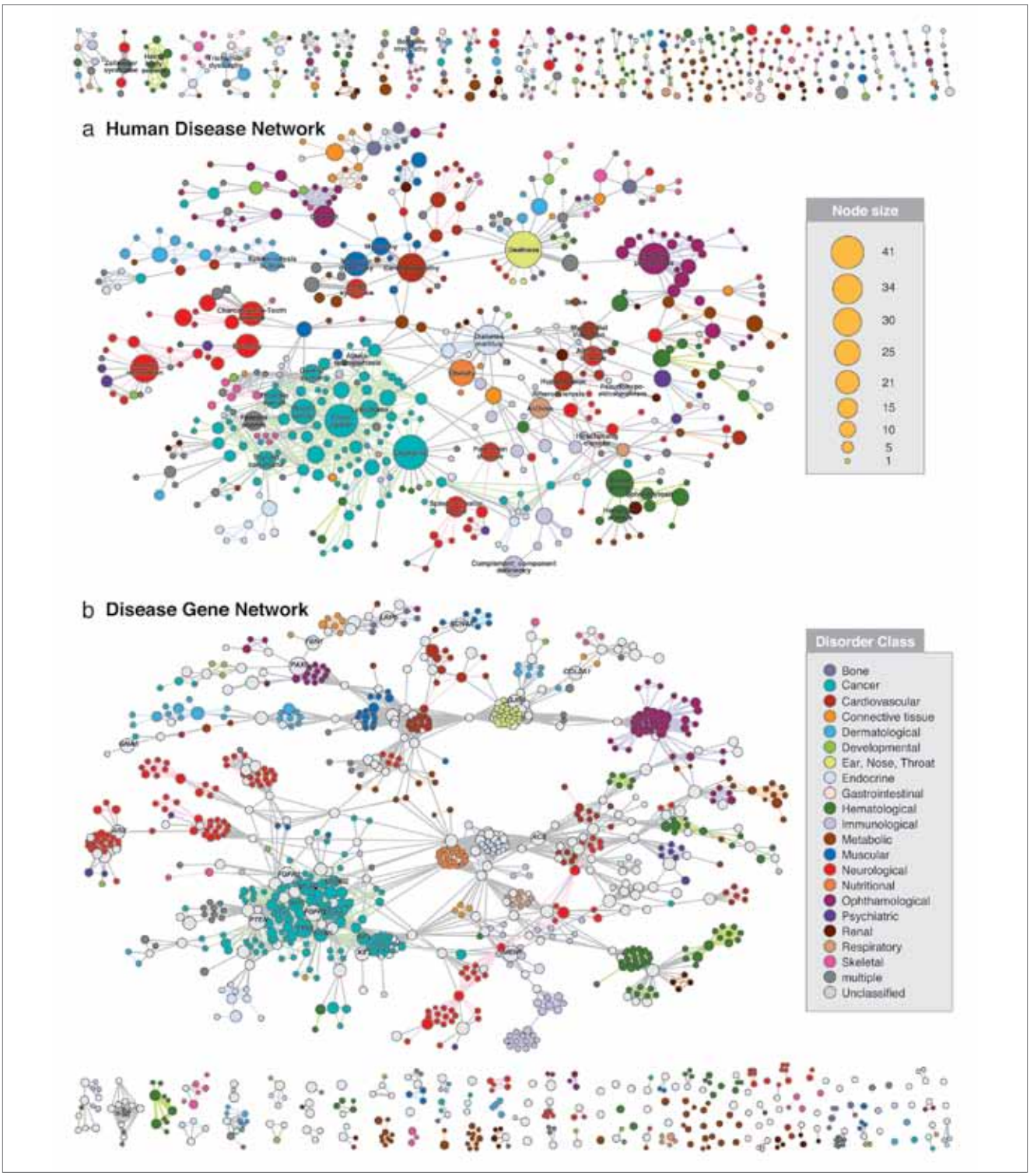


Figure 3.6: *Bipartite-graph representation of the diseasome.*
A disorder (circle) and a gene (rectangle) are connected if the gene is implicated in the disorder. The size of the circle represents the number of distinct genes associated with the disorder. Isolated disorders (disorders having no links to other disorders) are not shown. Also, only genes connecting disorders are shown.

Credit: (Goh et al., 2007)

such techniques are being used by researchers like Neo Martinez and Jennifer Dunne (Dunne et al., 2004) to better understand both dynamics and trophic interactions in ecosystems, see Figure 3.5. As geospatial mapping techniques achieve higher resolutions, better portray dynamic data streams and can be better correlated to trophic webs, patterns of species distribution and effects of human impacts should be easier to understand.

METAGENOMICS

Assaying of genomic materials for whole ecosystems has become a valuable tool for computational and molecular biology to gain a much deeper knowledge of ecological systems. Metagenomic data represent a vast wilderness of new information about how ecologies function at microbial scales (see Figure 3.7). They impact the complexity and interdependency of systems from those within the human gut to the Sargasso Sea. A National Research Council report from 2007 suggests that Metagemonics is a convergence of the study of microbes and genetics to form a new approach to understanding the ecology of microbial communities (National Research Council Board on Life Sciences, 2007).

Sharing of metagenomic data through high bandwidth networks and flexible visualization platforms is needed for collaboration and to allow large-scale visualizations of metagenomic data to be displayed and annotated. Systems, such as OptIportal, allow dynamic video streams to be added to create dynamic mash-ups of data and allow direct interactions amongst researchers on a variety of hardware platforms. Another example is the Free Space Manager in SAGE, developed at the Electronic Visualization Laboratory at the University of Illinois at Chicago, see Figure 3.8.

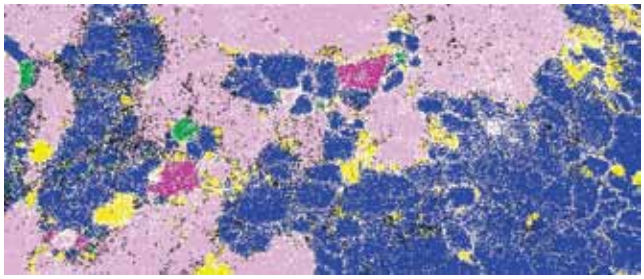


Figure 3.7: *Self Organizing Maps can be used to visualize metagenomic data. This sample is from metagenomic data from 1500 microbes of the Sargasso Sea color coded showing 40 Eukaryotes, 1065 Viruses, 642 Mitochondria and 42 Chloroplasts.*

Credit: (Abe et al., 2006)

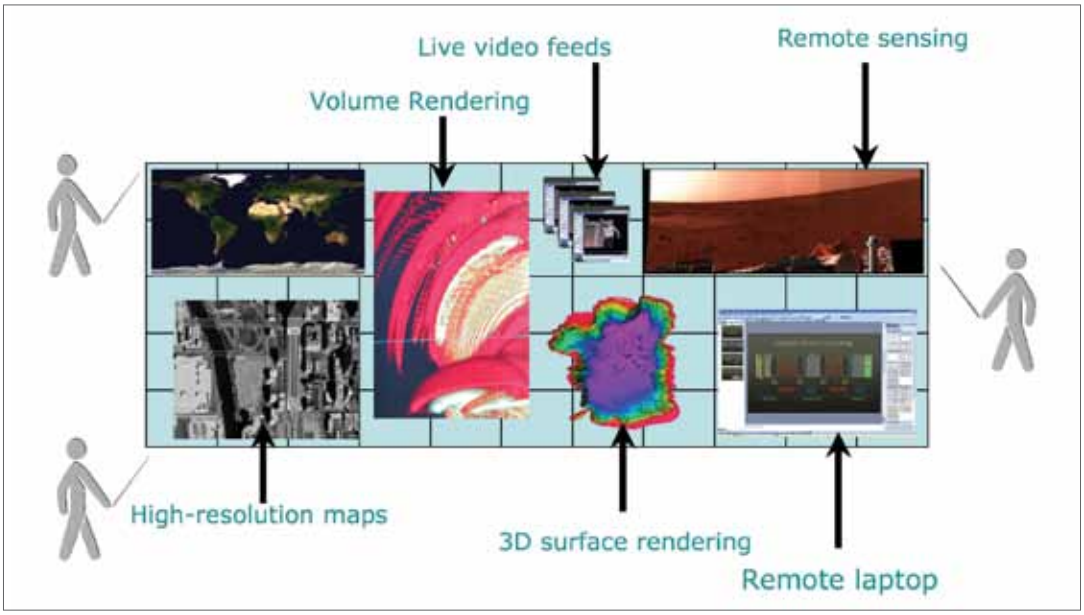


Figure 3.8: *The Free Space Manager provides an intuitive interface for moving and resizing graphics on a tiled display. When a graphics window is moved from one portion of the screen to another, the Free Space Manager informs the remote rendering clusters of the new destination for the streamed pixels, giving the illusion of working on one continuous computer screen, even though each of their systems may be several thousand miles apart.*

Credit: Electronic Visualization Laboratory, University of Illinois at Chicago, http://www.evl.uic.edu/cavern/sage/description.php.

EPIDEMIC MODELING AND FORECASTING

The ability to predict and monitor the trajectory of epidemics has gained enormously from the availability of global health data, the development of advanced complex systems models, and increased computer power, to both directly model the spread of contagions and also to model human social structure and behavior. Traditional geographic maps and network visualizations of human social interactions support the examination of large-scale field observations, travel records, and other data to understand the progression of densification and network clustering rendered as static maps or real-time animations. These techniques greatly aid the understanding of how epidemics, such as 2009 H1N1, progress and allow researchers to monitor virulence and morbidity at the "macro" scale.

Note that it has long been suggested that scientific ideas may spread by personal contact in analogy to what happens with contagious diseases (Garfield, 1980; Goffman, 1966; Goffman & Newill, 1964), connecting epidemic modeling to SoS research discussed in section 3.2. However, there are important differences: knowledge is desirable to acquire, scientific knowledge usually requires extensive education and training, and there are many structures created to promote contact between scientists (meetings, school, Ph.D. programs) and to extend the lifetime of scientific ideas (publication, archives). Recently, the features of contact that are common to the transmission of infectious diseases and those specific to science have lead to a new class of population models that apply to the temporal development of scientific fields (Bettencourt et al., 2006; Bettencourt et al., 2008).

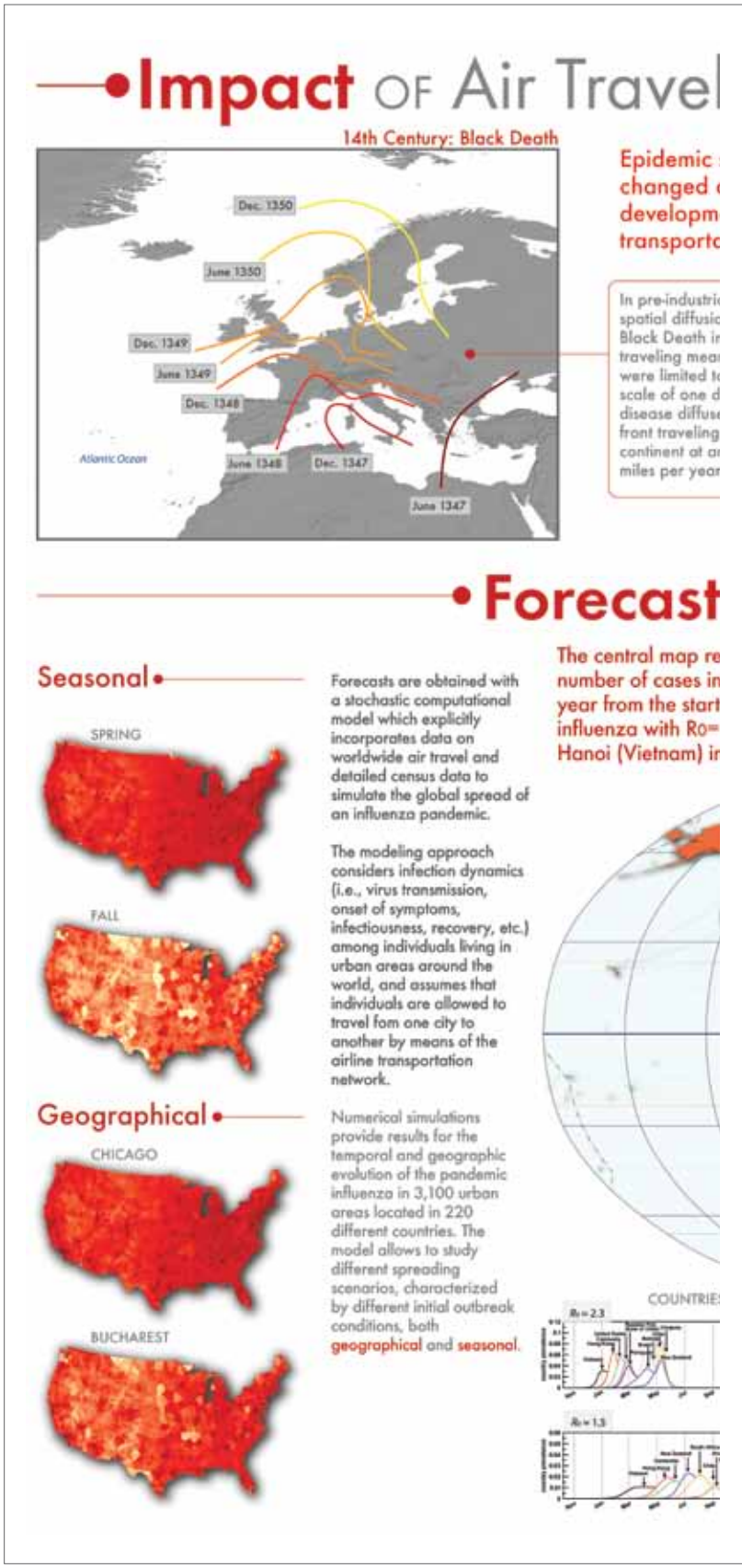


Figure 3.9: *Colizza,Vespignani and their colleagues apply statistical and numerical simulation methods in the analysis and visualization of epidemic and spreading phenomena. Recently, they developed a stochastic large-scale spatial-transmission model for the analysis of the global spreading of emerging infectious diseases. Detailed knowledge of the worldwide population distribution to the resolution scale of ¼° and of the movement patterns of individuals by air travel is explicitly incorporated into the model to describe the spatio-temporal evolution of epidemics in our highly interconnected and globalized world. Simulation results can be used to identify the main mechanisms behind observed propagation patterns, e.g., the patched and heterogeneous spreading of the SARS outbreak in 2002-2003, and to provide forecasts for future emerging infectious diseases, e.g., a newly emerging pandemic influenza. Maps showing different simulated scenarios of possible epidemics might be of crucial help in the identification, design, and implementation of appropriate intervention strategies aimed at possible containment.*

Credit: (Colizza, Barrat, Barthélemy, Valleron et al., 2007; Colizza, Barrat, Barthélemy, & Vespignani, 2007)

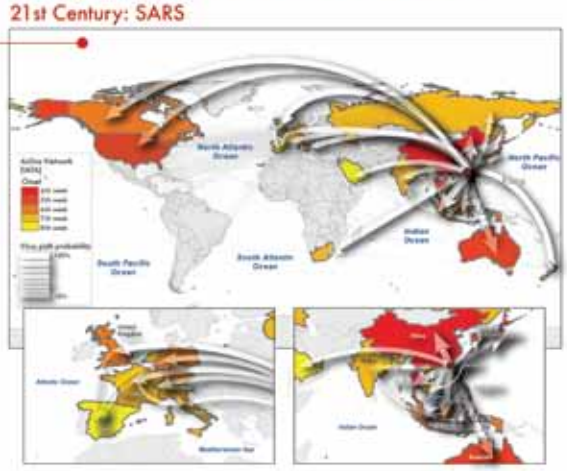
l ON Global Spread OF Infectious Diseases

spreading pattern dramatically after the ent of modern ation systems.

al times disease spread was mainly a n phenomenon. During the spread of the 14th century Europe, only few ns were available and typical trips relatively short distances on the time lay. Historical studies confirm that the d smoothly generating an epidemic as a continuous wave through the approximate velocity of 200-400

The SARS outbreak on the other hand was characterized by a patched and heterogeneous spatio-temporal pattern mainly due to the air transportation network identified as the major channel of epidemic diffusion and ability to connect far apart regions in a short time period. The SARS maps are obtained with a data-driven stochastic computational model aimed at the study of the SARS epidemic pattern and analysis of the accuracy of the model's predictions. Simulation results describe a spatio-temporal evolution of the disease (color coded countries) in agreement with the historical data. Analysis on the robustness of the model's forecasts leads to the emergence and identification of epidemic pathways as the most probable routes of propagation of the disease. Only few preferential channels are selected (arrows; width indicates the probability of propagation along that path) out of the huge number of possible paths the infection could take by following the complex nature of airline connections (light grey, source: IATA).

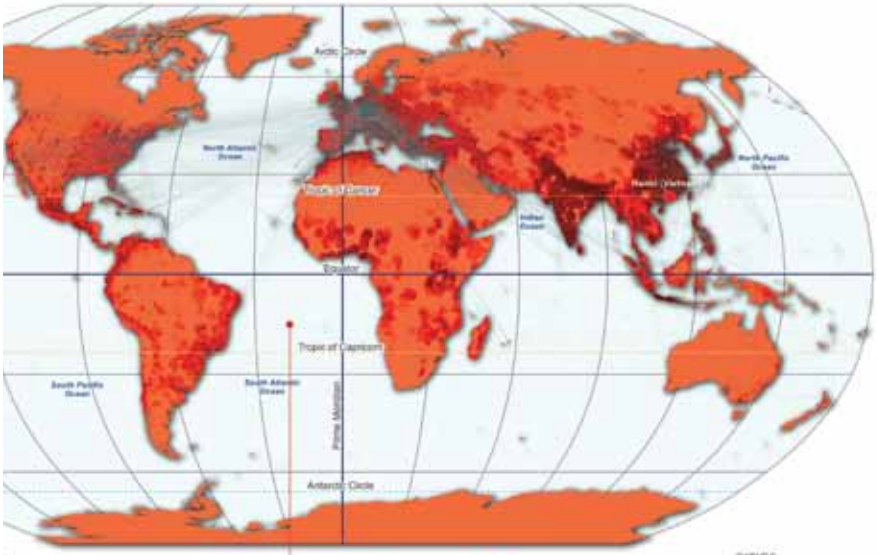
21st Century: SARS

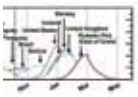


's OF THE Next Pandemic Influenza

presents the cumulative the world after the first of a pandemic 1.9 originating in the Spring.

The US maps focus on the situation in the US after one year, and show the effect of changes in the original scenario analyzed. Different color coding is used for the sake of visualization.

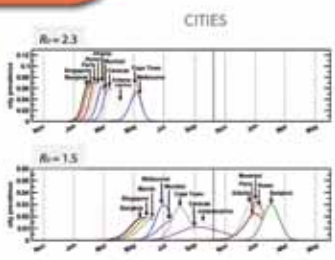




R_0=1.5 R_0=1.9 R_0=2.3

Time evolution of a pandemic starting in Hanoi (Vietnam) in the Fall in the no intervention scenario. Profiles of the fraction of infectious individuals in time (prevalence) are shown for some representative countries (left) and cities (right). Two different values of the reproductive number are considered: R_0=1.5, consistently with the values shown for the US map (top right), and R_0=2.3, in order to provide the comparison with faster spreading.

CITIES



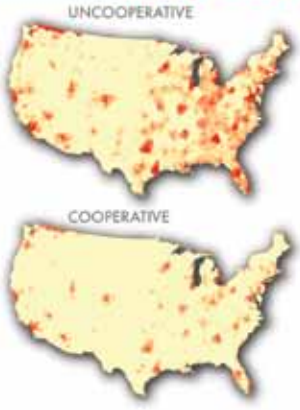
The model inlcudes the worldwide air transportation network (source: IATA) composed of 3,100 airports in 220 countries and E=17,182 direct connections, each of them associated to the corresponding passenger flow. This dataset accounts for 99% of the worldwide traffic and is complemented by the census data of each large metropolitan area served by the corresponding airport.

Additional spreading scenarios can be obtained by modeling different levels of infectiousness of the virus, as expressed in terms of the reproductive number R_0, representing the average number of infections generated by a sick person in a fully susceptible population.

Intervention strategies modeling the use of antiviral drugs can be considered. Two scenarios are compared: an uncooperative strategy in which countries only use their own stockpiles, and a cooperative intervention which envisions a limited worldwide sharing of the resources.

Reproductive Number (R_0)

Intervention



These also integrate population estimation techniques that allow the extrapolation of temporal trends and their statistical uncertainty into the future, so that it becomes possible to forecast numbers of authors and publications specific to a field. Spatial models and models that allow interaction and competition between fields are active research topics.

Monitoring the spread of disease is intimately related to social interactions at the "micro" scale. Complex systems approaches and network science have been used for a variety of applications, including the spread of obesity and happiness (Christakis & Fowler, 2007; Fowler & Christakis, 2008). Visualizations of activity patterns over evolving network structures are frequently used to communicate the results of simulations of empirical data analyses. Recent research aims to "sense" and display human activity patterns in real time, see Figure 3.10.

SciSIP Research and Science Infrastructure

Science of Science and Innovation Policy (SciSIP) is the title of a new program at NSF. The program supports research which expands models, analytical tools, data and metrics that can be applied in the science policy decision making process (National Science Foundation, 2009). Among others, SciSIP proposals and research "may develop behavioral and analytical conceptualizations, frameworks or models that have applications across a broad array of SciSIP challenges, including the relationship between broader participation and innovation or creativity. Proposals may also develop methodologies to analyze science and technology data, and to convey the information to a variety of audiences. Researchers are also encouraged to create or improve science and engineering data, metrics and indicators reflecting current discovery, particularly proposals that demonstrate the viability of collecting and analyzing data on knowledge generation and innovation in organizations." The program also supports "the

Figure 3.10: *The SocioPatterns Project being conducted by the ISI Institute aims to shed light on patterns in social dynamics and coordinated human activity. They develop and deploy an experimental social interaction sensing platform. This platform consists of portable sensing device and software tools for aggregating, analyzing and visualizing the resulting data. In the displays, the RFID stations are represented as labeled marks laid out in an oval configuration (left). The size of the dots is proportional to the summed strengths of the signals received from beacons. The links are interactions. These kinds of experiments track complex social interactions and can not only benefit the understanding of the spread of disease, but also of social networks.*

Credit: Photograph by Wouter Van den Broeck, courtesy of the SocioPatterns.org project (Cattuto et al., 2009).

collection, analysis and visualization of new data describing the scientific and engineering enterprise." A main user group is science policy makers, including campus level officials and program and division directors at governmental and private funding organizations. Note that SciSIP analyses results and their visual depictions as science maps are also valuable for other user groups, see section on Science Cartography on the next page.

A key prerequisite for the study of science and innovation by scientific means is high quality, high coverage empirical data. However, today's commercial and free data holdings do not have the quality and coverage needed for micro-level studies. For example, there is no database of all unique scientists in the U.S. or worldwide, all their publications, Ph.D. students and funding awards. In fact, many datasets are held in data silos that are not interlinked, e.g., funding is rarely interlinked with publication and patent data. Plus, there exists no algorithmic means to fully automate the unification of author names, interlink data records, and correct erroneous data. Instead, a combination of automatic mining and manual curation will have to be applied. To make this work, proper incentive structures need to be employed that promote sharing and cleaning of data in support of transparent science and technology (data). During the workshop, diverse science infrastructures and incentive structures were discussed.

Subsequently, SoS analysis and modeling are introduced, as well as the visual representation and communication of the structure and dynamics of science by means of science maps.

DATA ANALYSIS AND MODELING

Most leaps in science and technology are hard to predict. However, they share certain important contextual conditions, where empirical facts supporting new ideas become irresistible and, simultaneously, conceptual or technical frameworks exist to give birth to new theory. For example, in creating the theory of evolution, Charles Darwin was greatly inspired by observing changes in domesticated animal and plant breeds, as well as comparative studies of species throughout the world. But, the ultimate spark of inspiration that placed an enormous quantity of facts spanning the entire realm of biology into the synthesis we know as the theory of evolution by natural selection was his recent learning of the slow gradual changes in the Earth's plate tectonics by Charles Lyell.

The pursuit of a science of science is much more recent than the study of biology or many other sciences. It started in earnest around the 1960s, most notably with the general structures of scientific revolutions described by Thomas Kuhn (1962) and the empirical observations of Derek de Solla Price (1963), though many antecedents exist often shrouded in the veil of philosophical studies (Popper, 1959). Nevertheless, the last 50 years have seen an extraordinary increase in the availability of data, and the beginnings of the development of models and hypotheses about the general processes whereby science and technology develop and can be encouraged. This suggests that the virtuous cycle of empirical observation, modeling and theory development so characteristic of past discoveries may now start becoming possible in studies of science and technology. In its many forms, this is the greatest opportunity made possible by the torrents of data available to us today.

In this light, we can see the making of science and technology as a partly self-organized social process, made possible by the allocation of resources (research grants and labor), by the development of a culture that permits and rewards the pursuit of truth, and facilitated by the existence of specialized institutions (universities and research labs) and an advanced education system. Ultimately, the creation of science and technology are processes that will remain vital and may even accelerate if its internal social mechanisms can make use of inputs and create social and economic value. We discuss each of these three essential components, possible metrics and feedbacks briefly below, as well as associated research and technical challenges.

Science cannot exist without adequately educated and trained scientists, nor is it possible without research support for their time, and for necessary research and administrative facilities. Increasingly, metrics of research support, detailing project funding by federal agencies and less often by universities, foundations and other sources, are becoming available. Nevertheless, it remains fairly obscure to quantify how the activities of a group of researchers working in a specific area was supported. Better metrics of input resources are desirable if we are to understand important drivers of science.

Secondly, measures of scientific productivity have been a traditional target of bibliometric studies. These typically quantify numbers of publications (or patents), as well as impact factors (for journals or authors) based on citation counts. Recently, impact has started being measured in more informal ways, such as online clicks or download counts, e.g., those collected by MESUR (see textbox on the next page). These measures are becoming widely available; but, in the light of concepts of marginal return, must be integrated with available resources to that research area or author. Such metrics are still in the process of development and suffer from issues of data integration and standards.

From a different perspective, the single author or group is often the wrong unit of analysis when pursuing fundamental progress. It is the development of a new scientific field or

MESUR: MEtrics from Scholarly Usage of Resources

The MESUR database (http://www.mesur.org) developed at Los Alamos National Labs and now at Indiana University, contains 1 billion usage events (2002-2007) obtained from 6 significant publishers, 4 large institutional consortia and 4 significant aggregators. The collected usage data spans more than 100,000 serials (including newspapers, magazines, etc.) and is related to journal citation data that spans about 10,000 journals and nearly 10 years (1996-2006). In addition, the project has obtained significant publisher-provided counter usage reports that span nearly 2000 institutions worldwide.

technology by a group of scientists that ultimately generates societal and economic impact. This too, is starting to be the target of modeling, from mean-field population models to networks. It is the understanding of the social processes, whereby a good idea perhaps generated by a few scientists is explored and validated or falsified by many others. This is also the essence of both regular scientific practice, as well as the few and far between great discoveries. The study of these phenomena requires the analysis of temporal, spatial, topical, and social (network) aspects of scientific fields (Bettencourt et al., 2006; Bettencourt et al., 2008; Bettencourt et al., 2009) and remains fairly unexplored, mostly because it is difficult to isolate scientific domains at the microscopic, author level.

Finally, science will only persist if it creates demonstrably beneficial social and economic change. This is probably one of the most obvious, but also one of the least understood issues in the pursuit of a SoS, mostly because such benefits can be tenuous and extremely indirect. The study of patenting, mostly in economics, has revealed a few interesting lessons (Griliches, 1990) that also apply to scientometrics and bibliometrics: patents (and publications) are not a direct measure of the creation of new knowledge. This is because the propensity to publish varies enormously over time and from field to field, so that very small increments in knowledge can produce large amounts of publications and even citations. Many fields that have lost empirical relevance persist out of a self re-enforcing dynamics of new students coming in, new papers being produced, and inevitably, citations being generated.

Thus, metrics that capture the societal and economic value of research are imperative to overcome distortions in value perception. These must also not center exclusively on measures of economic impact (royalties on patents, startup firms, new economic activities) but also on advances of our fundamental knowledge that can inspire the public (via access to information) and stimulate a new generation to seek advanced education (through feedbacks on enrollment and career opportunities in scientific and technical areas).

Only the integrated understanding of the inputs, processes and outputs of science, and their feedbacks can ultimately lead to an understanding of the processes that enable new science and technology, and guide policy to encourage beneficial outcomes. This is the greatest challenge and opportunity we face today with an increasing availability of data that we must make sense of, and that must be enabled by aspects of technology, visualization, and dissemination.

SCIENCE CARTOGRAPHY

Results of SoS studies are frequently communicated visually. So called maps of science aim to serve today's explorers in navigating the world of science. These maps are generated through scientific analysis of large-scale, scholarly datasets in an effort to connect and make sense of the bits and pieces of knowledge they contain (Börner et al., 2003; Shiffrin & Börner, 2004). They can be used to objectively identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in a domain of interest. Local maps provide overviews of a specific area: its homogeneity, import-export factors, and relative speed. They allow one to track the emergence, evolution, and disappearance of topics and help to identify the most promising areas of research. Global maps show the overall structure and evolution of our collective scholarly knowledge. Science maps have been designed for diverse users and information needs as discussed subsequently.

Science Maps as Visual Interfaces to Scholarly Knowledge

The number and variety of available databases in existence today is overwhelming. Databases differ considerably in their temporal, geospatial, and topical coverage. Database quality reaches from 'downloaded from the Web' to 'manually curated by experts'. Visual interfaces to digital libraries provide an overview of a library's or publisher's holdings as well as an index into its records. They apply powerful data analysis and information visualization techniques to generate visualizations of large document sets. The visualizations are intended to help humans mentally organize, electronically access, and manage large, complex information spaces and can be seen as a value-adding service to digital libraries.

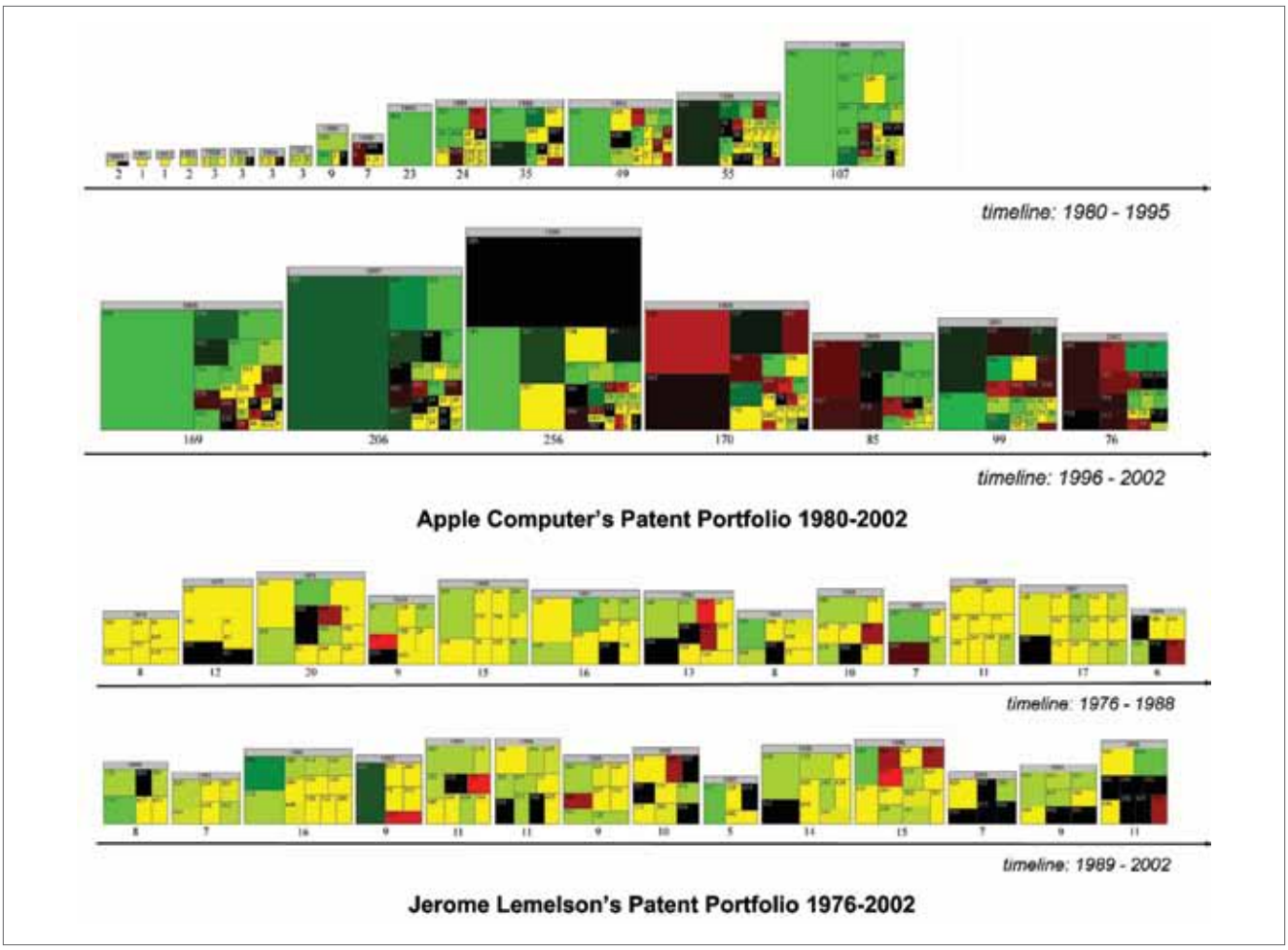


Figure 3.11: *Claiming Intellectual Property Rights via Patents. The evolving patent portfolios of Apple Computer (1980-2002) and Jerome Lemelson (1976-2002) are shown here. The number of patents granted per year matches the size of the square. Each square is further subdivided into patent classes which are color coded in green if the number of patents increased, in red if it decreased, and yellow if no patent was granted in this class in the last five years. While Apple Computer claims more and more space in the same classes, Lemelson's patent holdings are distributed more broadly over the intellectual space.*

Credit: (Shneiderman, 1992; Kutz, 2004)

Mapping Intellectual Landscapes for Economic Decision Making

A deep understanding of technology, governmental decisions, and societal forces is required to make informed economic decisions that ensure survival in highly competitive markets. Discontinuities caused by disruptive technologies have to be determined and relevant innovations need to be detected, deeply understood, and exploited. Companies need to look beyond technical feasibility to identify the value of new technologies, to predict diffusion and adoption patterns, and to discover new market opportunities as well as threats. The absorptive capacity of a company, i.e., its ability to attract the best brains and 'to play with the best' has a major impact on its survival. The importance of social networking tools and network visualizations increases with the demand to understand the "big picture" in a rapidly changing global environment.

Competitive technological intelligence analysis, technology foresight studies, and technology road mapping are used to master these tasks. Easy access to major results, data, tools and expertise is a key to success. But, exactly what is the competition doing, who makes what deals, and what intellectual property rights are claimed by whom?

Last, but not least, companies need to communicate their image and goals to a diverse set of stakeholders—to promote their products, to hire and cultivate the best experts, and to attract venture capital.

Science of Science Policy (Maps) for Government Agencies

Increasing demands for accountability require decision makers to assess outcomes and impacts of science and technology policy. There is an urgent need to evaluate the impact of funding/research on scientific progress: to monitor (long-term) money flow and research developments, to evaluate funding strategies for different programs, and to determine project durations and funding patterns.

Professional science managers are interested to identify areas for future development, to stimulate new research areas, and to increase the flow of ideas into products. Hence, they need to identify emerging research areas; understand how scientific areas are linked to each other; examine what areas are multi-disciplinary; measure collaborations and knowledge flows at the personal, institutional, national, and global level; identify and compare core competencies of economically competing institutions and countries; identify and fund central and not peripheral research centers, etc.

Professional Knowledge Management Tools for Scholars

Most researchers wear multiple hats: They are researchers, authors, editors, reviewers, teachers, mentors, and often also science administrators.

As researchers and authors, they need to strategically exploit their expertise and resources to achieve a maximum increase of their reputation. Expertise refers to the knowledge they already have or can obtain in a given time frame, but also expertise that can be acquired via collaborations. Resources refer to datasets, software, and tools, but also to people supervised or paid.

They need to keep up with novel research results; examine potential collaborators, competitors, related projects; weave a strong network of collaborations; ensure access to high quality resources; and monitor funding programs and their success rates. Last, but not least, they need to review and incorporate findings and produce and diffuse superior research results in exchange of citation counts, download activity, press, etc.

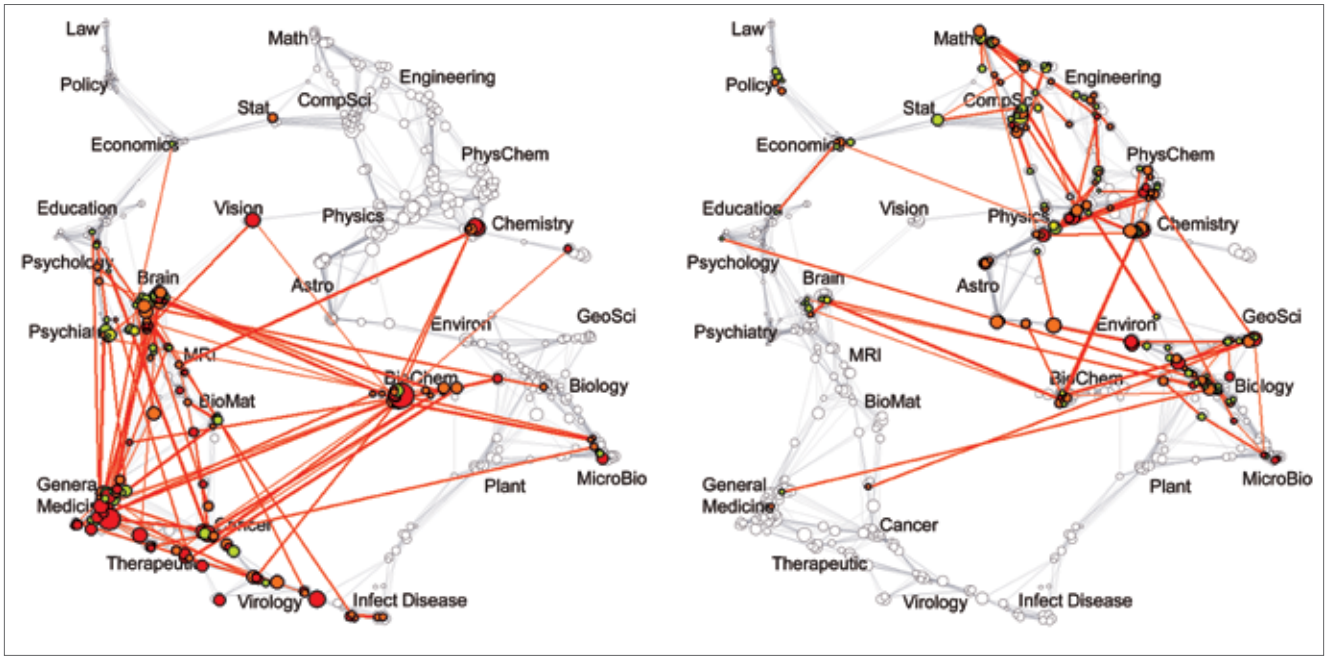


Figure 3.12: *Funding Profiles of NIH (left) and NSF (right).*
Using a base map of science, the core competency of institutes, agencies, or countries can be mapped and visually compared. Shown here are funding profiles of the National Institutes of Health (NIH) and the National Science Foundation (NSF). As the base map represents papers, funding was linked by matching the first author of a paper and the principal investigator using last name and institution information. A time lag of three years between funding of the grant and publication of the paper was assumed. While NIH mostly funds biomedical research, NSF focuses on math, physics, engineering, computer science, environmental and geo sciences, and education. Overlaps exist in chemistry, neuroscience, and brain research.

Credit: (Boyack et al., 2009)

As editors and reviewers, researchers act as gatekeepers of science. They need detailed expertise of their domain, as well as related domains of research, to ensure that only the most valuable and unique works become manifested in the eternal mountain of scholarly works.

As teachers and mentors, they provide students with a deep understanding of the structure and evolution but also the peculiarities of a domain of research and practice. They might give an overview of the teaching material to be covered first; then highlight major people, important papers, and key events; and provide a red thread or pathway and help students discover, understand, and interrelate details.

As science administrators, they are responsible for decisions regarding hiring and retention; promotion and tenure; internal funding allocations; budget allocation; outreach. Here, a global overview of major entities and processes and their temporal, spatial, and topical dynamics is important.

Science Maps for Kids

In school, children learn about many different sciences. However, they never get to see science 'from above'. Hence, they have a very limited understanding regarding the sizes of different sciences or their complex interrelations.

How will they be able to answer questions such as: What intellectual travels did major inventors undertake for them to succeed? Why is mathematics necessary to succeed in almost all sciences? How do the different scientists build on each other's work? Or how would children find their place in science and where would it be?

Imagine a map of our collective scholarly knowledge that would hang next to the map of the world in each classroom. Imagine that students could not only travel our planet online, but also explore the web of knowledge. How would this change the way we learn, understand, and create?

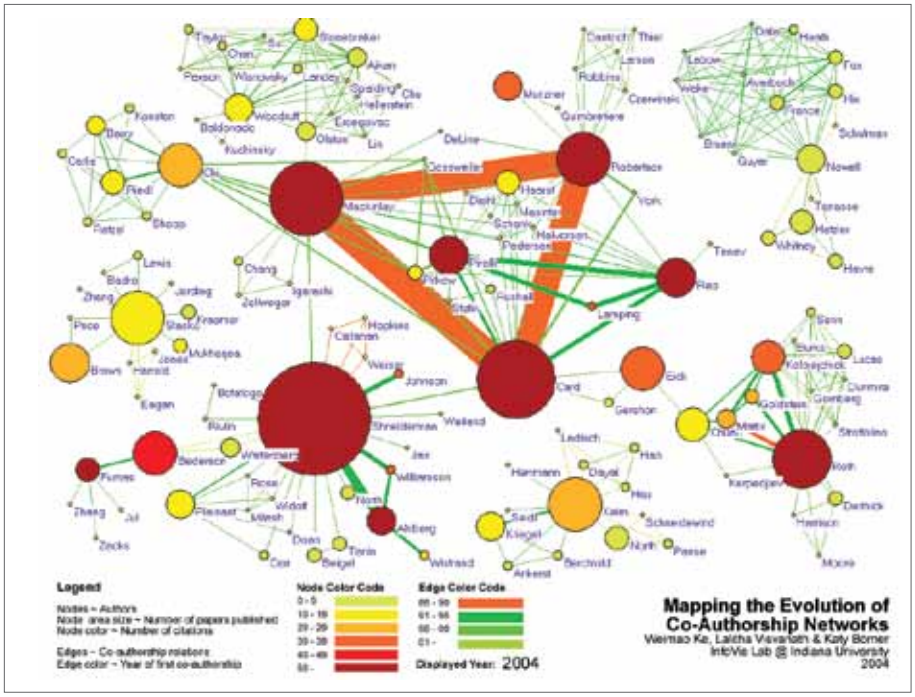


Figure 3.13: *Mapping the Evolution of Co-Authorship Networks.*
This is the last frame of an animation sequence that shows the evolution of authors (nodes) and their co-authorship relations (links) in the domain of information visualization. Node area size coded reflects the number of papers an author published, its color denoted the number of citations these papers received.

Link width equals the number of co-authorships and link color denotes the year of the first collaboration. Large, dark nodes are preferable. Large, light nodes indicate many papers that have not (yet) been cited. Shneiderman – working in a student dominated academic setting – has a very different co-author environment than Card, Mackinley, and Robertson, who are at Xerox PARC for most of the time captured here. An animated version of the above figure can be found at [http://iv.slis.indiana.edu/ref/iv04contest/Ke-Borner-Viswanath.gif.](http://iv.slis.indiana.edu/ref/iv04contest/Ke-Borner-Viswanath.gif)

Credit: (Ke et al., 2004)

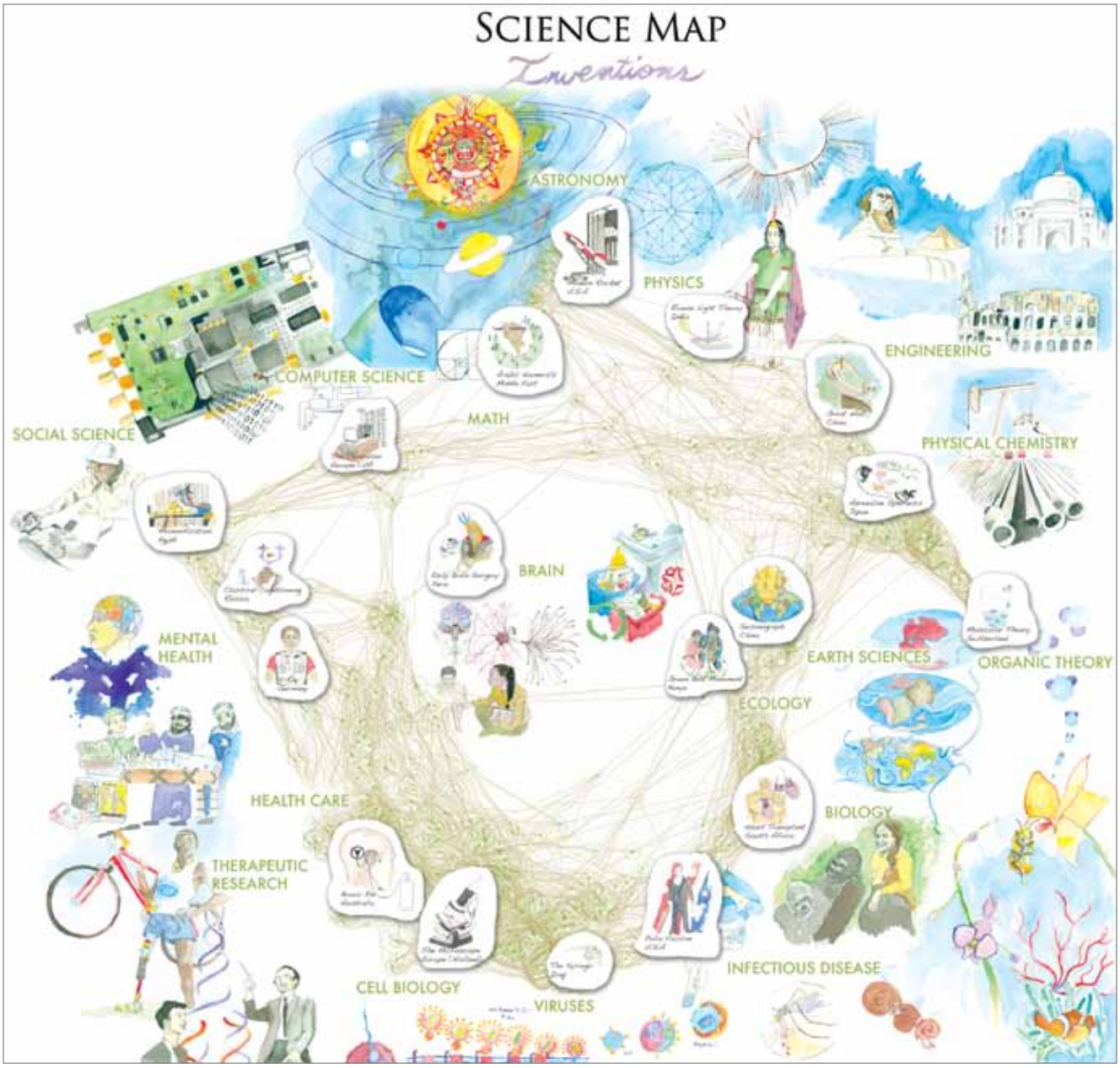


Figure 3.14: *Hands-on Science Maps for Kids.*
These maps invite children to see, explore, and understand science 'from above'. This map shows our world and the places where science is practiced or researched. Children and adults alike are invited to help solve the puzzle by placing major scientists, inventors, and inventions at their proper places. Look for the many hints hidden in the drawings to find the perfect place for each puzzle piece. What other inventors and inventions do you know? Where would your favorite science teachers and science experiments go? What area of science do you want to explore next? A large version of the map can be found at http://scimaps.org/flat/kids.

Credit: The base map is taken from the *Illuminated Diagram* display by Kevin Boyack, Richard Klavans, and W. Bradford Paley (Börner et al., 2009).

Section 4

Workshop Results

Main workshop results are presented here in form of: (1) user needs analyses conducted during Workshop I; and (2) project proposals generated by participants. The former is important to understand the concrete user needs, scientific culture (we attempted to keep the line of thinking as well as important terms and phrases), and desirable tools in a specific domain. The proposals present very specific solutions that could be implemented in the coming 5-10 years.

Not captured here, but available via the workshop web page at http://vw.slis.indiana.edu/cdi2008, are the:

- Brief Bio and (PR)2: Problems & Pitches – Rants & Raves. This form asked participants to briefly introduce themselves and to answer questions such as: What are your main interest(s) in attending the workshop? What information/ knowledge management needs do you have? What is the most insightful visualization of static or dynamic phenomena you know? What would you like to learn/ achieve at the workshop?

- Presentation slides of invited speakers.

Four parallel brainstorming sessions were held during Workshop I. Two aimed to identify user needs and possible "dream tools" for biomedical and ecological researchers. The other two groups attempted to do the same for SoS researchers and science policy makers. Each of the four groups was asked to answer five questions:

1. What data/services/expertise needs exist in the two domains?
2. Give a sketch of its visual user interface.
3. Provide two use scenarios.
4. Provide reasons why people would use it.
5. Envision how it may be sustainable.

The results of these brainstorming sessions—complemented by informal discussions during the workshop and in combination with the prior expertise of participants—formed the basis for the specific projects reproduced in Appendix D and discussed in this section, as well as for the general findings and recommendations given in the executive summary.

Biomedical/Ecological Research

The first brainstorming group was comprised of: Mark Gerstein (facilitator), Richard Bonneau, Neo Martinez, Stephen M. Uzzo (summarized results in this section) and Kei-Hoi Cheng, see Figure 4.1 on the next page. It focused on two subdisciplines: ecological and biomolecular (genomic) data systems and drew comparisons/synergies between them.

BRAINSTORMING: GENOMICS

Attendees focused primarily on problems of understanding relationships and annotations in genomic databases. This was considered a high priority, since problems of coordinating data organization and cleanup impede organizing the data. There are many disparate databases and people working on annotation. Demands on interrelationships amongst data along with large numbers of data entries are high. So a way to correlate all of these is needed to reveal relationships. This is further complicated by the fact that genes interact, the idea of one gene, one function is in the past. An interactive set of functions for genes is needed. Genomic databases have not caught up. Some concepts are tricky and it is complicated to fully deal with genomic functions as they relate to the molecules.

From the research standpoint, getting information about individual scientists' expertise into a common, shared community collaborative framework is needed to provide immediate access to experts (an intellectual NOT social "Myspace" for scientists). Such a resource would quickly allow the researcher to be identified as a domain expert and be targeted with specific kinds of questions, as well as interchange of ideas amongst other experts either within their domain or amongst an interdisciplinary cluster of domains. For instance, you should be able to ask an available expert about a protein you don't understand rather than doing 5 days of reading—publishing is not good enough. Also, there is a disincentive to make research available early, everything's a secret and embargos interfere with early communication of scientific research. But, like in the criminal justice system, "we're all in this together." Attendees in this group speculated that we need to move toward common goals more cooperatively than competitively.

Figure 4.1: *(from left to right) Mark Gerstein, agency representative, Stephen M. Uzzo, Weixia (Bonnie) Huang, Kei-Hoi Cheung, Neo Martinez, and Richard Bonneau.*

BRAINSTORMING: ECOLOGY

The group determined that the problems of gathering, storing and analyzing ecological data shares many of the same problems as that of biomedical data, such as genomics, with some important differences. A significant challenge is changing the way we look at ecological data so it can be federated and mined like other bioinformatic data. Hence, Table 4.1 compares and contrasts the needs of genomic data with that of ecological data. Metagenomics is, to some extent, providing opportunities to rethink how we look at ecological data both at the micro and macro scales. One of the key aspects of ecological data is the integration with GPS coordinates and traditional maps. Ecological data can be logged and mined like genomic data with the right data formatting and standardization. For instance, the analog for genomic interaction in ecological data is behavior and dynamics of individuals and populations. But, what is needed up front is an infrastructure to store and access interactions of ecological biodiversity. Then, we need to deal with the fact that ecological data use a format of publication that differs from genomic data. Much legacy data about species is in books and did not yet undergo optical character recognition (OCR), and hence is difficult to access. Biodiversity is more of an economic hook and driver than the framework for understanding ecological interactions that it needs to be.

How do you get people to look at the "long tail" rather than just at what is established and interesting? Sources must gain authority by peer review.

COMPARISON

Table 4.1 compares analogous problems in both genomic and ecological domains, hence the two columns. Arrows or crossing categories indicate a topic applying to both. Putting ecological data, systems and dynamics (including evolutionary) in comparative terms helps link the two and their needs and is a way that will hopefully help to recast the approach to solving bioinformatic problems in ecology and help stimulate interest in collaborative research between the two.

Table 4.1: *Comparing research challenges in genomics and ecology research*

GENOMICS		ECOLOGY
PROBLEMS		
Reconstructing accurate molecular descriptions internal to the cell and understanding how they respond to perturbations		Knowing the location and function of all the species on the planet (including descriptions, interactions, behavior, etc.)
DATA		
Activity (such as P peptide)		GPS coordinates and elevation
Function/sequencing > gene ontology data (e.g., cell adhesion or sub cellular location)		Habitat type
Phenotype >> causes cancer		Phylogeny
Internal interactions (yeast 2-hybrid, protein to protein)		Species traits (phenotype)
Metadata on methods of above		Function (what does it eat, where does it live?)
SERVICES		
Orthologs	<->	Construct phylogenic tree
Distribution of where it is expressed across organism		Provide spatial distribution
(Bi)clustering		List interactions (e.g., predator-prey), species status (e.g., endangered)
Functional prediction	<->	Perturbation analysis
Nearest neighbor, shortest path	<->	Generation & simulation network/interactions
Network inference		
Connections to R and Matlab		
Workflows to connect above		
EXPERTISE		
Software and database engineering		
Evolutionist (phylogenist)		
INTERFACE		
Loosely federated		Individual web services and mash-ups (workflows e.g., phylogenic and spatial)
API and web services (SOAP, JAVA RMI)		Google Earth tracks “map viewer”
Registry		
Firefox plug-ins (access via gene name)		
Genome browser	<->	Phylogenetic surfer (tree of life)
Network viewer		
USAGE		
Part in human genome (e.g., C1V associated w/disease)		Determine the impact of invasive species (location, determine interactions, generation and simulation to move to new location) and see what survives (“SimEarth”)

WHY USE IT?		
Tools already used, now easier		Basic research: see how phylogeny determines species distribution
Look for multiple data type support		Ecosystems managers and conservationists need to know this
Cope with data avalanche		
Prestige to institution, quality of data (citations)		
SUSTAINABILITY		
Open source/charity		
New mash-ups		
Enable comparisons/crosscheck		
Trust networks/confers respectability		
Enables collaboration		

PROJECT PROPOSALS

Based on the discussions, several concrete projects were proposed:

- Improve collaboration amongst biological researchers and develop **"Distributed Biomedical Data Repositories"** with common Web-based **"Interactive Visualizations"** to allow large-scale data to be easily browsed by researchers. (see page 65 in Appendix D)

- A **"Sequenomics Initiative"** is needed to allow mapping nucleic acid and protein sequence possibility spaces over time to create predictive, rather than descriptive, approaches to understanding molecular evolution. (see page 67 in Appendix D)

- Common data repositories and scalable data visualization interfaces of **"Environmental Sensor Data"** are needed to allow long-term trends to be charted, as well as provide opportunities for mash-ups with other data types, such as remote sensing, social/political, and other data sets. (see page 69 in Appendix D)

- An infrastructure/data repository to aid tracing the spatial and temporal **"Migration of Genetic Material in Metagenomic Data"** is needed to improve the understanding of diversity in microbial ecosystems and an evolutionary understanding of the "provenance" of gene structures in microbial biota. (see page 71 in Appendix D)

- Use semantic, social, and visual networks to build a **"cyberinfrastructure for enabling discovery and innovation in genome sciences and neurosciences"** to allow data and tools to be easily published, shared, discovered and linked for enabling scientific collaboration, knowledge discovery and technological innovation. (see page 60 and 80 in Appendix D)

- Using the **"Semantic Web to Integrate Ecological Data and Software"** would allow cataloging and integration of existing and evolving databases and software and create semantic web ontologies for accessing them through software "wrappers" that turn them into easily used and accessed Web services. This could also make ecological data available to the public in real-time and increase understanding and awareness of the biodiversity impacts of global change for policymakers, the public and formal and informal teaching and learning. (see page 66 in Appendix D)

- Developing intuitions and formalisms from "**social network analysis to understand the coupling of activity patterns between tissues in a diseased organ**" would make possible explorations of how genetic, epigenetic and environmental factors contribute to the disregulation of organs through a systems approach. (see page 63 in Appendix D)

SoS Research and Science Infrastructure

Analogous to the biomedical/ecological brainstorming sessions, the two SoS sessions focused on different users and their needs. One group selected NSF program officers as their main user group and tried to identify desirable features of science analysis tools. The other group focused on governmental officials and program managers, industrial partners, and venture capital investors and developed the plan for a "Science Observatory". Note that the tools, as well as the observatory, require an advanced data-algorithm-computing infrastructure in support of the study, management, and utilization of science. Both could potentially be useful for researchers, science policy makers, industry, government, and the general public at large.

BRAINSTORMING: SCIENCE ANALYSIS TOOLS

This brainstorming group was comprised of John T. Bruer (JSMF President), Kevin Boyack (Science Analysis and Mapping, SciTech Strategies), Lawrence Burton (NSF), Martin Storksdieck (Informal Learning Evaluation, ILI; now Director, Board on Science Education, National Academy of Sciences/National Research Council), Maria Zemankova (NSF) and Katy Börner (Scientometrics and Cyberinfrastructure Design, Indiana University), see Figure 4.2 and Appendix A for biographies.

DATA/SERVICE/EXPERTISE NEEDS

Science policy makers today face a steadily increasing flood of datasets and software tools relevant for their decision making. For example, Appendix C lists more than 100 relevant datasets and 110 tools.

The datasets come in diverse formats, with different temporal/geospatial/topical coverage, at many levels of granularity, in different quality, with different access rights and policies. The highest quality data, e.g., publication data by Thomson Reuters or Elsevier, has a high price tag. Previously, only Thomson Reuters data was used for bibliometric analyses. Today, Elsevier's Scopus and Google Scholar are viable alternatives for publication data. Patent, intellectual property claims, company profiles, and other technology data are also important for S&T studies.

The quality and functionality of existing tools is similarly complex. Some tools are open source and free, but might have little or no documentation. Others are commercially serviced 24/7, yet expensive to acquire and might be patented with little information on what particular algorithm and concrete parameter values are used. Typically, it takes several hours, or even days, to become familiar with a new tool. If the tool is unreliable, if it does not address the needs of science policy makers, or if it cannot be integrated into

Figure 4.2: *(from left to right) John T. Bruer, Kevin W. Boyack, Lawrence Burton, Martin Storksdieck, and Maria Zemankova.*

their daily decision making workflow, then the time spent on learning it is wasted. Given pressing deadlines, very few science policy makers experiment with new tools.

Specific insight needs identified during the brainstorming comprised:

- Given a set of new NSF proposals, place them in the topic space of existing proposals and awarded projects/papers/authors to identify potential reviewers/categories/novel ideas.
- Track changing expertise profiles of potential reviewers. This could be done automatically by analyzing the papers/proposals/projects/other scholarly results of reviewers. Alternatively or complementary, potential reviewers could be invited to submit/update expertise profiles.
- Identification of emergent research fields. Today, this is done using field surveys, expert workshops, and community feedback. Science of science tools might prove invaluable for identifying the structure and dynamics of a field, identifying bursts of activity, major experts, etc. in support of "forward thinking".
- Staying in contact with the scientific communities that a program officer cares about. Ideally, this is a 2-way communication that benefits both sides. Desirable is a system that provides **equal information to everybody**.
- Communicate NSF's success and budget needs to Congress— not via numbers, but via success stories.

The discussions made clear that there is an increasing demand for near real-time data analysis. Ideally, science policy makers would be able to monitor their research area or the progress of investigators as they happen. However, the data commonly used, e.g., NSF reports or investigators' annual progress reports, provide access to data that is one or more years old. For example, NSF's widely used *Science and Engineering Indicators* (National Science Board, 2008) are based on two-year-old citation data—it takes about two years before papers accumulate citations, plus another year for printing. As for NSF progress reports, an analogy to raising a child might be appropriate: Imagine the first time you hear or see your newborn is after one year—how likely would it be that your baby is still alive?

INTERFACE

The team extensively discussed science maps as visual interfaces to expertise, publications, patents, scholarly datasets and software, experiments, grants, etc. Just like in a Geographic Information System (GIS), different data layers could be turned on and off to facilitate comparisons, e.g.,:

- Topic coverage of solicitations vs. topic coverage of proposals.
- Topic coverage of proposals vs. topic coverage of already funded proposals.
- Topic coverage of proposals vs. expertise profiles of potential reviewers.
- Funding by NSF vs. funding by NIH, as shown in Figure 3.12 on page 16.
- NIH funding vs. Medline publications resulting from it.

Topical analysis over full text was identified as highly relevant—particularly across data silos, e.g., solicitations, proposals, resulting publications, and expert profiles.

USERS AND USE SCENARIOS

Besides the comparisons listed above, two other scenarios were identified:

- Anybody should be able to upload a 1-page proposal summary to see related projects, major experts, etc. based on link analysis and text analyses. This might reduce the number of proposals that have a high overlap with work that is already funded or proposals with insufficient expertise of investigators.
- Industry might like to cause "pull" on science by promising dollar amounts for certain results, e.g., see NetFlix competition (http://www.netflixprize.com). This would create a valuable bridge between academia and industry. Industry might also use maps of scientific results to find experts in a certain topic or to harvest research results for product development.

SUSTAINABILITY

Ideally, the system would be "mothered and fathered" by scientific communities. In fact, each scientific community might feel responsible for their own 'continent' on the map of science: The society for physics, astronomy, etc. with the usual boundary discussions. The American Association for the Advancement of Science (AAAS) and other national science societies might like to play a major role as well.

A base map of science can be created from Thomson Reuters/ Scopus data, see *Maps of Science: Forecasting Large Trends in Science, 2007* by Richard Klavans, Kevin Boyack and Figure 4.3 on the next page. Interactive maps with different data overlays can be explored at http://mapofscience.com. However, it is desirable to fold in other data sources, e.g., genes, proteins, subject data, books, experiments, protocols, jobs, etc., relevant for the different sciences.

It is highly desirable not to have this map owned by one corporate entity. That is, the team uniformly agreed that we should not simply wait until Google, Amazon or eBay generates it for us. Ideally, no commercial interest would interfere with the mission to generate the highest accuracy, highest coverage (all languages and all fields of scholarly endeavor) map. Seed funding by NSF, NIH, Mellon Foundation, MacArthur Foundation, and/or Ford Foundation, combined with a major endowment, would provide the financial stability to serve this map to the world. A 'service' model, such as the one used to generate global and local weather forecasts, might work as well.

Related efforts are Science.gov (http://science.gov) that searches over 40 databases and over 2000 selected Web sites, offering 200 million pages of authoritative U.S. government science information, including research and development results and WorldWideScience.org (http://worldwidescience.org), a global science gateway to national and international scientific databases and portals. Both services are actively developed and maintained by the Office of Scientific and Technical Information (OSTI), an element of the Office of Science within the U.S. Department of Energy. Maria Zemankova suggested collaborating with the National Library of Medicine (NLM), which successfully serves Medline (19 million publications in the biomedical domain) and is funded by NIH.

Plus, it might be beneficial to collaborate with publishers—finding reviewers for proposals has much in common with finding reviewers for papers; solicitation writing has analogies with writing calls for special issues.

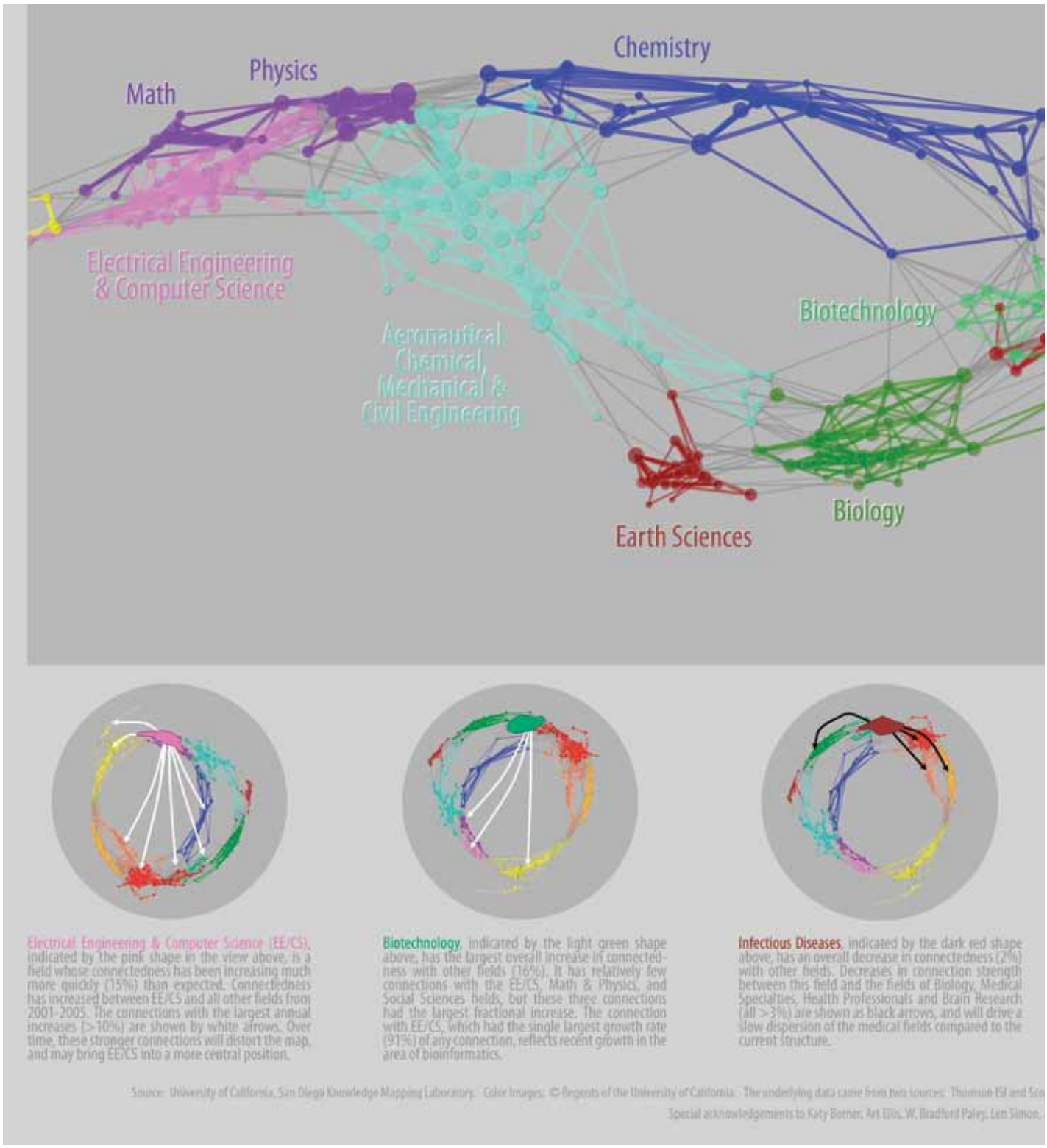
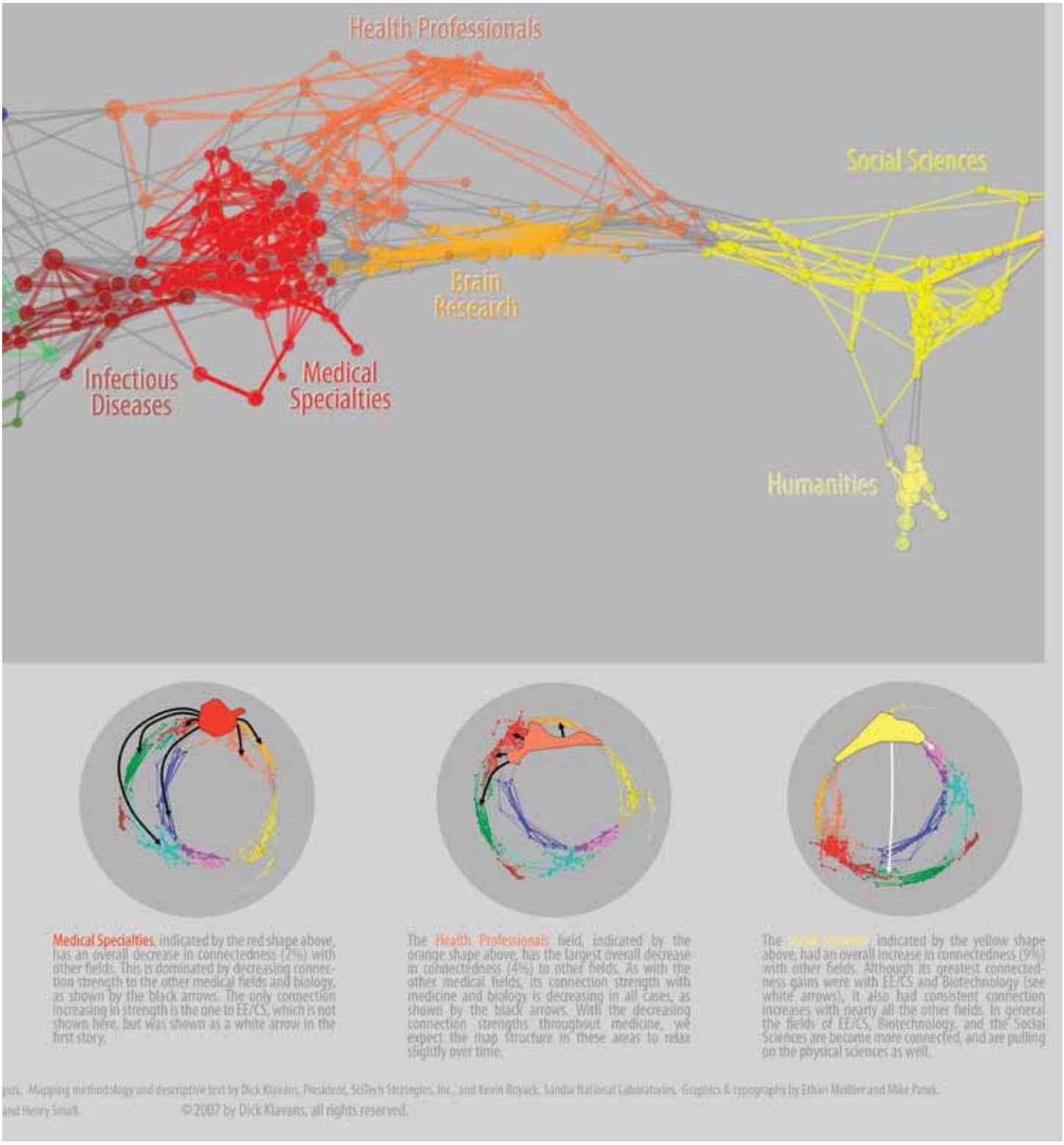


Figure 4.3: *A forecast of how the structure of science may change in the near future was generated by evaluating the change in the connectedness of various regions of the map over time (Klavans & Boyack, 2007). Groups of journals characterizing the disciplines on the map were defined using a metric based on a combination of the bibliographic coupling of references and keyword vectors.*



A three-dimensional layout of the disciplines (groups of journals) places those disciplines on a sphere, which is then unfolded using a Mercator projection to give a two-dimensional version of the map. This most recent map of science is based on the largest set of scientific literature yet mapped: 7.2 million papers and over 16,000 separate journals, proceedings, and series from a five year period, 2001-2005.

BRAINSTORMING: SCIENCE OBSERVATORY

The second brainstorming group included Peter van den Besselaar, Luís M. A. Bettencourt (summarized results in this section), Stefan Hornbostel, Masatsura Igami, and Alex Soojung-Kim Pang (facilitator), see Appendix A for biographies. The summary below has much overlap with the 2nd best idea write-ups by the same authors, see Appendix D.

This team's idea was to build a system that would encapsulate metrics of scientific and technological performance, together with biographic data, and potential drivers. The system would also be able to provide analysis tools for return on investment, useful for planning by: i) governmental officials and program managers; ii) industrial partners; and iii) venture capital investors.

For lack of a better word, we referred to this system as the 'Global Science Observatory (GSO)' – see Luís M. A. Bettencourt's 2nd best idea write up on page 72 in Appendix D – although we have also toyed with the metaphor of this system being the analog of a "global climate model for science". Given the shortcomings of the latter in its primary area of application, we felt the analogy may weaken the argument; but many of the difficulties in building one or the other system are certainly suggestive of interesting analogies.

DATA/SERVICE/EXPERTISE NEEDS

Data contained in the system must be both qualitative and quantitative. Specifically, it should include information on bibliometrics, biographies (careers, mobility and impact), funding (from states, industry, venture capital, etc), social networks (migration, citation and co-authorship), patents and national and institutional presence in various fields (publications and citations). It should also include more informal measures of interest in science and technology, such as usage data (downloads), media coverage, and educational impacts, wherever these can be quantified. The data should also include qualitative information about "Big Events" (discoveries and breakthroughs), field work, user groups and other informal utilization and references to science and technology results.

Crucially, the system would also be a laboratory for the development of models for quantifying and predicting the growth and development of S&T areas. Such models should make falsifiable predictions with quantified degrees of uncertainty (so they should be probabilistic), accounting for unknowns as well as risks and contingencies.

Figure 4.4: *(from left to right) Luís M. A. Bettencourt, Masatsura Igami, Peter van den Besselaar, Stefan Hornbostel, and Alex Soojung-Kim Pang.*

At present, there are no established approaches to modeling quantitative developments in science and technology. Several approaches have, however, started to use dynamical populations that account for the exchange of ideas, learning and training as the sources for increases in adoption and use of scientific ideas. Other studies analyze network structure of collaboration, citation, co-location, etc. to identify factors responsible for knowledge spillovers. A future integration of these kinds of approaches, with probabilistic formulation that can account for uncertainty and risk validated on new data, may generate the first models capable of reliable prediction of the course of science and technology.

INTERFACE

The greatest difficulty in designing an interface for the system would be to provide a high-level instrument for qualitative analysis (which may be of interest to science policy makers, venture capitalists, etc.), as well as being able to provide a transparent instrument for scientific analysis, containing all detail in the data and models.

This may require several different interfaces or at least a layered architecture, allowing the user to access the system at different levels of detail. We imagined a system with some sort of dashboard that would allow a simple, more qualitative interface for "end users," and a full blown detailed analysis for "power users". Clearly, details of such implementation would still have to be worked out.

TWO USE SCENARIOS

These are, in some sense, covered in point 2). In addition, it is important that different kinds of users, looking for very different forms of return on investment, can take advantage of the analyses afforded by the system. For example, a program manager working for a state or federal agency may place emphasis on education and training returns on investment, while a venture capitalist may look for a purely financial reward. An industrial analyst may lie somewhere in the middle, and search for comparative competitive analysis.

All such uses are predicated on progress in estimating the evolution of fields and the impact of funding decisions, etc. Such models should give an account of the evolutionary life cycles of fields from discovery to common place.

The system should also enable international and organizational competitive analyses (where are competencies in a given topic area?) and thus help identify opportunities for global cooperation.

USES AND USERS

There are two main reasons for the system to exist: 1) to provide a well curated repository of data and empirically sanctioned analyses of return on investment in S&T; and 2) to close the loop between drivers and consequences of scientific and technological development, thus enabling new predictive approaches to be tested and validated in a timely manner.

The first objective is an application, but may well be the single reason that would make a system of this nature sustainable. However, without progress in models that would allow projections into the future and the exploration of scenarios for investment etc., the system would remain purely an observatory, and would be a limited instrument for science policy.

SUSTAINABILITY

The system could exist as a non-profit, stand-alone organization or indeed as a for-profit company providing strategic analyses for investments in innovation in S&T. In its grandest implementation, it could take the status of a global observatory/laboratory, such as CERN, ESO, ITER, etc., funded as a joint international collaboration and serving the interests of many governments, firms, and individuals worldwide, at a much lower cost than the existing efforts. As the number of users increases, the costs per user will decrease.

PROJECT PROPOSALS

A set of concrete project proposals for advancing SoS research, specifically, and the scholarly research infrastructure, in general, were listed in the Executive Summary and are detailed in Appendix D. They comprise:

- A decentralized, free **"Scholarly Database"** is needed to keep track, interlink, understand and improve the quality and coverage of Science and Technology (S&T) relevant data. (see also page 76 and 77 in Appendix D)

- Science would benefit from a **"Science Marketplace"** that supports the sharing of expertise and resources and is fueled by the currency of science: scholarly reputation. (see page 74 in Appendix D) This marketplace might also be used by educators and the learning community to help bring science to the general public and out of the "ivory tower". (see page 89 in Appendix D)

- A **"Science Observatory"** should be established that analyzes different datasets in real-time to assess the current state of S&T and to provide an outlook for their evolution

under several (actionable) scenarios. (see page 72 in Appendix D)

- **"Validate Science Maps"** to understand and utilize their value for communicating science studies and models across scientific boundaries, but also to study and communicate the longitudinal (1980-today) impact of funding on the science system. (see page 81 in Appendix D)

- Design an easy to use, yet versatile, **"Science Telescope"** to communicate the structure and evolution of science to researchers, educators, industry, policy makers, and the general public at large. (see page 87 in Appendix D) The effect of this (and other science portals) on education and science perception needs to be studied in carefully controlled experiments. (see page 88 in Appendix D)

- **"Science of (Team) Science"** studies are necessary to increase our understanding and support the formation of effective research and development teams. (see page 78 and 82 in Appendix D).

- **"Success Criteria"** need to be developed that support a scientific calculation of S&T benefits for society. (see also page 88 in Appendix D)

- A **"Science Life"** (an analog to Second Life) should be created to put the scientist's face on their science. Portals to this parallel world would be installed in universities, libraries and science museums. (see page 80 in Appendix D) The portals would be "fathered and mothered" by domain, as well as learning experts. Their effect on education and science perception should be rigorously evaluated in carefully controlled experiments and improved from a learning science standpoint. (see page 91 in Appendix D)

Note that there are many synergies among these proposals that make the collective set much greater than the sum of the parts. High-quality and coverage of science and technology (S&T) relevant data can be shared via the science marketplace, that would be free for research, industry, educators, and the general public. Anybody would have easy access to tax financed scientific results. The science observatory would support the large-scale analysis and modeling of S&T. Results might be communicated via tables, trend graphs, or science maps and science telescopes. Science of (team) science studies would improve our understanding of the inner workings of science and help optimize tools and infrastructures that aim to improve S&T. Success criteria for S&T would help calculate and communicate the impact and importance of S&T for a nation, institution, and individual. Science portals would make science tangible, interesting and relevant for a large audience. The effect of novel means to communicate science to the general public would be rigorously evaluated in carefully controlled experiments and improved as needed.

Section 5

Prospective Timeline

An interactive timeline assembly was lead by Alex Soojun-Kim Pang with the intent to identify incremental and transformative socio-technical advances that will most likely make possible qualitatively new tools in biomedical/ecological, SoS and other areas of research and practice. The timeline, see Figure 5.1, together with a discussion of major emergent clusters as compiled by Stephen M. Uzzo, are presented here.

A closer examination of the timeline shows results in nine major clusters, as indicated by the color-coding on the timeline. These nine topic areas divide into two general regions of interest. The first has to do with the way institutions conduct research in relationship to each other and society. The other is how we translate, manipulate, understand and communicate data sets. Interestingly, these two areas do not seem to break along the lines of biomedical/ecological versus SoS areas, but seem to overlap significantly in both.

COLLABORATION

In the near term, it was determined that there was a need to identify and facilitate collaboration, as well as identify ways to leverage the scale of research in a sustainable way. This should especially focus on electronic methods, with the ultimate goal being to match electronic interaction with live human interaction.

SUSTAINABILITY

Sustaining research over the long-term will require significant effort and funding. For many of the initiatives recommended in this report, this means broad collaboration amongst institutions and funding agencies in terms of both leveraging existing infrastructure through improved interoperability and creating a new generation of shared infrastructure. Sustaining these efforts also means creating marketplaces for ideas, accountability and validation of the usefulness of spending initiatives, and ambitious long-range goals: for example, engineering cellular processes or science of science tools and observatories. An important aspect of maintaining long-term goals is providing incentives and rewards beyond both profits in the commercial sector, and publication and grants in the academic sector. Collaboration is key.

COMPETITIVNESS

Balancing collaboration with an entrepreneurial spirit stimulates innovation and helps create marketplaces for ideas. Institutions will need to be aggressive in maintaining their brain-trusts as new ideas and technologies emerge. Scientometric tools are beginning to be able to identify nascent research areas. Making these tools more predictive will allow better decision-making for funding.

ACCURACY/DATA STANDARDS

There needs to be a focus on setting data standards, which include the way such data are tagged, stored, retrieved, viewed and transported in data processing systems. This should not be an afterthought, but part of the data discovery and organization process. Also, because of the complexity of large-scale data sets, and dependencies of data discovery of many researchers on data sets created by others, the initial accuracy of baseline data sets is critical. Mechanisms to maintain data reliability and prevent accuracy from drifting is important, as a variety of ways are developed for visualizing and manipulating data sets.

INCREASE UNDERSTANDING (INCLUDING PUBLIC UNDERSTANDING)

The notion of making science understandable to the broadest possible audience is considered a priority. One aspect of this is broadening the understandability of data and models for scientists from multiple disciplines. The nature of interdisciplinarity demands common tools and vocabulary. The other is to make data and models understandable to non-scientists in government, education and the general citizenry. This will require that educational methods be developed and tested and are accessible and engaging in a variety of modalities and educational environments. Ultimately, in the long term, such approaches must be integrated across the board in K-20 education.

USER INTERFACE/DATA MANIPULATION TOOLS

The tools for accessing and manipulating data sets and the interface to these tools need to be more usable and have the ability to evolve to keep up with the kinds of problems these data are leveraged against. There is a sense that the data are accumulating much more rapidly and that computer tools are not keeping pace, but also that new kinds of tools are needed. The next generation of tools and interfaces to these tools

need to be much more powerful, flexible, multidimensional, and intuitive.

VISUAL (SPATIAL AND DYNAMIC) REPRESENTATION

Perhaps the approach most favored by attendees and seen as the most important way toward complex data understanding is visualization. While this has been an area of research and development for a while, the effectiveness of the traditional static graph needs to be addressed, along with incorporating more dynamic ways for representing such graphs. Also, the problem of representing, interpreting, and interacting with heterarchical and hierarchical graphs needs to be better addressed. Semantic webs need to be leveraged against this problem along with other ways of visualizing and manipulating dynamic interacting data sets. There is also a call for providing tools with which researchers can easily manipulate, change, and create whole conceptual models as they do the research to improve opportunities for data discovery.

MODELING

There needs to be a refinement of dynamic models used to organize and understand interactions in evolving systems. This applies primarily to molecular systems such as the genome, but dynamic models should also be generalizable to include ecosystems and relationships amongst research communities.

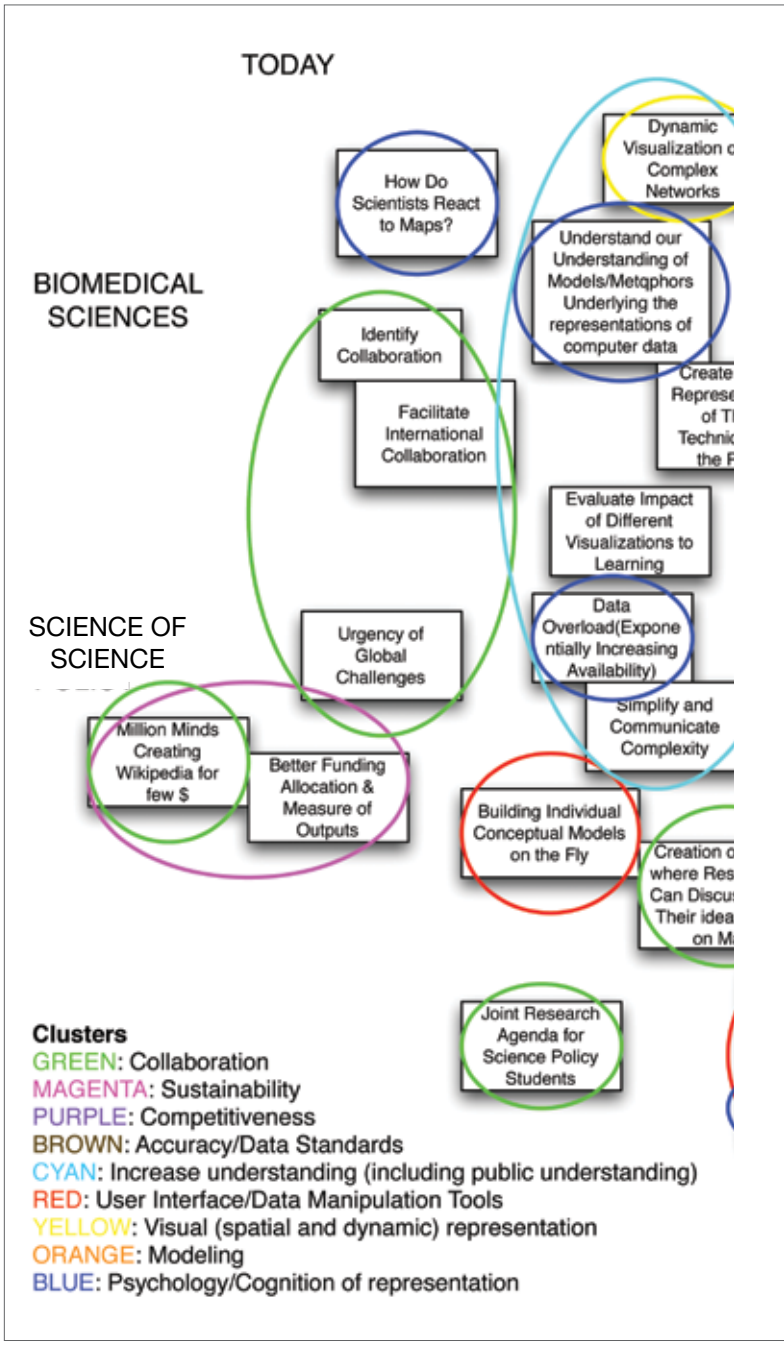
PSYCHOLOGY/COGNITION OF REPRESENTATION

Finding effective approaches to representing complex spatial and dynamic data sets is an ongoing problem. Significant effort needs to be expended to quantify the effectiveness of various approaches, what cognitive domains do they stimulate, and how does the resulting knowledge synthesize with science understanding? The development of analytical and visualization tools that truly exploit human visual perception and cognition is highly desirable. While the primary pathway of interest is visual, other modalities should be considered, particularly in light of broader educational needs. Finally, just as the 1990s were considered "the Decade of the Brain," the next ten years might be considered the "Decade of the Mind"; provided the resources are allocated to expand our understanding of how the mind works. Leveraging our increasing understanding of both learning and thought to the understanding of complex processes in nature may be the key to discovery in both.

CONCLUDING REMARKS

Taking a user-centric approach, the report aimed to capture recent developments in (1) biomedical/ecological research and (2) science of science research but also the need for (3) a national science infrastructure.

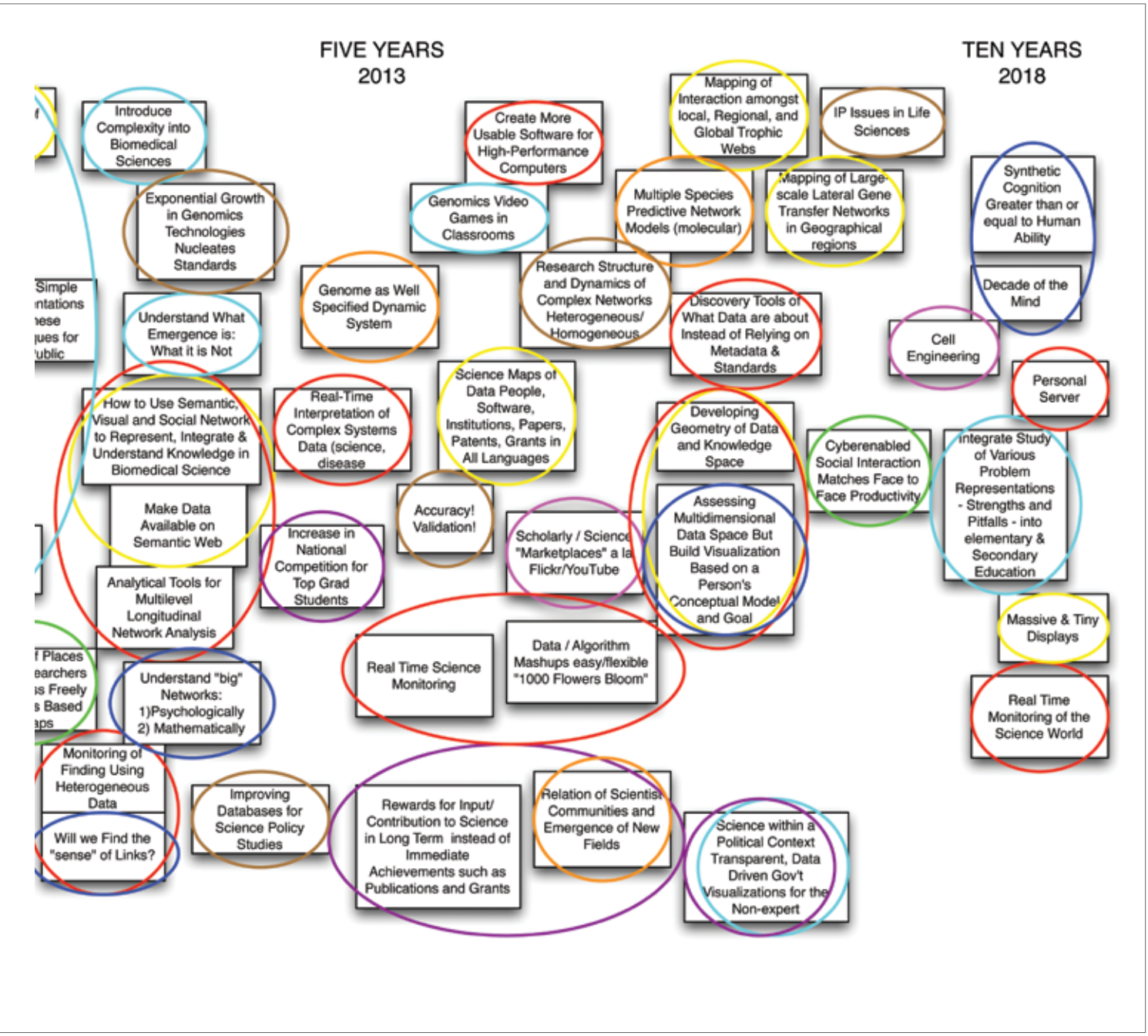
Based on these needs, the expertise of the workshop participants, and socio-technical opportunities, a concrete set of projects were outlined (see also Appendix D) that would lead to major progress in these three areas.



The printed version of this report will be shared with all workshop participants, governmental and private funding organizations, and other interested parties. A digital copy is available at http://vw.slis.indiana.edu/cdi2008.

It is our hope that many researchers, educators, and practitioners identify with the needs presented here and recognize the value of the proposed work and projects. We believe many of these suggested tools will come into existence over the next 10 years. They will be funded by government, as well as private and industry sources. Ideally, the proposed national science infrastructure will not be "owned" by one commercial entity but "mothered and fathered" by scientific communities.

Figure 5.1: *Timeline display of expected socio-technical advances is organized in a two-dimensional grid with time going linearly from the left (today) to right (within 10 years) and vertical categories: biomed (top), general (middle) and science of science (bottom).*



References

Abe, Takashi, Hideaki Sugawara, Shigehiko Kanaya, Toshimichi Ikemura. (2006). Sequences from Almost All Prokaryotic, Eukaryotic, and Viral Genomes Available Could be Classified According to Genomes on a Large-Scale Self-Organizing Map Constructed with the Earth Simulator. *Journal of the Earth Simulator, 6*, 17-23.

Association of Research Libraries. (2006). *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering: A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe*. Washington, DC. http://www.arl.org/bm~doc/digdatarpt.pdf (accessed on 12/30/2009).

Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, Margaret H. Wright. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure. *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. National Science Foundation. http://www.nsf.gov/od/oci/reports/atkins.pdf (accessed on 12/21/2009).

Bettencourt, Luís M. A., Ariel Cintrón-Arias, David I. Kaiser, Carlos Castillo-Chávez. (2006). The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models. *Physica A*(364), 513-536. http://arxiv.org/abs/physics/0502067 (accessed on 11/20/2008).

Bettencourt, Luís M. A., David I. Kaiser, Jasleen Kaur. (2009). Scientific Discovery and Topological Transitions in Collaboration Networks. *Journal of Informetrics, 3*, 210-221.

Bettencourt, Luís M. A., David I. Kaiser, Jasleen Kaur, Carlos Castillo-Chavez, David E. Wojick. (2008). Population Modeling of the Emergence and Development of Scientific Fields. *Scientometrics, 75*(3), 495-518.

Börner, Katy, Chaomei Chen, Kevin W. Boyack. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology (ARIST)* (Vol. 37, Ch pp. 179-255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Börner, Katy, Fileve Palmer, Julie M. Davis, Elisha F. Hardy, Stephen M. Uzzo, Bryan J. Hook. (2009). Teaching Children the Structure of Science. *Proceedings of SPIE-IS&T Visualization and Data Analysis Conference, January 19-20* San Jose, CA: SPIE, 7243, pp. 1-14.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: MIT Press.

Boyack, Kevin W., Katy Börner, Richard Klavans. (2009). Mapping the Structure and Evolution of Chemistry Research. *Scientometrics, 79*(1), 45-60.

Burns, Gully APC, Herr II, Bruce W., Newman, David, Ingulfsen, Tommy, Pantel, Patric & Smyth, Padhraic. (2007). A Snapshot of Neuroscience: Unsupervised Natural Language Processing of Abstracts from the Society for Neuroscience 2006 Annual Meeting. Society for Neuroscience 2007 Annual Meeting. http://scimaps.org/maps/neurovis (accessed 1/21/2010).

Cattuto, Ciro, Alain Barrat, Wouter Van den Broeck. (2009). *Contact Patterns, Part 2 SocioPatterns at the 25C3 Conference*. http://www.sociopatterns.org (accessed on 12/30/2009).

Christakis, Nicholas A., James H. Fowler. (2007). The Spread of Obesity in a Large Social Network Over 32 Years. *New England Journal of Medicine, 357*(4), 370-379.

Colizza, Vittoria, Alain Barrat, Marc Barthélemy, A.J. Valleron, Alessandro Vespignani. (2007). Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLoS Medicine, 4*(1), e13.

Colizza, Vittoria, Alain Barrat, Marc Barthélemy, Alessandro Vespignani. (2007). Epidemic Modeling in Complex Realities. *Comptes Rendus Biologies, 330*(4), 364-374.

de Rosnay, Joël, Robert Edwards (Trans.). (1979). *The Macroscope: A New World Scientific System*. New York: Harper & Row Publishers, Inc.

de Solla Price, Derek John. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Douguet, Dominique, Huei-Chi Chen, Andrey Tovchigrechko, Ilya A. Vakser. (2006). DOCKGROUND Resource for Studying Protein-Protein Interfaces. *Bioinformatics, 22*(21), 2612-2618.

Dunne, J.A., R.J. Williams, N.D. Martinez. (2004). Network Structure and Robustness of Marine Food Webs. *Marine Ecology Progress Series, 273*, 291-302. http://www.int-res.com/articles/meps2004/273/m273p291.pdf (accessed on 12/30/2009).

Fermi, G., M.F. Perutz. (1984). The Crystal Structure Of Human Deoxyhaemoglobin At 1.74 Angstroms *Journal of Molecular Biology 175*, 159-174.

Fowler, James H., Nicholas A. Christakis. (2008). Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis Over 20 Years in the Framingham Heart Study. *British Medical Journal, 337*, a2338.

Garfield, Eugene. (1980). The Epidemiology of Knowledge and the Spread of Scientific Information. *Current Contents, 35*, 5-10.

Gerstein, Mark, Werner Krebs. 2000. The Morph Server: A Standardized System for Analyzing and Visualizing Macromolecular Motions in a Database Framework. Nucleic Acids Res., 28: 1665-75. http://molmovdb.mbb.yale.edu/molmovdb (accessed 1/21/2010).

Goffman, W. (1966). Mathematical Approach to the Spread of Scientific Ideas: The History of Mast Cell Research. *Nature, 212*, 449-452.

Goffman, W., V.A. Newill. (1964). Generalization of Epidemic Theory: An Application to the Transmission of Ideas. *Nature, 204*, 225-228.

Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, Albert-László Barabási. (2007). The Human Disease Network: Supporting Information-Data Supplement. *PNAS, 104*(21), 8685-8690. http://www.pnas.org/content/104/21/8685/suppl/DC1#F13 (accessed on 1/5/2010).

Griliches, Zvi. (1990). Patent Studies as Economic Indicators: A Survey. *Journal of Economic Literature, 28*(4), 1661-1707.

Han, J.D., N. Bertin, T. Hao, D.S. Goldberg, G.F Berriz, Zhang L.V., D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal. (2004). Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network. *Nature, 430*(6995)), 88-93.

Harnad, Stevan. (2003). *Online Archives for Peer-Reviewed Journal Publications*. Routledge. http://www.ecs.soton.ac.uk/~harnad/Temp/archives.htm (accessed on 10/17/2008).

Hey, Tony, Stewart Tansley, Kristin Tolle (Eds.). (2009). *Jim Gray on eScience: A Transformed Scientific Method*. Redmond, WA: Microsoft Research.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein. (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science, 302*, 449-453.

Jones, Alwyn. 2010. O. http://xray.bmc.uu.se/alwyn/Distribution/ov11_12/ov12.html (accessed 1/21/2010).

Ke, Weimao, Katy Börner, Lalitha Viswanath. (2004, Oct 10-12). Major Information Visualization Authors, Papers and Topics in the ACM Library. *IEEE Information Visualization Conference* Houston, TX. http://iv.slis.indiana.edu/ref/iv04contest/Ke-Borner-Viswanath.gif (accessed 11/19/2008).

Klavans, Richard, Kevin W. Boyack. 2007. *Maps of Science: Forecasting Large Trends in Science*. Berwyn, PA and Albuquerque, NM. Courtesy of Richard Klavans, SciTech Strategies, Inc. In Katy Börner & Julie M. Davis (Eds.), 3rd Iteration (2007): The Power of Forecasts, *Places and Spaces: Mapping Science* http://scimaps.org (accessed on 01/05/2010).

Kuhn, Thomas Samuel. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kutz, Daniel. (2004). Examining the Evolution and Distribution of Patent Classifications. *Information Visualisation Conference*, London, UK: pp. 983-988.

Lee, I., Date S.V., Adai A.T., Marcotte E.M. (2004). A Probabilistic Functional Network of Yeast Genes. *Science, 306*(5701), 1555-1558.

Luscombe, N.M., M.M. Babu, H. Yu, M. Snyder, S.A. Teichmann, M. Gerstein. (2004). Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes. *Nature 431*, 308-312.

Marris, Emma. 2006. 2006 Gallery: Brilliant Display. *Nature*, 444, 985.

National Research Council Board on Life Sciences. (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC: The National Academies Press. http://books.nap.edu/openbook.php?record_id=11902&page=1 (accessed on 12/30/2009).

National Science Board. (2008). *NSF Division of Science Resources Statistics: Science and Engineering Indicators, 2008*. http://www.nsf.gov/statistics/seind08/ (accessed on 12/21/2009).

National Science Foundation. (2009). *Science of Science and Innovation Policy (SoS): Funding Opportunities and Program Guidelines*. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501084 (accessed on 01/05/2010).

National Science Foundation Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. Washington, DC http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf (accessed on 01/05/2010).

National Science Foundation, National Science Board. (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Washington, DC http://www.nsf.gov/pubs/2005/nsb0540/ (accessed on 12/30/2009).

National Science Foundation, Office of Cyberinfrastructure. (2009). *Software Infrastructure Committee Presentation*. https://nsf.sharepointspace.com/acci_public/ACCI Task Force Presentations December 2 2009/Software Task Force.pdf (accessed on 12/21/2009).

Nature Publishing Group. (2008). Nature Special: Big Data. *Nature, 455*(1) http://www.nature.com/news/specials/bigdata/index.html#editorial (accessed on 12/30/2009).

Parashar, Manish. (2009). Transformation of Science Through Cyberinfrastructure: Keynote Address. *Open Grid Forum* Banff, Alberta, Canada, October 14. http://www.ogf.org/OGF27/materials/1816/parashar-summit-09.pdf (accessed 12/21/2009).

Peterson, Michael, Gary Zasman, Peter Mojica, Jeff Porter. (2007). *100 Year Archive Requirements Survey*. Storage Networking Industry Association. http://www.snia.org/forums/dmf/programs/ltacsi/forums/dmf/programs/ltacsi/100_year/100YrATF_Archive-Requirements-Survey_20070619.pdf (accessed on 12/30/2009).

Popper, Karl R. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

President's Council of Advisors on Science and Technology. (2007). *Leadership Under Challenge: Information Technology R&D in a Competitive World: An Assessment of the Federal Networking and Information Technolgy R&D Program*. Washington, DC http://www.nitrd.gov/Pcast/reports/PCAST-NIT-FINAL.pdf (accessed on 12/30/2009).

Sayle, Roger. 1999. RasMol 2.6. http://www.openrasmol.org/Copyright.html (accessed 1/21/2010).

Schrödinger. 2010. PyMOL. http://www.pymol.http://www.pymol.org (accessed 1/21/2010).

Shiffrin, Richard, Katy Börner. (2004). Mapping Knowledge Domains. *PNAS, 101*(Suppl 1), 5183–5185.

Symyx Solutions, Inc. 2009. Chime Pro. http://www.symyx.com/products/software/cheminformatics/chime-pro/index.jsp (accessed 1/21/2010).

Thomas, James J., Kristin A. Cook. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. Washington, DC: IEEE Computer Society.

Tufte, Edward. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Wilson, E.O. (1998). *Consilience: The Unity of Knowledge*. New York: Knopf. 53.

Woese, C.R. (2004). A New Biology for a New Century. *Microbiology and Molecular Biology Reviews*, 68(2), 173-186.

Xia, Y., H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao, M. Gerstein. (2004). Analyzing Cellular Biochemistry in Terms of Molecular Networks. *Annu Rev Biochem*, 73, 1051-1087.

Yoon, I., R.J. Williams, E. Levine, S. Yoon, J.A. Dunne, N.D. Martinez. (2004). Webs on the Web (WoW): 3D Visualization of Ecological Networks on the WWW for Collaborative Research and Education. *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging, Visualization and Data Analysis* San Jose, CA: SPIE, 5295, pp. 124-132.

Zucker, Lynne G. and Michael R. Darby. (Forthcoming). Legacy and New Databases for Linking Innovation to Impact. In Fealing, Kay Husbands, Julia Lane, John Marburger, Stephanie Shipp, Bill Valdez (Eds.). Science of Science Policy: Handbook. http://scienceofsciencepolicyhandbook.net/contact.html (accessed 1/21/2010).

Appendices

Table of Contents

Appendices

Appendix A: Biographies of Participants

Organizers

Luís M. A. Bettencourt obtained his Ph.D. from Imperial College, University of London in 1996 for work on critical phenomena in the early universe, and associated mathematical techniques of Statistical Physics, Field Theory and Non-linear Dynamics. He held post-doctoral positions at the University of Heidelberg, Germany, as a Director's Fellow in the Theoretical Division at the Los Alamos National Laboratory (LANL), and at the Center for Theoretical Physics at MIT. In 2000, he was awarded the distinguished Slansky Fellowship at LANL for Excellence in Interdisciplinary Research. He has been a Staff Member at LANL since the Spring of 2003, first at the Computer and Computational Sciences Division (CCS), and since September 2005 in the Theoretical Division (T-7: Mathematical Modeling and Analysis). He is also External Research Faculty at the Santa Fe Institute (since July 2007), and a Research Professor at the Department of Mathematics and Statistics and the School of Human Evolution and Social Change at Arizona State University (ASU). Dr. Bettencourt carries out research in the structure and dynamics of several complex systems, with an emphasis on dynamical problems in biology and society. Currently he works on real-time epidemiological estimation, social networks of human communication, distributed sensor networks, information processing in neural systems and urban organization and dynamics. He maintains many diverse multidisciplinary collaborations at LANL, and beyond at the Santa Fe Institute, ASU, Princeton University, Harvard, MIT, Statistical and Applied Mathematical Sciences Institute (SAMSI), and University of California, Davis. He is a member of advisory committees for international conferences and referees for journals in physics, mathematics and computer science, and for international fellowship programs. In the last three years, he has co-advised four Ph.D. theses in epidemiology, complex networks, computational neuroscience and statistical mechanics. He is also a consultant for the Office Science and Technology Information of the U.S. Department of Energy on the subject of Scientific Innovation and Discovery.

Katy Börner is the Victor H. Yngve Professor of Information Science at the School of Library and Information Science, Adjunct Professor in the School of Informatics and Computing, Adjunct Professor at the Department of Statistics in the College of Arts and Sciences, Core Faculty of Cognitive Science, Research Affiliate of the Biocomplexity Institute, Fellow of the Center for Research on Learning and Technology, Member of the Advanced Visualization Laboratory, and Director of the Information Visualization Lab and the Cyberinfrastructure for Network Science Center (http://cns.slis.indiana.edu) at Indiana University. She received her Ph.D. in Computer Science from the University of Kaiserslautern, Germany in 1997. She is a curator of the Places & Spaces: Mapping Science exhibit (http://scimaps.org). Her research focuses on the development of data analysis and visualization techniques for information access, understanding, and management. She is particularly interested in the study of the structure and evolution of scientific disciplines; the analysis and visualization of online activity; and the development of cyberinfrastructures for large-scale scientific collaboration and computation. She is the co-editor of the Springer volume on "Visual Interfaces to Digital Libraries," a special issue of *PNAS* 101 (Suppl. 1) on "Mapping Knowledge Domains" (published in April 2004), and an editor of the *Journal of Informetrics*: *Special Issue on the Science of Science: Conceptualizations and Models of Science* (2009). She also co-edited a special issue on "Collaborative Information Visualization Environments" in *PRESENCE: Teleoperators and Virtual Environments*, (MIT Press, Feb. 2005), "Information Visualization Interfaces for Retrieval and Analysis" in the *Journal of Digital Libraries* (March 2005), and "Mapping Humanity's Knowledge" in *Environment and Planning B* (Sept. 2007). Her new book, *Atlas of Science: Visualizing What We Know* is forthcoming from MIT Press in 2010.

Mark Gerstein is the Albert L. Williams Professor of Biomedical Informatics at Yale University. He is Co-Director of the Yale Computational Biology and Bioinformatics Program, and has appointments in the Department of Molecular Biophysics and Biochemistry and the Department of Computer Science. He received his AB in Physics Summa Cum Laude from Harvard College in 1989 and his Ph.D. in Chemistry from Cambridge in 1993. He did post-doctoral work at Stanford and took up his post at Yale in early 1997. Since then he has received a number of Young Investigator Awards (e.g. from the Navy and the Keck foundation) and has published appreciably in scientific journals. His research is focused on bioinformatics, and he is particularly interested in large-scale integrative surveys, biological database design, macromolecular geometry, molecular simulation, human genome annotation, gene expression analysis, and data mining.

Stephen M. Uzzo is currently Vice President of Technology for the New York Hall of Science, where he works on a number of projects related to Science, Technology, Engineering and Mathematics (STEM) learning, sustainability and network science including, "Connections: the Nature of Networks," a public exhibition on network science that opened in 2004. He was also the local organizer for the 2007 International Conference and Workshop on Network Science. Dr. Uzzo serves on the faculty of the New York Institute of Technology (NYIT) Graduate School of Education, where he teaches STEM teaching and learning. During the 1980's, he worked on a number of media and technology projects. In 1981, he was part of the launch team for MTV and was appointed Chief Engineer for Video/ Computer Graphics Production and Distance Learning Networks for the NYIT Video Center in 1984. Other projects during that period included the first all digital satellite television transmission, best practices group for NBC Summer Olympic Games in Barcelona, and a team of scientists and engineers at Princeton's Space Studies Institute to develop and test lunar teleoperations simulators. During the 1990's, Dr. Uzzo served on numerous advisory boards for educational institutions, as well as facilitating major technology initiatives amongst K-12 public/private schools, higher education and government to improve STEM literacy. His work on various projects important to conservation include ecosystems studies that were instrumental in blocking offshore oil drilling in New York waters and a cross sound bridge in Oyster Bay, as well as cleanup planning for superfund sites. He has worked on preservation and open space projects on Long Island and the San Francisco Bay Peninsula. He holds a Ph.D. in Network Theory and Environmental Studies from the Union Institute.

Participants Workshop (I), Arlington, Virginia

Richard Bonneau is a faculty member at New York University's (NYU) Center for Comparative Functional Genomics, where he is a joint member of both the Biology and Computer Science Departments. His affiliate position with the Institute for Systems Biology offers many exciting possibilities for the exchange of ideas between the two centers, both focused on applying functional genomics to a wide range of systems. His research efforts are focused on making computational tools, algorithms and methods for systems-wide elucidation of biological systems. His research aims to develop computational methods at the intersection of two interrelated fields: protein structure and functional genomics. He is also currently the technical lead on two-grid computing collaborations with IBM—the first and second phases of the Human Proteome Folding Project. Dr. Bonneau did his doctoral work at the University of Washington in Seattle, working with David Baker on the the protein prediction platform, Rosetta. He is currently a member of the Rosetta-commons and continues to develop and work with Rosetta as part of several projects in the lab. Before joining NYU, he worked as a Senior Scientist with Leroy Hood at the Institute for Systems Biology. He also oversees TACITUS's (www.tacitus.com) approach to data gaming (3-D visualization) for all applications that focus on genomics, computational biology and cell biology.

Kevin W. Boyack recently joined SciTech Strategies, Inc. (www.mapofscience.com) after working for 17 years at Sandia National Laboratories. He received a Ph.D. in Chemical Engineering from Brigham Young University in 1990 and has worked in a variety of fields since that time. These fields include combustion (experimental and modeling), transport processes, socio-economic war gaming, and most recently, knowledge domain visualization or science mapping. At SciTech Strategies, he and Dick Klavans create maps of science for planning and evaluation purposes. His interests include metrics and indicators, text mining, data fusion, and expanding the application and uses of science mapping.

Olga Brazhnik is a Computer Scientists with the Division of Biomedical Technology of the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH). Olga leads programs in biomedical and clinical informatics, data and knowledge discovery, integration, and visualization, collaborative and semantic technologies, and computational biology. She works with research grants, biomedical technology research centers, and small business innovative research.

Dr. Brazhnik started her career as a physicist, applying theoretical and computational methods in biology and medicine. With the goal to develop a computational framework for creating coherent knowledge from diverse scientific data she launched a dual career as a scientists and an IT professional. She developed computer modeling and simulations, conducted theoretical and computational multidisciplinary research at the NIH, James Franck Institute at the University of Chicago, Virginia Tech, and the Institute of Applied Physics of the Russian Academy of Sciences.

Olga worked as the Chief Database Architect on projects of biomedical and clinical data integration for the US Air Force Surgeon General's Office, created bioinformatics databases at Virginia Bioinformatics Institute. She developed and taught graduate classes and professional workshops at George Mason University, VA Tech, and other schools. Olga has published in peer-reviewed journals, presented at numerous international scientific meetings, organized and led professional groups and activities. The most exciting part of her current work is in marrying the wisdom of crowds and individual creativity with cutting edge technology and solid science for the benefits of human health.

John Bruer has been President of the James S. McDonnell Foundation in St. Louis, Missouri since 1986. The foundation awards over $20 million annually to support biomedical research, education, and international projects. The Foundation has established programs in the areas of neuroscience, cancer research, education, and child health. Since 1999, the McDonnell Foundation has developed a specific program interest in complex systems research. Dr. Bruer hold degrees in Philosophy from the University of Wisconsin-Madison, Oxford University, and Rockefeller University. Bruer's book, *Schools for Thought: A Science of Learning in the Classroom* (MIT Press, 1993), received the 1993 Quality in Educational Standards Award and the 1994 Charles S. Grawemeyer Award in Education. Also, his book, *The Myth of the First Three Years* (Free Press, 1999), received the 2000 Eleanor Maccoby Award from the American Psychological Association. He is Adjunct Professor of Philosophy at Washington University and a member of the National Science Board. His current research interests include issues in cognitive neuroscience, emergence and reduction in the special sciences, and causal reasoning.

Kei Cheung is an Associate Professor at the Yale Center for Medical Informatics. He received his Ph.D. in Computer Science from the University of Connecticut. Dr. Cheung's primary research interest lies in the areas of bioinformatics databases and tool integration. Recently, he embarked on exploring the use of semantic Web technologies in the context of life sciences (including neuroscience) data integration. Dr. Cheung co-edited a book entitled, *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, which was published by Springer in January of 2007. He served as the chair of the First International Workshop on Health Care and Life Sciences Data Integration for the Semantic Web, which was co-located with the WWW2007 (World Wide Web Consortium) Conference. He is currently a Guest Editor of the upcoming Special Issue: "Semantic BioMed Mash-up", which will appear in the *Journal of Biomedical Informatics*. Dr. Cheung is an invited expert to the Semantic Web Health Care and Life Science Interest Group launched by the World Wide Web Consortium.

Stefan Hornbostel studied Social Sciences at the University of Göttingen in Germany. He completed his Ph.D. at the Freie Universität, Berlin. After his studies, he worked at the Universities of Kassel, Cologne, Jena and Dortmund, as well as at the Center of Higher Education Developement (CHE – Centrum für Hochschulentwicklung). Currently, Stefan is Professor in the Department of Social Sciences (science studies) at the Humboldt University of Berlin and Director of the Institute for Research Information and Quality Assurance in Bonn.

Masatsura Igami is a Senior Researcher of the National Institute of Science and Technology Policy (NISTEP). He received a Ph.D. from the University of Tsukuba, Japan in 2001. After receiving a Ph.D. in Engineering, he worked in the private industry, engaging in the development of computer simulation programs of molecular dynamics. He joined NISTEP in 2002. At NISTEP, he planned and conducted the 8th National Technology Foresight in Japan as part of Terutaka Kuwahara's group (FY2003-FY2004). He developed a bibliometric method to discover emerging research areas and to analyze the contribution of Japanese activities in these areas. This work is the origin of the *Science Map*, which is now bi-annually published by NISTEP. The latest report, *Science Map 2006,* was published in the Spring of 2008. Beginning in July 2005, he worked at the Organisation for Economic Co-Operation and Development (OECD) for two years, where he engaged in the developments of new science and technology indicators - especially indicators related to nanotechnology patent applications - within the organization. The mapping of nanotechnology patent applications is his latest work, which will be published in *Scientometrics*.

Neo Martinez is the Director of the Pacific Ecoinformatics and Computational Ecology Lab at the Rocky Mountain Biological Laboratory in Crested Butte, Colorado. He is an Affiliated Faculty Member of the Energy Resources Group at the University of California-Berkeley. He holds an MS and Ph.D. in Energy and Resources from the University of California, Berkeley and has made notable studies of food webs and the trophic niches they contain through research

supported by the NSF and the Ford Foundation Fellowship program. His interests and funded research projects have involved the construction of an Internet database on food webs and ecological networks, and the development, implementation and evaluation of prototype tools and applications within the semantic web for biocomplexity and biodiversity domains.

Alex Soojung-Kim Pang is a Research Director at the Institute for the Future (IFTF). At IFTF, he conducts research on the future of science, technology and innovation. He holds a Ph.D. in History and Sociology of Science from the University of Pennsylvania, and conducted research on the history of scientific representation and visualization. Before joining IFTF, he served as Managing Editor of the *Encyclopaedia Britannica*, where he oversaw its transition from print to electronic publication. He also taught at Williams College and the University of California, Davis and held post-doctoral fellowships at Stanford University and the University of California, Berkeley.

Martin Storksdieck is Director of the Board on Science Education at the National Academy of Sciences/National Research Council and a Research Fellow at the Institute of Learning Innovation (ILI), where he directs ongoing research studies on science learning in immersive environments; models involving researchers and scientists in science museums and science centers; and understanding the impact of science hobbyists (such as amateur astronomers) on the public understanding of science. He previously served as Director of Project Development and as Senior Researcher at ILI. He also was a science educator with a planetarium in Germany, where he developed shows and programs on global environmental change, served as editor, host, and producer for a weekly environmental news broadcast, and worked as an environmental consultant specializing on local environmental management systems. He holds a Masters in Biology from the Albert-Ludwigs University in Freiburg, Germany, a Masters in Public Administration from Harvard University, and a Ph.D. in Education from Leuphana University in Lüneburg, Germany.

Peter van den Besselaar is currently Head of Department and Research Director of the Science System Assessment Department at the Rathenau Instituut. He is also a Professor of Communication Science at the University of Amsterdam, with a special chair in e-Social Science. His research interests are the dynamics of science, technology and innovation; science and innovation policy; users and design; the effects of ICT on social change. Previously, he was Director of the Netherlands Social Science Data Archive (2002-2005) and an Associate Professor of Social Informatics at the University of Amsterdam (1986-2002). In the latter position, he directed a research program on ICT based innovation and on technology assessment. He has been a partner in a variety of large international research projects funded by the European Commission in the various Framework Programs. He obtained a degree in Mechanical Engineering (propedeuse; Technische Universiteit Eindhoven), a BSc Mathematics (kandidaatsexamen, Universiteit Utrecht), an MA in Philosophy (cum laude) and a Ph.D. from the Faculty of Psychology (both: Universiteit van Amsterdam).

Maria Zemankova is a Computer Scientist who is known for the theory and implementation of the first Fuzzy Relational Database System. This research has become important for the handling of approximate queries in databases. She is currently a Program Officer in the Intelligent Information Systems Division at the National Science Foundation. In 1992, she was the first recipient of the Special Insterests Group on Management of Data (SIGMOD) Innovations Award for her work in the conception of initiatives in research on scientific databases and digital libraries. She received her Ph.D. in Computer Science in 1983 from Florida State University for her work on Fuzzy Relational Database Systems.

Participants Workshop (II), New York City, New York

Kevin W. Boyack, see previous Workshop (I) listing.

Kyle Brown is founder and CEO of Innolyst, providing Web-based collaboration and management applications to the life-sciences and not-for-profit industries. Innolyst focuses on two major initiatives: ReseachCrossroads and PatientCrossroads. The public ResearchCrossroads Web site provides a comprehensive view of government, academic and nonprofit-funded research to spur collaboration and data sharing among academia, foundations, government and industry. PatientCrossroads is a patient registry and genetic testing platform that provides patient registration and self-report, genetic educational materials and access to genetic counselors and testing.

Noshir Contractor is the Jane S. & William J. White Professor of Behavioral Sciences in the School of Engineering, School of Communication and the Kellogg School of Management at Northwestern University. He holds a Ph.D. from the Annenberg School for Communication at the University of Southern California and a Bachelors Degree in Electrical Engineering from the Indian Institute of Technology in Madras (Chennai). He was on the faculty at the University of Illinois at Urbana-Champaign for twenty years prior to joining Northwestern in 2007. He is the Director of the Science of Networks in Communities (SONIC) Research Group—investigating factors that lead to the formation, maintenance, and dissolution of dynamically linked social and knowledge networks in communities. Specifically, his research team is developing and testing theories and methods of network science to map, understand and enable more effective: (i) disaster response networks; (ii) public health networks; and (iii) massively multiplayer online games (MMOs) networks. Professor Contractor has published or presented over 250 research papers dealing with communication. His book titled, *Theories of Communication Networks* (co-authored with Professor Peter Monge and published by Oxford University Press in English and scheduled to be published by China Renmin University Press in simplified Chinese in 2007), received the 2003 Book of the Year Award from the Organizational Communication Division of the National Communication Association. He is the lead developer of *IKNOW* (Inquiring Knowledge Networks On the Web), and its Cyberinfrastructure extension *CI-KNOW*, a network recommender system to enable communities using cyberinfrastructures, as well as *Blanche*, a software environment to simulate the dynamics of social networks.

Ingo Günther created sculptural works with video early in his career, which led him to more journalistically oriented projects that he later pursued in TV, print, and the art field. Based in New York, he played a crucial role in the evaluation and interpretation of satellite data gathered from political and military crisis zones. On an artistic level, the work with satellite data and mapping them for TV led to Günther's contribution to *documenta 8* (1987) and the installation *K4 (C3i)* (Command Control Communication and Intellience). Since 1989, Günther has used globes as a medium for his artistic and journalistic interests (see www.WorldProcessor.com). In 1989, he founded the first independent and non-commercial TV station in Eastern Europe–Channel X, Leipzig. He has contributed his work to numerous institutions, conferences, conventions and museums around the world, notably to: The National Galerie Berlin, 1983 and 1985; Venice Biennale, 1984; documenta, Kassel, 1987; P3 Art and Environment, Tokyo, 1990, 1992, 1996 and 1997; Ars Electronica, Linz, 1991; Centro Cultural de Belem, Lisbon, 1995; Hiroshima City Museum of Contemporary Art, 1995; Guggenheim Museum, New York, 1996; Kunsthalle Dusseldorf, 1998; Neues Museum Weserburg Bremen, 1999; World Economic Forum, Davos, 2000; V2 Rotterdam, 2003; Yokohama Triennale, Tokoyo, 2005; Kunstverein Ruhr, Essen, 2005; IFC/World Bank, Washington, DC; San Jose Museum of Art, San Jose, CA, 2006; and Siggraph, San Diego, CA, 2007. From 1990 to 1994, he was a Professor at the Academy of Media Arts in Cologne; from 2001 to 2003 he was a Professor at the University for Media, Art, and Design in Zurich; and from 2006 to 2007 he was a Visiting Professor at the Tokyo National University for Fine Arts and Music.

Sharon M. Jordan is Assistant Director for Program Integration in the Office of Scientific and Technical Information (OSTI, www.osti.gov) within the U.S. Department of Energy's (DOE) Office of Science. Her responsibilities include leading the department-wide Scientific and Technical Information Program, which includes representatives from each DOE laboratory and office, and developing and implementing scientific and technical information policy to ensure the broadest possible access to DOE's R&D findings, within applicable laws and regulations. Ms. Jordan was instrumental in transforming various systems used for acquiring and announcing R&D results to Web-based systems. To this end, OSTI has forged collaborations and opened science portals for efficient, one-stop searching by researchers and the general public. She also manages OSTI communications and outreach activities, as well as interagency e-gov initiatives, including www.science.gov, for which she received a Meritorious Service Award in 2005 from the interagency, Scientific and Technical Information (STI) managers group, known as CENDI. She has served in a range of information and program management positions with the DOE since 1974.

David Lazer is currently on the faculties of Political Science and Computer Science at Northeastern University, as well as continuing his previous appointment to direct the Harvard University Program on Networked Governance. He earned his Ph.D. in Political Science from the University of Michigan. He studies the political and organizational implications of different social networks, particularly those that are based on electronic connections, such as the Internet. A recent project,

for example, focused on the use of Web sites by members of Congress as a means to connect to constituents. His most recent book is a co-edited volume titled, *Governance and Information Technology: From Electronic Government to Information Government*.

Israel Lederhendler trained in comparative psychology and was active in behavioral neurobiology research for many years before moving into research administration at the National Institutes of Health (NIH). As a program official at the National Institute of Mental Health, he has organized conferences and workshops to foster communications and collaborations between NIH Institutes and Centers with external organizations. His activities as a scientist and administrator have focused on knowledge management, together with other data mining and reporting tools, as essential analytic activities that NIH will need to accurately reflect its scientific accomplishments and manage its many challenges. He has been notably involved and interested in merging scientific portfolio analysis with electronic tools for data integration and analysis. Recently, as Interim Project Manager of the Electronic Research Administration, Director of the Office of Research Information Services, and finally as Director of the Division of Information Services, he has tirelessly promoted the value of such tools for research, program management and decision support.

Bongshin Lee is a Researcher in the Visualization and Interaction Research Area at MicrosoftResearch. Her main research interest is human-computer interaction, with a focus on information visualization and interaction techniques. She is also interested in developing interfaces for mobile devices. Her Ph.D. work was centered on investigating how people interact with graph visualization systems to understand the large graph. Dr. Lee is currently focusing on understanding and visualizing uncertainty in data. She received her Bachelor's of Science in Computer Science from Yonsei University in Seoul, Korea. After spending a few years in the industry, she came to the U.S. to pursue a Ph.D. She earned her Master's of Science and Ph.D. in Computer Science from University of Maryland at College Park in 2002 and 2006, respectively.

Barend Mons obtained his Masters of Science (1981, Cum Laude) and his Ph.D. (1986) at the University of Leiden in The Netherlands, majoring in Cell and Molecular Biology. He has been extensively involved in biotechnological research and entrepreneurial applications of that research, from KREATECH-biotechnology, to Knewco, Inc. Since 2002, Dr. Mons has been an Associate Professor in Bio-Semantics at the Department of Medical Informatics, Erasmus Medical Centre at the University of Rotterdam and (since 2005) at the Department of Human Genetics at the Leiden University Medical Centre. He was one of the Founding Trustees of The Centre for the Management of Intellectual Property in Health Research and Development (MIHR), which assists people in developing countries to manage their critical Intellectual Property (IP) for the betterment of society. His present activities mainly focus on international networking to realize a completely new form of computer assisted distributed annotation and on-line knowledge discovery, in close collaboration between Rotterdam, Leiden and Knewco, and largely based on the Knewco Knowlet™ technology, combined with Open Access and Open Source Wiki-technology approaches.

James Onken is currently an Analyst in the Office of Research Information Systems (ORIS), a component of the National Institute of Health (NIH) Office of Extramural Research. He is responsible for the analysis and presentation of data on NIH research programs and research personnel for use in program evaluation and policy studies. He is also Program Manager for a new NIH Research Portfolio On-line Reporting Tool (RePORT) Web site, http://RePORT.nih.gov. Prior to working for ORIS, Dr. Onken worked at the National Institute of General Medical Sciences (NIGMS) for over 17 years, most recently as the Planning and Evaluation Officer and Assistant Director for Resource Allocation and Analysis for NIGMS's Division of Extramural Activities. Earlier in his career, he conducted research on human information processing, cognitive performance and mathematical models of decision-making; performed decision analysis for several federal agencies; and designed and developed computerized decision support systems. Jim holds a Masters and Ph.D. in Psychology from Northwestern University, and an MPH with a concentration in biostatistics from George Washington University.

Gregor Rothfuss works as a Tech Lead on the next generation of mapping technologies at Google. Gregor also serves as VP, Apache Lenya at the Apache Software Foundation and is a co-founder of the Open Source Content Management Organization (OSCOM). He attended the University of Zurich, and has continued developing his interests in Web architecture, scalability, mapping and open source systems development.

Ben Shneiderman (http://www.cs.umd.edu/~ben) is a Professor in the Department of Computer Science and Founding Director (1983-2000) of the Human-Computer Interaction Laboratory (http://www.cs.umd.edu/hcil) at the University of Maryland. He was elected as a Fellow of the Association for Computing Machinery (ACM) in 1997 and a Fellow of the American Association for the Advancement of Science (AAAS) in 2001. He received the ACM SIGCHI Lifetime Achievement Award in 2001. He is the author of *Software Psychology: Human Factors in Computer and Information Systems* (1980) and *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (4th ed. 2004). He pioneered the highlighted textual link in 1983, and it became part of Hyperties, a precursor to the Web. His move into information visualization helped spawn the successful company Spotfire (http://www.spotfire.com). He is a technical advisor for the HiveGroup and Groxis. With S. Card and J. Mackinlay, he co-authored *Readings in Information Visualization: Using Vision to Think* (1999). His books also include, *Leonardo's Laptop: Human Needs and the New Computing Technologies* (MIT Press), which won the IEEE Distinguished Literary Contribution award in 2004.

Eric A. Schultes trained in Organismal Biology at Oakland University (BS, 1992) Paleobiology (Ph.D., University of California, Los Angeles, 1997) and RNA Biochemistry (Post-doctoral, MIT, 2005). His work has involved knowledge management and visualization with the RNA simplex, a low-dimensional projection of the very-high dimensional sequence space of RNA, which grew out of his work with evolutionary theory, optical mineralogy and physical chemistry. More recently, Dr. Shultes has begun incorporating digital video and documentary ideas into his work as a private consultant, which he has pursued since 2005 through his company, Hedgehog Research. In 2009, he held Adjunct Research positions within the Department of Computer Science at Duke University, and a Visiting Scientist position within the Human Genetics Department at the Leiden University Medical Centre, in association with the Concept Web Alliance.

Eric Siegel is Director and Chief Content Officer at the New York Hall of Science. He leads the education, programs, exhibition development, science, and technology functions at the Hall of Science and is responsible for strategic planning and partnerships. He led the planning and institutional fundraising for the $92 million expansion that was completed in 2004, as well as directing the development of four major exhibitions. Eric has been in senior roles in art and science museums for 25 years and has consulted and published extensively in the museum field. He is President of the National Association for Museum Exhibitions; a member of the graduate faculty of the New York University Museums Studies program; Board Member of SolarOne, an urban environmental organization in New York City; and past Chairman of the Museums Council of New York City. Eric graduated from the CORO Leadership New York program, and holds an MBA in Arts Administration from State University of New York (SUNY), Binghamton.

Martin Storksdieck, see previous Workshop (II) listing.

Facilitators

Elisha F. Hardy holds a BA from Indiana University (IU) with a focus on Graphic Design and is currently a Masters candidate in Human-Computer Interaction Design at the School of Informatics and Computing at IU. She has been part of Katy Börner's InfoVis Lab and the Cyberinfrastructure for Network Science Center since June 2005, creating designs for many different projects, ranging from small flyers to national exhibitions. Since March 2008, she has been co-curator of the Places & Spaces: Mapping Science exhibit.

Weixia (Bonnie) Huang worked as a System Architect at the Cyberinfrastructure for Network Science Center at the School of Library and Information Science at Indiana University (IU), Bloomington when the workshops were run. She led the software development of the NSF funded, Network Workbench (NWB) and Cyberinfrastructure Shell (CIShell) projects. Currently, she is associated with the Kuali Project at IU. She is particularly interested in designing and developing software with sound extensibility, usability, and scalability. Before joining IU, she worked as a Research Staff Member at Xerox Wilson Research Center and as a Software Engineer at Sprint. She was the Architect and Programmer at Xerox, developing a Device-Centric Enterprise Service Platform for automated data transmission and remote diagnosis systems.

Attending Agency Staff

Robert Bell, National Science Foundation, Science Resources Statistics, Sociology

Lawrence Brandt, National Science Foundation, Division of Advanced Scientific Computing, Digital Government

Lawrence Burton, Social, Behavioral and Economic Sciences/ Science Resources Statistics

Patrick J. Clemins, National Science Foundation, Division of Biological Infrastructure, BioInformatics

Joanne D. Culbertson, National Science Foundation, Office of the Assistant Director for Engineering, Policy

Al DeSena, National Science Foundation, Division of Research on Learning in Formal and Informal Settings

Le Gruenwald, National Science Foundation, Division of Information and Intelligent Systems, Bioinformatics

Haym Hirsh, National Science Foundation, Division of Information and Intelligent Systems, Data Mining

Anne-Francoise Lamblin, National Science Foundation, Division of Biological Infrastructure, Bioinformatics

Julia Lane, National Sicence Foundation, Social, Behavioral and Economic Sciences, Economics

Terence D. Langendoen, Social, Behavioral and Economic Sciences/Behavioral and Cognitive Sciences, Linguistics

Mary Lou Maher, National Science Foundation, Computer and Information Science and Engineering, Human Centered Computing

William P. Neufeld, National Science Foundation, Division of Research, Evaluation and Communication

Frank Olken, National Science Foundation, Division of Information and Intelligent Systems

William W. Schultz, National Science Foundation, The Directorate for Engineering, Fluid Mech

Arlene M. de Strulle, National Science Foundation, Division of Research on Learning in Formal and Informal Settings, Learning Technology

Paola Sztuju, Educational Research

Grace Yuan, National Science Foundation, Division of Information Systems

Maria Zemankova, National Science Foundation, Division of Information and Intelligent Systems/Computer and Information Science and Engineering Directorate

Appendix B: Workshop Agendas

Workshop I

GOAL

The first two-day workshop brings together application holders from science of science studies and biology from the U.S. but also from Japan and Europe.

Science Policy Challenges relate to the study of:

- The evolution of scientific communities/fields – birth, growth, maturation, decline.
- Interactions among fields. Who 'eats' who's papers?
- Trends, patterns, or emergent research frontiers, feedback loops, etc.
- Interplay of competition and collaboration.
- Diffusion of people, ideas, skills, etc. in geospatial space and topic space.
- Effects of different funding models, e.g., few large vs. many small grants.

Biomedical/Ecological challenges comprise:

- Build the Encyclopedia of Life
- Prevent the Next Pandemic
- Creating an Inventory of Genotypes and Using it for Clean Energy and Nutrition.

DATE

March 10 & 11, 2008

MEETING PLACE

NSF, Room II-555, 4201 Wilson Boulevard, Arlington, VA

DAY 1:

12:00pm	Welcome by Organizers by Katy Börner
12:15pm	Introduction by Participants (5 min per person/ organization). Led by Stephen M. Uzzo
2:00pm	Break
2:15pm	Presentation of NSF CDI program by Mary L. Maher, NSF
2:30pm	Challenges and Opportunities by Luís M. A. Bettencourt
3:00pm	Breakout Sessions on "$10 Million SciPolicy and BioMed Challenge". Intro by Stephen M. Uzzo
4:00pm	Breakout Session Reports
4:30pm	Interactive Timeline Assembly - see connections and build on them. Led by Alex Soojung-Kim Pang
6:30pm	Adjourn
7:00pm	Joint Japanese dinner at Matsutake

DAY 2:

9:00am	Light Breakfast
9:30am	All the Data and Publications from Science on Web:A Vision for Harnessing this to Study the Structure of Science presentation by Mark Gerstein
10:00am	Breakout Sessions on "Envision and Draw your Dream Tool" Intro by Katy Börner
11:00am	Breakout Session Reports
11:30am	Science Mapping: Convergence, Consensus, Policy Implications presentation by Kevin W. Boyack
12:00pm	Joint Lunch
1:00pm	Write Description of 2nd Best Idea for CDI Grant Proposal. Led by Alex Soojung-Kim Pang
2:00pm	Presentation to Group
2:45pm	Break
3:00pm	Collective Exercise on "Who would like to collaborate with whom on what?" Lead by Katy Börner
4:00pm	Discussion of Next Steps, Funding Opportunities, etc.
5:00pm	Adjourn

Workshop II

GOAL

This workshop invites experts to brainstorm socio-technical infrastructures that address the needs and challenges identified in Workshop I. It will bring together about 20 leading experts in database research and digital libraries, cyberinfrastructure/ e-Science design, and social science of CI from the U.S. but also from Asia and Europe.

The goal is to identify:

- Key institutions around the world,
- Existing and evolving technologies, and
- Incentive structures

that address the needs identified in Workshop I without duplicating existing efforts in a sustainable fashion.

DATE

April 7 & 8, 2008

MEETING PLACE

New York Hall of Science, Queens, NY
Upper Level Auditorium and Lobby

DAY 1:

12:00pm	Welcome by Organizers by Katy Börner
12:15pm	Introduction by Participants (5 min per person/ organization = 15 slots) Led by Stephen M. Uzzo
2:15pm	Break
2:30pm	Workshop Goals
3:00pm	Challenges and Opportunities by Luís M. A. Bettencourt
3:45pm	Break
4:00pm	Breakout Sessions on "What SciPolicy and BioMed Challenge are solvable?" Intro by Stephen M. Uzzo
5:30pm	Breakout Session Reports & Interactive Timeline Assembly led by Katy Börner
6:00pm	Adjourn
7:00pm	Joint dinner at Deluge

DAY 2:

9:00am	Light Breakfast
9:30am	Presentation of Google Maps by Gregor J. Rothfuss
10:00am	Breakout Sessions on "Build the Dream Tool (on paper)" Intro by Katy Börner
11:00am	Breakout Session Reports
11:30am	Presentation of Worldprocessor Globes by Ingo Günther
12:00pm	Joint Lunch
1:00pm	Breakout Sessions on "Challenges and Opportunities"
2:00pm	Breakout Session Reports Intro by Katy Börner
2:45pm	Break
3:00pm	Collective Exercise on "Who would like to collaborate with whom on what? Led by Katy Börner
4:00pm	Discussion of Next Steps, Funding Opportunities, etc.
5:00pm	Adjourn

Appendix C: Listing of Relevant Readings, Datasets, and Tools

The resources listed here have been extracted from submissions by workshop participants, interviews with 38 science policy makers conducted in 2009, the *Places and Spaces: Mapping Science* exhibit web site (http://scimaps.org), and the *Atlas of Science* (Börner 2010). Several science and technology datasets were taken from Zucker & Darby, (forthcoming).

Readings

BOOKS ON DATA GRAPHICS

Bertin, Jaques. (1981). *Graphics and Graphic Information Processing*. Berlin: Walter de Gruyter.

Tufte, Edward R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, Edward R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press

MAPPING SCIENCE

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: MIT Press.

Börner, Katy, Chaomei Chen, Kevin W. Boyack. (2003). Visualizing Knowledge Domains. In Cronin, Blaise (Eds.), *Annual Review of Information Science & Technology* (Vol. 37, pp. 179-255), chapter 5, American Society for Information Science and Technology, Medford, NJ.

Boyack, Kevin W., Richard Klavans, Katy Börner. (2005). Mapping the Backbone of Science. *Scientometrics*, 64, 3:351-374.

Chen, Chaomei. (2002). *Mapping Scientific Frontiers*. London: Springer-Verlag.

Klavans, Richard and Kevin W. Boyack. (2006). Identifying a Better Measure of Relatedness for Mapping Science. *Journal of the American Society for Information Science and Technology*, 57, 2:251-263.

Small, H. G. (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science,* 24, 4:265-269.

MAPPING BIOMEDICAL/ECOLOGICAL ENTITIES

Burns, Gully A.P.C., Bruce W. Herr, II, David Newman, Tommy Ingulfsen, Patrick Pantel, Padhraic Smyth. (2006). *Society for Neuroscience Visual Browser*. http://scimaps.org/maps/neurovis (accessed 12/30/2009).

Curehunter Inc. (2009). Chi Precision Medical Data Mining: *Curehunter.org*. http://www.curehunter.com/public/showTopPage.do (accessed 12/30/2009).

Douglas, S.M., G.T. Montelione, M. Gerstein (2005) PubNet: a Flexible System for Visualizing Literature Derived Networks. Genome Biol 6: R80.

Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, Albert-László Barabási. (2007). The Human Disease Network. *PNAS* 104, 21:8685-8690.

Han JD, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J. Walhout, M.E. Cusick, F.P. Roth, M. Vidal. (2004). Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network. *Nature*, 430:88-93.

Huang N.N., D.E. Mootz, A.J. Walhout, M. Vidal, C.P. Hunter. (2002). MEX-3 Interacting Proteins Link Cell Polarity to Asymmetric Gene Expression in Caenorhabditis elegans. *Development*, 129:747-59.

John D. and Catherine T. MacArthur Foundation and the Alfred P. Sloan Foundation. (2009) Encyclopedia of Life (EOL). http://www.eol.org (accessed January 1, 2008).

Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, Doucette- L. Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, M. Vidal. (2004). A Map of the Interactome Network of the Metazoan C. elegans. *Science*, 303:540-3.

Rual, J.F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, M. Vidal. (2005). Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network. *Nature*, 437:1173-8.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF , Brasch MA, Thierry-Mieg N, Vidal M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, 287:116-22.

Yildriim, Muhammed A., Kwan-II Goh, Michael E. Cusick, Albert-László Barabási, and Marc Vidal. (2007). Drug-Target Network. *Nature Biotechnology* 25, 10:1119-1126.

Datasets

PUBLICATION DATASETS

American Psychological Association. (2009). *PsycINFO*. http://www.apa.org/pubs/databases/psycinfo (accessed on 12/17/2009).

Chemical Abstracts Service. (2007). *CAS Statistical Summary 1907-2007*. http://www.cas.org/ASSETS/836E3804111B49BFA28B95BD1B40CD0F/casstats.pdf (accessed on 12/17/2009).

Cyberinfrastructure for Network Science Center. (2009). *Scholarly Database*. http://sdb.slis.indiana.edu (accessed on 12/19/2009).

Elsevier B.V. (2009). *Scopus*. http://www.scopus.com (accessed on 12/17/2009).

Georgetown University. (2009). *Kennedy Institute of Ethics: Library and Information Services*. http://bioethics.georgetown.edu (accessed on 1/4/2010).

Google, Inc. (2009). *Google Scholar*. http://scholar.google.com (accessed on 12/17/2009).

ICMG, Ltd. (2007). *Nanomedicine Research*. http://www.nano-biology.net/literature.php (accessed on 1/4/2010).

Illinois Institute of Technology. (2009). Center for the Study of Ethics in the Professions: *NanoEthicsBank*. http://ethics.iit.edu/NanoEthicsBank (accessed on 1/4/2010).

International Council on Nanotechnology. (2009). *EHS Database*. http://icon.rice.edu/research.cfm (accessed on 1/4/2010).

International Information System for the Agricultural Sciences and Technology. (2009). *AGRIS/CARIS: Agricultural Literature Repository*. http://www.fao.org/agris/search/search.do (accessed on 12/17/2009).

ITHAKA. (2009). *JSTOR: Trusted Archives for Scholarship*. http://www.jstor.org (accessed on 12/17/2009).

Ley, Michael. (2009). *The DBLP Computer Science Bibliography*. http://www.informatik.uni-trier.de/~ley/db (accessed on 12/17/2009).

Los Alamos National Laboratory. (2009). *Library Without Walls project*. http://library.lanl.gov/lww (accessed on1/04/2010).

Loyola University. (2009). *Nanobase*. http://www.wtec.org/loyola/nanobase (accessed on 1/4/2010).

The National Bureau of Economic Research. (2009). NBER Scientific Papers Database: Data. http://www.nber.org/data (accessed on 12/31/2009).

National Federation of Advanced Information Services. (1990). *NFAIS Abstract Dataset, 1957-1990*. http://www.nfais.org (accessed on 4/30/2009).

National Federation of Advanced Information Services. (2009). *NFAIS: Serving the Global Information Community*. http://www.nfais.org (accessed on 1/5/2010).

National Institutes of Health, U.S. National Library of Medicine. (2009). *PubMed.gov*. http://www.ncbi.nlm.nih.gov/pubmed (accessed on 12/17/2009).

Office of Science and Technical Information. *WorldWideScience.org: The Global Science Gateway*. http://worldwidescience.org (accessed on 1/4/2010).

Ovid Technologies, Inc. (2009). *PsycINFO*. http://www.ovid.com/site/catalog/DataBase/139.jsp (accessed on 4/30/2009).

ProQuest LLC. (2009). *Dialog: Authoritative Answers Enriched by Proquest*. http://www.dialog.com/about (accessed on 4/30/2009).

Stanford Linear Accelerator Center. (2009). SPIRES-HEP Database. http://www-spires.fnal.gov/spires/about (accessed on 4/30/2009).

Thomson Reuters. (2009). *ISI Web of Knowledge*®. http://apps.isiknowledge.com (accessed on 4/30/2009).

Thomson Reuters. (2009). *ISI Web of Knowledge: ISI Highly Cited*. http://isihighlycited.com (accessed on 12/31/2009).

Thomson Reuters. (2009). *RefViz: Explore Research Literature…visually!* http://www.refviz.com (accessed on 10/02/2008).

Thomson Reuters. (2009). *Social Sciences Citation Index*. http://thomsonreuters.com/products_services/science/science_products/a-z/social_sciences_citation_index (accessed on 5/1/2009).

Zucker, Lynne G., Michael R. Darby. (2007) *Nanobank*. http://www.nanobank.org (accessed on 1/4/2010).

PATENT DATASETS

Cambia. (2009). *Patent Lens: Initiative for Open Innovation*. http://www.patentlens.net/daisy/patentlens/patentlens.html (accessed on 4/30/2009).

École Polytechnique Fédérale de Lausanne. (2009). *CEMI's PATSTAT Knowledge Base*. http://wiki.epfl.ch/patstat (accessed on 4/30/2009).

European Commission. (2009). *Eurostat: Your Key to European Statistics*. http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home (accessed on 12/17/2009).

European Patent Office. (2009). *EPO Global Patent Index*. http://www.epo.org/patents/patent-information/subscription/gpi.html (accessed on 12/17/2009).

United States Patent and Trademark Office. (2010). *Patents Search*. http://www.uspto.gov (accessed on 1/4/2010).

United States Patent and Trademark Office. (2010). *Patent Full-Text and Full Page Image Databases*. http://patft.uspto.gov (accessed on 1/4/2010).

U.S. Department of Energy. (2009). Office of Scientific and Technical Information, *DOE Patents*. http://www.osti.gov/doepatents (accessed on 12/31/2009).

World Intellectual Property Organization. (2007). *WIPO Patent Report: Statistics on Worldwide Patent Activities.* http://www.wipo.int/export/sites/www/freepublications/en/patents/931/wipo_pub_931.pdf (accessed on 12/17/2009).

FUNDING DATASETS

Federal Government of the United States of America. (2009). *Office of Management and Budget: Earmarks.* http://earmarks.omb.gov (accessed on 12/17/2009).

Federal Government of the United States of America. (2009). *Office of Management and Budget: MAX Homepage*. https://max.omb.gov/maxportal (accessed on 12/17/2009).

Federal Government of the United States of America. (2009). *USAspending.gov*. http://www.usaspending.gov (accessed on 12/17/2009).

National Institutes of Health. (2009) Electronic Research Administration: IMPAC II Introduction. http://era.nih.gov/impacii (accessed on 12/17/2009).

National Institutes of Health. (2009). *NIH Data Book*. http://report.nih.gov/nihdatabook (accessed 12/31/2009).

National Institutes of Health. (2009). *Query View Report System*. http://staffecb.cit.nih.gov/login.cfm (accessed on 4/30/2009

National Science Foundation. (2009). *Find Funding*. http://nsf.gov/funding (accessed on 12/17/2009).

National Science Foundation. (2010). *Division of Science Resources Statistics: Science and Engineering Statistics.* http://www.nsf.gov/statistics (accessed on 1/4/2010).

National Science Foundation. (2010). *WebCASPAR: Integrated Science and Engineering Resources Data System.* http://webcaspar.nsf.gov (accessed on 1/4/2010).

United States Department of Defense. (2009). *Defense Advanced Research Projects Agency: DARPA Budget.* http://www.darpa.mil/budget.html (accessed on 12/31/2009).

United States Department of Defense. (2010). *Defense Advanced Research Projects Agency: Research and Development Descriptive Summaries* (RDDS). http://www.dtic.mil/descriptivesum (accessed 1/4/2010).

United States Department of Health and Human Services. (2009). *Office of Inspector General: Fraud Exclusions Program*. http://www.oig.hhs.gov/fraud/exclusions.asp (accessed on 12/17/2009).

United States Small Business Administration. (2009). *TECH-Net*. http://tech-net.sba.gov/tech-net/public/dsp_search.cfm (accessed on 1/4/2010).

U.S. Department of Health and Human Services. (2009). *National Institutes of Health Research Portfolio Online Reporting Tool (RePORT)*. http://projectreporter.nih.gov (accessed on 12/17/2009).

TECHNOLOGY DATASETS

Deloitte Recap LLC. (2009). *Recap by Deloitte*. http://www.recap.com (accessed on 12/31/2009).

Ewing Marion Kauffman Foundation. (2009). *The Kauffman Firm Survey*. http://sites.kauffman.org/kfs (accessed on 12/31/2009).

Ewing Marion Kauffman Foundation. (2009). *Kauffman Index of Entrepreneurial Activity*. http://www.kauffman.org/research-and-policy/kauffman-index-of-entrepreneurial-activity.aspx (accessed on 12/31/2009).

Internal Revenue Service. (2009). SOI Tax Stats: Corporation Tax Statistics. http://www.irs.gov/taxstats/bustaxstats/article/0,,id=97145,00.html (accessed on 1/4/2010).

Jarmin, Ron S., Javier Miranda. (2009). *The Longitudinal Business Database*. IDEAS, Department of Economics, University of Connecticut. http://ideas.repec.org/p/cen/wpaper/02-17.html (accessed on 12/31/2009).

National Bureau of Economic Research. (2004). Historical Cross-Country Technology Adoption (HCCTA) Dataset. http://www.nber.org/hccta (accessed on 12/31/2009).

National Opinion Research Center, University of Chicago. (2009). NORC Data Enclave. http://www.norc.org/DataEnclave/Aboutus (accessed on 1/4/2010).

PIPRA. (2009). *PIPRA: Enabling Access to Public Innovation*. http://www.pipra.org (accessed on 12/17/2009).

Thomson Reuters. (2010). *SDC Platinum*. http://thomsonreuters.com/products_services/financial/financial_products/deal_making/investment_banking/sdc (accessed 1/4/2010).

The University of Michigan. (2009). *Panel Study of Entrepreneurial Dynamics*. http://www.psed.isr.umich.edu/psed (accessed 1/4/2010).

BIOMEDICAL/ECOLOGICAL DATASETS

Biotechnology Information Institute. (2009). *Biopharma: Biopharmaceutical Products in the U.S. and European Markets*. http://www.biopharma.com (accessed on 1/4/2010).

Cheung, K.H., K.Y. Yip, A. Smith, R. Deknikker, A. Masiar, M. Gerstein (2005) YeastHub: A Semantic Web Use Case for Integrating Data in the Life Sciences Domain. *Bioinformatics* 21 Suppl 1: i85-96.

Global Biodiversity Information Facility. Global Biodiversity Information Facility Data Portal. http://www.gbif.org.

Knewco Inc. 2009 Concept Web by Knewco: *Wikiproteins*. http://www.wikiprofessional.org (accessed on12/30/2009).

Lam, H.Y., E. Khurana, G. Fang, .P Cayting, N. Carriero, K.H. Cheung, M.B. Gerstein (2009) Pseudofam: The Pseudogene Families Database. *Nucleic Acids Res.* 37: D738-43.

Maddison, D. R. and K.-S. Schulz (eds.) 2007. The Tree of Life Web Project. http://tolweb.org.

National Library of Medicine, National Institutes of Health. (2009). *NCBI: GenBank*. http://www.ncbi.nlm.nih.gov/Genbank (accessed on 4/30/2009).

NCBI. (2009). *Plant Genomes Central: Genome Projects in Progress*. http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html (accessed on 12/17/2009).

Smith, A.K., K.H. Cheung, K.Y. Yip, M. Schultz, M.K. Gerstein (2007) LinkHub: A Semantic Web System That Facilitates Cross-Database Queries and Information Retrieval in Proteomics. *BMC Bioinformatics* 8 Suppl 3: S5.

United States Department of Agriculture. (2009). *Natural Resources Conservation Service Plants Database*. http://plants.usda.gov (accessed on 12/17/2009).

U.S. Geologic Survey. National Map Viewer. http://nmviewogc.cr.usgs.gov/viewer.htm.

UTEK Corporation. (2009). *UTEK Knowledge Express*. http://www.knowledgeexpress.com (accessed on 12/31/2009).

Wikimedia Foundation, Inc. (2009). *Wikispecies*. http://species.wikimedia.org (accessed on12/30/2009).

GENERAL DATASETS AND FEDERAL REPORTS

Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, Gavin Wright (Eds). (2009). *Historical Statistics of the United States: Millenial Edition Online*, Cambridge University Press. http://hsus.cambridge.org/HSUSWeb/HSUSEntryServlet (accessed on 12/17/2009).

Central Intelligence Agency. (2009). *The World Factbook*. https://www.cia.gov/library/publications/the-world-factbook (accessed on 12/17/2009).

Computing Research Association. (2007). *CRA Taulbee Survey*. http://www.cra.org/statistics (accessed on 1/4/2010).

Illuminating Engineering Society. (2009). *Awards and Services*. http://www.iesna.org/programs/awards.cfm (accessed on 12/17/2009).

Inter-University Consortium for Political and Social Research. (2009). *National Science Foundation Surveys of Public Attitudes toward and Understanding of Science and Technology, 1979-2001*. http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/04029 (accessed on 12/31/2009).

National Research Council. (2007) The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington D.C.: National Academies Press. http://www.nap.edu/catalog.php?record_id=11902. (accessed on 1/6/10)

National Science Board. (2006). *Science and Engineering Indicators 2006*. Arlington, VA: National Science Foundation. http://www.nsf.gov/statistics/seind06 (accessed on 1/4/2010).

National Science Foundation. (2007). *eJacket Privacy Impact Assessment, Version 1.0*. http://www.nsf.gov/pubs/policydocs/pia/ejacket_pia.pdf (accessed on 4/30/2009).

National Science Foundation. (2009). *Division of Science Resources Statistics (SRS): Science and Engineering Statistics*. http://www.nsf.gov/statistics (accessed on 12/17/2009).

National Science Foundation. (2009). *Research.gov: Powering Knowledge and Innovation*. http://www.research.gov (accessed on 12/17/2009).

National Science Foundation. (2009). *S&E State Profiles*. http://www.nsf.gov/statistics/states (accessed 1/4/2010).

National Science Foundation. (2009). *Science and Technology Pocket Data Book*. http://www.nsf.gov/statistics/databook (accessed on 1/4/2010).

National Science Foundation. (2009). *Survey of Earned Doctorates*. http://www.nsf.gov/statistics/srvydoctorates (accessed on 1/4/2010).

Organisation for Economic Co-operation and Development. (2009). *OECD: Data Warehouse and Databases*. http://www.oecd.org/statsportal/0,3352,en_2825_293564_1_1_1_1_1,00.html (accessed on 12/17/2009).

Project on Emerging Nanotechologies. (2010). *Consumer Products: An Inventory of Nanotechnology-Based Consumer Products Currently on the Market*. http://www.nanotechproject.org/inventories/consumer (accessed on 1/4/2010).

Socioeconomic Data and Applications Center. (2008). *Gridded Population of the World and the Global Rural-Urban Mapping Project*. Columbia University. http://sedac.ciesin.columbia.edu/gpw (accessed on 10/16/2008).

United States Census Bureau. (2009). *Fact Finder: Your Source for Population, Housing, Economic and Geographic Data*. http://factfinder.census.gov/home/saff/main.html?_lang=en (accessed on 12/17/2009).

United States Census Bureau. (2009) Federal, State, and Local Governments: *Federal Assistance Award Data System*. http://www.census.gov/govs/www/faads.html (accessed on 12/17/2009).

United States Department of Agriculture. (2009*). Agricultural Research Service: USDA National Nutrient Database for Standard Reference.* http://www.ars.usda.gov/main/site_main.htm?modecode=12-35-45-00 (accessed on 12/17/2009).

United States Department of Commerce. (2009). *Bureau of Economic Analysis*. http://www.bea.gov (accessed on 12/17/2009).

United States Department of Labor. (2009). *Bureau of Labor Statistics*. http://www.bls.gov (accessed on 12/17/2009).

U.S. Department of Labor Office of Foreign Labor Certification, State of Utah. (2009). *Foreign Labor Certification Data Center: Online Wage Library*. http://www.flcdatacenter.com (accessed on 12/17/2009).

U.S. Environmental Protection Agency. (2009). *National Environmental Policy Act (NEPA): Environmental Impact Statement (EIS) Database*. http://www.epa.gov/oecaerth/nepa/eisdata.html (accessed on 12/17/2009).

Zucker, Lynne G. and Michael R. Darby. (Forthcoming). Legacy and New Databases for Linking Innovation to Impact. In Fealing, Kay Husbands, Julia Lane, John Marburger, Stephanie Shipp, Bill Valdez (Eds.). Science of Science Policy: Handbook. http://scienceofsciencepolicyhandbook.net/contact.html (accessed 1/21/2010).

Tools

TEMPORAL

Macrofocus GmbH. (2008). *Macrofocus*. http://macrofocus.com (accessed on 12/31/2009).

GEOLOCATION

Gahegan, M., M. Takatsucka, M. Wheeler, F. Hardisty. (2002). GeoVISTA Studio: An Integrated Suite of Visualization and Computation Methods for Exploration and Knowledge Construction in Geography. *Computer, Environment and Urban Systems, 26*(4), 267-292.

Kapler, Thomas, William Wright. (2005). GeoTime Information Visualization. *Information Visualization, 4*(2), 136-146.

Regents of University of California. (2008). *Tobler's Flow Mapper*. Center for Spatially Integrated Social Science. http://csiss.ncgia.ucsb.edu/clearinghouse/FlowMapper/ (accessed on 10/01/2008).

Takatsuka, M., M. Gahegan. (2002). GeoVISTA Studio: A Codeless Visual Programming Environment for Geoscientific Data Analysis and Visualization. *The Journal of Computers & Geosciences, 28*(10), 1131-1144. http://www.geovistastudio.psu.edu/jsp/index.jsp (accessed on 12/28/2009).

TOPICAL/TEXT MINING

CASOS. (2009a). *AutoMap: Extract, Analyze and Represent Relational Data from Texts*. http://www.casos.cs.cmu.edu/projects/automap (accessed on 4/28/2009).

Gerstein, M, M Seringhaus, S Fields (2007) Structured Digital Abstract Makes Text Mining Easy. *Nature* 447: 142.

Knewco Inc. (2008). *Knewco: Knowledge Redesigned*. http://www.knewco.com (accessed on 10/02/2008).

Lin, Xia. (2008). *Visual Concept Explorer*. http://cluster.cis.drexel.edu/vce (accessed on 10/02/2008).

Paley, W Bradford. (2002). *TextArc*. http://www.textarc.org (accessed on 10/02/2008).

Persson, Olle. (2008). *BIBEXCEL*. http://www.umu.se/inforsk/Bibexcel (accessed on 10/02/2008).

Rzhetsky, A,, M. Seringhaus, M. Gerstein (2008) Seeking a New Biology Through Text Mining. *Cell* 134: 9-13.

Search Technology, Inc. (2008). *Vantage Point: Turn Information into Knowledge*. http://www.thevantagepoint.com (accessed on 12/31/2009).

Thinkmap Inc. (2008). Thinkmap Visual Thesaurus. http://www.visualthesaurus.com (accessed on 4/30/2009).

NETWORKS

AbsInt Angewandte Informatik GmbH. (2009). *aiSee: A picture is worth a thousand words*. http://www.aisee.com (accessed on 4/28/2009).

Adar, Eytan. (2006). *GUESS: A Language and Interface for Graph Exploration.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries & Gary Olso (Eds.), New York: ACM Press, pp. 791-800.

Adar, Eytan. (2007). *GUESS: The Graph Exploration System*. http://graphexploration.cond.org (accessed on 10/02/2008).

Advanced Technology Assessment. (2009). *Software: DyNet*. http://www.atalab.com/software (accessed on 4/28/2009).

Analytic Technologies. (2009b). *KeyPlayer 1.44*. http://www.analytictech.com/keyplayer/keyplayer.htm (accessed on 04/28/2009).

Analytic Technologies. (2009a). *E-Net: Ego-Network Analysis*. http://www.analytictech.com/e-net/e-net.htm (accessed on 4/28/2009).

Analytic Technologies. (2009c). *NetDraw*. http://wwwanalytictech.com/downloadnd.htm (accessed on 4/28/2009).

AT&T Research Group. (2008). *Graphviz-Graph Visualization Software*. http://www.graphviz.org (accessed on 12/17/2009).

Auber, David. (2003). *Tulip: A Huge Graph Visualisation Framework.* Berlin: Springer-Verlag.

Batagelj, Vladimir, Andrej Mrvar. (1998). Pajek - Program for Large Network Analysis. *Connections, 21*(2), 47-57.

Bender-deMoll, Skye, Daniel A. McFarland. (2009). *SoNIA: Social Network Image Animator*. http://www.stanford.edu/group/sonia/documentation/index.html (accessed on 04/28/2009).

Bergström, Peter, Jr. Whitehead, E. James. (2006). *CircleView: Scalable Visualization and Navigation of Citation Networks*. http://www.cs.ucsc.edu/~ejw/papers/bergstrom_ivica2006_final.pdf (accessed on 10/16/2008).

Borgatti, S. P., M. G. Everett, L. C. Freeman. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. http://www.analytictech.com/downloaduc6.htm (accessed on 10/2/2008).

Brandes, Ulrik, Dorothea Wagner, Project Leaders. (2008). *Visone: Analysis and Visualization of Social Networks*. http://visone.info (accessed on 10/02/2008).

CAIDA: Cooperative Association for Internet Data Analysis. (2009). *Otter: Tool for Topology Display*. http://www.caida.org/tools/visualization/otter (accessed on 4/28/2009).

CASOS. (2009b). *ORA: Organizational Risk Analyzer*. http://www.casos.cs.cmu.edu/projects/ora (accessed on 4/28/2009).

Chen, Chaomei. (2007). *CiteSpace*. http://cluster.cis.drexel.edu/~cchen/citespace (accessed on 4/28/2009).

CiNii. (2004). *Researchers Link Viewer*. http://ri-www.nii.ac.jp/ComMining1/index.cgi (accessed on 10/02/2008).

Cummings, Jonathon N. (2009). *NetVis Module-Dynamic Visualization of Social Networks*. http://www.netvis.org (accessed on 4/28/2009).

Cytoscape Consortium. (2008). *Cytoscape*. http://www.cytoscape.org (accessed on 4/28/2009).

Douglas, Shawn M., Gaetano T. Montelione, Mark Gerstein. (2005a). PubNet: A Flexible System for Visualizing Literature-Derived Networks. *Genome Biology, 6*(9), R80.

Douglas, Shawn M., Gaetano T. Montelione, Mark Gerstein. (2005b). *PubNet: Publication Network Graph Utility*. http://pubnet.gersteinlab.org (accessed on 10/02/2008).

Fekete, Jean-Daniel. (2004). *The Infovis Toolkit.* Proceedings of the 10th IEEE Symposium on Information Visualization: IEEE Press, pp. 167-174.

FMS, Inc. (2009). *Sentinel Visualizer: The Next Generation of Data Visualization*. http://www.fmsasg.com/Products/SentinelVisualizer (accessed on 4/28/2009).

Graphviz Team. (2009). *Graphviz: Graph Visualization Software*. http://www.graphviz.org (accessed on 04/28/2009).

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Martina Morris. (2003). *statnet: Software Tools for the Statistical Modeling of Network Data* http://statnetproject.org (accessed on 4/28/2009).

Harzing, Anne-Wil. (2008). *Publish or Perish: A citation analysis software program*. http://www.harzing.com/resources.htm (accessed on 4/22/08).

Heer, Jeffrey, Stuart K. Card, James A. Landay. (2005). *Prefuse: A Toolkit for Interactive Information Visualization.* Proceedings of the 2005 Conference on Human Factors in Computing Systems CHI 2005, Gerrit C. van der Veer & Carolyn Gale (Eds.), New York: ACM Press, pp. 421-430.

HistCite Software LLC. (2008). *HistCite: Bibliometric Analysis and Visualization Software*. http://www.histcite.com (accessed on 10/03/2008).

Huisman, M., M. A. J. V. Duijn. (2003). StOCNET: Software for the Statistical Analysis of Social Networks. *Connections, 25*(1), 7-26.

Hunscher, Dale. (2009). *UrlNet Python Library*. http://code.google.com/p/urlnet-python-library/ (accessed on 4/28/2009).

Information Visualization Laboratory. (2005). *InfoVis Cyberinfrastructure*. School of Library and Information Science, Indiana University. http://iv.slis.indiana.edu (accessed on 10/03/2008).

Kalamaras, Dimitris V. (2009). *SocNetV*. http://socnetv.sourceforge.net/news.html (accessed on 4/28/2009).

Kautz, Henry, Bart Selman. (1997). *Welcome to ReferralWeb*. http://www.cs.rochester.edu/u/kautz/referralweb (accessed on 4/28/2009).

Keyhubs, LLC. (2009). *The Company Behind Your ORG Chart: A Simple Online Tool for Mapping Informal Networks*. http://www.keyhubs.com (accessed on 4/28/2009).

Kintip: Traitement Informatique de la Parenté. (2009). *PUCK*. http://www.kintip.net/content/view/55/21 (accessed on 4/28/2009).

Krackhardt, David. (2006). *KrackPlot Software*. http://www.andrew.cmu.edu/user/krack/krackplot.shtml (accessed on 4/28/2009).

Krebs, Valdis. (2008). *InFlow - Software for Social Network Analysis & Organizational Network Analysis*. Orgnet.com. http://www.orgnet.com/inflow3.html (accessed on 10/02/2008).

Leonard, Steve. (2009). *UNISoN: UseNet Incorporates Social Networks*. http://unison.sleonard.co.uk (accessed on 4/28/2009).

Leydesdorff, Loet. (2008). *Software and Data of Loet Leydesdorff*. http://users.fmg.uva.nl/lleydesdorff/software.htm (accessed on 10/01/2008).

MelNet. (2005). *PNet*. http://www.sna.unimelb.edu.au/pnet/pnet.html (accessed on 4/28/2009).

Müller, Hans-Georg. (2005). *UDrawGraph: Welcome*. http://www.informatik.uni-bremen.de/uDrawGraph/en/index.html (accessed on 4/28/2009).

Noldus Infomation Technology. (2009). *MatMan*. http://www.noldus.com/human-behavior-research/products/matman (accessed on 04/28/2009).

NWB Team. (2006). Network Workbench Tool. Indiana University, Northeastern University and University of Michigan. http://nwb.slis.indiana.edu (accessed 12/31/2009).

O'Madadhain, Joshua, Danyel Fisher, Scott White. (2008). *JUNG: Java Universal Network/Graph Framework*. University of California. http://jung.sourceforge.net (accessed on 10/02/2008).

On, Josh. (2004). *They Rule*. http://www.theyrule.net (accessed on 10/02/2008).

Opsahi, Tore. (2009). *tnet: Analysis of Weighted and Longitudinal Networks*. http://opsahl.co.uk/tnet/content/view/15/27 (accessed on 4/28/2009).

Optimice. (2009). *ONA Surveys*. http://www.onasurveys.com (accessed on 04/28/2009).

Richards, William D. (1995). *Negopy 4.302*. http://www.sfu.ca/personal/archives/richards/Pages/negopy.htm (accessed on 4/28/2009).

Richards, William D., Andrew J. Seary. (2007). *The Vancouver Network Analysis Team: FATCAT*. http://www.sfu.ca/personal/archives/richards/Pages/fatcat.htm (accessed on 04/28/2009).

Seary, Andrew J. (2005). MultiNet: An Interactive Program for Analyzing and Visualizing Networks. Ph.D. Thesis, Faculty of Applied Science, Simon Fraser University. http://ir.lib.sfu.ca/handle/1892/2451 (accessed 12/31/2009).

Shneiderman, Ben. (1992). Tree Visualization with Tree-Maps: 2-D Space-Filling Approach. *ACM Transactions on Graphics, 11*(1), 92-99. http://www.cs.umd.edu/hcil/treemap-history (accessed on 10/1/2008).

Siek, Jeremy, Lie-Quan Lee, Andrew Lumsdaine. (2002). *The Boost Graph Library: User Guide and Reference Manual*. New York: Addison-Wesley.

Snijders, Tom A. B. (2009). Social Network Analysis: Snowball. http://stat.gamma.rug.nl/snijders/socnet.htm (accessed on 4/28/2009).

Sonivis. (2009). Sonivis. http://www.sonivis.org (accessed on 4/28/2009).

SourceForge, Inc. (2009). EgoNet. http://sourceforge.net/projects/egonet (accessed on 4/28/2009).

Tanglewood Research, Inc. (2009). Network Genie. https://secure.networkgenie.com (accessed on 4/28/2009).

Technical University of Dortmund, University of Cologne, oreas GmbH. (2007). OGDF: Open Graph Drawing Framework. http://www.ogdf.net/ogdf.php (accessed on 4/28/2009).

TECLAB. (2009). *IKNOW: Inquiring Knowledge Networks on Web*. http://sonic.ncsa.uiuc.edu/software.htm (accessed on 4/28/2009).

The igraph Project. (2009). *The igraph Library*. http://igraph.sourceforge.net (accessed on 4/28/2009).

TouchGraph LLC. (2007). *TouchGraph Google Browser*. http://www.touchgraph.com/TGGoogleBrowser.html (accessed on 10/02/2008).

Trier, Matthias. (2008). *Commetrix: Dynamic Network Visualization and Analysis*. http://www.commetrix.de (accessed on 4/28/2009).

Tsuji, Ryuhei. (2008). *Ryuhei Tsuji's Software: PermNet Ver0.94*. http://rtsuji.jp/en/software.html (accessed on 4/28/2009).

Tulip Software Team. (2009). *Tulip*. http://www.tulip-software.org (accessed on 4/28/2009).

Usher, Abe. (2006). *libsna: The Library for Social Network Analysis*. http://www.libsna.org (accessed on 4/28/2009).

Van Eck, N.J., L. Waltman. (2009) *VOSviewer*. http://www.vosviewer.com/documentation (accessed on 10/16/2009)

Vicsek, Tamás, Dániel Ábel, Balázs Adamcsek, Imre Derényi, Illés Farkas, Gergely Palla, Péter Pollne. (2009). *CFinder: Clusters and Communities – Overlapping Dense Groups in Networks*. http://www.cfinder.org (accessed on 4/28/2009).

yWorks. (2009). *yWorks: The Diagramming Company*. http://www.yworks.com (accessed on 4/28/2009)

BIOMEDICAL

Burba, A.E, ,U. Lehnert, E.Z. Yu, M. Gerstein (2006) Helix Interaction Tool (HIT): a Web-Based Tool for Analysis of Helix-Helix Interactions in Proteins. *Bioinformatics* 22: 2735-8.

Flores, S.C, K.S. Keating, J. Painter, F. Morcos, K. Nguyen, E.A. Merritt, L,A, Kuhn, M,B, Gerstein (2008) HingeMaster: normal mode hinge prediction approach and integration of complementary predictors. Proteins 73: 299-319.

Gerstein, M, S.M. Douglas (2007) RNAi Development. *PLoS Comput Biol* 3: e80.

Korbel, J.O., A. Abyzov, X.J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, M.B. Gerstein (2009) PEMer: A Computational Framework with Simulation-Based Error Models for Inferring Genomic Structural Variants from Massive Paired-End Sequencing Data. *Genome Biol* 10: R23.

OmniViz Inc. (2008). *biowisdom*. http://www.biowisdom.com/content/omniviz (accessed on 4/30/2009).

Richardson Lab Kinemages. (2009). *3D Analysis: "The Mage Page"*. http://kinemage.biochem.duke.edu/kinemage/magepage.php (accessed on 4/28/2009).

Rozowsky, J., G. Euskirchen, R.K. Auerbach, Z.D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, M.B. Gerstein (2009) PeakSeq Enables Systematic Scoring of ChIP-Seq Experiments Relative to Controls. *Nat Biotechnol* 27: 66-75.

Seringhaus, M.R., P.D. Cayting, M.B. Gerstein (2008) Uncovering Trends in Gene Naming. *Genome Biol* 9: 401.

Theoretical and Computational Biophysics Group. (2009). VMD: Visual Molecular Dynamics. Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign. http://www.ks.uiuc.edu/Research/vmd (accessed on 12/31/2009).

Yip, K.Y., H. Yu, P.M. Kim, M. Schultz, M. Gerstein (2006) The tYNA Platform for Comparative Interactomics: A Web Tool for Managing, Comparing and Mining Multiple Networks. *Bioinformatics* 22: 2968-70.

Yu, H., R. Jansen, G. Stolovitzky, M. Gerstein (2007) Total Ancestry Measure: Quantifying the Similarity in Tree-Like Classification, with Genomic Applications. *Bioinformatics* 23: 2163-73.

Zhang, Z, N. Carriero, D. Zheng, J. Karro, P.M. Harrison, M. Gerstein (2006) PseudoPipe: An Automated Pseudogene Identification Pipeline. *Bioinformatics* 22: 1437-9.

GENERAL

Butts, Carter. (2009). *Carter's Archive of S Routines for the R Statistical Computing Environment*. http://erzuli.ss.uci.edu/R.stuff (accessed on 4/28/2009).

CERN - European Organization for Nuclear Research (Hoschek). (2004). *Colt: Open Source Libraries for High Performance Scientific and Technical Computing in Java*. http://www-itg.lbl.gov/~hoschek/colt (accessed on 10/01/2008).

Heijs, Anton. (2008). *Treparel*. Treparel Information Solutions B.V. http://www.treparel.com (accessed on 10/02/2008).

Hewlett-Packard Development Company, L.P. (2009). *Zoomgraph*. http://www.hpl.hp.com/research/idl/projects/graphs/zoom.html (accessed on 4/28/2009).

Hunter, John. (2008). *Matplotlib*. http://matplotlib.sourceforge.net (accessed on 10/30/2007).

Infospace, Inc. (2009). *WebCrawler: The Web's Top Search Engines Spun Together*. http://www.webcrawler.com (accessed on 12/17/2009).

Kartoo. (2008). *Kartoo Visual Meta Search Engine*. http://www.kartoo.com (accessed on 10/02/2008).

Oculus Info Inc. (2006). *Oculus Excel Visualizer*. http://www.oculusinfo.com/papers/ExcelVizWhitepaper-final.pdf (accessed on 10/02/2008).

The Morphix Company. (2009). *MetaSight*. http://www.morphix.com/1_MetaSight/introduction.htm (accessed on 4/28/2009).

The R Foundation. (2008). *The R Project for Statistical Computing*. http://www.r-project.org (accessed on 10/02/2008).

University of Illinois at Chicago. (2007). *Arrowsmith Project Homepage*. http://arrowsmith.psych.uic.edu (accessed on 10/16/2008).

Williams, Thomas, Colin Kelley. (2008). *gnuplot homepage*. http://www.gnuplot.info (accessed on 7/17/08).

Appendix D: Promising Research Projects

During both workshops, participants were asked to take one hour of their time to write down their "2nd Best Idea" for a project proposal. The request for the 2nd instead of 1st best ideas addresses the fact that many of the participants are commercial or scientific competitors. The resulting set of proposed science and technology "sweet spots" and corresponding research projects demonstrates the expertise, innovativeness, and proposal writing skills of those participants that went for the challenge. With permission of the "intellectual property" owners, some of these project proposals are reproduced here. Contact information is provided for each author and we highly encourage you to contact the original authors if you are interested to collaborate on or to fund one of these projects.

Proposals are divided into projects that address biomedical/ecological research challenges and proposals that aim to advance SoS research and the scientific infrastructure. Within each section, proposals are organized by last author name. Minor edits were made to improve readability.

W1 **Attended Workshop 1** W2 **Attended Workshop 2**

Biomedical/Ecological Research

SoS and Science Infrastructure Research

Kei Cheung

Use of Semantic, Social, and Visual Networks to Build a Cyberinfrastructure for Enabling Discovery and Innovation in Genome Sciences and Neurosciences

Center for Medical Informatics
Yale University School of Medicine
333 Cedar Street, PO Box 208009
New Haven, CT 06520, USA

kei.cheung@yale.edu

W1

POTENTIAL COLLABORATORS

Mark Gerstein, Katy Börner, Neo Martinez, Richard Bonneau, Stephen Miles Uzzo, Luís M. A. Bettencourt, Kevin Boyack and John T. Bruer.

SUMMARY

As the quantity/diversity of biomedical data and the number of tools that are used to analyze and visualize such data are growing rapidly on the Internet, there is an urgent need to develop a cyberinfrastructure that allows these data and tools to be easily published, shared, discovered and linked for enabling scientific collaboration, knowledge discovery and technological innovation in the biomedical domain. To this end, we will collaborate closely with experts in the genomics and neuroscience fields in exploring how to creatively use the semantic and social Web, as well as graph networks to help genome/neuro-scientists to describe and integrate data and tools of their interest to mine and visualize data for gaining new scientific knowledge.

INTELLECTUAL MERIT

The proposed activity involves interdisciplinary research spanning biomedical sciences, database, Semantic Web, social science, sustainability, and policy research. Not only does it help advance knowledge and understanding within each of these fields, but it also produces advancement across different fields. For example, data and tool integration helps genome/neuroscience researchers to explore their studies in a broader context, generating new hypotheses based on integration of data derived from other studies featuring related hypotheses. Such hypothesis-driven data/tool integration helps expand the frontier of informatics research, including Semantic Web and Social Web, as the complexity and diversity of data, as well as the nature of collaboration involved in genome science or neuroscience presents challenging issues to Semantic Web and Social Web.

BROADER IMPACTS

The proposed cyberinfrastructure potentially benefits the genome and neuroscience communities. In addition, such a cyberinfrastructure can be generally applied to other fields of science. It can also serve as a platform for education. Teachers and students (including K-12, undergraduate, and graduate programs) can access educational materials through the cyberinfrastructure.

TRANSFORMATIVE POWER

The proposed research transforms the ways computers can help genome/neuro scientists interact and do research via the Internet. In addition, it transforms the way scientists publish, share, and access data/tools via the Web. It also transforms the way students learn about genome science and neuroscience via the Internet.

Mark Gerstein

Developing Intuitions and Formalisms from Social Network Analysis to Understand the Coupling of Activity Patterns Between Tissues in a Diseased Organ

Albert Williams Professor of Biomedical Informatics,
Molecular Biophysics & Biochemistry, and Computer Science
Yale University
MBB, PO Box 208114
New Haven, CT 06520, USA

Mark.Gerstein@Yale.edu
http://bioinfo.mbb.yale.edu

W1

POTENTIAL COLLABORATORS

M. Rubin (Cornell), Kei Cheung (Yale) and Social Network and Map of Science Person (to be determined).

SUMMARY

An organ contains many different tissues, which, in turn, contains populations of cells with complex yet distinct gene activity patterns and molecular networks of interactions. In a sense, an organ is like a country containing many cities, which, in turn, have a distinct activity pattern that shares some common themes.

In a disease, the activity patterns (i.e. gene expression profiles) of these different tissues become dis-regulated to varying degrees. In almost all cases, the detailed molecular mechanisms and biological systems underlying the pathological effects of organ disfunction remain unknown. The key question addressed here is how varying activity patterns in different tissues are coupled together as they become disregulated. We aim to understand using a systems level approach.

We hypothesize that genetic differences between the cells in each tissue (polymorphisms and mutations), epigenetic differences and environmental factors play a pivotal role in the poorly characterized pathological sequence that leads to the complex, multifactorial and lethal disorder. We would propose to decipher the genomic expression patterns that form the phenotypic determinants. One potential model system to focus on would be prostate cancer. In this we would plan on collaborating with M. Rubin at Cornell Medical School in New York. We would focus on the differences between lethal prostate cancer (which grows quickly), indolent prostate cancer (which grows slowly and for which it is inadvisable to operate), and normal non-cancerous tissue.

Our overall approach to understanding organ dis-regulation faces two main informatics challenges:

1. We need to be able to characterize, in a precise fashion, the different phenotypes of the different affected tissues. To address this we would propose to work on ontology construction within a targeted domain. In this system, we would want a way of hierarchically grouping the tissues into those that were more or less similar to each other and also to characterize which ones "touched" each other (i.e. were in spatial proximity in the organ). We would use this ontology to guide our characterization of the soft coupling of the gene expression networks.

2. We need a new paradigm for thinking about how coupled activity patterns in different proximal networks affect each other. For this, we would look to the social sciences and theories of complex systems. We would study how networks of activity (e.g. in different scholarly disciples or in different cities) can be softly coupled and how this coupling affects how they evolve. There is much data available on this from bibliometric and economic databases. Recent research in this field, for instance, has illustrated the different publishing patterns of different disciplines and how epidemiological models (using differential equations) can be used to model the spread of ideas between communities. We would not expect these characterizations for models to predict organ function, per se; rather our thrust would be to use them to provide intuitions and simple, easy-to-understand toy problems that we could abstract key metrics and principles from, which then could be applied to biomedical challenge at hand. Furthermore, we would expect that these problems would be ideal platforms to develop tools and visualizations that could then be applied to molecular networks.

Specifically, we would utilize expression profiling data (RNA and micro-RNA-based platforms) of key tissues involved (i.e. the prostate). There is much publicly available data available from the National Cancer Institute's (NCI) Cancer Genome Anatomy Project (CGAP) database. We would couple these

human molecular networks available from the Human Protein Reference Database (HPRD) and recent ChIP-chip experiments (data available from the UCSC genome browser).

We will next conduct hypothesis-driven, comparative bioinformatics analyses and modeling of the molecular networks in each of the key tissues, derive predictions of dysregulated biomolecular systems (genes, pathways, biological processes), and validate these results via targeted gene expression experiments – to see if we can predict an expression profile in a particular coupled tissue system that latter could be validated. The data for the latter, of course, would have to be furnished through an experimental collaboration. The key question we would strive to answer is how the activity patterns (i.e. expression patterns) in one tissue are coupled to those in another.

BROADER IMPACTS

Ultimately, we believe tissue-specific genetic signatures will lead to identification of novel targets for pharmacological intervention. Our research design combines relevant tissues and high throughput validation with ontology-anchored phenotypic integration and mining techniques, well-organized genome-phenome datasets, and association formalisms (such as molecular networks and phenotypic ontologies) to reverse engineer gene regulation. Data mining of ontology-anchored networks has only rarely been systematically and effectively applied to disease. Thus, this application will utilize biological, genomic, and computational systems biology approaches in tandem with the theory of complex systems to generate hypotheses and assess the salient features of molecular networks underpinning organ level phenotypes through biological modeling and validation. The specific biological aims of this application are:

i. To utilize systems biology approaches to define molecular signatures for the tissue-specific involvement
ii. To perform multiscale computational analyses of molecular mechanisms of phenotypes to derive mechanistic models and predictions
iii. To validate the molecular predictions through the use of targeted gene expression and RT-PCR assays

Mark Gerstein

Developing an Infrastructure for Creating and Browsing Multimodal Abstracts of Scientific Papers

Albert Williams Professor of Biomedical Informatics,
Molecular Biophysics & Biochemistry, and Computer Science
Yale University
MBB, PO Box 208114
New Haven, CT 06520, USA

Mark.Gerstein@Yale.edu
http://bioinfo.mbb.yale.edu

W2

POTENTIAL COLLABORATORS

We would like to involve authors in the preparation of these summaries, which act both as a computerized snapshot of pertinent facts in a particular article and as a gateway to future access.

SUMMARY

Overview

We propose to develop an approach for better organizing all the scientific literature – proactively, instead of retrospectively – through the creation of a multimodal abstract. The multimodal abstract would have a number of parts:

1. The traditional technical abstract
2. A structured digital abstract meant to be parsed by computers
3. A mainstream abstract meant to be understandable to the mainstream educated public
4. A non-textual abstract (i.e. graphical, video, and audio), making the content of the paper available in a non-textual fashion.

Each version of the abstract would seek to the make the content of a traditional scientific publication useful and accessible to a different constituency. After developing standards for creating these abstracts we would propose tools for searching through and browsing large numbers of them, perhaps in a graphical fashion.

Motivation

Structured Abstract

Scientific literature is expanding at a staggering rate. It would take five years to read the scientific material produced in a 24 hour period, and 80% of this information is stored as unstructured text. The current curation process is a bottleneck and will become more so as this volume of literature continues to increase. The only viable solution is to employ the vast pool of article writers to help keep up with curation. Why curate? As the boundary continues to blur between scientific journals and databases, it is important to ensure smooth integration of scientific knowledge between text-based and digital forms. Currently, we publish traditional manuscripts and depend upon teams of curators – who necessarily lack intimate familiarity with every paper they process – to extract meaningful findings and transplant them to databases. We propose embedding this curatorial step in the publication process itself, with author input and subject it to peer review.

Mainstream Abstract

A concomitant problem with the growth of scientific literature is the increasing jargonization and balkanization of scientific terminology. Most scientists can not understand each other. This is the motivation for trying to create an abstract in simple lay language.

Non-textual abstract

Finally, much of the information in scientific papers is in itself fundamentally not textual. Most scientific writing results from the desire to explain mathematical concepts or observations about nature. Fundamentally, these efforts use language to transform complex scientific observations and mathematical ideas into a textual representation. Often, papers are built around figures and tables, with the exact text being only a secondary consideration. Scientists are trained to be cautious and as accurate as possible in their writing, as colleagues need accurate text to then translate the results back into their mathematical or visual observations. The exact bounds of the factual observations are of primary importance and central to the process. This is the motivation for the graphical and video abstracts.

Author and Journal Involvement

We would like to involve authors in the preparation of these summaries, which act both as a computerized snapshot of pertinent facts in a particular article and as a gateway to future access.

For Structured Digital Abstract

By assigning digital identifiers to genes, proteins and other biological entities up front and with author input, we reduce the need for challenging back-end curation. The most challenging step in text-mining is often spotting proper names and identifiers among reams of text, and mandatory author input in picking out such identifiers should greatly ease the process. The authors of a study are both the most versed in its contents, and most vested in seeing their results deposited accurately in databases. One potential idea would be to use the automated markup system in wikiprofessional.org to help generate an initial version of a structured abstract that authors could then fix.

For Mainstream abstract

The mainstream abstract could be written by the authors; though it would probably be helpful to get strong oversight in this process by an editor or even a PR person, to ensure that the language was as simple as possible.

For Non-textual Abstract

Furthermore, we would imagine that the authors of an article would be most interested in making it broadly accessible in mainstream and graphical formats. We would like the graphic, video and auditory abstracts to build on the existing SciVee effort. SciVee (http://www.scivee.tv) caters to the "YouTube generation" of video consumers; after all, they are the next Nobel Laureates. Using PLoS and other content taken from PubMed Central, SciVee provides a video-on-demand service that mashes up video provided by the authors and the paper content into what is called a "pubcast." Who better to provide this intermediate view than one of the authors by giving a five-to-ten-minute video presentation of the content of the paper? We imagine that the video can be made using a webcam and software standard on a PC or Mac, or done more professionally. Podcasts may be what the reader is seeking when video seems like overkill. Perhaps a podcast of the traditional journal issue is desirable: while jogging or walking to the laboratory you could get an overview of the latest issue of this journal, presented either by the authors of papers in that issue or by a journal editor. This takes eToCs to a new level and medium. It seems that every student walking around campus has the means in their hands and ears to take advantage of this today. This could also benefit scientists with disabilities. Science, Nature, and other journals are using podcasts regularly, and they seem to be well received.

Development of a Querying, Browsing, and Visualization Infrastructure

A major part of this process would be developing appropriate tools for making use of the structured abstracts, i.e. being able to search more quickly for ideas and to browse large bodies of literature. Visualization will be important in the latter. Obviously, this discussion is a little short and much more could be described in relation to this.

Conclusion

We think that the above will greatly improve electronic availability of scientific information, and help make this vast body of knowledge readily available to the researchers who need it most. We envision applying current text-mining, curatorial, graphical, and video approaches at the pre-print stage, in concert with an author-completed web form, to generate early versions of a multimodal abstract. This will then be subject to peer review, updated and finally approved by authors and editors. The input of a dedicated curator will also be helpful when generating this abstract, particularly in the pilot phase. We hope production of the multi-modal abstract will become a standard step in the pre-print publication process.

Bongshin Lee

Interactive Visualizations for Distributed Biomedical Data Repositories

Visualization and Interaction Research (VIBE)
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

bongshin@microsoft.com

W2

POTENTIAL COLLABORATORS

- Biomedical researchers who can describe their needs and desires
- Social scientists who can analyze the network of biomedical researchers and their publications/data
- Computer scientists who can provide ways to clean up and transform the legacy data, and build distributed data repositories and a set of visualization systems to access them

SUMMARY

In the field of biomedical research, we have a huge amount data–the raw experiment data as well as the publications from previous research. However, we do not have a good understanding of the research (e.g., research topics, research trends, and expert researchers for a specific topic) and many researchers often make redundant efforts. This is mainly because previous research, especially the experiment data, is not effectively shared. Biomedical researchers are not willing to share their data because they do not benefit from doing so. Furthermore, they do not have good distributed data repositories that can be easily searched and browsed.

We need to develop two main components; not only to facilitate the effective sharing of biologists' data, but also to encourage them to collaborate. First, we need to build distributed biomedical data repositories along with appropriate recognition and rewarding mechanisms. For example, it is important to develop ways to recognize sharing the source of research (i.e., good raw data), as well as sharing the result/output of the research (i.e., publications). Second, we need to develop a set of interactive, web-based visualizations to help biologists effectively search, browse, and understand those large, distributed collections of data.

BROADER IMPACTS

The infrastructure of data repositories and visualization systems developed should not be specific to the biomedical research domain. Also, biomedical research is one of the biggest research areas, in terms of the number of research/researchers, their publications, and raw data. Therefore, we should be able to apply the same techniques to other domains.

What Challenges and Opportunities do you foresee?

- Data cleaning/transformation: Given that we have a large amount of legacy data that may not be in the right format, it is essential to provide ways to clean the data. The data cleaning itself still remains a challenge even though there have been some research efforts on this topic. A good systematic solution would benefit other research areas.
- Visualization scalability: As is often the case with most visualization systems, it is a great challenge to scale up visualization for a huge dataset, such as biomedical data, especially when the data is distributed.
- Visualization generalizability: There is often a tradeoff between generalizability and power of visualization systems. It will be also a challenge to develop visualization systems that can be generalizable, but still powerful. The generalizability of visualization is especially important when the techniques developed for the biomedical research are applied to other domains.

Neo Martinez

Using the Semantic Web to Integrate Ecological Data and Software and Increase Understanding and Awareness of the Biodiversity Impacts of Global Change

Director
Pacific Ecoinformatics and Computational Ecology Lab
1604 McGee Avenue
Berkeley, CA 94703

neo@PEaCElab.net
www.foodwebs.org

W1

POTENTIAL COLLABORATORS

Google, Microsoft, Long-Term Ecological Research sites, San Diego Super Computing Center, etc.

SUMMARY

Ecologists need to know the structure and function of the ecological systems they study. Comprehensive study of such systems is greatly limited by the heterogeneity (messiness) of the data, analyses, simulations, and locations of information describing these systems. Solutions to these problems are also messy to the point of most ecologists knowing little of the databases and software available to conduct ecological analyses. We propose to develop an evolving catalogue of ecological databases and software, describe these information tools using semantic web ontologies, and create "wrappers" for this software to turn these tools into Web services that can be integrated to increase the functionality of ecological analysis. This could, for example, integrate EstimateS software with Google Earth to monitor changes in biodiversity as the planet warms over the next few years and decades, based on continuing biodiversity inventories created by a wide range of academic, governmental, and nongovernmental researchers and oganizations.

INTELLECTUAL MERIT

Ecological data and software are greatly limited by the technological and sociological silos in which they are embedded. This project will break down those silos in order to facilitate the use and integration of these tools for greater scientific understanding of the ecological consequences for biodiversity of global change, including global warming and habitat modification.

BROADER IMPACTS

This research will demonstrate how scientific understanding can be greatly advanced by integrating information tools using semantic Web technologies. This understanding will also be made available to the public in real-time to increase public awareness of the ecological consequences for biodiversity of global change, including global warming and habitat modification.

TRANSFORMATIVE POWER

This project will usher in a new ability to conduct ecological research on global changes that are currently restricted to local and disparate analyses that are exceedingly difficult to generalize and inform larger-scaled impacts and concerns.

Erik Schultes

The Sequenomics Initiative: Mapping Nucleic Acid and Protein Sequence Spaces

Hedgehog Research
1500 Duke University Road
Box J2B
Durham, NC 27701, USA

eschultes@hedgehogresearch.com
http://www.hedgehogresearch.info

W2

POTENTIAL COLLABORATORS

Thomas H. LaBean, Duke, Chemistry & Computer Science
Peter T. Hraber, LANL T13, and possible corporate/private organization partnerships.

SUMMARY

There are two principle classes of information-carrying biopolymers found in nature: nucleic acids (DNA and RNA) and proteins. Understanding how information encoded at the sequence level determines the physical properties of biopolymer structure and function is the core question driving nearly every biological, biomedical, and biotechnological enterprise and promises a new predictive (rather than descriptive) approach to understanding molecular evolution in nature and in vitro.

Despite the differences in their backbone and monomer chemistries, both classes of polymers have identical formal descriptions: each are linear sequences composed of arbitrary orderings of their constituent monomers. This formal description of nucleic acids and proteins as sequences implies the existence of a discrete, high-dimensional space of sequence possibilities. It is evident that the number of sequence possibilities vastly exceeds the number of sequences in existing databases, and indeed, in life on earth. Not only does this mean that vast regions of sequence space are left unexplored, but it is clear that biological sequences compose an exceeding limited and profoundly biased "sampling" of sequence space. Without a more systematic and controlled sampling of sequence possibilities it is logically impossible to formulate general principles of biopolymer folding, structure, and evolution.

The overarching goal of this proposal is to perform the first systematic and comprehensive sampling of nucleic acid and protein sequence spaces, producing "maps" locating existing biological data sets in the larger context of sequence possibilities. Referred to as the Sequenomics Initiative, the proposed research brings together diverse expertise from computational, biological and biotechnological disciplines to capitalize on the unique opportunity afforded by readily available sequence and structure databases, fast and powerful algorithms for nucleic acid and protein folding, and laboratory techniques that permit the synthesis of arbitrary sequences and pools of unevolved sequences. Organizing these diverse technologies into four distinct, yet interlocking working groups (theory, visualization, database, experiment), the Sequenomics Initiative seeks to create a unified framework for biopolymer research (encompassing both nucleic acids and proteins).

The Sequenomics Initiative is a large-scale, yet highly-tractable project. Although foundational maps that will guide all further investigations will be available within 12 to 24 months, like geographic cartography, the Sequenomics Initiative will be an ongoing project with new and refined maps being produced as time, resources, and technology permit. As such, the Sequenomics Initiative will take advantage of Wiki-based, open source/open access tools to disseminate massive datasets, analytical software, bibliographic and Wikipedia information related to all facets of mapping sequence space.

INTELLECTUAL MERIT

Intellectual merit stems from the promise to reveal previously hidden, universal rules governing biopolymer folding, stimulate new approaches to the development of structure prediction algorithms, suggest novel experiments using individual sequences, pools and microarrays. While the vast majority of research effort is devoted to particular instances of biopolymer sequences, it is the intelligent and efficient sampling from the ensemble of unevolved sequences beyond biological examples, that most distinguishes Sequenomics Initiative.

BROADER IMPACTS

A deep understanding of the intrinsic symmetries and correlations of sequence space promise to have a disruptive

impact on structure prediction, biopolymer design, and combinatorial searches–leading to the next generation of advances in biotechnology and biomedicine that promises significant societal and economic benefits. By virtue of its unique approach to evolutionary biology (placing life as we know it in the context of life as it could be), the Sequenomics Initiative will inject an entirely new knowledge domain into long-standing, yet still contemporary, public policy debates regarding evolution and public education.

TRANSFORMATIVE POWER

The potential impact of the proposed research is transformative for entire research fields such as computational biology, evolutionary biology, genomics and structural proteomics.

What Challenges and Opportunities do you foresee?

There are three key challenges to the Sequenomics Initiative which are deeply rooted in computer science: (1) the formulation of a rigorous analytical description of biopolymer sequence spaces, drawing on well-established and as yet uncharacterized mathematical properties of high-dimensional regular graphs; (2) the development of informative representations of sequence spaces, especially visualization tools; (3) the development of sampling methods that are efficient and informative, despite being necessarily sparse (e.g., perfect sampling algorithms). Developing resources and solutions for these challenges will have wide application in the information sciences.

The Sequenomics Initiative's integrated Working Groups (Theory, Database, Visualization, and Experiments) will stimulate significant interdisciplinary research and education outcomes that could not be otherwise achieved. The Sequenomics Initiative will catalyze far-reaching research explorations well beyond its initial funding through the development of tools and methods that enable researchers to explore sequence space for themselves.

Erik Schultes received a Research Award from the Foundational Questions Institute Boston: Experimentally Probing the Origins of Macromolecular Structure in 2008.

Visit: http://www.fqxi.org.

Stephen M. Uzzo

Towards a Terrestrial Environmental Data Sensing, Visualization and Processing System

VP, Technology
New York Hall of Science
47-01 111th Street
Queens, NY 11368, USA

suzzo@nysci.org
http://nyscience.org

W1

POTENTIAL COLLABORATORS

Would require partners who had expertise in the following areas:

- Database and front-end development
- Knowledge of the manipulation and representation of complex data sets
- Interest in climatalogical, ecological and environmental data
- Knowledge and experience in cognition, epistemology and learning

Potential Partners:

- New York Hall of Science (or other Science or Natural History Museum)
- NASA
- National Ecological Observatory Network (NEON)
- National Science Foundation (NSF)
- SRI
- Institute of Learning Innovation (ILI), TERC, Concord Consortium (or other education research entity)
- MIT Media Lab
- University partner (such as Indiana University) interested in making such a project an aspect of a CDI research proposal

SUMMARY

Overview: Needs

One of the most daunting problems facing environmental researchers is the difficulty in making sense of large-scale, high-resolution, multivariate data on environmental change over prolonged periods. Remote sensing is increasingly used to monitor a variety of environmental parameters, but the resolution for many data sets is less than 30 meters, which cannot adequately detect small-scale changes in populations, distribution, long-term population shifts and effects of invasive species. Nor can it adequately resolve local changes in parameters affecting niche populations of some rare or highly localized species. These small-scale effects can be important indicators for predicting more widespread changes and influences to the environment (even regional effects). This is especially problematic in areas such as the San Francisco Bay Peninsula due to the complexity and dynamics of microclimates.

Aircraft imaging can provide higher resolution data than that of satellite borne instrumentation, but is expensive to deploy regularly enough to provide adequate temporal resolution. In addition, reduction in data resolution, caused by boundary layer turbulence is problematic at low altitudes; thus negating some of the benefit of reduced backscatter and luminance veiling through low-altitude imaging.

Reliance on the use of extensive ground truthing to refine data from airborne and spaceborne methods poses logistical problems is expensive to implement, and can be highly invasive–impacting ecosystems under study as well as the quality of data collected. Population and behavioral studies are among the obvious cases for reducing impact. These effects can be most pronounced in areas that would paradoxically benefit most from increased study.

The NSF-NEON project, attempts to create an infrastructure of local/regional data collection systems which can be linked nationally allowing many parameters to be monitored in addition to airborne imaging and LIDAR to provide comprehensive data and imaging characterizations about local ecology. NEON includes support for informal and formal education programs to make use of the data collection systems and the scientific interpretation of these data. What is missing is a deep investment in visualization and environmental understanding, which resides at the intersection of ecology, epistemology, cognition, database design and software user interface design. Novel ways of combining a variety of data sets and leveraging cyberinfrastructure approaches to create new tools for data discovery and understanding will be needed to provide ways for both researchers and citizenry to understand the massive

amounts of data which NEON and related initiatives will gather (along with the continuing input from remote sensing programs).

A similar case in which the promise did not lead to successful education programs was NASA's Mission to Planet Earth. Huge amounts of data have been gathered from the wide variety pf instrumentation deployed, but bringing these data into formal and informal learning environments did not provide extensive changes in the way science is taught, nor did it result in any significant usable approaches or interfaces in formal and informal science learning environments. The methods for using these data required extensive translation and the tools were very cumbersome.

Solution

A CDI approach to preventing such a missed opportunity with NEON is to develop extensible data protocols and tools that integrate both live and stored parameters from in situ instrumentation, along with up-to-date airborne and remote sensing data sets. They would be specifically designed to couple with flexible and modular tools for the manipulation and visualization of these data sets. The visualization tools would need to be designed to be easily modified and prototyped (use of open software standards and off-the-shelf, well supported proprietary software). A test-bed for developing a variety of approaches for prototyping user interfaces and various combinations of data mash-ups could then be tested rigorously with users to determine the impact on environmental understanding.

An optimized design for online tools and visualization and immersive approaches could then be deployed to provide an effective portal and suite of learning tools for use in formal and informal learning environments.

INTELLECTUAL MERIT

It would further the field of science understanding and the efficacy of real-time data integration and visualization for use in environmental/climate research and education and the public understanding of ecology. The data integration and visualization aspect of the project would be a novel area of research, both in the areas of climate modeling, as well as teaching and learning.

BROADER IMPACTS

Through continued development, this project would result in conventions and frameworks that could be deployed in many different kinds of science in which many data points and parameters are involved in data interpretation and result in a large-scale monitoring system from which to gather trends in climate and ecological change. By carefully developing interfaces that can be used at many different intellectual levels (including K-12 schools, as well as university and professional researchers), data will be available to all and be usable in a variety of formal and informal educational settings.

TRANSFORMATIVE POWER

Changing the paradigm by which we view the environment could help us to understand and impact the human perception of the environment in a much more profound way.

Stephen M. Uzzo

Tracing the Spatial and Temporal Migration of Genes in Metagenomic Data

VP, Technology
New York Hall of Science
47-01 111th Street
Queens, NY 11368, USA

suzzo@nysci.org
http://nyscience.org

W2

POTENTIAL COLLABORATORS

People sequencing metagenomic data such as Institute for Genomics Research;

People analyzing genomics (such as Mark Gerstein's group) and lateral gene transfer (such as James Lake, Maria Rivera and Peter Gogarten).

SUMMARY

There is increasing interest in the understanding of metagenomic data. The realization that lateral gene transfer is, to some degree, the rule in the environment rather than the exception opens up a whole field of potential discovery in understanding the geospatial nature of lateral gene transfer and the nature of metagenomics in general. This project would expand on ideas, like the Genographic Project, to create a data analysis infrastructure to help us understand the nature and patterns of gene migration, which may prove vital to better understanding all living things. This would be a large-scale project in several stages:

1. Fund the development of more efficient ways that sequences can be compared amongst organisms for the purpose of detecting transferred genes and trace them geographically in the environment.

2. Develop a framework for geospatially tracing how genetic material migrates through populations and what happens to it over time.

3. Develop visualization tools to trace patterns of migration that are scalable into multiple domains of inquiry, including ecology, genomics and teaching at all levels, as well as informal science learning.

INTELLECTUAL MERIT

This project will help frame the understanding of lateral gene transfer in spatial and temporal terms and provide a better understanding of evolution in microecologies.

BROADER IMPACTS

Could increase understanding of changes in macroecologies and may be able to help detect small-scale effects and their impact on greater ecological processes, including possibly predicting metabolic effects of climate change.

TRANSFORMATIVE POWER

We may be able to develop metabolic "weather reports" to understand not only biological and evolutionary processes, but better model atmospheric and geologic interactions and effects.

What Challenges and Opportunities do you foresee?

The ability to derive predictors from metagenomic data that quantify and validate gene transfer as well as identify the vectors of genes in the ecology of organisms.

Luís M. A. Bettencourt
The Global Science Observatory

Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545, U.S.A.

http://math.lanl.gov/~lmbett

W1 W2

POTENTIAL COLLABORATORS

Peter van der Besselaar, Johan Bollen, Katy Börner (Indiana University, similar idea), Kevin Boyak, Kyle Brown (Innolyst), Noshir Contractor (Northwestern University), Mark Gerstein, Stefan Hornbostel, Masatsura Igami, David Lazer (Harvard University), Barend Mons (Erasmus University, Netherlands & KnewCo, Inc. Wikiproteins), Ben Shneiderman (University of Maryland), Martin Storksdieck (Project Development Institute for Learning Innovation, similar idea), and Stephen M. Uzzo.

SUMMARY

The Global Science Observatory (GSO) will integrate scientific and technological information to assess the current state of scientific fields and provide an outlook for their evolution under several (actionable) scenarios. It will integrate standard (real-time streaming), bibliometric data from publication databases, with patent records from patenting offices world-wide, and with information such as biographic data, research investment from public agencies and private entities, linkages with the private sector, usage records for access to publications online, media coverage of research, and its documented uses in education. The system will integrate categorical information (see above) with organizational and geographic systems, allowing the construction of maps of science and technology according to different coordinate reference systems, including but not exclusively, world (geographic), disciplinary and organizational (universities, firms) maps.

Data across all these and other dimensions will be integrated into the system by/from federal funding agencies (DOE, NSF, NIH, CORDIS, etc), individual scientists, publishers and other stake-holders (as in a Wiki-environment) and can be made transparent and vetted by all parties. An API environment will be implemented that enables users of the system to create mash-up applications by combining data, analysis tools and visualizations across data dimensions. Privacy and proprietary issues may require special attention and balance against the open spirit of the observatory and will follow best practices of analogous (virtual) communities.

An important component of the system will be an associated 'laboratory', dedicated to the generation of forecasts of future scientific and technological activity, payoffs and of potentially transformative events. To do this, data will be analyzed retrospectively for validation of models of the evolution of fields, and special effort will be placed on the identification of signatures of optimal conditions for acceleration, including studies of field lifecycles and susceptibility to 'paradigm changes'. An emphasis will be placed on the development of predictive (probabilistic) models that generate an outlook for science and technology with quantified uncertainty. New observations that fall outside confidence bounds of model predictions will be used to falsify models, to flag novel developments and/or possibly reset parameter values. Models will focus on observable quantities with relevance for users of the system, including public funding agencies, private firms and venture capital investors. Comparative analyses at different levels of aggregation and units of analysis will be enabled in the front end so that, for example, performance levels (as measured by publication output, citations, persons trained, or value added to production) can be assembled for institutions or nations, and correlated with investments (incoming funding, labor hours, venture capital invested, infrastructure improvements). These analyses will use extrapolations from previous data and analogous situations. They will be confronted with results in order to generate an environment for successive improvement of models and theory.

INTELLECTUAL MERIT

Innovation in science and technology is the engine of economic growth in modern societies. Increasing globalization, in terms of intellectual capital and economic activity, is opening up unprecedented opportunities to accelerate scientific and technological breakthroughs that can address global sustainable development, while

at the same time decentralizing and lowering the bar for productive participation in the mechanisms that underlie innovation and change. The social dynamics and regimes of incentives that make these efforts successful are still poorly understood theoretically, and as a consequence, actionable insights that can predictably accelerate scientific and technological development are largely lacking. Recent progress in information technology has the potential for large-scale transformation on how scientific and technological investments are allocated and in making their results broadly available to users and the public—increasing transparency and rationality in the management of science and inspiring education and life long learning. The system proposed here, the Global Science Observatory, will provide an integrated solution to how science is promoted and the evaluation of its results. While providing a practical tool for investment and risk assessment, the system would constitute a laboratory for the development of models and theory for endogenous economic growth and innovation processes, making it possible to test fundamental ideas from economics and the social sciences against a large body of past data and future observations accessible to all.

BROADER IMPACTS

The Global Science Observatory will improve the transparency and rationality of scientific and technical investment as a public good, while promoting the identification and diffusion of innovation to the private sector and public at large. Many new forms of computational thinking are being developed that address the challenge of data collection, integration and analysis that will underlie the GSO. Research programs in maps of science, population models for the evolution of fields and data assimilation, tagging and knowledge discovery systems - such as: Mesur, http://www.mesur.org/MESUR.html); Wiki Professionals, http://www.wikiprofessional.org/conceptweb; and ResearchCrossroads, http://www.researchcrossroads.com - are already providing proof of principle in the way data can be integrated via mash-ups and collaborative participation (in a Wiki model) to generate a system such as the GSO. Integration of efforts and wider participation are, however, still necessary to validate and extend these approaches and to render the information truly useful in terms of comparative and strategic analyses. When enabled, comparative analysis of scientific status and development of new ideas should encourage new and productive collaborations among individuals, institutions and nations. It should also provide an accessible entry point for the non-technical public to learn about developing scientific and technological activity. As such, the GSO may constitute as an important instrument for science education, one that conveys the impression of science and technology as living social movements.

TRANSFORMATIVE POWER

The Global Science Observatory will provide the first "perspective from above" of science and technology as dynamical, complex systems that are affected by societal events and flows of resources (money, people) from one activity to another. It will create a scientific discovery accelerator by creating an unprecedented test for a quantitative and predictive understanding of innovation processes and endogenous mechanisms of (economic) growth; simultaneously providing a real-time testing environment (laboratory) for models of the social and cognitive sciences and of economics. It will also constitute a constantly improving collaborative environment and a virtual community – through streaming data, data revisions, and progress in predictive theory – dedicated to the predictive analysis of return on investment and risk, of value to public science managers, private enterprise and venture capital investments, and for the information of the public at large.

Katy Börner
Towards a Scholarly Marketplace that Scholars Truly Deserve

Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
Wells Library 021, 1320 E. 10th St
Bloomington, IN 47405, USA

http://ella.slis.indiana.edu/~katy
katy@indiana.edu

W1 W2

POTENTIAL COLLABORATORS

Leads of existing "product" databases, e.g., Medline, JSTOR, Citeseer, arxiv, myExperiment, BioCrossroads, NIH, KnewCo, CiteULike, and many others.

Researchers that study "Social design", i.e., the design and inner workings of incentive structures in a proactive, invasive way. David Lazer and Noshir Contractor would be highly relevant.

Research teams/communities, but also funding agencies/ publishers interested in being beta testers.

SUMMARY

This is an (cyber)infrastructure development proposal with a major research component on social design, i.e., the real-time design, implementation, evaluation and optimization of incentive structures.

Software Engineering — CI Design Component

We propose to develop a "Flickr/YouTube-like" scholarly marketplace (SciMart). A SciMart is a Web service that can be used by a single research lab, center, university or an entire scientific community or nation to share expertise, publications, patents, datasets, algorithms, protocols, workflows, and other "products" of scientific research. This marketplace will also host information on "resource intake" such as funding, people, and products generated by others (e.g., cited papers, downloaded data, software).

To ease the setup and adoption of SciMart, it can be pre-filled (and continuously updated) with existing data, (e.g., Medline, USPTO, BioCrossroads funding data, etc.). Major existing and new standards for scholarly information exchange (such as DOI, OpenURL, OAI, OCLC's WikiD, FOAF, and other "open science" standards) will be supported to ease the plug-and-play of new data sources. Wiki-like functionality will support the adding, modification, linking, and deletion of prefilled/new data (ee KnewCo system).

We plan to collaborate and/or adopt KnewCo's strategy of sharing modified/cleaned data with the source data providers.

Just like Wikipedia, SciMart is an open data, open source, and open standards Web service. The origin of all data, software, and standards used are easily identifiable so that credit can be given appropriately.

SciMart provides many "basic" services. Besides easy upload and download, SciMart supports that:

- Science "products" can be interlinked via semantic associations (e.g., this tool was used to analyze these two datasets and the results are published in paper X).
- Any "product subset" can be tagged, and folksonomies can be created.
- Any "product subset" can be rated and favored
- "Resources" (e.g., funding) can be interlinked with "products".

"Value added services" will be provided as 'mash-ups' that might show among others:

- Recommendations based on topic similarity, link similarity or their combination.
- Interest groups, e.g., to keep track of "products" relevant for a specific project. Note that many "products" will be relevant for several projects.
- The increase in the number of different "products" over time.
- Geospatial origin of contributions and users but also the geospatial coverage of "products".
- Topical coverage or usage of "products".
- All contributions and the usage patterns for any user.
- (Topic) Networks of user/product types for specific time spans.
- Bursts, dynamics.

Upload, download, rating, and annotation of "products" can be done anonymously or by using any user name. Users can create profiles which might be their true research credentials or anything they would like to reveal.

The proposed project goes beyond what arXiv, Citeseer, myExperiment, SciVee, and Sourceforge or science news sites serve today, as it supports the sharing, interlinkage, tagging, and rating of any kind of "product". In fact, it creates a kind of "meta-index" of existing "product" and "resource" databases.

Multiple SciMarts can be easily merged if desired and meta-analyses can be performed.

Research Component

SciMart will only succeed if the (evolving) needs and desires of its different user communities can be addressed.

- This requires a **detailed user and task analysis** of who needs what, in what context, with what priority, etc.
- A formal study of what incentive structures are at work in Flickr, YouTube and Wikipedia. Which ones cause what effects? How do they work in concert? What can we learn from them for the design of successful scholarly marketplaces?
- Real-time implementation and evaluation of different incentive structures in the life SciMart System – very similar to how Google optimizes its search and other services.

INTELLECTUAL MERIT

SciMart applies Web 2.0 technologies to ease the sharing of scientific results. Easy access to not only publications but also datasets and applications will:

- Enable "resource" and "product" discovery across scientific and language barriers.
- Increase repeatability of scientific results.
- Help avoid duplication of research, reimplementation of the same algorithms, repeated cleaning of the same datasets, etc.
- Improve transfer/combination of science products from one area of science to others as well as to industry (products).
- Support new types of searches (e.g., find all papers that used dataset X).
- Maps provide a global view of existing "products" in one or more SciMarts and their interlinkages, ratings, etc. in support of product navigation and management.

BROADER IMPACTS

Speeding up the diffusion of 'products' across disciplinary boundaries is expected to improve the innovative capacity of a nation. SciMarts can also be used for educational purposes, e.g., by linking educational resources to respective scientific products, easing the transfer of scientific results to the classroom. Industry might like to add information on 'grand challenges', prices (e.g., NetFlix), or jobs to the map to cause an 'industry pull' on researchers. Ultimately, SciMart will serve as a "visual index/interface" to existing science product holdings, helping to direct relevant users to the resources they truly want.

TRANSFORMATIVE POWER

SciMart will support the easy sharing/collection, interlinkage, annotation, and rating of scholarly products. The resulting meta-index will interlink products that have much higher coverage/quality than any database available today. Many of the proposed value added services are not in existence today or are only as proprietary tools/services. Maps of science that are created based on this meta-index will help communicate science needs, resources and results among researchers, industry, educators, and society.

CHALLENGES

Privacy

If all my/anybodies complete scholarly activity is recorded, scientists will be "naked". While corporations find this desirable in today's world (see *WIRED* magazine, issue 15.4 from March 2007), would private persons like this? Note that today's teens, patients-like-me, or who-is-sick do not seem to mind sharing very intimate data. Is this our future?

Social Design

What incentive structures are at work in Flickr, YouTube and Wikipedia that make these sites "tick"? What can we learn from them for scholarly marketplaces?

NIH RECENTLY FUNDED TWO U24 AWARDS:

The National Research Network: VIVO: Enabling National Networking of Scientists. NIH 1U24RR029822-01 award (Michael Conlon, UF, $12,300,000) Sept. 09 - Aug. 11. The proposed work will establish national networking of scientists by providing a new software system (VIVO) and support for scientists using VIVO. Scientists using VIVO will be able to find other scientists and their work. Conversely, scientists using VIVO will be found by other scientists doing similar or complimentary work.

Networking Research Resources Across America, NIH 1U24RR029825-01 award, Nadler, L. M., Harvard U, $14,000, Sept. 09 - Aug. 11. Nine institutions of different sizes, geographical location, and culture have come together to build and implement a Federated National Informatics Network that will allow any investigator across America to discover research resources that are presently invisible.

Kevin W. Boyack

Comparing Open and Paid Literature Sources from Coverage and Metrics Viewpoints

SciTech Strategies, Inc.

kboyack@mapofscience.com
http://mapofscience.com

W1

POTENTIAL COLLABORATORS

Katy Börner, Johan Bollen and OpenSource Database Curators.

SUMMARY

Current science metrics are based primarily on citation data from Thomson Scientific. Open source databases, such as Medline, are becoming increasingly available - and although most do not contain citation data - a combination of data from open sources might provide an alternative to Thomson's data for activity-based metrics. We propose work to test this hypothesis. Appropriate open source literature data sources will be identified and mapped against Thomson's data to show potential coverage. More detailed analyses will determine if Thomson-based activity metrics (publication counts, collaboration, etc. by country, region, etc.) can be mimicked using open source literature data.

INTELLECTUAL MERIT

No comparative mapping of open literature sources (such as Medline, Citeseer, arXiv, etc.) vs. paid literature data sources (such as Thomson Scientific or Scopus) currently exists. A mapping of these literature databases will show if open sources can be used in place of paid sources for certain purposes. In addition, if coverage of open sources is found insufficient for the purposes of metrics, this study will identify those areas in which open sources are deficient, and thus provide a strategy for the further development of open source literature databases.

BROADER IMPACTS

Dissemination of the results of this study have the potential to publicize and promote additional use of open source literature databases. Although this might not increase the actual availability of such data, it would certainly increase the usage of open source literature – and thus has the potential to positively impact the entire science system through increased usage of currently available open source literature.

If a combination of open sources is found to be a suitable alternative to paid literature sources for the purposes of metrics, it would make the generation and use of metrics more readily available to a large population of researchers worldwide.

TRANSFORMATIVE POWER

If a combination of open sources is found to be a suitable alternative to paid literature sources for the purposes of metrics, it has the potential to change the cost structure for the generation and use of science metrics. It also would enable the development of tools that could lead to real-time generation and an update of science metrics and indicators.

Kevin W. Boyack

Conceptualization of the Overall Data Space Pertinent to Science Policy

SciTech Strategies, Inc.

kboyack@mapofscience.com
http://mapofscience.com

W2

POTENTIAL COLLABORATORS

Kyle Brown, Johan Bollen, Online Computer Library Center (OCLC), Library of Congress, Wikipedia, and OpenSource Database Curators.

SUMMARY

A variety of data sources are currently used for science policy. Others are not currently used, but could be very useful. The sum total of the concept space comprised by these sources is currently not known. If this space were characterized, even qualitatively, it could lead to the proposal and testing of new metrics more attuned to today's science. Characterization of this space is also fundamental to the proposal and production of any "dream tool" for science policy.

INTELLECTUAL MERIT

No comprehensive mapping of policy input sources (funding [grant, foundation, corporate, VC, etc.], contract, publication, patent, corporate data) exists on either a quantitative or qualitative level. Nor are the overlaps or potential linkages between these sources known. A qualitative (and quantitative, where possible) characterization of the sum of these sources would provide a basis for a new generation of metrics and science policy. It would also enable the design of tools to more fully exploit the available policy and metric space.

BROADER IMPACTS

A comprehensive mapping of sources would benefit people other than policy makers. It has the potential to inform institutions and individual researchers of opportunities, potential collaborations, and competition at a scale that is not possible today. It also has the potential to uncover bridges between fields and topics more quickly than what happens in the current state of incremental research that is prevalent today, and thus speed the growth of interdisciplinary science and engineering.

TRANSFORMATIVE POWER

See broader impacts.

What Challenges and Opportunities do you foresee?

Three potentially large challenges are readily apparent. First, one must identify all (or a large subset of "all") policy-relevant-literature-like data sources. Second, one must procure samples (whether quantitative or qualitative) of each sufficient enough to broadly describe the overall concept space and overlaps between sources. Third, once data is procured, one must determine what facets of the different data sources can be used to seam them together. This may not be straight-forward, and is likely to require seaming different portions of the space using different facets. The accuracy of this work may be difficult to address. Despite these potential challenges, even a preliminary result would be useful as a stepping stone to more detailed and comprehensive work in this area.

Noshir Contractor
Team Science Exploratorium

Jane S. & William J. White Professor of Behavioral Sciences
Department of Industrical Engineering & Management Sciences, Communication Studies, and Managment & Organizations
Northwestern University
2145 Sheridan Road, TECH D241
Evanston, IL 60208-3119

nosh@northwestern.edu
http://nosh.northwestern.edu

W2

POTENTIAL COLLABORATORS

Scholars/organizations with "instruments" that capture large-scale streaming data sources scientometric and bibliometric sources, location-based and telecommunication-based devices;

Methodologists with visual-analysis techniques for cleaning, collating, aggregating, updating, and modeling, large-scale datasets/streams;

Computer scientists who can implement efficient peta-scale-ready algorithms for manual and automated data annotation, data integration, validation and analysis;

Theorists who can develop explanations for phenomena using concepts/variables that might not have been proposed previously because collecting the data required would have been impossible (without "instruments" employed by 1, see above);

Supercomputing/Grid facilities with petascale computing and storage support infrastructures.

SUMMARY

There is growing recognition that the societal grand challenges confronting us will need to be addressed by teams of scientists working over time and space and who are embedded within a multidimensional knowledge network that includes digital resources–such as documents, datasets, analytic tools and tags. Many of these teams will be multidisciplinary, interdisciplinary and transdisciplinary. Given the large amount of investments being made in the support of these research initiatives, it is critical that we make every effort to maximize the likelihood of developing socio-technical systems that offer effective strategies for scientists to leverage the multidimensional knowledge networks while in the process of creating and maintaining these teams.

This effort seeks to develop the infrastructure that helps us advance our understanding of the socio-technical drivers of effective team science.

There is a large body of findings and well-developed methodologies for studying interactions and behavior of teams in social psychology, communication, organization studies, as well as in economics, sociology, and organization studies. A NSF-funded review identified nine distinct theoretical perspectives that have been applied in more than one discipline: functional theory; psychodynamic theory; social identity theory; socioevolutionary theory; the temporal perspective; the conflict-power-status perspective; the symbolic-interpretive perspective; the network perspective; and the feminist perspective (Poole & Hollingshead, 2005). Thus, there is a surfeit of theoretical ideas that can be applied to groups.

However, without exception, the explanatory variables in our theories are informed by the type of data we can collect. Within the past decade, a handful of theorists have begun to propose more ambitious and complex theoretical conceptualizations (Arrow, McGrath & Berdahl, 1998; Katz, Lazer et al; Monge & Contractor, 2003). There is clearly an intellectual apparatus emerging to develop a complex systems theory of team science that has the potential to provide a deeper understanding of the dynamics and outcomes of teams engaged in complex situations, such as transdisciplinary research. However, a major challenge confronting these theoretical developments is the inability to collect the high resolution, high quality and high volume, data necessary to test and extend such theories.

There are several important primary data sources that contain digital traces of Team Science activities within the scientific community; these include bibliographic sources (e.g., Web of Science, Pub Med), patent databases, funding databases such as CRISP and at NSF, project web sites, etc. There are several web crawling and text-mining tools that can be used

to generate secondary network data that provides relations among the entities in these multidimensional networks. In addition, participation in recent e-science/cyberinfrastructure hubs (such as MyExperiment and nanoHUb), provide logs of usage activities by people who are uploading, retrieving or tagging a wide variety of digital resources such as documents, data sets, analytic tools, instructional materials, etc.

Unfortunately, the data sources and analytic tools are currently "silo-ed." Hence, any effort to capture, model, visualize and analyze the drivers of effective team science are wrought with duplicated, redundant, and non-reusable efforts.

What is needed is the development of a shared cyberinfrastructure that will help the various stakeholders interested in advancing Team Science. We call this the Team Science Exploratorium.

INTELLECTUAL MERIT

- Advance the socio-technical understandings of the Science of Team Science
- Advance scalable data mining techniques required to infer various relations in the multidimensional networks
- Advance scalable algorithsm required to model large scale, multidimensional networks

BROADER IMPACTS

- Develop socio-technical tools (technologies designed by taking into account social incentives) to assist the scholarly community in exploring and utilizing the potential for Team Science
- Help program officers and reviewers to identity and evaluate – prospectively and retrospectively – the likelihood of a team engaging in effective Team Science
- Help science policy makers and portfolio managers identify the strengths and weaknesses of their investment strategies

TRANSFORMATIVE POWER

This project has the potential to have quantum improvements in addressing large societal challenges that are being pursued by current investments in Team Sciences.

What Challenges and Opportunities do you foresee?

Challenges in assembling a transdisciplinary team that will each see intellectual merits in various facets of this effort.

Additionally, an opportunity to develop a new Transdisciplinary Science of Team Science.

Noshir Contractor is one of the organizers of the *Science of Team Science Conference* at Northwestern University in Chicago, IL on April 21-23, 2010.

Web page: http://scienceofteamscience.northwestern.edu

Ingo Günther
The Science Space Machine

Artist
WorldProcessor Globes
88 West Broadway
New York, NY 10007, USA

i-gun@refugee.net

W2

POTENTIAL COLLABORATORS

Architects: Kevin Boyack, Barend Mons, Katy Börner, Second Life and Game designers, etc.

SUMMARY

Building a 3-D model of the world of science. The governing principal is the innovative/intellectual space that scientists occupy, defend and expand - then navigate and use. But this should also work for non-scientists.

Principles

- It would put a face on the scientists, or rather the scientist's face on their science
- It would be a bit like "home steading" as the scientists can occupy the space and design it themselves
- It would transcend text as the core carrier of memorable (and operational) knowledge

Obviously a good structure, building codes and construction laws should be published–and the standards would be necessary to adhere to. The space would be navigable, both via Web based interface and physical (as a in 3-D structure), and would possibly consists of (a yet to be invented) dynamically expanding and contracting solid portion (perhaps inflatable) and projections. The mental template would initially be based on the notion of memory before script - as put forth in Frances Yates' book, *The Art of Memory* - and then that notion would have to be expanded and possibly obliterated in the process. The space would mirror the field of science—a parallel world.

Scalability is essential; so that anybody can get an overview. But, then drilling down on the one hand and de-selecting relationships on the other is equally important. Ideally, these structures would be placed in universities, libraries, science museums, etc.

[INTELLECTUAL] MERIT AND BROADER IMPACTS/TRANSFORMATIVE POWER

Experience science in a physical and comprehensive way. Turning science into a playing field that appeals to almost everyone, either by content or by a visual experience.

What Challenges and Opportunities do you foresee?

- Perhaps this would require a new profession - the science architect; a person, perhaps somewhat between a science journalist and an information architect
- Actual building materials that conform to multidimensional, spacial requirements
- Motivation and incentive structures
- Ego may be nice, but not enough
- A possibility to purchase or lease futures based on fields/area/spaces might be necessary
- Venture capital could invest here
- Considering how much a rampant commercialization of science may also be compromising and perverting it
- Establishing a science currency (like milage or credits) to fund and facilitate operation
- Building the Web interface for this - could be Science Life (as in Second Life); much can be learned from this parallel world

Masatsura Igami

Create Maps of Science That Help Science Policy

National Institute of Science and Technology Policy
16th Floor, Central Government Building. No. 7 East Wing
3-2-2, Kasumigaseki, Chuyoda-ku, Tokyo 100-0013, Japan
igami@nistep.go.jp

W1

POTENTIAL COLLABORATORS

A community of scientists, funding agency (NSF, JST, etc), and patent offices (USPTO, JPO, EPO, etc).

SUMMARY

Make reliable maps and develop models for forecasting.

- Establish reliable way to track evolution of science quantitatively. Generate a series of science maps backwardly (probably from 1980s to 2007) and track the history of science. Create an open environment where scientists can exchange their views on the maps freely and modify the mapping algorithm based on their opinions.
- Develop models to forecast the incremental development of science, not breakthroughs (e.g. convergence of scientific fields, etc.).

Assess impacts of science policy on dynamism of science

- Overlay the information of funding on the maps developed is the first step and assess impacts of funding on dynamism of science, countries' competitiveness, and productivity of individual organization like universities.

Measure knowledge flows from science to technology

- Link science map and patents, based on citations or inventors/author links, and assess the influence of scientific discovery on inventions, ultimately innovations.

INTELLECTUAL MERIT

There is no study that shows the impact of public funding on the dynamics of science. This project will provide an assessment of the impact that funding has had over the past 20~30 years on the development of science and will also show the role of public funding in the national innovation system (e.g. counties' competitiveness, and productivity of individual organizations such as universities, and the changing of its role over time).

Measurement of knowledge flows from science to technology is another way to measure how science plays a role in enhancing counties' competitiveness. This will give some evidence that shows the importance of basic research in national innovation systems.

BROADER IMPACTS

How to utilize the maps of science for Science Policy is a big issue to be solved. The first step of this project is dedicated to assess validity of the maps and to create an arena where scientists can freely share their view on the dynamics of science. One difficulty in dialogues among scientist is that they share their opinions based mainly on their background. To share a common background, like maps, would be useful to facilitate creative talking.

David Lazer

Producing Knowledge on Sharing Knowledge/The Architecture of Knowledge Creation

Associate Professor of Public Policy
Director, Program on Networked Governance
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street, T362
Cambridge, MA 02138, USA

David_Lazer@harvard.edu

W2

POTENTIAL COLLABORATORS

This would require a team of collaborators, including the people building the relevant infrastructure and social scientists.

SUMMARY

Theory/Hypotheses

The essential question is whether an infrastructure built to enable communal production of knowledge within a scientific community actually has had an impact on how people do their research. This requires collecting some baseline data on who works with whom, how easily they can find what is happening in relevant fields, who is collaborating with whom, etc. This, in turn, would require analysis of some behavioral data—e.g., careful tracking of how these tools were used at the individual level, bibliometric data on who has collaborated with whom—as well as some survey data at multiple points in time. The survey data would capture some of the individual level factors expected to be relevant (e.g., personality traits, offline network, norms). Potential hypotheses include:

1. More cross-silo connections will be facilitated with such an infrastructure. For example, there will be more collaboration across universities and disciplines.
2. There will be important individual level factors that drive participation in the system, including: age, psych factors such as extroversion, views around "generalized reciprocity".
3. A key driver of participation across research communities will be the development of norms that reward contributors to the system. These norms will disseminate through the network, and thus there will be variation based on field, institution, etc.
4. Important design elements will include the development of tools to navigate and connect materials and management of identity. Issues around identity include: credit and responsibility for contributions, but also external social control (e.g., hierarchy may exert itself if identity is transparent).
5. Such infrastructures, if successful, will produce some intellectual convergence on questions and methods used to pursue particular research endeavors (increasing exploitation but reducing exploration) vs. such infrastructures that will facilitate serendipitous collisions that create innovative ideas/methods.
6. Such infrastructures will create access/visibility for some of the "have nots" in the system vs. such infrastructures that will reinforce the privileged position of the highest status actors.

Research Design

Data Required

- Multiple rounds of surveys around attitudes, networks, etc.
- Behavioral data on usage, ideally as detailed as possible (down to click level)
  Bibliometric data on papers, citation patterns, linkage of papers to data, etc.
- Information on individuals, such as institutions over time, degree, advisor, career trajectory, etc.

Mode of Analysis/Tools

The core of the analysis would be around longitudinal data on networks. How do the networks evolve? A variety of tools around dynamic network would be necessary, such as p* (which might not be robust enough for the relevant analyses). Robust, easy to use, and malleable tools will be needed to deal with reducing/analyzing/visualizing highly detailed data. There are considerable privacy issues, because many of the relevant data should be considered confidential (e.g., usage), but will be essential to addressing the above questions. How to balance openness and confidentiality issues?

INTELLECTUAL MERIT

Knowledge is collectively constructed. The emergence of online resources creates digital traces that allows the examination of the process of construction.

BROADER IMPACTS AND TRANSFORMATIVE POWER

The practical implications are considerable. This research should allow assessment of the impact of these infrastructures, as well as the paths through which the infrastructure has an impact. These results should inform future decisions to create such infrastructures, and as well as design choices.

What Challenges and Opportunities do you foresee?

1. Ideally data collection will be built into the infrastructure, plus baseline data should be collected before the inception of such a system. This requires involvement of the social scientists from the very beginning.
2. Given such high levels of involvement in the creation of the system, and collaboration with the designers of the infrastructure, it will be very difficult to keep a scholarly distance from the data. How should the research endeavor be designed so as to reduce the amount of bias that will slip into the research process?
3. See privacy issues under “Mode of Analysis/Tools”.

Israel Lederhendler
Incentives for Collaborative Intelligence

Director, Division of Information Systems
National Institutes of Heath
Building 1 - Shannon Building, 134
1 Center Dr
Bethesda, MD 20814, USA

lederhei@od.nih.gov

W2

POTENTIAL COLLABORATORS

Workshop participants and others.

SUMMARY

During the last several decades Science has simultaneously benefited and struggled with the ready availability and growth of massive amounts of data. The World Wide Web has become a medium that is particularly well suited to scientists who are among the principal generators of this unprecedented creation of data and information. Many areas of science are shifting towards greater attention to intersections among knowledge domains and there is also a growing imperative to see the bigger picture. Movement toward synthesis and integration is countered by the pressures of information overload, ambiguity, and redundancy. To deal with these pressures and intellectual limitations of the individual, many scientists retreat to increased specialization within isolated silos of knowledge. Instead of knowledge silos, it is more necessary than ever to enhance collaborative relationships and cooperation among scientists. This is not a requirement that can be imposed from the "top," although institutional leadership for this change is critical. Nor can we expect current leaders in science to voluntarily change their culture and freely share their ideas with unknown competitors.

The World Wide Web has moved beyond being simply a source of information and data. It has become a dynamic medium that actually creates new information because of its wide accessibility and underlying social networks (e.g. Flickr, Facebook). The web is a medium where many individuals can collaborate more easily than ever before. As a social medium, the web brings value and innovation that a simple collection of individuals could not achieve. The experience of Wikipedia illustrates the power of open social networks to build an unprecedented knowledge-base relevant to its users.

INTELLECTUAL MERIT

The development of visualization tools for complex data sets allows for substantially improved appreciation of metadata and relationships among ideas. With better understanding of complex data, opportunities for collaboration and sharing of intellectual property increase. In the best of possible worlds, this produces a very rich culture of scientific entrepreneurialism (generic use). The age of the individual scholar who can command many areas of science (e.g. Darwin) is no longer possible. Instead, scholars must increasingly work collaboratively to solve problems that reach across boundaries.

BROADER IMPACTS

In society today, there is an increased interdependence between researchers, public and private funding agents, commercial enterprise, and makers of public policy. It is virtually impossible for the traditional model of the single independent scientist or lab to establish mastery in this highly diversified system. But informatics and new computational approaches are creating exceptional opportunities – not only for the scientists doing research, but for society as a whole.

What Challenges and Opportunities do you foresee?

A more intense culture of shared knowledge has the potential to transform science and society. People will benefit from an enhanced discovery and invention process. At the same time, research is needed to understand how the fundamental drivers of scientists for visibility, recognition, and reward are met in a new collaborative intellectual environment where credit is more distributed. To move toward collaborative intelligence, research on how science is done will be needed that focuses on the drivers and incentives of scientists at different career stages. It's clear that a more open and collaborative environment is required to meet society's immediate research challenges.

Jim Onken

Science of Science Infrastructure Design and Development

Analyst, Office of Research Information Systems
National Institutes of Health
Office of Extramural Research
Building 1 - Shannon Building, 134
1 Center Dr.
Bethesda, MD 20814, USA

james.onken@nih.gov

W2

POTENTIAL COLLABORATORS

Social Scientists, Mathematicians, Demographers, Computer Scientists, Cognitive Psychologists, Representatives of funding agencies, Members of Scientific Communities (e.g., Physics, Epidemiology) with an interest in network analysis and studies of science

SUMMARY

The ultimate goal would be developing the intrastructure needed to perform studies of science: the databases, analysis methods and tools, and collaboration spaces.

1. A federated database of research programs (funding sources, information on the specific initiatives undertaken; funding levels and mechanisms of support); research projects funded (grant number, title, abstract, keywords, concepts and tags); the products of that research (publications, patents or other intellectual property); and the (characteristics of) people and resources involved in the research (investigators, students, de-identified patients, subjects and gene expression profiles, etc.)
2. Visualization and analysis tools to understand and explore this information. Tools would include:
    a. Visual mapping in a variety of types of spaces (a general tool that would allow different stakeholders/users to select different representations of the same of data) including:
        i. Geographical space
        ii. Research concept spaces
        iii. Organizational space (i.e., map what agencies fund what research, and what organizations conduct it)
        iv. People space (map who is doing what, in what area, and their interconnections)
    b. Exploration tools (search, drill-down, etc); a variety of different methods for search and retrieval from the database. The visual maps described in (a.) could also serve as one of several possible front-end search tools.
    c. Analysis methods and tools:
        i. Clustering algorithms for data reduction and mapping in hyperdimensional spaces (with projections onto two dimensions)
        ii. Network analysis tools
        iii. Tools for longitudinal analyses of networks
        iv. Tools to model changes in scientific output, progress and workforce dynamics
        v. Tools to map critical events onto time series
3. A clearinghouse for sharing Science of Science (SoS) datasets, tools, results and comments.

Initially, begin with proof-of-concept projects in well-defined areas that can be integrated and built upon as interest builds.

While the focus would be on developing the intellectual resources for conducting SoS studies, the tools developed should also target other audiences, such as the lay public and members of the research communities under study. For example, search tools should be constructed so that the public can find information on particular topics of interest, and network analyses should be done in such a way in order to assist researchers in content domains find collaborators.

BROADER IMPACTS

The integration of databases and tools to analyze them should provide a more coherent view of current scientific efforts, the allocation of resources to different areas of sciences, and future needs for research personnel. Making the products of this research more intelligible and readily available to the public might increase interest in and support for the sciences, increase the scientific literacy of the public, recruit more students, and allow the public to more fully benefit from investments in science.

TRANSFORMATIVE POWER

Could be what's needed to move beyond more "conventional" bibliometric studies, which I think have generally not been perceived to be that meaningful or useful to policymakers and scientific communities.

What Challenges and Opportunities do you foresee?

Skepticism from the research communities concerning the complexity of the endeavor and whether the information to be generated will be so "high-level" as to not be meaningful to them. Some well-chosen demonstrations initially might be critical to success.

SCIENCE AND TECHNOLOGY IN AMERICA'S REINVESTMENT (STAR)

Measuring the EffecT of Research on Innovation, Competitiveness and Science (METRICS)

STAR METRICS will create a reliable and consistent inter-agency mechanism to account for the number of scientists and support staff that are on research institution payrolls supported by federal funds. STAR METRICS will build on this information in future to allow for measurement of science impact on economic outcomes, such as job creation, on scientific outcomes, such as the generation and adoption of new science, often measured by citations and patents, as well as on social outcomes such as public health. Second, it builds on and expands existing research tools with minimal burden for government agencies and grantees.

Source: http://nrc59.nas.edu/star_info2.cfm

Ben Shneiderman
Telescopes for the Science Observatory

Professor, CS, ISR, UMIACS, Founding Director HCIL
A. V. Williams Building
Department of Computer Science
University of Maryland
College Park, MD 20742, USA
ben@cs.umd.edu

W2

POTENTIAL COLLABORATORS

Catherine Plaisant, Ben Bederson, Georges Grinstein, Martin Wattenberg, Noshir Contractor, Saul Greenberg, Chris North, Adam Perer, Mary Czerwinski and Marc Smith.

SUMMARY

Science Observatories will enable science policy analysts, sociologists of science, scientometricians, corporate research planners, and venture capitalists to explore the rich sources of data about scientific grants, research projects, published journals and conference papers, Web sites, awards, institutional affiliations, topics, patents, etc. To fully apprehend the multiple sources gathered by the Science Observatory, sufficiently powerful telescopes are needed. The intricate interlinking among these data sources means that traditional database query languages, form-fill-in, and tabular lists are only partially effective in exploration. To support exploration and discovery, visualization combined with statistical data methods is necessary to identify patterns, trends, clusters, gaps, outliers, and anomalies. Traditional visualizations present information from a single relational table source, but users increasingly wish to explore multiple linked relational tables. Paired visualizations, such as a geographic map with a journal publication database, would enable users to select a geographic area and see what kinds of research was being done, or to select a topic area and see where the authors are located. Another possibility is to have a granting agency's hierarchy linked to a publication's and patent's database to see how agency funding has produced meaningful outcomes. To accomplish this goal, multi-variate, temporal, hierarchic, and network visualizations are needed – plus a linking mechanism to connect datasets.

INTELLECTUAL MERIT

The complex network of relationship in science data presents a challenge for data exploration and discovery. New tools to deal with data cleaning, missing data, ambiguity and uncertainty are only the starting point. Exploring the richly interconnected data sources will require powerful, new primitives for information visualization tools, that go well beyond the multiple coordinated views with brushing and linking, that are the current state of the art. Snap-Together Visualizations are a starting point, but breakthrough ideas are needed to enable systematic, yet flexible exploration. Evaluating the efficacy of these new designs will also promote new techniques that go beyond the Multi-dimensional In-depth Long-term Case studies (MILCs) that have been applied in the past.

BROADER IMPACTS

Better tools for exploring science research data sets will enable more efficient and effective allocation of resources by government granting agencies, corporate research groups, and venture capitalists. Furthermore, the methods developed in this project will be widely applicable to those who wish to integrate diverse datasets that exist in many domains (such as healthcare, emergency response, online communities, sustainability, legal, finance, etc).

TRANSFORMATIVE POWER

A deeper understanding of the processes of science could change the way science is done, ensuring that researchers in similar or closely related areas are aware of each others work. This could accelerate science production and make more efficient allocation of resources. More importantly, this project could change the way science is conducted, shortcutting the publication process, putting potential collaborators in contact, discovering surprising insights from distant disciplines.

What Challenges and Opportunities do you foresee?

Data collection is the easy part. Then the complexity of cleaning and linking it will be greatly facilitated by visualization and data mining tools. The challenge will be to make comprehensible user interfaces that enable many people to grasp the rich relationships that exist. The opportunity within science and beyond is great, since making relationships visible supports discovery of patterns, trends, clusters, gaps, outliers, and anomalies. The goal of visualization is insight, not pictures.

Eric Siegel

Do Visualizations Tell a Story of Science?

Director & Chief Content Officer
New York Hall of Science
47-01 111th Street
Queens, NY 11368, USA

esiegel@nysci.org
http://nyscience.org

W2

POTENTIAL COLLABORATORS

Katy Börner, Ingo Günther, Martin Storksdiek, Steve Uzzo and Luís M. A. Bettencourt.

SUMMARY

I am interested in creating a project that tests the value and effectiveness of interactive visualizations with the public. How can we make visitors understand 1/1000 of the importance of the topics being discussed here in their daily lives (with regard to security, climate change, finance, health, risk, and other topics of direct interest to the general public). We can test various 2-D and 3-D methods of visualization and a range of interactive mechanics.

INTELLECTUAL MERIT

Through these discussions, I have become aware of two very interrelated major changes in the way science is being pursued. The first is the transition from not having enough data to having too much data. The second is the transition from the study of nature to the study of data. I am also interested in how an individual learner can move from the anecdotal understanding of phenomena (it is cold this winter) to the scientific understanding of phenomena (there is a decades-long upward trend in temperature). It appears that visualization can offer tools to address this discontinuity in the public's understanding of science. Exploring ways in which the public might engage with the sea of data surrounding them involves some testing and evaluation of public responses to different visualizations. The methodology for this research would be developed in collaboration with ILI, and the range of topics that we could explore could be addressed through collaboration with Luís M. A. Bettencourt, who is very polymathic in his interest in this field.

BROADER IMPACTS

Clearly research on how the public understands science is a central concern to science policy makers, scientists, educators, and above all the public themselves. Our hope would be to identify effective means of engaging the public with rich visualizations of relevant scientific (including social science) phenomena and that we could shape this into a program, exhibition, or electronically distributed education resource that would be experienced by millions of users in free choice and school-based learning experiences. If we can identify effective uses of visualization for the public, and identify generalizable characteristics of good visualizations, then this could inform further public education programming in schools, science centers, and other venues.

TRANSFORMATIVE POWER

I am a recent convert to the power of the kinds of visualization tools that are being discussed here. Like many converts, I have maybe extreme and uninformed views. That being said, I think there is something fundamental in the transformation from anecdotal to statistical—which is absolutely critical to establishing a scientific habit of mind. I would be fascinated to see if these tools can effect such a transformation, and I think it would be a very basic change in the way we envision learning about the world.

What Challenges and Opportunities do you foresee?

I think that the Informal Science Education program at NSF will be interested in such a project, and we are beginning to plan this project now. The challenge is the possibility of no one else being interested in supporting this research and public program, at least until NSF has funded a proof of concept.

Martin Storksdieck

Changing Public Discourse: A Tool to Bring Science to the Masses

Director of Project Development
Institute for Learning Innovation
3168 Braverton Street, Suite 280
Edgewater, MD 21037, USA

storksdieck@ilinet.org

W1

POTENTIAL COLLABORATORS

Major scientific and science-rich organizations and organizations that represent the public (from AAAS to NAAEE, from NSTA to science journalists), a few as Co-PI, and others in advisory roles;

A data-based system and visualization organization, maybe in collaboration with commercial and other existing structures;

Science mapping and data mining experts;

A science policy, science history/philosophy, science sociology Co-PI (such as Bruce Lewenstein);

A relevant expert in Public Understanding (John Falk, Jon Miller, Rich Borchelt, etc.);

An evaluation and assessment group with broad experience in program and product evaluation. Likely a team of different evaluators who bring in different strengths.

SUMMARY

Why would the public fund science? Is there a need for direct access to funded science in the United States and elsewhere? What transparency do we need so that "the public" is served well?

Who is the public?

Not everyday interested people, for the most part (or citizens and consumers – two distinct roles we play), but those who claim to represent the public (such as the media, political groups and parties, specific or special interest groups, environmental organizations, etc); however, there are also informal and formal educators, serious leisure/hobbyists and citizen scientists (which are mostly considered as part of the former group, but I like to name them separately). The latter group, not only provides an important social link between science (STEM) and the (science attentive and science inattentive) public, but they also act as informal science communicators (see for instance amateur astronomy outreach to the public, e.g. www.galaxyzoo.org).

I would also include everybody who is forced to learn about science, beginning at what developmental psychologists would consider an adult age: 12 and up.

What do we believe these groups need?

Part of the process and nature of science (and hence an element in understanding science as a human endeavor, not as one that just happens) is a better understanding of how science is supported, where science occurs, who is part of science and what the products of science are, both historically and currently. While scientists in their sub-disciplines have access to this information as part of their professional networks, their working infrastructure, the "public" does not. Science is an incredibly complex and confusing human enterprise or system for most of the public. The project will, therefore, provide an access portal that allows the public to explore these aspects of science through a database that allows relevant and simple queries and provides easy to understand and annotated visualizations. The system would be fed by experts or by a parallel, more complex system that is being developed simultaneously through the science community itself. The system would mesh data sets as they become available and create data analysis and visualization tools that allow rare and occasional users entry into an exploration and provides them with knowledge and understanding (implicit narratives).

The project would start with a large-scale, front-end evaluation of assumptions, needs, knowledge and (mis)conceptions, expectations and needs of a variety of stakeholder groups (see above).

Then an early proto-type would be developed and tested thoroughly with stakeholder groups through usability analysis and formative evaluation. The system (about which I say fairly little here because in order to describe it I would need my collaborators) would probably have to be developed as an open,

Wiki-like system that can organically grow and become more refined over time, adding more science systems and ultimately countries. Once launched, a remedial evaluation will be used to improve the in situ system.

The system would include an embedded evaluation system that tracks user feedback and impact of the system on the public perception and discussion of STEM, as well as the public's understanding of STEM (as PUS and PUR), particularly in areas of science that are of particular societal relevance.

The system would be funded by CDI, but placed into the ownership of an independent consortium of relevant organizations who would find the sustained funding source. The group would vote on membership and is charged to grow over time and become increasingly more inclusive. An important element of the system would be that it tracks how STEM resonates in society and in the public sphere.

INTELLECTUAL MERIT

The system would provide a new way for people to think about science, and it will combine traditional science mapping with science infrastructure and embed it into society. We will use an ecosystems approach, or one from sociology, to ensure that the focus is not on the increasing knowledge-base of science, but on science as a process and how science is resonating in society. In that sense, the system would teach its user about society and social structures itself. Since it would serve scientists and the public, it would also broaden scientists' understanding of their own role in society.

The team is highly qualified because its trans- and interdisciplinary composition will foster creative tension and debate. [In fact, the project will likely need a professional outside facilitator to minimize the risk of dangerous levels of tension].

BROADER IMPACTS

The broader impacts lie in assessing whether a central support system can change the relevance of STEM in the societal discourse about science. The broader impacts would be measured by a variety of indicators, but one would want to claim that the system should have a systemic effect: it needs to raise public awareness and understanding of the process and nature of science as well as raise the quality of the societal discourse on science. I would not claim that the outcome of the system is higher "acceptance" for STEM: we should leave that to the societal discourse itself; but the system would provide an infrastructure for accountability, openness and transparency; which, in the long run, should serve STEM positively.

TRANSFORMATIVE POWER

The system will likely change the way we think about science. When I say "we" I mean not just the public. Incidentally, most scientists are unreflective about science as system or enterprise. For better or worse, we will change the way science is placed within a larger societal context.

MARTIN STORKSDIECK BECOMES DIRECTOR, BOARD ON SCIENCE EDUCATION, NATIONAL ACADEMY OF SCIENCES/NATIONAL RESEARCH COUNCIL

The Board on Science Education (BOSE) is a standing board within the Center for Education which is part of the Division of Behavioral and Social Sciences and Education at the National Research Council, the operating arm of the National Academies. The Board meets biannually and its membership reflects expertise in a variety of domains within science and science education, such as natural and learning scientists, educational researchers, policy professionals, science education practitioners, teacher educators, philanthropic corporate stakeholders. BOSE:

- Provides advice on the research and best practices for science education at all levels in school and non-school settings. For example, one recent report (Taking Science to School) was targeted at grades pre-K through 8 and another (Learning Science in Informal Environments) will address life-long, out of school learning as well.
- Interprets research findings to inform practice. For example, Taking Science to School has a companion practitioner's volume entitled Ready, Set, Science!
- Develops future directions for science education research and practice. For example, a workshop was recently held on the status of mathematics and science standards, their implications, and the future research agenda.

Source: http://www7.nationalacademies.org/bose

Martin Storksdieck

Understanding the Pipeline of Science: A Community of Practice Based Systems to Create, Analyze and Predict Science Careers and Science Career Decisions

Director of Project Development
Institute for Learning Innovation
3168 Braverton Street, Suite 280
Edgewater, MD 21037, USA

storksdieck@ilinet.org

W2

POTENTIAL COLLABORATORS

A group that establishes the Science Observatory and Network;

An educational entity composed of a formal and informal research and development group (ILI, Upclose in Pittsburgh, Exploratorium, TERC, EDC, Robert Tai from University of Virginia, etc.);

Sociologists and psychologists involved in Positive Youth Development (Possible Selves, SAI, etc.);

Citizen science community, possibly through the Cornel Lab of Ornithology (Rick Bonney) or the Astronomical Society of the Pacific (Jim Manning or Suzy Gurton).

SUMMARY

This project idea is ultimately an add-on to a comprehensive system of registering and connecting scientists and their research, funding, dissemination and communication/cooperation (see Working Group Idea). While the original system allows science policy makers, funders and users to analyze trends in science, and scientists to create a more effective way of conducting research, this tool's purpose is to predict science careers and new development in science based on individual career choices. The project's theoretical framework would lie in the Community of Practice and Possible Selves and Self-Efficacy, and Situated (or enacted) Identity literature. The main idea is to add to the scientist and researcher community additional people, consisting of emerging scientists or scientists in training (graduate students and post-docs), high school and college students, middle school students, upper elementary students and formal and informal science educators/communicators. Young and emerging scientists or science enthusiasts would be invited to join the community of scientists with their own Intellectual Science ID, and share "their science" and scientific accomplishments, just like practicing scientists; science educators and communicators could register under similar conditions. While the system would serve as a teaching tool for understanding the nature, process and institution of science and research, it would also be constructed as a tool for fostering and sustaining interest in and fascination with science. This project would also provide opportunities for citizen and apprenticeship science projects that team researchers with "scientists in training" and young citizen scientists". By linking the data set of practicing scientists with emerging ones , it would provide the first analysis and forecast tool for STEM careers–one that could be analyzed by disciplines, science domains, countries.

The "teaching and apprenticeship" tool needs little explaining. However, active participation by young science enthusiasts is needed to create the data base for projecting and analyzing science career interest and choices, and to assess the potential for a future science attentive public.

The analysis tool would work fundamentally differently for emerging and established scientists. Scientists' future career path prediction would only be predictable after the system has been established for a few years, since historic data modeling is needed to validate the predictive power of the system.

Predicting established scientists' career path: based on existing career, past career, publication record, collaboration with other scientists (weighing their disciplines) and current research grants/work, the tool would estimate the likely deviation of an existing scientist's career path, (say from a theoretical physicist working on quantum-based calculations of molecules to a theoretical biochemist who applies these and similar calculations to key macromolecules in the cell environment). These shifts would be visualized in interactive cladogram-like charts.

The Pipeline function would be applied to upper elementary students and higher, allowing "What-if" scenarios based on historically validated models and current and past "science

contributions" and expressed science interest of participating youth. The system would likely use a variety of incentives to encourage sharing (visibility, recognition and reward), but may also, on occasion, use standardized psychometric surveys to gauge science interest, attentiveness, participation and attitude, as well as career awareness and goals. However, the system would mostly rely on (self-reported) behavioral data for younger participants.

INTELLECTUAL MERIT

The system provides a tool for fostering and sustaining interest in science and science careers and a tool for mapping and forecasting and self-assessing existing careers. The beauty of the system is that its self-reflection and assessment capabilities provide not only useful feedback for individual participants, but also the data can be aggregated to provide science policy makers and funders with predictions on the science (or STEM) pipeline. Additionally, in its most mature version, the system can predict where investments in science interest, attitudes and career aspirations might show the strongest results.

BROADER IMPACTS

- A system for personal science career monitoring and management
- A system for predicting future science trends based on individual career choices
- A system that actively uses Community of Practice and Situated Identity theory and Possible Selves Theory to feed the science career pipeline
- A system to predict the power of the pipeline
- A system to link science enthusiasts with scientists for the purpose of mentoring and apprenticing
- A system that provides opportunities for increased citizen scientists in close collaboration with existing endeavors

TRANSFORMATIVE POWER

- Allows prediction of science development, based on individual career choices
- Provides a means to predict the science career pipeline
- Identity-based science career and science interest creation through community of practice creates a strong affinity to science in young people ("it's real")

What Challenges and Opportunities do you foresee?

This project needs the overall system of science observatory and science community and is an add-on. It requires national and international community outreach to have educators (children), parents, and emerging scientists join. It also needs neutral and well respected organizations to host and manage. It has the potential to create privacy issues, and because children would be involved, it could create internet security issues as well.

Indiana University
Los Alamos National Laboratory
Yale University
New York Hall of Science