NSF Workshop on Discovery Informatics

February 2-3, 2012

Arlington, VA

Final Workshop Report

August 31, 2012

This workshop was sponsored by the Division of Information and Intelligent Systems of the Directorate for Computer and Information Sciences at the National Science Foundation under grant number IIS-1151951.

This report can be cited as:

"Final Report on the 2012 NSF Workshop on Discovery Informatics." Yolanda Gil and Haym Hirsh (Eds). National Science Foundation Workshop Report, August 2012. Available from http://www.discoveryinformaticsinitiative.org/diw2012.

Table of Contents

Workshop Participants

Workshop Chairs

Yolanda Gil, University of Southern California, Information Sciences Institute

Haym Hirsh, Rutgers University

Invited Participants

Cecilia Aragon, University of Washington

Phil Bourne, University of California San Diego

Elizabeth Bradley, University of Colorado

Will Bridewell, Stanford University

Paolo Ciccarese, Harvard University

Susan Davidson, University of Pennsylvania

Helena Deus, Digital Enterprise Research Institute

Clark Glymour, Carnegie Mellon University

Carla Gomes, Cornell University

Alexander Gray, Georgia Institute of Technology

Larry Hunter, University of Colorado Denver

David Jensen, University of Massachusetts Amherst

Kerstin Kleese van Dam, Pacific Northwest National Laboratory

Vipin Kumar, University of Minnesota

Pat Langley, Arizona State University

Hod Lipson, Cornell University

Huan Liu, Arizona State University

Yan Liu, University of Southern California

Miriah Meyer, University of Utah

Andrey Rzhetsky, University of Chicago

Steve Sawyer, Syracuse University

Alex Schliep, Rutgers University

Christian Schunn, University of Pittsburgh

Nigam Shah, Stanford University

Karsten Steinhaeuser, University of Minnesota

Alex Szalay, The Johns Hopkins University

Loren Terveen, University of Minnesota

Raul E. Valdes-Perez, Vivisimo Inc.

Evelyne Viegas, Microsoft Research

Cognizant Program Officer

Dr. Vasant Honavar, NSF CISE/IIS

Government Observers

Dr. Josh Alspector, IDA

Dr. Mitra Basu, NSF CISE/CCF

Dr. Bonnie Dorr, DARPA

Dr. Le Gruenwald, NSF CISE/IIS

Dr. David Jakubek, OSD

Dr. Jia Li, NSF MPS/DMS

Dr. Mark Luker, NCO NITRD

Dr. Wen Masters, ONR

Dr. Michael Nelson, Georgetown University

Dr. Grace Peng, NIH NIBIB

Dr. Marc Rigas, NSF OD/OCI

Dr. Edwina Rissland, NSF CISE/IIS

Dr. Tom Russell, NSF OD/OIA

Dr. Carey Schwartz, ONR

Dr. Abdul Shaikh, NIH NCI

Dr. Julia Skapik, AAAS Science and Technology Fellow

Dr. George Strawn, NCO NITRD

Dr. Kenneth Whang, NSF CISE/IIS

Dr. Maria Zemankova, NSF CISE/IIS

Dr. Fen Zhao, NSF CISE/CCF

Executive Summary

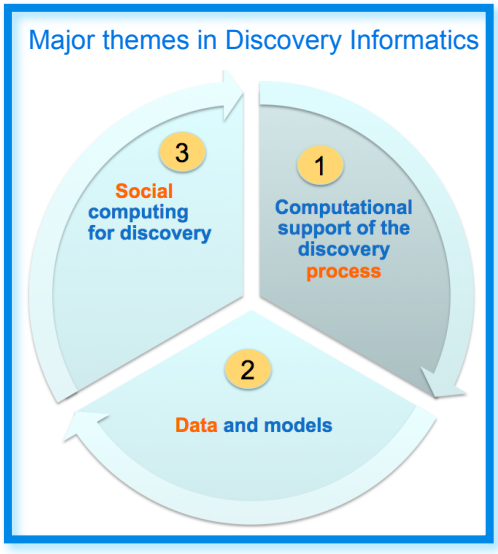
Advances in computing are transforming nearly all areas of science and engineering. In turn, the pursuit of new discoveries has resulted in innovations across all areas of computing. We are now facing the limits of our ability to gain insight from the volume, variety, and velocity of available data, posing fundamental challenges that can only be addressed through symbiotic advances in computing. Our ability to understand and gain insight from data of unprecedented complexity could be greatly increased with appropriate intelligent assistance and automation.

The Workshop on Discovery Informatics was convened to articulate the research challenges concerned with the management of knowledge and of the complex processes involved in scientific discovery. Workshop participants identified an expansive range of fundamental research challenges for information and intelligent systems brought into focus around three themes:

1. *New computational approaches are needed to manage the complexity of discovery processes that surpass human cognitive abilities.* This complexity often hampers scientists' knowledge and ability to analyze the large amounts of data at their disposal. We have reached a point where cognitive limitations are constraining scientific progress. We need to make scientific processes easily inspectable and reproducible. Innovations are needed to augment human abilities to analyze complex data through sophisticated processes, and enable understanding and insight.

2. *New computational approaches are needed to increase the connections between knowledge and data and exploit them to facilitate scientists' understanding of complex phenomena.* Data leads to new scientific knowledge, but the connection between knowledge and data is often not explicitly preserved in existing computational frameworks. As more complex data becomes available with increasing volume, variety, and velocity, the exploration of models becomes unmanageable, hurting our ability to do science effectively. We must develop general mechanisms for automated data-driven model refinement, data collection guided by models, and model-driven data analysis.

3. *New computational approaches are needed to flexibly combine diverse human abilities to tackle science problems that may not be otherwise considered possible.* New opportunities for discovery lie in the amalgamation of human expertise and effort. Although collaborations among scientists are common we have limited ability to facilitate unplanned, cross-disciplinary collaborations. In addition, we need better mechanisms to bring to bear human creativity to complement brute force computation, and open up science to valuable problem solving from massive amounts of volunteer contributors.



Existing relevant research efforts are scattered across disciplines and lack the critical mass needed to make a significant impact on these challenging aspects of science. Advances in these areas will transform the practice of science in two ways: 1) improving *existing* discovery processes that are unmanageable and suffer from human cognitive limitations, and 2) developing *new* discovery processes that increase our ability to understand challenging scientific phenomena. Further, outcomes in these areas are not domain specific, and can be leveraged across different science and engineering disciplines, having multiplicative returns, avoiding the inefficient, redundant development of computing innovations that would otherwise be repeated in specific disciplines (e.g., bio-, geo-, eco-informatics).

Discovery Informatics focuses on computing advances aimed at identifying scientific discovery processes that require knowledge assimilation and reasoning, and applying principles of intelligent computing and information systems in order to understand, automate, improve, and innovate any aspects of those processes. A new initiative in Discovery Informatics would enable and catalyze the transformational innovations needed to have a broad impact on the improvement and innovation of scientific discovery processes.

Discovery Informatics would require advancing basic research in many areas of computing, including: information extraction and text understanding to process publications and lab notebooks; synthesis of models from first principles, hypotheses, or data analysis; dynamic and adaptive design of data analysis methods; design, execution, and steering of experiments; selective data collection; data and model visualization; theory and model revision; collaborative activities that improve data understanding and synthesis; intelligent interfaces for scientists; design of new processes for scientific discovery; and computational mechanisms to represent and communicate scientific knowledge to colleagues, researchers in other disciplines, students, and the public.

Discovery Informatics will accelerate 21st century science and will have outcomes vital to the nation in numerous ways. National security is in severe need of better technologies for data analysis, noticing the unusual, and discovering patterns. Personal health and preventive medicine depend on our ability to enable people to contribute to the scientific enterprise in meaningful ways, by contributing data, analysis, personal histories, and sensor data. Our future relies on a better understanding of environmental and sustainability factors that is well beyond our current abilities. Our national competitiveness will be significantly boosted by a significant push in our nation's capabilities as a knowledge economy that would result from a renewed strength in Discovery Informatics. Discovery Informatics will advance the frontiers of computing, particularly in emerging areas of information and intelligent systems, while enabling new discoveries and innovations in all areas of science and engineering.

Participants stressed the need to act immediately. There is no doubt that our ability to generate and share data has surpassed our ability to analyze it. There is no doubt that there is data available or ready to be collected that could lead to many great discoveries of societal importance. We should strive to be in a position where not only can we harness the vast amounts of data at our disposal, but we are also able to pose increasingly complex questions that current methods do not even allow us to begin to imagine.

1 Introduction

Written by Yolanda Gil and Haym Hirsh

Computing has been a crucial enabling force for science in recent decades, creating in turn numerous opportunities for fundamental research in computer science. Ongoing investments in cyberinfrastructure have a tremendous impact on scientific discoveries [ACCI 2011]. Cyberinfrastructure today provides important capabilities such as high-performance computing, distributed services, shared high-end instruments, data management services, and support for virtual organizations. These investments have radically changed many sciences, and opened new doors to discovery and innovation.

However, scientists in all disciplines openly acknowledge their inability to exploit all the data and information that is already available to them and that continues to expand so rapidly (e.g., [Science 2011]). The volume, variety, and velocity of the data already available across all areas of science and engineering are already surpassing existing analytic capabilities to understand complex phenomena. Three hallmarks of 21st century science highlight major challenges for discovery:

1. **Discovery processes are increasingly complex.** This complexity results from having to integrate diverse data, software, expertise, results, etc. Literature search to synthesize what is known is one example of an increasingly unmanageable process given the ever-increasing size of the published record. Data analysis is another example, where complexity often hampers scientists' knowledge and ability to analyze the large amounts of data at their disposal. Unfortunately, many discovery processes are still largely human-driven activities. We have reached a point where cognitive limitations are constraining scientific progress. *New computational approaches are needed to manage the complexity of discovery processes that surpass human cognitive abilities.*
2. **Tight connections between knowledge and data are central to discovery processes around complex phenomena.** Data leads to new scientific knowledge, but the connection between knowledge and data is often not explicitly preserved in existing computational frameworks. This scientific knowledge is captured in a variety of forms: publications, influence networks, taxonomies, Bayesian models, etc. Keeping knowledge and data separate makes it harder for scientists to keep track of what hypotheses have been considered, what data supports them, what models have been created from the data, and how new hypotheses are formulated from those models. As more complex data becomes available with increasing volume, variety, and velocity, the exploration of models becomes unmanageable. *New computational approaches are needed to increase the connections between knowledge and data and exploit them to facilitate scientists' understanding of complex phenomena.*
3. **Innovative social processes can enable new discoveries.** New opportunities for discovery lie in the amalgamation of human expertise and effort. Although collaborations among scientists are common we currently lack the ability to facilitate unplanned, cross-disciplinary collaborations. A researcher addressing a complex scientific question in one field often only realizes the need for expertise in

another field during the course of the work. In addition, the public's participation in science makes it possible to have massive contributions of effort that result either in precious data that would not otherwise be available or in valuable problem solving that only humans can perform. *New computational approaches are needed to flexibly combine diverse human abilities to tackle science problems that may not be otherwise considered possible.*

A major research initiative focused on understanding and improving scientific discovery processes would have a profound impact on all sciences, accelerating the pace of scientific advances and innovation. Fundamentally new computational frameworks to address these challenges would make those processes significantly more manageable, enabling scientists to explore more complex phenomena than ever before. Those processes could also be made more efficient, making scientists significantly more productive. Moreover, new processes that do not exist today could be designed, enabling innovations to the scientific process that open doors to new discoveries.

Although there is some existing relevant research, the work is scattered across several disciplines and will not achieve the critical mass required to have a significant effect on scientific discovery. In computer science, there is relevant work in information management, intelligent interfaces, workflows, text extraction, visualization, machine learning, theory formation, collaborative systems, and social computing. There is also relevant work in the social sciences to understand the processes of scientific discovery, innovation, and collaboration. Researchers with common goals and complementary expertise are separated by disciplinary boundaries. Moreover, in the domain sciences these topics are addressed in a variety of informatics groups: bioinformatics, geoinformatics, ecoinformatics, astroinformatics, etc. As a result, advances have been piecemeal, with limited impact.

A new initiative in Discovery Informatics could bring critical mass to the improvement and innovation of scientific discovery processes. **Discovery Informatics focuses on computing advances aimed at identifying scientific discovery processes that require knowledge assimilation and reasoning, and applying principles of intelligent computing and information systems in order to understand, automate, improve, and innovate any aspects of those processes**. Discovery Informatics would encompass a broad spectrum of basic research in areas such as information extraction and text understanding to process publications and lab notebooks; synthesis of models from first principles, hypotheses, or data analysis; knowledge representation and reasoning for all forms of scientific knowledge; dynamic and adaptive design of data analysis methods; design, execution, and steering of experiments; selective data collection; data and model visualization; theory and model revision; collaborative activities that improve data understanding and synthesis; intelligent interfaces for scientists; design of new processes for scientific discovery; and computational mechanisms to represent and communicate scientific knowledge to colleagues, researchers in other disciplines, students, and the public.

The NSF Discovery Informatics Workshop was convened to explore the core research challenges for scientific discovery that concern information and intelligent systems. Many disciplines were represented at the workshop, with many attendees doing work that crosses disciplinary boundaries. Invited workshop participants included computer science researchers from academia and industry, scientists in several areas of science, and social scientists studying cognitive and social aspects of science. Major outcomes of the workshop were identifying the importance of Discovery Informatics, outlining an initial agenda for basic research in this area, and creating the seeds for a more cohesive community.

Table 1. A research agenda for Discovery Informatics

Discovery Informatics Goals

Computing advances aimed at identifying scientific discovery processes that require knowledge assimilation and reasoning, and applying principles of intelligent computing and information systems in order to understand, automate, improve, and innovate any aspects of those processes.

Key Challenges

- ➔ Information extraction and text understanding
- ➔ Model synthesis from first principles, hypothesis, and data analysis
- ➔ Reasoning with all forms of scientific knowledge
- ➔ Dynamic and adaptive design of data analysis methods
- ➔ Experiment design, execution, and steering
- ➔ Model-guided data collection
- ➔ Data and model understanding leading to insight
- ➔ Evolution of scientific models and theories
- ➔ Collaborative synthesis of new knowledge
- ➔ Meaningful participation of the public in science tasks

Areas of Basic Research

INFORMATION AND KNOWLEDGE
- Knowledge representation and reasoning
- Semantics and ontologies
- Data and information integration
- Model and theory revision
- Knowledge and information management
- Problem solving and constraint reasoning
- Process and workflow management
- Uncertainty reasoning
- Natural language processing

INTERACTION
- Cognitive aspects of scientific discovery
- Intelligent user interfaces
- Human computer interaction
- Collaboration and communication
- Visualization of models and data
- Social computing
- Innovation and creativity
- Tutoring and education frameworks

AUTONOMY
- Integrated intelligence
- Distributed intelligence
- Model-driven learning
- Intelligent control
- Adaptive and robust intelligence
- Robotics

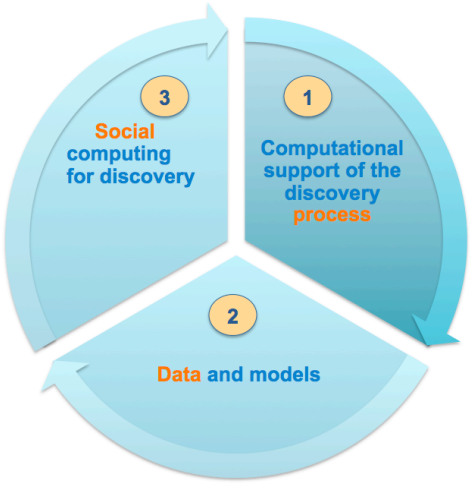


Figure 1. Three major themes in Discovery Informatics

Table 1 summarizes the goals, associated challenges, and resulting research agenda for Discovery Informatics. The rest of this reports describes in detail the rationale for these challenges and research goals.

Three overarching themes were selected to discuss a research agenda for Discovery Informatics at the workshop: 1) computational support of the discovery process; 2) integration of data and models; 3) social computing for discovery. These three themes are highlighted in Figure 1.

The rest of this report summarizes the discussions of workflow participants. The next section provides an overview of motivating scenarios in a variety of domain sciences contributed by the participants based on their work. These selected scenarios illustrate the kinds of challenging science discovery questions that necessitate significant improvements in the current state of the art. For each of the three broad research themes identified, we discuss the science needs, the state of the art, and the challenging research questions in computer science that need to be addressed. This report does not include exhaustive literature reviews to present the state of the art; instead, we highlight success stories contributed by the participants based on their own work. Next, we discuss general observations that emerged during the various sessions of the workshop. We also present the arguments put forward by the attendees regarding the urgency and timeliness of addressing this research area. We include two position statements by two participants reflecting on the workshop from their perspectives as experts in very different scientific fields. We end with recommendations for pushing a significant effort in Discovery Informatics that will inspire computer science researchers and will also benefit science across all domains.

2 Motivating Scenarios

During the workshop, several scenarios were discussed that illustrate the many opportunities for improving scientific discovery processes. Workshop participants work in collaboration with scientists in many areas; therefore, these scenarios illustrate the breadth and diversity of opportunities across scientific disciplines. This section presents three representative scenarios from the fields of social sciences, geosciences, and biology.

2.1 Education for better science, better citizens, and better communities

Written by Steven Sawyer and Susan Davidson

Despite decades of research, social scientists struggle to provide actionable guidance as to the efficacy of particular education choices or the long-term effect of various educational approaches. It is clear that more education is better than less, but what type of educational system is best for a particular individual inhabiting a particular social structure? With the rising cost of college education and decreasing federal and state funds available to offset this cost, how can online learning be used effectively and learning outcomes validated so as to be useful for potential employers?

Developing insight into these and other broad questions about education may now be within reach. We have moved from an environment in which digital information about local communities – e.g., quality indicators of schools, environmental information, local governance regulations, economic facts – was sparse to an environment in which there is a wealth of online information, in official forums as well as unofficial forums such as blogs. Such information, while locally important, could be studied at a national level to understand trends and correlations. Likewise, the data on educational approaches (such as curricular models, student performance, and student-produced materials) is increasingly available in digital form and could be collected, tagged, and used as a large-scale dataset for analysis and discovery.

For this opportunity to be realized, many barriers must be overcome. For example, data may be of poor quality, since it is often not curated or validated when entered, and may lack meta-data on context or provenance. It may also be incomplete with respect to the new questions that are being asked, which could be quite different from the ones anticipated when the data was collected. Interrelated data may be segmented across different datasets, and stored using incompatible formats and different terminologies. Relevant data may be in text form (e.g., blogs or descriptions), and thus not easily queryable. Furthermore, privacy and regulatory issues may arise when correlating data across datasets, or as a result of storing provenance information.

Advances in Discovery Informatics could help in many respects. Volunteer contributors could be guided through social computing platforms to contribute personal data and experiences, leading to significant improvement and expansion of data availability. In addition to the data, volunteers could be guided to contribute valuable meta-data and provenance information that would allow the interpretation and integration of data from multiple sources. Advanced models and analytic techniques need to be developed to exploit the diversity and volume of relevant data. Collaboration frameworks are needed to enable ad-hoc collaborations to integrate findings and analysis methods that are currently segmented across different intellectual communities. Intelligent support for developing models and understanding is needed, in frameworks that can be used not only by social scientists but also by researchers in other disciplines and by decision makers.

2.2 Forensic Paleoclimatology

Written by Liz Bradley and Karsten Steinhaeuser

Paleoclimatology is currently stymied by data-analysis challenges that could be solved with the assistance of scientific discovery tools. Cores, for instance, are used to sample

glaciers, trees, caves, and sediments at the bottom of oceans and lakes, among other things. There are vast archives of raw paleoclimate data lying around waiting for analysis. The World Data Center for Paleoclimatology archive at NOAA, for instance, contains millimeter-by-millimeter measurements of up to 13 variables in cores from 7,000 sites, some of which are thousands of meters in length. Without computational assistance, needless to say, this is not a humanly possible task.

The first step in analyzing the data contained in these archives is to create an age model: a curve that relates the depth in the core to the age of the material at that point. Some cores have discernible layers, but in many cases they have been obscured by intermixing, shifts, or other geological activity. Where annual layers exist, one can deduce the core's timeline by counting them. Where they do not, one must resort to forensic reasoning about the processes that created the core, and that affected it between formation and collection, in order to create the age model. This is not a trivial process; ocean sediment cores, for instance, are "bioturbated" by marine organisms, or glacial folding near their bases. For these reasons, deeper parts of the core may contain younger material. Worse yet, there are very few gold-standard measures of time; 14C's half-life is known, for example, and thus it theoretically makes a good "clock." However, its timescale is comparatively short and its levels in the atmosphere have varied over that time span. Occasional broad-scale events, such as volcanic eruptions and reversals of the Earth's magnetic field, can leave traces in cores; outside of that, independent synchronization marks are rare. As a result, building those models requires significant effort by a trained expert.

Advances in Discovery Informatics could provide new approaches to significantly improve and automate the analysis of cores and other paleoclimate data. New approaches are needed help automate labor-intensive tasks, enable new analyses, facilitate collaborations, and improve dissemination of results and findings to a broader audience. These approaches could likewise benefit other Earth science disciplines, including climate, ecology, and environmental science, among others.

2.3 Mass Phenotyping

Written by Helena Deus, Larry Hunter, and Nigam Shah

The rise of high-throughput technologies and the drop in price of genome sequencing have led to massive amounts of genotype information being produced and even managed and freely shared by its owners.[1] One of the primary challenges in making sense of the dramatic increase in human genotype data is finding suitable phenotype information for correlational analyses. Surprisingly, one of the most popular uses of this data has been in discovering long-lost relatives through phylogenetic analysis. Until recently, such phenotype data was primarily derived from assays or measurements made in clinical or research laboratories. However, laboratory phenotyping is expensive and low-throughput, and a variety of promising alternatives has arisen. For example, initiatives such as the "quantified self" allow tool makers and users with an interest in self-tracking to wear inexpensive sensors that collect data over extended periods of time; the data is then collaboratively analyzed and correlated through a crowdsourcing approach. A second potential source of phenotype data is that of behavior and epidemiological modeling through analysis of data from social networks. Mass behavior can be modeled, even predicted, through harnessing the data made available by the advent and popularity of the

[1] See, for example, http://www.23andme.com and http://www.patientslikeme.com

social web. These models could be used for market analysis, public health or even predicting the outcome of democratic elections. Furthermore, the dynamism in these networks ensures that the models can be constantly adapted and their accuracy improved. Finally, there is great potential in mining large numbers of scientific note-taking tools, such as electronic lab-books or electronic medical records, since these can reveal the inefficiencies and bottlenecks in the scientific discovery process.

Mass phenotyping is the process of collecting and integrating massive amounts of phenotypical information in order to discover patterns which would be invisible otherwise, and to correlate them with genotypical information. There are many applications where mass phenotyping would have a large impact. Diet patterns and obesity, for example, have been found to be correlated with a higher incidence of several cancers [Calle and Thun 2004]; mass phenotyping through integrating this genotypical information with patient behavior and physiological parameters could potentially be used to discover other such correlations that would otherwise remain unknown. Another example is mental disorders, which constitute 13 percent of the global burden on disease, surpassing both cardiovascular disease and cancer [Collins et al. 2011], and tend to be more prevalent in the ageing population. The most obvious symptom of mental diseases is behavioral changes, which can very easily be tracked through wearable or fixed sensors. However, not all behavior changes translate to disease: mass phenotyping can be used to identify which behavior changes are likely to be correlated with disease. Geno-phenotyping can be used to further validate and weight this likeliness of disease. Finally, public health policies may be put in place in order to monitor and prevent certain mass behaviors that could result in the spread of disease. In [Ferguson et al. 2005], the authors showed that elimination of nascent pandemics may be feasible using a combination of geographically targeted prophylaxis and social distancing measures. The availability of patterns from mass phenotyping in those rare situations may enable the easier identification and handling of risk groups, either through genetically identifying those individuals who are more likely to be affected or monitoring risk behaviors.

Discovery Informatics could enable the representation and integration of massive amounts of diverse phenotype information. Novel discovery and analytic techniques would help uncover complex behavior, social, and disease patterns. The collection on a large scale of detailed phenotype and other relevant data directly from individual volunteers would significantly expand and enrich the data available to researchers.

3 Computational Support of the Discovery Process

Written by Yolanda Gil and Kerstin Kleese-Van Dam

While advances in computing have transformed science, they have done so in tandem with a significant increase in the complexity of science practice. The depth and diversity of skills required to analyze the data available are hampering our ability to discover new complex phenomena. Many aspects of scientists' work are still labor intensive. Obtaining insight and understanding is increasingly hard in light of the growing complexity of science endeavors.

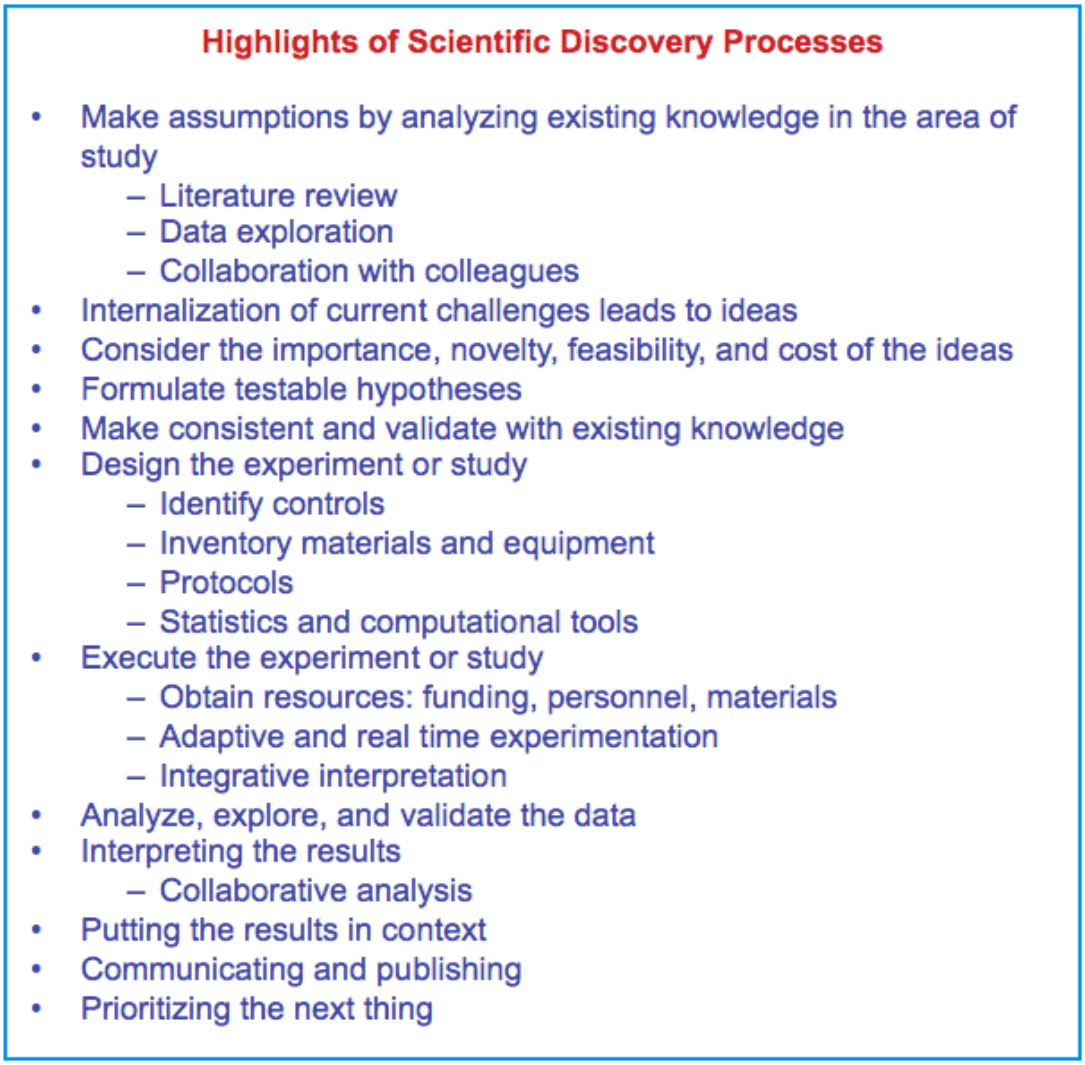


Figure 2. A high-level view of scientific discovery processes

Many of these obstacles to accelerating scientific discoveries pose fundamental research challenges for computer science. Interpreting results from complex data, gaining insights, and managing all the information and knowledge available are increasingly more challenging tasks. In addition, the overhead of interacting with many separate elements (software systems, data, people) significantly curtails productivity. Scientists could be much more productive and creative if new approaches were developed to manage the complexity of a broad spectrum of scientific processes.

3.1 Scientific Research and Discovery Processes

Figure 2 gives a high-level view of major aspects of scientific discovery processes. Scientists invest a considerable amount of their time understanding the state of the art in their area of investigation, by reading the literature, analyzing data, and discussing with colleagues. This understanding results in ideas that lead to the formulation of hypotheses and the design of experiments to test and evaluate them. The nature of experiments varies

widely across different areas of science, but preparing an experiment typically includes identifying controls, procuring instruments and other resources, designing protocols and techniques to collect the data, and selecting statistical methods and computational tools to analyze the data and confirm or refute hypotheses. Carrying out the experiments may take a long period of time, and may require monitoring and real-time analysis and subsequent adjustment of the instrumental apparatus. The data obtained is then analyzed and interpreted, typically by performing data cleaning and quality control steps, integrating data from additional sources, and then running combinations of analytic software into an end-to-end analysis method (e.g., simulation models, statistical routines, data mining). Interpreting the results then involves creating explanations for observed phenomena and examining the original hypotheses in light of the experimental results. This is often done in consultation with colleagues, and is an important component of the scientific publication process that ensues. Finally, scientists reflect on the work and prioritize what might be the most promising directions to pursue next.

3.2 Success Stories

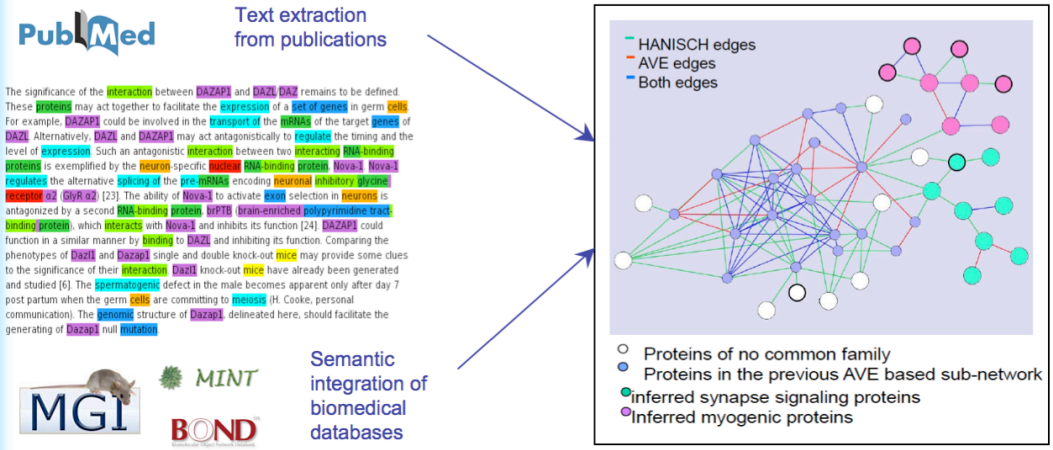
Research in recent years has shown the impact of improving scientific discovery processes. The research of the workshop participants represented several major aspects of work in this area.

Creating integrated models of existing knowledge from publications: The focus is on creating structured knowledge about what is known in particular areas of science, so that scientists can very efficiently review background knowledge relevant to their research. In biomedical research, for example, thousands of databases are created manually for this purpose [Galperin and Fernández-Suárez 2011]. Some research focuses on automatic methods to create knowledge bases from the literature. General-purpose text extraction techniques have been adapted to tackle particular types of facts and to integrate them with other available knowledge [Leach et al. 2009]. Reasoning algorithms have been developed to support inference of new hypotheses from a given body of knowledge, and to evaluate alternative hypotheses about biological process models by presenting to the user the assumptions and relationships that must hold in order for their model of a biological process to be true [Callahan et al. 2011]. Other research focuses on representing and relating scientific claims in different publications, the evidence to support them, and their relationships to other claims [Ciccarese et al. 2012].

Workflows to analyze data efficiently and record provenance: Workflows offer explicit representations of computational methods, and have long been recognized as a crucial element of scientific discourse [Gil et al. 2007]. Workflows represent explicitly how data is processed by software components, and as a result workflow systems can manage complex data analysis processes and keep automatic records of the provenance of new results obtained. This makes scientific methods and processes more reusable, inspectable, and reproducible. Shared workflow repositories and provenance standards are beginning to emerge. Workflow systems can automatically explore the space of possible experiments and customize the data analysis to the data [Gil et al. 2011].

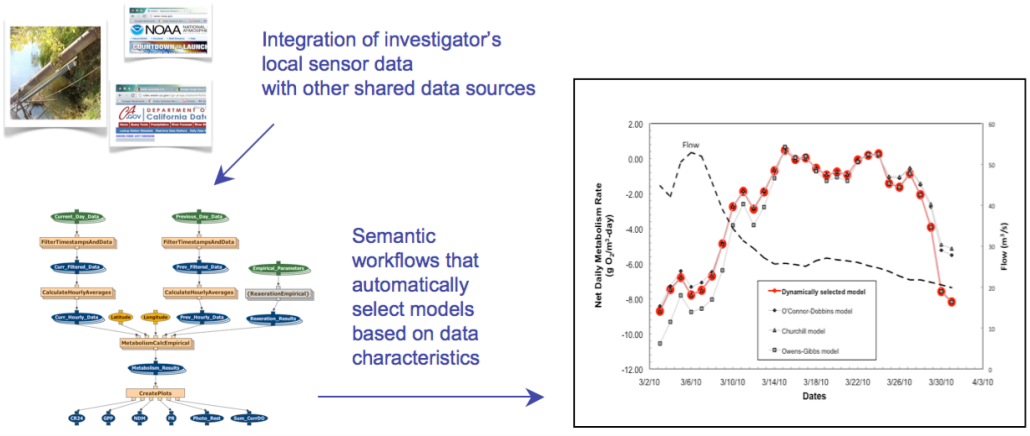
Success Story Highlight:
Discoveries through Automated Synthesis and Assisted Analysis of Scientific Publications

Efficient discovery of genes involved in mouse craniofacial development with the Hanalyzer system [Leach et al 09]. The system creates assertions based on co-occurrences in PubMed articles using open software for text extraction. It uses semantic web infrastructure to integrate assertions from existing biomedical databases. The system then reasons about the resulting semantic network to create novel correlations in the network. Scientists can visualize the augmented network and create hypotheses that can be tested in lab experiments. The bright green nodes and edges in the visualization shown here were inferred by the system, and suggested to scientists several new genes that were expressed but not previously detected experimentally. More details at http://hanalyzer.sourceforge.net.



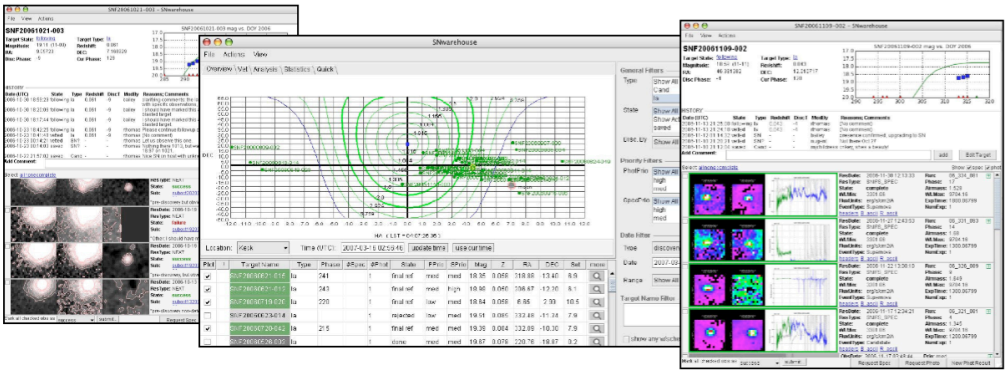
Success Story Highlight:
Efficient Data Analysis through Automatic Workflow Configuration

Semantic metadata and provenance are used throughout the data analysis process in the Karma data integration and the Wings workflow system to automatically choose models for water reaeration depending on river flow conditions. Karma integrates data from sensors with data from regional and national sources, generating metadata that is attached to the integrated datasets. Wings uses the flow, velocity, and reach geomorphology for each day of the period of analysis to choose an appropriate model, effectively configuring a different workflow to be run every day. The red dots highlight how the model chosen changes over time. More details at http://wings-workflows.org.



Success Story Highlight:
User-Centered Problem-Driven Visualizations

Visualizations designed to incorporate usability principles, to take into account cognitive loads of users, centered on the scientific tasks can have a positive impact on data-intensive science. Sunfall is a collaborative visual analytics system in operation at a large-scale astrophysics project, the Nearby Supernova Factory (SNfactory). Sunfall reduced the number of false positive supernova candidates by a factor of 10 and reduced scientist scanning workload by 90%.The project has discovered over 1,000 supernovae, of which some 600 are spectroscopically confirmed and nearly 400 of which are the critical Type Ia SNe so useful for cosmological research. The design principles developed for Sunfall could be applied to future scientific large-scale automated transient alert pipelines, such as NASA's Joint Dark Energy Mission (JDEM) or the Large Synoptic Survey Telescope (LSST). More details at http://snfactory.lbl.gov.



Information visualization to gain insights: A major focus of scientific visualization work has traditionally been the presentation of large datasets, addressing algorithmic and scaling issues. More recently, a growing trend in visualization is on user-centered, problem-driven work that emphasizes the selection, integration, and presentation of information in the context of a scientist's task [Meyer et al. 2009; Meyer et al. 2010]. This new line of research emphasizes the combination of data with models and other information, and the design of interactive interfaces that lead to sensemaking and gaining insight for real-world problems. Furthermore, this approach strongly relies on close collaboration between visualization researchers and scientists to ensure that the resulting visualization tools effectively support complex analysis tasks within the scientific discovery process.

3.3 Shortcomings of the Current State of Affairs

Many aspects of the scientific discovery process would be significantly more manageable through intelligent assistance and, in some cases, automation.

Some activities in science are only supported in very basic ways. For example, reviewing the literature to understand the state of the art in an area remains largely a human driven process. There have been many improvements, such as the availability of capable search engines, publishers' annotations of articles with metadata, explicit networks of citations, and many others. We mentioned above the ongoing research on creating declarative knowledge bases to amalgamate what is otherwise scattered across publications. These resources are extremely valuable, but they are created manually and thus will not scale [Baumgartner et al. 2007]. Automatic extraction from text has good performance for very particular types of tasks, such as entity co-reference. However, it has many limitations in terms of extracting more sophisticated structured information from

articles. Ontologies provide appropriate structures for the knowledge bases, but each one is developed with a particular focus, and integrating them is a challenge. The reasoners used have limitations in the kinds of inferences that they make. Deductive inferences are useful to add implied facts that are not explicitly stated, but more sophisticated reasoning is required to generate explanations and propose hypotheses.

Data collection, integration, and analysis processes include many repetitive steps that are still done manually, for example, data reformatting and conversion routines. As data is analyzed by different software (various simulation models, statistic analyses, etc), it must be converted to the particular formats required by each software tool. The entire process is typically driven manually, with the scientist selecting the software to compose complex methods, and configuring parameters each step of the way. Even workflow systems that support the process are not proactive in suggesting appropriate methods for a scientist's problem and data at hand.

Scientific collaborations are common but collaborative processes are far from adequately supported, particularly those of an opportunistic nature. Increasingly, scientific research is conducted by multi-institutional and interdisciplinary project teams, processing exponentially vaster and more complex data flows. Science collaboratories aim to bridge this gap by allowing scientists to share, reuse, and refine their computational workflows. However, tools for making ad hoc cross-disciplinary collaborations more commonplace are lacking, as is the fluid and efficient exchange of knowledge among researchers.

3.4 Research Challenges

Scientific processes must be made more explicit, allowing computers to manage them. Those processes should be described in such a way that domain-specific algorithms and software have well-defined roles. Aspects of those processes that focus on related activities should be easier to integrate with one another.

More knowledge about the context of each of the scientist's activities must be captured, so that systems can be more proactive and participatory in the processes. Formal representations of models, as well as appropriate metadata, would also facilitate the management of the processes.

Capturing scientific processes pervasively will enable cost-effective reproducibility. Open software and provenance standards to capture the processes used to generate scientific results will enable broad sharing and reuse of methods, enable inspectability of published results, and facilitate integration of research results even across domains. Science is steadily moving toward more open and shared resources, and we will need new approaches to discover such resources and exploit them. Privacy mechanisms for data and processes must be taken into account to respect personal and sensitive information, particularly in tools for biomedical and social sciences.

There are many examples of approaches designed to assist with various aspects of the discovery process, but there is much room for further investigation. Research on the generality of those approaches and on their broader uptake is needed. This research must involve both scientists and computer science researchers, so that both cutting-edge basic research and science impact can be ensured. We must understand well the tradeoff between generalized approaches and targeted approaches in terms of their effectiveness and usability. A better understanding of adoption of scientific software must be developed.

The use of data and information throughout scientific processes must be better supported. Finding information and data sources that are relevant to a problem should be done in terms of meaningful to a scientist. Data and information integration must be greatly improved, particularly through higher level concepts that allow cross-disciplinary research.

Existing social and collaborative approaches are insufficient to support the fluid exchanges of data and knowledge that are increasingly needed for scientific discoveries. Although many technologies make the sharing of information very easy, knowledge is still difficult to transfer, because it is often hard to represent and changes rapidly. However, common understanding can be negotiated. Social software could play a bigger role in developing common ground in relation to knowledge artifacts.

Many research challenges remain in visualization and intelligent user interfaces. Examining and exploring data and models interactively can help scientists gain insights into a problem. The design of interfaces for scientific tasks is an area that needs to receive more attention. Many lessons learned are scattered across science disciplines and may not be well studied or even reported. The principles behind effective integrated information presentations and interactive visualizations are not well developed.

4 Connecting Data and Models

Written by Pat Langley and Yolanda Gil

In a world flooded with data, there is a natural tendency to focus on data-centered science. Discovery Informatics research would bring models to the forefront and emphasize the interplay between models and data. Science has always involved an iterative process where the collection and analysis of data leads to models, and where model predictions and anomalies encourage collection of more data. Models also play a role in the design of new measuring instruments that produce new observations, and in the transfer of knowledge across sciences and into the engineering disciplines. New basic research is needed on approaches to design discovery systems that can exploit the interplay between data and models, closing the loop between data-guided model revision and model-guided data collection.

4.1 Models and Scientific Discovery

The use of models to communicate knowledge, to generate explanations of phenomena, and to turn science knowledge into engineering principles is a distinguishing characteristic of science. The use of models is absent in purely theoretical disciplines, such as philosophy, and purely empirical ones that mine data without attempting to understand general laws.

Figure 3 illustrates the variety of forms that models take across different scientific fields. For example, qualitative causal models are widely used in biology. In ecology, differential equations and Bayesian models are more common. In psychology, rule-based systems are used to represent cognitive skills. Physicists rely heavily in mathematical formalisms. In addition to the different forms of the models, there is also variability in terms of how explicit they are. Some models are simply expressed in the text of scientific articles, others are captured in transparent notations (e.g., a causal network), and others are embodied in artificial artifacts (e.g., software for complex simulations).

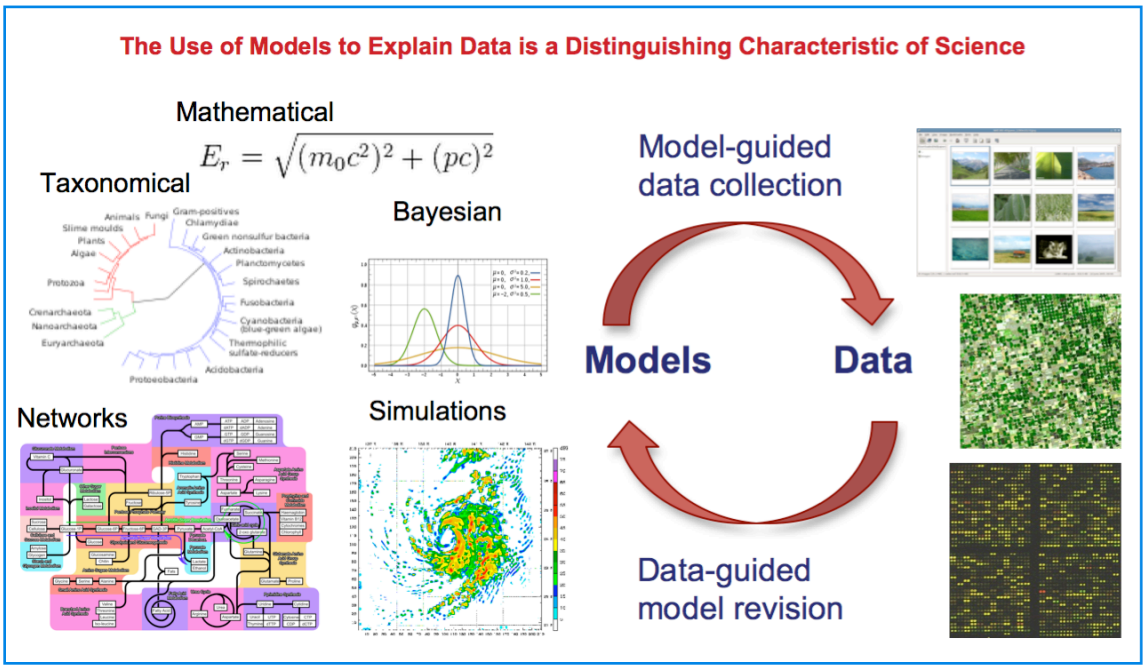


Figure 3. The interplay between data and models

Figure 3 also highlights the interplay between models and data in science. The generation, evaluation, and selection of models are all informed by observations, whether experimental or observational in character. Conversely, the collection and interpretation of data is informed by candidate models. A commonly used metaphor in computer science involves the notion of searching through a space of alternatives. If we apply this metaphor to the scientific enterprise, we can view scientists as carrying out search through two spaces that are connected but distinct. Search through the space of models is constrained by theoretical knowledge and by data, since each model aims to fit and/or explain the latter. Search through the space of data is constrained by current models, since observations are most useful when they distinguish among alternative accounts. These search spaces are very complex, and often heuristic knowledge guides scientists toward more promising areas of the search space. Together, these two interactions produce an iterative loop between data collection and model construction/revision that drives much of the scientific process.

The history of science suggests that the precise relationship between models and data can change over time. In a discipline's early stages, scientists are often content to find empirical relations that describe, summarize, and predict data collected through experimentation or observation. In contrast, more advanced fields often expect their models to move beyond simple description and to provide causal accounts or explanations that use conceptual terms familiar to scientists.

Computational tools that relate data and models would have great impact on discovery. Such computational aids can improve both the speed and the accuracy of model construction, data analysis, and model evaluation. This in turn will let scientists develop more comprehensive models that connect to larger datasets, which means they will be able to study and understand more complex phenomena effectively. Moreover, because these computational mechanisms embody general principles of model development, data

analysis, and evaluation, they will support transfer of knowledge and best practices among different scientific groups and even across separate fields. They would help make key aspects of the scientific method accessible to the wider community, and thus broaden participation in the overall scientific enterprise. They would also help disseminate sound scientific practices to government and industry practitioners interested in data analysis and data-intensive computing.

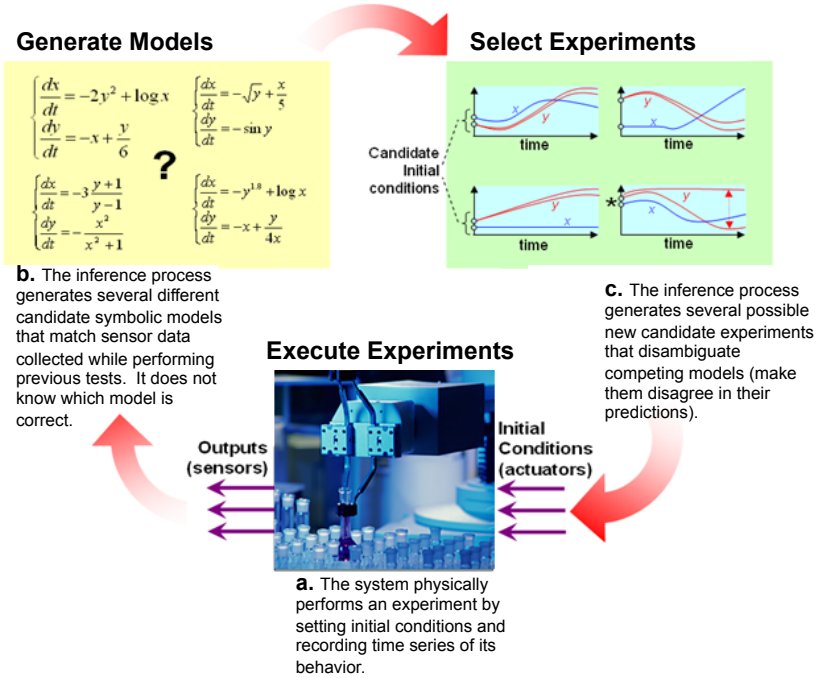
4.2 Success Stories

Computational discovery tools that integrate models and data have shown the potential for computers to contribute to the understanding of complex scientific phenomena.

Computational discovery tools that use scientific notations: The notations for models used in computational tools are often designed with computational tractability in mind, leading to far less expressive power than found in traditional scientific formalisms. Approaches that aim to discover knowledge in established scientific notations, whether qualitative causal models or differential equations or reaction pathways, make the models easier to communicate and understand [Shrager and Langley 1990; Dzeroski et al. 2007].

Success Story Highlight:
Automated Experimentation and Discovery of Natural Phenomena

Automated discovery systems formulate hypotheses and remote control an experiment to validate or refute those hypotheses. Initially the experiments are random, but as data are collected the system carries out experiments that test the most promising hypotheses. In a cycle of experimentation, hypotheses generation, and experiment design, the system can vary multiple parameters at once and work with multiple experimental systems simultaneously to rapidly converge on a small set of hypotheses for expert human inspection. For example, in genomics this approach scales better than traditional high throughput experiments that scan through multiple parameters in brute force. More details at http://www.eureqa.org.



b. The inference process generates several different candidate symbolic models that match sensor data collected while performing previous tests. It does not know which model is correct.

c. The inference process generates several possible new candidate experiments that disambiguate competing models (make them disagree in their predictions).

a. The system physically performs an experiment by setting initial conditions and recording time series of its behavior.

Robot scientists to automate end-to-end discovery: Robot scientists are systems that automate experimental execution and observation, and combine it with experimental design and model revision. Recent examples have been demonstrated in fields as diverse as physics [Schmidt and Lipson 2009] and cell biology [King et al. 2011].

Discovery of causal models: While there are many approaches to establishing simple associations or correlations in data, for many scientific phenomena the interest is in the discovery of causal models [Glymour 2004; Ramsey et al. 2010]. The availability of large datasets has made it possible to test causal hypotheses efficiently [Jensen et al. 2008], accelerating the cycle of discovery through bypassing the execution of sometimes costly experiments.

4.3 Shortcomings of the Current State of Affairs

As noted above, computers have been used successfully in science for decades, both to utilize formal models for prediction and explanation, on one hand, and to generate candidate hypotheses by analyzing data, on the other. However, these two movements have generally remained distinct, with work on model representation and simulation having only weak connections to observations, and with work on data analysis and hypothesis generation having few links to modeling traditions. Both computational approaches have offered many benefits to scientists, but, as long as they remain isolated from each other, they cannot reach their full potentials.

One drawback concerns representation. There exist a variety of computational environments that let users create, visualize, and simulate scientific models, especially in the fields of biology and environmental sciences. Some of these frameworks encode models in fairly simplistic terms (e.g., sets of causal links or sets of numeric equations), but others support richer conceptions of models that incorporate higher level content. Unfortunately, most of the work on computational discovery of models from data has focused instead on formalisms developed by computational researchers rather than domain scientists. There is also an inherent tradeoff between the expressive power of data or knowledge representations and their usability. In some areas of science there has been significant work on elaborate ontologies and data formats, yet users tend to gravitate toward the simplest ones that may lack expressiveness but are easiest to use for the task at hand. Many scientific disciplines that rely heavily on modeling and simulation also have a significant body of a priori knowledge, for example, in the case of Earth science there are physical principles governing fluid dynamics, heat transfer, and so on. Domain knowledge should be taken into account when analyzing data generated by such models.

Another limitation of most computational scientific research is that it assumes unidirectional processing. Some methods utilize models to generate predictions, at most using the match to observations for evaluation purposes. Other approaches utilize data to generate candidate models, but they assume one-pass processing in the opposite direction. Neither framework reflects the iterative, closed-loop character of the scientific process that has served many disciplines so well for centuries.

The recent and increasing availability of very large datasets in many areas of science has altered traditional processes for model formulation. Historically, the scientific context at a time in a discipline has suggested particular causal questions, and an experiment, or related sequence of experiments has been designed to answer them. Recently, the interest has been on automated or semi-automated methods for searching very large datasets for patterns that indicate more than accidental correlations. These search methods need development—

for example, methods are needed that can reliably extract causal cascades and their feedbacks from brain imaging data.

The availability of large datasets also presents issues of scale in formulating appropriate models. Many techniques scale poorly in the number of variables and in spatio-temporal granularity, and many methods for data-guided model induction scale poorly in the amount of data. The availability of data at different resolutions and different scales, and via different instruments (as they are upgraded over time), also presents important challenges to the understanding of the underlying phenomena.

Computational discovery methods have the potential for application to many different scientific fields, but their generality has not been well investigated. Clearly, there is a need for computational environments that scale well to all the facets of complexity that arise in science and that are accessible to a broad community of researchers.

4.4 Research Challenges

We should develop computational formalisms for scientific models, and their relationships to data, that support the full range of notations encountered in the sciences. This will require increasing the representational expressiveness of formalisms for both models and data beyond those typically used in computational work. It will also mean taking seriously the need to map these digital formalisms onto notations already in use by scientific communities for publication, instruction, and other forms of communication. The aim is to provide computational support for the full variety of scientific activities without limiting researchers' ability to express content.

We should address issues of generality and usability by identifying equivalence classes of scientific tasks that enable reuse of computational methods across many disciplines. The mechanisms that we develop for these equivalence classes should scale effectively (ideally, in a linear fashion) to sources of complexity in both models and data. One natural approach is to take advantage of the feedback loop between data collection and model revision, as discussed earlier. Moreover, these techniques should provide explicit support for reusing models, datasets, and operations performed over them.

To further ensure usability, we should embed these representations and mechanisms into interactive software environments that support the construction and revision of models, the collection and explanation of observations, and the relations among these processes. These integrated systems should incorporate not only efficient and general algorithms, but also principles of human-computer interaction to ensure they are widely accessible. We need systems that assist individual scientists in creating and updating models, collecting and interpreting data, and other key activities and processes that focus on the interplay of models and data. We also need explicit interlinking of science products in scientific communities that supports sharing not only of annotated datasets and comprehensive models, but also relations among them and links to relevant literature. This should include links to the data analysis and model creation processes in forms that make scientific results readily inspectable and easily reproducible.

The scale and complexity of the space of possible models is daunting for many scientific phenomena. Computational discovery tools will be crucial to make strides in these areas. Looking to the future, there will be many science questions that would require model creation well beyond human ability. Designing systems that can operate in a true

partnership with scientists and be trusted to explore and discover on their own will be key to deciphering many long standing scientific problems.

Finally, to let us determine whether our community is making progress, we must develop methods for evaluating both component algorithms and integrated discovery systems. Some techniques should draw on experiences with actual scientific models and data, to establish relevance, despite the complication that, in science, we can never be certain of "ground truth." However, we can complement such studies with evaluation methods that utilize synthetic models and data, which can provide known targets and also allow systematic experimentation [Langley 1996]. Together, these will let us study the robustness of our computational methods to factors such as model and data complexity, incomplete knowledge, and measurement noise.

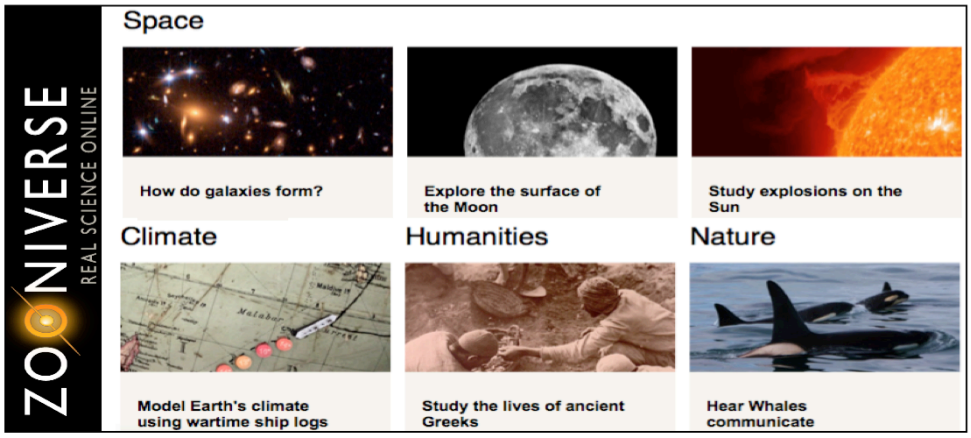
5 Social Computing for Science

Written by Yolanda Gil and Haym Hirsh

Scientific questions requiring overwhelming amounts of labor seem to be within reach thanks to the many unskilled volunteers that offer their services to science. These *citizen scientists* are contributing daily by collecting, labeling, and even analyzing massive amounts of data points [Savage 2012]. In addition, there is emerging evidence of collective intelligence resulting from group work [Woolley et al. 2010]. Harnessing people's ability to contribute to science is one of the most exciting approaches for innovating science processes and enable discoveries that were once out of our reach.

Success Story Highlight:
Social Computing for Scientific Discovery

Galaxy Zoo went live in 2007 and enlisted 175,000 citizen scientists to contribute to the classification of the morphology of one million galaxies from the Sloan Digital Sky Survey (SDSS). This task is impossible for computers and unmanageable for the science community, but can be tackled by thousands of lightly trained volunteers. A schoolteacher in Holland, Hanny van Arkel, discovered a strange green object below the galaxy in one of the pictures, now known as Hanny's Voorwerp (Dutch for object) and became co-author in the published article. Today, several sister projects in space, humanities, biology, and climate have engaged more than 630,000 people under the Zooniverse platform. More details at http://www.zooniverse.org.



5.1 The Role of Volunteer Contributors in Scientific Discovery

Science questions are becoming increasingly ambitious, and researchers do not always have the resources to address them. Contrast this with the strong interest of the public at large in science. Volunteers find tremendous appeal in contributing to science in a meaningful way. But they may be motivated for many other reasons. People also like to be able to contribute to solving problems of societal interest, for example, by contributing local geospatially tagged observations or personal medical data. We need more creative ways to harness people with a proven ability to make meaningful contributions to science. We need to broaden public participation in science, including students, motivated citizens, and younger scientists.

5.2 Success Stories

In recent years, a number of systems have been developed that successfully use volunteer contributions for a variety of science tasks. These systems explore particular points of what could be a very large space of possibilities in terms of volunteer contributions.

Success stories in this area exemplify different types of volunteer contributions, each with thousands or hundreds of thousands of participants. One approach is to take a large task and decompose it into very small subtasks that can be distributed to massive numbers of volunteer contributors who will each complete their task in no time. An example is the eBird project, where people are contributing bird sightings in their backyards giving scientists large amounts of data that they can use to study bird migrations (http://ebird.org/). These volunteers are naturally geographically distributed, so it takes them very little effort to provide this kind of data. Another way to harness volunteer effort is to give them tasks that are beyond a computer's abilities and can be better done by people. An example is GalaxyZoo, where people tag images taken from telescopes, and have provided astronomers with labeled observations of different kinds of galaxies (http://www.galaxyzoo.org/). Current image processing algorithms are not able to generate accurate labels, so here humans are performing computations that are not possible for computers. The Zooniverse system is a generalization of GalaxyZoo that is being applied to other astronomy problems, as well as historical and biology research (http://www.zooniverse.org). In other cases, collaborating in a task brings out people's ingenuity and creativity to accomplish things that they could not do individually. For example, conjectures in mathematics have been proven in very short amounts of time by collaborating volunteers (http://polymathprojects.org).

5.3 Shortcomings of the Current State of Affairs

Citizen science could be used in many more science areas if their social dynamics were better understood. Tales of discoveries aided by high school teachers, K-12 students, gamers, and crowds have become the talk of the town. We have little understanding of what science tasks might benefit from volunteer contributors, and how to make such volunteer efforts commonplace across different fields of science.

We have only an initial and very limited understanding of the principles for designing these systems so that they have appeal to contributors and keep them engaged in the long run. Polymath has very specific rules of engagement that facilitate collaboration. In other

cases, casting the work as a game is crucial to popularity. The key features in the design of all these kinds of systems that led to their success are not well understood.

Social computing systems for science are part of the overall science discovery process. It should be easy for other scientists to identify the aspects of their processes that could be aided by social computing; however, this is not well understood.

5.4 Research Challenges

Much research is needed to understand how to create effective human-computer teams. We must analyze existing approaches and develop a taxonomy of approaches with many modalities for human participation and a variety of forms of contribution. The role of human computation in larger computing contexts must be better studied. The collaborative creation of knowledge is an open research question, particularly as regards the accommodation of ad-hoc collaborations and unanticipated uses of data and information. Human creativity and ingenuity are a crucial resource in science, and can drive brute-force computation that systems can best carry out. Defining synergistic systems that combine human contributions and computation will innovate scientific processes and can lead to discoveries in areas where traditional methods have not made sufficient strides.

An open area of research is the design of such social computing systems. They must be designed so that the goals and beliefs of both humans and systems can be tracked and mutually understood in the context of the problem at hand and as the interactions progress. Participants may have a variety of backgrounds and expertise, so their roles and types of contributions must be defined and evolved over time. Defining tasks, decomposing them appropriately, and making appropriate assignments to either teams or individuals remains a challenge. The incentives that motivate people to participate, to sustain training, to change roles and types of contributions, and generally to stay engaged are not well understood.

New social computing paradigms could be developed that significantly augment what has been done to date. This could include new ways of producing, communicating, and reviewing scientific results, possibly redesigning many social aspects of traditional scientific processes.

6 Discovery Informatics: A Research Agenda for Intelligent Systems

The goals, associated challenges, and resulting research agenda for Discovery Informatics were highlighted in Table 1. Three major areas of research are key to meeting the challenges of Discovery Informatics:

1. *Information and knowledge*. This includes research on new approaches to knowledge representation, algorithms for reasoning about abduction and about constraints, reasoning about uncertainty in all forms, process modeling and reasoning techniques, non-monotonic theory revision, ontology development and exploitation, natural language processing, information and knowledge management, data and information integration.

2. *Interaction*. Research includes new approaches to human-computer interaction, intelligent assistance, collaboration and communication interfaces, visualizations of both models and data, social computing in science, cognitive aspects of discovery including innovation and creativity processes, and tutoring and education frameworks.
3. *Autonomy*. This area of research includes new approaches to integrating intelligent capabilities, adaptive and robust intelligence, distributed intelligence, model-driven learning, robotics, and intelligent control.

7 General Observations

Written by Miriah Meyer, Karsten Steinhaeuser, and Yolanda Gil

Several general observations and recurring themes emerged from the discussions. We summarize them here.

Big Data

Discovery Informatics will tackle unique big data challenges that would otherwise remain unaddressed. The volume, variety, and velocity of data is surpassing our ability to interpret and understand observations and derive comprehensive models that lead to new discoveries.

First, Discovery Informatics will address volume through the development of new approaches that integrate intelligent capabilities to reason with sophisticated scientific knowledge, explore large hypotheses spaces, fully automate the design and execution of experiments, and dynamically learn and adapt models to changing phenomena. These advanced intelligent capabilities will be required to mine vast quantities of data to understand complex phenomena.

Second, Discovery Informatics will address data variety by enabling the aggregation and analysis of smaller datasets, giving rise to new kinds of longitudinal big data. Moreover, many exciting prospects result from the integration of big data with local datasets collected by individual investigators (sometimes called "dark data" [Heidorn 2008]). Big data can provide breadth to smaller datasets to aid understanding of local phenomena in the context of the broader bigger picture.

Third, Discovery Informatics will enable coping with the velocity of data collection. Real-time data processing requires adaptive and flexible intelligent systems that can keep up with the pace of the data available, harness large temporal and spatial extent of complex phenomena, and design new collection apparatus that incorporate model-based control and experimentation.

Innovating Science Practice for Individual Scientists

A very large number of individual scientists who are studying small datasets would bring their research to a new level by integrating it with larger datasets about related, broader phenomena. There are myriads of such single investigators, possibly with the help of a few graduate students, working on problems of their own choosing. This type of research complements the science done by large collaboration teams, and is vital in engaging young scientists and building a sustainable workforce. It is critical to

acknowledge and balance all of these modes of scientific research. Discovery Informatics should address the spectrum of discoveries across the board.

Scientific Workflow

Within the traditional paradigm of the scientific method, observation generally leads to the formulation of hypotheses, which in turn are tested using controlled, repeatable experiments. However, advances in computational tools may radically transform the scientific discovery process. For one, Discovery Informatics can leverage abundant data in conjunction with powerful analysis tools for *hypothesis generation*. Moreover, the long-term vision (which is already being pioneered in some scientific domains, e.g., biology) includes comprehensive frameworks that can not only generate scientific hypotheses, but automatically *design and carry out experiments* to test them. Depending on the scientific discipline, these tasks may be performed autonomously (e.g., bench experiments) or in collaboration with human scientists (e.g., control of instruments).

Fading Boundaries between Computer Science and Domain Sciences

Many discoveries will be enabled by fading the boundaries between computer science and the domain sciences. Many rich problems and intuitions brought to bear by scientists can only be investigated through tools and innovations brought about by collaborations with computer scientists. However, these collaborations remain challenging to establish and to maintain. Scientists do not always know what they need, and do not have good connections with social scientists and computer science researchers who understand how to design discovery systems for ill-defined problems. Scientists often see computer scientists as providers of computing services, or as developers of research prototypes that are not ready for real use. Conversely, computer scientists often do not value the contributions to computer science brought about by scientists in other disciplines. In the end, the sciences are more open to computing than computing is open to the sciences. It is common to see computer scientists hired by science departments; it is extremely rare for scientists to become part of computer science departments. We need to do a better job of blurring the boundaries between disciplines, recognizing contributions that occur when those boundaries are crossed.

Adoption of Tools for Discovery across Sciences

Many discovery tools have been developed in different sciences, but rarely trespass into other science domains. In addition, the adoption of computer science tools for discovery is very uneven, as the number and nature of discovery tools is very diverse across sciences. There is much to be gained from automating routine scientific tasks across all sciences. We need to highlight success stories to encourage adoption and dissemination of ideas in Discovery Informatics across disciplines.

Discovering the New versus Discovering the Old

Many tools for discovery assist scientists by exposing and connecting what they have already discovered. Although they have value in their own right, more emphasis is needed on tools that discover new laws and provide new insights. As the complexity of scientific problems grows, the dimensionality of the hypothesis spaces will go beyond human abilities. We need to develop systems that can tame that complexity.

Building a Community for Discovery Informatics Research

Many workshop participants were surprised to find that they had many common interests, and yet they had never met one another before. They found new colleagues who already shared views about the challenges and approaches to science discovery, but who had never interacted before. The forums where the various workshop participants publish are very diverse and mostly non-overlapping. There is no common forum for sharing problems and learning from one another's experiences. More efficient investments will occur when researchers are able to build on one another's work.

Identifying Success Stories and Lessons Learned

Many aspects of Discovery Informatics research are empirical and practical in nature, and positive results are better documented than efforts that did not lead to strong successes, which are just as important but seldom reported. Success stories need to be identified and highlighted to better articulate the potential benefits of this area of research.

8 Why Now?

Written by Yolanda Gil and Haym Hirsh

Workshop participants stressed the need to act immediately. There is no doubt that our ability to generate and share data has surpassed our ability to analyze it. There is no doubt that we have data available or ready to be collected that could lead to many great discoveries. We should strive to be in a position where not only can we harness increasing amounts of data, but we will have developed the capability to pose increasingly complex questions that current methods do not even allow us to begin to imagine.

Addressing these challenges will require fundamental basic research that will significantly raise the bar on the intelligent capabilities of computational frameworks for science. Advancing our understanding of intelligence skills to supporting scientific discovery will bring information processing to a whole new level. These basic research advances will permeate all areas of computing.

Enabling discoveries is not just desirable for the sake of science, but is a necessity, as discoveries address problems of national and societal importance. National security is in severe need of better technologies for data analysis, noticing the unusual, and discovering patterns. Personal health and preventive medicine depend on our ability to enable people to contribute to the scientific enterprise in meaningful ways, by contributing data, analysis, personal histories, and sensor data. Our future relies on a better understanding of environmental and sustainability factors that are well beyond our current abilities. Our national competitiveness would be significantly boosted by a significant push in our nation's capabilities as a knowledge economy that would result from a renewed strength in Discovery Informatics.

Investments in Discovery Informatics would have a multiplicative effect in several dimensions. First, by addressing the human bottleneck in our data-rich world, advances in this area would help increase the rate of discoveries. Furthermore, they would enable investigations that we cannot even dare to pose today. In addition, advances in this area could be leveraged across all science and engineering disciplines. Organizing a community

would address current redundancy and inefficient compartmentalization in the domain informatics (e.g., bio/geo/eco/…).

Discovery Informatics would also benefit the individual science researcher while benefiting larger science collaborations. Single investigators working on local problems would find their activities better supported in analyzing their personal data. Moreover, Discovery Informatics would greatly facilitate the analysis of their local data in combination with large, shared datasets and big data initiatives that would otherwise not be incorporated into their work in practice.

Science is a costly enterprise, and engaging the public would enable scientists to harness massive amounts of volunteer effort from people who could make meaningful contributions. Discovery Informatics could inspire budding scientists of all ages, from energetic young students to retired professionals with interest and ability to volunteer time and resources.

By opening the scientific process, Discovery Informatics would engage, educate, and empower students and the public to innovate and to improve their lives. Personal data collected by individuals would give rise to "personal science", where people could study, for example, their own health, improve their neighborhoods, and monitor their local ecosystem.

Discovery Informatics would enable lifelong learning and training of the future workforce. The development of usable tools that encapsulate, automate, and disseminate important aspects of state-of-the-art scientific practice would allow: K-12 students to access important aspects of science research; undergraduates to become more involved in research projects, as they would be more accessible; post-doctoral and young researchers to be more productive in building their careers in science and engineering; and seasoned researchers to learn about new disciplines in a hands-on practical manner, significantly facilitating cross-disciplinary work.

9 Reflecting on the Workshop: Scientist Perspectives

Two prominent scientists attended the workshop and were invited to provide personal perspectives on the potential of Discovery Informatics to impact science.

Phil Bourne is a Professor in the School of Pharmacy at the University of California, San Diego. In addition to his contributions to computational biology, he is widely known for his leadership in the Protein Data Bank, one of the most widely used resources in the biomedical community. He is also founding editor of *PLoS Computational Biology*, a driving force in open scientific data sharing and publications.

Alex Szalay is an astrophysicist at the Johns Hopkins University. He collaborated for many years with Jim Grey on handling big data in astronomy. He has co-led several large community efforts including the Sloan Digital Sky Survey and the National Virtual Observatory, and the GalaxyZoo volunteer effort.

9.1 A Biologist's Perspective, by Phil Bourne

From my perspective as a basic researcher in computational biology, a maintainer of a major biological database, the Protein Data Bank (PDB), and the Editor in Chief of a high-profile, open-access journal (*PLoS Computational Biology*), the data deluge and how to address it in the best interest of science has been on my mind for some time. Somehow the term Discovery Informatics put my various thoughts into a larger and more exciting perspective. Prior to the workshop, many of us had been discussing various aspects of improving the scholarly lifecycle in a new medium in various forums, but the idea, perhaps obvious in retrospect, of improving the rate and depth of scientific discovery as a driver bought it all together. This happened in part because of the breadth of expertise in the room, all of which will be needed to make a difference.

"As this openness further pervades other disciplines and science itself becomes more cross-disciplinary the raw material for change is there. Right now, much of that raw material is stovepiped in individual data resources and journals, and the tool of discovery across those resources (with one or two exceptions) is a search engine. We need meaningful and automatic discovery across resources through deep search and analysis."

I was reminded of similar meetings some 20 years ago, as bioinformatics began to emerge. I can remember discussions with computer scientists around structures. On one occasion it took 10 minutes before I realized they were talking about data structures and I protein structures. We have come a long way since then. Computation is an integral part of modern day biomedical sciences research of any kind and biological scale — from atom to population. Being an integral part of scientists' daily activities will be true for Discovery Informatics, hopefully in much less than 20 years. In my opinion, the "tipping point" that got bioinformatics started was the advent of the human genome. Is there such a tipping point to foster in the era of discovery bioinformatics?

In Malcolm Gladwell's thesis, the tipping point may not be something obvious: the removal of all graffiti from the New York City Subway System leading to a city renaissance comes to mind. For Discovery Informatics I would like to think that open science is the catalyst. This is certainly a major factor in the biomedical sciences. In fact, bioinformatics careers, including my own, have been built, not from generating our own data, but by using the free and open data and knowledge generated by others. As this openness further pervades other disciplines and science itself becomes more cross-disciplinary, the raw material for change is there. Right now, much of that raw material is stovepiped in individual data resources and journals, and the tool of discovery across those resources (with one or two exceptions) is a search engine. We need meaningful and automatic discovery across resources through deep search and analysis. We need the ability to simulate living complex systems and share those models and outcomes. We need professional and non-professional scientists to be part of the process. They can be. One of the most interesting pandemic modeling studies I have seen recently was performed by a 15-year-old high school student. Empowerment through knowledge can have exciting and unexpected consequences. To date, the Discovery Informatics Workshop is the most exciting way forward I have seen to achieve these outcomes.

9.2 An Astrophysicist's Perspective, by Alex Szalay

"It is clear that computers will have an ever larger role in our daily lives as scientists. […] Some of our experiments will be designed by algorithms, some of our astronomical observing strategies will be optimized by clever workflows. Through new technologies we will see a much broader engagement of the public in deep science."

Astronomy has always been a data-driven discipline. We cannot do experiments with celestial objects; our only option is to observe them, and then do our best to interpret these observations. And observe them we did – for thousands of years astronomers have collected data which led to an increasingly sophisticated understanding of gravity, celestial mechanics, then nuclear physics, and more recently, even particle physics. The most accurate constraint on the mass of the neutrino, one of most elusive elementary particles, comes from astrophysical observations.

Arguably, the data explosion in modern science began with particle physics and astrophysics. As imaging detectors have become better and better, our telescopes have collected ever more data. Astronomers have always been accustomed to identifying extremely rare objects among the many typical ones; it still surprised everyone how rapidly the community has embraced the new technologies to look at ever more data, by running complex database queries. The Sloan Digital Sky Survey's database has rapidly become the world's most used astronomy facility.

It is clear that astronomy is generating some big datasets. At the same time, there is a "long tail": for every 100-terabyte dataset there are 100 1-terabyte collections, and hundreds of thousands of gigabyte-sized data collections. These smaller datasets represent a much more complex analysis challenge, due to their heterogeneity. The Virtual Astronomical Observatory is successfully emerging as a grass-roots effort to create an environment where scientists can combine their own small datasets with the big collections.

At the same time, so far the community has not found an easy way to either preserve or extract new knowledge from the aggregation of this "long tail." It is hard not to see the potential in bringing together many seemingly unrelated datasets into a single big collection, in which self-organization by similarities will reveal new, unexpected connections: consider the success of Facebook or YouTube. These examples show that even very light metadata tagging can still result in new connections and new meaning.

Looking at the sky is very appealing for a much broader audience than just professional astronomers. There are more than a hundred thousand amateur astronomers, with quite serious telescopes in their backyards. GalaxyZoo has attracted several hundred thousand people who spent millions of hours looking for strange objects at the website. We have seen the emergence of "Internet Scientists," who have made several major discoveries in the GalaxyZoo data. It led us to understand that there is a "long tail" not only in scientific datasets, but in the scientists themselves.

Over the centuries we have also learned to distinguish detection from discovery. Computers can help us to "detect" rare objects, yet it takes a human, understanding the context of the detection at more than one level, to see whether the detection is a truly

significant new discovery. Many supernovae have been detected by use of various telescopes over the last century, many of them by amateurs, yet it took Adam Riess and Brian Schmidt to recognize that the properties of some of the high redshift supernovae observed in the images taken by the Hubble Space Telescope have a profound implication about the ultimate fate of our Universe – and this insight led to their Nobel Prize.

It is clear that computers will have an ever larger role in our daily lives as scientists. Data-driven discoveries will be the norm soon, in many other areas of science beyond astronomy. Some of our experiments will be designed by algorithms, some of our astronomical observing strategies will be optimized by clever workflows. Through new technologies we will see a much broader engagement of the public in deep science. Shortly, most scientists will be as much at home in data analytics and statistics as in their own disciplines. By bringing together a rich mix of computer scientists, psychologists, machine learning experts, physical and life scientists, and sociologists, this workshop has shown the potential of this emerging brave new world we are about to enter.

10 Recommendations

Written by Yolanda Gil and Haym Hirsh

Critical mass and strategic thinking will only occur in a climate of sustained funding programs and a strong, synergistic community. The main recommendations from the workshop participants are:

- **Significant investments must be made in basic research in Discovery Informatics in order to create a critical mass that can make a significant impact in this area.** Basic research in information management, natural language processing, knowledge-centered data analysis and machine learning, model-based reasoning, robotics, education frameworks, collaborative systems, social computing systems, intelligent interfaces, and design is needed. Integrated intelligent capabilities will be required to address the intricacies of scientific discovery processes.
- **General principles and methodology in Discovery Informatics must be broadened across domain sciences**. The characterization of domains and facets that impact current Discovery Informatics practices is still not understood. This would help identify equivalent classes of tasks and problem domains across sciences. Methodologies to approach new domains, problems, processes, and users need to be developed. This kind of work cannot be done by domain scientists or computer scientists or social scientists alone. These disciplines need to come together on an equal footing to address these challenging and still ill-defined problems.
- **Creative mechanisms are needed to break the barriers across fields and subfields where key expertise to advance Discovery Informatics is widely scattered.** There are pockets of research in the social sciences, the domain sciences, and computer science, but there is virtually no communication across these disciplines. Some sciences, notably biology, have strong informatics communities, but many do not. Within computer science, the research is scattered across many areas, including machine learning, knowledge technologies and semantic web, human-computer interaction, natural language, databases, planning, and

collaboration research. There is also a need to involve social scientists to analyze science processes, understand requirements, and facilitate adoption of these technologies.

- **Basic research to advance Discovery Informatics needs to be facilitated and rewarded.** Students and young researchers should be trained and supported to pursue this research area. Discovery Informatics activities will require developing sustained collaborations with scientists. Traditional computer science criteria for research merit do not transfer well to Discovery Informatics research. Appropriate criteria need to be developed to encourage and reward research involving finding good problems, designing innovative approaches, evaluating and understanding the impact of those approaches in science practice, and generalizing the results to other science domains.
- **The impact of Discovery Informatics advances over time should be measurable.** We have seen a steady progress in the dimension of scale in computation in science, moving from terabytes to petabytes to exabytes and beyond. New dimensions for progress need to be articulated along other complex aspects of the scientific endeavor. Identifying success stories and significant advances in this area will help shore up the vision and the potential impact of pursuing this research agenda.

11 Conclusions

We envision Discovery Informatics providing the impetus for synergistic advances across multiple sub-areas of information and intelligent systems. Science will provide a unique testbed for developing integrative models of intelligence that will include model formulation, automated experimentation, learning, planning, reasoning, dynamic adaptation, human-computer interaction, and collaboration. It will also lay the groundwork for the development of the next generation exploratory apparatus, formal theories, and computational frameworks to not only accelerate discovery but to enable new modes of discovery to tackle questions that are currently well beyond our reach.

The broader impacts of Discovery Informatics research include the facilitation of interdisciplinary research at the interface between computer and information sciences and the various biological, physical, mathematical, health, social sciences and engineering.

These new discovery frameworks will result in enhanced modes of teaching and learning in science, technology, engineering, and mathematics (STEM) disciplines. The engagement of citizen scientists with varying levels of expertise and ability in scientific research will transform the scope and reach of science research in ways otherwise not possible.

Collectively, these activities are likely to not only fundamentally transform the practice of science across all disciplines, but also contribute to multiple areas of national priority such as healthcare, security, and sustainability, with significant impact on national competitiveness.

Discovery Informatics research has the potential to transform the scientific endeavor, and bring it to realms that would otherwise not be reachable.

Acknowledgements

This work was sponsored by the Division of Information and Intelligent Systems of the Directorate for Computer and Information Sciences at the National Science Foundation under grant number IIS-1151951.

References

[ACCI 2011] Final Reports of the Task Forces of the NSF Advisory Committee for Cyberinfrastructure (ACCI), March 2011. Available from http://www.nsf.gov/od/oci/taskforces/

[Baumgartner et al 2007] Baumgartner, W.A., Jr., et al. "Manual curation is not sufficient for annotation of genomic databases." Bioinformatics, 2007. 23(13): p. i41-8.

[Callahan et al 2011] Callahan, A., M. Dumontier, and N.H. Shah. "HyQue: evaluating hypotheses using Semantic Web technologies." Journal of Biomedical Semantics, 2011. 2 Suppl 2: p. S3.

[Calle and Thun 2011] Calle EE, Thun MJ. "Obesity and cancer." Oncogene 2004, 23:6365-78.

[Ciccarese et al 2012] Ciccarese P, Shotton D, Peroni S, Clark T. "CiTO + SWAN: The Web Semantics of Bibliographic Records, Citations, Evidence and Discourse Relationships." Semantic Web Journal, 2012.

[Collins et al 2011] Collins PY, Patel V, Joestl SS, et al. "Grand challenges in global mental health." Nature 2011, 475:27-30.

[Dzeroski et al 2007] Dzeroski, S., Langley, P., & Todorovski, L. (2007). "Computational discovery of scientific knowledge." In S. Dzeroski & L. Todorovski (Eds.), Computational discovery of communicable scientific knowledge. Berlin: Springer.

[Ferguson et al 2005] Ferguson NM, Cummings DAT, Cauchemez S, et al. "Strategies for containing an emerging influenza pandemic in Southeast Asia." Nature 2005, 437:209-14.

[Galperin and Fernández-Suárez 2011] Michael Y. Galperin and Xosé M. Fernández-Suárez. "The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection," Nucl. Acids Res. (2011) doi: 10.1093/nar/gkr1196 First published online: December 5, 2011

[Gil et al 2007] Gil, Y.; Deelman, E.; Ellisman, M. H.; Fahringer, T.; Fox, G.; Gannon, D.; Goble, C. A.; Livny, M.; Moreau, L.; and Myers, J. "Examining the Challenges of Scientific Workflows." IEEE Computer, 40(12):24-32, 2007.

[Gil et al 2011] Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. "Wings: Intelligent Workflow-Based Design of Computational Experiments." IEEE Intelligent Systems, 26(1), 2011.

[Glymour 2004] C. Glymour, "The Automation of Discovery." Daedalus, Winter (2004), pp 69-77.

[Heidorn 2008] P. Bryan Heidorn. "Shedding Light on the Dark Data in the Long Tail of Science." Library Trends, 57(2), pp. 280-299, Fall 2008.

[Jensen et al 2008] David D. Jensen, Andrew S. Fast, Brian J. Taylor, and Marc E. Maier. "Automatic Identification of Quasi-Experimental Designs for Discovering Causal Knowledge." Proceedings of SIGKDD, Las Vegas, NV, 2008.

[King et al 2009] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, Amanda Clare. "The Automation of Science", Science Vol. 324, 3 April 2009.

[Langley 1996] Langley, P. "Relevance and insight in experimental studies." IEEE Expert, October 1996.

[Leach et al 2009] SM Leach, H Tipney, W Feng, WA Baumgartner Jr, P Kasliwal, RP Schuyler, T Williams, RA Spritz, and L Hunter. "Biomedical Discovery Acceleration, with Applications to Craniofacial Development." PLoS Computational Biology 2009, 5(3): e1000215. doi:10.1371/journal.pcbi.1000215

[Meyer et al 2009] Miriah Meyer, Tamara Munzner, and Hanspeter Pfister. "MizBee: A Multiscale Synteny Browser." IEEE Trans. Visualization and Computer Graphics 15(6):897-904 (Proc. InfoVis 09), 2009.

[Meyer et al 2010] Miriah Meyer, Bang Wong, Tamara Munzner, Mark Styczynski and Hanspeter Pfister. "Pathline: A Tool for Comparative Functional Genomics." Computer Graphics Forum (Proc. EuroVis 2010), 29(3), 2010.

[Ramsey et al 2010] J. Ramsey et al., "Six Problems for Causal Inference from fMRI." NeuroImage, 2010 Jan 15;49(2):1545-58.

[Savage 2012] Neil Savage, "Gaining Wisdom from Crowds", Communications of the ACM, March 2012.

[Schmidt and Lipson 2009] Schmidt M. and Lipson H. "Distilling Free-Form Natural Laws from Experimental Data," Science, Vol. 324, no. 5923, pp. 81 – 85, 2009.

[Science 2011] "Challenges and Opportunities." Science, Vol. 331 no. 6018, pp. 692-693, 11 February 2011.

[Shrager and Langley 1990] Shrager, J., & Langley, P. (Eds.) "Computational Models of Scientific Discovery and Theory Formation." San Francisco: Morgan Kaufmann, 1990.

[Woolley et al 2010] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. "Evidence for a Collective Intelligence Factor in the Performance of Human Groups". Science 29 October 2010, Vol. 330 no. 6004 pp. 686-688, DOI: 10.1126/science.1193147