AAAI Technical Report FS-12-03
Discovery Informatics: The Role of AI Research in Innovating Scientific Processes

Discovery Informatics: AI Opportunities in Scientific Discovery

Yolanda Gil
Information Sciences Institute
University of Southern California
gil@isi.edu

Haym Hirsh
Department of Computer Science
Rutgers University
hirsh@cs.rutgers.edu

Abstract

Artificial Intelligence researchers have long sought to understand and replicate processes of scientific discovery. This article discusses Discovery Informatics as an emerging area of research that builds on that tradition and applies principles of intelligent computing and information systems to understand, automate, improve, and innovate processes of scientific discovery.

Introduction

Computing has been a crucial enabling force for science in recent decades. 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 had a tremendous impact on scientific discoveries [ACCI 2011], have radically changed many sciences, and opened new doors to discovery and innovation.

But advances in computing have also imposed new challenges to fully utilizing computing in scientific discover. 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. Human cognitive abilities to make discoveries are limited and greatly challenged by the overwhelming availability of data. These challenges have prompted a rich research agenda for Artificial Intelligence (AI) and information systems [Muggleton 2006; Waltz and Buchanan 2009; Gil and Hirsh 2012].

This paper discusses Discovery Informatics, an emerging area of research focused on computing advances that target scientific discovery processes requiring knowledge assimilation and reasoning, and applying principles of intelligent computing and information systems to understand, automate, improve, and innovate any aspects of those processes. AI has a long tradition of analysis and replication of scientific processes [Simon 1969; Langley et al 1987]. Discovery Informatics builds and expands on this tradition to innovate and improve scientific discovery processes.

The paper begins with a discussion of the potential of Discovery Informatics for two important current topics in science: "big data" and "the long tail" of science. It then focuses on two illustrative areas of research for information and intelligent systems: workflows of scientific processes and citizen science. They provide two examples of the potential for information and intelligent systems to improve and innovate scientific discovery processes.

Discovery Informatics

Three hallmarks of 21st century science highlight major challenges for discovery:

1. **Discovery processes are increasingly complex**. This complexity results from having to integrate a great range of resources, such as multiple diverse data sources, software systems of variable interoperability and usability, and participants with diverse expertise. For example, it has become increasingly unmanageable to conduct effective literature searches to identify and synthesize what is known in an area of interest given the ever-increasing size of the published record. Data analysis is another example, wherein the complexity of both the data and analysis tools often hampers scientists' ability to leverage effectively the large

amounts of data at their disposal. As a further example, discovery processes are still largely human-driven activities, and cognitive limitations increasingly constraining scientific progress. New computational approaches are needed to manage such complexities that are arising in contemporary discovery processes.

2. **Discovery processes for complex phenomena require tight connections between knowledge and data**. Science is becoming increasingly data-centered, with data leading to new scientific knowledge disseminated in such forms as publications, taxonomies, Bayesian models, and influence networks. However, the connection between that knowledge and the original data is often not thoughtfully captured and preserved in existing computational frameworks. This separation between knowledge and data makes it difficult 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 capture of 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 [Young 2010]. 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. Discovery Informatics could bring critical mass to the improvement and innovation of scientific discovery processes.

Discovery Informatics research 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.

Importantly, Discovery Informatics can also drive research in AI and information systems as it poses challenges to our capabilities in areas such as natural language understanding, model formulation and revision, knowledge representation, automated reasoning, semantics and ontologies, problem solving, constraint reasoning, uncertainty reasoning, process modeling and execution, robotics, intelligent control, distributed intelligence, adaptive and robust intelligence, model-driven learning, intelligent user interfaces, cognitive aspects of discovery, collaboration and communication, tutoring and education frameworks, integrated intelligence, and social computing.

Tackling Big Data

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. The availability of unprecedented amounts of data, sometimes referred to as "big data," will require new approaches to tackle the complexity of the underlying phenomena. Discovery Informatics could have great impact in our ability to analyze and understand big data.

First, Discovery Informatics would address volume through the development of new approaches that integrate intelligent capabilities to reason with sophisticated scientific models, explore large hypotheses spaces, fully automate the design and execution of experiments, and dynamically learn and adapt models to changing phenomena. Big data often requires new methods different from methods that work for smaller data, and often also requires large amounts of trial and error until appropriate methods are found for the data at hand. These advanced intelligent capabilities will be required to mine vast quantities of data to understand complex phenomena.

Second, Discovery Informatics would address data variety by enabling the aggregation and analysis of smaller datasets, giving rise to new kinds of longitudinal big data. In addition, 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 would 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 data collection, harness the large temporal and spatial extent of complex phenomena, and design new collection methods that incorporate model-based control and experimentation.

Uncovering the Long Tail of Dark Data

In recent years we have seen a surge of "collaboratories," where groups of scientists share very large datasets, expensive instruments, and supercomputing facilities. The Large Hydron Collider and the Compact Muon Solenoid experiment in physics and the International Virtual Observatory and the Sloan Digital Sky Survey in astronomy are good examples. In contrast, there is a "long tail" distribution [Anderson 2006] with a very large number of individual scientists who focus on collecting and studying small datasets that are seldom shared and are known as the "dark data" of science [Heidorn 2008]. These datasets and the contributions from these scientists are key to understanding important global questions, particularly in ecology and geosciences.

Discovery Informatics would have great impact on the long tail of science. Data preparation and analysis processes are often very costly, and tools to support these processes would greatly augment their productivity. Moreover, these tools could also assist in the integration of these smaller datasets with data from repositories from shared observatories. They would also enable the aggregation of individual datasets, and create big data for related, broader phenomena. This type of research complements the science done by large collaboration teams. Discovery Informatics should address the spectrum of discovery processes across the board, from big data to the long tail.

Modeling Discovery Processes: Workflows and Beyond

In this section we discuss some of the opportunities for automating and improving discovery processes through workflow systems.

Workflows have emerged as a useful paradigm to describe, manage, and share complex scientific analyses [Taylor et al 2007; Gil et al 2007b]. Workflows represent declaratively the software components that need to be executed in a complex application, as well as the data dependencies among those components. Workflow systems exploit workflow representations in order to manage the efficient execution of workflows in distributed environments. Workflow systems sometimes include a significant amount of analytics tools targeted to specific types of data, such as social sciences [Schmerl et al 2011] genomics [Reich et al 2006; Giardine et al 2005], and neuroscience image processing [Dinov et al 2009]. We describe here some workflow systems and some of the AI research enabled by these projects. Many more challenges remain that Discovery Informatics research could address [Gil 2009].

The Pegasus workflow system manages mapping and execution of computational workflows in distributed shared resources that may be highly heterogeneous [Deelman et al 2005; Deelman et al 2003]. To map workflow tasks, Pegasus uses descriptions of the execution requirements of each of the codes, and finds available hosts in the execution environment that satisfy those requirements. It includes several algorithms for optimizing the selection of execution resources not only based on task performance but also on minimizing queuing delays and data movement times. It also has facilities to recover from execution failures that may occur, due to bugs in the application codes, memory faults in the execution host, network failures, and other unexpected errors that are commonplace in distributed architectures. The Wings workflow system extends these capabilities with semantic workflows to enable the automatic elaboration of high-level workflows into executable ones [Gil et al 2011a].

Wings uses semantic descriptions of datasets to automatically configure parameters of the methods in the analysis and customize them to the data [Gil et al 2011b]. Semantic frameworks to describe domain-specific workflow components are beginning to emerge [DiBernardo et al 2008]. Workflows also have a clear role in enabling reproducibility of computational methods [Mesirov 2010]. There are many opportunities in workflow systems for AI, including constraint reasoning, intelligent user interfaces, and process modeling and execution management.

Taverna [Oinn et al 06; Hull et al 06] focuses on workflows for bioinformatics applications. In this area, there are thousands of services that are made available over the network for access by a wide community of scientists. Taverna workflows are composed from these services, and are cast in a simple and intuitive workflow language. Workflows from Taverna and other systems can now be shared in the myExperiment social site [De Roure et al 2009]. Scientists can post workflows, tag workflows to enable discovery and reuse, and rate the workflows. Many open problems remain in order to facilitate widespread collaboration and sharing of scientific workflows that would benefit from AI, including process representation, abstractions, collaboration, and distributed intelligence.

Social Computing for Discovery: Opening Science and Broadening Participation

Major innovations in scientific discovery processes are occurring in the area of citizen science, creating opportunities for Discovery Informatics. Citizen science projects are already inspiring budding scientists of all ages, from energetic young students to retired professionals with interest and ability to volunteer time and resources [Savage 2012]. Science is a costly enterprise, and engaging the public enables scientists to harness massive amounts of volunteer effort from people who are able to make meaningful contributions. Citizen science projects range widely in terms of the complexity of the contributions. We describe here a range of citizen science projects and some of the AI research done in these projects. There are many additional opportunities for Discovery Informatics that could amplify the impact of the work in this area.

Simple contributions can be sensing activities contributed by citizens. In the eBirds project [McCaffrey 2005] (http://www.ebirds.org) participants report on bird sightings in their local environment to enable scientists to track bird migrations. The vast amounts of data thereby available open the door to machine learning techniques. For example, [Kelling et al 2012] describes an approach to improve the quality of both human and machine contributions in the system.

Another way to harness citizen scientists effort is to give them tasks that are beyond a computer's abilities and can be better done by people. An example is GalaxyZoo (http://www.galaxyzoo.org/), wherein participants are given images of distant galaxies and must assign tags to them, thereby providing important information concerning astronomical images that would otherwise never have been inspected by astronomers [Lintott et al 2010]. 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 social sciences and biology research (http://www.zooniverse.org). AI techniques, in particular social computing, can be developed to design the best mechanisms to exploit the synergies between human contributions and computing. [Kamar et al 2012] use Bayesian models to predict the answers to a task when contributions may contain errors, and use these models to assign tasks to contributors. Assignment of tasks to contributors can be improved through grouping contributors into classes that have similar expertise levels [Psorakis et al 2011]. Recent approaches make these systems more robust to contributions of varying quality while taking maximum advantage of all contributors [Simpson et al 2011].

The FoldIt project exemplifies citizen science projects focused on collaborative contributions through serious games (http://fold.it). FoldIt enables contributors to form teams and compete to create the most optimal protein fold through complex geometrical reasoning [Khatib et al 2011]. Teams have proposed protein foldings that are superior to the best performing science algorithms. EteRNA is a more recent game where volunteers explore foldings of RNA molecules, where the best proposals are actually synthesized in the lab (http://eterna.cmu.edu/).

A more complex kind of contribution occurred in the Polymath project [Nielsen 2011]. It provided a massively collaborative online site wherein mathematicians collaborate with high-school teachers, engineers, and other volunteers to solve mathematics conjectures and open problems by decomposing, reformulating, and contributing to all aspects of a problem. This project uses common Web infrastructure for collaboration, interlinking public blogs for publishing problems and associated discussion threads with wiki pages that are used for write-ups of basic definitions, proof steps, and overall final publication (http://polymathprojects.org/, http://michaelnielsen.org/polymath1). Interactions among contributors to share tasks and discuss ideas are regulated by a simple set of guidelines that serve as social norms for the collaboration (http://polymathprojects.org/general-polymath-rules/). Tracking and assigning credit is central

to these social norms. It is unknown whether the simplicity of these norms will be preserved as this relatively young project evolves, or whether they will evolve similarly to Wikipedia's editing practices, which started with simple guidelines that became increasingly more complex over time.

Citizen scientists have also created their own science questions based on personal motivations, and used science-grade data and tools to make published contributions in first-rate journals [Rocca et al 2012]. By making scientific processes more explicit, Discovery Informatics would enable more ubiquitous occurrences of these kinds of self-organized citizen projects.

Discovery Informatics could bring about innovations in social computing to organize volunteer contributors of complementary skills and insights to be more effective and to solve increasingly more challenging science tasks.

Conclusions

Discovery Informatics has the potential to catalyze AI researchers to make significant contributions to scientific discoveries. This would require fundamental research advances in all areas of AI. Scientific discovery offers a challenging testbed for intelligent systems, with potential for inspiring new generations of researchers and for having very broad societal impact.

Acknowledgements

We would like to thank the participants of the NSF Workshop on Discovery Informatics for many useful discussions. This research was supported in part 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, "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/.

Anderson, C.. *The Long Tail*: How Endless Choice is Creating Unlimited Demand. London: Random House, 2006.

Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Blackburn, K., Lazzarini, A., Arbree, A., Cavanaugh, R., and Koranda, S. "Mapping Abstract Workflows onto Grid Environments." Journal of Grid Computing, Vol. 1, No. 1, 2003.

Deelman, E., Singh, G., Su, M., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Berriman, G. B., Good, J., Laity, A., Jacob, J. C., and D. S. Katz. "Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems". Scientific Programming Journal, Vol 13(3), 2005.

De Roure, D., Goble, C., and R. Stevens. "The design and realisation of the virtual research environment for social sharing of workflows." Future Generation Computer Systems, 25(5):561 – 567, 2009.

DiBernardo, M, Pottinger, R., and M. Wilkinson. "Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework." Journal of Biomedical Informatics, 41:837-847, 2008.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." Genome Research. 15(10):1451-5, 2005.

Gil, Y. From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. Scientific Programming, Volume 17, Number 3, 2009.

Gil, Y. and H. Hirsh (Eds). "Final Report of the NSF Workshop on Discovery Informatics." National Science Foundation project report, August, 2012.

Gil, Y., Ratnakar, V., Deelman, E., Mehta, G. and J. Kim. "Wings for Pegasus: Creating Large-Scale Scientific Applications Using Semantic Representations of Computational Workflows." Proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI), Vancouver, British Columbia, Canada, July 22-26, 2007.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. and J. Myers. "Examining the Challenges of Scientific Workflows," IEEE Computer, vol. 40, no. 12, pp. 24-32, December, 2007.

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.

Gil, Y., Szekely, P., Villamizar, S., Harmon, T. C., Ratnakar, V., Gupta, S., Muslea, M., Silva, F., Knoblock, C. A. (2011). Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows. Proceedings of the International Semantic Web Conference (ISWC), 2011.

Heidorn, P.B. "Shedding Light on the Dark Data in the Long Tail of Science." Library Trends, Vol. 57, No. 2, Fall 2008.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M., Li, P., and T. Oinn. "Taverna: A Tool for Building and Running Workflows of Services", Nucleic Acids Research, Vol 34, 2006.

Kamar, E., Hacker, S. and E. Horvitz. "Combining Human and Machine Intelligence in Large-scale Crowdsourcing," International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Valencia, Spain, June 2012.

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W., Yu, J., Damoulas, T. and C. P. Gomes. "eBird: A Human/Computer Learning Network for Biodiversity Conservation and Research." Proceedings of the Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence, Toronto, Ontario, Canada, July 22-26, 2012.

Khatib, F., F. DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski, and D. Baker. "Crystal structure of a monomeric retroviral protease solved by protein folding game players." Nature Struct Mol Biol. Sep 18;18(10):1175-7, 2011.

Langley, P., Simon, H.A., Bradshaw, G.L., Zytkow, J.M. "Scientific Discovery: Computational Explorations of the Creative Processes." Cambridge, MA: The MIT Press, 1987.

Lintott, Chris, Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, Jan Vandenberg. "Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies". Monthly Notices of the Royal Astronomical Society, 2010.

Mesirov, J. P. "Accessible Reproducible Research." Science, 327:415, 2010.

McCaffrey, R. E. "Using Citizen Science in Urban Bird Studies." Urban Habitats, 3, 1, 70-86, 2005.

Nielsen M. "Reinventing Discovery." Princeton University Press, 2011.

Oinn, T., Greenwood, M., Addis, M., Nedim Alpdemir, M., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., et al. "Taverna: Lessons in creating a workflow environment for the life sciences." Concurrency and Computation: Practice and Experience, Special Issue on Workflow in Grid Systems, Volume 18, Issue 10, August 2006.

Psorakis, I, Roberts, S. J., Ebden, M., and Sheldon, B. "Overlapping Community Detection using Bayesian Nonnegative Matrix Factorization." Physical Review E, 83(6), 2011.

McCaffrey, R. E. "Using Citizen Science in Urban Bird Studies." Urban Habitats, 3, 1, 70-86, 2005.

Muggleton, S. H. "2020 Computing: Exceeding human limits," Nature, Special Issue on 2020 Computing, Vol. 440, pp. 413-414, 2006.

Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. "GenePattern 2.0". Nature Genetics 38(5):500-501, 2006.

Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, et al. "Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach." PLoS ONE 7(7), 2012.

Savage, N. "Gaining Wisdom from Crowds." Communications of the ACM, Vol. 55 No. 3, Pages 13-15, March 2012.

Schmerl, B., Garlan, D., Dwivedi, V., Bigrigg, M. and K. M. Carley. "SORASCS: A Case Study in SOA-based Platform Design for Socio-Cultural Analysis." Proceedings of the 33rd International Conference on Software Engineering., Hawaii, USA, 2011.

Science editorial: Challenges and Opportunities. Special Issue on Dealing with Data. Science, 11 February 2011.

Simpson, E., Roberts, S. J., Smith, A., and C. Lintott. "Bayesian Combination of Multiple, Imperfect Classifiers." Proceedings of the Twenty-fifth Annual Conference on Neural Information Processing Systems (NIPS), 2011.

Simon, H. A. The Sciences of the Artificial. MIT Press, Cambridge, Mass, 1st edition, 1969.

Waltz, D. and B. Buchanan. "Automating Science." Science, Vol. 324 no. 5923 pp. 43-44, 3 April 2009.

Young, J.R., "Crowd Science Reaches New Heights," Chronicle of Higher Education 56, no.37, June 4, 2010.