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

Invited Talks

Timothy W. Clark,[1] William Cohen, [2] Lawrence Hunter, [3]
Chris Lintott, [4] Hod Lipson, [5] Jude Shavlik[6]
[1]Harvard University; [2]Carnegie Mellon University; [3]University of Colorado, Denver;
[4]University of Oxford; [5]Cornell University; [6]University of Wisconsin, Madison

Pervasive Semantic Annotation of Biomedical Literature Using Domeo

Timothy W. Clark

Despite the fact that we now have access to almost all peer reviewed publications on the Web, these publications appear to us in a linear form which is a replica of print journals. At the same time there are increasingly attractive opportunities to surface data and concepts directly on the web, using semantic organization. This talk will discuss how — for biomedical researchers — the web of documents and the web of data / concepts can be bridged and integrated, using the Domeo Web Annotation Toolkit and the Annotation Ontology (AO). Domeo and AO can be used to annotate any HTML document whether or not it is under update control of the user. The AO annotation can be selectively shared and exchanged and is orthogonal to any specific biomedical domain ontology. We believe this approach will be extremely useful in drug discovery to break down information silos, increase information awareness and sharing, and integrate terminologies and data with documents and text, both public and private. We will discuss applications we are currently developing in collaboration with a major pharma.

Timothy W. Clark is director of bioinformatics, at the MassGeneral Institute for Neurodegenerative Disease and an instructor in neurology of the Harvard Medical School. He trained as a computer scientist at Johns Hopkins, and began his work in life science informatics as one of the initial developers of the National Center for Biotechnology Information (NCBI) Genbank and a collaborator on the initial NCBI prototype of PubMed. He subsequently served as vice-president of Informatics at Millennium Pharmaceuticals, where his team built one of the first integrated bio- and chemi-informatics software platforms in the pharmaceutical industry. He is a founding editorial board member of the journal *Briefings in Bioinformatics,* an advisory committee member of the World Wide Web Consortium, and an advisory board member for the Neuroscience Information Framework (nif.nih.gov). Clark's research program focuses on multimodal semantic integration of biomedical web communities, scientific discourse and experimental results. He is the principal investigator of the semantic web applications in neuromedicine (SWAN) (swan.mindinformatics.org) and science collaboration framework (www.sciencecollaboration.org) projects. His informatics group built the reusable software platform for Stembook (www.stembook.org), an online review of stem cell biology published by the Harvard Stem Cell Institute, and created the Parkinson's Disease (PD) Online Research website (pdonlineresearch.org) in collaboration with the Michael J. Fox Foundation for Parkinson's Research.

Reasoning with Data Extracted from the Scientific Literature

William Cohen

The growing size of the scientific literature has led to a number of attempts to automatically extract entities and relationships from scientific papers, and then to populate databases with this extracted information. In my group we have been exploring techniques for using this sort of extracted information for specific tasks, including "bootstrapping" to improve the coverage of an extraction system, retrieval tasks involving the scientific literature, and modeling protein-protein interaction data. This is joint work with Ramnath Balasubramanyan, Dana Movshovitz-Attias, and Ni Lao.

William Cohen received his bachelor's degree in computer science from Duke University in 1984, and a PhD in computer science from Rutgers University in 1990. From 1990 to 2000 Cohen worked at AT&T Bell Labs and later AT&T Labs-Research, and from April 2000 to May 2002 Cohen worked at Whizbang Labs, a company specializing in extracting information from the web. William Cohen is president of the International Machine Learning Society, an action editor for the *Journal of Machine Learning Research,* and *ACM Transactions on Knowledge Discovery from Data.* He is also an editor, with Ron Brachman, of the AI and Machine Learning series of books published by Morgan Claypool. In the past he has also served as an

action editor for the journals *Machine Learning, Artificial Intelligence,* and *Journal of Artificial Intelligence Research.* He was general chair for the 2008 International Machine Learning Conference; program cochair of the 2006 International Machine Learning Conference; and cochair of the 1994 International Machine Learning Conference. Cohen was also the cochair for the 3rd International AAAI Conference on Weblogs and Social Media, and was the program cochair for the 4rd International AAAI Conference on Weblogs and Social Media. He is a AAAI Fellow, and in 2008, he won the SIGMOD "Test of Time" Award for the most influential SIGMOD paper of 1998. Cohen's research interests include information integration and machine learning, particularly information extraction, text categorization and learning from large datasets. He holds seven patents related to learning, discovery, information retrieval, and data integration, and is the author of more than 180 publications.

The First Artificial Mind Will Think about Molecular Biomedicine

Lawrence Hunter

Biomedicine, particularly as informed by genome-scale instrumentation, provides a unique domain for artificial intelligence and discovery informatics research. There are at least three phenomena that contribute to its status as a good domain for AI. First, there are several important characteristics of the domain, including (a) the knowledge-based (rather than lawlike) nature of scientific explanation in biomedicine, (b) the modest role that common sense knowledge plays in biological reasoning, and (c) the possibility of embodiment of programs in the context of powerful automated experimental instrumentation. Second, there are a variety of highly significant resources available to researchers developing AI systems, including (a) extensive, continuously expanding, publicly available, and incompletely analyzed dataset of high value; (b) carefully constructed ontological resources constructed and maintained by diverse communities of experts, and (c) many specific use cases that offer clearly defined and highly significant problems that AI techniques have high potential to address. Finally, there is an extensive community of biomedical researchers and practitioners highly motivated to exploit and interact with computational systems that increase the quality, speed or ease of their scientific insights. The power of this combination of eager user community, valuable existing resources and appropriate domain characteristics is already clear from existing work in biomedical discovery informatics, but, as this talk will try to argue, the future is even brighter.

Lawrence Hunter is the director of the computational bioscience program and of the Center for Computational Pharmacology at the University of Colorado School of Medicine, and a professor in the departments of pharmacology and computer science (Boulder). He received his Ph.D. in computer science from Yale University in 1989, and then spent more than 10 years at the National Institutes of Health, ending as the chief of the Molecular Statistics and Bioinformatics Section at the National Cancer Institute. He inaugurated two of the most important academic bioinformatics conferences, ISMB and PSB, and was the founding president of the International Society for Computational Biology. Hunter's research interests span a wide range of areas, from cognitive science to rational drug design. His primary focus recently has been the integration of natural language processing, knowledge representation and machine learning techniques and their application to interpreting data generated by high throughput molecular biology.

Efficient Crowdsourcing: How to Do Science with 600,000 Participants

Chris Lintott

Citizen science in the form of crowdsourcing has to proved to be an effective response to the growing size of scientific datasets. This talk will present strategies and results from the Zooniverse, a collection of projects that have enabled more than 500,000 people to help scientists classify galaxies, discover planets, sort through whale songs and even transcribe ancient papyri. As datasets continue to grow, these projects must adapt, and the talk will concentrate on methods that move beyond the current naive random task assignment model. A dynamic Bayesian classification of volunteers, applied to a supernova hunting project, was able to achieve much greater efficiency in classification, while dividing classifiers into communities based on their ability and behavior. Future development of such systems will need to incorporate such analysis into their methodology, allowing user behavior to guide task allocation, training and perhaps even collaboration.

Chris Lintott is a researcher in the department of physics at the University of Oxford where he is also a junior research fellow at New College. As PI of Galaxy Zoo and chair of the Citizen Science Alliance, he leads a large, transatlantic and multidisciplinary team of developers, scientists and educators with the aim of building the widest possible range of projects that enable meaningful public participation in science. His own research focuses on the formation of the present day population of galaxies, and he is a strong advocate of public understanding of science. In this latter role he serves on the board of trustees of Royal Museums Greenwich, and is copresenter of the long-running BBC series *The Sky at Night.*

The Robotic Scientist: Distilling Natural Laws from Experimental Data, from Particle Physics to Computational Biology

Hod Lipson

Can machines discover scientific laws automatically? For centuries, scientists have attempted to identify and document analytical laws that underlie physical phenomena in nature. Despite the prevalence of computing power, the process of finding natural

laws and their corresponding equations has resisted automation. This talk will outline a series of recent research projects, starting with self-reflecting robotic systems, and ending with machines that can formulate hypotheses, design experiments, and interpret the results, to discover new scientific laws. While the computer can discover new laws, will we still understand them? Our ability to have insight into science may not keep pace with the rate and complexity of automatically-generated discoveries. Are we entering a post-singularity scientific age, where computers not only discover new science, but now also need to find ways to explain it in a way that humans can understand? We will see examples from psychology to cosmology, from classical physics to modern physics, from big science to small science.

See Schmidt M., Lipson H. (2009) "Distilling Free-Form Natural Laws from Experimental Data," *Science,* 324(5923): 81–85. Try it on your own data.

Hod Lipson is an associate professor of mechanical and aerospace engineering and computing and information science at Cornell University in Ithaca, NY. He directs the Creative Machines Lab, which focuses on novel ways for automatic design, fabrication and adaptation of virtual and physical machines. He has led work in areas such as evolutionary robotics, multi-material functional rapid prototyping, machine self-replication and programmable self-assembly. Lipson received his Ph.D. from the Technion - Israel Institute of Technology in 1998, and continued to a postdoc at Brandeis University and MIT. His research focuses primarily on biologically-inspired approaches, as they bring new ideas to engineering and new engineering insights into biology. For more information visit www.mae.cornell.edu/lipson.

Human-in-the-Loop Machine Learning

Jude Shavlik

Machine learning has made tremendous progress over the past several decades. It has become one of today's most important technologies for discovery and its future impact is likely to grow rapidly for the foreseeable future. However, to use the powerful capabilities offered by machine learning, domain experts typically need to find a collaborator who is a highly trained computer scientist possessing substantial experience with machine learning. This greatly limits the impact of this powerful technology. We are addressing the important challenge of reducing the barrier to entry for using machine learning by allowing domain experts to more directly communicate their expertise to machine learning algorithms. Providing such domain expertise in an effective manner promises to democratize machine learning, more quickly spreading this valuable technology to tasks where it can have a substantial impact. We are focusing on allowing domain experts to do more than providing (a) the features used to describe examples and (b) the desired outputs for training examples. We are creating learning algorithms that accept naturally expressed "advice" whenever a domain expert has some knowledge that he or she wishes to provide. The human-provided advice need not be 100% correct since our learning algorithms are robust in the presence of imperfect advice.

Jude Shavlik is a professor of computer sciences and of biostatistics and medical informatics at the University of Wisconsin - Madison, and is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI). He has been at Wisconsin since 1988, following the receipt of his PhD from the University of Illinois for his work on explanation-based learning. His current research interests include machine learning and computational biology, with an emphasis on using rich sources of training information, such as human-provided advice. He served for three years as editor-in-chief of *AI Magazine* and serves on the editorial board of about a dozen journals. He chaired the 1998 International Conference on Machine Learning, cochaired the First International Conference on Intelligent Systems for Molecular Biology in 1993, cochaired the First International Conference on Knowledge Capture in 2001, was conference chair of the 2003 IEEE Conference on Data Mining, and cochaired the 2007 International Conference on Inductive Logic Programming. He was a founding member of both the board of the International Machine Learning Society and the board of the International Society for Computational Biology. He coedited, with Tom Dietterich, *Readings in Machine Learning*. His research has been supported by DARPA, NSF, NIH (NLM and NCI), ONR, DOE, AT&T, IBM, and NYNEX.