Advancing Science through Mining Libraries, Ontologies and Communities

James A. Evans1,2 and Andrey Rzhetsky2,3

1Department of Sociology, University of Chicago, Chicago, Illinois 60637, USA
2Computation Institute, University of Chicago and Argonne National Laboratory, Chicago, Illinois 60637, USA
3Departments of Medicine and Human Genetics, Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA

Life scientists today cannot hope to read everything relevant to their research. Emerging text-mining tools can help by identifying topics and distilling statements from books and articles with increased accuracy. Researchers often organize these statements into ontologies—consistent systems of reality-claims. Like scientific thinking and interchange, however, text-mined information—even when accurately captured—is complex, redundant, sometimes incoherent, and often contradictory: It is rooted in a mixture of only partially consistent ontologies. We review work that models scientific reason and suggest how computational reasoning across ontologies and the broader distribution of textual statements can assess the certainty of statements and the process by which statements become certain. With the emergence of digitized data regarding networks of scientific authorship, institutions and resources, we explore the possibility of accounting for social dependencies and cultural biases in reasoning models. Computational reasoning is starting to fill out ontologies and flag internal inconsistencies in several areas of bioscience. In the not too distant future, scientists may be able to use statements and rich models of the processes that produced them to identify under-explored areas, resurrect forgotten findings and ideas, deconvolute the spaghetti of underlying ontologies, and synthesize novel knowledge and hypotheses.

A vast and rapidly growing volume of text traces the succession of findings and ideas that constitute modern science. Extrapolating from global library data, we estimate the world hosts at least a trillion scholarly pages. An incomplete inventory (see Figure 1), divided into biological, social and physical sciences, contains 400, 200, and 65 billion pages, respectively (see supplementary online material). From the Western invention of the printing press in 1453, scientific knowledge has grown, increasingly become published in English, and shifted from books to journals (Figure 1A). Published knowledge has accumulated fastest in eras of peace and prosperity; it grows much more slowly in turmoil (Figure 1B). Until recently, access to this knowledge required going to a library. The complete collection of science, however, is distributed so widely across libraries that to assemble all knowledge on any broad topic would require lifelong travel (Figure 1C). More than a quarter of the world’s basic and applied science books percent appear in less than ten libraries. Google and the Google Book settlement, which finalized November, 2009, is beginning to reverse this trend by making millions of these books available for search and reading through the Internet. This follows the massive migration of scientific journals online over the past decade (Figure 1D).

With the emergence of new journals born-digital and novel ways of expressing findings and hypotheses in science tweets, blogs and online databases—and with more scientists producing science than ever before—researchers can catalog only a vanishing fraction of what is relevant to their work by traditional reading and note-taking. In response, scientists in many fields have begun to use computation not only to search and browse scientific texts (1), but to read and reason about them (2). Numerous obstacles remain, but the possibility of enlisting computation in discovery as well as analysis has inspired a growing body of knowledge and tools whose use and development have been nowhere more active than in the molecular life sciences.

Processing Natural Language

The process by which text is refigured into standardized, machine-readable representations of meaning is often called semantic analysis. The expressive richness and ambiguity of natural language, however, make automatically extracting statements from scientific text a formidable challenge. For example, consider the difficulty involved in extracting all information from the text: “NCOA3 in turn acylates histones which makes downstream DNA more accessible.” As a result, information extraction (IE), a robust approach to semantic analysis, currently avoids attempting to process every phrase in text (e.g., by simply extracting “NCOA3 acetylates histones”). IE assumes a relatively simple, fixed template of expected information. Researchers then fill semantic slots with information from text through a series of steps.
First, what is to be extracted is narrowed using a supplied lexicon of semantically classified terms (e.g., genes, enzymes, co-factors, small molecules) (3). The researcher then identifies these classes in text, a step called named-entity recognition (NER) using deterministic rules or computational techniques that statistically learn from human-coded data. The accuracy of the best NER in some domains rivals that of human coders at over 90% (4). The researcher then assembles these mentions into basic and then complex noun, verb and prepositional phrases. Finally, semantic entities and events are recognized, inserted into the template and merged if they are determined to share a referent. Many biomedical research teams in recent years have extracted protein-protein interactions from text using an IE approach and filling a simple template which involves two or more proteins, an interaction, and occasionally evidence associated with the claim. Following from our previous example, NCOA3 acetylates histones as determined with high-resolution mass spectrometry.

A related but more involved method of robust semantic analysis builds on formal grammars that model how the mind generates language. Several exist, including transformational, relational, dependency, construction and categorial grammars (5). Transformational or constituency grammars (TGs) are the most commonly used and comprise a set of symbols representing constituent words and phrases and the rules by which they are substituted to create language. Parsing a sentence, given a TG, boils down to reconstructing the substitution steps that most likely generated the sentence. Context-free grammars (CFGs) are currently the TG most commonly implemented as they offer the most affordable compromise between expressive power and computational cost. A syntactic CFG renders a sentence through symbols that signify structures like noun and verb phrases, which are each, in turn, substituted for word classes like nouns, verbs, adjectives and prepositions, which are ultimately replaced with words like “bright” or “phosphorylate.”

Similarly, a semantic grammar operates with symbols that stand for semantic categories such as enzyme, substrate, or enzymatic reaction, and constitutes rules including selection restrictions that limit the ways in which entities can appear in semantically meaningful statements within the domain (5). Semantic grammars in science rely on the notion that scientific domains are characterized by specific scientific sublanguages, which can be characterized by a finite set of rules (6). In robust semantic grammars, one typically mixes syntactic and semantic rules, and may implement them deterministically or probabilistically. In a probabilistic grammar, each rule is associated with a probability that can be used to infer the most likely genesis or parse of the sentence. Collectively, these approaches to robust semantic analysis are used extensively to extract meaningful statements in computer science, business intelligence, biology and medicine. They have been particularly fruitful in fields at the interface of chemistry and biology, where the fundamental importance of molecules and reactions have driven more linguistic conventions than many other areas of science and scholarship. They are also beginning to enter many other natural and social sciences.

Assembling Ontologies

If computation could distill articles into a database of statements, it would still be too large for a researcher to browse. Imagine all possible statements of the abstract form molecule A is a substrate for enzyme B. Or consider a scientist interested in the relationship between human gene p53, which regulates the cell cycle, and cancer, a cell proliferation disorder, who finds 36,589 relevant articles in PubMed. The path to computational reasoning commonly involves aggregating statements from text into singular declarations ordered into a consistent system or ontology. Modern scientific ontologies derive from two traditions. In philosophy, ontology—historically a branch of metaphysics—is the study of existence. For modern science, this translates into a specification of empirical entities and their organization into statements of fact that highlight essential qualities, parts and relationships. Ontologies emerged in computer science as an approach to knowledge representation in the early 1990’s (7), preceded by practical classification systems like the 19th century Dewey Decimal system and the International Classification of Diseases. First used in artificial intelligence, ontologies now permeate software engineering and database theory, where consistent content schemas facilitate the interoperability of data stores. They are also becoming an integral part of the World Wide Web (WWW). In 2004, the WWW Consortium endorsed the Web Ontology Language (OWL), which supports the formal semantics required by ontologies, as the centerpiece of its Semantic Web framework.

Modern scientific ontologies, often classified into light- and heavyweight, contain a controlled vocabulary of concepts and relationships that link them. Lightweight ontologies comprise terminologies or simple taxonomies with little or no information about the entities or relationships. For example, all known enzymes organized into a list or a simple taxonomy (e.g., a kinase is-an enzyme) would serve as a valuable data resource. In contrast, heavyweight ontologies like
Cyc, a massive schema of common sense knowledge, or the Foundational Model of Anatomy (FMA) add formal axioms and constraints to characterize entities and relationships distinctive to the domain. A formal axiom in biology might specify that gene encode proteins; but proteins cannot encode genes. Light and heavyweight ontologies may draw upon reference or upper-level ontologies to characterize their parts—abstract entities and processes as in Cyc or concrete elements like human bones within the FMA. They then subclass these elements and link them with domain-specific information. The most frequently cited ontology in science, the Gene Ontology (GO), is a structurally lightweight taxonomy that comprises 22,000 entities biologists use to characterize gene products (8). GO statements are concept-relation-concept triplets like “oxidative phosphorylation is a metabolic process” and “photosynthesis, light harvesting is part of photosynthesis, light reaction.”

Knowledge bases of text-mined statements similarly draw upon ontologies for the lexicon used to extract entities and relations from text, but duplication and contradiction are permitted as they are in articles (9).

After computer and information science, ontologies are most used in biomedicine, but also increasingly in astronomy and diverse areas of engineering, government and business. Recall the recent ontological debate over whether Pluto is a planet, an asteroid, or a dwarf planet—the controversial appellation eventually contrived by the XXVIth General Assembly of the International Astronomical Union in Prague 2006. Among applied ontologists, there is broad agreement that ontologies should primarily be understood as precise data structures to facilitate sharing and reuse—a kind of object-oriented content (10). Ontologists are divided, however, over whether to promote one ontology that enables/constrains the interoperation of all others or to let a thousand flowers bloom and encourage a wide range customized to scientific usage. “So much of scientific knowledge is not absolute—it is constructed—it is context-dependent. Ontologies can provide an impression of certainty that may not always be appropriate” described Mark Musen, co-editor of Applied Ontology and author of the popular ontology editor Protégé. Consider, for example, the multiple, co-existing definitions of gene concurrently used in biology: a unit of inheritance, a chunk of DNA, and a template for a group of proteins. Others believe ontologies are and should only be built on “settled parts of science” (11). But by creating one ontological compendium—one description of existence—scientists necessarily preclude others.

Many ontology communities routinely update changes in their systems, and recent work in artificial intelligence is beginning to assist ontology evolution by automatically comparing ontologies and designing repair plans that split functions and add arguments to make them commensurable (12). Some ontologies are also beginning to incorporate uncertainty (13). These amendments suggest an emerging interpretation of ontologies in science, not simply as truth statements or data sharing structures, but representations of mental constructs through which we organize our growing understanding about the world.

Computational Reasoning
The culminating step of computational reasoning involves building a reasoner or agent that infers new knowledge from the existing statements of an ontology. Many reasoners use variants of unambiguous proposition or first-order predicate logic to make inferences and prove theorems. Automated theorem proving (ATP) has advanced in recent years and is used intensively in circuit and software design, aerospace and related industries where verifying a particular operation under all possible conditions is critical. The WoframAlpha Computational Knowledge Engine takes a similar approach to computational question answering by using rules to assemble systematic knowledge across a variety of domains and fit it with algorithms to produce on-demand analysis. Another approach takes computational inference still further by using extracted published knowledge to condition models of chemical and biological agents, and then enables these agents to interact in order to simulate and predict higher-level cellular, tissue and organism outcomes.

Scientific disagreement, alternative views of the world, and uncertainty become most consequential in the reasoning process. For example, are three amino acids in sequence a “small molecule” or a “macromolecule”—a very short protein. Crisp logical reasoners have been unleashed on biomedical ontologies in order to extend them, for example, by associating genes with biochemical pathways (14, 15). Logical approaches have been most successful, however, not at generating new knowledge but flagging internal contradictions for repair (16, 17). Because of contradictions and self-references, many of the most used ontologies—like the SNOMED medical ontology and GO—are insufficiently restricted to be expressed with protocols like OWL and cannot take advantage of many of the reasoners available for consistent systems. This has suggested to others the need to incorporate error into the reasoning process.
Early approaches to probabilistic logic dealt with the challenge of inducing support for a proposition from evidence. Belief networks combine proposition probabilities to model uncertainty within a broader domain (18). Belief networks include nodes, which represent variables, interconnected with arcs, which signify probabilistic influences. In biology, a node might constitute the active or inactive state of a gene, and an arc the probability that the gene is active given the state of the protein that regulates it. The strength of influence between nodes is propagated along the graph via forward conditional probability. Bayesian networks are the most commonly implemented form of belief network. They use Bayesian conditioning as the basis for probability updating, and so emphasize the relationship between assumptions and the accuracy of evidence in assessment (19). Consider the use of Bayesian networks to diagnose illness. Given some set of observed symptoms, one can compute probabilities along arcs of the network to compare the likelihood of each disease (20). Several implementations of Bayesian networks have been developed to handle particular types of inference: Dynamic Networks for dynamic systems, Causal Networks for causal processes, and Influence Diagrams that add values and decisions to the belief network in order to optimize choices (21).

Probabilistic reasoners allow researchers to relax the assumption that statements within an ontology or associated knowledge base are certain and universal (22). Instead of using absolute certainty statements like “NCOA3 acetylates histones,” a probabilistic reasoner might hypothetically soften this to “NCOA3 acetylates histones 45% of the time.” This strategy reduces the precision of conclusions, but can reduce the influence of isolated mistakes and make computational inference possible even in the presence of contradiction. In developing ontologies from literature, however, researchers called curators often excise repetitive, contradictory and incommensurable statements. They recognize article statements are unequal, but rather than rank their certainty, curators have tended to censor “uncertain” ones. This is changing. Gene function annotation using the GO now allows for the distinction between experimental findings and structural inference (17).

A few researchers have begun to use probabilistic reasoners on the complete collection of statements extracted from literature: Retaining uncertain statements has allowed them to explore how statements become certain. For example, two recent papers explicitly model the process by which statements and citations in molecular biomedical articles respond to each other in information cascades (23, 24). An information cascade is a chain of collective reasoning that degenerates into repetition (25-27). Figure 2a illustrates this process by showing a sequence of experiments about the same phenomenon (e.g., NCOA3 acetylates histones). The first researcher to investigate the relationship interprets his experiment directly. The second interprets his experiment taking both his own research and the previously published interpretation into account. The third scientist accounts for still more published history and proportionally discounts her findings. This may not appear irrational to the individual scientist. She is acting like a Bayesian statistician: the more prior collective knowledge, the more that prior should influence her interpretation. The process is not collectively rational, however, because the sequence of prior published experiments were not independent. The first had much more influence on the resolution than the last.

A recent investigation into the claim that β amyloid, a protein concentrated in the Alzheimer’s brain, is produced by and injures skeletal muscle in patients with inclusion body myositis (IBM) illustrates this process and its potential costs (24). The relationship was first published in five 1992 and 1993 articles produced by two research groups. Before 1996, six articles critical of the claim were also published, two from the same laboratory that had produced four of the five original supportive articles. This is consistent with the hypothesized “Proteus Phenomena” whereby early findings are subjected to contradiction (29). Nevertheless, two-hundred and forty-two new analyses and reviews came out before 2008 that explored the relationship, and all but one disproportionately cited the positive articles and amplified the claim within the community. Some reviews even made the claim firmer and more general than it had been in its initial context. As consequence of this cascade of support, interpretation of new experiments neither doubted the claim, nor did it explore other possible roles of β amyloid. A very recent paper takes a wider view, and shows that β amyloid may beneficially function as an antimicrobial peptide in the innate immune system (28). This insight raises the possibility that Alzheimer’s is an infectious disease and suggests novel treatment strategies. Despite this new claim’s efficacy, the information cascade surrounding β amyloid’s deleterious effect almost certainly prolonged experimental consideration of its possible beneficial role in Alzheimer’s and immunity.

Building on this work could allow analysts to identify the scope of convergence and divergence processes in biology, chemistry and related fields. Other work has highlighted how research communities are much more aware of some findings—those
appearing in their own journals—than in others published about identical topics (30). Ultimately, reasoning across the distribution of untidy statements could allow us to examine the dynamic nature of research attention in the sciences.

**Incorporating Social Structure and Culture**

The complete distribution of published research statements—even perfectly parsed and analyzed—would still ignore dependencies between statements induced by communication. To produce research, scientists engage in multi-party conversations that span university hallways, workshops, conferences and libraries. Research on the social production of science has begun to trace these linkages using article bylines and acknowledgements to understand how authors and resources organize into teams (31), networks (32), institutions and regions (33) around research problems. In large article collections, author identification has been a challenge, but recent approaches that use many article features are accurate at over 98% even for large article collections like PubMed (34).

Improved models of scientific production could allow reasoning models to statistically deconvolute social, geographic, and funding-related dependencies among published statements (35). For example, offline discussions about “the best” techniques for inferring protein interactions will likely constrain the range of methods deployed. The projection of social, geographic and funding networks onto the network of published statements expose correlations—common repetitions likely induced by the communication of ideas. Figure 2b illustrates this correlation between personal relationships and scientific interpretations. A finding corroborated by articles from three laboratories with no known association, from distant locales using different methods, is much more persuasive than a one repeated by a Ph.D. advisor and his students. Computational models conditioned on social dependencies might weight independent statements with greater confidence.

Incorporating social structure could also improve computational prediction. In a classic analysis, Don Swanson observed that the community studying Raynaud’s disease noted blood viscosity as a common symptom and that the socially disconnected nutrition community published how dietary fish oil reduced blood viscosity. Swanson hypothesized fish oil could be beneficial for Raynaud’s patients and it was found successful in an independent randomized clinical trial (36). Where scientific elements (e.g., blood viscosity) cross the social boundaries between communities, co-occurring elements, problems and solutions from one domain can be connected with those in another. This approach is the scientific equivalent of market arbitrage: Accelerated by computation, it could facilitate “conversations” between contemporary and historical or orphan ideas that were underappreciated in the scientific context of their debut (37).

Beyond social relations, scientists sort into fields with differing knowledge cultures—methods of reasoning, evidentiary standards, and styles of articulation (38). Textual clues provide insight into the patterns that distinguish these cultures. A meta-analysis of oncology papers found that methods sections in those citing industry support were systematically more vague than those citing only government funding (39). Articles and patents, even for the same discovery, are also quite different, with the article emphasizing continuity with previous research and the patent highlighting distinction. Uneven contact between related fields leads to concepts from one domain having imperfect analogs in another. The gene of genetics—the basic unit of heredity—historically had an emphasis and theoretical function distinct from the gene of biochemistry—an encoding segment of DNA—but the two have become mangled with contact. When scientists move between fields or draw from multiple domains, meanings intermix. Established terms attract new meanings over time as the context and concerns of research shift (40). In this way, fields, communities and even individuals within papers host multiple clustered but distinct symbolic systems—their texts draw concepts from a mixture of ontologies.

Following this logic, reasoners could be trained to account for cultural bias. By computationally classifying published statements and estimating the likelihood that each class will enter scientific discourse, reasoners could reweight certainty in those under-represented. For example, the results of successful experiments appear in print much more frequently than unsuccessful ones. This approach could quantify that likelihood and begin to correct for it. It might also enable prediction by suggesting that logically possible but never published negative statements (e.g., human gene miR-96 does not increase susceptibility to heart disease) are more likely than unpublished positive ones in densely crowded research areas, but less so in sparse ones between disciplines where many questions remain unasked.

Reasoners that account for the socio-cultural dispersion of statements could enable us to recover the mix of cognitive ontologies that gave rise to them. This could highlight—and subject to testing—higher-level ontological disputes in science that lie above the level of most ad hoc hypotheses. Theoretical progress is often associated with the reconciliation of ontologies from multiple theories. Consider the merger of
evolution and genetics that precipitated modern evolutionary theory (41) or the common impulse in physics to generate a unified theory. Separating, formalizing, and explicitly comparing the conceptual ontologies that give rise to statements may create novel opportunities to blend ontologies for theoretical experimentation and improvement (12).

Conclusion
The rapidly increasing volume and electronic availability of published science can seem overwhelming to the modern bioscientist. It also poses a unique opportunity. Recent advances in natural language processing, ontology construction, and reasoning models are being brought together by scientists to computationally read and reason. As analysts begin to extract more of the richness from texts, computational reasoners may become capable of modeling certainty and generating predictions based on the full range of factors scientists have always considered, including the socio-cultural processes through which science is produced. Figure 3 diagrams relationships between published science, ontologies, and reasoning in the scientific system. It points to this expansion of features that researchers are beginning to engage in their models of science and their tools to analyze and advance it.

Many obstacles remain, including the need for better models of the production of language and science, more efficient algorithms and faster computation. Nevertheless, computational extraction and reasoning with ontologies have already begun to help scientists overcome some of the limitations of working within a distinct community. By expanding the set of useful documents through text mining, scientists enlarge the distribution from which they can sample ideas and extend the length and potential of their inferences. Reasoning models have allowed scientists to complete paradigms, like working out the genetic details of hypothesized bio-pathways. It is also beginning to enable ontology comparison, which could flag opportunities for theoretical recombination that punctuate scientific advance.

References
10. The authors carried out personal interviews with ten leaders in the field of applied ontology, including Mark Musen, Peter Karp, Barry Smith, Judith Blake, Larry Hunter, Lynette Hirschman, Bob Carpenter, and Jeff Shrager, sampling across the diversity of perspectives.
The mathematics of belief networks are closely related to that of game theory where winning a game parallels the verification of a scientific conclusion, and the uncertainty of competitive moves maps to uncertainty about scientific evidence and assumptions.


This approach can be captured by second-order logic in the principle of bivalence, but cannot exist in first-order statements that demand universal relationships.


We thank all those who agreed to be interviewed for this project (as listed in references) and for their helpful comments on the manuscript, as also John Goldsmith and Ian Foster. We are especially grateful to OCLC for allowing us to use their WorldCat data, ThomsonReuters for allowing us to use their Citation data, and FSO for granting us access to their online access data for the analysis in Figure 1.
Figure 1. (A) Estimated number of distinct pages from the Online Computer Library Center’s *WorldCat* database of books and journals in 71,000 libraries across 121 countries, split by manuscripts and journals, broad subject area, and the most common eight languages from 1450 to present; (B) Manuscript pages, by language, mapped against major historical events of the 20th Century; (C) Distribution of volumes across libraries: Number of volumes plotted against the number of OCLC libraries in which each are held, split by manuscript and serials, subject, and language. All distributions feature a spiking tail suggesting a core collection of books that appears in nearly all libraries; (D) Growth in the number and publication age of journal pages available via the Internet, and freely on the Internet (without institutional subscription), 1998-2006.
Figure 2. (A) A hypothetical temporal sequence of experimental findings (1’s and 0’s in the beakers) and published articles (1’s and 0’s in the papers, see (23)). Early findings are reflected accurately in publications, while scientists’ interpretation of later findings incorporate the history of publication into account; (B) The broader social network in which 2A scientists live. The positive correlation between social ties in 2B and propositional agreement in 2A suggests that communication induces accord.
Figure 3. Elements, context and process of scientific reasoning. Arrows represent causes or influences. The figure emphasizes the scope of first-generation computational reasoning and the emerging, second-generation reasoning we suggest in this review.