Research Challenges and Opportunities in Knowledge Representation

NSF Workshop

February 7-8, 2013

Arlington, VA

Final Workshop Report

Edited by

Natasha Noy, Deborah McGuinness

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-1217638.

This report can be cited as:

"Final Report on the 2013 NSF Workshop on Research Challenges and Opportunities in Knowledge Representation." Natasha Noy and Deborah McGuinness (Eds). National Science Foundation Workshop Report, August 2013. Available from http://krnsfworkshop.cs.illinois.edu/final-workshop-report/KRChallengesAndOpprtunities_FinalReport.pdf

Table of Contents

Workshop Participants

Workshop Chairs

Natasha Noy (chair), Stanford University
Deborah McGuinness (co-chair), Rensselaer Polytechnic Institute
Eyal Amir (co-chair), University of Illinois, Urbana-Champaign

Workshop Participants

Chitta Baral, Arizona State University, US
Michael Beetz, Technical University of Munich, Germany
Sean Bechhofer, University of Manchester, UK
Craig Boutilier, University of Toronto, Canada
Anthony Cohn, University of Leeds, UK
Johan de Kleer, PARC, US
Michel Dumontier, Carleton University, Canada
Tim Finin, University of Maryland, Baltimore County, US
Kenneth Forbus, Northwestern University, US
Lise Getoor, University of Maryland, US
Yolanda Gil, Information Science Institute at University of Southern California, US
Jeff Heflin, Lehigh University, US
Pascal Hitzler, Wright State University, US
Ian Horrocks, Oxford University, UK
Craig Knoblock, Information Science Institute at University of Southern California, US
Henry Kautz, University of Rochester, US
Yuliya Lierler, University of Nebraska, US
Vladimir Lifschitz, University of Texas Austin, US
Peter Patel-Schneider, Nuance, US
Christine Piatko, John Hopkins University, US
Doug Riecken, Columbia University, US
Mark Schildhauer, NCEAS, US

Cognizant Program Officer

Vasant Honavar, National Science Foundation, CISE/IIS

Government Observers

Murray Burke, DARPA
Bonnie Dorr, DARPA
James Donlon, National Science Foundation
Frank Olken, National Science Foundation
Michael Pavel, National Science Foundation

Executive Summary

Modern intelligent systems in every area of science rely critically on knowledge representation and reasoning (KR). The techniques and methods developed by the researchers in knowledge representation and reasoning are key drivers of innovation in computer science; they have led to significant advances in practical applications in a wide range of areas from natural-language processing to robotics to software engineering. Emerging fields such as the semantic web, computational biology, social computing, and many others rely on and contribute to advances in knowledge representation. As the era of "Big Data" evolves, scientists in a broad range of disciplines are increasingly relying on knowledge representation to analyze, aggregate, and process the vast amounts of data and knowledge that today's computational methods generate.

We convened the Workshop on Research Challenges and Opportunities of Knowledge Representation in order to take stock in the past decade of KR research, to analyze where the major challenges are, to identify new opportunities where novel KR research can have major impact, and to determine how we can improve KR education as part of the core Computer Science curriculum. The main outcome of the workshop is a set of recommendations both for KR research and for policy-makers to enable major advancements that will have broad impact in science, technology, and education.

Successes of the past decade

Workshop participants identified remarkable successes of the past decade. Lightweight KR systems got deployed in many large-scale applications outside of academia. IBM's Jeopardy-winning system Watson, Apple's Siri, Google's Knowledge Graph and Facebook Graph Search would not have been possible without the advances that KR researchers have made in the past decade. For the first time, we now have international standards for knowledge representation, developed by the World-Wide Web Consortium. Both researchers in academia and practitioners in industry are widely adopting these standards. Scientists in almost any field today consider formal methods indispensable for dealing with big data. Researchers in robotics, computer vision, and machine learning are beginning to realize the power and new opportunities that they gain by integrating KR techniques into their systems. Finally, theoretical advances in KR have been remarkable, with researchers developing extremely scalable reasoning methods and achieving deep understanding of new and far more expressive models and formalisms.

Areas where we expect considerable advancement

These successes have laid the groundwork for the considerable advances that we expect to see in the next decade. When workshop participants brainstormed what main breakthroughs to expect, we agreed that the potential advances that we can look forward to include the following: The **large-scale data analysis** will enable

scientists to process their big data, and, critically, to extract *knowledge* from the reams of data that they collect. **Real-life question answering** will move to deep natural-language understanding and will enable far more advanced interactions with robots and other computing systems than we do today. **Advances in analytics that drive markets, personalization, and manufacturing** will add knowledge and reasoning to models that engineers use today. We believe that KR will enable **scientific advances** in life sciences, physics, astronomy, and other scientific disciplines. Finally, KR methods have huge potential in **education** by supporting learning and enabling new learning modalities by helping students build arguments and understanding the process of scientific thinking.

Challenges

Naturally, there are many scientific challenges that KR researchers must address to enable these advances. The changing landscape of the past few years enables us to tackle many challenges that appeared unattainable before. We can harness the big data for analytical and learning methods, we can use social mechanisms of bringing together the power of citizen scientists and the crowd that have only recently become available, and we can rely on the computational power available today that allows a robotic device to perform more processing "on board" than before. We hope to use these advances in order to make significant advances in developing hybrid representation and reasoning methods, dealing with heterogeneity at many different levels, capturing knowledge in an entirely new ways. Finally, with the KR methods entering the everyday toolbox of practitioners in today's knowledge-intensive economy, we must make these methods accessible and usable for those who are not experts in KR.

Recommendations and grand challenges

Participants have developed three grand challenges in areas of big data, education, and scientific knowledge processing. We designed these grand challenges in a way that will require significant new advances in knowledge representation and reasoning, but will ground the research in practical applications. Such grounding will provide both the framework for the research and a way to evaluate and measure success.

Participants stressed the need for stronger ties to other communities within computer science. KR researchers can benefit from these ties and also provide a formal framework that will help researchers in such fields as natural-language processing, machine learning, and robotics to advance their research.

Finally, workshop participants agreed that we must highlight the role of KR in computer science curriculum. In today's knowledge-based economy, we need scientists who are comfortable with knowledge representation and semantics. Expanding the curriculum recommendation to address the KR topics explicitly will guide educators on the important topics in KR and information systems.

1 The Workshop Background and Motivation

Written by Eyal Amir, Deborah McGuinness, Natasha Noy

We organized the workshop in order to discuss the new challenges and opportunities that arise from the explosion of data and knowledge, increased reliance of scientists on computational data, its heterogeneity, and new modes of delivering, storing, and representing knowledge. This radical shift in the amount of data, in the way that scientists distribute, store, and aggregate this data, precipitates new challenges for knowledge representation. KR researchers must address scalability of their methods for representation and reasoning on entirely different scale. The distributed and open nature of the data-intensive science requires representation and reasoning about provenance, security, and privacy. The increased adoption of semantic web technologies and the rapid increase in the amounts of structured knowledge that is represented on the semantic web creates its own set of challenges. The increased use of KR methods in computer vision, robotics, and natural-language processing emphasizes the opportunity for practitioners in those fields to affect directions in which KR research proceeds. As all of us get more accustomed to social mechanisms for creating and sharing data, it is incumbent upon the KR researchers to study how these new interaction and knowledge creation paradigms affect the field.

Specifically, the following developments made such discussion particularly timely:

- increased reliance of scientists on knowledge representation methods to "tame" the explosion of big data in a distributed and open world
- increased availability of linked open data and the need to develop principled methods to represent and use this data in applications
- increased need for background knowledge in processing images, video, and natural-language text and for integration tasks in robotics and other fields

The workshop participants included researchers from a wide range of subfields of knowledge representation--from semantic web, to uncertainty reasoning, to robotics. This broad coverage allowed us both to address challenges and opportunities from various KR directions and to represent all points of view from within the KR community. Additionally, the workshop included leading scientists in non-KR fields, such as biomedicine, earth sciences, computational biology, biodiversity, evolutionary biology, physics, and others. We invited scientists who are both experts in these fields and are either experts in KR or are keenly aware of the challenges and opportunities that their fields bring to KR. These scientists had first-hand expertise in what it would mean to address these challenges. Inviting scientists who need KR to solve real-world problems, allowed us to stay focused on these real-world problems and use cases as we discussed the challenges that KR can and should help solve.

Our goal was to discuss the opportunities and challenges in training the new generation of scientists and researchers who can address these research topics: what are the tutorials and workshops that need to be organized? What are the new courses that can be developed that bring together the KR challenges and the real world? Anticipating higher demands for knowledge engineers in the near future, we must design the key courses and curricula to train future workers in the knowledge-based economy and consider the ways to bring our advances into everyday practice.

Report outline

The workshop consisted mostly of interactive brainstorming sessions with the goal of producing a report that outlines the key challenges and opportunities for knowledge representation and reasoning. In the rest of this report, we discuss the highlights, challenges, and results of the workshop. We start by discussing the major success of KR in the past decade. These successes have touched every area of human endeavors and have made great strides in solving difficult reasoning problems (Section 2). We then discuss our vision for the next decade and where we think new advances are likely to happen. Our discussion was firmly grounded in challenging use cases from other fields that must drive KR advances (Section 3). Finally, we discuss why these advances are difficult and what are the major scientific challenges that researchers in knowledge representation and reasoning must address (Section 4). We conclude by sketching out three "grand challenges" that can drive these advances (Section 5) and set of recommendations both to funding and policy bodies and to the KR community (Section 6).

2 Successes of the Past Decade

The past decade has seen advances in knowledge representation and reasoning power systems and fields as diverse as the Space Shuttle operations, media and publishing, smart phones, advances in biology and life sciences, and many others. Indeed, just as advances in Artificial Intelligence (AI) were often not considered to belong to AI once they became part of mainstream systems, many scientists are not aware that they are relying on advances that KR researchers have made. We distinguish three types of KR successes:

1. **Success of KR standardization efforts:** For the first time since the advent of KR research, the last decade saw the development, and, more important, wide acceptance of international standards for describing data and ontologies on the Web and for reasoning with them. These advances lead to broad availability of structured data in standard formats for KR researchers to user and consume.
2. **"KR-Lite" in deployed systems and standards:** The class of what we call "KR-Lite" applications are the applications that use simple knowledge representation and reasoning. These successes include Watson and Siri, Google knowledge graph and Facebook graph search, international

standardization of the Semantic Web technologies and their adoption by the industry, ubiquity of linked data.

3. **Applications of advanced KR methods outside of KR:** This group of successful applications in other subfields of artificial intelligence and outside of academia uses advance knowledge representation and reasoning techniques. These systems include, for example, robots, such as cooking robots that rely on advances in computer vision and reasoning technology. Scientists in disciplines such as biology, medicine, environmental sciences now consider formal ontologies to be indispensable for dealing with big data. The Space Shuttle uses a reasoning system to maneuver the aircraft.
4. **Theoretical and practical achievements within KR:** Finally, this third group of successes are the achievements within the field of KR itself. These achievements include scalable and competent reasoners that can now process orders of magnitude more data in a fraction of the time; the use of satisfiability solvers in areas such as software and hardware verification; the theoretical and practical advances in such paradigms as Answer Set Programming for modeling and solving search and optimization problems.

In the rest of this section, we provide the details on some of these advances.

2.1 Standardization of infrastructure and languages and their wide adoption

Written by Natasha Noy

Formal knowledge representation has entered the mainstream with international standards bodies developing standards for KR languages and many enterprises using these standards to represent their data. Particularly notable are the standards for the Semantic Web languages that were developed by the World-Wide Web Consortium (W3C) and have found wide adoption (Figure 1). These standards include the Resource Description Framework (RDF) and the RDF Schema for representing basic classes, types, and data on the Web. The Web Ontology Language (OWL) is the W3C standard for representing ontological data on the web. These languages define the standards for representing formal models and data in the (Semantic) Web environment. KR researchers have contributed to these standards, and, indeed, have largely driven and informed their development. The OWL specification in particular builds on the rich tradition of knowledge representation in general and, specifically, description logics. OWL provides formal semantics that enable reasoning. The availability and large scale of data and knowledge represented in OWL and RDF has in turn spurred new developments in scalable reasoning (cf. Section 2.4.1).

Finally, SPARQL is another W3C standard—the one for querying RDF data. The W3C SPARQL standard includes not only the syntax for the language, but also formal semantics for several entailment regimes—specifications of formal reasoning mechanisms that SPARQL query engines may support. These standards represent the first and the most significant step in creating international and widely used knowledge-representation standards.

While the existence of such standardization for the first time is in itself a success story, it is its wide adoption by major players in different industries that has really taken lightweight knowledge representation into the mainstream. For example, Google uses "rich snippets," and RDFa representation (RDF embedded in HTML) of the key structures in a web site, to provide structured search for people, products, business, recipes, events, and music. The New York Times publishes its 150-year archive as linked open data. The NY Times "Semantic API" provides access to 10,000 subject headings and metadata on people, organizations, and locations. BBC uses semantic web technology to power many of its web sites; BBC also consumes RDF data published elsewhere on the Semantic Web. BestBuy has been using RDFa to annotate its products and claims to have higher click-through rates from search engines for the annotated products. These are just a few examples of mainstream large industry organizations adopting these standards and improving their bottom line, their offering to customers and their competitive position.

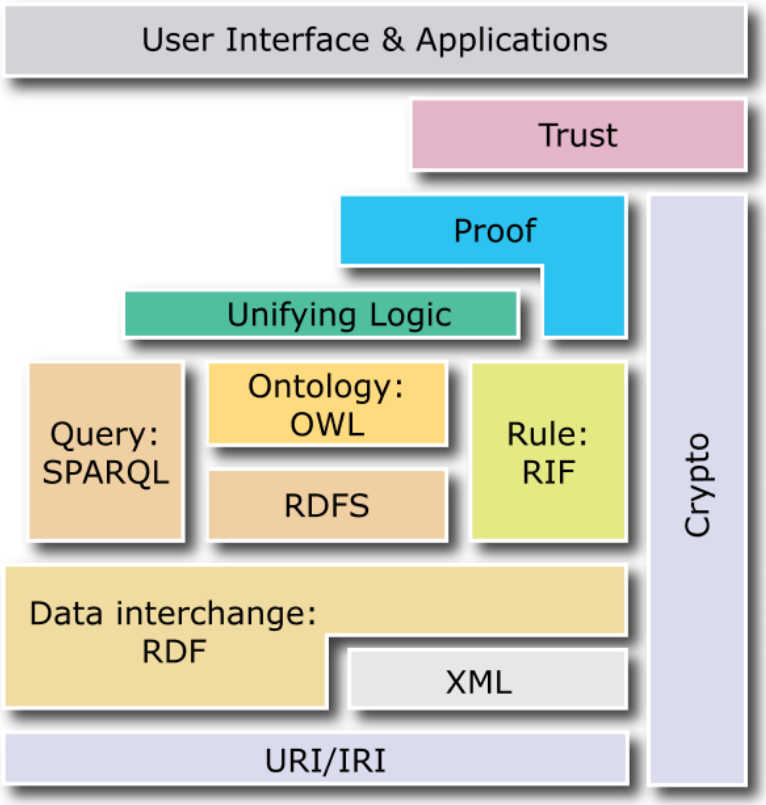

Figure 1. "Semantic Web Layer Cake": The layered architecture of knowledge representation standards adopted by the World-Wide Web Consortium. These standards include RDF for data interchange, RDF Schema and OWL for representing domain models, SPARQL for querying RDF data and RIF for representing rules. Image source: http://www.w3.org/2007/03/layerCake.png (W3C).

2.2 Lightweight KR In Deployed Systems and Standards

One of the most exciting developments over the past decade is seeing the KR technologies becoming an integral part of everyday applications. Many of these KR applications are what workshop participants called "KR Lite": lightweight technologies that came out of the KR communities having big impact on everyday applications outside of academia.

2.2.1 KR-Lite for fielded systems Watson, Siri, Google Knowledge Graph, and others

Written by Deborah McGuinness

Applications are continuing to emerge that use some amount but not necessarily a deep amount of knowledge representation and reasoning. Some of these applications have received a fair amount of usage (e.g., Siri), or a fair amount of coverage (e.g., Watson) or both. Siri is an intelligent assistant that is embedded in the newer Apple phones. Its roots are in the DARPA Personal Assistant that Learns (PAL) program and specifically from the Cognitive Assistant that Learns and Organizes (CALO) project. Within that large sponsored research project, there was much foundational work using deep knowledge representation and reasoning, however the portions of the effort that migrated to Siri appear to use some background representation of knowledge required to handle the primary Siri tasks - knowledge navigation and particular tasks such as making calls, scheduling meetings, etc. Siri does other things such as speech recognition and natural language processing but it does not appear to depend heavily on detailed knowledge representations and sophisticated reasoning. Many would consider it a success for knowledge representation and reasoning since it does show off value of some background knowledge and reasoning to answer a very wide range of questions.

IBM's Watson system is a question answering system that the authors claim requires a synthesis of information retrieval, natural language processing, knowledge representation and reasoning, machine learning, and computer-human interfaces to address its task of providing answers to questions. It achieved wide recognition when it defeated two world-class Jeopardy champions by producing questions in response to answers following Jeopardy rules. It builds on research in many areas of artificial intelligence and has foundations in several government-sponsored research programs such as the Novel Intelligence for Massive Data (NIMD) program by Advanced Research Development Activity (ARDA). As part of NIMD, IBM, Stanford, Battelle and other institutions used IBM's unstructured information management architecture (UIMA) to take unstructured text and extract facts that would be used along with background ontologies and reasoners to answer questions - in that setting for intelligence analysts. In the Watson setting, researchers performed significant additional work and the enhanced question answering system covered much broader domains and again used text analytic techniques, knowledge models and reasoning to answer questions. One of the claims of this system is that it integrates shallow and deep knowledge. The AI Magazine

article provides a number of examples where it uses knowledge representation and reasoning to do things such as help with scoring and disambiguation for example.

2.2.2 Open government data

Written by Deborah McGuinness

Many governments are making data more broadly available. High profile administrations such as those in the United States and the United Kingdom, just to name a few, are publishing large amounts of government funded data and at the same time requiring many organizations (typically government funded organizations) to make their data available online. Further administrations are requiring government organizations to identify "high value datasets". This is not just happening in large first world countries, but countries all over the world are putting their data online at staggering rates. As a result the Linked Open Data cloud is growing significantly (Figure 2). There is a proliferation of sites such the United States' data.gov effort and the UK's data.gov.uk effort where anyone can obtain an increasing amount of data. Many academic organizations are partnering with the leaders in open government data not only to create portals such as the RPI's site or Southhampon's site, but also to add a large array of tools to ingest, manipulate, visualize, and explore open linked data are appearing.

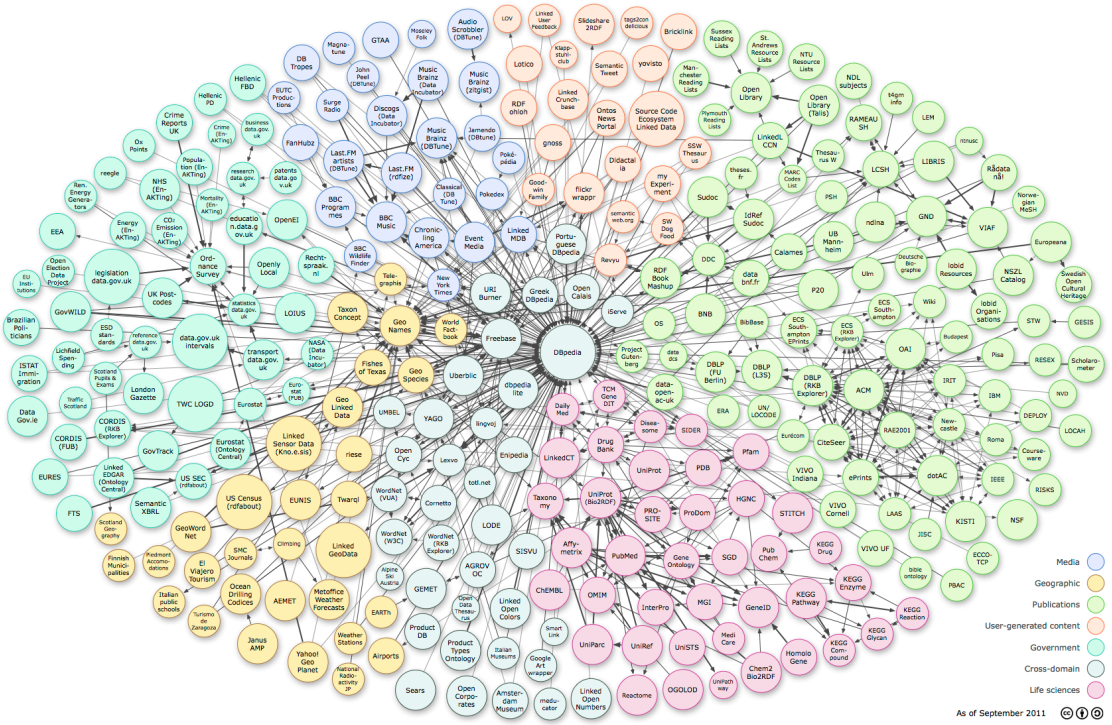

Figure 2. The Linked Open Data cloud: The collection of publicly available datasets on the Semantic Web. This collection has grown from only 12 datasets in 2007 to 295 in 2011 (this diagram). These 295 datasets comprise 31 billion triples. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

The linked open data world is also generating new success stories as well as new challenges. It is not uncommon to find a billion of formally expressed facts (triples) in open data endpoints. With the enhanced availability and access has come the need for knowledge representation environments that can represent and reason with a broad range of data and can also function at scale.

2.3 Applications of Advanced KR Methods

In addition to "KR-Lite" appearing in many industrial applications, a number of applications have taken full advantage of recent advances in scalable reasoning capabilities.

2.3.1 Ontologies and big data in science

Written by Natasha Noy

In biology and biomedical informatics today scientists cannot imagine managing their data without widely adopted ontologies, such as the Gene Ontology (GO). GO is a collaboratively developed ontology for annotating genes and gene products across species. At of March 2013, GO contains almost 40,000 classes and more than 60,000 relationships describing biological processes pertinent to the functioning of cells, tissues, organs, and organisms; components of cells and their environment; and molecular functions of gene products. Scientists use GO widely to aggregate and analyze their data. Today, there are hundreds of thousands of manually and automatically created GO annotations for dozens of species, with 345,000 annotations for *Homo Sapiens* alone.

In recent years, scientists in many other disciplines have started developing ontologies as they see these artifacts as indispensable components in their pipelines to process big data (Figure 3). Indeed, we consider this pull from disciplines such as biology, environmental sciences, health informatics, and many others as another success of KR. Rather than viewing KR with skepticism, organizers of such meetings as the annual meetings of the American Geophysical Union (AGU) or American Chemical Society (ACS) schedule special sessions on semantics. Most major initiatives supported by the National Science Foundation (e.g., EarthCube for earth and geosciences, DataONE for environmental sciences, iPlant for plant biology) have working groups focusing on semantic technologies and all rely on ontologies to represent their data and services.

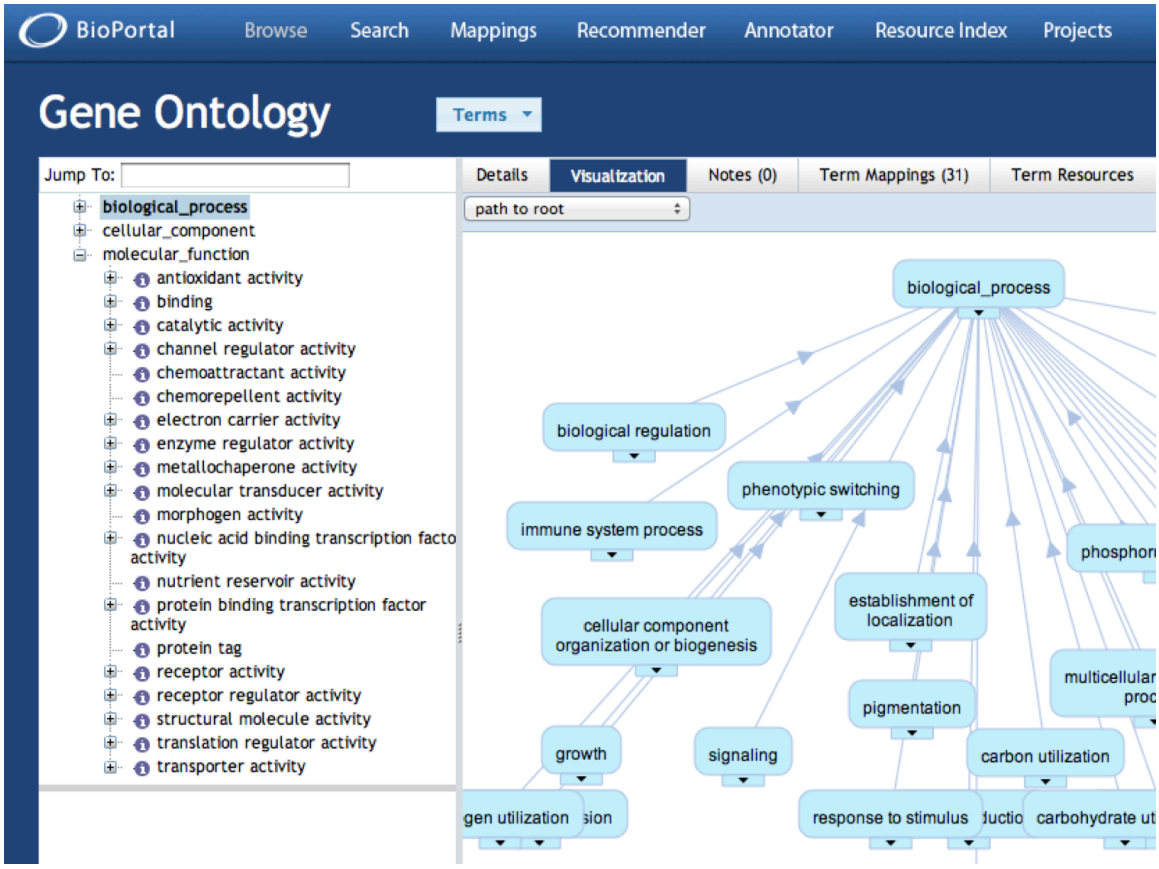

Figure 3. The Gene Ontology in the NCBO BioPortal. Scientists have developed hundreds of ontologies to represent concepts in their domains. There are more than 350 publicly available ontologies in the National Center for Biomedical Ontology BioPortal. There are more than 5 million terms in the ontologies in BioPortal—in the biomedical domain alone. The Gene Ontology, shown in the figure, is used widely for aggregation and analysis of data.

2.3.2 Applications based on formal models

Written by Yulia Lierler

For many applications, both inside and outside AI, KR provides the mechanism to define formally reasonable and desirable behaviors of agents and systems. Impact of those formal KR models is greatest in three main categories: AI-Planning (including Robotics), Natural-Language Processing (NLP), and Diagnosis of physical systems and processes.

In the last decade AI Planning has looked for breakthroughs in KR as guides for building new planners with extended capabilities. That research built on the well-understood connection between formal models of action and efficient automated planners. The RCS/USA-Advisor for the Reaction Control System (RCS) in the Space Shuttle is an example of such breakthrough KR advance applied in the context of real-world application. The RCS system is responsible for maneuvering the aircraft while it is in space. The RCS/USA-Advisor is a part of a decision support system for

shuttle controllers. It is based on a reasoning system and a user interface. The reasoning system is capable of checking correctness of plans and finding plans for the operation of the RCS. This application would not be possible without advances made in the 1990s and early 2000s in formal models of action.

Also in the last decade, breakthroughs in the natural language processing (NLP) have built on formal models developed in the 1990s and early 2000s. Application systems, such as Nutcracker (Balduccini et al., 2008), that understand narratives and instructions in the context of formal descriptions of the world applied established (circa 2001) formal models of action, logical representation, and taxonomy. Without those formal models the natural-language understanding tasks of question answering and entailment would be confined to statistical methods such as bag-of-words which provide poor performance. Most important, current established methods in NLP, such as semantic parsing, would not exist in their current form without their KR foundations.

Finally, many current federal-government efforts would be inefficient or suffer failures if KR-based diagnosis systems were not established in the 1990s and the past decade. For example, efforts by the DOE to overhaul the Nuclear Reactor diagnosis systems (e.g. Argonne National Lab's PRODIAG) would be impossible without advances in KR-based Diagnosis. Formal models developed in the 1990s and 2000s have shown how to scale models of physical systems. Without those, model-based diagnosis would not exist today, and diagnosis of many vital systems would be impossible.

2.4 Theoretical and practical advances within KR

In this section, we highlight the key advances within the KR field itself that have successfully tackled problems that appeared intractable before.

2.4.1 Availability of scalable and competent reasoners

Written by Peter Patel-Schneider and Ian Horrocks

Formal reasoning is a complex task. Reasoning has high worst-case computational complexity or is undecidable in many common representation languages. In the past, reasoning systems had to be limited to small examples or carefully controlled so that they did not consume excess computational resources.

Over the last decade or so, reasoners that work effectively in many or most cases of reasonable size have been developed for many representation languages. These reasoners have been made possible by new ways of thinking about reasoning (such as reasoners that are sound but not complete), by new theoretical algorithms for reasoning, by combining optimizations initially from different kinds of reasoners, by better implementation techniques, and by increases in processing speed and main memory size. Currently, competent reasoners (such as Glucose (Audemard and Simon, 2009)) exist for hard SAT problems with thousands of variables, for simple

ontologies with hundreds of thousands of concepts (Kazakov et al., 2011), and for complex ontologies with thousands of concepts (Motik et al., 2009). Researchers have developed methods to handle successfully even first-order problems of reasonable size. New techniques for scalable reasoning include reductions to simpler kinds of problems. For example, reasoning in ontology languages with limited expressive power can be reduced to querying over relational databases.

Many of the advances in reasoning were stimulated by competitions, such as the annual competitions between first-order reasoners, the competition for propositional modal reasoners, and the SAT competitions.

This is not to say that reasoners can successfully process any problem in a representation language. However, for particular representation languages it is no longer necessary to maintain close control of reasoners, even for large problems.

2.4.2 Advances in satisfiability and answer set programming

Written by Yulia Lierler

Declarative problem solving is another area of significant algorithmic and representation advances in the past decade. The best example in this area is Answer Set Programming (ASP, for short). Answer Set Programming (Brewka et al., 2011) is a declarative programming paradigm stemming from knowledge representation and reasoning formalism based on the answer set semantics of logic programs. Answer set programming offers a simple, yet powerful, modeling language for optimization and search problems. It is particularly useful in solving search problems where the goal is to find a solution among a finite, but very large, number of possibilities. Problems of this kind are encountered in many areas of science and technology. Typically, determining whether such a problem is solvable is NP-hard. Indeed, answer set programming has close connections to another prominent field of knowledge representation—satisfiability (Gomes et al., 2008). Satisfiability and answer set programming in the past decade have seen ever faster computational tools, and a growing list of successful practical applications. For example, satisfiability solvers are used as general purpose tools in areas such as software and hardware verification, automatic test pattern generation, planning, and scheduling (Gomes et al., 2008). Advances in algorithmic techniques developed for satisfiability then enable advances in other areas of automated reasoning including answer set programming, satisfiability modulo theory, first order model building, constraint programming. At the same time, answer set programming is increasingly leaving its mark in tackling applications in science, humanities, and industry (Brewka et al., 2011).

3 What Can KR Do for You? The Application Pull

Knowledge representation will play a key role in assuring success in many of the challenges that the United States faces in its data- and knowledge-driven economy in the coming decade. Extracting knowledge from data, creating new knowledge-driven applications, and generating new expressive knowledge will likely lead to advances in many areas. Almost any domain that has any data to process into knowledge will benefit from advances in KR. Indeed, KR is already being applied in biomedicine, health care and life sciences, oil and gas industry and sustainable energy, engineering, open government initiatives, earth and environmental sciences, defense, autonomous robotics, education, digital humanities, social sciences (census and decision making), museums and cultural collections, finance, defense, material and geosciences, and personal assistants.

In these fields, KR methods underpin information management and retrieval, data analysis and analytics, machine learning, processing of sensor data, agents and multi-agent collaboration, representation of engineering systems, natural language processing and understanding, representation of preferences, human-human and human-machine collaboration (human augmentation).

We have collected several use cases and challenges from these different areas in order to highlight the opportunities that KR provides.

3.1 Scientific discovery

With scientists producing ever increasing volumes of data, they must go from the "big data" to knowledge and scientific insights. The KR methods provide representation formalisms to describe the data, common ontologies to share these description, mechanisms for formulating and processing complex queries over heterogeneous sources, methods to overcome heterogeneity and variety of data, approaches for both cognitive scalability in understanding the data and scalability of reasoning over the increasing volumes of data, and formalisms to describe provenance of the data and its context.

3.1.1 Use case: environmental sustainability

Written by Mark Schildauer with edits from Yolanda Gil and Deborah McGuinness

Consider one Grand Challenge science question: "How will Climate Change impact the sustainability of the world's ecosystems"?

This Grand Challenge question requires clarifying influences and interactions among a number of processes that are traditionally the focus of distinct disciplines: coupling complex, multi-scale models from earth, atmosphere, hydro, and ocean domains representing processes that almost certainly have complex feedback loops; and integrating these with models and data that factor in human dimensions as well.

Data sources range from industrial reports of energy consumption and emissions from burning fossil fuels, to time-series of global land-use coverage from remote-sensed images; to a wealth of on-the-ground measurements representing observations and measurements from distributed, uncoordinated researchers and sensors, as well as systematic monitoring efforts such as the nascent NEON program.

The semantic challenges here are clear and prevalent. First, the semantic challenge of assisting with the discovery and integration of highly heterogeneous data-- representing an incredibly diverse set of fundamental measurements of earth features and phenomena, taken at many resolutions across a range of spatial scales, using multiple methodologies, and preserved in a variety of informatics frameworks, ranging from relatively unstructured spreadsheets on local hard drives to larger, well-modeled databases, none of which however, offer consistent semantics for interoperability. Semantic technologies can help tame terminological idiosyncrasies that currently abound within earth science domains-- ranging from non-standardized use of terms that are often context- or discipline-dependent, to imprecise terms, and a wealth of synonyms, hypernyms and hyponyms that are used in uncoordinated and unreferenced ways that severely compromise the ability to discover, interpret, and re-use information with scientific rigor. Knowledge representation techniques can motivate the development and use of standardized terminologies for the Earth Sciences, with obvious advantages of helping to unify and disambiguate semantic intention.

Earth Science researchers employ a number of different statistical and modeling approaches to investigate and predict a huge range of natural phenomena. KR techniques can greatly enhance the comparability and re-use of analyses, models, and workflows, by clarifying the semantic dimensions for appropriate inputs, providing for more nuanced interpretation of the outputs, and clarifying how these components are linked. By deploying best practices in ontology construction, KR techniques can enable far more than simple terminological harmonization, through advanced inference capabilities possible through the use of logically rich vocabularies processed by increasingly powerful reasoners. Ontologies additionally can lead to stronger community convergence and interoperability, via standardization in the construction of models, and through the promotion of rigorous specifications of model inputs and outputs that can lead to greater efficiency in data collection efforts. In addition, the logical expressivity of modern KR languages, especially when implemented in accordance with emerging Web standards, enable expression of detailed provenance information and other metadata, that are increasingly important in determining suitability for use-- of data as well as analytical results or other products. Finally by constructing community-based, cross-disciplinary ontologies, KR methods can escalate prospects for trans-disciplinary communication, by reducing semantic ambiguity when results are reported in the literature, or the broader applicability of findings are discussed or potentially used to support policy.

With the continued pace of anthropogenic energy use, land transformation, and resource extraction activities, integrated earth science investigations are becoming critical to inform society about how to sustain basic human needs-- for adequate food, water, shelter, and clean air-- for ourselves and future generations. Planktonic life in the ocean, and the world's great forests absorb massive amounts of carbon from the atmosphere, and help offset human emissions of carbon into the atmosphere; but these systems are currently undergoing rapid changes in function and extent. In this industrial age, we must be able to understand the impacts that human activity can and will have on the earth system, and especially how our current activities might impact future prospects for human viability and quality of life. Added to this are concerns about less easily quantifiable concerns, such as preserving the world's rich biodiversity, e.g. coral reefs and areas of untrammeled forest, or even having places where penguins, elephants, salmon, and tigers can exist in the wild. The KR&R community can assist the earth sciences at this critical time, by helping the field to better organize and adapt its data and modeling resources, as well as its communication of results, to a digital, networked information environment. KR solutions will be prime enablers for Grand Challenge questions in the Earth Sciences, where they will not only accelerate our understanding of complex, interlinked phenomena, but also help inform critical policy decisions that will impact environmental sustainability in the future.

3.1.2 Use case: Biomedical and pharmaceutical research

Written by Michel Dumontier

Advances in biomedical and pharmaceutical research are built on prior knowledge, and require easy and effective access to information that is buried in scientific publications or in small and large partially annotated datasets. Significant effort is currently spent to curate articles into simple facts that provide insight into component functionality. Similarly, much work goes into massaging data from an arbitrary collection of formats into a common format and then cleaning, integrating and consolidating data into meaningful information. A major aspect of modern scientific data management to create useful data for query answering and analysis lies in the use of ontologies to create machine-understandable representations of knowledge. With hundreds of ontologies now available for semantic annotation and formal knowledge representation, there are new opportunities and challenges for biomedical and pharmaceutical research and discovery.

The most recognized use of ontology in biomedical research is enrichment analysis. The goal of enrichment analysis is to find a set of attributes that are significantly enriched in a target set over some background set also sharing that attribute. With over 30,000 terms and millions of genes and proteins annotated in terms of functions, localization and biological processes, the Gene Ontology has been used to bring insight into thousands of scientific experiments. While new research bears the plethora of ontologies to the automatic annotation and enrichment of text-based descriptions such as scientific articles, scientists uncover new associations between

previously unlinked entities. However, while such experiments are relatively easy to perform, a major outstanding challenge lies in being able to reconcile these associations with prior knowledge and establishing the degree to which we are confident about any assertion found in a web of data in which broad scientific claims must be reconciled with experimental facts arising from specific methodologies executed over model systems. Clearly more research must be directed towards accumulating evidence to provide plausibility, confidence and explanation in the face of incompleteness, uncertainty or contradiction.

3.1.3 Use case: Advancing healthcare

Written by Natasha Noy

Continuous and large-scale analysis of data will lead to new insights in many areas. For example, analysis of shared medical profiles may shed light on drug safety and efficacy that exceeds the power of expensive clinical trials. Recently, the web site PatientsLikeMe enabled patients with amyotrophic lateral sclerosis (ALS) to organize a self-reported clinical trial (Wicks et al., 2011). We will need KR in order to aggregate information, and to match automatically patients and patients to clinical trials. Similarly, we will be able to address the challenges of personalized medicine, using knowledge representation and reasoning to develop personalized treatment plans, identify individuals with similar rare diseases, leverage data from tests, literature and common practice. The IBM Watson team is moving in that direction already.

3.2 Education

Written by Kenneth Forbus

One of the major success stories of AI and Cognitive Science has been the rise of intelligent tutoring systems. Intelligent tutoring systems and learning environments incorporate formally represented models of the domain and skills to be learned. Such systems have already been shown to be valuable educationally in a variety of domains, such as learning algebra, and are currently used by over a half-million students in the United States every year. The potential for such systems to revolutionize education, by offering anytime, anywhere feedback has been recognized in prior NSF studies. Such automatic methods of providing interactive feedback offer benefits across all types of education, ranging from traditional classrooms to massive open on-line courses. A key bottleneck in creating such systems is the availability of formally represented domain knowledge. Moving into supporting STEM (Science, Technology, Engineering, and Mathematics) learning more broadly will require new kinds of intelligent tutoring systems. For example, helping students learn to build up arguments from evidence, thereby understanding the *process* of scientific thinking, not just its results, requires a broad understanding of the everyday world, the informal models students bring to instruction, and the likely trajectories of conceptual change. Commonsense knowledge is both important

for interacting with people via natural language, and because many student misconceptions are based on everyday experience and analogies with systems encountered in everyday life. Intelligent tutoring systems, fueled by substantial knowledge bases that incorporate both specialist knowledge and commonsense knowledge, could revolutionize education.

3.3 Robotics, sensors, computer vision

3.3.1 Household robots

Written by Michael Beetz, Leslie Park Kaebling

From the early days of KR, the control of robotic agents has been a key motivating topic—if not the holy grail—of Artificial Intelligence research. In AI-based robot control, plans are sequences of actions to achieve a given goal. A robot reasons about which actions to execute in which order and abstracts away from *how* to execute the actions.

In recent years we have seen a number of robotic agents performing human-scale everyday manipulation activities such as cleaning an apartment, making salad, popcorn, and pancakes, folding towels, and so on. The physical capabilities of robots have recently made huge strides. The actuators are reasonably safe and reliable and the sensing is sufficiently accurate to recognize known objects in somewhat complex arrangements.

Thus, it is time to reconsider the role of knowledge representation and reasoning in representing and reasoning about actions and environment. There are three critical areas: methods for representing knowledge, methods for updating the robot's internal knowledge representation based on percepts and actions (*belief-state update*), and methods for planning, execution, and execution monitoring (*action selection*). Additionally, we can consider the problem of learning, which is often distinguished from belief-state update because the knowledge being acquired or updated may be more abstract or variable over a longer time scale.

Reasoning about actions: Robotic agents cannot perform everyday activities as vague as "clean up," "set the table," and "prepare a meal" without comprehensive knowledge-processing capabilities. Thus, we have to investigate and develop knowledge processing methods that, given a vague task, are capable of inferring the information needed to do the appropriate action to the appropriate object with the appropriate tools in the appropriate way. If robotic agents are to be that competent, their reasoning must not stop at actions such as pick up an object. Even for a simple action such as pick up an object, the robot has to decide where to stand, which hand(s) to use, how to reach for the object, which grasp type to apply, where to place the fingers, how much grasp force to apply, how much lift force to apply. These decisions are context- and task-dependent. How to grasp a bottle might depend on whether I want to fill a glass with it or whether I want to put it away. How to grasp a

glass might depend on whether or not it is filled. If knowledge processing does not reason about these aspects of robot activity, it misses great opportunities for having substantial impact on robot performance.

Uncertainty: There are three major challenges for operating in a domain where, for example, a household robot operates: a mixed continuous- discrete state space, the dimensionality of which is unbounded, substantial uncertainty about the current state of the world and about the effects of actions, and very long time-horizons for planning. For example, to find pickles in the back of a refrigerator, a robot must do some combination of moving objects (possibly removing them temporarily, or pushing them aside to get a better view and selecting viewpoints for look operations). The robot needs a representation of its belief about the current state of the refrigerator, such as what objects are likely to occur in the refrigerator and where they are likely to occur. All of these actions ultimately take place in real, continuous space, and must be selected based on the robot's current belief state.

3.3.2 Understanding spatial and spatio-temporal data

Written by Anthony Cohn

Cameras and other spatially located sensors, such as GPS transceiver in mobile agents, produce enormous volumes of spatial data. There are many applications that require sophisticated understanding of such data, or can be usefully augmented with it, from surveillance, to mobile assistance, to environmental monitoring. The use of qualitative spatial representations not only provides some relief from both the volume and noisiness of such data, but also enables integration of different kinds of spatial knowledge (topological, orientation, size, distance, etc.).

Understanding visual data has been a challenge for the Computer Vision community for decades, and whilst much progress has been made in methods which attempt to understand such data using continuous/numerical techniques, it is only recently that interest has (re)started in trying to extract symbolic high-level representations from video data. Because video data is inherently noisy (e.g. owing to changing lighting conditions) the high variance in the presentation of activities visually, extracting symbolic hypotheses is highly challenging. The challenge is made that much harder by the sheer volume of data, both already "out there on the web" or acquired in real time; but on the other hand this sheer volume of *Big Data* also provides mitigation for the problem since there is often redundancy (e.g. through multiple kinds of sensors, or spatially overlapping sensors). Another form of mitigation can come in the form of background knowledge about how the world is, and how activities progress, so as to help understand missing data, correct noisy data, and to help integrate and fuse conflicting data. In turn, this brings the challenge of where such background knowledge comes from, and in particular whether it all has to be specified manually, or whether it can be automatically acquired, through data mining or machine learning techniques.

Another opportunity in this area is to combine data acquired from sensors with language data – consider for example a cooking show, and the commentary from the chef and other people on screen and the visual images of the ingredients being prepared; there is some temporal synchronicity here, but it is not perfect; there is extra information in both data streams and "superfluous" information (e.g. where they first tasted this particular ingredient).

3.4 From Text To Knowledge

Written by Chitta Baral

Question answering systems are systems that answer questions with respect to a collection of documents, where the questions are often in natural language and the documents include text and may include other forms of information such as video and web pages. Question-answering systems are useful in many domains, including analysis of intelligence documents, answering questions with respect to medical transcripts, answering questions with respect to research literature, looking for answers from past law cases and finding answers from patent databases.

At the top level, question answering involves understanding questions, understanding text and other forms of information, and formulating answers. Knowledge Representation and reasoning plays important roles in each of these steps. Thus, there is a significant opportunity for KR in the building of question-answering systems.

Understanding questions and text is mainly about natural language understanding. Building natural language understanding systems involves translating text to a KR formalism, augmenting it with various kinds of knowledge that include common-sense knowledge, domain knowledge and linguistic knowledge; and reasoning with all of them to come up with components of answers that then need to be glued together to form answers.

In formulating answers to questions, for certain kinds of questions straightforward database operations such as joins are sufficient. But for many other kinds of questions such as "Why", "How" and "What-If" questions one needs to first formalize what answers to such questions are. Researchers have done some formalizations of this kind in some contexts, such as answers to prediction, explanation, diagnostic, and counterfactual questions with respect to simple action domains. We need to do a lot more. For example, in a biological domain, the knowledge involves a combination of (a) ontological information about entities, their components, their properties, classes and sub-classes (b) and events and sub-events, when then can happen and their impact. Answering "Why", "How" and "What-If" questions in such a domain remains a challenge.

3.5 Why KR?

What does knowledge representation bring to these challenges? Figure 4 summarizes some of the key contributions of formal knowledge representation for the applications that we highlighted throughout this section. Reasoning and inferring new facts is the first natural contribution. The lightweight KR, which can provide simple hierarchical inference (which is already used in Watson and Google Knowledge graph) is already a win in itself. Query expansion and query answering is another form of reasoning that provides knowledge that may not have been stated implicitly in response to a query. Ontologies and other formal descriptions of the domains can inform the natural-language understanding tools about the semantics of the domain of discourse. KR languages can serve as a logic form for the target representation of the results of natural-language understanding. Knowledge representation provides a "lingua franca" for integrating diverse resources. For example, ontology-based data access uses ontologies as an entry point to access many different databases (Calvanese et al., 2011). For software agents and robots, KR provides a flexible approach to represent information and to discover implicit information.

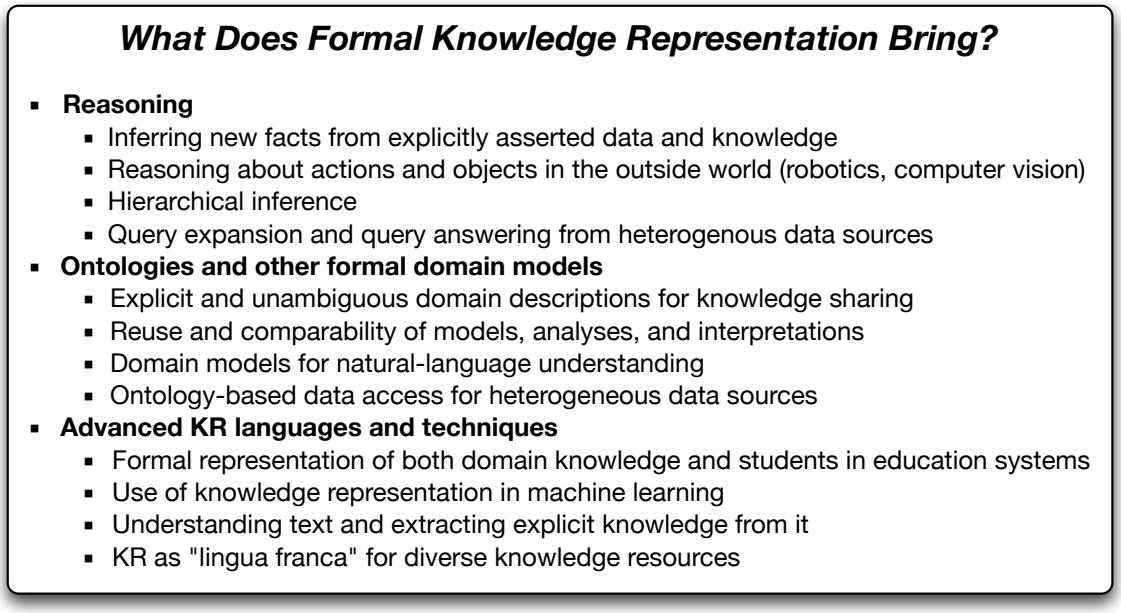

Figure 4. Summary of key elements that formal knowledge representation brings to the applications in various fields. The applications described throughout Section 3 rely on these elements of formal knowledge representation to enable applications in a variety of domains.

4 Why is it difficult? Challenges for the KR Community

The application areas that we highlighted in the previous section present both opportunities and challenges to KR researchers. We can group these challenges along the following four dimensions:

1. KR languages and reasoning
2. Dealing with heterogeneity of knowledge
3. Knowledge Capture
4. Making KR accessible to non-experts

Addressing these four challenges will enable us to make significant strides in addressing the opportunities in other domains. Indeed, many of the opportunities in the previous section rely on solutions to the same challenges (e.g., representing uncertainty, capturing knowledge, combining different types of reasoning).

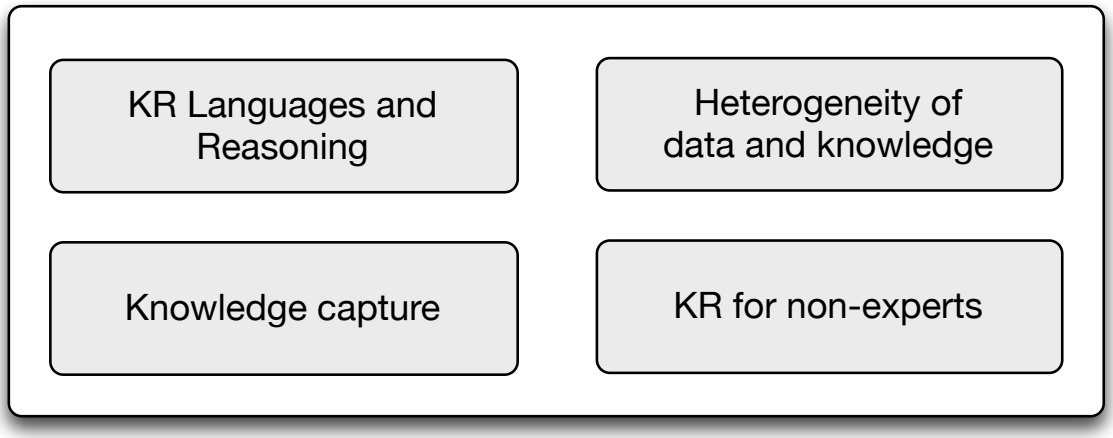

Figure 5. The key areas of research in KR for the next decade. Throughout Section 4, we discuss the key research challenges in these four areas of research. The areas focus both on the challenges in the representation and reasoning per se, as well as the use of KR methods by non-KR experts and knowledge capture from text, from experts, and from novel sources.

4.1 KR Languages and Reasoning

Written by Peter Patel-Schneider

Knowledge representation languages and reasoning methods are naturally at the core of the KR research. KR languages enable engineers to describe their domains formally, with clear semantics. Over the years, scientists have developed many different representation languages for the effective representation of different kinds of information---propositional logics for boolean combinations of atomic facts, first-order logic for general quantified information, Horn rules for particular kinds of inference, modal logics for contexts, temporal logics for time, description logics for ontologies, graphical languages for relationships between objects, probabilistic

logics for non-boolean information, nonmonotonic logics for overridable information, and so on. We have built and optimized reasoning engines for these languages, often with effective performance on problems of moderate size or complexity. Other kinds of languages and techniques have also been developed for storing or transforming information, for example databases for storing and accessing large numbers of simple facts and statistical and related methods for detecting commonalities in large amounts of information.

Today, these trade-offs become easier to manage, as researchers develop competent reasoners for increasingly complex formalisms. The community is coalescing around knowledge-representation standards and semantics for complex reasoning tasks. There is greater availability of data that we can "lift" into knowledge, thus both informing our approaches and applying them in practice. The key challenge today is finding the right balance between more complex formalisms and the lightweight KR. But more critical is the task of integrating different formalisms and approaches to develop hybrid approaches that get the "best of all worlds". In the rest of this section we highlight these key challenges that the researchers will need to address in the coming decade.

4.1.1 Hybrid KR

Written by Peter Patel-Schneider, Yulia Lierler

Using one particular language limits us to the problems that we can effectively represent (and reason with) in that language. Thus, if we want to gain some or all of the benefits of multiple languages, we must try to combine languages, for example combining description logics and temporal logics to represent changing ontologies, or combining rules and databases to permit simple reasoning over large numbers of facts. However, combining two languages often results in an increase in the complexity of reasoning. For example, combining description logics and rules in the obvious manner generally results in a language with undecidable reasoning, even though both components have decidable reasoning. Developing systematic means for combining (a) heterogeneous KR languages and (b) various reasoning techniques under one roof is by no means a solved issue.

There are some recent initial successes in this direction. For example, advances in satisfiability modulo theories (Barrett et al., 2009) and constraint answer set programming (Brewka et al., 2011) demonstrate a potential for this direction of research. For instance, constraint programming (Rossi et al., 2008) is an efficient tool for solving scheduling problems, whereas answer set programming (Brewka et al., 2011) is effective in addressing elaborate planning domains. Constraint answer set programming that unifies these two KR sub-fields is best for solving problems that require both scheduling and planning capabilities of underlying tools.

Similarly, Description Logic Rules combine description logics and rules but limits the scope of the rules to obtain decidable reasoning. In this way, we can obtain most

of the benefits of the two (or more) languages, while still retaining the desirable features of the component languages. We need to perform this analysis for each combination of languages—a formidably difficult task.

We can also consider producing a loose combination, where the two languages exist mostly independently, with separate reasoners, communicating via some sort of lingua franca or common sub-language and using some sort of intermediary to translate between or otherwise control the separate reasoners. This sort of loose combination can also be used to combine several reasoners over the same language, where the reasoners handle, or are complete or effective on, different but potentially overlapping sub-languages. Scientists have used these loose combinations for quite some time, starting with blackboard systems and continuing up to modern performance systems like Watson (Ferrucci et al., 2010). Nevertheless, many problems remain in producing such combination systems, ranging from issues of allocating resources to issues related to characterization of the capabilities of the combination.

Bridging open-world knowledge and closed-world data

Written by Pascal Hitzler

Some major knowledge representation languages, such as those around the Web Ontology Language OWL, which is based on description logics, adhere to the open-world assumption, which appears to be appropriate for many application contexts such as the Semantic Web. If a system uses an open-world assumption, it commonly assumes that a statement is *true*, unless it has information to conclude otherwise. For instance, if we do not know the temperature under which a specific sample was collected, our system can assume a interpretation with *any* temperature value might be correct. However, we usually implicitly understand database content as adhering to the closed world assumption. In our example, any specific statement of the temperature that we cannot infer from the data will be false. Using content that adheres to the closed-world assumption together with open-world knowledge bases can thus easily lead to undesired effects in systems utilizing deductive reasoning.

In order to avoid such effects, we need to develop practically useful languages and reasoning algorithms that combine open-world and closed-world features – this kind of combination is known as local-closed-world modeling. Scientists have recently made some advances in this respect (the recent paper (Knorr et al., 2012) provides an entry point to the state of the art), in particularly driven by borrowing from the field of non-monotonic reasoning, which is closely related to the closed-world assumption. The proposed languages, however, are usually rather unwieldy for application purposes, and arguably attempt to address the closed-world data access problem by means which are too sophisticated for the problem at hand, and thus seem to make unnecessarily strong demands on resources used for knowledge modeling, reasoning, or knowledge base maintenance.

The KR community will need to address this issue by developing simple, intuitive, light-weight solutions for local closed-world knowledge representation and reasoning.

Bridging KR and Machine Learning

Written by Pascal Hitzler, Lise Getoor

In the age of big data and information overload, there is a growing need for research which bridges work in knowledge representation and machine learning. As the data available becomes richer and more intricately structured, machine learning (ML) research needs rich knowledge representations that can capture information about the structure in the data, the data sources and other important aspects that affect quality. ML research also need models and hypotheses that are complex and can represent different types of objects, their relationships, and how these may change over time.

At the same time, research in knowledge representation and knowledge acquisition can benefit from newly emerging machine learning methods which discover hidden or latent-structure in data. From topic modeling, which discovers a single latent variable, to richer statistical relational models that can discover hidden relations and hierarchical models as well, these structure discovery methods can be useful bottom-up approaches to the acquisition of new knowledge.

And, most importantly, there is the opportunity to close the loop between data-driven knowledge discovery and knowledge-based theory refinement: by using richer knowledge representation languages to be able to search over the space of features used in a machine learning algorithm to discover new structures which can be added into the knowledge representation and used in further structure and knowledge discovery.

In order to close the loop, we need systems that can mix logic and probabilities, and perform a mix of deductive and statistical reasoning. Emerging research subareas such as statistical relational learning and probabilistic logic programming are promising directions, which aim to make use of both statistical and logical representations and reasoning methods.

Mixing data and simulations

Written by Lise Getoor

There is a growing need for methods that can exploit and integrate data that is produced from the simulation of physical models and observational data that is gathered from sensors and other data collection methods. In many cases, different people build the models and collect the data; they do it at different time, make different modeling assumptions, use different languages, operate at different times scales. Nonetheless, the ability to fuse the information from these different models

and to make more informed decisions based on both models and data is important. Indeed, as observational data becomes increasingly accessible and diverse, we need to address new challenges of fusing this data with models to extract the knowledge. For example, many of the ecological models often describe a single species. Before, if one had data, it would be for just a single species as well. Now, we can get population data from cameras and other sensors about co-occurrence of species. Integrating this data with ecological models will provide new insights on the interaction of species and effects of the environment.

There is a great need for modeling languages that can handle multiple models and data in a robust, extendable and interpretable manner. Such languages will have applications in climate modeling, environmental modeling, power grids, ecological models, manufacturing systems, health and medical systems.

4.1.2 Representing inconsistency, uncertainty, and incompleteness

Written by Chitta Baral, Natasha Noy

A number of recent developments give rise to the growing body of knowledge bases that contain incomplete or inconsistent knowledge. These knowledge bases include the knowledge bases that are acquired automatically, at least in part, or built by a distributed community of users. These knowledge bases include linguistic knowledge bases such as WordNet, Verbnet, and FrameNet and world knowledge bases such as ConceptNet, Google Knowledge Graph, DBpedia, and Freebase. Similarly, the body of knowledge created by a distributed, often uncoordinated "crowd." This knowledge is inevitably inconsistent and incomplete. Users have different contexts and views and represent knowledge at different levels of abstraction. The knowledge that we extract from the "big data" that is produced by scientists may also appear inconsistent or incomplete---for example, because we do not have the provenance metadata that could explain the differences in measurements. Developing reasoners that will perform and *scale* robustly given incomplete and inconsistent knowledge is one of the key challenges today.

4.1.3 Challenges in reasoning

Written by Peter Patel-Schneider

The worst-case computational complexity of complete reasoning is dismal for even representation languages of moderate expressive power, such as the W3C OWL Web Ontology Language. For languages of higher expressive power, reasoning is undecidable. Nevertheless, there are many reasoning systems for various languages, including propositional logic, OWL (W3C OWL Working Group, 2009), rules, constraint satisfaction, and Datalog variants, that have good expected resource consumption in most cases or even almost all cases. Even with these successes, improving expected performance remains a vital issue in reasoning, for example building first-order reasoners that can perform well over the amounts of

information required to support general common-sense reasoning in broad domains.

Robust reasoning

Written by Peter Patel-Schneider

The current level of reasoner performance is often not adequate, for example in server-based applications. Instead, we need robust scalable reasoning, i.e., little resource consumption on all natural inputs, even when reasoning over large ontologies or large numbers of rules and in the presence of very many facts. There are some reasoning systems on languages of limited expressive power that approach this level of performance, for example, many RDF systems (such as Sesame (Bock et al., 2007)) and systems that reason over limited description logics (Calvanese et al., 2010). A major goal is to improve the performance of these systems to rival that of the fastest storage and querying systems. Another major goal is to provide similar levels of performance for more expressive languages.

Parallel computation and distributed storage can help in improving the performance of reasoners, but do not in themselves provide a complete solution. We must pay careful attention to all aspects of performance, including not just wall-clock time but also communication costs, memory footprint, and the effect of memory hierarchies. Parallel reasoning in expressive languages is a difficult problem due to the sophisticated centralized control needed for effective performance.

Effective human-scale reasoning

Written by Kenneth Forbus

One of the major differences between today's AI reasoning systems and human reasoning is that human reasoning tends to become more efficient and effective as people learn more. AI systems, in contrast, typically require careful hand-crafted optimization to achieve high performance. For example, while IBM's Watson used knowledge gleaned by reading, the processing pipeline for both learning and question-answering was carefully constructed and crafted by human designers for the Jeopardy! task. Understanding the space of reasoning tasks and architectures that handle them optimally is one interesting research question. Another interesting research question is how to achieve the desirable properties of human reasoning in software. Human reasoning is robust, flexible, and operates over broad domains of knowledge. Understanding how to create software that operates similarly would be revolutionary, both in terms of our scientific understanding of human cognition and in terms of economic benefits. Some promising approaches currently being explored include partitioning knowledge (Amir and McIlraith, 2005), parallel processing (Urbani et al., 2012), and analogical processing .(Forbus et al., 2009)

4.1.4 Lightweight KR

Written by Pascal Hitzler

We use the term "Lightweight" Knowledge Representation to refer to methods and solutions that have low expressivity, including class hierarchies, fragments of RDFS, or logic-based languages of polynomial or lower time complexity. The recently increased focus on lightweight KR is driven both by theoretical advances concerning such languages, and by application successes of the likes of IBM's Watson, Apple's Siri, or Google's Knowledge Graph.

This development stands somewhat in contrast to the highly expressive logics that KR researchers often investigate. Dealing with lightweight paradigms thus poses new questions that we must address. In particular, we need the principled development of lightweight languages, algorithms and tools from both an application perspective and a theoretical angle. At the same time, we need viable pathways for bootstrapping light-weight solutions in order to move to more powerful use of formal semantics and automated reasoning. In particular, to help with the uptake of heavier-weight solutions, it would be very helpful to develop knowledge modeling languages and interfaces which have a light-weight appearance, e.g. through the use of modeling patterns and macros, while reasoning algorithms in the background may actually be deep and involved to meet application needs.

4.2 Dealing with heterogeneity of data and knowledge

One of the biggest challenges—and opportunities—for KR lies in integrating heterogeneous data and knowledge sources. Heterogeneity comes in many different forms:

- heterogeneity of knowledge models, such as ontologies, vocabularies, levels of abstraction, accuracy, etc.
- heterogeneity of data and information artifacts, in syntax (xls vs xml vs Unicode), in structure (table vs csv vs vector), in semantics (e.g., actual measurements from sensors)
- heterogeneity of data items (e.g., different ids for the same data objects)
- dynamic data and models that change over time
- data acquired from rapidly proliferating sensors.

In many cases, integrating diverse objects or data and knowledge sources results in whole that is larger that the sum of its parts. We can gain valuable insights by integrating data produced by different scientific experiments or by bringing together observations from different species. Robots integrate diverse instructions and inputs--that may lead to new actions or instruct the robot to acquire additional knowledge.

Scientists work on different modes of integrating heterogeneous data and models, from tight coupling and integration to the loose integration of only those components that the task requires. While solving the heterogeneity problem remains a holy grail of KR research and still poses many challenges (see the rest of this section), a number of recent developments present new opportunities that make us hopeful that we can make significant progress in the coming years:

- *New incentives to share data*: Many government programs now mandate sharing of data. It is becoming more common to get academic credit from data citations. Sharing data results in more collaborative and integrative opportunities.
- *Crowdsourcing technology:* Systems like Freebase, DBpedia (via Wikipedia) and (soon) Wikidata allow people to manually integrate their own information into a shared model.
- *Better tools* -- increasing ease with which researchers and citizens can contribute to global knowledge bases.
- *Increasing capabilities of KR backends*: Moore's law, cheap storage, parallel hardware and software, better reasoners, allow us to scale to billions of triples. Only in past few years has storage, bandwidth, and computational power become sufficient to enable rapid, effective data-sharing, enhancing possibilities for collaborative efforts
- *Potential is becoming acknowledged.* KR is the most promising way to identify and document (for community-based sharing) *objects* and *events* identified in others' models and analyses run on Big Data cloud. Scientists increasingly acknowledge the need for robust "semantic annotation" of model outcomes and images that enable interrogation of resources *across systems.*
- *Big data* renders useful both high precision, low recall approaches and low accuracy techniques useful in many cases

These new developments will help us address integration challenges both at the model (ontology) and data level. We summarize the challenges at different levels of representation—from raw data to formal ontologies—in Figure 6.

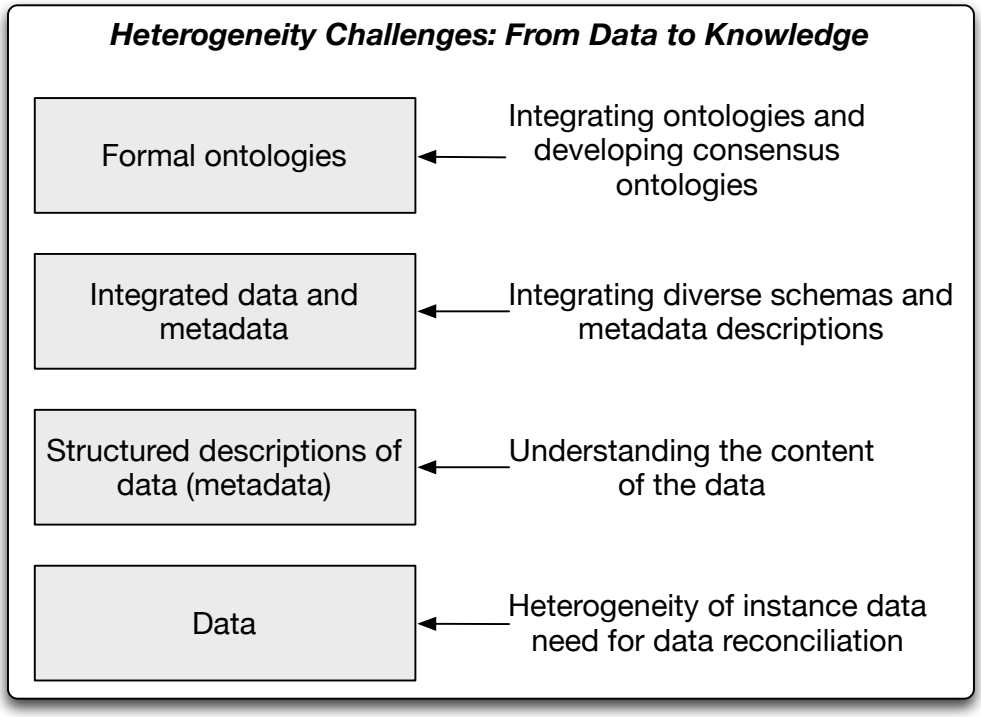

Figure 6. Heterogeneity challenges at a variety of levels between data and formal knowledge. Section 4.2 discusses challenges and possible approaches in dealing with heterogeneity at various levels, from raw data to metadata, to formal ontologies.

4.2.1 Closing the Knowledge--Data Representation Gap

Written by Craig Knoblock

The KR community has developed sophisticated languages and ontologies for representing the knowledge in diverse subjects, yet the amount of data that is actually represented in a KR system continues to shrink as an overall percentage of data available. Consider just the growing body of data available as part of the Linked Open Data cloud (Bizer et al., 2009). While this information is published in RDF, much of the data is published in RDF using only the schema of the original data source so there is no useful semantic description of the data. While there are rich ontologies for some of the Linked Data sources, these are the exceptions and not the rule. Then there is the rest of the Web, which provides the majority of the available data. On the Web, the data and services are available in any of a variety of formats and there is no attempt to provide any semantic description of the data at all. The challenge and the opportunity are to bring the rich set of KR languages and ontologies to the vast amount of data that is available today.

Solving the knowledge–data representation gap will lead to huge advances in our ability to exploit diverse sources of knowledge. Consider the domain of biology where there are huge investments in research, equipment, and data collection. The

ability to find and reuse data is extremely limited because it is a largely manual process to find, understand, and use data generated by other researchers. But if all of the data within this domain were published and described with respect to shared domain ontologies, then researchers could quickly discover relevant data sources and then exploit this knowledge to more effectively conduct their research.

Closing this gap requires developing new methods, tools, and incentives to represent the huge amount of data that is available today. The core research problems are as follows:

- **Automatic Modeling:** we need methods to build semantic descriptions of the growing amount of data that is being produced. Given there is already a huge amount of data that lacks the semantic metadata that describes it, we need automatic methods to support the semantic description of this legacy data (Parundekar et al., 2012).
- **Data Transformation:** we need methods that can quickly and easily (and perhaps automatically) transform data between alternative representations since different representations of the same data are often needed for different purposes (Noy, 2004).
- **Data Linking:** we need to go beyond simply understanding data sources at the schema level, we also need to understand how information is linked at the data level. As such, we need tools to support the automatic or semi-automatic linking of data across sources (Bizer et al., 2009).
- **Source Publication:** given that the amount of data is so vast and dispersed and the knowledge of what it contains is highly distributed, we need easy-to-use open-source tools that enable the users of the data sources to describe their own datasets (Taheriyan et al., 2012).
- **Incentives:** Finally, we need incentives in the form of an immediate return in the time and effort invested in publishing semantic descriptions of data to encourage the use of such tools. These incentives could be in the form of useful software tools that provide capabilities that are enabled by the semantic descriptions of the sources.

By bringing knowledge representation techniques and tools to the data and services that are already being published on the Web, we have the opportunity to start a revolution in representing, discovering, and exploiting the vast amount of data available today.

4.2.2 Heterogeneity: The Ontology Perspective

Written by Jeff Heflin

The flexibility of KR languages makes them well-suited for describing a diverse collection of ontologies. We can use axioms to explicitly specify the relationships between terms from different ontologies, or we can define the terms using common vocabularies and infer the relationships between them. Of course integration may

be manual, automated or some combination of the two. There has been significant progress on automated ontology alignment, but the vast majority of the approaches only produce subclass or equivalent class alignments. However, real-world heterogeneity often requires complex axioms to resolve, and requires the use of negation and disjunction among other things. Furthermore, concepts not typically found in KR languages, such as arithmetic to perform unit conversions or string manipulations (e.g., to map *fullName* to *firstName* and *lastName*) are necessary to achieve practical integration. Can these conversions be learned?

Noise and quality become critical issues when considering multiple ontologies and data sources. When there are multiple ontologies for a given domain, how can we determine which are of the highest quality and which are the best fit for a given modeling problem? How do we decide which data sources appear to be the most complete and contain the fewest errors? If we determine that there is an error in the conclusions reached by an integrated KR system, how do we debug it? Can we automate data cleaning, and in what way does data cleaning for KR differ from data cleaning in databases? If data is contradictory (whether due to errors, untruths, timeliness, or different perspectives), how can useful conclusions be drawn from the data? Can we bridge the gap between logical approaches and human ability to handle noise?

Another integration question is essentially the centralized vs. distributed storage model. The vast majority of Semantic Web systems are centralized. These have the advantage of being able to answer queries quickly, but require significant disk space and are only as current (i.e., fresh) as the last time data was crawled. The biggest issues facing these systems are how to continue to scale them, e.g., by parallelism, and how to perform truth-maintenance in a scalable way, since most are forward-chaining reasoners. More recently, there has been work on federated query systems. These systems attempt to use multiple distributed, knowledge bases (e.g., RDF files or SPARQL end points) to answer questions. Such systems have the advantage that they can provide fresh results, do not need massive local storage, and are not subject to legal or policy restrictions on storage of data. However, this comes at the price of high latency. The main questions for these systems include: what is an appropriate indexing mechanism for storage? Should it be manually created and provided by site owners, or can it be produced automatically via crawling? What is the optimal abstraction of the data source that should be stored in the index? Alternatively, should discovery be completely dynamic and rely on no index at all? How can we reduce the number of sources that must be polled, and thereby reduce both bandwidth usage and total response time? How does distributed query optimization interact with inference? To what extent does the topology of ontologies impact query optimization strategies?

4.2.3 Developing consensus ontologies

Written by Mark Schildhauer

The lack of relevant cyberinfrastructure to enable community-wide uses and benefits in KR is one of the most significant current impediment to innovations in KR and associated reasoning methodologies to enable a new age of discovery and interoperability for integrative natural sciences. Specifically, outside of the genomics community, ontologies and other controlled vocabularies are not well established, nor accepted and validated by the research community for most fields. For example, in earth sciences, there are many modest to significant vocabularies that have been constructed, yet closer examination reveals these are often inconsistent with one another, with incompatible axiomatic structures, displaying disciplinary quirks in representation if not outright errors, critical gaps in content, and typically having unclear or simplistic inferencing utility. For the earth sciences, a dedicated, community-based ontology construction effort is desperately needed, that allows for researcher input and vetting, working in close conjunction with KR experts, with a commitment to backward compatibility so that current investments will be useable, though not necessarily as powerful, as KR languages and reasoning engines continue to improve. We believe that a similar situation exists in other sciences, and earth sciences are just one example of a community requiring shared consensus ontologies.

4.3 Knowledge capture

Intro written by Yolanda Gil

Significant trends in recent years offer new opportunities to advance knowledge capture:

1) **The availability of people to contribute significant amounts of knowledge**. People are volunteering their expertise and time to contribute meaningful knowledge to on-line repositories. Web sites populated by volunteers collect encyclopedic knowledge, how-to knowledge, travel advice, product recommendations, etc. This knowledge is not necessarily in structured form, but it is in digital form and it is continuously growing. In addition, the very large numbers of contributors provide enough scale to aggregate information and to use redundancy to improve the quality of the knowledge that we collect. A variety of citizen-scientists projects demonstrated that contributors can carry out sophisticated tasks if the system offers an appropriate framework to contribute their knowledge or skills. If we develop the right interfaces and incentives, we will be able to capture vast amounts of structured knowledge from volunteer contributors.

2) **The continuously improving performance of text extraction approaches.** Today's text-extraction systems are growing ever more sophisticated. We can fine-tune them to extract particular kinds of knowledge from text: entities, properties,

events, etc. Although their quality varies depending on the extraction target, we can use text-extraction systems in practice for a variety of applications. Moreover, question-answering systems that extract answers directly from text have made significant advances, as demonstrated by the Watson system. With further improvements in the performance of these systems, text extraction could become a broadly used approach to knowledge capture.

3) **The availability of data at unprecedented scale enabling the discovery of new knowledge.** Automated algorithms have demonstrated the extraction of useful patterns from data. Advanced algorithms for extracting complex patterns and knowledge from large datasets could be key to mining "big data".

4) **The widespread use of sensors and other cyberphysical systems that collect continuous and detailed data about dynamic phenomena.** These systems can observe and collect data about physical entities and processes over long periods of time that can be mined to develop new models of the world and ground knowledge on those models.

However, with these new opportunities come the new challenges in understanding how best to use these novel and promising forms of knowledge capture.

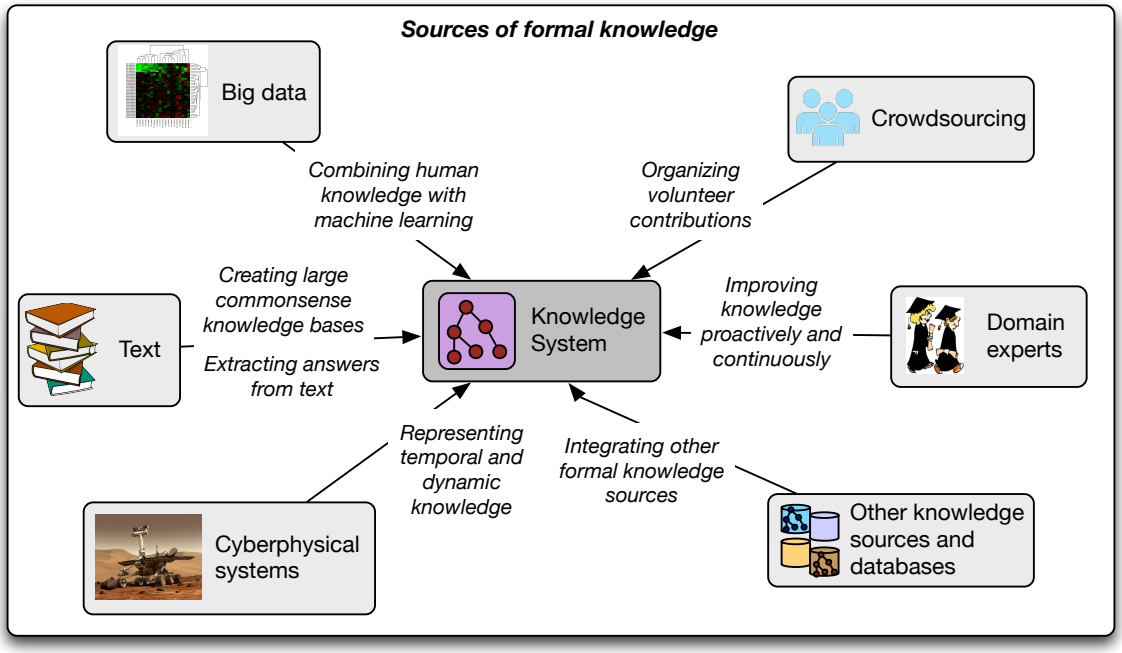

Figure 7. Knowledge capture. Section 4.3 discusses challenges and opportunities in capturing knowledge from diverse and novel sources..

4.3.1 Social knowledge collection

Written by Yolanda Gil

There are many challenges in the social acquisition of knowledge. What kinds of knowledge can we collect effectively in a crowdsourcing collaborative way? What are appropriate knowledge acquisition tasks that contributors can handle? How can people detect and correct misconceptions in a knowledge base? How can systems learn from several people who are providing overlapping and perhaps incompatible or even contradictory information? What are the most effective editorial processes to organize contributors? What training may be needed to support advance forms of knowledge acquisition? What mechanisms can be used to validate contributions? What are successful ways to reach and recruit potential contributors to maintain a reasonable community over time? What are the right incentives and rewards to retain contributors? What are appropriate mechanisms to manage updates and changes?

In current approaches, the systems are quite passive and the contributors largely manage contents and extensions to the knowledge base. We need further research in order to enable the knowledge collection framework to take a more active role in guiding the acquisition process. We will need significant advances in meta-reasoning architectures to assess missing knowledge, to estimate confidence on what is known, and to design strategies to seek new knowledge.

We foresee that knowledge repositories are likely to be interconnected and draw from knowledge that has been collected from different groups of contributors. For example, a repository of genomics knowledge and a repository of biodiversity knowledge could be interconnected to relate genomic information to specific species. The provenance of knowledge sources will be crucial to propagate updates throughout the knowledge bases and to assess trust and resolve conflicting views.

4.3.2 Acquiring Knowledge from people

Written by Yolanda Gil

We need intelligent systems that can acquire knowledge from people, whether new ways to do tasks or simply people's preferences for how the system should behave. Acquiring knowledge directly from people will always be a necessary skill for intelligent systems, even if they are able to acquire much of their knowledge through machine learning approaches.

Key research questions in this area include: How can people extend the knowledge in a system? How can people understand what a system has learned on its own and help it to extend that knowledge? How can people correct misconceptions in a knowledge system? How can intelligent systems learn from several people who are providing overlapping information?

4.3.3 Capturing knowledge from text

Written by Chitta Baral

Extracting relational facts from text has a long history where researchers have used methods based on manually encoded patterns, machine learned patterns and combinations of both. Most of these methods, however, require that we fix the relations *a priori*. In many domains, the same text may have information about a variety of relations. For example, various kinds of biological relations such as protein-protein interactions, gene-disease relationships, gene-drug relationships, gene-variant relationships, drug-mutation relationships and so on can be extracted from a collection of biological text. We need novel methods that can extract arbitrary relations, perhaps when the user specifies the relation as a query "on the fly."

An even bigger challenge in capturing knowledge from text, is to go beyond extraction of relational facts and to obtain more general information. Addressing this challenge will essentially involve translating text to a knowledge representation formalism. Such translation is necessary in many applications such as in developing a system (a) that can understand commands and directives given to it in natural language (e.g., robots in human--robot interaction), (b) that can compare the correctness of students' answers with respect to gold standard answers in an intelligent tutoring system, (c) that can read statements about a scenario and hypothesize about missing information as needed in helping make discoveries using known information and unexplainable observations, and (d) that can answer questions based on the system's understanding of text.

Making significant advances in capturing knowledge from text will also require developing KR formalisms that are particularly well suited for knowledge extraction from text, such as formalisms with temporal and dynamic logic connectives or nonmonotonic formalism. The choice of a particular formalism may depend on the type of text that we are processing and we need to develop a general methodology to find the appropriate formalism.

4.3.4 Building large commonsense knowledge bases

Written by Kenneth Forbus

One of the important lessons from artificial intelligence and cognitive science research is that human commonsense reasoning rests on a vast accumulation of knowledge. This knowledge ranges from high-level abstractions (e.g. concepts of number) to concrete, everyday facts (e.g. that water flows downhill). Our broad base of experience enables us to quickly ascertain when things do and don't make sense. Without such knowledge, for example, question-answering programs can give nonsensical answers. Endowing software with these same reasoning abilities is important for overcoming brittleness, making them more autonomous, and facilitating trust in their operations. Creating large commonsense knowledge bases,

since they can be used across multiple systems for multiple purposes, provides a new knowledge infrastructure for cyber and cyber-physical systems. Thus understanding what is needed in large commonsense knowledge bases, how to build them to human-scale, and how to use and maintain them effectively become key challenges for the community. This section discusses each challenge in turn.

Kinds of knowledge needed: The point of commonsense knowledge is that it can be used for many tasks. Simple kinds of questions can be answered directly, and everyday knowledge provides the background needed to describe situations and frame problems for professional reasoning. For example, engineering teams requires the full panoply of their technical knowledge to design an e-reader that fits in a jacket pocket, but it is their everyday knowledge that informs them about how large jacket pockets tend to be. Experience to date indicates that a wide range of knowledge is needed, ranging from abstract ontological frameworks to masses of concrete, specific information. However, we are still working to understand the representation and reasoning requirements for different tasks. For example, IBM's Watson showed that structured, relational representations led to factoid Q/A performance that far surpassed what was possible with purely statistical, word-based representations. Interestingly, Watson's representations were also very shallow, encoding the contents of particular sentences at linguistic levels. By contrast, a Northwestern learning by reading experiment (text plus sketches) showed that deeper Cyc-based representations were useful for answering textbook problems (Lockwood and Forbus, 2009). More experimentation with large-scale systems that integrate rich knowledge resources, high-performance reasoning, and learning at scale are needed.

Building human-scale large knowledge bases: We have learned much from efforts to build large knowledge bases by hand, and those efforts have provided useful resources for the research community (e.g., OpenMind, ResearchCyc and OpenCyc). However, building beyond where we are now requires continuing and expanding the movement to automatic and semi-automatic learning already underway. For example, learning by reading (Barker et al., 2007, Carlson et al., 2010) is one promising approach. Other modalities, such as sketch understanding, vision, and robotics, are also reaching the point to where they can be used to accumulate everyday knowledge. No matter what the modality, crowdsourcing, ranging from web-based games to trained volunteers and hobbyists, can be harnessed to provide both raw information and feedback.

Maintaining human-scale knowledge bases: No real-world process of constructing large-scale artifacts is perfect, and errors are an inevitable. For human-scale knowledge bases, the software itself must become an active curator of its knowledge. This includes monitoring its own performance, identifying problems and gaps, and taking proactive steps to repair and improve its knowledge and reasoning abilities. For example, in Learning Reader's rumination process the system asked itself questions off-line, and the reasoning it performed in trying to answer them improved subsequent interactive Q/A performance (Forbus et al.,

2007), and NELL uses statistical methods to evaluate the quality of the relations it extracts from the web (Carlson et al., 2010). Understanding how to put most of the burden of maintenance onto the software itself, albeit with human oversight for trust, is an important question.

4.3.5 Knowledge discovery from big data

Written by Yolanda Gil

There is a long tradition in Artificial Intelligence of extracting valuable knowledge from data. The approaches range from explanation-based learning based on applying knowledge to describe examples, to pattern extraction from large amounts of data. Automated techniques to extract knowledge from data have always been valuable, but they become crucial when dealing with large and complex data. In many complex domains, embracing "big data" will expose the limitations of automated methods, which often lack deep knowledge that humans possess or their insight to pose the right questions in the first place. Science is an example of such a complex domain, a mixture of data-rich but also knowledge-rich problems. Automated algorithms can discover new patterns, but those patterns must be related to current scientific knowledge and models. A crucial area of research is how to effectively combine human knowledge with automated algorithms so that their separate strengths can be mutually magnified. How can systems effectively assist people to formulate appropriate questions and design problems and features? How can the space of possible hypotheses be designed so people can direct algorithms in what they believe are promising areas of the search space? How can algorithms effectively communicate their discoveries to people? How can people turn their domain knowledge and expertise into effective guidance for the system? How can a system help users get insights into a problem? How can we design a tighter loop between autonomous exploration and the reasoning that people do to set up the system for the next exploration cycle? This area of research will enable discoveries that will otherwise be out of reach in knowledge-rich "big data" problems.

4.4 Making KR accessible to non-experts

We have argued that KR brings huge benefits to scientists and practitioners in many fields. Indeed, many of them are turning to structured representation of data and knowledge as museums and media companies publish their data as Linked Open Data, and scientists develop ontologies in many domains. Yet, the entry barrier to KR is very high. Today, it is impossible for someone who is not familiar with KR to build an ontology and to use it to explore their datasets "in an afternoon." They might know that there is a potential huge win for them, but there is no place to start.

Enabling non-experts to use KR tools is a two-fold challenge: First, we need to provide tools and recipes and examples to enable them to turn their data to knowledge easily and to see at least the initial benefits quickly ("KR in an

afternoon"). Visualizing and exploring the massive quantities of data that are becoming available is another critical challenge.

4.4.1 KR in the afternoon

Written by Sean Bechhofer

Recent work has seen the development of standards for knowledge representation languages, in particular web-based representation such as RDF, RDF(S), OWL and associated technologies such as SPARQL. This development has been accompanied by the development of an ecosystem of tools for creating and manipulating representations, including editors (Protégé, Topbraid Composer, NeOn Toolkit, etc.), APIs (OWL-API, Jena, etc.) and a number of reasoning engines (Pellet, FaCT++, Hermit, etc.).

While these tools exist, there is a lack of introductory materials that would introduce novice users to the potential benefits of using such representations. If we consider analogies of text processing or machine learning, tools often come with simple example applications that allow a user to quickly explore the technologies ("in an afternoon"). For example, UIMA comes with a number of test scripts that allow the user to run a simple document analysis example "out of the box". Such packaging tends to be absent with KR tools. For example, a user who downloads Protégé can spend time developing an ontology, but then lacks a suitable example application within which the ontology can be deployed and the benefits of descriptive modeling, classification, inference etc. observed.

We need simple, prototypical examples that will illustrate the benefits of KR, as well as the features of the languages and their usage in applications. A possibility would be to exploit the increasing number of resources being exposed as Linked Data (itself an activity that has been facilitated by the presence of standardized infrastructure). For example, cultural heritage organizations such as the Smithsonian have been publishing collection information as Linked Data. This data can then be hooked through to other informational resources such as the New York Times or the BBC. We can build sample applications around small subsets of these data collections, for example encouraging users to build a small ontology that they can then map to those sources.

Note that the intention here would not be to provide materials that teach KR from the ground up (this would be a somewhat ambitious aim in an afternoon), but to provide motivating examples as to what one can *do* with KR.

4.4.2 Visualization and data exploration

Written by Jeff Heflin

A significant challenge for knowledge representation in the twenty-first century is enabling ordinary users to investigate the data. It is obvious that the majority of

users will never learn sophisticated logical formalisms, nor will they learn query languages like SPARQL. Consider SQL as an example. It is the query language for relational databases that is designed for developers' use. These developers then design application-specific interfaces that have various widgets to allow users to express specific kinds of queries. Although this approach might be sufficient for very specific applications of KR, one of the promises of KR is to integrate diverse data from different domains and allow serendipitous discoveries. This discovery is not possible with pre-defined queries.

How can we develop approaches for querying and exploring data regardless of the ontology or ontologies that describe it? Many KR approaches, including RDF, have graph models, however as Karger and schraefel (Schraefel and Karger, 2006) argue, "big fat graphs" are not good for displaying large amounts of data, distinguishing between different kinds of nodes, or grouping things in ways that are intuitive for humans. Furthermore, queries are limited to looking for specific nodes/links to focus on or creating a query-by-example subgraph. Another alternative is natural language query interfaces. However, natural language database query systems have been around since the 70s, and yet the technology is still not in common use. The advent of systems like Siri is encouraging, but Siri is limited to a selected set of sources and does not handle unexpected queries well. Perhaps the semantics available in KR will lead to higher accuracies and overcome the limitations of these systems, but one should consider other alternatives in case progress is slow. Even if we had perfect natural language query technology, it would still be difficult for the user to inspect an unfamiliar knowledge base. Access to the ontology does not ensure that users understand how the ontology is used, or to what extent it is populated with instances. The KB may be sparse in some areas and dense in others. We need approaches that allow users to get high-level views of the data and drill down to inspect details once interesting relationships are discovered. The knowledge base should enable views that allow the level of inference to be adjusted, so that users can evaluate the degree to which incorrect knowledge has impact on the system.

For the KR research community, there is a question of how to evaluate visualization contributions. KR experts are often not familiar with the evaluation approaches generally accepted by the user interface (UI) community. Such experiments can be more costly and difficult to set up than running a system against a standard benchmark, and are often avoided. Can some middle ground be reached, or how can the KR community be encouraged to learn and practice UI experimental methodology?

5 Grand Challenges

We propose a number of grand challenges that can both drive KR research and provide significant advances in other fields. We describe three challenges in detail: analyzing big data; improving STEM Education; and capturing scientific knowledge.

Each of these challenges can serve as a basis for solicitations or larger programs to drive research.

5.1 Grand Challenge: From Big Data to Knowledge

Written by Pascal Hitzler

In this report, we gave many examples of "big data." Big data comes from scientists who make their data and experiment results public, from sensors on robots, from data being collected on the Web. We believe that transforming this data into *knowledge* both poses a great opportunity and frames new challenges for knowledge representation research. Many of these challenges either did not exist in the past at all or existed at a completely different scale. We highlight some of the main aspects:

- **Scalable algorithms:** Big Data requires scalable algorithms to a new order of magnitude. In particular, reasoning algorithms have to be developed which can process data and knowledge bases in real-time. Parallelized, shared-memory algorithms are also required, as well as distributed reasoning capabilities on distributed memory. Techniques for partitioning knowledge and thus partitioning reasoning may also be needed.
- **Processing of data streams:** Big Data includes high-volume data streams, such as those coming from sensors and social networks. Principled and practical methods are required to deal with such streams, which includes an appropriate dealing with their temporal and belief revision aspects. Knowledge representation and reasoning aspects regarding spatial, temporal, causal and meronymic information, etc., will need to be developed both as principled-based approaches and as practically useful systems.
- **Understanding sensor data and other numeric data:** Some Big Data is heavily numeric, such as sensor data, or involves numeric data, such as quantifiable aspects of physical objects or results of scientific experiments. This calls for a seamless integration numeric processing and representation of such quantitative knowledge with logic-based knowledge representation and reasoning.
- **Dealing with uncertainty and inconsistency:** Big Data is noisy in the sense that it contains errors, omissions, overspecializations, vagueness, etc. Current reasoning approaches are unable to handle with this kind of noise in large datasets, and new theories and methods need to be developed to meet this need. Researchers use formal representation of provenance as one way to address uncertainty and inconsistency, but provenance does not provide full solution to representing and reasoning with contradictions in scientific discourse.
- **Dealing with heterogeneity:** Big Data is inherently heterogeneous, i.e. such data, even on the same overall topic, can be created with very different perspectives, underlying theories, biases, modeling rationales, etc. We will need to develop knowledge representation methods that can capture such

aspects at scale, and lead to corresponding reasoning algorithms and systems.

- **Bridging KR and other disciplines:** In order to deal with Big Data, it will be required to close the representation gap to Machine Learning and Data Mining approaches and to information extraction from texts, whose capabilities to extract higher-level features from data have so far only been of limited usefulness for deduction-based intelligent systems. Knowledge representation models, techniques, and best practices are needed that can handle the various levels of abstraction also with the various levels of "cleanliness" or dirtiness of the data that is generated by a wide range of techniques from potentially a wide range of authors.
- **Security and trust:** Knowledge representation and reasoning capabilities are required which satisfactorily address and incorporate issues of security, privacy, and trust.

Addressing these challenges will allow us to go from data to knowledge and information, to understand and interpret the data, and to make it actionable.

5.2 Grand Challenge: Knowledge Representation and Reasoning for Science Technology Engineering and Math (STEM) Education

Written by Kenneth Forbus

One of the major success stories of AI and Cognitive Science has been the rise of intelligent tutoring systems. Intelligent tutoring systems and learning environments incorporate formally represented models of the domain and skills to be learned. Scientists have shown such systems to be valuable educationally in a variety of domains. Over half-million students in the United States use such systems every year. Prior NSF studies have recognized the potential for intelligent tutoring systems to revolutionize education, by offering anytime, anywhere feedback. Such automatic methods of providing interactive feedback offer benefits across all types of education, ranging from traditional classrooms to massive open on-line courses (MOOCs). A key bottleneck in creating such systems is the availability of formally represented domain knowledge. Moving into supporting STEM learning more broadly will require new kinds of intelligent tutoring systems. For example, helping students learn to build up arguments from evidence, thereby understanding the process of scientific thinking, not just its results, requires a broad understanding of the everyday world, the informal models students bring to instruction, and the likely trajectories of conceptual change. Commonsense knowledge is both important for interacting with people via natural language, and because many student misconceptions are based on everyday experience and analogies with systems encountered in everyday life. Intelligent tutoring systems, fueled by substantial knowledge bases that incorporate both specialist knowledge and commonsense knowledge, could revolutionize education. Thus the time is right to construct large-scale knowledge bases and environments to support creating intelligent tutoring systems that operate across the range of STEM learning, from K-16.

Why now:

Progress in artificial intelligence and cognitive science more broadly has led to a deeper understanding of how to represent aspects of human mental life, including events, causality, and explanations. For example, qualitative reasoning research has led to educational systems that have been used in a variety of domains, including thermodynamics, physics, and ecology. Such systems demonstrate that scaling up to a broader range of domains could provide broader impacts in terms of new educational systems. Moreover, progress in research on learning by reading is reaching the point that scaling up in terms of total amount of knowledge in systems is becoming possible (see Section 4.3.4), as well as different kinds of knowledge.

Creating formal representations of scientific domains, at multiple age-appropriate levels, will stimulate research into human mental models and conceptual change. Such formalisms are typically constructed via cognitive task analysis, performed by hand using professional cognitive scientists, making it an expensive process. The combination of educational data mining and the ability to extract formal representations from natural interaction (see Section Large KBs) offers the potential for making this process more automatic.

A major limitation in today's intelligent tutoring systems is the means of interacting with them. Typically specialized form-based interfaces are used, which are not very natural and provide a barrier to student use. Today's dialogue-based tutors are limited by the need to hand-craft language processing. Larger knowledge bases can provide off-the-shelf semantics for the everyday world, whose construction is amortized across a wide variety of educational systems. Since spatial learning is crucial in many STEM domains, progress in sketch understanding and computer vision is needed to provide spatial modalities for interaction.

One of the long-term visions of research in intelligent tutoring systems is to provide Socratic tutoring. Socratic tutoring requires being able to understand student-proposed examples. That is, the tutor must use a combination of common sense reasoning and professional knowledge to understand what will actually happen in the situation that the student proposes, compare that understanding with the student's explanation, and then use the differences between them to provide corrective feedback. Building robust Socratic tutors that can operate in a broad range of STEM domains would be a grand challenge for KR&R, that might be achievable in a decade, given where the field is right now, with enough resources.

5.3 Grand Challenge: Develop Knowledge Bases that Capture Scientific Knowledge

Written by Yolanda Gil

The scientific literature continues to grow at unmanageable rates. Scientists often find it hard to keep up with publications within their discipline. Moreover, many research problems require understanding and incorporation of findings from

related fields and integration of knowledge across disciplinary boundaries. Ongoing research efforts aim to address these problems through diverse approaches to create knowledge bases from the scientific literature. On the one hand, well-trained curators use knowledge-engineering approaches to create knowledge repositories of published work that have high quality and precision. On the other hand, automated text extraction approaches can be used to develop knowledge bases that have lower precision but are less expensive to create and maintain. At the same time, human volunteers are creating significant repositories of scientific knowledge, including scientific portals attached to Wikipedia as well as citizen science efforts for data collection.

All these efforts are being pursued by very separate communities that approach the development of knowledge repositories in complementary ways. They also are often detached from significant bodies of background knowledge about what is known in science. A grand challenge for knowledge representation is to develop knowledge bases of scientific knowledge extracted from the published literature. These knowledge bases need to be broadly available and need to retain access to provenance - the metadata that encodes where the information comes from and how it may have been manipulated to put it into the knowledge base. Scientific knowledge bases will be very large and diverse, and as a result they will be challenging to create and complex to manage.

Addressing this grand challenge will advance the field of knowledge representation in several significant areas:

- **Extending and combining knowledge capture modalities**: Successful approaches to creating scientific knowledge bases would require a combination of manual knowledge engineering, text extraction, and interactive knowledge capture from volunteer contributors. These alternative modalities of knowledge capture have mostly been studied separately, and only combined in small-scale problems. Research challenges center around the system-mediated integration of human abilities with algorithmic abilities to represent scientific knowledge.
- **Enabling and understanding contributions provided in natural form by people:** Crowdsourcing and human computation provides novel approaches to knowledge capture that builds on the way humans interact and contribute online. Studying the modalities of human computation and understanding the most efficient ways of combining it with traditional knowledge capture promises the opportunity to scale up significantly the rate of knowledge capture from humans.
- **Integrating diverse knowledge representation frameworks:** Scientific knowledge bases need to organize knowledge from literature in the context of current scientific theories and knowledge. Representing this knowledge would push the state of the art in KR, as scientific knowledge is diverse in nature (spatial, network, quantitative, etc.) and requires the incorporation of uncertainty, abstraction, and qualitative reasoning.

- **Enabling question answering with uncertain and diverse knowledge models**: Scientific knowledge bases require the ability to answer questions with appropriate explanations, to take a proactive role in seeking new knowledge, and to manage alternative models and theories. This would challenge our ability to represent sophisticated questions, reason under uncertainty, and integrate diverse knowledge representation modalities. It would also require significant research in provenance representation and reasoning, explanation modeling, generation, and presentation, capture of context, and proactive knowledge acquisition.
- **Managing knowledge representation and reasoning at scale:** Scientific knowledge bases are currently very large and they are growing. They are also increasingly distributed and are often maintained by separate organizations, much like current scientific data sources are. This will pose new challenges in terms of managing large-scale distributed knowledge bases. Semantic web research has focused on distributed representations of knowledge, but the scale of the content and the reasoning tasks will require significant extensions to current approaches. Managing and propagating updates will pose new challenges, as will the distributed allocation of reasoning tasks. Scientific knowledge bases would significantly impact the pace of discoveries. Hunter suggests that the biomedical research may be a well-scoped challenge domain for AI, with accessible knowledge well captured in textual form. Additionally, it may require limited common sense reasoning and it has immediate potential in accelerating discoveries with broad societal impact.

While advancing the field of KR in these areas, we will also achieve the goal of extracting knowledge from scientific scale and enabling new discoveries on a completely new scale.

6 Recommendations

In this final section, we summarize a set of recommendations, both to the funding bodies and the knowledge representation researchers that have emerged from the workshop. These challenges include research, education, and infrastructure recommendations.

- **Develop research infrastructure:** A set of shared resources with query benchmarks, well described heterogeneous datasets, and shared ontologies would enable researchers (1) to compare their tools on the same sets of inputs; and (2) to get access to these datasets for development and testing. Research infrastructure must also include tools that enable researchers unfamiliar with formal knowledge representation to deploy knowledge representation methods quickly and easily in their scientific projects ("KR in the afternoon"). Such infrastructure is expensive to create and to maintain, and we need resources, similar to the NSF CRI, to support these efforts.
- **Include knowledge representation challenges in solicitations:** Many solicitations, such as the "big data" solicitations from NSF and NIH focus on the data, but largely ignore the challenges in extracting knowledge from this data. Without focusing on knowledge and knowledge representation challenges, however, big data initiatives will fall short of their promise to deliver qualitatively new advances based on big data.
- **Ground knowledge representation challenges in applications**: The knowledge representation researchers have not always been successful in grounding their research in advances in applications that need to use these advances. We encourage knowledge representation researchers to consider "application pull" in order to provide motivation, requirements, and evaluation framework for their innovations. In this report, we provided a number of application scenarios from many fields that can drive knowledge representation research.
- **Strengthen the potential impact of knowledge representation through the web**: There is great potential for broader impact of knowledge representation and ontologies through the increasing use of semantic web technologies. We should foster and expand current efforts. Knowledge capture from volunteers or crowdsourcing can be a phenomenal resource to create multitudes of formal knowledge repositories in all areas of human endeavor.
- **Develop programs around grand challenges:** We have outlined three different grand challenges that will both drive knowledge representation research and significantly advance other human endeavors (education, science, big data analysis). We propose developing programs around these grand challenges.
- **Develop stronger ties with other communities in Computer Science:** In this report, we have repeatedly highlighted the mutual benefits of integrating KR methods from NLP or Machine learning. We must build bridges with

other communities, such as databases, human-computer interaction, and cyber-physical systems. A good step in bringing researchers from these, and other, communities together will be organization of interdisciplinary workshops and development of grand challenges that will require advances in different fields. Indeed, the grand challenges that we highlighted in this report will require such interdisciplinary collaboration.

- **Highlight the role of knowledge representation in curriculum development:** ACM and IEEE publish recommended curricula in Computer Science. The proposed curriculum revision (CS 2013) mentions knowledge representation and ontologies only briefly. However, in today's knowledge-based economy, we need scientists who are comfortable with knowledge representation and semantics. We must strengthen both graduate and undergraduate education in knowledge representation. Expanding the curriculum recommendation to address the knowledge representation topics explicitly will guide educators on the important topics in knowledge representation and information systems.

References Cited

Amir E and McIlraith S (2005) Partition-based logical reasoning for first-order and propositional theories. *Artificial intelligence* **162**, 49-88.

Audemard G and Simon L (2009) Predicting Learnt Clauses Quality in Modern SAT Solver. In *Twenty-first International Joint Conference on Artificial Intelligence (IJCAI'09)*. (ed.), Vol. pp.

Balduccini M, Baral C, and Lierler Y (2008) Knowledge representation and question answering. In *Handbook of knowledge representation*. van Harmelen F, Lifschitz V, andPorter B (ed.), Vol. pp. 779–819. Elsevier,

Barker K, Agashe B, Chaw SY, Fan J, Friedland N, Glass M et al. (2007) Learning by reading: A prototype system, performance baseline and lessons learned. In *The National Conference on Artificial Intelligence*. (ed.), Vol. 22, pp. 280, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999,

Barrett CW, Sebastiani R, Seshia SA, and Tinelli C (2009) Satisfiability Modulo Theories. *Handbook of satisfiability* **185**, 825-85.

Bizer C, Heath T, and Berners-Lee T (2009) Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)* **5**, 1-22.

Bock J, Haase P, Ji Q, and Volz R (2007) Benchmarking OWL Reasoners. In *VLDB'07*. (ed.), Vol. pp. Vienna, Austria.

Brewka G, Niemelä I, and Truszczynski M (2011) Answer Set Programming at a Glance. *Communications of the ACM* **54**, 92-103.

Calvanese D, De Giacomo G, Lembo D, Lenzerini M, Poggi A, Rodriguez-Muro M et al. (2011) The MASTRO system for ontology-based data access. *Semantic Web Journal* **2**, 43-53.

Calvanese D, Kharlamov E, Nutt W, and Zheleznyakov D (2010) Evolution of DL-lite Knowledge Bases. In *9th International Semantic Web Conference (ISWC'10)*. (ed.), Vol. pp. Shanghai, China.

Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka Jr ER, and Mitchell TM (2010) Toward an Architecture for Never-Ending Language Learning. In *AAAI*. (ed.), Vol. pp.

Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA et al. (2010) Building Watson: An overview of the DeepQA project. *AI Magazine* **31**, 59-79.

Forbus KD, Klenk M, and Hinrichs T (2009) Companion cognitive systems: Design goals and lessons learned so far. *Intelligent Systems, IEEE* **24**, 36-46.

Forbus KD, Riesbeck C, Birnbaum L, Livingston K, Sharma A, and Ureel L (2007) Integrating natural language, knowledge representation and reasoning, and analogical processing to learn by reading. In *The National Conference on Artificial Intelligence*. (ed.), Vol. 22, pp. 1542, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999,

Gomes CP, Kautz H, Sabharwal A, and Selman B (2008) Satisfiability Solvers. In *Handbook of Knowledge Representation*. Harmelen Fv, Lifschitz V, andPorter B (ed.), Vol. pp. Elsevier,

Kazakov Y, Krötzsch M, and Simančík F (2011) Concurrent Classification of EL Ontologies. In *10th International Semantic Web Conference (ISWC 2011)*. (ed.), Vol. pp. Bonn, Germany.

Knorr M, Hitzler P, and Maier F (2012) Reconciling OWL and Non-monotonic Rules for the Semantic Web. In *20th European Conference on Artificial Intelligence (ECCAI 2012)*. De Raedt L, Bessiere C, Dubois D, Doherty P, Frasconi P, Heintz F and Lucas P (ed.), (ed.), Vol. 242, pp. IOS Press, Montpellier, France.

Lockwood K and Forbus K (2009) Multimodal knowledge capture from text and diagrams. In *Proceedings of the fifth international conference on Knowledge capture*. (ed.), Vol. pp. 65-72, ACM,

Motik B, Shearer R, and Horrocks I (2009) Hypertableau Reasoning for Description Logics. *Journal of Artificial Intellgence Research (JAIR)* **36**, 165--228.

Noy NF (2004) Semantic Integration: A Survey Of Ontology-Based Approaches. *SIGMOD Record* **33**,

Parundekar R, Knoblock CA, and Ambite JL (2012) Discovering concept coverings in ontologies of linked data sources. In *The 11th International Semantic Web Conference (ISWC)*. (ed.), Vol. pp. 427-43, Springer, Boston, MA.

Rossi F, Van Beek P, and Walsh T (2008) Constraint programming. *Foundations of Artificial Intelligence* **3**, 181-211.

Schraefel M and Karger D (2006) The pathetic fallacy of RDF. In *International workshop on the semantic web and user interaction (SWUI)*. (ed.), Vol. 2006, pp.

Taheriyan M, Knoblock CA, Szekely P, and Ambite JL (2012) Rapidly integrating services into the Linked Data cloud. In *The Semantic Web–ISWC 2012*. Vol. pp. 559-74. Springer,

Urbani J, Kotoulas S, Maassen J, Van Harmelen F, and Bal H (2012) WebPIE: A Web-scale parallel inference engine using MapReduce. *Web Semantics: Science, Services and Agents on the World Wide Web* **10**, 59-75.

W3C OWL Working Group (2009) OWL 2 Web Ontology Language Document Overview. In (ed.), Vol. pp. W3C Recommendation,

Wicks P, Vaughan TE, Massagli MP, and Heywood J (2011) Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat Biotechnol* **29**, 411-4.