Thoughtful Artificial Intelligence:

Forging A New Partnership for Data Science and Scientific Discovery

Yolanda Gil

Information Sciences Institute and Department of Computer Science
University of Southern California
4676 Admiralty Way
Marina del Rey 90292, USA
gil@isi.edu

Abstract

Artificial intelligence will play an increasingly more prominent role in scientific research ecosystems, and will become indispensable as we tackle more integrative science questions. While in recent years computers have propelled science by crunching through big data, qualitatively different scientific advances will result from advanced intelligent technologies for crunching through knowledge and ideas. We propose seven principles for developing thoughtful artificial intelligence, which will propel intelligent systems to become partners for scientists. We present a research agenda for thoughtful artificial intelligence, and discuss its potential for in data science and scientific discovery.

1. Introduction

Scientific research is accomplished by an ecosystem of contributors. From principal investigators that propose insightful problems, to graduate students that go deep into a specific question, to lab assistants that patiently sit through experiments, to undergraduates that contribute to simpler mundane tasks, there are a range of contributions made by people with different abilities and levels of experience. What kind of role could intelligent machines have in this ecosystem?

In the last few decades, advances in data-intensive computing have pushed the envelope in the scale of the phenomena that can be studied. Well-designed data structures, efficient algorithms, and distributed computation work at unison to process large-scale data, leading to spectacular discoveries in diverse areas such as high-energy physics, biomedicine, and geosciences. In recent years, the incorporation of intelligent techniques for data mining and machine learning has given rise to data science, bringing powerful data-driven discovery capabilities to scientists [Science 2011]. Indeed, a recent cover of *Science* states "artificial intelligence transforms science" [Science 2017a].

However, the role of artificial intelligence systems, particularly in machine learning, has been limited to solving a well-defined task where the data and techniques are given to them by the scientist. Confining intelligent machines to this data-intensive computing realm is severely limiting our ability to truly harness the potential of machines to make richer contributions. This will be particularly important for scientific applications of data science, where the complexity of the data, the questions, and the tasks will routinely challenge our ability to make discoveries.

I propose new research on *thoughtful artificial intelligence systems (ThAIS)*, which will provide significant new capabilities for data science and scientific discovery. ThAIS are capable of seeking and using knowledge necessary to do a task in a rational, ethical, and proactive manner, and are designed to interact with people, with other sources of knowledge, and with other systems. In this paper, I put forward this vision and articulate seven principles for thoughtful AI, and describe how they will bring data science to a new level. Providing these capabilities poses phenomenal challenges for artificial intelligence research.

I also posit that future scientific endeavors will require partnerships of scientists and thoughtful artificial intelligence systems, where machines will pursue independently substantial aspects of the research and contribute their own discoveries. These intelligent systems should be capable of taking on significant problems by formulating their own research goals, proposing and testing hypotheses, designing theories, debating alternative options, and synthesizing new knowledge. They should be able to explain their reasoning, compare their lines of inference to other possible paths, and situate their findings. Intelligent systems should be able to communicate with scientists with different levels of expertise and understanding in a topic. To form a true partnership, they should be able to take guidance from scientists as well as provide guidance to them in turn.

The paper begins with a discussion on the nature of the complexity of major unresolved questions in science, and why machine intelligence will be necessary to make headways. After a short overview of artificial intelligence research for scientific discovery, I propose seven principles for the developing of ThAIS. I then discuss a research agenda for artificial intelligence to achieve this vision.

2. Science Challenges in the New Millennium

Our capabilities to do research need to be augmented, as scientific questions become significantly more complex. Compare the challenges of finding a cure for polio with finding a cure for cancer. Polio, a scourge that has affected humanity for millennia, was cured through a vaccine that was discovered by one scientist. We are now faced with understanding diseases such as glioblastoma, a brain cancer that takes very few months to go to advanced stages and is very hard to detect and to treat. Scientists with different specialties have different pieces of the puzzle, such as genomics, proteomics, transcriptomics, MRIs, clinical and surgical, therapies and drugs, etc.

Important research questions are increasingly collaborative, requiring dozens, hundreds, and in some cases thousands of people in different disciplines to work together for months or years to produce results. This progression is described very well in [Barabási 2005], where the individual researchers were the norm in the past (Galileo, Newton, Darwin) giving way to dual collaborations early in the 20th century (Watson and Crick) and transitioning to larger projects such as the human genome project. These collaborations have now become even more sophisticated and complex. For example, the discovery of the Higgs boson demonstrated a major mobilization of the particle physics community to build the Large Hadron Collider (LHC), carry out the Atlas and Compact Muon Solenoid (CMS) experiments, and analyze the resulting data, leading to articles with thousands of authors [Atlas 2012]. Massive amounts of information were created and consumed by different subgroups of researchers in a painstaking and time-consuming manner. As a result, this kind of significant discovery only occurs occasionally and with significant coordination effort.

As we look to the future, science challenges have become overwhelming for collaborating scientists. In particle physics, the LHC will increase its collision rate by a factor of 10, and more sophisticated instruments such as the Long-Baseline Neutrino Facility (LBNF) will soon be a reality [Science 2017b; Cho 2017]. Other examples of these science challenges include understanding the Earth as a system of systems, studying the brain from the molecular to the cellular to the organ level, detecting and managing natural hazards, protecting our environment and using it sustainably, curing cancer and other degenerative diseases, personalizing learning and on-demand training, and designing materials with desired properties. Other important scientific questions are not even being posed, since they are far out of reach. The diversity and complexity of the data available for analysis, the amount of information to track, the amount of possible hypotheses and models to explore, and the amount of coordination across expertise areas all add up to closely related but fragmented information space that is extremely difficult to explore and manage. Even if some important aspects of these problems could potentially be solved given our current technology, we would also want faster turnaround in the work and see new results in days or weeks rather than years of research.

Computers already play a significant role in this collaborative scientific research ecosystem, albeit mostly by crunching through the large amounts of data. Through high-end computing capabilities, we now see machines do peta-scale computations routinely. Through scalable databases and information retrieval capabilities, we know machines can store large amounts of information and help scientists sift through it. Through advances in machine learning, new discoveries have been made in climate research, ecosystems, materials science, and social sciences. The scientific advances due to these technologies would be inconceivable or impossible without them.

3. Artificial Intelligence in Science

Artificial intelligence has a long tradition in tackling scientific research as a problem-solving activity [Simon 1969; Langley et al 1987; Lindsay et al 1993]. Figure 1 highlights some of the aspects of scientific discovery that have been addressed by artificial intelligence research. The description highlights the particular science targeted where the impact was most significant.

Figure 1. Major activities in scientific discovery that have been a focus of artificial intelligence research.

	Tasks	**Artificial intelligence techniques and applications**

Problem formulation	• Awareness of related work in the literature
	• Connecting relevant published information
	• Generation of new plausible hypotheses
		• Extracting knowledge from publications about geological samples
		  ○ GeoDeepDive [Peters et al 2014]
		• Linking knowledge from diverse sources in biology
		  ○ Bio2RDF [Belleau et al 2008]
		• Formulation of hypotheses given a knowledge base of known facts about biological entities
		  ○ Hanalyzer [Leach et al 2009], HyQue [Callahan et al 2011]

Experimentation and data collection	• Experiment design and execution
	• Autonomous data collection
	• Extracting knowledge from sensor readings
		• Hypothesis-driven design and execution of experiments in biology
		  ○ Robot Scientist [King et al 2004]
		• Robots for autonomous sensing and experimentation in field geology
		  ○ Nomad [Wettergreen et al 1999]
		• Extracting knowledge from noisy classifications of remote sensing data for hydrology
		  ○ [Jia et al 2016]

Data analysis	• Explore the space of hypotheses	• Explore the space of hypotheses consistent with data in organic chemistry
		  ○ Dendral [Lindsay et al 1993]
		• Bayesian classification of infrared astronomy data
		  ○ AutoClass [Cheeseman et al 1996]
		• Causal networks for econometric modeling
		  ○ TETRAD [Scheines et al 1998]
		• Integrating domain knowledge into equation discovery for population dynamics
		  ○ LAGRAMGE [Todorovski et al 2007]
		• Searching with genetic algorithms for laws that fit data in physics
		  ○ EUREKA [Schmidt and Lipson 2009]

Model revision	• Updating existing theories and models
	• Integrating prior knowledge
	• Understanding significance of results
		• Process models to revise hypotheses in paleoclimatology
		  ○ ACE [Anderson et al 2014]
		• Integration of knowledge from diverse sources about molecular mechanism for diseases
		  DiseaseConnect [Liu et al 2014]

This is not a comprehensive survey, but rather a sample of pioneering work on artificial intelligence for scientific discovery. However, there are several things to note. There is a wide diversity both in artificial intelligence research areas as well as in scientific domains. We also note that the scientific challenges uncovered research challenges for artificial intelligence, and as a result new advances in artificial intelligence enabled significant new advances in science. For more comprehensive overviews see [Gil and Hirsh 2012; Gil and Pierce 2015; Karpatne et al 2017].

Today, many artificial intelligence technologies are familiar to scientists. They include machine learning, natural language processing, semantic networks and ontologies, causal reasoning, robotics, and image processing among others. There are numerous uses of these technologies across diverse areas of science. The impact of artificial intelligence in science is palpable and continues to expand. Many scientific advances have only been possible through the pursuit of new lines of research in artificial intelligence.

In the coming years, I expect that we will see broader dissemination of these artificial intelligence techniques as well as further automation of scientific tasks. These intelligent systems will become an essential part of the scientific research ecosystem.

But I envision a much more expanded role for artificial intelligence in scientific discovery. Scientists will need intelligent systems with much more ambitious capabilities for independent inquiry, proactive learning, and deliberative reasoning. I envision a new generation of artificial intelligence systems that will enable a true partnership between scientists and machines. This partnership will be essential to tackle the science quests of our generation and many generations to come.

4. On Humans and Machines

ThAIS will become effective partners that will significantly enhance human abilities. Humans are clearly capable of being magnificent researchers, but even scientists with the best reputation make mistakes due to a number of factors:

- **Resource limitations:** Humans are resource limited in time, attention span, memory, and computation. These cognitive limitations of humans lead to situations where someone misses something that is present in the data, because they run out of time to analyze it, or found something else more interesting, they forgot it, or could not figure out the solution. One obvious place where this happens is in reading the literature, where we are limited to absorbing only a small fraction of the published record as it grows faster than we can handle. [Peters et al 2014] describes an automated system that extracted more information about the rock record than a 10-year manual effort.
- **Errors:** Humans make mistakes, and we have all come to expect it. They may overlook something, or record the wrong thing, or ignore some important data, or reach the wrong conclusion. These errors can lead to limited coverage of the observations, or to misleading findings. Just recently, a graduate student failing to reproduce a key piece of research about economic growth and debt contacted the authors and discovered that the data for some countries had been omitted by mistake [Herndon et al 2013].

- **Biases:** Humans are biased. They will use a method that they know well even if newer methods work better, just because it takes effort to learn new things and to change the research infrastructure often. There is also the well-known phenomenon of cognitive bias, where people tend to look for theories and interpretations that suit their own beliefs. [Anderson et al 2014] discusses a system that analyzed data in published papers and generated hypotheses that the authors had missed.
- **Poor reporting:** Humans are not great at recall and recounting. When people write scientific papers, they tend to focus on the big picture and not include details. Sure some details are not important, but others are and should be mentioned. Many studies have shown that this makes reproducibility extremely challenging for the vast majority of published scientific articles [Garijo et al 2013].

Intelligent systems can help counter these human shortcomings, and we have already referred to some examples that were used to detect them (e.g., [Peters et al 2014] and [Anderson et al 2014]). Intelligent systems can be systematic, covering all the space of choices without ignoring any details. They are correct, in that they follow instructions to the letter. They are unbiased, considering all the plausible interpretations however unlikely. They can also do rigorous reporting, recording and presenting every aspect of their work with justifications and verifications if needed. These are all superb qualities that we certainly want in a capable research collaborator.

At the same time, intelligent systems have important limitations in areas where humans excel. Because intelligent systems are always built to perform specific tasks, they have very narrow knowledge. Therefore, they cannot put ideas in a broader context and understand the importance of a new result. They cannot think out of the box and change perspectives or reframe problems. They also cannot envision novel forms of thinking about a problem. These limitations could be lifted as intelligent systems research progresses, but even with these limitations the intelligent systems that we could develop to support science would have important capabilities as we just discussed. Moreover, the areas where intelligent systems have limitations are precisely areas where humans shine: ingenious and creative perspectives, unconventional insights, priorities of science questions and goals, and awareness of the importance of results and ideas.

In sum, humans have well-recognized limitations in areas where intelligent machines could be very effective. As our science endeavors grow in ambition, their complexity will exacerbate human shortcomings and limitations. Intelligent systems indeed have the potential to play an important role in the research ecosystem and become effective research partners. What would it take to build such intelligent machines?

5. Thoughtful Artificial Intelligence

I propose a research agenda on a new generation of approaches that I will refer to as *thoughtful artificial intelligence systems (ThAIS)* that will become effective partners in data science and scientific discovery. ThAIS are defined by the following principles:

1. **Rationality principle**: ThAIS behave according to expectations for an artificial intelligence, that is, their behavior to accomplish any task is governed by the knowledge that they possess.

2. **Ethical principle**: ThAIS incorporate responsible and ethical behaviors, in particular the ability to recognize and convey their limitations in making decisions and taking action.
3. **Thoughfulness principle**: ThAIS use knowledge and resources that would be considered important about the context of their task in order to guide their behavior particularly in difficult or unusual cases. That is, their knowledge is not confined to the scope of their specific task. This comes with the ability to set out to expand their knowledge and seek to acquire it.
4. **Initiative principle**: ThAIS seek new knowledge proactively, and can use a variety of mechanisms to acquire it (taught by others, learned from data, extracted from text, obtained by experimenting with the world, etc). They are not just passive recipients of data or knowledge that is selected and prepared for them by people.
5. **Networked principle**: ThAIS are connected to a network of resources (documents, services, sensors and effectors, people), which gives them the ability to seek and access new knowledge or capabilities needed for doing a task.
6. **Articulation principle**: ThAIS can understand guidance and questions posed to them and respond not just with appropriate behavior but also with appropriate response back to the requester.
7. **Systems principle**: ThAIS have basic engineering design properties (such as compositionality, abstraction, and connectivity) that support integration with other systems.

Current AI systems satisfy one or only a few of these principles, but not all. For example, all artificial intelligence and machine learning systems satisfy the rationality principle, conversational interfaces comply with the articulation principle, and semantic web systems have the networked principle at their core. Today, we do not have ThAIS that embody all these principles.

ThAIS will exhibit significant new capabilities that would be key to bringing data science to a new level. Scientists expect them from research partners, and AI systems will have to grow in all these directions in order to take a much-needed role as partners in data science and scientific discovery.

6. A Research Agenda for Thoughtful AI and its Potential Impact in Data Science and Scientific Discovery

There are many research challenges in developing ThAIS. This section summarizes the challenges and some of our ongoing work towards this vision.

6.1 Ethical ThAIS: Awareness of Limitations

ThAIS should understand their limitations, and behave accordingly. This involves reasoning about their confidence on the completeness and quality of their knowledge, their ability to accomplish any task requested, and the responsibilities of being an authority in a particular subject domain. Their ethical behavior will hinge on whether they can stop themselves from taking action when they are not qualified or able to do so.

Research is needed in the area of representations of hypotheses, claims, and evidence. A biomedical repository with an entry about the interaction between a protein and a gene can cite several papers as evidence, but researchers would like more nuanced representations of that evidence so they can make informed decisions about how to use it: were the experiments done for humans or another organism, were they reliable mass spectrometry experiments or simple fluorescence spectroscopy, were any of the results replicated, and what were the p-value ranges obtained. This kind of meta-knowledge and provenance is crucial to determine the confidence on the cumulated scientific record. I have contributed to the development of provenance standards as an important enabler for this line of research [Gil and Miles 2013; Moreau et al 2011].

6.2 Networked ThAIS: Towards a Scientific Knowledge Web

Increasingly, knowledge about scientific entities of interest is captured in shared catalogs with metadata descriptions that enable scientists to find, compare, and relate those entities. Many ontologies reflect community agreements on standard ways to represent them. There is significant adoption of semantic web technologies in science. However, the establishment of links among scientific entities and the meaning of those links is largely done manually, as is the processing and understanding of the linked items. Networked ThAIS would need semantic links that they could find, interpret, exploit, and enrich.

Current structured representations for shared scientific knowledge focus on ontological models that represent objects and their properties. Ontologies are widely used in data repositories to specify metadata in order to find and integrate datasets by reasoning about their descriptions. But scientific knowledge is much more than objects and properties, and scientists need support beyond querying and aggregating data based on basic properties. In order for ThAIS to access comprehensive scientific knowledge, we need to augment our current representations and capture more complex aspects of scientific research such as data analysis processes, hypotheses and claims, and synthesis of evidence from data. We have focused on computational processes for data analysis, developing semantic workflow representations to capture knowledge about how the properties of datasets affect how they should be analyzed, and how the assumptions of data analysis algorithms constrain their use in a data analysis method [Gil et al 2011]. The WINGS intelligent workflow system incorporates a variety of computational techniques that use those representations to assist scientists in: 1) finding appropriate workflows, 2) customizing existing workflows, 3) exploring the space of alternative workflow configurations, 4) correlating the results of exploratory executions, and 5) detecting commonly used workflow fragments [Gil 2014]. We have also promoted their publication as web objects following linked open data principles [Garijo et al 2017].

There is significant ongoing work on capturing more explicitly additional forms of knowledge that are currently scattered in publications, lab notebooks, emails, presentations, and other documentation. As more structured representations of additional kinds of scientific knowledge become available, ThAIS can access increasingly more information and resources about science.

Through all this work, I see not only limitations of current representations but also the need for research on knowledge capture systems that help scientists to create them. The research involves understanding how to embed new computational approaches for method representations and abstractions into routine practice while minimizing the scientist's effort in specifying them. This will require novel intelligent user interfaces that interconnect data, software, people, instruments, and other scientific resources to effectively create meaningful chunks of a scientific knowledge web.

Given access to the data, models, and methods in this semantically enriched scientific knowledge web, ThAIS could find all the relevant information, reason about the interconnections, discover new relevant data or results across different disciplines, establish new connections and generalizations, suggest and potentially resolve inconsistencies, and generally manage the intricacies of this immensely rich and incredibly powerful scientific knowledge web. Particularly effective would be intelligent systems that could help scientists connect and collaborate across disciplines, something that currently takes significant effort and often times serendipitous connections and yet leads to transformative approaches to tackle old problems.

6.3 Thoughtful ThAIS: Context beyond their Scoped Tasks

Data analysis occurs in a much larger context of discovery: scientists are driven by models and hypotheses, and the analysis results must be converted to evidence and related back to those hypotheses. The meta-reasoning processes involved in deciding how to explore and revise such hypotheses computationally are largely unexplored. Data analytics and machine learning methods focus on finding patterns in the data, but today the overarching processes of hypothesis formulation and revision are fully carried out manually by scientists. ThAIS could largely automate this process, starting with initial hypotheses and automatically finding relevant data, analyzing it, and assessing the results. We have developed DISK, a system that uses a meta-reasoning approach that starts from research hypotheses, triggers lines of inquiry to map hypotheses to data queries and computational workflows, and uses meta-workflows to combine workflow results and discern what is interesting to report back to the researcher [Gil et al 2016a; Gil et al 2017a].

6.4 Initiative-Driven ThAIS: Access to the Scientific Record

In order for ThAIS to acquire new knowledge about science, the scientific record should be made more accessible. Current published articles do not contain the information necessary to understand what was done in enough detail that they can be reproduced [Garijo et al 2013].

Many scientists would like to improve the transparency and reproducibility of their papers, but the best practices remain difficult to understand and follow in practice. Inspired by and partnering with early career researchers, I developed the Scientific Paper of the Future (GPF) Initiative to teach scientists how to write papers that describe and cite explicitly not just data but also software and methods (workflows and provenance) [Gil et al 2016b]. The Initiative includes a special issue of a journal where submissions discuss the scientist's motivation for structuring and reporting their research products more thoroughly. We have now trained hundreds of scientists, including many center directors and principal investigators, who have changed their practices as a result. Improved publication of computational experiments will provide ThAIS with a more accessible scientific record.

6.5 Articulate ThAIS: Describing Scientific Results to Different Audiences

Articulating and communicating scientific findings will be a strong requirement for ThAIS. The new findings have to be placed in perspective of what is already known in the literature. In addition, appropriate explanations need to be generated. We are investigating the use of semantic representations of data analysis workflows to generate alternative narrative accounts in DANA, a system to customize the methods section of an article to different readers depending on their interest and expertise levels [Gil et al 2017b]. In order to be true partners in the scientific research enterprise, ThAIS will have to communicate not just how they analyzed data, but the reasons and context for the analysis as well as the significance of the results. Perhaps by reading ThAIS-generated descriptions, human scientists will adopt more precise language to describe scientific results and improve the reproducibility of the work.

6.6 System Design for ThAIS: Compositionality, Abstraction, and Connectivity

We need to develop architectures for intelligence systems that can be easily extended with new knowledge and integrated with others. Unless ThAIS have compositionality, abstraction, and connectivity, it will be hard to combine the capabilities above as they grow in breadth and depth.

7. Conclusions

Thoughtful artificial intelligence goes beyond exhibiting rational behavior, and incorporates ethical considerations, expanded context, initiative, networked connection, articulate communication, and compositional system design. Thoughtful artificial intelligence systems will result in a new generation of intelligent capabilities that will be a game changer for data science and scientific discovery. There are significant challenges ahead, and we must focus not only on developing thoughtful AI systems but also on developing infrastructure and mechanisms to enable them. With thoughtful AI systems as partners for data science and scientific discovery, scientists will be better equipped to tackle discoveries that are now unimaginable.

Acknowledgments

We gratefully acknowledge support from the Defense Advanced Research Projects Agency through the SIMPLEX program with award W911NF-15-1-0555, and from the National Institutes of Health under award 1R01GM117097.

References

[Anderson et al 2014] K. Anderson, E. Bradley, L. Rassbach de Vesine, M. Zreda, and C. Zweck, "Forensic Reasoning about Paleoclimatology: Creating a System that Works," Advances in Cognitive Systems 3:221-240, 2014.

[Atlas 2012] The Atlas collaboration. The Higgs Boson. Science Vol 338, no 6114, 21 Dec 2012. DOI: 10.1126/science.338.6114.1558.

[Barabási 2005] A. Barabási. Network theory-the emergence of creative enterprise. Science 308, 639, 2005.

[Belleau et al 2008] Francois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41:5 p706-716, 2008.

[Cho 2017] Adrian Cho. Excavation starts for U.S. particle physicists' next giant experiment. Science, Jul. 21, 2017.

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

[Cheeseman et al 1996] P. Cheeseman, J. Stutz, "Bayesian Classification (AutoClass): Theory and Results", in Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, & Ramasamy Uthurusamy, Eds. AAAI Press/MIT Press, 1996.

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

[Garijo et al 2013] Garijo, D.; Kinnings, S.; Xie, L.; Xie, L.; Zhang, Y.; Bourne, P. E.; and Gil, Y. Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. PLOS ONE, 2013. DOI: 10.1371/journal.pone.0080278.

[Garijo et al 2017] Garijo, D.; Gil, Y.; and Corcho, O. Abstract, Link, Publish, Exploit: An End to End Framework for Workflow Sharing. Future Generation Computer Systems, 2017.

[Gil and Hirsh 2012] Gil, Y. and H. Hirsh (Eds). "Final Report of the NSF Workshop on Discovery Informatics." National Science Foundation, Arlington, VA, 2012.

[Gil and Miles 2013] Gil, Y.; Miles, S A Primer for the PROV Provenance Model. World Wide Web Consortium (W3C), 2013.

[Gil and Pierce 2015] Gil, Y.; and Pierce, S. Final Report of the 2015 NSF Workshop on Intelligent Systems for Geosciences. National Science Foundation, Arlington, VA, 2015.

[Gil et al 2011] "A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs." Gil, Y.; Gonzalez-Calero, P. A.; Kim, J.; Moody, J.; and Ratnakar, V. Journal of Experimental and Theoretical Artificial Intelligence, 23(4), 2011.

[Gil et al 2016a] Automated Hypothesis Testing with Large Scientific Data Repositories. Gil, Y.; Garijo, D.; Ratnakar, V.; Mayani, R.; Adusumilli, R.; Boyce, H.; and Mallick, P. In Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS), Evanston, IL, 2016.

[Gil et al 2016b] Gil, Y.; David, C. H.; Demir, I.; Essawy, B. T.; Fulweiler, R. W.; Goodall, J. L.; Karlstrom, L.; Lee, H.; Mills, H. J.; Oh, J.; Pierce, S. A; Pope, A.; Tzeng, M. W.; Villamizar, S. R.; and Yu, X. Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance. Earth and Space Science, 3. 2016.

[Gil et al 2017a] Towards Continuous Scientific Data Analysis and Hypothesis Evolution. Gil, Y.; Garijo, D.; Ratnakar, V.; Mayani, R.; Adusumilli, R.; Boyce, H.; Srivastava, A.; and Mallick, P. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, 2017.

[Gil et al 2017b] Gil, Y.; and Garijo, D. Towards Automating Data Narratives. Proceedings of the Twenty-Second ACM International Conference on Intelligent User Interfaces (IUI-17), Limassol, Cyprus, 2017.

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

[Herndon et al 2013] Thomas Herndon, Michael Ash and Robert Pollin. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge Journal of Economics, 2013. doi:10.1093/cje/bet075

[Jia et al 2016] Xiaowei Jia, Xi C. Chen, Anuj Karpatne, Vipin Kumar. Identifying dynamic changes with noisy labels in spatial-temporal data: A study on large-scale water monitoring application. IEEE International Conference on Big Data 2016.

[Karpatne et al 2017] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop R. Ganguly, Shashi Shekhar, Nagiza F. Samatova, Vipin Kumar. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Transactions on Knowledge and Data Engineering, 2017. DOI: 10.1109/TKDE.2017.2720168

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

[King et al 2004] King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; Oliver, S. G. (2004). "Functional genomic hypothesis generation and experimentation by a robot scientist". Nature. 427 (6971): 247–252. PMID 14724639. doi:10.1038/nature02236.

[Langley 1981] Pat Langley. Data-driven discovery of physical laws. Cognitive Science, 5, 31-54, 1981.

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

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

[Lindsay et al 1993] Lindsay, Robert K., Bruce G. Buchanan, E. A. Feigenbaum, and Joshua Lederberg. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. Artificial Intelligence 61, 2 (1993): 209-261.

[Liu et al 2014] Chun-Chi Liu, Yu-Ting Tseng, Wenyuan Li, Chia-Yu Wu, Ilya Mayzus, Andrey Rzhetsky, Fengzhu Sun, Michael Waterman, Jeremy J. W. Chen, Preet M. Chaudhary, Joseph Loscalzo, Edward Crandall, Xianghong Jasmine Zhou; DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic Acids Res 2014; 42 (W1): W137-W146. doi: 10.1093/nar/gku412

[Moreau et al 2011] Moreau, L.; Clifford, B.; Freire, J.; Futrelle, J.; Gil, Y.; Groth, P.; Kwasnikowska, N.; Miles, S.; Missier, P.; Myers, J.; Plale, B.; Simmhan, Y.; Stephan, E.; and denBussche , J. V. The Open Provenance Model Core Specification (v1.1). Future Generation Computer Systems, 27(6). 2011.

[Scheines et al 1998] Scheines R, Spirtes P, Glymour C, Meek C, Richardson T. The TETRAD Project: Constraint Based Aids to Causal Model Specification. Multivariate Behavioral Res. 1998 Jan 1;33(1):65-117. doi: 10.1207/s15327906mbr3301_3.

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

[Science 2011] Special Issue on “Dealing with Data. Science.” Science, Vol. 331 no. 6018, pp. 692-693, 11 February 2011.

[Science 2017a] Special Issue on “Artificial Intelligence transforms science.” Science, Vol 357, no. 6346, 7 July 2017.

[Science 2017b] AI is changing how to we science. Science, 2017. DOI: 10.1126/science.aan7049

[Spirtes et al 2001] Spirtes,P., Glymour,C. and Scheines,R. (2001) Causation, Prediction and Search. MIT Press, Cambridge, 2001.

[Todorovski et al 2007] Ljupčo Todorovski and Sašo Džeroski. Integrating Domain Knowledge in Equation Discovery. In: Džeroski S, Todorovski L, Eds. Computational Discovery of Scientific Knowledge: Introductions, Techniques and Applications in Environmental and Life Sciences. vol. 4660. Springer; 2007. p. 69–97.

[Wettergreen et al 1999] Wettergreen, D., Bapna, D., Maimone, M., Thomas, G. Developing Nomad for robotic exploration of the Atacama Desert, Robotics and Autonomous Systems. 26(2-3) 127–148, 1999.