Data Science 1 (2017) 119–129
DOI 10.3233/DS-170011
IOS Press

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, Marina del Rey, CA, USA
E-mail: gil@isi.edu; ORCID: https://orcid.org/0000-0001-8465-8341

Editor: Michel Dumontier (https://orcid.org/0000-0003-4727-9435)
Solicited reviews: Ross King (https://orcid.org/0000-0001-7208-4387); Xander Wilcke (https://orcid.org/0000-0003-2415-8438); James McCusker (https://orcid.org/0000-0003-1085-6059)

Received 24 July 2017
Accepted 23 August 2017

Abstract. Artificial intelligence will play an increasingly more prominent role in scientific research ecosystems, and will become indispensable as more interdisciplinary science questions are tackled. While in recent years computers have propelled science by crunching through data and leading to a data science revolution, 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 turn intelligent systems into partners for scientists. We present a personal perspective on a research agenda for thoughtful artificial intelligence, and discuss its potential for data science and scientific discovery.

Keywords: Data science, scientific discovery, thoughtful artificial intelligence

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?

Though I pose this question from the point of view of science, I believe it applies to data science just the same. Scientific discovery is arguably the hardest form of data science, where the goal is to understand rather than correlate, where the phenomena are deeply interacting rather than independent, and where new findings must be integrated with existing theories.

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

2451-8484 © 2017 – IOS Press and the authors.

This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

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 [33]. Indeed, a recent cover of Science states "artificial intelligence transforms science" [35].

However, the role of artificial intelligence systems, particularly in machine learning, has been reduced to solving a well-defined task where the data and techniques are given to them by the scientist. This limits our ability to tackle problems where the complexity of the data, the questions, and the tasks challenge our human capabilities to make discoveries. Confining intelligent machines to this data-intensive computing realm is severely limiting our ability to truly harness the potential of machines to enable us to take on larger problems.

I propose new research on *thoughtful artificial intelligence*, which will provide significant new capabilities for data science and scientific discovery. Thoughtful artificial intelligence systems would be 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 artificial intelligence, 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 ones, 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 development of thoughtful artificial intelligence. I then discuss a personal perspective on a research agenda 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 information about the problem, with disparate data on genomics, proteomics, transcriptomics, MRIs, clinical, 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 [2], 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 [37]. 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.

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 petascale 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.

As we look to the future, science challenges will be ever more complex, posing unprecedented challenges for 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 [7,32]. Other examples of these significant future 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 with sustainable policies, 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 given our current capabilities. 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 important aspects of these problems can be solved, we want faster turnaround in the work and see new results in days or weeks rather than years of research.

3. Artificial intelligence in science

Artificial intelligence has a long tradition in tackling scientific research as a problem-solving activity [24,26,34]. Table 1 highlights examples of the aspects of scientific discovery that have been addressed by artificial intelligence research, mentioning a particular science targeted where the impact was most significant. This is not meant to be a comprehensive survey, but rather a sample of pioneering work on artificial intelligence for scientific discovery for readers unfamiliar with this literature. For more detailed overviews see [16,18,21,23,36]. The goal here is to highlight two important things. First, the scientific challenges uncovered research challenges for artificial intelligence, and as a result new advances in artificial intelligence enabled significant new advances in science. Second, there is a wide diversity both in artificial intelligence research areas as well as in scientific domains.

Today, many artificial intelligence technologies are familiar to scientists. They include machine learning, natural language processing, semantic networks and ontologies, causal reasoning, robotics, and

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

	Tasks	Artificial Intelligence approaches 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 [29] • Linking knowledge from diverse sources in biology ○ Bio2RDF [3] • Formulation of hypotheses given a knowledge base of known facts about biological entities ○ Hanalyzer [25], HyQue [4]
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 [22] • Robots for autonomous sensing and experimentation in field geology ○ Nomad [39] • Extracting knowledge from noisy classifications of remote sensing data for hydrology ○ [20]
Data analysis	• Explore the space of hypotheses • Machine learning • Causal models	• Explore the space of hypotheses consistent with data in organic chemistry ○ Dendral [26] • Bayesian classification of infrared astronomy data ○ AutoClass [6] • Causal networks for econometric modeling ○ TETRAD [30] • Integrating domain knowledge into equation discovery for population dynamics ○ LAGRAMGE [38] • Searching with genetic algorithms for laws that fit data in physics ○ EUREKA [31]
Model revision	• Updating existing theories and models • Integrating prior knowledge • Understanding significance of results	• Process models to revise hypotheses in paleoclimatology ○ ACE [1] • Integration of knowledge from diverse sources about molecular mechanism for diseases ○ DiseaseConnect [27]

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 be embedded in all aspects of the scientific research ecosystem.

I envision a much more expanded role for artificial intelligence in scientific discovery. Scientists will need intelligent systems with the ability to do 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 a new generation of science quests.

4. On humans and machines

AI systems 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:

- *Cognitive limitations*: Humans are resource limited in attention span, memory, and processing time. 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. [29] describes an intelligent system that extracted more information about the rock record than an earlier 10-year manual effort.
- *Errors*: Humans make mistakes, and we have all come to expect it. They may overlook something, record the wrong thing, 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 [19].
- *Biases*: Humans are biased. They will use a method that they know well even if newer better methods become available, just because of the required 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. [1] discusses an intelligent 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. 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 [10].

As our science endeavors grow in ambition, their complexity will exacerbate human shortcomings and limitations. Intelligent systems can help counter these human shortcomings, and we have already referred to some examples that were used to address them (e.g., [29] and [1]). 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 ideal qualities for capable research collaborators.

At the same time, today's 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. Interestingly, these are areas where humans shine, through ingenious and creative perspectives, unconventional insights, priorities of science questions and goals, and awareness of the importance of results and ideas.

These limitations of intelligent systems will be lifted as intelligent systems research progresses. Today's intelligent systems provide very helpful but very specific and narrow capabilities. Tomorrow, they will become research assistants to scientists. But very soon they will develop into colleagues that can do valuable research contributions independently. Intelligent systems indeed have the potential to play an important role in the research ecosystem and become effective research partners for scientists. 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. *Context principle*: ThAIs seek 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.
3. *Initiative principle*: ThAIs learn new knowledge proactively, and can use a variety of mechanisms to acquire it (e.g., 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.
4. *Networking 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.
5. *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.
6. *Systems principle*: ThAIs have basic engineering design properties (such as compositionality, abstraction, and connectivity) that support integration with other systems.
7. *Ethics principle*: ThAIs incorporate responsible and ethical behaviors, in particular the ability to recognize and convey their limitations in making decisions and taking action.

These principles are summarized in Table 2. Current artificial intelligence systems aim to satisfy only one or a few of these principles, but not all. For example, game playing systems care about the rationality principle, conversational interfaces are designed with the articulation principle in mind, and semantic web systems have the networking principle at their core. Significant new capabilities will come from the combination of all these principles.

ThAIs will represent a new generation of artificial intelligence systems that will be more resilient and resourceful, behave more responsibly, and have a sense of their role within a larger science ecosystem. ThAIs will be key to bringing data science to a new level. Scientists expect these principles from research partners, and intelligent 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.

Table 2
Defining Principles for Thoughtful Artificial Intelligence Systems

	Principle	Description
1	Rationality	Behavior is governed by knowledge
2	Context	Seek to understand the purpose and significance of tasks
3	Initiative	Proactively new learn knowledge relevant to their task
4	Networking	Access external sources of knowledge and capabilities
5	Articulation	Respond with persuasive justifications and arguments
6	Systems	Facilitate integration and collaboration with other systems
7	Ethics	Behavior that conveys scope and limitations

6. A research agenda for thoughtful artificial intelligence and its potential impact in data science and scientific discovery

This section discusses major research challenges in developing ThAIs, and highlights some of our ongoing work towards this vision. I leave out the rationality principle because it has been discussed in depth elsewhere.

6.1. 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 [3]. 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 [8]. I am personally interested in computational processes for data analysis, and have developed 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 [15]. 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. We publish workflows as web objects following linked open data principles [9]. This enables other systems to examine, compare, and reuse these workflows, effectively building on another system's scientific processes. This is a key building block towards allowing ThAIs to acquire knowledge from other networked sources about how to do data analysis.

Further research is needed 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.

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 relevant information for their tasks, reason about new 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 will 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.

6.2. Context-driven ThAIs: Task scoping

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.

I am very interested in enabling intelligent systems to reason about hypotheses to create goals and define their own problems to pursue. We are developing 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 [13,14].

Much work remains to be done to represent scientific hypotheses, to capture meta-reasoning strategies that scientists pursue, and to analyze results. Ongoing work in cognitive science and philosophy of science is articulating such processes [5]. We need to incorporate them in artificial intelligence systems.

6.3. 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 [10].

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 (SPF) Initiative to teach scientists how to write papers that describe and cite explicitly not just data but also software and methods (workflows and provenance) [11]. 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.

This line of work complements ongoing efforts to extract information from the scientific literature through text mining [29]. Most of the work focuses on extracting facts and claims. Much remains to be done to address the challenges of extracting information about processes and methods.

6.4. Articulate ThAIs: Describing scientific results to different audiences

Communicating scientific findings will be a strong requirement for ThAIs. New findings have to be placed in perspective of what is already known in the literature. In addition, appropriate explanations and compelling arguments need to be generated.

I am very interested in using computational workflows to generate automatically explanatory text that could be included in scientific papers 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 [12]. 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.

6.5. System design for ThAIs: Compositionality, abstraction, and connectivity

We need to develop architectures for intelligence systems that can be easily extended with new capabilities 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.

6.6. 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 more nuanced representations of that evidence are needed in order to 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 [4]. I have contributed to the development of provenance standards as an important enabler for this line of research [17,28].

7. Conclusions

Thoughtful artificial intelligence goes beyond exhibiting rational behavior, and incorporates expanded context, initiative, networked connection, articulate communication, compositional system design, and ethical considerations. Thoughtful artificial intelligence systems will result in fundamentally new capabilities that will be a game changer for data science and scientific discovery.

There are significant research challenges ahead, and we must focus not only on developing thoughtful artificial intelligence but also on developing infrastructure and mechanisms to enable them. With thoughtful artificial intelligence systems as partners for data science and scientific discovery, we will be better equipped to tackle discoveries that are now unreachable.

Acknowledgements

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. We would like to thank Daniel Garijo, Hiroaki Kitano, and Parag Mallick for many thoughtful discussions.

References

[1] 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** (2014), 221–240. http://www.cogsys.org/pdf/paper-9-3-39.pdf.

[2] A. Barabási, Network theory – The emergence of creative enterprise, *Science* **308** (2005), 639. doi:10.1126/science.1112554.

[3] F. Belleau, M. Nolin, N. Tourigny, P. Rigault and J. Morissette, Bio2RDF: Towards a mashup to build bioinformatics knowledge systems, *Journal of Biomedical Informatics* **41**(5) (2008), 706–716. doi:10.1016/j.jbi.2008.03.004.

[4] A. Callahan, M. Dumontier and N.H. Shah, HyQue: Evaluating hypotheses using semantic web technologies, *Journal of Biomedical Semantics* **2**(Suppl 2) (2011), S3. doi:10.1186/2041-1480-2-S2-S3.

[5] S. Chandrasekharan and N.J. Nersessian, Building cognition: The construction of computational representations for scientific discovery, *Cognitive Science* **39**(8) (2015), 1727–1763. doi:10.1111/cogs.12203.

[6] P. Cheeseman and J. Stutz, Bayesian classification (AutoClass): Theory and results, in: *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds, AAAI Press/MIT Press, 1996. https://dl.acm.org/citation.cfm?id=257954.

[7] A. Cho, Excavation starts for U.S. particle physicists' next giant experiment, *Science* (2017). doi:10.1126/science.aan7143.

[8] P. Ciccarese, D. Shotton, S. Peroni and T. Clark, CiTO + SWAN: The web semantics of bibliographic records, citations, evidence and discourse relationships, *Semantic Web Journal* **5**(4) (2012), 295–311. doi:10.3233/SW-130098.

[9] D. Garijo, Y. Gil and O. Corcho, Abstract, link, publish, exploit: An end to end framework for workflow sharing, *Future Generation Computer Systems* **75** (2017), 271–283. doi:10.1016/j.future.2017.01.008.

[10] D. Garijo, S. Kinnings, L. Xie, L. Xie, Y. Zhang, P.E. Bourne and Y. Gil, Quantifying reproducibility in computational biology: The case of the tuberculosis drugome, *PLoS ONE* **8**(11) (2013), e80278. doi:10.1371/journal.pone.0080278.

[11] Y. Gil, C.H. David, I. Demir, B.T. Essawy, R.W. Fulweiler, J.L. Goodall, L. Karlstrom, H. Lee, H.J. Mills, J. Oh, S.A. Pierce, A. Pope, M.W. Tzeng, S.R. Villamizar and X. Yu, 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**(10) (2016), 388–415. doi:10.1002/2016EA000201.

[12] Y. Gil and D. Garijo, Towards automating data narratives, in: *Proceedings of the Twenty-Second ACM International Conference on Intelligent User Interfaces (IUI-17)*, Limassol, Cyprus, 2017. https://doi.org/10.1145/3025171.3025193.

[13] Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce and P. Mallick, Automated hypothesis testing with large scientific data repositories, in: *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS)*, Evanston, IL, 2016. http://www.cogsys.org/papers/ACS2016/Papers/Gil_et.al-ACS-2016.pdf.

[14] Y. Gil, D. Garijo, V. Ratnakar, R. Mayani, R. Adusumilli, H. Boyce, A. Srivastava and P. Mallick, Towards continuous scientific data analysis and hypothesis evolution, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, San Francisco, CA, 2017. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14839.
[15] Y. Gil, P. Gonzalez-Calero, J. Kim, J. Moody and V. Ratnakar, A semantic framework for automatic generation of computational workflows using distributed data and component catalogs, *Journal of Experimental and Theoretical Artificial Intelligence* **23**(4) (2011), 389–467. doi:10.1080/0952813X.2010.490962.
[16] Y. Gil and H. Hirsh (eds), Final report of the NSF Workshop on Discovery Informatics. Report, National Science Foundation, Arlington, VA, 2012. http://www.isi.edu/~gil/diw2012/NSFDiscoveryInformatics2012-FinalReport.pdf.
[17] Y. Gil and S. Miles, *A Primer for the PROV Provenance Model. World Wide Web Consortium (W3C)*, 2013. http://www.w3.org/TR/prov-primer/.
[18] Y. Gil and S. Pierce, Final report of the 2015 NSF Workshop on Intelligent Systems for Geosciences. Report, National Science Foundation, Arlington, VA, 2015. https://dl.acm.org/citation.cfm?id=2856633.
[19] T. Herndon, M. Ash and R. Pollin, Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff, *Cambridge Journal of Economics* **38**(2) (2013), 257–279. doi:10.1093/cje/bet075.
[20] X. Jia, X.C. Chen, A. Karpatne and V. Kumar, Identifying dynamic changes with noisy labels in spatial-temporal data: A study on large-scale water monitoring application, in: *IEEE International Conference on Big Data*, 2016. https://doi.org/10.1109/BigData.2016.7840738.
[21] A. Karpatne, G. Atluri, J.H. Faghmous, M. Steinbach, A. Banerjee, A.R. Ganguly, S. Shekhar, N.F. Samatova and V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, *IEEE Transactions on Knowledge and Data Engineering* **29**(10) (2017), 2318–2331. doi:10.1109/TKDE.2017.2720168.
[22] R.D. King, K.E. Whelan, F.M. Jones, P.G.K. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell and S.G. Oliver, Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature* **427**(6971) (2004), 247–252. doi:10.1038/nature02236.
[23] H. Kitano, Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery, *AI Magazine* **37**(1) (2016), 39–49. doi:10.1609/aimag.v37i1.2642.
[24] P. Langley, H.A. Simon, G.L. Bradshaw and J.M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Processes*, MIT Press, Cambridge, MA, 1987. https://mitpress.mit.edu/books/scientific-discovery.
[25] S.M. Leach, H. Tipney, W. Feng, W.A. Baumgartner Jr., P. Kasliwal, R.P. Schuyler, T. Williams, R.A. Spritz and L. Hunter, Biomedical discovery acceleration, with applications to craniofacial development, *PLoS Computational Biology* **5**(3) (2009), e1000215. doi:10.1371/journal.pcbi.1000215.
[26] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, DENDRAL: A case study of the first expert system for scientific hypothesis formation, *Artificial Intelligence* **61**(2) (1993), 209–261. doi:10.1016/0004-3702(93)90068-M.
[27] C.C. Liu, Y.T. Tseng, W. Li, C.Y. Wu, I. Mayzus, A. Rzhetsky, F. Sun, M. Waterman, J.J.W. Chen, P.M. Chaudhary, J. Loscalzo, E. Crandall and X.J. Zhou, DiseaseConnect: A comprehensive web server for mechanism-based disease–disease connections, *Nucleic Acids Research* **42** (2014), W137–W146. doi:10.1093/nar/gku412.
[28] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan and J.V. denBussche, The open provenance model core specification (v1.1), *Future Generation Computer Systems* **27**(6) (2011), 743–756. doi:10.1016/j.future.2010.11.020.
[29] S.E. Peters, C. Zhang, M. Livny and C. Ré, A machine reading system for assembling synthetic paleontological databases, *PLoS ONE* **9**(12) (2014), e113523. doi:10.1371/journal.pone.0113523.
[30] R. Scheines, P. Spirtes, C. Glymour, C. Meek and T. Richardson, The TETRAD project: Constraint based aids to causal model specification, *Multivariate Behavioral Research* **33**(1) (1998), 65–117. doi:10.1207/s15327906mbr3301_3.
[31] M. Schmidt and H. Lipson, Distilling free-form natural laws from experimental data, *Science* **324**(5923) (2009), 81–85. doi:10.1126/science.1165893.
[32] Science News Staff, AI is changing how we do science, *Science* (2017). doi:10.1126/science.aan7049.
[33] Science Staff, Challenges and opportunities, *Science* **331**(6018) (2011), 692–693. doi:10.1126/science.331.6018.692.
[34] H.A. Simon, *The Sciences of the Artificial*, MIT Press, 1969. https://mitpress.mit.edu/books/sciences-artificial.
[35] Special Issue on Artificial Intelligence Transforms Science, *Science* **357**(6346) (2017). http://science.sciencemag.org/content/357/6346.
[36] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction and Search*, MIT Press, Cambridge, MA, 2001. https://mitpress.mit.edu/books/causation-prediction-and-search.
[37] The Atlas Collaboration, The Higgs boson, *Science* **338**(6114) (2012), 1558–1559. doi:10.1126/science.338.6114.1558.
[38] L. Todorovski and S. Džeroski, Integrating domain knowledge in equation discovery, in: *Computational Discovery of Scientific Knowledge: Introductions, Techniques and Applications in Environmental and Life Sciences*, S. Džeroski and L. Todorovski, eds, Lecture Notes in Artificial Intelligence, Vol. 4660, Springer, 2007, pp. 69–97. doi:10.1007/978-3-540-73920-3_4.
[39] D. Wettergreen, D. Bapna, M. Maimone and G. Thomas, Developing Nomad for robotic exploration of the Atacama Desert, *Robotics and Autonomous Systems* **26**(2–3) (1999), 127–148. doi:10.1016/S0921-8890(99)80002-5.