Proceedings of the 44th Annual Conference of the Cognitive Science Society (CogSci), Toronto, Ontario, July 27-30, 2022.

Towards Capturing Scientific Reasoning to Automate Data Analysis

Yolanda Gil, Deborah Khider, Maximiliano Osorio, Varun Ratnakar, Hernan Vargas, Daniel Garijo
University of Southern California
{gil,khider,mosorio,varunr,hvargas,dgarijo}@isi.edu

Suzanne Pierce
University of Texas at Austin
pierce@tacc.utexas.edu

Abstract

This paper describes an initial cognitive framework that captures the reasoning involved in scientific data analyses, drawing from close collaborations with scientists in different domains over many years. The framework aims to automate data analysis for science. In doing so, existing large repositories of data could be continuously and systematically analyzed by machines, updating findings and potentially making new discoveries as new data becomes available. The framework consists of a cycle with six phases: formulating an investigation, initiating the investigation, getting data, analyzing data, aggregating results, and integrating findings. The paper also describes our implementation of this framework and illustrates it with examples from different science domains.

Keywords: scientific discovery, automated data analysis, cognitive scientists, AI scientists

Introduction

Over the last decades, many scholars have shed light on the diverse and rich processes involved in scientific reasoning, from discovering laws [Simon 1977], to understanding causal mechanisms [Craver and Darden 2013; Pearl 2018], to collaboration [Trickett et al 2015], to producing paradigm shifts [Kuhn 1962]. The development of cognitive models that reflect how scientists think is indeed a daunting task. Our goals are much narrower, focusing very specifically on capturing the scientific reasoning that we observed through many years of working with scientists in data analysis as we represented their tasks, implemented their computational methods, and supported their collaborative work.

Our focus is on scientific research that revolves around data analysis, in particular observational science where data is abundant. There are many other types of scientific research that are not directly linked to the analysis of data (though eventually they can be). Some are designed to gain some understanding on how to tackle an open problem, perhaps by assembling information about the state-of-the-art in relevant publications or by coming up with new ways to frame a problem that can lead to new research avenues. Other investigations are designed to be exploratory in nature in terms of trying out possible directions through informed guesses to gather more information about the problem. These eventually lead to data analysis which is the current focus of our work.

There is prior work on developing frameworks for scientific data analysis. Others have focused on automating the extraction of findings from the literature [Tshitoyan et al 2019], the exploration of complex search spaces [Senior et al 2020], the formulation of hypotheses [Callahan et al 2011], or the design of laboratory experiments [Groth and Cox 2017]. Our focus is on scientific reasoning involved in data analysis where there are significant data resources that enable the pursuit of many research problems, where we find current analyses are done in a piecemeal manual way. This is the case in many science domains, where extensive amounts of data are available including biomedical, geosciences, and social sciences datasets. Today, their analysis is driven by researchers with limited time and resources and much of the data is underutilized. *Our work is the first to focus on capturing the scientific reasoning that can lead to the automated continuous analyses of these vastly underutilized data resources.*

This article draws on our work with scientists in very diverse domains, developing a variety of platforms to support their data analysis work. We have worked on population genomics, clinical omics, water quality, cancer multi-omics, neuroscience, hydrology, agriculture, climate, wildfire, disease spread, and economics [Zheng et al 2015; Gil et al 2019; Gil et al 2011; Gil et al 2017; Gil et al 2021; Khider et al 2020]. The contributions of this work are twofold:

- A proposed general framing of scientific reasoning for data analysis as a cognitive framework with six distinct phases, based on our observations in several scientific domains
- A description of our implementations that capture and represent the knowledge involved in four of the six phases in different domains and our use of this approach in different frameworks and science domains

We begin by framing the kinds of reasoning involved in scientific data analysis, including formulating an investigation, initiating the investigation, getting data, analyzing data, aggregating results, and integrating findings. Then, we illustrate how we capture the knowledge required and carry out the reasoning involved in our implementations for different science domains. We close with conclusions and directions for future work.

Framing Scientific Reasoning for Data Analysis

A long-term research project targets a broad set of open problems, which are tackled by breaking them down into smaller problems that can be accomplished in a reasonable way with whatever resources are available for the project (people, computing, time, data, etc). An open problem is broken down into subproblems, perhaps creating tasks and decomposing them in turn into subtasks. Eventually they become tractable enough that they prompt specific research activities that can be undertaken to reach some conclusion given the resources available. Here, we focus on such activities and refer to them as *investigations*. More specifically, we consider here investigations that have a concrete goal that can be accomplished through data analysis whose results address that aim.

Figure 1 shows an overview of the cognitive framework that we have developed and the major phases that we have identified. They are described in the rest of this section.

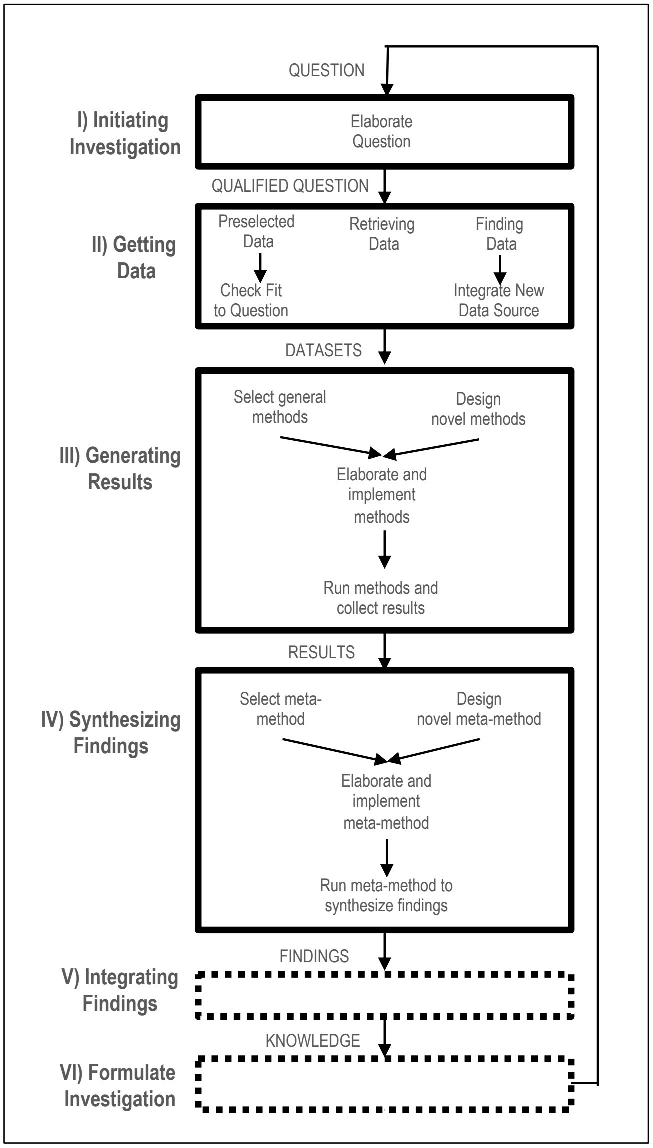

Figure 1. A cognitive framework for scientific data analysis.

Phase I: Initiating Investigations

The aim of an investigation is to answer a *question*, which is most useful if it is well scoped to be tractable given the time and resources available. An example of a question that we have seen in neuroscience is whether the effect of a specific genotype on the prefrontal cortex size is associated with age. Questions can be posed as testing a *hypothesis*, which would be a statement that can be supported or dismissed based on data analysis. For example, in cancer omics we have analyzed data to test hypotheses that a particular protein known to be associated with colon cancer is present in a specific colorectal cancer patient sample. Hypotheses can be implied, but when explicitly declared they are more effective in guiding the investigation.

Sometimes the formulation of questions or hypotheses is itself the aim of an investigation. Scientific endeavors can be arbitrarily complex, and as we mentioned we focus here on investigations that can be resolved through data analysis. We will consider a question or hypothesis as the starting point.

We have noticed general patterns in the types of questions that are posed in investigations for any given domain or context, and many of these patterns appear across domains. This is not surprising, since statistics, inference, and induction are general methodologies across sciences.

Figure 2 illustrates broad categories of scientific questions. Some questions have to do with associations between observed variables. For example, in neuroscience we can ask if a brain characteristic, such as hippocampus size, is associated with a disorder such as ADHD. Other questions are concerned with characterizing the data at hand. For example, a question posed about a timeseries could be whether it has seasonality or longer periodicities. There are many important questions that target a causal understanding of a complex system [Pearl 2018]. These are concerned with changes in its state over time (due to actions or events) that lead to other (causally related) changes.

Table 2 shows examples of the kinds of questions that our collaborating scientists have set out to address.

We have found that scientists spend significant time narrowing down the types of questions that they will focus on. This is a process that can be considered an investigation in itself as we mentioned earlier.

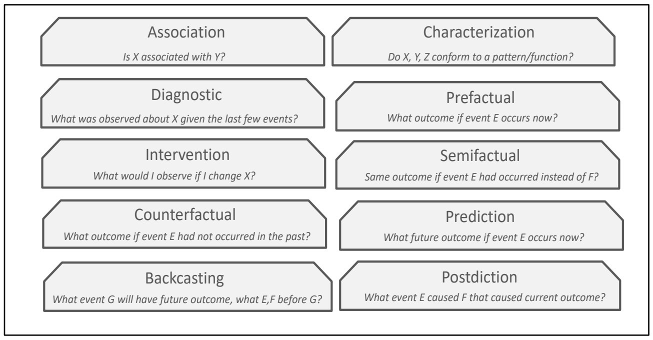
Association
Is X associated with Y?
Characterization
Do X, Y, Z conform to a pattern/function?
Diagnostic
What was observed about X given the last few events?
Prefactual
What outcome if event E occurs now?
Intervention
What would I observe if I change X?
Semifactual
Same outcome if event E had occurred instead of F?
Counterfactual
What outcome if event E had not occurred in the past?
Prediction
What future outcome if event E occurs now?
Backcasting
What event G will have future outcome, what E,F before G?
Postdiction
What event E caused F that caused current outcome?

Figure 2. Broad categories of scientific questions about dynamic systems.

Table 2. Examples of general types of scientific questions.

Question Category	Example
Association	*Is <brainCharacteristic> associated with <neurologicalDisorder> in comparison to healthy controls?*
Characterization	*Does <timeseries> have <period> seasonality?*
Diagnostic	*What local areas are subject to frequent flooding?*
Intervention	*What could be the yield of <crop> if <fertilizerSubsidies> are given this year?*
Counterfactual	*What would have been the yield of <crop> if precipitation had been <number> times <larger/smaller> last year?*
Backcasting	*What <plantingWindow> could have lead to increased crop yield this year over last year?*
Prefactual	*Can we expect <flooding> in <area> in the rainy season this coming year?*
Prediction	*How will future <climateScenarios> impact <waterResources> in <region>?*

Questions can be elaborated by adding *qualifiers* to narrow down the scope of the question. In our examples, this might involve selecting a specific time period (e.g., going from "future" to "the next 10 years"), a region (e.g. the Northern side of a river basin), or a threshold (e.g., flooding is defined as water covering the soil by more than 2 inches for more than 1 day.). In some cases the qualifiers are determined by the resources available. For example if computational resources are limited a smaller time frame will be chosen.

Alternatively, questions can be decomposed into further subquestions whose results need to be aggregated to answer the original question. For example, to determine the yield of a crop for given planting dates one might have to consider different values for fertilizer use. Each possible amount of fertilizer that will be considered would result in a subquestion. In this kind of exploration of the solution space, sometimes cast as parameter sweeping, different values of a parameter are tried and then all the results are combined (through an average or some other function) into an answer for the overall question.

In more complex cases, a question can lead to a collection of subtasks. For example, a question concerning fertilizer subsidies could lead to a subtask to find an agroeconomic model that combines agriculture yield predictions with socioeconomic aspects of the likelihood of uptake of fertilizers by farmers based on market prices. Finding such a model may not be straightforward, and the scientist may need to consider if it is possible to build it given the time and other resources available. If the question is important and requires developing new sophisticated models, it may take many months to get that subtask done before any data analysis can take place. Another example is questions concerning flooding or water resources where a preexisting model may provide a starting point but needs to be calibrated to adapt it to the region at hand. In those cases, the calibration subtask may also take significant effort but is required before any meaningful data is available for analysis.

Phase II: Getting Data

As we mentioned earlier, for some investigations there are significant amounts of data available in shared repositories. In certain cases, simulations or other predictive models can be used to generate the data needed. But in other cases, the data is already preselected and is the center of the investigation. We discuss each situation in this section.

Preselected Data Sometimes a question is posed about a specific dataset that is provided along with the question. In those cases, there may be subtasks concerning whether the dataset is right for the question asked.

Although there may be no need to seek out data, it may still be useful to still try to find relevant data and alert the scientist that other data is available that they are not aware of.

Retrieving Data Given a question that can be addressed with available datasets, the investigation proceeds by retrieving the relevant data. When using data repositories, this involves mapping each question to a query that describes the relevant datasets.

Finding Data Some questions require running simulations of a complex dynamic system. Datasets must be found to set up the initial state. For example, an economic model may need the current market price of different crops. In addition, datasets may be needed to support the analysis of different scenarios. For example, different drought conditions can be explored if datasets are found that contain climate forecasts. It may not be trivial to locate data repositories that contain the desired datasets.

In all these cases, if the necessary data are not found, then the question cannot be pursued and has to be adjusted or abandoned.

Phase III: Generating Results

To analyze the data available, a data analysis method is applied to the data. In rare cases, a new method may need to be developed as one of the subtasks to address the question posed. But in most cases, there are widely-used proven methods that can be applied to the data at hand. The analytic method is often discipline specific, but it can be general such as statistical methods or machine learning algorithms. For example in population genomics a common method is association testing, while in document analysis a common method is hierarchical clustering. In the case of simulations, setting up models is a key part of the method.

Multi-Step Methods Over the years, we have implemented and/or executed hundreds of scientific data analysis methods. They always consist of multiple interdependent steps. The steps are executed in turn, often started by the scientist through simple interfaces such as a command line invocation or a map-based application.

General Methods We have found that scientists typically follow a method as described in a publication or a method that is commonly used. At the same time, it is rare that all labs have the same software: some labs prefer Python and others R or Matlab. Therefore, even though the execution of the data analysis software is done with particular software, scientists have a concept of *general methods* in an implementation-independent manner. General methods are abstract plans. Abstract plan steps are decomposed into several substeps or specialized into more specific substeps that eventually bottom out in data analysis software that can be executed. The main steps in the general method could involve a simulation model or an empirical model. A main step in the method could also be a statistical function.

Data Preparation General methods do not typically mention steps that are not critical to the method but are necessary for running its implementation. For example, data transformation steps may be added as the general method is elaborated. Data preparation steps are always needed to pre-process data so it fits the requirements of the software used to implement the main steps of the method.

Results Once data analysis is executed, there will be a collection of *results* that needs to be aggregated in the next stage. Associations may have been analyzed for each of the datasets available, and now the individual results need to be combined. In other cases, hundreds of simulations using different parameter values may have been run, whose individual results would need to be combined in order to answer the initial question. We have also seen cases where several alternative analysis methods are run (i.e., an ensemble method), and combining their results helps increase performance or reduce uncertainty.

Phase IV: Synthesizing Findings

The individual results of the data analysis phase are then synthesized into a set of overall *findings*. As was the case in the prior phase, this requires identifying an existing method or designing one. This might be as easy as taking an average of the results, or some other statistical function. This is sometimes called meta-analysis when the datasets were collected independently (e.g., in different studies). When the investigation starts with a hypothesis, the finding must be in support or against it (for example with a confidence value being high or low).

In extreme cases, analysis results are hard to aggregate into conclusive findings of an investigation. For example, in cancer omics we may look at data from several hundred patients but there are different types of data available for each (for one it may be only genomic data, for another it may be genomic and also proteomic data from mass spectrometry, and for yet another it may be genomic data and proteomic data from fluorescence imaging). In such extreme cases, scientists often consider the evidence separately for each type of data, making the meta-analysis a straightforward aggregation for each type of data.

We found an interesting case in neuroscience, where the scientists do not run the data analysis, only the meta-analysis. They worked with several data providers who did not want to share their data, but were willing to run the analysis in their respective sites and share the results. The scientists then did the meta-analysis over those results.

Phases V and VI: Integrating Findings with Current Knowledge and Formulating the Next Investigation

The findings from data analysis are integrated with existing knowledge or theories, leading to revisions or extensions. The final phase includes prioritizing problems or questions based on their potential impact, refining problems into subproblems or subquestions, and ultimately initiating new investigations so the cycle is back to Phase I.

These two phases have not been the focus of our work so far. They have been studied by others [Thagard 2012; Samuels and Wilkenfeld 2019; Addis et al 2016; Chandrasekharan, S. & Nersessian 2015].

Capturing Scientific Reasoning for Data Analysis Inquiries

This section describes the representations and reasoning in our implementation of four of the six phases of our cognitive framework, namely: initiating the investigation, getting data, analyzing data, and aggregating results. We leave out two phases that will be subject of future work: the last phase of integrating findings with what is known, and the subsequent phase that re-iterates the cycle by formulating the next investigation. We have used these representations in different systems that address different science domains and purposes [Gil et al 2021; Gil et al 2017; Gil et al 2011].

In describing these representations, we provide examples using a simplified, more readable format. In our implementations, we use semantic web representation standards from the World Wide Web Consortium (W3C), including OWL, RDF, SPARQL, SWRL, and PROV [World Wide Web Consortium 2022]. These languages have expressive limitations, but come with open-source tools and efficient off-the-shelf reasoners. Their limitations have not been an issue for our research so far, and there are many benefits to doing our work on an open-source substrate. In addition, many scientists are familiar with these languages, as they are widely used in biomedicine and increasingly used in other scientific disciplines.

Initiating Investigations

In our framework, scientists provide the initial hypotheses and questions that initiate the investigations.

We create a *question ontology* that includes classes of objects or concepts in the domain that can be used to formulate questions. For the cancer omics domain, our question ontology included classes such as <protein>, <gene>, and <patient-sample>. For the neuroscience domain, a question ontology can include classes such as <genotype>, <brainCharacteristic>, and <demographicAttribute>. Some of the terms may appear in existing domain ontologies, but they have to be agreed upon as valid terms for expressing hypotheses.

We also formulate *question templates*, which are logic expressions that includes variables of a type already included in the question ontology, in addition to text that expresses the question being posed. Examples in neuroscience include:

> Is the effect of <genotype> in <brainTrait> associated with <demographicAttribute>?
>
> Is the effect size of <genotype> on <brainRegionTrait> of <brainRegion> associated with <demographic>?

Then, scientists create *question statements* by specializing question templates. For example, from the last question template above the following question statement could be formulated:

> Is the effect of APOE in HyppocampalVolume associated with Age?

As we mentioned, questions may need to be further specified. We do this with domain-specific *qualifiers* that need to be defined using a *qualifier ontology*. For example, this question:

> What will be the increase in <crop> yield if there are <item> subsidies?

may prompt the following qualifiers:

> What will be the increase in <crop> yield if there are <item> subsidies measured as <potentialCropProduction>, from <beginDate> to <endDate>, in <region>?

We have noticed that often times the scientists do not have a choice in these qualifiers, as they may be determined by the datasets that are available. For example, if data is only available for certain years, then the years selected will have to be within that range. Therefore, for elaborating questions to include all necessary qualifiers, we have developed in the past user interfaces to elicit those qualifiers from users.

Getting Data

Once the question is specified, it can be used to formulate the right queries that will get relevant data.

Retrieving Data For retrieving data, we need to represent *data queries*. Those queries are then issued against the data repository to retrieve the data. For that reason the queries are formulated using *metadata attributes* defined for the data repository, otherwise an ontology mapping or translation step would be needed. The following is an example of a SPARQL query to retrieve data for the second question template above:

```
SELECT ?dataset WHERE {
  ?cohort a Cohort .
  ?cohort HasGeneticDataType ?Genotype .
  ?cohort HasDataset ?dataset . }
```

which requests a dataset from a study cohort that has the desired genotype specified in the question. Note that the data repository needs to offer appropriate metadata so that queries like these can be formulated. Not all data repositories do, and in that sense our work creates new requirements for scientific repositories in order to support automated data analysis.

Finding Data For finding data, we start with an elaborated question. For example, a question to generate crop yield in a region with fertilizer subsidies would require running an agriculture model. Different agriculture models have different data needs, but generally they would require data about soils and slopes and weather predictions. We represent the data requirements for each model, in this example in terms of physical variables needed for soil and atmosphere:

```
MODEL {Cycles} MODEL_REQUIRES {
    soilThickness, surfaceSlope, soilMoisture, dailyMaxTemp,
    atmosphSaturation, dailyPrecipitationpVolume }
```

and we know this is a useful model for the query because it generates the measurement that the query requires:

```
MODEL {Cycles} MODEL_GENERATES {
     potentialCropProduction }
```

Now that we know what data is required, we issue data queries accordingly. For example, the following is a JSON query to retrieve data for the model above:

```
QUESTION_REQUIRES
    Variables {
      soilThickness, surfaceSlope, soilMoisture, dailyMaxTemp,
      atmosphSaturation, dailyPrecipitationpVolume }
   SpatialCoverage_Intersects { ?region}
   TemporalCoverage_Intersects { ?beginDate ?endDate }
```

We have found that if the scientist selects a specific model first, that narrows down the data needs to those of that model alone. But in some cases, it may be preferrable to run as many models as there is data for. This is a good example of how these phases are not necessarily done sequentially, and scientists do backtrack and change some of their decisions based on what they see happening downstream. For example, they may have a question in mind but the data to run the necessary models cannot be found, and if so the question will be changed to adapt it to the data available. This back and forth is often done manually, doing web searches to find relevant data that might support the desired model.

Generating Results

Once the data is located, a method can be selected.

Multi-Step Methods As we mentioned, we have found that methods typically consist of multiple interdependent steps. Most methods can be represented as a *workflow*, with input data and output data. Methods may involve iterating over some steps while tweaking their setup and manually checking the results until the scientist is satisfied. If there is no need for manual inspection, methods can be cast as workflows so they are automatically executed by a workflow system. We use a workflow language to represent methods and their execution [Gil et al 2011].

General Methods General methods can be represented as abstract plans. What we have found is that in most cases the general methods are quite prescriptive, leaving little room for decomposition or elaboration. That is, most general methods can be described as *skeletal plans*, a special form of abstract plans where there is only specialization and no decomposition [Friedland and Iwasaki 1985]. That is, each step in the skeletal plan is specialized into a more specific step and the steps in the final plan have a one-to-one correspondence with the steps in the original skeletal plan. We have adapted these ideas to develop a workflow management system that can specialize workflow templates into executable workflows, and can incorporate reasoning about constraints in the process. This is described in detail elsewhere [Gil et al 2011]. For example, an abstract step might indicate the use of an agriculture model which can be specialized to use specific models, or an abstract step could be to detrend a time series which can be specialized to linear detrending or polynomial detrending.

Data Preparation We also cast data preparation steps as workflows. We find that they do not tend to use general methods, instead they are implemented by cutting corners to save time. The cost of implementing data preparation often deters scientists from examining questions thoroughly. Their automation appears very feasible and very desirable.

Synthesizing Findings

We synthesize findings through meta-workflows. Our meta-workflows are implemented using the same workflow representations, except they take as inputs the workflows that generated the results. This is important, since the reasoning to synthesize and aggregate results needs to take into account the method and the implementation used to generate them. Our semantic web representations allow us to publicly *publish* workflows by posting them on the Web with a unique persistent URL, and to access them through that URL as well as all its constituent steps and intermediate datasets.

Integrated Reasoning through Lines of Inquiry

A key innovation in our work is capturing how questions, methods, and meta-analysis are connected together to allow the inquiry to be triggered and proceed. A *line of inquiry (LOI)* is the mechanism that we use to make such a connection. It includes:

1. A question template, with variables that will be bound by the values provided by the user when the formulate a question based on the template
2. A query template, which contains the variables in the question as well as additional variables to describe the desired characteristics of the dataset. Running a query template returns dataset identifiers
3. One or more workflow identifiers, with their input and output data as variables that can be linked to the dataset identifiers that result from running the query template
4. A meta-workflow with its input data as variables that can be linked to workflow identifiers

In cases when an LOI is created for a hypothesis, the finding returned by the meta-workflow must be in support or against the hypothesis or else a refinement of the hypothesis if evidence was found for it.

When the user poses a question, it is matched against the question template in all available LOIs. The LOIs that match are then triggered, which results in their query being executed, then the workflows, and then the meta-workflow. LOIs are fundamental knowledge structures for science. They are often constructed as the investigation proceeds, and if properly captured they can be reused for subsequent investigations.

Conclusions

We presented a cognitive framework with six distinct phases to model key aspects of scientific reasoning for data analysis. We also showed with examples our representations of the knowledge used about the data, methods, and meta-analysis in the four phases that we have addressed with our work. The framework needs to be extended further to cover the important phases of integrating findings and formulating investigations. Its application to new domains and problems would be needed to evaluate it, as well as characterizing the nature of its limitations and designing appropriate future extensions.

The proposed framework can be used to guide the development of user interfaces and systems for scientific data analysis in new domains, by guiding requirements elicitation, ontology design, and process flow. The framework can also be used to characterize the role of different systems and tools that support scientific data analysis. These benefits are applicable beyond science, as the framework is relevant more generally to the emerging field of data science as a methodological guide. The ultimate goal of this research is automating scientific data analysis tasks in areas of science where significant amounts of data are available, to enable continuous analysis of data to update findings and accelerate discoveries.

Acknowledgements

We would like to thank our collaborators over the years, particularly Emmanuel Johnson, Parag Mallick, Ravali Adusumilli, Neda Jahanshad, Alice Yang, Scott Peckham, Armen Kemanian, and Kelly Cobourn. We gratefully acknowledge support from the US Office of Naval Research through award N00014-21-1-2437, the National Institutes of Health through award 1R01AG059874-01, and the Defense Advanced Research Projects Agency through award W911NF-18-1-0027.

References

Addis, M., Sozou, P.D., Lane, P.C., Gobet, F. (2016). Computational Scientific Discovery and Cognitive Science Theories. In: Müller, V.C. (eds) Computing and Philosophy. Synthese Library, vol 375. Springer, Cham. https://doi.org/10.1007/978-3-319-23291-1_6

Callahan, A., Dumontier, M. & Shah, N.H. (2011). HyQue: evaluating hypotheses using Semantic Web technologies. Journal of Biomedical Semantics 2, S3.

Chandrasekharan, S. & Nersessian, N.J. (2015). Building Cognition: The Construction of Computational Representations for Scientific Discovery. Cognitive Science 39:1727-1763.

Craver, C. F., and L. Darden. (2013). In Search of Mechanisms: Discoveries across the Life Sciences. University of Chicago Press.

Friedland, P.E. and Iwasaki, Y. (1985). The concept and implementation of skeletal plans. Journal of Automated Reasoning 1: 161–208.

Garijo, D.; Fakhraei, S.; Ratnakar, V.; Yang, Q.; Endrias, H.; Ma, Y.; Wang, R.; Bornstein, M.; Bright, J.; Gil, Y.; and Jahanshad, N. (2019). Towards Automated Hypothesis Testing in Neuroscience. Proceedings of the Fifth Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH), held in conjunction with the 45th International Conference on Very Large Data Bases (VLDB).

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

Gil, Y.; Szekely, P.; Villamizar, S.; Harmon, T.; Ratnakar, V.; Gupta, S.; Muslea, M.; Silva, F.; and Knoblock, C. (2011) Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows. Proceedings of the Tenth International Semantic Web Conference (ISWC).

Gil, Y.; McWeeney, S.; and Mason, C. E. (2013). Using Semantic Workflows to Disseminate Best Practices and Accelerate Discoveries in Multi-Omic Data Analysis. AAAI Workshop on Expanding the Boundaries of Health Informatics using AI (HIAI), held in conjunction with the Conference of the Association for the Advancement of Artificial Intelligence (AAAI).

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

Gil, Y.; Garijo, D.; Khider, D.; Knoblock, C. A.; Ratnakar, V.; Osorio, M.; Vargas, H.; Pham, M.; Pujara, J.; Shbita, B.; Vu, B.; Chiang, Y.; Feldman, D.; Lin, Y.; Song, H.; Kumar, V.; Khandelwal, A.; Steinbach, M.; Tayal, K.; Xu, S.; Pierce, S. A.; Pearson, L.; Hardesty-Lewis, D.; Deelman, E.; da Silva, R. F.; Mayani, R.; Kemanian, A. R.; Shi, Y.; Leonard, L.; Peckham, S.; Stoica, M.; Cobourn, K.; Zhang, Z.; Duffy, C.; and Shu, L. (2021). Artificial Intelligence for Modeling Complex Systems: Taming the Complexity of Expert Models to Improve Decision Making. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(2).

Groth, P., J. Cox. (2017). Indicators for the use of robotic labs in basic biomedical research: a literature analysis. PeerJ.

Khider, D.; Athreya, P.; Ratnakar, V.; Gil, Y.; Zhu, F.; Kwan, M.; and Emile-Geay, J. (2020). Towards Automating Time Series Analysis for the Paleogeosciences. Proceedings of the Sixth Workshop on Mining and Learning from Time Series (MiLeTS), held in conjunction with the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'20).

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

Pearl, J. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books Publishers.

Samuels, R. and Wilkenfeld (eds.) (2019). Advances in Experimental Philosophy of Science. London, UK: Bloomsbury.

Senior, A. W., Evans, R., Jumper, J. et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature 577.

Simon, H. A. (1977). Models of Discovery and Other Topics in the Methods of Science. Springer.

Thagard, P. (2012) The Cognitive Science of Science: Explanation, Discovery and Conceptual Change. Cambridge, MA: MIT Press.

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2005). Puzzles and peculiarities: How scientists attend to and process anomalies during data analysis. In M. E. Gorman, R. D. Tweney, D. Gooding, & A. Kincannon (Eds.), Scientific and Technological Thinking (pp. 97-118). Mahwah, NJ: LEA.

Tshitoyan, V., Dagdelen, J., Weston, L., et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98.

World Wide Web Consortium. (2022). Semantic Web Standards. Retrieved from https://www.w3.org/standards/semanticweb/.

Zheng, C. L; Ratnakar, V.; Gil, Y.; and McWeeney, S. K. (2015) Use of Semantic Workflows to Enhance Transparency and Reproducibility in Clinical Omics. Genome Medicine, 7(73).