

Towards Automated Hypothesis Testing in Neuroscience

Daniel Garijo^{*1}, Shobeir Fakhraei^{*1}, Varun Ratnakar¹, Qifan Yang³, Hanna Endrias², Yibo Ma¹, Regina Wang¹, Michael Bornstein³, Joanna Bright³, Yolanda Gil^{1,2}, and Neda Jahanshad³

¹ Information Sciences Institute, University of Southern California
{dgarijo,shobeir,varunr,yiboma,gil}@isi.edu

² Department of Computer Science, University of Southern California
{endrias}@usc.edu

³ Imaging Genetics center, University of Southern California
{qifan.yang,mbornste,joannabr,neda.jahanshad}@usc.edu

Abstract. Scientific data generation in the world is continuous, and often a result of a collaborative process. However, scientific data analyses tend to be focused on a small fragment of recent data acquisitions. In order to leverage this incoming flow of data, we present ENIGMA-DISK, an end-to-end framework to continuously process and organize available data and update the assessment of a given hypothesis as new data become available. Our scope is within the ENIGMA consortium, a large international collaboration for neuro-imaging and genetics whose goal is to understand brain structure and function. ENIGMA-DISK includes an ontology to organize datasets, cohorts, researchers, tools, working groups and organizations participating in ENIGMA via the Organic Data Science framework, and uses an automated discovery framework to continuously test hypothesis through the execution of scientific workflows. We illustrate the usefulness of our system with an example in the neurosciences that assesses confidence of a predefined hypotheses when new patient data become available over time.

Keywords: Hypothesis Evaluation, Scientific Workflow, Ontology, Automated Discovery, Neuroscience

1 Introduction

Medical discoveries are based on hypothesis testing and rigorous data analysis. Such analysis is often time-consuming to perform, error-prone, and includes steps that are hard to systemically reproduce in general settings. Moreover, the designed hypothesis is tested only once against the acquired data sample and later archived, with a slim chance of future re-evaluations. Interestingly, in empirical sciences such as medicine, it is not uncommon for a hypothesis to yield

^{*} equal contribution

contradictory results when re-evaluated against new data samples. In our data-driven world, potentially related data to a hypothesis is being continuously generated but is often not gathered and studied to its full potential for hypothesis re-evaluation. Lack of an integrated system to constantly monitor all the hypotheses of interest and update the underlying analysis when new data become available is one of the challenges for automatic hypothesis re-evaluation. Having such a framework that can keep the hypotheses alive depends on systematically capturing the knowledge about the data and analytics involved in the hypothesis which is often heterogeneous and compartmentalized.

In this paper we propose our solution to address the above challenges based on our previous work for Automated DIScovery of Scientific Knowledge (DISK) [1], and report our ongoing progress. We have extended DISK to explore neuroscience hypothesis and data by generalizing the ability to the system to connect to external knowledge bases. To illustrate our approach, we have cooperated with the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA)⁴ consortium [2], a neuroscience collaboration where projects span many contributors from different institutions around the world. In our proposed solution we address challenges of *data*, *analytics*, and *hypothesis* complexity. The *data* shared through ENIGMA consortium includes multiple levels of heterogeneity and is regularly expanding in volume. The *analytics* related to such data requires the use of dozens of interconnected tools each of which may require substantial domain knowledge. The underlying *hypotheses* span the range of multi-modal neurological and genetic data which could be collected in multiple datasets.

2 Related Work

On the surface, two related research areas in machine learning to our methods are online algorithms [3] (algorithms that revise their models when new data become available), and data stream specific models [4] (that deal with challenges of reprocessing portions of prior data to scale to large data streams). A major advantage of our work over these methods is that our analysis steps do more than just learning from data. For example, some of our steps may include integrating the relevant cohort properties. Another important difference is that our system can react when new kinds of data become available and invoke new analytic tools or algorithms different from the original ones. In addition, distinctive to active agents such as Robot Scientist [5], our method simply listens and reacts to the data that others collect. Moreover, in contrast to other hypothesis evaluation solutions, such as EXPO [6] and HELO [7], our approach represents supporting evidence for hypotheses as reproducible computational components, records their evolution in reaction to new data, and updates their confidence intervals.

⁴ <http://enigma.usc.edu>

3 Background

In this section we describe our domain of focus as well as the sub-components we leverage to develop our solution.

3.1 The ENIGMA Consortium

The ENIGMA consortium [2] is an international network connecting researchers in imaging genomics, neurology and psychiatry, in order to understand brain structure and function, based on multi-modal imaging and genetic data collected from various patient populations. One of the major ambitions of the consortium is to combine various datasets made available via its international partners into larger samples necessary to detect minute gene effects on complex traits that are otherwise not confidently identifiable with smaller isolated samples. Major goals of ENIGMA network include: creating a network of scholars with similar interests in imaging genetics, ensuring reproducibility of major findings through member collaborations, and facilitating information, algorithms and data sharing.

Members of the consortium constantly share new datasets and results, and run experiments and analysis across all available related data. The challenges involved in this global and dynamic collaborative platform, highlights a need to systematically organize its heterogeneous resources to facilitate identification and retrieval of entities of interest. The ENIGMA network would also benefit from a solution to capture the hypotheses under investigation by its members and their related analysis workflows to make them reproducible, especially if such solution could automatically find the related data and dynamically update the analysis results when new data becomes available. In this paper we layout the overall architecture and components of such solution for the ENIGMA consortium and report on our developed prototype.

3.2 Organic Data Science Platform

The ENIGMA consortium is a collaborative effort where users should be able to access, extend and curate past and ongoing efforts. We use the Organic Data Science framework (ODS) [8] for collecting and managing information about ENIGMA (ENIGMA-ODS). ODS is built on Semantic MediaWiki, which uses W3C standards such as RDF and SPARQL to represent its contents in a structured manner. Each wiki page represents a different resource (e.g., a researcher, a project, an organization, etc.) and contains a table with the most relevant properties of that resource’s class. For example, the wiki page of an organization will have *name* and *address* properties. Wiki pages can be filled out by users, who contribute to population and curation of the ENIGMA-ODS knowledge base.

ODS-ENIGMA is structured based to the ENIGMA Ontology network⁵ [9], which extends popular vocabularies such as Schema.org⁶ and includes a representation for datasets, cohorts, persons, organizations, protocols, instruments,

⁵ <https://w3id.org/enigma>

⁶ <http://schema.org/>

software and working groups together with their more common relationships. However, users may extend the ontology with their own properties and categories whenever necessary. The ODS platform is accessible online.⁷

3.3 The DISK framework


DISK [1, 10] is a framework designed to test and revise hypotheses via automatic analysis of dynamic scientific data. DISK evaluates and revises an input hypothesis via continuously examining related data as they become available. It also triggers new kinds of analyses and workflows with the availability of new kinds of data, tracking the provenance of revised hypothesis and its related details. DISK operates based on the description of available data repository metadata, recorded in a data catalog, expressed using domain ontologies with the W3C OWL and RDF Semantic Web standards. Each dataset has a set of metadata assertions, defined in triples of the form $\langle \textit{subject}, \textit{property}, \textit{value} \rangle$, where the *subject* identifies the resource being described (e.g., a dataset), the *property* refers to the aspect of the subject we want to describe (e.g., creation date) and the *value* identifies the value of the property for a resource (e.g., creation date is 2-2-2020). The data catalog supports W3C SPARQL queries to specify the desired metadata properties of datasets.

In order to use DISK, a user defines the hypothesis of interest through a GUI which transforms the hypothesis statements into a machine-readable representation. To evaluate a hypothesis, DISK relies on a library of *Lines of Inquiry* (LOI). A Line of Inquiry includes a hypothesis pattern, a relevant data query pattern, a set of related scientific workflows and one or more meta-workflows to combine workflow results and generate revised confidence values or hypotheses. If a hypothesis matches a Line of Inquiry, the system will search for the appropriate data to pass to compatible workflows for execution. Workflows are executed via WINGS [11], a semantic workflow system for designing scientific computational experiments which specifies the steps and configuration of data processing by software components. The execution results and their corresponding provenance trace are then stored in a Linked Data repository. Finally, the associated meta-workflows explore this repository and revise the original hypothesis if necessary. More details about the DISK framework can be found in [1, 10].

4 The ENIGMA-DISK Framework

We have extended DISK for neuroscience data exploration, analysis execution, and hypothesis testing. The framework integrates the ENIGMA-ODS platform for data search, which has access to all available information from datasets, cohorts, protocols and working groups. Our extension enables newly added and curated datasets to ENIGMA-ODS to be used in assessing existing or new Lines of Inquiry. We validate our framework by testing the hypothesis: “*Is the effect*

⁷ http://organicdatacuration.org/enigma_new/index.php/Main_Page


Lines of Inquiry

Short Description

The BrainCharacteristic of gene is associated with demographic characteristic

Long Description


The BrainCharacteristic of gene is associated with demographic characteristic


Hypothesis Pattern (Ctrl-Space for suggestions)

- 1 ?Gene hyp:hasBrainCharacteristic ?BrainCharacteristic .
- 2 ?BrainCharacteristic hyp:associatedWith ?Demographic .
- 3 ?GeneCharacteristic a neuro:BrainCharacteristic .
- 4 ?Demographic a neuro:DemographicCharacteristic

Data Query Pattern (Ctrl-Space for suggestions)

- 1 ?cohort a ?cohortClass .
- 2 ?cohortClass rdfs:label "Cohort (E)" .
- 3 ?cohort ?datasetProp ?dataset1 .
- 4 ?dataset1 ?featureProp ?Demographic .
- 5 ?dataset1 ?schemaProp ?schema1 .
- 6 ?schema1 ?urlProp ?url1 .
- 7 ?cohort ?datasetProp ?dataset2 .
- 8 ?dataset2 ?featureProp ?BrainCharacteristic .
- 9 ?dataset2 ?schemaProp ?schema2 .
- 10 ?schema2 ?urlProp ?url2 .
- 11 ?datasetProp rdfs:label "HasDataset (E)" .
- 12 ?featureProp rdfs:label "HasFeature" .
- 13 ?schemaProp rdfs:label "Schema:Distribution (E)" .
- 14 ?urlProp rdfs:label "Schema:ContentUrl (E)" .

Workflows to Run 


meta

Variable Bindings: {age = ?url1, hv = ?url2}

Fig. 1. An example of a line of inquiry for assessing if the brain characteristic of a gene such as its effect size is associated with a demographic characteristic such as age.

size of the APOE₄ gene on Hippocampus volume associated with age?". This hypothesis is important in Alzheimer's disease studies, which is the most common neuro-degenerative disorder and severely impacts patients' daily behaviors, thinking, and memory over a wide range of ages [12]. Hippocampus, the most compelling imaging bio-marker [13], has been shown to be particularly vulnerable to Alzheimer's disease pathology, and already considerably damaged at the time clinical symptoms first appear. Several imaging-genetic studies found the correlation between *apolipoprotein E gene allele E₄* (APOE₄), a major genetic risk factor for Alzheimer's disease, and higher rates of hippocampal volume loss [14].

Our hypothesis triggers the Line of Inquiry shown in Figure 1, which studies the correlation between *a genetic effect on a brain characteristic* and *a demographic value*. This hypothesis meets the requirements listed in the *Hypothesis Pattern* section of Figure 1, i.e., APOE₄ being a gene, *effect size* a brain char-

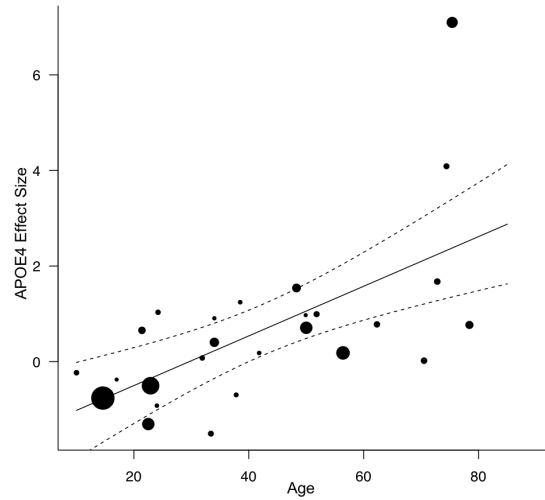


Fig. 2. Meta-analysis for age and the effect size of risk gene linked to Alzheimer’s Disease. Once the effect of APOE4 is observed, age is positively associated with APOE4 effect size per cohort from the ENIGMA consortium ($P = 0.0001$).

acteristic and *age* a demographic value. Once the hypothesis pattern is met, ENIGMA-DISK will aim to find the appropriate datasets to run the workflows associated with the LOI. DISK uses the information under the *Data Query Pattern* section to issue a SPARQL query to the ENIGMA-ODS platform. The query pattern aims to retrieve the dataset URLs (*schema:contentURL*) belonging to the same cohort that contain the target brain characteristic and demographic value. DISK then uses the resultant data URLs, as input to the associated workflow in the LOI (i.e., the “*meta*” workflow in Figure 1). The workflow consists of a meta-analysis to determine whether the magnitude of the target gene (APOE4) effect on a phenotype, is driven by the target demographic (age).

The underlying data of this analysis is based on imaging phenotypes and genotypes of 13,417 participants obtained through ENIGMA consortium [2], scanned at different cohorts worldwide. To configure the workflow, we selected the most significant single nucleotide polymorphisms (SNP) loci closest to the target gene region, *rs283812*, and built a dosage regression model [15] to get the target gene effect size with respect to mean bilateral hippocampal volume as the trait of interest. Age, age^2 , sex, intracranial volume, 4 genetic principal components, disease status and scanner effect (if applicable) were included in the covariates. After a dosage regression model was applied at cohort level, the resulting target gene effect size and mean age of each cohort were used for meta-regression [15, 16]. Figure 2 shows the results of our meta-regression analysis, automatically generated via the ENIGMA-DISK framework after selecting APOE4 as our target gene. Age is positively associated with APOE4 effect size ($P = 0.0001$), which validates the original hypothesis.

5 Conclusion and Future Work

In this paper we described ENIGMA-DISK, a framework to automatically test hypotheses in the neurosciences domain. Our framework integrates the ENIGMA-ODS platform as an external crowd-sourced knowledge base, allowing further testing on previous hypotheses whenever a user contributes new datasets in the system. ENIGMA-DISK is still on its early infancy, but has shown its potential for continuous hypothesis evaluation in this domain. We are working towards addressing three main challenges: improving synchronization between ENIGMA-ODS and ENIGMA-DISK to make the system more adaptive to triggering all compatible Lines of Inquiry with addition of new datasets; facilitating the query patterns to make them more accessible and easy to use for users without SPARQL knowledge; and automatically re-evaluating new hypotheses based on generated workflow results.

Acknowledgment

We are grateful to the KAVLI foundation for their support of ENIGMA Informatics (PIs: Jahanshad and Gil). We also acknowledge support from the National Science Foundation under awards IIS-1344272 (PI: Gil), ICER-1541029 (Co-PI: Gil), and IIS-1344272(PI: Gil), and from the National Institutes of Healths Big Data to Knowledge Grant U54EB020403 for support for ENIGMA (PI: Thompson) and High resolution mapping of the genetic risk for disease in the aging brain grant AG059874 (PI:Jahanshad).

Bibliography

- [1] Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, and Parag Mallick. Automated hypothesis testing with large scientific data repositories. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS)*, 2016. 2, 4
- [2] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, 8(2):153–182, 2014. 2, 3, 6
- [3] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. 2
- [4] Joao Gama. A survey on learning from data streams: current and future trends. *Progress in Artificial Intelligence*, 1(1):45–55, 2012. 2
- [5] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. The automation of science. *Science*, 2009. 2
- [6] Larisa N Soldatova and Ross D King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006. 2

- [7] Larisa N Soldatova, Andrey Rzhetsky, Kurt De Grave, and Ross D King. Representation of probabilistic scientific knowledge. In *Journal of biomedical semantics*, volume 4, page S7. BioMed Central, 2013. 2
- [8] Yolanda Gil and Varun Ratnakar and Paul C. Hanson. Organic data publishing: A novel approach to scientific data sharing. In *Second International Workshop on Linked Science: Tackling Big Data (LISC), held in conjunction with ISWC*, Boston, MA, 2012. 3
- [9] M Jang, Tejal Patted, Yolanda Gil, Daniel Garijo, Varun Ratnakar, Jie Ji, Prince Wang, Aggie McMahon, Paul M Thompson, and Neda Jahanshad. Towards automatic generation of portions of scientific papers for large multi-institutional collaborations based on semantic metadata. In *CEUR workshop proceedings*, volume 1931, pages 63–70, 2017. 3
- [10] Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, Arunima Srivastava, and Parag Mallick. Towards continuous scientific data analysis and hypothesis evolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 4
- [11] Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro Gonzalez-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1):62–72, 2010. 4
- [12] Alzheimer’s Association et al. 2018 alzheimer’s disease facts and figures. *Alzheimer’s & Dementia*, 14(3):367–429, 2018. 5
- [13] Jason L Stein, Sarah E Medland, Alejandro Arias Vasquez, Derrek P Hibar, Rudy E Senstad, Anderson M Winkler, Roberto Toro, Katja Appel, Richard Bartecek, Ørjan Bergmann, et al. Identification of common variants associated with human hippocampal and intracranial volumes. *Nature genetics*, 44(5):552, 2012. 5
- [14] N Schuff, N Woerner, L Boreta, T Kornfield, LM Shaw, JQ Trojanowski, PM Thompson, CR Jack Jr, MW Weiner, and Alzheimer’s; Disease Neuroimaging Initiative. Mri of hippocampal volume loss in early alzheimer’s disease in relation to apoe genotype and biomarkers. *Brain*, 132(4):1067–1077, 2009. 5
- [15] Derrek P Hibar, Hieab HH Adams, Neda Jahanshad, Ganesh Chauhan, Jason L Stein, Edith Hofer, Miguel E Renteria, Joshua C Bis, Alejandro Arias-Vasquez, M Kamran Ikram, et al. Novel genetic loci associated with hippocampal volume. *Nature communications*, 8:13624, 2017. 6
- [16] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer’s disease. *Nature genetics*, 45(12):1452, 2013. 6