Towards Continuous Scientific Data Analysis and Hypothesis Evolution

Yolanda Gil[1], Daniel Garijo[1], Varun Ratnakar[1], Rajiv Mayani[1],
Ravali Adusumilli[2], Hunter Boyce[2], Arunima Srivastava[2], Parag Mallick[2]

[1] Information Sciences Institute, University of Southern California,
4676 Admiralty Way, Marina del Rey CA, 90292, USA. {gil, dgarijo, varunr, mayani}@isi.edu

[2]Stanford School of Medicine, Canary Center for Early Cancer Detection, Stanford University
1265 Welch Road, Stanford CA 94305, USA. {ravali, hboyce, arus, paragm}@stanford.edu

Abstract

Scientific data is continuously generated throughout the world. However, analyses of these data are typically performed exactly once and on a small fragment of recently generated data. Ideally, data analysis would be a continuous process that uses all the data available at the time, and would be automatically re-run and updated when new data appears. We present a framework for automated discovery from data repositories that tests user-provided hypotheses using expert-grade data analysis strategies, and reassesses hypotheses when more data becomes available. Novel contributions of this approach include a framework to trigger new analyses appropriate for the available data through lines of inquiry that support progressive hypothesis evolution, and a representation of hypothesis revisions with provenance records that can be used to inspect the results. We implemented our approach in the DISK framework, and evaluated it using two scenarios from cancer multi-omics: 1) data for new patients becomes available over time, 2) new types of data for the same patients are released. We show that in all scenarios DISK updates the confidence on the original hypotheses as it automatically analyzes new data.

Introduction

In many areas of science, sensors and instruments are continuously collecting data. Yet most research projects analyze data at a particular point in time, and once articles are published they are rarely revisited to account for new data. In some cases, this makes sense since more data may only be tangentially related, and thus may not be relevant to include in a joint analysis. However, in many cases the availability of additional data may significantly affect prior results, by confirming with additional evidence or invalidating them. In addition, the new data may enable new types of analyses, leading to important revisions of prior findings or to entirely new findings.

Our goal is to automatically and continuously analyze scientific data as it becomes available, so scientists can be alerted if their prior studies are affected or if new results are gleamed. In prior work, we developed an approach to represent hypotheses and link them to relevant data to be analyzed, test them through *lines of inquiry* that capture expert-grade data analysis strategies, and aggregate the analysis results through *meta-reasoning* to combine the evidence gathered and generate revised hypotheses and confidence values [Gil et al 2016]. We implemented our approach in the DISK (Automated DIscovery of Scientific Knowledge) framework, and demonstrated that DISK could reproduce the results from a seminal cancer article that reported on comprehensive analyses of two open data repositories.

In this paper, we report on new work to address automated and continuous hypothesis revision as new data becomes available. The novel contributions of this work are: 1) a framework to continuously select appropriate data to test the hypotheses under consideration and to launch appropriate analyses, 2) a hypothesis representation that can represent hypothesis evolution along with supporting evidence, and 3) an implementation of our approach in DISK. We present a preliminary evaluation in cancer multi-omics, where new data and new kinds of observations become available over time and affect prior findings.

Motivating Scenarios from Multi-Omics

In many science domains, new data continuously becomes available to researchers. This is the case with cancer multi-omics, where data for new patients and new kinds of data are generated continuously. Multi-omic analysis enables the study of the genome (genomics data), its products, which include expressed RNAs and proteins (transcriptomics and proteomics data respectively), and how those products interact amongst themselves and with the genome to drive cell behavior (phenotypic data) [Ritchie et al 2015].

Understanding these relationships is crucial to uncover the mechanisms that lead to cancer and other diseases.

Projects like The Cancer Genome Atlas (TCGA) [Tomczak et al 2015] and the associated Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Rudnick et al. 2016] are creating large repositories of omics data that are rapidly approaching petabyte scale. Data is generated by dozens of sites for thousands of patients (and non-patients) with different types of cancer. The data is collected in a well specified and relatively uniform way that facilitates aggregate analysis. These repositories include diverse omics data, such as multiple types of genomic data (DNA sequencing, RNA transcriptomics, epigenetic) and proteomics, as well as pathologic data from biopsy (H+E), radiomic data (CT, MRI), and extensive clinical annotations. Analyses of both TCGA and CPTAC data constantly appear in the literature, but those studies are not commonly revised in light of new data even though their results may be changed in significant ways. For example, TCGA has been growing at a rate of 30 TB/year since its initial construction [Robbins et al 2013, Stephens et al 2015]. Our goal is to develop a framework that can automatically reconsider the hypotheses and results of the initial published studies given the availability of new data. We describe two common scenarios that influence hypothesis testing as new data appears.

Availability of Additional Data of the Same Type

Adjusting Confidence on Hypotheses. High-throughput omics data is continuously being generated. It is often the case that at the beginning of a study, researchers only have access to an initial dataset from a small set of patients. Later on when data for additional patients becomes available, the hypotheses might need to be re-evaluated to update the confidence estimates in light of the new cases. For example, researchers are often interested in testing the hypothesis where a protein of interest is abundant within a given tumor type. If a small set of cases is available, they may find weak evidence for the association. But as more cases are acquired, the scientists may become either increasingly or decreasingly confident in the association.

Testing New Hypotheses. More data also allows scientists to test additional hypotheses. For example, an initial, small dataset may reveal two patient subgroups each defined by a particular set of shared genes. As more cases are added to the study, it may become possible to hypothesize additional subgroups given the larger genetic variations when there are more individuals.

Availability of New Kinds of Data

Considering New Analyses. Different types of data effectively represent alternative sources of observations that can be combined to get stronger evidence for a hypothesis. A challenge in large-scale studies is that data of diverse types will often arrive incrementally and unevenly. For example, in the TCGA and CPTAC studies of colon cancer, transcriptomic data was collected for nearly three years before proteomic data. An initial analysis of the transcriptomic data available could fail to conclusively demonstrate the association of a protein with a tumor type. With both transcriptomics and proteomics data available, a joint analysis could reveal that while the native form of the protein may not be expressed, a mutant form may be. There is an analogous situation when proteomics data is available first. Researchers would look for the standard reference protein instead of a mutant form of the protein as suggested by the genomics data, which would lead to a lack of support for the expression of the protein. Notably, it is quite common for proteomics and genomics data to be poorly correlated [Maier et al 2009]. As multiple types of data become available, the evidence for the hypothesis can change.

Related Work

Machine learning algorithms have some commonalities with our work in that they do some form of hypothesis generation and hypothesis revision [Mitchell 1997]. In general, learning algorithms explore a large hypothesis space and are designed to generate and revise hypotheses (models) as they run. Many algorithms have an *online* version that updates the learned model when new data becomes available [Shalev-Shwartz 2012]. For example, latent Dirichlet allocation (LDA) [Blei et al 2003] is a popular algorithm for topic modeling that can process all the documents available [McCallum 2002]. Its online version builds the topic models as it processes documents incrementally [Hoffman et al 2010; Řehůřek 2009; Langford 2011]. Online algorithms generate an initial model (a hypothesis) and then revise it as they process new data. A major difference with our work is that our analysis steps do more than just learning from data, for example some steps may match a patient's data with a reference human dataset or assembling a group of peptides into a single protein. Another major difference with our work is that we need the ability to formulate new goals and carry out new types of analyses when new data and new kinds of data become available, which may involve new analytic tools or algorithms different form the original ones. We need a framework that can formulate the data analysis and learning goals that can be pursued with the data available, and to change what those goals and analyses are when new data arrives. In addition, given the size of the hypothesis space in the science domains we are tackling, we need the ability to direct the system with some initial hypotheses to be tested, and to be able to decide what data is relevant to these

hypotheses in the first place. Usually in machine learning it is assumed that the system will process all the data that it is given, and in this sense our system is adding a meta-reasoning layer to set up its own analytic and learning goals [Cox and Raja 2011; Kim et al 2011]. Our user hypotheses are akin to meta-level goals, and our lines of inquiry and workflows akin to problem-solving strategies. Our novel contribution to meta-reasoning is the formulation of meta-reasoning goals and strategies for scientific discovery.

Another closely related area of research is learning from streaming data [Gama 2012]. These are systems that process through large amounts of data that is continuously coming in, perhaps in several streams of different types. They are often given a type of pattern (or hypothesis), and their goal is to learn how that pattern manifests in the data. The main challenge addressed in this line of research is memory management, that is, how much prior data to reprocess and how much to simply drop in order to scale to the very large data sizes being streamed. While these approaches always use the same data analysis algorithm, we need a framework that can pursue different kinds of data analysis that are appropriate for the available data.

The Robot Scientist [King et al 2009] is able to formulate hypotheses, test them through physical experiments, and revise them based on the experiment results. The system formulates and generates experiments to collect new data, while in our work we are simply recipients of data that others collect. While the Robot Scientist does one kind of analysis for the same kind of data, we are interested in situations where there may be different kinds of data and different kinds of analyses done.

Other work has proposed models for hypothesis representation. EXPO [Soldatova and King 2006] and nanopublications [Groth et al 2010] define general classes to annotate static hypothesis statements from papers. The HyQue ontology [Callahan et al 2011] describes an event-based model to represent and assess hypotheses created by users, which are then evaluated against a knowledge base integrating information from multiple sources. HELO [Soldatova et al 2013] extends hypothesis statements with probabilities and reasons over them. In contrast, our model represents not just the hypotheses but also their supporting evidence as computational provenance records, as well as their evolution as new data is added and hypothesis statements and confidence scores are updated.

Background on DISK: Automated Hypothesis Testing with Large Data Repositories

DISK offers a novel framework for automating hypothesis analysis over large repositories of scientific data [Gil et al 2016]. We give a brief overview of DISK through a simple example.

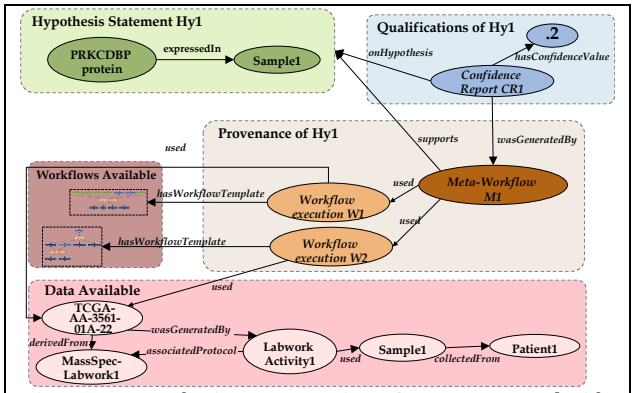

Figure 1. Hypothesis statements in DISK are annotated with statement qualifiers as well as provenance.

DISK assumes that all the datasets available in a data repository are described with metadata in a Data Catalog. Each dataset has a set of metadata assertions. DISK also assumes that these metadata are expressed using domain ontologies represented in the W3C OWL [McGuiness and van Harmelen 2004] and RDF [Manola and Miller 2004] Semantic Web standards. Metadata assertions are triples of the form <object property value>, using namespaces to indicate the source ontology for each object and term. The Data Catalog supports queries that specify the desired metadata properties of the desired datasets. These queries are expressed in the W3C SPARQL query standard [Prud'hommeaux and Seaborne 2008].

DISK is given an initial hypothesis to be investigated. For the science problems that we are interested in, and certainly cancer omics, the size of the hypothesis space (thousands of genes and hundreds of thousands of proteins and all their mutations result in myriad combinations) requires that the system is directed with some initial hypotheses to be tested. Domain ontologies of entities and relationships are used to create logic predicates and assertions to formulate hypotheses. The user provides a hypothesis statement by first selecting a class of hypotheses, and then fill-in-the-blanks to instantiate a specific hypothesis. It is straightforward to add additional types of hypotheses. In DISK, a hypothesis consists of: 1) a *hypothesis bundle*, containing individual *statements*, 2) *statement qualifiers* that are attached to a statement or a group of statements of the hypothesis and typically contain a confidence value, 3) a *hypothesis provenance* with the information of the analyses carried out to gather evidence about the hypothesis, and 4) a *hypothesis history* with links to previous hypotheses that were revised to produce the current one.

Figure 1 illustrates our hypothesis representation with an example, where ovals are objects and arrows are relationships between them. In this case, there is a single hypothesis statement, which is that protein kinase C delta-binding

protein (PRKCDBP) is expressed in a patient sample. The concepts for protein and patientSample both come from a biology ontology, and the expressedIn relation between them is from an ontology that we have created to form statements in this domain. The qualifier shows a confidence value of 0.2. The provenance shows that the confidence value was obtained by running two workflows, W1 and W2, and their results were combined with one meta-workflow, M1. The properties used, derivedFrom, and wasGeneratedBy are from the W3C PROV provenance standard [Lebo et al 2013], and hasWorkflowTemplate from the OPMW ontology to represent workflows [Garijo et al 2016]. We will illustrate the representation of hypothesis history in a later section.

To test a given hypothesis, DISK draws from a library of *lines of inquiry*. A line of inquiry captures a possible approach to hypothesis testing, and includes: 1) a *hypothesis pattern* that specifies what types of hypothesis statements the line of inquiry is designed to test, 2) a *data query pattern* to retrieve relevant data, 3) a set of *computational workflows* that specify the steps to analyze the data selected, and 4) a set of *meta-workflows* that specify how to combine the results of the analyses and generate a revised hypothesis and/or confidence values. For example, to test the type of hypothesis that a protein is associated with a certain tumor type, we could create a line of inquiry for proteomic analysis. The line of inquiry would have a query pattern to retrieve proteomic data from samples taken from patients that have that tumor type, a workflow that would use that data to do a proteomics analysis to look for likely proteins that appear in the samples, and a meta-workflow to generate a confidence value based on the amount of data and the type of algorithms used in the analysis.

Figure 2 illustrates this example of a line of inquiry. Note that the hypothesis pattern draws terms from two namespaces: bio, for a general biology ontology, and hyp, for a small ontology that we have developed to express hypotheses in the domain of multi-omics. The variables in the hypothesis pattern and in the data query pattern are indicated with a question mark (and do not have namespaces), and when the patterns are matched against the hypothesis statement and Data Catalog entries respectively then the variables are bound to matching objects. The query pattern is a SPARQL query that can be sent to the Data Catalog. The workflows and meta-workflow include *bindings*, which express how the variables corresponding to datasets in the Data Catalog should be mapped to input variables of the workflow in question. For meta-workflows, we include additional information about which input variable of the meta-workflow corresponds to the hypothesis, and which output variable of the meta-workflow corresponds to the revised hypothesis.

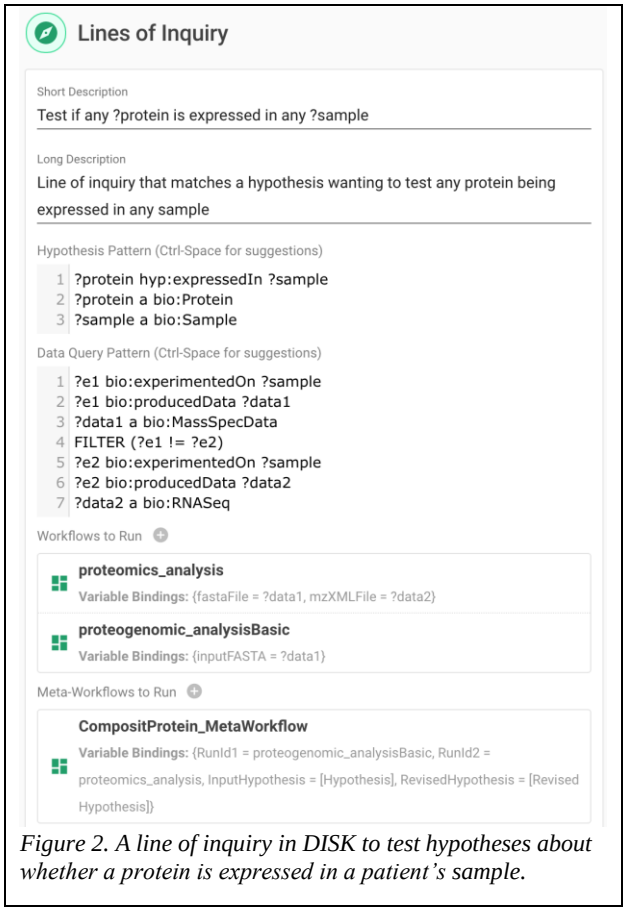

Figure 2. A line of inquiry in DISK to test hypotheses about whether a protein is expressed in a patient's sample.

To test a hypothesis, we match the given hypothesis statement against the hypothesis pattern of all the lines of inquiry available to see which ones are relevant. We then run the query patterns for each to see which ones match with some dataset in the Data Catalog. Those that have matching datasets can be triggered. When a line of inquiry is triggered, its workflows are executed and their results are input to its meta-workflows. DISK uses the WINGS workflow system [Gil et al 2011a; Gil et al 2011b] for both workflows and meta-workflows. The result of a meta-workflow is always a hypothesis revision, which can be either an updated confidence value for the initial hypothesis or a different hypothesis. DISK can formulate new hypotheses by refining user-provided hypothesis statements. For example, in the omics domain, DISK is aware of the concept of mutations. As part of its automated analysis, DISK tests both the user-defined hypothesis and mutation-related revisions thereof. In the future, we plan to extend DISK so it can explore a wider class of new hypotheses based upon examination of user-defined hypotheses. For example, in this domain, DISK might observe that a novel protein (not mentioned by the user) better supports their hypothesis. For example, a proteogenomic analysis may reveal that a mutation of PRKCDBP is present in the

samples, rather than PRKCDBP itself as originally hypothesized.

Confidence values for the hypotheses are set up in the meta-workflows. Each sample analysis leads to counts for each protein. These counts are converted to a probability value (range 0-1) by a domain-knowledge informed posterior of P(obs|count). P(obs|count) can be thought of as an estimator of the probability of a protein being present in a sample given a number of spectral counts that have been measured for it. It is derived from a null model in which spectral counts are measured for proteins that are known not to be present in a particular sample. Multi-sample confidence values are computed by aggregating individual sample p-values relative to a null distribution of proteins known to not be in any of multiple samples.

The DISK framework is designed to be general and applicable to other domains. New domain ontologies, workflows, and meta-workflows would need to be designed. DISK poses strong requirements for the Data Catalog in order to support automated analysis. DISK requires that data is collected in a well-defined and relatively uniform way that facilitates aggregation of the results. It also requires that the Data Catalog has proper metadata, so that the workflows can express constraints about their data requirements. For TCGA and CPTAC all patient samples are collected with standardized protocols, and the datasets contain appropriate metadata.

More details about the general framework and underlying algorithms in DISK can be found in [Gil et al 2016]. The next section describes how we have extended this framework to support continuous data analysis and hypothesis evolution.

Continuous Data Analysis and Hypothesis Revision

The framework that we have described offers a capability to test hypotheses by using available data. However, we need additional capabilities to address the availability of new data or new types of data over time in order to address the motivating scenarios shown earlier. This poses two major challenges:

1. Selection of relevant data and analyses: As new data appears, only data that is relevant to the hypothesis should prompt new analyses. If the new data is irrelevant, nothing should happen. Once data is selected, appropriate analytic methods should be run on the data to reassess the initial hypothesis. Because the new data may represent a small increment over prior data, we need to be mindful use of execution resources and not re-execute analyses unnecessarily.
2. Tracking hypothesis evolution: Given that many potentially independent analyses can be carried out over time, the evolution of the original hypotheses must be documented. This is important for inspectability, reproducibility, and explanation.

There are several aspects that must be addressed. As new data becomes available, there can be potentially many possible analyses that are relevant. Given that there are always limited computing resources, what are appropriate analyses to pursue? There are many nuanced situations. First, combining different types of data is almost always better than analyzing a single type of data. For example, if both proteomics and genomics data are available for the same set of patients it is in principle possible to do a proteomics analysis, a genomics analysis, or a combined proteogenomics analysis, but the combined analysis is always the best course of action. Second, analyzing all the data available is better even if it is unclear how to combine the results. Consider a case where there is proteomics data for a large amount of patients and both proteomics and genomics data only for a different but smaller set of patients. It may be useful to do a proteomics analysis on the former and at the same time a proteogenomic analysis on the latter. It may not be clear how to integrate the results from both analyses, but each analysis could contribute meaningful evidence to the initial hypothesis. We need to track how each dataset and analysis supports the hypotheses, and how the hypotheses evolve over time. Third, some analyses may be preferred to others. For example, in proteomics fluorescence data is easier to obtain but provides weaker evidence than mass spectrometer data, so if both are available then the latter is preferred. Fourth, past analyses done to data should not be re-run when those analyses are independent of the new data. In other words, past analyses should be tracked and the results reused as much as possible rather than unnecessarily executed so that execution resources are made available for other analyses. This requires tracking what data was analyzed by what workflow to test what hypothesis. In summary, the challenge in all these scenarios is to arbitrate and select which analyses to run and how to reflect the results in a revised hypothesis.

Selection of Relevant Data and Analyses

Building on our prior work on representing data analysis strategies as lines of inquiry, we create a selection and prioritization approach to accommodate continuous data analysis and hypothesis revision as follows.

We design lines of inquiry for different combinations of types of data, so that the meta-workflow in each line of inquiry has the method to combine the results of the individual workflows that each analyze some subset of the types of data in that line of inquiry. These meta-workflows are challenging to design and this is a new area of research altogether. When it is unclear how to integrate the analyses of two types of data, then we do not create a com-

bined line of inquiry and instead we will have two separate ones. For example it is unclear how to combine evidence from pathology data and genomics data, so we would create two separate lines of inquiry to analyze each type and each would generate a revised hypothesis reflecting the evidence provided by that type of data.

Lines of inquiry are matched against the original hypothesis rather than against the latest revised hypothesis. The idea is that the new analysis would include the data that was available before plus the new data, so we can produce a new revision that by definition is based on more data than any prior revisions of the initial hypothesis. If more data of the same type is available, then the same lines of inquiry will be triggered, but more data will be matched.

We annotate lines of inquiry based on the coverage of their analyses, and trigger only the broader ones. For example, the analysis done in a proteogenomics line of inquiry is more comprehensive than the analysis done in a proteomics line of inquiry. These coverage annotations are now manually done, but they can be automated through workflow matching techniques that we developed in prior work [Garijo et al 2014]. Essentially, if the workflows of a line of inquiry are subworkflows of the workflows in another one then the latter is more comprehensive.

In summary, lines of inquiry are first matched with their hypothesis pattern, determined applicable if there are datasets retrieved by the query pattern, and then selected for execution based on their coverage annotations.

Prior execution results must be recorded and tracked to detect any opportunities for reuse. For example, a few initial steps of a proteogenomics workflow may be the same as the steps in a proteomics workflow that was run before, and if so those computations should not be repeated. Another situation is cumulative data of the same type. If samples from 50 patients have already undergone proteomics analysis and 20 more patients are added, then the original 50 do not need to be processed again as long as the earlier results are available to be combined with the results of the new 20 samples. We do this by maintaining a provenance catalog where each analysis run is described in detail, including appropriate metadata for the data that was used (e.g., what patient, what kind of omics data, what cancer type). This requires that the initial datasets are well described with metadata in a Data Catalog, as is the case with the cancer multi-omic data in TCGA and CPTAC.

After execution, each of the triggered lines of inquiry results in a revised hypothesis, possibly with a new hypothesis statement (or several) and a revised confidence value together with the workflows and meta-workflows that form their provenance. If only one line of inquiry was triggered (whether to analyze one or several types of data) then we will have a single revised hypothesis. But if, as we mentioned above, there are types of data whose analyses we do not know how to combine then an independent

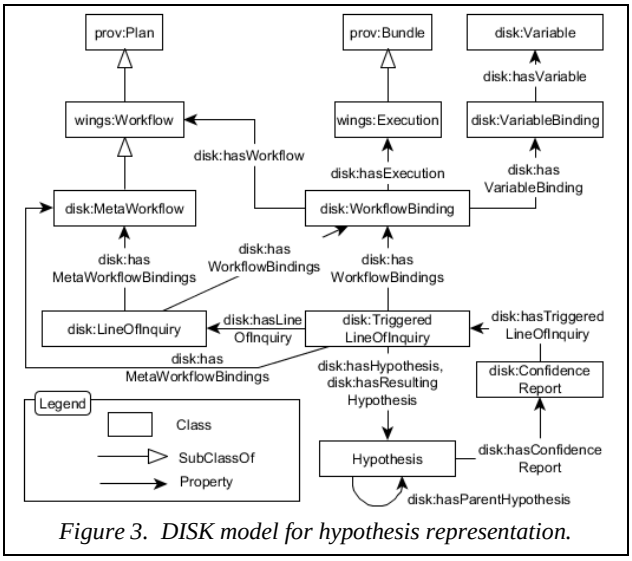

Figure 3. DISK model for hypothesis representation.

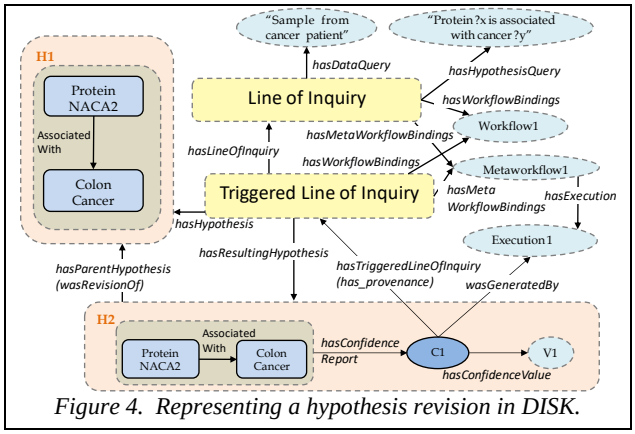

Figure 4. Representing a hypothesis revision in DISK.

line of inquiry will be run for each type of data. Each run will result in a revised hypothesis, each with a potentially different statement and confidence value. All of them will be added as revisions of the initial hypothesis. In all cases, we document the provenance of every revision as we describe next.

A Model for Representing Hypothesis Evolution

Figure 3 gives an overview of the main classes and properties of the model we have developed for representing hypotheses. The namespace prefixes indicate whether a term is reused from a provenance ontology (in our case prov) or from a workflow ontology (in our case wings). The top of the figure shows the representation of the supporting evidence as provenance records of workflows, meta-workflows, and lines of inquiry. These are the concepts that were used in the example of Figure 1. The bottom of the figure focuses on the hypothesis evolution aspects.

Figure 4 sketches an example of a hypothesis revision using our model. The initial hypothesis statement H1 is linked to all the triggered lines of inquiry, each linked in

turn to the original line of inquiry with its hypothesis pattern and its data query pattern as well as the datasets that were retrieved and how they were used as bindings for the analysis workflows. The triggered line of inquiry's meta-workflow produces a revised hypothesis, possibly with a new statement or a new confidence value. This gives the revised hypothesis a detailed provenance record. Finally, the revised hypothesis is linked to the initial hypothesis.

Hypotheses may be composed of several statements, and a different confidence value may be attached to each statement. This is done through the *confidence report* concept, shown as C1 at the bottom of Figure 4. The confidence report links a hypothesis statement to its confidence value and the triggered line of inquiry that generated it. The confidence report also allows assigning different confidence values to the same statement, each resulting from different lines of inquiry. Recall that this is the case when we do not know how to combine the evidence from the analysis with different types of data, so we create separate lines of inquiry that are independently triggered and generate different revised hypotheses.

Hypothesis Evolution through Continuous Data Analysis in DISK

We have implemented our approach in the DISK framework [Ratnakar 2016]. DISK uses a Data Catalog to match the query patterns of its lines of inquiry against the datasets available. When new datasets are available, they are added to this catalog. DISK is now able to run continuously and check its Data Catalog for new datasets in case new ones are added. In addition, DISK now includes a coverage annotation for lines of inquiry, which indicates which lines of inquiry have broader scope so only the broader ones are triggered. We have also extended DISK with a Provenance Catalog that records all the results of workflow executions so that they can be reused in subsequent analyses.

The formal domain-independent ontology to represent DISK hypotheses, lines of inquiry, and the provenance of any revised hypotheses is available online [Ratnakar et al 2016]. As the concepts on the top of Figure 3 show, this vocabulary extends the W3C PROV-O standard provenance model [Lebo et al 2013], which makes it interoperable with other provenance records coming from the data repositories themselves.

In the rest of this section, we describe how DISK addresses continuous analysis of new data and hypothesis revision using data drawn from a seminal TCGA genomic and TCGA/CPTAC proteogenomic study of colon cancer [Zhang et al 2014]. This study spanned nearly 6 years. The data analysis component of the study required more than a year of effort from a team of 6 bioinformaticists using dozens of software tools and analysis scripts. At the time, it was the very first proteogenomic study ever published. This year, 3 additional studies of this scope are being published by authors across the globe. Each of these studies focused on just a small portion of the now available data. DISK will make it possible for continuing ongoing re-analysis of such studies as new data is collected in TCGA and CPTAC.

We use 84 datasets of genomic and proteomic data from 42 different patient samples [Adusumilli 2016], which represents almost half of the cohort used in [Zhang et al 2014]. This enables us to replicate a significant amount of the work, but requires less computation than using all the data from that study. The size of each genomic dataset is around 2 GB, while proteomic files are 6 GB. We have 3 different lines of inquiry with workflows that include popular omics analysis tools such as X!!Tandem [Bjornson et al 2008] and TopHat2 [Kim et al 2013], customProDB [Wang and Zhang 2013], SAMtools [Li 2009], PeptideProphet [Keller et al 2002], and ProteinProphet [Nesvizhskii et al 2003]. When executed linearly, the workflows in the analysis take 336 CPU hours on a single machine. When parallelized, the CPU time is approximately 35 hours. Just the intermediate and final data generated by each sample are 50 GB, with the output of a single run on the 42 samples around 2TB of data. To demonstrate how DISK is able to continuously analyze new data and revise hypotheses, we create several scenarios where different slices of data become available in increments over time. We present now results from those scenarios. Extensive documentation about the data, software, and workflows for the results reported here is in [Adusumilli et al 2016].

Availability of Additional Data of the Same Type

In Scenario I, DISK was given an initial hypothesis that the NACA2 protein is associated with colon cancer. DISK was given initially 20 datasets with 10 RNA-seq files (genomic) and 10 mass-spectrometer files (proteomic) data for the same 10 patients at time t1. Then, DISK was given 20 additional datasets each at time t2, t3, and t4. Table 1 summarizes the scenario, showing the time points when data became available, the line of inquiry triggered, and the revised confidence for the hypotheses. The analysis was not re-executed for data that was analyzed earlier. Note that the confidence value of the hypothesis was reassessed at each time point, but the same value was obtained. When the last dataset was added, the confidence increased. The confidence increased because there is strong protein-level evidence that a mutated form of the NACA2 protein is expressed in the dataset. In addition, because the sample number had increased substantially, it is less likely that the results arose at random, thus increasing the confidence value.

Table 1: Additional data of the same type results in revised confidence values for the initial hypothesis.

Scenario I			
	Datasets	Triggered LOIs	Confidence
t1	20	Proteogenomic	0.5
t2	+20	Proteogenomic	0.5
t3	+20	Proteogenomic	0.5
t4	+24	Proteogenomic	0.845

Table 2: New types of data becoming available results in revised confidence values and perhaps revised hypotheses.

Scenario II				
	Datasets added	Data type	Triggered LOIs	Confidence value
t1	20	RNA-seq	Genomic	0.297
t2	+20	Mass-spec	Proteogenomic	0.5
t3	+22	RNA-seq	Genomic	0.297
t4	+22	Mass-spec	Proteogenomic	0.845

Scenario III				
	Datasets added	Data type	Triggered LOIs	Confidence value
t1	20	Mass-spec	Proteomic	0
t2	+20	RNA-seq	Proteogenomic	0.5
t3	+22	Mass-spec	Proteomic	0
t4	+22	RNA-seq	Proteogenomic	0.845

Availability of New Kinds of Data

DISK was given the same initial hypothesis, and then was given new kinds of data over time. We show two scenarios where the data was given in a different order: Scenario II where genomic data appeared first and then proteomic data, and Scenario III where data appears in reverse order.

Table 2 summarizes the main events in each scenario, showing the time points when new data is available, lines of inquiry triggered, and the revised confidence value for the hypothesis. We describe Scenario II here in detail. The explanation for Scenario III is analogous.

The first datasets available at time t1 were RNA-seq data for 20 patients. This triggered a genomic line of inquiry, which resulted in a confidence value of 0.297.

Next, mass spectrometer data from samples of the same 20 patients became available at time t2. This triggered three lines of inquiry: a genomics one (the same one that was just run but with the additional data), a proteomics one, and a proteogenomics one. Since the proteogenomics one is annotated as having broader coverage than the two others, it was the only one triggered. This analysis resulted in a confidence value of 0.5.

Later on, 22 additional datasets containing RNA sequencing data for 22 new patients became available at time t3. DISK automatically triggered a genomic analysis that included those 22 patients plus the earlier 20. The confidence value did not increase.

Note that at this point, DISK had two revised hypotheses, both with the same statement (NACA2 is associated with colon cancer) but one with a confidence report that had a value of 0.297 linked to the genomic analysis going back to 42 datasets, and another with a confidence report that had a value of 0.5 linked to the proteogenomics analysis based on only 20 datasets.

At t4, mass spectrometer data from the second batch of 22 patients became available. DISK matched three lines of inquiry again, each with the 22+20 datasets, and triggered only the proteogenomics one since it is annotated as having broader coverage. The analysis produced a revised confidence value of 0.845.

Note that the final confidence value is the same in all scenarios because DISK sees the same cumulative data.

Conclusions and Future Work

DISK is a novel framework to test and revise hypotheses based on automatic analysis of scientific data repositories that grow over time. Given an input hypothesis, DISK is able to search for appropriate data to test it and revise it accordingly, and does this continuously as new data becomes available. DISK is also capable of triggering new kinds of analyses when new kinds of data become available. The provenance of the revised hypotheses is recorded, with all the details of the analyses. We have demonstrated DISK using multi-omics data from a seminal cancer study.

Future research includes extending DISK to generate interactive explanations for scientists based on provenance records, developing a general approach to the design of meta-workflows, handling more complex hypotheses, and exploring the use of this approach in other areas of science.

Acknowledgments. We gratefully acknowledge support from the Defense Advanced Research Projects Agency through the SIMPLEX program with award W911NF-15-1-0555, and from the National Institutes of Health under awards 1U01CA196387 and 1R01GM117097. We also acknowledge support from the Canary Foundation.

References

Adusumilli, R. “Datasets used in [Gil et al 2016] for AAAI 2017”. *Zenodo*. 2016. http://doi.org/10.5281/zenodo.180716.

Adusumilli, R., Ratnakar, V., Garijo, D., Gil, Y., and Mallick, P. “Additional materials used in the paper ‘Towards Continuous Scientific Data Analysis and Hypothesis Evolution’ of the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17).” *Zenodo*. 2016. http://doi.org/10.5281/zenodo.190374

Bjornson, R. D., Carriero, N. J., Colangelo, C., Shifman, M., Cheung, K. H., Miller, P.L. and Williams, K. “X!!Tandem, an Improved Method for Running X!Tandem in Parallel on Collec-

tions of Commodity Computers." *Journal of Proteome Research.* 7 (1), 293-299. 2008.

Blei, D., Ng, A., and M. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3, 2003.

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

Cox, M. T., and A. Raja (Eds). *Metareasoning: Thinking about Thinking*. MIT Press, 2011.

Gama, J. "A survey on learning from data streams: current and future trends". *Progress in Artificial Intelligence*, 1(1), 2012. http://doi.org/10.1007/s13748-011-0002-6

Garijo, D., Corcho, O., Gil, Y., Gutman, B. A., Dinov, I. D.,Thompson, P., and Toga, A. W. "FragFlow: Automated fragment detection in scientific workflows". *Proceedings of the 2014 IEEE Tenth International Conference on e-Science*, 2014.

Garijo, D., Gil, Y., and Corcho, O. "Abstract, Link, Publish, Exploit: An End-to-End Framework for Workflow Sharing". To appear in *Future Generation Computer Systems*, 2016.

Gil, Y., Ratnakar, V., Kim, J., Gonzalez-Calero, P. A., Groth, P., Moody, J., and Deelman, E. "Wings: Intelligent Workflow-Based Design of Computational Experiments." *IEEE Intelligent Systems*, 26(1). 2011.

Gil, Y., Gonzalez-Calero, P. A., Kim, J., Moody, J., and Ratnakar, V. "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.

Gil, Y., Garijo, D., Ratnakar, V., Mayani, R., Adusumilli, R., Boyce, H., and Mallick, P. "Automated Hypothesis Testing with Large Scientific Data Repositories". *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS)*, 2016.

Groth, P., Gibson, A., Velterop, J. "The anatomy of a nanopublication". *Information Services and Use*, 30, 1-2: 52-56, 2010.

Hoffman, M., Blei, D., and F. Bach. "Online Learning for Latent Dirichlet Allocation." *Proceedings of the Twenty-Fourth Conference on Neural Information Processing Systems (NIPS)*, 2010.

Keller A., Nesvizhskii A.I., Kolker E., Aebersold R. "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." *Analytical Chemistry*. 74:5383-92, 2002.

Kim, J., Myers, K., Gervasio, M., and Y. Gil. "Goal-directed Metacontrol for Integrated Procedure Learning". In *Metareasoning: Thinking about Thinking*, Cox, M. T., and A. Raja (Eds), MIT Press, 2011.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg S. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". *Genome Biology*. 14:R36. 2013.

King, D. R., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., et al. "The Automation of Science." *Science*, Vol. 324, 3 April 2009.

Langford, J. Vowpal Wabbit. Available from https://github.com/JohnLangford/vowpal_wabbit, 2011.

Lebo, T., McGuiness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. "PROV-O: The PROV Ontology". *W3C recommendation*, 30 April 2013. https://www.w3.org/TR/prov-o/.

Li, H. et al. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25, 2078–2079, 2009.

Maier, T., Güell, M., Serrano, L. "Correlation of mRNA and protein in complex biological samples". *FEBS Letters* 583(24), 17 December 2009, http://dx.doi.org/10.1016/j.febslet.2009.10.036.

Manola, F. and Miller, E. "RDF Primer". *W3C Recommendation*, 10 February 2004. http://www.w3.org/TR/rdf-primer/.

McGuinness, D. and van Harmelen, F. (Eds). "OWL Web Ontology Language Overview." *W3C Recommendation*, 10 February 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

Mitchell, T. *Machine Learning*, McGraw Hill, 1997.

Nesvizhskii A.I., Keller A., Kolker E., Aebersold R. "A statistical model for identifying proteins by tandem mass spectrometry." *Analytical Chemistry*. 75:4646-58, 2003.

Prud'hommeaux, E. and Seaborne, A. (Eds). "SPARQL Query Language for RDF". *W3C Recommendation*, 15 January 2008. https://www.w3.org/TR/rdf-sparql-query/

Ratnakar, V. "DISK software" (v1.0.0). *Zenodo*. 2016. http://doi.org/10.5281/zenodo.168079

Ratnakar, V., Garijo, D. and Gil, Y. "The DISK Ontology" (v1.0.0). 2016. Available from http://disk-project.org/ontology/disk#.

Řehůřek, R. gensim. 2009. Available from http://radimrehurek.com/gensim/.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S.A, and Kim, D. "Methods of integrating data to uncover genotype–phenotype interactions." *Nature Reviews Genetics* 16,85–97. 2015.

Robbins, D. E., Grüneberg, A., Deus, H. F., Tanik, M. M., & Almeida, J. S. "A self-updating road map of The Cancer Genome Atlas." *Bioinformatics*, 29(10), 1333–1340. 2013. http://doi.org/10.1093/bioinformatics/btt141.

Rudnick P. A., Markey S. P., Roth J., Mirokhin Y., et al. "A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline." *J. Proteome Res*, 15(3), 2016.

Shalev-Shwartz, S. "Online Learning and Online Convex Optimization", *Foundations and Trends in Machine Learning* (4)2, pp 107-194, 2012. http://dx.doi.org/10.1561/2200000018.

Soldatova, L. and R.D. King. "An Ontology of Scientific Experiments". *Journal of the Royal Society Interface* (in press). 2006.

Soldatova, L.N., Rzhetsky, A., Grave, D. K., King, R. D. "Representation of probabilistic scientific knowledge". *Journal of biomedical semantics.* 4(Suppl 1): S7. 2013.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., Robinson, G. E. "Big Data: Astronomical or Genomical?" *PLoS Biology*, 13(7), 2015. e1002195. http://doi.org/10.1371/journal.pbio.1002195.

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". *Contemporary Oncology*. 19(1A): A68-A77. 2015.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., et al. "Proteogenomic characterization of human colon and rectal cancer." *Nature* 513,382–387, 2014. http://doi.org/10.1038/nature13438.

Wang X. and Zhang B. "customProDB: an R package to generate customized protein databases from RNA-Seq data." *Bioinformatics*. 29, pp. 3235-3237, 2013.