Data Analytics as Data: A Semantic Workflow Approach

Kristin P. Bennett, John S. Erickson, Hannah de Los Santos,
Spencer Norris, Evan Patton, John Sheehan, Deborah L. McGuinness
Rensselaer Polytechnic Institute
bennek@rpi.edu, erickj4@rpi.edu, delosh@rpi.edi,
norris@rpi.edu, ewpatton@gmail.com, sheehj4@rpi.edu, dlm@cs.rpi.edu

Abstract

By treating the end-to-end data science workflow as data itself and through the conceptual modeling of the goals and functional intent of the data analyst, the entire process of data analytics becomes open and accessible to the powerful tools of artificial intelligence, machine learning, statistics, and data mining. We examine the fundamental questions and capabilities that must be addressed to realize capturing and reasoning over workflows as well as interpreting and contextualizing their results. Our approach focuses on capturing key components of complete workflow processes, making explicit the "deep" semantics of the workflow plan; the analysis performed; the structure and sub-components of the workflow; and intermediate and final data products. Our goal is to provide sufficient detail to facilitate practical workflow and work product integration, interpretation, reuse, reproducibility, recommendation, and search. The structure for this workflow-as-data view is formalized by an extensible, reusable ontology that we are creating that applies to all aspects of the workflow representation and reasoning process. We report on our exploration and reuse of existing methods, tools and ontologies as well as our semantic analytics contributions to real world projects addressing childhood health challenges.

1 Extended Abstract

Applied data science projects typically employ end-to-end workflows composed of multiple steps and potentially many data sources which may themselves be products of other workflows. Capturing these workflows is essential for scientific rigor, reproducibility, reliability, and provenance[1]. Beyond their capture, reasoning over workflows will enable rapidly fundamental advances in data science. If we treat the end-to-end data science workflow as data itself, the entire process of data analytics becomes open and accessible to the powerful tools of artificial intelligence, machine learning, statistics, and data mining. We can study the effectiveness of data analytics methods. We can support automatic end-to-end data analytics through sharing, recommendation and planning of new workflows. We can support workflow and data reuse. We can enable collaborative workflows. We can capture and query knowledge embedded in workflows and data. We can infer and understand relationship between workflows. Through the application of explicit rules, we can analyze and access the security risk and vulnerability of workflows. We examine the fundamental questions and capabilities that must be addressed to realize capturing and reasoning over workflows as well as interpreting and contextualizing their results. We discuss how the pervasive application of semantics can play a foundational role in creating this capacity in transparent, sharable ways. With the pervasive use of data analytics pipelines by many types of users, including and especially non-data scientists, it is imperative that we address the needs to capture, automate, and optimize data analytics pipelines in an integrated fashion.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

What is needed to fully capture and reason over data analytics workflow? Workflows are the embodiment of sequences of operations that apply mathematical principles and scientific concepts, producing results including intermediary data products, visualizations, statistical analytics, and interpretations. As with any machine learning problem, data representation is everything; appropriate representation is the key to effective analysis. What then is the appropriate representation of workflows as well as intermediate results, so that they can be effectively used and interpreted? The following questions must be addressed:

- What are the workflow components, how do they fit together in end-to-end data analytics?
- How can these components be represented to yield meaningful understanding?
- How can the semantics of workflows be captured naturally, in practical settings, without inhibiting the data analytics process?
- What concepts are needed to enable sharing and merging of workflows and resources?
- What are the shared concepts of data and workflow products that are domain dependent or independent?
- How do we express these shared concepts?
- How can we leverage existing tools and best practices including workflow management systems, scientific "notebook" frameworks, code markdown and standard ontologies?

Our work builds on artificial intelligence research on semantics and ontology and research work and semantic workflows[4, 2] in combination with our expertise in machine learning and analysis[6, 8]. Current work on semantic workflows has focused on creating tools to capture the structural aspects of data analytic processes and facilitate their reuse and reproduction, but do not deal with the first-principles synthesis of those workflows, including the sequencing of concept application, the selection of components, the optimization of hyper-parameters, the production of results, and the capturing of work products. With the increasing use of machine learning pipelines by practitioners, it is imperative that a bridge be found to connect their goals with learning models and workflows via optimized, fully-implemented data analytics pipelines.

Our focus is on capturing key components of the complete workflow process, making explicit the key "deep" semantics of the workflow plan; the analysis performed; the structure and sub-components of the workflow; and intermediate and final data products, with sufficient detail to facilitate workflow integration, interpretation, and reuse. We examine an exemplar semantic workflow that captures a particular analysis of a child health informatics dataset. We highlight the semantics of the selection and data preparation processes, the choice and tuning of the analysis methods and hyperparameters, as well as code-level annotations that may be critical for understanding and interpreting the results. We are producing an extensible, reusable ontology for use in all aspects of the workflow representation and reasoning process. To date we have generated a prototype version of the ontology and its potential for guiding the annotation process, and we are developing requirements for a framework for facilitating workflow annotations, linking, and analyses. We will report on our exploration and reuse of some of the existing methods, tools and ontologies (e.g. WINGS[3], Taverna[11], YesWorkflow[7], PROV-O[5], ProvONE[9], Jupyter Notebooks[10]) as well as our semantic analytics contributions to real world projects. We examine our success and challenges to date by applying these methods in multiple projects that examine childhood health challenges.

Acknowledgments

This work was supported by NIH grant U2CES026555-01, NSF grant 1331023, and internal Rensselaer Polytechnic Institute and RPI Tetherless World Constellation funding.

References

[1] John S Erickson, John Sheehan, Kristin P Bennett, and Deborah L McGuinness. Addressing scientific rigor in data analytics using semantic workflows. In *International Provenance and Annotation Workshop*, pages 187–190. Springer, 2016.

[2] Yolanda Gil, Pedro A Gonzalez-Calero, Jihie Kim, Joshua Moody, and Varun Ratnakar. A semantic framework for automatic generation of computational workflows using distributed data and component catalogues. *Journal of Experimental & Theoretical Artificial Intelligence*, 23(4):389–467, 2011.

[3] Yolanda Gil, Varun Ratnakar, and Christian Fritz. Assisting scientists with complex data analysis tasks through semantic workflows. In *AAAI Fall Symposium: Proactive Assistant Agents*, 2010.

[4] Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro A González-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1):62–72, 2011.

[5] Timothy Lebo, Satya Sahoo, Deborah L. McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV ontology. *W3C Recommendation*, 30, 2013.

[6] Deborah L. McGuinness and Kristin P. Bennett. Integrating semantics and numerics: Case study on enhancing genomic and disease data using linked data technologies. In *Proceedings of SmartData 2015, August 18-20 2015, San Jose, CA*, 2015.

[7] Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, Fernando Chirigati, Saumen Dey, Juliana Freire, et al. YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. *arXiv preprint arXiv:1502.02403*, 2015.

[8] Evan W Patton, Elisabeth Brown, Matthew Poegel, Hannah De Los, Chris Fasano Santos, Kristin P Bennett, and Deborah L McGuinness. SemNExT: A framework for semantically integrating and exploring numeric analyses.

[9] DataONE Scientific Workflows Provenance Working Group. ProvONE: a PROV extension data model for scientific workflow provenance. W3C unofficial draft, 27 March 2014.

[10] Helen Shen. Interactive notebooks: Sharing the code. *Nature*, 515(7525):151–152, 2014.

[11] Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, et al. The taverna workflow suite: Designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic acids research*, page gkt328, 2013.