Artificial Intelligence Buzzword Explained: Scientific workflows

Daniel Garijo
Information Sciences Institute
and Department of Computer Science
University of Southern California
dgarijo@isi.edu

The reproducibility of scientific experiments is crucial for corroborating, consolidating and reusing new scientific discoveries. However, the constant pressure for publishing results [1] has removed reproducibility from the agenda of many researchers: in a recent survey published in Nature (with more than 1500 scientists) over 70% of the participants recognize to have failed to reproduce the work from another colleague at some point in time [2]. Analyses from psychology and cancer biology show reproducibility rates below 40 % and 10% respectively [3] [4]. As a consequence, retractions of publications have occurred in the last years in several disciplines [5] [6], and the general public is now skeptical about scientific studies on topics like pesticides, depression drugs or flu pandemics [7].

Reproducing the results of a previous study can be a challenge, as even when the original datasets and end results are available, a significant investment in time may be required [8]. Fortunately, the community has started to pay attention to initiatives for preserving the data and software used in scientific publications (e.g., Zenodo[1], Github[2], etc.). In computational sciences, scientific workflows were proposed in the last decade as a means to address reproducibility. A scientific workflow defines the set of computational tasks and dependencies needed to carry out *in silico* experiments [9]. Typically, scientific workflows are represented as directed graphs, where the nodes represent computational tasks and the edges represent their dependencies. Figure 1 shows an example with two workflows, one for text analytics on the left and another one for neuro-image analysis on the right.

Scientific workflows have been used in many domains, including astronomy [10], brain image analysis [11] and bioinformatics [12]. Besides improving reproducibility, scientific workflows have also proved to be helpful in teaching new users to visualize the overall structure of a method, save time when reusing an existing method and debug or inspect and modularize scientific experiments [13], [14].

There are many challenges associated to scientific workflows. During the last decade plenty of systems have been designed to efficiently represent and execute them in both local and distributed environments (e.g., [12], [15]–[22], etc.). Different approaches have focused in optimizing workflow execution (e.g., [23]) and their results (e.g., [24]). Other works have addressed workflow reuse [14], [25], recommendation [26] [27] and discovery [13], [28], as building on previous findings is considered to be critical to push science forward. Here we overview those aspects of workflows related to reproducibility, i.e., workflow preservation, traceability of the results and workflow sharing.

[1] https://zenodo.org/
[2] http://github.com/

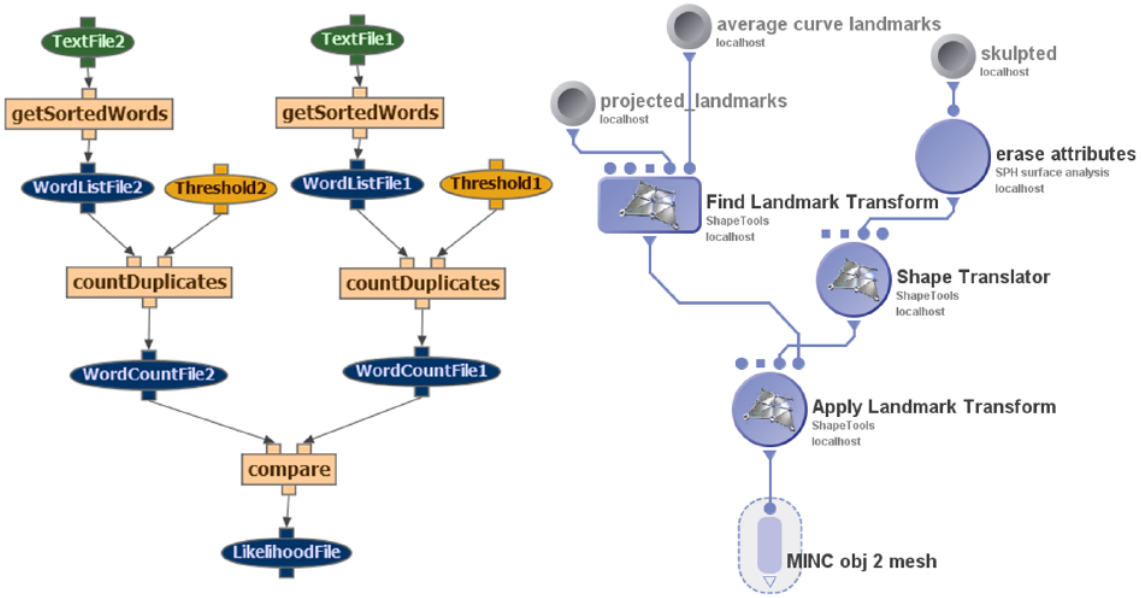

Figure 1: Two scientific workflows from two different workflow systems. The one on the left represents tasks as rectangles and data with ovals, while the one on the right represents task in blue and inputs in grey.

There are two ways in which a workflow may be preserved. The first way is by documenting the method captured by the workflow itself, i.e., providing enough details on each of the tasks of the workflow for anyone to be able to understand their functionality [29], [30]. The rationale is simple: given the pace at which software and data evolve, it is difficult to ensure that within five, ten or twenty years the whole workflow will still be reusable. This is common in domains where scientific workflows rely on external web services and evolving community-built datasets (e.g., the Protein Data Bank[3] in bioinformatics). New releases of software, changes to the existing APIs or new data discoveries may supersede existing resources, making them outdated and sometimes incompatible with the rest of the tasks in the workflow. Therefore, documentation approaches tend to contextualize, describe and generalize the functionality of every dataset and task used in the workflow. Documentation approaches are usually complemented with sample data, pointing to archived versions of the software to facilitate understanding the original method. Another key feature of these approaches includes documenting the *provenance* of the results of a workflow. The provenance of a result aims to capture its creation process, i.e., all the steps that contributed to its outcome, including the original datasets and intermediate data. A provenance record also attributes credit to the scientists responsible for producing the result. There is a standard model for provenance publishing on the web [31], and related work has extended it to publish scientific workflow metadata[4] [29], [30], [32]. Once a workflow is documented, it may be included as part of a repository [33]–[35] for others to reuse.

The second way to preserve workflows is by capturing their functionality in containers (e.g., Docker[5]) or virtual machines. This way the workflow becomes a black box that performs the experiment functionality, including inputs, software and dependencies for execution. The challenge relies in the creation process of such containers. Approaches like [36] monitor the execution of the experiment to create a virtual machine, while approaches like [37] depend on the authors to document the infrastructure details for the workflow. Recent work has proposed a more flexible approach, capturing each of the steps of the

[3] http://www.rcsb.org/pdb/home/home.do
[4] http://vcvcomputing.com/provone/provone.html
[5] https://www.docker.com/

workflow as an independent container [38]. Finally notebooks[6] are gaining a lot of momentum as an alternative lightweight method to encapsulate and test script based experiments.

Scientific workflows have demonstrated to be useful to re-execute, reuse and share the methods and tasks commonly used in a community [14]. Workflows should be treated as first class citizens in cyberinfrastructure [39], since they provide the means of transparent and reproducible work. There are still open challenges in workflows, and venues like eScience[7] and Super Computing[8] discuss and publish new research every year.

REFERENCES

[1] D. Fanelli, "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data," *PLoS ONE*, vol. 5, no. 4, 2010.

[2] Baker, Monya, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604.

[3] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251.

[4] Begley, C.Glen and Lee, M.Ellis, "Drug development: Raise standards for preclinical cancer research," vol. 483, pp. 531–533, Mar. 2012.

[5] A. Marcus and I. Oransky, "Top retractions of 2014," *The Scientist*, Dec. 2014.

[6] J. D. Rockoff, "Amgen Finds Data Falsified in Obesity-Diabetes Study Featuring Grizzly Bears," *Wall Str. J.*, Sep. 2015.

[7] S. American, "In Science We Trust: Poll Results on How you Feel about Science," *Sci. Am.*, Oct. 2010.

[8] D. Garijo *et al.*, "Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome," *PLoS ONE*, vol. 8, no. 11, p. e80278, 2013.

[9] I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, *Workflows for e-Science: Scientific Workflows for Grids*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[10] J. E. Ruiz, J. Garrido, J. D. Santander-Vela, S. Sánchez-Expósito, and L. Verdes-Montenegro, "AstroTaverna: Building workflows with Virtual Observatory services," *Astron. Comput.*, vol. 7–8, pp. 3 – 11, 2014.

[11] I. D. Dinov *et al.*, "Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline," in *Frontiers in Neuroinformatics*, 2009, vol. 3.

[12] K. Wolstencroft *et al.*, "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud," *Nucleic Acids Res.*, 2013.

[13] A. Goderis, "Workflow re-use and discovery in Bioinformatics," School of Computer Science, The University of Manchester, 2008.

[14] D. Garijo *et al.*, "Workflow Reuse in Practice: A Study of Neuroimaging Pipeline Users," in *10th IEEE International Conference on eScience 2014*, 2014.

[15] Y. Gil *et al.*, "Wings: Intelligent Workflow-Based Design of Computational Experiments," *IEEE Intell. Syst.*, vol. 26, no. 1, pp. 62–72, 2011.

[16] E. Deelman *et al.*, "Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems," *Sci. Program.*, vol. 13, no. 3, 2005.

[17] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo, "Vistrails: Visualization meets data management," in *In ACM SIGMOD*, 2006, pp. 745–747.

[18] B. Ludäscher *et al.*, "Scientific workflow management and the Kepler system," *Concurr. Comput. Pract. Exp.*, vol. 18, no. 10, pp. 1039–1065, 2006.

[6] http://jupyter.org/

[7] http://escience-2016.idies.jhu.edu/

[8] http://www.supercomp.org/

[19] R. Filgueira *et al.*, “eScience Gateway Stimulating Collaboration in Rock Physics and Volcanology,” in *e-Science (e-Science), 2014 IEEE 10th International Conference on*, 2014, vol. 1, pp. 187–195.
[20] B. Giardine *et al.*, “Galaxy: a platform for interactive large-scale genome analysis,” *Genome Res.*, vol. 15, no. 10, pp. 1451–1455, Oct. 2005.
[21] I. Taylor, “Triana Generations,” in *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing*, Washington, DC, USA, 2006, p. 143–.
[22] T. Fahringer *et al.*, “ASKALON: A Development and Grid Computing Environment for Scientific Workflows,” in *Workflows for e-Science*, I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, Eds. Springer, 2007, pp. 450–471.
[23] E. Deelman *et al.*, “Pegasus: Mapping Scientific Workflows onto the Grid,” in *Grid Computing*, vol. 3165, M. Dikaiakos, Ed. Springer Berlin / Heidelberg, 2004, pp. 11–20.
[24] S. Holl, “Automated Optimization Methods for Scientific Workflows in e-Science Infrastructures,” University of Bonn, 2014.
[25] A. Goderis, U. Sattler, P. Lord, and C. Goble, “Seven Bottlenecks to Workflow Reuse and Repurposing,” in *The Semantic Web – ISWC 2005*, vol. 3729, Springer Berlin Heidelberg, 2005, pp. 323–337.
[26] J. Zhang, W. Tan, J. Alexander, I. Foster, and R. Madduri, “Recommend-As-You-Go: A Novel Approach Supporting Services-Oriented Scientific Workflow Reuse,” in *In proceeding of IEEE International Conference on Services Computing, SCC*, Washington, DC, 2011, pp. 48–55.
[27] J. Starlinger, B. Brancotte, S. Cohen-Boulakia, and U. Leser, “Similarity Search for Scientific Workflows,” *Proc. VLDB Endow.*, vol. 7, no. 12, pp. 1143–1154, 2014.
[28] R. Bergmann and Y. Gil, “Similarity assessment and efficient retrieval of semantic workflows,” *Inf. Syst.*, vol. 40, pp. 115–127, 2014.
[29] D. Garijo and Y. Gil, “A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data,” in *Proceedings of the 6th workshop on Workflows in support of large-scale science*, Seattle, 2011, pp. 47–56.
[30] K. Belhajjame *et al.*, “Using a suite of ontologies for preserving workflow-centric Research Objects,” *Web Semant. Sci. Serv. Agents World Wide Web*, 2015.
[31] T. Lebo *et al.*, “The PROV ontology, W3C Recommendation,” WWW Consortium, Apr. 2013.
[32] P. Missier, S. Dey, K. Belhajjame, V. Cuevas-Vicenttín, and B. Ludäscher, “D-PROV: Extending the PROV Provenance Model with Workflow Structure,” in *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance*, Lombard, Illinois, 2013, p. 9:1–9:7.
[33] D. D. Roure, C. A. Goble, and R. Stevens, “The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows,” *Future Gener. Comp Syst*, vol. 25, no. 5, pp. 561–567, 2009.
[34] P. Mates, E. Santos, J. Freire, and C. T. Silva, “CrowdLabs: Social Analysis and Visualization for the Sciences,” in *23rd International Conference on Scientific and Statistical Database Management (SSDBM)*, 2011, pp. 555–564.
[35] K. Belhajjame *et al.*, “A workflow PROV-corpus based on Taverna and Wings,” in *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, Genoa, Italy, 2013, pp. 331–332.
[36] F. Chirigati, D. Shasha, and J. Freire, “ReproZip: Using Provenance to Support Computational Reproducibility,” in *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance*, Lombard, Illinois, 2013, p. 1:1–1:4.
[37] I. Santana-Pérez and M. Pérez-Hernández, “Towards Reproducibility in Scientific Workflows: An Infrastructure-Based Approach,” *Sci. Program.*, vol. 2015, p. 11, 2015.
[38] Qasha, Wawaa, Cala, Jacek, and Watson, Paul, “A Framework for Scientific Workflow Reproducibility in the Cloud,” in *IEEE 12th International Conference on eScience*.
[39] Y. Gil *et al.*, “Examining the Challenges of Scientific Workflows,” *Computer*, vol. 40, no. 12, pp. 24–32, Dec. 2007.