A Roadmap to Robust Science for High-throughput Applications: The Scientists' Perspective

M. Taufer*, E. Deelman†, R. Ferreira da Silva†, T. Estrada‡, and M. Hall§
*U. Tennessee Knoxville, †U. Southern California,
‡ U. New Mexico, §U. Utah
Email: *taufer@utk.edu, †{deelman, rafsilva}@isi.edu, ‡trilce@unm.edu, §mhall@cs.utah.edu

2021 IEEE 17th International Conference on eScience (eScience) | 978-1-6654-0361-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/eScience51609.2021.00044

***Abstract*—This poster presents our first steps to define a roadmap to robust science for high-throughput applications used in scientific discovery. These applications combine multiple components into increasingly complex multi-modal workflows that are often executed in concert on heterogeneous systems. The increasing complexity hinders the ability of scientists to generate robust science (i.e., ensuring performance scalability in space and time; trust in technology, people, and infrastructures; and reproducible or confirmable research). Scientists must withstand and overcome adverse conditions such as heterogeneous and unreliable architectures at all scales (including extreme scale), rigorous testing under uncertainties, unexplainable algorithms in machine learning, and black-box methods. This poster presents findings and recommendations to build a roadmap to overcome these challenges and enable robust science. The data was collected from an international community of scientists during a virtual world café in February 2021.**

***Index Terms*—Performance Scalability, Trustworthiness, Reproducibility**

I. Problem and Contributions

High-throughput applications, such that application workload consists of a large ensemble of self-contained tasks and application performance is measured by the number of tasks completed per unit of time, are vital for scientific discovery. High-throughput applications combine multiple components into increasingly complex multi-modal workflows (i.e., data generation; data collection and merging; data pre-processing and feature extraction; data analysis and modelling; and data verification, validation, and visualization) that are executed in concert on large-scale heterogeneous systems including high performance computing, multi-task computing, and cloud platforms. These increasing complexities hinder the **ability of scientists to generate robust science**, which we define as the capacity of high-throughput applications to withstand and overcome adverse conditions such as heterogeneous, unreliable architectures at all scales including extreme scale, rigorous testing under uncertainties, unexplainable algorithms (e.g., in machine learning), and black-box methods [1].

There are three key requirements to achieve robust science:

- **Performance scalability**: high-throughput applications must meet both hardware and software performance expectations when executed despite heterogeneous resources and large scale systems.
- **Trustworthiness**: individuals must trust technology (i.e., hardware and software), people (e.g., collaborators across scientific domains), and organizations hosting the applications' execution and data (e.g., a cloud provider such as IBM, AWS, or Google hosting scientific data) to behave as specified or expected.
- **Reproducibility**: individuals must be able to draw the same scientific conclusions using the knowledge encapsulated in the original computational experiment.

These three requirements are the driving metrics in the work presented in this poster. Specifically, this poster presents findings and recommendations that support designing and implementing robust science across critical high-throughput applications. The findings and recommendations were collected through one virtual mini-workshop in February 2021 [2] called virtual world café based on the world café method. In the virtual world café, we engaged application communities to share needs and recommendations through structured conversational processes in which participants were distributed across several breakout sessions in an online meeting, with participants switching sessions periodically and getting introduced to the previous discussion at their new session by a session lead.

II. Findings and Recommendations

We sort the findings and recommendations in four categories: high-throughput applications; scalability, trustworthiness and reproducibility; machine learning for high-throughput computing workflows; and workforce development.

High-throughput Applications *Findings:* Data-driven and HPC scientific simulations are amenable to high-throughput computing thanks to being easily divided into digestible chunks for concurrent processing. Examples of such applications include: the Transitory Exoplanet Sky Survey (TESS), reproducing GW150914 (i.e., the first observation of gravitational waves from a binary black hole merger), molecular dynamics simulations, neural network architecture search, and connectomics. These applications have defined methods for dividing data and processing chunks in parallel with little to no interaction between chunks. *Recommendations:* (a) Work with the communities to create a taxonomy of data-driven applications and develop standards for data manipulation. (b) Consider developer time and costs when prioritizing the applications to target.

The work in this posters is funded by the National Science Foundation (NSF) under grants #2028881, #2028923, #2028930, #2028955, and #2028956.

978-1-6654-0361-0/21/$31.00 ©2021 IEEE
DOI 10.1109/eScience51609.2021.00044

Scalability, Trustworthiness, and Reproducibility *Findings:* Scalability, trustworthiness, and reproducibility are closely connected. Reproducibility is a requirement for trustworthiness and both affect scalability. The human aspect of trustworthiness also limits scalability. Definitions for these three metrics may change across domains; the lack of an explicit definition in interdisciplinary projects engaging application experts, computer scientists, and CI experts may slow down collaborations. Scalability in high-throughput applications is limited by hardware (e.g., memory bandwidth restrictions, I/O bottlenecks, network bandwidth) and software (e.g., lack of parallelism caused by complex or inefficient communication between processes/nodes/sites, or algorithms that were not well designed). Scalability issues are alleviated by effective data manipulation (e.g., discarding data, selecting compression techniques, storing patterns rather than raw data). Reproducibility is difficult if not impossible at large scale. Resources, funding, and workforce support for reproducibility are often limited. Current incentives to share and publish artifacts include federal agencies' investment (e.g., NSF call for reproducibility in neuroscience), journals and conference initiatives (e.g., Cambridge University Press Experimental Results), and technical society badges (e.g., ACM badges). Those incentives are not sufficient yet. Models and executions are trustworthy if explainable. The end-user trusts simple and understandable models over more complex models even when such models have better accuracy or performance since the complexity determines how easily users can reason on and understand results. Annotated executions are vital for determining trustworthiness in disciplines with rapidly evolving models and data. Stochasticity and "messy" data are major challenges in trustworthiness of high-throughput applications. *Recommendations:* (a) Collect and categorize the meaning of scalability, trustworthiness, reproducibility from different points of view. Work with the community to identify and develop standards. Engage the community to constantly curate the definitions and make sure they are still relevant. (b) Create processes and mechanisms for deciding what data should be kept or thrown out. Store recurring patterns in data rather than the full data set to drastically reduce data storage costs. (c) Establish trust and reproducibility in published work now to avoid building on false results. Make reproducibility studies a standard acceptable component of journals and conferences, removing any stigma of "just reproducing another group's works." Share intermediate results for validation. Design and disseminate tools and APIs that support workflow traceability and are easy to adopt across communities.

Machine Learning for High-throughput Computing Workflows *Findings:* Scientific data from different sources (e.g., biological, astrophysics, materials science) is messy, unsteady, varying, and lacks a general annotation format. It is hard to extract the data/information AI developers need from what is given by the scientists, leading to trust issues in the data and model. Moreover, in most AI applications, there is insufficient validation data. The inner stochasticity in AI (e.g., drop-out layers, random seeds for NN weights) hinders the trustworthiness and reproducibility of the applications. *Recommendations:* (a) Create common annotation standards across applications and scientific domains. Automate the end-to-end pipeline of data generation and analysis (traceability) and executions (explainability), including the sources and manipulation methods of raw data. (b) Provide validation datasets and benchmarks for every high-throughput application. (c) Explain AI models in order to trust them; it is better to have simple and understandable models with comprehensive records more than complex architectures that the end-user cannot trust or replicate the results.

Workforce Development *Findings:* The frantic pace of the academic community exposes the challenging trade-off between performance and scalability vs. trustworthiness and reproducibility. The human participation in raw data processing, analysis of results, or other stages in high-throughput applications impose a bottleneck especially in terms of performance scalability and trustworthiness. Students in particular tend to focus on performance and scalability to the detriment of trustworthiness and reproducibility. Students are often pressured to produce impactful results under tight time constraints for their degree. *Recommendations:* (a) Work with communities to curate standards. Invest resources to train students and scientists in best practices and standardization. (b) Provide necessary infrastructure, such as repositories and tools, to make applications reproducible, scalable, and trustworthy. (c) Foster collaboration from cloud alternatives, GitHub (repositories), ACM Badges, and XSEDE resources to promote scalability, trustworthiness, and reproducibility.

III. Conclusions and Relevance of the Work

The discussions presented in this poster are the first steps to establish a vibrant next generation community that works together to define, design, implement, and use these sets of solutions for robust science. Next steps include building a set of scalable solutions for robust science across and within five critical areas of high-throughput applications: architecture; systems; high performance computing; programming models and compilers; and algorithms and theory. We will combine these areas into a integrated continuum through AI-orchestrations, policies, and practices.

Acknowledgment

The authors want to thank the participants in the February 2021 virtual world café for the vibrant discussions. The findings and recommendations in this manuscript are the results of those discussions.

References

[1] M. Livny, J. Basney, R. Raman, and T. Tannenbaum, Todd "Mechanisms for High Throughput Computing" In SPEEDUP Journal, 11(1), 36–40, 1997.

[2] M. Taufer, E. Deelman, R. Ferreira da Silva, T. Estrda, and M. Hall "Performance Scalability, Trust, and Reproducibility: A Community Roadmap to Robust Science in High-throughput Applications," In https://robustscience.org/