Metrics2023 Conference Report

Submitted by: Ritu Arora, Wayne State University

1. Executive Summary

The Metrics2023 Conference was held from November 11- November 12, 2023, in Denver, Colorado. The main goal of the Metrics2023 conference was to bring together colleagues from a diverse range of disciplines and organizations to network, collaborate, and work towards defining a taxonomy of CyberInfrastructure (CI) projects and the appropriate metrics for their short-term and long-term evaluation. The conference agenda included talks, panel/s, working sessions involving brainstorming in small groups, and presentations from the working groups.

Abstracts for presentation at Metrics2023 and full-papers for journal publication were invited for peer-review. Eight journal papers were accepted and published in a special issue of the Springer Nature computer Science journal, and these are accessible from the following link: https://link.springer.com/journal/42979/topicalCollection/AC_e8e45e9fd5dab093110050bfa0211846 . Slides from some of the presentations/talks at Metrics2023 have been made publicly accessible through the conference website: https://sites.google.com/view/metrics2023 .

An overview of the brainstorming sessions held at Metrics2023 conference and a discussion of the future directions is included below.

2. Ice-Breaking Session

There was an ice-breaking session in the morning of November 11, 2023, during which the event participants introduced themselves to each other and shared responses to the following questions:

- What does success mean to you?
- If you could have only one type of food everyday, what would that be?

3. Brainstorming Sessions

The participants formed small groups to discuss the following topics/questions after lunch on November 11, 2023.

Question 1: How do we build trust in the data related to the metrics for success?

Question 2: What are the challenges related to gathering metrics of success?

Question 3: How are we currently using the data related to our project metrics, to inform or revise our processes/actions/decision-making?

Question 4: What are the considerations for developing or improving on the "responsible metrics" for assessing the CI projects and defining benchmarks for success?

Approximately one hour was allocated for the discussion on each of the aforementioned questions and each group presented the summary of their discussions the next day (on November 12, 2023). All the groups worked independently and were encouraged to pick a moderator/note-taker. All the group members were encouraged to contribute to the discussions. Each question was introduced to the group and clarifications about the context of the questions were provided as needed. For example, the term “Responsible Metrics” [1] was described in the context of question 4. In a nutshell, “Responsible Metrics” imply ethical and appropriate use of the “quantitative indicators in the governance, management and assessment of research” [1].

Each group prepared slides for sharing the key points from their discussions and these slides are available at the following link: https://sites.google.com/view/metrics2023/agenda .

4. Summary of the Brainstorming Sessions

Question 1: How do we build trust in the data related to the metrics for success?

- To a certain extent “trust” is a luxury - even getting data is difficult in many circumstances!
- There can be multiple kinds of trust, such as: do we believe where the data came from, do we believe it is correct, do we believe it is relevant?
- It is important to have robust, mechanistic data collection.
- Information from multiple sources is helpful - consistency between metrics/datasets is a more useful goal than guaranteeing reliability.
- There are multiple perspectives and multiple related metrics.
- There should be transparency about sources of data, how analysis was done, caveats, reproducibility especially in different types of projects.
- A scoring model that accounts for multiple input sources for data and uses combinations of metrics is important.
- Auditing input data is in principle useful, but time-consuming and tedious. Spot checking, identifying suspicious inputs/users/data streams may be more useful.
- It is crucial to be clear about how data is collected and processed.
- How much we care about “trust” depends on the questions we are asking and the audience - it’s important to try to provide qualitative/quantitative estimates of uncertainty.
- “Alt metrics” like social media engagement are valuable as a composite, but unreliable on their own.
- Regarding the validity of citation data, a question arises whether various publications are equally valuable or not.
- There is a need for community standards and international standards or best practices for creating metrics.
- Develop metrics taxonomies and a baseline metrics ontologies—or less formal “playbooks” for common metrics (e.g., demographics, participants or people) - perhaps create a library. As an example, refer to the training material ontology presented in [2].
- There should be advocacy for the use of FAIR practices [3].
- Any metric will be gamed if stakes are high enough.

- Looking at other impact models, such as, Baldrige criteria, can be useful.
- It is important to complement quantitative metrics with qualitative measurements — e.g., by gathering feedback from direct engagement with specific users over time.
- Recognize that the data are imperfect. Perfection is the enemy of progress. There can be missing data or data with errors. For example, the data related to publications or computational jobs run on a system could be missing.
    - Communicate, share best practices and methods for collecting publications.
- Practices for citing/acknowledging CI components should be shared with the community — the computational facilities are not always included in acknowledgements or references of the publications. NSF PAR is a repository for papers that are associated with grants. The computational resources used for producing data/results can be acknowledged/cited in the publications associated with NSF grants and these publications should be deposited in NSF PAR.
- It is important to engage with publishers — the publishers could require authors to identify the facilities used in producing the data/results described in their manuscripts.

Question 2: What are the challenges related to gathering metrics of success?

- Self-reporting is biased and incomplete (e.g., survey data, PI self-reporting of information). Web scraping/searching may also not be perfect if you are looking for citations/acknowledgments.
- Defining the research question (or ”success”) is a crucial step to choosing the metrics; this is not always obvious at the outset and is sometimes iterative.
- Knowing who your stakeholders are and what they care about is crucial, and impacts the gathering of metrics. (And can be a moving target if your stakeholders don’t quite know what they want.)
- We are all struggling with ROI (and possibly the ROI of defining metrics and collecting data in support of them).
- There is a considerable effort across the HPC center community to collect acknowledgements, and more (automated) effort is needed towards this end. There is difficulty in getting human response though and hence there is value in automation.
- Software citation challenges: there are cultural issues / lack of a community standard on citations and insufficient knowledge sharing about the limitations of citations.
- Demonstrating software use via downloads is challenging.
- Managing the different types of data that you have to collect from different resources is a challenge.
- Keeping data safe can be challenging: some data are sensitive, and should not be public.
- It can be challenging to handle large volumes of data, data formats, and data interoperability.
- “Home runs” (success stories) are not always captured by day-to-day operational metrics.
- Practices for citing/acknowledging CI components— information on facilities used is not always included in publications.

- Some metrics are easy to get (e.g., social media mentions) but are not the best measure of what we really care about.
- Number of publications is not a sufficient metric of how much science is being done per watt / core — publications aren't always equal, contributions to publications aren't always equal.
- We can't tell how much productivity is happening just by looking at system usage.
- Collecting metrics can easily fall second to getting the actual work done.
- Long-term impacts are potentially massive but can be difficult to collect metrics for.
- There can be a tendency to overvalue the metrics that we can easily get.
- If you depend on others to collect the data, you have to incentivize getting it.
- There is trouble getting management buy-in to collect metrics we aren't already collecting.
- It can take a lot of effort, e.g., experimental studies, even to determine which metrics should be gathered.
- Projects may be unwilling to report on metrics that might not look good.

Question 3: How are we currently using the data related to our project metrics, to inform or revise our processes/actions/decision-making?

- Data is being used for informing future acquisitions. It helps in getting ahead of what users need rather than reacting.
- Using data for informing future proposals (internal and external), software development efforts, and outreach efforts and strategy.
- Targeted surveys can inform strategic choices about use of funding, but there's a tension between breadth and specificity in terms of number of responses.
- User questions inform public software roadmap.
- User feedback is used for shaping training curriculum (before/after surveys, 3 month follow up).
- System usage data - e.g., job size (multinode vs not) - is used for informing the need for infiniband on future clusters.
- As an example, the experience shared by an HPC evangelist trying to on-board non traditional HPC users led to the adoption of desktop tools, and additional resources were dedicated to that style of use.
- Metrics data has helped in identifying researchers using resources poorly (e.g., idle components of nodes) and working with them.
- Should there be a firewall between projects and stakeholders to allow projects to use data honestly to improve?
- Looking at metrics at greater frequency and more consistently helps reprioritize projects, redirect efforts and costs, see that initiatives are taking too long to start up and would tie up resources and people, and shut-down projects sooner.
- The data is being used to find users that need help, and also to determine which systems aren't working well.

Question 4: What are the considerations for developing or improving on the "responsible metrics" for assessing the CI projects and defining benchmarks for success?

- We need to make sure that the metrics we are collecting can inform the right people at the right level of detail (e.g., CFO vs. research support staff).
- The entire process, from data collection to analysis to presentation, needs to be transparent and clear in order to build trust. Considerations: is the data being collected elsewhere? What are the possible biases? Is there stakeholder buy-in for collecting additional information to inform metrics (this is not cost-free)? Can we get information from alternate sources, even if it requires a lot of data transformation?
- Software packages: it would be helpful for funding agencies to define a core set of metrics they care about and that PIs have to report (e.g., NSF REU or NIH shared instrumentation grants).
- One size doesn't fit all.
- Context is key for understanding and adapting metrics strategy.
- Transparency of metrics—data collection, definitions and results some form of a reality check — make sure the results are realistic and make sense — this is individual and internal to projects.
- It is important to develop a set of Best Practices and share those with the community.
- Automated/repeatable processes for data collection can be developed instead of using manual/ad hoc methods for collecting metrics related data.
- Include a professional evaluator in project proposals.
- It is important to have multiple dimensions and metrics, otherwise it is too easy to game the system.
- It is important to be transparent about error bars/uncertainty in/limitations of the metrics.
- Different kinds of responsibility: social/ethical responsibility, academic responsibility, choosing the right solutions for the right problems is important.
- Qualitative analysis should supplement quantitative assessments.
- Not punishing lack of experience or lack of publications when evaluating researchers for allocations is important. When evaluating allocation proposals, it can be useful to work with the researchers to be more successful next time rather than rejecting them flat-out.
- Don't cherry-pick the data to give the result you want.
- Have checklists and trainings on how to present and evaluate metrics when submitting proposals or reports and when evaluating these.
- Proxy metrics can be responsible or irresponsible, and they can hide the true metric you're looking for.

5. Future Directions

There is a clear need for defining standards or practices for creating metrics for projects. It is also important to ensure that the metrics data is accurate and fair, and is both quantitative and qualitative. To ensure this, it is important to have transparent and automated mechanisms for collecting the data related to the metrics of interest. Metrics2023 provided a forum for discussing such thoughts and fostering collaborations. Planning for the next conference in the Metrics series has begun.

Some of the Participants in the Brainstorming Sessions

Abani Patra (Tufts University)
Anita Orendt (University of Utah)
Ann Backhaus (Pawsey Supercomputing Research Center)
Aaron Weeden (University at Buffalo)
Brian O'Shea (Michigan State University)
Brett Milash (University of Utah)
Dave Hart (National Center for Atmospheric Research (NCAR))
David Bernholdt (Oak Ridge National Laboratory)
D.K. Panda (Ohio State University)
Gil Speyer (Arizona State University)
Greg Dean (University at Buffalo)
Julien Langou (University of Colorado Denver)
Kadidia Konate (Lawrence Berkeley National Lab)
Kerk Kee (Texas Tech University)
Laura Lindzey (University of Washington Applied Physics Laboratory)
Layla Freeborn (University of Colorado-Boulder)
Marlon Pierce (National Science Foundation)
Mary Thomas (San Diego Supercomputer Center)
Matthew Curry (Sandia National Lab)
Minhaz Uddin (Texas Tech University)
Nikolay Simakov (University at Buffalo)
Opeyemi Lawal (Texas Tech University))
Preston Smith (Purdue University)
Roshan Lal Neupane (University of Missouri-Columbia)
Richard Gerber (NERSC / Lawrence Berkeley National Laboratory)
Sukrit Sondhi (Macmillan Learning)
Torey Battelle (Arizona State University)
Weslley Pereira (University of Colorado Denver)
Winona Snapp-Childs (Indiana University)

References

[1] The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management: https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/

[2] N. Hoebelheinrich et al., "Recommendations for a minimal metadata set to aid harmonised discovery of learning resources RDA Supporting Output Education and Training on Handling of Research Data IG," 2022.

[3] M. D. Wilkinson et al., "Comment: The FAIR Guiding Principles for scientific data management and stewardship," Sci. Data, vol. 3, pp. 1–9, 2016.