From Volunteer Computing to Data Democratization: A Personal Journey in Citizen Science and Scientific Discovery

Michela Taufer
University of Tennessee, Knoxville

January 2025

Volunteer Computing and Scientific Discovery

In the early 2000s, computational scientists faced a significant obstacle: acquiring sufficient computing resources to enable groundbreaking scientific discoveries for a glowing community without access to supercomputers. During this period, a groundbreaking paradigm emerged known as volunteer computing (VC). Numerous names have been used to refer to this paradigm, including global computing, crowd computing, and desktop computing. Volunteer computing effectively deploys the unused computational capacity of numerous machines worldwide, democratizing access to computational resources and actively involving the public in scientific discovery.

The effort to deploy volunteer computing resources was initially spearheaded by SETI@home, a groundbreaking project to analyze signals in the search for extraterrestrial life [2]. My research expanded upon this pioneering work by addressing several computational and precision-related obstacles with the goal of extending the applicability of volunteer computing to a broader array of scientific applications. A major breakthrough in my efforts was the successful integration of high-performance computing (HPC) software such as CHARMM [3] for molecular dynamics simulations into a volunteer computing framework, specifically through the Predictor@home project [4]. Predictor@home marked an important achievement as it became the first project to be fully executed using the BOINC (Berkeley Open Infrastructure for Network Computing) platform [1]. This achievement effectively transformed MD simulations, traditionally executed on monolithic supercomputers, to run on both Windows and macOS systems, constituting a distributed supercomputing network for the general public.

Predictor@home emerged as the pioneering initiative in volunteer computing focused on the prediction of protein structures, playing a piloting role in the CASP (Critical Assessment of Protein Structure Prediction) competition. By harnessing the collective power of public resources and engaging the participants owing those resources, the project enabled expansive simulations of protein folding from unknown sequences of amino acids to fully folded ternary structures. Predictor@home, launched in 2004, attracted nearly 100,000 volunteers who contributed more than 264,000 registered devices to protein structure prediction efforts, resulting in 3.3 million hours of computation (equivalent to 380 years) over the 3-month CASPO6.

In my study, I also thoroughly explored the intricate technical difficulties associated with maintaining the reliability of scientific computations across diverse volunteer computing platforms. I introduced the homogeneous redundancy algorithm in the Predictor@home project, which was crucial to achieving computational precision despite the intrinsic variability of the public computing infrastructure [5]. The development of homogeneous redundancy laid a solid foundation for the realization of reproducibility in results in other volunteer computing projects, further facilitating the seamless incorporation of citizen science into large-scale scientific projects and advancing collaborative research by leveraging distributed computing resources.

From Computing to Data: The Transformation of Citizen Science

The establishment of graphics processing units (GPUs) in the mid-2010s started a transformative shift in computational sciences by drastically reducing the financial burden associated with computational processes. However, this enhanced accessibility has been accompanied by a significant downside: the data output has accelerated exponentially. The domain of research that builds on simulations has experienced a profound transformation, moving from a phase where progress was

mainly hindered by limited computational resources to an era now primarily driven by immense volumes of data. This paradigm shift has redefined the dynamics and obstacles associated with citizen science initiatives. Currently, the community has vocally advocated for the data to be openly accessible. However, it is now increasingly apparent that simply open access to generated data is not enough. There is a pressing need for data to be not only open but also easily interpretable and operational, thus streamlining the process of extracting insights and enabling scientific discoveries. This can be achieved through the development of intuitive interfaces and the integration of advanced analytical tools.

My latest research focuses on adding to the ongoing challenges associated with data by innovating systems designed to democratize both access to and use of data. In particular, projects such as the National Science Data Fabric (NSDF) [6] aim to close the gap that exists between the generation of large-scale datasets and their ease of use. The NSDF specifically seeks to establish a federated data-sharing ecosystem that incorporates modular and containerized environments. These environments are intended to reduce barriers to collaboration while simultaneously upholding strict standards of data quality, interoperability, and reproducibility.

For example, large datasets like NASA's 1.8 PB DYAMOND or the 2.8 PB ECCO ocean model highlighted the need for efficient data handling tools, including FAIR (Findable, Accessible, Interoperable, Reusable) principles. These principles were embedded in user-centric tools such as interactive dashboards, Jupyter Notebooks, and metadata-driven catalog systems. Beyond accessibility, these tools empowered researchers by enabling remote queries and real-time data analysis without requiring bulk data transfers.

Enabling Citizen Science Through Data and Computing

Citizen science, increasingly driven by data, has transformed research methodologies by focusing on accessibility and usability. This shift has broadened participation in scientific discovery, particularly engaging communities and institutions traditionally underrepresented in research. However, the growing reliance on data also highlights a critical gap: open science (OS) and open data (OD) infrastructures must evolve to explicitly support citizen science projects. Accessibility alone is inadequate; these infrastructures must also prioritize usability, workforce development, and technical scalability to ensure meaningful engagement.

Addressing these challenges requires a multi-faceted research agenda built on innovative design and pragmatic development, making OS/OD resources more accessible, interpretable, and actionable through:

- Expand OS and OD repositories for citizen science by developing intuitive tools for seamless data contribution, validation, and retrieval, as well as embedding FAIR principles to ensure data is Findable, Accessible, Interoperable, and Reusable, alongside automated data quality metrics for usability and reproducibility.
- Build workforce development and capacity for citizen science by creating training modules and best-practice guidelines to equip project managers and citizen scientists with the skills to leverage OS/OD infrastructures effectively, as well as establishing interdisciplinary collaborations between citizen scientists, professional researchers, and data scientists to strengthen engagement and technical expertise.
- Leverage advanced computing for scalable citizen science by integrating AI/ML-driven tools for automated data processing, anomaly detection, and real-time feedback to improve data integrity, as well as providing interactive dashboards and cloud-based analytics to empower citizen scientists with meaningful insights from their research.

References

[1] D. Anderson. BOINC: a system for public-resource computing and storage. In *Fifth IEEE/ACM International Workshop on Grid Computing*, pages 4–10, 2004.

[2] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer. Seti@home: an experiment in public-resource computing. *Commun. ACM*, 45(11):56–61, Nov. 2002.

[3] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[4] M. Taufer, C. An, A. Kerstens, and C. L. B. III. Predictor@Home: A "Protein Structure Prediction Supercomputer" Based on Global Computing. *IEEE Trans. Parallel Distributed Syst. (TPDS)*, 17(8):786–796, 2006. 10.1109/TPDS.2006.110.

[5] M. Taufer, D. P. Anderson, P. Cicotti, and C. L. B. III. Homogeneous Redundancy: a Technique to Ensure Integrity of Molecular Simulation Results Using Public Computing. In *Proceedings of the 19th International Parallel and Distributed Processing Symposium (IPDPS)*, pages 9 –, Denver, CO, USA, April 2005. IEEE Computer Society.

[6] M. Taufer, H. Martinez, J. Luettgau, L. Whitnah, G. Scorzelli, P. Newel, A. Panta, T. Bremer, D. Fils, C. R. Kirkpatrick, and V. Pascucc. Enhancing Scientific Research with FAIR Digital Objects in the National Science Data Fabric. *IEEE Computing in Science and Engineering (CiSE)*, 25(5):39–47, 2023. 10.1109/MCSE.2024.3363828.