2023

Cyberinfrastructure Deployments on Public Research Clouds Enable Accessible Environmental Data Science Education

Tyler L. McIntosh
University of Colorado Boulder

Erick Verleye
University of Colorado Boulder

Jennifer K. Balch
University of Colorado Boulder

Megan E. Cattau
Boise State University

Nayani T. Ilangakoon
University of Colorado Boulder

See next page for additional authors

Publication Information

McIntosh, Tyler L.; Verleye, Erick; Balch, Jennifer K.; Cattau, Megan E.; Ilangakoon, Nayani T.; Korinek, Nathan; . . . and Wessman, Carol A. (2023). "Cyberinfrastructure Deployments on Public Research Clouds Enable Accessible Environmental Data Science Education". In R. Sinkovits, A. Romanella, S. Knuth, K. Hackworth, and J. Pummill (Eds.), *PEARC '23: Practice and Experience in Advanced Research Computing* (pp. 367-373). Association for Computing Machinery. https://doi.org/10.1145/3569951.3597606

Cyberinfrastructure deployments on public research clouds enable accessible Environmental Data Science education

Tyler L. McIntosh
tyler.l.mcintosh@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA

Erick Verleye
erve3705@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA

Jennifer K. Balch
jennifer.balch@colorado.edu
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Department of Geography, University of Colorado Boulder
Boulder, Colorado, USA

Megan E. Cattau
megancattau@boisestate.edu
Human-Environment Systems, Boise State University
Boise, Idaho, USA

Nayani T. Ilangakoon
ginikanda.ilangakoon@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA

Nathan Korinek
nako1890@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Department of Geography, University of Colorado Boulder
Boulder, Colorado, USA

R. Chelsea Nagy
chelsea.nagy@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA

James Sanovia
james.sanovia@colorado.edu
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA

Edwin Skidmore
edwins@arizona.edu
BIO5 Institute, University of Arizona
Tucson, Arizona, USA

Tyson L. Swetnam
tswetnam@arizona.edu
BIO5 Institute, University of Arizona
Tucson, Arizona, USA

Ty Tuff
ty.tuff@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA

Nathan Quarderer
nathan.quarderer@colorado.edu
Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA
Environmental Data Science Innovation and Inclusion Lab, University of Colorado Boulder
Boulder, Colorado, USA

Carol A. Wessman
carol.wessman@colorado.edu
Ecology & Evolutionary Biology Department, University of Colorado Boulder
Boulder, Colorado, USA
Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Modern science depends on computers, but not all scientists have access to the scale of computation they need. A digital divide separates scientists who accelerate their science using large cyberinfrastructure from those who do not, or who do not have access to the compute resources or learning opportunities to develop the skills needed. The exclusionary nature of the digital divide threatens equity and the future of innovation by leaving people out of the scientific process while over-amplifying the voices of a small group who have resources. However, there are potential solutions: recent advancements in public research cyberinfrastructure and resources developed during the open science revolution are providing tools that can help bridge this divide. These tools can enable access to fast and powerful computation with modest internet connections and personal computers. Here we contribute another resource for narrowing the digital divide: scalable virtual machines running on public cloud infrastructure. We describe the tools, infrastructure, and methods that enabled successful deployment of a reproducible and scalable cyberinfrastructure architecture for a collaborative data synthesis working group in February 2023. This platform enabled 45 scientists with varying data and compute skills to leverage 40,000 hours of compute time over a 4-day workshop. Our approach provides an open framework that can be replicated for educational and collaborative data synthesis experiences in any data- and compute-intensive discipline.

PEARC '23, July 23–27, 2023, Portland, OR, USA

ACM ISBN 978-1-4503-9985-2/23/07.
https://doi.org/10.1145/3569951.3597606

CCS CONCEPTS

• **Computer systems organization** → **Cloud computing**; • **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *Collaborative and social computing*.

KEYWORDS

cyberinfrastructure, open science, digital equity, team science

ACM Reference Format:
Tyler L. McIntosh, Erick Verleye, Jennifer K. Balch, Megan E. Cattau, Nayani T. Ilangakoon, Nathan Korinek, R. Chelsea Nagy, James Sanovia, Edwin Skidmore, Tyson L. Swetnam, Ty Tuff, Nathan Quarderer, and Carol A. Wessman. 2023. Cyberinfrastructure deployments on public research clouds enable accessible Environmental Data Science education. In *Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3569951.3597606

1 INTRODUCTION

Within today's research community there is a digital divide where some researchers regularly use cyberinfrastructure[1] (CI) resources for computationally-intensive science, while others do not or cannot, due to lack of access to CI itself or lack of training opportunities in how to deploy it [1, 2]. Without key CI resources for enabling data-intensive science, some educators and researchers have been excluded from the cloud computing revolution of the last decade. This digital divide in CI mirrors a more broadly understood divide in society driven by financial disparities between educational and governmental institutions, exclusionary policies against underserved communities, and delayed arrival of transformative technologies

[1] In the spirit of accessibility, a glossary of computer science and open science terms used here is provided in Appendix A

like electricity or high speed internet [1, 2]. However, recent advancements in public research CI and an open science revolution around data and software are creating readily available online educational and CI resources for researchers regardless of their home institution or funding portfolios [9]. As research tools, analysis-ready data, and public research CI mature, these resources can be leveraged to enable accessible workshops and working groups. Such learning environments represent a significant step toward enhanced equity, scientific innovation, and ultimately bridging the digital divide.

This divide crosses certain disciplinary boundaries, in particular, sciences that traditionally have not involved large computational problems, e.g. ecology, social sciences. The steep learning curve associated with CI and a widespread lack of foundational digital literacy further exacerbates the divide, creating a need for accessible education and training in those skills. The environmental sciences exemplify this divide, and fulfilling related needs will empower the scientific community to better meet the scale of urgent global environmental challenges [6]. Environmental data science (EDS) is an emerging data-intensive subdiscipline that bridges the computational, biological, environmental, and social sciences, and which has the potential to shape and be shaped by emerging open science resources and tools. EDS is enabled by an explosion in Earth observation data.

This paper demonstrates a successfully deployed and easily-scalable open-source CI architecture and its use during an EDS working group at the University of Colorado (CU) Boulder Earth Lab and the Environmental Data Science Innovation and Inclusion Lab (ESIIL), a next-generation National Science Foundation (NSF) synthesis center. The working group provides a case study of leveraging publicly funded research CI and developments from the open science revolution to bridge the digital divide with a community of environmental scientists, enabling activities previously impossible without direct access to a high performance computing cluster and training to deploy it.

2 A CASE STUDY IN ENVIRONMENTAL DATA SCIENCE EDUCATION AND DATA SYNTHESIS: THE 2023 FOREST RESILIENCY DATA SYNTHESIS WORKING GROUP

The Forest Resiliency Data Synthesis Working Group was held in person at CU Boulder in February 2023. Objectives included improving participants' data skills, providing training in data- and compute-intensive workflows, and promoting synthesis science capabilities and data-driven inquiry. The 4-day working group followed the ESIIL Working Group model, designed to simultaneously integrate the data training institute and synthesis working group models [7].

The working group was attended by 32 participants who represented a wide range of academic institutions, forest resiliency-related disciplines, geographic regions, industries, organizations, and career stages. Participants also had a wide range of data science skills and exposure to computationally-intensive workflows.

Ecosystem resiliency is a subject ideally suited for data-intensive inquiry conducted by a CI-enabled working group. Ecosystem resilience is the ability of a system to experience shocks while retaining function, structure, feedback capabilities, and therefore identity [13]. Efforts to untangle complex interactions between multiple temporally- or spatially-overlapping disturbances at scale can benefit from integrating multiple sources of big data. Applying EDS approaches to research on the resiliency of forested ecosystems is an ongoing process (e.g. complimenting in-situ field measurements with remotely sensed datasets). Computationally-intensive research in this space, paired with knowledge of theory and practice, have the potential to drive improvements in sustainable forest management and benefits for forest-reliant and forest-proximate communities.

3 A REPRODUCIBLE, OPEN-SOURCE CYBERINFRASTRUCTURE FOR COMPUTATIONALLY-INTENSIVE EDUCATION AND DATA SYNTHESIS

The working group used a reproducible and scalable CI deployment consisting of a system of virtual machines (VMs) with copies of a single software container (Fig 1). The CI was composed of public research resources and open source tools orchestrated in partnership with CyVerse [4]. The deployment included 45 VMs, 384 CPUs, 1.4 TB of RAM, and 4.4 TB of collective disk space.

To address the need for accessible CI resources, the NSF has invested heavily in public research computing infrastructure through its TeraGrid (2001-2011), XSEDE (2012-2022) [11], and now ACCESS (2023-) framework. Commercial cloud platforms are highly available but their deployment comes with a significant technical learning curve and a high financial cost that may be out of reach for educators or researchers from small or underserved institutions. NSF ACCESS public platforms have significantly reduced barriers of accessing computational resources and serve as a democratizing force within the scientific computing ecosystem [10]. Jetstream2, an NSF ACCESS public cloud, provided the cloud resources used here [8].

CyVerse and Jetstream2's Cloud Automation and Continuous Analysis Orchestration (CACAO) web platform deployed a multi-instance JupyterHub with Python Data Science Notebooks and RStudio Geospatial Jupyter notebooks. CACAO utilizes Infrastructure as Code (IaC) frameworks: it uses Terraform templates to provision resources from Jetstream2, and Ansible to install and configure the necessary environment. CACAO's simple user interface (UI) abstracts the need for interfacing with these frameworks directly, allowing users to easily configure deployment VMs, shared storage, user authentication, and Docker images. The size of the Docker image and number of requested VMs have the greatest impact on subsequent Jetstream2 provisioning start-up time.

A single Main Node and multiple Worker Nodes were established within the Jetstream2 allocation to host the JupyterHub server. The server was deployed using Project Jupyter's Zero to JupyterHub deployment orchestrated by Kubernetes. The Docker container running on each Worker Node was built using the Jupyter Docker Stacks jupyter/r-notebook:3.0.0 base Image. The Image was augmented with various Python, R, and Linux geospatial packages.

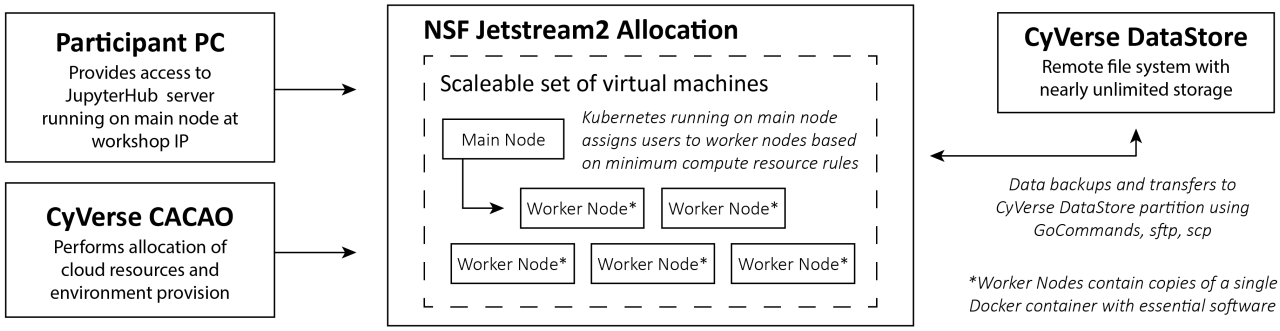

Figure 1: A visual representation of the deployed cyberinfrastructure

End users could write code in a Jupyter Notebook, on an RStudio server, or run Shiny apps from the JupyterHub. Kubernetes pulls and runs the custom Docker image onto each VM ahead of time so as to decrease user wait time. Kubernetes provides hardware constraint rule settings to control user node assignments. For this workshop, each user was given their own Worker Node with 30 G of RAM, ensured by a constraint of at least 20 G of RAM per user in the Kubernetes manifest file.

A Jetstream2 allocation provides limited storage. CyVerse iRODS Data Store was used for project data storage and sharing. GoCommands, sftp, and scp were used to back up and transfer data between the CyVerse Data Store partition and the Jetstream2 allocation. A cron job on the Main Node executed a Bash script which used the Kubernetes kubectl exec command to transfer data backups from Worker Node's home directories to the Data Store.

R and Python scripts for educational modules were pushed to a central GitHub repository and cloned within the deployment. An easily accessible collection of open data was prepared by downloading relevant datasets from Google Earth Engine [5], allowing participant incubation groups to focus on synthesis and analysis rather than data collection. Data was uploaded to the CyVerse Data Store for rapid transfer to shared storage and worker node VMs during the event.

Working group participants were introduced to open science, CI basics, and the deployed CI on the first day of the event. Participants were then immediately able to access the cloud environment with familiar UIs using a JupyterHub server IP address. Instructors performed a CI walkthrough to demonstrate use of the system and provided an explanation of the datasets available.

4 CASE STUDY OUTCOMES

4.1 Cyberinfrastructure integration with working group activities

In the ESIIL working group model participants alternate between learning and practicing data science. Our CI deployment achieved two goals: a) it supported the learning of data science by acting as a platform for individuals to code along with an instructor during interactive modules, and b) supported the practice of data science by providing a large collaborative research space for breakout groups to launch projects, generate ideas, harmonize data, and develop analyses in real time. Co-housed data and computation made the provided datasets quick and easy to import, circumventing many of the network constraints that have plagued previous workshops.

Participants developed data skills and were exposed to new workflows while running real-time, computationally-intensive analyses during two-hour educational modules. Modules involved a mix of running prepared code blocks in R and Python and participating in live coding sessions. The analyses performed would have been impossible or time-prohibitive to complete on individual PCs.

Collaborative, project-focused incubation groups pursued a diverse set of data-oriented projects using the available data datasets and CI. Initial synthesis brainstorming sessions guided by prepared questions jumpstarted these groups. Facilitators synthesized outcomes from early sessions into a set of cross-cutting themes from which groups selected actionable projects. Facilitators emphasized a respectful code of conduct; the value of contributions from all group members; and tractable, data-oriented ideas. These emphases were key to creating a collaborative cohort able to take advantage of the opportunities presented by the deployed CI.

4.2 Reproducibility and Accessibility

Interactive data education environments have a need for reproducible and easily deployed infrastructures. This case study leverages public platforms and contributions from the open science revolution [12]. Open source Infrastructure as Code (IaC) models were used to bypass manual configuration of CI to define and deploy the consistent, stable infrastructure, while Docker was used to create the runtime environment on each VM. The Docker containers used for the event are hosted on Docker Hub with public access. The event contributed to open science through the development of a collaborative open science network and open code-based educational modules that will be hosted on CU Boulder Earth Lab's open EDS education site, EarthDataScience.org.

The deployed CI was highly accessible for users. By bypassing the installation and setup of software on individual PCs with an IP-accessed CI, participants were able to focus on skill development, data synthesis, and science during the working group. Through the development of data science skills, participants were able to see the potential of CI-enabled science and how it could be applied to their own research. To enable future transitions to CI-enabled science for these participants, all participants have registered CyVerse accounts and received free enrollment to CyVerse trainings around foundational open science skills and container orchestration.

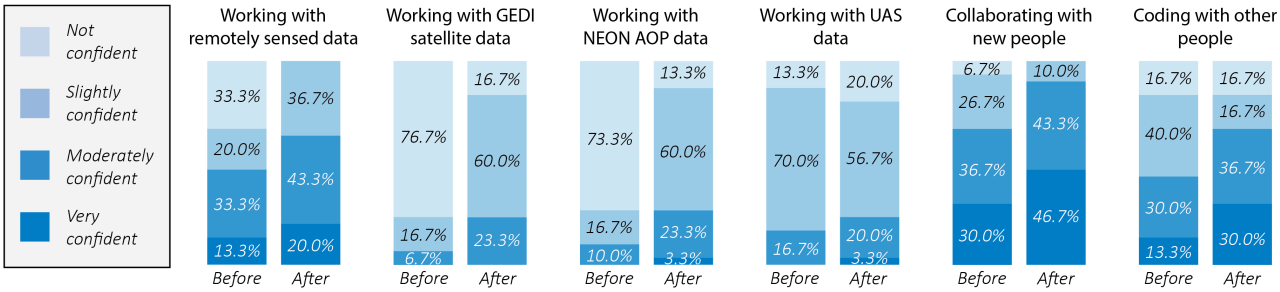

Figure 2: Participant confidence completing tasks independently before and after the working group

4.3 Evaluation

We evaluated working group success[2] with a post-event participant survey. All participants were "extremely" or "very" satisfied with their experience. Respondents identified development of data analysis skills as a top outcome from the event, including learning about relevant workflows, available data, and the potential of data-heavy analysis in their field. Participants highlighted the value of an inclusive and diverse participant cohort,[3] the educational modules, and the synthesis structure that provided opportunities for both brainstorming and concrete research projects.

Participants reported impressive increases in confidence using big data and applying data analytics to their research (Fig 2). 83.3% of participants were "extremely" or "very" satisfied with the deployed CI's ease of use. Fifty percent of project groups have continued meeting after the event with the intent to publish collaborative papers relying on computation-intensive methods.

An identified area of improvement related to the transfer of open science and CI skills to unrelated work and projects. Although participants found the developed system effective and easy to use during the event, many were unsure how to use the same publicly-available research computing resources on their own. This gap is critical to correct in future iterations of similar events. Additionally, the ability to re-assign worker node compute resources mid-deployment would increase the computation capacity of incubation project groups.

4.4 Conclusion

The CI deployed in this case study demonstrates a novel method of leveraging public research computing resources and open source tools to help bridge digital divides in the scientific community. The deployed system enabled a diverse cohort of participants to develop data science and analytics skills, a step toward the larger goal of empowering communities that have been excluded or absent from the scientific computing revolution.

The success of this endeavor is evident in positive evaluations from the event and documented increases in participant confidence using relevant skill sets. The stable CI formed the foundation for both skill development and collaborative incubation activities with a library of readily-available data, enabling the unique ESIIL working group model of simultaneous training and synthesis.

A next step in developing the system deployed here will be testing its scalability and improving delivery of information related to the transferable use of CI and public research computing resources. The outlined infrastructure has already been scaled from use in a prior 10-participant workshop to the 32-participant event described here. The ESIIL and CyVerse teams intend to re-deploy the same system for an event with over 200 participants.

We see the model described here as a reproducible and scalable solution for collaborative data- and compute-heavy learning in any discipline, and hope that it can provide a framework for others. Our success also demonstrates the benefits derived from NSF CI investments, which can be magnified by leveraging them in tandem with open science resources and intentional facilitation.

[2]Qualitative measures of success in the form of participant quotations are provided in Appendix B

[3]Participant demographics and experience with relevant data science skills are described in Appendix C

ACKNOWLEDGMENTS

This work used Jetstream2 at Indiana University through allocation BIO220085 from the NSF ACCESS program, supported by NSF grants OAC-2138259, OAC-2138286, OAC-2138307, OAC-2137603, and OAC-2138296. CyVerse is based upon work supported by the NSF under Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442.

The working group was supported by the Environmental Data Synthesis Innovation and Inclusion Lab (ESIIL, NSF award DBI-2153040) and additional NSF grants DEB-2017889 and DEB-1846384. Additional funding was provided by Earth Lab through the CU Boulder's Grand Challenge Initiative and the Cooperative Institute for Research in Environmental Sciences (CIRES).

Author contributions to this article and the associated working group are outlined using an adjusted CRediT contributor role taxonomy [3]. Writing - Original Draft: T.L.M. (lead) and E.V., T.T., T.L.S. (supporting); Writing - Review and Editing: All authors; Conceptualization: J.K.B., M.E.C., C.A.W., T.L.S., T.T., N.Q., R.C.N., J.S.; Software: E.V., T.L.S., E.S., T.T.; Project Administration: T.L.M.; Visualization: T.L.M. and E.V.; Supervision: J.K.B., R.C.N., T.L.S.; Funding Acquisition: J.K.B., M.E.C., C.A.W., T.L.S., T.T., N.Q., R.C.N; Educational Modules - Development & Teaching: N.T.I. N.K., J.S.; Event facilitation: J.K.B., M.E.C., C.A.W., T.L.M., J.S.

REFERENCES

[1] Nicole A. Buzzetto-Hollywood, Hwei C wang, Magdi Elobeid, and Muna E Elobaid. 2018. Addressing Information Literacy and the Digital Divide in Higher Education. *Interdisciplinary Journal of e-Skills and Lifelong Learning* 14 (2018), 077–093. https://doi.org/10.28945/4029

[2] D. Atkins, K. Droegemeier, Sue Feldman, H. Garcia-Molina, M. Klein, David Messerschmitt, Paul Messina, J. Ostriker, and M. Wright. 2003. Revolutionizing Science and Engineering Through Cyberinfrastructure. *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January* NA, NA (Jan. 2003), 1–82.
[3] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing* 28, 2 (April 2015), 151–155. https://doi.org/10.1087/20150211 Table 1: CRediT - contributor role taxonomy.
[4] CyVerse. 2023. CyVerse: The Open Science Workspace for Collaborative Data-driven Discovery. https://cyverse.org/
[5] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202 (Dec. 2017), 18–27. https://doi.org/10.1016/j.rse.2017.06.031
[6] Stephanie E. Hampton, Matthew B. Jones, Leah A. Wasser, Mark P. Schildhauer, Sarah R. Supp, Julien Brun, Rebecca R. Hernandez, Carl Boettiger, Scott L. Collins, Louis J. Gross, Denny S. Fernández, Amber Budden, Ethan P. White, Tracy K. Teal, Stephanie G. Labou, and Juliann E. Aukema. 2017. Skills and Knowledge for Data-Intensive Environmental Research. *BioScience* 67, 6 (June 2017), 546–557. https://doi.org/10.1093/biosci/bix025
[7] Stephanie E. Hampton and John N. Parker. 2011. Collaboration and Productivity in Scientific Synthesis. *BioScience* 61, 11 (Nov. 2011), 900–910. https://doi.org/10.1525/bio.2011.61.11.9
[8] Jetstream2. 2023. Jetstream2. https://jetstream-cloud.org/
[9] Alondara Nelson. 2022. Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. http://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf
[10] Craig A. Stewart, Amy Apon, David Y. Hancock, Thomas Furlani, Alan Sill, Julie Wernert, David Lifka, Nicholas Berente, Thomas Cheatham, and Shawn D. Slavin. 2019. Assessment of non-financial returns on cyberinfrastructure: A survey of current methods. In *Proceedings of the Humans in the Loop: Enabling and Facilitating Research on Cloud Computing*. ACM, Chicago IL USA, 1–10. https://doi.org/10.1145/3355738.3355749
[11] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* 16, 5 (Sept. 2014), 62–74. https://doi.org/10.1109/MCSE.2014.80
[12] Ruben Vicente-Saez and Clara Martinez-Fuentes. 2018. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research* 88 (July 2018), 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043
[13] Brian Walker, C. S. Holling, Stephen R. Carpenter, and Ann P. Kinzig. 2004. Resilience, Adaptability and Transformability in Social-ecological Systems. *Ecology and Society* 9, 2 (2004), art5. https://doi.org/10.5751/ES-00650-090205

A GLOSSARY OF OPEN SCIENCE AND COMPUTER SCIENCE TERMS

In the spirit of accessibility we provide an alphabetized glossary of open science and computer science terms used.

- Ansible: An open source automation software simplifying application deployment. Allows developers to create scripts to provision runtime environments on computing resources so that deployments can be made reproducible
- CACAO: CyVerse's Cloud Automation and Continuous Analysis Orchestration, one of multiple web-based user interfaces for using and containerizing a Jetstream2 allocation
- Cluster: A collection of virtual machines, usually consisting of a main VM and several worker VM
- Container: A standalone package of software that includes everything needed at runtime to execute an application
- Cyberinfrastructure (CI): The collection of interconnected computing systems, networked devices, software, and data that enable the exchange, processing, storage, and analysis of digital information.
- CyVerse: An NSF-funded project with the mission to provide life scientists with computational infrastructure to handle large datasets and complex analyses
- Disk space: The maximum amount of data that can be added and stored by a disk drive
- Docker: An open source platform that allows developers to build, run, and manage application containers
- Docker Image: Template of instructions for building a docker container that can run on the Docker platform
- GitHub: A collection of git repositories hosted on the web, often used for version control
- Infrastructure as Code (IaC): IaC models are used to define and deploy consistent, stable infrastructures such as networks and virtual machines (VMs). IaC bypass manual configuration of cyberinfrastructure, allowing for the creation of repeatable environments
- Jetstream2: The latest iteration of NSF's distributed cloud computing infrastructure for science and engineering.
- JupyterHub: A framework for hosting a Jupyter Notebook environment server for multiple users. Handles authentication and offers an administration interface.
- Kubernetes: An open-source container orchestration system. For example, a main task of Kubernetes is balancing load across several worker nodes running containerized applications.
- Main node: Node used for orchestrating and balancing the load on a cluster. The main node can be responsible for assigning users to each worker node.
- Node: Any virtual machine inside of a cluster
- Open science: The idea that scientific knowledge (of all kinds and where appropriate) should be openly accessible, transparent, and reproducible
- Open science revolution: A widespread phenomenon in modern science characterized by the adoption of principles and behaviors that promote open science.
- Terraform: An open source Infrastructure As Code (IaC) framework. Allows developers to create scripts that provision cloud resources so that infrastructure deployments can be made reproducible
- User Interface: How a user interacts with an application. This is most commonly a graphical interface with buttons, menus, and image assets.
- Version Control: A system for keeping software with many versions and configurations well organized
- Virtual machine (VM): A computer system that is connected to remotely from a local computer
- Worker node: Node used for running the main application and reporting to the main node

B QUALITATIVE PARTICIPANT SATISFACTION

All participants in the working group presented here as a case study were satisfied with the event, with 73% "Extremely Satisfied" and 27% "Very Satisfied." In addition to quantitative measurements of success, participants provided written feedback. Here we provide a set of quotes from this feedback as they demonstrate the wide-ranging success of the CI-dependent working group and the potential for similar digital divide-bridging events.

- "Learning about the possibilities of big data felt like going from cooking on a camp stove to a state of the art kitchen"
- "We went from conceptual ideas to tangible workflows and even a big data-derived product within a day."
- "The outputs that each group came up with were a testament to the success of the workshop."
- "The amount of ideas generated within such a short amount of time was impressive and inspiring."
- "This truly was the best workshop I have been to."
- "The leaders did a great job of balancing making sure everyone was heard and all ideas were out there with providing time for making tangible progress on the ideas and projects."
- "The combo of tutorials and syntheses was well balanced so that we could think broadly and build off of our skills as we developed them."

C WORKING GROUP PARTICIPANT DEMOGRAPHICS AND EXPERIENCE

The working group described here as a case study was composed of a cohort of 32 participants, many of whom referenced the various dimensions of the group's diversity as a factor contributing to its success.

Most participants (70%) identified as women, while the rest (30%) identified as men. One person identified as both Woman and Gender non-conforming. A majority of participants identified as White (77%). Participants also identified as Hispanic or Latinx (17%) or Asian (10%). (Percentages total higher than 100% because some participants selected multiple identities). Ten percent of participants identified as LGBTQ+, and 3% indicated that they preferred not to answer. Three percent of participants indicated that they have a disability or are neurodiverse, while another 3% preferred not to answer. The majority of participants (77%) reported having one or more parents or guardians who completed a Bachelor's degree or higher. Most participants (97%) were located in the United States, while 3% were located outside of the United States.

In terms of education, most participants had earned a doctoral degree (67%). The remainder had earned either a master's degree (17%) or bachelor's degree (16%). More than half of participants identified their career stage as early career scientists (56%). The remaining 44% of participants were non-academic scientists (27%), mid-career scientists (10%), or senior scientists (7%).

The majority of participants also worked for academic institutions (74%). However, there was also representation from government agencies (13%), non-profit research and development organizations (10%), and private, for-profit organizations (3%). Among those participants who were affiliated with an academic institution, more than half (60%) indicated that their institution was research intensive. About a third (27%) described their institution as serving both undergraduate and graduate students, while 7% reported primarily serving undergraduates. Thirteen percent reported their institution as Hispanic-serving, 3% as American or Native American Pacific Islander-serving, and another 7% as a minority-serving institution. (Percentages total higher than 100% because some participants selected multiple responses to describe their institutional affiliation).

On a self-reported scale from 1-5 (1: low, 5: high), participants had low levels of experience with Python and AWS or other cloud computing options (primarily 1s and 2s), but a high level of experience working in R (primarily 4s and 5s). Participants had widely varying experience with version control systems, remote sensing, and integrating diverse data sources.