CyVerse: Cyberinfrastructure for Open Science

Tyson L. Swetnam[1*], Parker B. Antin[1], Ryan Bartelme[1,12], Alexander Bucksch[1], David Camhy[6], Greg Chism[1], Illyoung Choi[1], Amanda M. Cooksey[1], Michele Cosi[1], Cindy Cowen[1], Michael Culshaw-Maurer[1,11], Robert Davey[3,11], Sean Davey[1], Upendra Devisetty[1,7], Tony Edgin[1], Andy Edmonds[1], Dmitry Fedorov[4], Jeremy Frady[1], John Fonner[2], Jeffrey K. Gillan[1], Iqbal Hossain[1], Blake Joyce[1], Konrad Lang[5], Tina Lee[1], Shelley Littin[1], Ian Mcewen[1], Nirav Merchant[1], David Micklos[8], Andrew Nelson[10], Ashley Ramsey[1], Sarah Roberts[1], Paul Sarando[1], Edwin Skidmore[1], Jawon Song[2], Mary Margaret Sprinkle[1], Sriram Srinivasan[1], Jonathan D. Strootman[1], Sarah Stryeck[5,6], Reetu Tuteja[1,7], Matthew Vaughn[2], Mojib Wali[6], Mariah Wall[1], Ramona Walls[1,9], Liya Wang[8], Todd Wickizer[1], Jason Williams[8], John Wregglesworth[1], & Eric Lyons[1]

1 The University of Arizona, Tucson, Arizona, United States of America
2 Texas Advanced Computing Center, Austin Texas, United States of America
3 Earlham Institute, Norwich, United Kingdom
4 ViQI Inc. Santa Barbara, California, United States of America
5 Know-Center GmbH, Graz, Austria
6 Graz University of Technology, Graz, Austria
7 Greenlight Biosciences, Durham North Carolina, United States of America
8 DNA Learning Center, Cold Spring Harbor Laboratory, Long Island New York, United States of America
9 Critical Path Institute, Tucson, Arizona, United States of America
10 Boyce Thompson Institute, Ithaca, New York, United States of America
11 The Carpentries, Oakland, California, United States of America
12 Pivot Bio, Berkeley, California, United States of America

Current Address: 1657 E Helen St, University of Arizona, Tucson, Arizona, United States of America
* Corresponding Author email: tswetnam@arizona.edu

Abstract

CyVerse, the largest publicly-funded open-source research cyberinfrastructure for life sciences, has played a crucial role in advancing data-driven research since the 2010s. As the technology landscape evolved with the emergence of cloud computing platforms, machine learning and artificial intelligence (AI) applications, CyVerse has enabled access by providing interfaces, Software as a Service (SaaS), and cloud-native Infrastructure as Code (IaC) to leverage new technologies. CyVerse services enable researchers to integrate institutional and private computational resources, custom software, perform analyses, and publish data in accordance with open science principles. Over the past 13 years, CyVerse has registered more than 110,000 verified accounts from 160 countries and was used for over 1,600 peer-reviewed publications. Since 2011, 45,000 students and researchers have been trained to use CyVerse. The platform has been replicated and deployed in two countries outside the US, with additional private deployments on commercial clouds for US government agencies and multinational corporations. In this manuscript, we present a strategic blueprint for creating and managing SaaS cyberinfrastructure and IaC as free and open-source software.

Introduction

CyVerse, a combination of the words 'Cyber' and 'Universe', is the result of 15 years of continuous software development and 13 years in production as a free cyberinfrastructure platform for public research [1, 2]. Representing the largest and longest-running public investment in Plant and Life Science research cyberinfrastructure (117 million USD to date), it operates on free and open-source software (FOSS) and commodity hardware [3–5]. CyVerse's mission is to design, deploy, and expand a national cyberinfrastructure for life sciences research and to train scientists in its use. Here, the term 'cyberinfrastructure' encompasses software, hardware, and the people who can use and train others to operate these systems [6] (see glossary S1 Table).

Initially funded as "The iPlant Collaborative" in 2008 by the United States National Science Foundation (NSF) Directorate for Biological Infrastructure (DBI) for plant sciences and genomics research, the project expanded to support all life sciences in 2013 and rebranded as "CyVerse" in 2016. Today, CyVerse caters to various scientific domains and accelerates data-driven research by connecting public and commercial cloud and high-performance computing (HPC) platforms. It promotes open and collaborative science, reproducible scientific computing [7], and the FAIR data principles [8].

Scientific discoveries increasingly rely on data analysis using complex software systems to manage large datasets and dynamic computational workflows [9]. However, challenges arise due to the lack of specialized programming skills among domain scientists [10–12], barriers to the reproducibility of work by others [7, 13, 14], unequal access to cloud computing [15, 16], and difficulty understanding computational results [8, 17, 19]. To solve some of these problems, researchers have transitioned to cloud and HPC centers, and public research datasets have moved to commercial cloud storage [11, 12, 20–23].

The adoption of cloud-native technologies and techniques for research is currently hindered by a scarcity of educators with necessary cyberinfrastructure skills [10]. Cloud-native science involves building and storing analysis-ready data in cloud-optimized formats [12] and using cloud-native software to analyze the data on distributed computational platforms. Hybrid approaches may involve various data analysis combinations, regardless of software, operating systems, virtualization, or containers, which can run as workflows on HPC, high-throughput computing (HTC), or cloud resources [24]. Researchers must fundamentally change how they conduct and publish their data, code, and results to become cloud-native scientists.

Peer-reviewed reproducible research requires original data, software (dependencies), and instructions for running the analyses [7, 18, 25, 26]. Publishing software and analytical code are additional compliments beyond the requirement of publishing final datasets as part of funding agreements [27, 28]. Analysis-ready data, executable software environments (containers), and analytical code form the basis of digital 'research objects' for replicating methods or building new research on prior results and analyses [29–32]. Cyberinfrastructure, which is capable of hosting research objects, enables replication with new data and promotes reproducible science [14] (Fig 1). CyVerse focuses on two areas of open science: (i) developing spaces for creating research objects, and (ii) operating public cyberinfrastructure to reduce the effort needed to adopt cloud-native science practices.

Scientific Overview

CyVerse is one of only a few publicly available cyberinfrastructure projects which are fully open source (S2 File). Open source cyberinfrastructure levels the field, allowing contributors from small or underserved institutions to contribute directly to the same

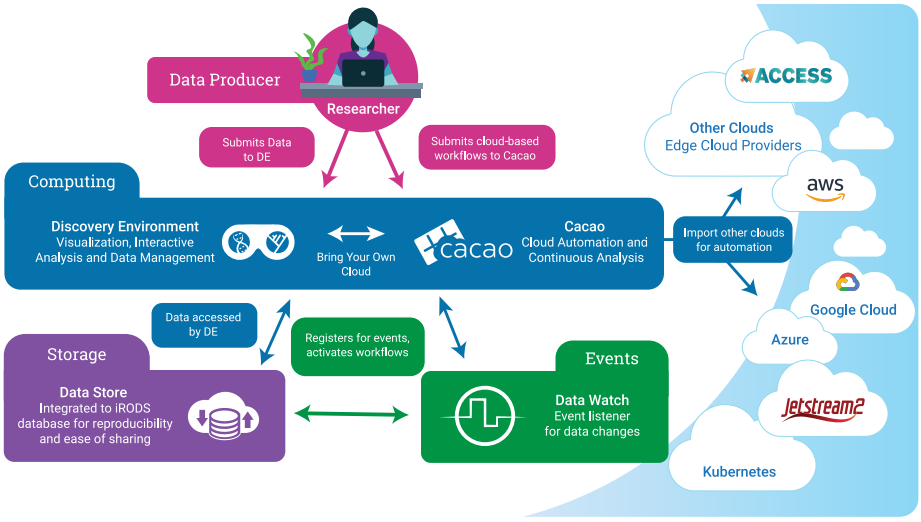

Fig 1. Resources designed for scientists CyVerse offers researchers diverse compute, storage, and event-based workflow options. Data and associated metadata are uploaded and managed in an iRODS Data Store, accessible through various transfer protocols. CyVerse delivers moderate computing power for analytical research via its Discovery Environment data science workbench and connects to public and commercial cloud through the Cloud Automation and Continuous Analysis Orchestration (CACAO) platform. Researchers can register and activate workflows in CyVerse using event-based triggers through the Data Watch API.

science as the most famous and well funded R1 scientists. CyVerse services are designed to manage research data for its entire life cycle [13, 33, 34] and to empower the FAIR [8] and CARE (collective benefit, authority to control, responsibility, and ethics) data principles [35]. By running open source software, researchers can verify every aspect of their computational workflows on the platform. By coordinating continuous analyses with development and operations ('DevOps') Fig 2, CyVerse infrastructure enables researchers to focus on their science and not on the design and deployment of bespoke infrastructure, which have higher creation and operating costs and less reusability. CyVerse education, outreach, and training (EOT) focuses on developing researchers' data science skills, the creation of "research objects", and conducting open science [36–38] via the services and products CyVerse provides.

Technical Overview

CyVerse is both Software as a Service (SaaS) and Infrastructure as Code (IaC), as well as a data hosting service [39, 40] (Fig 1). SaaS relies on the internet to connect users to centrally hosted applications. IaC uses machine-readable definition files (code) to declaratively provision computing and data storage resources rather than having to physically configure hardware and software. CyVerse uses IaC workflows based on Terraform, Ansible, and Kubernetes to enable its interactive data science workbench and cloud environments. As a SaaS, US public CyVerse is run entirely on commodity hardware on-premise (on-site), although it can also be deployed entirely on commercial cloud service providers. As IaC, CyVerse can be dynamically provisioned into larger deployments by leveraging federated Kubernetes clusters and public research cloud (e.g., Jetstream2) resources. CyVerse provides the instructions for deploying its cloud-native services via Ansible Playbooks [41], Argo Workflows [42], Kubernetes (K8s) [43], and Terraform [44] from its public code repositories (S2 Table).

CyVerse utilizes HPC resources at Texas Advanced Computing Center (TACC) and HTC through the OpenScienceGrid, each of which are connected over Internet2 [45] (S1 Fig). CyVerse is also connected to and manages commercial resources on AWS, Google Cloud, and Azure through its external partnership programs. By managing itself as SaaS and IaC, replicated versions of CyVerse can be deployed in a matter of hours. Managing both SaaS and IaC has benefits from an economy of scale perspective and reduces more costly and time consuming manual deployment processes on physical hardware.

CyVerse's customizable, multi-platform data science workbench, called the 'Discovery Environment' (DE), provides a single web-based interface for running executable, interactive, HPC and HTC applications. The DE leverages the CyVerse Terrain API, which is based on Swagger and OpenAPI for most of its functionality. The DE supports dozens of research software analysis pipelines via its HTCondor-based [46] jobs, as well as all popular integrated development environments (IDE), e.g., RStudio [47], JupyterLab [48], and Visual Studio Code [49], which are run via a Kubernetes (K8s) cluster [43]. Once container images are cached on the compute nodes, users can launch their preferred IDE in under 30 seconds.

Design and Implementation

All CyVerse APIs, SaaS, and IaC source code are provided via its public GitHub and GitLab organizations (S2 Table). DevOps and User documentation are hosted online along with manuals, tutorials, and walkthroughs (S3 Table). CyVerse featured software stack can be best described as a 'layer cake' (Fig 2). Conceptually, each layer targets a different set of users and use cases relative to specific scientific objectives. Functionally, most researchers interact with the Products, shown in blue (Fig 2) which are supported by graphic interfaces via a web-browser. Research Software Engineers and/or DevOps personnel with advanced programming expertise can take advantage of the foundational services, shown in purple (Fig 2) and hardware resources (shown in green) for customized applications. In the Supporting Information section we provide additional details about CyVerse physical resources, services, and third party SaaS products which CyVerse helps manage for its user community.

Results

CyVerse has been cited in 1,695 peer-reviewed articles as of 2023 (S1 File). CyVerse has collaborated on or enabled over 50 externally funded projects in the last five years through its "Powered by CyVerse" framework (S4 Table). Each of these leverage different aspects of CyVerse's available SaaS and IaC (S3 Table,S4 Table,S5 Table).

See S1 Appendix for statistics about CyVerse's tools, integrated apps (S4 Table, data store usage (S6 Fig,S6 Table), geographic user distribution (S11 Fig).

Education Outreach & Training

The CyVerse Learning Center, an internal team created in 2017, has taught 11 professional workshops (4 in-person; 7 remote) to 456 principal investigators, early-career faculty, post-doctoral researchers, graduate students, and professionals. Since 2015, Cyverse and community experts have presented 79 webinars on platform, software, and science topics to 2,300 attendees (6,046 registrants), and have amassed more than 83,100 views on YouTube (Fig 3).

A core component of CyVerse's user-support success is rapidly answering questions from researchers, students, and learners of all career stages via a content management

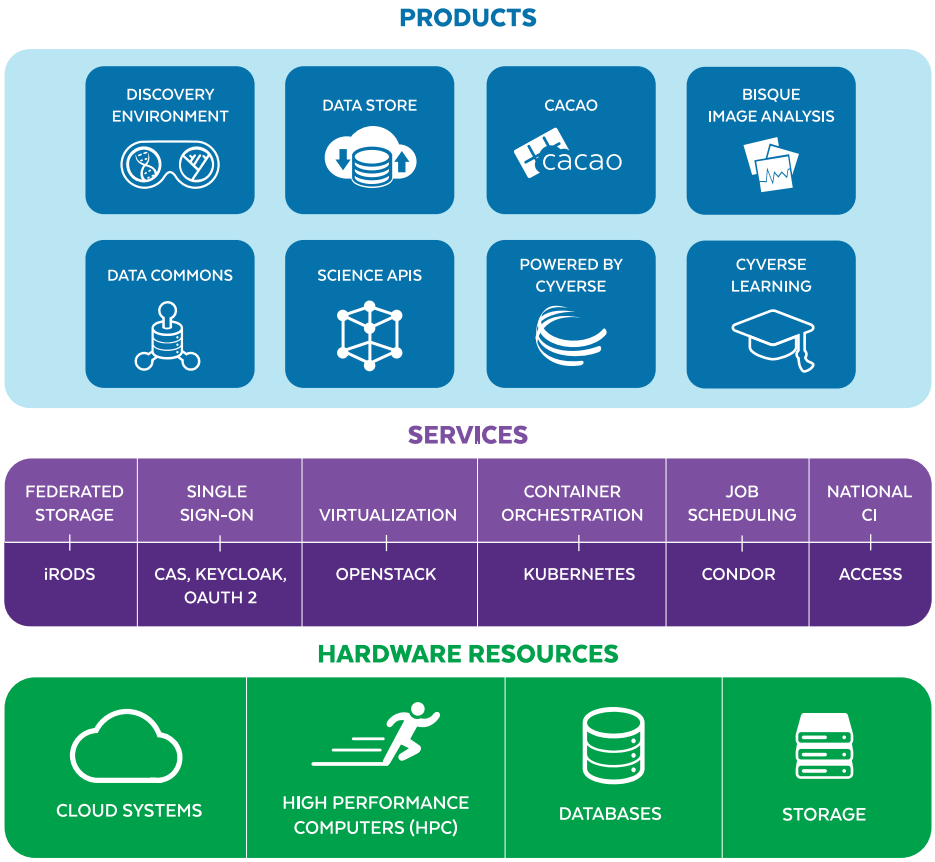

Fig 2. Layers of technology CyVerse software technology and cyberinfrastructure hardware components form a "layer cake" with hardware supporting services and software products. In general, the top layer is easiest to use but least flexible, while the bottom layers have the most power and utility but are least user-friendly.

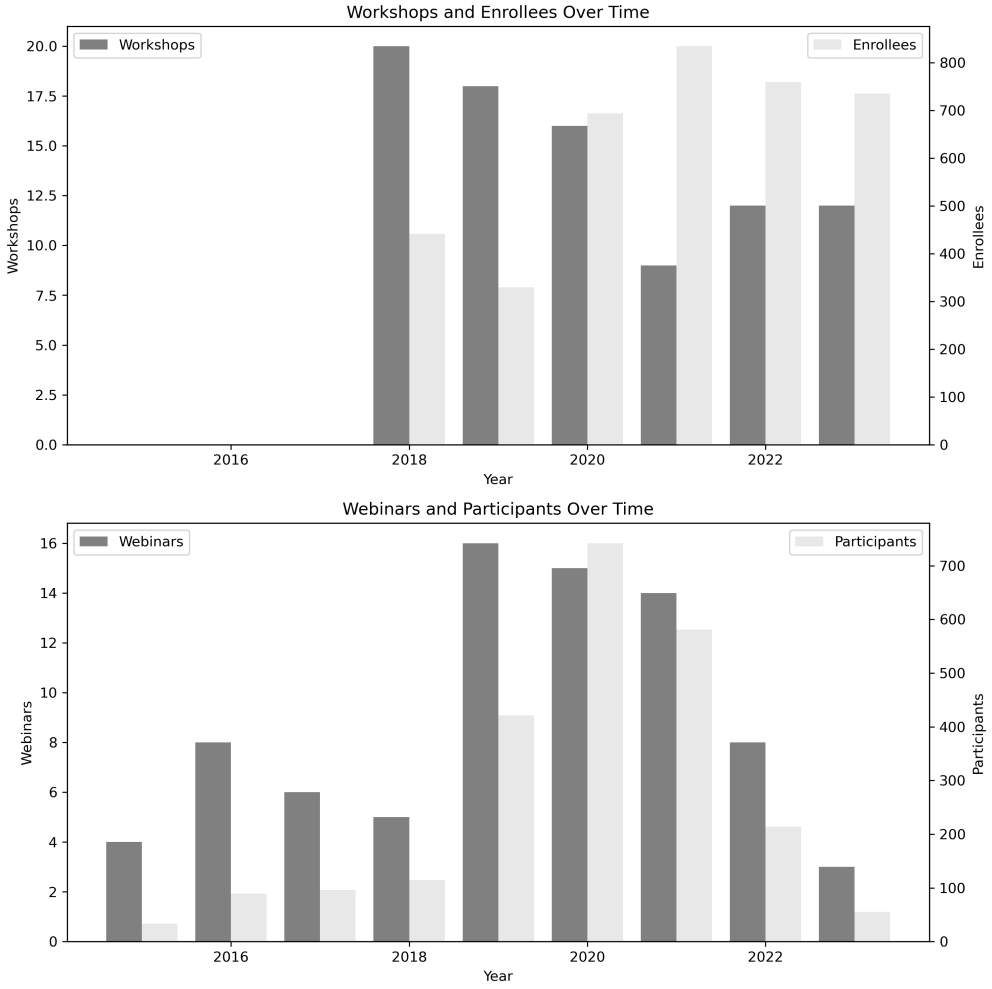

Fig 3. Workshop and webinar participants In-person and Virtual Workshops and total enrollees, Virtual Webinars given and participants who RSVP'd.

system (CMS). CyVerse does pay for an Intercom.io [50] account, which is integrated across CyVerse web services and into its documentation websites. Registered users can request support directly through Intercom's chat feature, which also generates a ticket in the Intercom system and an email to the user. While support tickets are often automatically binned for the appropriate team to respond, CyVerse staff regularly reviews tickets to ensure the correct team will answer. This internal network of communication results in CyVerse having an average response time of less than four hours to resolution for new tickets during normal working hours (Monday - Friday). Intercom also allows CyVerse staff to monitor user demographics and user experience. CyVerse staff received and answered approximately two thousand support tickets a year over the last three years.

CyVerse was an early supporter of Software Carpentry [51, 52], now called "The Carpentries" [19, 53]. We continue to collaborate with Carpentries staff and trainers in developing novel data science training materials and in hosting workshops. Digital literacy is currently at the forefront of many universities' education and research programs, with "Data Science" institutes, colleges, departments, and degree programs being created globally to meet this demand.

Asynchronous training materials are maintained through the Learning Center website (https://learning.cyverse.org), while in-person and virtual workshops, entitled "Foundational Open Science Skills," "Container Basics," and "Advanced Containers," are offered for data scientists, educators, researchers, IT professionals, and students on a recurring basis. The workshops are designed to prepare investigators and their teams, both new and established, to meet the growing expectations of funding agencies, publishers, and research institutions for scientific reproducibility and data accessibility. Since 2018, 345 early career faculty and post-doctoral researchers have attended CyVerse Foundational Open Science Skills workshops. Numerous attendees have gone on to write successful NSF proposals, start their own research labs, or integrate open science principles into their curricula or departmental courses, using skills and knowledge gained during the workshop. A recent phenomenon includes humanities researchers enrolling in our workshops, revealing interest in these foundational computational skills from a wide range of disciplines.

User Demographics

Most CyVerse users are from a life sciences background (plant science, biology, genomics); however, over 23% of users are from sciences-other-than life science (Fig 4). In the last 10 years CyVerse was demonstrated to approximately 45,000 attendees at in-person or virtual workshops and webinars. Over 121,000 accounts have been created since its initial public launch in 2011. In the last four years, 37,310 unique users from over 160 countries have logged into the platform (S11 Fig). The majority of users are from non-Research 1 Universities (92,275 accounts) and self-acknowledged as being students (Fig 4). Remarkably, 3.8% more users identify as female than male (Fig 4). Additional anonymized user demographic information is provided in the Supporting Information section.

There are typically 40 ± 10 unique users in web-sessions on the platform at any given time. Peak usage is observed during workshops or educational courses, with over 300 concurrent users in the Discovery Environment observed without interruption. In a typical 30-day period $1{,}100\pm100$ registered users will log into the platform and use it at least once. Between August 2017 and August 2021 the top ten users (excluding CyVerse staff and affiliated users) launched over 1,000 web sessions; the top 100 users had 300 web sessions; the top 500 had 100 web sessions.

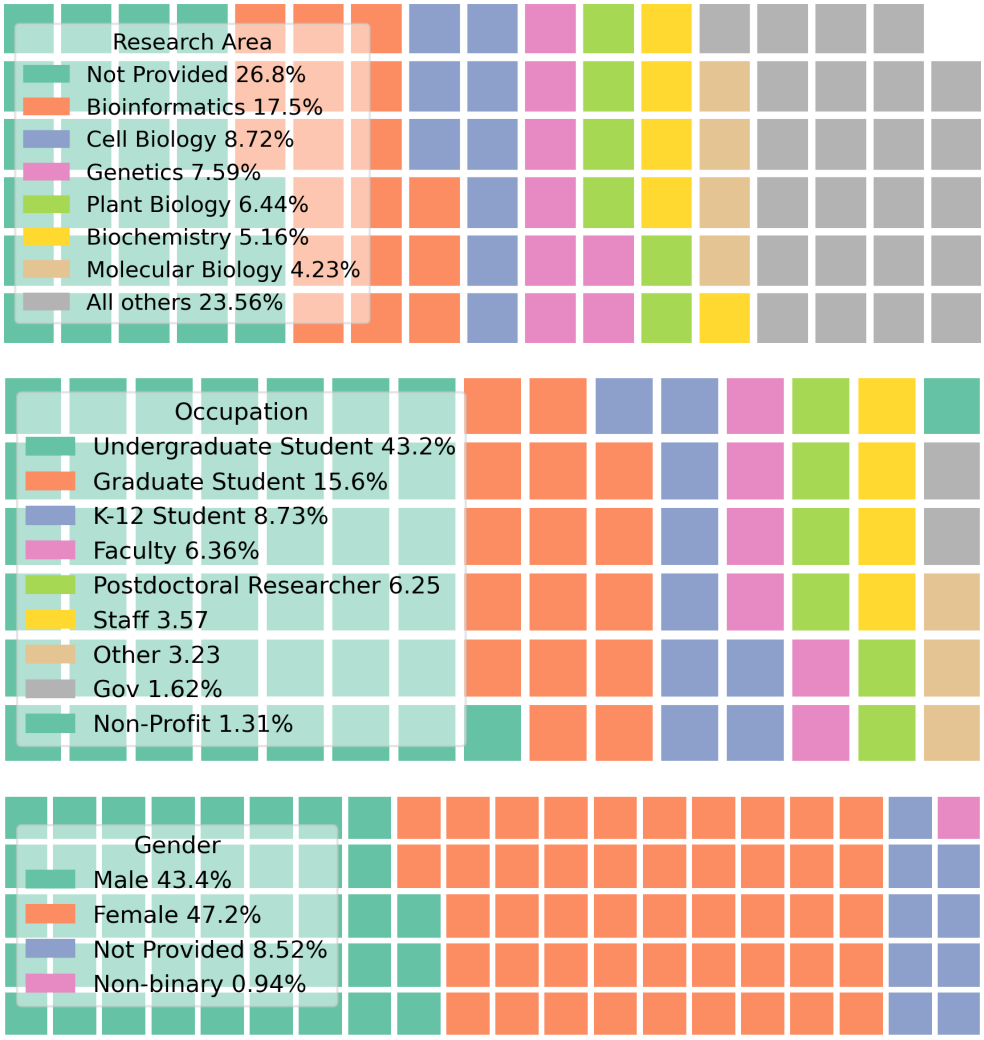

Fig 4. Demographics registered users' research area, occupation, and genders.

Scientific Impacts

CyVerse has been referenced in over 1,600 scientific publications including peer-reviewed publications, masters theses, and Ph.D. dissertations (S1 File). In addition, CyVerse has been mentioned in over 125 NSF awards' public abstracts ($257M awarded funding). CyVerse staff have written over 400 letters of support for grant proposals and have been a collaborating partner on over 50 federal awards to date. We can generate a rough estimate of the total number of NSF award dollars supported by CyVerse by calculating the percentage of awards for which CyVerse wrote a letter of support and who mentioned CyVerse in their public abstract (25%) as well as the number of awards mentioning CyVerse in their public abstract without requesting a letter of support (73%) (Fig 5). If only 25% of awards requesting a letter from CyVerse mentioned CyVerse in the public abstract, CyVerse likely supported $700M to $1B of NSF awards. Specifically, this does not include projects supported by other US agencies (regional, state, federal, private) and international research projects. The multiplier on research support from NSF's $117M investment in CyVerse has likely resulted in 6× to 9× return on investment (ROI).

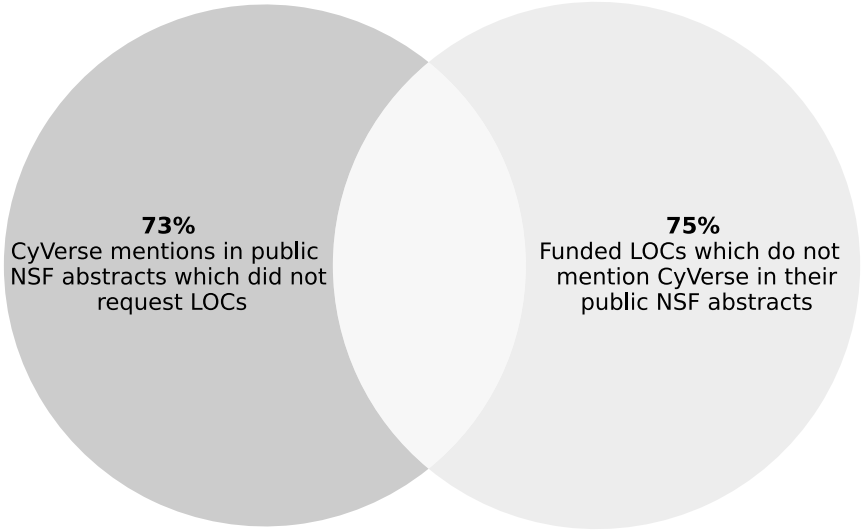

Fig 5. Multiplier effects Only 25% of CyVerse-powered NSF-funded projects mentioned "CyVerse" or "iPlant Collaborative" in their public abstract, conversely, 73% of funded projects which requested Letters of Collaboration (LOCs) do not mention using CyVerse. Ergo, CyVerse may have indirectly supported between $700M and $1B of NSF research awards over the last 15 years. The primary $117M awarded to CyVerse has resulted in a 6× to 9× return on investment (ROI) to the NSF.

International expansion

In 2018, CyVerse UK was launched at the University of Nottingham [54]. CyVerse UK leverages the foundational services (described in the Design and Implementation section from CyVerse US) while maintaining their own Data Store and Discovery Environment [55]. In 2019, CyVerse Austria (CAT) was launched at Graz University of Technology (TUG) [56,57]. CAT was initiated through the BioTechMed initiative [58] to foster collaborative, reproducible life science research in Graz. CAT connects TUG

with the Medical University of Graz and the University of Graz. The system adheres to the EU-GDPR and ensures that universities comply with their respective research data management policies [59, 60], meaning that research data is stored on-site with secured access. Through the iRODS-based Data Store, researchers can work together in collaborative projects within BioTechMed. CAT has quickly grown in visibility, and recently the Austrian ministry funded a follow-up project [61] to further develop and expand CAT within Austria. In addition, given that CyVerse US, CyVerse UK, and CyVerse CAT use the same foundational technologies, our teams and user communities are able to rapidly redeploy and reuse workflows across deployments, decreasing the time it takes researchers to leverage new analytical methods. Additional international collaborations with CyVerse US are currently underway in Canada and Australia.

Impact beyond the life sciences

Science projects beyond the life sciences leveraging CyVerse include: astronomy [62, 63], atmospheric science, national defense, earth science, environmental health [64], geology & geoscience, health science, hydrology [65], natural resources, lunar and planetary science, ocean science, oceanography, pedagogy, pollution monitoring, and space sciences [66–69] (see S1 File for the list of 1,600 publications which reported using CyVerse). As of 2023, CyVerse has provided DOIs to 153 data sets, which range from viral and plant genomes to astronomical black hole images.

CyVerse partners with numerous organizations and institutions to help develop new tools and research outside its core life science focus. Prominent examples include the DesignSafeCI for Natural Hazards, whose design was inspired by the CyVerse Discovery Environment [70], CUAHSI's HydroFrame and HydroShare which also rely on iRODS [65, 71, 72], and open workflows which leverage OpenTopography.org and the OSG [73]. CyVerse has International Traffic in Arms Regulations (ITAR) compliance for software, and has been successfully deployed on secured commercial cloud (AWS GovCloud) for defense and space situational awareness projects. CyVerse Health Information System is a HIPAA Privacy Program certified with Authority to Operate until 2026. Beyond these first degree influences that CyVerse has had, its affiliate projects, e.g., Jetstream and Jetstream2, are enabling ever more external research outcomes outside the life sciences [11, 74]. For example, the Atmosphere/Jetstream UI was adopted by the Massachusetts Open Cloud (MOC) [75] and is used for research and education.

Availability & Future Directions

All SaaS and IaC templates developed by CyVerse are available on its GitHub and GitLab organizations (see S2 Table) and are licensed under Open Source Initiative (OSI) compliant licenses. CyVerse developed software (e.g., iRODS CSI Driver, DE, & CACAO) are released under the BSD v2 Clause License, except in specific cases where other OSI licenses may supersede. All training material and platform documentation are licensed under the Creative Commons (CC BY 4.0) License. CyVerse operates multiple public services from the *.cyverse.org domain name service (S4 Table).

Tomorrow's Challenges

A survey of over 700 National Science Foundation (NSF) principal investigators found that a lack of skills in the use of cyberinfrastructure and in training opportunities was the greatest bottleneck to leveraging existing investments in research cyberinfrastructure by the life sciences community [10]. More broadly, this digital divide

disproportionately impacts minorities and individuals working at smaller, underserved, or rural institutions, who may lack access to high speed internet and consequently, the ability to do data-intensive science [16, 76–78]. CyVerse was designed to meet this need head on, with the minimal requirement that the student or researcher has at least limited access to the internet.

The August 2022 White House Office of Science and Technology Policy (OSTP) "Nelson Memo" made clear that open science and open data are to be requirements of all federally-funded research beyond 2025 [79]. CyVerse's user registrations represent over 160 countries and dozens of scientific disciplines (Fig 3), further revealing a global demand for FOSS cyberinfrastructure. CyVerse is ready to help facilitate future open science research at any scale, having already developed services and resources explicitly around FOSS, FAIR & CARE data principles, and open access data.

After 15 years of core support from the National Science Foundation, CyVerse will be one of the largest NSF projects ever to transition to a self-sustaining revenue model. CyVerse addressed this challenge by ensuring adherence to its core vision/mission, identifying its user community, and working with established partners for advice and guidance. CyVerse has successfully launched four revenue streams. First, through an NSF supported partnership with Phoenix BioInformatics 501(c)3 in 2021 CyVerse developed and implemented a subscription system for individuals and institutions. CyVerse also continues to partner with large federally-funded research proposals in an infrastructure support role through its Powered by CyVerse program and provides on-premises deployments for institutions, organizations, and companies through its Professional Services program. CyVerse also receives funding support at the state-level from its host institutions. Together, these diversified revenue streams ensure long-term project financial stability while meeting the needs of CyVerse's diverse user communities.

As CyVerse pivots towards a sustainable model which relies in part on subscriptions for services, researchers from underserved groups are likely to be excluded at higher rates than other groups, thus widening the digital divide in research computing [76, 78]. Therefore, our goal remains to continue offering all of our services with enforced computing and storage limits for free. By taking in a diversified revenue stream we intend to maintain a 'basic' free tier for all students and for anyone interested in testing CyVerse for their work.

Treating scientific software as infrastructure, rather than as part of research, would help address the ongoing issue of sustainability of FOSS cyberinfrastructure like CyVerse, Jetstream2, and the National Research Platform. Organizations which create and support FOSS SaaS and IaC are providing services in the same vein as traditional research computing centers, which are considered necessary. If only large and wealthy institutions have the capacity and capital to invest in on-premise hardware and the people to administer it, or to pay the significant commercial cloud services fees that modern research computing requires, inequities will persist or even grow [80–83]. Therefore, it is our belief that unfettered access to research and educational cyberinfrastructure is vital to ensuring a diverse, equitable, and inclusive society [84, 85].

Streaming toward the edge

Data intensive science from streaming data and edge computing [86–88] is one frontier at which CyVerse envisions itself in the next decade. Moving computations to the edge with the Internet of Things (IoT), Machine Learning (ML), and generative AI for remote sensing using platforms such as sUAS, and integrated sensor networks streaming real and near real-time data are all areas where CyVerse is already involved. Applying CyVerse's cyberinfrastructure capabilities to the most pressing challenges our society faces include, but are not limited to: adapting to and developing better strategies for

resilience to climate change, exploring Genotype by Environment = Phenotype (G×E=P) in both agricultural and natural settings [89, 90], using ML and AI for monitoring Earth system processes and studying human health, and developing precision medicine and synthetic biological approaches to life science (See S1 Appendix for explicit examples).

Supporting information

Appendix

S1 Appendix. Description of CyVerse Core & Cloud Services. Additional details about CyVerse featured platforms, core services, and cloud native services.

S1 Table. Glossary. Frequently used abbreviations and acronyms with descriptions.

S2 Table. Version Control. Public and private version control organizations on GitHub and GitLab for CyVerse Software, Public Container Registry, and Education.

S3 Table. Interfaces. Websites under the *.cyverse.org DNS address.

S4 Table. Powered by. External projects supported by CyVerse within the last 5 years. Resources include (Web) Hosting, Compute, Data Storage, Discovery Environment (DE), & API access.

S5 Table. University of Arizona Hardware. On-premises resources maintained by CyVerse at the University of Arizona. DE = Discovery Environment, VICE = Virtual Interactive Compute Environment.

S6 Table. Benchmarking Data Store transfers. Duration of seconds within CyVerse, as well as from (download) and to (upload) other research HPC (TACC), cloud (XSEDE Jetstream2) and commercial cloud services (AWS, Google Cloud). *Transfer duration represents a rounded number of seconds as a geometric mean of n=30 runs.

S7 Table. DE Applications. How they run, where they run, and popular applications.

S1 Fig. Where CyVerse US operates. CyVerse connects with the ACCESS-CI major facilities (diamonds) and the OpenScienceGrid (OSG) (circles) via Internet2 and regional high speed networks (solid lines). Jobs running on CyVerse's OSG applications are distributed across the country, based on resource availability.

S2 Fig. Authentication. Users log in through a Web Browser where they submit their credentials through either KeyCloak and CILogon. These authentication requests are accepted by OAuth2.0 (black) and returned. UML template adapted from GitHub user JMBarbier.

S3 Fig. User login screen. CyVerse authentication uses Keycloak with CILogon, GitHub, Globus Auth, or Google credentials. Users can log in with their unique CyVerse username or from their preferred single-sign on service.

S4 Fig. User Portal. Provides access to all other CyVerse platforms and services, as well as requests for additional data storage, cloud resources, and workshops.

S5 Fig. Data Store. The iRODS data store is accessible from a variety of multi-client access end points. The resource servers that make-up the data store include on-premises storage servers at UArizona, as well as federated storage on commercial and public research clouds.

S6 Fig. Data Store traffic. Data Store transfers by CyVerse users total (TiB/month) data download and upload from the Data Store, the number of files (million/month) downloaded and uploaded.

S7 Fig. Discovery Environment User Interface. The DE uses a table of contents menu (left side) with a collapsable hamburger menu. The Data Store, Apps, and Analyses can be viewed in the mainframe. Help, updates, and user profile features are visible in the upper right corner. Administrator accounts can approve access requests (VICE), edit public Apps and Tools, approve DOI Requests, and edit Reference Genomes in the table of contents (lower left).

S8 Fig. Featured container deployment. CI/CD workflow for featured container applications in the Discovery Environment. Image recipes (Dockerfiles) are hosted on GitHub/GitLab and use build triggers with automation servers (GitHub Actions) to build and tag images. Tested images are pushed to public and private registries on DockerHub and self-hosted Harbor. Images are cached on the DE production servers (nodes) for rapid deployment as containers at runtime.

S9 Fig. Discovery Environment Interactive. Interactive jobs include GUI apps like RStudio and JupyterLab. The DE manages interactive jobs through Kubernetes (K8s) and its Terrain API. Access to apps are managed by an Ingress Controller (e.g., NGINX). The analysis service shows whether the app is deployed, loading, or currently running and loads the UI for the analysis. Central authentication is managed by CAS. Users can load data from the iRODS datastore into their running containers. LDAP manages the user's secure authentication. Data Store is mirrored nightly at TACC from UArizona.

S10 Fig. Discovery Environment Executable. Executable (command line interface driven) Apps are managed by HTCondor and the Terrain API. Jobs are submitted through the DE user interface where they trigger a job submission service managed by HTCondor with Advanced Message Queueing Protocol (AMQP). Once the job runs, it is sent to a node where a program called RoadRunner uses Docker-Compose to manage the execution. Data are copied back to the iRODS data store when the app completes using a program called porklock. A PostgreSQL database monitors all job status and outcomes.

S11 Fig. Global and USA distribution of CyVerse user accounts. CyVerse registered accounts, by country (top) and by US state (bottom).

S1 File. Publications. Peer-reviewed research citing the use of resources from the iPlant Collaborative (2008-2017) and CyVerse (2017-Present). Also see https://cyverse.org/publications for the latest update.

S2 File. CyVerse vs others. CyVerse's services versus other public research and commercial cyberinfrastructure. Some services offered by commercial cloud have free tiers as well as paid subscriptions. (link to table)

Acknowledgments

This material is based upon work supported by the US National Science Foundation under Grants DBI-0735191, DBI-1265383, and DBI-1743442. Work on CyVerse Austria was funded by the Austrian Infrastructure Program 2016/2017, Bundesministerium für Bildung, Wissenschaft und Forschung Austria, BioTechMed/Graz Hochschulraum-Strukturmittel 'Integriertes Datenmanagement.' The project was supported by Digitale TU Graz (Graz University of Technology).

References

1. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Front Plant Sci. 2011;2:34.
2. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol. 2016;14:e1002342.
3. Crowston K, Howison J. The social structure of free and open source software development. First Monday. 2005 [cited 15 Aug 2021]. doi:10.5210/fm.v10i2.1207
4. von Krogh G, von Hippel E. The Promise of Research on Open Source Software. Manage Sci. 2006;52:975–983.
5. Scacchi W, Feller J, Fitzgerald B, Hissam S, Lakhani K. Understanding free/open source software development processes. Softw Process Improv Pract. 2006;11:95–105.
6. Stewart TA. Intellectual Capital: The new wealth of organization. Crown; 2010.
7. Peng RD. Reproducible research in computational science. Science. 2011;334:1226–1227.
8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
9. Understanding Data Motion in the Modern HPC Data Center. [cited 9 May 2023]. Available: `https://ieeexplore.ieee.org/abstract/document/8955242`
10. Barone L, Williams J, Micklos D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Comput Biol. 2017;13:e1005755.
11. Gentemann CL, Holdgraf C, Abernathey R, Crichton D, Colliander J, Kearns EJ, et al. Science storms the cloud. AGU Advances. 2021;2. doi:10.1029/2020av000354
12. Abernathey RP, Augspurger T, Banihirwe A, Blackmon-Luca CC, Crone TJ, Gentemann CL, et al. Cloud-Native Repositories for Big Scientific Data. Computing in Science Engineering. 2021;23:26–35.

13. Buck S. Solving reproducibility. Science. 2015;348:1403.

14. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Front Neuroinform. 2017;11:76.

15. Fairlie RW. Race and the Digital Divide. Contrib Econ Analysis Policy. 2004;3. doi:10.2202/1538-0645.1263

16. Norris P. The digital divide. Routledge; 2020.

17. Kitchin R. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE; 2014.

18. Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, et al. The Tao of open science for ecology. Ecosphere. 2015;6:1–13.

19. Michonneau F, Paul D. Scaling Up Data Literacy and Computing Skills Training in Biodiversity Science, Lessons Learned from The Carpentries. Biodiversity Information Science and Standards; Sofia. 2019. doi:10.3897/biss.3.35108

20. Kratzke N, Quint P-C. Understanding cloud-native applications after 10 years of cloud computing - A systematic mapping study. J Syst Softw. 2017;126:1–16.

21. Ramachandran R, Bugbee K, Murphy K. From open data to open science. Earth Space Sci. 2021;8. doi:10.1029/2020ea001562

22. Understanding Data Motion in the Modern HPC Data Center. [cited 9 May 2023]. Available: `https://ieeexplore.ieee.org/abstract/document/8955242`

23. Boomija MD, Raja SVK. Securing medical data by role-based user policy with partially homomorphic encryption in AWS cloud. Soft Computing. 2022;27:559–568.

24. Understanding Data Motion in the Modern HPC Data Center. [cited 9 May 2023]. Available: `https://ieeexplore.ieee.org/abstract/document/8955242`

25. Mesirov JP. Computer science. Accessible reproducible research. Science. 2010;327:415–416.

26. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1:0021.

27. NIH Data Sharing Policy and implementation guidance. [cited 10 Sep 2021]. Available: `https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm`

28. Open Data at NSF. [cited 10 Sep 2021]. Available: `https://www.nsf.gov/data/`

29. Belhajjame K, Corcho O, Garijo D, Zhao J, Missier P, Newman DR, et al. Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. SePublica@ ESWC. users.ox.ac.uk; 2012. pp. 1–12.

30. Hettne KM, Dharuri H, Zhao J, Wolstencroft K, Belhajjame K, Soiland-Reyes S, et al. Structuring research methods and data with the research object model: genomics workflows as a case study. J Biomed Semantics. 2014;5:41.

31. Edmunds SC, Li P, Hunter CI, Xiao SZ, Davidson RL, Nogoy N, et al. Experiences in integrated data and research object publishing using GigaDB. International Journal on Digital Libraries. 2017;18:99–111.

32. Palma R, Garcia-Silva A, Gomez-Perez JM, Krystek M. A Research Object-Based Toolkit to Support the Earth Science Research Lifecycle. 2018 IEEE 14th International Conference on e-Science (e-Science). ieeexplore.ieee.org; 2018. pp. 50–57.

33. Bucksch A, Das A, Schneider H, Merchant N, Weitz JS. Overcoming the Law of the Hidden in Cyberinfrastructures. Trends Plant Sci. 2017;22:117–123.

34. Sahneh F, Balk MA, Kisley M, Chan C-K, Fox M, Nord B, et al. Ten simple rules to cultivate transdisciplinary collaboration in data science. PLoS Comput Biol. 2021;17:e1008879.

35. Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE principles for indigenous data governance. Data Sci J. 2020;19. doi:10.5334/dsj-2020-043

36. Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, et al. Data-intensive Science: A New Paradigm for Biodiversity Studies. Bioscience. 2009;59:613–620.

37. Faris J, Kolker E, Szalay A, Bradlow L, Deelman E, Feng W, et al. Communication and data-intensive science in the beginning of the 21st century. OMICS. 2011;15:213–215.

38. Wolf F, Hobby R, Lowry S, Bauman A, Franza BR, Lin B, et al. Education and data-intensive science in the beginning of the 21st century. OMICS. 2011;15:217–219.

39. Choudhary V. Software as a Service: Implications for Investment in Software Development. 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). 2007. p. 209a–209a.

40. Morris K. Infrastructure as Code: Managing Servers in the Cloud. "O'Reilly Media, Inc."; 2016.

41. Ansible RH. Ansible is Simple IT Automation. [cited 11 Sep 2021]. Available: `https://www.ansible.com/`

42. Argo Workflows - The workflow engine for Kubernetes. [cited 11 Sep 2021]. Available: `https://argoproj.github.io/argo-workflows/`

43. Bernstein D. Containers and Cloud: From LXC to Docker to Kubernetes. IEEE Cloud Computing. 2014;1:81–84.

44. Terraform by HashiCorp. [cited 11 Sep 2021]. Available: `https://www.terraform.io/`

45. Beck M, Moore T. The Internet2 Distributed Storage Infrastructure project: an architecture for Internet content channels. Computer Networks and ISDN Systems. 1998;30:2141–2148.

46. Thain D, Tannenbaum T, Livny M. Distributed computing in practice: the Condor experience. Concurr Comput. 2005;17:323–356.

47. Team R, Others. RStudio: integrated development for R. RStudio, Inc, Boston, MA URL `http://www.rstudio.com`. 2015;42.

48. Perez F, Granger BE. Project Jupyter: Computational narratives as the engine of collaborative data science. Retrieved September. 2015;11:108.

49. Sole AD, Del Sole A. Introducing Visual Studio Code. Visual Studio Code Distilled. 2019. pp. 1–17. doi:10.1007/978-1-4842-4224-7_1

50. Conversational relationship platform. [cited 9 Sep 2021]. Available: https://www.intercom.com/

51. Wilson G. Software Carpentry: Getting Scientists to Write Better Code by Making Them More Productive. Computing in Science Engineering. 2006;8:66–69.

52. Wilson G. Software Carpentry: lessons learned. F1000Res. 2014;3:62.

53. Pugachev S. What are "the carpentries" and what are they doing in the library? Portal. 2019;19:209–214.

54. CyVerse UK. [cited 11 Sep 2021]. Available: https://cyverseuk.org/

55. Minotto A, Van Den Bergh E, Davey RP. CyVerse UK: Widening the Scope to the UK and Beyond. Plant and Animal Genome XXVI Conference (January 13-17, 2018). PAG; 2018. Available: https://pag.confex.com/pag/xxvi/meetingapp.cgi/Paper/31449

56. Lang K, Stryeck S, Bodruzic D, Stepponat M, Trajanoski S, Winkler U, et al. CyVerse Austria—A Local, Collaborative Cyberinfrastructure. Math Comput Appl. 2020;25:38.

57. Wieser F, Stryeck S, Lang K, Hahn C, Thallinger G, Feichtinger J, et al. A local platform for user-friendly FAIR data management and reproducible analytics. Journal of Biotechnology. 2021. doi:10.1016/j.jbiotec.2021.08.004

58. BioTechMed-Graz. [cited 11 Sep 2021]. Available: https://biotechmedgraz.at/de/

59. RDM - TU Graz Framework Policy for RDM. [cited 11 Sep 2021]. Available: https://www.tugraz.at/sites/rdm/policies/tu-graz-framework-policy-for-rdm/

60. Research Data Management. [cited 11 Sep 2021]. Available: https://ub.uni-graz.at/en/services/publication-services/research-data-management/

61. Austrian DataLAB and Services - Cluster Forschungsdaten. 11 May 2020 [cited 11 Sep 2021]. Available: https://forschungsdaten.at/adls/

62. The Event Horizon Telescope Collaboration. First M87 EHT results: Calibrated data. CyVerse Data Commons; 2019. doi:10.25739/G85N-F134

63. Morzinski KM, Close LM, Males JR, Kopon D, Hinz PM, Esposito S, et al. MagAO: Status and on-sky performance of the Magellan adaptive optics system. Adaptive Optics Systems IV. International Society for Optics and Photonics; 2014. p. 914804.

64. Ramírez-Andreotta MD, Walls R, Youens-Clark K, Blumberg K, Isaacs KE, Kaufmann D, et al. Alleviating Environmental Health Disparities Through Community Science and Data Integration. Front Sustain Food Syst. 2021;5. doi:10.3389/fsufs.2021.620470

65. Olschanowsky C, Maxwell RM, Condon LE, Strout M, Altintas I, Purawat S, et al. Hydroframe: A Software Framework to enable Continental Scale Hydrologic Simulation. 2019. p. A11A–01.

66. Furfaro R, Linares R, Gaylor D, Jah M, Walls R. Resident space object characterization and behavior understanding via machine learning and ontology-based Bayesian networks. Advanced Maui Optical and Space Surveillance Tech Conf(AMOS). amostech.com; 2016. Available: https://amostech.com/TechnicalPapers/2016/SSA-Algorithms/Furfaro.pdf

67. Walls RL, Gaylor D, Reddy V, Furfaro R, Jah M. Assessing the IADC Space Debris Mitigation Guidelines: A case for ontology-based data management. AMOS Paper. 2016. Available: https://amostech.com/TechnicalPapers/2016/SSA/Walls.pdf

68. Reddy V, Linder T, Linares R, Furfaro R, Tucker S, Campbell T. RAPTORS: Hyperspectral Survey of the GEO Belt. AMOS Technologies Conference, Maui Economic Development Board, Kihei, Maui, HI. amostech.com; 2018. Available: https://amostech.com/TechnicalPapers/2018/NROC/Reddy.pdf

69. Carlson O, Hohenstein S, Bui J, Tanquary H, Fritz C, Gross DC. Human Factors in the Unified Architecture Framework Applied to Space Situational Awareness. 2019 IEEE International Systems Conference (SysCon). ieeexplore.ieee.org; 2019. pp. 1–7.

70. Rathje EM, Dawson C, Padgett JE, Pinelli J-P, Stanzione D, Adair A, et al. DesignSafe: New cyberinfrastructure for natural hazards engineering. Nat Hazards Rev. 2017;18:06017001.

71. Tarboton DG, Idaszak R, Horsburgh JS, Heard J, Ames D, Goodall JL, et al. HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing. International Congress on Environmental Modelling and Software. 2014. Available: https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/7/

72. Purawat S, Olschanowsky C, Condon LE, Maxwell R, Altintas I. Scalable Workflow-Driven Hydrologic Analysis in HydroFrame. Computational Science – ICCS 2020. Springer International Publishing; 2020. pp. 276–289.

73. Swetnam TL, Pelletier JD, Rasmussen C, Callahan NR, Merchant N, Lyons E, et al. Scaling GIS Analysis Tasks from the Desktop to the Cloud Utilizing Contemporary Distributed Computing and Data Management Approaches: A Case Study of Project-based Learning and Cyberinfrastructure Concepts. Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale. New York, NY, USA: ACM; 2016. pp. 21:1–21:6.

74. Hancock DY, Stewart CA, Vaughn M, Fischer J, Lowe JM, Turner G, et al. Jetstream-Early operations performance, adoption, and impacts: Early Jetstream Performance and Results. Concurr Comput. 2018;57:e4683.

75. Mass open cloud – an open cloud exchange public cloud. [cited 11 Sep 2021]. Available: https://massopen.cloud/

76. Jackson LA, Zhao Y, Kolenic A 3rd, Fitzgerald HE, Harold R, Von Eye A. Race, gender, and information technology use: the new digital divide. Cyberpsychol Behav. 2008;11:437–442.

77. Sisneros L, Sponsler BA. Broadband access and implications for efforts to address equity gaps in postsecondary attainment. Education Commission of the States. 2016 [cited 19 Jun 2021]. Available: http://files.eric.ed.gov/fulltext/ED565437.pdf

78. Brown V. Technology Access Gap for Postsecondary Education: A Statewide Case Study. In: Promoting Global Competencies Through Media Literacy. IGI Global; 2018. pp. 20–40.

79. Nelson A. Office of science and technology policy (OSTP) memorandum on access to federal research. 2022 [cited 25 Mar 2023]. Available: https://policycommons.net/artifacts/3159884/08-2022-ostp-public-access-memo/3957772/

80. Stewart CA, Hancock DY, Wernert J, Link MR, Wilkins-Diehr N, Miller T, et al. Return on Investment for Three Cyberinfrastructure Facilities: A Local Campus Supercomputer, the NSF-Funded Jetstream Cloud System, and XSEDE (the eXtreme Science and Engineering Discovery Environment). 2018 IEEE/ACM 11th International Conference on Utility and Cloud Computing (UCC). IEEE; 2018. pp. 223–236.

81. Stewart CA, Apon A, Hancock DY, Furlani T, Sill A, Wernert J, et al. Assessment of non-financial returns on cyberinfrastructure: A survey of current methods. Proceedings of the Humans in the Loop: Enabling and Facilitating Research on Cloud Computing. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1–10.

82. Stewart CA, Hancock DY, Wernert J, Furlani T, Lifka D, Sill A, et al. Assessment of financial returns on investments in cyberinfrastructure facilities: A survey of current methods. Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning). New York, NY, USA: Association for Computing Machinery; 2019. pp. 1–8.

83. Chalker A, Hillegas CW, Sill A, Broude Geva S, Stewart CA. Cloud and on-premises data center usage, expenditures, and approaches to return on investment: A survey of academic research computing organizations. Practice and Experience in Advanced Research Computing. New York, NY, USA: Association for Computing Machinery; 2020. pp. 26–33.

84. Atkins DE. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. National Science Foundation; 2003.

85. Hacker TJ, Wheeler BC. Making research cyberinfrastructure a strategic choice. Educause Quarterly. 2007;30:21.

86. Shi W, Cao J, Zhang Q, Li Y, Xu L. Edge Computing: Vision and Challenges. IEEE Internet of Things Journal. 2016;3:637–646.

87. Satyanarayanan M. The Emergence of Edge Computing. Computer. 2017;50:30–39.

88. Willis C, Lambert M, McHenry K, Kirkpatrick C. Container-based Analysis Environments for Low-Barrier Access to Research Data. Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact. New York, NY, USA: ACM; 2017. pp. 58:1–58:4.

89. Orgogozo V, Morizot B, Martin A. The differential view of genotype-phenotype relationships. Front Genet. 2015;6:179.

90. Gonzalez EM, Zarei A, Hendler N, Simmons T, Zarei A, Demieville J, et al. PhytoOracle: Scalable, modular phenomics data processing pipelines. Front Plant Sci. 2023;14:1112973.