

Transforming SCIENCE  
through DATA-DRIVEN DISCOVERY



**CYVERSE<sup>®</sup>**

SERVICES OVERVIEW

# TABLE OF CONTENTS

<b>Welcome to the Cyber Universe</b>	<b>3</b>
What is Cyberinfrastructure?	<b>4</b>
Our Mission and Vision	<b>6</b>
<b>The CyVerse Data Cycle</b>	<b>7</b>
Bring Data into CyVerse	<b>8</b>
Analyze Data and Images	<b>9</b>
Share with Collaborators	<b>10</b>
Security and Accessibility	<b>11</b>
<b>Become a Data Scientist</b>	<b>12</b>
<b>Build Global Communities</b>	<b>14</b>
<b>Transform Science</b>	<b>15</b>



# WELCOME TO THE CYBER UNIVERSE

CYVERSE FILLS A NICHE CREATED BY THE COMPUTING ERA AND A RAPIDLY EVOLVING WORLD. Developing solutions for today's grand scientific challenges means understanding how the organisms that contribute to our food, fuels, and ecosystem are shaped by interactions with their environment. CyVerse provides life scientists with powerful computational infrastructure to handle massive datasets and complex analyses, thus enabling data-driven discovery and collaborations. Our powerful extensible platforms provide full-path data management, bioinformatics tools, image analyses, cloud services, APIs, and

## OUR PRODUCTS

---



**Discovery Environment**  
Web workbench



**Science APIs**  
Programmatic access



**Atmosphere**  
Cloud-computing platform



**Data Store**  
Secure sharing



**DNA Subway**  
Genome analysis



**BisQue Image Analysis**  
Image analysis environment



**Powered by CyVerse**  
Third-party interoperability



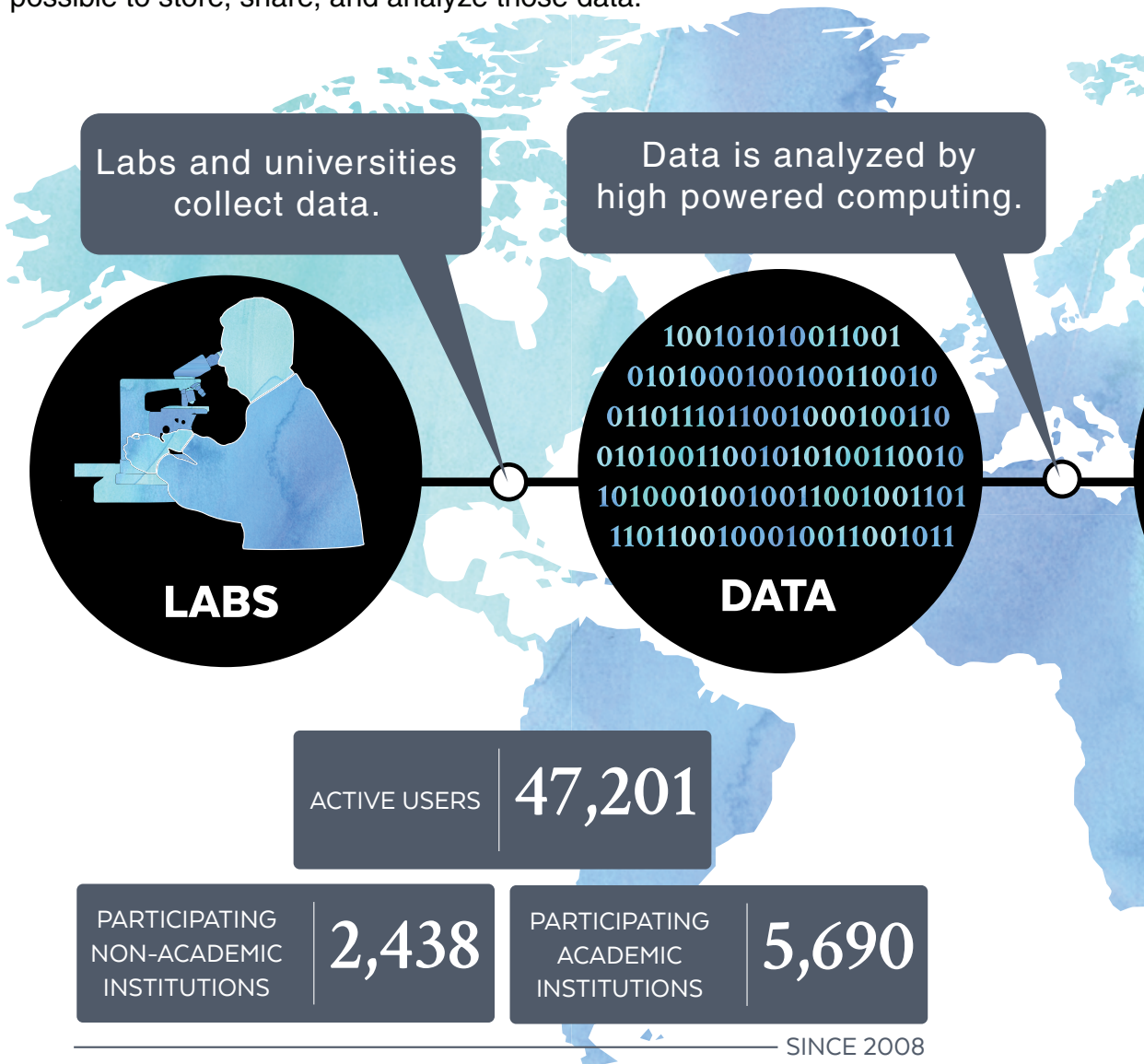
**Learning Center**  
Resources for product use



**Data Commons**  
Search, discover, and reuse

# WHAT IS CYBERINFRASTRUCTURE?

**CYBERINFRASTRUCTURE** (also known as CI or computational infrastructure) provides solutions to the challenges of large-scale computational science. Just as physical infrastructure, such as laboratories, DNA sequencing centers, and greenhouses, make it possible to collect and generate huge amounts of data, the hardware, software, and people that comprise cyberinfrastructure make it possible to store, share, and analyze those data.



**CYBERINFRASTRUCTURE DEVELOPMENT** is supported by the National Science Foundation to empower all scientists in the use of high-performance computing and large-data resources. Using cyberinfrastructure, teams of researchers can attempt to answer questions that previously were unapproachable because the computational requirements were too large, too complex, or simply unknown to some researchers.

Data is accessed, analyzed, and shared by people from anywhere in the world through cloud computing.



**ANALYSIS**



**PEOPLE**

**CYVERSE CYBERINFRASTRUCTURE INCLUDES:**

- A data storage facility
- An interactive, web-based analysis platform
- Cloud infrastructure to use remote servers for computation, analysis, and storage
- Web authentication and security services
- Support for scaling computational algorithms to run on large, high-speed computers
- Education and training resources in how to use cyberinfrastructure
- People with expertise in all of the above

# OUR MISSION AND VISION



**VISION:** Transforming Science through Data-driven Discovery

**MISSION:** Our mission at CyVerse is to design, deploy, and expand a national cyberinfrastructure for life sciences research, and to train scientists in its use.

We work to remove the infrastructure barriers for researchers and provide provenance for data science, streamlining the computational aspects of your projects so you can focus completely on your research.

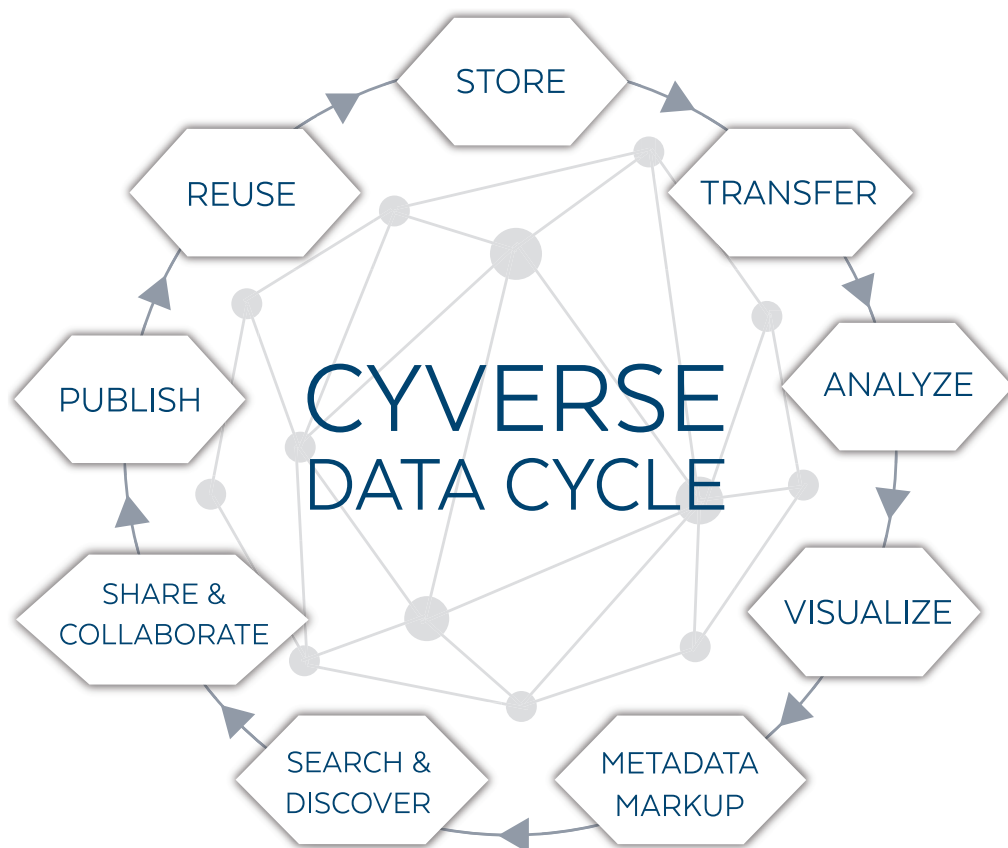
CyVerse's mission is shaped by community-driven needs. We rely upon your feedback to develop the infrastructure you need most to advance your science and educational agenda.

CyVerse is funded by the National Science Foundation's Directorate for Biological Sciences. We are a dynamic virtual organization led by the University of Arizona to fulfill a broad mission that spans our partner institutions: Texas Advanced Computing Center, Cold Spring Harbor Laboratory, and the University of North Carolina at Wilmington.

# THE CYVERSE DATA CYCLE

WITH A SUITE OF PLATFORMS that enables analysis and sharing of data, analysis methods, and results, CyVerse provides all the components to address your data science needs:

- Online access from anywhere with internet service
- Easy data management and analysis tools
- An extensible framework that lets you integrate your own software tools
- Secure storage for your work
- Password-protected sharing, so you can work with colleagues around the world
- Training in computational thinking, methodologies, and best practices
- Assistance moving your datasets to publication readiness



# BRING DATA INTO CYVERSE



## GETTING STARTED WITH CYVERSE IS EASY.

Just go to [www.cyverse.org](http://www.cyverse.org) and create an account with a username (it's usually simplest to use your institutional email address) and password.

Every user has access to 100 GB of data storage in CyVerse's Data Store, with the ability to request more space for special research projects. Whether your data are small or large – kilobytes to terabytes, or structured or unstructured – CyVerse's cloud-based Data Store can help you with your data storage needs. Our user-friendly web platform enables researchers to upload files from anywhere with an internet connection, over any device capable of going online.

Based upon a technology called iRODS, CyVerse's data hosting gives you great flexibility and control over your data. Choose from different interfaces – from web services to mountable file systems to high-speed command-line transfers – to connect your data to every CyVerse platform.

# ANALYZE DATA AND IMAGES



## Discovery Environment

All aspects of your bioinformatics workflow – data management, analysis, sharing large datasets, and more – can be handled within our Discovery Environment, or DE, platform. The DE provides a modern web interface for powerful computing, data, and app resources for your scientific analyses. By providing a consistent user interface to access the tools and computing resources needed for specialized scientific analyses, the DE facilitates data exploration and scientific discovery. In the DE, you can use hundreds of command-line analytical tools, utility tools, and scripts, without the need to learn the command line.

You can run analytical apps and workflows on high-performance computing nodes. You can also integrate most Linux command-line tools, scripts or executables into the DE or create a new workflow by chaining one app to another. Because much of the complexity is hidden from the user, the DE makes it easy for non-technical users to get the job done simply and efficiently; no need to master command-line syntax to use the analysis tools.



## BisQue Image Analysis

You can use BisQue (Bio-Image Semantic Query User Environment) to store, visualize, organize, and analyze images in the cloud. BisQue was developed for the exchange and exploration of biological images, but can support image data from other domains. Centered around a database of images and metadata, BisQue supports search and comparison of image datasets. Novel semantic analyses are integrated into the system, allowing high-level semantic queries and comparison of image content.

BisQue is integrated with CyVerse's authentication system, the Data Store, and computational infrastructure for scalability and ready access to a large set of downstream analysis options. Developers can integrate existing applications or create new ones, leveraging BisQue's rich set of custom visualizations, image handling routines, and APIs (Application Programming Interfaces) for building scalable, web-based image analysis applications.



# SHARE WITH COLLABORATORS



## Atmosphere

One of our goals is to empower global research collaborations by making it easy for research partners to share datasets securely. The CyVerse Atmosphere platform provides an isolated virtual machine (VM) image and software in a cloud environment, which can be used and shared among teams across institutions. You can also control sharing settings for data files within the Data Store and Discovery Environment, allowing select research partners access to all, or some, of your datasets.

Our objective is to continually improve the means by which we facilitate global scientific collaboration, thereby enhancing the ability of researchers and developers to share data vital to their projects, and arrive more efficiently at new discoveries.

“The fact that Atmosphere is a shared resource means that students can go back to their home institutions and can still work together even if they're in very different places – that's powerful.”

– Gavin Conant of the University of Missouri teaches an undergraduate computational biology course using CyVerse

DATA PRODUCED, SHARED, AND ACCESSED WITHIN CYVERSE

2016



27,924 VIRTUAL IMAGES LAUNCHED



2,500 USER DATA IN THE DATA STORE (TB)



158,374 JOBS RUN WITH APIs ON HPC RESOURCES



# SECURITY AND ACCESSIBILITY



## Data Store

Within the Data Store, you manage your data security by setting permissions to keep your data private only to you, share it with specific collaborators, or make it accessible to the general public (so no account is required). You can download some of these publicly accessible data for your own research purposes, or you can choose to make your datasets open-access within the Data Store at any time.

### DATA SECURITY

In the digital age, we understand the need to keep your research data secure. We continually investigate and evaluate novel methods to improve data security. Our authentication system provides secure single sign-on, giving you exclusive access and control over your data security settings.

### OPEN SOURCE

With our open source policy, CyVerse open source code is available on GitHub. Developers who are interested in replicating CyVerse's infrastructure framework or developing their own similar framework may access all of our infrastructure coding and diagrams.

“By working across disciplines as immunologists, mathematicians, and computer scientists, we were able to tackle a problem that was untenable to any discipline alone. We've created an analytical infrastructure with CyVerse that allows for all the data to be stored and analyzed by researchers everywhere.”

– Adam Buntzman of the University of Arizona discusses leading a team to become the first to sequence the human immunome, a breakthrough in the field of immunology

# BECOME A DATA SCIENTIST

PEOPLE ARE THE PRINCIPAL COMPONENT OF CYBERINFRASTRUCTURE, and we consider everyone learning about or working within the life sciences to be our stakeholders. CyVerse Education, Outreach, and Training (EOT) accomplishes its mission by offering targeted strategies that move each stakeholder group – especially those lacking the skills or access to best leverage bioinformatics – towards increasing knowledge, involvement, and use of cyberinfrastructure. EOT accomplishes its objectives through a variety of activities.



## WORKSHOPS AND SEMINARS

Our popular two-day workshops provide a comprehensive look at platforms, tools, and services provided by CyVerse for large-scale data analysis. Workshop topics build progressively to cater to the needs of general and advanced audiences, and are customized according to participant interests. Hands-on demos and guided exercises cover

popular topics (e.g., genome assembly, genome annotation, RNA-Seq, GWAS, etc.). Workshop participants also are able to consult with instructors and bring their own data to get started on a project.

The Data-2-Publication Workshop pairs faculty groups with CyVerse domain experts to move important datasets to publication readiness, and also serves as a focal point for solving cyberinfrastructure challenges such as integrating new tools and workflows. Additionally, CyVerse is proud to partner with Software Carpentry™ and Data Carpentry™ to provide additional workshops to train new data scientists around the world.



## WEBINARS AND ONLINE LEARNING

We offer Online and remote learning options to maximize accessibility of our educational tools. Monthly ‘Get Started with CyVerse’ webinars introduce new users to CyVerse services. Other webinars cover topics such as new software releases, with step-by-step demonstrations for how to use our tools, allowing participants to ask questions and further hone their skillsets.

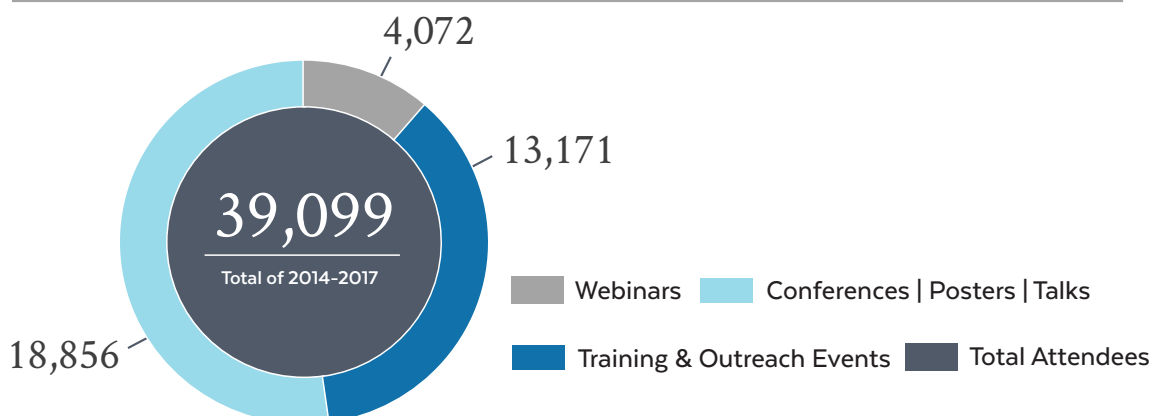


Our Learning Center offers in-depth tutorials covering popular science topics in how to conduct analyses with CyVerse tools, and best practices that help users apply computational thinking and avoid common bioinformatics pitfalls.

## EDUCATIONAL CYBERINFRASTRUCTURE

Our DNA Subway platform bundles research-grade bioinformatics tools and databases into intuitive workflows, presenting them in an appealing interface modeled on the metaphor of a subway map. “Riding” on different DNA Subway lines, students can predict and annotate genes in up to 150 KB of DNA (Red Line), identify homologs in sequenced genomes (Yellow Line), analyze DNA barcodes and construct phylogenetic trees (Blue Line), and much more.

## CYVERSE WORKSHOP PARTICIPATION (Number of Attendees Per Type of Event)



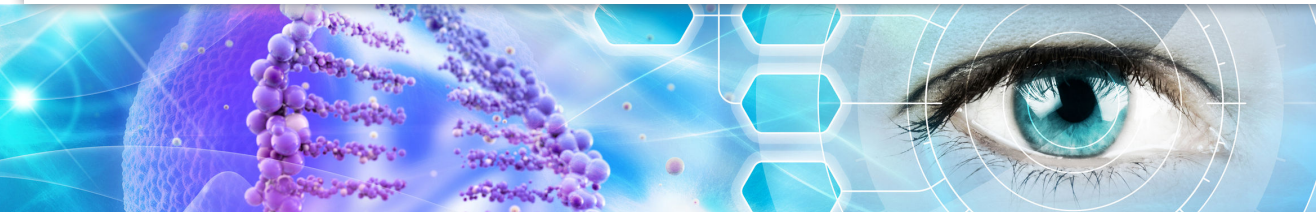
# BUILD GLOBAL COMMUNITIES



CYVERSE UNIQUELY ALLOWS RESEARCH PARTNERS TO WORK ACROSS INSTITUTIONAL BOUNDARIES, bringing global collaborative communities together to tackle complex interdisciplinary projects. Our extensible framework seamlessly integrates disparate programs and tools, so that you can easily bring your data all the way through their processing cycle without the need to continually reformat them. All of the tools you need are available within CyVerse, or can be readily incorporated into our framework so that you have no need to download and install new software for your analyses. You can also bring your own computational algorithms and workflows into the CyVerse infrastructure, and make them readily available for your own or other people's projects.

Through our partnership with the NSF's XSEDE (Extreme Science and Engineering Discovery Environment) project, the most powerful publicly accessible system of supercomputing resources in the U.S., you will be able to do more with your data, greatly expanding upon the results achievable with standard institutional computing capabilities.

# TRANSFORM SCIENCE

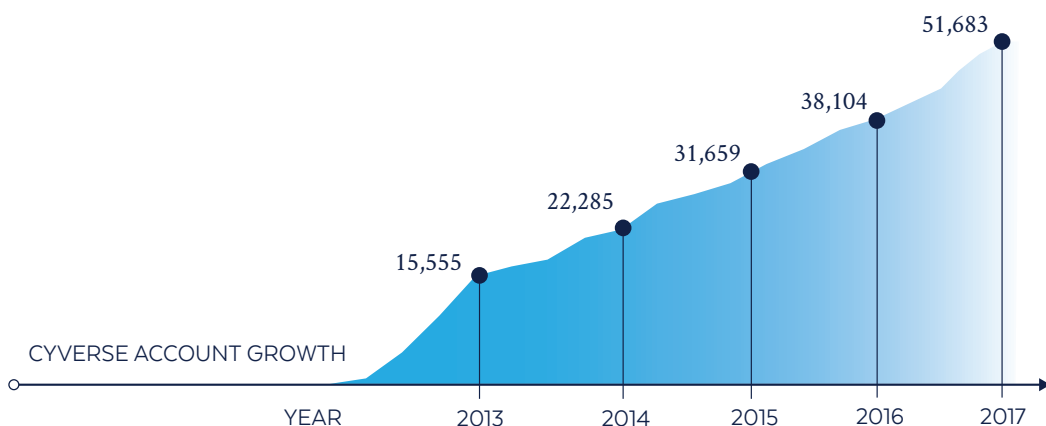


CYVERSE EMPLOYS THE MOST NOVEL METHODOLOGIES AND PRACTICES to bring you streamlined data-driven discovery – from secure data storage, management, analysis, sharing, and collaboration, to education and training resources – to empower you as a data scientist regardless of your scientific discipline.

We are continually expanding our infrastructure capabilities and resources, partnering with third parties and comparable projects to transform data science in the digital age.

Through our Powered by CyVerse program, third-party projects can leverage our cyberinfrastructure to provide services to their users, including a linked authentication system allowing secure single sign-on between their application and all CyVerse services, access to the Data Store, and the ability to execute analyses on high-performance computing resources.

## THE CYVERSE COMMUNITY CONTINUES TO EXPAND



# CONTACT US

Transforming SCIENCE  
through DATA-DRIVEN DISCOVERY



---

[www.cyverse.org](http://www.cyverse.org)

[info@cyverse.org](mailto:info@cyverse.org)



[@CyVerseOrg](https://twitter.com/CyVerseOrg)



[CyVerse.org](https://www.google.com/+CyVerseOrg)



[CyVerse.org](https://www.youtube.com/CyVerse.org)



[github.com/cyverse/](https://github.com/cyverse/)



[CyVerse.org](https://www.linkedin.com/CyVerse.org)



[facebook.com/CyVerse.org](https://facebook.com/CyVerse.org)