



CYVERSE™

Transforming Science Through Data-driven Discovery

Accelerating Plant Science with CyVerse Genomics Workflows, Data Management, and Training

Jason Williams – Lead, CyVerse – Education, Outreach, Training

Cold Spring Harbor Laboratory

williams@cshl.edu [ @JasonWilliamsNY]



Cold Spring Harbor Laboratory



Slides and Tutorials



www.cyverse.org/sol2016



Cold Spring Harbor Laboratory



CyVerse Evolution



CyVerse 2016

Transforming Science
Through Data-Driven
Discovery

Vision:

Transforming science through data-driven discovery

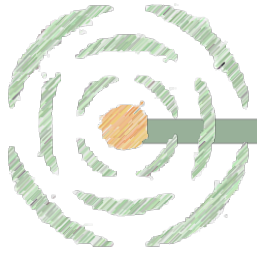
Mission:

Design, develop, deploy, and expand a national **cyberinfrastructure** for life science research, and train scientists in its use

More than 30K users, PB of data, and hundreds of publications, courses, and discoveries



CyVerse Evolution



iPlant 2008
Empowering a New
Plant Biology



iPlant 2013
Cyberinfrastructure for
Life Science



CyVerse 2016
Transforming Science
Through Data-Driven
Discovery

CyVerse is Built for Data



Plant / Microbial



Animal



Biomedical



Ecological/Climate

CyVerse supports all domains of life science

CyVerse Evolution



DBI-0735191 and DBI-1265383

We are funded by the National
Science Foundation

- We are your colleagues and collaborators!
- \$100 Million in investment
- Freely available to the community
- Spur national/international collaboration
- Cite CyVerse:
CyVerse.org/acknowledge-cite-cyverse



What is Cyberinfrastructure?



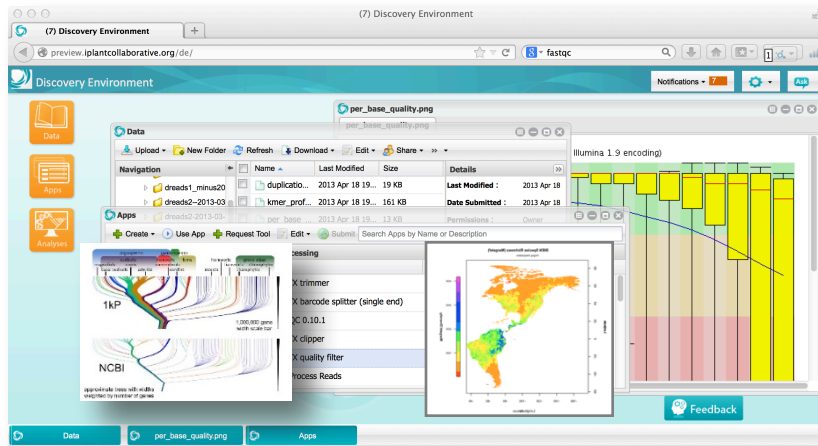
Cyberinfrastructure for Data-intensive Biology

- Data storage
- Software
- High-performance computing
- People

organized to solve problems of size and scope not otherwise solvable.



Cyberinfrastructure for Data-intensive Biology



Platforms, tools, datasets



Storage and compute



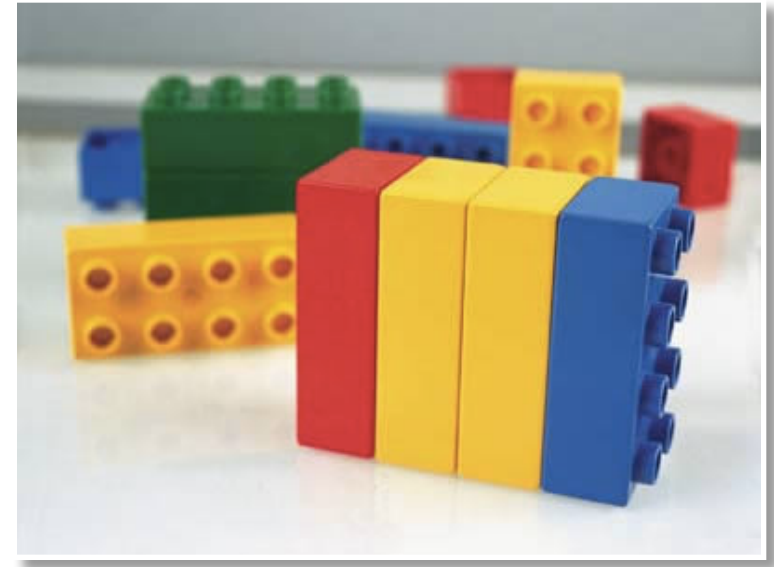
Training and support

What is available in CyVerse?



CyVerse Products

- We strive to be the **CI Lego blocks**
- Danish 'leg godt' - '**play well**'
- Also translates as '**I put together**' in Latin
- If a solution is not available you can craft your own using CyVerse CI components



CyVerse Features



Get Science Done

- Ability to access and manage data
- Software to analyze data
- Computing resources
- Skills and help to use software and interpret results



Ensure Reproducibility

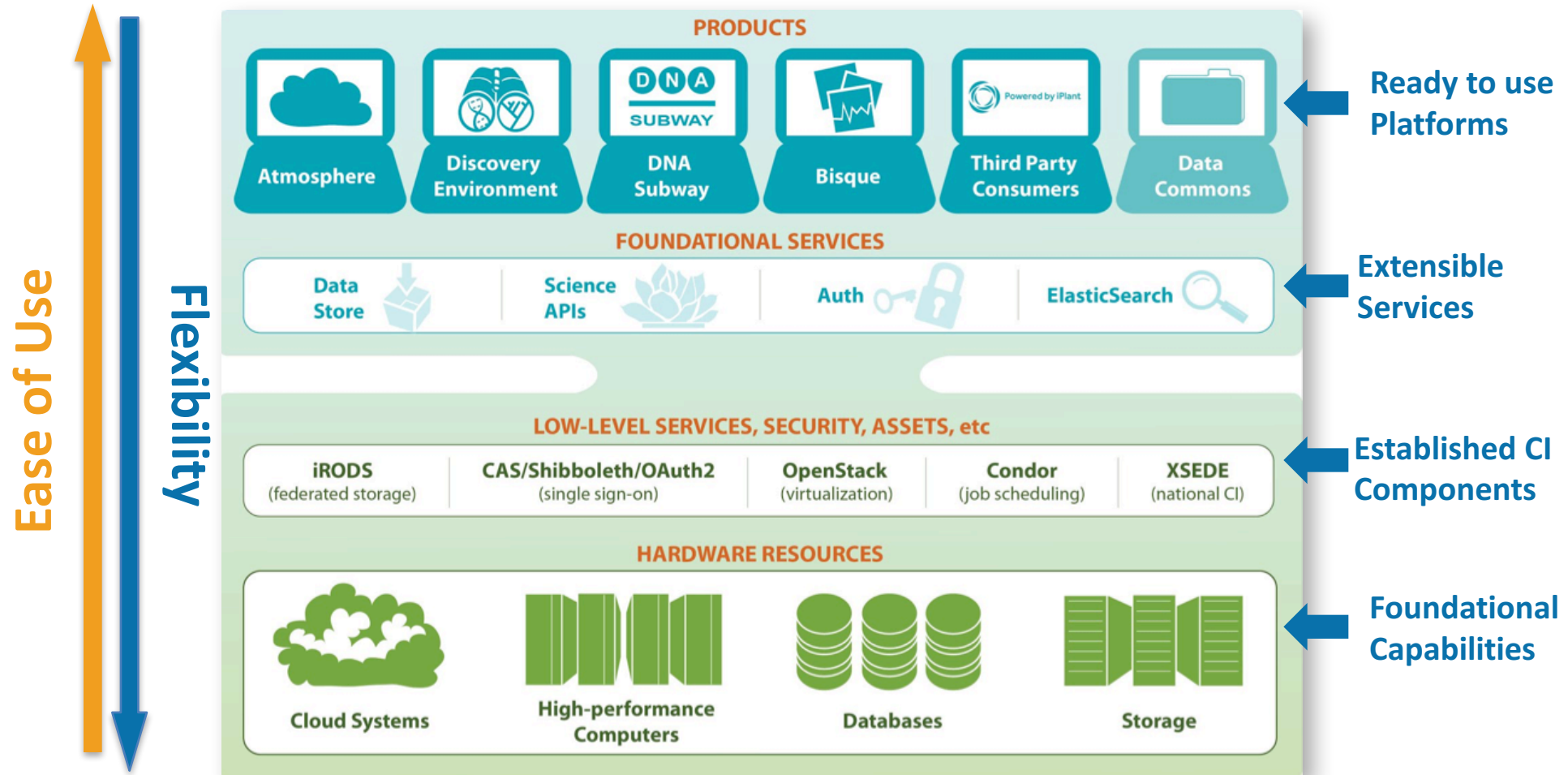
- Metadata management
- Ability to share data and workflows
- Open source sustainable tools

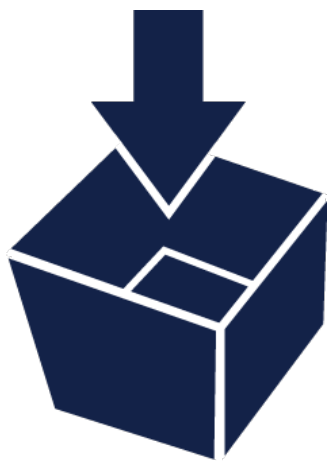


Increase Productivity

- High-performance and scalable computing
- Ability automate and collaborate
- Funding spent on science, not software or hardware

CyVerse product stack





Data Store

The resources you need to share and manage data with your lab, colleagues and community

- ✓ Initial 100 GB allocation – TB allocations available
- ✓ Automatic data backup
- ✓ Easy upload /download and sharing



Discovery Environment

Hundreds of bioinformatics Apps in an easy-to-use interface

- ✓ A platform that can run almost any bioinformatics application
- ✓ Seamlessly integrated with data and high performance computing
- ✓ User extensible – add your own applications





Atmosphere

Cloud computing for the life sciences

- ✓ Simple: quick access to hundreds of virtual machine images
- ✓ Flexible: fully customize your software setup
- ✓ Powerful: integrated with CyVerse computing and data resources





Science APIs

Fully customize CyVerse resources

- ✓ Science-as-a-service platform
- ✓ Define your own compute, and storage resources (local and *CyVerse*)
- ✓ Build your own app store of scientific codes and workflows





DNA Subway

Educational workflows for Genomes, DNA Barcoding, RNA-Seq

- ✓ Commonly used bioinformatics tools in streamlined workflows
- ✓ Teach important concepts in biology and bioinformatics
- ✓ Inquiry-based experiments for novel discovery and publication of data





Bisque

Image analysis, management, and metadata

- ✓ Secure image storage, analysis, and data management
- ✓ Integrate existing applications or create new ones
- ✓ Custom visualization and image handling routines and APIs

How can you do science in CyVerse?





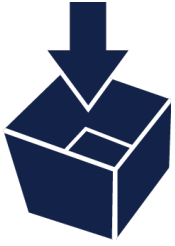
CyVerse Data Commons

Making data discoverable and reusable

Data Store architecture is built for metadata management

- Data within CyVerse maintains provenance and can be labeled using a variety of metadata templates or using custom schemas
- Direct submission of sequencing data to NCBI Sequence Read Archive
- Request DOIs and ARKs (Archival Resource Key) for permanent data identification
- Support for elastic search, data sharing, and anonymous access

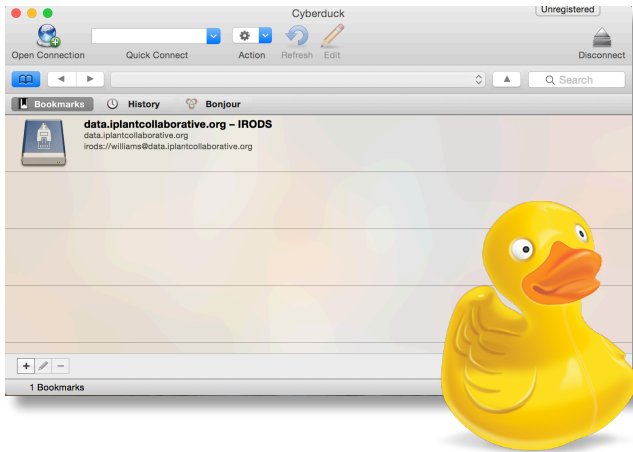




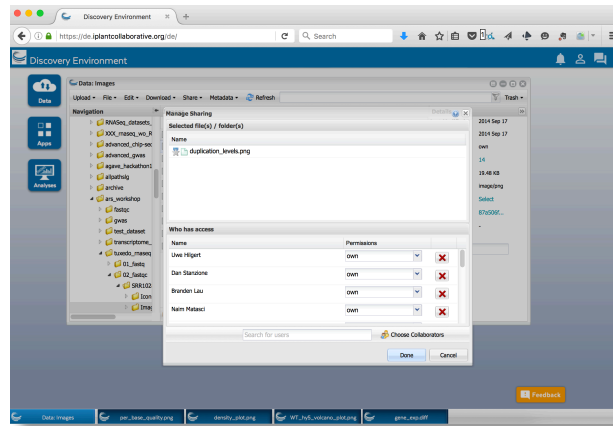
CyVerse Data Commons

Supporting the lifecycle of data

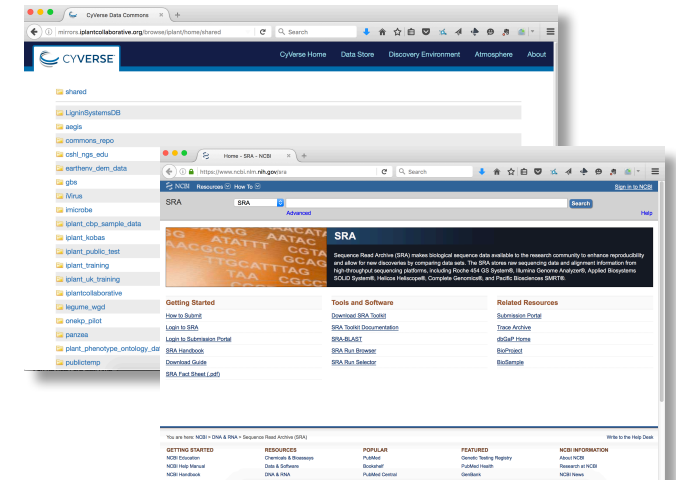
Simple upload



markup, search, share



publish

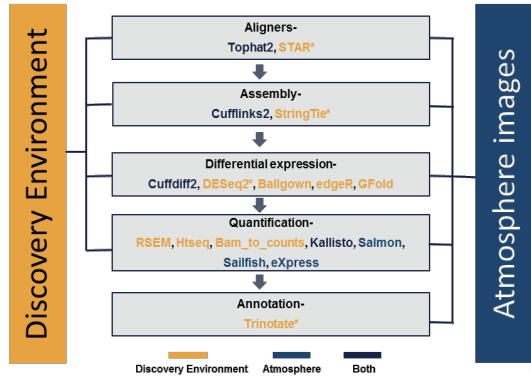




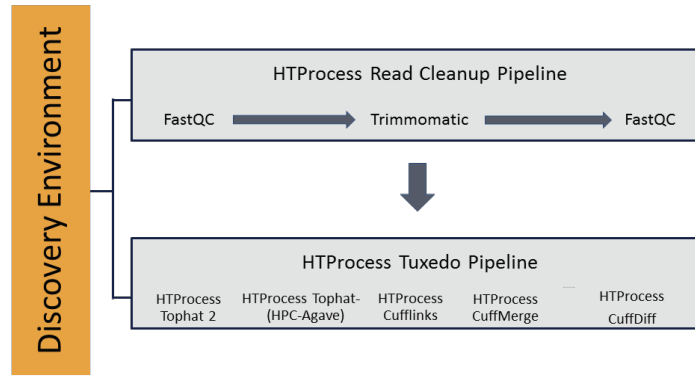
Discovery Environment

End-to-end workflows for genomics

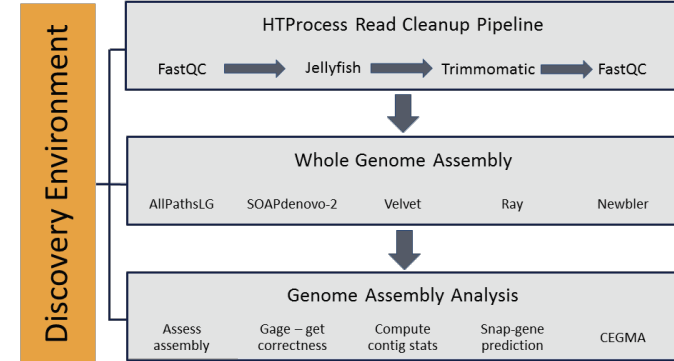
RNA-Seq 1: Differential Expression



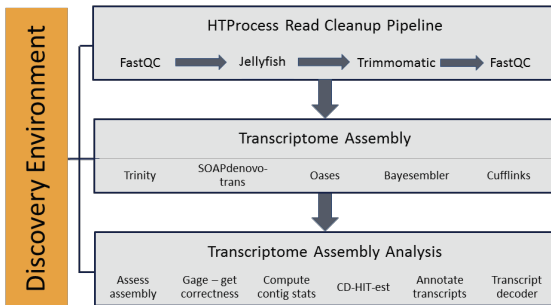
RNA-Seq 2: High-throughput



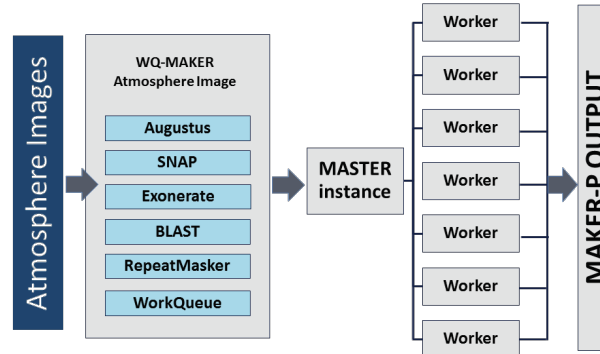
Genome Assembly



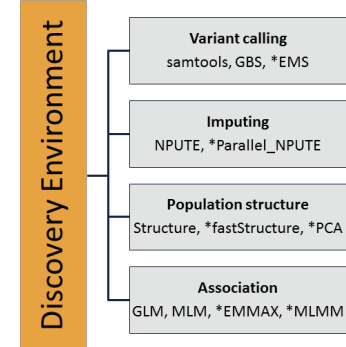
Transcriptome Assembly



WQ-Maker Annotation



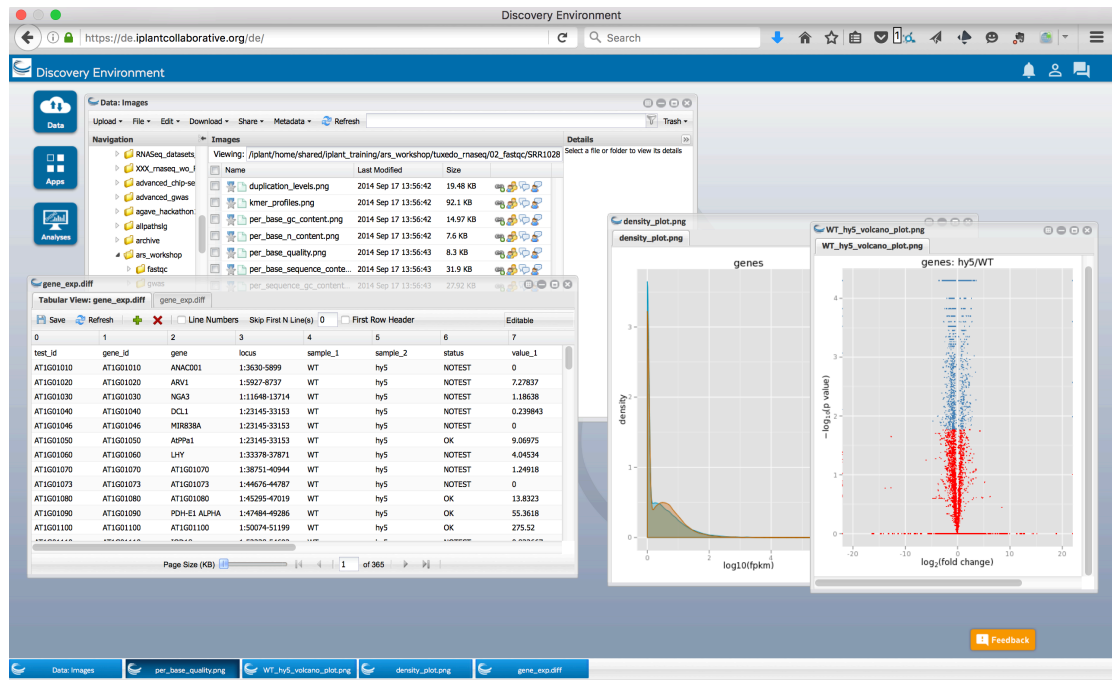
Association Studies





Discovery Environment

User friendly, developer friendly



CyVerse Platforms



The Agave Platform
API



docker



XSEDE Supercomputing



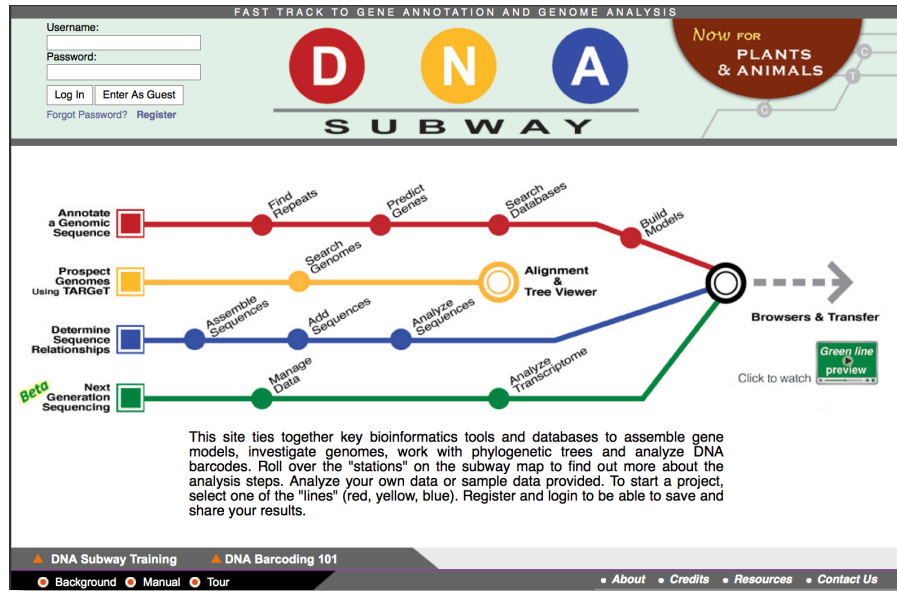
External cloud systems



DNA Subway

SUBWAY Classroom applications

RNA-Seq for the
Next Generation



- Classroom-friendly RNA-Seq workflows
- More than 900Gbs Student-faculty projects analyzed in 36 genomes
- Online educational materials



- DNA Barcoding workflow
- ~10,000 student-analyzed barcodes and 125 novel NCBI submissions
- Online educational materials



How can I use CyVerse?



Hands-on training

Workshops that reach users where they are

Tools & Services



- Two days; targeted to researchers
- Hands-on learning modules tailor to interests
- Individual consultations
- 1026 participants since 2011

Genomics in Education

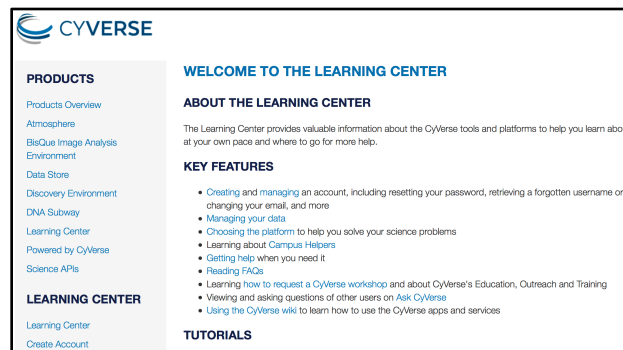


- Two days; targeted to educators
- Pair bioinformatics with classroom labs
- Help for generating lesson plans
- 748 participants since 2010

Online learning materials

Documentation, tutorials, and help

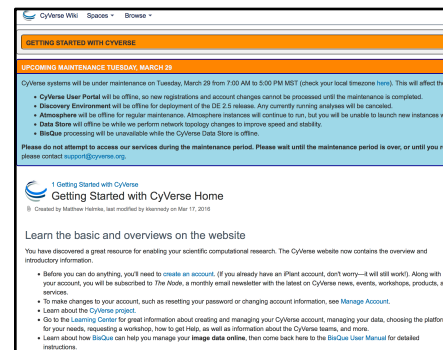
Learning Center



The screenshot shows the CyVerse Learning Center homepage. It features a navigation menu on the left with sections for 'PRODUCTS' (including Atmosphere, BioQueue Image Analysis Environment, Data Store, Discovery Environment, DNA Subway, Learning Center, Powered by CyVerse, and Science APIs) and 'LEARNING CENTER' (including Learning Center and Create Account). The main content area is titled 'WELCOME TO THE LEARNING CENTER' and 'ABOUT THE LEARNING CENTER', followed by 'KEY FEATURES' and 'TUTORIALS'. The 'KEY FEATURES' section lists several bullet points: 'Creating and managing an account, including resetting your password, retrieving a forgotten username or changing your email, and more', 'Managing your data', 'Choosing the platform to help you solve your science problems', 'Learning about Campus Helpers', 'Getting help when you need it', 'Reading FAQs', 'Learning how to request a CyVerse workshop and about CyVerse's Education, Outreach and Training', 'Viewing and asking questions of other users on Ask CyVerse', and 'Using the CyVerse wiki to learn how to use the CyVerse apps and services'.

- cyverse.org/learning-center
- Tutorials and quick-starts

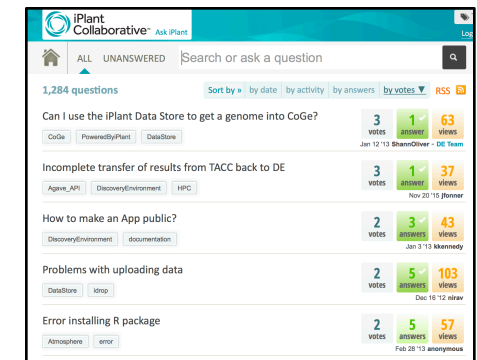
Wiki



The screenshot shows a page from the CyVerse Wiki titled 'Getting Started with CyVerse Home'. The page includes a maintenance notice for Tuesday, March 23, stating that CyVerse systems will be under maintenance from 7:00 AM to 5:00 PM MET. Below the notice, there is a section titled 'Getting Started with CyVerse Home' with a sub-header 'Learn the basic and overviews on the website'. The text explains that the website is a great resource for enabling scientific computational research and provides a list of links for account management, learning about the project, and requesting workshops.

- wiki.cyverse.org
- Documentation

Forum



The screenshot shows the iPlant Collaborative Ask iPlant forum page. It features a search bar at the top with the text 'Search or ask a question'. Below the search bar, there are several question entries with their respective statistics (votes, answers, views) and dates. The questions include: 'Can I use the iPlant Data Store to get a genome into CoGe?', 'Incomplete transfer of results from TACC back to DE', 'How to make an App public?', 'Problems with uploading data', and 'Error installing R package'.

- ask.iplantcollaborative.org
- Community help/Q&A





CYVERSE™

Transforming Science Through Data-driven Discovery

CyVerse Executive Team



THE UNIVERSITY
OF ARIZONA®

Parker Antin
Nirav Merchant
Eric Lyons



TEXAS ADVANCED
COMPUTING CENTER

Matt Vaughn



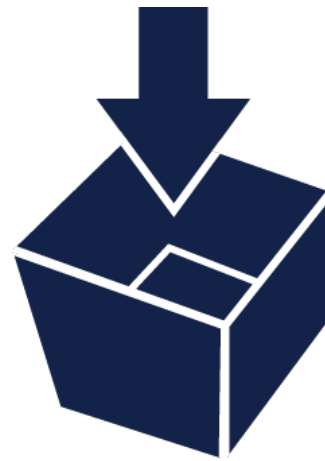
Cold
Spring
Harbor
Laboratory

Doreen Ware
Dave Micklos

CyVerse is supported by the National Science Foundation under Grant No. DBI-0735191 and DBI-1265383.



Getting Data into CyVerse





CyVerse Data Store

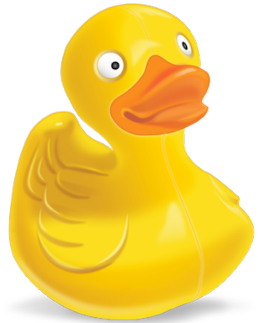
- Store any type of file related to your research
- Move files seamlessly between CyVerse platforms
- Automate file transfers
- Share files with lab members, collaborators, and communities



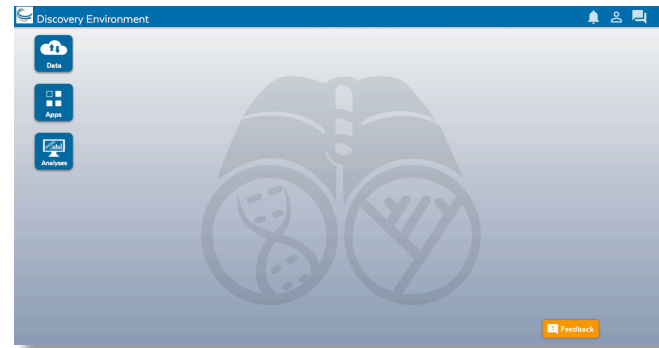
CyVerse Data Store

Multiple ways to access

Point-and-click



Cyberduck



Discovery Environment

Command line

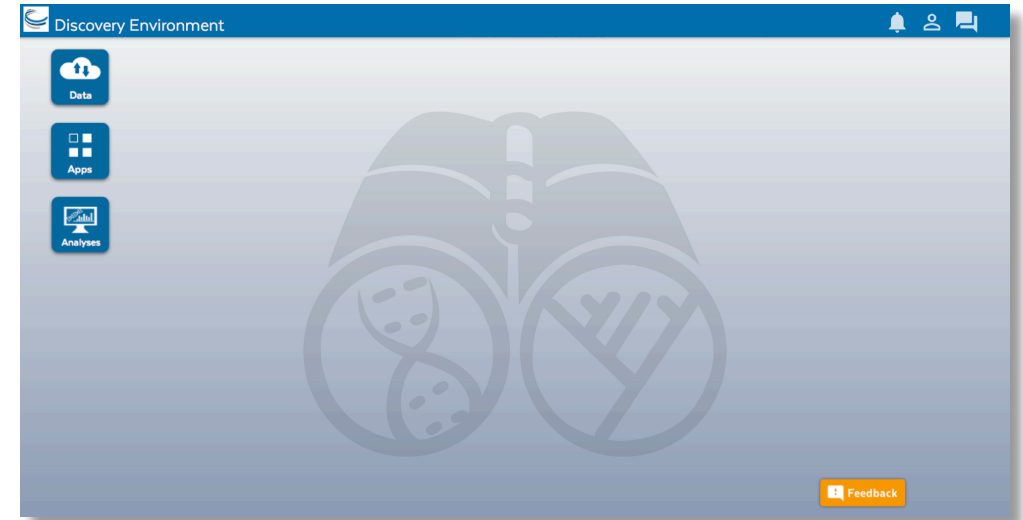
```
vector jasonwilliams -- bash -- 44x11
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
C- /iplant/home/williams/analyses/TASSEL_4
.3.0_MLM_analysis1-2013-09-11-20-17-30.232
C- /iplant/home/williams/analyses/TASSEL_4
.3.0_MLM_analysis1-2013-09-12-14-52-35.844
C- /iplant/home/williams/analyses/Test_of_
New_App_analysis1-2013-10-25-14-40-49.857
```

iCommands



Discovery Environment

- Simple upload/download for small files
- Bulk upload files and folders (<10GB)
- Import from URL (no size limit)

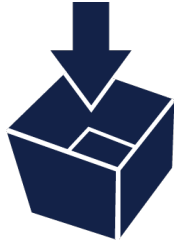


Advantage +

Covers most upload/download sharing needs

Disadvantage -

Some size/speed limitations

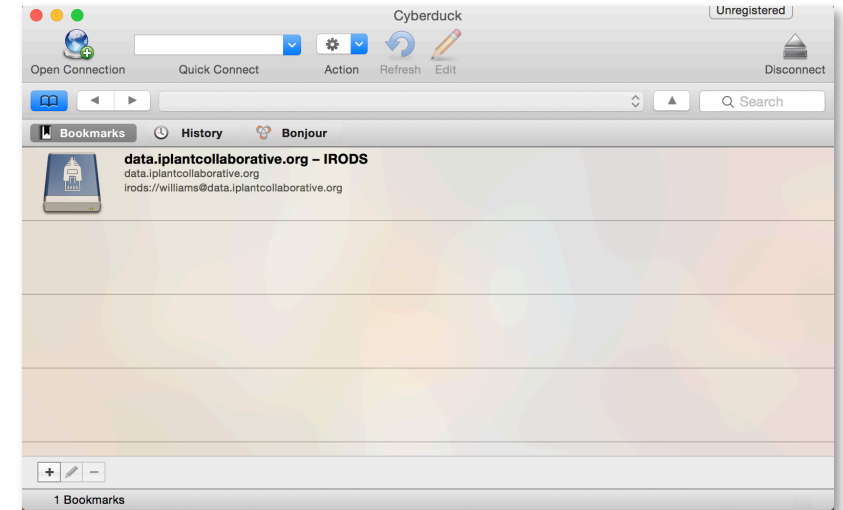


Cyberduck

- Drag and drop files and folders
- No size limit, file editing/previews
- Easy Desktop functionality

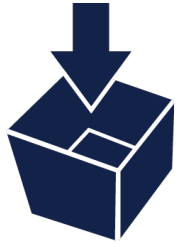
Advantage +

More like desktop file systems



Disadvantage -

No permissions/metadata control



iCommands

- Full flexibility
- Ability to script and automate
- Access from terminal/server

Advantage +

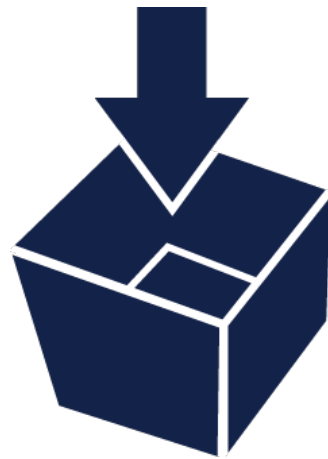
Customizability

```
jasonwilliams — bash — 44x11
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
C- /iplant/home/williams/analyses/TASSEL_4
.3.0_MLM_analysis1-2013-09-11-20-17-30.232
C- /iplant/home/williams/analyses/TASSEL_4
.3.0_MLM_analysis1-2013-09-12-14-52-35.844
C- /iplant/home/williams/analyses/Test_of_
New_App_analysis1-2013-10-25-14-40-49.857
```

Disadvantage -

Requires some command line
expertise

Cyberduck and iCommands Demo



Discovery Environment



Discovery Environment

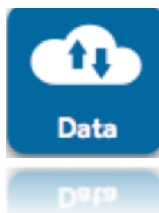
- ✓ A platform that can run almost any bioinformatics application
- ✓ Seamlessly integrated with data and high performance computing
- ✓ User extensible – add your own applications





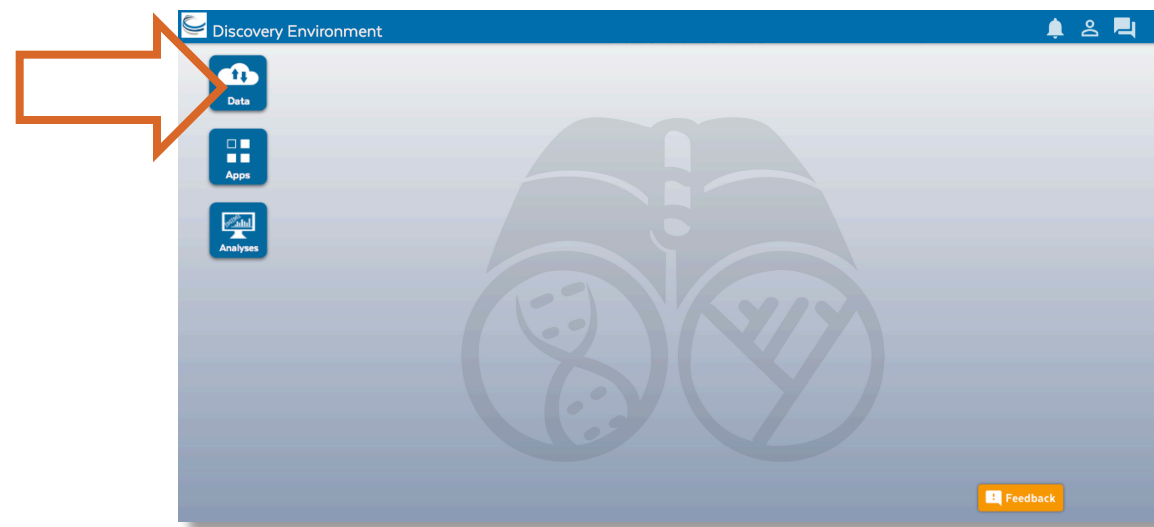
Discovery Environment Overview

Manage data



Data

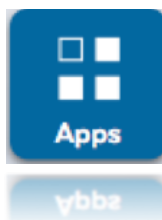
- Upload / Download files and folders
- Share files via URL (Public Links)
- Share files/folders with other users





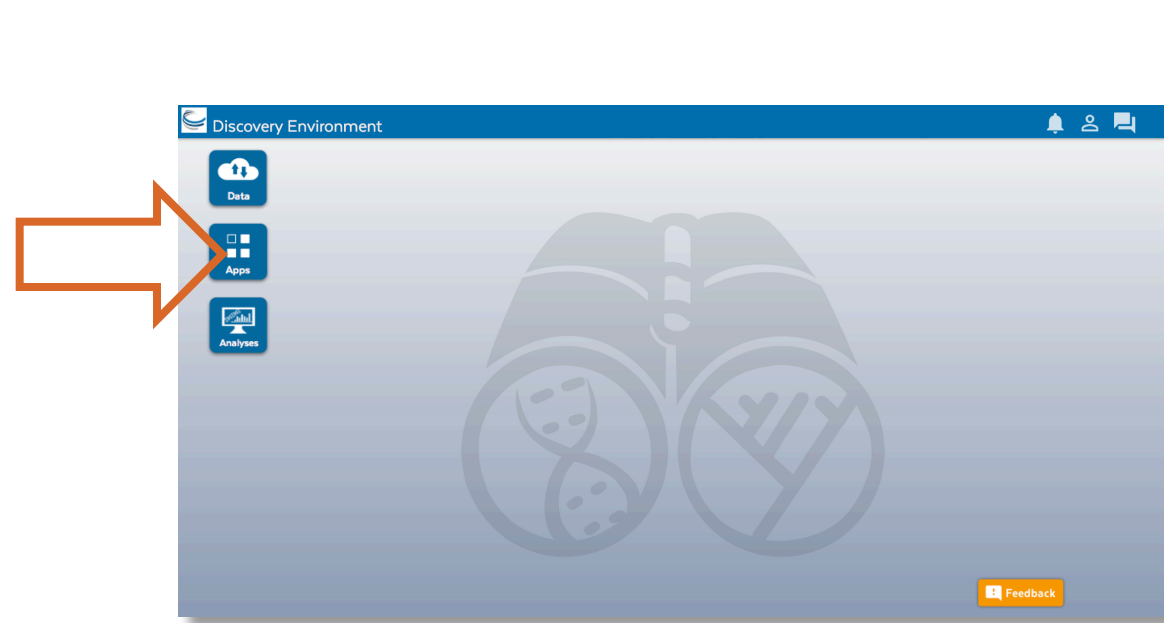
Discovery Environment Overview

Analyze data and customize Applications



Apps

- Run hundreds of bioinformatics Apps
- Build automated workflows
- Modify Apps or integrate new ones





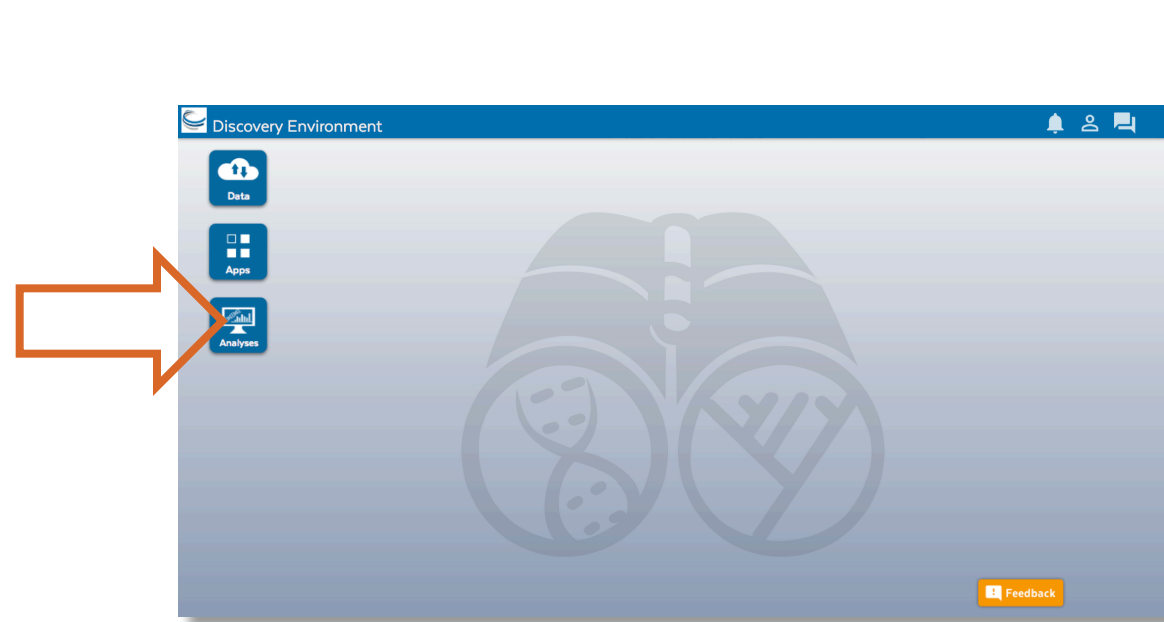
Discovery Environment Overview

View history, find results, reproduce analyses, optimize parameters



Analyses

- Monitor job status and find results
- Cancel jobs or re-launch jobs
- Detailed job history



Discovery Environment Demo





Discovery Environment

Demo analysis – sequence alignment using MUSCLE

Task: Take unaligned DNA sequences in FASTA format and create a multiple alignment

- ✓ View sample data in iPlant Data Store
- ✓ Launch a job using the MUSCLE sequence alignment app
- ✓ Monitor the job progress and view results



Atmosphere





Atmosphere

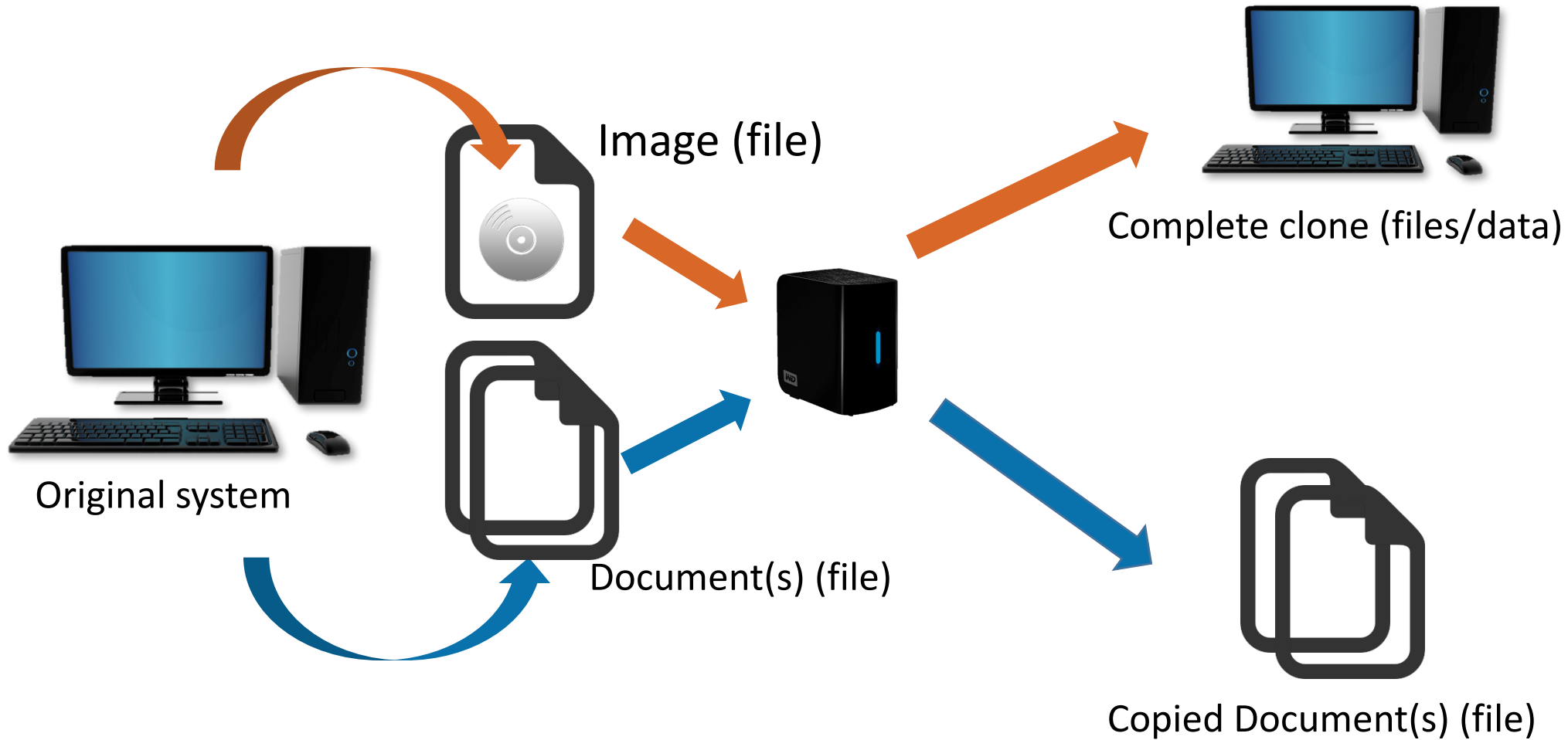
- ✓ Simple: 1-click access to hundreds of virtual machine images
- ✓ Flexible: Fully customize your software setup
- ✓ Powerful: Integrated with CyVerse computing and data resources





What is Cloud Computing?

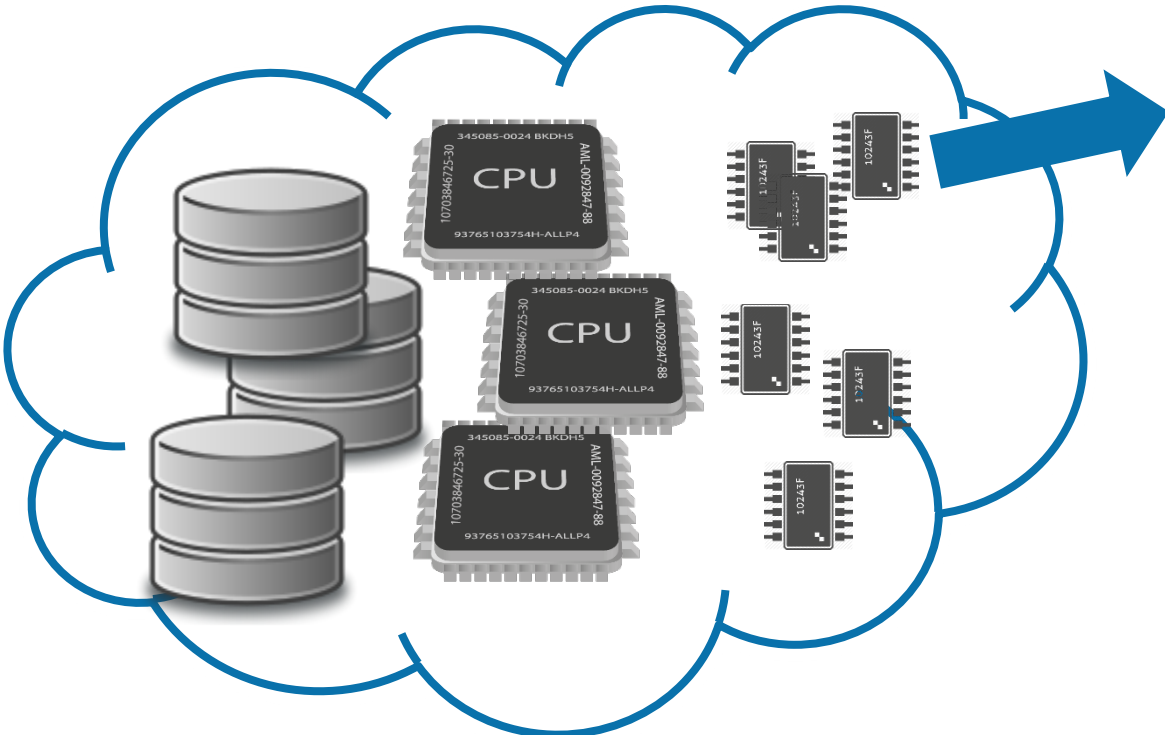
Important concepts: Image



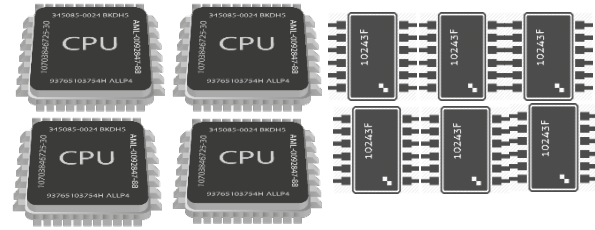


What is Cloud Computing?

Important concepts: Instance



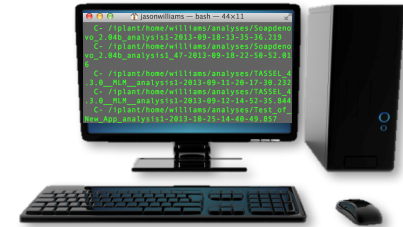
CyVerse Cloud



(Disk + CPU + Memory) + (Image)



128.196.34.158



Atmosphere Instance
(virtual machine)





Atmosphere Overview

Largest, easiest to use cloud for Life Sciences

CYVERSE BETA | Dashboard | Projects | Images | Providers | Help

Getting Started

- Launch New Instance**
Browse Atmosphere's list of available images and select one to launch a new instance.
- Browse Help Resources**
View a video tutorial, read the how-to guides, or email the Atmosphere support team.
- Change Your Settings**
Modify your account settings, view your resource quota, or request more resources.

- Choose an existing image or customize
- Instances up to 16-Core / 128 GB RAM
- Access via shell or VNC
- Share you image with selected users, or make them public





Atmosphere

Cloud computing for life sciences: sample use cases

- Run the software and data that are monopolizing your laptop/desktop
- Use desktop enabled images to run visually oriented programs (GUI)
- SUDO access – manage complex dependencies
- Uniform computing setups for your lab, collaborators, and students
- Make your own software available to a larger user community



Atmosphere Demo





Atmosphere

Cloud computing for life sciences

Windows



VNC Viewer



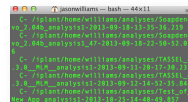
PuTTY

Mac



128.196.65.175

VNC Viewer

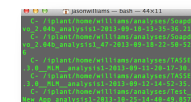


Shell/terminal

Linux



VNC Viewer



Shell/terminal

VNC Viewer:

PuTTY:

www.realvnc.com/download/viewer

www.putty.org



Where to go from here:



Learning Center

- Get Started Guide
- Tutorials and Videos
- Documentation

Upcoming Events

- Workshops
- Webinars



Slides and Tutorials



www.cyverse.org/sol2016

