CYVERSE™

Transforming Science Through Data-driven Discovery

CyVerse Overview and Tutorial

CSHL 2016 Cereal Genomics Course

Jason Williams – Education, Outreach, Training Lead

Cold Spring Harbor Laboratory

williams@cshl.edu @JasonWilliamsNY

NSF

THE UNIVERSITY OF ARIZONA

TACC TEXAS ADVANCED COMPUTING CENTER

CSH Cold Spring Harbor Laboratory

UNCW

Slides and Tutorials

www.cyverse.org/cereal2016

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

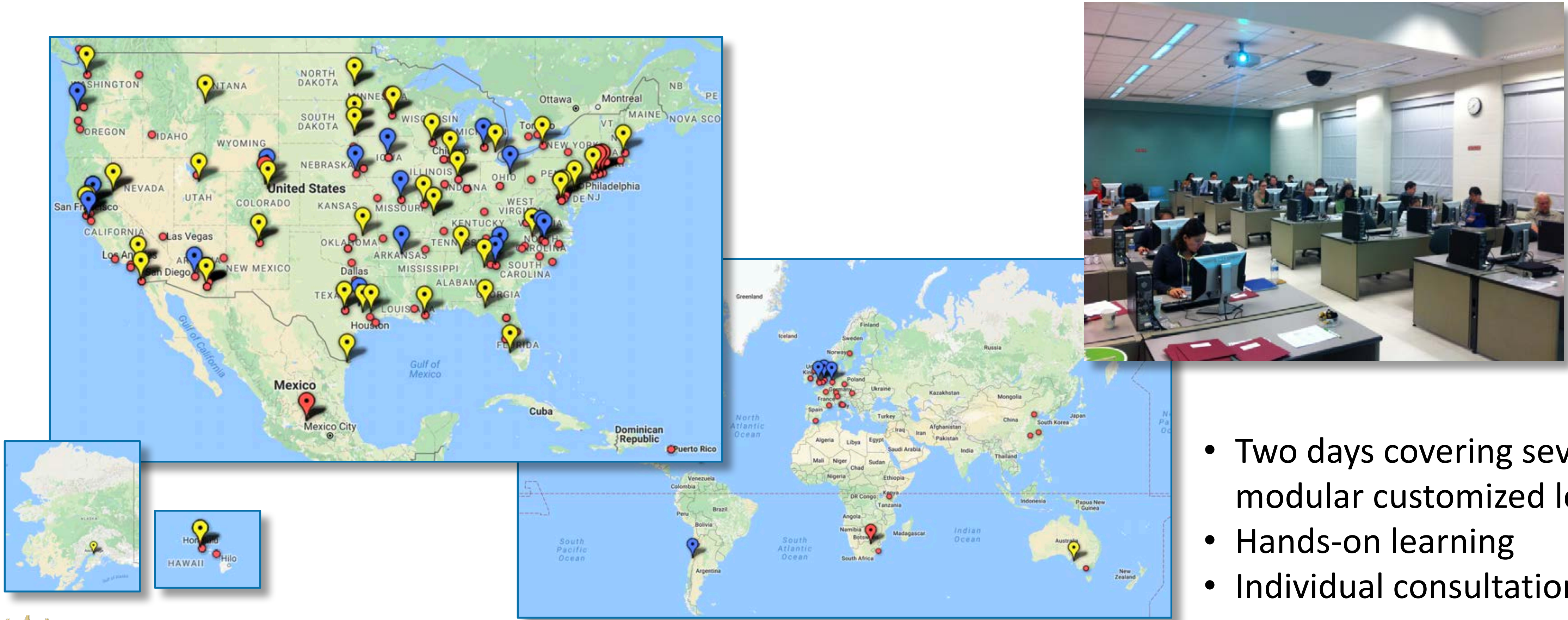
More than 30K users, PB of data, and hundreds of publications, courses, and discoveries

Apologies...

"I had the feeling I have been exposed to many bioinformatics tools but I would be unable to use any of them on my own."

There is help (lots)

CyVerse Workshops can come to you



- Two days covering several modular customized lessons
- Hands-on learning
- Individual consultations

Community-driven learning

software carpentry

DATA CARPENTRY

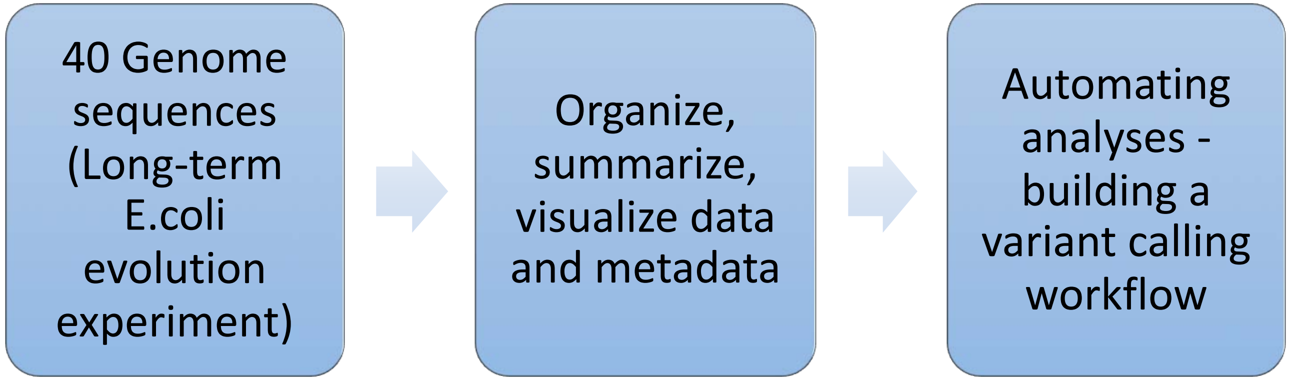
MAKING DATA SCIENCE MORE EFFICIENT

Goal: provide basic lab skills for research computing; “get more done in less time and with less pain.”

Goal: provide researchers training on the fundamentals and best practices in data analysis and management.

- **Scientists teaching scientists**
- **All-volunteer Instructors (>500 world-wide)**
- **Community-maintained lessons**
- **No assumptions of knowledge for learners**

Genomics Lesson Narrative



40 Genome sequences (Long-term E.coli evolution experiment)

Organize, summarize, visualize data and metadata

Automating analyses - building a variant calling workflow

Cover the ‘unspoken’ protocols make for effective, reproducible research

Hands-on lessons run from the cloud

Some learning goals

Interacting with Computers

- Could Computing
- Connecting to remote computing (SSH/PuTTY)
- File Transfer (FileZilla, other command-line tools: scp, rsynch, wget, etc.)

Data Management and Organization

- Open source
- Metadata and reproducibility
- Important genomics file formats (CSV/TSV, FastQ, SAM/BAM, VCF, etc.)
- Organizing a filesystem for computational projects (Linux)
- Unix Shell (command-line: ls, cd, mkdir, cp, rm, wc, grep, cut, columns, head, tail, less etc,)
- R: Creating projects, scripts, and examining data

Automation and scripting

- R scripting
- 'For' loops
- Building automated pipelines
- Using multithreaded applications

Data Cleaning and visualization

- R: various packages and functions
- R: dplyr
- R: ggplot
- FastQC - quality control of high-throughput sequence data
- Trimmomatic - filtering and trimming of high-throughput sequence data
- Integrated Genome Viewer

CyVerse Evolution

iPlant 2008
Empowering a New Plant Biology

iPlant 2013
Cyberinfrastructure for Life Science

CyVerse 2016
Transforming Science Through Data-Driven Discovery

CyVerse Evolution

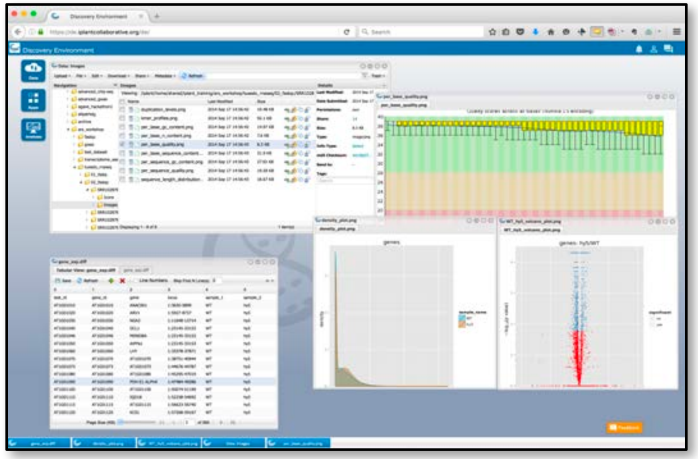
We are funded by the National Science Foundation

DBI-0735191 and DBI-1265383

- 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?

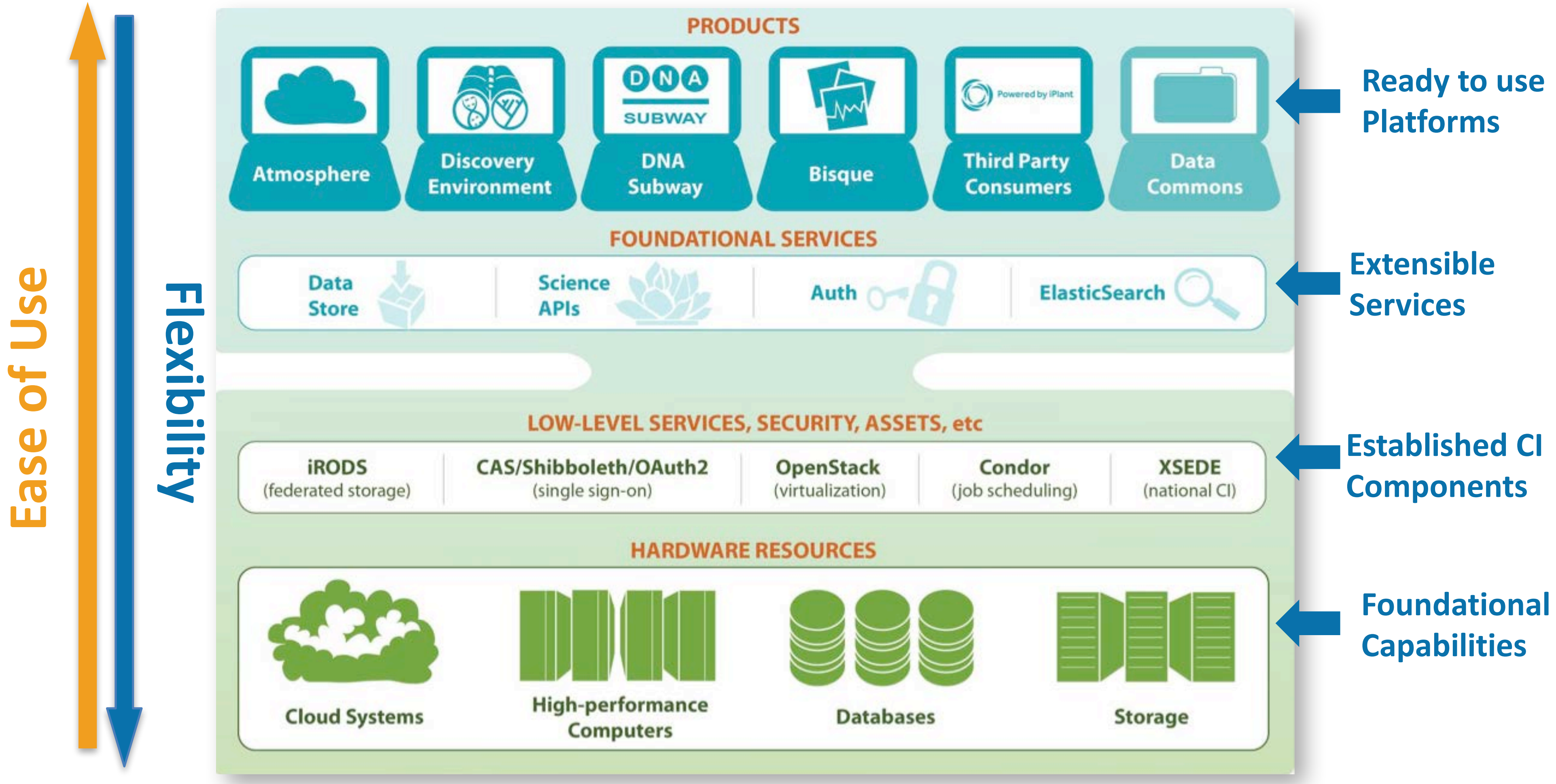


Platforms, tools, datasets

Storage and compute

Training and support

CyVerse product stack



CyVerse Institutions

THE UNIVERSITY OF ARIZONA

TACC
TEXAS ADVANCED COMPUTING CENTER

CSH
Cold Spring Harbor Laboratory

UNCW

CyVerse is a collaborative virtual organization

CyVerse UK

WARWICK
THE UNIVERSITY OF WARWICK

UNIVERSITY OF LIVERPOOL

BBSRC

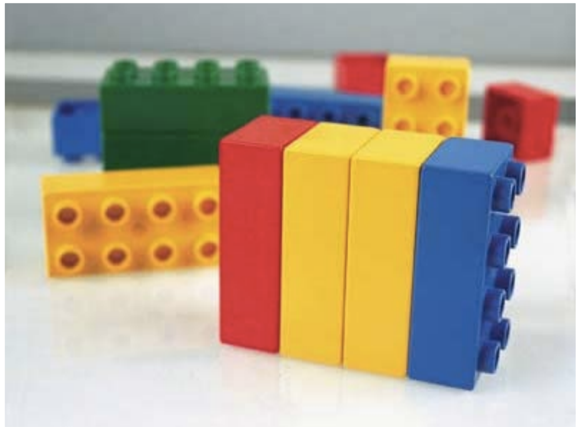
The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Earlham Institute

NSF

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



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: 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

Getting Data into CyVerse

CyVerse Data Store

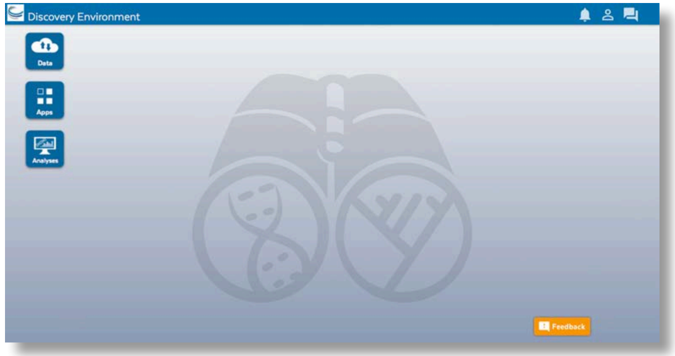
- Store any type of file related to your research
- Move files seamlessly between CyVerse platforms
- Automate file transfers
- Share files with lab members, collaborators, and communities

CyVerse Data Store

Multiple ways to access

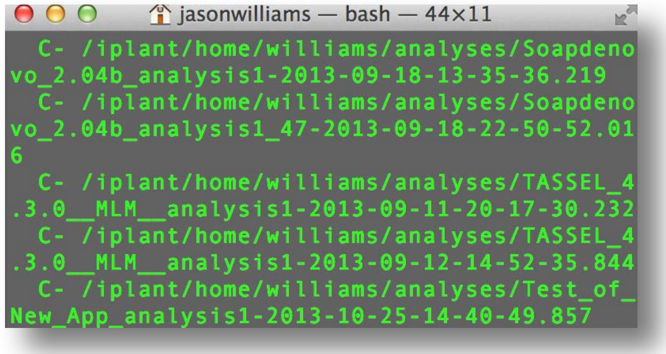
Point-and-click

Cyberduck



Discovery Environment

Command line

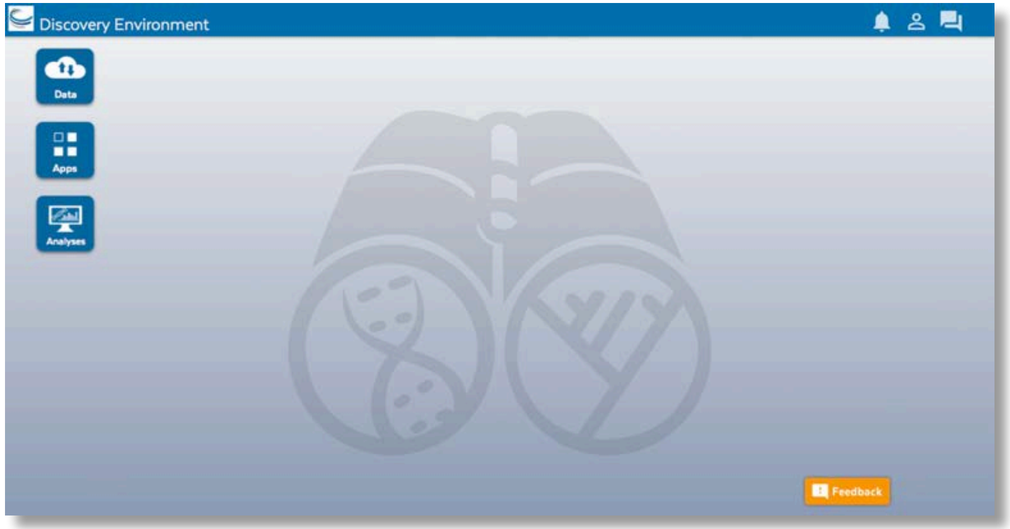


```
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
  C- /iplant/home/williams/analyses/TASSEL_4
.3.0__MLM__analysis1-2013-09-11-20-17-30.232
  C- /iplant/home/williams/analyses/TASSEL_4
.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 +

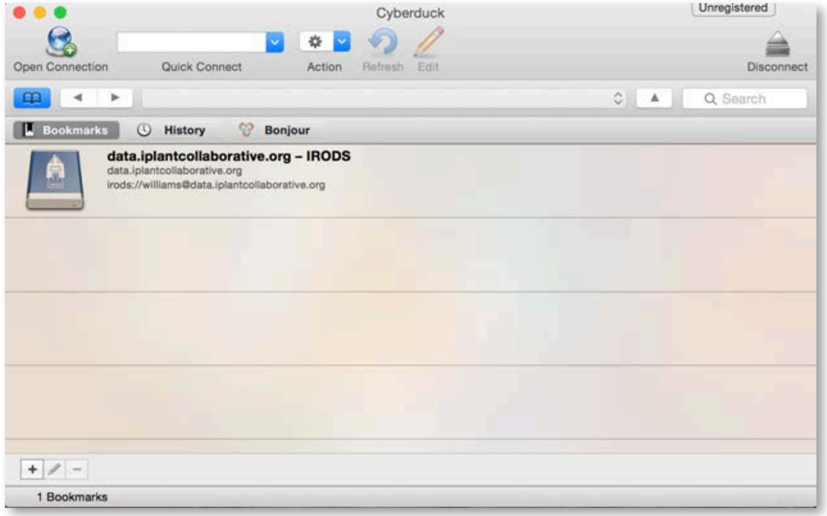
Covers most upload/download sharing needs

Disadvantage -

Some size/speed limitations

Cyberduck

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

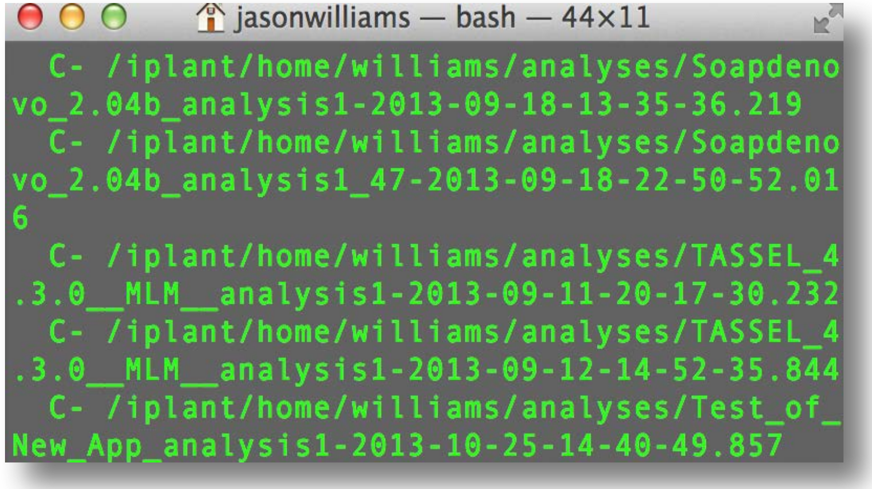


data.iplantcollaborative.org – IRODS
data.iplantcollaborative.org
irods://williams@data.iplantcollaborative.org

Advantage +	**Disadvantage -**
More like desktop file systems	No permissions/metadata control

iCommands

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



```
  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
```

Advantage +

Customizability

Disadvantage -

Requires some command line expertise

Cyberduck and iCommands Demo

Discovery Environment

Discovery Environment

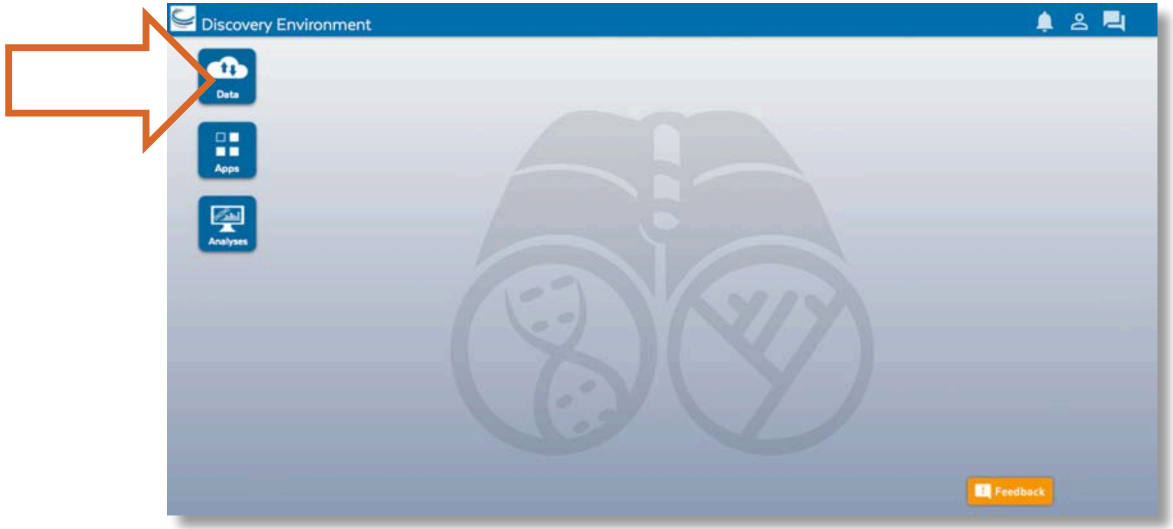
- ✓ A platform that can run almost any bioinformatics application
- ✓ Seamlessly integrated with data and high performance computing
- ✓ User extensible – add your own applications

Discovery Environment Overview

Manage data

Data

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

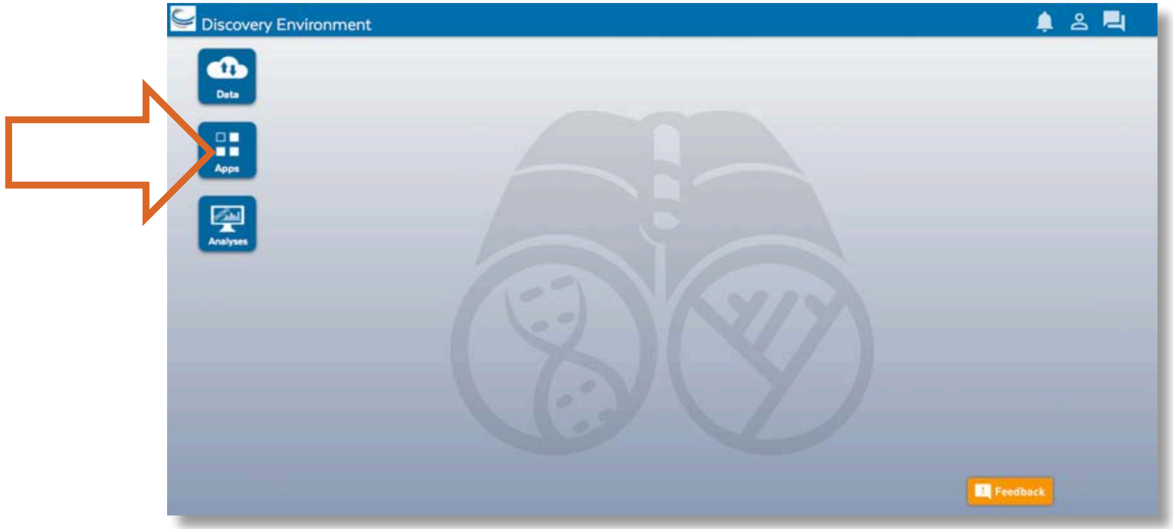


Discovery Environment Overview

Analyze data and customize Applications

Apps

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

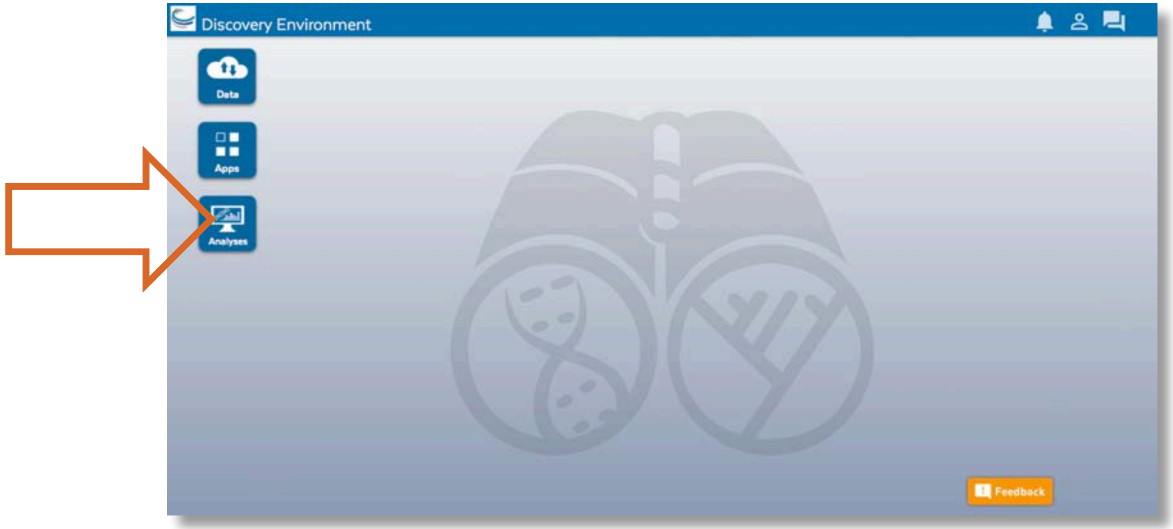


Discovery Environment Overview

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 Data Store
- ✓ Launch a job using the MUSCLE sequence alignment app
- ✓ Monitor the job progress and view results

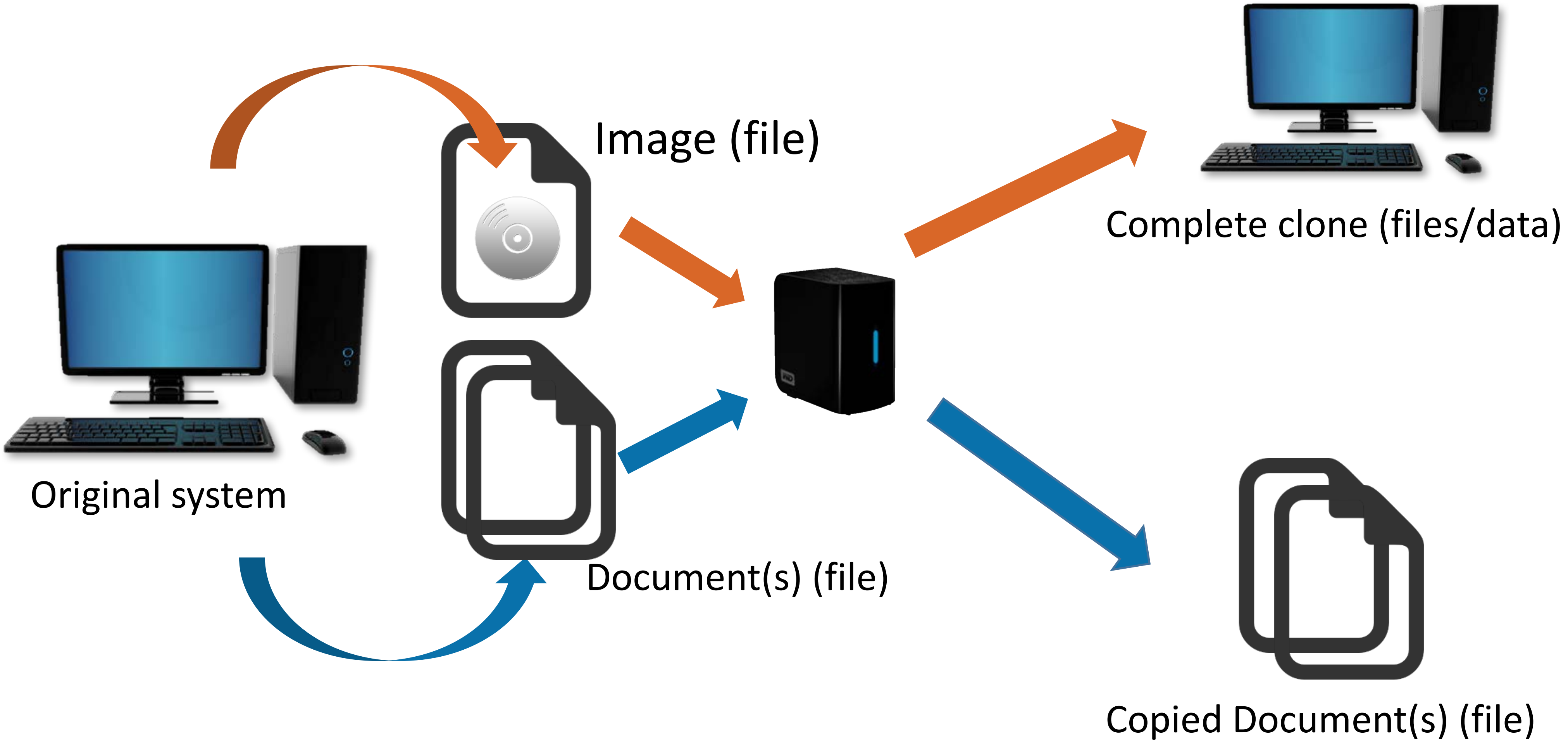
Atmosphere

Atmosphere

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

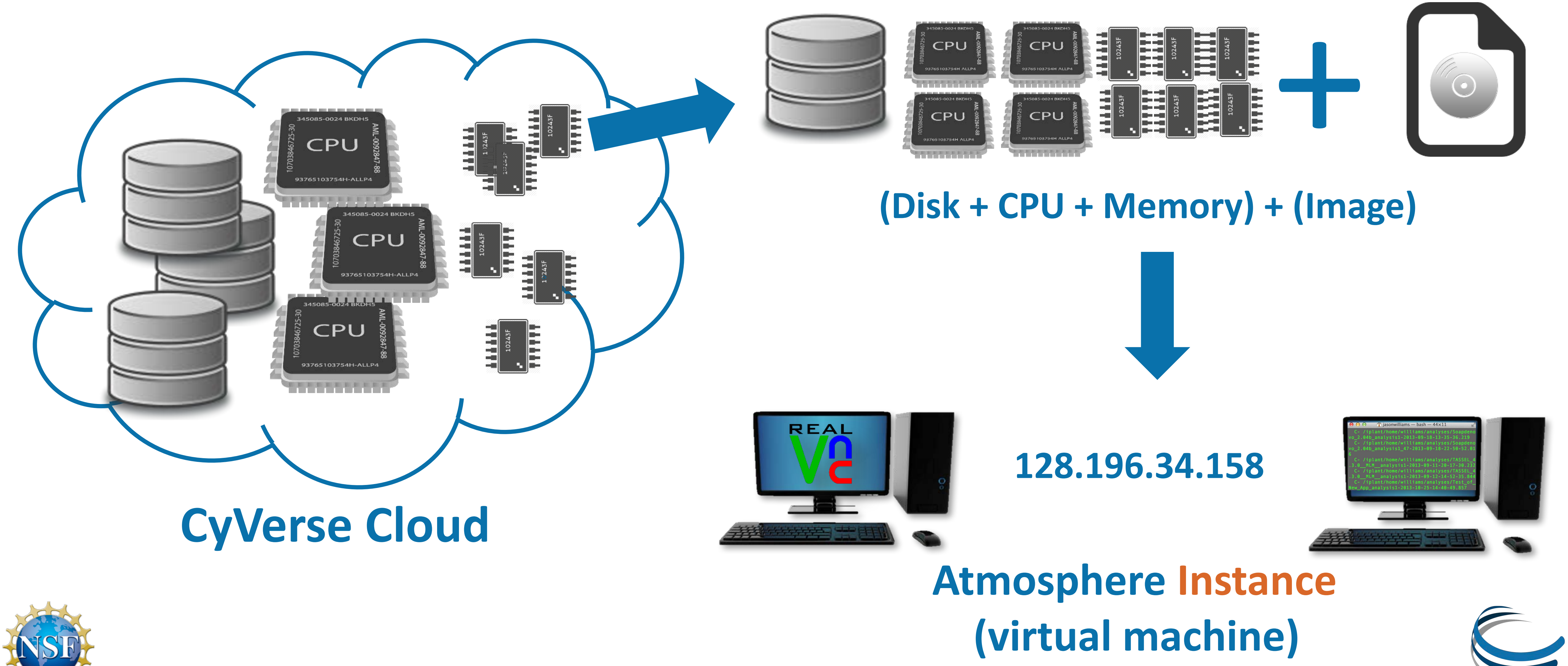
What is Cloud Computing?

Important concepts: Image



What is Cloud Computing?

Important concepts: Instance



CyVerse Cloud

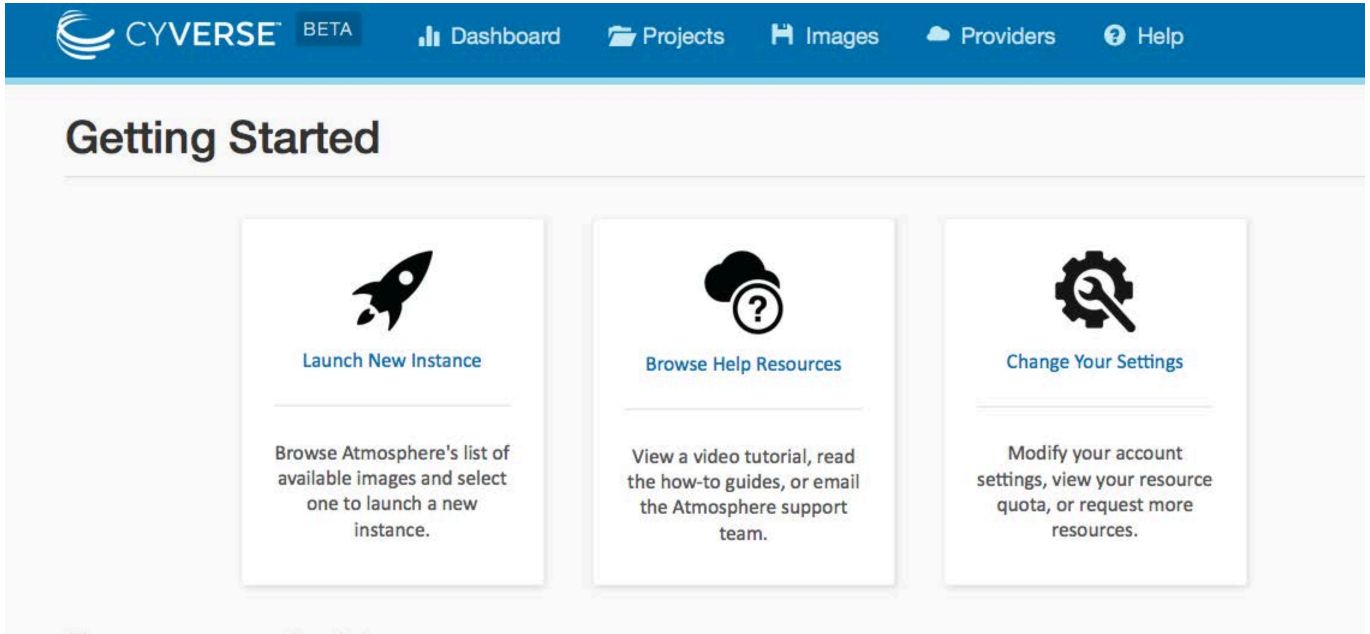
(Disk + CPU + Memory) + (Image)

128.196.34.158

Atmosphere Instance
(virtual machine)

Atmosphere Overview

Largest, easiest to use cloud for Life Sciences



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

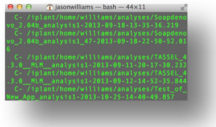
Windows

VNC Viewer

PuTTY

Mac

VNC Viewer

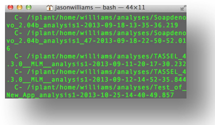


```
jasonwilliams — bash — 44×11
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
  C- /iplant/home/williams/analyses/TASSEL_4
.3.0__MLM__analysis1-2013-09-11-20-17-30.232
  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
```

Shell/terminal

Linux

VNC Viewer

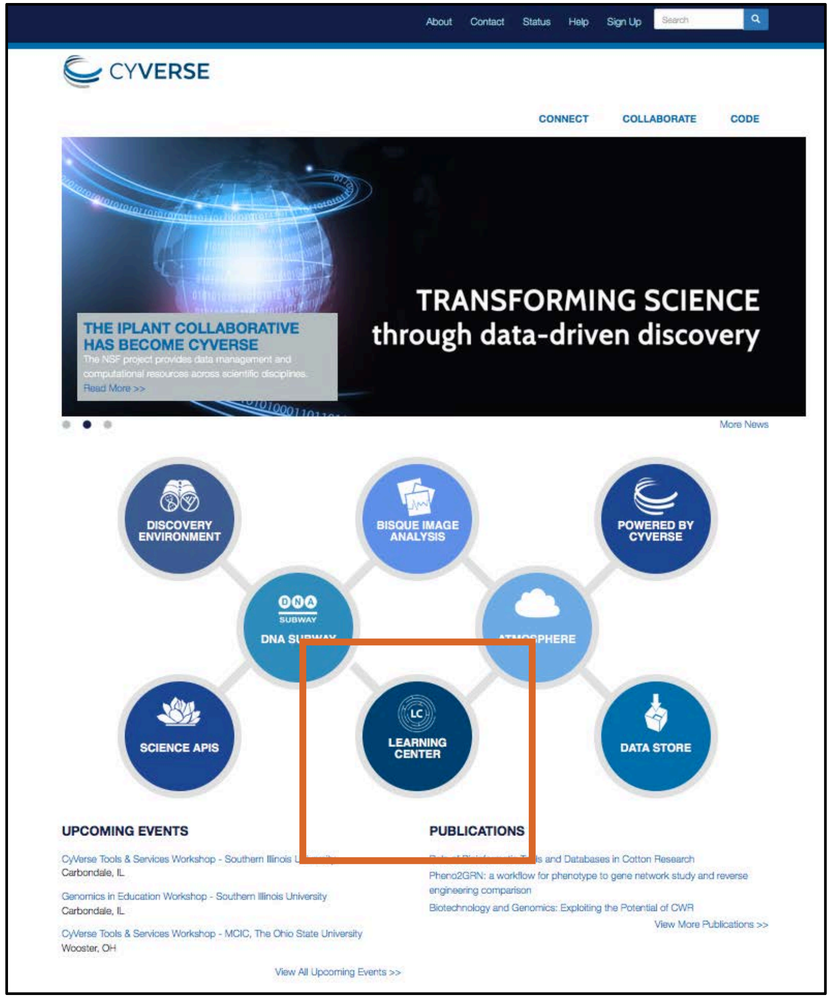


```
jasonwilliams — bash — 44×11
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
  C- /iplant/home/williams/analyses/TASSEL_4
.3.0__MLM__analysis1-2013-09-11-20-17-30.232
  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
```

Shell/terminal

VNC Viewer:	www.realvnc.com/download/viewer
PuTTy:	www.putty.org

Where to go from here:



TRANSFORMING SCIENCE through data-driven discovery

THE IPLANT COLLABORATIVE HAS BECOME CYVERSE

The NSF project provides data management and computational resources across scientific disciplines.

Read More >>

More News

UPCOMING EVENTS

CyVerse Tools & Services Workshop - Southern Illinois U
Carbondale, IL

Genomics in Education Workshop - Southern Illinois University
Carbondale, IL

CyVerse Tools & Services Workshop - MCIC, The Ohio State University
Wooster, OH

View All Upcoming Events >>

PUBLICATIONS

ls and Databases in Cotton Research

Pheno2GRN: a workflow for phenotype to gene network study and reverse engineering comparison

Biotechnology and Genomics: Exploiting the Potential of CWR

View More Publications >>

Learning Center

- Get Started Guide
- Tutorials and Videos
- Documentation

Upcoming Events

- Workshops
- Webinars

CYVERSE

Transforming Science Through Data-driven Discovery

Executive Team

THE UNIVERSITY OF ARIZONA

Parker Antin
Nirav Merchant
Eric Lyons

TACC
TEXAS ADVANCED COMPUTING CENTER

Matt Vaughn

CSH Cold Spring Harbor Laboratory

Doreen Ware
Dave Micklos

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

Slides and Tutorials

www.cyverse.org/cereal2016

NSF

THE UNIVERSITY OF ARIZONA

TACC
TEXAS ADVANCED COMPUTING CENTER

CSH
Cold Spring Harbor Laboratory

UNCW