# CyVerse Data Store

TM

## Managing Your ‘Big’ Data

THE UNIVERSITY OF ARIZONA

TACC
TEXAS ADVANCED COMPUTING CENTER

CSH Cold Spring Harbor Laboratory

UNCW

# Download Slides and Follow Along

mcbios.readthedocs.org

# Welcome to the Data Store

Manage and share your data across all CyVerse platforms

# Working with Big Data

## Challenges: the scope and scale of life sciences data continue to grow

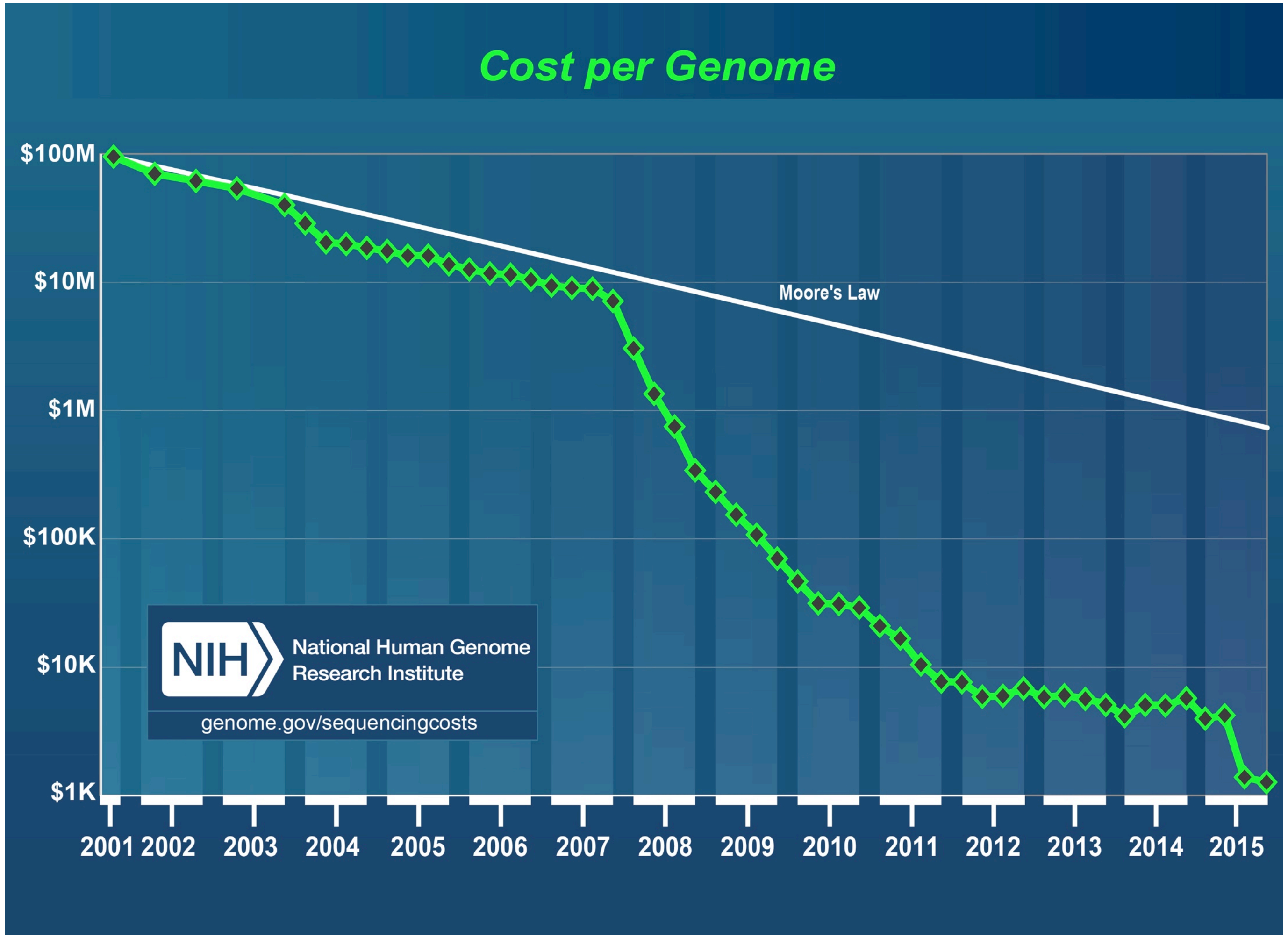
The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Science
information
climate
research
data
nature
BIG DATA
SCIENCE IN THE PETABYTE ERA

- Big data a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time
- Big data sizes are a constantly moving target currently ranging from a few dozen terabytes (TB) to many petabytes of data in a single data set.

"'Big Data': Big gaps of knowledge in the field of Internet". *International Journal of Internet Science* **7**: 1–5.

# Working with Big Data

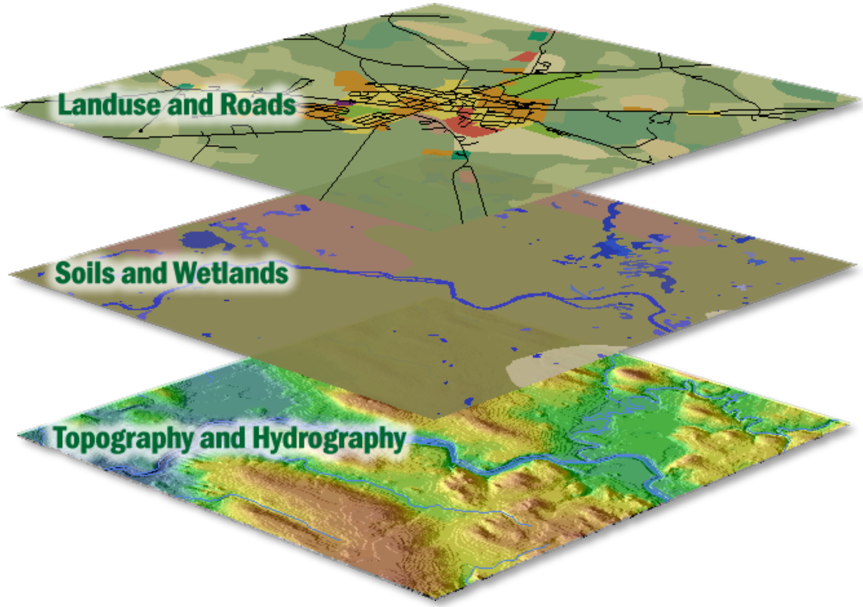
## Challenges: data generation is cheaper and faster

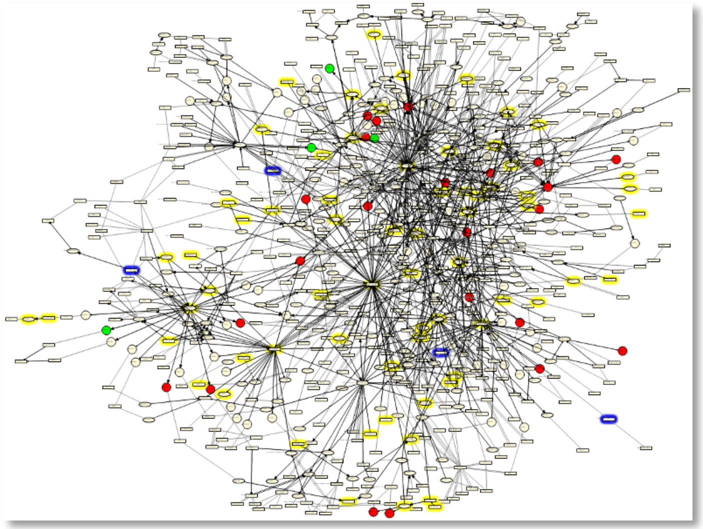


http://www.genome.gov/sequencingcosts/

# Working with Big Data

## Challenge: biology encompasses more than sequence data





Advanced Imaging

Geospatial

Network

Biologists work with and require access to diverse data types

# Working with Big Data

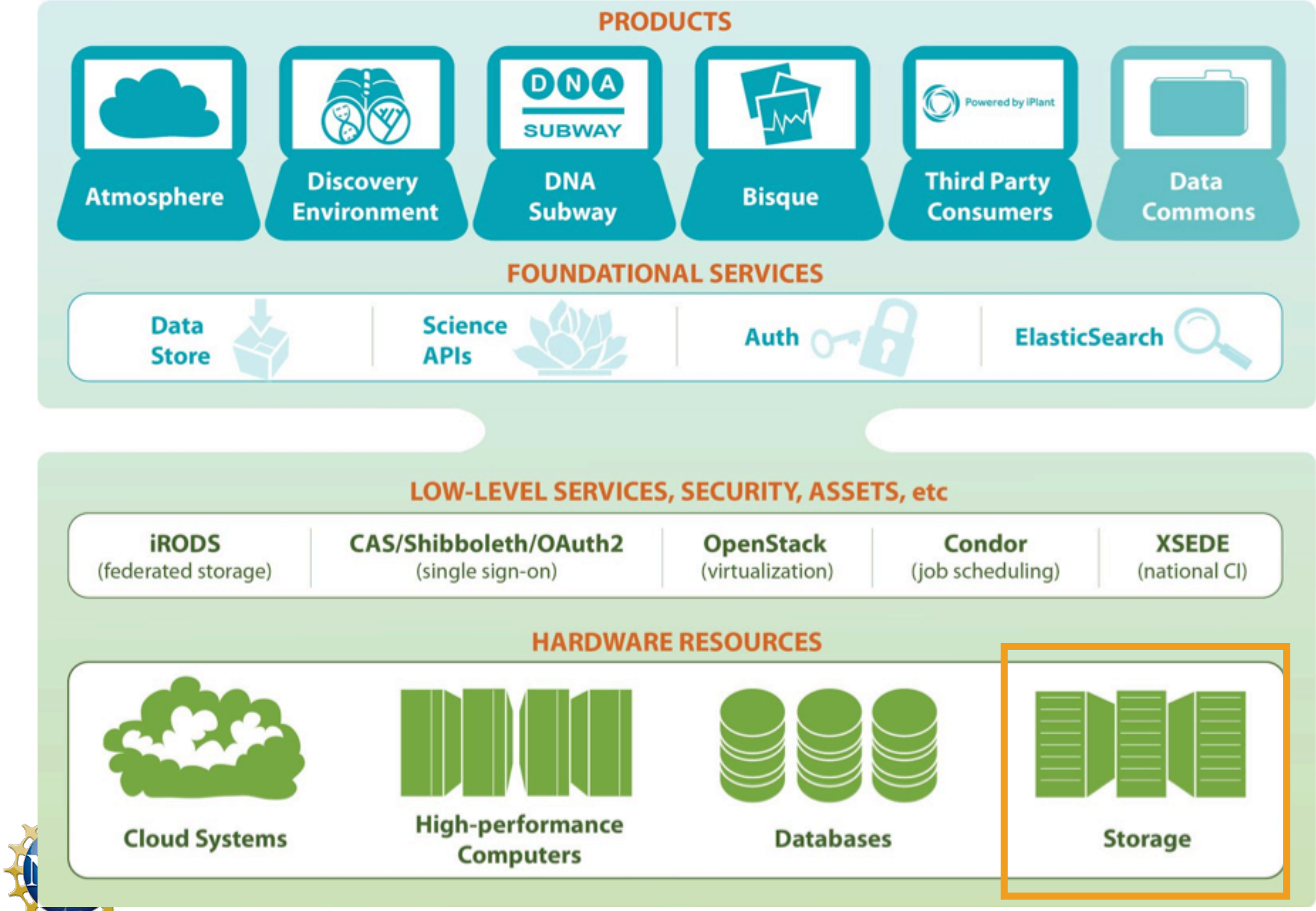
## Challenges: changes in data require changes in tools

- Difficult / slow transfers
- Expense for storage / backup
- Difficult to share and publish
- Analysis
- Metadata (What Is metadata?)

Changes in scale introduce quantitative and qualitative complications

# Data Store Overview

## The Data Store services all CyVerse platforms



PRODUCTS

Atmosphere | Discovery Environment | DNA Subway | Bisque | Third Party Consumers | Data Commons

FOUNDATIONAL SERVICES

Data Store | Science APIs | Auth | ElasticSearch

LOW-LEVEL SERVICES, SECURITY, ASSETS, etc

iRODS (federated storage) | CAS/Shibboleth/OAuth2 (single sign-on) | OpenStack (virtualization) | Condor (job scheduling) | XSEDE (national CI)

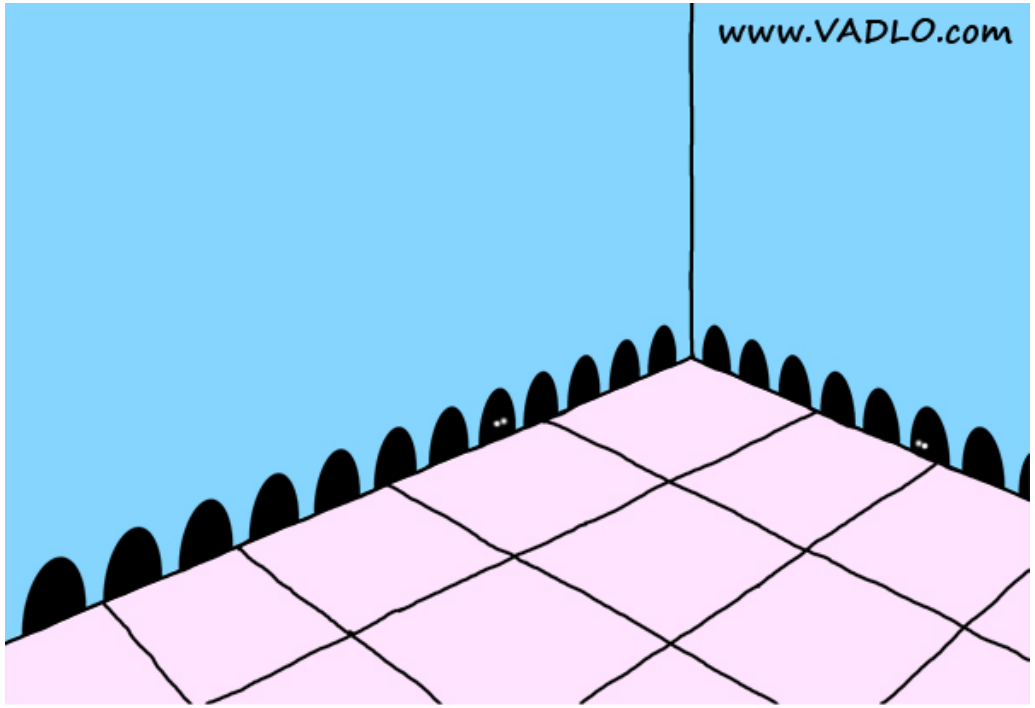
HARDWARE RESOURCES

Cloud Systems | High-performance Computers | Databases | Storage

- Access your data from multiple CyVerse services
- Automatic backup (redundant between University of Arizona and University of Texas
- Default 100 GB allocation, > 1 TB allocations available with justification

# Data Store Overview

## Avoid reinventing the wheel



Mouse house that did not receive infrastructure funding.

- iRODS (integrated Rule-Oriented Data System) is an established, scalable, open-source data management sytem
- iRODS supports many data intensive projects
- iRODS abstracts data services from data storage to facilitate executing services across heterogeneous, distributed storage systems

# Data Store Overview

## Benefits

Get Science Done

- Store any type of files related to your research
- An evolving "Data Commons" lets you access important datasets

Reproducibility

- Metadata captures information needed for reproducibility
- Automatic backup and accessibility support your data management plan

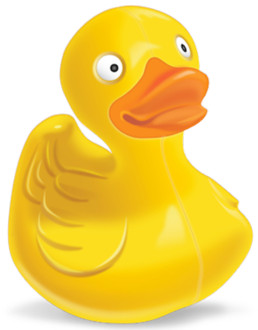
Productivity

- iRODS makes high-speed transfers possible (100 GB in ~30 min)
- Share data instantly with collaborators within CyVerse

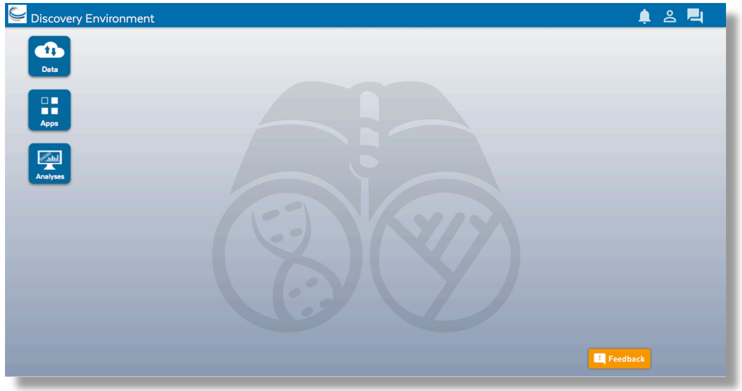
# Data Store Overview

## Multiple ways to access

## Point-and-click

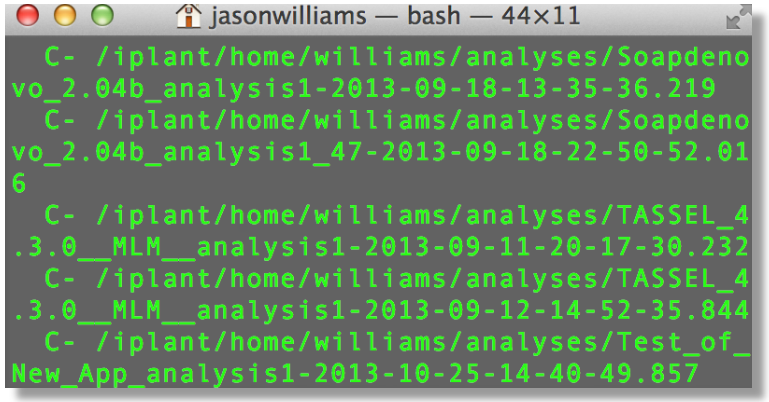


Cyberduck



Discovery Environment

## Command line



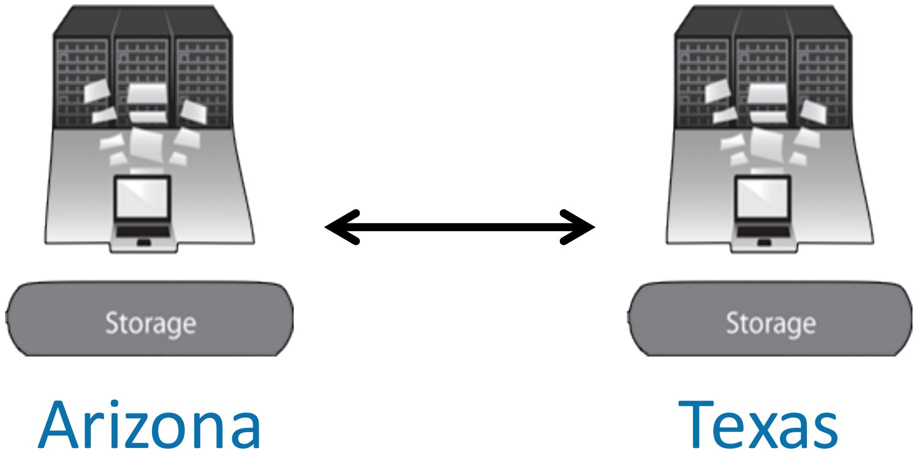
```
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1-2013-09-18-13-35-36.219
  C- /iplant/home/williams/analyses/Soapdeno
vo_2.04b_analysis1_47-2013-09-18-22-50-52.01
6
  C- /iplant/home/williams/analyses/TASSEL_4
.3.0__MLM__analysis1-2013-09-11-20-17-30.232
  C- /iplant/home/williams/analyses/TASSEL_4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

# Data Store Overview

## Some important things we will not "see" in the demo

### Data Backups



Key component of your data management

Worry-free

### Data Transfer

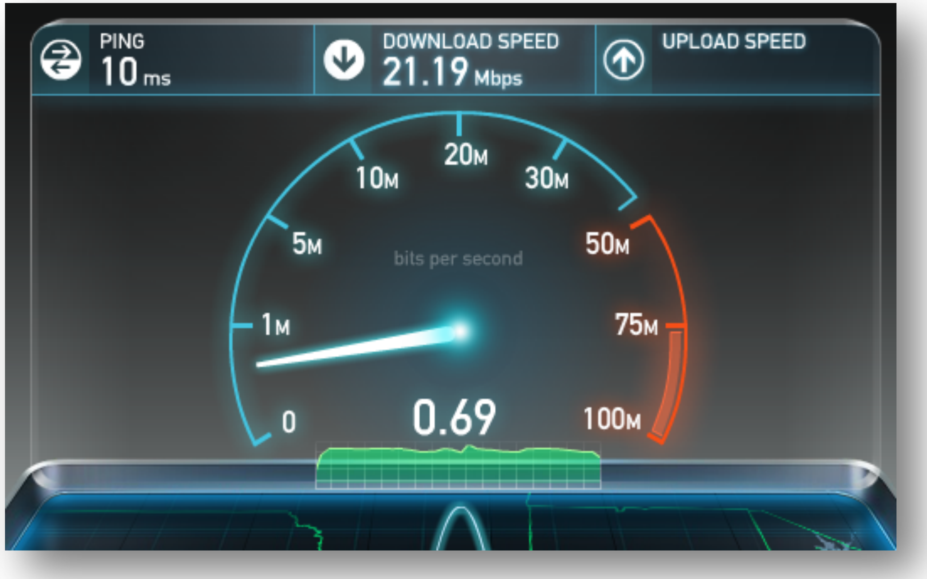
| Source | Destination | Copy Method | Time (seconds) |
|---|---|---|---|
| CD | My Computer | cp | 320 |
| Berkeley Server | My Computer | scp | 150 |
| External Drive | My Computer | cp | 36 |
| USB 2.0 Flash | My Computer | cp | 30 |
| Data Store | My Computer | iget | 18 |
| My Computer | My Computer | cp | 15 |

Closer to optimum conditions: transfers between University of Arizona and UC Berkeley

100 GB: 26m15s, 1 GB 17.5s

# Data Store Overview

## Some important things we will not “see” in the demo



http://www.speedtest.net/

Local connections and institutional policies limit data transfer

# Data Store Overview

Hands-on demo

# Data Store Overview

## User perspectives and potential applications

### Bench Scientist

- Uploads all of his .fastq files along with 50GB of root growth videos
- Shares all his analyses results with his thesis advisor

### Bioinformatician

- Created a metadata template for assembled genomes her students and collaborators will place in a shared folder
- Uses public links in the supplemental materials of her publications

### Core Facilities

- Developed a script to automate transfer of data to core users
- Uses a shared folder to make large datasets accessible

Images from personas based on: Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies
PLOS Biology DOI: 10.1371/journal.pcbi.1003496

# Data Store Overview

## Time for Summaries and Tips

# Tips for any transfer method

## Spaces / Special Characters

- Many software packages are sensitive to spaces in files names and/or the special characters below
- Rename uploaded files before using them in an analysis

~ ` ! @ # $ % ^ & * ( ) + =

{ } [ ] | \ : ; " ' < > , ? /

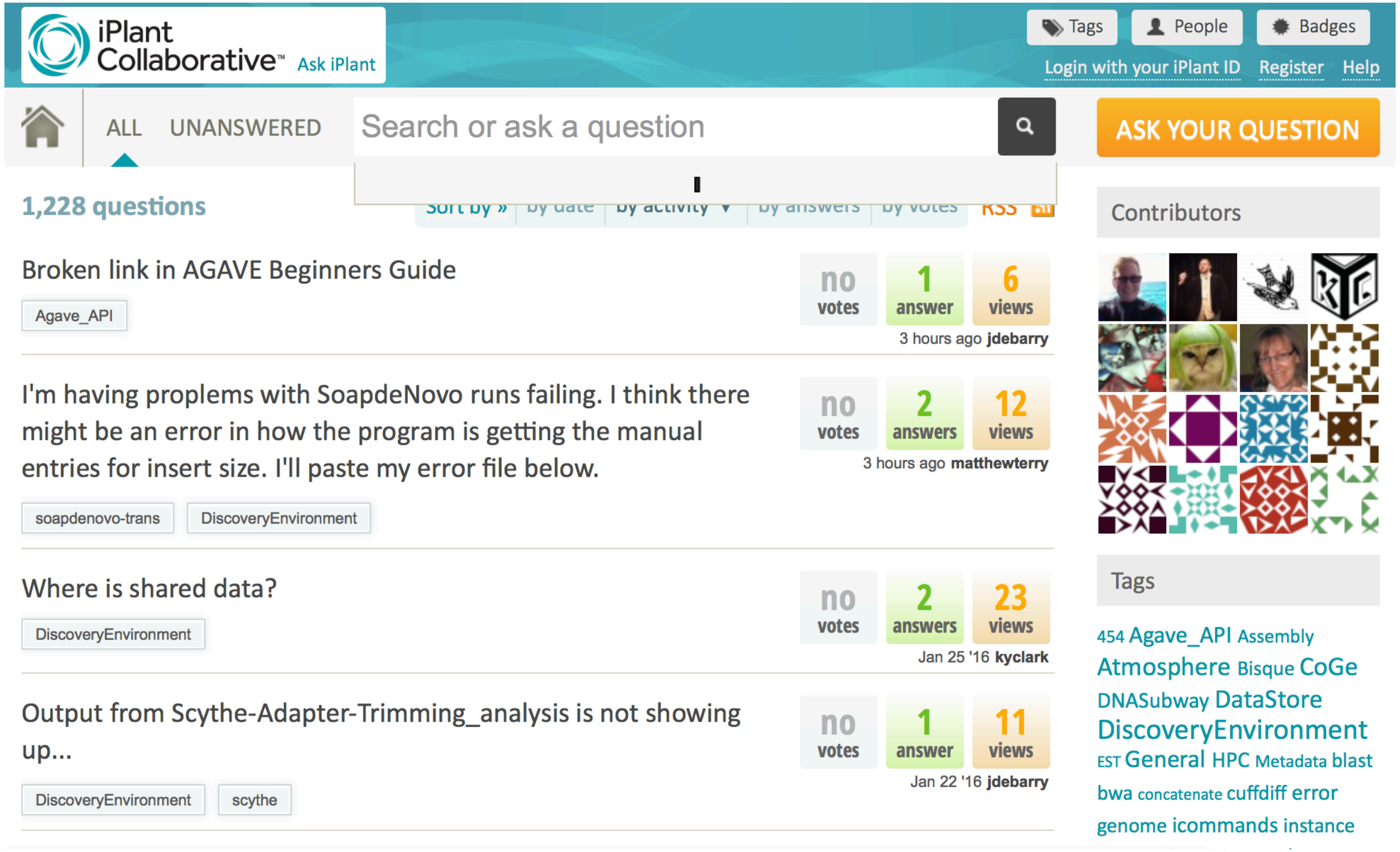
# Tips

## When sharing, use this chart to decide appropriate permissions

| Permission | Read | Download | Metadata | Rename | Move | Delete |
|---|---|---|---|---|---|---|
| Read | x | x |  |  |  |  |
| Write | x | x | x |  |  |  |
| Own | x | x | x | x | x | x |

# Keep asking: ask.iplantcollaborative.org

## Detailed instructions with videos, manuals, documentation in Learning Center



iPlant Collaborative™ Ask iPlant
Tags
People
Badges
Login with your iPlant ID
Register
Help
ALL
UNANSWERED
Search or ask a question
ASK YOUR QUESTION
Sort by » by date by activity ▼ by answers by votes RSS
Contributors
Tags

1,228 questions

Broken link in AGAVE Beginners Guide
Agave_API
no votes | 1 answer | 6 views
3 hours ago jdebarry

I'm having proplems with SoapdeNovo runs failing. I think there might be an error in how the program is getting the manual entries for insert size. I'll paste my error file below.
soapdenovo-trans | DiscoveryEnvironment
no votes | 2 answers | 12 views
3 hours ago matthewterry

Where is shared data?
DiscoveryEnvironment
no votes | 2 answers | 23 views
Jan 25 '16 kyclark

Output from Scythe-Adapter-Trimming_analysis is not showing up...
DiscoveryEnvironment | scythe
no votes | 1 answer | 11 views
Jan 22 '16 jdebarry

Tags: 454 Agave_API Assembly Atmosphere Bisque CoGe DNASubway DataStore DiscoveryEnvironment EST General HPC Metadata blast bwa concatenate cuffdiff error genome icommands instance