



OPEN SCIENCE DATA CLOUD



# Using Open Science Data Cloud



OPEN CLOUD CONSORTIUM



THE UNIVERSITY OF  
CHICAGO

**FIU**

FLORIDA  
INTERNATIONAL  
UNIVERSITY

**UIC**

UNIVERSITY OF ILLINOIS  
AT CHICAGO

cdis

CENTER FOR  
DATA INTENSIVE SCIENCE



OPEN SCIENCE DATA CLOUD

An open-source, cloud-based infrastructure that provides the scientific community with resources for managing terabyte and petabyte-scale scientific datasets





OPEN SCIENCE DATA CLOUD

## ***Storing data***

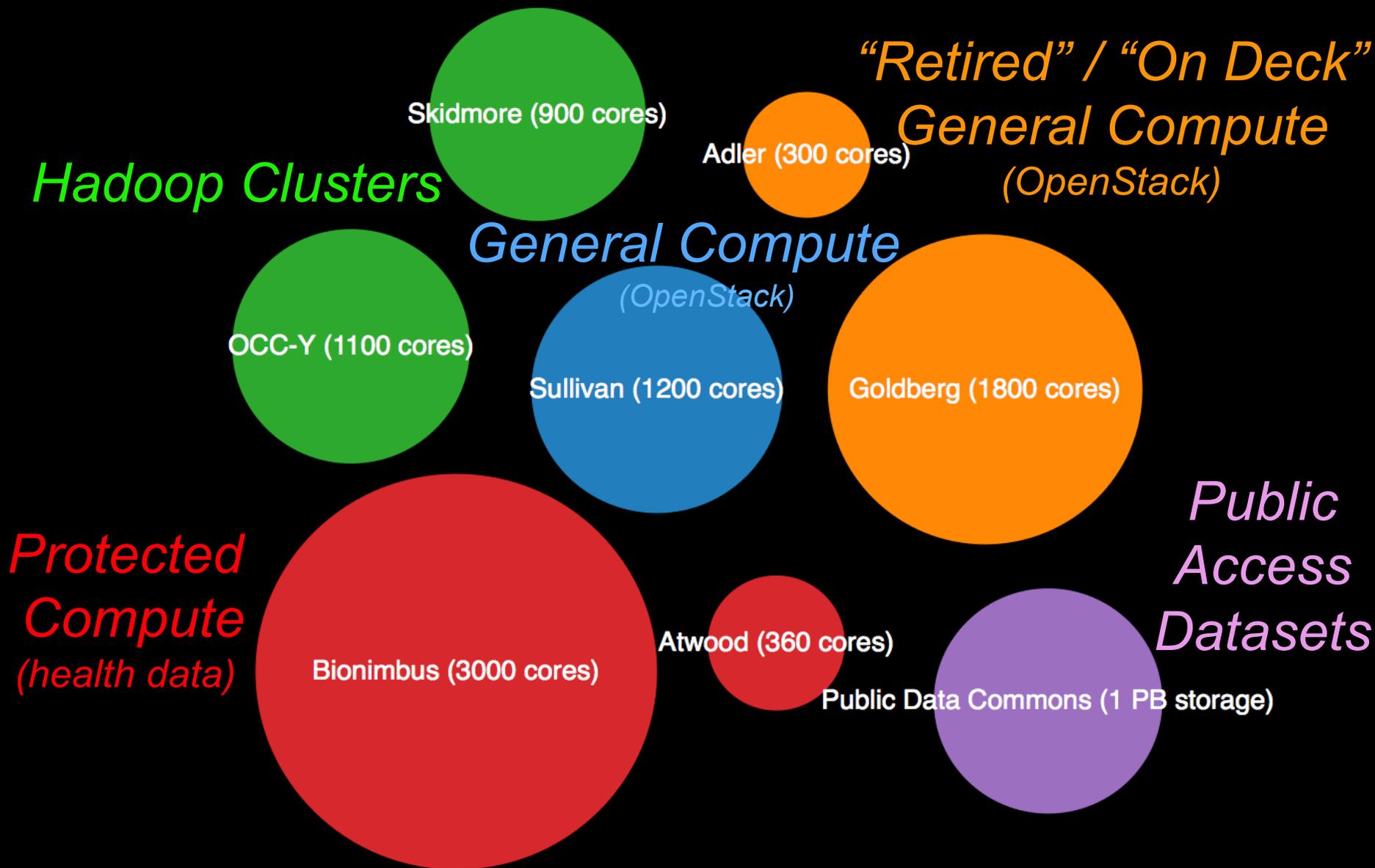
- 10 PB of storage across all resources
- Public and protected infrastructures for sensitive data
- 1 PB Public Data Commons of popular scientific datasets

## ***Analyzing data***

- About 9000 compute cores
- Both OpenStack and Hadoop-based clouds

## ***Sharing data and analysis tools***

- Access to shared group storage for collaborations
- Access to a public pool of virtual machine snapshots
- Authenticated log-in using University/Institution credentials
- Connected to high-performance networks for data transfer



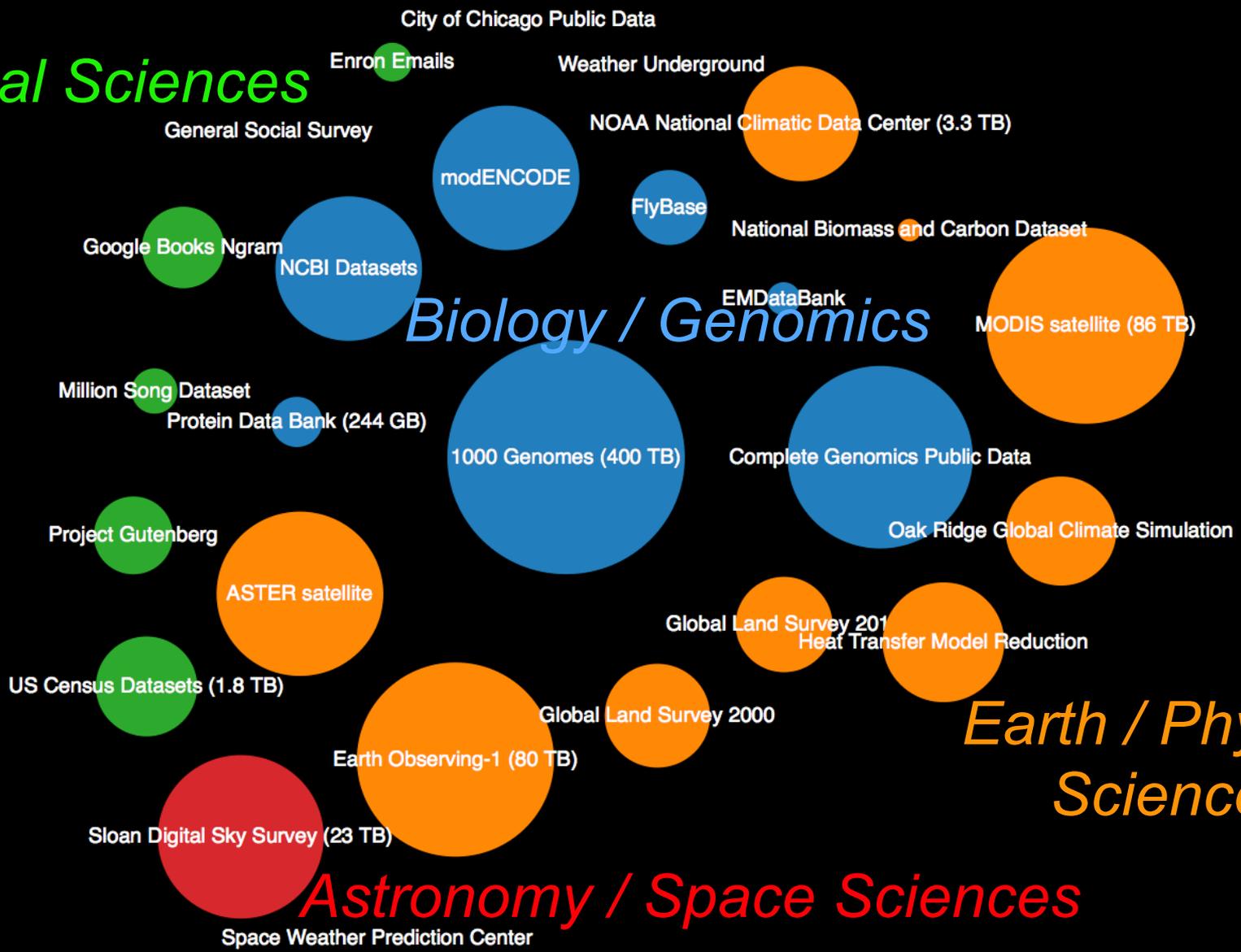


*Social Sciences*

*Biology / Genomics*

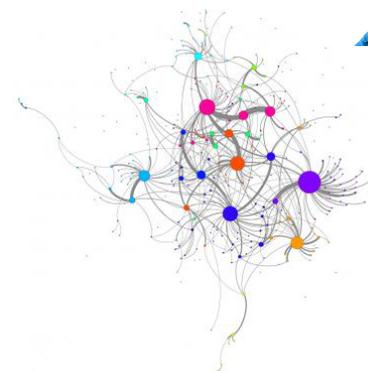
*Earth / Physical Sciences*

*Astronomy / Space Sciences*



# The OSDC is an interdisciplinary hub.

- Over 700+ total unique user accounts
- Users from 140+ different institutions
- Each month, on average
  - About 200 unique users
  - 1.8 million core hours
  - 800 TB user data stored



## • Researchers in

ability active algorithms allows analysis  
 analyzer annotation application automatic  
 compare comparison contain cover custom data  
 databases designer different documents  
 explorer gene genome given high-value  
 information integration  
 knowledge life mining nominator online organization  
 perform pipelines predictor products  
 provide public recognition repository  
 scientific sequence serve suite system  
 text tool trinity update user

- Biology
- Medicine
- Computer Science
- Mathematics
- Earth Science
- Social Science
- Urban Science
- Digitized Humanities

Protected data cloud (PDC) for analyzing human genomic data

- Collaboration with Institute for Genomics and Systems Biology (IGSB) at UChicago
- Allows users authorized by NIH to compute over human genomic data from dbGaP in a secure and compliant fashion
- Contains data from The Cancer Genome Atlas (TCGA)

**[bionimbus-pdc.opensciencedatacloud.org](https://bionimbus-pdc.opensciencedatacloud.org)**

## ***Bionimbus-PDC***

 **PDC**

BIONIMBUS PROTECTED DATA CLOUD



---

Institute for  
Genomics &  
Systems Biology

---

Web-based tool hosted by OSDC for visualizing trends in repositories of digitized texts

# Bookworm

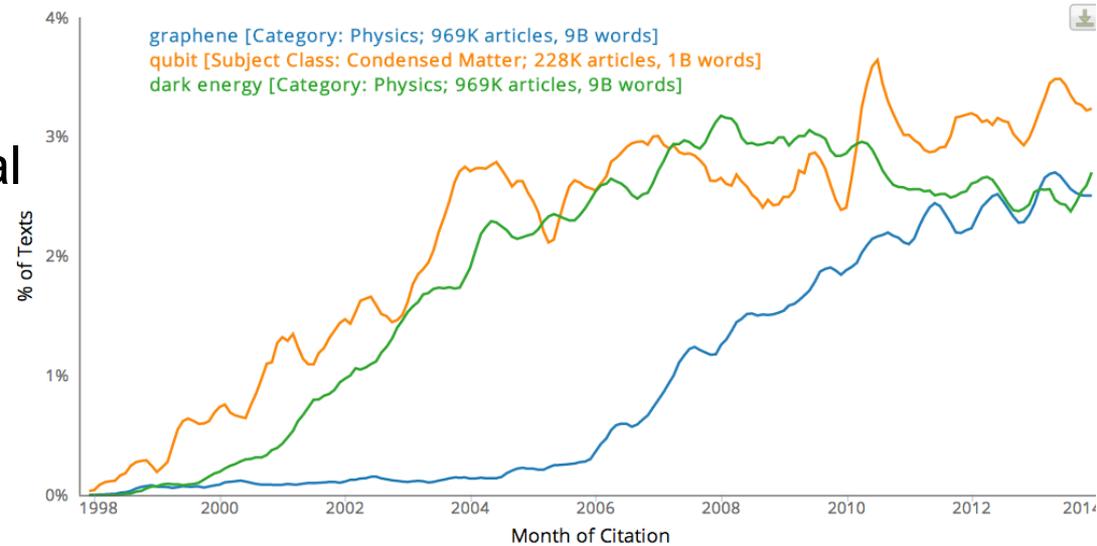
- Open Library books
- arXiv science publications
- Chronicling America historical newspapers
- US Congress bills
- Social Science Research Network paper abstracts
- Create your own bookworm

bookworm: [arXiv](#)

Search for trends in hundreds of thousands of articles at [arxiv.org](#)



in Category: Physics - +  
 in Subject Class: Condensed Matter - +  
 in Category: Physics - + Search





## ***Project Matsu***

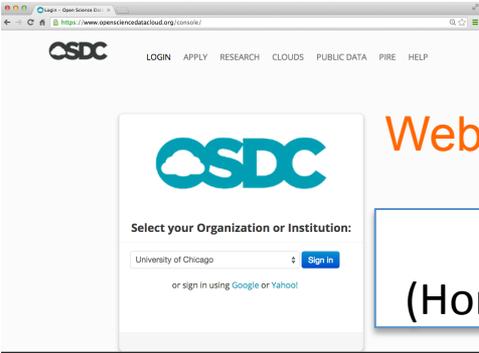


Collaboration with NASA to develop open source technology for cloud-based processing of satellite imagery to support earth sciences.

The OSDC is used to process Earth Observing 1 (EO-1) satellite imagery from the Advanced Land Imager (ALI) and the Hyperion instruments and to make this data available to interested users.

- Namibia flood dashboard
- Hadoop-based Matsu “Wheel” system for processing all data  
**[matsu.opensciencedatacloud.org](http://matsu.opensciencedatacloud.org)**

# OSDC Infrastructure



Web Browser

Users

Console  
(Horizon-based)

SSH

OSDC Cloud

Login Node

SSH

Cloud Controller  
(Nova, Glance and  
Keystone)

Manage VMs  
(Nova)

Compute Node

VM

VM

VM

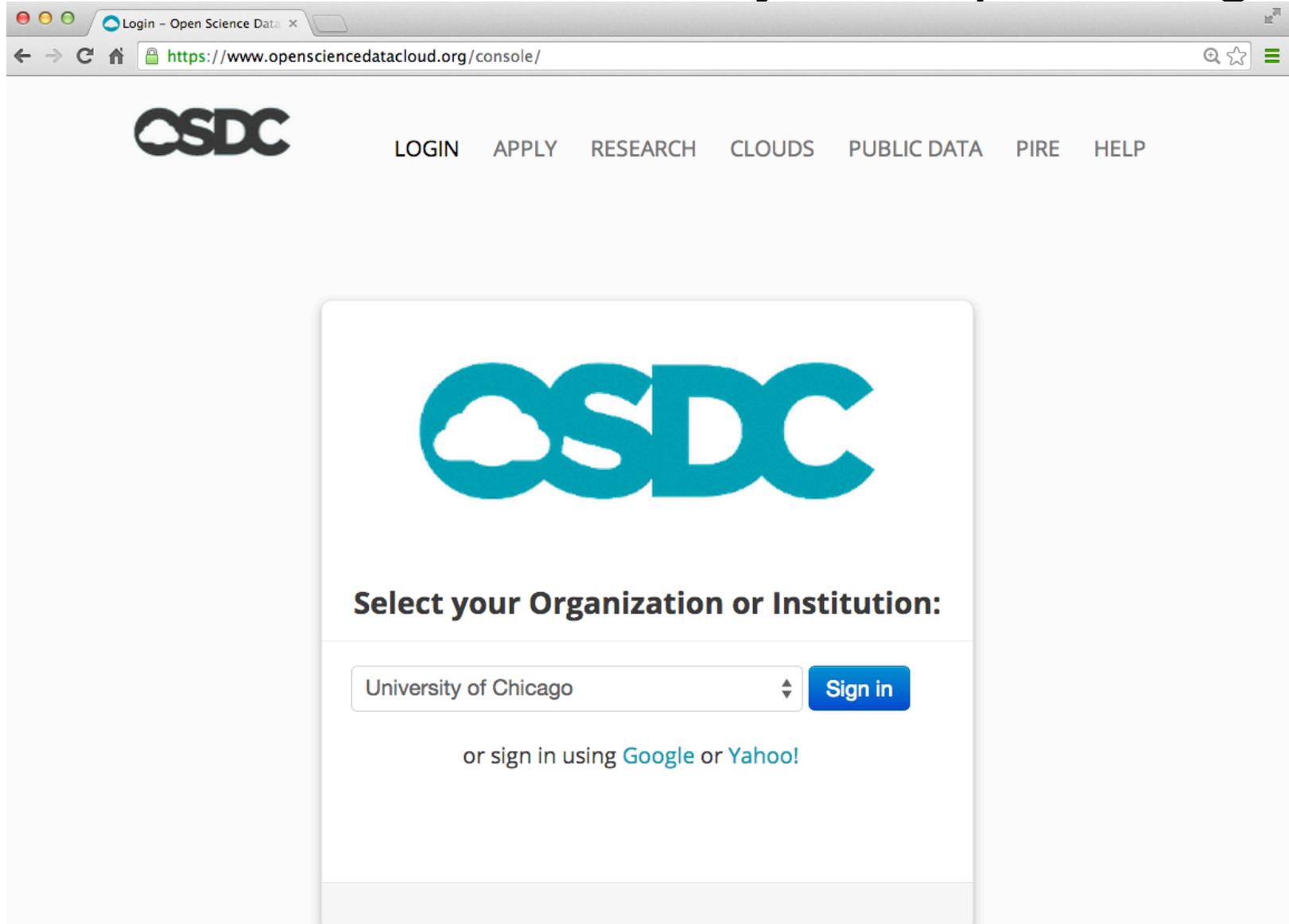
Compute Node

Compute Node

Authenticated  
GlusterFS mount



# User view of OSDC “Tukey” web portal log-in



[www.opensciencedatacloud.org/console/](https://www.opensciencedatacloud.org/console/)

# Users log-in to web portal with University credentials

The University of Chicago

https://shibboleth2.uchicago.edu/idp/Authn/MCB

 THE UNIVERSITY OF CHICAGO

 OPEN CLOUD CONSORTIUM

## Sign In

You are logging in to: **Open Science Data Cloud Console**  
We provide and support cloud computing and storage services for the scientific research community. The OSDC is run by the Open Cloud Consortium, a non-profit organization whose primary goal is to support scientific advances by working with researchers in a variety of disciplines.

**CNetID:**  [Hospital Employee?](#)

**Password:**  [Forgot your password?](#)

Signing in allows you to access multiple University of Chicago web applications while entering your CNetID and password only once. To end your session, simply close your browser.

**Questions?** Contact the IT Services Service Desk by phone at 2-5800 (773-702-5800), via email at [itservices@uchicago.edu](mailto:itservices@uchicago.edu), or get walk-in help at the TECHB@R on the first floor of Regenstein Library during reference desk hours <http://hours.lib.uchicago.edu/>.

**Alumni** account holders may contact [alumni-support@uchicago.edu](mailto:alumni-support@uchicago.edu) or call 1-877-292-3945 between 9 AM and 3 PM CST with any questions.

Authentication powered by Shibboleth™ 

# User view of OSDC "Tukey" web portal



LOGIN APPLY RESEARCH CLOUDS PUBLIC DATA PIRE HELP

Overview

Instances

Images & Snapshots

Access & Security

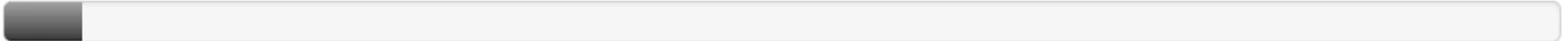
Logged in as: mpatterson

Settings

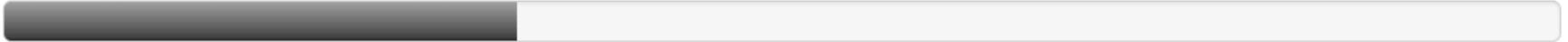
Sign Out

## Quota Summary

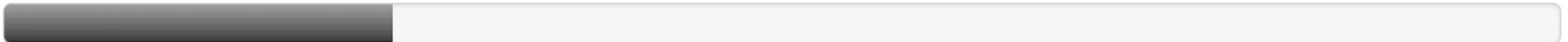
Used 2 of 42 Available Instances



Used 16 of 48 Available vCPUs



Used 32,768 MB of 132,072 MB Available RAM



Select a month to query its usage:

November 2014 Submit

This Month's Cloud Core Hours: - This Month's Cloud Disk Usage (GB): -

This Month's Hadoop Disk Usage (GB): - This Month's Hadoop Job Hours: -

Download CSV Summary

# User view of OSDC “Tukey” web portal



LOGIN APPLY RESEARCH CLOUDS PUBLIC DATA PIRE HELP

Overview **Instances** Images & Snapshots Access & Security Logged in as: mpatterson [Settings](#) [Sign Out](#)

## Instances

Launch Instance

Terminate Instances

<input type="checkbox"/>	Instance Name	IP Address	Size	Keypair	Status	Task	Power State	Cloud	Actions
<input type="checkbox"/>	<a href="#">GeoServer_20141112</a>	172.16.1.81	m1.xlarge   16GB RAM   8 VCPU   20GB Disk	mpatterson	Active	None	Running	Sullivan	<a href="#">Create Snapshot</a> ▼
<input type="checkbox"/>	<a href="#">atwood_lme4_mcmc8</a>	172.16.1.20	m1.xlarge   16GB RAM   8 VCPU   20GB Disk		Active	None	Running	Atwood	<a href="#">Create Snapshot</a> ▼

Displaying 2 items

# User view - launching a VM

## Images

Filter

<input type="checkbox"/>	Image Name	Type	Status	Public	Format	Cloud	Actions
<input type="checkbox"/>	new-cube-20141030	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	Ubuntu-14.04-LTS-v1.0	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	CentOS-5.10-v1.0	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	CentOS-6.0-v1.0	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.3	Image	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	Ubuntu-14.04-LTS-v1.0	Image	Active	Yes	QCOW2	Atwood	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.6	Image	Active	Yes	QCOW2	Atwood	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.5_ia32-libs	Image	Active	Yes	QCOW2	Atwood	Launch
<input type="checkbox"/>	Ubuntu-12.04-LTS-v1.5	Image	Active	Yes	QCOW2	Atwood	Launch

Displaying 9 items

## User Snapshots

Filter

<input type="checkbox"/>	Image Name	Type	Status	Public	Format	Cloud	Actions
<input type="checkbox"/>	pireweb2	Snapshot	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	pireweb	Snapshot	Active	No	QCOW2	Sullivan	Launch
<input type="checkbox"/>	OSDC_PIREtools	Snapshot	Active	Yes	QCOW2	Sullivan	Launch
<input type="checkbox"/>	OSDC_ToolExplorer_UDR	Snapshot	Active	Yes	QCOW2	Sullivan	Launch

### Launch Instance

Details Access & Security Post-Creation

**Instance Source**  
Image Specify the details for launching an instance.  
The chart below shows the resources used by this project in relation to the project's quotas.

**Image**  
Ubuntu-14.04-LTS-v1.0

**Instance Name**  
m1.tiny  
m1.small  
m1.medium  
m1.large  
✓ m1.xlarge  
m1.xxlarge  
m2.xxlarge  
m2.xlarge

**Flavor Details**

Name	m1.xlarge
VCPUs	8
Root Disk	20 GB
Ephemeral Disk	0 GB
Total Disk	20 GB
RAM	16,384 MB

**Project Quotas**

**Number of Instances (1)** 9 Available

**Number of VCPUs (8)** 8 Available

**Total RAM (16,384 MB)** 50,152 MB Available

# User view – log on to VM, install tools, ready for analysis

Instances – Open Science <https://www.opensciencedatacloud.org/project/instances/>

Overview **Instances** Images & Snapshots Access & Security Logged in as: mpatterson Settings Sign Out

## Instances

Launch Instance Terminate Instances

Instance Name	IP Address	Size	Keypair	Status	Task	Power State	Cloud	Actions
<input type="checkbox"/> GeoServer_20141112	172.16.1.81	m1.xlarge   16GB RAM   8	mpatterson	Active	None	Running	Sullivan	<input type="button" value="Create Snapshot"/>

atwood\_lme4\_mcmc8

Displaying 2 items

The screenshot shows an R Graphics terminal window titled "R Graphics: Device 2 (ACTIVE)". It displays four network graphs arranged in a 2x2 grid. The top-left graph is labeled "Triangulation" and shows a dense network of black lines forming a complex shape. The top-right graph is labeled "SOI graph" and shows a similar network but with a different internal structure. The bottom-left graph is labeled "Gabriel" and shows a network with a different internal structure. The bottom-right graph is labeled "Relative" and shows a network with a different internal structure. The terminal window also shows R code for plotting these graphs and a terminal window showing system usage and commands.

```
mpatterson@i-000001ed: ~ (ssh)
Usage of /: 13.5% of 19.69GB Users logged in: 1
Memory usage: 4% IP address for eth0: 172.16.1.1
Swap usage: 0%
Graph this data and manage this system at https://landscape.canonical.com/
Get cloud support with Ubuntu Advantage Cloud Guest
http://www.ubuntu.com/business/services/cloud
mpatterson@i-000001ed:~$ ssh -AX 172.16.1.1
mpatterson@i-000001ed:~/medical/geo/ASDAR/scripts_data
> plot(Syracuse, border="grey60")
> plot(Sy5_nb, coords, add=TRUE, pch=".")
> text(bbox(Syracuse)[1,1], bbox(Syracuse)[2,2], labels="SOI graph", cex=0.7)
> plot(Syracuse, border="grey60")
> plot(Sy6_nb, coords, add=TRUE, pch=".")
> text(bbox(Syracuse)[1,1], bbox(Syracuse)[2,2], labels="Gabriel", cex=0.7)
> plot(Syracuse, border="grey60")
> plot(Sy7_nb, coords, add=TRUE, pch=".")
> text(bbox(Syracuse)[1,1], bbox(Syracuse)[2,2], labels="Relative", cex=0.7)
> par(oopar)
>
mpatterson@i-000001ed: ~
Swap usage: 0%
Graph this data and manage this system at https://landscape.canonical.com/
Get cloud support with Ubuntu Advantage Cloud Guest
http://www.ubuntu.com/business/services/cloud
mpatterson@i-000001ed:~$ sudo apt-get upgrade
```

# Tips

- Public datasets are available and automatically mounted to your virtual machines at `/glusterfs/osdc_public_data/`
- Your home directory is in `/glusterfs/users/username`
  - Anything in there is accessible from all virtual machines
  - Anything you store anywhere else will go away when your VM is terminated
- Create a keypair under “Access and Security” and choose “All resources”.
- Remember to make two ‘hops’ and carry your key on both hops to get to your virtual machine
  - `ssh-add yourkeypair.pem`
  - `ssh -A username@sullivan.opensciencedatacloud.org`
  - `ssh -A ubuntu@ipaddress`
- Check the status page: [www.opensciencedatacloud.org/status/](http://www.opensciencedatacloud.org/status/)
- Read the docs: [www.opensciencedatacloud.org/support/](http://www.opensciencedatacloud.org/support/)

# ID and Metadata Services for Data Commons

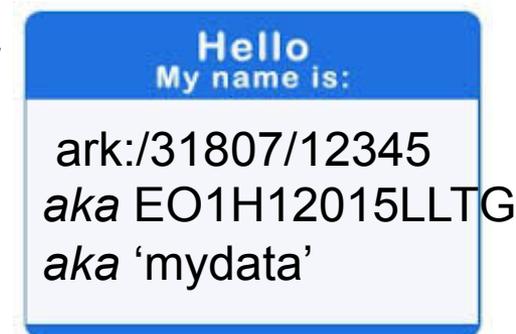
# Challenges

Significant concerns for the scientific and research community in storing data long-term in an accessible and usable manner

- Preserving data provenance
  - *What are these data, how were they produced?*
- Maintaining scientific reproducibility and workflows
  - *Can I access these data the same way I used to?*

Challenges:

- Data moves, location changes.
  - *Hardware dies/changes, no home for data.*
- Identifiers
  - *Variety of identifiers- DOI, ARK, UUID, etc*



# ID and metadata services

## ID services goal:

- Flexibility to support data in multiple locations/ access points
- Flexibility to support multiple identifiers

## Metadata services goal:

- For all data (cross-discipline), support simple core metadata
- Support or allow discipline-specific metadata for search capabilities as needed per field

# ID service user demo with EO-1 data: *Signpost*

