OPEN SCIENCE DATA CLOUD

Using Open Science Data Cloud

OPEN CLOUD CONSORTIUM

cdis CENTER FOR DATA INTENSIVE SCIENCE

THE UNIVERSITY OF CHICAGO

FIU FLORIDA INTERNATIONAL UNIVERSITY

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

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

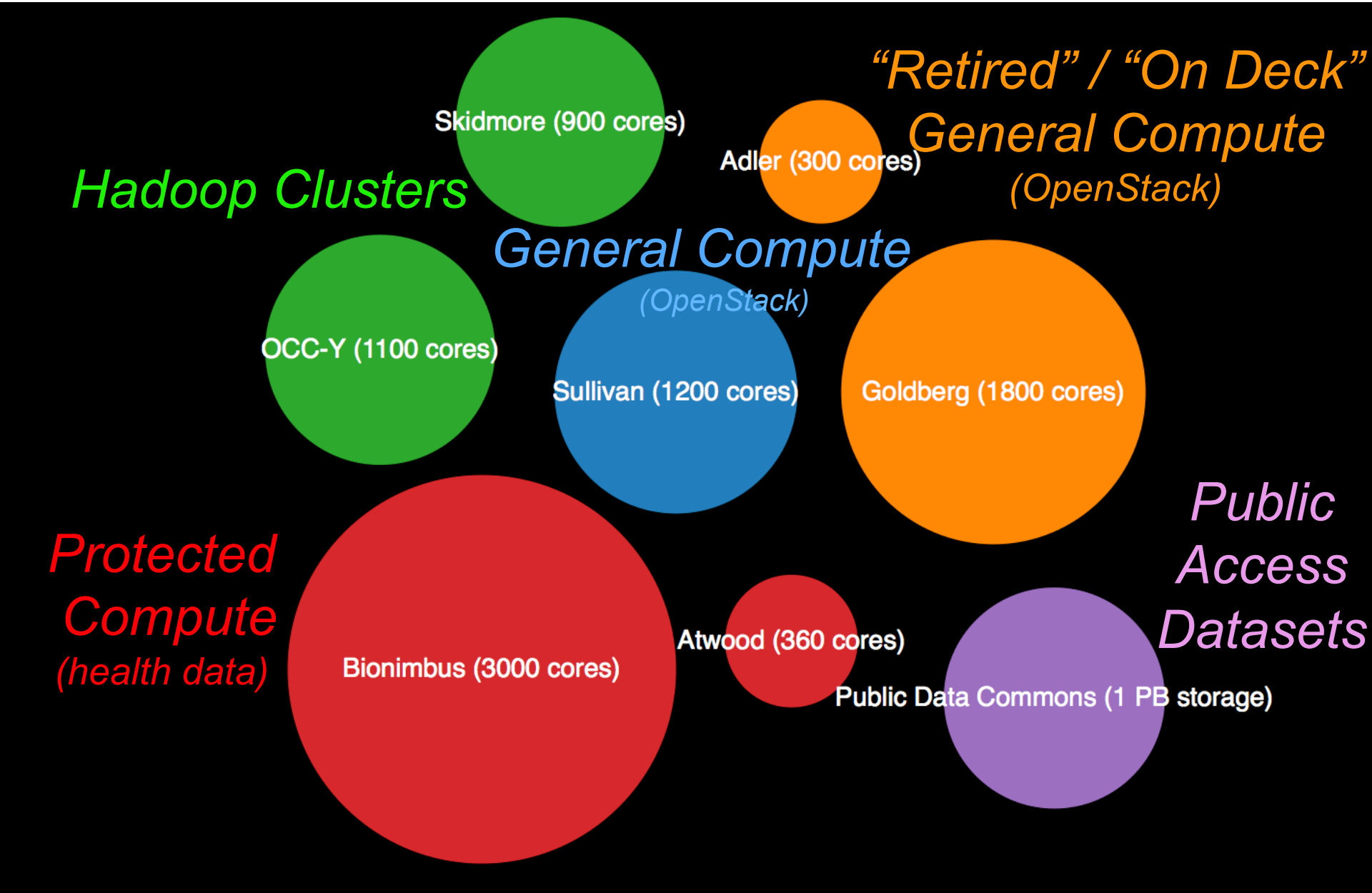
- About 9000 compute cores
- Both OpenStack and Hadoop-based clouds

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

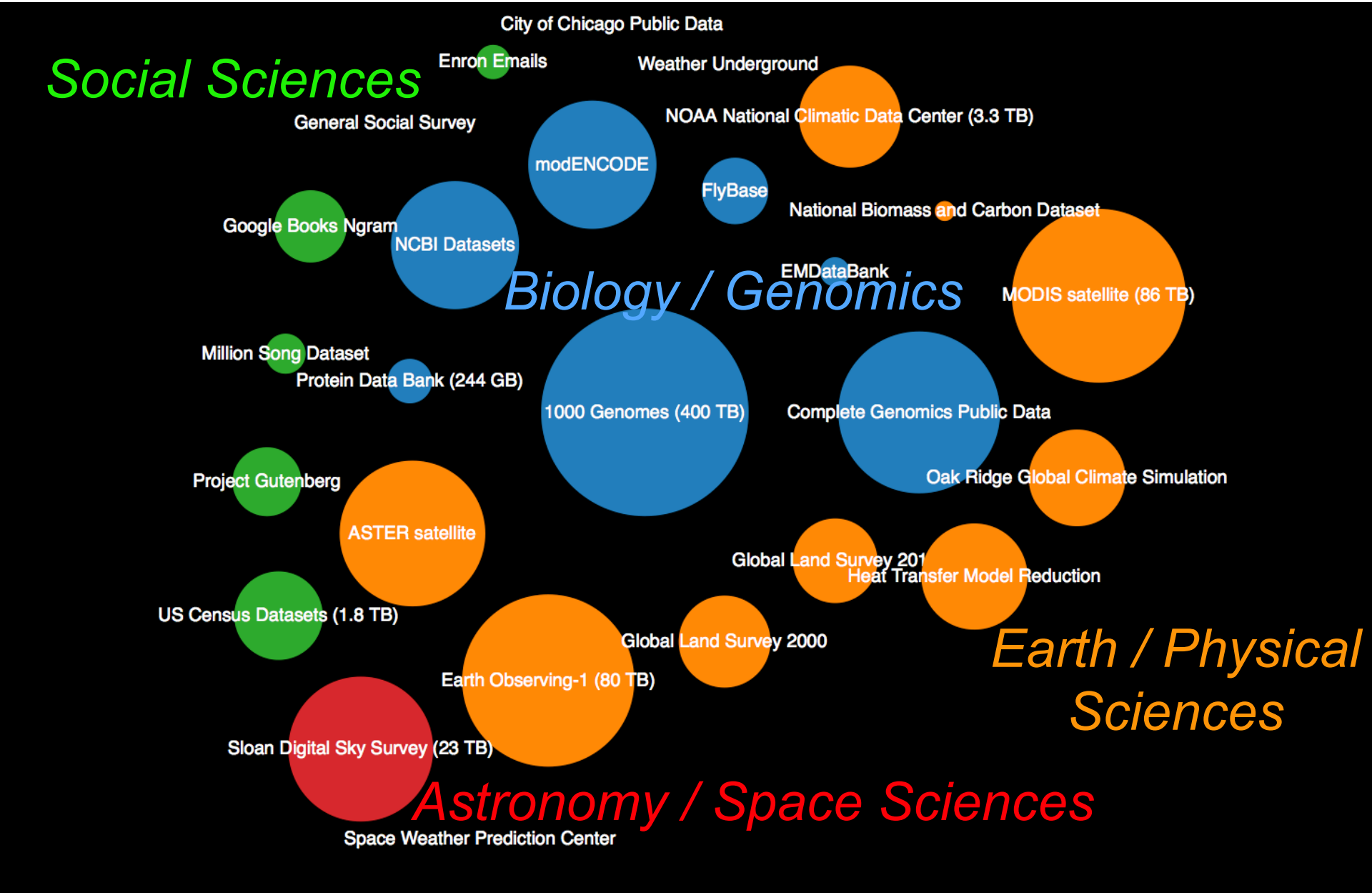
OPEN SCIENCE DATA CLOUD

Cloud Resources by cluster size / type



OPEN SCIENCE DATA CLOUD

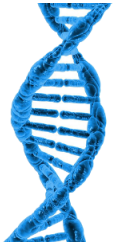
Public Data Commons by data size / type



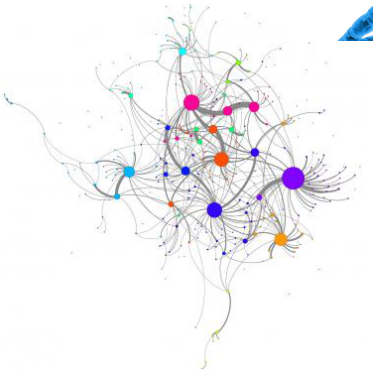
OPEN SCIENCE DATA CLOUD

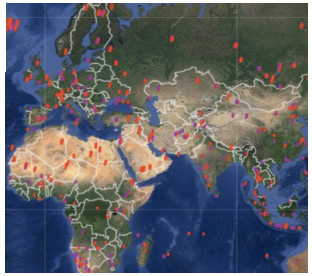
Community user stats

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

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

Protected data cloud (PDC) for analyzing human genomic data

Bionimbus-PDC

PDC

BIONIMBUS PROTECTED DATA CLOUD

- 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)

Institute for Genomics & Systems Biology

bionimbus-pdc.opensciencedatacloud.org

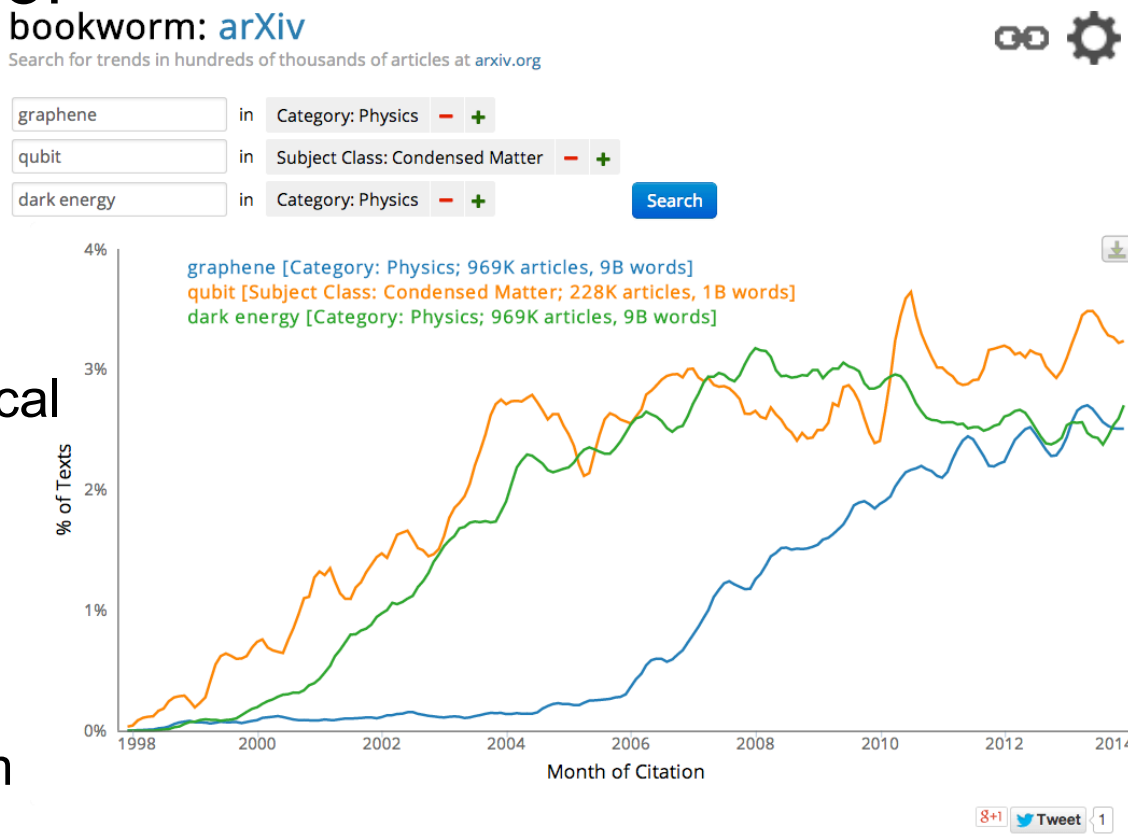
OPEN SCIENCE DATA CLOUD

Featured Projects: Text Data

Bookworm

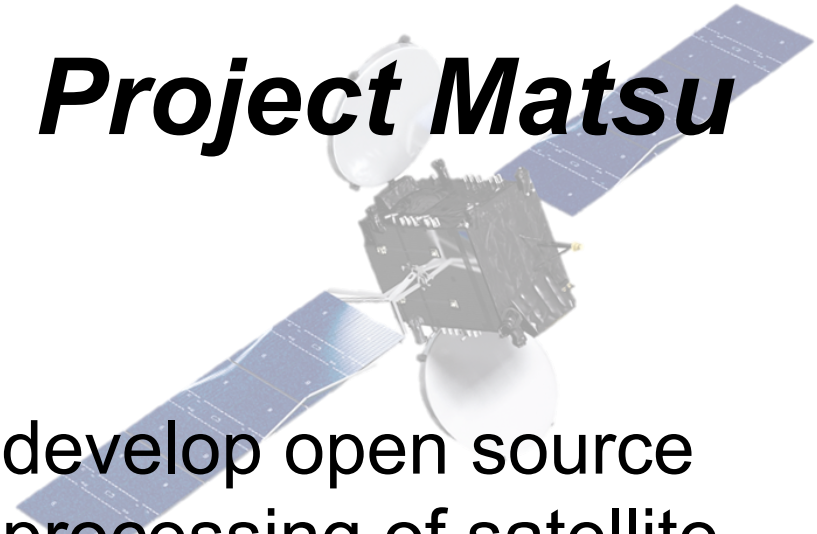
Web-based tool hosted by OSDC for visualizing trends in repositories of digitized texts

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



bookworm.culturomics.org

Project Matsu



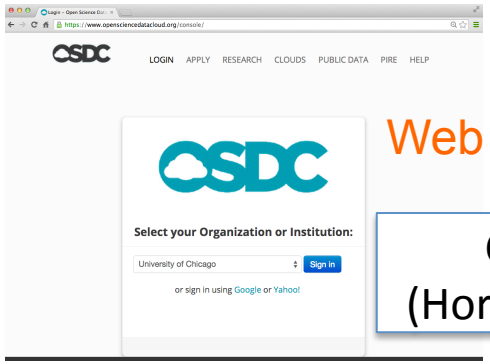
Collaboration with NASA to develop open source technology for cloud-based processing of satellite imagery to support earth sciences.

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

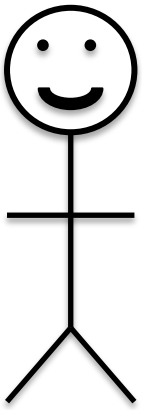
matsu.opensciencedatacloud.org

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

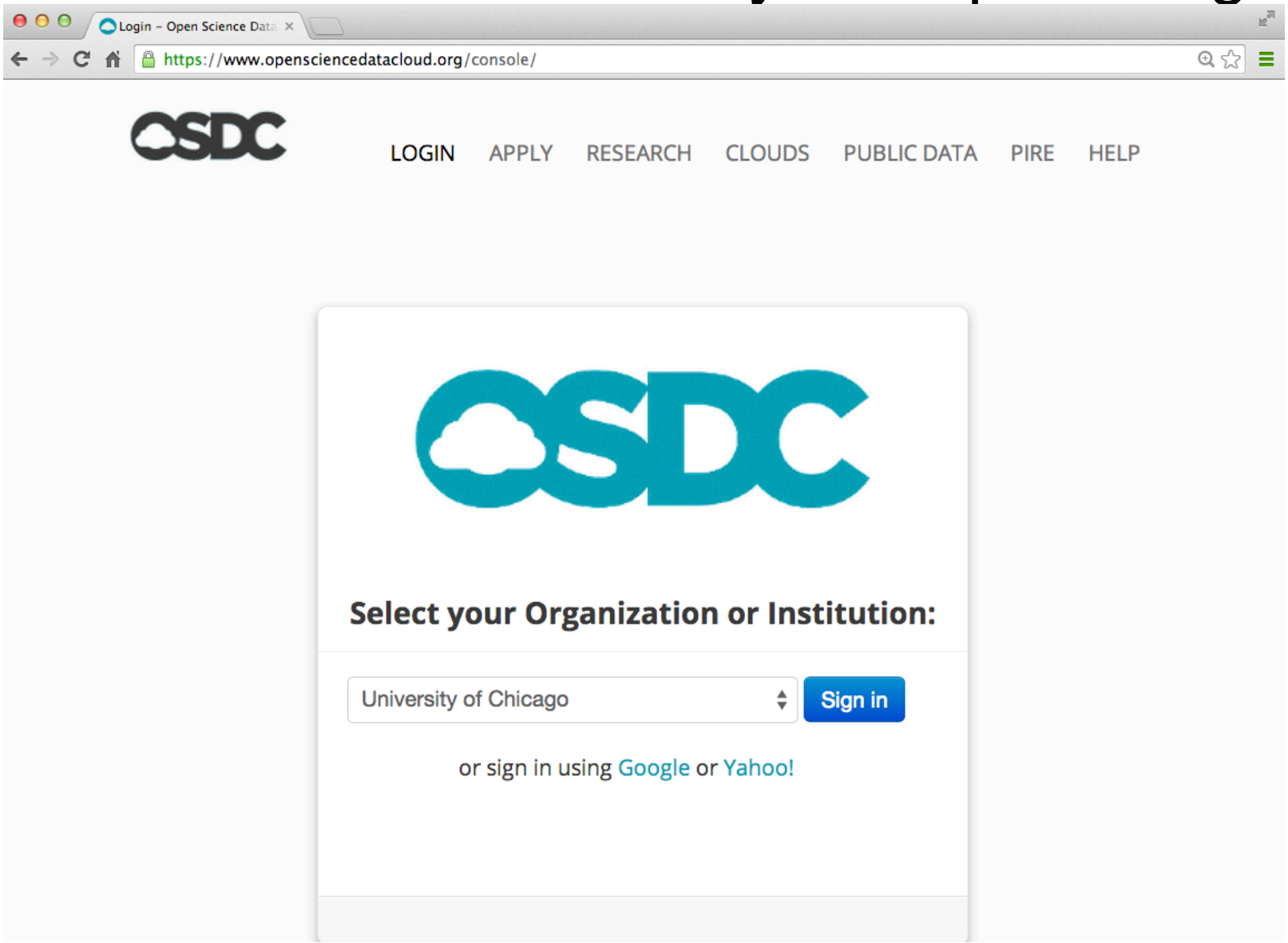
Compute Node

Compute Node

VM VM VM

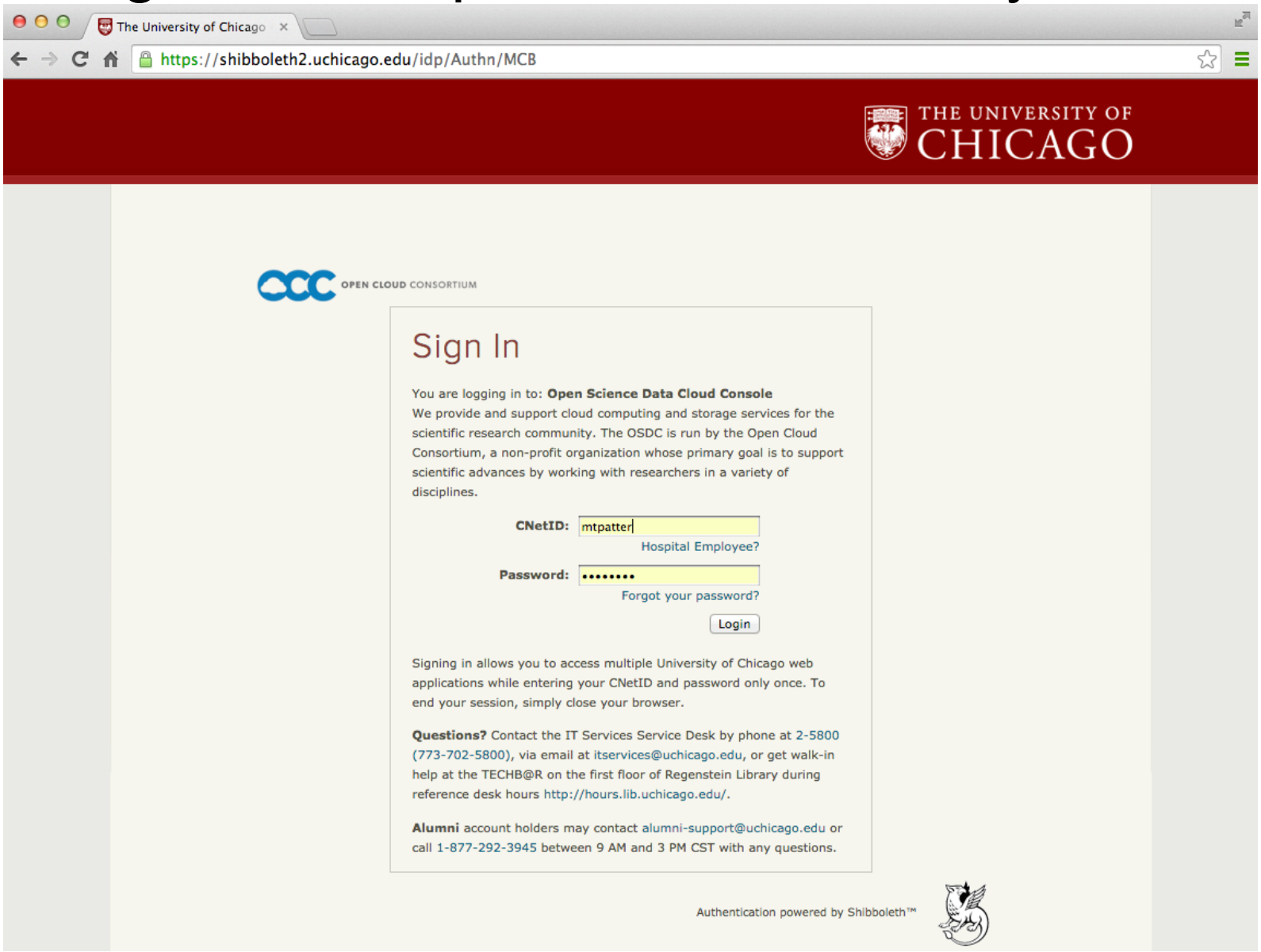
Authenticated GlusterFS mount

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

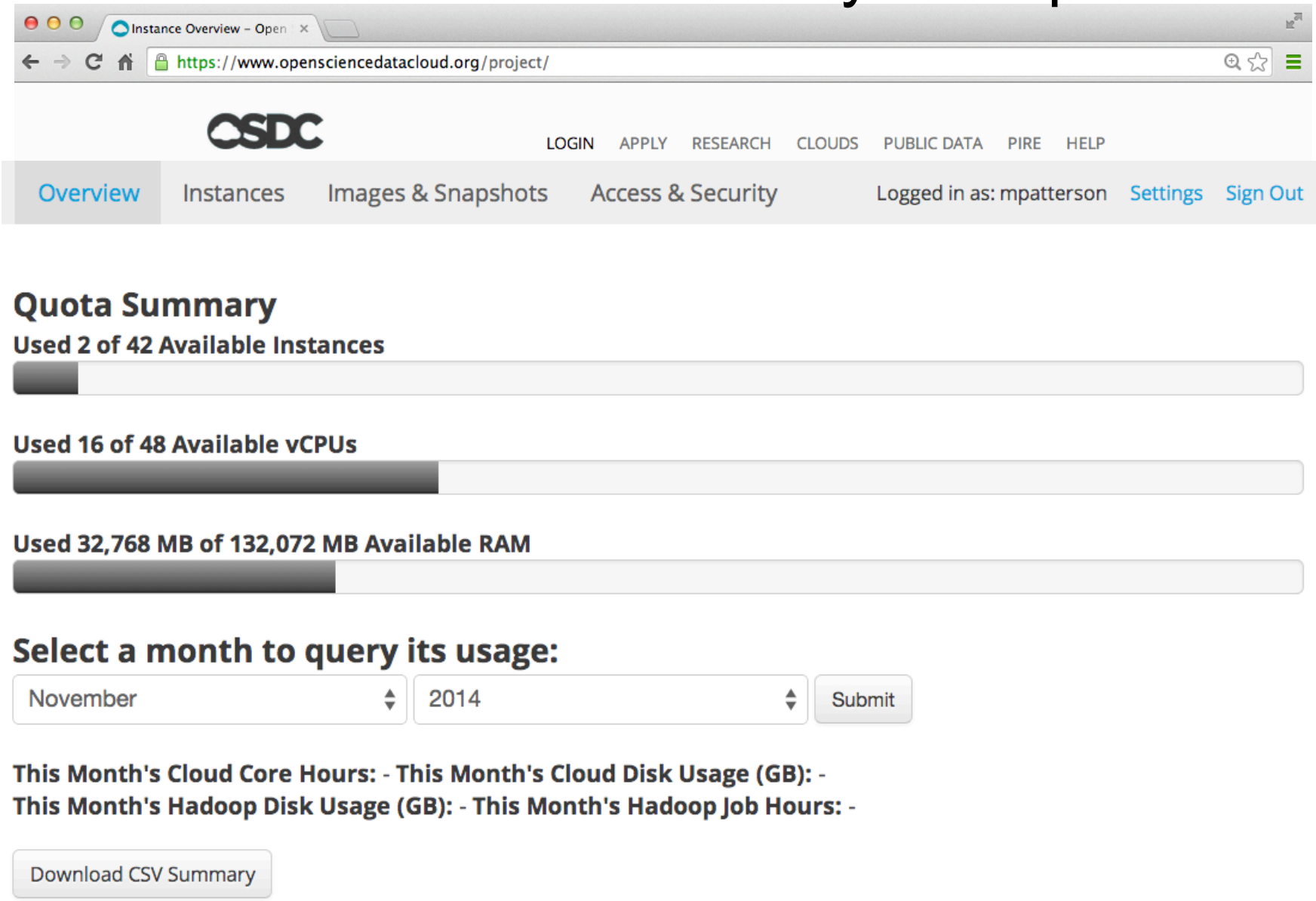


www.opensciencedatacloud.org/console/

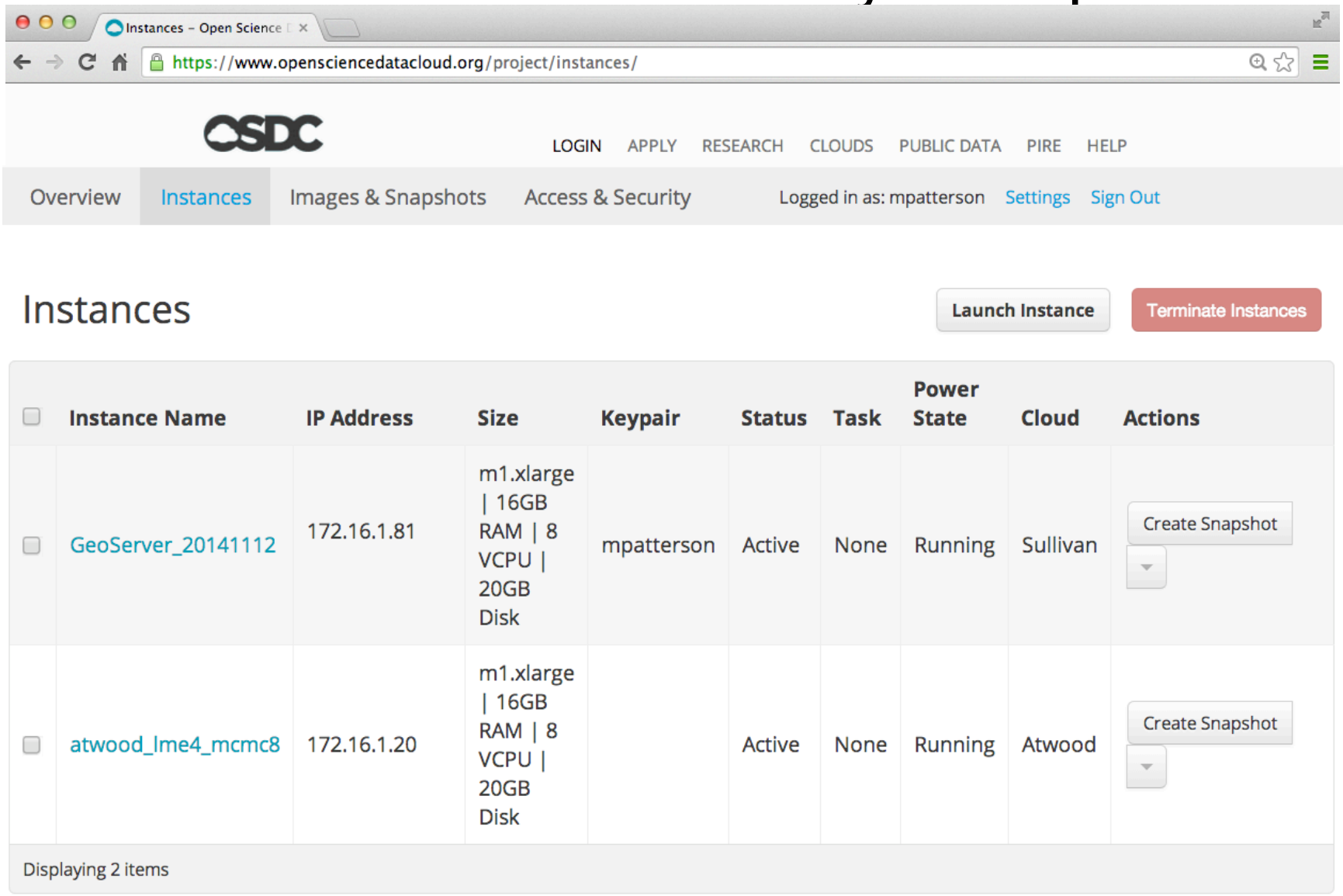
Users log-in to web portal with University credentials



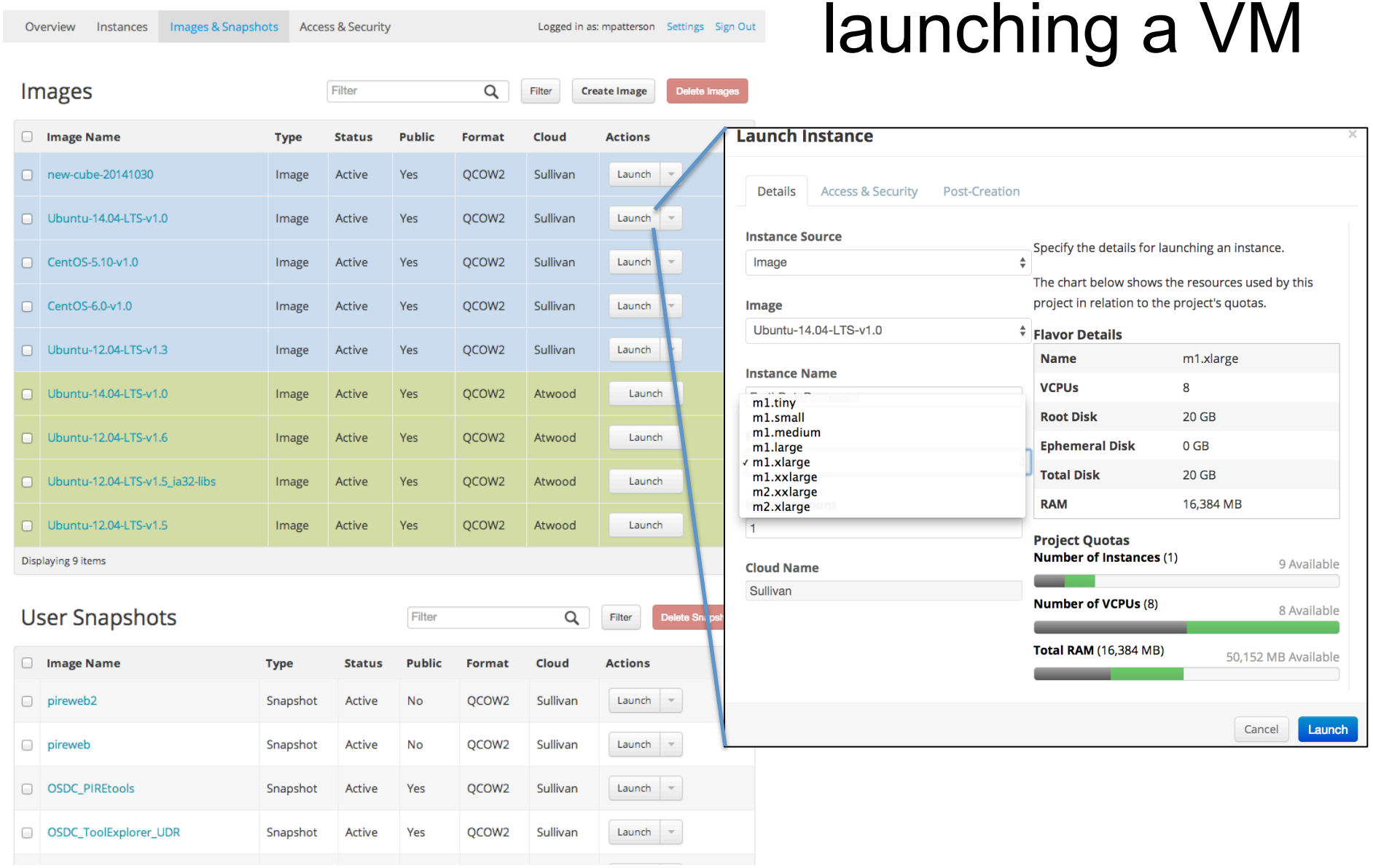
User view of OSDC “Tukey” web portal



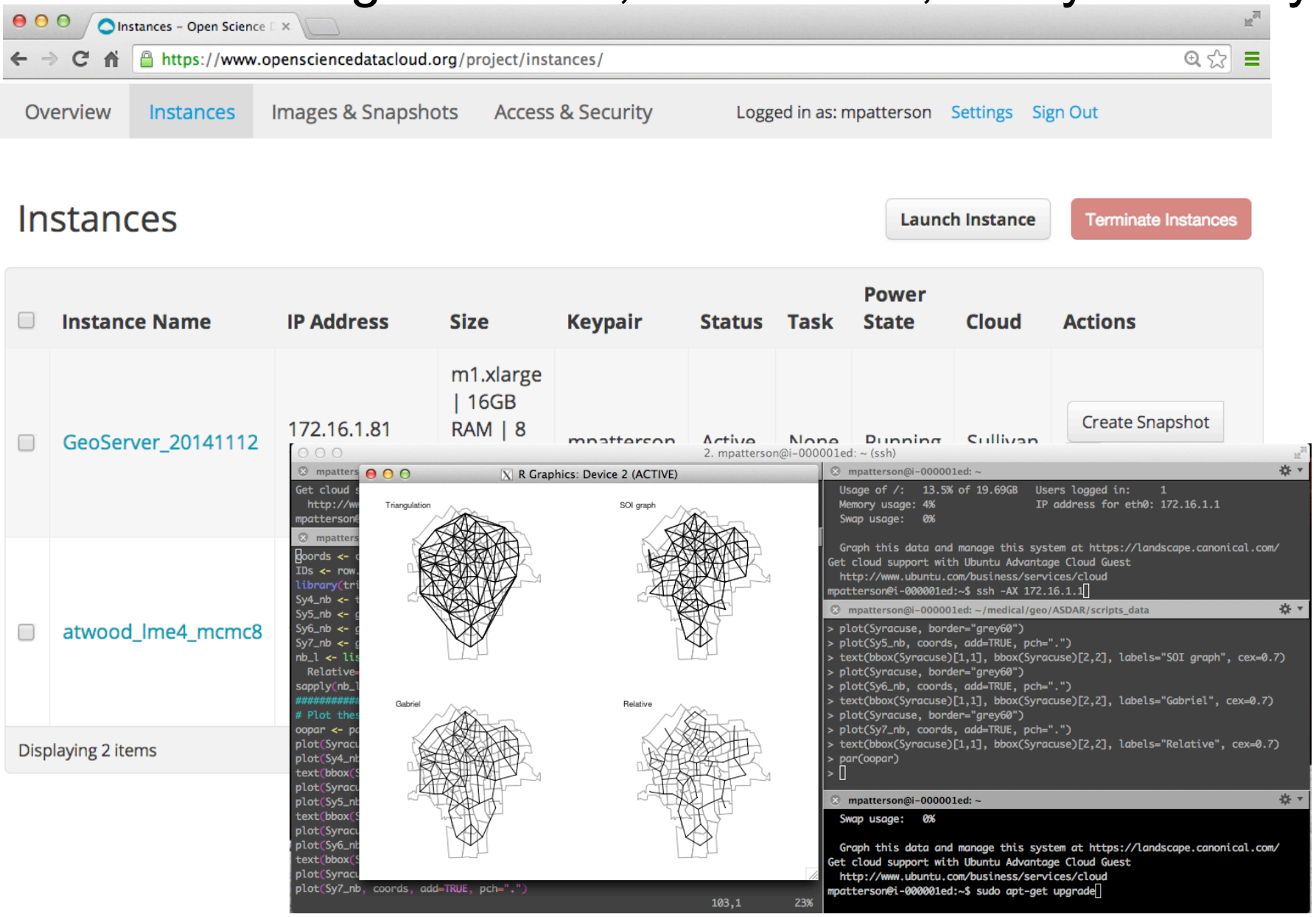
User view of OSDC “Tukey” web portal



User view - launching a VM



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



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/
- Read the docs: www.opensciencedatacloud.org/support/

OPEN SCIENCE DATA CLOUD

ID and Metadata Services for Data Commons

THE UNIVERSITY OF CHICAGO

CENTER FOR DATA INTENSIVE SCIENCE

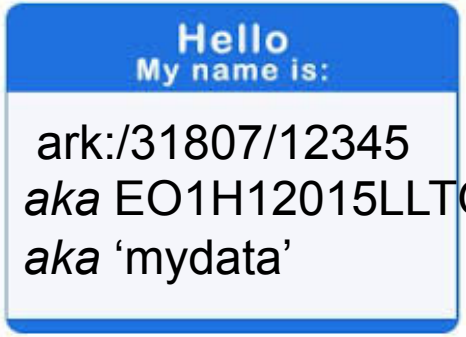
Challenges

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:

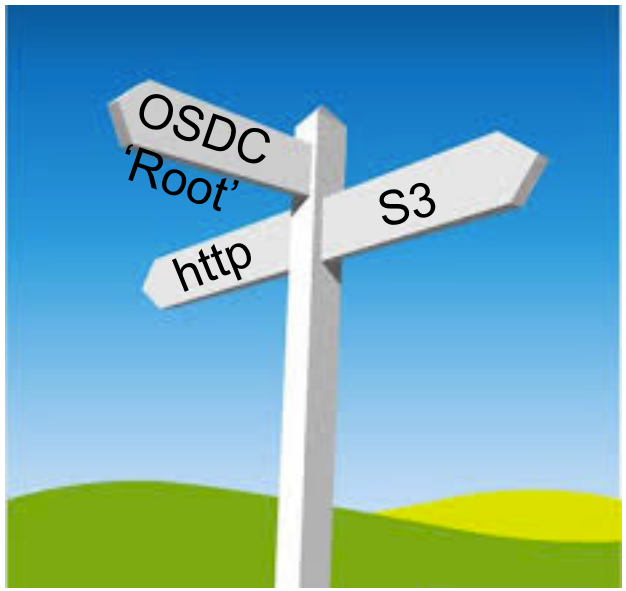
- Flexibility to support data in multiple locations/ access points
- Flexibility to support multiple identifiers

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

OSDC
OPEN SCIENCE DATA CLOUD

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



THE UNIVERSITY OF CHICAGO

cdis CENTER FOR DATA INTENSIVE SCIENCE