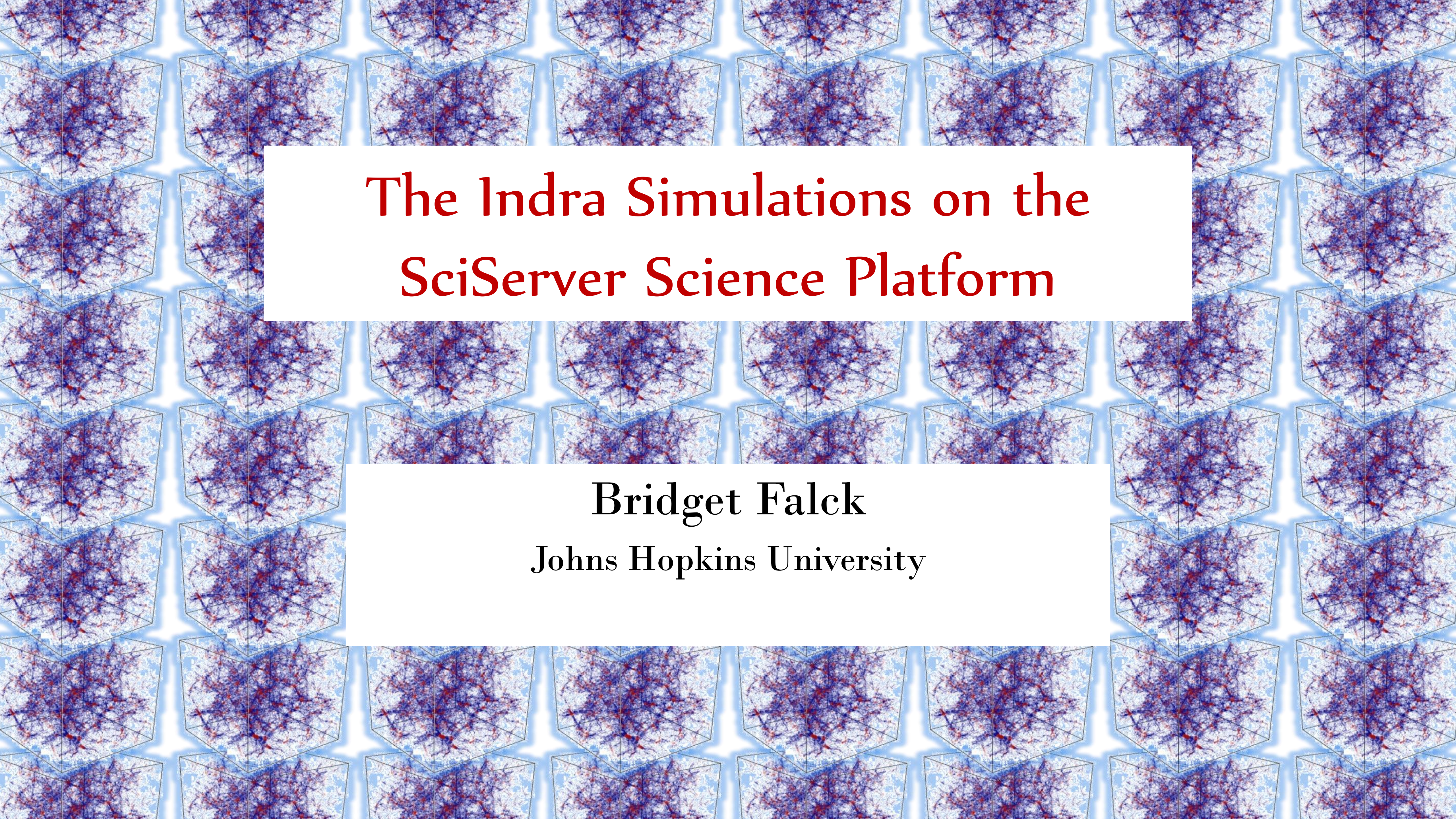


The Indra Simulations on the SciServer Science Platform

Bridget Falck

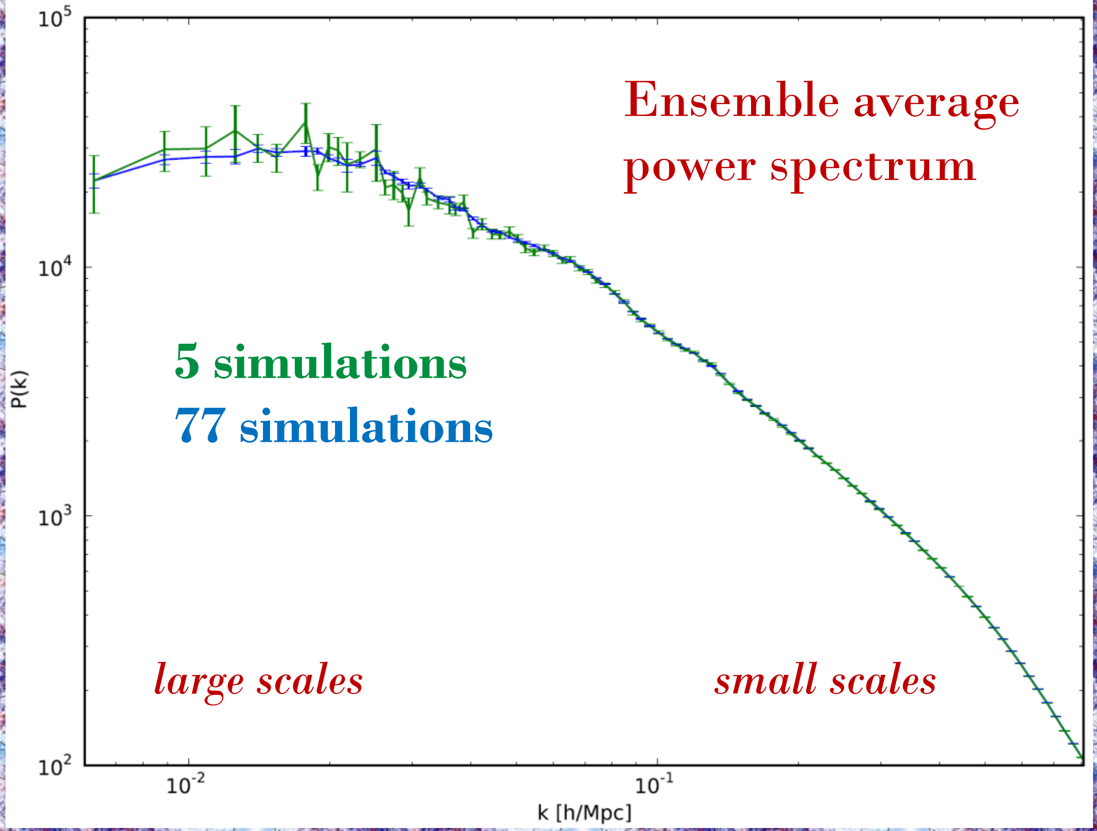
Johns Hopkins University

Motivation: Cosmic Variance

- Theoretical predictions of large-scale structure require numerical simulations, but we can't simulate our observable Universe exactly, only its statistical properties

- Both simulations and observations have good statistics on small scales but poor statistics on large scales – the cosmic variance limit

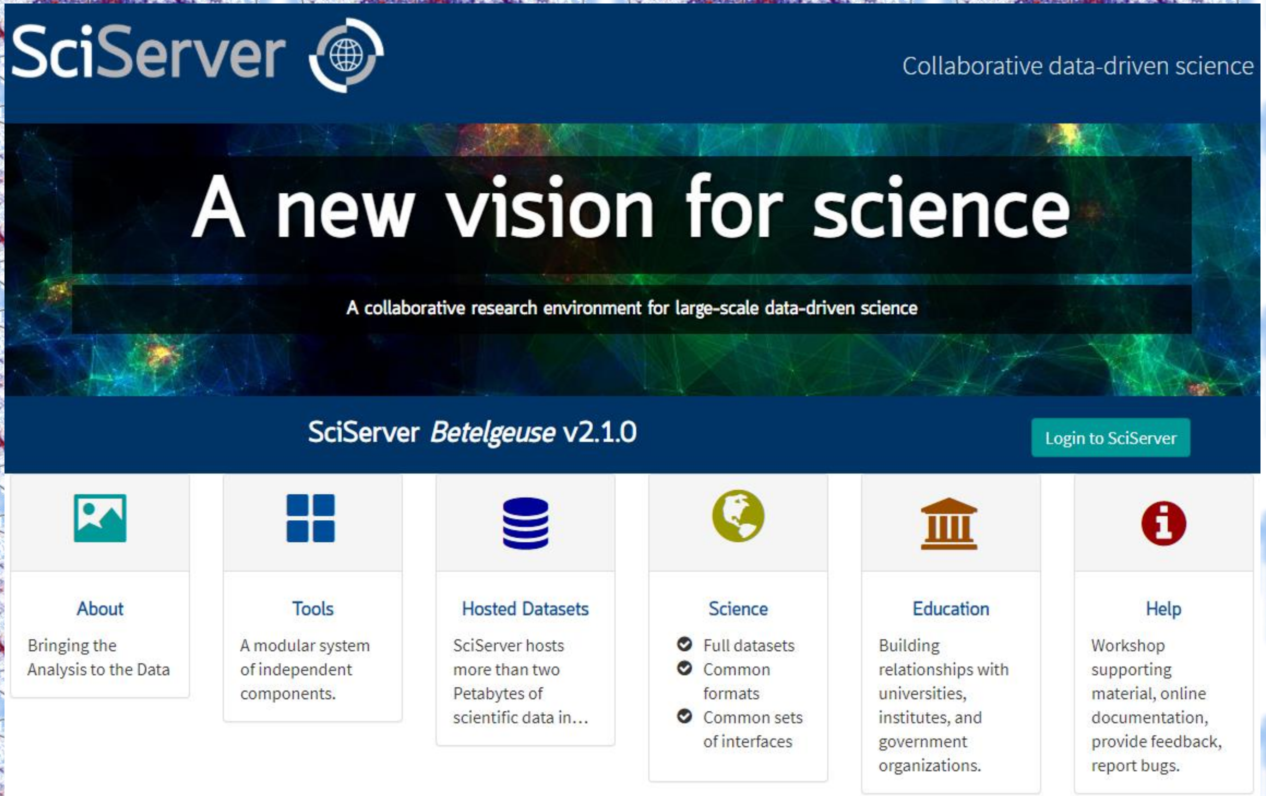
- Need to run many simulations with different initial conditions but *same model and parameters*



The Indra Simulations

- Suite of 384 simulations with the same cosmology
  - Each a 1 Gpc/h-sided box with 1024^3 dark matter particles, run with L-Gadget code
  - Output: 64 snapshots of particle positions and velocities, FOF/SUBFIND halo catalogs, and 505 time-steps of Fourier-space density grids
- 750 TB of data, available to the public and *computationally-accessible* via the SciServer
  - Ensemble averages and covariances, conditional and extreme statistics, mock galaxy catalogs and lightcones, etc.
  - Test-bed for new data architectures and analysis tools
- Other simulation suites beyond Indra are needed and being produced
  - Vary cosmological models and parameters, include hydrodynamical effects, etc.
  - Variety of scientific questions and codes means no standard outputs

SciServer Science Platform



SciServer

Collaborative data-driven science

A new vision for science

A collaborative research environment for large-scale data-driven science

SciServer *Betelgeuse* v2.1.0

Login to SciServer

About

Bringing the Analysis to the Data

Tools

A modular system of independent components.

Hosted Datasets

SciServer hosts more than two Petabytes of scientific data in…

Science

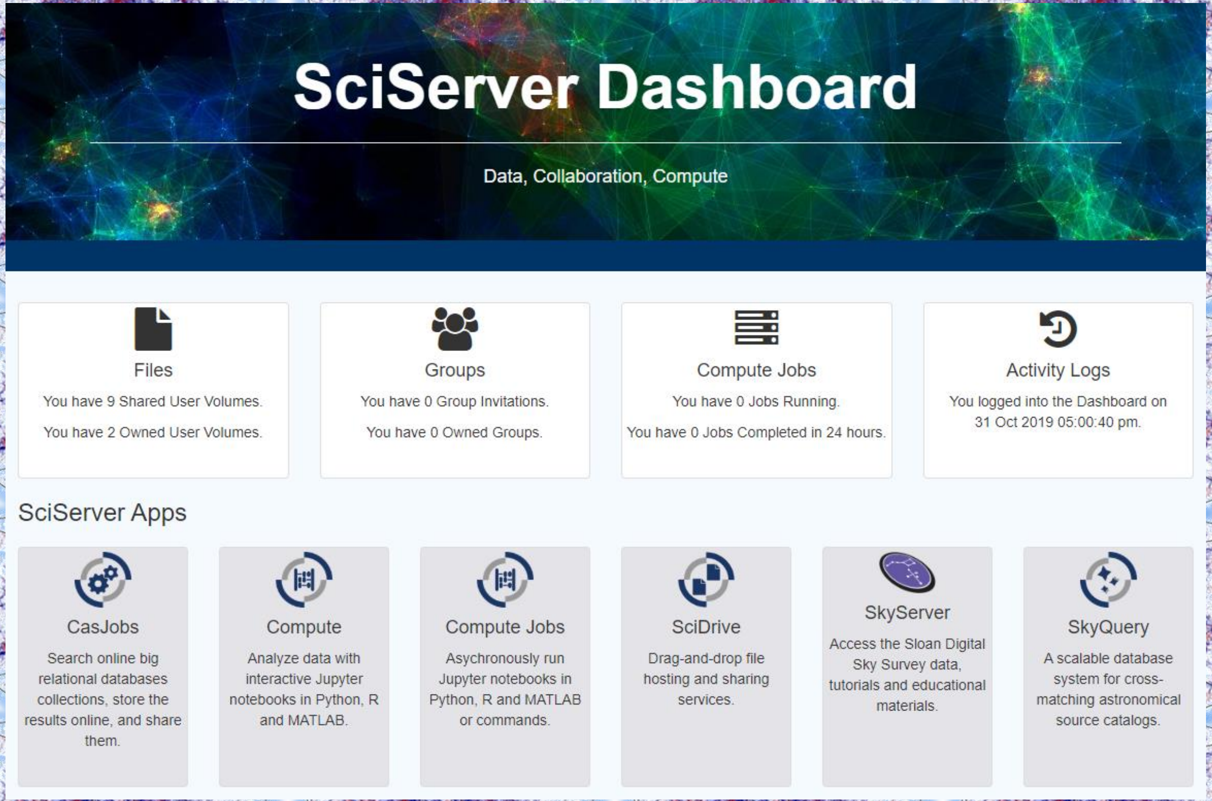
- ✓ Full datasets
- ✓ Common formats
- ✓ Common sets of interfaces

Education

Building relationships with universities, institutes, and government organizations.

Help

Workshop supporting material, online documentation, provide feedback, report bugs.



SciServer Dashboard

Data, Collaboration, Compute

Files

You have 9 Shared User Volumes.

You have 2 Owned User Volumes.

Groups

You have 0 Group Invitations.

You have 0 Owned Groups.

Compute Jobs

You have 0 Jobs Running.

You have 0 Jobs Completed in 24 hours.

Activity Logs

You logged into the Dashboard on 31 Oct 2019 05:00:40 pm.

SciServer Apps

CasJobs

Search online big relational databases collections, store the results online, and share them.

Compute

Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.

Compute Jobs

Asychronously run Jupyter notebooks in Python, R and MATLAB or commands.

SciDrive

Drag-and-drop file hosting and sharing services.

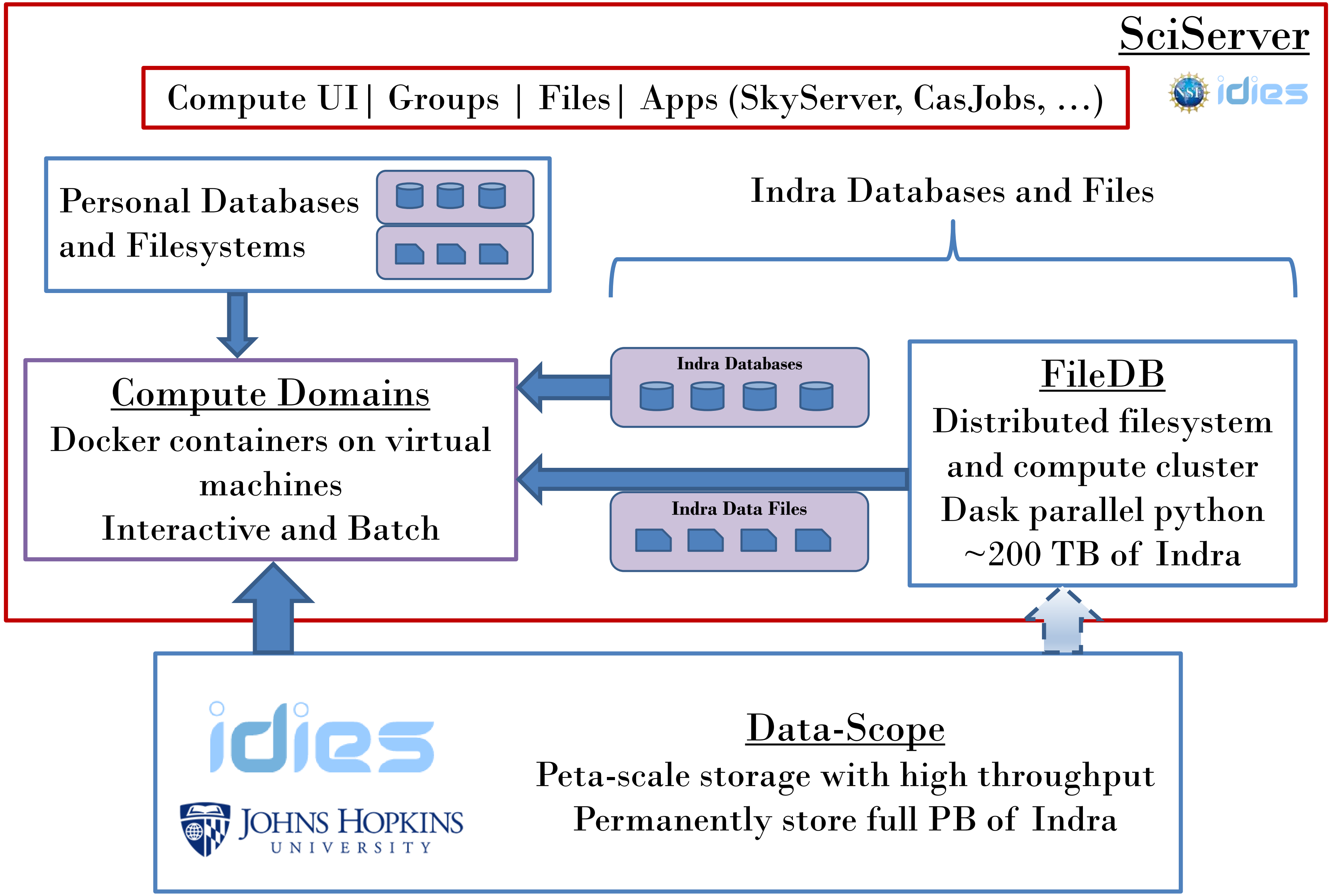
SkyServer

Access the Sloan Digital Sky Survey data, tutorials and educational materials.

SkyQuery

A scalable database system for cross-matching astronomical source catalogs.

Indra Infrastructure

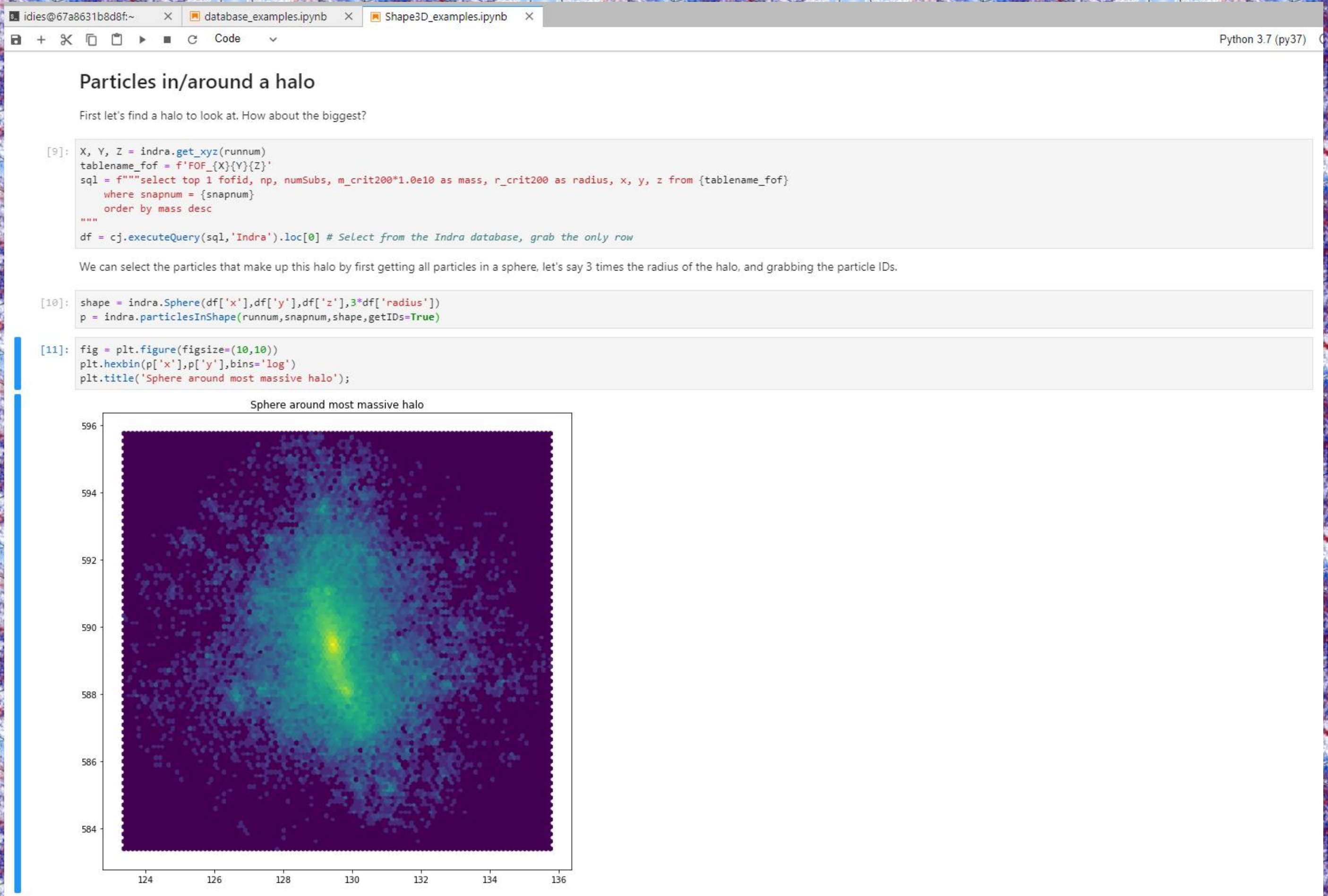

SciServer
Compute UI | Groups | Files | Apps (SkyServer, CasJobs, …)
Personal Databases and Filesystems
Indra Databases and Files
Compute Domains
Docker containers on virtual machines
Interactive and Batch
Indra Databases
Indra Data Files
FileDB
Distributed filesystem and compute cluster
Dask parallel python
~200 TB of Indra
Data-Scope
Peta-scale storage with high throughput
Permanently store full PB of Indra

- Permanent read-only storage connected to compute domain
- Distributed filesystem for parallel computation
- Leverage relational databases
- Accessible through SciServer with its collaboration tools and hosted astrophysical datasets

The indra-tools software library

- Python library to read and interact with data
  - Don't assume expert users
  - Make it easier for experts
  - Hide the file system

- Example notebooks
  - Make it easy to get started
  - Show sample database queries
  - Explain advanced features (e.g., Shape3D)



Particles in/around a halo

First let's find a halo to look at. How about the biggest?

```
[9]: X, Y, Z = indra.get_xyz(runnum)
     tablename_fof = f'FOF_{X}{Y}{Z}'
     sql = f"""select top 1 fofid, np, numSubs, m_crit200*1.0e10 as mass, r_crit200 as radius, x, y, z from {tablename_fof}
         where snapnum = {snapnum}
         order by mass desc
     """
     df = cj.executeQuery(sql,'Indra').loc[0] # Select from the Indra database, grab the only row
```

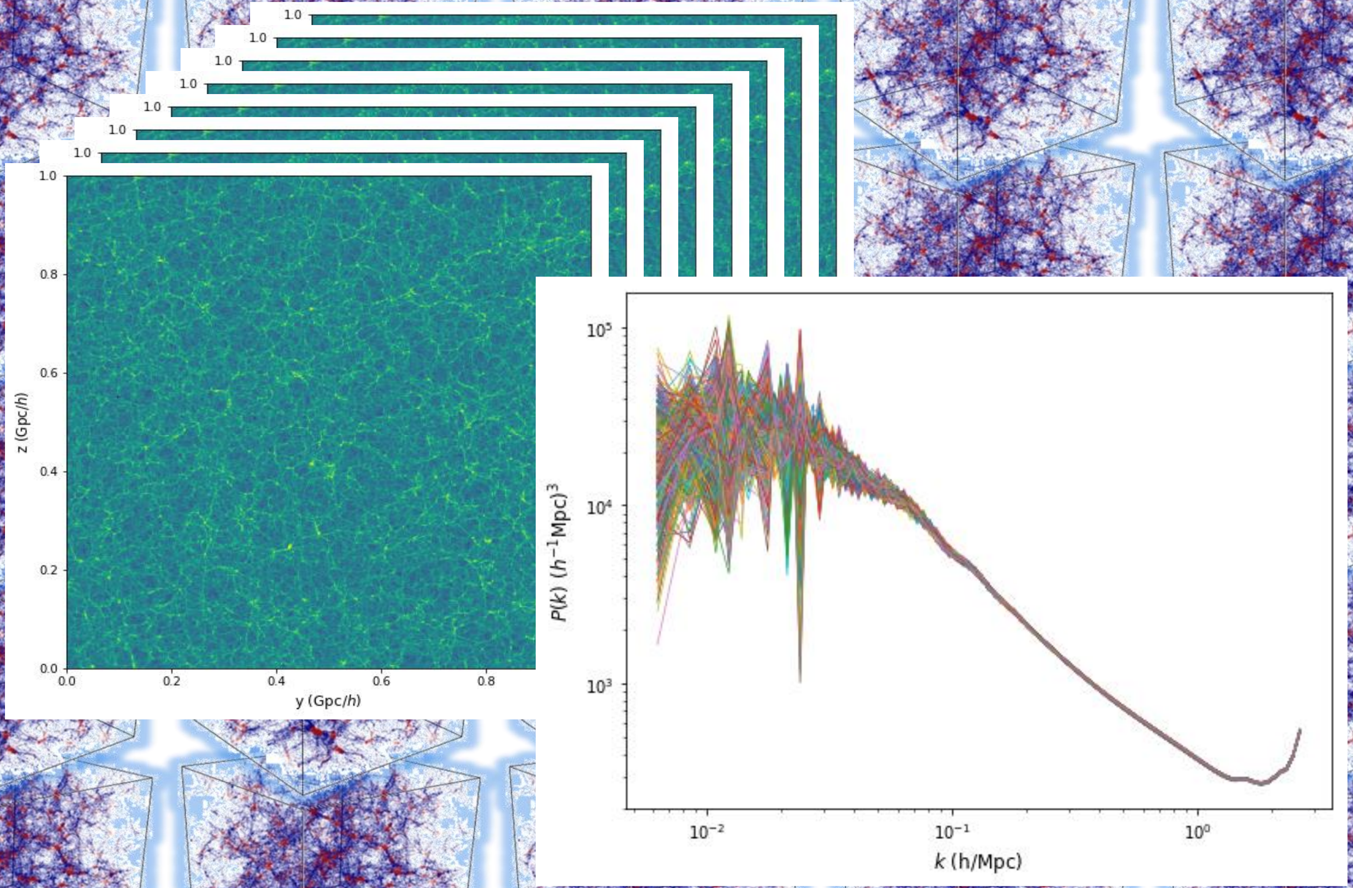
We can select the particles that make up this halo by first getting all particles in a sphere, let's say 3 times the radius of the halo, and grabbing the particle IDs.

```
[10]: shape = indra.Sphere(df['x'],df['y'],df['z'],3*df['radius'])
      p = indra.particlesInShape(runnum,snapnum,shape,getIDs=True)
```

```
[11]: fig = plt.figure(figsize=(10,10))
      plt.hexbin(p['x'],p['y'],bins='log')
      plt.title('Sphere around most massive halo');
```

Infrastructure that enables heavy computation

- FileDB distributed filesystem with a Dask parallel python cluster
- 448 Cloud-In-Cell density grids calculated in 2 hours!
- 481 billion particles total
- Still testing different use cases and building job submission capability



Discussion Questions

- What hardware and technologies are required to host large public data sets and make them computationally accessible?
- What are the unique requirements or challenges of hosting simulated data vs. observational archives?
- When we plan archives for large missions, how do we ensure that we don't leave theory behind?
- …