# What should the cosmo+CMB data repository look like?

Stephen Bailey, DESI Data Management

**Context**: Unlike NSF and NASA, DOE lacks a data archive center dedicated to curating datasets and simulations beyond the lifetime of individual projects, and facilitating their joint analysis.  This is especially relevant for the Cosmic Frontier, where combining public datasets and external simulations is the norm.

This is an opportunity for LBNL / NERSC, where we have a prototype in the "Cosmology Data Repository"
(primarily used internally)

Non-technical challenges:

- Buy-in across the DOE labs to make it useful for everyone within DOE
- What would be useful for *us*?

Panel Discussion:

- Julian Borrill, CMB
- Joe DeRose, Simulations
- Debbie Bard, NERSC / LSST-DESC
- Peter Nugent, CRD

# What should the cosmo+CMB data repository look like?

Julien Borrill, CMB

- Explicitly named as the DOE Cosmic Frontier data archive - CMB is cosmology too!
- Supporting data explorations as well as transfers
- Mirrored across multiple DOE sites for data security and access reliability
    - Which sites are already primary/secondary data centers for the big Cosmic Frontier experiments?
        - CMB (including S4): NERSC + (planned to be ALCF)
- Granular access controls
    - Public
    - Collaboration
    - Working Group
- Coupled to project data distribution services
    - Access permissions assigned to each dataset
    - Access groups auto-populated by projects
- Integrated into the DOE computing ecosystem
    - Superfacility
    - Federated identity
    - Seamless data staging
    - Local analysis resources

# What should the cosmo+CMB data repository look like?

Joe DeRose, Simulations

- Accessible from DOE compute facilities
  - Ease of access from compute nodes for production level analysis, and things like Jupyterhub at NERSC for development.
- Database-like structure for complex simulation outputs
  - Hierarchical data association is common.
  - Often useful to be able to rapidly associate objects of various types in a single simulation. E.g. galaxies and their host halos.
- Make small subset of common analysis tasks as simple to perform in-situ as possible.
  - E.g. correlation function/power spectrum estimation on subsets of simulation objects.

# What should the cosmo+CMB data repository look like?

Debbie Bard, NERSC / LSST-DESC

This could be done easily, or it could be done well.

**Easy approach**:

- data server with simple interface (a la SDSS)
  - Data take-out, scientists do processing etc at home

**Comprehensive approach**:

- data portal with search and processing capabilities, with well curated, annotated data.
  - Metadata: provenance, what analysis has been applied, relevant cuts/selections, uncertainties
  - Would require buy-in from the community on standardized units, formats and metadata
  - Include not just data, but code/tools/models (eg repo of ML models applicable across fields)
  - This would be a *data user facility* and should be resourced as such
    - Simply standing up a bunch of disks with a web interface is not sufficient to get value out of datasets
    - Hardware is specialised, requires expertise to maintain.
    - Need very robust fat pipes to the outside world, if it is to be queried from multiple sites
    - Needs to handle fluctuating demand (multi-site? Redundant copies? )
    - Needs excellent documentation and support

# What should the cosmo+CMB data repository look like?

Peter Nugent, CRD

- Should provide a framework in which joint analysis of various data sets could be done easily
  - e.g. LSS+CMB, LSST+DESI, etc.
  - This would also include simulations+data.
- For the future and LSST, there could, and should, be some real-time component linked to transient science
  - SNe, MMA, Strong Lensing.