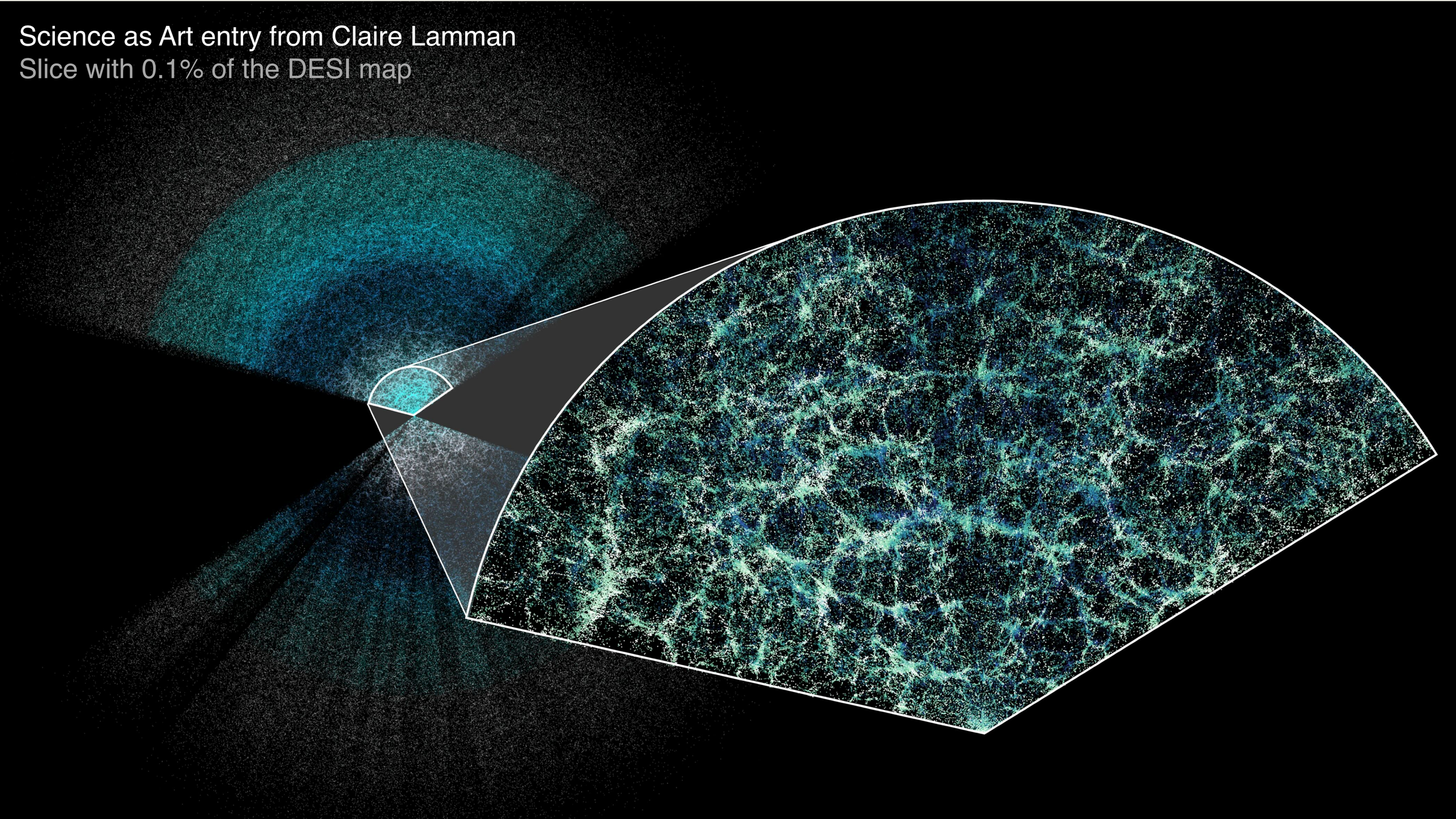
Making a 3D Map of the Universe at NERSC with the Dark Energy Spectroscopic Instrument (DESI)

Stephen Bailey
LBNL Physics Division
September 30, 2024

Science as Art entry from Claire Lamman

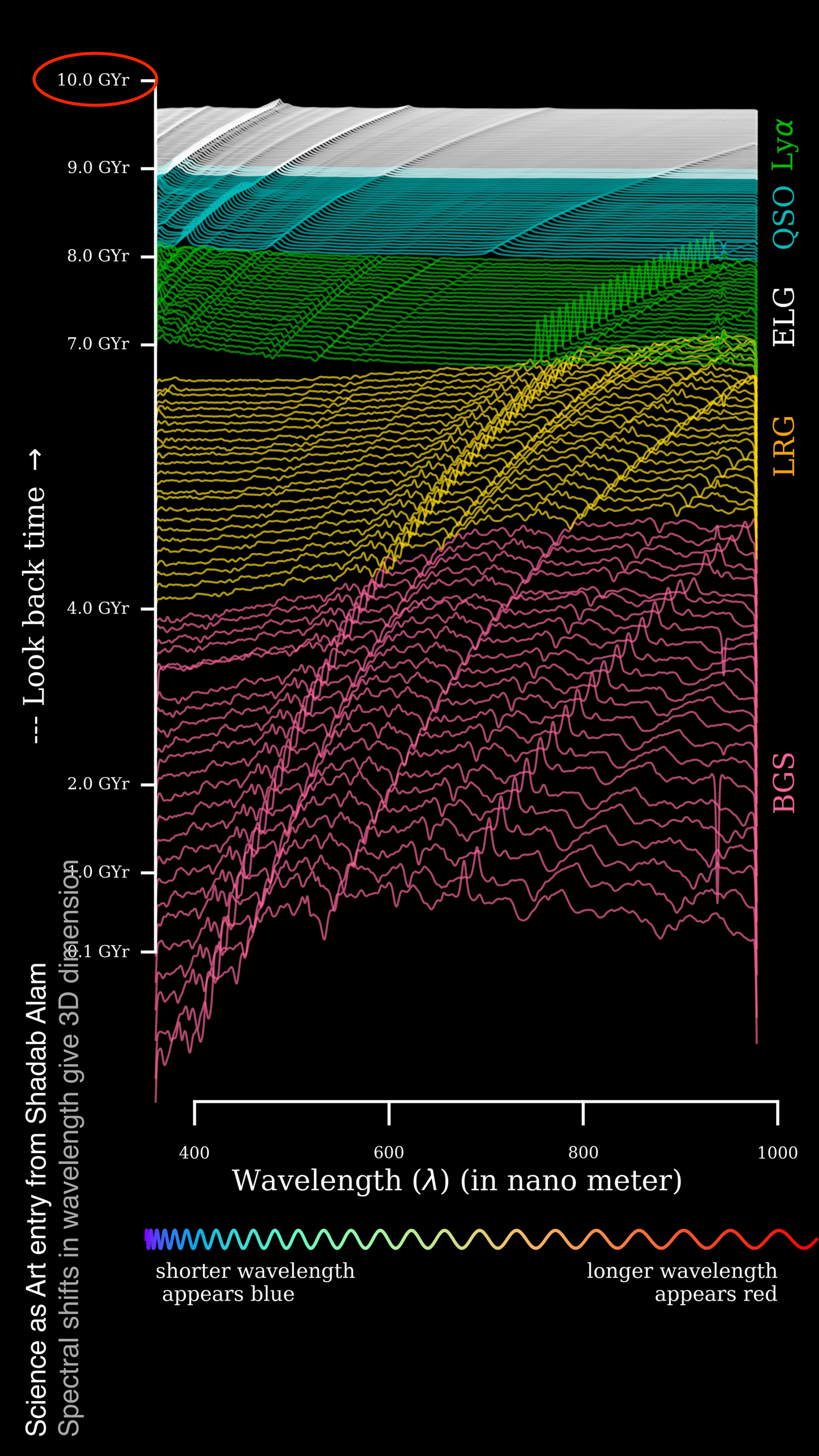
Slice with 0.1% of the DESI map

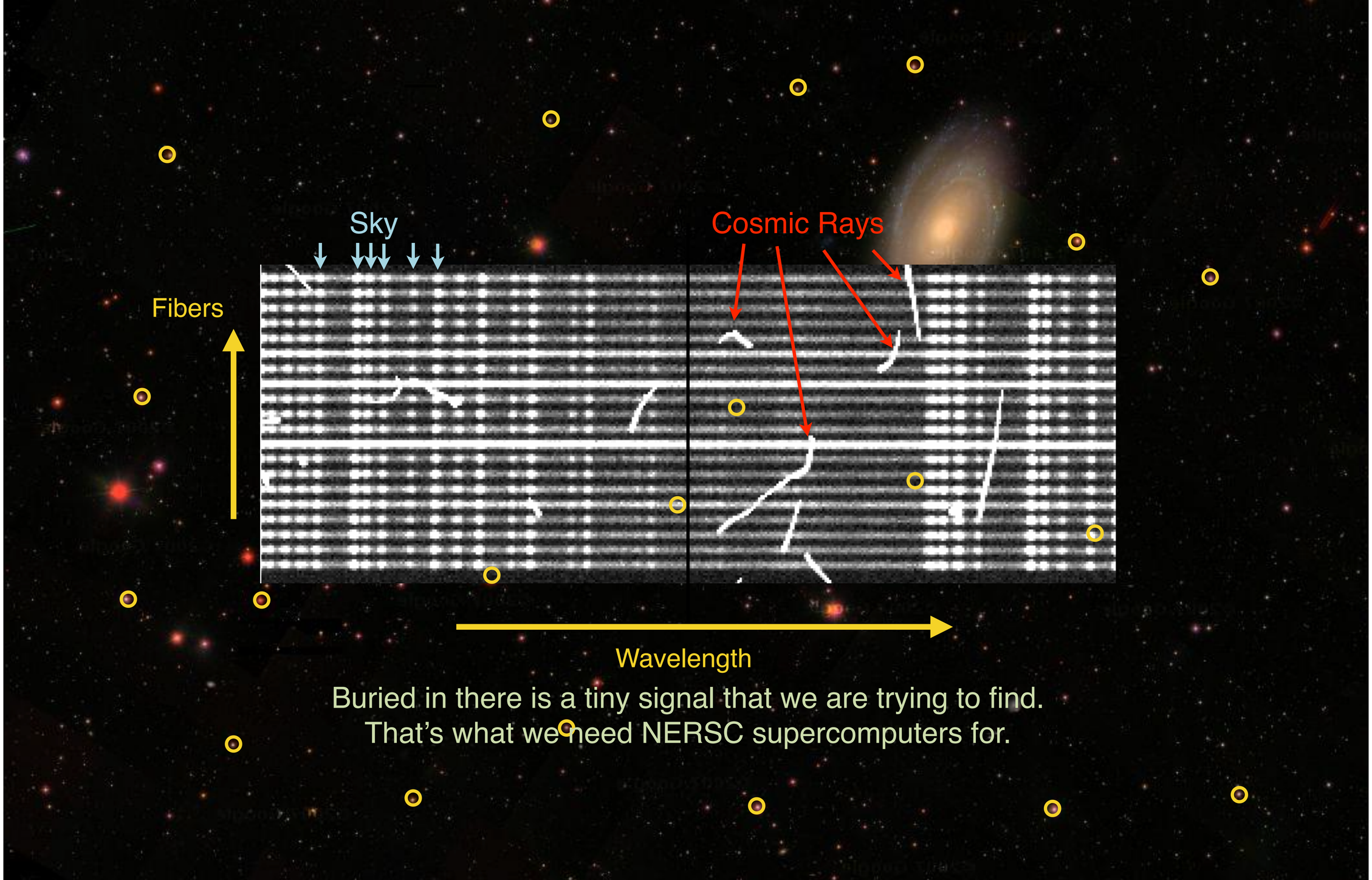


Start by Making a 2D Map of the Universe

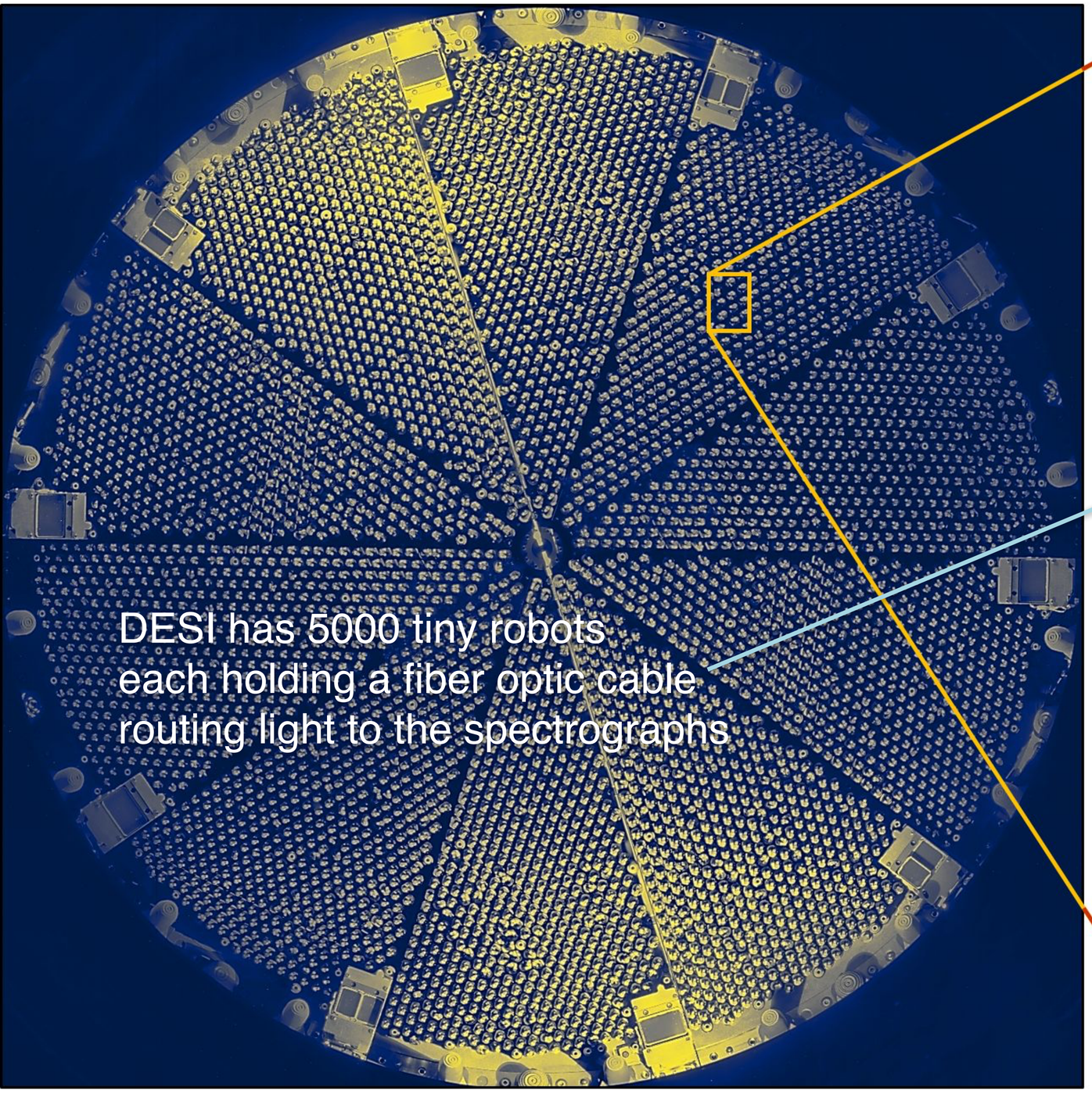
- Simultaneously fit data from
  - 4 telescopes on Earth
  - 2 telescopes on satellites
  - 6 funding agencies
  - 4 continents
  - 4 data portals
- Co-locating data at NERSC enabled joint processing
  - 2.8 billion objects identified in the 2D map
  - >50% of these are other galaxies (not stars in our own galaxy)
- Generations of processing spanned Edison, Cori, and now Perlmutter
- Public data releases hosted by NERSC at https://legacysurvey.org
  - 671 papers using these data (as of September 27 2024)


RA,Dec = 10.0377, 2.9520, zoom 12
+
−
legacysurvey.org/viewer
+ Images
+ Overlays
arcmin
Contrast: 1
Brightness: 1
Jump to object: NGC 5614
Custom catalog upload (FITS or CSV; RA,Dec,[name,color,radius]):
Choose File
no file selected
Upload
Tips & Tricks | Leaflet | Source | © Legacy Surveys / D.Lang (Perimeter Instit


10.0 GYr
9.0 GYr
8.0 GYr
7.0 GYr
4.0 GYr
2.0 GYr
1.0 GYr
0.1 GYr
--- Look back time →
QSO Lyα
ELG
LRG
BGS
400
600
800
1000
Wavelength (λ) (in nano meter)
shorter wavelength
appears blue
longer wavelength
appears red
Science as Art entry from Shadab Alam
Spectral shifts in wavelength give 3D dimension



Buried in there is a tiny signal that we are trying to find.
That's what we need NERSC supercomputers for.


DESI has 5000 tiny robots
each holding a fiber optic cable
routing light to the spectrographs

Daily (nightly!) DESI Operations

Telescope at Kitt Peak near Tucson

NERSC

Semi-realtime data transfer

realtime queue

regular queue

Feedback
— semi-realtime QA
— survey ops next night

Nightly Processing

Yearly Data Assemblies

this is primary motivation for using NERSC / HPC center

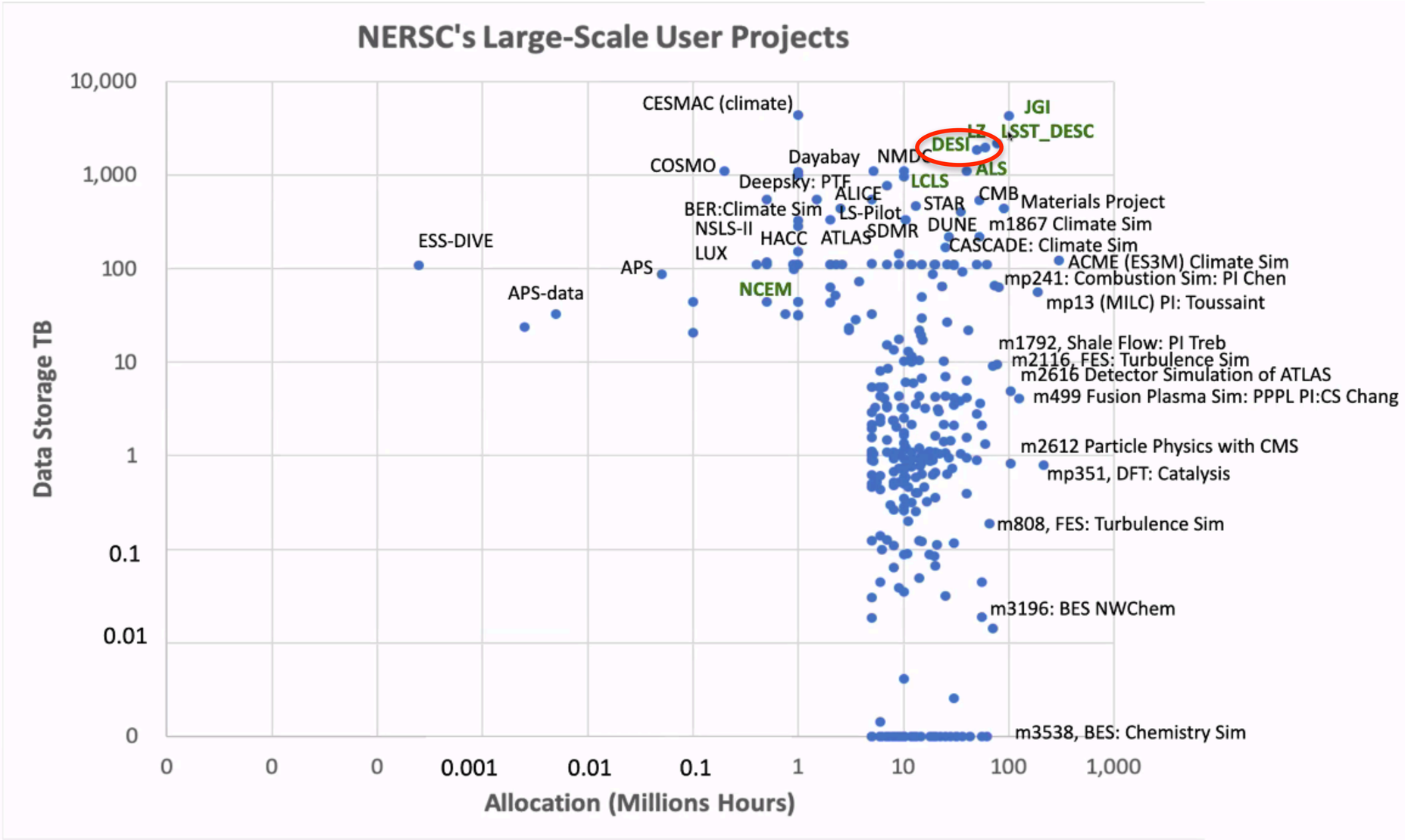
Cosmology Papers!

5k spectra
every ~15 minutes
every night
for 5 years
= 10s of millions of spectra!

DESI Collaboration (hundreds of scientists, worldwide)

DESI @ NERSC in context

From 2020



Among the largest (data x compute)

DESI is ~10% of all NERSC users

DESI uses the full NERSC ecosystem

- Compute
  - Realtime for nightly processing
  - Big Iron for quarterly/yearly reprocessing and science analyses
- I/O
  - CFS, scratch, HPSS
  - Data Transfer Nodes, Globus, rsync, spin container with nginx (data.desi.lbl.gov)
- Workflow
  - scronjobs
  - Databases
- Analysis
  - Jupyter
  - Interactive, debug, and regular queues
- QA monitoring
  - Spin, more scronjobs
- NERSC liaisons to facilitate communication, User services for account management + help desk

Compute capacity + everything else is why we are at NERSC.
Having **all at one location** is the unique benefit for NERSC.

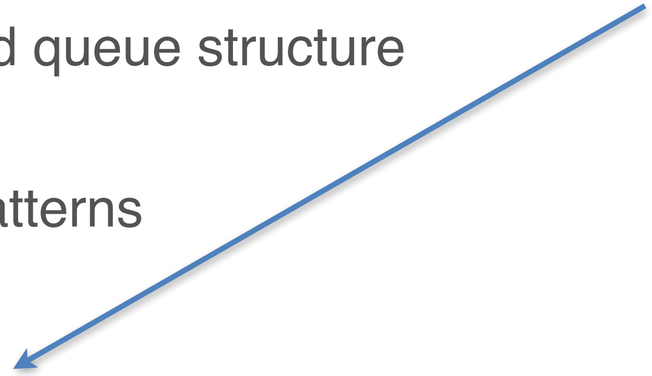
A (formerly) unusual mix

- Designed for HPC from the start
- Written in Python from the start

- This has been a very effective combination for DESI

Designed for HPC from the start

- Instead of porting legacy serial code and trying to get it to fit into HPC, decided to rewrite from scratch
  - Take ideas / algorithms / design philosophy, but no actual code
  - Fresh start to implement new ideas
  - Modernize code practices for better maintainability
- "Designed for HPC" meant
  - Parallelism considered in code design from the start
  - Accepted reality of memory/core and queue structure
- "Designed for HPC" did *not* mean
  - Always using classic HPC design patterns
  - Using a classic HPC language
  - Accepting NERSC "as-is"
    - Constructive collaboration with NERSC



Early adopters of
– Spin
– Collaboration accounts
– Workflow nodes (Cori)
– scronjobs (Perlmutter)
– read-only CFS mount /dvs_ro

Ongoing advocacy for robustness

Example of non-traditional HPC usage

- Context
  - Telescope produces new exposures every ~15 minutes
  - Logically these can be processed independently of each other
- Initial design
  - Bundle N>>1 exposures into a single massively parallel job
  - Computationally most efficient, most HPC-like
  - Problem: failure of any single exposure impacts all other unrelated exposures
    - Our problem (algorithmic, data quality) or NERSC's problem (bad node, I/O hiccup); same effect
- Current model
  - N>>1 independent small jobs (mostly single node for ~20 minutes)
  - Most robust to individual failures, easiest to recover
  - Problem: Less efficient, exceeding queue submit limits means more job hand-holding
- As code gets faster and machine more stable, bundling becomes more attractive again

Another example of non-traditional HPC design

- Underlying algorithms composed of steps that can be run on a laptop
- Wrapped by MPI parallelism and job workflow, while separating algorithms from parallelism
- GPU-optional

- Designed for scaling up on HPC, but HPC isn't required just to run the code
  - parallelism optional
  - GPUs optional

- Result: Efficient model for algorithm development and debugging

Written in Python from the start

- ~20 years ago
  - Using Python to coordinate job submission and parse logs
  - Needed some additional library so asked NERSC to install it (pre-conda / docker / virtual env...)
  - Help desk refused, citing that Python was “not an HPC language”
- Today
  - Python has first-class support at NERSC
  - More users login via Jupyter (dominantly Python) than ssh
  - 3rd party libraries make Python an effective HPC(-lite) language
- DESI Python toolkit
  - **mpi4py** for parallelism
  - **numpy**, **scipy** to leverage core compiled algorithms written by others
  - **numba** for JIT compilation of DESI Python –> compiled code
  - **cupy** for GPU while maintaining CPU-only option for other sites

Coming full circle on HPC + Python

- One of our remaining C++ codes is still CPU-only, taking ~15% of production time
- Exploring porting it from C++ to Python so that we can use cupy to port to GPUs while maintaining a CPU-only path for non-NERSC usage

Yay, NESAP!

NERSC Science Acceleration Program

- NESAP round 1: optimizing for Cori CPUs
  - 10x faster on Cori Haswell, made Cori KNL viable to use

Laurie Stephey
Rollin Thomas

- NESAP round 2: porting to Perlmutter GPUs
  - *additional* 17.6x faster

Daniel Margala
Rollin Thomas

- NESAP round 3: porting additional codes for GPUs
  - ongoing

Soham Ghosh
Daniel Margala

- NESAP has been a game-changing partnership between NERSC and DESI

DESI Data Releases

	Year 1	Year 3
Galaxies + Quasars in map	14.7M	33.6M
Jobs	34k	55k
Files	5M	10M
TB	212 TB	453 TB
Walltime	2.5 weeks	2 weeks
CPU + GPU node-hours	~1200+6000	2000+9500
Public Release	April 2025	2026

Using Perlmutter enabled us to double our dataset while *reducing* the walltime to reprocess

252 papers from DESI science collaboration so far (September 29 2024)
~100 additional papers using the preview “Early Data Release”
Hundreds of citations to DESI results

Thanks for 10+ years of NERSC+DESI partnership

- NESAP program
- Storage Group
- User services
- Superfacility program and liaisons
- Management
- DESI users