How big is big data?

Working with large astronomy datasets

150,000,000 rows (of 11 double-precision columns) would take ~2 hours over typical home connection, ~10 seconds over internal 10 Gb/s connection, and ~1 s to pull out of RAM from local machine.

We have tables with 15 billion rows (200 hours)

1 TB can be transferred in 20 minutes over internal DL network, pulled from RAM in 1 minute, ~1 week over typical home connection

Science Platforms

Why?

- Growing data volume *and* complexity;
- New mode(s) of doing research that are Data-driven and/or Archive-driven

Goals:

- Maximize scientific output of community: legacy science, versatile tools, collaborative workspaces, joint analysis of large datasets, serendipitous discoveries
- Framework for robust science: data quality, reproducible workflow, data longevity

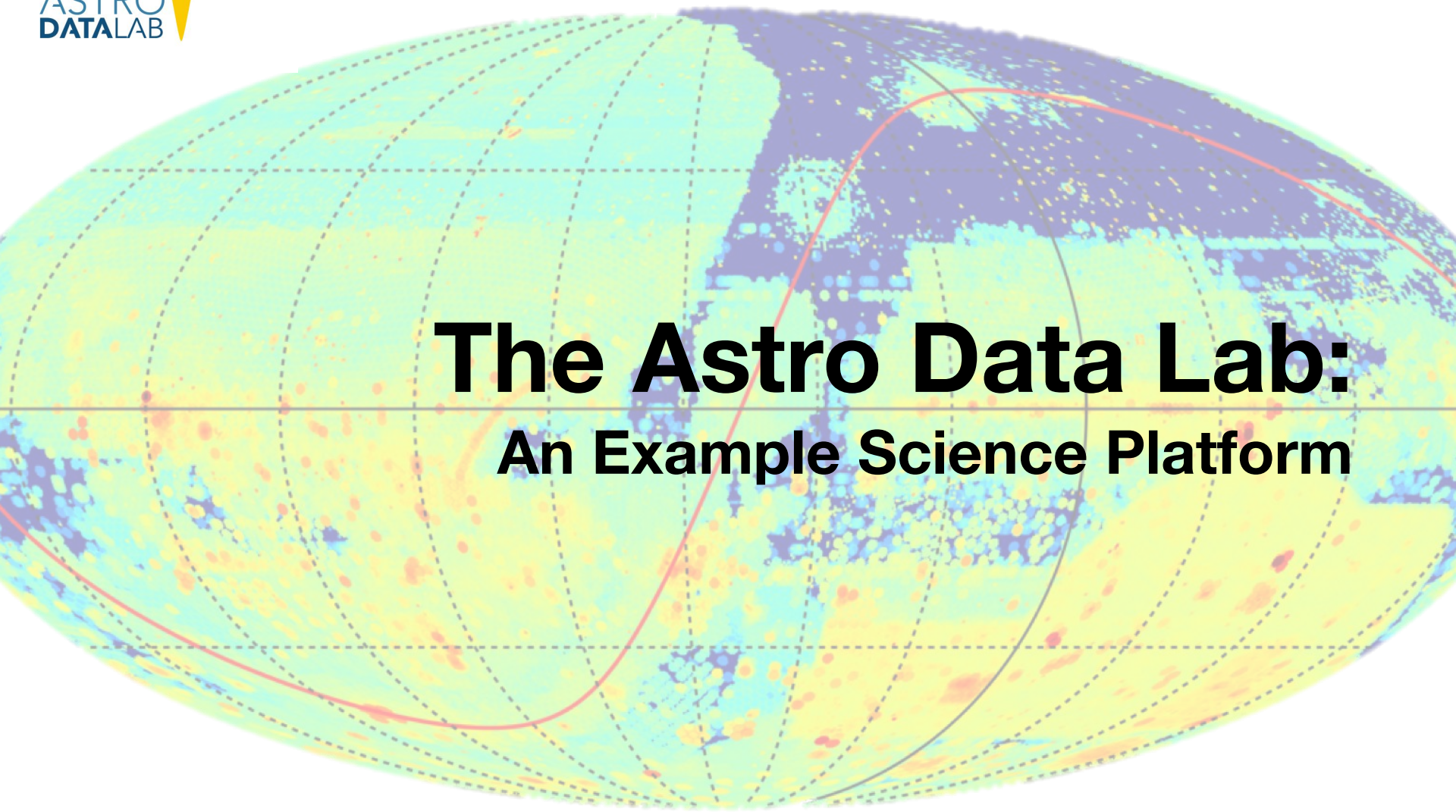
Science Platforms

Roles of science platforms:

- Bring the analysis to the data (software tools)
- Lower the barrier of entry: **user-friendly** interfaces, training of workforce at all career stages, tutorials and collaboration with educational institutions or across disciplines
- Coordinate among science platforms: share expertise & lessons learnt, similar interfaces/technologies, data/code transfer

What?

- Software (& hardware) with a backend and a client side
- Standard protocols (e.g., VO) “under the hood”
- Documentation and tutorials (user manual, helpdesk ++)
- “Everything up to writing the paper”



The Astro Data Lab:
An Example Science Platform

NSF's National Optical-Infrared Astronomy Research Laboratory

AURA

@DataLabAstro

Data Lab Team

Cross-section of astronomy, data science & computer science

Current team:

- Adam Bolton, Acting Program Coordinator
- Mike Fitzpatrick, Lead Software Engineer
- Benjamin Hauger, System Administrator
- Wendy Huang, Lead Web Developer, Technical Manager
- Stephanie Juneau, Data Scientist
- David Nidever, Data Scientist
- Robert Nikutta, Project Scientist
- Adam Scott, Database Architect
- Benjamin Weaver, Data Scientist

Astro Data Lab

Goal:

- Efficient exploration and analysis of **large** astronomy datasets

Approach:

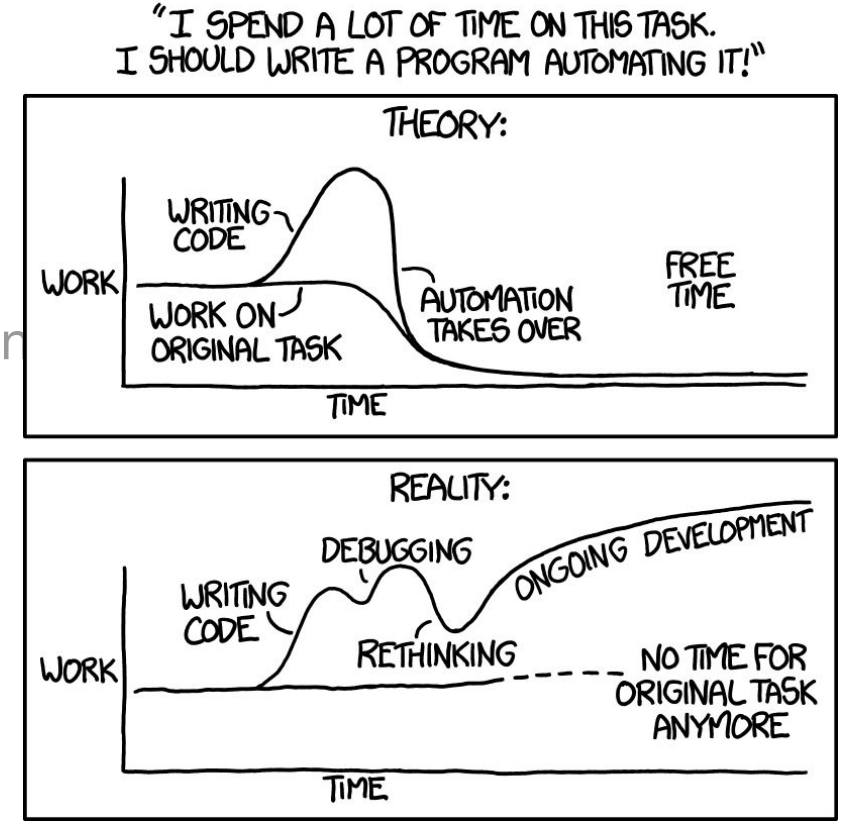
- Science question
- Data discovery
- Build intuition through interaction with selected catalog and image set(s)
- Develop workflow (automated)
- Scientific discovery!

Goal:

- Efficient exploration and analysis of **large** astronomy datasets

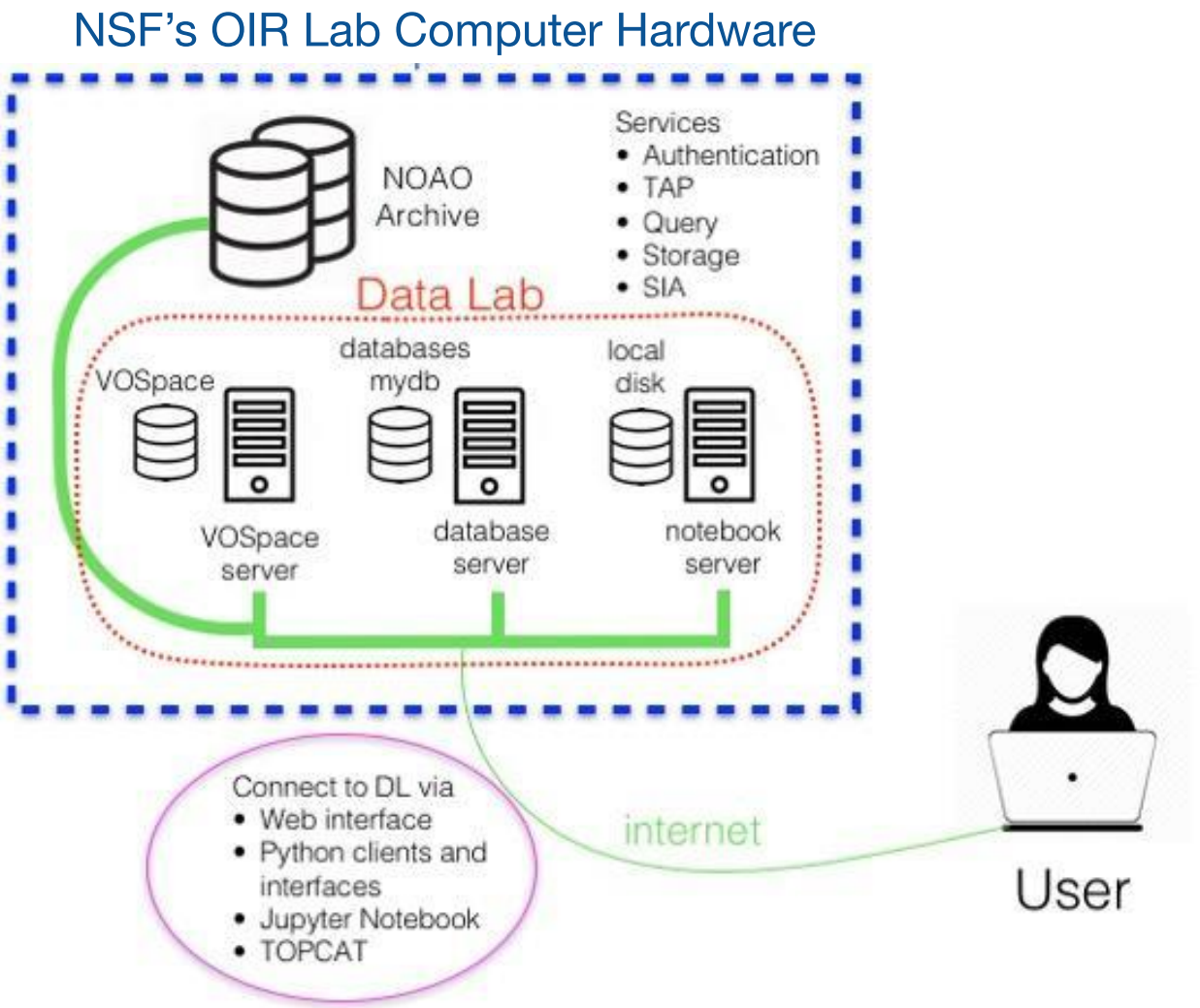
Approach:

- Science question
- Data discovery
- Build intuition through interaction with image set(s)
- **Develop workflow** (automated)
- Scientific discovery!



(xkcd.com/1319/)

Working close to the data



Data Lab in a Nutshell

Large Catalogs – TB-scale databases: DES, DESI imaging, NSC, AllWISE, SDSS, Gaia, etc.

Pixel Data – Get images and cutouts from OIR Lab's Science Data Archive

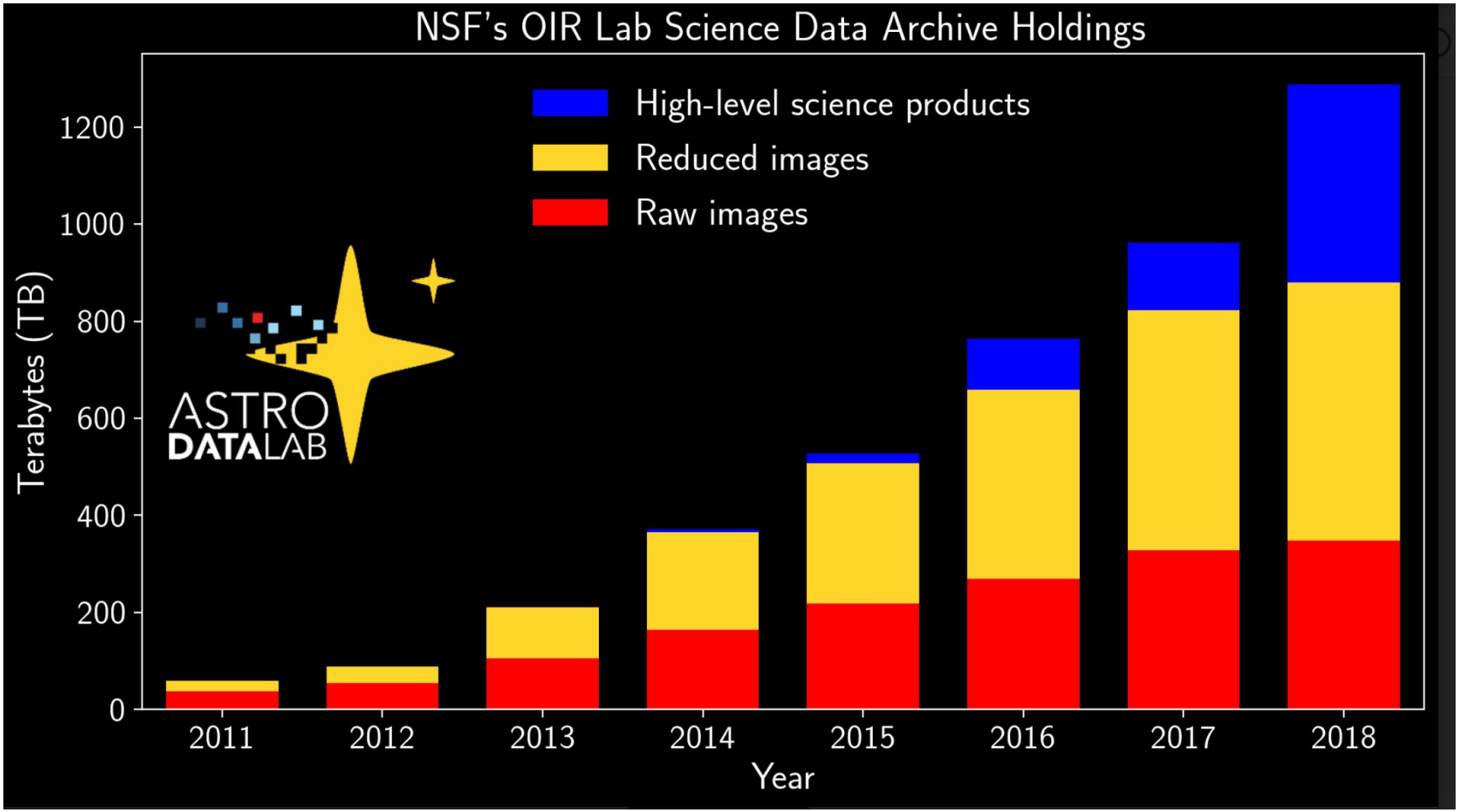
X-matching – Cross-match your catalog with any of ours, very fast

Visualization – Explore data interactively

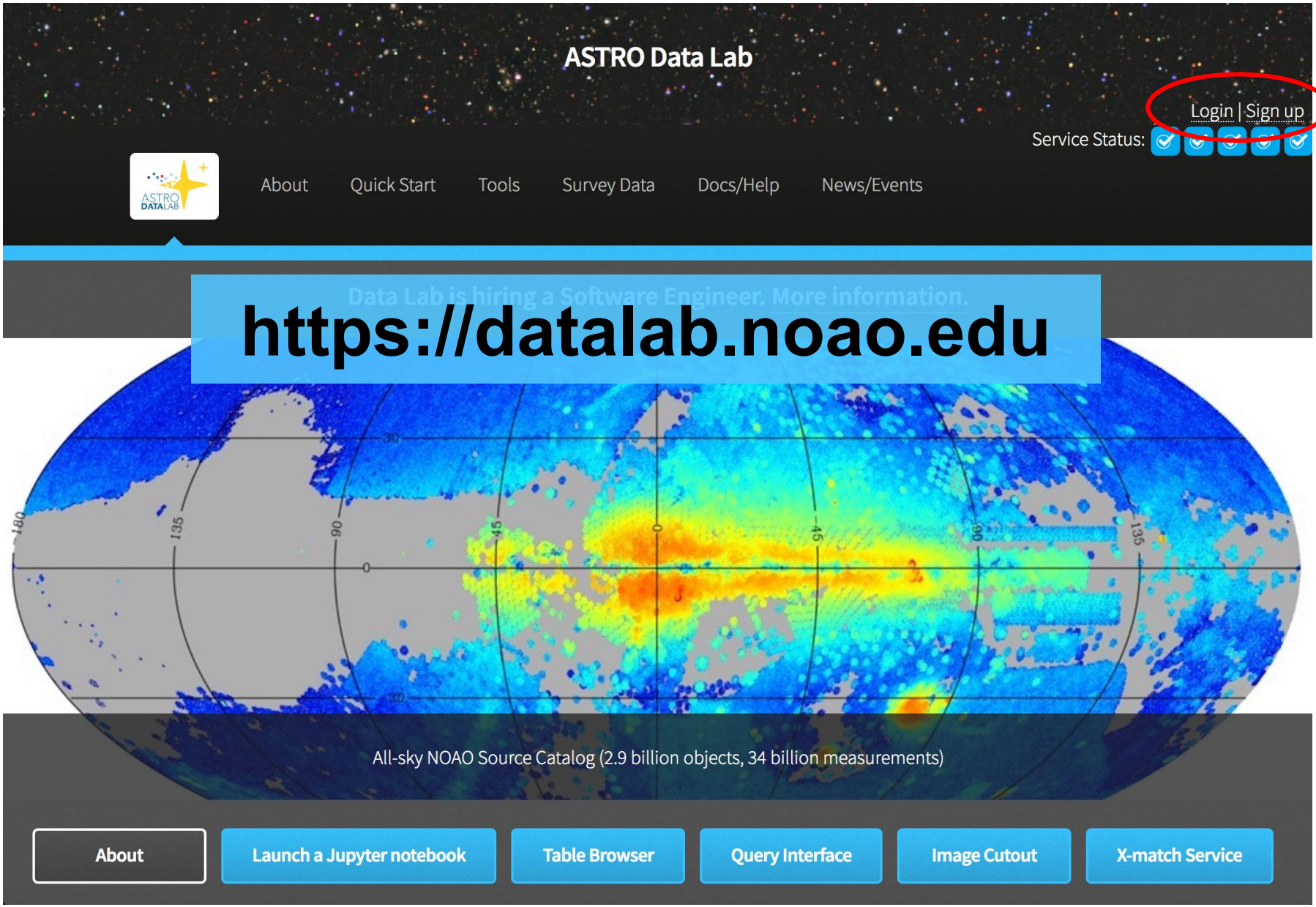
Analysis – Run Python code on our Jupyter Notebook server

Virtual Storage – Store data & results: 1 TB in VOSpace and 250 GB in myDB per user!

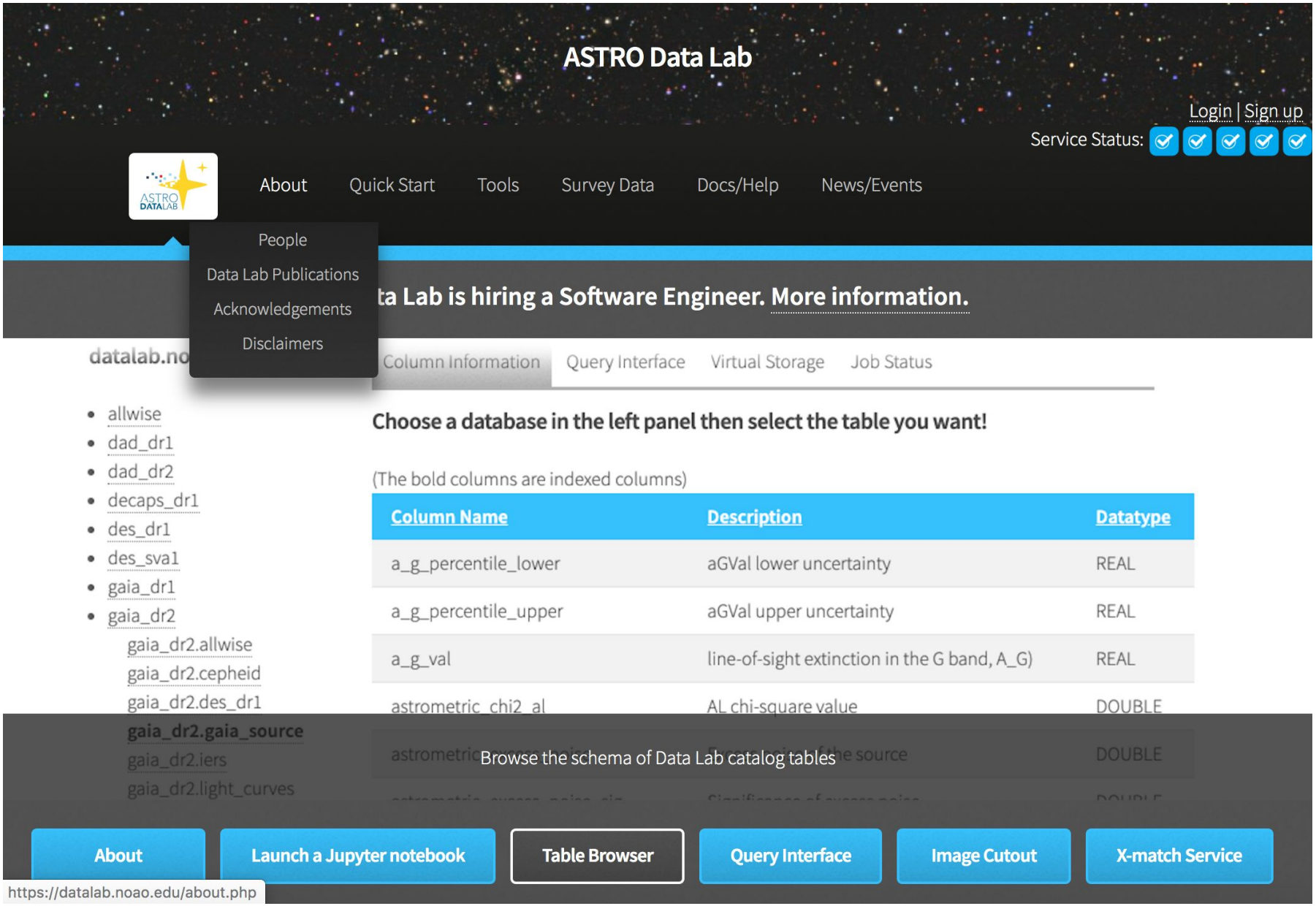
Data Holdings at OIR Lab Science Archive



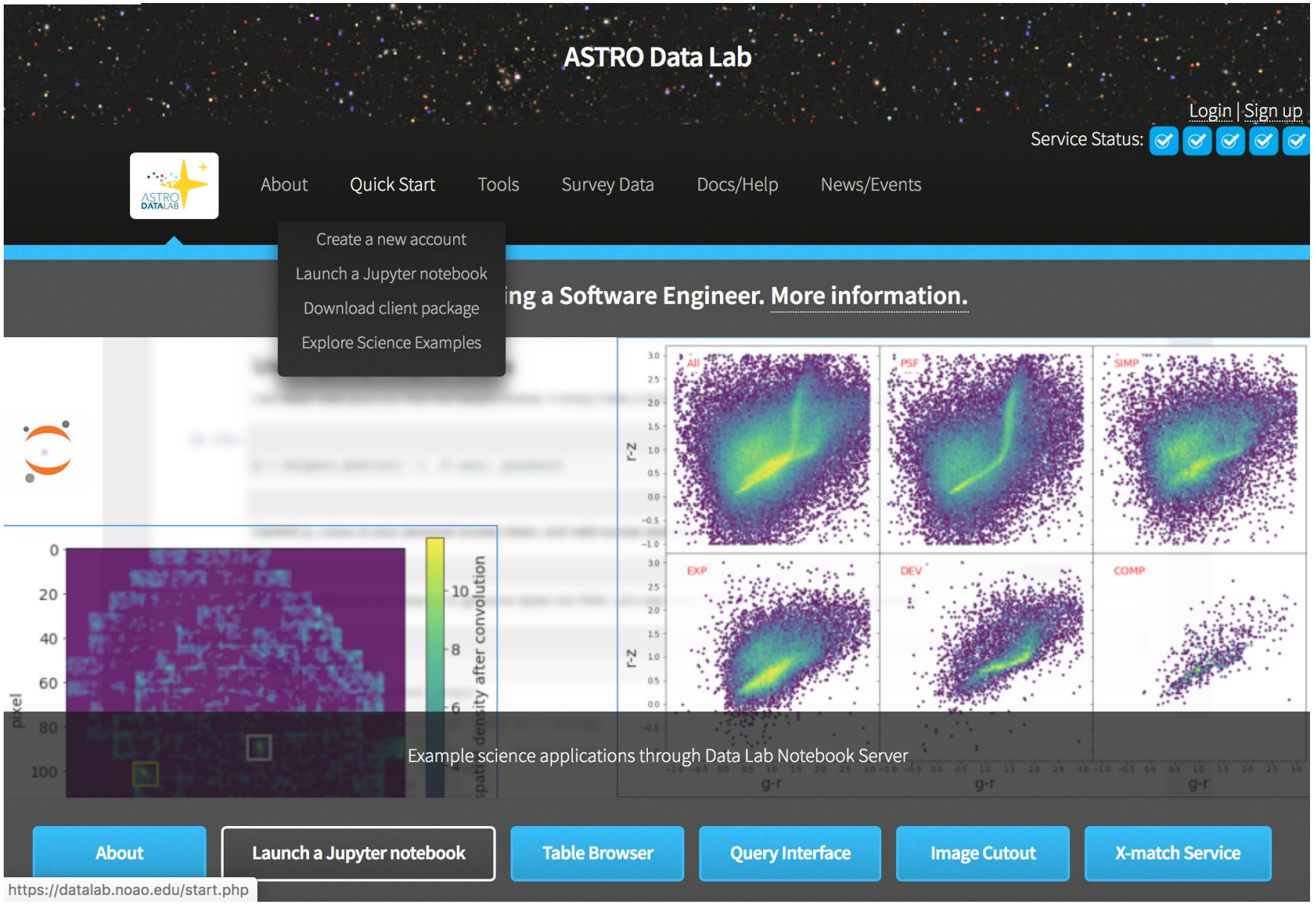
Using the Data Lab



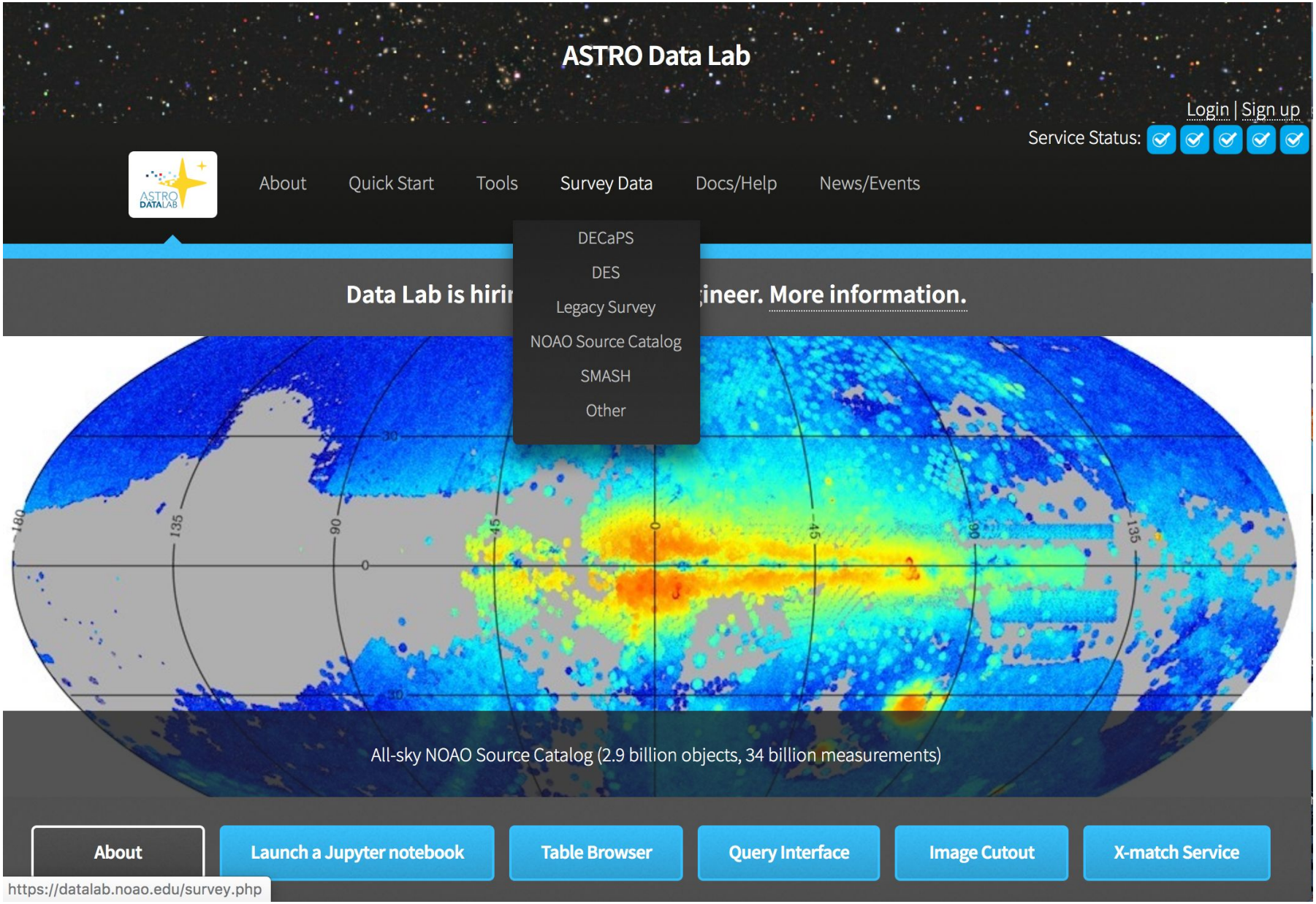
Using the Data Lab



Using the Data Lab



Using the Data Lab



Using the Data Lab: Survey Data

NSF's OIR Lab facilities

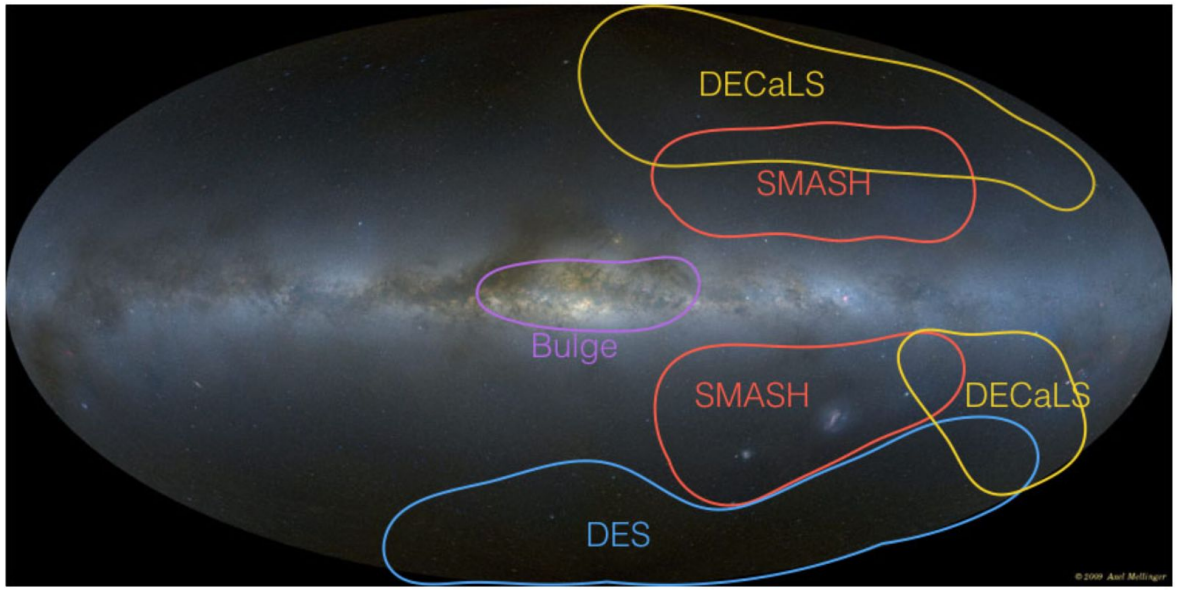
- DECaPS
- DES
- Legacy Survey
- NOAO Source Catalog
- SMASH

External Datasets

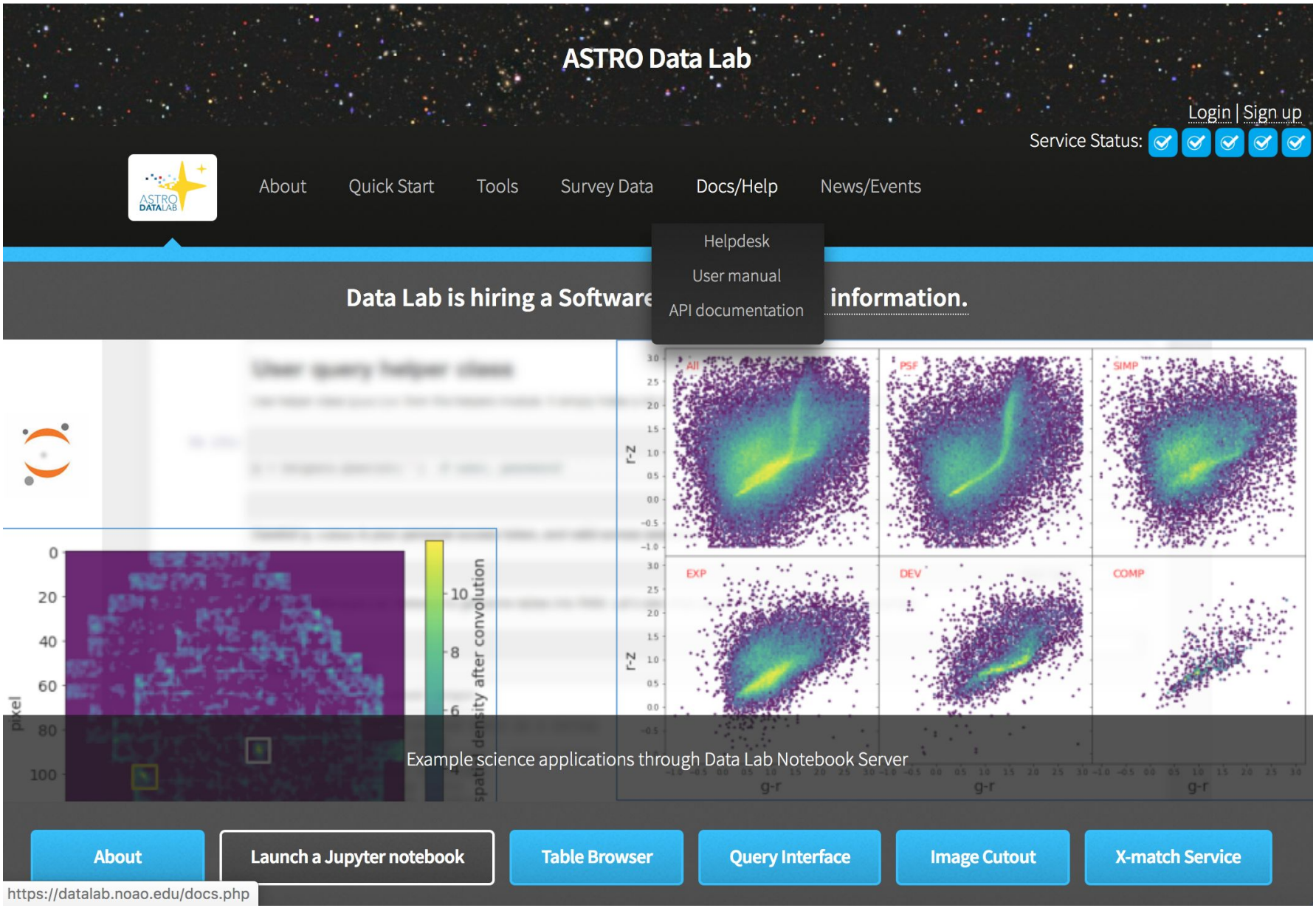
- DECam Asteroid Database
- DES SVA1
- GAIA
- PHAT
- SDSS DR13
- SPLUS
- UNWISE
- USNO

Survey Coverage

The map below shows the areas covered by surveys with catalog data currently available through the ASTRO Data Lab (SMASH, DECaLS) and those expected within approximately a year (DES, Bulge). Hover over an outline to see the survey name or click on an outline or on the sidebar links to go to the page for that survey. For pure image data from these and other observing programs, visit the NOAO Science Archive.



Using the Data Lab



Datasets available at the Data Lab

NSF's OIR Lab facilities – MzLS (Mayall 4m, Kitt Peak), DECaLS, SMASH, DES (Blanco 4m, Chile), NSC (all)

Steward Obs. – BASS (Bok 2.3m, Kitt Peak)

Public surveys – SDSS DR13-16, GAIA DR1-2, AllWISE, unWISE, USNO and ++

Hundreds of TB more coming

Total holdings at PB scale (including pixel data)

Datasets available at the Data Lab

OIR Lab/Steward Facilities Featured Surveys:

- DESI imaging Legacy Survey (LS): ~1.6 billion objects in DR8
- SMASH: ~370 million objects in DR2
- DES: ~400 million objects in DR1
- DECaPS: ~2 billion objects
- NOAO All-Sky Source Catalog (NSC): ~2.9 billion objects
- NSC (+ single-epoch): ~3 billion objects (~30 billion measurements)

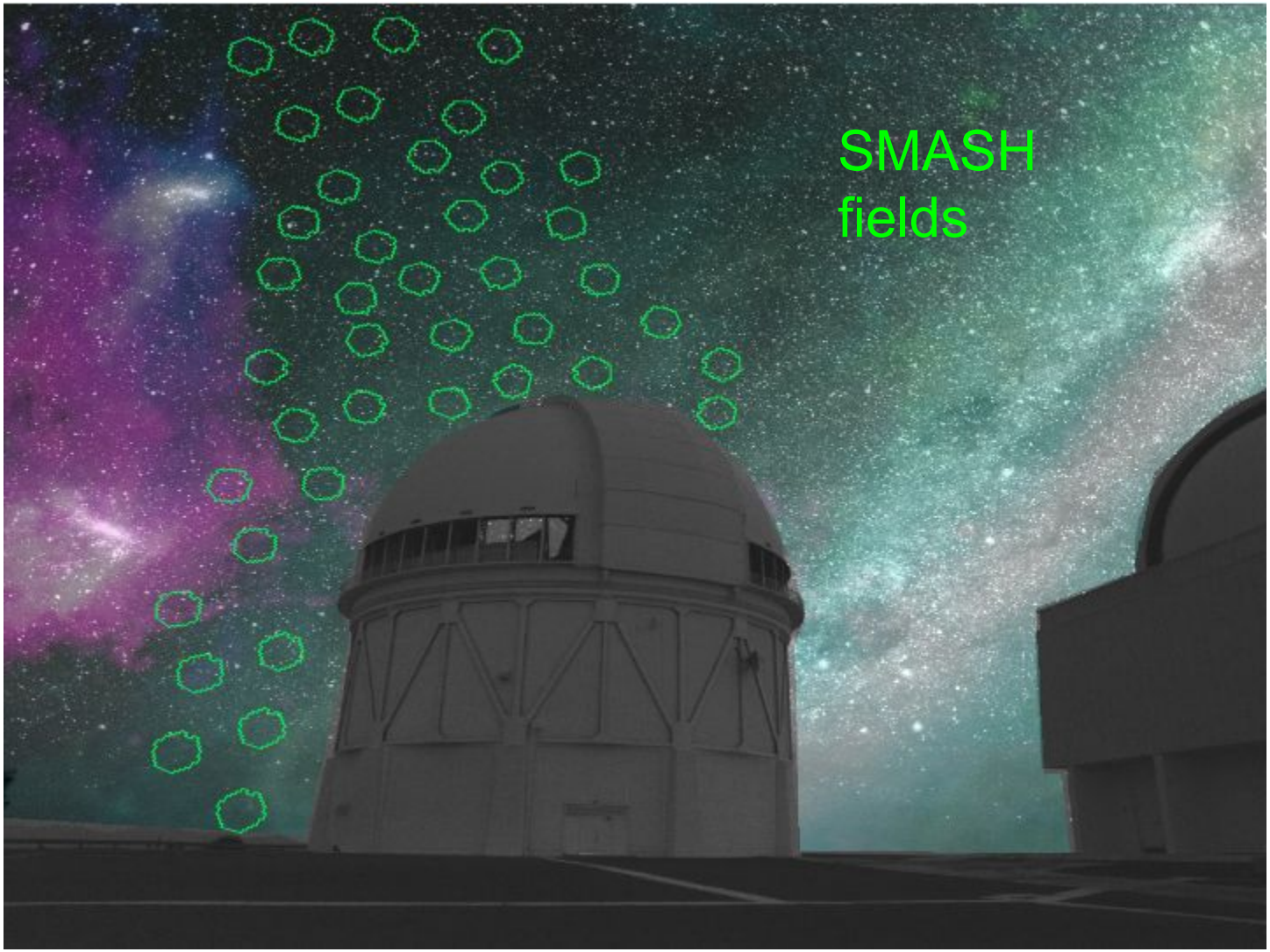
Additional Surveys:

- selected tables from SDSS/BOSS DR13, 14 & 16, All-WISE, unWISE, GAIA DR1 & DR2, DES SVA1, the Allen NEO catalog (DAD), and USNO-A2/B, etc.

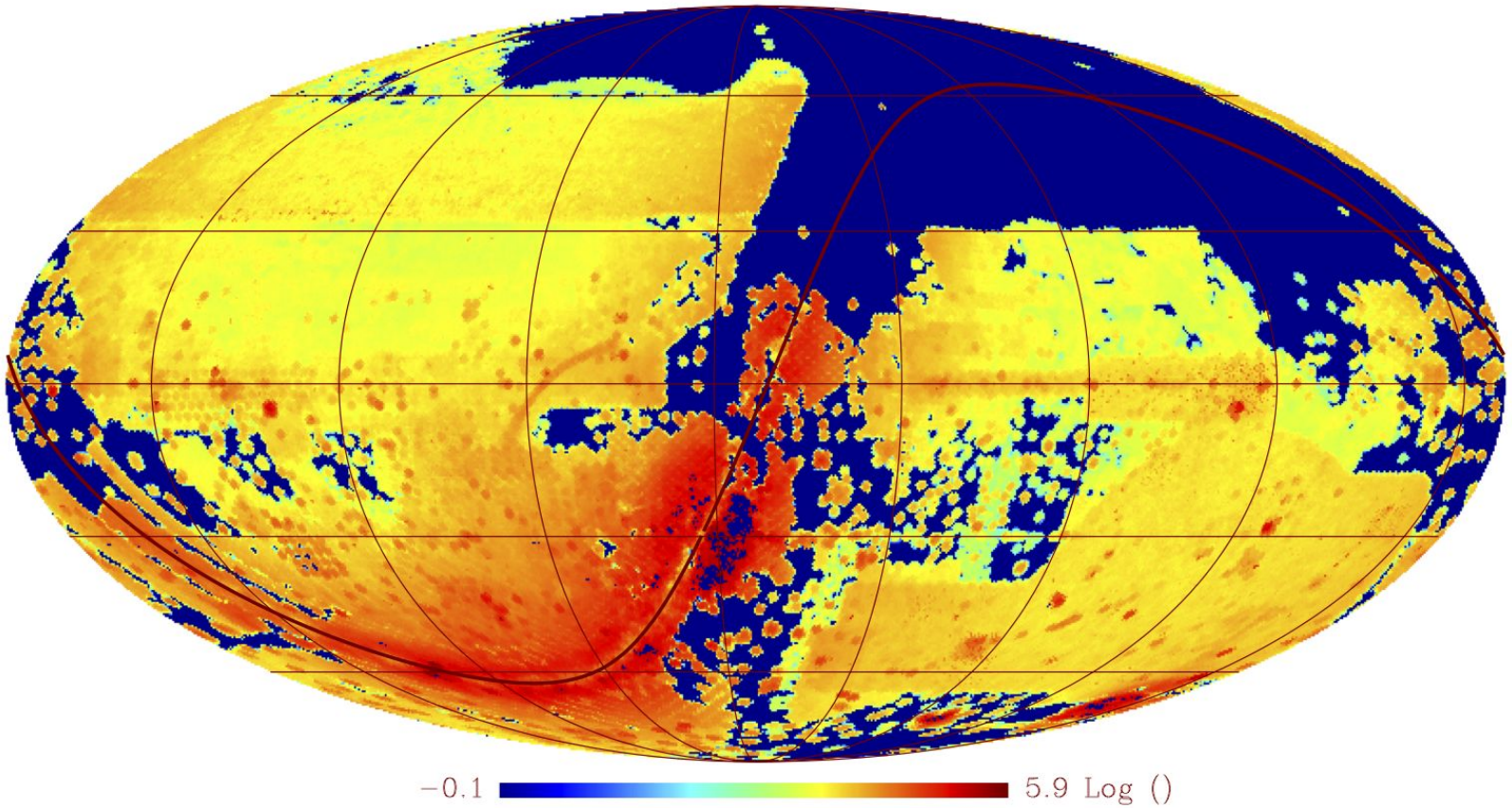
Mayall 4m & Bok 2.3m telescopes on Kitt Peak, AZ

Blanco 4m telescope in Cerro Tololo, Chile

Blanco 4m telescope in Cerro Tololo,Chile



All-Sky NOAO Source Catalog: NSC



- ~2.9 billion objects, ~30 billion measurements; basic aperture photometry
- DR1 available (Nidever et al 2018, AJ); DR2 in progress

Simulated dataset at the Data Lab

TRILEGAL simulated LSST catalog:

- produced by Leo Girardi and Giada Pastorelli
- 19 billion stars for use in investigating LSST science cases

datalab.noao.edu/tap

- allwise
- dad_dr1
- dad_dr2
- decaps_dr1
- des_dr1
- des_sva1
- gaia_dr1
- gaia_dr2
- hipparcos
- hipparcos2
- ls_dr3
- ls_dr4
- ls_dr5
- ls_dr6
- ls_dr7
- ls_dr8
- lsst_sim
  - **lsst_sim.simdr1**
- nsc_dr1
- phat_v2
- sdss_dr13
- sdss_dr14
- skymapper_dr1
- smash_dr1
- smash_dr2
- splus_dr1
- splus_edr
- stripe82
- tap_schema

Column Information | Query Interface | Virtual Storage | Job Status

Choose a database in the left panel then select the table you want!

(The bold columns are indexed columns)

Column Name	Description	Datatype
av	V-band Extinction	REAL
c_o	Surface C/O ratio by number, n_C/n_O	REAL
cexcess	Carbon excess at surface, log(n_c-n_O)-log(n_H)+12, -1 for O-rich stars	REAL
comp	Flag for binarity, single star=0, primary=1, secondary=2, entire binary=3	SMALLINT
g_bpmag	Gaia G-BP color (Vegamag)	REAL
g_rpmag	Gaia G-RP color (Vegamag)	REAL
gaia_gmag	Gaia G magnitude (Vegamag)	REAL
galb	Galactic latitude	DOUBLE
gall	Galactic longitude	DOUBLE
gc	Galactic component, thin disk=1, thick disk=2, halo=3, bulge=4, MCs=5	SMALLINT
gmag	LSST g-band magnitude (ABmag)	REAL
htm9	HTM index (order-9 -> ~10 arcmin size)	INTEGER
imag	LSST i-band magnitude (ABmag)	REAL

DESI Dark Energy Spectroscopic Instrument

- 14,000 square degrees
- >35 million spectra of galaxies and quasars!
- 10 million spectra of stars
- Commissioning started in 2019 (survey 2020-2025)

Object Class	Number of Spectra	Redshift Range
bright galaxies, r < 19.5	10 million	0 < z < 0.4
luminous red galaxies (LRGs)	4.2 million	0.4 < z < 1.0
emission line galaxies (ELGs)	18 million	0.6 < z < 1.6
quasars (QSOs)	2.4 million	0.5 < z < 3.5
Milky Way stars	10 million	---

→ desi.lbl.gov

DESI at the Data Lab

- Host DESI imaging Legacy Surveys (DECaLS, BASS/MzLS)
  → Databases (latest: ls_dr8)
  → Images in Science Archive (raw + processed)

now!

- Host a copy of DESI targets
  → Database for final, public set of targets

- Host DESI redshifts
  → Database for public releases of redshift catalogs
  → Tools for spectra visualization/analysis

future

- Create example Notebooks & workflows

- Users can work with all data products

Summary of Current Functions

Function	Method
Sky exploration	Image discovery tool Catalog overlay tool Catalog visualization tool (prototype)
Authentication	Web interface datalab command Python authClient, DL interface
Catalog query	**Web interface** datalab command line (CLI) **Python queryClient**, DL interface TOPCAT
Image query	Simple Image Access (SIA) service Web Cutout Tool
Query result storage	myDB Virtual storage space
File transfer	datalab command and Virtual storage space
Analysis	Jupyter notebook server

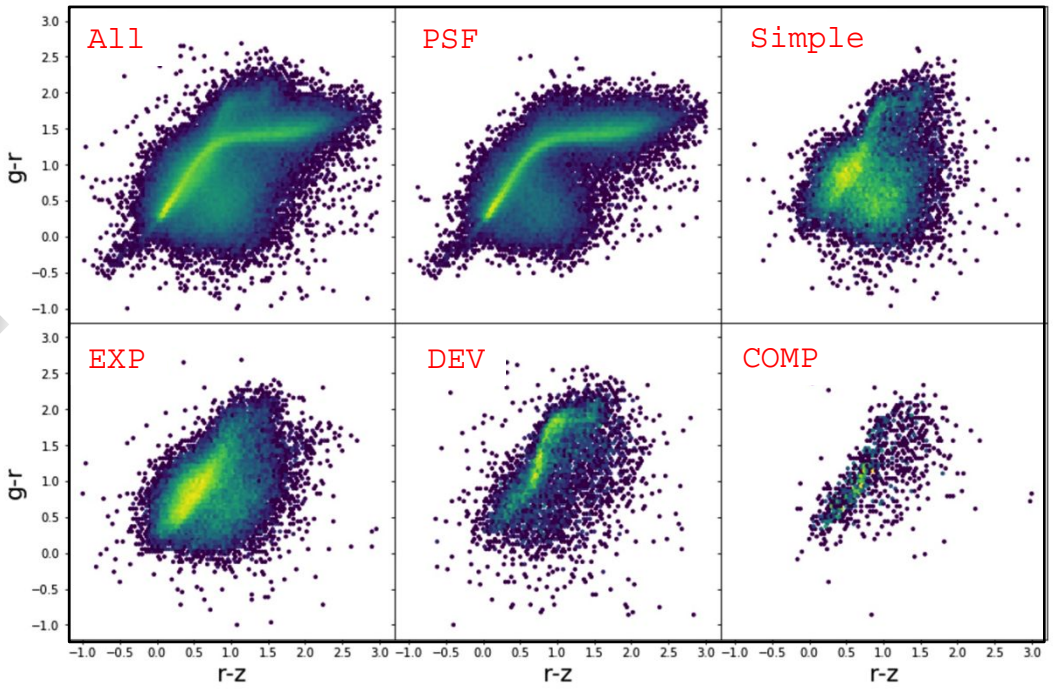
Query to database:

magnitudes and object shape (type)

```
query = """
        SELECT dered_mag_g as gmag, dered_mag_r as rmag,
               dered_mag_z as zmag,
               dered_mag_w1 as w1mag, dered_mag_w2 as w2mag, type,
               snr_g, snr_r, snr_z, ra, dec
        FROM ls_dr3.tractor_primary
        WHERE (snr_g>3 and snr_r>3 and snr_z>3)
        LIMIT 200000"""

# dered_mag_g,r,z  = AB mag in DECam g,r,z bands corrected
#                    for Galactic reddening
# dered_mag_w1,w2  = AB magnitudes in WISE bands W1 & W2
#                    corrected for Galactic reddening
# type             = object type (PSF, SIMP, EXP, DEV, COMP)
# snr_g,r,z        = signal-to-noise ratios (S/N) in g,r,z bands
# ra,dec           = celestial coordinates
#
# WHERE: requirement that S/N>3 in each DECaLS band
# LIMIT: returns 200,000 rows that satisfy the query
```

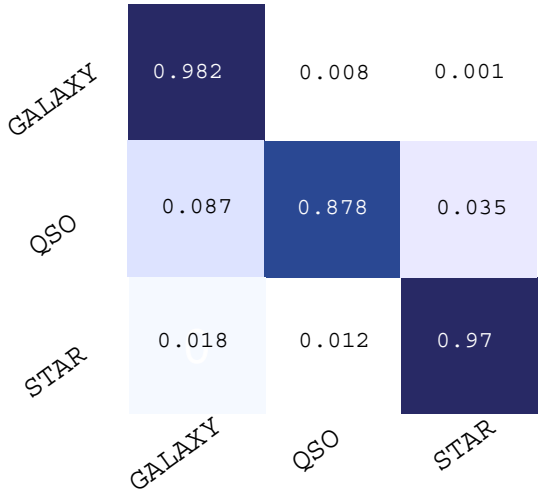
Analysis: *color-color plot per type*



Example Workflow

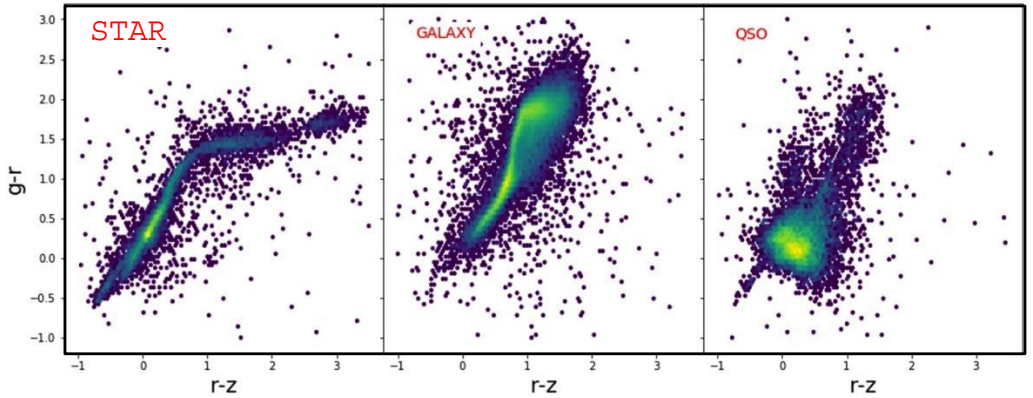
Machine-Learning:

Confusion matrix (spectroscopic training set)



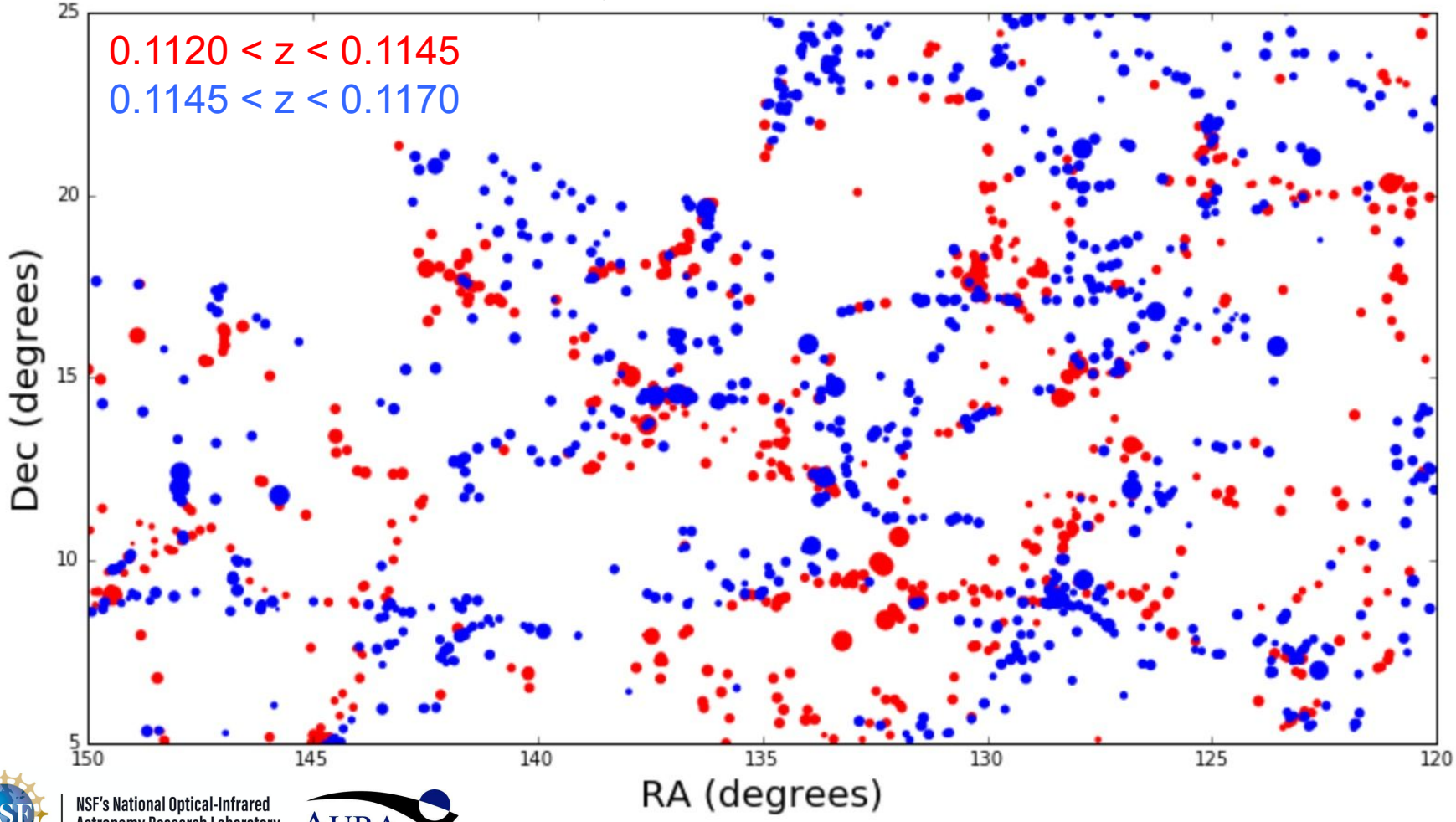
Joint query:

cross-match with SDSS spectroscopic class



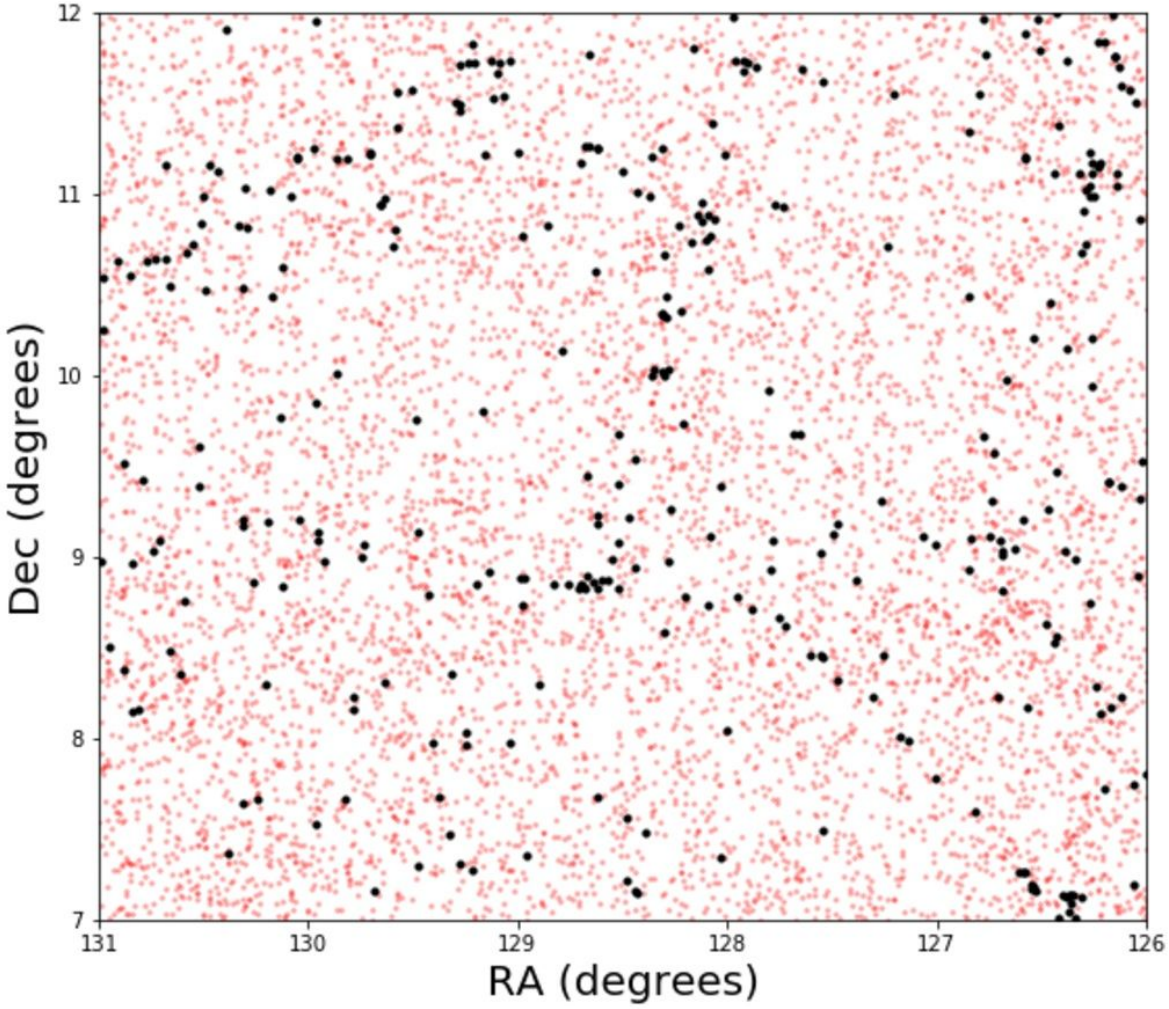
Example Application: Large Scale Structures

- Query SDSS (specObj DR13)
- Galaxies in 2 thin redshift slices (750 km/s)
- Point-size coded to velocity dispersion



Example Application: Large Scale Structures

- Joint Query SDSS (specObj DR13) – ls_dr3

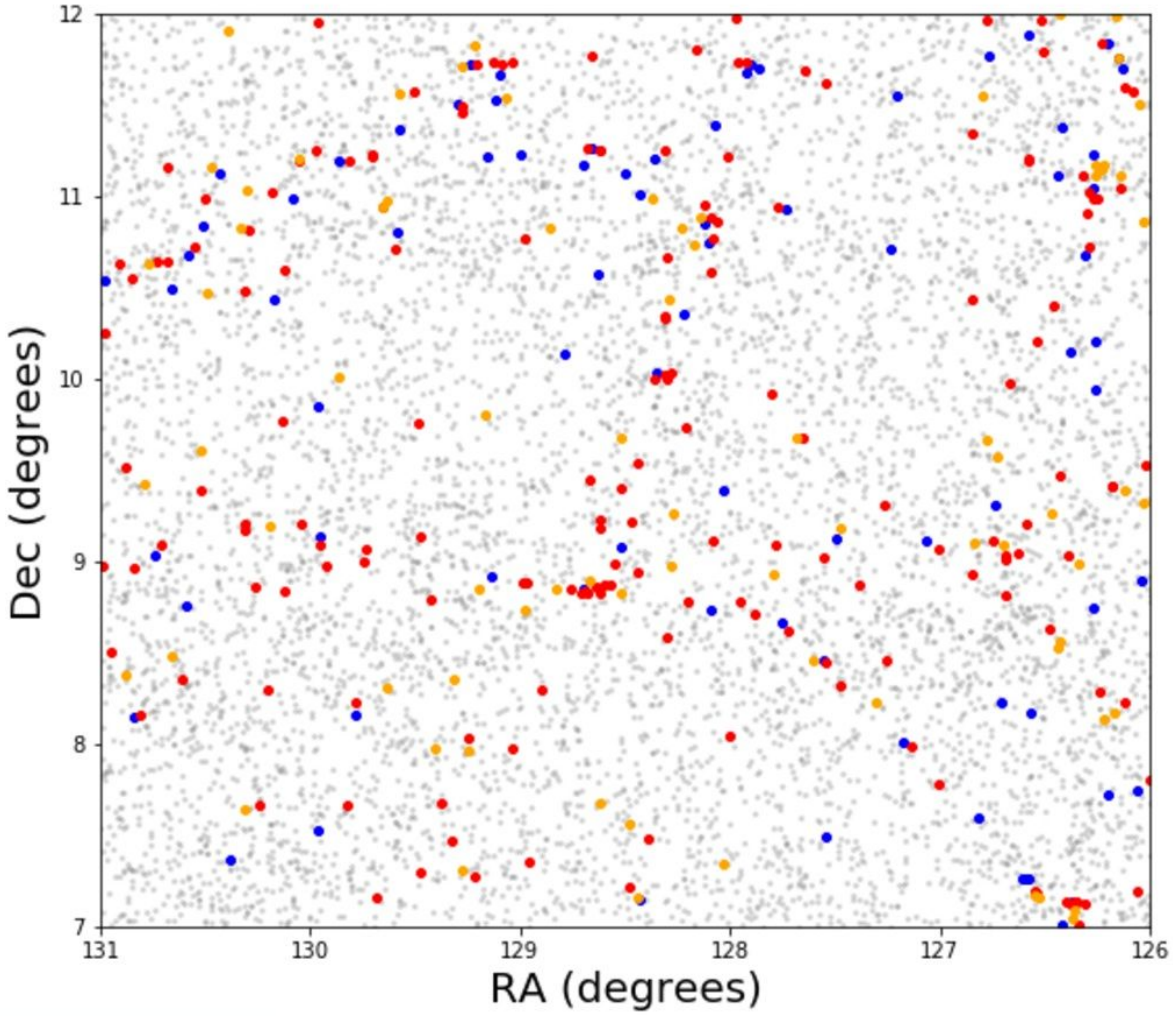


All spec-z

0.105 < z < 0.125

Example Application: Large Scale Structures

- Joint Query SDSS (specObj DR13) – ls_dr3



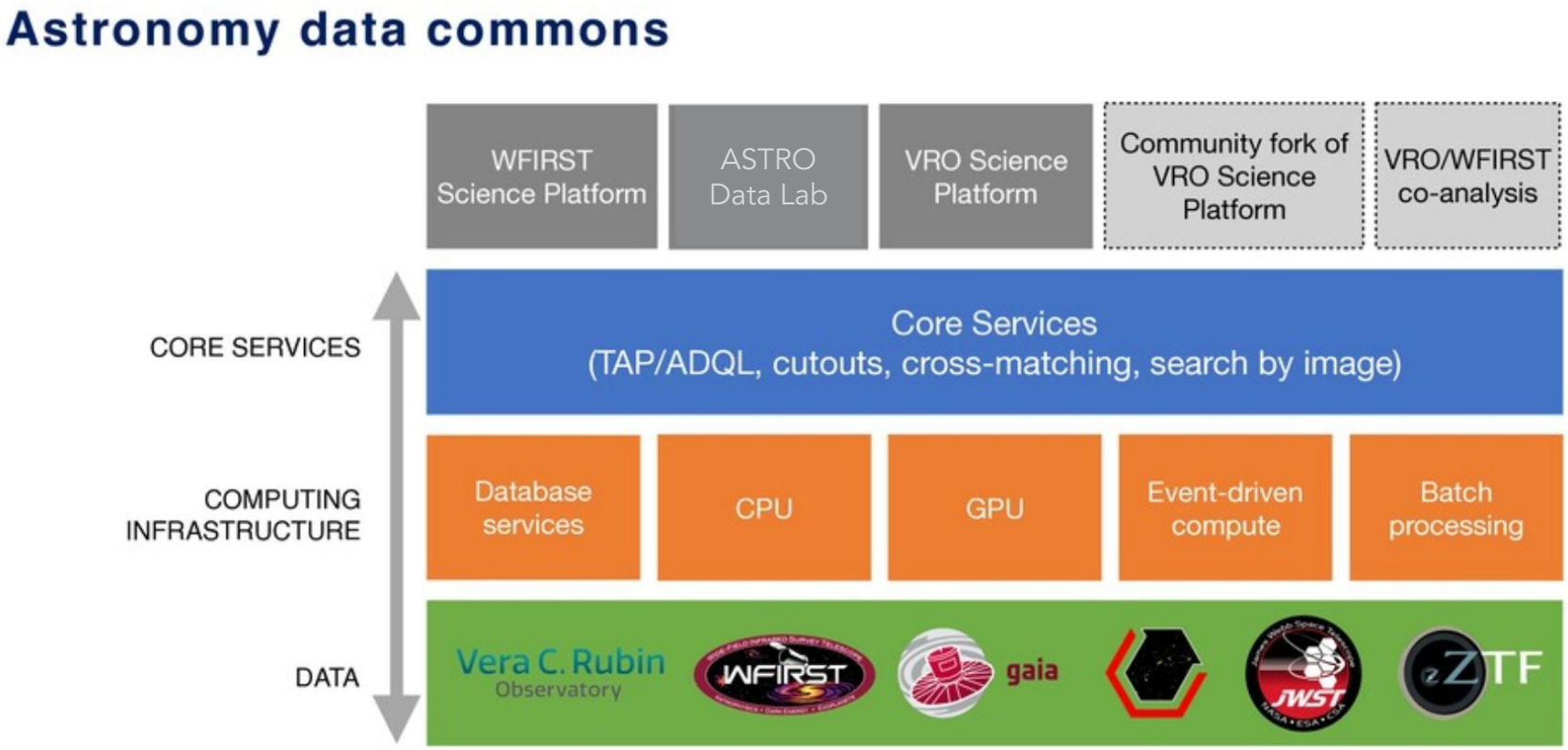
LS Type:

COMP

DEV

EXP

Future Astronomy Archives / Science Platforms



*…co-locate **data** with **cloud computing** infrastructure and commonly used **software services, tools & apps** for managing, analyzing and sharing data to create an **interoperable resource**…*

Slide by Arfon Smith (STSci)

Try it out and get in touch!

- Web: datalab.noao.edu
- Email: datalab@noao.edu
- GitHub: https://github.com/noaodatalab
- Twitter: @DataLabAstro