Astro Data Lab Science Platform

The OG Data Lab

Robert Nikutta

NOIRLab

for the Data Lab team

IVOA Interop Bologna / May 2023

datalab.noirlab.edu

ASTRO DATALAB

What is Data Lab today?

Human side:

- Community of now over 2500 users *(including many of you!)*
- Small team of S/W engineers, scientists, data analysts, sysads
  (develop, make it work, enhance datasets, engage with users a lot)

Data and data services:

- Open-access science platform for big-data astronomy
- Repository for large datasets *(catalogs/**TAP**, images, and new: spectra)*
- Website / authenticated portal; Python API library; command-line client
- Exploratory tools *(e.g., survey coverage/MOC, catalog overlay, Aladin/**HiPS**,...)*
- Services to access data *(SQL/ADQL, **TAP**, **SIA**, ...)*
- Analysis facilities *(Jupyter notebook server, common-task helpers)*; jailed terminal
- Cross-matching service
- Image cutout service
- Remote storage **(VOSpace**, *MyDB)*; 64 file services (read-only **VOSpace**s)
- Visualization *(all of Python, APIs)*

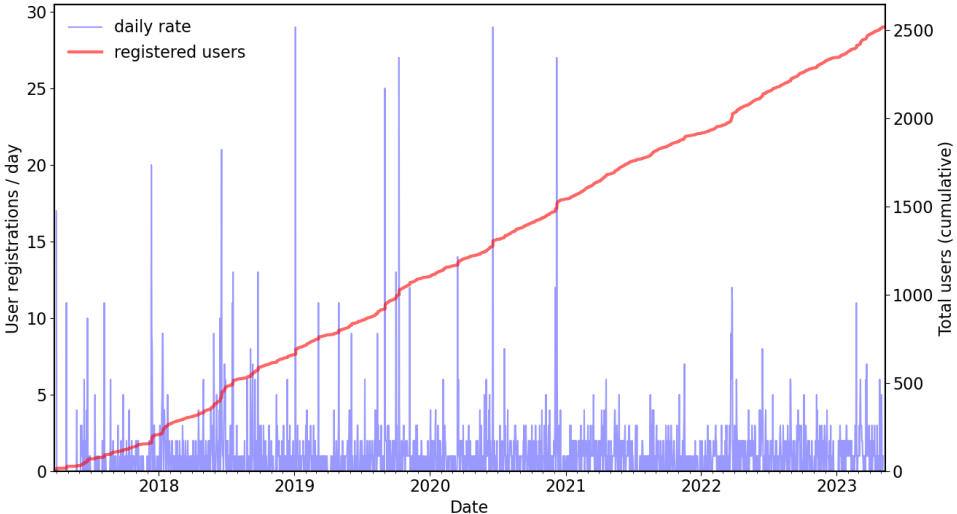
Contact: robert.nikutta@noirlab.edu

Why Data Lab?

- *Originally:* to exploit NOAO (now NOIRLab) public data; *specifically:* DES
- Bring users' analyses to the data (remote compute)
- Joint analysis with other co-located datasets (ingested by us, or uploaded by user)
- Preparation of users for PB-scale data now, readiness for LSST-era surveys
- *Today:* DES just one of 20+ major surveys hosted at Data Lab
- **New:** science platform also for large spectroscopic data (*SPARCL*; hosts SDSS, BOSS, and soon DESI spectra)

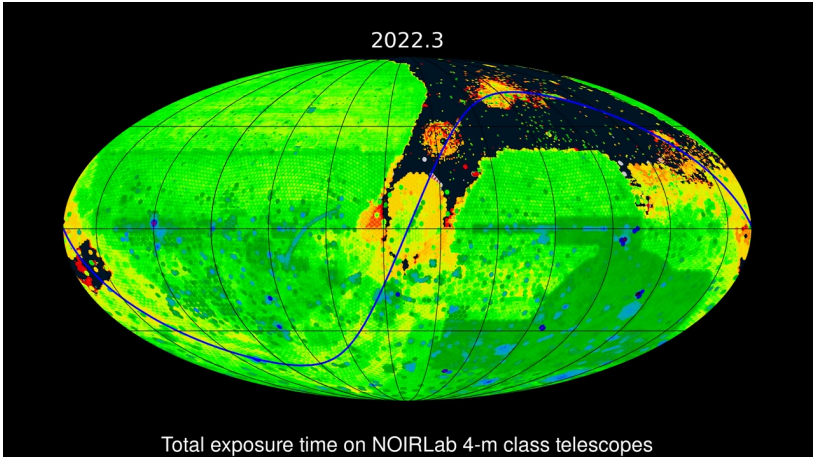
Contact: robert.nikutta@noirlab.edu

User community



Contact: robert.nikutta@noirlab.edu

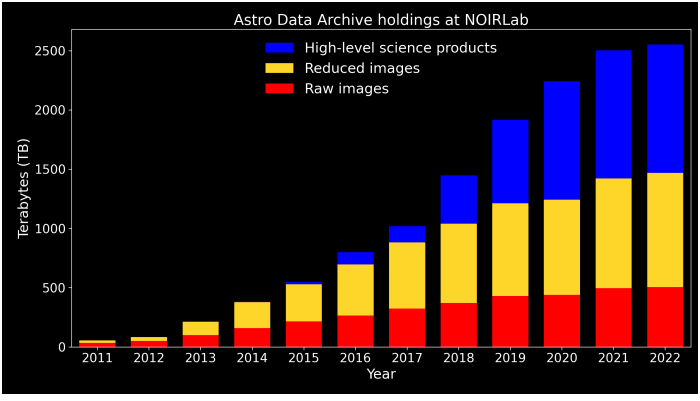
Sky coverage over time



Much of it we also combine to a single catalog.
NSC DR2 (PI: David Nidever) has 4B objects, 69B measurements, up to 1000 epochs.
DR3 will use PSF photometry and have more data.

Contact: robert.nikutta@noirlab.edu

Some numbers on data holdings



- At Data Lab: 150 TB of catalog data, 250+ billion rows, plus 93B rows in crossmatch tables, plus 2B rows in image search tables (SIA)
- Images, masks, coadds, etc.: over 3 PB (mostly at Data Archive)

Contact: robert.nikutta@noirlab.edu

Some hosted datasets

- DES DR1, DR2 (691e6 objects in DR2)
- DECaLS DR 3 - 10 (2.8e9 objects in DR10)
- DECaPS DR1 + DR2 (5.5e9 objects total)
- NOIRLab Source Catalog DR1, DR2 (3.9e9 objects, 68e9 measurements in DR2)
- Gaia DR1, DR2, EDR3, DR3 (1.8e9 objects in DR3)
- PHAT (7.5e9 objects)
- AllWISE (750e6 objects) & unWISE DR1 (2.2e9 objects) & CatWISE2020 (1.9e9 objects)
- SMASH DR1, DR2 (360e6 objects in DR2)
- SDSS DR 12, 13, 14, 16, 17 (5.1e6 objects in DR17)
- S-PLUS DR1 + DR2 (35e6 total)
- SkyMapper DR1, DR2 (500e6 objects, 5e9 measurements in DR2)
- UKIDSS DR11plus (1.2e9 objects)
- VHS DR5 (1.4e9 objects)
- DELVE DR1 + DR2 (3e9 objects total)
- Various Gemini LLP HLSPs
- SDSS+BOSS spectra; soon DESI
- LSST simulated catalogs DR2, MW + LMC + SMC (11.4e9 objects)
- Buzzard DR1 simulated catalog (3.1B objects, 20+ filters)

Temporal information in: NSC, SMASH, ...

Contact: robert.nikutta@noirlab.edu

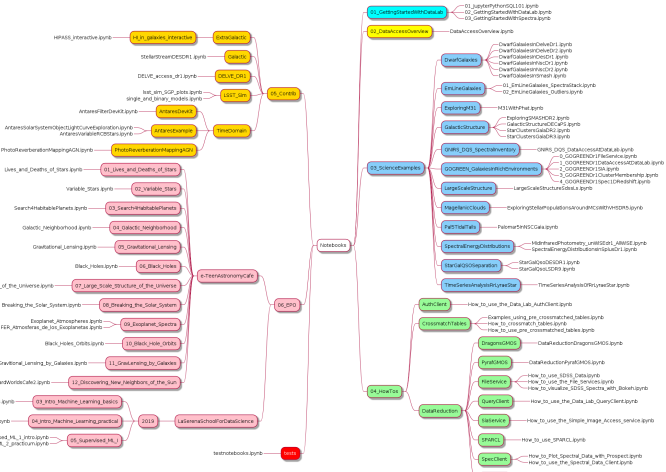
Near/mid-term future datasets

- DESI spectra (Dark Energy Spectroscopic Instrument); EDR in just weeks
  → *40M galaxies with precise redshifts*
- LS DR11
  → *Deeper and wider*
- DELVE DR3 & DR4
  → *Stacks and catalogs*
- unTimely DR2 (single-epoch WISE/NEOWISE catalog)
  → *Full lightcurves from all of WISE/NEOWISE*
- More Gemini LLP high-level science products
- Your data?
  → *HLSP data publication service*

Contact: robert.nikutta@noirlab.edu

Curated default notebooks

https://github.com/astro-datalab/notebooks-latest/



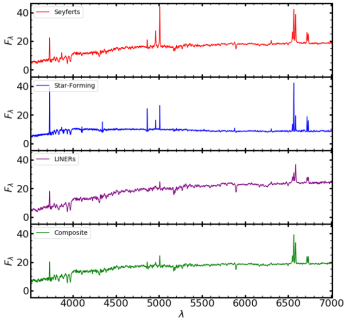
Notebooks on

- Basic intros
- Data access
- Technical How-Tos
- Complete science examples
- ANTARES filter dev kit
- EPO
- Community-contributed
- Tests
- ...

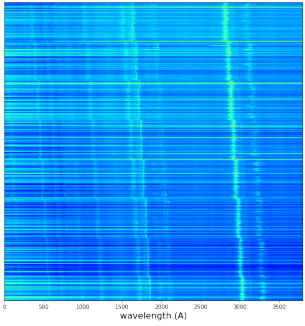
Contact: robert.nikutta@noirlab.edu

New: Spectroscopy services

- Data Lab now has a fast spectroscopic service in production: SPARCL
  https://datalab.noirlab.edu/sparcl/
- Serving SDSS + BOSS DR16
- Data discovery by properties
- Fast retrieval of thousands of spectra (and ivar, models, etc.)
- Scalability to DESI tested already, DESI EDR public in about 2 weeks
- Not yet VO-compliant **Also: Can we form an IVOA Spectroscopy IG?**



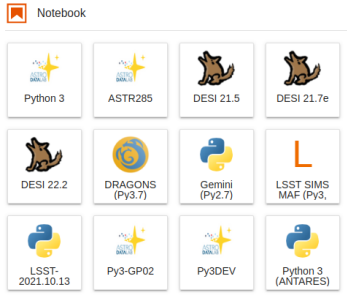
SPARCL



Contact: robert.nikutta@noirlab.edu

Updates on the backend

- Finally(!), unit testing (framework with auto-templates, etc.)
- Containerization of services
  - started with NB server
  - now entire system (testing and development envs first, not yet prod)
- DB machines SSD expansion (double), probably good for 3 years
- Custom kernels for data analysis with Gemini IRAF and DRAGONS



Contact: robert.nikutta@noirlab.edu

Data Lab in the IVOA context (Possible trigger alert.. Outside POV)

What works well for us?

TAP wonderful, SIA nice, VOSpace (with some work), UCDs super useful → **Thank you!!**

What is too hard / complex?

- Registering services is too complex
  IVOA doesn't do tooling, but tooling out there insufficient / too complex. Official options:
  - "Use web browser": go to NAVO **(defunct?)**, or to Euro-VO **(nice but complex UI)**, watch videos, "create VOResource" **(by hand?)**, validate
  - "Use purx" → **What?**
  - "Run your own publishing registry" → **How?**
    Some publication software has code to do that built in **(Which S/W?)***; if you run something else, ask around on the registry mailing list* **(That's not good documentation)***, as multiple data centres have already implemented at least the OAI-PMH part* **(Sounds like no full implementation?)***. Even when running off-the-shelve software, at least skim the underlying VO standard, Registry Interfaces, to have an idea of how all this is supposed to work together.* **(I came to 'How to get services into the Registry', but they send me elsewhere now..)**

Contact: robert.nikutta@noirlab.edu

Continued..

- Some formats are too verbose (e.g. XML for VOTable) → Efficient standard table formats available in the industry (e.g. Parquet); **Make them usable in the VO context?**

What feels not useful (controversy alert!)

- For developers new to astronomy the IVOA can be an opaque ecosystem (can't see the forest for all the trees)
- Feels like it could need severe pruning
  E.g.: *"StandardRegExt-1.1: an extension of the VOResource XML schema to register standard documents produced by the IVOA TCG"*

What is needed?

- A **much** easier onboarding process/program for newcomers (scis and engineers)
  - Not just a quick "How to use a few VO-compliant tools", but also: IVOA vision, organization, modes of engagement, rules, state of the IVOA around the world
  - And then technical: for scientists, for developers, for managers(!)
- More... next time!

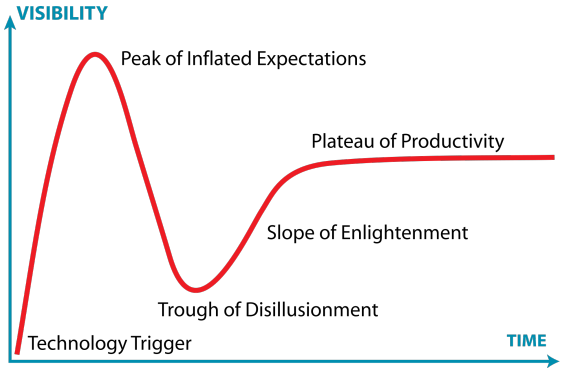
Contact: robert.nikutta@noirlab.edu

Data Lab in the near to mid-term future

- Registering our services in the Registry(!)
- Going gradually to the cloud
  (better on-demand scalability, easier dev, deployment & maintainability)
- HLSP publication facility *(BYOD)*
- Dynamic user-administered groups/projects
- Batch compute capabilities *(job manager)*

Contact: robert.nikutta@noirlab.edu

Re: "AI" and chatGPT



Newsletters *The Atlantic*

IDEAS

AI Is a Waste of Time

The newest AI tools are accelerating basic research and scaring the general public. But many people are simply using them as toys.

By Derek Thompson

Apr 28, 2023

Contact: robert.nikutta@noirlab.edu

Write a query, chatGPT

Prompt (after some engineering): *We're writing an SQL query against astronomical data. First,* **select 1000 rows** *from a table called des_dr2.main within 1 degree of right ascension 60 and declination -30. Then* **crossmatch** *the resulting table with another table, called gaia_source in the gaia_dr3 schema. Crossmatch them positionally using right ascension and declination with a matching radius of 0.1 degrees, and* **keep only the nearest match**. **Use Q3C** *for Postgres.* **Return all columns from the gaia table**, *and only* **the mag auto columns for all bands from des**.

Output:

```
SELECT gaia.*, des.mag_auto_g, des.mag_auto_r, des.mag_auto_i, des.mag_auto_z, des.mag_auto_y
FROM (SELECT * FROM des_dr2.main WHERE q3c_radial_query(ra, dec, 60, -20, 1) LIMIT 1000) AS des
INNER JOIN gaia_dr3.gaia_source AS gaia
ON q3c_join(gaia.ra, gaia.dec, des.ra, des.dec, 0.1)
ORDER BY q3c_dist(gaia.ra, gaia.dec, des.ra, des.dec);
```

Next day, with same prompt:

```
SELECT *
FROM gaia_dr3.gaia_source AS gaia
INNER JOIN Q3C_JOIN(des_dr2.main AS des, gaia.ra, gaia.dec, 0.1)
ON des.id = gaia.nearest_match
WHERE Q3C_RADIUS(des.ra, des.dec, 60, -30) <= 1
AND des.nmag_auto > 0
```

Contact: robert.nikutta@noirlab.edu

Contact the Data Lab team

datalab@noirlab.edu

datalab.noirlab.edu

github.com/astro-datalab

@DataLabAstro

Thank you!

Hacking together on great science with Data Lab

Contact: robert.nikutta@noirlab.edu