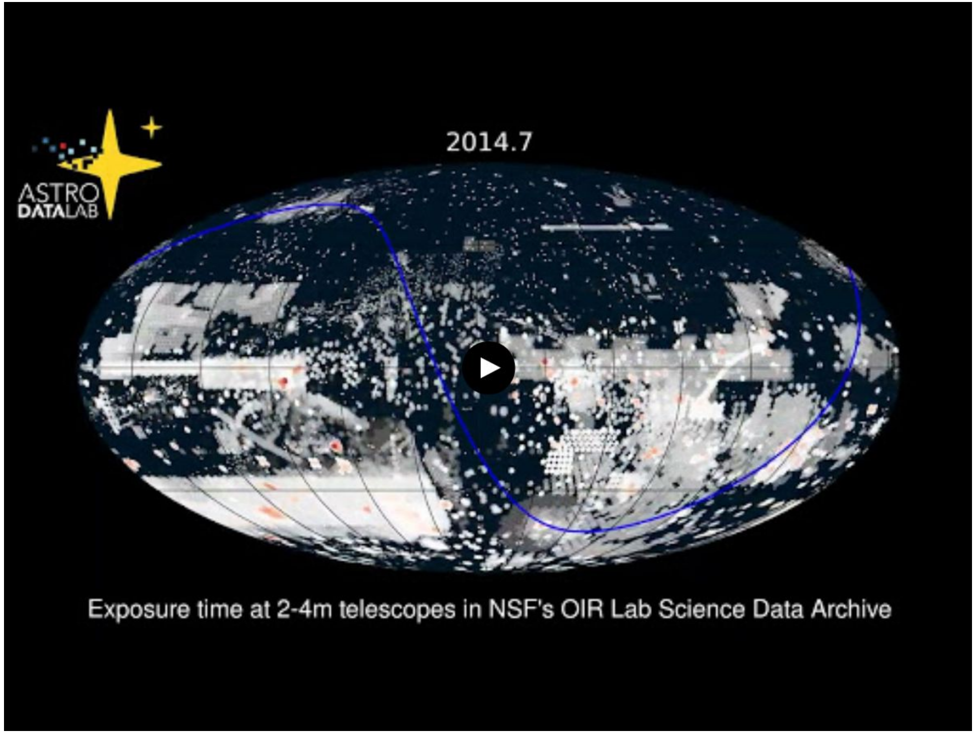
Astro Data Lab

An Open-Access and Open-Data Science Platform

Robert Nikutta & Data Lab Team
NOIRLab
robert.nikutta@noirlab.edu

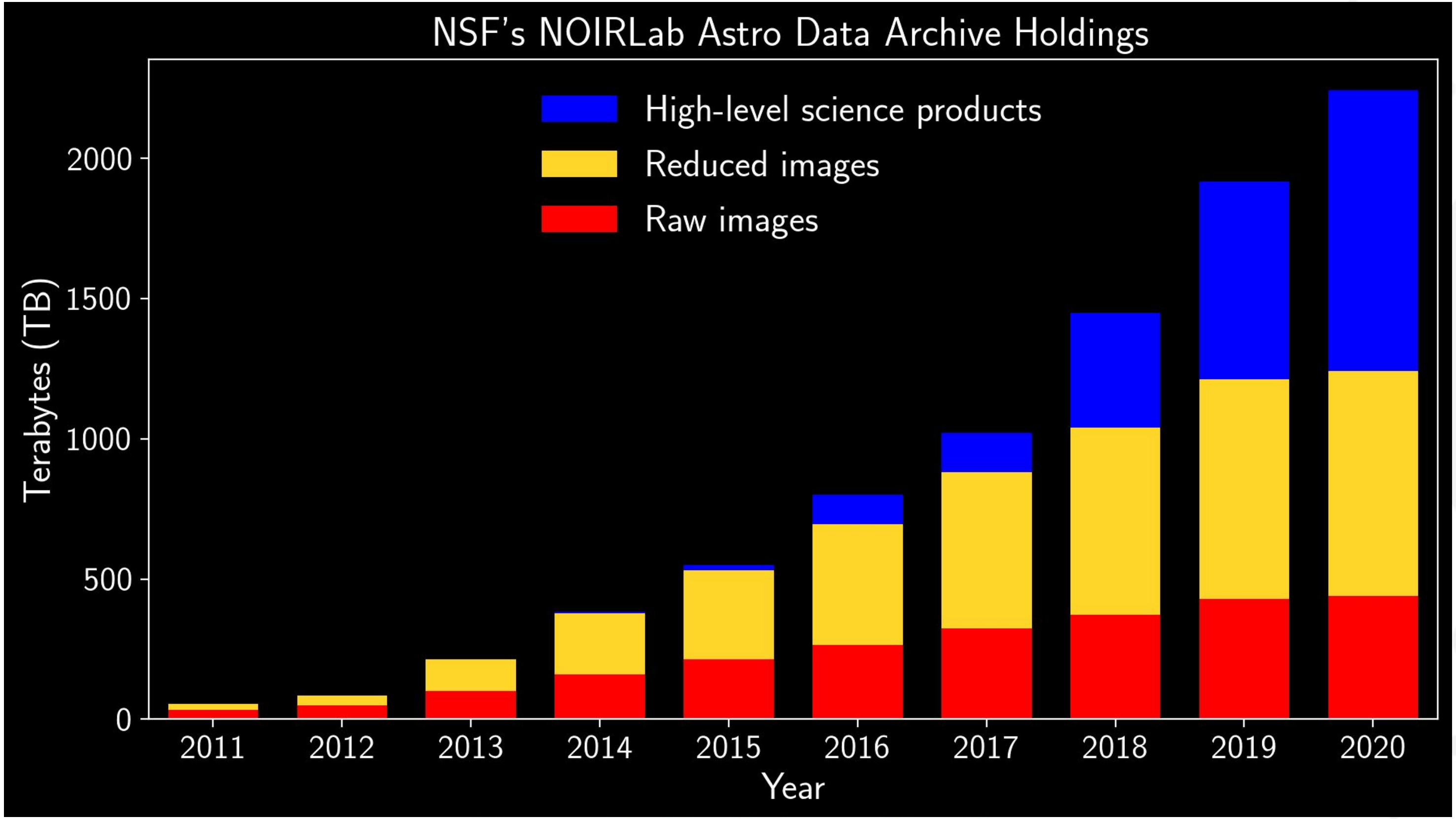
Webinar @ LIneA, June 10, 2021

Motivation - The Data Avalanche


ASTRO DATALAB
2014.7
Exposure time at 2-4m telescopes in NSF's OIR Lab Science Data Archive

https://youtu.be/IbRWdOqWrEk

Data growth



Discovering Our Universe Together

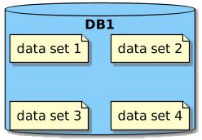
From data set to Science Platform

1

Data Set

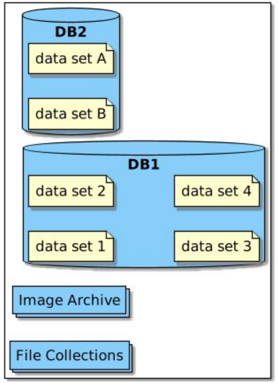
2

Database



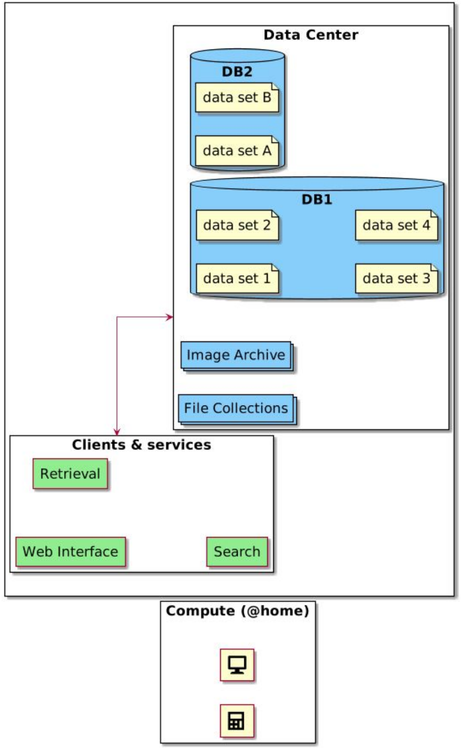
3

Data Center



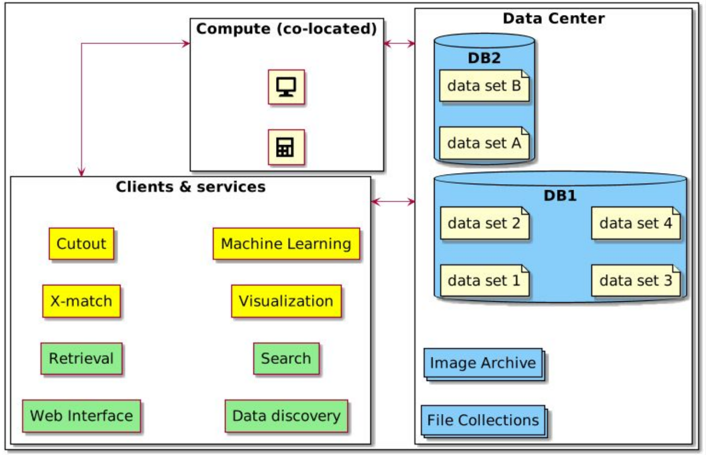
4

Science Archive



5

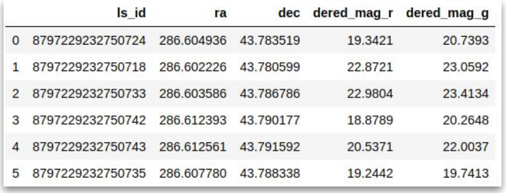
Science Platform



Data types at Data Lab

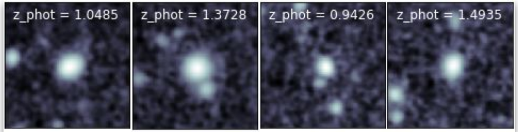
Currently

Catalogs
(2D tables)

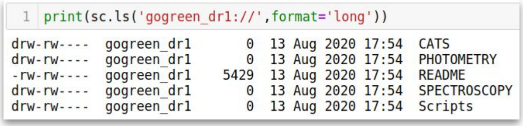


	ls_id	ra	dec	dered_mag_r	dered_mag_g
0	8797229232750724	286.604936	43.783519	19.3421	20.7393
1	8797229232750718	286.602226	43.780599	22.8721	23.0592
2	8797229232750733	286.603586	43.786786	22.9804	23.4134
3	8797229232750742	286.612393	43.790177	18.8789	20.2648
4	8797229232750743	286.612561	43.791592	20.5371	22.0037
5	8797229232750735	286.607780	43.788338	19.2442	19.7413

Images
(2D arrays)



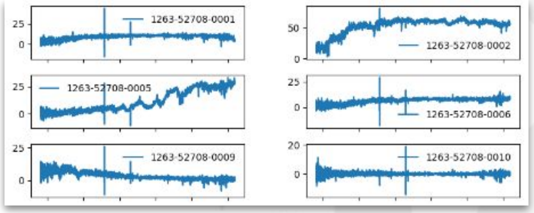
Heterogeneous
data collections
(file service)



```
print(sc.ls('gogreen_dr1://',format='long'))
drw-rw----   gogreen_dr1       0  13 Aug 2020 17:54  CATS
drw-rw----   gogreen_dr1       0  13 Aug 2020 17:54  PHOTOMETRY
-rw-rw----   gogreen_dr1    5429  13 Aug 2020 17:54  README
drw-rw----   gogreen_dr1       0  13 Aug 2020 17:54  SPECTROSCOPY
drw-rw----   gogreen_dr1       0  13 Aug 2020 17:54  Scripts
```

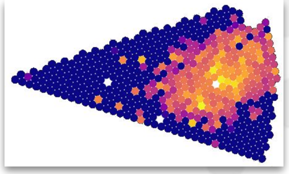
Currently

1D spectra
(queryable)



Soon

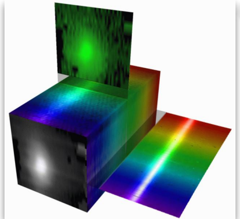
DESI



Future

2D spectra
IFU cubes &
complex data

7500
8000
8500
9000
Observed frame wavelength (A)



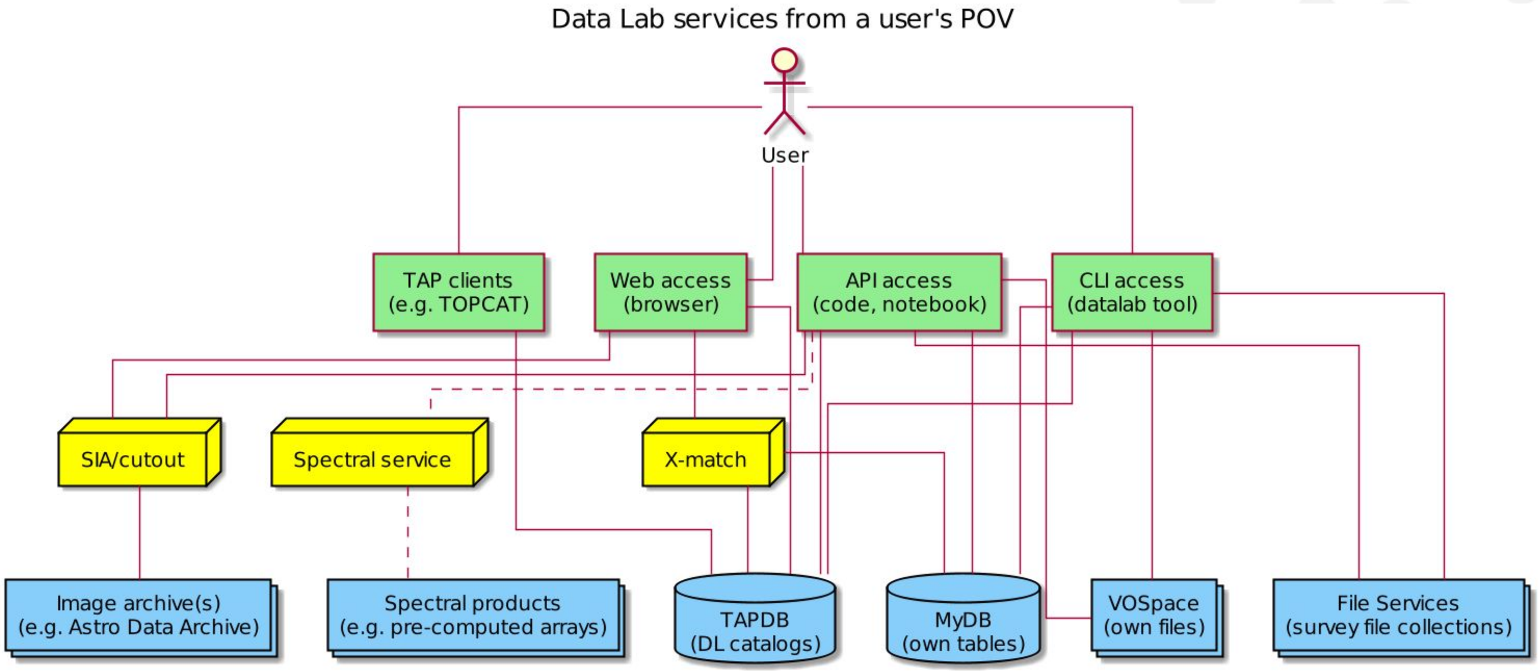
GMOS-IFU,
MaNGA,
US-ELTs,
...

Scientific services at Data Lab

Use these front-ends...

...to interface these services...

…to/or access these data


Data Lab services from a user's POV
User
TAP clients (e.g. TOPCAT)
Web access (browser)
API access (code, notebook)
CLI access (datalab tool)
SIA/cutout
Spectral service
X-match
Image archive(s) (e.g. Astro Data Archive)
Spectral products (e.g. pre-computed arrays)
TAPDB (DL catalogs)
MyDB (own tables)
VOSpace (own files)
File Services (survey file collections)

High-level goals

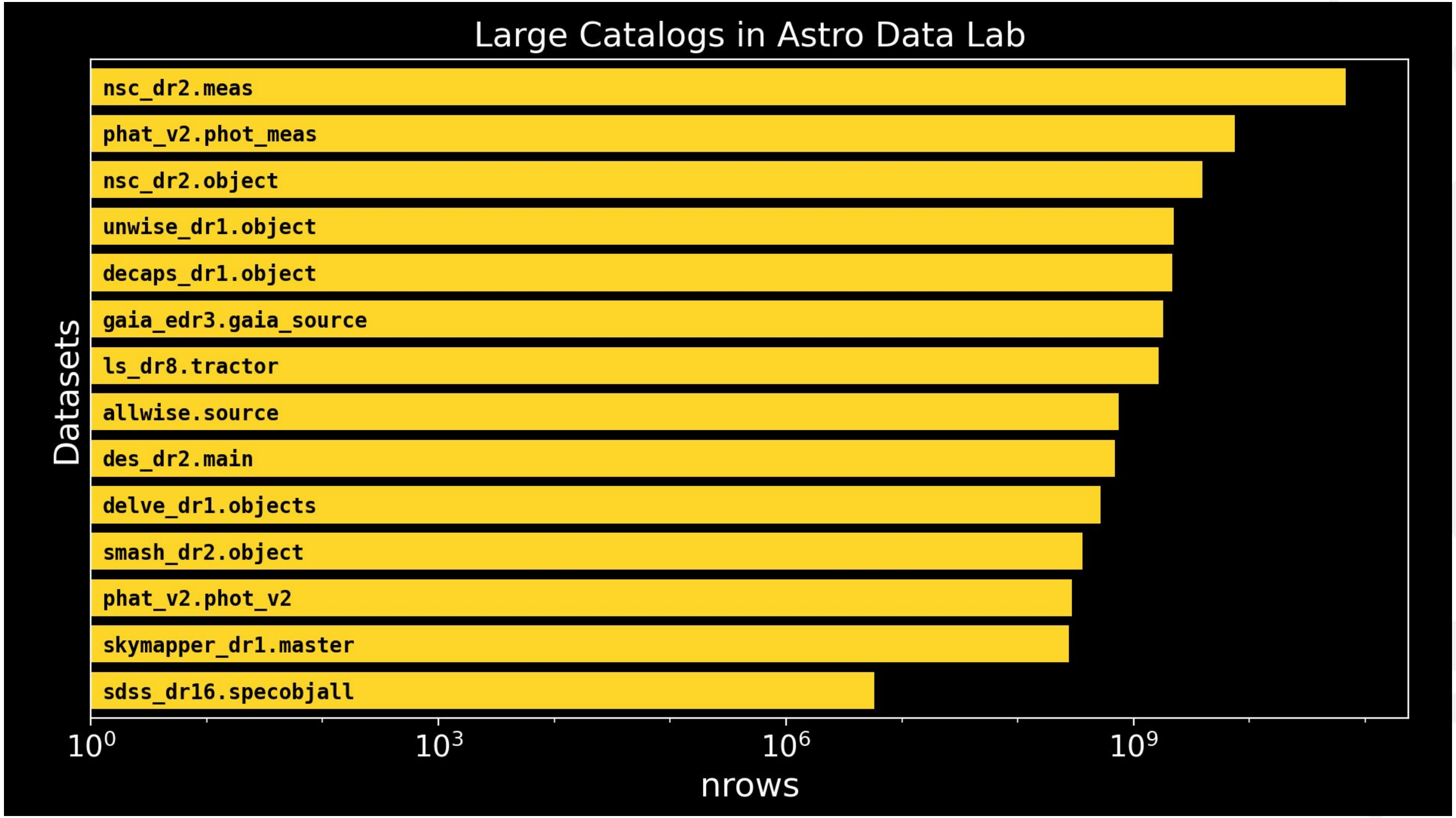
- Enable easy exploration of very large data holdings
  → catalogs, pixels, spectra, survey file collections…

- Connect the various data products, joint analysis
  → e.g. find interesting objects in catalogs, now find good images of them

- Enable remote analysis
  → Bring your code & algorithms & your data to the Big Data;
  execute code on our servers; analyze; visualize; publish

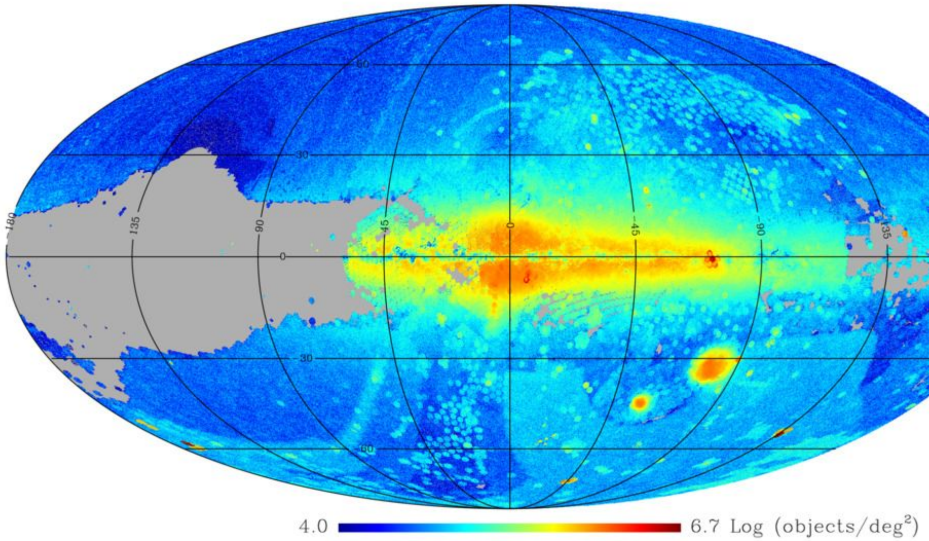
- Enable easy user collaboration
  → sharing of query results, data sets, notebooks, group databases & storage

Some large tables at Data Lab



Discovering Our Universe Together

NOIRLab Source Catalog (NSC)



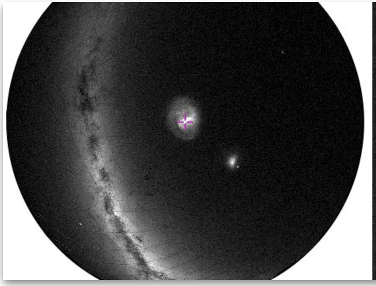
Nidever+2021 (AJ)

- 35,273 sq deg
- u,g,r,i,z,U,VR bands
- 412,116 exposures
- 3.9 billion objects
- 69 billion measurements
- 100s of epochs in some regions
- 21-23.6 mag median depth
- 0.99-1.35 arcsec median seeing
- Photometric calibration accuracy 1-2%
- Astrometric calibration accuracy 2 mas

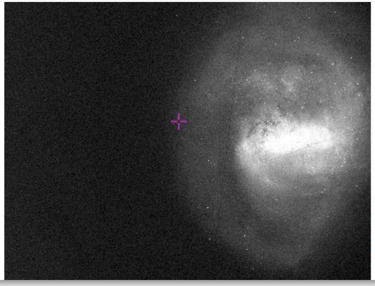
- DR3 with PSF photometry mid-2022

(Visual) data exploration

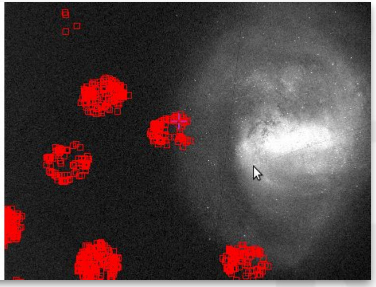
Web survey viewer
(based on Aladin Lite)



Survey footprint navigation



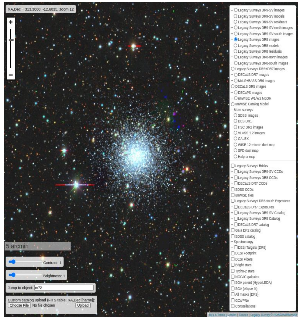
Pan & Zoom
(here on the LMC)

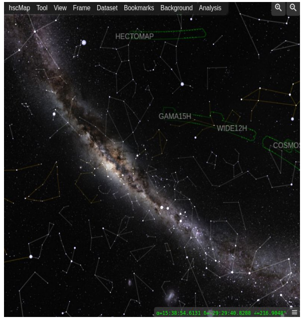


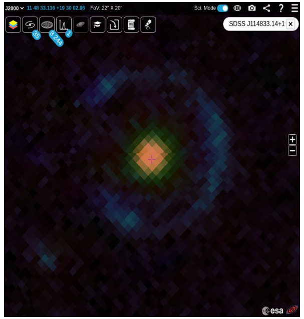
Catalog overlays
(here SMASH fields)

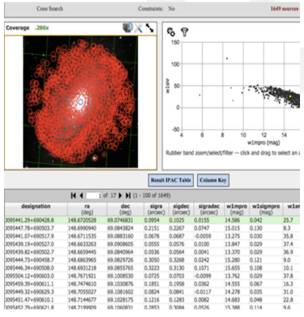
Also looking into other viewers, e.g.

- *Legacy Sky Viewer*
- *hscMap*
- *ESA Sky*
- *Firefly*
- *WWT*
- *...*





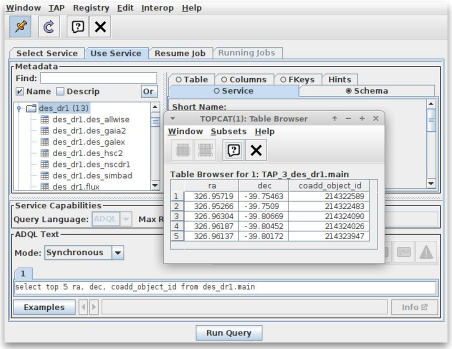




Getting to catalog data (1)

- SQL-like queries via TAP→ PostgreSQL and ADQL
- Both sync and async queries→ Submit & wait / Submit & check later
- Can query both DL catalog holdings and user’s MyDB
- Clients:

TAP-aware (e.g. TOPCAT)



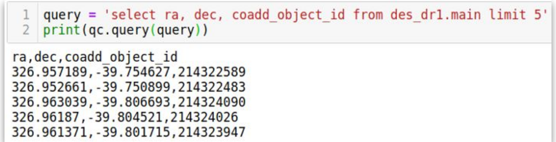
Mode: Synchronous

```
select top 5 ra, dec, coadd_object_id from des_dr1.main
```

Table Browser for 1: TAP_3_des_dr1.main

	ra	dec	coadd_object_id
1	326.95719	-39.75463	214322589
2	326.95266	-39.7509	214322483
3	326.96304	-39.80669	214324090
4	326.96187	-39.80452	214324026
5	326.96137	-39.80172	214323947

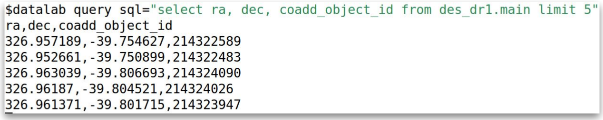
queryClient.py (notebooks, scripts)



```
query = 'select ra, dec, coadd_object_id from des_dr1.main limit 5'
print(qc.query(query))
```

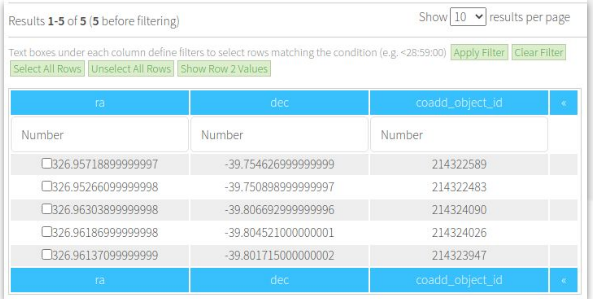
```
ra,dec,coadd_object_id
326.957189,-39.754627,214322589
326.952661,-39.750899,214322483
326.963039,-39.806693,214324090
326.96187,-39.804521,214324026
326.961371,-39.801715,214323947
```

datalab CLI (on local computer)



```
$datalab query sql="select ra, dec, coadd_object_id from des_dr1.main limit 5"
ra,dec,coadd_object_id
326.957189,-39.754627,214322589
326.952661,-39.750899,214322483
326.963039,-39.806693,214324090
326.96187,-39.804521,214324026
326.961371,-39.801715,214323947
```

Query from DL website



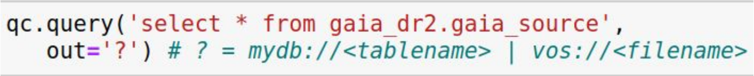
Results **1-5** of **5** (**5** before filtering)

Show 10 results per page

Text boxes under each column define filters to select rows matching the condition (e.g. <28:59:00)

ra	dec	coadd_object_id	«
Number	Number	Number	
□326.95718899999997	-39.754626999999999	214322589	
□326.95266099999998	-39.750898999999997	214322483	
□326.96303899999998	-39.806692999999996	214324090	
□326.96186999999998	-39.804521000000001	214324026	
□326.96137099999999	-39.801715000000002	214323947	
ra	dec	coadd_object_id	«

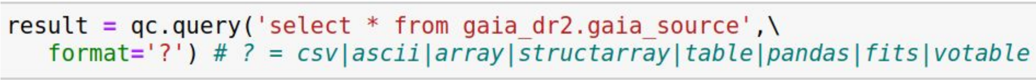
Getting to catalog data (2)



```
qc.query('select * from gaia_dr2.gaia_source',
   out='?') # ? = mydb://<tablename> | vos://<filename>
```

- Query results can be either:
  - Streamed back to client → *Convert yourself, according to your needs*
  - Loaded as table straight to user's MyDB → *Great for subsequent x-matching*
  - Saved as file to user's VOSpace → *Great for sharing as a file with others*

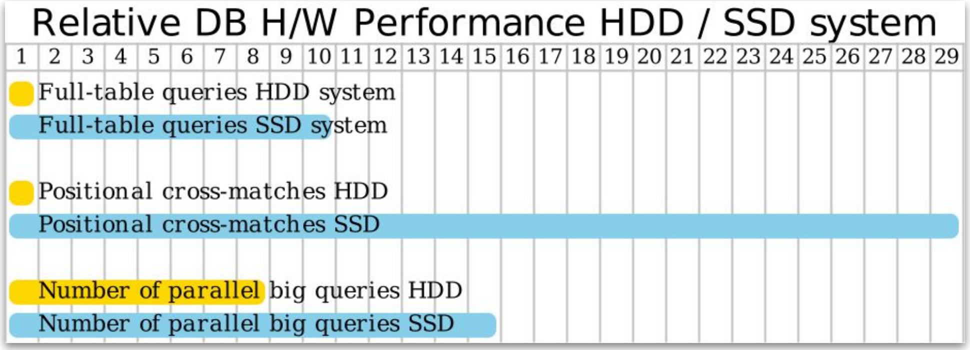
- Output can be streamed back in various formats, e.g.
  - CSV stream
  - Pandas data frame
  - VOTable, AstropyTable
  - Numpy array, record array, ...



```
result = qc.query('select * from gaia_dr2.gaia_source',\
   format='?') # ? = csv|ascii|array|structarray|table|pandas|fits|votable
```

Getting to catalog data (3)

- Query performance is key
- Optimized PostgresSQL (config), optimized tables (index cols, dtype stacking)
- Fast H/W is paramount (throughput)
- In summer 2019 we switched to SSD-based systems



We are now purely CPU-bounded up to ~15 sustained large queries running in parallel; a good place to be!

- Expanding SSD storage again now (to ~150 TB, RAID-6)

Connecting catalogs with catalogs:

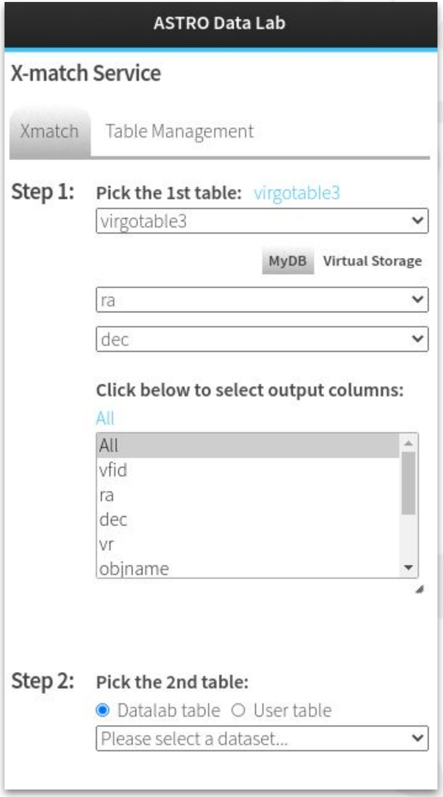
Cross-matching

Python API
(e.g. in Jupyter notebooks)

Positional cross-matching web tool
(uses same API)

On the backend: (Quad Tree Cube, Q3C)
Very fast! *(Koposov & Bartunov 2006)*

Tens of millions of rows in user table
are no problem



Step 1: Pick the 1st table: virgotable3

virgotable3

ra

dec

Click below to select output columns:

All

All
vfid
ra
dec
vr
objname

Step 2: Pick the 2nd table:

● Datalab table ○ User table

Please select a dataset...

Connecting catalogs with catalogs:

Cross-matching / Pre-computed x-match tables

- Originally: on a what-makes-sense-per-survey basis
- Now working toward an (almost) automated mechanism:

  - For each new data set, x-match against all reference sets

  Astrometry: latest Gaia `gaia_edr3.gaia_source, gaia_dr2.gaia_source`

  Photometry: latest NSC `nsc_dr2.object, unwise_dr1.object`

  Spectroscopy: latest `sdss_drN.specobjall (currently N=16)`

  - Default matching radius (1.5”), single nearest neighbor, no empty rows

  - Match table has few columns: `id1,ra1,dec1,id2,ra2,dec2,angular distance`

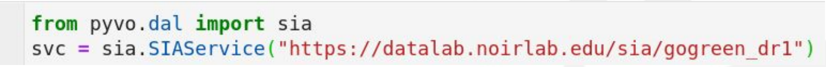
  - Re-compute when reference data sets are updated

Connecting catalogs with pixels:

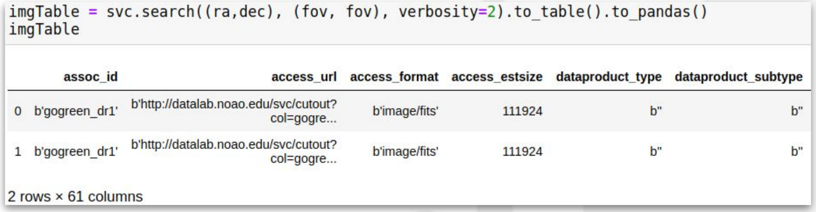
SIA service and image cutout

Simple Image Access:

- Query metadata DB about images that contain RA/Dec position
- Constraints are possible (i.e. exposure time, product type, ...)
- **`access_url`** field has link to FITS file
- **fov** argument used to compute a (usually) square cutout
- We query own image holdings & those at the NOIRLab *Astro Data Archive*



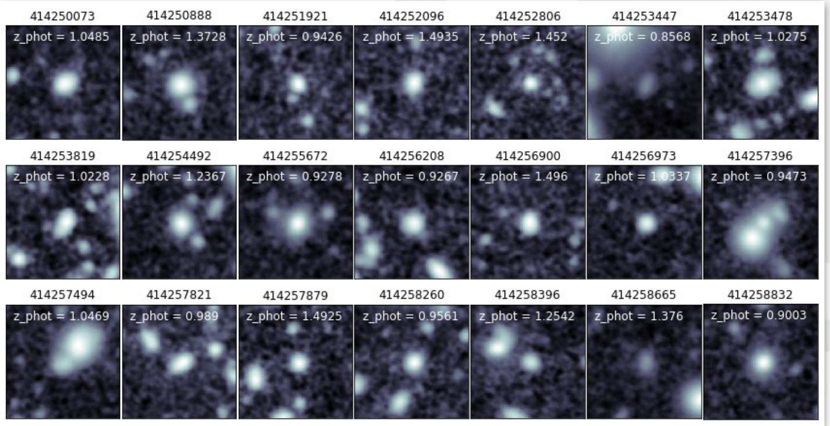
```
from pyvo.dal import sia
svc = sia.SIAService("https://datalab.noirlab.edu/sia/gogreen_dr1")
```



```
imgTable = svc.search((ra,dec), (fov, fov), verbosity=2).to_table().to_pandas()
imgTable
```

	assoc_id	access_url	access_format	access_estsize	dataproduct_type	dataproduct_subtype
0	b'gogreen_dr1'	b'http://datalab.noao.edu/svc/cutout?col=gogre...	b'image/fits'	111924	b''	b''
1	b'gogreen_dr1'	b'http://datalab.noao.edu/svc/cutout?col=gogre...	b'image/fits'	111924	b''	b''

2 rows × 61 columns



From the recent GOGREEN-Gemini-LLP & Data Lab joint DR1

User data storage services

Co-location of user data and DL holdings

User file storage: VOSpace

- 1 TB / user (soft quota)
- read/write, access via *`storeClient.py`* and *`datalab`* CLI
- *`public/`* subdirectory to share files with other users
- read-only linked in user's Jupyter notebook space

User database: MyDB

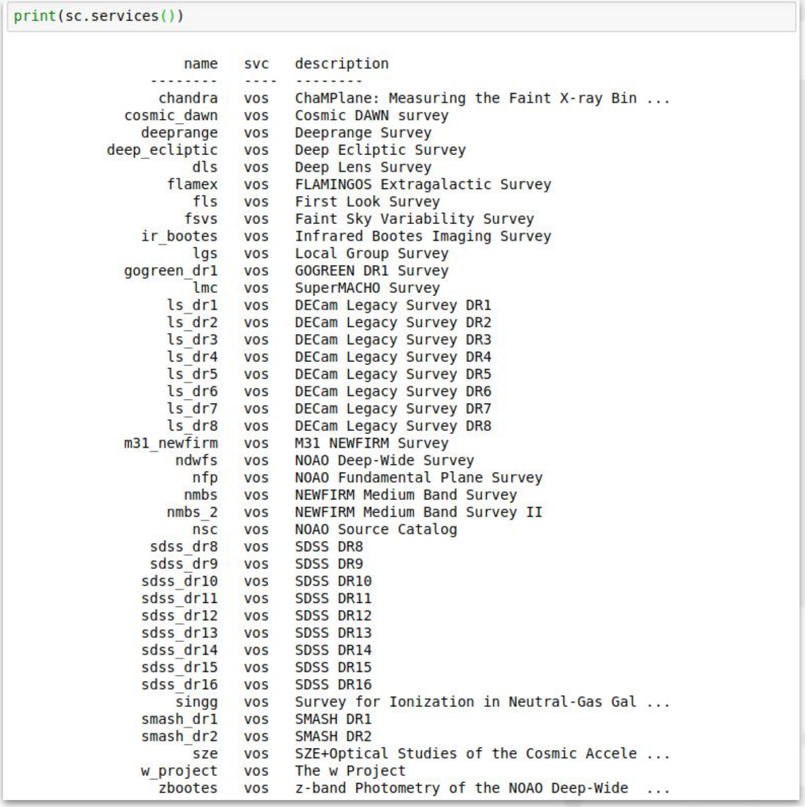
- 250 GB / user (soft quota)
- read/write, access via *`queryClient.py`* and *`datalab`* CLI
- also used for very fast positional cross-matching

File services: public VOSpace

Public (read-only) file services to serve heterogeneous survey file collections, e.g.

- Arbitrary directory structure
- Weight masks, images, catalog files
- Documentation files
- “Aux” files… anything goes

Access through *`storeClient.py`* and *`datalab`* CLI



```
print(sc.services())
```

```
         name  svc   description
     --------  ----  --------
      chandra  vos   ChaMPlane: Measuring the Faint X-ray Bin ...
  cosmic_dawn  vos   Cosmic DAWN survey
    deeprange  vos   Deeprange Survey
deep_ecliptic  vos   Deep Ecliptic Survey
          dls  vos   Deep Lens Survey
       flamex  vos   FLAMINGOS Extragalactic Survey
          fls  vos   First Look Survey
         fsvs  vos   Faint Sky Variability Survey
    ir_bootes  vos   Infrared Bootes Imaging Survey
          lgs  vos   Local Group Survey
  gogreen_dr1  vos   GOGREEN DR1 Survey
          lmc  vos   SuperMACHO Survey
       ls_dr1  vos   DECam Legacy Survey DR1
       ls_dr2  vos   DECam Legacy Survey DR2
       ls_dr3  vos   DECam Legacy Survey DR3
       ls_dr4  vos   DECam Legacy Survey DR4
       ls_dr5  vos   DECam Legacy Survey DR5
       ls_dr6  vos   DECam Legacy Survey DR6
       ls_dr7  vos   DECam Legacy Survey DR7
       ls_dr8  vos   DECam Legacy Survey DR8
  m31_newfirm  vos   M31 NEWFIRM Survey
        ndwfs  vos   NOAO Deep-Wide Survey
          nfp  vos   NOAO Fundamental Plane Survey
         nmbs  vos   NEWFIRM Medium Band Survey
       nmbs_2  vos   NEWFIRM Medium Band Survey II
          nsc  vos   NOAO Source Catalog
     sdss_dr8  vos   SDSS DR8
     sdss_dr9  vos   SDSS DR9
    sdss_dr10  vos   SDSS DR10
    sdss_dr11  vos   SDSS DR11
    sdss_dr12  vos   SDSS DR12
    sdss_dr13  vos   SDSS DR13
    sdss_dr14  vos   SDSS DR14
    sdss_dr15  vos   SDSS DR15
    sdss_dr16  vos   SDSS DR16
        singg  vos   Survey for Ionization in Neutral-Gas Gal ...
    smash_dr1  vos   SMASH DR1
    smash_dr2  vos   SMASH DR2
          sze  vos   SZE+Optical Studies of the Cosmic Accele ...
    w_project  vos   The w Project
      zbootes  vos   z-band Photometry of the NOAO Deep-Wide  ...
```

Next: massively-multiplexed spectro

Next frontier are data products from massively-multiplexed spectroscopic surveys:

- Past and now: *SDSS*, including *SDSS-V*
- Now and soon: *DESI*
- Future: *MSE*, *FOBOS*, ...

Issues:

- Traditional spectra access via FITS files is painfully slow (>1s)
- Overhead in file I/O.
- Not feasible for many spectra.

⇒ **Develop fast spectro service.**

⇒ How do we access millions of spectra quickly?

DESI Dark Energy Spectroscopic Instrument

U.S. DEPARTMENT OF ENERGY | Office of Science

ASTRO DATALAB

- ★ 14,000 square degrees
- ★ 40 million spectra of galaxies and quasars!
- ★ 10 million spectra of stars
- ★ Commissioning/early SV data in 2019-20
- ★ 5-year survey to start soon

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	---

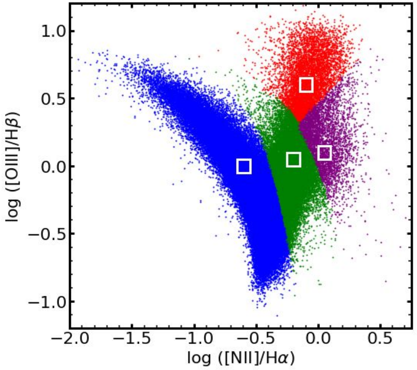
Mayall 4m (Kitt Peak, AZ)

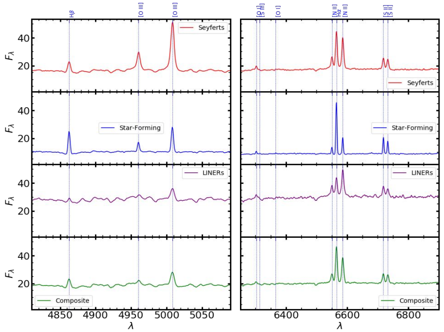
AURA

Slide courtesy S. Juneau

Science Example - Stacking spectra

Stacking SDSS Spectra of Galaxies Selected from the BPT Diagram





- With Data Lab SDSS file service, the notebook executed in **7.5min using 4x100 spectra**

- Using the current DL Spectral Service demonstrator, notebook completes in 25 sec (**8 sec for spectrum access**)

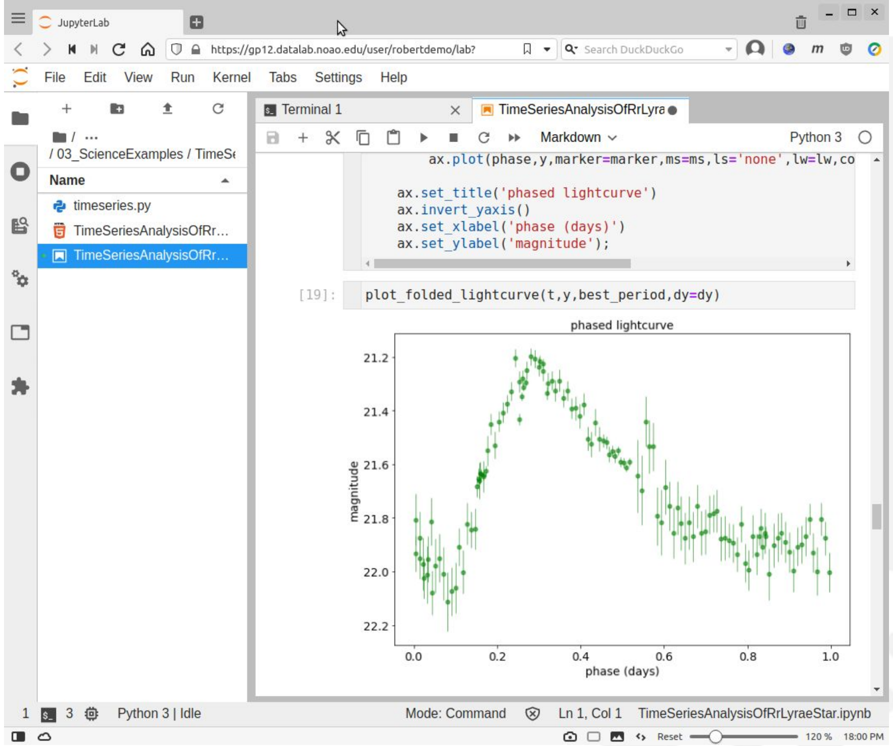
- Currently developing production-level system that can **scale to DESI**

Discovering Our Universe Together

Slide courtesy M. Fitzpatrick

Bring your analysis to the big data

- Remote computing, co-location with data
- No installation required (just a browser)
- Jupyter notebook server
- DL-curated NB suite + user-contributed NBs
- Full astro S/W stack installed
- Planned: users own their containers, e.g.
  - can install S/W
  - can start from scratch any time
- Interfaces to data, to services, and to user storage (DB and VOSpace)



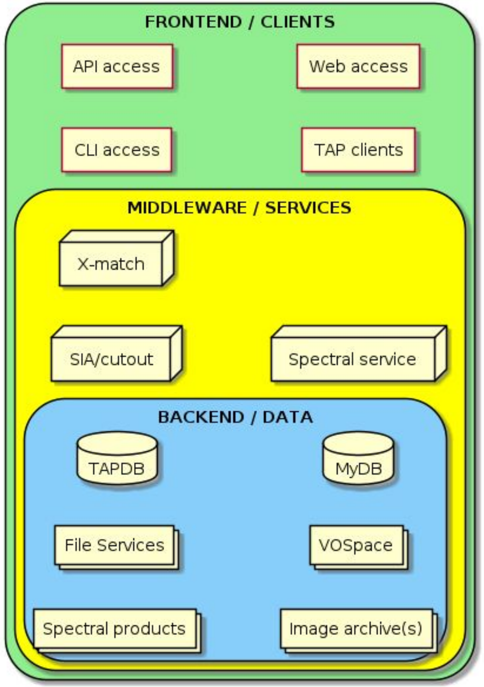
```
        ax.plot(phase,y,marker=marker,ms=ms,ls='none',lw=lw,co

    ax.set_title('phased lightcurve')
    ax.invert_yaxis()
    ax.set_xlabel('phase (days)')
    ax.set_ylabel('magnitude');
```

```
[19]: plot_folded_lightcurve(t,y,best_period,dy=dy)
```

Many ways to connect to the data

- Support standard IVOA protocols wherever possible
  e.g. TAP, SIA, ADQL, (SSA)
- Develop purposefully custom mechanisms where needed
  e.g. PostgreSQL, Q3C
- Translate between layers where needed
  e.g. Data Lab queryClient, storeClient
- Support various access modes
  e.g. sync and async queries
- Support various (popular) access methods
  e.g. Python API (notebooks, scripting), CLI, TOPCAT, Web portal services

Data Lab layers



Planned: self-managed user groups

To enable easy user collaboration and data sharing:

- User can create group → *Spontaneous collaborations*
- Can invite/remove others → *Membership management by group itself*
- Can assign roles → *E.g. admin, write+read, read-only*
- Can attach storage → *Group-owned VOSpace and MyDB instances*
- Admin/owner can dissolve group again → *Wrap it up when done*
- Projects can mint DOIs → *E.g. for a paper manuscript*

Solutions exist already out there. Adapt and/or emulate what’s great, e.g.:

- Sci Server’s group management + storage volume attachment
- CADC VOSpace 2.0 with DOI minting capability

Also a service: User support

Sign up for a free Data Lab account: https://datalab.noirlab.edu/

Get help from the DL team (**we solve every help request**)

Email: datalab@noirlab.edu

Helpdesk: https://datalab.noirlab.edu/help

User Manual: https://datalab.noirlab.edu/docs/manual

Code base: https://github.com/noaodatalab/

Suite of example NBs: https://github.com/noaodatalab/notebooks-latest/

Ping us on Twitter: @DataLabAstro

If needed: install Data Lab clients and CLI on local computer:

```
pip install noaodatalab  ← Still ‘noao’ in the name
```

The three things to take away

- A Science Platform combines big data, co-located (remote) compute, data discovery, easy data access, analysis, visualization, and collaborative working.

- As part of the larger NOIRLab data ecosystem, the *Data Lab Science Platform* does all these, while hosting one of the largest collections of photometric data, image datasets, and adding spectroscopic capabilities. Importantly, users can *access all data products from raw to HLSP*.

- We are a *community* Science Platform: users-first, open-source (client code and NB collection), open-access (most data sets), open-standards (supporting IVOA standards and interoperability).

Obrigado.
Thank you.

robert.nikutta@noirlab.edu

datalab.noirlab.edu