NSF's National Optical-Infrared Astronomy Research Laboratory

AURA

The Future of Astrophysical Archives: Science Platforms

Stéphanie Juneau

Associate Astronomer & Data Lab Scientist

NSF's OIR Lab

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
- Roles:
  - 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

Example: Astro Data Lab - datalab.noao.edu


Mayall 4m (Kitt Peak, AZ)


Blanco 4m (Cerro Tololo, Chile)

Mission

Efficient exploration and analysis of large astronomy datasets with an emphasis on NSF's OIR Lab (NOAO) wide-field 4-meter telescopes

NSF's National Optical-Infrared Astronomy Research Laboratory

AURA

Summary of Current Data Lab Functions

Function	Method
Sky exploration	Image discovery tool Catalog overlay tool
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
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)

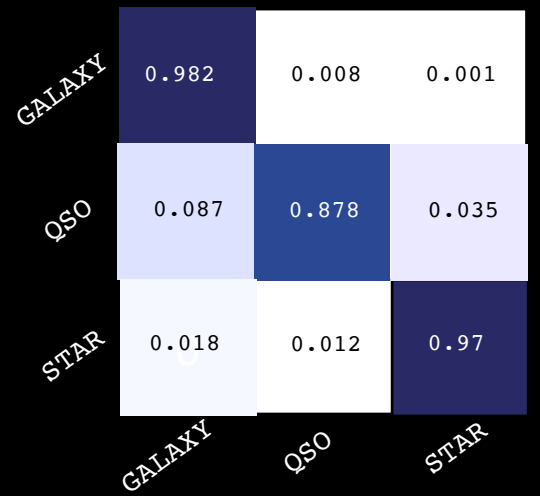
```
query = """
        SELECT dered_mag_g as gmag, dered_mag_r as rmag,
               dered_mag_z as zmag,
               dered_mag_w1 as w1mag, dered_mag_w2 as w2mag, type,
               snr_g, snr_r, snr_z, ra, dec
        FROM ls_dr3.tractor_primary
        WHERE (snr_g>3 and snr_r>3 and snr_z>3)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

Example Workflow

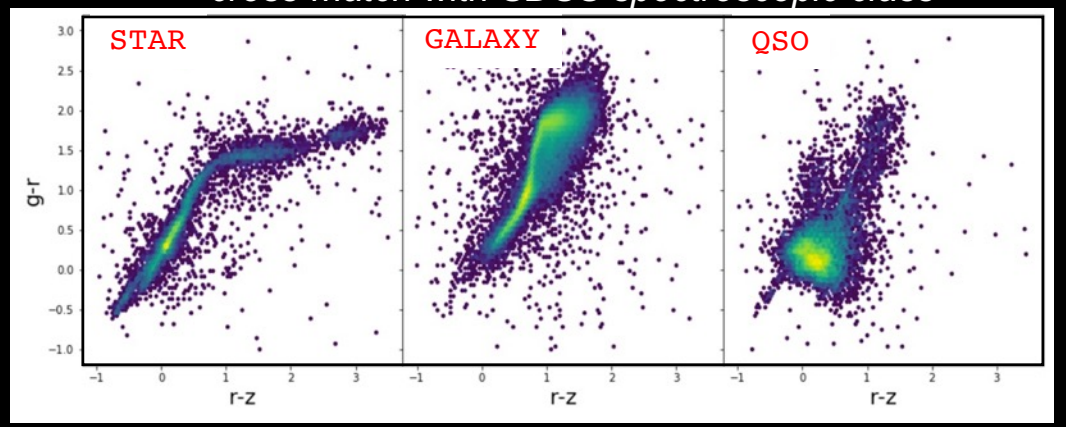
Machine-Learning:

Confusion matrix (spectroscopic training set)

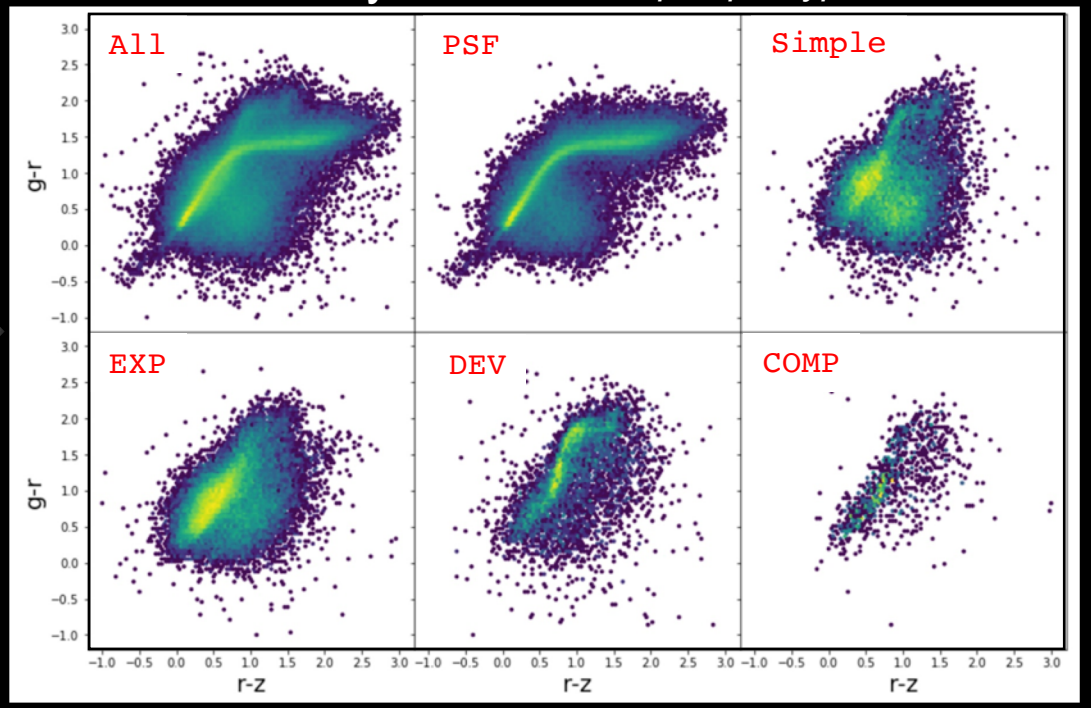


Joint query:

cross-match with SDSS spectroscopic class



Analysis: *color-color plot per type*



Big & Small (not a *one-size-fits-all* approach)

- Range of archive/platform scales to fulfill a range of needs:
  - Large-scale centers/clouds for, e.g., frequently used large datasets (+cross-analysis)
  - Small-scale centers for local needs (e.g., small observatory) and for, e.g., faster response to new or very specific needs (plus, small teams are more agile)
- Common, nationally/collaboratively supported cyberinfrastructure
  - Backbone infrastructure with toolkits (e.g., portable software containers)
  - Training to create/maintain science platforms in addition to training the user base
- Distributed workforce with expertise
  - Retain in-house expertise with missions (past + current)
  - Create an expertise network that is spread/diverse to maximize creativity & problem-solving
  - Ensure a viable career path with possible growth/promotions along the way for technical or dual (science + computation) career

Interoperability ←→ connectivity

- *It's all in the layers*
  - Adopting standards not only (necessarily) in data models but in API/connectivity
  - Can learn/adopt strategies from the web community (e.g., easily configured APIs that can serve out complex data models) → web service layer = common interface
- Distributed computing
  - Techniques/software solutions to work across the data archives/science platforms
  - With increased connectivity, we no longer need to be under the same roof to use the same data
- Need-driven development rather than top-down design
  - Start from **use cases**: what is needed? what is the demand? (+ be responsive as these will evolve)
  - Cast a **wide net**: who gets to sit at the table and voice their needs?

Recommendations

1) **Science Platforms**: software/tools/tutorials co-located with the data

2) **Big & Small**: Variety of data archive/platform scales but with a common, nationally/collaboratively supported infrastructure (including software toolkits & training to achieve a vast network of expertise & broad userbase)

3) **Interoperability** from adopting standards not only (necessarily) in data models but in API/connectivity, and driven by needs of the astro community at large to optimize their scientific output

Appendix: Q&A

Q1: What do you see as the future of archives?

There are at least three different major functions of an astronomical archive: 1) enabling the main experimental analysis for which the data were obtained; 2) providing long-term storage and curation of the data, and 3) enabling scientific exploitation of the data beyond its original experimental context. The latter calls for the development of Science Platforms (more than simply storing data; they include server-side analysis tools so that the users can work close to the data). User friendliness will remain key: i.e., we cannot solely rely on powerful tools “under the hood”, in particular as the number and size of large datasets will continue to grow (data flood). There will be a strong need for adequate training, including tutorials, but also classroom-adapted material to start training the next generation sooner and make sure they are well-versed in data science in addition to astronomy.

Q2: What are the challenges to enabling interoperability between archives?

Heterogeneous data formats and choices of technologies; network speed can still be a bottleneck for large datasets; different interfaces may also act as a barrier to user accessibility in addition to developer accessibility between archives. Dealing with mission/project-specific specs can also hinder interoperability if they impose choices that diverge between archives (e.g., serving flat files versus databases). Interoperability requires some standardization. Standardization can make it easier to borrow solutions from other archives, but can also get in the way of innovation and adoption of new industry technology, and make it harder for an archive to meet its personal development goals. So achieving a balance in terms of standardization would be necessary.

Q3: How can we facilitate interoperability between the NASA and NSF-funded archives?

By adopting standards (including existing standards, such as the CAOM where applicable - Common Archive Observation Model); by working collaboratively to solve common, shared problems; by developing example use cases for potential users that showcase successful workflows employing and combining NASA and NSF-funded archives. Some cases such as cross-analysis may further benefit from co-located data. This can be achieved with multiple (consistent) copies/mirrors of particular datasets and/or using common tools to retrieve data from multiple sources, or yet implementing a system for distributed computing, which could be a longer term, versatile solution.

Q4: What are the advantages/disadvantages of a central archive?

Advantages: removes the need (and latency) for data transfer; optimizes cross-analysis of large datasets; might make it easier to standardize datasets (which would still need substantial dedicated effort); might make it easier to converge on a technological solution

Disadvantages: ideas/technologies more likely to stall due to being less diverse; slower to adapt to big changes or new directions (smaller teams/smaller centers are more agile); more geographically limited expertise in the astronomy community (risk of losing in-house expertise and knowledge at important nodes), which works against knowledge retention – in particular for mission-specific and project-specific knowledge. Risks of alienating archives from the experiments that feed them, and depend on them.

Q5: How should we assure long-term access to data?

Funding that is not strictly tied to a mission/project (or missions/projects have much longer enforced funding end date beyond their “completion”); the legacy of astronomy datasets for data-mining driven-science should be a core goal for data centers; should not require a new project to continue data access from a previous project (as is done for SDSS as there is uncertainty at every round: will SDSS-V get funded and take over SDSS-IV? Then will there be SDSS-VI? etc.). Strategies could be adopted from experts in other disciplines such as archivists and librarians to define a persistent archive (e.g., the San Diego Supercomputer Center has an implementation of a persistent archive).

Q6: What information do we need to maintain from projects?

Information on data acquisition including observation conditions, data format (metadata), data properties (including calibration data and processing), documentation on data access, data pipeline used (tracking software versions), data analysis software used (for reproducibility), ideally a bibliographic database as well as working examples and tutorials. We also need to maintain expertise (people), and ensure that the data are not only discoverable but useable with adequate training. For heavy data reduction pipeline, it may not be feasible or practical to serve compute resources to run the pipeline but information on the pipeline source (and versions used) should be available.

Q7: Should we archive simulations?

Yes, but there needs to be considerations for the types of simulations, their purposes/use cases, and the data formats (e.g., snapshots of cosmological simulations versus derived catalogs of a simulated light-cone are very different “beasts”), and whether they need to be co-located with observational datasets or not. The most obvious/natural segue with observational data archives might be to archive simulated catalogs that are in formats akin to observations (e.g., simulated LSST or DESI photometry catalogs). For more simulation-specific archives, if there were portable science platform software “toolkits”, one could launch such a dedicated science platform as part of a greater network/cyberinfrastructures of data centers (science platforms).

Q8: How do we enable astronomers to process some/all of the relevant data on local machines?

Concept of science platform with server-side analysis, combined with multiple ways to access subsets of data including download or access through browser, TAP schema, programmatic API access, etc. A number of different ways can adapt to the demand to use the data in different workflows (rather than expecting a “one-size-fits-all” scenario). Science platforms should aim to allow users to customize their software environments, so that they can bring their local software close to the data in addition to the tools provided by the science platform itself. Conversely, Science Platforms could aim to be deployable elsewhere, such that users could deploy the infrastructure on local hardware, with a slice of the data that they need. Container technology makes this much easier. Then they could mix and match the science platform software and tools with their own.

Q9: What should the relationship be between astronomical archiving and other national-scale archiving challenges? E.g. in geophysics, biology, physics, etc.

Communication to establish best practices and share lessons learnt would be valuable. In terms of the scientific content, perhaps there are some multidisciplinary questions (e.g., astrobiology, planetary science) that could further benefit from a co-located or shared effort. Overall, the specific needs might differ sufficiently to argue against a centralized archive.

If using cloud computing and cloud data storage, there can be concerns with leadership / governance if we were to move everything to a commercial cloud solution or to a multi-disciplinary cloud/archive. The best solution might be a compromise to maintain flexibility to respond to the current variety of scientific needs.