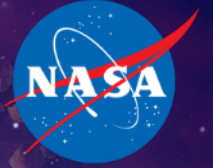


National Aeronautics and
Space Administration



EXPLORE SOLAR SYSTEM & BEYOND


Astrophysics Advisory Committee

March 31, 2022

Linda Sparke, Roopesh Ohja

Astrophysics Division

Science Mission Directorate

 [@NASAUniverse](https://twitter.com/NASAUniverse) [@NASAEoplanets](https://twitter.com/NASAEoplanets)



Astrophysics Science Platform Project

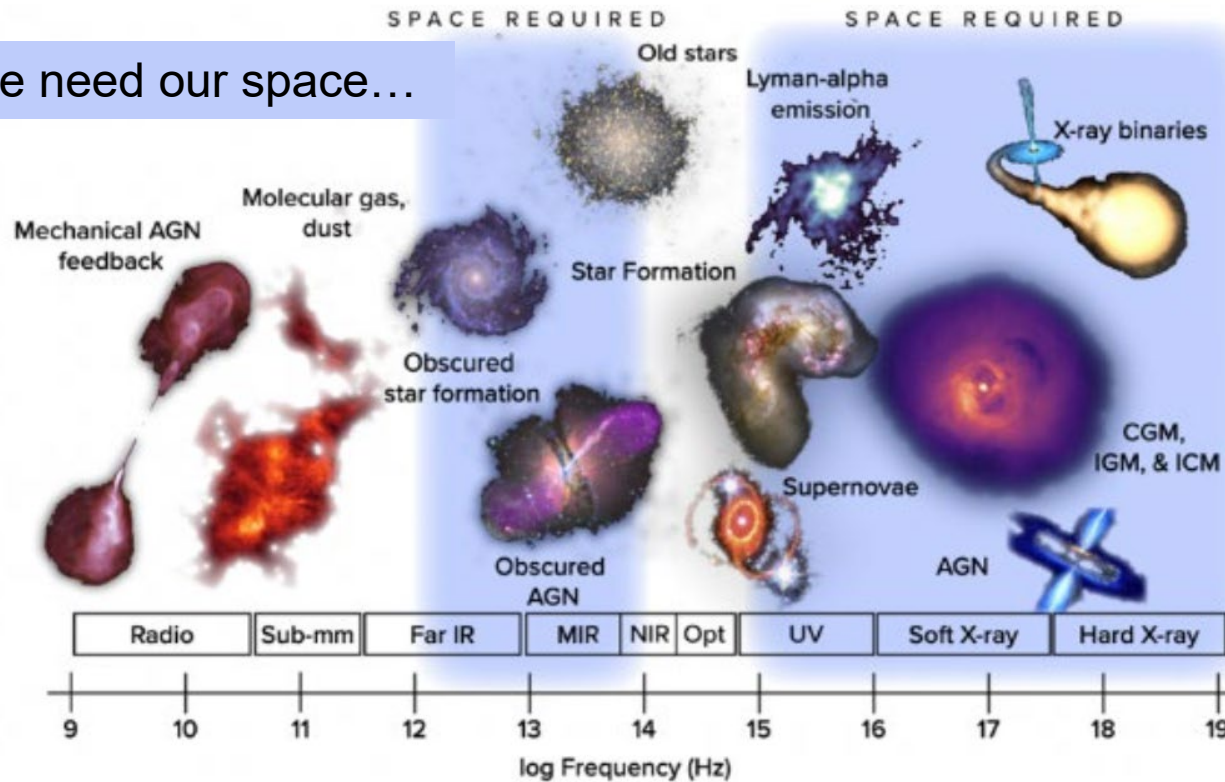
Astro2020 Decadal Survey Section 4.5.1 on Data Archiving, Curation, and Pipelines:

“The importance of joint analysis of observations from different facilities and wavelengths, and of sophisticated archiving with associated science platform tools, will grow dramatically over the next decade.”

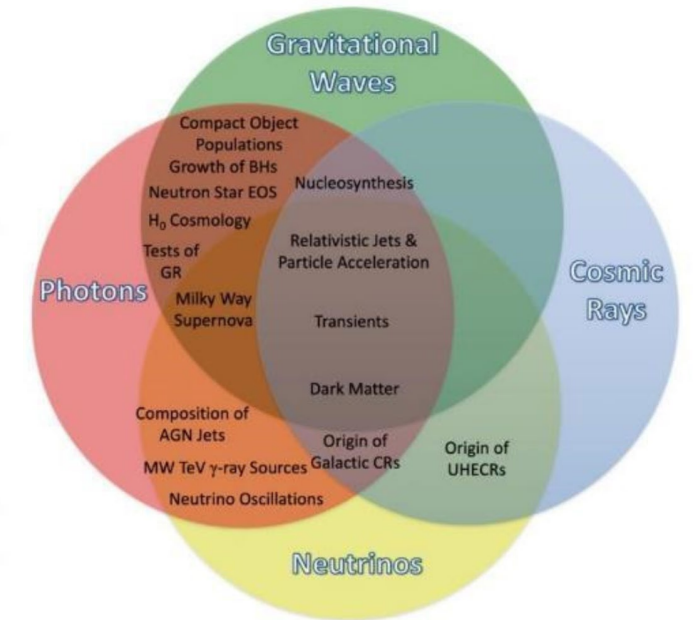
In October 2021, APAC advised the Astrophysics Division to “strategically identify appropriate cloud-based infrastructure options to facilitate analysis and theoretical modeling of the large data volumes from missions like Euclid by the wide community.”

Astrophysics Data Landscape in 2030

We need our space...



“All priority science areas require multiwavelength observations.” - Astro2020 Decadal Report



“Exploring the cosmos in the multi-messenger and time domains is a key scientific priority for the coming decade”. – Astro2020 Decadal Report

Astro2020: multi-wavelength and multi-messenger astronomy

Data Analysis Landscape in 2030

More and better data will demand new Astrophysics Archive capabilities

- **Data volumes** in the Astrophysics Archives **double roughly every 2 years**
- **Large data sets are on the way:**
 - **Euclid** plans to release its first survey data in ~FY25 (Quick-Release1 ~30TB), with 1.5PB in FY27, and a final release of 4.5PB
 - **Roman** expects 0.75PB of level-2 science data (scan-by-scan) and level-3 data (co-added) after the first 6 months (~FY28), then 1.5PB per year
 - **Rubin Observatory (NSF-DoE)** will release 20TB/night from ~FY24, roughly 1.5PB per year, with a final data release of 15PB after 10 years of operation
- Multiwavelength astronomy requires **analyzing data sets jointly**; science often requires analyzing **data along with simulations**
- **Advanced data science tools** (e.g., AI/Machine Learning) can extract new knowledge, given adequate **computing resources close to the data**

Big Data Analytics and Machine Learning



Astrophysics Science Platform: 3 components

1. Cloud-based interface and supporting system rooted in open-source software, providing compute resources “proximate” to the data (low-latency access)
2. A rich collection of notebooks and pre-configured software containers for multi-wavelength and big-data science
3. Advanced data services that enable fast, location-agnostic services for data held on the platform, in NASA Astrophysics archives, and beyond.

Community users:

scientists needing server-side analysis for large amounts of data;

those needing computational facilities for big data analytics, Machine Learning, etc;

collaborators at different institutions wishing to share a computational environment;

scientists who cannot easily build the software they want to use; etc.



1. Cloud-based interface

Users should be able to

- log in without requiring NASA credentials (PIV card);
- easily obtain a login to use modest free resources (storage and computation) and pre-configured software for common astronomy codes;
- access both NASA and external astronomy data in the cloud;
- run science analysis through a web browser;
- use the platform to collaborate worldwide;
- apply for an account that allows more resources, running a user's own code, etc.;
- bring own resources in the same cloud to compute on Astrophysics data.



2. Notebooks and software containers

Users should be able to

- access a library of notebooks ready to run on the platform, to benefit from mission expertise encoded in working examples without having to “know the right person”;
- use software containers (pre-configured environments) for common tasks, without having to undertake complex software builds;
- access Big Data efficiently by using example notebooks developed by experts for the specific platform and dataset;
- start doing more sophisticated statistical and machine-learning analyses quickly and easily.

3. Advanced data services

Platform users should be able to

- use data discovery tools on the platform;
- use graphical user interfaces to data discovery tools such as archive web portals, and connect the results to their platform analysis;
- use context-aware software to fetch Astrophysics data automatically on the cloud or from on-premises archive services as appropriate;
- access data from a variety of sources outside Astro archive holdings with standard APIs*.

*Application Programming Interfaces – intermediary for machine-to-machine queries

Near-term goals for FY22-23

- Make some high-priority NASA datasets available in AWS* S3** buckets – e.g. Spitzer mosaics, Chandra & Swift imaging data, IRSA catalog products, MAST HST and GALEX data
- Develop a Python software layer for cloud-based data access, integrated with standard tools such as astroquery and/or PyVO;
- Prototype a public user interface based on JupyterHub on the NASA AWS cloud;
- Develop a first set of example Jupyter notebooks^ that use the standard tools with data across the NASA archives -- both data in the cloud and data held on-premises.

* Amazon Web Services

** Simple Storage System (fast access for data that you want to compute with)

^ A web-based interactive computing platform

5-year plan: under development

A public JupyterHub platform for science community login, with

- Pre-configured software containers* (grab and run...)
- Some limited free data and compute resources,
- An environment to build and run the user's own code,
- Collaboration tools such as file sharing.

Hopefully also:

- Single-sign-on login, federated across the platform and the archives (sign in on the platform, and MAST knows who you are...)

*A container packages up code with all its dependencies, so the application runs quickly and reliably when moved from one computing environment to another



Backup Slides

Multi-Mission Use of Cosmological Simulations: a Pilot Project

Trillion-particle Outer Rim simulation of Heitmann et al. 2019:
Over 5PB of gravitational N-body simulation output: a data-handling challenge!

from which Korytov et al. 2019 made a synthetic galaxy catalog for LSST Dark Energy Science Second Data Challenge (DESC DC2):

- **2 billion rows, 5TB of information** on mock galaxies: stellar mass, morphology, SEDs, broadband magnitudes, host halo information, weak lensing shear
- Covers 440 square degrees, redshift $z < 3$, to magnitude 28 in r -band

IRSA serves the mock-catalog with “standard” user tools

Expected in FY2022: simulated sky images “as seen by” the Roman High-Latitude Imaging Survey. IRSA will serve the images with tools.

- Images will cover 20 sq deg: Troxel et al (2022, in preparation)

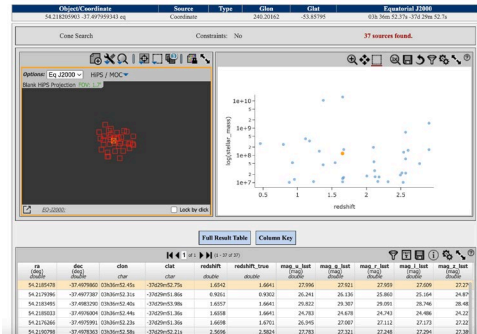
Tri-Project Team: Alina Kiessling (JPL, PI)

Mock Catalogs: Andrew Benson (Carnegie Obs), Andrew Hearin, Katrin Heitmann, Eve Kovacs (Argonne)

Simulated Images: Mike Jarvis (U Penn), Michael Troxel (Duke U); Archive: Vandana Desai, Harry Teplitz (IRSA)

CosmoDC2 Mock Catalog Access Suite

<https://irsa.ipac.caltech.edu/Missions/cosmodc2.html>



IRSA's standard interactive tool for searching, exploring, visualizing catalogs



Machine-to-machine: standard VO API for tables allows access by Python and third-party interfaces (e.g. TOPCAT)



Analysis-ready bulk download in format compatible with industry-developed big-data toolkits (in test)



IVOA = International Virtual Observatory (VO) Alliance;
NASA archives use VO protocols to make their data searchable and retrievable