Lawrence Berkeley National Laboratory
LBL Publications

Title

Jupyter-Enabled Astrophysical Analysis Using Data-Proximate Computing Platforms

Permalink

Journal

ISSN

Authors

Juneau, Stéphanie
Olsen, Knut
Nikutta, Robert
et al.

Publication Date

DOI

Peer reviewed

Jupyter-enabled astrophysical analysis using data-proximate computing platforms

S. Juneau
NSF's NOIRLab

K. Olsen
NSF's NOIRLab

R. Nikutta
NSF's NOIRLab

A. Jacques
NSF's NOIRLab

S. Bailey
Lawrence Berkeley National Lab

Abstract—**The advent of increasingly large and complex datasets has fundamentally altered the way that scientists conduct astronomy research. The need to work closely to the data has motivated the creation of online science platforms, which include a suite of software tools and services, therefore going beyond data storage and data access. We present two example applications of Jupyter as part of astrophysical science platforms for professional researchers and students. First, the Astro Data Lab is developed and operated by NOIRLab with a mission to serve the astronomy community with now over 1,400 registered users. Second, the Dark Energy Spectroscopic Instrument (DESI) science platform serves its geographically-distributed team comprising about 900 collaborators from over 90 institutions. We describe the main uses of Jupyter and the interfaces that needed to be created to embed it within science platform ecosystems. We use these examples to illustrate the broader concept of empowering researchers and providing them with access to not only large datasets but also cutting-edge software, tools, and data services without requiring any local installation, which can be relevant for a wide range of disciplines. Future advances may involve science platform networks, and tools for simultaneously developing Jupyter notebooks to facilitate collaborations.**

IN ASTRONOMY, and other disciplines, the need for online science platforms is driven by both rapidly increasing data volume and by the complexity of datasets, which require highly specialized and diverse software libraries to be co-located with the data. In terms of data rates, 1 TB/night is typical of current data-intensive facilities, while the incipient Rubin Legacy Survey of Space and Time (Rubin LSST) will collect 20 TB of raw data and send a stream of 10 million alerts[1] of brightness changes each night. The size of teams working collaboratively on data analysis is also growing, routinely comprising tens to several hundreds of researchers and students. Clear advantages arise from being able to connect Jupyter Notebooks or JupyterLab to large datasets to perform analysis and data visualization efficiently both to support collaborative work and to empower a wide array of researchers including from institutions that may lack local data storage or compute resources.

Historically, observational research was primarily conducted by individuals and small teams able to use a telescope or satellite to make observations. They would process and analyze the data using software packages on their local computers, before publishing results in astronomical journals. Larger teams and collaborations started changing the landscape by adding a need to share data among several people. Another big change came with publicly accessible data archives, which can be used to address new science questions without new observations. Researchers can download the needed data, and process and/or analyze them. For instance, statistics from the Hubble Space Telescope show that discoveries from archival data outpace those made for the original intended purpose of the observations[2]. Even more recently, another major change took place as some data sets are becoming prohibitively large or complex such that researchers no longer have the means to efficiently download the data onto their local computers. Instead, one needs to bring the computing workflows to the data.

Below, we showcase two astronomy projects as ongoing successful example applications. First, the Astro Data Lab ([1], [2]) at NSF's National Optical-Infrared Astronomy Research Laboratory (NOIRLab) is an online astronomy science platform currently serving large public astronomical datasets including databases with tables ranging from 500,000 to 65 billion rows. Data-proximate analysis and visualization is key to use databases containing over 100 billion rows; transferring even 150,000,000 rows with a small subset of the columns would take ~ 2 hours over a 14 Mb/s connection, compared to ~ 10 seconds over NOIRLab's internal 10 Gb/s connection. It is even more relevant when considering that the databases are supplemented with Petabytes of astronomical images. While this calculus will change with availability of 5G networks, internal network upgrade opportunities will favor the code-to-data approach for the foreseeable future. Since opening its doors in 2017, the Data Lab now has over 1,500 registered users from different countries, ranging from students to faculty/researchers, and educators.

Second, we describe how the Dark Energy Spectroscopic Instrument (DESI; [3], [4]) team employs Jupyter as the primary way to enable their geographically distributed collaboration (nearly 900 researchers from more than 90 institutions) to access petabytes of data with pre-installed software libraries. Jupyter Notebooks are also used to train new team members, and as a quick way to perform integration testing of software releases. For each science platform, we give an overview before presenting use cases for research, training, and education. We will then consider common elements that may apply to science platforms in general, and possible improvements from enhanced Jupyter capabilities.

ASTRO DATA LAB

Overview and implementation

The goal of the Astro Data Lab is to enable efficient archival use of massive astronomical survey data, particularly those produced on the facilities of NSF's National Optical-Infrared Astronomy Research Laboratory (NOIRLab). As shown in **Figure 1**, the data volume from NOIRLab facilities has grown explosively over the last decade, with much of that growth owed to

---

[1] Alerts are sent when a change in brightness is detected by the software comparing the new observation of a given object to previous measurements so that researchers can investigate the cause (e.g., new supernovae, variable stars).

[2] https://archive.stsci.edu/hst/bibliography/pubstat.html

CiSE (accepted)

wide-angle images taken with the Dark Energy Camera[3]. These surveys of the sky result in large maps with a multitude of stars, as well as galaxies that are billions of light-years from Earth. These maps are typically processed through automated measurement pipelines, yielding tables of billions of astronomical objects. The value of these processed data is enormous, not just for the teams taking the data for their own purposes; much of the value lies in the potential for the broad astronomical community, and the public, to make unanticipated discoveries.

Astro Data Lab enables several kinds of archival science cases, including 1) discovery of unusual, rare, or otherwise interesting objects selected through queries to databases; 2) classification of objects in large samples, potentially including image cutouts and aided by machine learning techniques; and 3) discovery in the time domain, where time series data of selected objects are used to create new understanding or reveal new phenomena. All of these science cases are supported by services provided by Astro Data Lab, with Jupyter Notebooks as the key technology that brings their power to users. For example, users can query databases directly within a Notebook (via a *QueryClient*), and save their output to storage associated with their Data Lab accounts, which is currently one Terabyte per user (via a *StoreClient*). **Figure 2** illustrates how these pieces connect together to make Jupyter an integral part of the Astro Data Lab science platform, and we include examples of the code syntax to use the Clients within a Notebook in the Appendix.

The Astro Data Lab science platform runs on premises using commodity hardware (1U and 2U general-purpose servers with typically two Xeon CPUs and 20 cores each). The various data services and the Jupyter notebook server are currently deployed as virtual machines, with the available resources shared by all users. To facilitate resource allocation, and permit users to maintain their own software stack, we are currently working to implement containerization based on Docker and Kubernetes. Specialized hardware consisting of three machines (production, hot standby, staging) is deployed for the cat-

alog databases, and is optimized for CPU/RAM throughout and backed by fast SSD storage.

Jupyter for Research

The majority of Astro Data Lab users are professional astronomers and graduate students. They tend to have topical expertise in their subject of research but varied levels of coding experience and knowledge. To remove the need for local installation, the Jupyter Notebook server includes pre-installed software packages and Python libraries (e.g., numpy, matplotlib, pandas), and is populated with example Jupyter notebooks directly within the user accounts. The default installation includes astronomy-specific packages such as Astropy[4] ([5]) and the datalab clients[5].

The Astro Data Lab Jupyter notebook server provides a convenient way to quickly query database tables, visualize their content, and perform detailed analyses. As an example scientific use case, we consider the Survey of the Magellanic Stellar History (SMASH[6] [9]), which mapped a wide area in and around the periphery of the Magellanic Clouds, two of the nearest companion galaxies to the Milky Way. The SMASH database table contains millions of stars forming the Magellanic Cloud galaxies as well as millions of distant, background galaxies, totaling nearly 360 million unique objects. Discoveries from SMASH include a ripple in the disk plane of the Large Magellanic Cloud (LMC; [10]) − likely the result of an interaction with its neighbor galaxy the Small Magellanic Cloud − and a previously unknown faint dwarf galaxy companion to the Milky Way, named Hydra II ([11]). The study of such faint companion galaxies may hold clues to the nature of the unknown "dark matter".
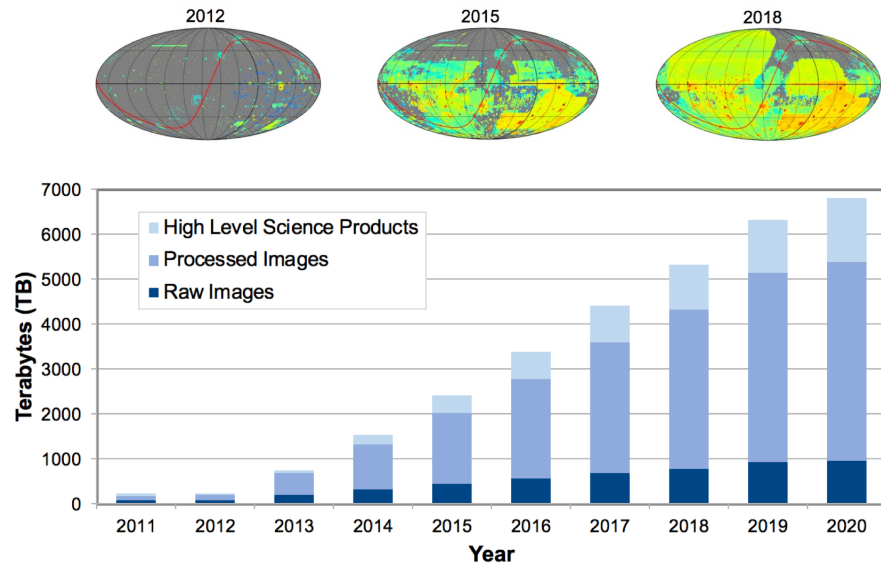
The Hydra II dwarf galaxy finding is featured in a science example Jupyter Notebook provided to users, giving them a demonstration of a discovery from beginning to end, and a starting point for their own explorations. It shows the chain of steps for research from how the database query is constructed and issued to the server, how to display the measurements that are returned, and how to process them to highlight

---

[3]https://www.darkenergysurvey.org/the-des-project/instrument/

[4]https://www.astropy.org/
[5]https://github.com/noaodatalab/datalab
[6]https://datalab.noirlab.edu/smash/smash.php

Figure 1. Evolution of the sky coverage from NOIRLab's observatories with telescope diameter of 2-4 meters during the years 2012, 2015 and 2018. The color scheme indicates the exposure time on a logarithmic scale, mostly ranging from 900 seconds (15 minutes; green) up to nearly 1 million seconds (277 hours; dark red). The rapid increase of coverage starting around 2013 coincides with astronomical surveys conducted with wide-field cameras such as the Dark Energy Camera (DECam). These full-sky maps were generated with an equal-area *Mollweide* projection, and the Milky Way Galactic Plane is shown with a solid red line. The bar chart illustrates the growth in data volume stored at NOIRLab's Astro Data Archive, split between unprocessed (raw) images, processed images, and high-level data products such as image mosaics.
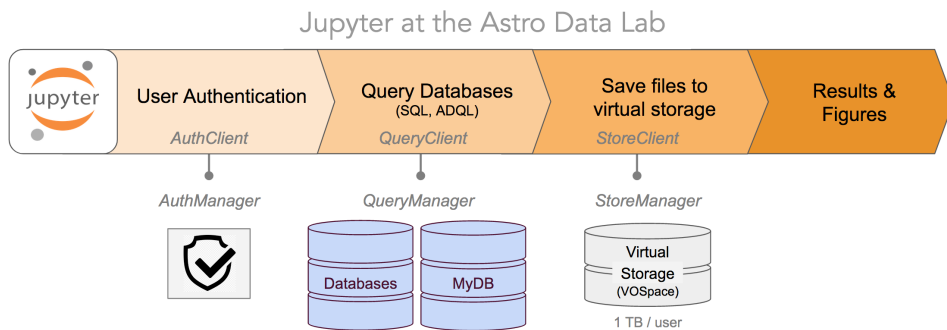
Figure 2. Illustration of some of the key Clients and Managers that are part of the Data Lab platform, and which can be used with Jupyter to connect the users to the data, and enable them to conduct their analysis using Notebooks. User authentication is only required for users to access their MyDB or virtual storage space. The queryClient allows users to send a database query, and retrieve the results directly within a Notebook. The overall goal is for users to be able to conduct their entire analysis up to generating publication-quality figures on the Jupyter server.

4

the newly discovered galaxy. Researchers and students can also browse other example notebooks from within their Astro Data Lab account, or on the "03_ScienceExamples" GitHub repository[7].

Researchers who conduct their analysis on the Astro Data Lab platform are asked to include a formal acknowledgement in all resulting journal publications, and to use the *facilities* command "Astro Data Lab" when available, such as for the American Astronomical Society journals. They may further propose to add their Jupyter Notebook to our contributed Notebook collection[8]. User-contributed notebooks that show applications not yet covered in our collection are especially welcome. The "05_Contrib" folder is divided into astronomy topics, and includes step-by-step instructions as well as an Astro Data Lab Notebook template for guidance. The team then reviews the proposed notebook, and may iterate with the contributor until the Notebook is approved to be added to the notebook collection, which will consist in merging the pull-request (PR) on GitHub. The vetting process requires manual interventions and topical knowledge. Once the GitHub PR is merged, the distribution and testing of the new contributed notebook is automated as described below.

Jupyter for Training

Training is crucial at all career levels. This is partly due to the important changes on the software, tools, and methods being developed and utilized. For example, several experienced researchers are newly learning the Python programming language. More fundamentally, this is also due to the increasing prominence of data-driven modes of conducting science. Training presents new challenges as we need to convey and teach new, data-intensive methods, but it also presents new opportunities. Data Lab in combination with Jupyter helps make research fairer and more equitable. Improving access to quality data is an important part of building an inclusive community by lowering the barrier of entry, thus leveling the playing field for all ([6]).

The Astro Data Lab recognizes the importance of providing researchers, students, and professors with adequate technical training and support in data science to empower them to take advantage of the increasing volume and complexity of datasets. In terms of documentation and tutorials, the team produced a User Manual[9] using Sphinx Documentation[10] as well as numerous example Jupyter Notebooks, which are available as a reference at any time. In addition, team members give live presentations and webinars, conduct online and/or in-person demos as well as hands-on tutorials, during which participants can interactively follow along. Most of the demos and tutorials involve going through Jupyter Notebooks, which range from entry-level to advanced science examples to technical "How-To" notebooks, and which are described in a README file. The procedure to maintain, and automatically update the collection of Notebooks is described below.

Jupyter for Education

Jupyter Notebooks facilitate education about topics in astronomy and/or data science, including computer coding, data manipulation and analysis. We currently support the Python language, and have two primary sets of notebooks under the Education and Public Outreach (EPO) umbrella.

The Teen Astronomy Cafe Program[11] consists of hands-on computer activities and discussions led by volunteer scientists in addition to the program director. The target audience is 12 to 18 year-old students. Jupyter Notebooks are produced ahead of time, and may include interactive widgets (sliders, and buttons) and/or instructions to teach students to modify code cells. Each activity is designed to cover a different astronomy topic for which data science techniques can be employed. As of 2020, these EPO Notebooks can be run either on Google Colab or directly on the Astro Data Lab platform.

At the level of undergraduate and graduate students, the Astro Data Lab is adding a collection of notebooks adapted from the La Serena School for Data Science[12]. These notebooks were originally written by the guest lecturers, and cover more advanced topics involving, e.g., statistical

---

[7]https://github.com/noaodatalab/notebooks-latest/tree/master/03_ScienceExamples

[8]https://github.com/noaodatalab/notebooks-latest/tree/master/05_Contrib

[9]https://datalab.noirlab.edu/docs/manual/index.html

[10]https://www.sphinx-doc.org

[11]http://www.teenastronomycafe.org/

[12]http://www.aura-o.aura-astronomy.org/winter_school/

tools applied to large datasets, high performance computing, image analysis, and machine-learning techniques. The goal is for students to learn hands-on skills that they can then apply to their own research.

Lastly, Data Lab notebooks have been used in the classroom upon requests from professors and teachers, including semester-long graduate-level astronomy courses. In some instances, we create temporary *demo* accounts for students. These in-house demo accounts are generated via a script, and allow the users to save their work in progress during the duration of the fixed-term account. After they expire, they are wiped clean, and deactivated using an automated script.

Notebook Deployment and Testing

The Data Lab team develops a suite of notebooks for all users. These range from notebooks introducing Python, SQL, and Data Lab, over technical How-Tos, to complete science cases that reproduce interesting results from the astronomical literature. Astronomy and data science education notebooks complete the picture. Data Lab also assists users who wish to contribute a notebook to the collection by providing them with a template and step-by-step instructions.

Notebooks are developed either on the Data Lab notebook server[13], or locally (with the `datalab` package[14] installed). All version control occurs through GitHub[15]. Pull requests are reviewed by the Data Lab team. When merged to the master branch, a webhook is triggered which sends a beacon to the Data Lab servers. If the beacon is successfully verified, the server pulls the latest version from GitHub and re-populates a specific directory with the updated notebooks. This `notebooks-latest/` directory is read-only mounted in all users' notebook spaces, and it can be browsed both using a browser and a terminal. Users can also pull a whole new read & write copy of it using a custom shell command (`getlatest <targetdir>`).

The notebook suite is now around 40 notebooks strong and growing. We must ensure that all notebooks continue to execute correctly at all times, despite the backend datasets and middleware services being updated often. To facilitate automated pass/fail testing we employ `nbconvert` wrapped in a custom Python package. Testing can be performed through a dedicated notebook (part of the notebook suite), or in command line mode. The testing supports globbing for `.ipynb` files, exclusion patterns, dynamic notebook kernel switching, and summary reporting. Test failures do not interrupt the testing suite, and the tracelog of all tested notebooks is available for inspection. Currently the test suite completes in 35 minutes on the Data Lab notebook server.

DESI

Overview and implementation

The DESI instrument ([4]) will be used to survey one third of the sky (14,000 square degrees), and obtain spectra for 35 million galaxies and quasars in order to construct the largest three-dimensional map of the universe to date [3]. By comparing this map to predictions from numerical models, the DESI team will constrain the nature of dark energy, the dominant yet mysterious force driving the expansion of the universe. In addition to its main cosmology goals, which rely on spectroscopy of very distant galaxies and quasars, DESI will obtain spectra for 10 million stars from the Milky Way, and an additional 10 million brighter, less distant galaxies. These added-value data will allow astronomers to answer a slew of astrophysical questions about stars, galaxies, and their evolution.

Currently installed on the four-meter Mayall telescope at Kitt Peak National Observatory, Arizona, DESI can obtain spectra of up to 5000 targets in a single observation, by allocating a single optical fiber per target. The DESI team first needs to measure the sky positions (coordinates) of all the galaxies, quasars, and stars to observe with DESI, select which ones can be observed simultaneously with 5000 fibers, and compute how to align each remotely-controlled fiber with its target (*fiber assignment*). The targets are selected from the pre-imaging Legacy Surveys[16], which are now complete, and consist of optical and infrared images which were used to measure

---

the position and fluxes of over 1.6 billion unique objects.

DESI generates $\mathcal{O}(1\ \text{PB})$ of data per year, including raw telescope data, processed results, and computer simulations needed for final science analyses. This amount of data together with the sophisticated and specialized software were motivators to use a Jupyter server running close to the data at the National Energy Research Scientific Computing Center (NERSC), where team members can sign in and launch an instance of JupyterLab with pre-installed software. The JupyterHub servers are maintained by NERSC as a service to all users, not specific to DESI, but they are significantly used by DESI collaborators using DESI-maintained custom kernels with DESI-specific software.

NERSC deploys its JupyterHub service using a container-as-a-service platform based on Kubernetes, and uses wrapspawner[17] and batchspawner[18] to enable JupyterLab sessions to spawn on shared or exclusive (batch) nodes on CPU or GPU resources. Most DESI collaborators using Jupyter access it via a node shared with other NERSC users for non-CPU-intensive exploratory analysis and visualization. More intensive Jupyter-based computing is available through distributed analytics platforms like Spark or Dask, or MPI-enabled applications leveraging IPyParallel, managed from notebooks through script-based adaptations to high-performance computing environments.

Below, we present example use cases of Jupyter for the DESI collaboration of nearly 900 members including research staff, postdoctoral researchers, graduate students and engineers. These include examples for the purpose of research, training, education, and software testing. The DESI software is maintained on GitHub (desihub[19]), and includes 59 repositories with 107 members.

Jupyter for Research

In the case of DESI, research includes survey planning, and data processing, exploration, and analysis. The research applications themselves are fairly broad as they include both the primary cosmology goals centered around dark energy as well as a slew of possible other projects on galaxies, stars, black holes, etc. Therefore, having a suite of tools and analysis workflows will maximize the outcome from the DESI observations. Jupyter enables interactive exploration and analysis of petabytes of data by remote collaborators without requiring them to download the data locally. It also provides kernels with pre-installed software releases for DESI-specific data analysis.

Jupyter is frequently used by DESI collaborators for initial data exploration and algorithm development (**Figure 3**), before deploying those algorithms as Python packages to be used in non-Jupyter batch jobs using hundreds to thousands of nodes. Jupyter notebooks are also used for collating the results of those batch jobs and visualizing the final results.

A specific example of data exploration is the interactive spectral visualization tool `prospect`[20], which is developed on GitHub, and used by the DESI team at NERSC to visually inspect DESI spectra to debug data processing problems. It is based on Bokeh, and works in the JupyterLab environment. **Figure 4** gives an overview of a version of `prospect` that is being implemented at the Data Lab, showing an example application with a Sloan Digital Sky Survey (SDSS) spectrum of a galaxy.

Jupyter for Training

Jupyter Notebooks are used to create DESI Tutorials[21] to train new DESI collaboration members, and to teach existing members new tasks or functionalities. They are maintained and can be browsed on GitHub. However, DESI team members must run them at NERSC after authenticating, cloning the notebooks from GitHub, and launching a JupyterLab instance on the Jupyter server. The underlying DESI software is written in Python, and is pre-installed. The user has the ability to select DESI kernels to test different software versions if needed, or simply run the default kernel. Tutorials range from basic functionality to generating mock data to planning observations via fiber assignment calculations, and investigating
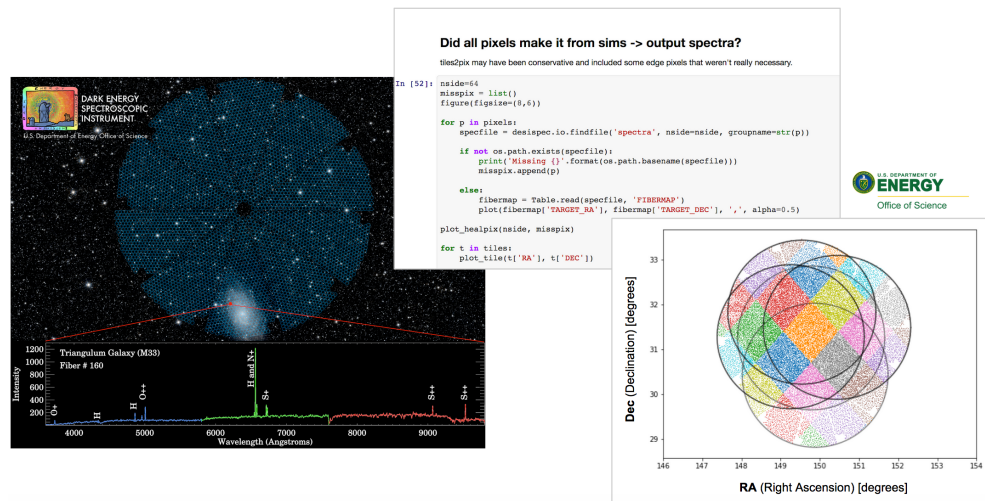
[17]https://github.com/jupyterhub/wrapspawner
[18]https://github.com/jupyterhub/batchspawner
[19]https://github.com/desihub
[20]https://github.com/desihub/prospect
[21]https://github.com/desihub/tutorials

Figure 3. *(Left)* Overlay of the DESI focal plane on the sky with 5000 fiber patrol regions. The example test spectrum at the bottom was obtained from a fiber positioned on a part of the Triangulum Galaxy (red dot) as part of early commissioning of the instrument in October 2019, and shows spectral features as labeled (Credit: DESI Collaboration; Legacy Surveys; NASA/JPL-Caltech/UCLA); *(Middle)* Section of a Jupyter Notebook using the DESI software to plot the fiber location to visualize how overlapping DESI pointings can result in repeated observations, which can be used to obtain more signal. *(Right)* Resulting plot shown inline in the Notebook. The black circles illustrate the full size of a given DESI pointing (focal plane), while each colored dot marks the location of a fiber. The color scheme is mapped to HealPix indices, which are defined to tile the sky with nside=64 in this example. The color saturation level is keyed to the number of repeated fiber assignments (darker colors signify more overlap).

available spectra from early observations. They are meant to train users in a modular fashion, and to enable them to re-use and combine parts from different tutorials in order to create their own workflow.

Jupyter for Education

One of the educational efforts using Jupyter is a program called *DESI High*[22], which aims to teach 14 to 18 year-old students to understand Dark Energy with actual data from the DESI project and Jupyter notebooks. The latter are made available on GitHub and can be run with Binder. The DESI High program also includes a lot of links to useful material and references such as videos, websites, tutorials. An underlying motivation of this program, which is shared with some of the initiatives mentioned previously, is to inspire and teach the next generation of (data) scientists.

Jupyter for Software Testing

Integration testing of DESI software releases is performed by a Jupyter Notebook on the Jupyter server at NERSC. The "minitest.ipynb" notebook in the GitHub `desitest`[23] repository runs an end-to-end integration test from simulating input data to processing it with the data pipeline to performing final quality assurance and analysis. This notebook includes several steps, some of which spawn batch jobs and maintain logfiles outside of the notebook itself. Under normal activity levels on the NERSC servers, the minitest notebook runs end-to-end in about two hours and uses approximately 500 core-hours of computing resources, a scale that would not be viable for typical continuous integration services. The output from this notebook serves as the "reference run" for the software release, providing specific examples of the data outputs, formats, and performance for that release.

---

[22]https://github.com/michaelJwilson/DESI-HighSchool

[23]https://github.com/desihub/desitest
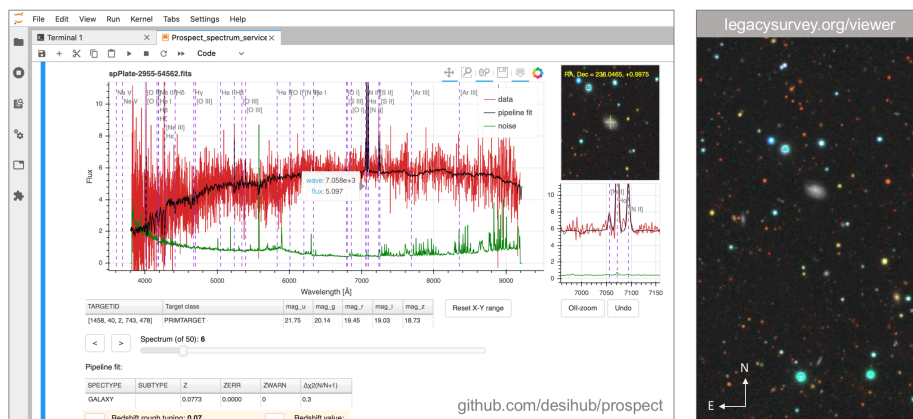
Figure 4. Interactive spectral visualization tool `prospect`, which is developed on GitHub, and used by the DESI team at NERSC. It is based on Bokeh, and works in the JupyterLab environment. Above is a version of `prospect` that is being implemented at the Data Lab, showing an example application with a SDSS spectrum of a galaxy (in red), with the best-fit model overlaid in black, and the error spectrum in green. Vertical dashed lines mark the location of expected spectral features such as emission lines from ionized gas (better seen in the bottom right *zoom box* inset, which zooms around the location of the cursor by capturing the mouse-over information). The target galaxy can be seen in the top right image along with its celestial coordinates in Right Ascension (RA) and Declination (Dec). The interactive tool allows users to zoom, pan, and reset the plot by using the Bokeh buttons at the top of the viewer, and displays tabulated information about the object being plotted (6th object of 50 in the example above as indicated by the slider). Clicking on the image thumbnail leads to the location of the galaxy in the interactive Sky Viewer (https://www.legacysurvey.org/viewer; screenshot on the right; with the North and East direction indicated with arrows).

JUPYTER IN SCIENCE PLATFORMS

We now consider similarities and differences between the Astro Data Lab and DESI examples, as well as other astrophysical science platforms. Both the Data Lab team and the DESI team take advantage of the user-friendly format of Jupyter Notebooks to create examples and tutorials to train new users, and to illustrate key functionalities. They both maintain their tutorials and examples on GitHub for version control and contributions. In the case of Data Lab, each user account comes with the latest collection of example Notebooks which range from "Getting Started" novice level, to more technical "How To" notebooks, complete scientific use cases, as well as educational notebooks and contributed notebooks from the astronomy community. In the case of DESI, team members can follow instructions to clone the notebooks from GitHub and start using them.

The main difference may be the target audience, in the sense that the Astro Data Lab is primarily serving publicly available data to the general astronomy community, while the DESI team has restricted access to proprietary data. Furthermore, the DESI effort is centered around the instrument's primary cosmology mission. In contrast, the Astro Data Lab aims to maximize the scientific output of the community including all topics in astronomy and/or cosmology.

Toward a Network of Science Platforms

Other astronomy science platforms are either active or being developed, and all rely on Project Jupyter to enable their users to perform data proximate analysis. For instance, the Space Telescope Science Institute (STScI) is utilizing Project Jupyter by 1) curating Jupyter notebooks as training materials[24] for astronomy mission data services; 2) building services for science users to discover notebooks and utilize Jupyter-based tools such as the `Jdaviz`[25] package of astronomical data analysis visualization tools, and 3) deploying science platforms in the Amazon

---

[24]https://github.com/spacetelescope/notebooks
[25]https://jdaviz.readthedocs.io/en/latest/

Web Service (AWS) that provide these notebooks and compute resources to users in proximity to public mission data. STScI develops these platforms to support community engagement through workshops as well as through archive services, for example as an entry point for users to interact with data from the upcoming Roman Space Telescope mission.

Another example is the Canadian Astronomy Data Center (CADC[26]), which is currently developing a Jupyter Notebook server to be deployed on the Canadian Advanced Network for Astronomical Research (CANFAR[27]) in 2021, and which will utilize authentication and storage (VOSpace[28]) protocols and group permission already implemented at CADC, in addition to providing users with access to data archives, and software.

Also in development, the Vera C. Rubin Observatory uses JupyterLab as the core technology of the Notebook Aspect of its Rubin Science Platform[29] (RSP). The Rubin Observatory, currently under construction, will feature a wide-field 8.4-meter telescope, a gigapixel camera, and a data management system capable of handling the hundreds of petabytes that its 10-year Legacy Survey of Space and Time (LSST) will collect. The Notebook Aspect will be the key interface allowing the Rubin user community to access the massive dataset and run the LSST science pipelines. Other major projects in development planning their own science platforms include the Maunakea Spectroscopic Explorer[30], an 11-meter spectroscopic survey telescope, and the U.S. Extremely Large Telescope Project, which will have a share of both the Thirty Meter Telescope[31] and the 25-m Giant Magellan Telescope[32].

Challenges that may limit the adoption of science platforms include keeping the barrier of entry as low as possible for onboarding new users with no prior experience, the potential cost or agency limitations if platforms rely on commercial cloud solutions, and the complexity or heterogeneity of datasets and services. The latter can be exacerbated if each platform deploys different, non-standard solutions. A tighter communication across platforms and disciplines including sharing experiences, designing for inter-operability, and discussing and adopting common core technologies (e.g., Jupyter, Docker, Kubernetes) would alleviate some of these expected challenges.

While the pace of technological upheaval in software and data science is rapid, Jupyter has clearly discovered a need in the realm of astronomical research for a flexible, extensible, and easily deployable interactive software environment, and thus is well-positioned to remain a core technology. This need is likely to expand as astronomical data volumes grow further. Indeed, it is not difficult to imagine a future in which the bulk of astronomical data analysis moves to the Cloud, in part powered by Jupyter, perhaps through a system as that envisioned by the Science Platform Network ([8]).

CONCLUSION

In this paper, we considered how Jupyter can play a central role in astrophysical science platforms by connecting large volumes of data with a local analysis host. Namely, we presented two example cases that employ a Jupyter server to provide researchers and students with access to Petabytes of data, with pre-installed software for advanced data analysis, and with tutorials. We first described NOIRLab's Astro Data Lab, which is operating based on a mission to maximize the scientific outcome of the astronomy community from publicly-available wide-field astronomical surveys.

Beyond performing research, Astro Data Lab's efforts extend to training users and contributing to education in astronomy and data science. We then presented how the DESI team utilizes Jupyter for a range of activities such as survey preparation, software development, research, training, and education. Online science platforms share the common goal of co-locating software, tools, and tutorials with the data. This is increasingly needed as datasets grow both in volume and in complexity.

Science platforms can be leveraged to improve on the robustness and on the reproducibility of research results by making entire analysis workflows available. For instance, Jupyter Notebooks

[26] http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/
[27] https://www.canfar.net/en/
[28] https://www.ivoa.net/documents/VOSpace/
[29] https://nb.lsst.io
[30] https://mse.cfht.hawaii.edu
[31] https://www.tmt.org
[32] https://www.gmto.org

could be added to (or cited by) peer-reviewed journal publications. In addition to posting Notebooks on an open source platform such as in a GitHub repository, one can create a Release for the repository, and register a persistent digital object identifier (DOI) with the open repository Zenodo[33]. This approach ensures the longevity of published Notebooks. Additional coordination between science platforms and professional journals may further strengthen this good practice.

Another interesting future outlook is the opportunity to facilitate collaboration between researchers. In particular, users have already often expressed their interest in being able to collaboratively write Jupyter notebooks in real-time as part of online platforms. This may be achieved with recent, ongoing developments, such as Jupyter Real-Time Collaboration (Jupyter-RTC). Such enhanced capability would further strengthen the impact of Jupyter in accelerating the progress of astrophysical discoveries.

Taking an even broader perspective, there are many parallels to draw with other scientific disciplines where similar data proximate computing platforms are developed and operated. For instance, CyVerse[34] is an interdisciplinary platform heavily used for biology and bioinformatics. While the type of data differs (e.g., genome sequence rather than stars and galaxies), there are similar needs in terms of filtering, analyzing, and visualizing data. Other examples include Pangeo[35], a platform for open, reproducible, and scalable geoscience. Many of the challenges faced and solutions adopted by these projects are transferable to astrophysical data science, creating a wealth of opportunity for future cross-disciplinary collaboration.

APPENDIX

We introduce how to use the Data Lab Clients within a Jupyter Notebook, which were illustrated in Figure 2. First, the authClient works by creating a token to recognize that a user is properly authenticated, which is required to access virtual storage (VOSpace) and personal database space (MyDB). This step can be skipped if a user does not need to access those in a given Notebook, or

if the user has recently authenticated as the token persists in the account.

```
from dl import authClient as ac
from getpass import getpass

token = ac.login(
        input("Enter user name:"),
        getpass("Enter password:"))
if not ac.isValidToken(token):
    raise Exception("Token is not valid.
    Please check your usename/password and
    execute this cell again.")
```

Secondly, we show how the queryClient can be used to send a SQL (and ADQL,[36] Astronomical Data Query Language) query and retrieve the results in a Jupyter Notebook, formatted as a pandas DataFrame in the example below, and run synchronously (the optional keyword `async_` can be set to True to run the query asynchronously, in which case a jobId is returned, which can be used to check on the status of the query, and retrieve the results after the query has completed).

```
from dl import queryClient as qc

query = "SELECT * FROM ls_dr8.tractor
    LIMIT 1000"
df = qc.query(sql=query, fmt='pandas')
```

Thirdly, we show how the storeClient can be used to store files to and retrieve from VOSpace. VOSpace is an implementation of network-attached storage, and each user account comes with a dedicated storage area. The result of a query can be written out directly to VOSpace:

```
res = qc.query('SELECT * FROM
    ls_dr8.tractor LIMIT 1000',
    out='vos://lsdr8_sample.csv')
```

The file can be later retrieved:

```
from dl import storeClient as sc
from dl.helpers.utils import convert

res = sc.get('vos://lsdr8_sample.csv')
df = convert(res,'pandas')
```

To share files with others, users can place files within the `public/` subdirectory in their VOSpace, e.g.

```
sc.mv(fr='vos://lsd8_sample.csv',
    to='vos://public/lsdr8_sample.csv')
```

Another user can then access it, and convert the content to, e.g., a pandas DataFrame:

```
res = sc.get(
    'otheruser://public/lsdr8_sample.csv')
df = convert(res,'pandas')
```

[33] https://zenodo.org/
[34] https://www.cyverse.org/
[35] https://pangeo.io/

[36] https://www.ivoa.net/documents/ADQL/

ACKNOWLEDGMENT

■ REFERENCES

1. M.J. Fitzpatrick, K. Olsen, F. Economou, et al., "The NOAO Data Laboratory: a conceptual overview," *Society of Photo-Optical Instrumentation Engineers (SPIE)*, vol. 9149, Proc. SPIE id. 91491T, 2014.

2. R. Nikutta, M. Fitzpatrick, A. Scott, and B.A. Weaver, "Data Lab: A community science platform," *Astronomy and Computing*, vol. 33, p. 100411, 2020.

3. DESI Collaboration et al., "The DESI Experiment Part I: Science, Targeting, and Survey Design," arXiv:1611.00036, 2016.

4. DESI Collaboration et al., "The DESI Experiment Part II: Instrument Design," arXiv:1611.00037, 2016.

5. Astropy Collaboration et al., "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package," *The Astronomical Journal*, Volume 156, Issue 3, article id. 123, 19 pp., 2018. (journal)

6. D. Norman, "Can Big Data Lead an Inclusion Revolution?," *AstroBeat; Astronomical Society of the Pacific,* No. 162, 2018.

7. K. Olsen, A. Bolton, A., S. Juneau, S., et al., "The Data Lab: A science platform for the analysis of ground-based astronomical survey data," *Bulletin of the American Astronomical Society,* Vol. 51, p. 61, 2019.

8. V. Desai, M. Allen, c. Arviset, et al., "A Science Platform Network to Facilitate Astrophysics in the 2020s," *Bulletin of the American Astronomical Society,* Vol. 51, p. 146, 2019.

9. D. L. Nidever, K. Olsen, A. R. Walker, et al. 2017, AJ, 154, 199. doi:10.3847/1538-3881/aa8d1c

10. Y. Choi, D. L. Nidever, K. Olsen, et al. 2018, ApJ, 869, 125. doi:10.3847/1538-4357/aaed1f

11. N. F. Martin, D. L. Nidever, G. Besla, et al. 2015, ApJL, 804, L5. doi:10.1088/2041-8205/804/1/L5

Stéphanie Juneau, Associate Astronomer at NSF's NOIRLab and Data Scientist at the Astro Data Lab (in Tucson, AZ). Dr. Juneau received a Ph.D. in Astronomy from the University of Arizona. Her scientific research is focused on galaxies and supermassive black holes. She is interested in applying statistical methods, advanced visualization techniques, and in improving astrophysical science platforms. She is a member of the DESI and Euclid collaborations, and of the American Astronomical Society (AAS). Contact her at stephanie.juneau@noirlab.edu.

Knut Olsen, CSDC Principal Data Scientist at NSF's NOIRLab (in Tucson, AZ). Dr. Olsen received his Ph.D. in Astronomy from the University of Washington. He works on stellar populations in nearby galaxies. Contact him at knut.olsen@noirlab.edu.

Robert Nikutta, Project Scientist for Astro Data Lab at NSF's NOIRLab (in Tucson, AZ). Dr. Nikutta received his Ph.D. from the University of Kentucky. His scientific interests range from the nuclear regions of Active Galactic Nuclei to astroinformatics and big data

research. Contact him at robert.nikutta@noirlab.edu

Alice Jacques, Data Analyst at NSF's NOIRLab and Astro Data Lab (in Tucson, AZ). Alice received her B.S. degree in Physics and Computer Science from Georgia College and State University and her M.S. degree in Physics from the University of Louisville. Contact her at alice.jacques@noirlab.edu.

Stephen Bailey, Senior Software Developer in the Physics Division at Lawrence Berkeley National Lab (in Berkeley, CA). Dr. Bailey received his Ph.D. in Physics from Harvard University and currently leads the DESI Data Management team. He enjoys converting raw data into useful data for the benefit of large science collaborations, and has been doing so for over 25 years. Contact him at StephenBailey@lbl.gov.