Data Lab FAQ

What is the Data Lab?

The Data Lab is a new NOAO facility to allow researchers to work, in an efficient and well-integrated way, with datasets too big and unwieldy for typical desktop computers. The Data Lab will allow users to:

- Explore the contents of large catalogs and their relationship to images, spectra, and other large datasets.
- Experiment with workflows to perform quantitative analysis on large datasets.
- Collaborate with groups of other users through all stages of analysis using a common workspace.
- Share derived data products and results with the entire astronomical community.
- Export those workflows to large scale computing facilities as necessary.

Is the Data Lab the new NOAO Archive? Is it a new computing center?

No. Although the Data Lab will provide some storage and compute resources, as well as access to other public datasets hosted at NOAO, it works in partnership with the NOAO Science Archive (NSA) responsible for the raw and pipelined image data. The Data Lab's proximity to the NSA and large catalogs will give it the fastest possible data access for workflows using its computing resources.

How do I access the Data Lab?

The Data Lab will consist of a number of components (published data services, virtual storage, analysis and visualization tools, etc). These will all be accessible, to some degree, using web browsers, desktop applications, and scriptable interfaces. Additionally, shell-level access will allow power users a command-line interface for custom analysis.

What types of data can I analyze with the Data Lab?

The initial focus of the Data Lab will be DECam catalogs and images. However, the Data Lab will have the capability to provide access services for general catalog, image and spectral data, including data that users choose to upload for use in their personal storage. Time-series, data cubes, and event data are some examples of what can be included as part of a workflow. In some cases, users will want to import custom analysis code to use these additional data.

What types of analysis can I do with the Data Lab?

The Data Lab will provide a core set of tools for data manipulation, visualization and catalog and pixel-level analysis to augment the code the astronomer might import for their particular science interest. The number of science tools will grow over time and will include tools developed by users, either using core Data Lab tools as a base or by extending other users' applications with new capabilities. Scriptability and batch execution will allow for *workflows* to be created if needed.

Can I use my own code in the Data Lab?

Definitely. Users will be able to configure custom development or runtime environments in their user shell to run interactive applications as they would on their desktop. The difference between running on the desktop vs. the Data Lab is that the Data Lab will take advantage of faster access to the data services and provide additional compute resources. These same custom environments will be able to be packaged with tasks deployed for use in batch processing of the data at other computing facilities or on the user's desktop.

Can't I do this already on my desktop machine?

Up to a point, yes. However, PI observing programs can easily generate several TB of imaging data or catalogs with tens of millions of objects, and survey programs produce significantly more. However, data transfer times or computationally expensive algorithms will in many cases reach the limit of what common desktop systems can handle efficiently. For certain types of analysis, the Data Lab will offer the *option* of faster data access and more computing power for more efficient processing of your data.

Can't I do this (and more) at a large super-computing center?

Certainly, but not everyone has ready access to these centers, and jobs too big for your desktop won't necessarily require that scale of computing resources. The Data Lab will offer:

- Additional computing power for those users and science cases that need it
- An environment for doing the open-ended exploration, experimentation, and collaboration that is needed in the early stages of projects involving large datasets
- A platform for developing and refining detailed workflows at smaller scale before moving them to large computing facilities

Can I share my results with collaborators?

The Data Lab's virtual storage will allow users to create groups of other users that can have shared access to their data. Members of a survey team will be able to use shared access to a centralized data store to divide the work (e.g. nightly runs) between members or to share results. Secured access to databases may similarly be shared within a group.

Can I make my results publicly available?

Once a final set of data products is produced, the Data Lab will provide a publishing mechanism to make the data (catalogs, images or spectra) available using standard Virtual Observatory (VO) protocols. These publishing tools can be run on the user's own machine to share the data, or in some cases the Data Lab can host the data as a more permanent service.

Do I have to upload all of my local data to use it in the Data Lab?

Not always. Some components of the Data Lab will be exportable and able to be installed on your local machine to reproduce the same functionality. For example, processing tools or virtual storage could be installed to work with local data, yet still allow

workflows to remotely access large catalog data that may be needed in the workflow. Data Lab Virtual Machines will be available to download and run on your own hardware, or pre-packaged containers of specific components may be used on your machine (or in the cloud) when needed.

Can I still get my KPNO/CTIO data from the NSA and work entirely on my desktop?

Sure, and for some programs this may still be the best approach. Use of the Data Lab is entirely optional.

When will the Data Lab be available to use?

We are still in a design and ramp-up phase. However, we expect to have a Science Demo available by the January 2016 AAS meeting and a first public release by 2017. Parts of the system may become available in a soft roll-out before then as we work with collaborators on the public release.

The Data Lab project sounds really ambitious. Will it really happen?

Ambitious, yes, but achievable in large part because we can adapt and extend existing code to build much of the needed infrastructure (e.g. virtual storage and data services) and focus on integration of the system components. With this foundation, science capabilities will grow over time as user-developed code becomes available. The project is moving towards an external review to ensure we deliver a credible project plan, but during system design we haven't identified any major technical challenges that would keep us from succeeding.

Can I/my group get involved with Data Lab development?

All Data Lab software will be completely Open Source and we are happy to discuss potential collaborations or how we might be able to help your project.

Where can I get more information?

The project website is available at

http://datalab.noao.edu

This site will be updated as the project evolves. Interested users should also contact Mike Fitzpatrick (fitz@noao.edu) or Knut Olsen (kolsen@noao.edu).