Birds of a Feather session on Science platforms

William O'Mullane[1], Megan Sosey[2], Hassan Siddiqui[4], Gregory Dubois-Felsmann[3], Gerard Lemson[5], Christophe Arviset[6], Mike Fitzpatrick[7], Ivelina Momcheva[2], Sebastien Fabbro[8], Brian Major[8],
[1]*Large Synoptic Survey Telescope, Tucson, AZ, USA;* `womullan@lsst.org`

[2]*Space Telescope Science Institute*

[3]*IPAC, California Institute of Technology, Pasadena, CA, U.S.A.*

[4]*Vega for Gaia/ESAC*

[5]*The Johns Hopkins University*

[6]*European Space Astronomy Centre*

[7]*NOAO*

[8]*CADC*

Abstract. How users will interact with data in the future is always unclear. Currently we see Jupyter Notebooks or JupyterLab emerging in many places as the way forward for one aspect of this. This BoF explored some topics around providing and environment for doing science.

1. Introduction

It seems timely to consider how we might offer users a smoother experience as they move between data providers. Current VO services allow one to send queries to multiple centres but in the notebook environment one may wish to do something more sophisticated. We should consider whether users can send requests from one centre to another or whether the same notebooks should be runnable in different centres. How do we deal with batch processing - large jobs? How do we manage resources/quotas (disk/memory/cpu)? How can we enable users to share their work (both notebooks and data) and create ad-hoc scientific collaborations? We had a few short presentations:

- LSST Approach (Dubois-Felsmann): science platform will give access to the data and visualization tools and documentation, it will allow collaboration and allow for added value processing close to the data using Jupyter. A question caused clarification that you could write C++ or any other language in that system.

- SciServer Approach (Lemson): SciServer is format agnostic storage with extensible tools (query and analysis), it allows hosting and sharing datasets. Near data access is provided with Jupyter. Notebooks can be executed in batch mode but no MPI.

- European Space Science Data Centre (Arviset):Science Exploitation and Preservation Platform at ESAC intends to enable data processing where the data (Jupyter), also for small data and for legacy software.
- NOAO approach (Fitzpatrick): Datalab provides full sky exploration of catalogs and local dataspace, it allows workflows to run close to the data. Providing Jupyter and legacy code execution.
- STScI DSMO (Momcheva): increase science output from holdings, shorten turn around time, connect multi wavelength, considering a Jupyter hub system deployed on Amazon.
- CADC (Fabbro):raw OpenStack portal with vanilla VMs, some projects using Jupyter, intending to containerize.
- CADC/IVOA (Major):IVOA and remote computing grid and web services working group, working with knowledge discovery group to define use cases. Goal is fast interoperable computing services close to the data, support Machine Learning.

Q: vision for provisioning resources (disk space, cpu time... scaling into the future)
A: groups provide their own hardware, example: JHU, you can create groups and decide who has access tot he data and how much is provisioned. Right now they are in the process of building trial functionality, they aren't checking what the actual goal of the computations are, but they don't want to be a generic compute center

2. Principles

A few good principles were mentioned:

- multiple entry points into the system (web, notebooks, cmdline tools, scripting apis)
- language agnostic (python flask micro-services architecture, restful interface)
- enable user developed tools
- established standards with hidden complexity for friendly interfaces
- provide access to external data/services vs local ingest