USER TOOLS FOR LEVEL 1 DKIST DATA

Stuart J. Mumford[2,1], Fraser Watson[1], Alisdair Davey[1]

1. National Solar Observatory, 2. University of Sheffield

DKIST Level 1 Data

The DKIST Data Centre will be providing level one calibrated data for download by the scientific community. In addition to this a set of Python tools are being developed to facilitate the use of these data.

Due to the very high data rates for level 1 data (over 2 Petabytes per year) these data provide a new set of challenges for both the data center and the users of the data. The user tools aim to provide easy access to the level one data in Python.

The main features of the user tools will be:

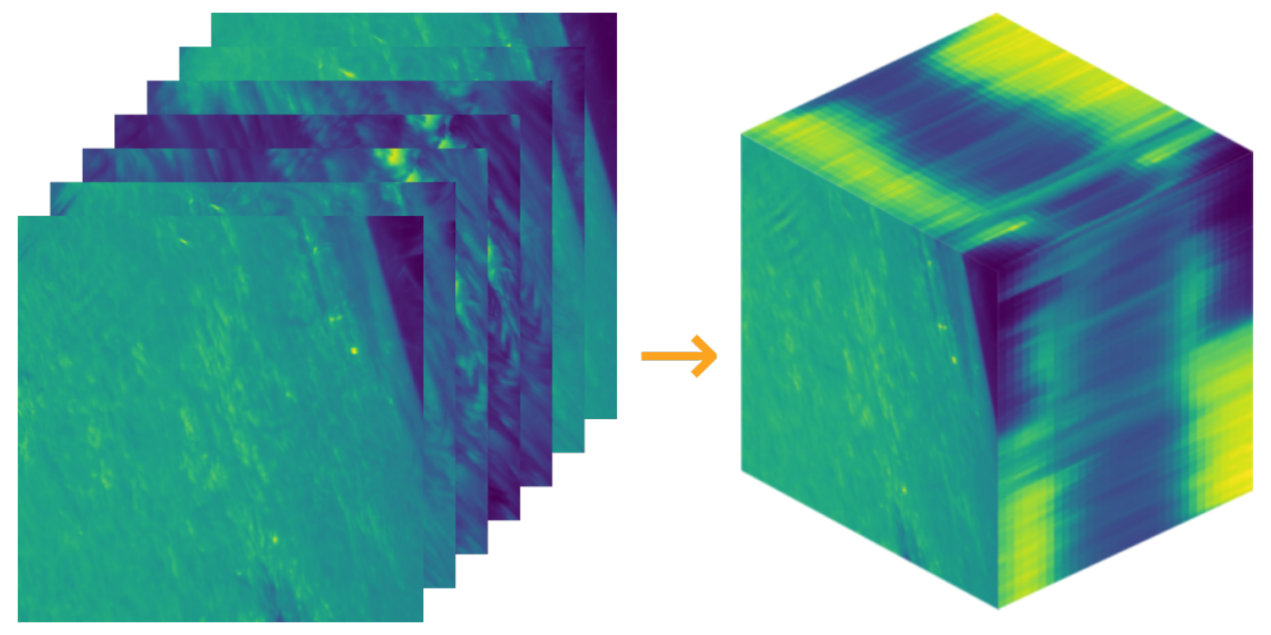
- Dataset search.
- Dataset download.
- Loading of datasets.
- Providing coordinate aware representation of datasets.
- Enabling the use of the Scientific Python ecosystem on DKIST level 1 data.

DKIST Datasets

Level 1 DKIST data will be provided as "datasets" which divide the data up into observations from one instrument and one pass band from one observing program.

These datasets will be divided into many individual FITS files, such that each is a single "calibrated frame", for example for the VISP slit spectrograph each FITS file would be a two dimensional space-wavelength array. This means every dataset will be comprised of many tens or hundreds of thousands of FITS files.

One of the primary functions of the user tools is to expose these individual frames as a contiguous cube without rewriting the data or having to open the files until the array values need to be read.



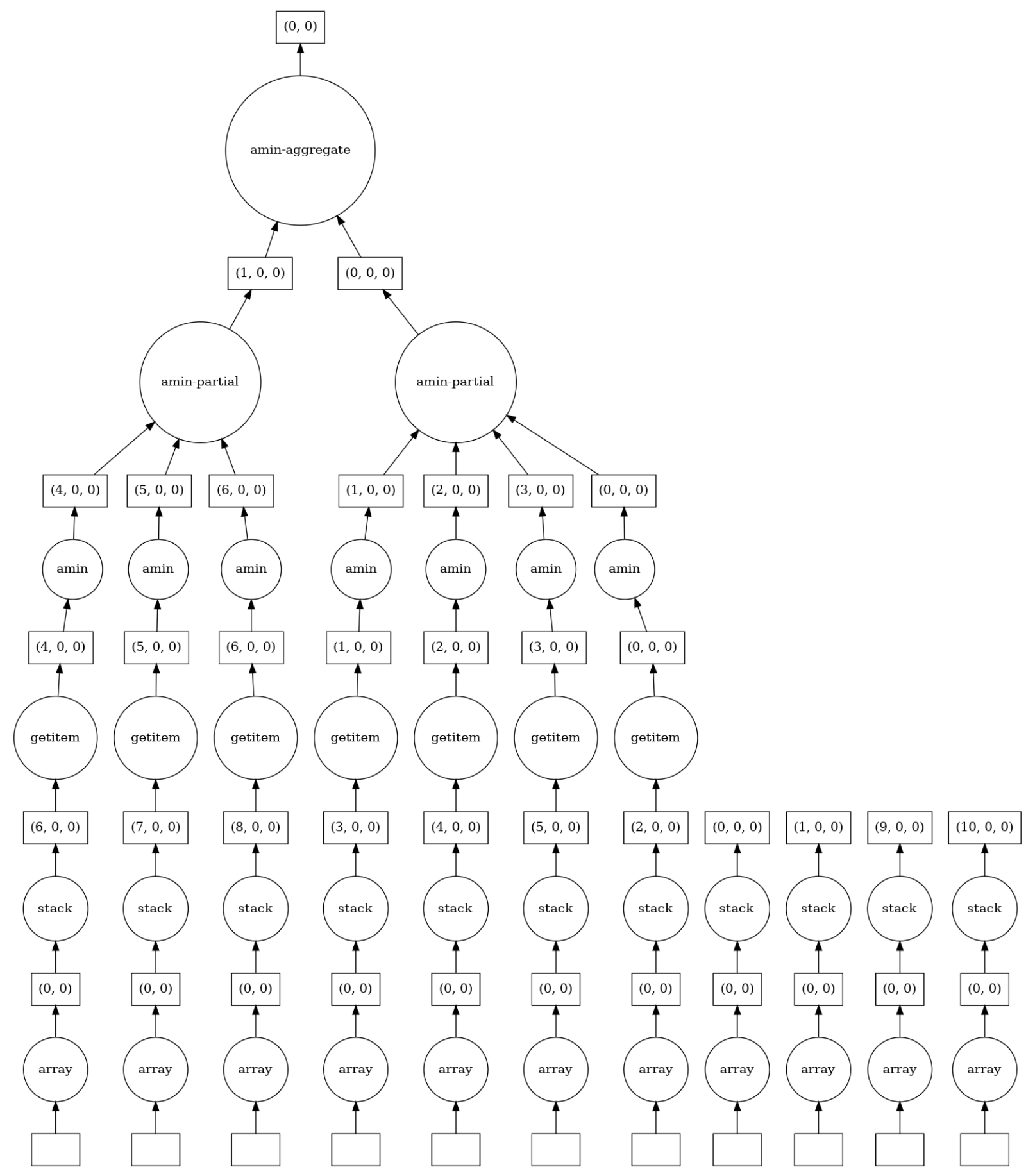
To make this more efficient an Advanced Scientific Data Format (asdf) file containing all the metadata is provided for each dataset. This asdf file contains:

- The ordering information for the FITS files.
- A copy of all the FITS headers for all files in the dataset.
- A World Coordinate System for the whole reconstructed dataset.
- Some additional metadata about the dataset.

Dask

Dask is a Python library for larger than memory and distributed computation. The DKIST user tools are utilising dask to enable the on demand loading of data from many FITS files. Representing the dataset as a Dask array also provides users lots of potential for parallel computation on a variety of hardware from laptops to HPC or cloud computing.

```
>>> arr = arr[2:-2]
>>> arr = np.min(arr, axis=0)
>>> arr.visualize(optimize_graph=False)
```



As shown above, when doing operations on a Dask array, a graph of operations is generated, and only when the values of array are required is the graph executed. This means that complex operations can be done on a large distributed, multi-dimensional dataset, while opening the minimum number of files.

Loading a Dataset

The user tools load datasets from asdf files. If the FITS files referenced by the asdf file are not present, the dataset can still be loaded and the metadata explored.

```
>>> from dataset import Dataset

>>> ds = Dataset.from_asdf("VTF_20450812T090801.asdf")
>>> ds
<dkist.dataset.dataset.Dataset object at 0x7f3245c965c0>
dask.array<stack, shape=(4, 128, 19, 966, 980), dtype=float32, chunksize=(1, 1, 1, 966, 980)>
WCS<pixel_axes_names=(stokes, scan number, wavelength position, spatial y, spatial x),
    world_axes_names=(stokes, time, wavelength, latitude, longitude)>
```

This facilitates the inspection and subsetting of the dataset before having to transfer the large arrays. This can be used to only transfer subsets of the whole dataset, for example, only Stokes I or a specific time window. Only whole FITS files can be transferred as there is no processing before download at the data center.

In the example below we select the 0th stokes profile (I) and the 50th wavelength scan from the dataset:

```
>>> partial = ds[0,50]
>>> partial
<dkist.dataset.dataset.Dataset object at 0x7f1a40215c88>
dask.array<getitem, shape=(19, 966, 980), dtype=float32, chunksize=(1, 966, 980)>
WCS<pixel_axes_names=(wavelength position, spatial y, spatial x),
    world_axes_names=(wavelength, latitude, longitude)>
```

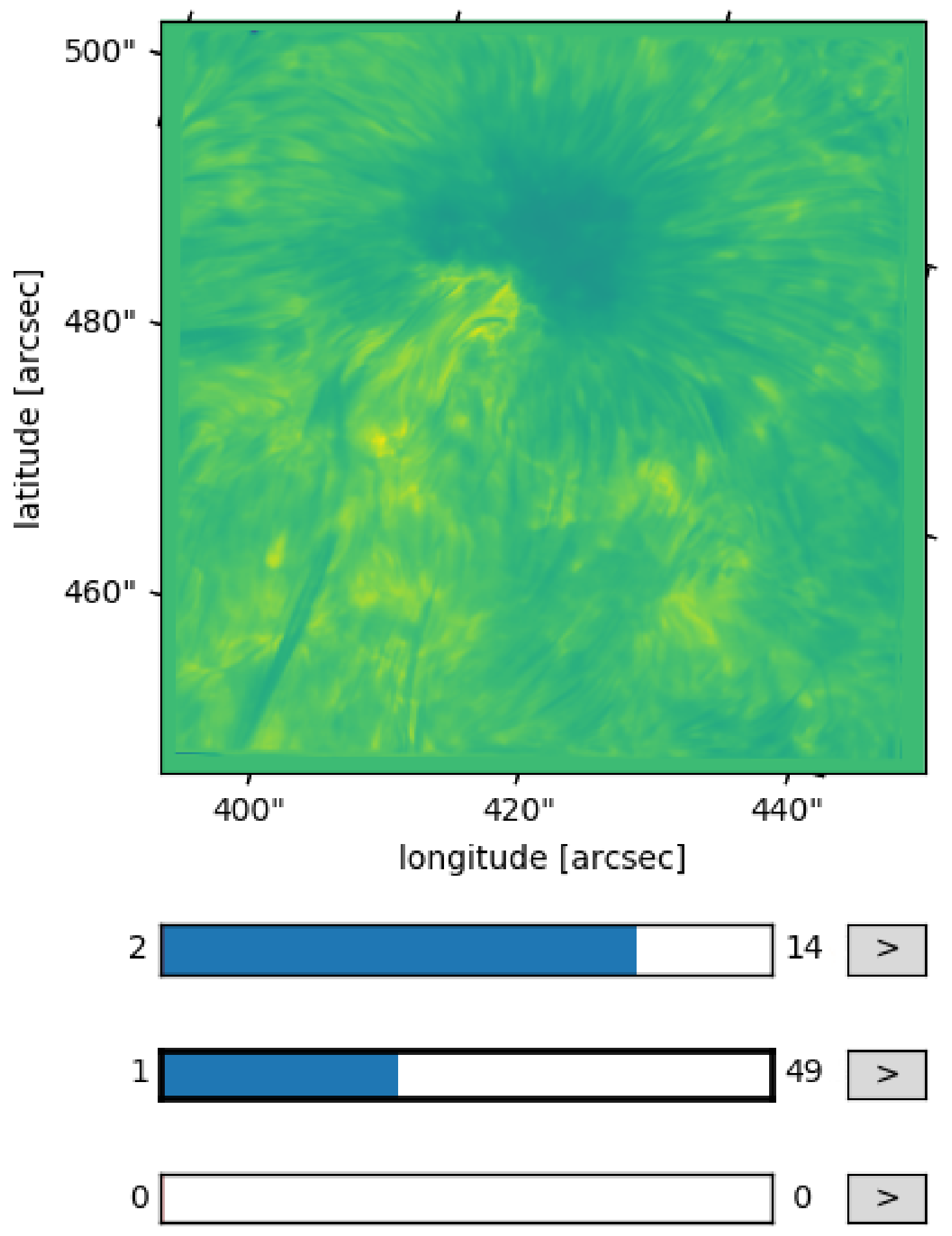
This subset can then be downloaded, only transferring 19 FITS files out of a total of 9728.

```
>>> partial.download()
```

Plotting

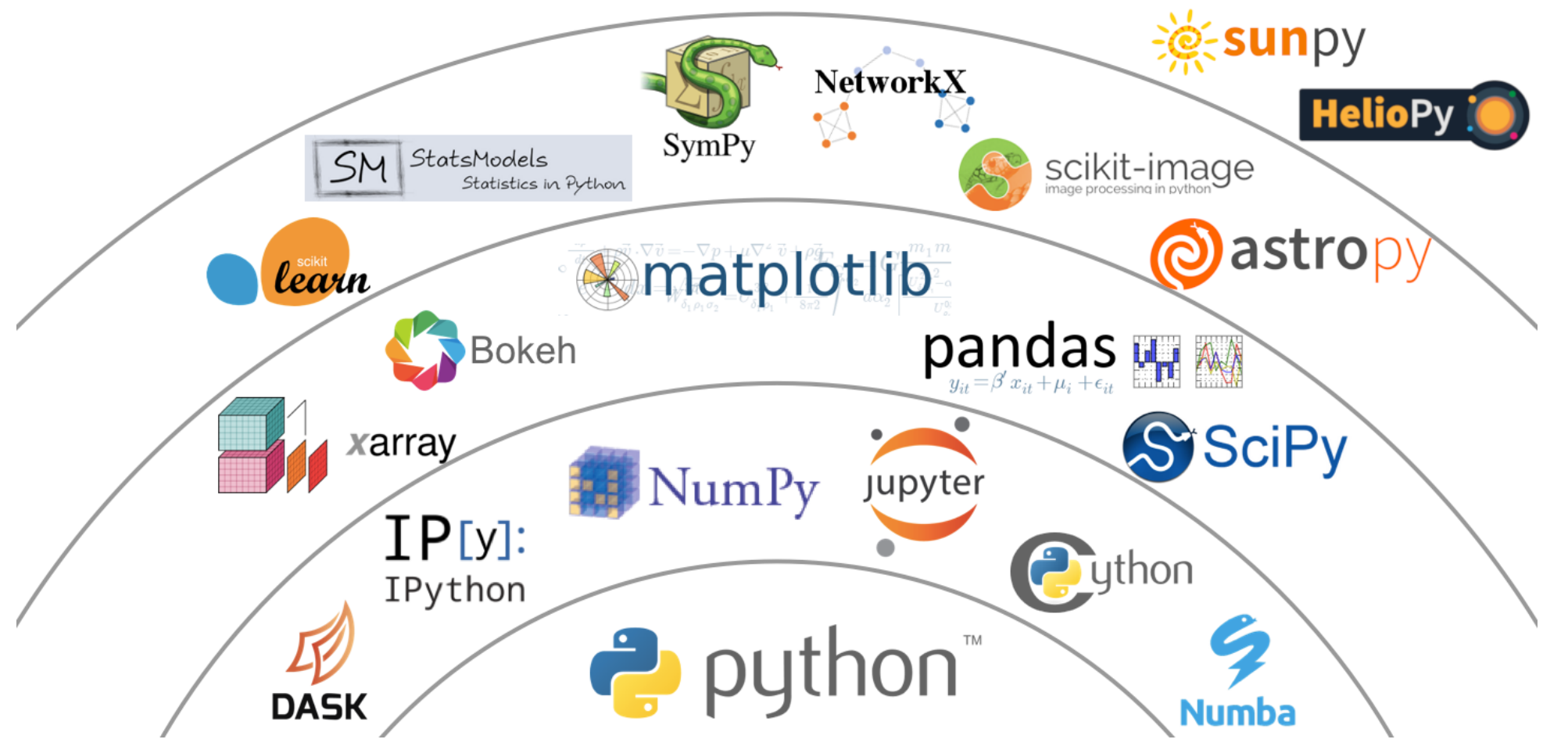
A dataset object is able to display a summary plot for all dimensionalities of data. Here we plot our original five dimensional dataset, and three sliders are displayed, for wavelength, time and stokes profile.

```
>>> ds.plot()
```



Scientific Python

The Scientific Python ecosystem is comprised of a large number of packages which specialise in different functionality. The objective of the DKIST user tools is to enable the use of this ecosystem with DKIST data.



The DKIST user tools uses a large number of these different packages, to provide interoperability with the whole ecosystem the key packages used are Dask for the array, matplotlib for visualisation, SunPy and Astropy for coordinate transformations and units, and gWCS for representation of the world coordinates.