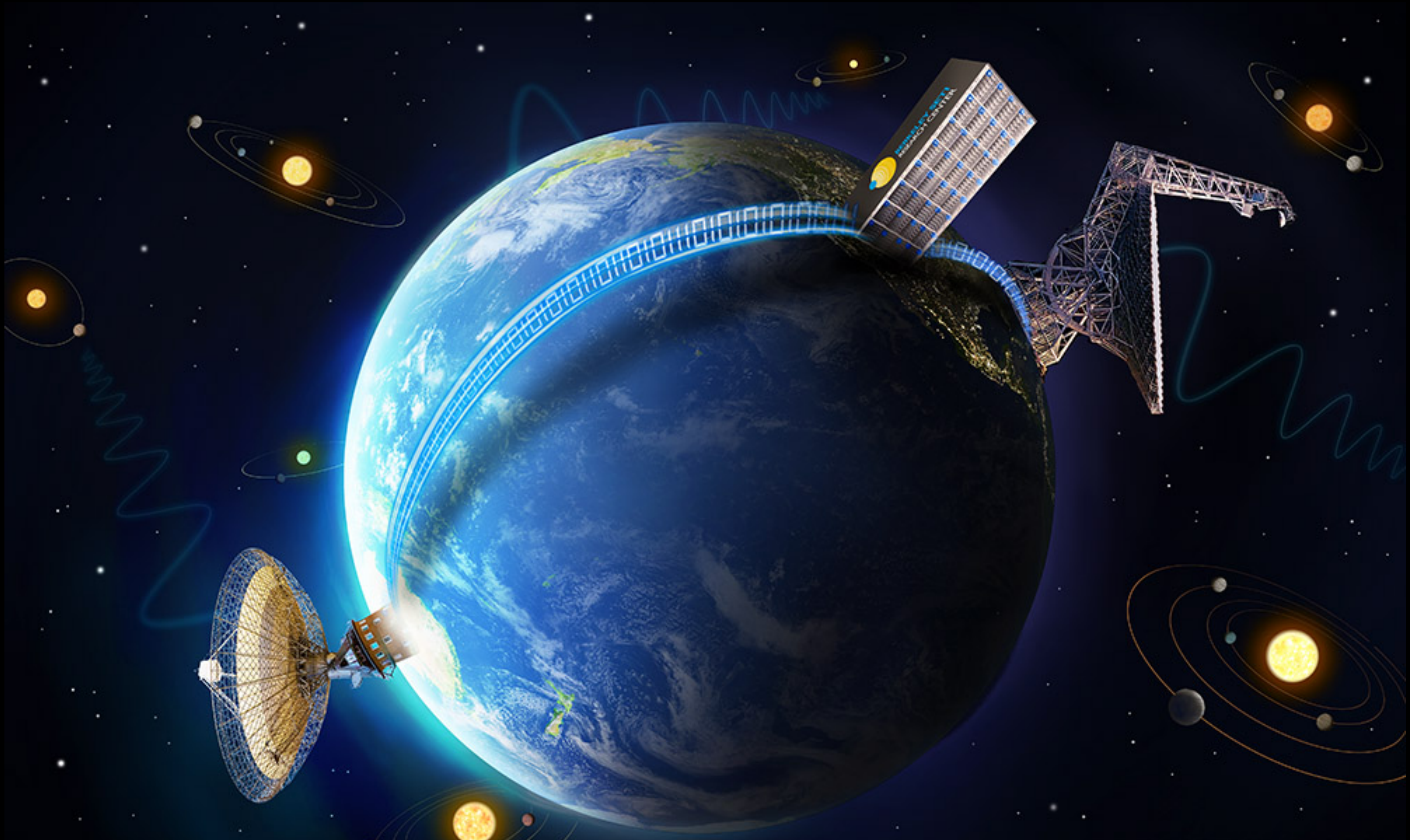
Breakthrough Listen: Finding Life beyond Earth With HDF5



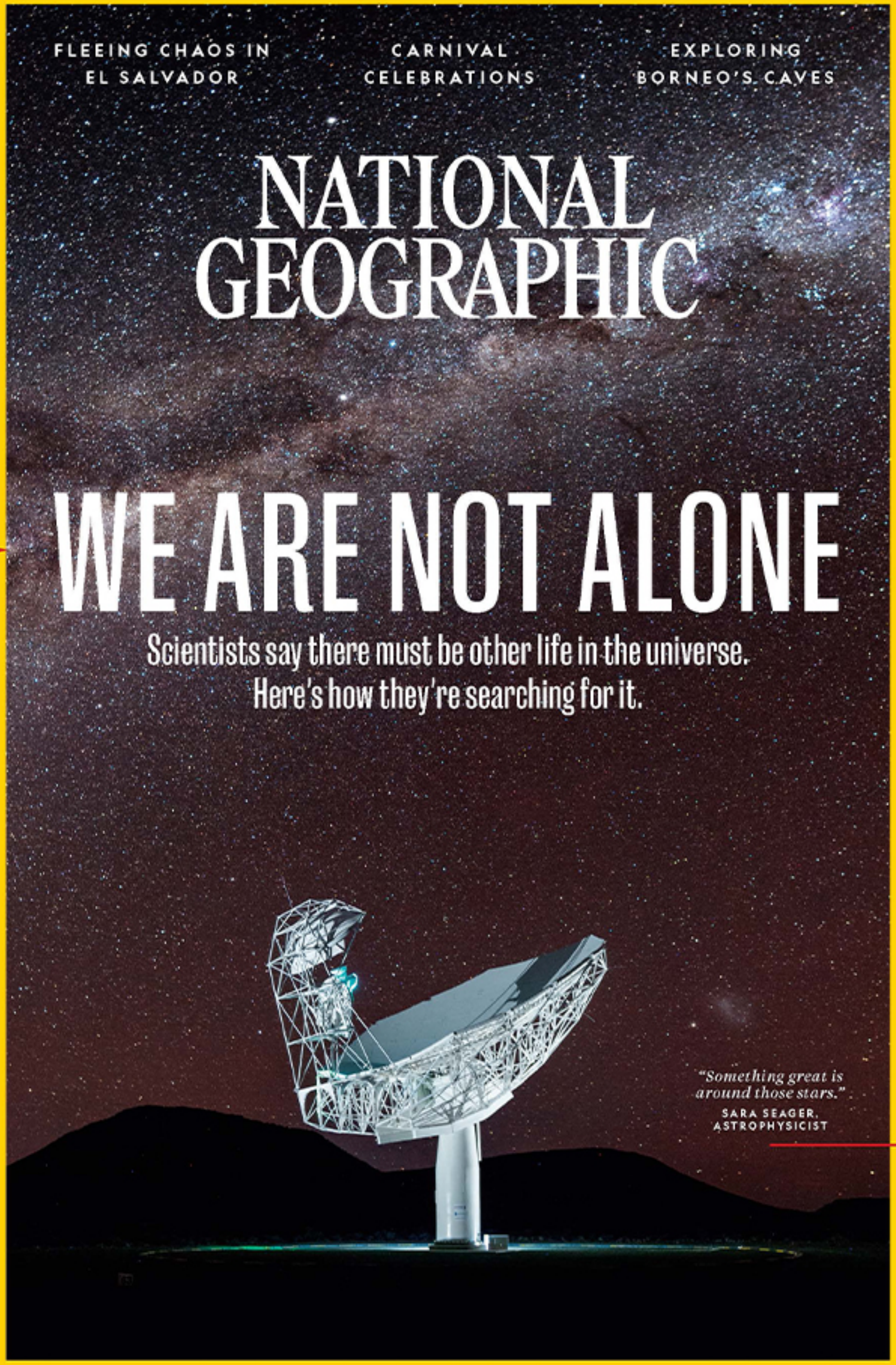
Dr Danny Price, ICRAR / UC Berkeley

European HDF Users Group Summer 2021

BREAKTHROUGH
LISTEN

ICRAR

03.2019


FLEEING CHAOS IN EL SALVADOR
CARNIVAL CELEBRATIONS
EXPLORING BORNEO'S CAVES
NATIONAL GEOGRAPHIC
WE ARE NOT ALONE
Scientists say there must be other life in the universe. Here's how they're searching for it.
"Something great is around those stars."
SARA SEAGER, ASTROPHYSICIST

300,000,000,000

Number of stars in the Milky Way (approx)

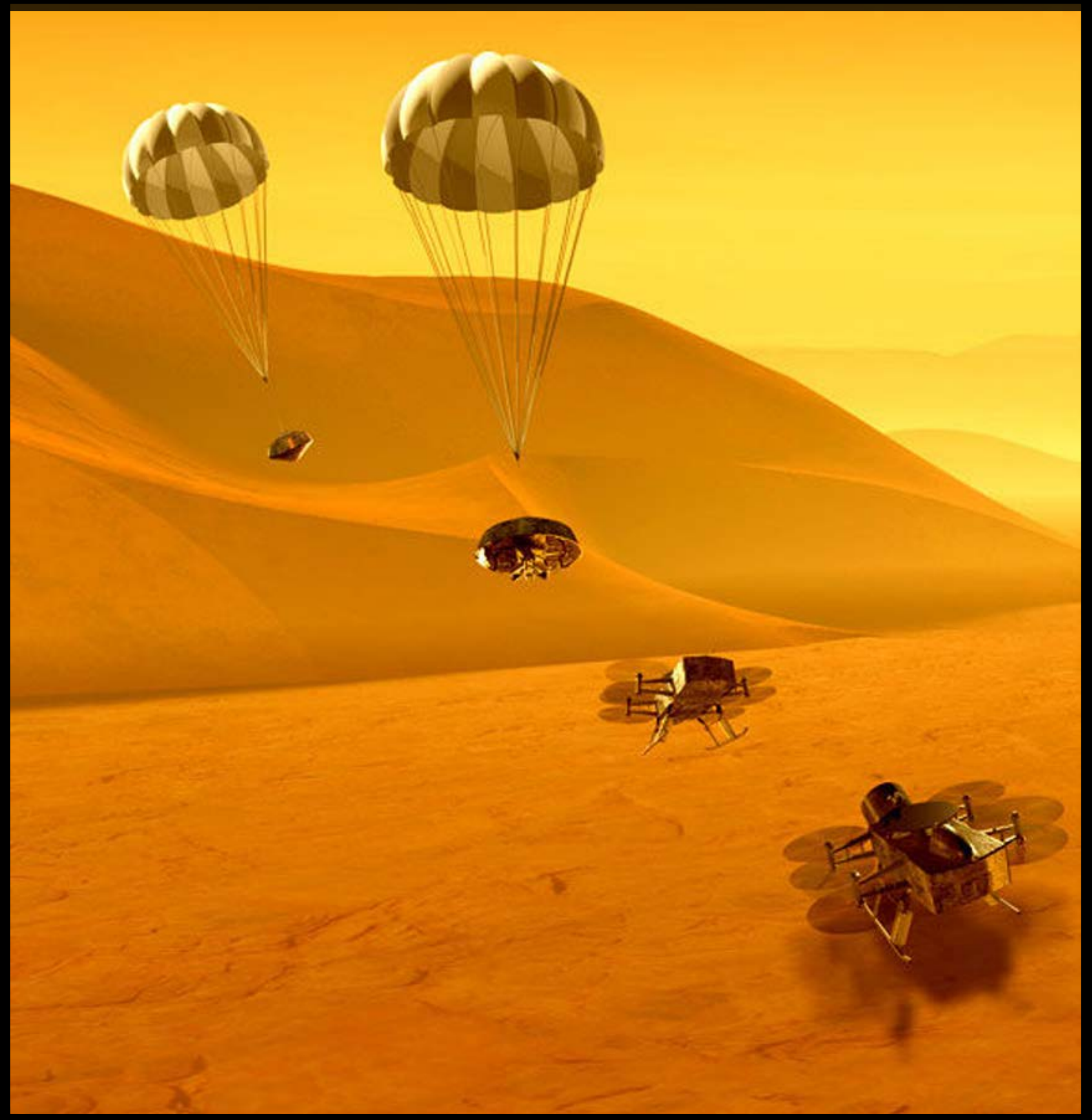
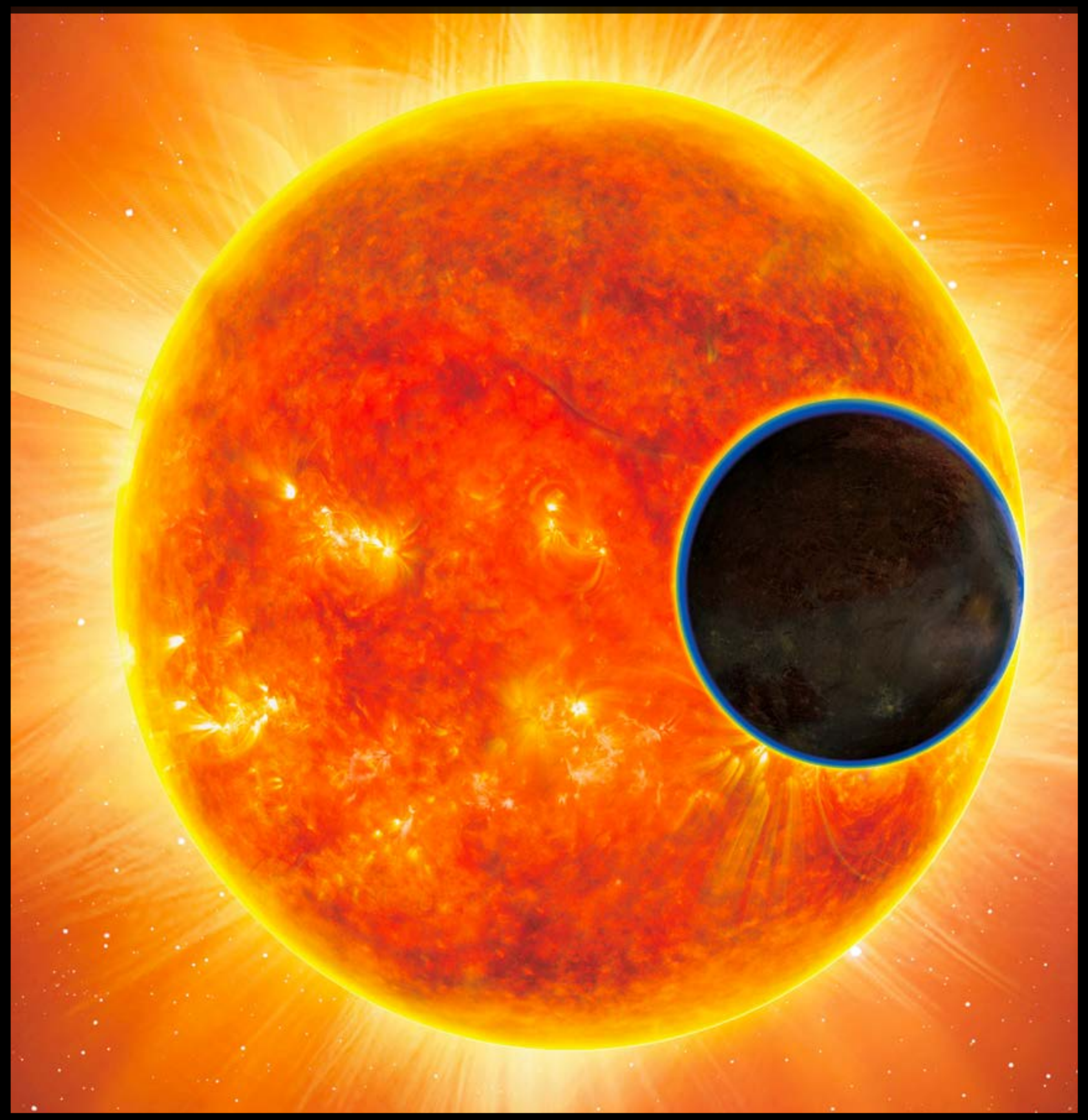
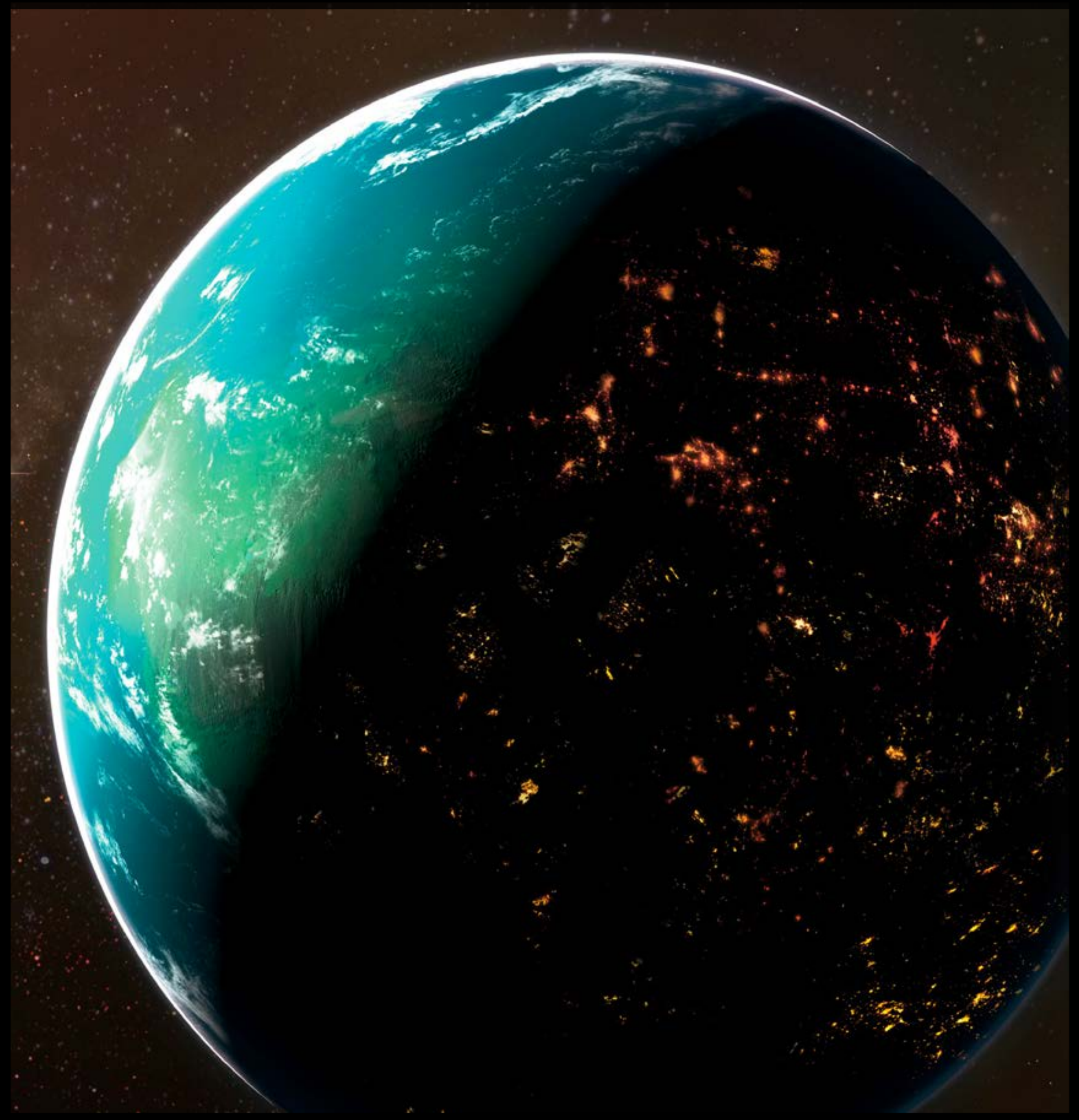
>2,000,000,000,000

Number of galaxies in observable Universe (approx)

>200000000000000000000000
(2 x 10^{23})
Estimated number of stars in the Universe

How do we find life beyond Earth?

How do we find life beyond Earth?

$$$$$	$$$$	$
		
In situ (we go there, boldly)	Atmospheric biosignature (chemical disequilibrium)	Technosignature detection (SETI)
$N_{stars} = 1$	$N_{stars} \sim 10$	$N_{stars} \sim 10^{23}$

Images: NASA, Nat Geo

BREAKTHROUGH
LISTEN

BREAKTHROUGH LISTEN

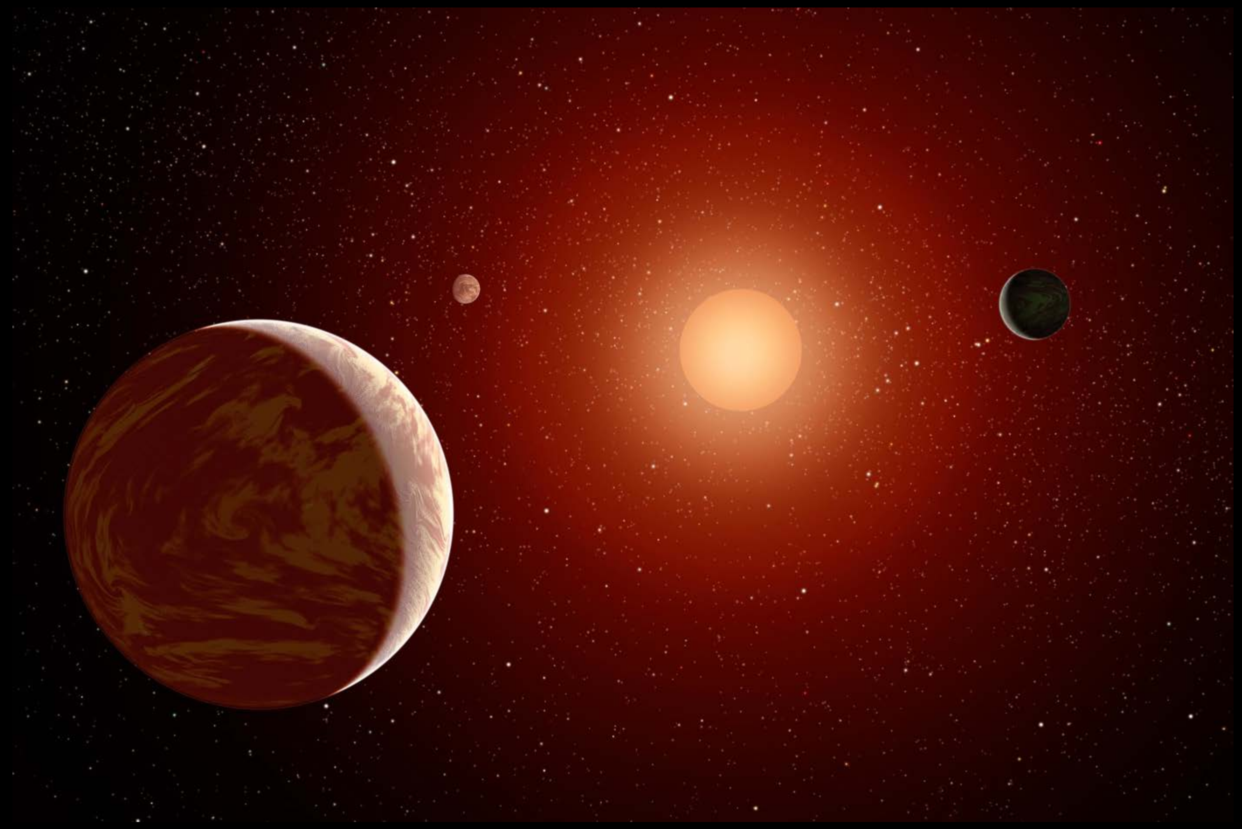
"THE APOLLO PROGRAM OF SETI"

- E. Enriquez

Breakthrough Listen is the largest ever scientific research program aimed at finding evidence of intelligent life beyond Earth.

The Breakthrough Listen Initiative:

Overview

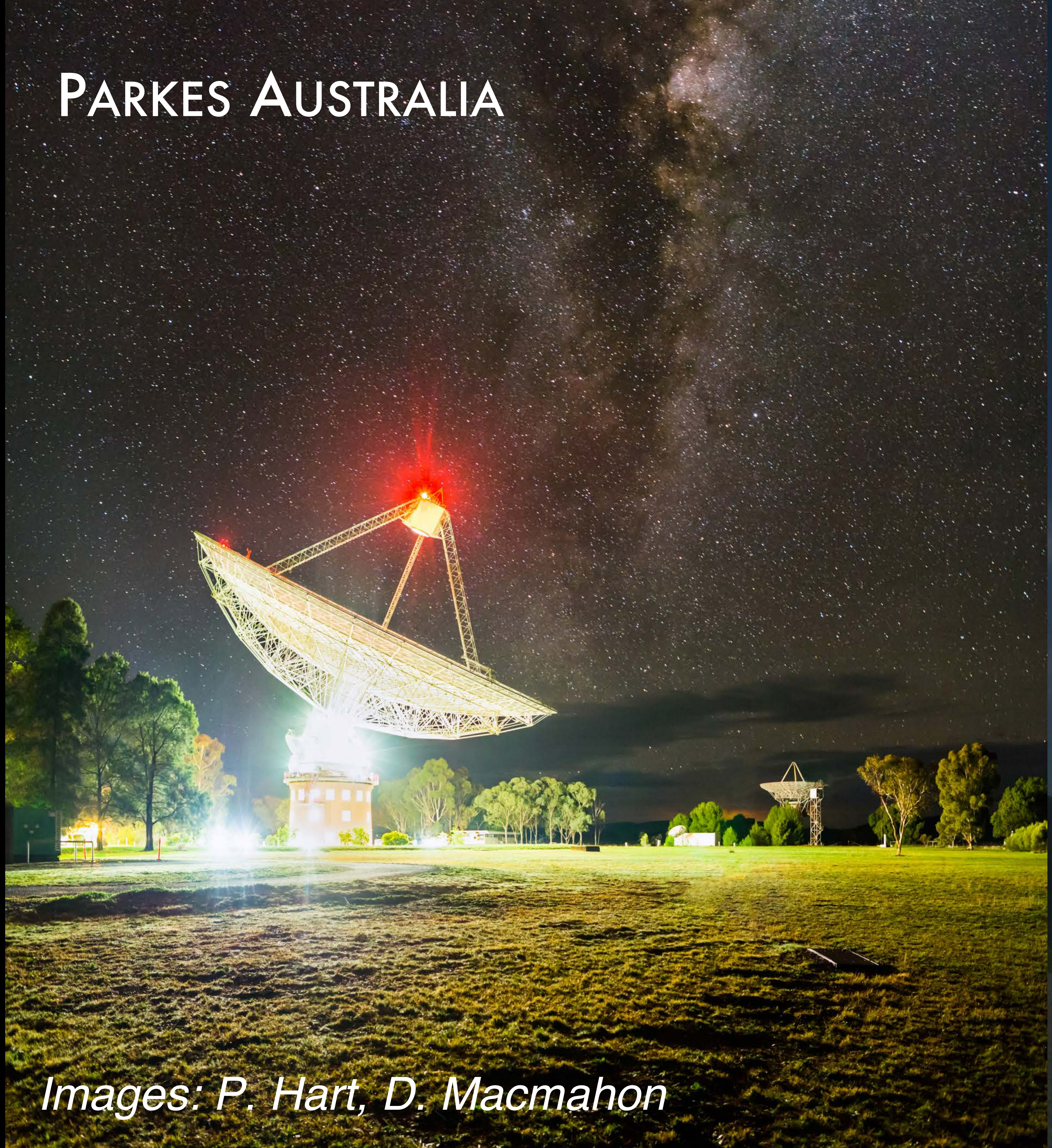


1 Million Stars

MW Galactic Plane Survey

100 Galaxies

Open data and open source

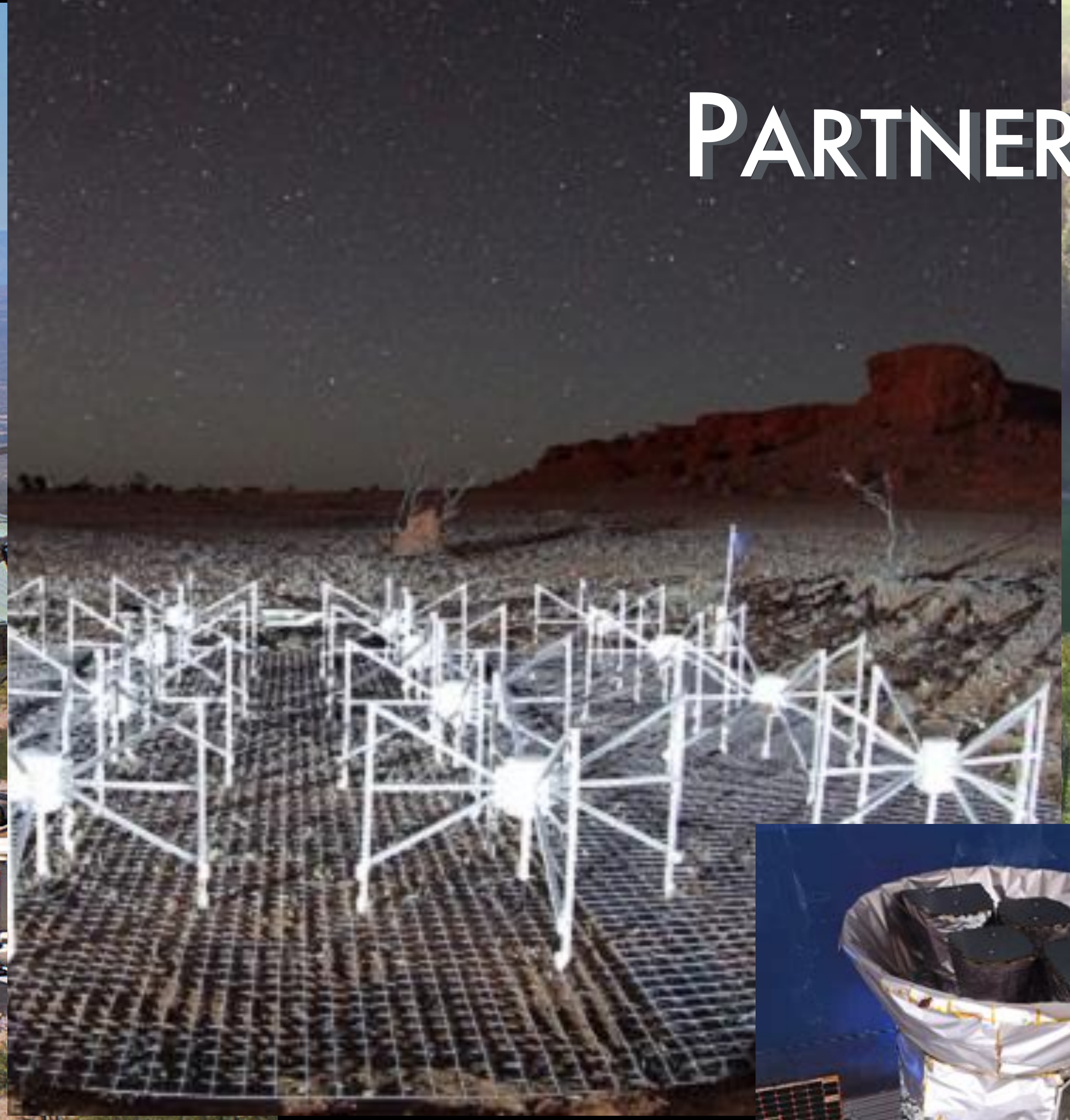

PARKES AUSTRALIA
Images: P. Hart, D. Macmahon

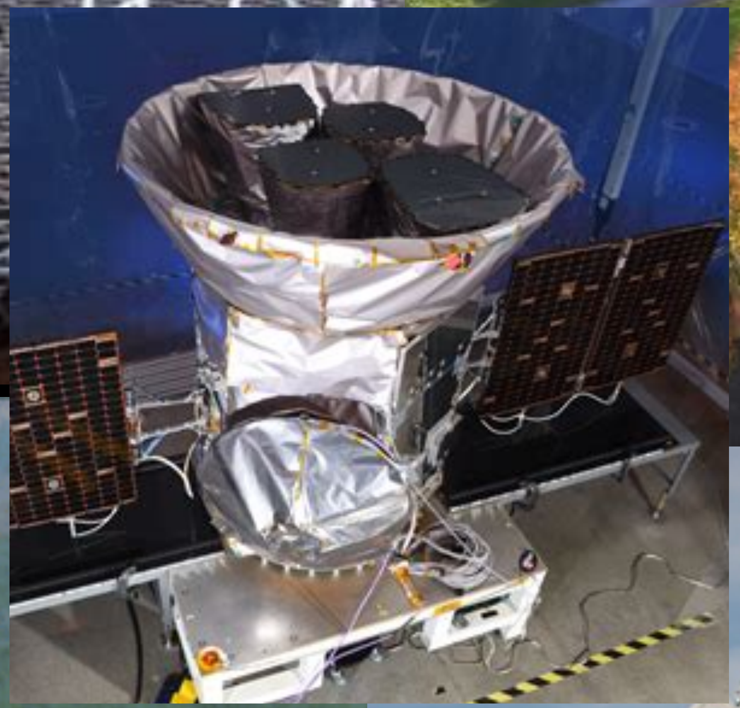
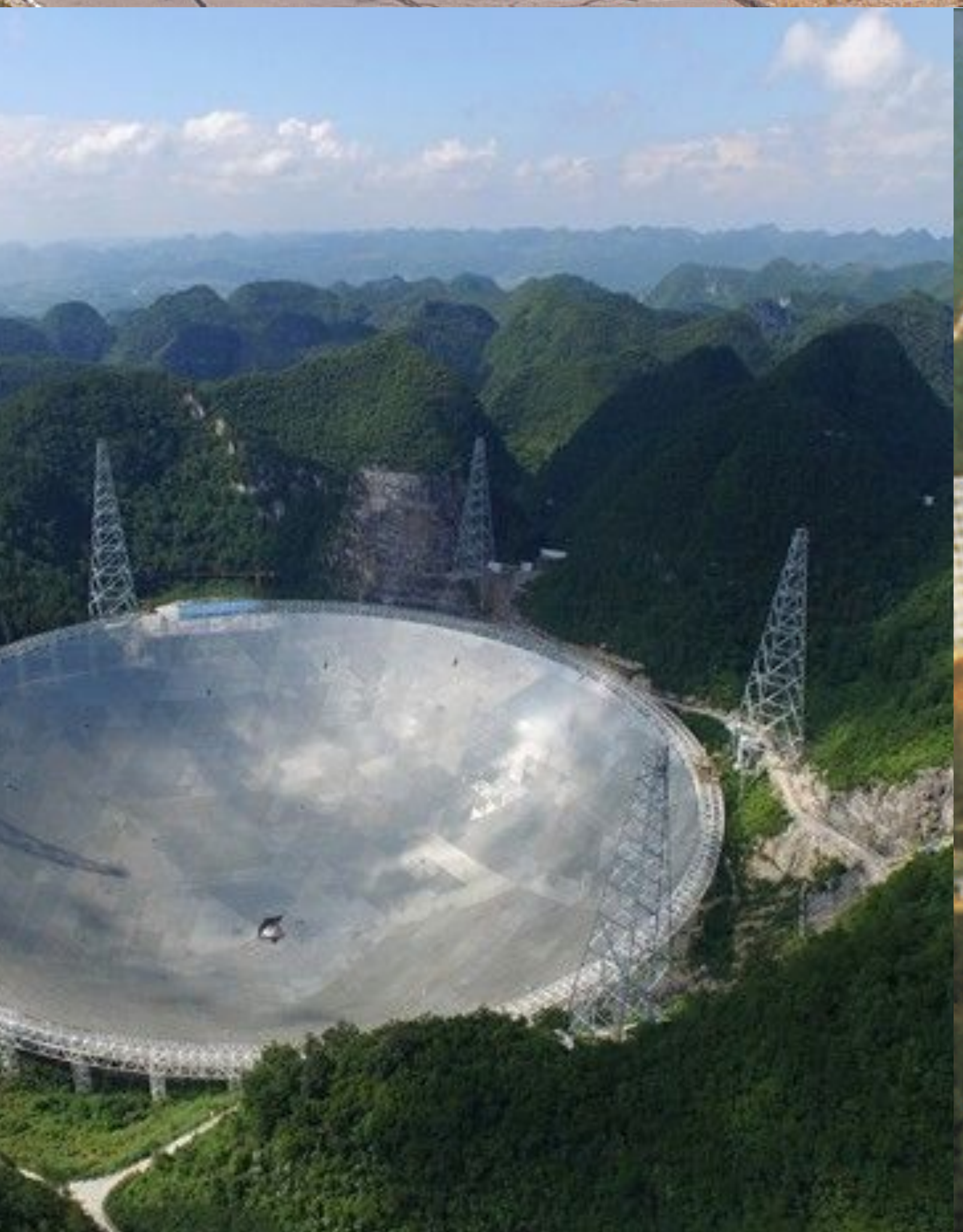
GREEN BANK USA

Automated planet finder, Lick observatory


MEERKAT TELESCOPE, SOUTH AFRICA
Image: SARAO

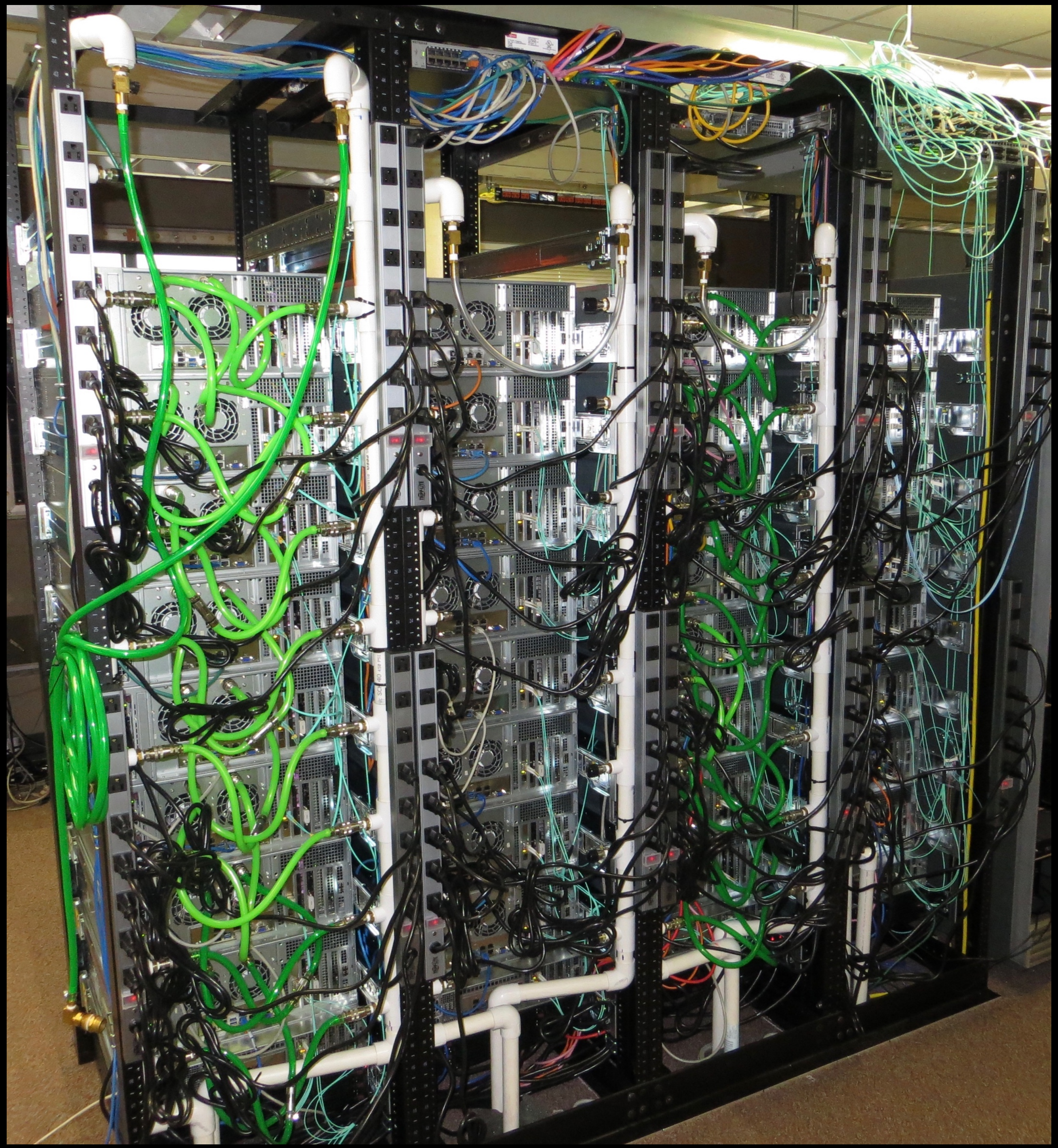
Partner Facilities

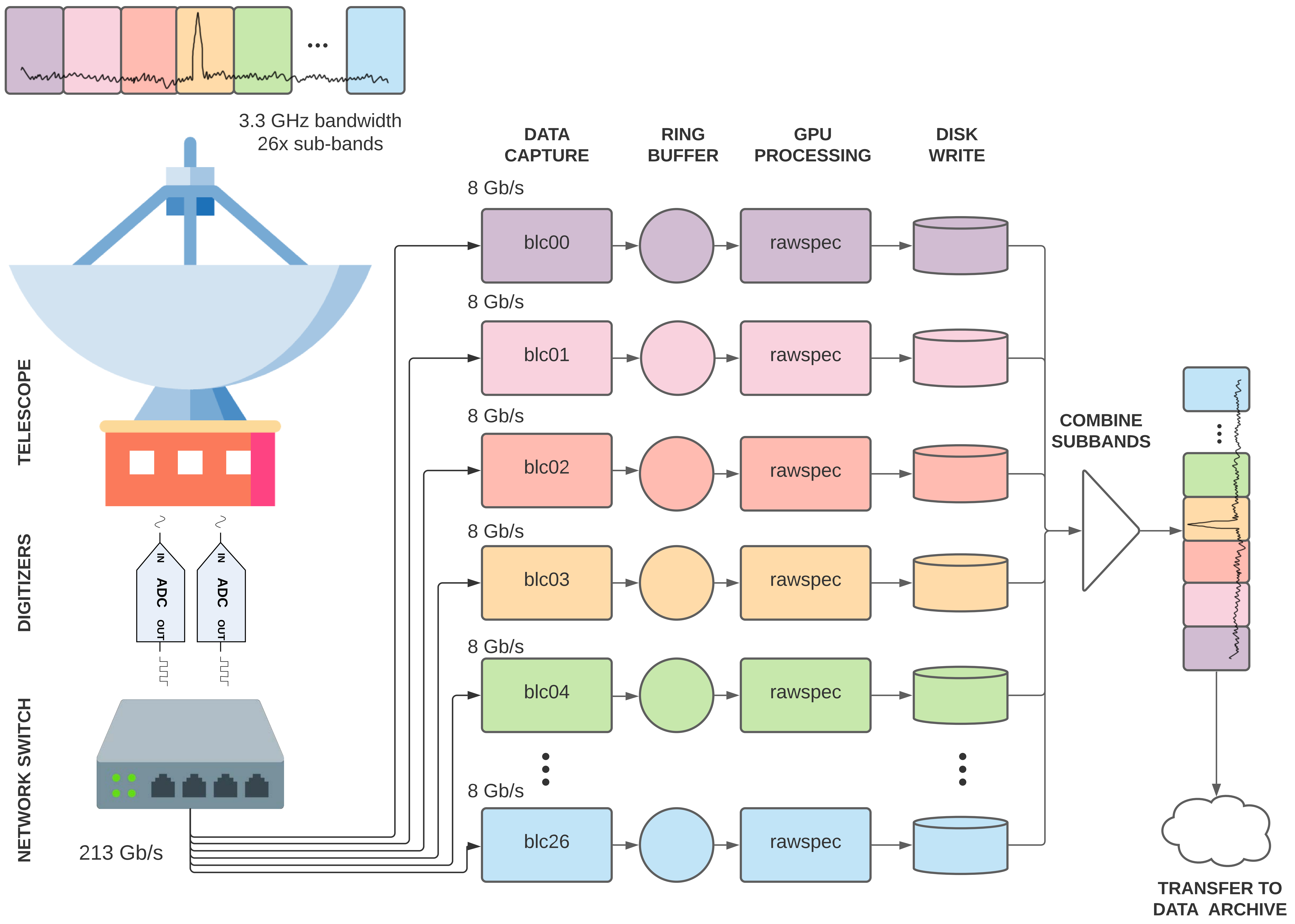
Step 1: Observing & Recording Data

Parkes Australia

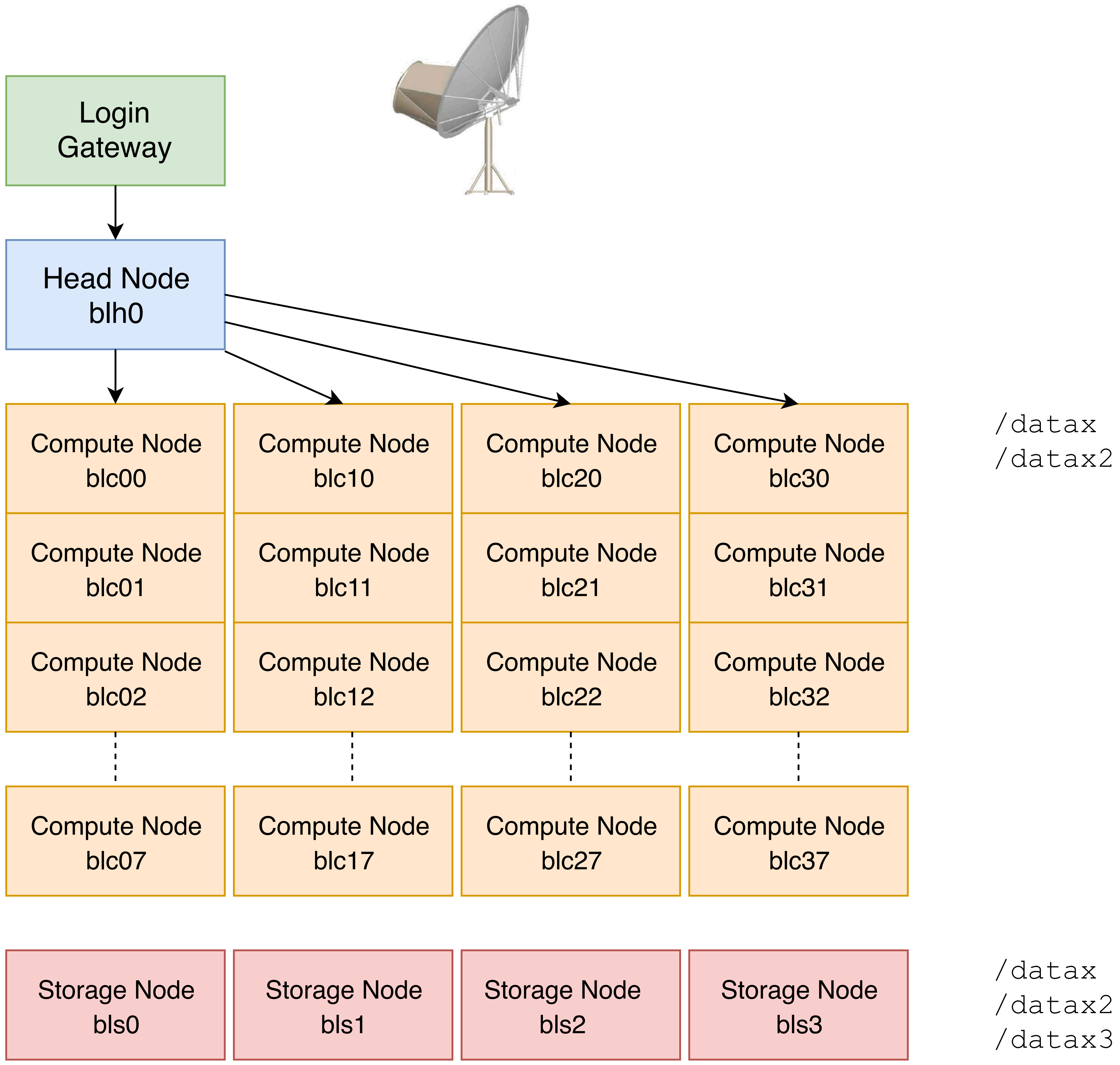


Green Bank USA

Parkes Data Flow Diagram



BL Data Recorders



- Currently (Jun 21) 13.6 PB of data stored on disk.
- 9 PB of storage at Green Bank
  – 65x compute nodes (w. GPU)
  – 9x storage nodes (x36 3.5″ disks)
- 3.5 PB of storage at Parkes
  – 27x compute nodes
  – 6x storage nodes
- 5 PB of data currently hosted at Berkeley, available at seti.berkeley.edu/opendata

Data Challenges

- High-speed UDP data capture meant we did not use distributed filesystem.
- Started using JBOD (just a bunch of disks), now moving data into a newly-commissioned gluster cluster for archiving.
- The cloud has remained too expensive (about 4x self hosting), but this is improving.
- Started using 4 TB drives (2015), 16 TB drives are now reasonably priced.

Step 2: Storing & Analyzing Data

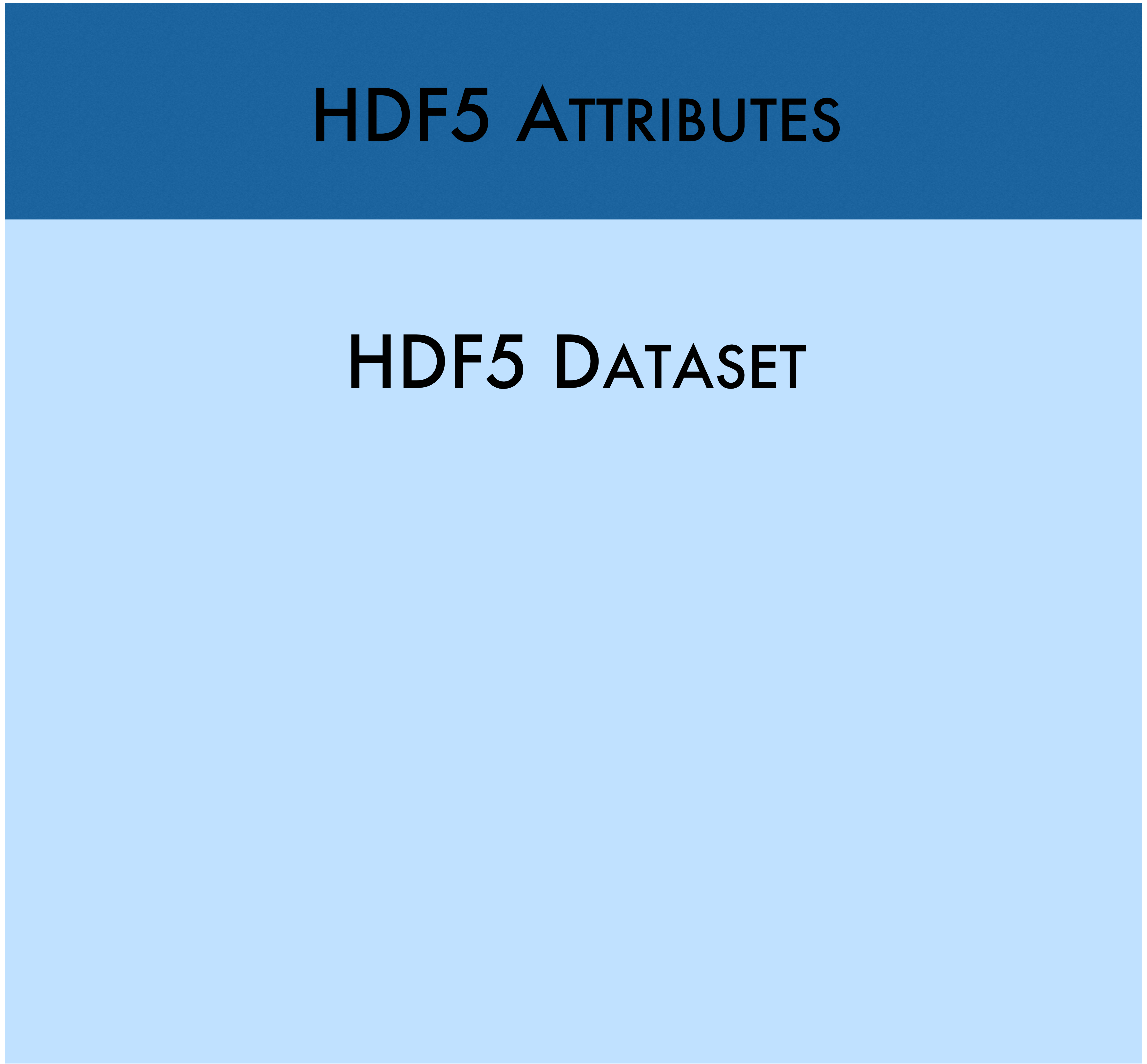
SIGPROC FILTERBANK FORMAT

HEADER
DATA BLOB

- A very simple format with a header followed by a data payload.

```
--- File Info ---
    telescope_id :                                4
           nbits :                               32
            fch1 :                           1361.5
          tstart :                    58643.2839468
       data_type :                                1
          nchans :                            45056
           ibeam :                               12
           tsamp :                   0.898779428571
     rawdatafile : guppi_58643_24533_053837_G238.12-3.44_0001.0002.fil
            foff :                   -0.00341796875
         src_raj :                     7:23:29.904
         src_dej :                     -24:18:46.8
          nbeams :                               13
        az_start :                              0.0
     source_name :                     G238.12-3.44
        za_start :                              0.0
      machine_id :                               20
            nifs :                                4

Num ints in file :                              334
      File shape :                  (334, 4, 45056)
--- Selection Info ---
Data selection shape :                  (334, 4, 45056)
Minimum freq (MHz) :                         1207.5
Maximum freq (MHz) :                         1361.5
```

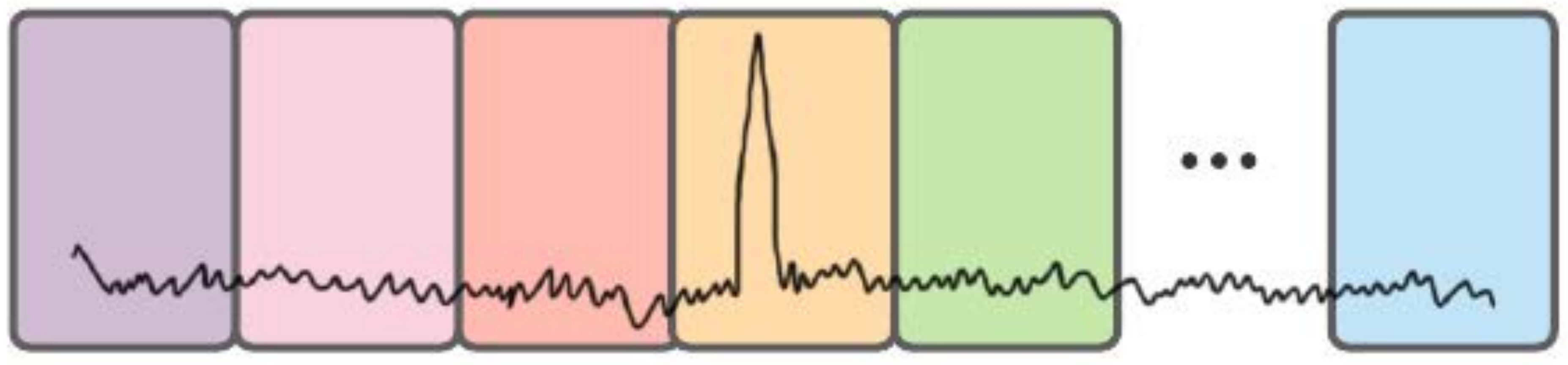
HDF5 Filterbank Format



- Filterbank header converted into set of HDF5 attributes.
- Data stored in a HDF5 dataset.
- Applying bitshuffle compression (designed for radio data).
- We use a python package called `blimpy` to interact with sigproc + HDF5 data, uses `h5py` and `hdf5plugin`

```
pip install blimpy
```

Virtual Dataset Opportunities

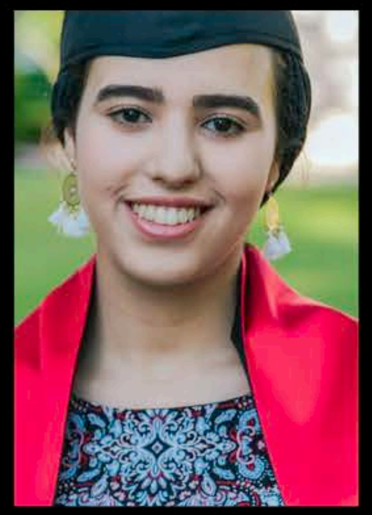


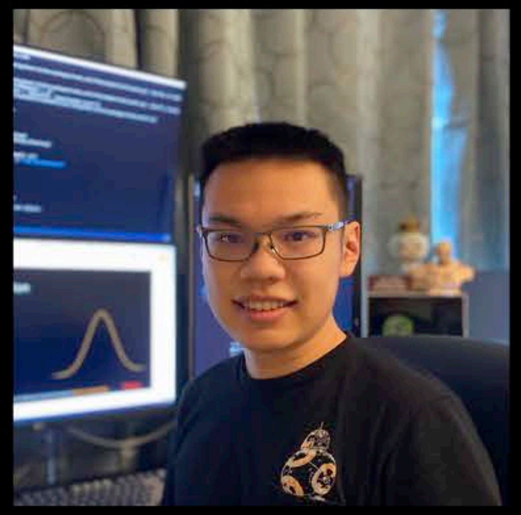
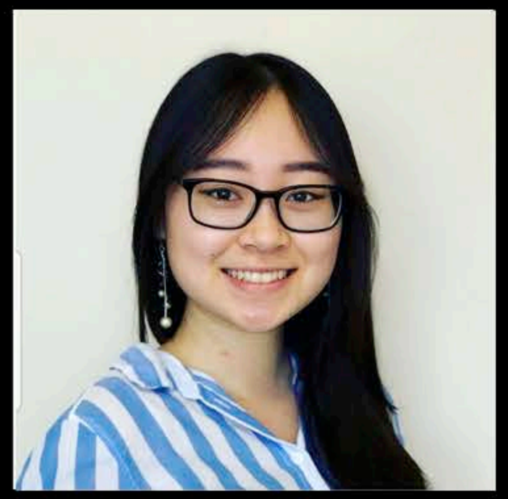
- Currently most observations are spread across multiple files (one file per sub-band).
- Can use HDF5 Virtual Datasets to combine sub-bands without moving data.

```
import h5py

nodes          = ['blc0%i'%i for i in range(8)]
n_nodes        = len(nodes)
n_timestep     = 92
n_chan_per_sub = 256
filename       = 'guppi_58948_45245_6051771717_J1019-5749_S_0001.0001.h5x'
vsources = []

layout = h5py.VirtualLayout(shape=(n_timestep, n_chan_per_sub * n_nodes), dtype='<f4')
for ii, node in enumerate(nodes):
    vsource = h5py.VirtualSource(f'collate/{node}/{filename}', 'bp_xx', shape=(92, 256))
    layout[:, ii*n_chan_per_sub:(ii+1)*n_chan_per_sub] = vsource

with h5py.File("VDS_TEST.h5", 'w', libver='latest') as f:
    f.create_virtual_dataset('data', layout, fillvalue=0)
```

BL@Scale

- Our 2020 REU student team created a cloud-based SETI pipeline: BL@Scale.
- Being used to prototype new machine learning algorithms and test cloud workflows.
- We would like to bring these techniques to our on-premises data archive + HPC cluster.




Fatima Zaidouni | Peter Ma | Yuhong Chen | Shirley Wang | Rachel Zhong

Progress and Future Steps

- We have recently rolled out gluster and are expanding our gluster storage capacity.
- We are using Jupyterhub to serve notebooks on co-located GPU servers.
  We would like to write a SSH spawner to serve notebooks across servers at different sites.
- Can we run HSDS on gluster, and modify our tools to use `h5pyd`?
- Also considering a SLURM + Singularity processing approach.


THANK YOU
@berkeleyseti
BERKELEY SETI
RESEARCH CENTER
ICRAR
Image: P. Hart