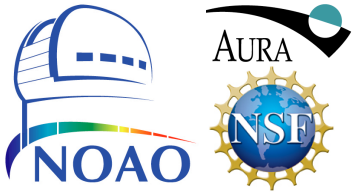
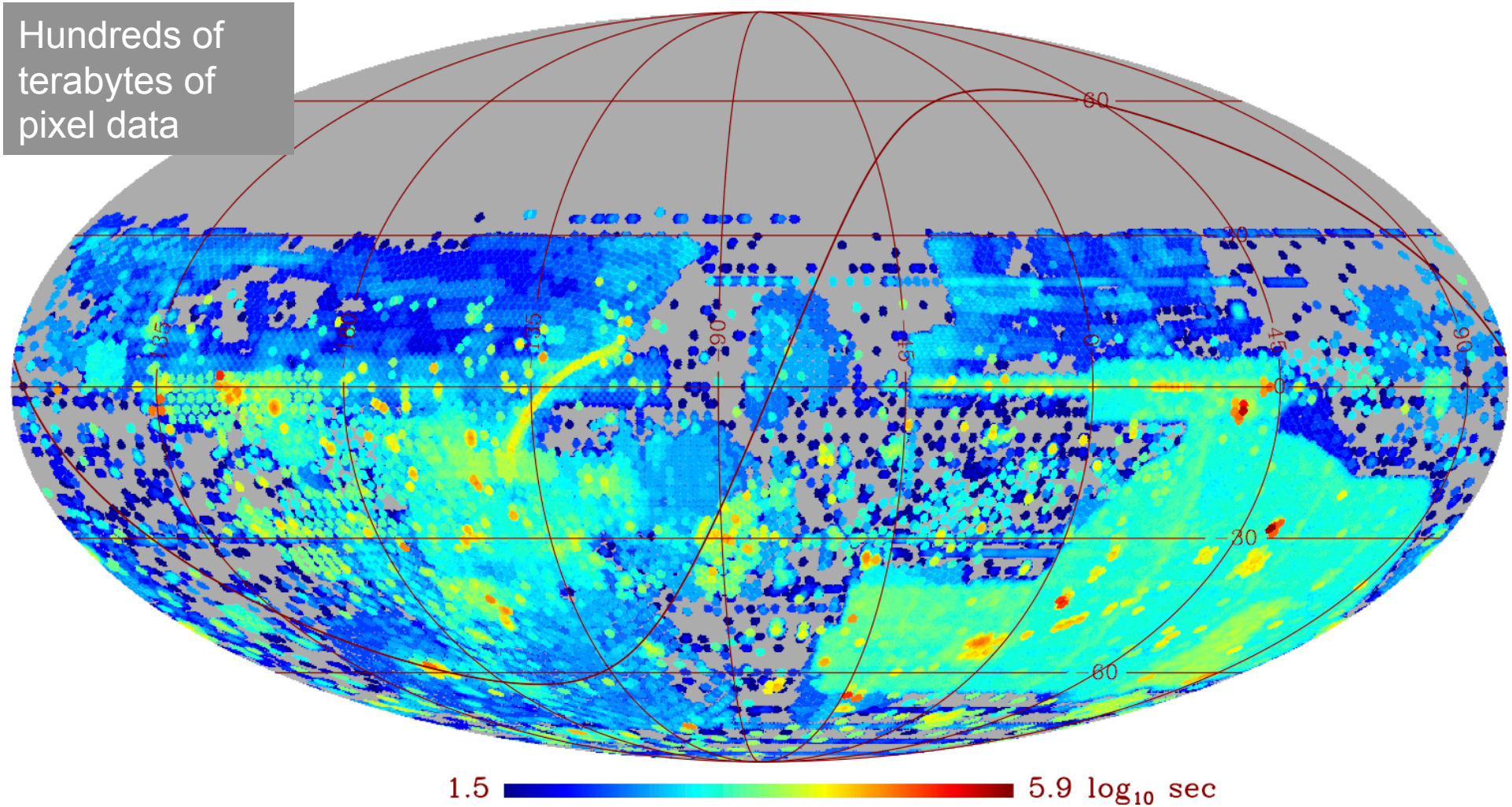


Hunting Dwarf Galaxies: A Preview of the NOAO Data Lab

Data Lab team: Knut Olsen, Mike Fitzpatrick, Matthew Graham, Ken Mighell, Betty Stobie, Pat Norris, and Steve Ridgway

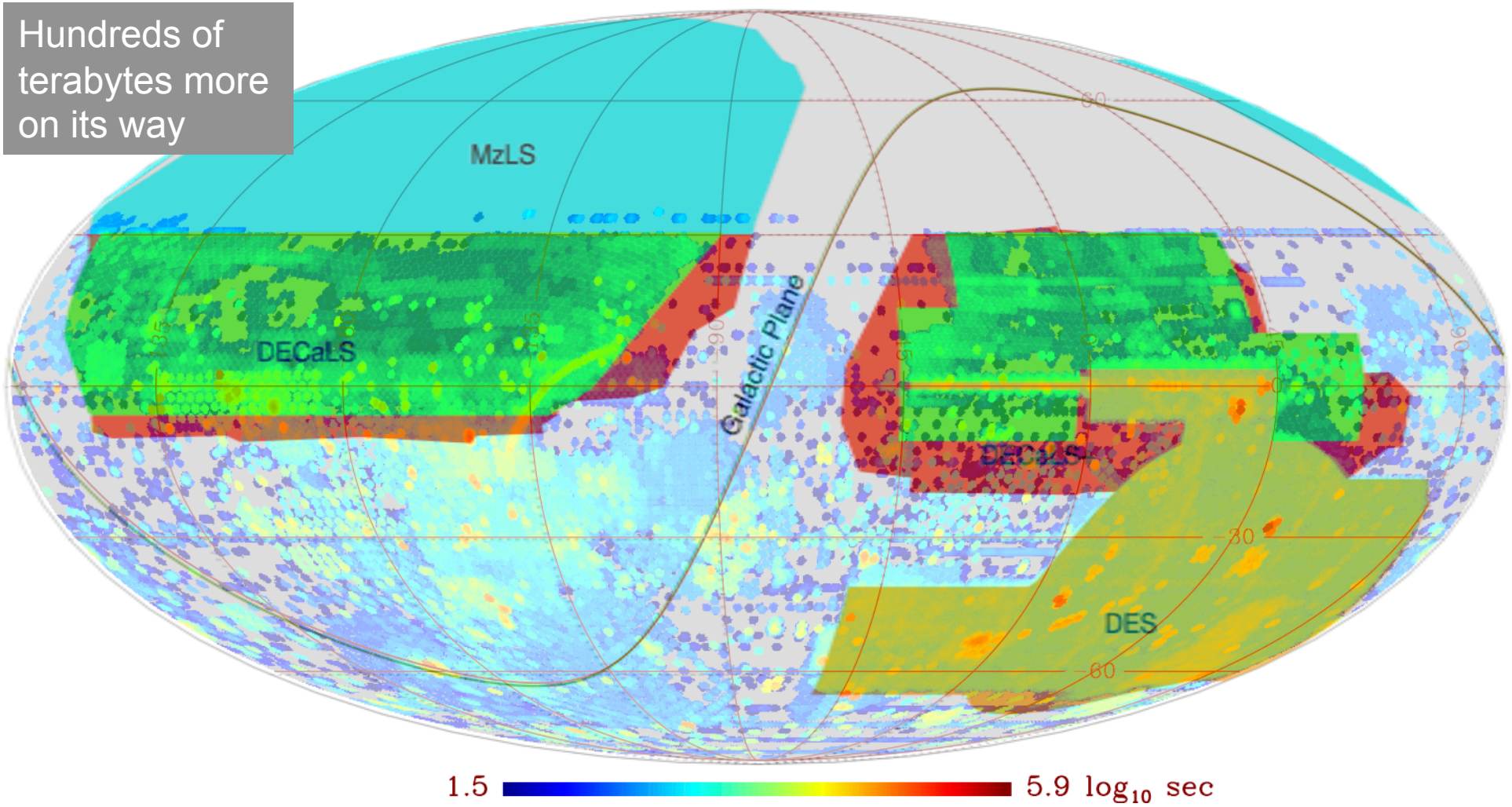
Big Data at NOAO

Hundreds of terabytes of pixel data



Big Data at NOAO

Hundreds of terabytes more on its way



Big Data at NOAO

350 TB (December 2015) of on-target imaging data (t_{exp}>30s) currently from:

- Dark Energy Survey
- DECaLS and DESI Targeting Survey
- Community DECam programs and surveys

Hundreds of TB more coming

Total holdings of several PB

Large catalogs coming:

- Dark Energy Survey – 45 TB
- DESI Targeting Survey – ~5 TB
- Community programs and surveys – up to several TB each

NOAO Data Lab

Goal:

Efficient exploration and analysis of the large datasets being generated by instruments on NOAO wide-field 4-m telescopes

A Science Case: Satellites of the Milky Way

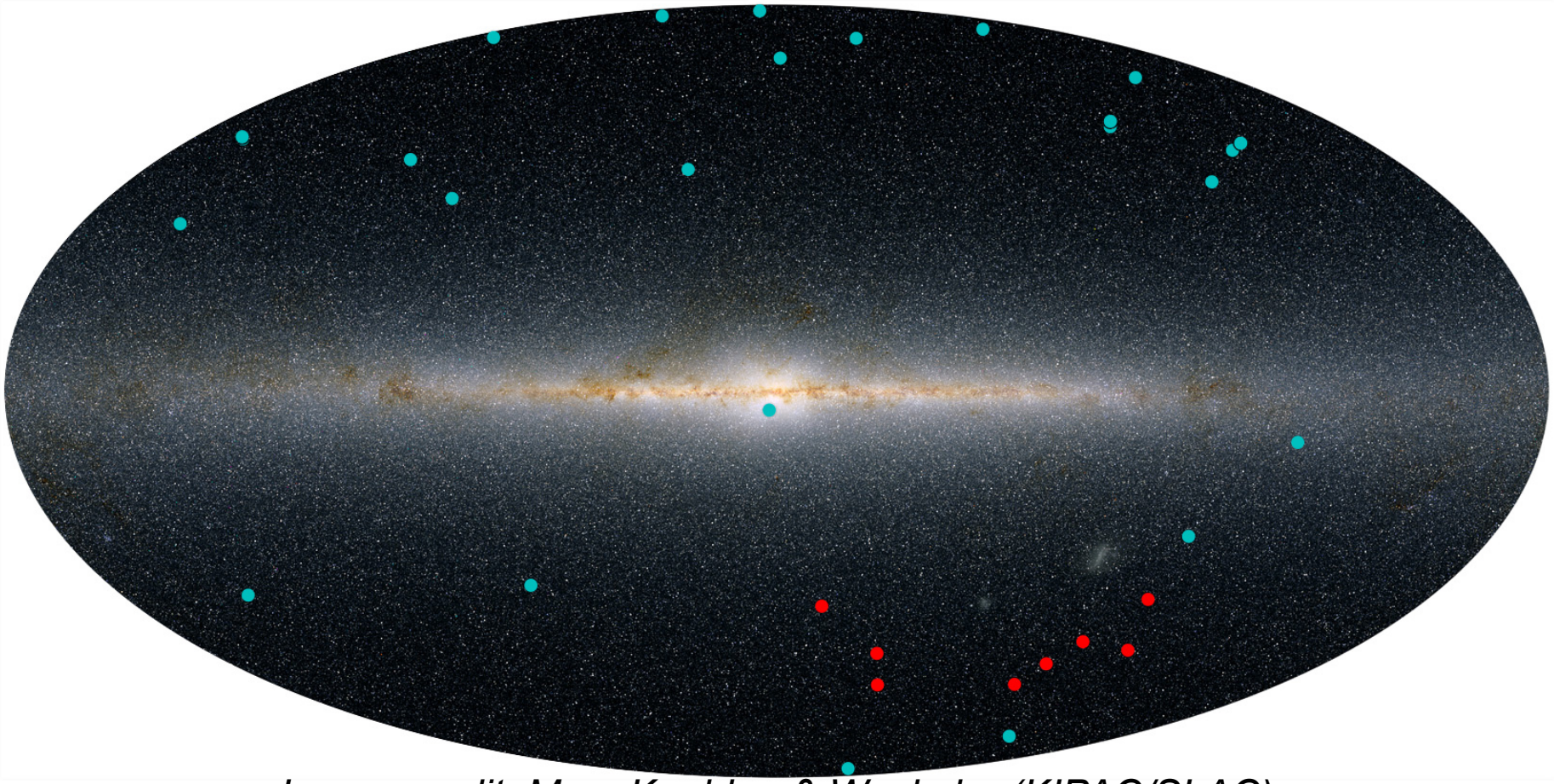
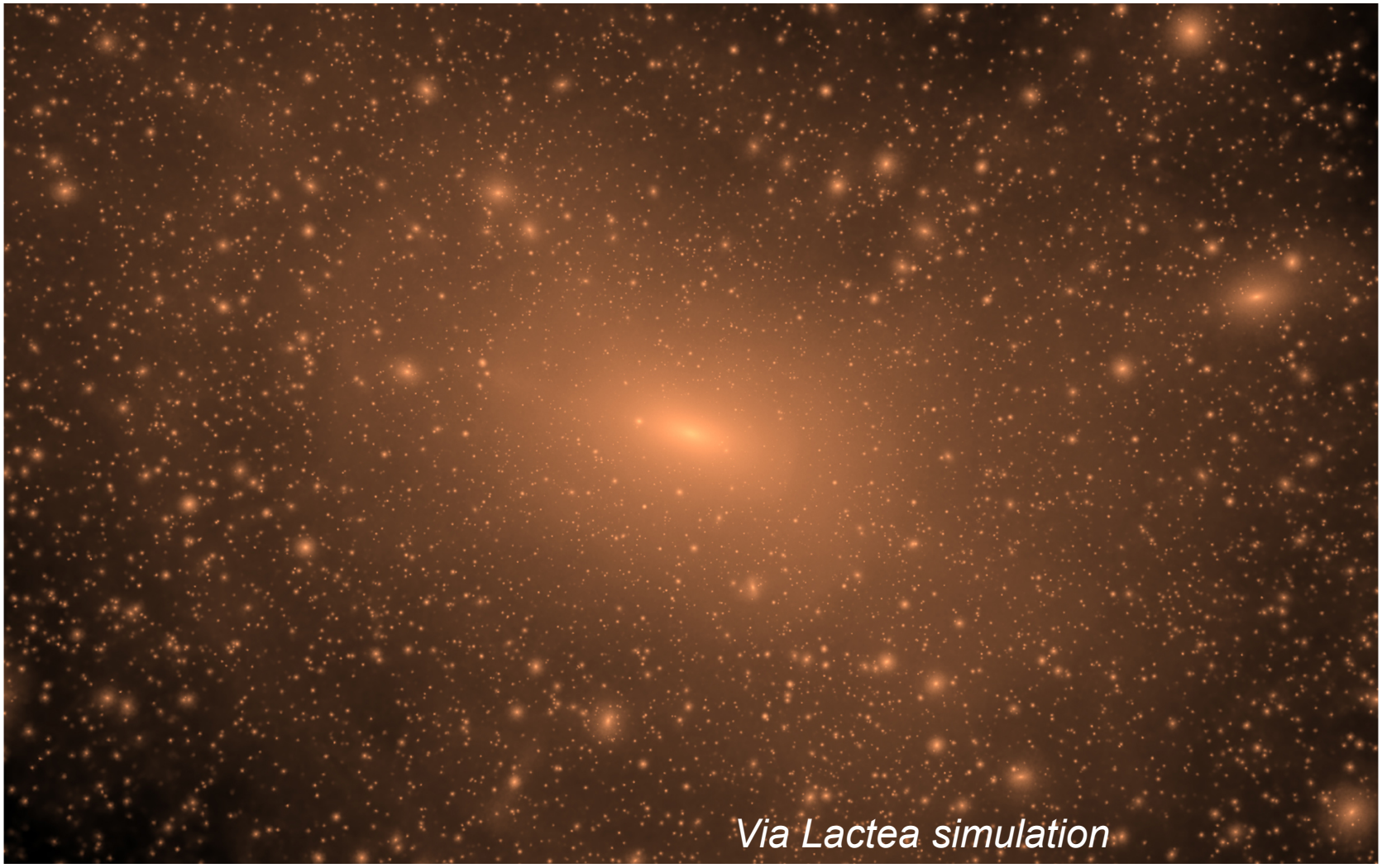
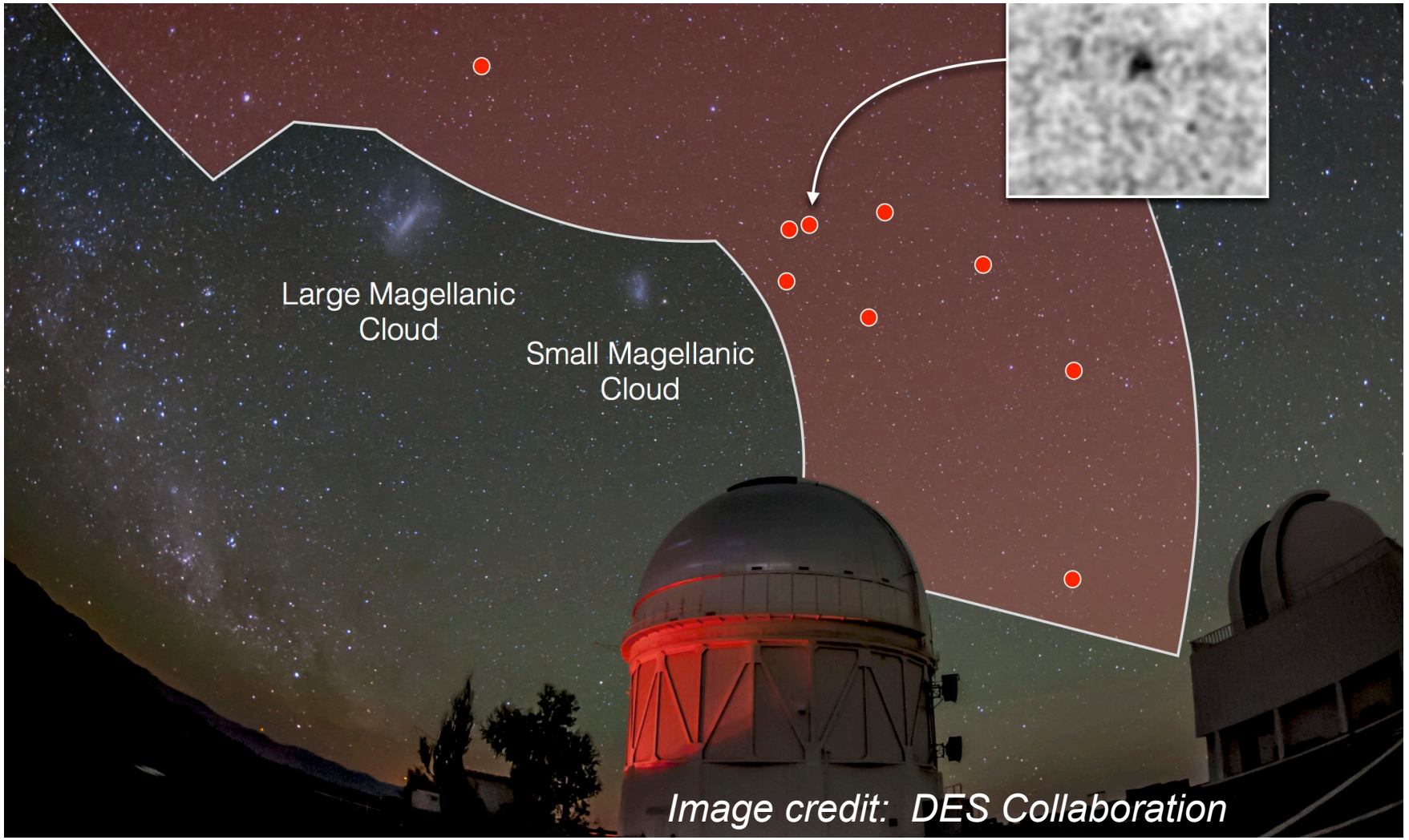


Image credit: Mao, Kaehler, & Wechsler (KIPAC/SLAC)

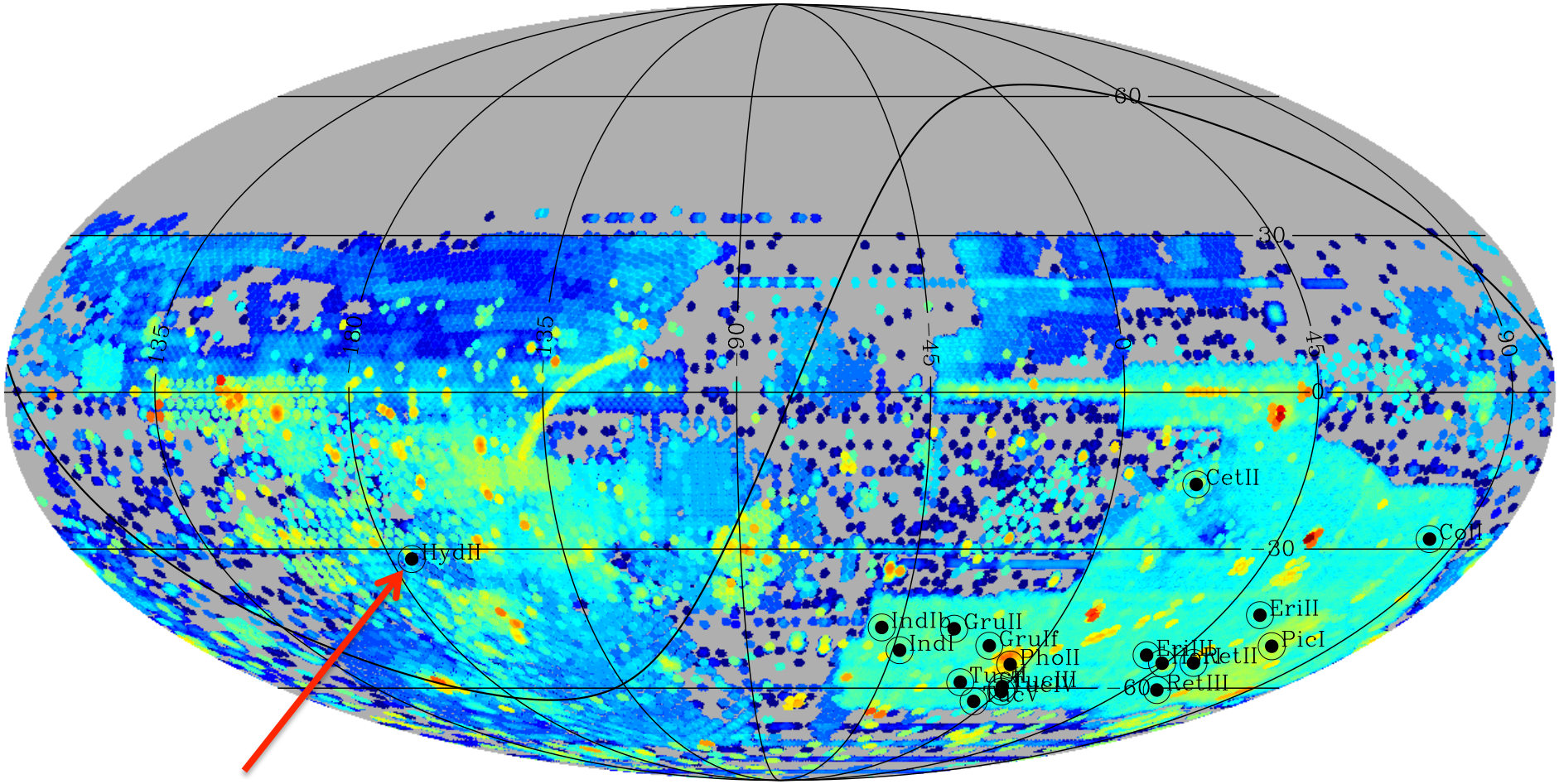
The “Missing Satellite” Problem



Satellites of satellites?



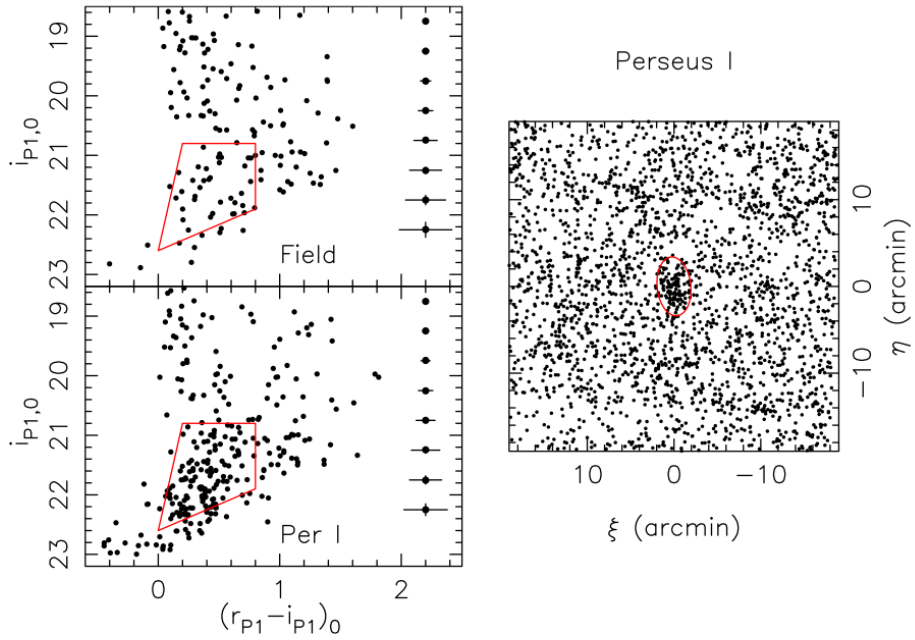
DECam Sky Coverage



Hydra II discovered by Martin et al. (2015) from SMASH survey (Nidever et al. 2015)

Scientific Approaches Using the Data Lab

- Catalog science
  - Example use case: search for Galactic substructure through photometric selection of candidate populations
  - Data Lab will provide access to large catalog databases, query interface, personal storage, and visualization capability



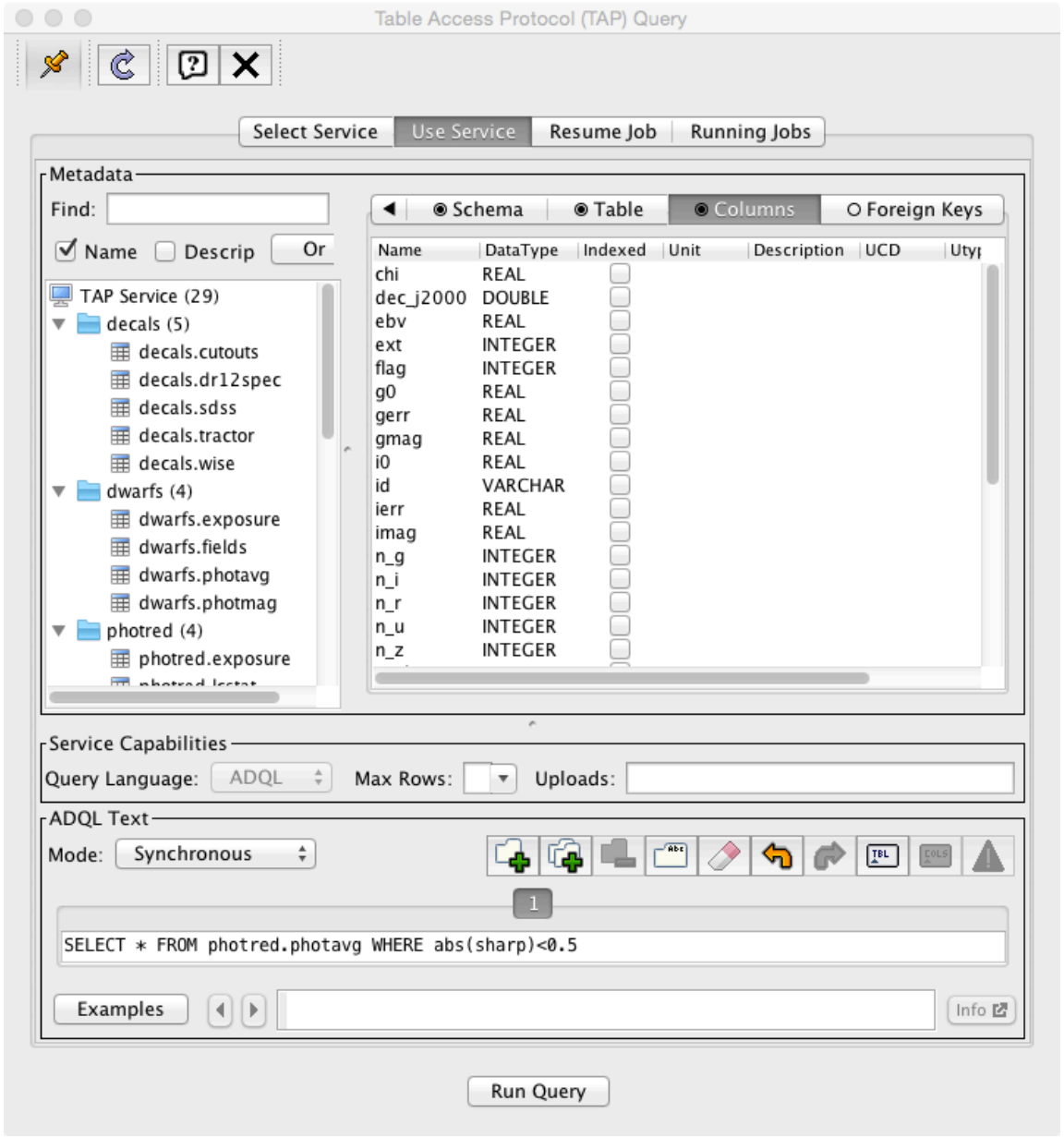
Martin et al. (2013)

Interactive approach

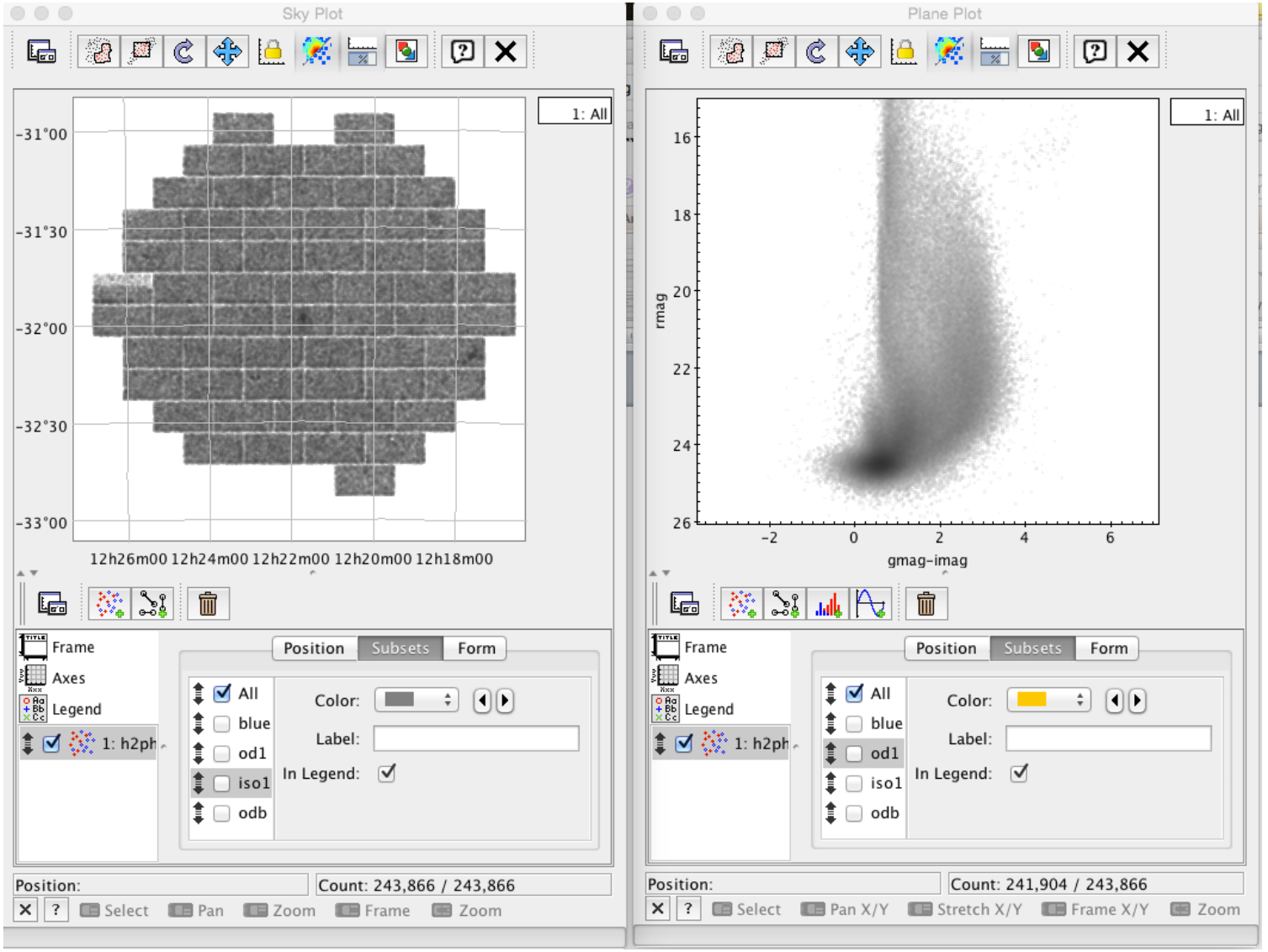
Data Lab TAP service provides access to database of catalog data

Hundreds of millions of rows

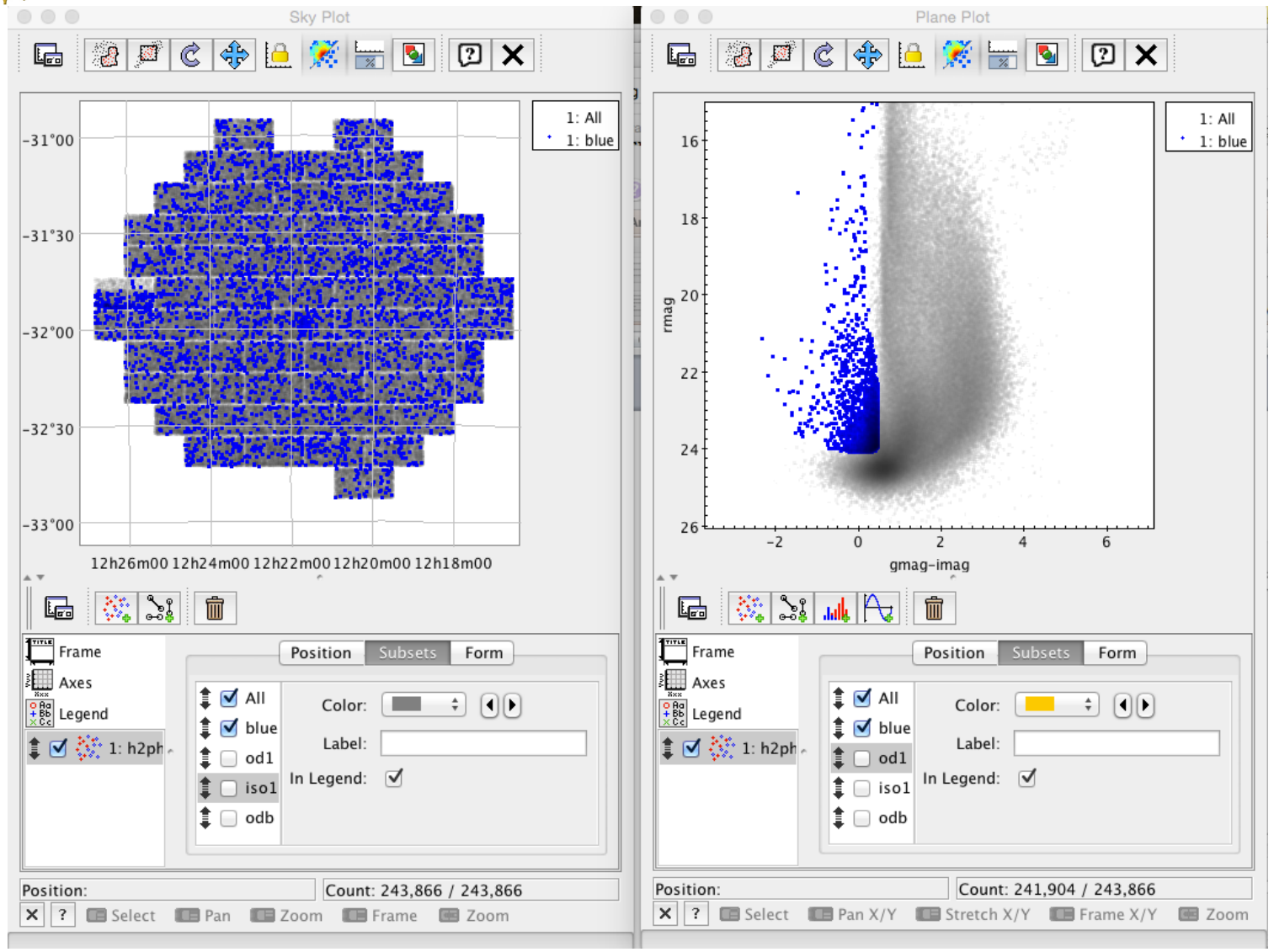
Accessible via TOPCAT



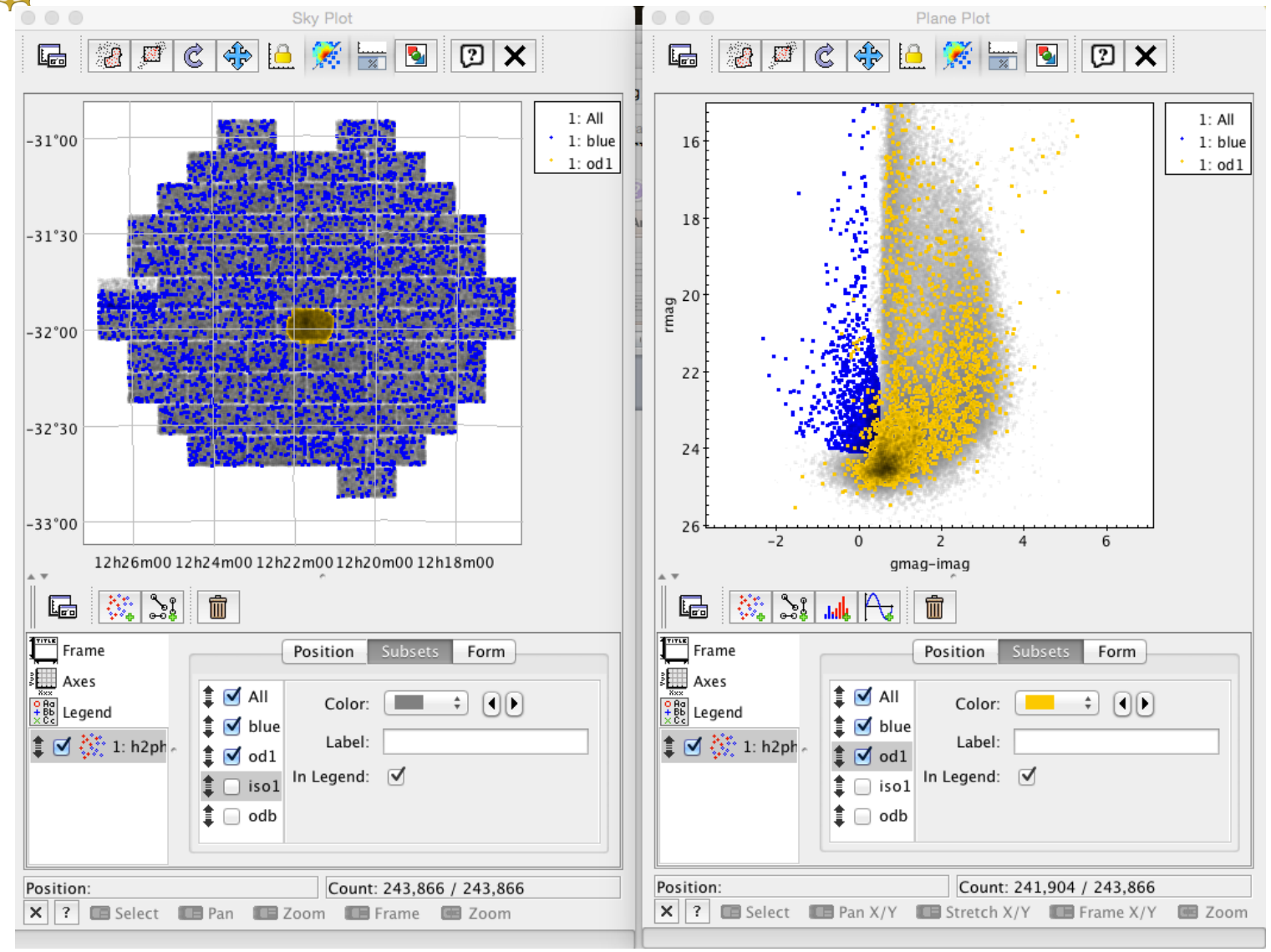
Interactive approach



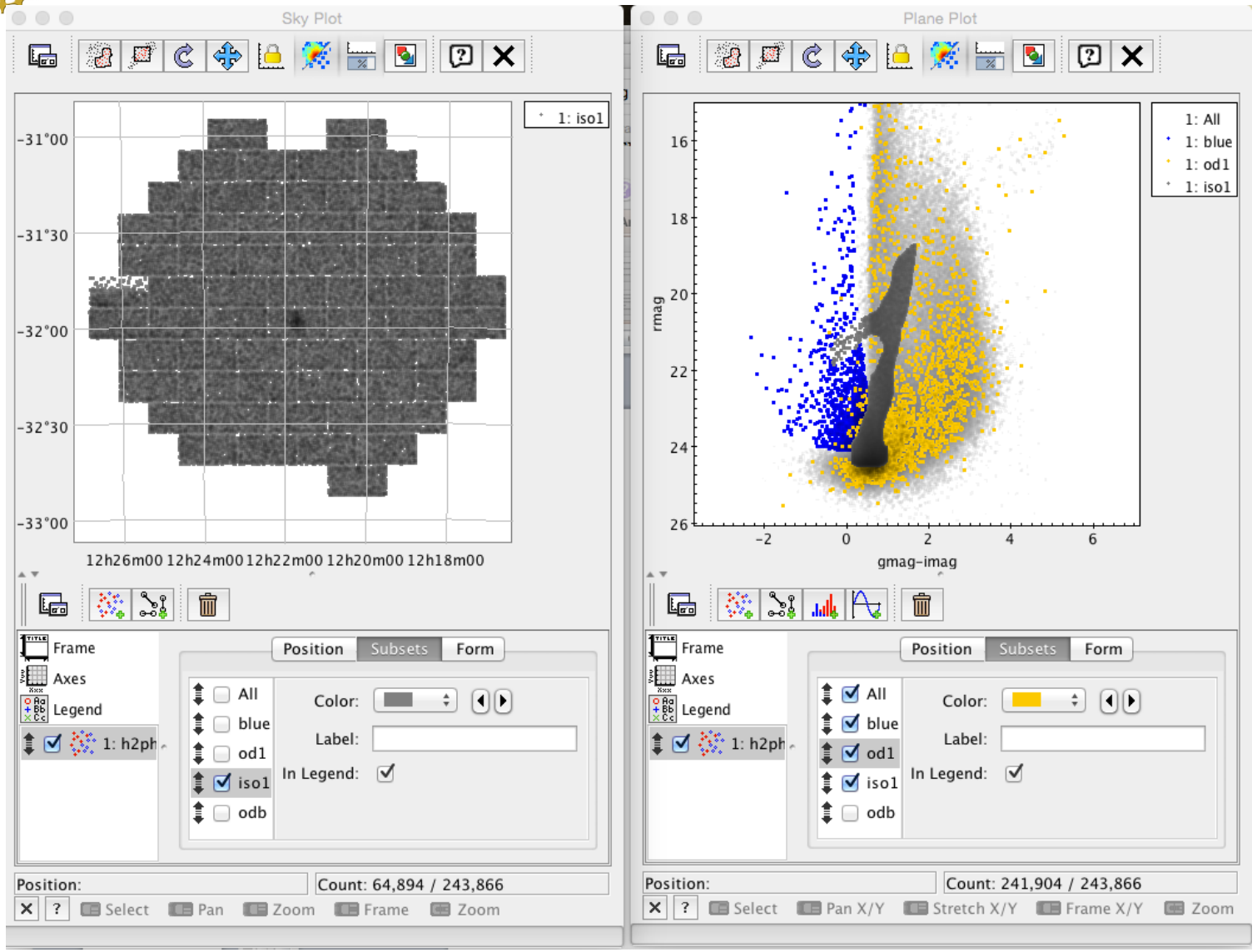
Interactive approach



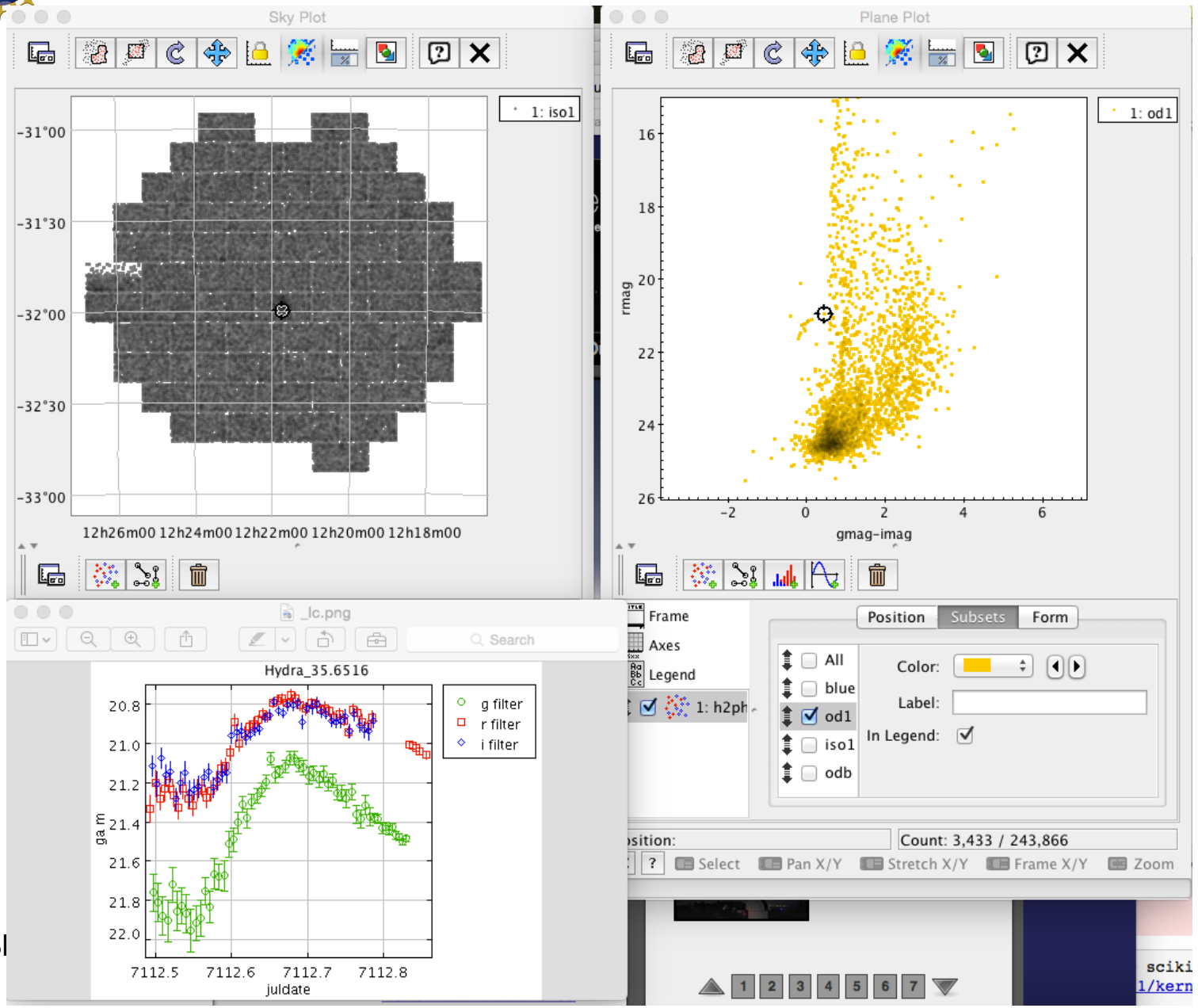
Interactive approach



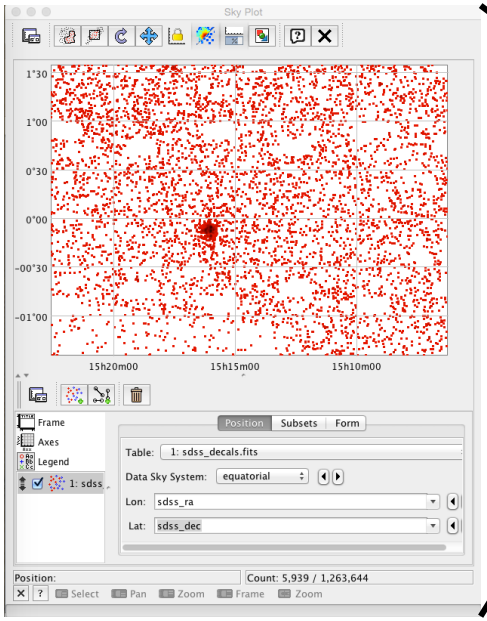
Interactive approach

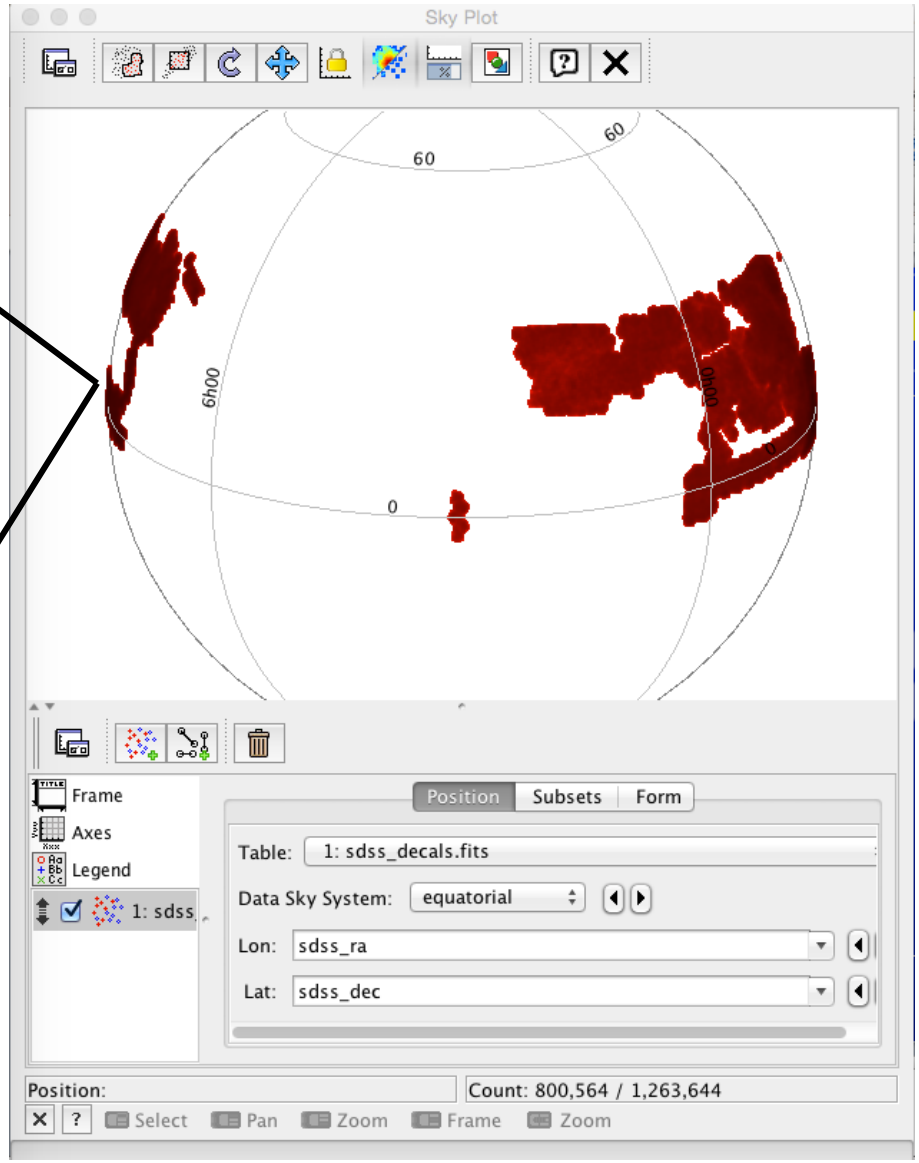


Interactive approach



Beyond interactive exploration





DECaLS DR1 sources
matched to SDSS
Objects with blue *g-i* colors

Beyond interactive exploration

- Key infrastructure:
  - Table Access Protocol interface to database
  - Virtual storage space
  - Multiple programmatic interfaces (datalab command, STILTS, Python APIs)
  - Image cutout service
  - Compute service

A prototype scripted approach

This notebook is about finding dwarf galaxies in SMASH data by identifying spatial and/or isochronal overdensities and then looking for associated RR Lyrae in multipass data.

This notebook requires the gavo, gatspy, scikit-learn and ipywidgets packages are installed. Original by Matthew Graham. Modified slightly by Knut Olsen.

First we retrieve the data for Hydra II from DataLab with a TAP query:

In [2]:
```
#%matplotlib inline
%matplotlib notebook
import numpy as np
import matplotlib.pyplot as plt
from gavo import votable
from lxml import etree
from cStringIO import StringIO
import mpld3

def get_data():
  '''Retrieve the data via TAP and put into a numpy array'''
  accessURL = "http://dldb1.sdm.noao.edu:8080/ivoa-dal/tap"
  query = "select ra_j2000, dec_j2000, gmag, gmag - rmag as gr, id from smash.photavg where gmag between 0 and 25 and r
  raw = votable.ADQLSyncJob(accessURL, query).run().openResult()
  # We now need to parse the VOTable XML
  xml = etree.parse(raw)
  data = []
  for row in xml.findall('//TR'):
    data.append([td.text for td in row.findall('TD')])
  data = np.array(data).T
  return data

def get_lightcurve(id):
  '''Retrieve the SMASH light curve for the specified id and filter via TAP'''
  accessURL = "http://dldb1.sdm.noao.edu:8080/ivoa-dal/tap"
  query = "select mjd, mag, magerr, filter from smash.photmag where id = '%s'" % id #, filter)
  vot = votable.ADQLSyncJob(accessURL, query, userParams = {"FORMAT": "csv"}).run().openResult().read()
  data = np.genfromtxt(StringIO(vot), delimiter = ",", skip_header = 1, dtype = None)
  return data
```

iPython notebook by Matthew Graham

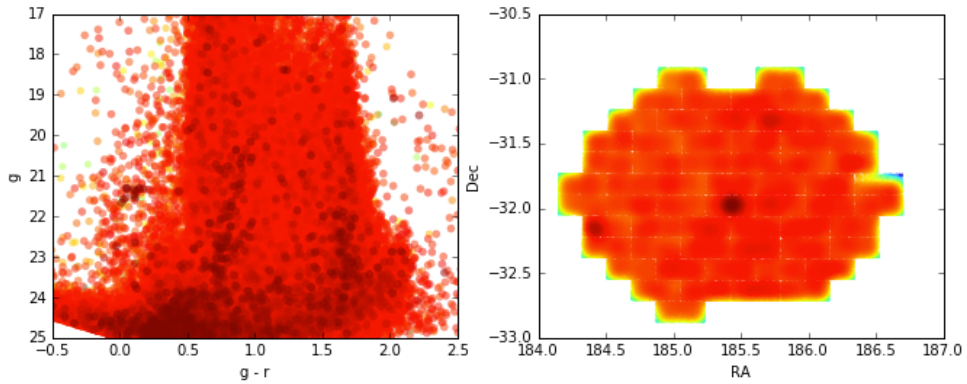
A prototype scripted approach

```
In [2]: from sklearn.neighbors.kde import KernelDensity
        data = get_data()
```

```
In [19]: # Calculate the KDE for RA, Dec
         kde = KernelDensity(algorithm = 'ball_tree', kernel = 'gaussian', bandwidth = 0.05, rtol = 1.e-8).fit(data[0:2].T)
         #colors = np.exp(kde.score_samples(data[0:2].T))
         colors = kde.score_samples(data[0:2].T) # This gives log results
         idx = colors.argsort()
```

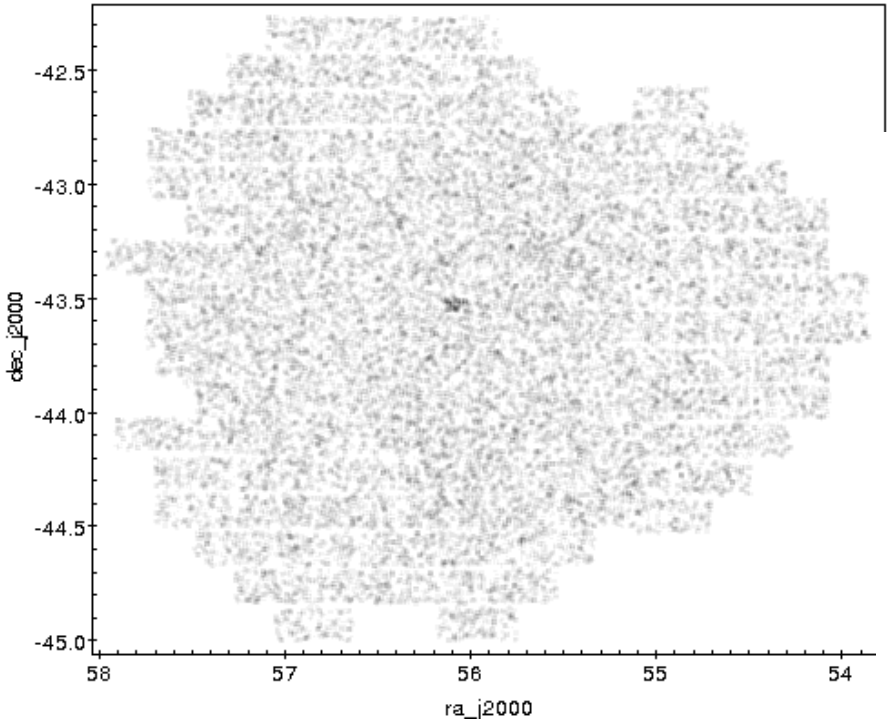
```
In [52]: def plot(colors, idx):
           fig, (ax1, ax2) = plt.subplots(1, 2, figsize = (11.0, 4.0))
           ax1.scatter(data[3][idx], data[2][idx], marker = '.', c = colors[idx], s = (colors[idx]-min(colors[idx])+1)*40, alpha
           ax1.set_xlim(-0.5, 2.5)
           ax1.set_ylim(17., 25.,)
           ax1.set_ylim(ax1.get_ylim()[::-1])
           ax1.set_xlabel("g - r")
           ax1.set_ylabel("g")
           ax2.scatter(data[0][idx], data[1][idx], marker = '.', c = colors[idx], s = 10, alpha = 0.5, edgecolor = '')
           ax2.set_xlabel('RA')
           ax2.set_ylabel('Dec')
           plt.show()
           return fig, ax1, ax2

         plot(colors, idx)
```

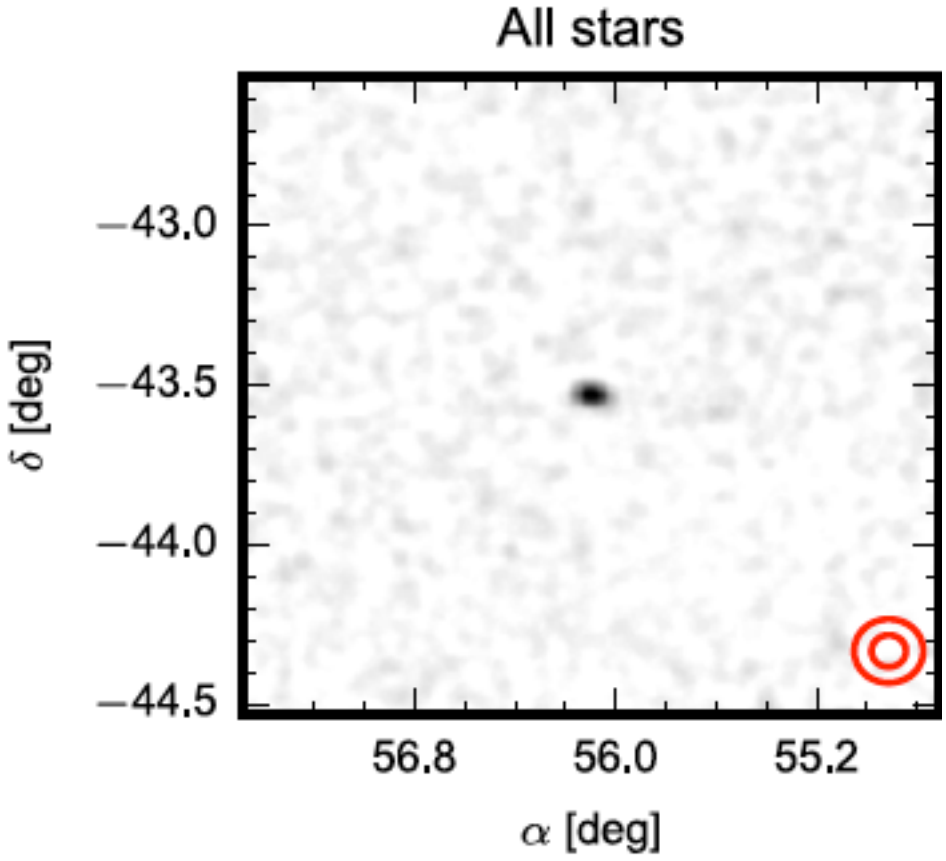


iPython notebook by Matthew Graham

Proceeding from image data

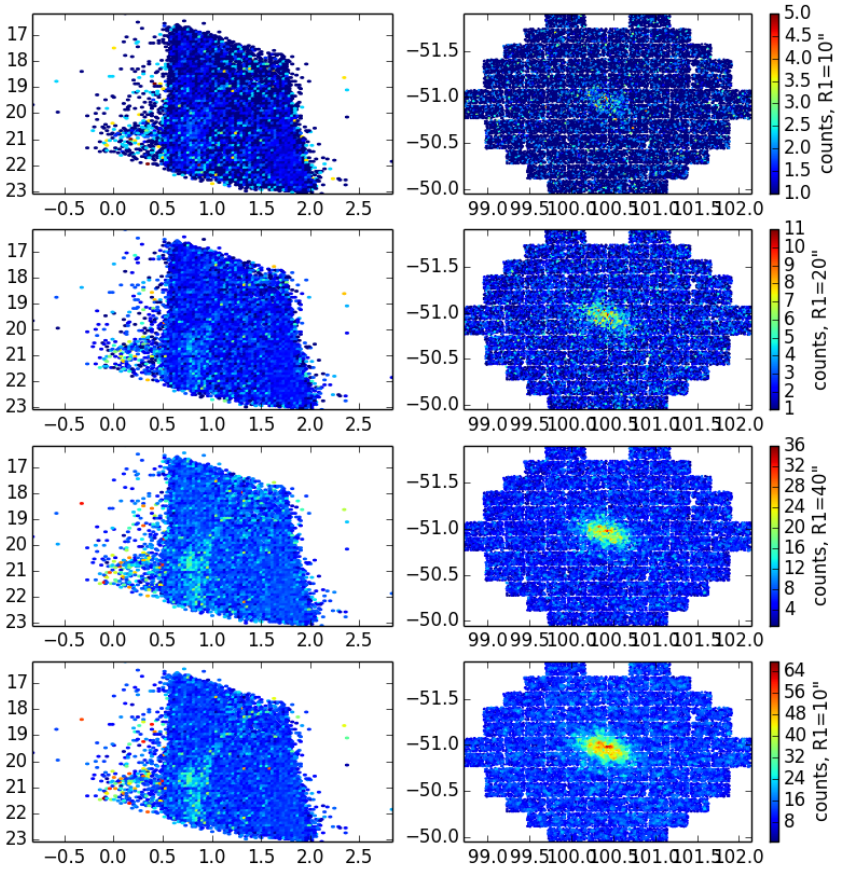


Eridanus II dwarf from NIKE quick-look pipeline (Ken Mighell)



Beyond interactive exploration

- Use *datalab* command to execute query for blue sources and store results in a virtual space (VOSpace)
  - Alternatives: scriptable STILTS tapquery or GAVO TAP python interface, followed by local analysis
- Use *datalab* command to execute density estimator code
  - Alternative: use VOSpace Capability to run code and generate graphics
- Use *datalab* command to retrieve image cutouts and store in VOSpace
  - Alternative: Python API
- Use STILTS or *datalab* command to retrieve full photometric data in region around promising overdensities, store in VOSpace or transfer to local machine
- For subset of sources, use TAP upload tool to make queries for time series data, store output in VOSpace


counts, R1=10"
counts, R1=20"
counts, R1=40"
counts, R1=10"

The Data Lab in a Nutshell

Large Catalogs – Data Lab will serve TB-scale databases

Pixel Data – Data Lab will connect users to images and spectra in NOAO Science Archive

Virtual Storage – Minimizes data transfer

Visualization – Data Lab will enable data exploration

Compute Processing – Data Lab will allow workflows to run close to the data using significant compute resources

Additional features – Access to published datasets and external data services, data publication, exportable workflows, distributable software

The Data Lab is coming!

- Demo of range of capabilities at June AAS meeting
- Public release in mid-2017
- Open positions available! Data Scientists and Developers
- Learn more at http://datalab.noao.edu