The NOAO Data Lab Project Introduction

Knut Olsen

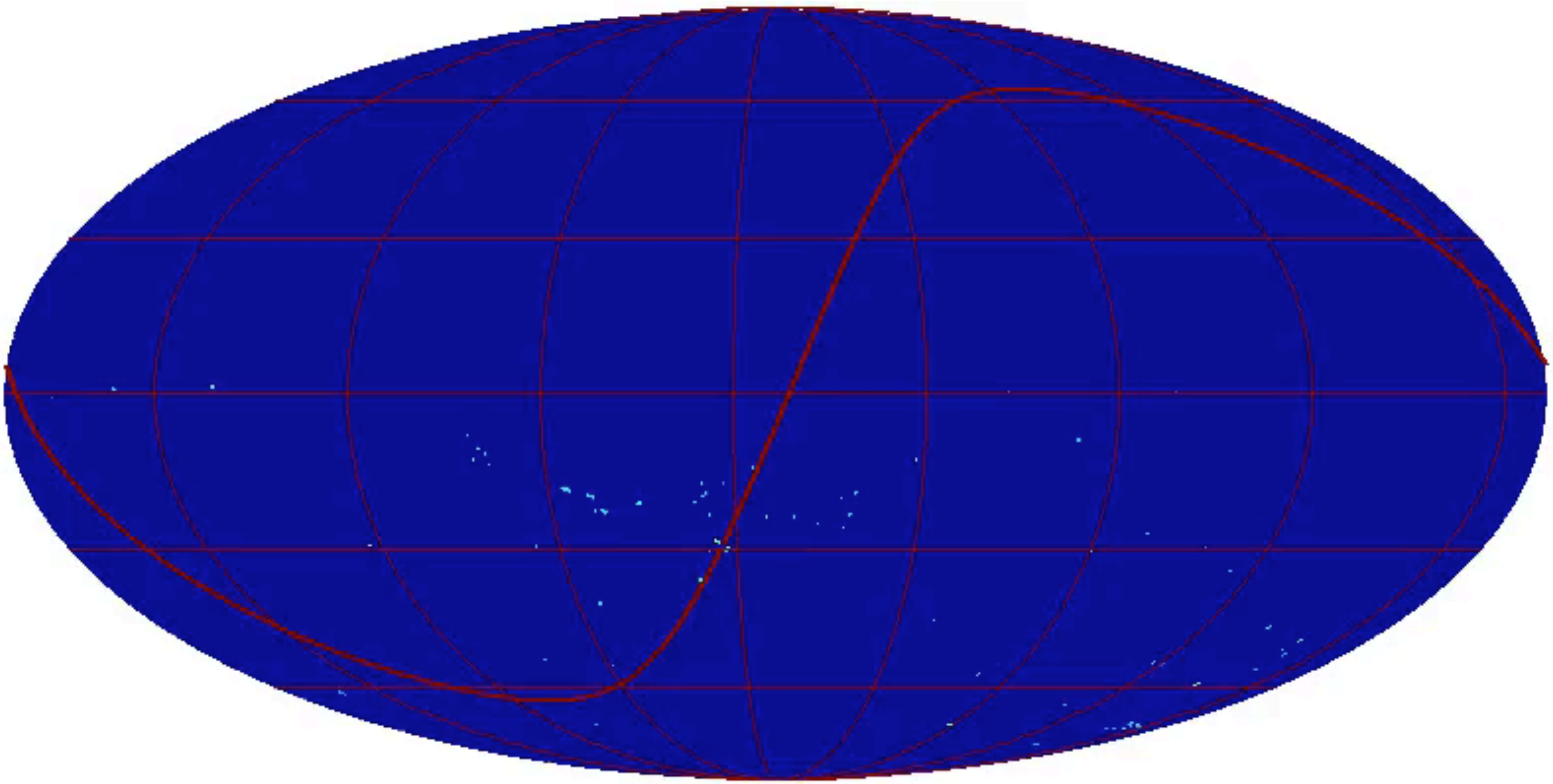
for the Data Lab team

Data Lab Team

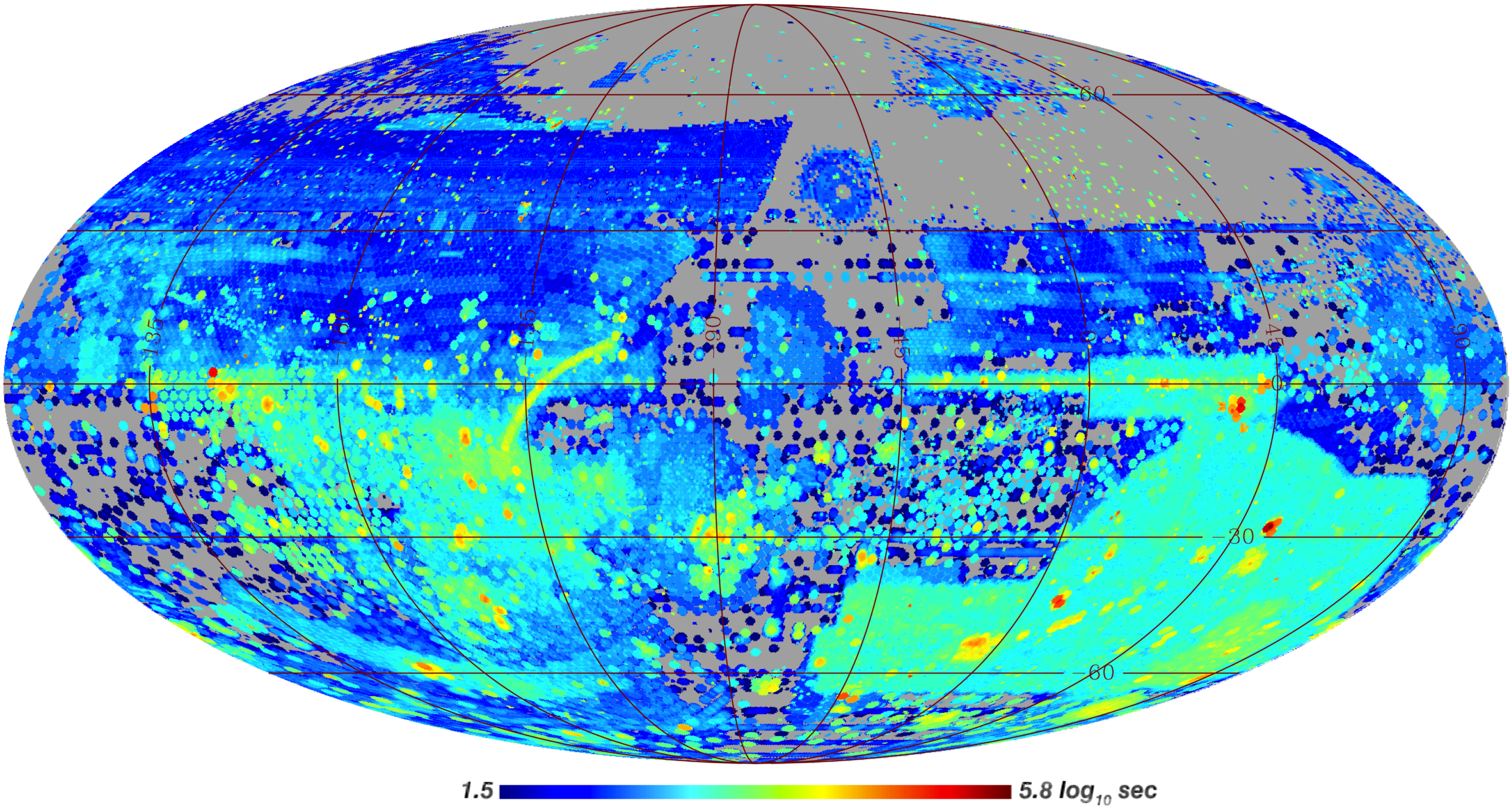
Current team:

- Mike Fitzpatrick, Lead Developer
- Matthew Graham, Scientist/Developer
- Wendy Huang, Software Engineer
- Stephanie Juneau, Data Scientist
- David Nidever, Data Scientist
- Robert Nikutta, Data Scientist
- Pat Norris, Test Engineer
- Knut Olsen, Project Scientist
- Steve Ridgway, Scientist
- Pete Wargo, System Administrator

NOAO wide field imaging data over time



DECam and Mosaic data in May 2016



Data Volume and Complexity

500 TB (January 2017) of on-target imaging data (t_{exp}>30s) currently from:

- Dark Energy Survey
- Legacy Surveys for DESI Targeting
- Community DECam and Mosaic 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
- Approach:
  - Catalogs and images linked to catalog objects
  - Data discovery
  - Developing intuition through interaction with selected catalog and image set of known objects
  - Automation of analysis to aid discovery of unknown objects

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

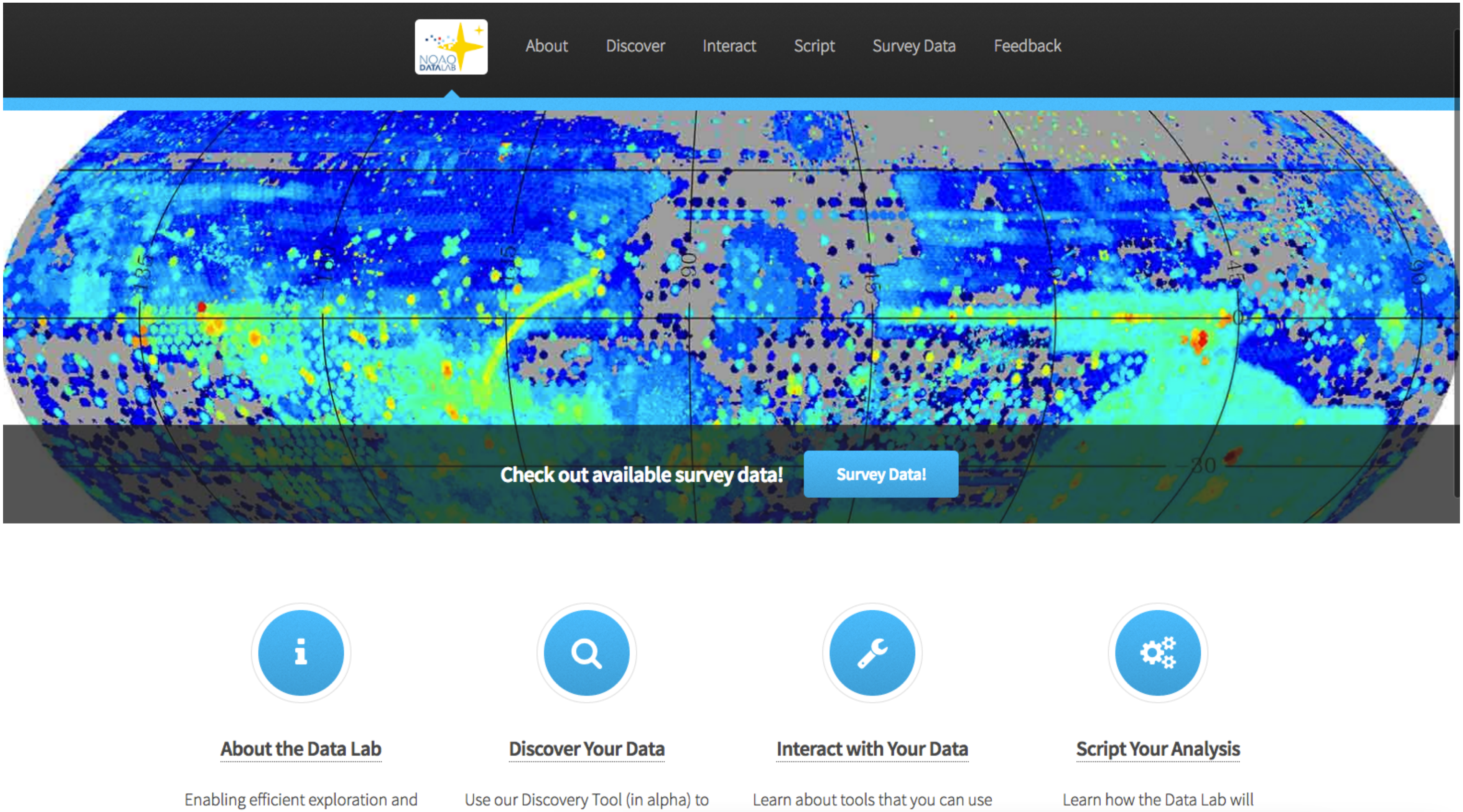
Compute Processing – Data Lab will allow workflows to run close to the data

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

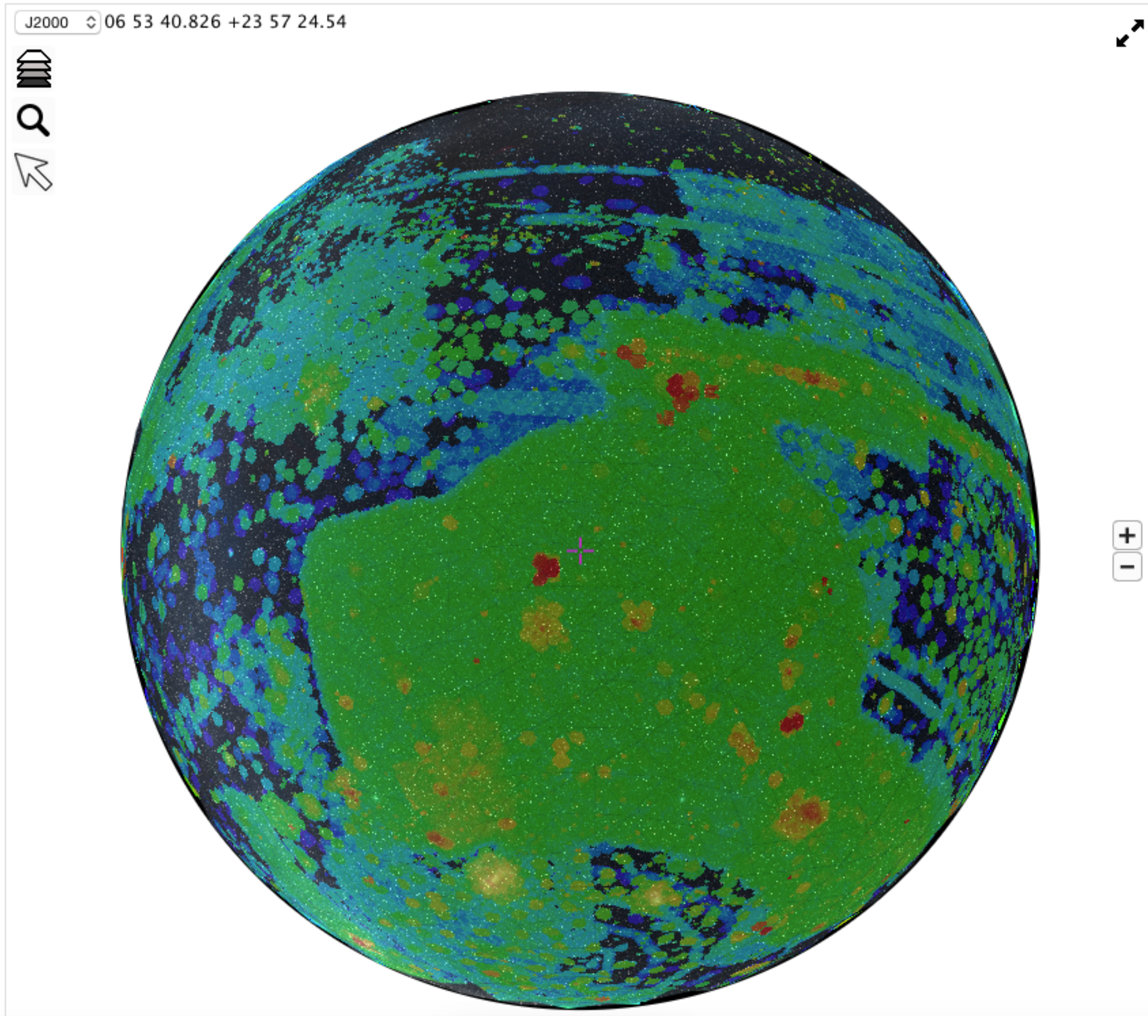
Timeline

- March 2015: Conceptual Design Review
  - Lisa Storrie-Lombardi (Chair), Severin Gaudet, Zeljko Ivezic, Connie Rockosi, Beth Willman reviewed Science Case & Requirements, System Architecture, Operations Concept & Requirements, and Schedule
- Fall 2015 hiring campaign
- June 2016 San Diego AAS Demo
- August 2016 Interim Review
  - Lisa Storrie-Lombardi (Chair), Severin Gaudet, Zeljko Ivezic, Ed Olszewski, Beth Willman, and Dennis Zaritsky reviewed progress and Year 2 plan
- January 2017 AAS SMASH DR1 and DECaLS DR3
- Summer 2017 first public release
- End 2017/Early 2018 DES DR1

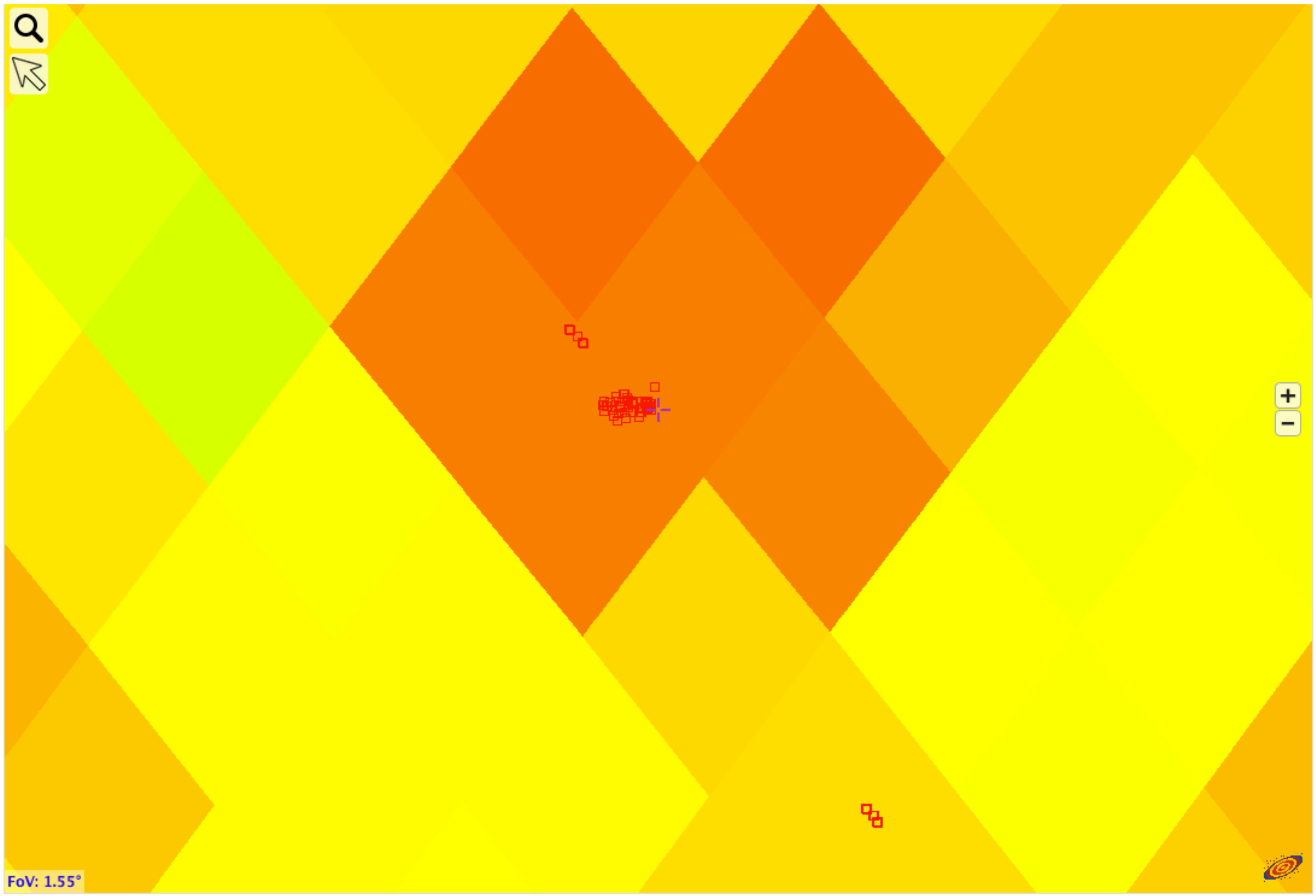
datalab.noao.edu



Data discovery



Data discovery



REFERENCE	FITS_EXTENSION	OBJECT	SURVEY	SURVEYID	PROP_ID	START_DATE	RA
c4d_151112_003815_ori.fits.fz	1	NGC1399_t1_d3_short			2015B-0602	2015-11-11 00:00:00	54.5602958333333
c4d_151112_020032_ori.fits.fz	1	NGC1399_t1_d3_short			2015B-0602	2015-11-11 00:00:00	54.5602041666667
c4d_151111_020907_ori.fits.fz	1	NGC1399_t1_d11_short			2015B-0602	2015-11-10 00:00:00	54.56775
c4d_151111_074627_ori.fits.fz	1	NGC1399_t1_d13_short			2015B-0602	2015-11-10 00:00:00	54.5581208333333
c4d_151112_001200_ori.fits.fz	1	NGC1399_t1_d1_short			2015B-0602	2015-11-11 00:00:00	54.5554208333333
c4d_151112_011749_ori.fits.fz	1	NGC1399_t1_d1_short			2015B-0602	2015-11-11 00:00:00	54.5571291666667
[illegible]	[illegible]	[illegible]			[illegible]	[illegible]	54.5552541666667

Survey data

About Discover Interact Script **Survey Data** Feedback

SMASH

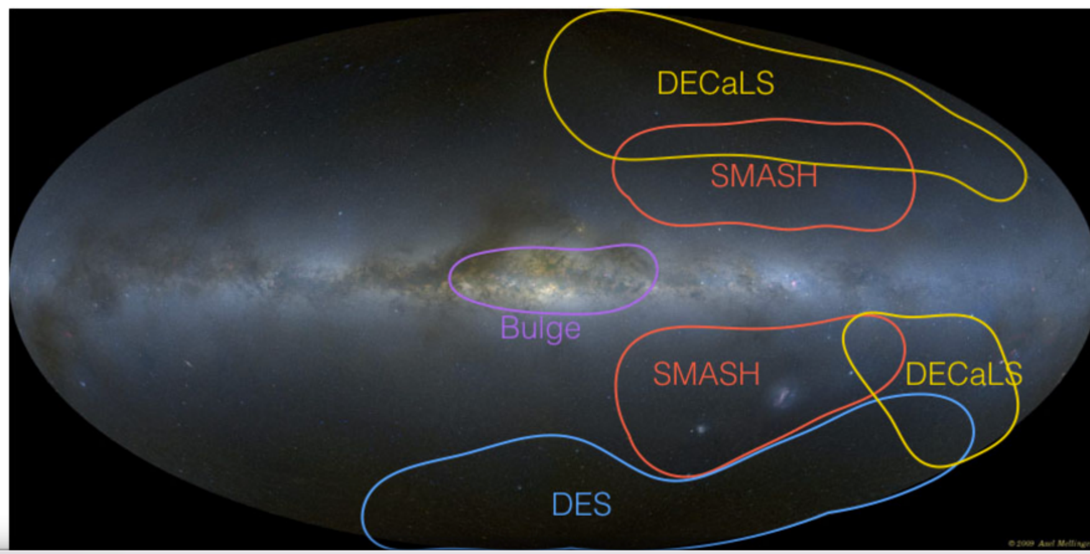
DECaLS

DES
(coming soon)

Bulge
(coming soon)

Survey Coverage

The map below shows the areas covered by surveys with catalog data currently available through the NOAO Data Lab (SMASH, DECaLS) and those expected within approximately a year (DES, Bulge). Hover over an outline to see the survey name or click on an outline or on the sidebar links to go to the page for that survey. For pure image data from these and other observing programs, visit the NOAO Science Archive.



SMASH DR1

Survey of the Magellanic Stellar History (SMASH)

About | Discover | Interact | Script | **Survey Data** | Feedback

Description

- Overview
- Goals
- First Data Release
- Data Reduction and Calibration

Data Access

Analysis

Explore

Results

The SMASH Survey

Overview

The Survey of the Magellanic Stellar History (SMASH) is using DECam to map 480 square degrees of sky to depths of ugriz~24 with the goal of identifying broadly distributed, low surface brightness stellar populations associated with the stellar halos and tidal debris of the Magellanic Clouds. It will eventually contain measurements of approximately 250 million objects distributed in discrete fields spanning an area of about 2400 square degrees. The first data release (DR1) contains ~100 million objects from 61 observed fields. Browse these pages to learn more about SMASH and to access the data. The SMASH overview paper (Nidever et al. 2017) describes the survey in detail, including its goals, survey strategy, reduction, and calibration.

SMASH DR1 Data Access

Description

Data Access

Analysis

Explore

Results

Data Access

The SMASH data are accessible by a variety of means:

Data Lab Table Access Protocol (TAP) service
TAP provides a convenient access layer to the SMASH catalog database. TAP-aware clients (such as TOPCAT) can point to *http://datalab.noao.edu/tap*, select the *smash_dr1* database, and see the database tables and descriptions. *smash_dr1* contains six tables: *chip, exposure, field, object, source,* and *xmatch*. These are described in the schema page.

Data Lab Query Manager

The Query Manager is available as part of the prototype Data Lab software distribution. The Query Manager client provides a Python API to Data Lab database services. For the SMASH DR1 release, these services include only anonymous access through synchronous queries of the catalog made directly to the database. The full public release of the Data Lab Query Manager in the summer of 2017 will include authenticated access, synchronous and asynchronous queries, TAP queries, personal database storage, and storage through the Data Lab VOSpace.

Image cutouts

The Data Lab Simple Image Access (SIA) service provides a fast way to retrieve cutouts from SMASH images. For an example of how to use the SIA service, see this Jupyter notebook.

Jupyter Notebook Server

The Data Lab Jupyter Notebook server contains examples of how to access and visualize the SMASH catalog.

FTP access

SMASH DR1 Data Analysis

Survey of the Magellanic Stellar History (SMASH)

Description

Data Access

Analysis

Explore

Results

Analysis

Jupyter Notebook Server

We have set up a public Jupyter Notebook server to allow anonymous access and exploration of the SMASH catalog and images. By clicking this link, you will start an instance of this server running. You can make changes to the example notebooks, but note that these changes will disappear once you close the page or the browser.

Example notebooks in the Data Lab Notebook server

You can view static versions of the example notebooks contained on the Jupyter Notebook server by selecting a notebook from the list below:

- Basic access (field list, avg. photometry of a field, single-source light curve)
- Interactive filtering and plotting (Hydra II dwarf galaxy discovery demonstration)
- Making an interactive source density map
- Identifying ugr dropout candidates (Simple Image Access search and retrieval)
- Demonstrating criteria for separating stars and galaxies in the SMASH catalog (visualization of millions of points)

Identifying *r*-dropouts

Catalog query

For our query, we will look for objects that are undetected or have large errors in u, g, and r, but are detected and have small errors in i and z. We will only keep objects that have a match in the ALLWISE catalog. Using subqueries to limit the object and xmatch tables using indexed columns makes the query run much faster than it would otherwise.

In [2]:
```
%%time
db1='smash_dr1.object' # the SMASH object table with average magnitudes
db1sel='db1.fieldid,db1.id,db1.ra,db1.dec,db1.umag,db1.gmag,db1.rmag,db1.imag,' +\
    'db1.zmag,db1.uerr,db1.gerr,db1.rerr,db1.ierr,db1.zerr,db1.depthflag' # select ID, coordinates, and mags
db2='smash_dr1.xmatch' # the SMASH cross-match table, which contains cross-matches to ALLWISE
db2sel='db2.wise_id,db2.wise_w1mag,db2.wise_w1err,db2.wise_w2mag,db2.wise_w2err' # ALLWISE W1&W2 mags
db1where='(db1.ndetu=0 or db1.uerr>0.3) and ' + \
        ' (db1.ndetg=0 or db1.gerr>0.3) and ' + \
        ' (db1.ndetr=0 or db1.rerr>0.3) and ' + \
        ' (db1.ndeti>0 and db1.ierr<0.1) and ' + \
        ' (db1.ndetz>0 and db1.zerr<0.1)' # pick ugr dropouts
db2where='(db1.id=db2.id)' # only pick dropouts that are found in ALLWISE W1

# Create the query string.
query = 'SELECT '+db1sel+','+db2sel+' FROM (SELECT * FROM '+db1+' WHERE depthflag > 1) AS db1, '+ \
    '(SELECT * FROM '+db2+' WHERE wise_match=1) AS db2 ' +\
    'WHERE ('+db2where+' and '+db1where+')'

print "Your query is:", query
print "Making query"

# Call the Query Manager Service
response = queryClient.query(token, adql = query, fmt = 'csv')
df = pd.read_csv(StringIO(response))

print len(df), "objects found."
```

```
Your query is: SELECT db1.fieldid,db1.id,db1.ra,db1.dec,db1.umag,db1.gmag,db1.rmag,db1.imag,db1.zmag,db1.uerr,db1.ger
r,db1.rerr,db1.ierr,db1.zerr,db1.depthflag,db2.wise_id,db2.wise_w1mag,db2.wise_w1err,db2.wise_w2mag,db2.wise_w2err FR
OM (SELECT * FROM smash_dr1.object WHERE depthflag > 1) AS db1, (SELECT * FROM smash_dr1.xmatch WHERE wise_match=1) A
S db2 WHERE ((db1.id=db2.id) and (db1.ndetu=0 or db1.uerr>0.3) and  (db1.ndetg=0 or db1.gerr>0.3) and  (db1.ndetr=0 o
r db1.rerr>0.3) and  (db1.ndeti>0 and db1.ierr<0.1) and  (db1.ndetz>0 and db1.zerr<0.1))
Making query
5769 objects found.
CPU times: user 52.3 ms, sys: 11.1 ms, total: 63.3 ms
Wall time: 39.2 s
```

NOAO DATALAB

Identifying *r*-dropouts

Displaying the cutouts

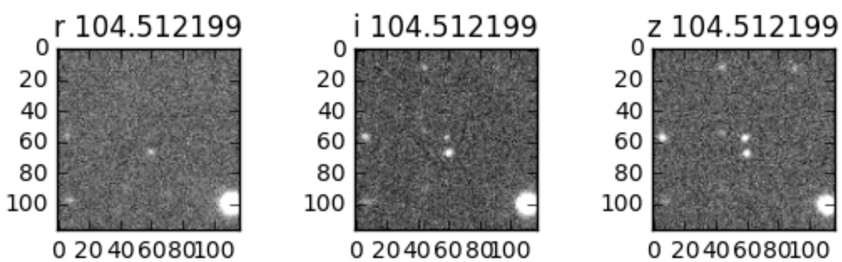
Now let's show the cutouts. The object in question is indeed invisible in the r-band image, but is visible in both i and z, and appears point-like.

In [12]:
```
a1=plt.subplot2grid((2,8),(0,0),rowspan=2,colspan=2)
imgplot = plt.imshow(rimg)
a1.set_title('r '+id1.astype('string'))

a2=plt.subplot2grid((2,8),(0,3),rowspan=2,colspan=2)
imgplot = plt.imshow(iimg)
a2.set_title('i '+id1.astype('string'))

a3=plt.subplot2grid((2,8),(0,6),rowspan=2,colspan=2)
imgplot = plt.imshow(zimg)
a3.set_title('z '+id1.astype('string'))
```

Out[12]:
```
<matplotlib.text.Text at 0x7fc06272fe90>
```



To go through the whole list of cutouts, the code from this notebook would be best put into a Python script and run from the command line, saving the images or making a figure showing all of the candidate objects at once.

Discovering Hydra II dwarf

jupyter SMASH_HydraII_interactive Last Checkpoint: Last Monday at 11:37 PM (autosaved)

Control Panel | Logout

File Edit View Insert Cell Kernel Widgets Help | Python 2

Code | CellToolbar

Query the SMASH DR1 database

We will query the averaged photometry table from the SMASH catalog and select Field 169, which we know contains the Hydra II dwarf.

In [2]:

```
field = 169                      # SMASH Field Number to query
depth = 1                       # minimum depth
raname = 'ra'
decname = 'dec'
mags = 'gmag,rmag'
dbase='smash_dr1.object'
fid = 'fieldid'

# Create the query string.
query = ('select '+raname+','+decname+','+mags+',depthflag from '+dbase+ \
        ' where ('+fid+' = \'%d\' AND' \
        ' (depthflag > %d) and ' + \
        ' (abs(sharp) < 0.5) and ' + \
        ' (gmag is not null) and ' + \
        ' (gmag between 9 and 25) and ' + \
        ' ((gmag-rmag) between -1.5 and 3.0))') % \
         (field, depth)

print "Your query is:", query
```

```
Your query is: select ra,dec,gmag,rmag,depthflag from smash_dr1.object where (fieldid = '169' AND (depthflag > 1) and
  (abs(sharp) < 0.5) and  (gmag is not null) and  (gmag between 9 and 25) and  ((gmag-rmag) between -1.5 and 3.0))
```

We issue the query through the Query Manager, which connects directly to the database.

In [3]:

```
%%time
print "Making query"
# Call the Query Manager Service
response = queryClient.query(token, adql = query, fmt = 'csv')
df = pd.read_csv(StringIO(response))

print len(df), "objects found."
```

```
Making query
297788 objects found.
CPU times: user 307 ms, sys: 57.1 ms, total: 364 ms
Wall time: 12.5 s
```

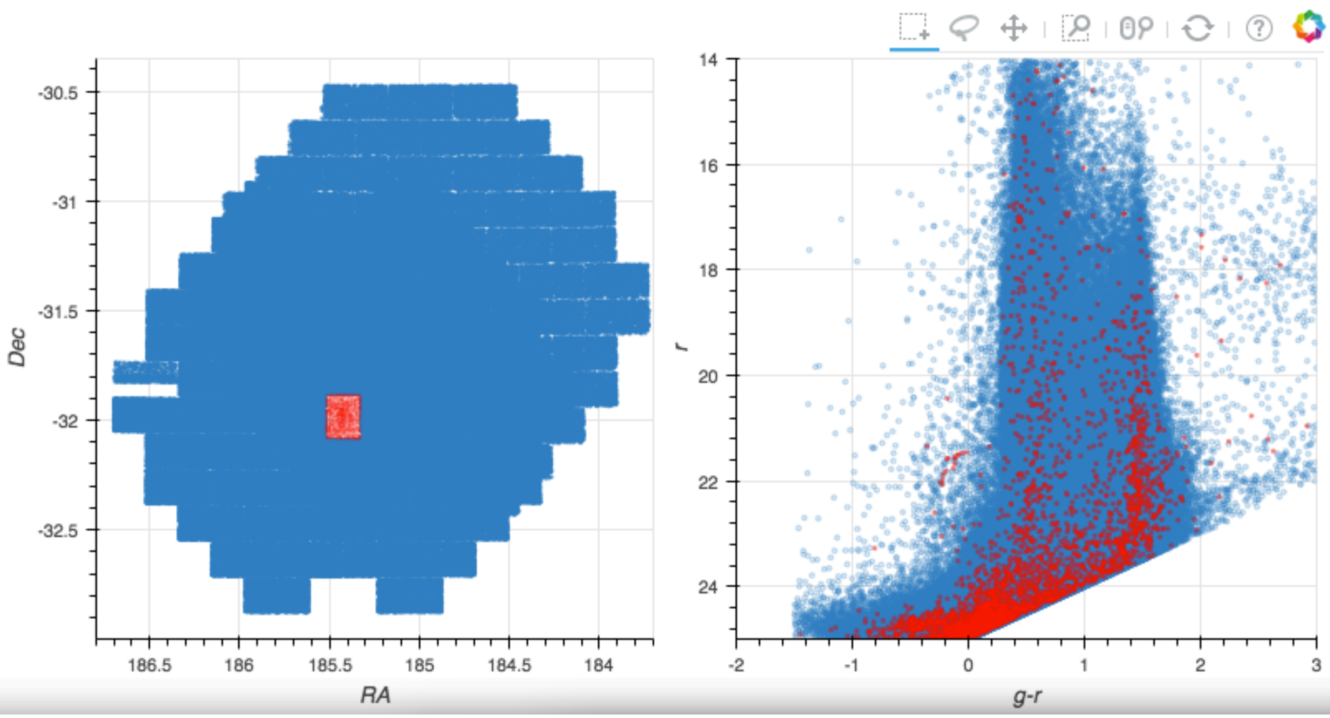
NOAO DATALAB

Discovering Hydra II dwarf

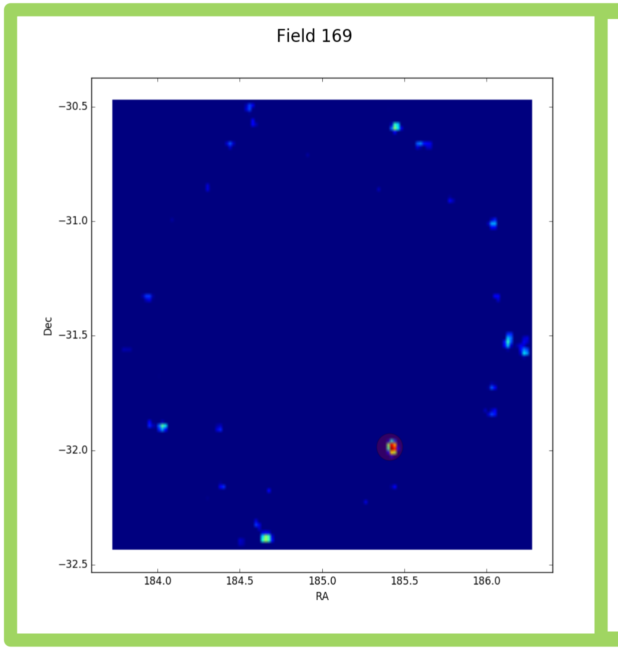
The plots

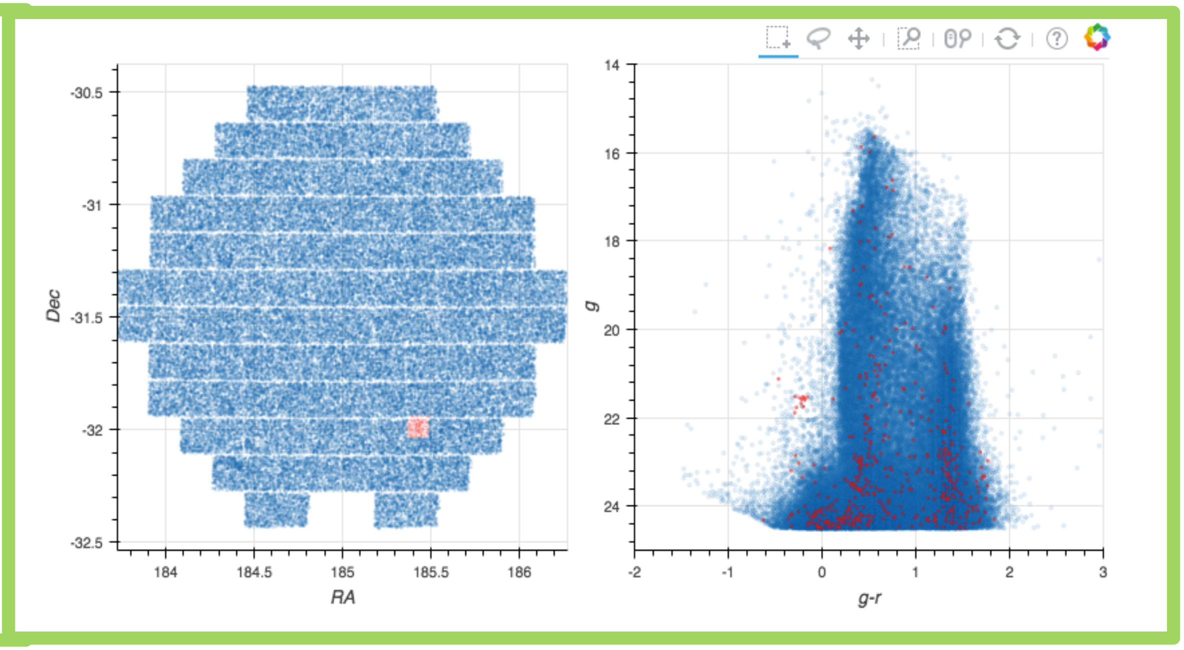
Finally, we render the plot. The figures are interactive, with ability to pan, zoom, and select samples of data that are then updated in the other plot. With the large number of points used here, the interaction can be a little slow, depending on browser and hardware. Try Box Select on the clump of points at lower left, where Hydra II is lurking.

```
In [12]: show(p)
```



Automation of workflow





From Poster 154.25

Coming in 2017

- Authentication
- Asynchronous queries and myDB through Query Manager
- Virtual storage and disk allocation
- Compute service
- Feedback? Visit datalab.noao.edu or contact us at datalab@noao.edu