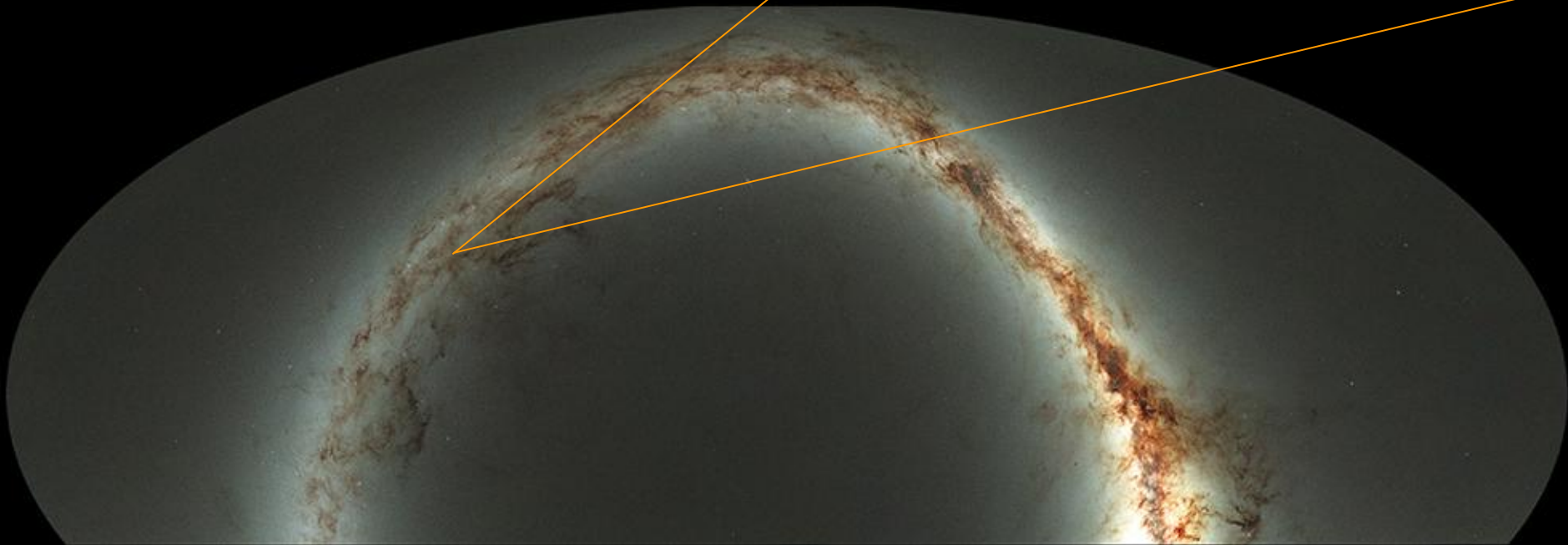


Using Astronomy Big Data and National Cyberinfrastructure to Drive AI Access and Innovation

Curt Dodds - Institute for Astronomy
University of Hawaii, Manoa

Pan-STARRS Milky Way Gigapan



Astronomy Big Data

Sources of Big Data

Observation (ground, space)

Simulation

Surveys

Long duration time-series telescope observations

Moore's Law

Increased image size and dimensionality

Increased simulation grid resolution and step frequency

Astronomy Big Data

Solar System

Sun, asteroids, comets, planets

Galactic

Stars

Exoplanets

Extragalactic

Galaxies, quasars

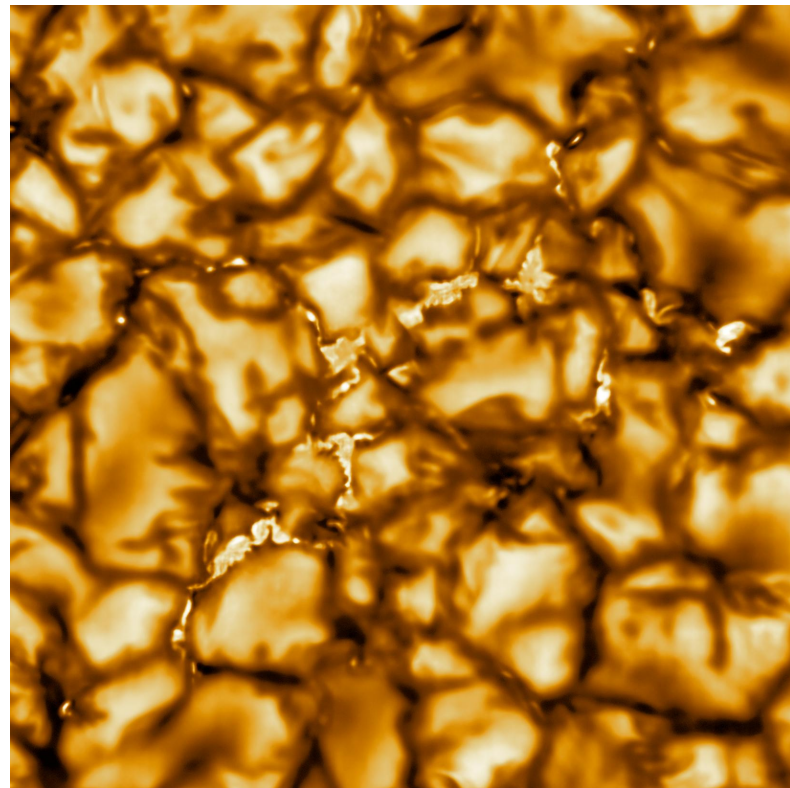
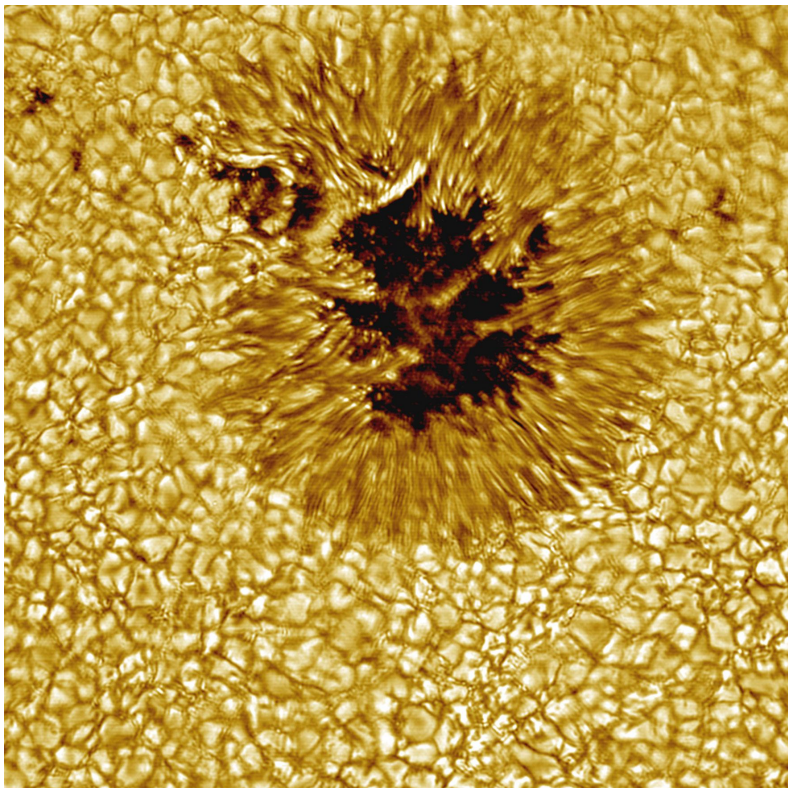
Cosmology

The Sun

Daniel K Inouye Solar Telescope (DKIST)

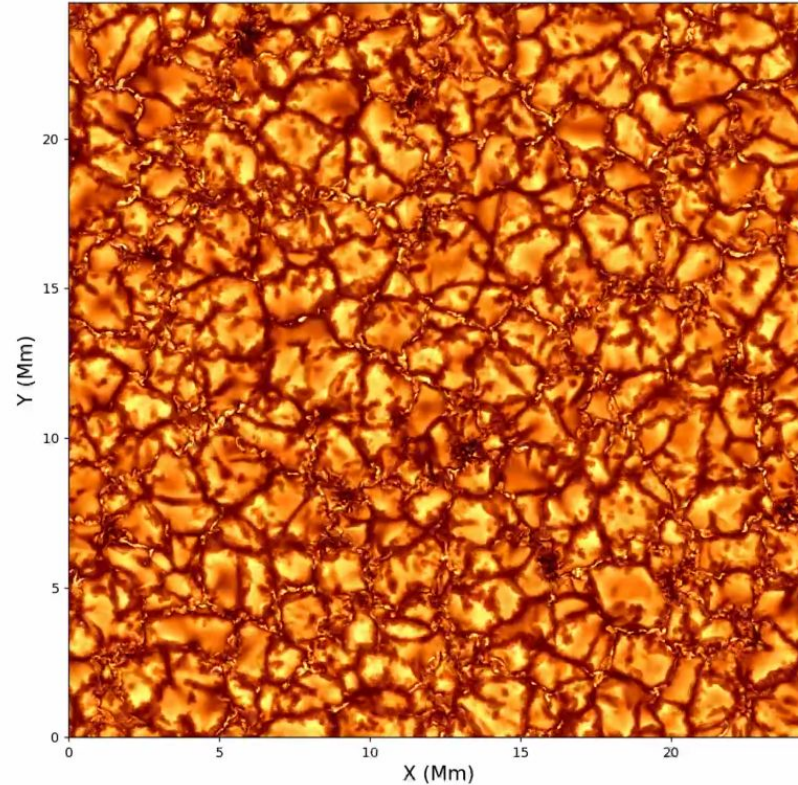


Daniel K Inouye Solar Telescope (DKIST)

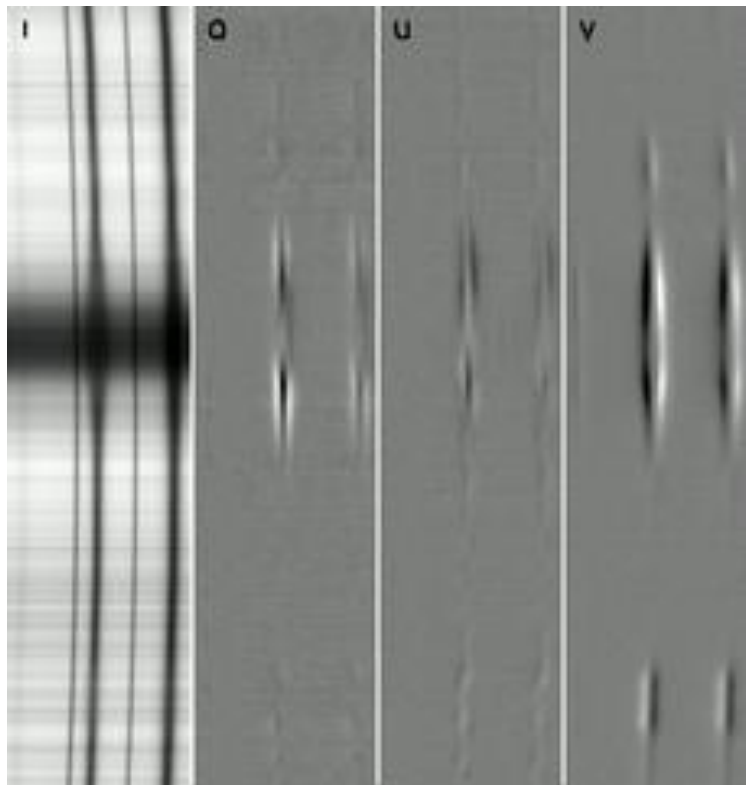


Spectropolarimetric Inversion in 4-Dimensions (SPIN4D)

SPIN4D SSD 200G, Continuum Intensity, time=1.200 hours



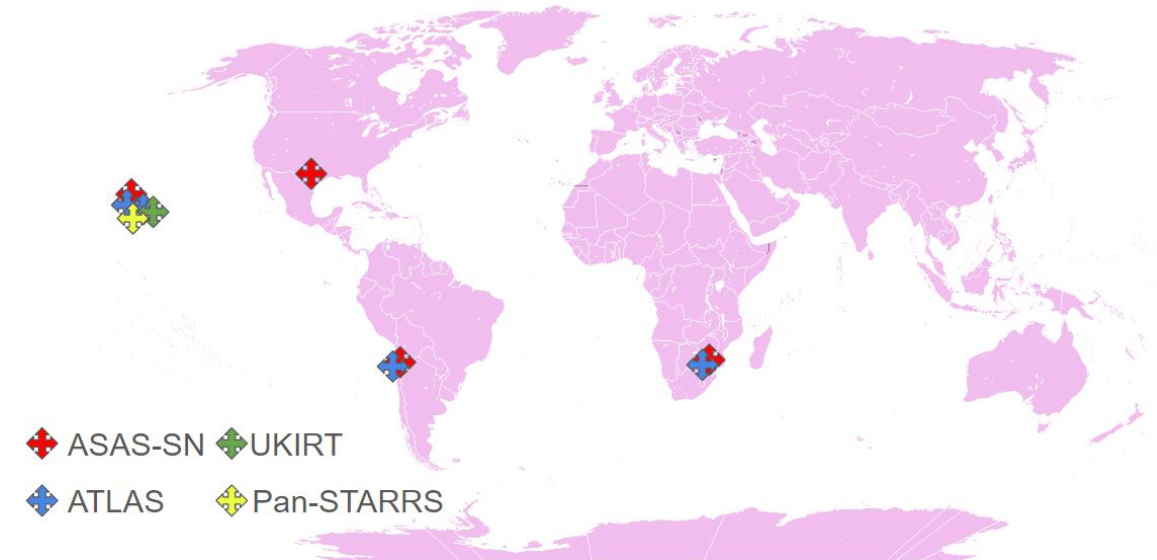
Hinode Solar Optical Telescope Spectropolarimeter



All Sky Surveys

All-Sky Surveys

- [All-Sky Automated Survey for Supernovae](#) (ASAS-SN)
- [Asteroid Terrestrial-impact Last Alert System](#) (ATLAS)
- [Panoramic Survey Telescope and Rapid Response System](#) (Pan-STARRS)



All-Sky Surveys

Time-domain Astronomy

- Variable stars
- Supernovae (exploding stars)
- Solar flares and coronal mass ejections (CME)

Object Classification

- Galaxy, quasar, star, asteroid, comet, supernova, variable star type

Regression

- Estimated photometric redshift (distance from Earth)

ASAS-SN Sky Patrol 2.0 light curve service

ASAS-SN Sky Patrol ID: 103079449737 (317.63137737, 47.36427584)

ATLAS-REFCAT2

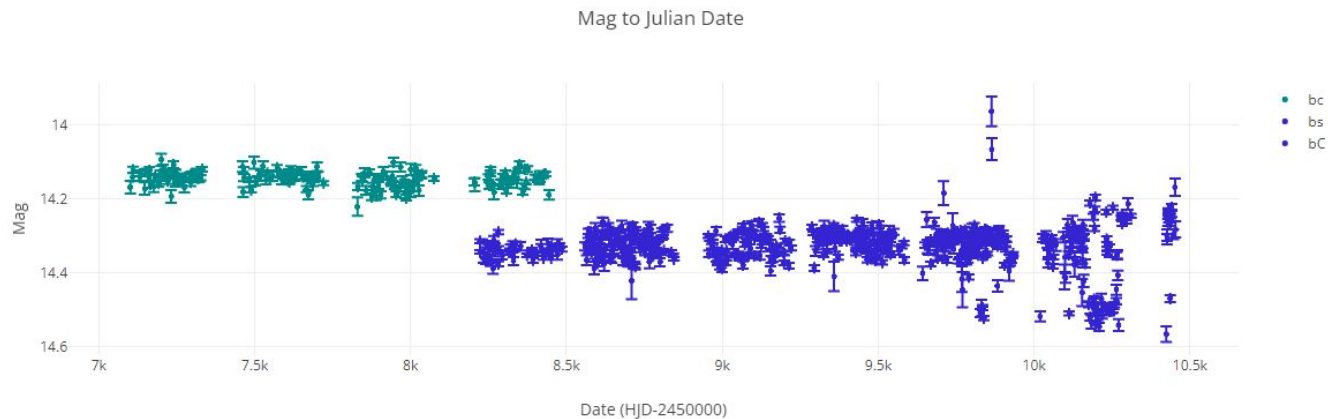
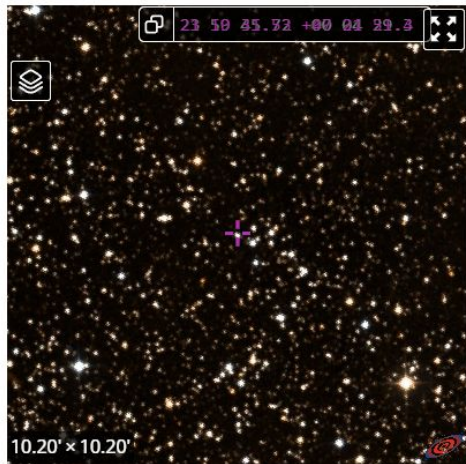
164833176313777790

Gaia DR2

2164493303358851840

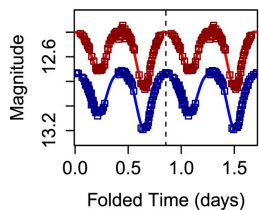
TIC

136701111

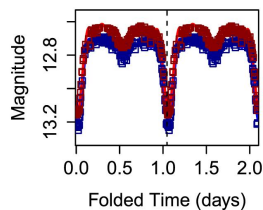


ATLAS-VAR data release of light curves with classification

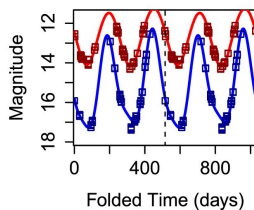
V1014 Cas (contact binary)



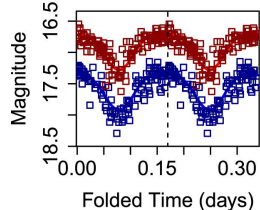
V0495 Aur (detached binary)



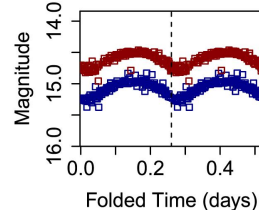
AL Vul (Mira star)



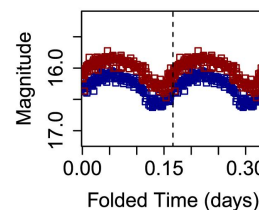
ATO J072.4005+49.1182



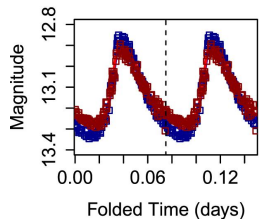
ATO J095.2167+02.8272



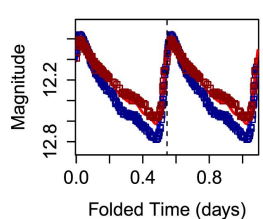
ATO J107.3917-29.8918



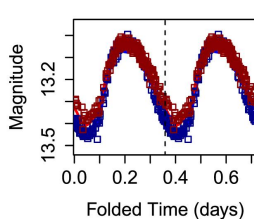
V0460 And (delta Scuti star)



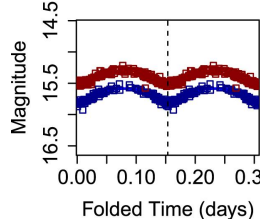
EZ Cnc (RRab)



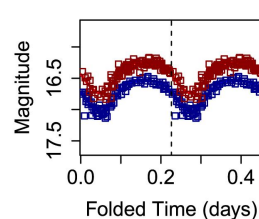
AM LMi (RRc)



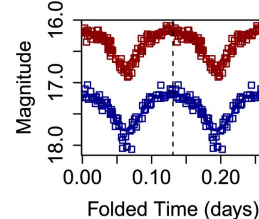
ATO J124.0019-16.9703



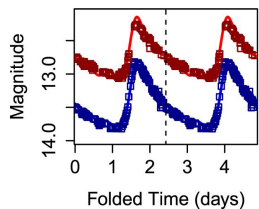
ATO J272.0595+24.3011



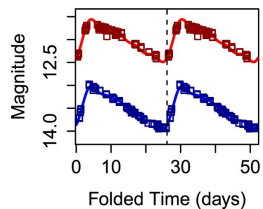
ATO J287.2866+18.0397



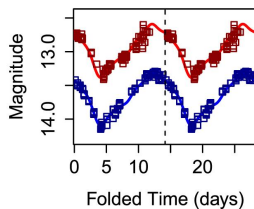
V0913 Mon (classical Cepheid)



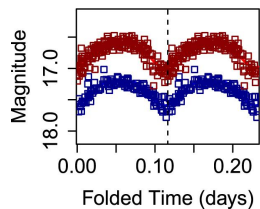
OT Per (classical Cepheid)



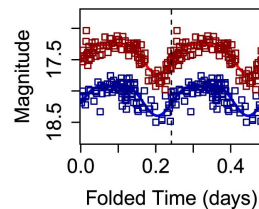
V0801 Aql (W Virginis star)



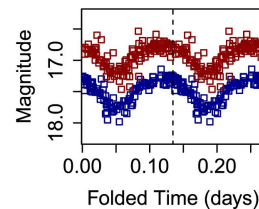
ATO J297.3013+44.7629



ATO J312.0597+59.4165



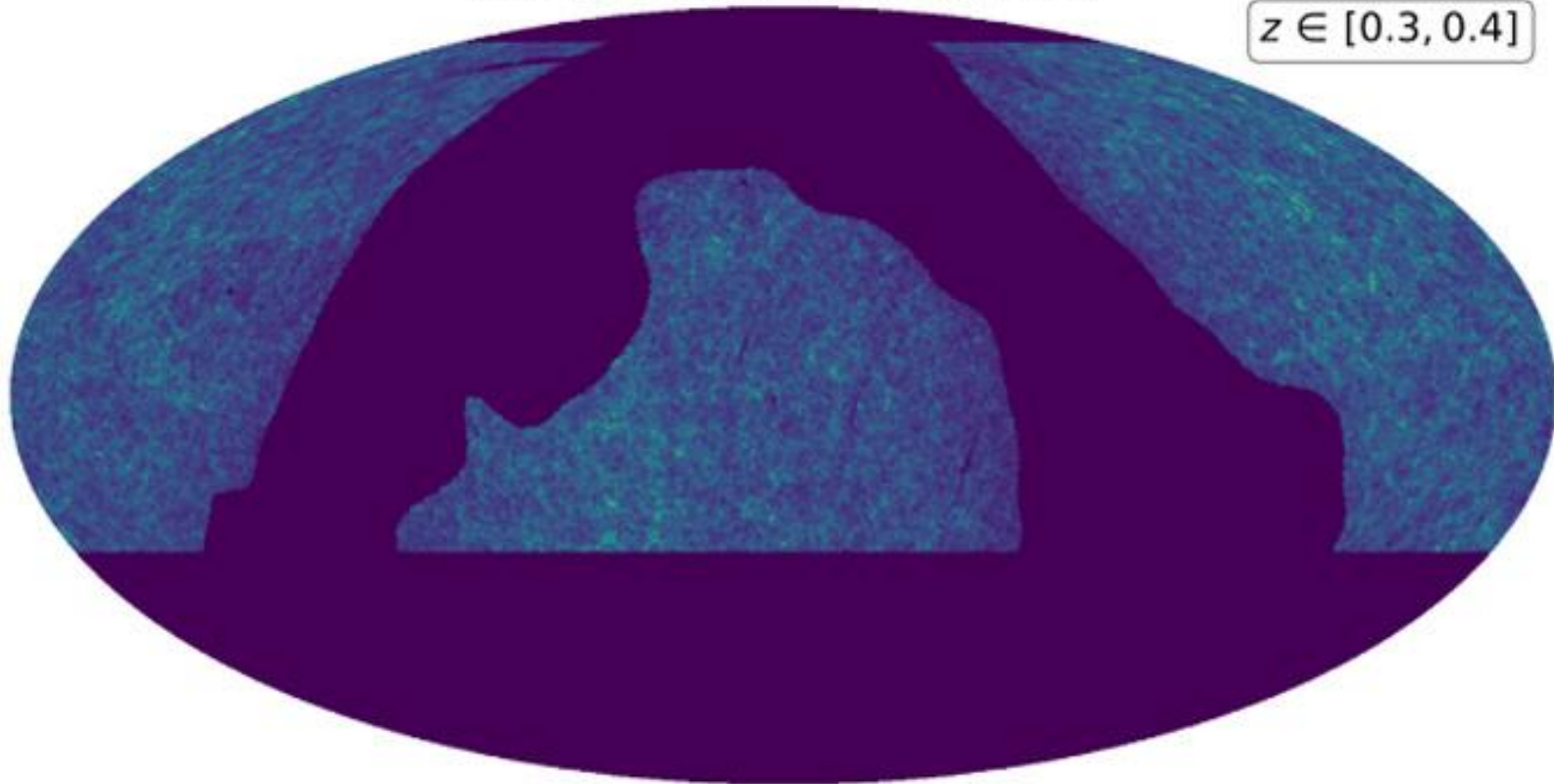
ATO J346.5811+39.4864



Pan-STARRS WISE-PS1-STRM Catalog

WISE-PS1-STRM, galaxies: 9,001,424

$z \in [0.3, 0.4]$



National AI Cyberinfrastructure

National AI Cyberinfrastructure

ACCESS

Open Science Grid

Open Science Data Federation (OSDF) / Pelican Platform

National Research Platform

Commercial cloud providers: EC2, GCP, Azure, etc.

National AI Research Resource (NAIRR) pilot

National Data Platform (NDP)

National Science Data Fabric (NSDF)

Campus HPC, Science DMZ, DTNs

National AI Cyberinfrastructure

ACCESS

Open Science Grid

Open Science Data Federation (OSDF) / Pelican Platform

National Research Platform

Commercial cloud providers: EC2, GCP, Azure, etc.

National AI Research Resource (NAIRR) pilot

National Data Platform (NDP)

National Science Data Fabric (NSDF)

Campus HPC, Science DMZ, DTNs

National Astronomy Cyberinfrastructure

National Astronomy Data

NASA archives were not designed for AI/ML

- Designed before the AI renaissance

- SQL queries with extremely limited result sizes

- Typically $\ll 10$ Gbps bandwidth from archive sites

- Large N^2 crossmatch queries unsupported (but important!)

- Image cutout services are not performant or scalable

Friction prevents researchers (grad students!) from working at scale

Tools and services are fragmented and heterogeneous

Some recent projects have addressed these issues in part (ASAS-SN, DKIST, LSST)

Legacy Data Access

ATLAS Photometry Server

“Next, submit an RA and Dec coordinate to the server to obtain a URL for checking the status. **Note that our request may be throttled if we make too many in a short time.**”

Mikulski Archive for Space Telescopes (MAST)

(Hubble Space Telescope, Pan-STARRS, JWST, Kepler, TESS)

3GB MyDB for query results (**to query 150TB** Pan-STARRS DR2 catalog)

You can retrieve 0.002% of the data BY DESIGN!

Legacy Data Access Patterns

Example: Download ATLAS Variable Stars from MAST

<https://archive.stsci.edu/hlsp/atlas-var> (Heinze et al. 2018)

Shard 360deg into 180x 2deg partitions each 100MB < x < 2GB

Had to use trial and error to determine partition limits

Manually write a download script

Wait 5 days for download of 29GB of data to finish

Legacy Data Access

Example: “A catalog of broad morphology of Pan-STARRS galaxies based on deep Learning”, Hunter Goddard (MS thesis)

The galaxy images were then downloaded using Pan-STARRS *cutout* service. The images are in the JPG format and have a dimensionality of 120×120 pixels as in¹. Pan-STARRS *cutout* provides JPG images for each of the bands. Here we use the images of the g band, as the color images using the y, i, and g bands are in many cases noisy, and do not allow effective analysis of the morphology. The process of downloading the data was completed in 62 days.



<https://krex.k-state.edu/bitstream/handle/2097/41353/HunterGoddard2021.pdf>

Python Example Script

There is a simple [Jupyter notebook script](#) that shows how to access the `ps1filenames.py` and `fitscut.cgi` scripts from Python. It is straightforward to retrieve both FITS image cutouts and JPEG/PNG cutouts. The script also shows an example of retrieving a color JPEG image. The script includes some simple helper functions for retrieving both FITS and JPEG/PNG images and shows how to display the images.

Bulk Image Download Python Script

It is fairly simply to download PS1 cutout images in bulk. Here is a Python script that queries the interfaces described above using a list of RA, Dec positions and extracts 1 arcmin (240 pixel) FITS cutouts for each position and filter. This script can extract more than 4 cutouts per second if you have a sufficiently fast internet connection. It can be easily modified for special requirements (e.g., downloading different kinds of images, or download JPEG images instead of FITS images).

NOTE: If you modify this script to download images in multiple threads, please do not use more than 10 simultaneous threads for the download. The `ps1images` service is a shared resource, and too many requests from a single user can cause the system to be unresponsive for all users. If you attempt to download images at an excessive rate, eventually you will find your downloads blocked by the server.

Sample JPEG Images

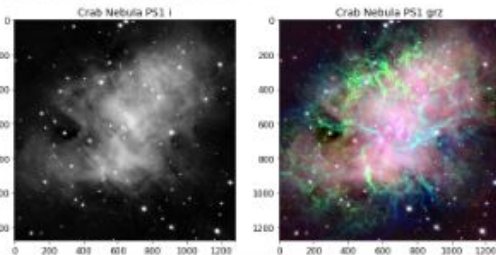
This gets single-band grayscale and color JPEG images at the position of the Crab Nebula. The extracted region size is 1280 pixels = 320 arcsec.

```
# Crab Nebula position
ra = 81.633210
dec = 22.014460
size = 1280

# grayscale image
qln = getqzain(ra,dec,size=size,filter='i')
# color image
qlc = getcolorin(ra,dec,size=size,filters='grs')

pylab.rcParams.update({'font.size':12})
pylab.Figure(1,(2,6))
pylab.subplot(121)
pylab.imshow(qln,origin='upper')
pylab.title('Crab Nebula PS1 i')
pylab.subplot(122)
pylab.title('Crab Nebula PS1 grs')
pylab.imshow(qlc,origin='upper')

Downloading https://ps1images.stsci.edu/cgi-bin/ps1filenames.py?ra=81.63321&dec=22.01444&size=1280&format=fits&filters=i (Done)
Downloading https://ps1images.stsci.edu/cgi-bin/ps1filenames.py?ra=81.63321&dec=22.01444&size=1280&format=fits&filters=grs (Done)
<matplotlib.image.AxesImage at 0x11807e820>
```




New Data Access

https://github.com/asas-sn/skypatrol

NSF PS1 Quasar Wikis Wireless .htaccess LDAP Aut... A Brief Guide to Cali... admon Amazon.com: Super...

README GPL-3.0 license



tests passing Docs failing

ASAS-SN SkyPatrol client

The SkyPatrol pyasasn client allows users to query the ASAS-SN input catalog and retrieve light curves from our photometry database. These light curves are regularly updated with continuous photometry run on nightly images. Read the docs [here](#).

Installation

The easiest way to get started is with pip, using Python 3.6+:

```
python -m pip install skypatrol
```

To build from source:

```
git clone https://github.com/asas-sn/skypatrol.git
pip3 install skypatrol/
```

New Data Access

Alert Brokers

Information for Rubin Observatory Alerts and Community Brokers

What is an Alert Broker?

One of the results of Rubin Observatory's real-time difference image analysis processing is a world-public stream of alerts containing data about transient, variable, and moving sources (find out more about Rubin Observatory processing and products like alerts in the [Data Products Definition Document](#)). These alerts will be distributed to community brokers: software systems that will ingest, process, and serve astronomical alerts from Rubin Observatory and other surveys to the broader scientific community. Typical broker functionality includes cross-match association with archival catalogs, the identification and prioritization of objects in need of follow-up observations, and photometric classification based on light-curve analysis.

How are the Rubin Observatory Alert Brokers Selected?

Due to the anticipated high bandwidth of the Rubin Observatory alert stream, only a limited number of brokers can receive the stream directly from Rubin Observatory. The broker selection process is detailed in [LDM-612](#); all decisions are the responsibility of Rubin Observatory Operations. In May 2019 the process began with Letters of Intent submitted from 15 teams, and in June 2019 a [Community Brokers Workshop](#) was held in Seattle, WA. In August 2019 [the SAC announced that all LOI teams were encouraged to submit full proposals](#). In December 2020, nine teams submitted full proposals. The SAC recommendation that seven brokers receive direct access to the full stream and two operate downstream of the selected brokers was accepted by Rubin Operations.

Where can I find information about pre-LSST alert streams, real and simulated?

First, check out questions one and two of the Rubin Observatory [Answers to Community Broker FAQs](#). For more information about ZTF alerts, see Patterson et al. (2019): [The Zwicky Transient Facility Alert Distribution System](#). An archive of ZTF alerts made available by the University of Washington can be accessed [here](#). ZTF alerts are being received, processed, and made publicly available by [ANTARES](#) (see the [ANTARES FAQ](#) and [ANTARES Devkit](#)), [ALeRCE](#), [Fink](#), and [Lasair](#).

New Data Access

Page Contents

MAST Queries ([astroquery.mast](#))

- [Introduction](#)
- [Getting Started](#)
- [Accessing Proprietary Data](#)
- [Additional Resources](#)
- [Reference/API](#)
 - [astroquery.mast Package](#)
 - [MAST Query Tool](#)
 - [Classes](#)

MAST Queries ([astroquery.mast](#))

Introduction

The Mikulski Archive for Space Telescopes (MAST) is a NASA funded project made to collect and archive a variety of scientific data to support the astronomical community. The data housed in MAST includes science and engineering data, with a primary focus on data sets in the optical, ultraviolet, and near-infrared parts of the spectrum, from over 20 space-based missions. MAST offers single mission-based queries as well as cross-mission queries. Astroquery's `astroquery.mast` module is one tool used to query and access the data in this Archive.

`astroquery.mast` offers 3 main services: `MastClass`, `CatalogsClass`, and `Cutouts`. `MastClass` allows direct programatic access to the MAST Portal. Along with `ObservationsClass`, it is used to query MAST observational data. The `Catalogs` class is used to query MAST catalog data. The available catalogs include the Pan-STARRS and Hubble Source catalogs along with a few others listed under the Catalog Queries section of this page. Lastly, `Cutouts`, a newer addition to `astroquery.mast`, provides access to full-frame image cutouts of Transiting Exoplanet Survey Satellite (TESS), MAST Hubble Advanced Product (HAP), and deep-field images, through `TesscutClass`, `HapcutClass`, and `ZcutClass` respectively. For a full description of MAST query options, please read the [MAST API Documentation](#).

Driving AI Innovation

Driving AI/ML Innovation

Reduce time to get started

Data discovery as a service

Data exploration as a service

Data ready for AI/ML training

Preprocessing adjacent to data origin

High throughput data distribution optimized for Pytorch, Keras

Transparent data caching

Eliminate sources of friction

Driving AI/ML Innovation

- Support novel data access patterns
- Online training data for AI/ML on time-series
- Real-time data sources
- AI/ML inference applications
- Data exploration without data movement
- Data preprocessing without data movement
- Move only the data you want
- Transparent caching for efficiency and performance

OSDF/Pelican

Hawaii OSDF Data Origins

- Participate in OSDF/Pelican
- Deploy data origin service on UH/IfA DTNs
- Deploy data origin service on CC* HPC storage
- Internal outreach to researchers
 - Who produce data
 - Who consume data

Hawaii OSDF Data Origins

IfA DTNs

dtn-itc

- [Hinode SOT SP](#) solar observations and inversions mirror from High Altitude Observatory in Boulder, CO
- Critical Early DKIST Science: Spectropolarimetric Inversion in Four Dimensions with Deep Learning ([SPIN4D](#))
- [ATLAS-VAR](#) variable star light curves

dtn-max (Baltimore)

dtn-naoj (Tokyo)

dtn-hurp (Hilo, Hawaii)

dtn-uk (planned - London)

Hawaii OSDF Data Origins

UH CC* KoaStore data origin (new):

- CC* UH 800TB set aside for data federation using OSDF

Datasets (work in progress)

- ASAS-SN - light curves for any source
- SPIN4D - solar photosphere simulation
- Hinode SOT SP - solar spectropolarimetric survey
- ATLAS-VAR - variable stars
- StePS - cosmological N-body simulation

NRP

Institute for Astronomy K8s/NRP

Heterogeneous K8s cluster in Hawaii

- 640 CPU cores
- 8x L40S GPU, 2x V100GPU

Federate to NRP

Storage integration

- on-premise project storage clusters (ATLAS, ASAS-SN, SPIN4D, Pan-STARRS, H20)
- campus HPC Lustre storage cluster
- IfA DTNs

Data Services

Vision: to make siloed astronomy data from Hawaii available for ML training on NAIRR, NDP, NRP, OSG and other HPC resources.

Objectives:

Dataset discovery service on OSDF data origin, UH DTNs

Dataset discovery/exploration on OSG, NRP resources (Jupyter Notebook)

Dataset streaming service on OSDF data origin, UH DTNs

Dataset client streaming to OSG, NRP resources (Jupyter Notebook, PyTorch, Keras)

Extract-Transform-Distribute (ETD)

ETD Data Discovery and Streaming Service is deployed adjacent to a data source using containers, (Docker, K8s).

Discovery - enumerate available datasets, file exploration and access

Extract - select, slice and sample from data sources

Transform - process extracted examples for AI training, e.g.

[torch.utils.data.DataLoader](#) and [tf.data.Dataset](#)

Distribute - asynchronous parallel streaming

Proof of Concept at Univ. of Hawaii using DTNs, NRP, OSDF

Applications for education, training, transfer-learning, real-time inference

Resources

[ACCESS](#)

[Open Science Grid \(OSG\)](#)

[Open Science Data Federation \(OSDF\)](#)

[Pelican Platform](#)

[National Research Platform \(NRP\)](#)

[National AI Research Resource \(NAIRR\) pilot](#)

[National Data Platform \(NDP\)](#)

[National Science Data Fabric \(NSDF\)](#)

[Science DMZ](#)

[Data Transfer Node \(DTN\)](#)

Contact Information

Curt Dodds

Institute for Astronomy, University of Hawaii, Manoa

dodds@hawaii.edu