



CASU@IRIS: enabling wide area survey science

Nicholas Walton et al

Institute of Astronomy,
University of Cambridge

Astronomical Data Intensive Science @ IoA

CASU, Gaia, PLATO and AstroMedical

- Develop and run advanced data analysis pipelines for astronomical data
 - Science data fills major archives: e.g. ESO SAF, ESA Gaia Archive
 - CASU operational repository for internal science team releases
- Generation of efficient code
 - Design, implementation, testing, standards, code repositories
 - e.g. CASU code release
- Significant data management and data distribution
 - Interfaces, standards, documentation, hardware configuration and support
- Collaboration in a range of project consortia, large and small
 - Space based: ESA / Ground based: ESO/ ING etc
- Participation in the science programs of the projects
 - Science survey team leadership and membership

CASU: Data for Data Intensive Science

Significant compute infrastructure in place

Installation in the APM includes:

- 20 racks
- 1500 cores
- 2 PB RAID disk
- All connected via 2x1Gbit links to the main UCam backbone
- Recent significant infrastructure upgrades (IMAXT/ WEAVE/ 4MOST/ LSST) to West Cambridge Data Centre)
- Pilot access to CASU@IRIS resources



IRIS @ CASU: Services in build 2020

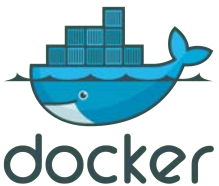
Running pipelines and services in containers



- [Jupyter](#) allows for interactive data access and exploratory analysis close to where the data are located



- [Dask](#) natively scales python, supports parallelism



- [Docker](#) provides stable and reproducible environments in containers



kubernetes

- [Kubernetes](#) (K8s) allows for deployment, scaling and management of containers in a cluster



openstack™

- [OpenStack](#) manages large pools of compute resources

CASU

IRIS

Applications running in containers

- Web archive servers, postage stamp services, on-demand data analysis
- Prototype architecture running pipelines in containers in a Docker swarm cluster
- CASU JupyterHub running in the swarm cluster and spawning analysis notebooks in containers allows interactive analysis closer to where the data are
- Parallel and distributed pipelines.
- Architecture upgraded to Kubernetes, running on top of OpenStack (deployed to IRIS at Cambridge CSD3)

Imaging Science Pipelines

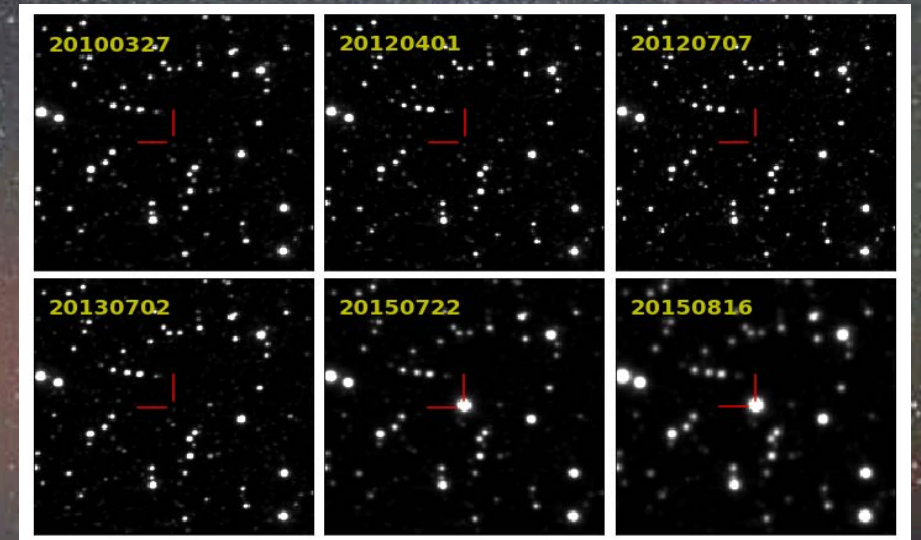
Monthly delivery of data products to Public Survey PIs
Preparation and delivery of Phase 3 data products to ESO
PSF photometry development and deployment for selected surveys (eg Bulge and Magellanic Cloud surveys)



VST + OmegaCAM

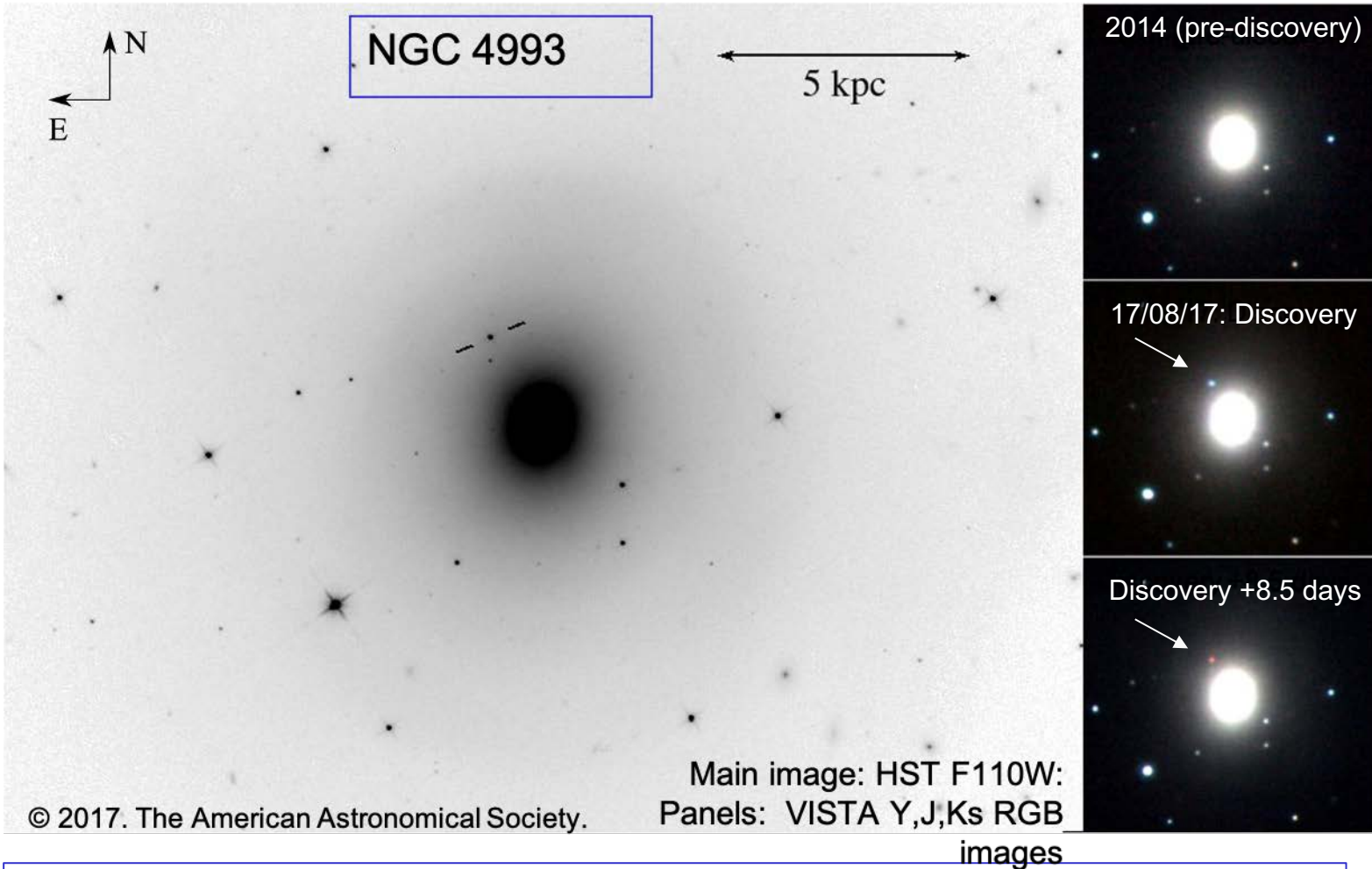


VISTA + VIRCAM

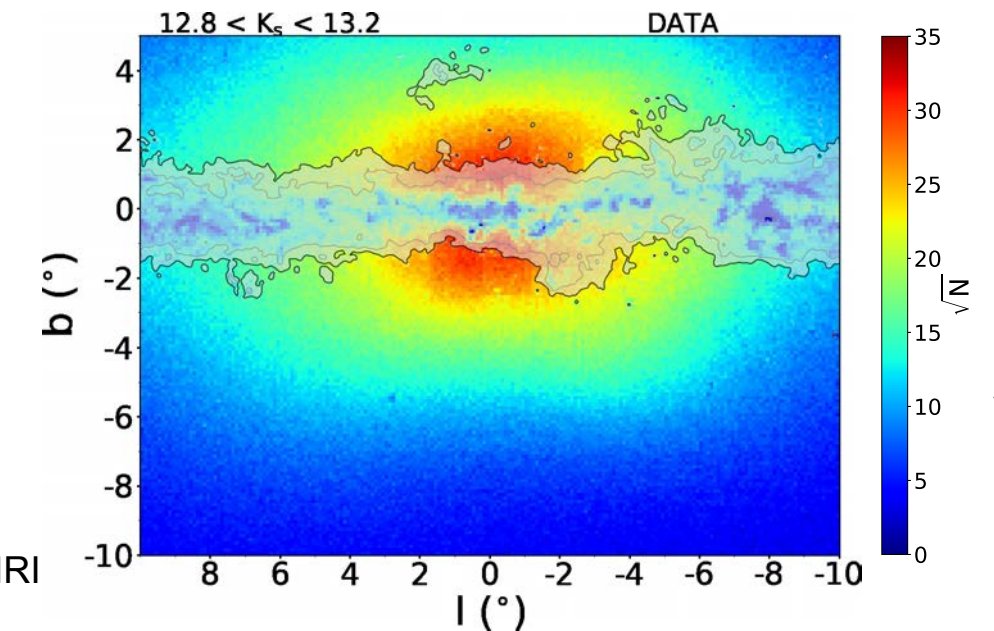
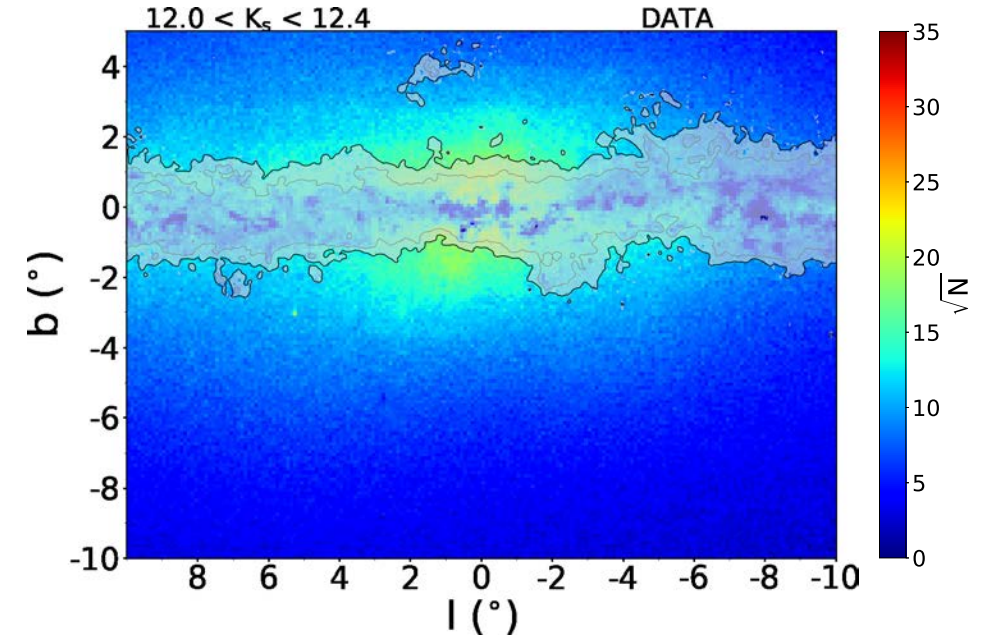


VISTA Alerts: rapid processing

Imaging Surveys: VISTA



IoA developed and operate pipeline processing for VISTA IR surveys



Simion et al, 2017

The Emergence of a Lanthanide-rich Kilonova Following the Merger of Two Neutron Stars: Tanvir, Levan, Gonzalez-Fernandez et al, 2017

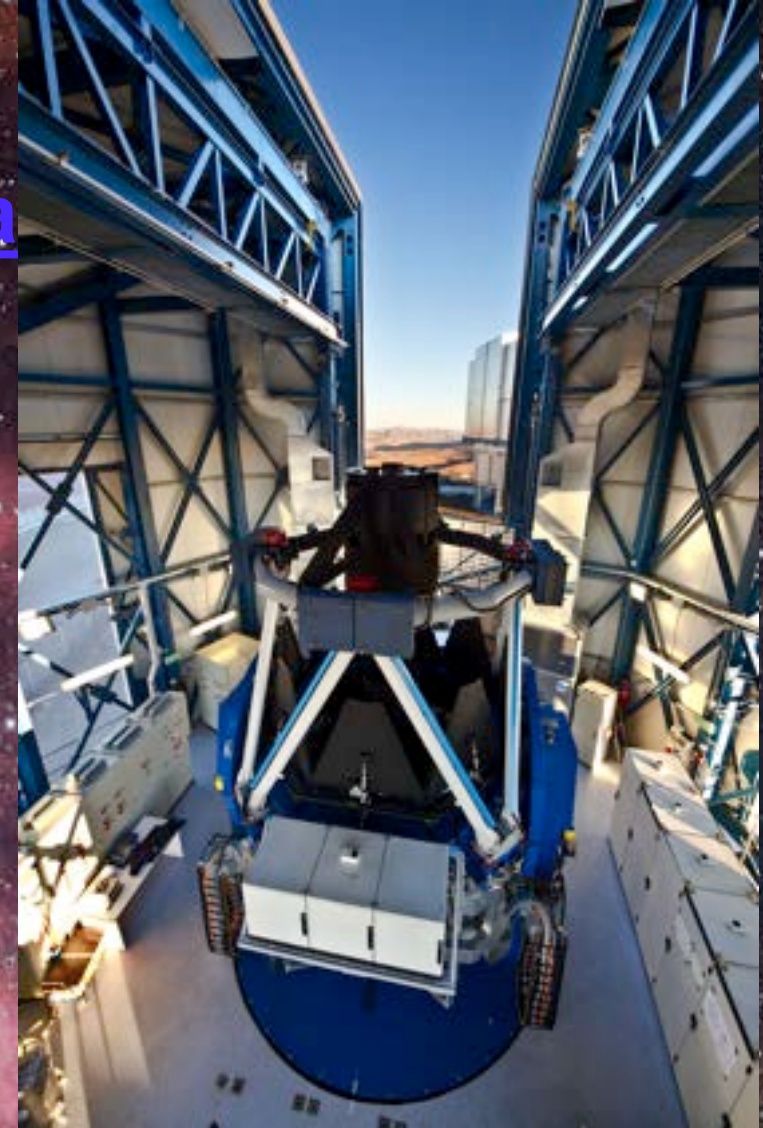
Imaging Surveys: VST

<http://casu.ast.cam.ac.uk/surveys-projects/vista>

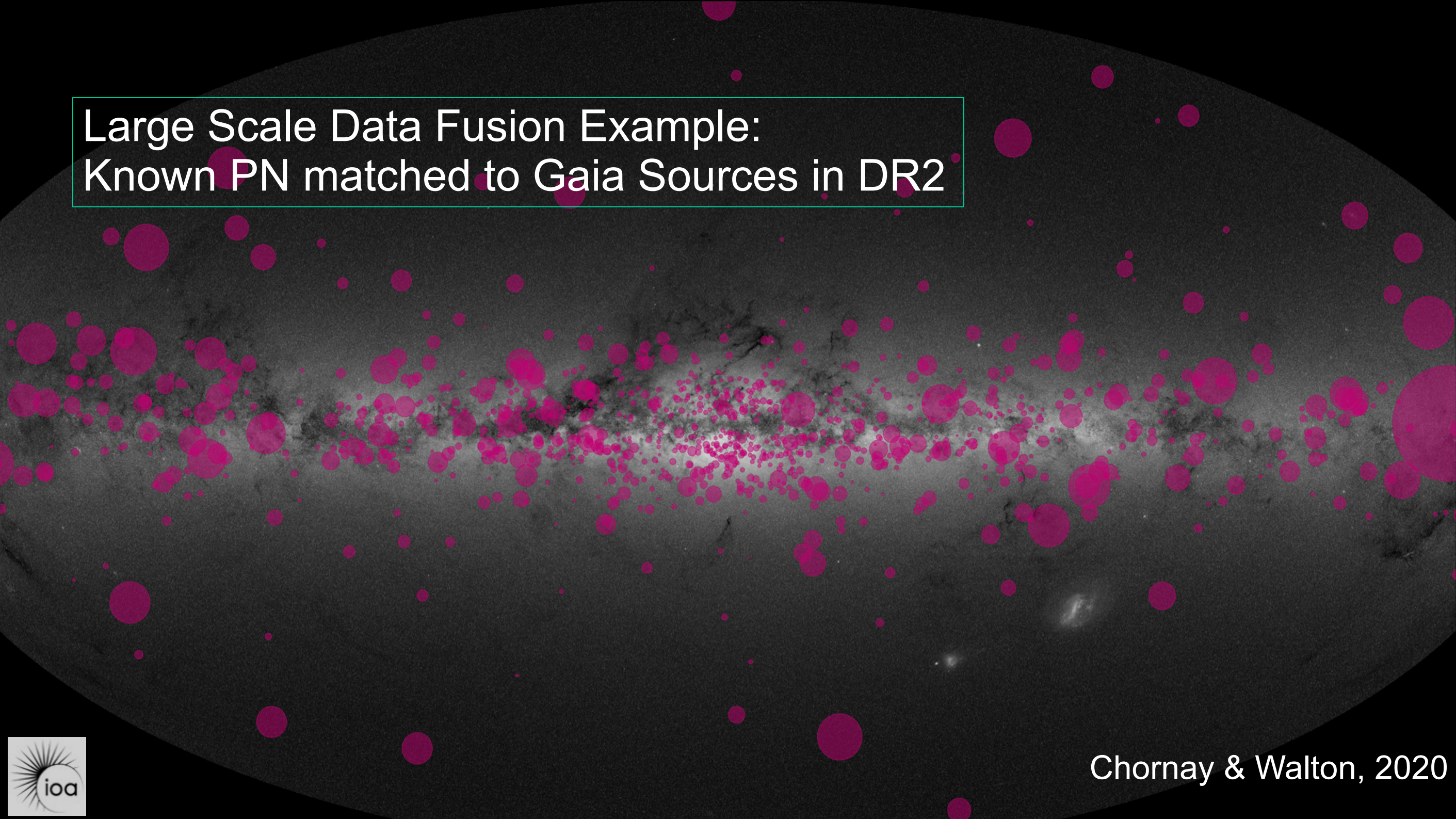
Why? Range of ESO public surveys, e.g. ATLAS, VPHAS, etc

What? IoA developed and operate pipeline processing for VST (Optical) surveys plus support of ESO Phase 3 releases

When? Surveys ongoing, 5+ years

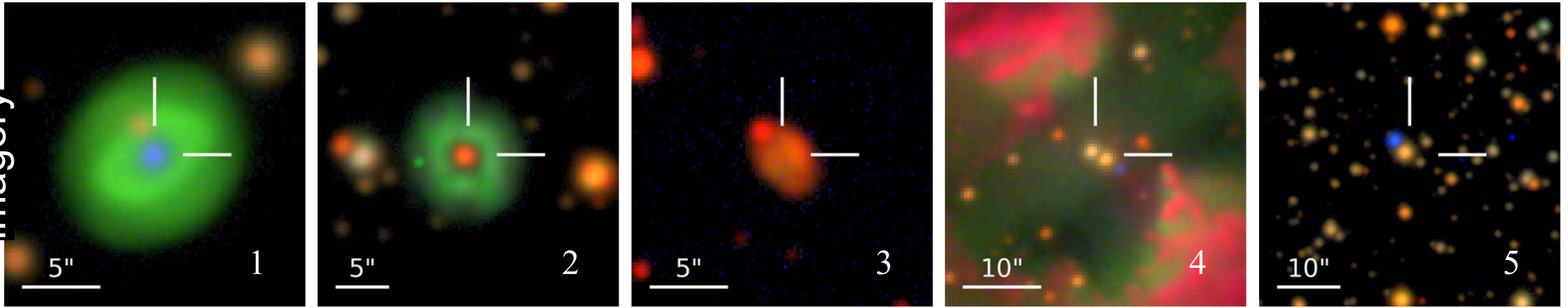


Large Scale Data Fusion Example:
Known PN matched to Gaia Sources in DR2

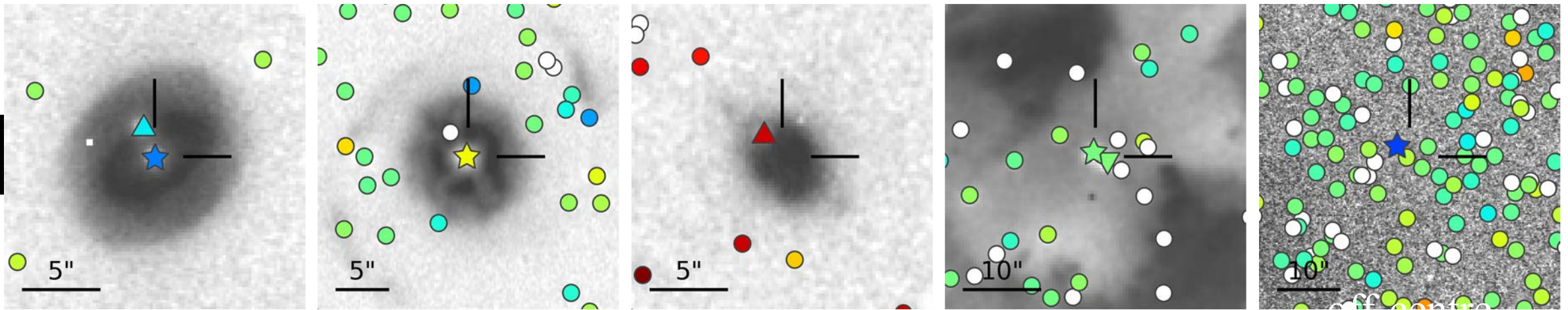


Chornay & Walton, 2020

VPHAS+ ugr
imagery



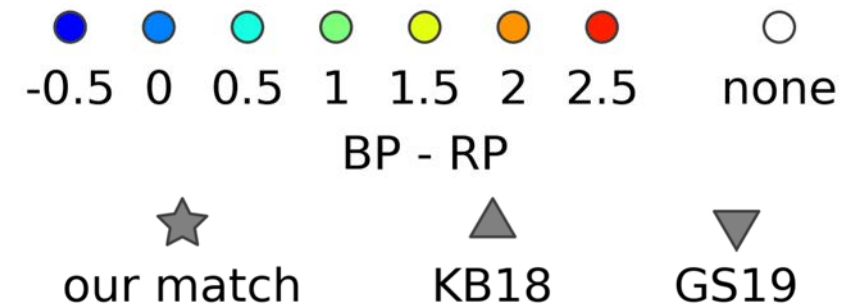
Gaia



Chornay & Walton, 2020

CSPN Match Examples

Automation of cross match and cut out service, access to Gaia catalogues and VST images via Jupyter@CASU@IRIS





World class image analysis powering UK data intensive astronomy

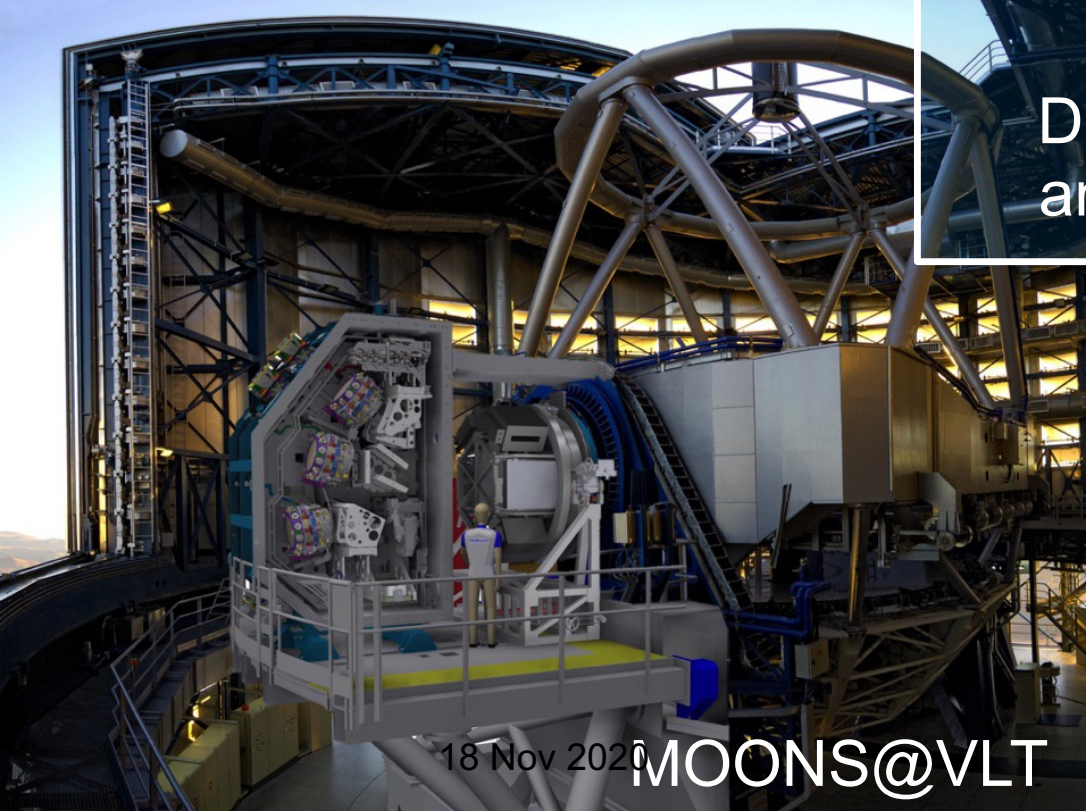
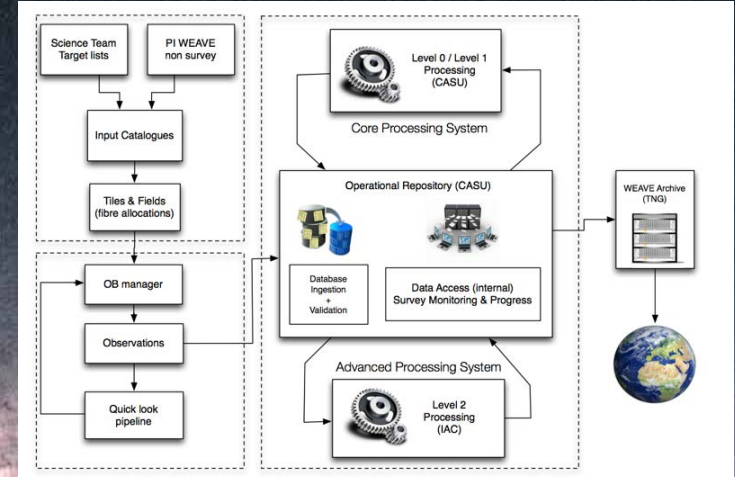
Spectroscopic Pipelines

Design of complete data system architecture

Development of QC pipeline

Development of Operational Repository and spectral extraction pipelines

Design of interfaces to spectral analysis pipelines and Archive



4MOST

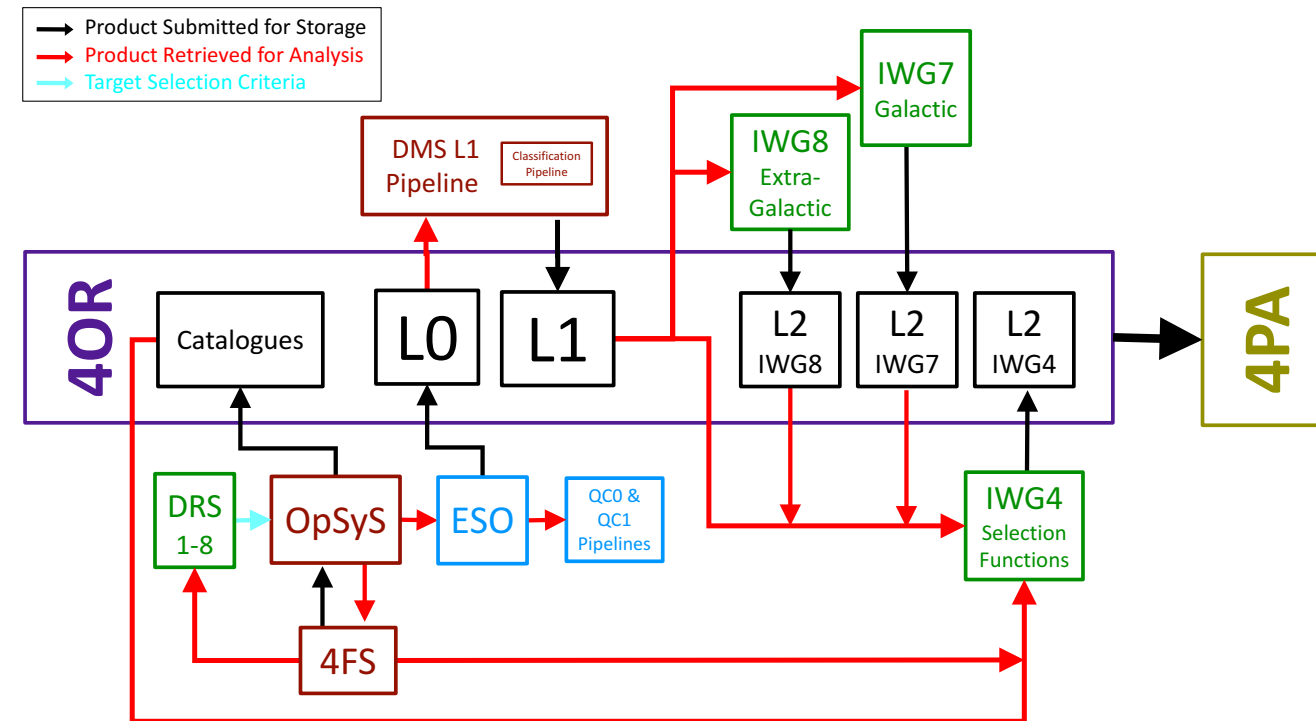
<https://www.4most.eu>

Why? Galactic structure (Gaia complement), High energy sky (eROSITA complement), Cosmology

What? IoA lead development (and future operations) of the complete (QC/L0/L1/L2 Op Repository) processing system

When? Surveys start 2021 with Operations phase 5+5 years/ plus likely 5 year additional survey operations

Science case	S/N / Å	r _{AB} -mags	Targets (Millions)
S1 Milky Way Halo LR Survey	10	16–20.0	1.5
S2 Milky Way Halo HR Survey	140	12–15.5	0.08
S3 Milky Way Disk and Bulge LR Survey	10–30	14–18.5	10.7
S4 Milky Way Disk and Bulge HR Survey	140	14–15.5	1.8
S5 Galaxy Clusters Survey	4	18–22.0	1.1
S6 AGN Survey	4	18–22.0	0.5
S7 Galaxy Evolution Survey (WAVES)	4	18–22.5	1.4
S8 Cosmology Redshift Survey	4	20–22.5	9.4
Total			>25



WEAVE:

<http://www.ing.iac.es/weave>

~1000 fibres (+mIFU and IFU)
over $\sim\pi$ deg²
at R up to 20,000
for $\lambda \sim 366-959$ nm

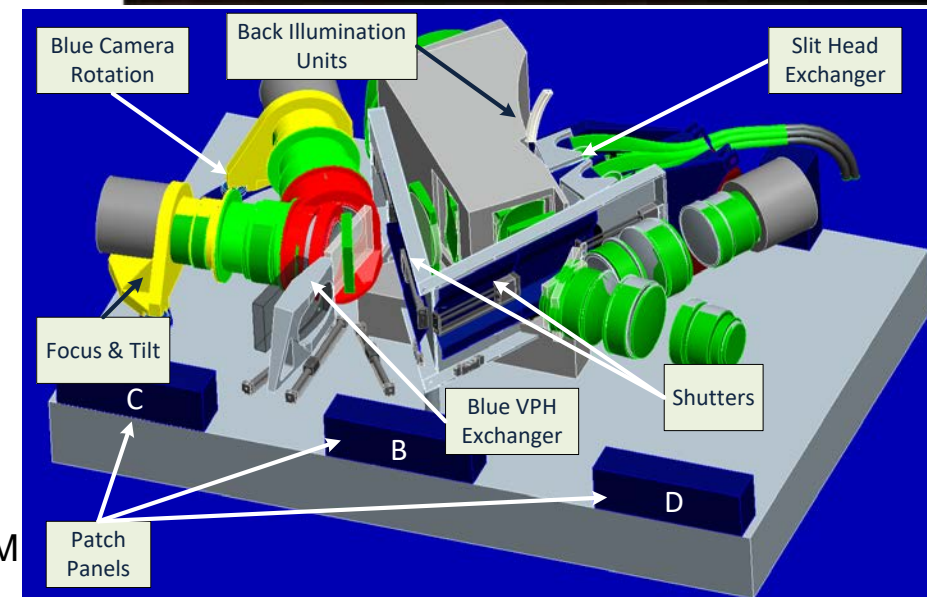
WEAVE 7 year surveys commence
early 2020



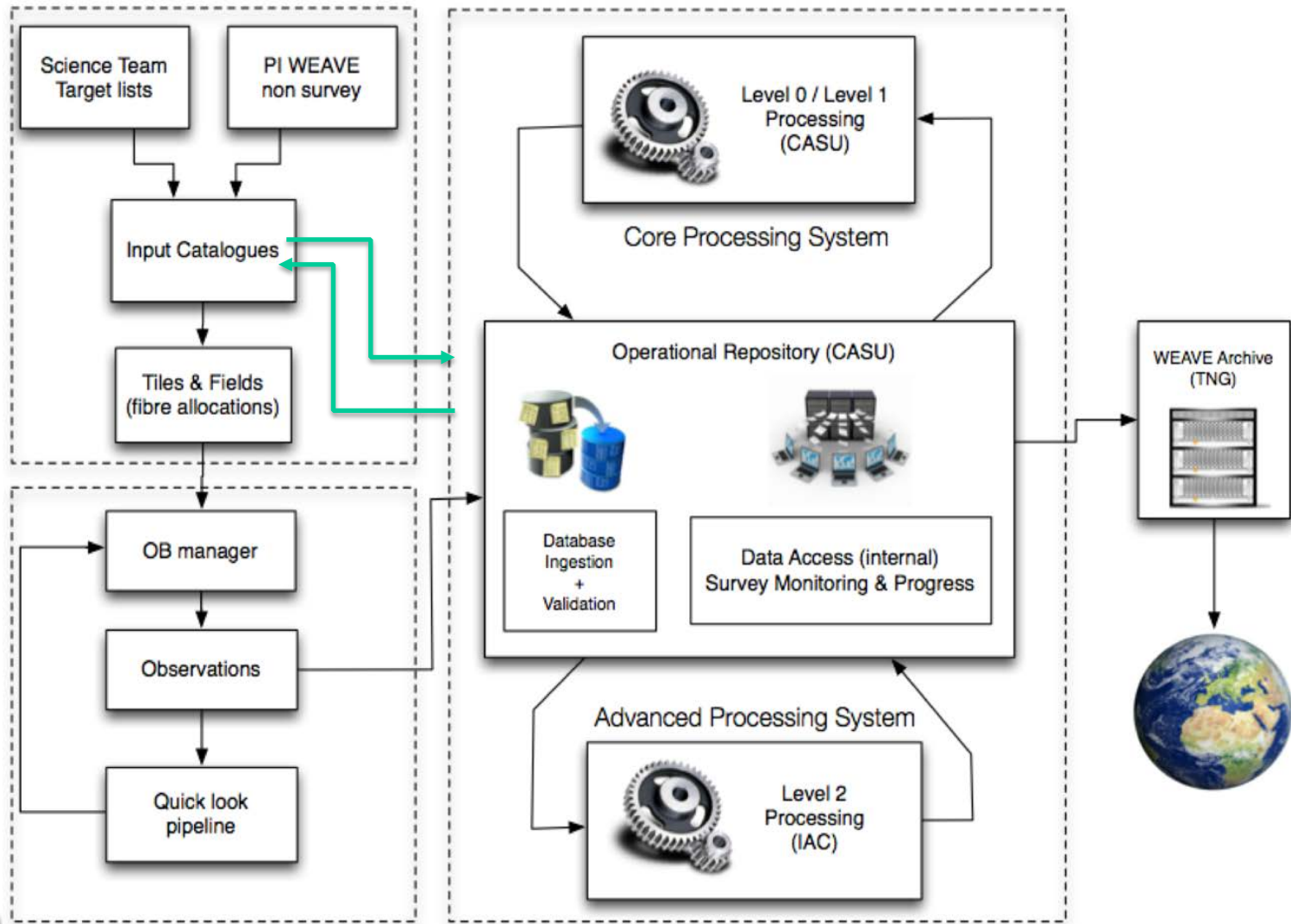
18 Nov 2020



Nic Walton - CASU@IRIS Science @ IRIS AHM



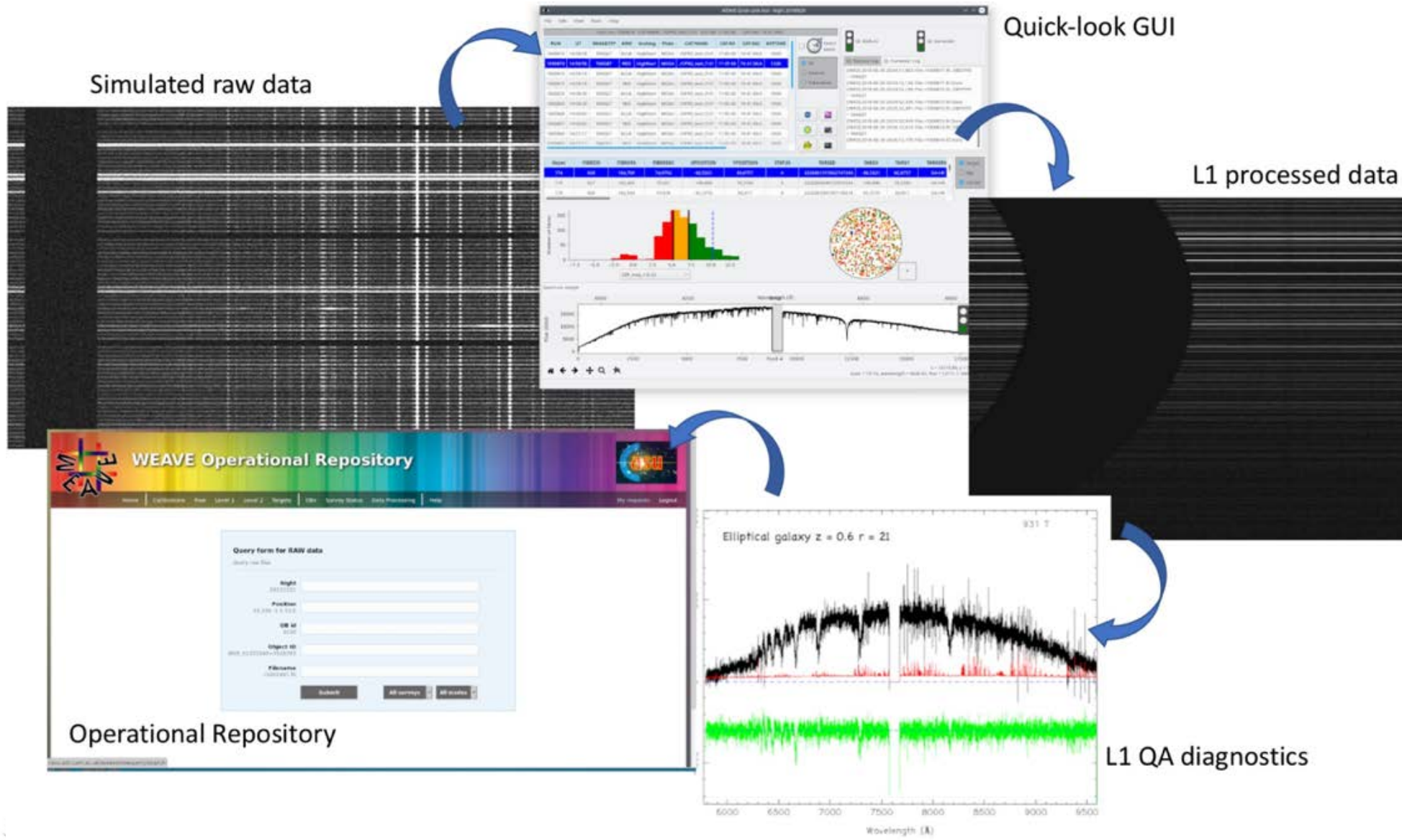
WEAVE Data Flow (simplified)



Goals: Galactic structure (Gaia complement), Galaxy evolution (LOFAR complement) loA lead development (and future operations) of the Core (QC/L0/L1/ Op Repository) processing system Surveys start 2021 with Operations phase 7 years



Spectroscopy: WEAVE data full system tests



Output from OpR3 allowed full assessment of:

- Survey scheduling
- Quality of L1 and L2 data
- WEAVE archive
- Overall system performance
- Operations both WEAVE and Science teams

CASU Archives and Operational Repositories

4OR: supporting internal releases

Query for RAW data

Query for OB progress monitor

Field: CCG_NGC6791_LR_F1W1

Observing Block filter options

OB id	Night	OB name	PROCTEMP	OBSTEMP	Made (binning)	Targets	SRV1	SRV2	SRV3	Author	Trimester	Status
1123	20160903	CCG_NGC6791_LR_F1W1	11331	F80E	LR MOS (1)	884	CCG			lara.kucak@impa.fr	54	Completed
1136	20160917	CCG_NGC6791_LR_F1W1	11332	CMC8	LR MOS (2)	838	CCG			alan.andrews@ulb.ac.be	54	Completed
1133	20160908	0773-053.0	11331	CMC8	LR MOS (1)	815	CA-IRDoc			arnoud.vanderschueren@ulb.ac.be	54	Completed

Gaia-ESO Survey Archive

Home Overview Observations Objects Processing Help

My requests Logout

Welcome to the Gaia-ESO archive at CASU

Gaia-ESO is a public spectroscopic survey, targeting ≥ 105 stars, sampling all major components of the Milky Way, from halo to star forming regions, providing the first homogeneous overview of the distributions of kinematics and elemental abundances. This alone will revolutionise knowledge of Galactic and stellar evolution. When combined with Gaia astrometry the survey will quantify the formation history and evolution of young, mature and ancient Galactic populations.

From this archive you will be able to search and download processed data. Data are in multi-extension FITS files which contain both images with spectral data and tables with meta-data and derived information about each object. Processing information will all be written to the relevant FITS header. The database allows users to receive data in two forms:

- A FITS file with several image extensions and some table extensions. Each row in the image represents a spectrum taken in a particular OB. This may be a single exposure or it might be a summed (stacked) spectrum. But it will be all the spectra for a given configuration.
- A FITS file with a single image extension and several table extensions. This will be the format for a spectrum of a single object. Again, this may be from a single exposure of that object or a summed spectrum. There will be several versions of the spectrum representing different levels in the reduction.

In order to download data you need a username and password. If you don't have one please contact the project PI.

Cambridge Astronomy Survey Unit – Institute of Astronomy

Quick links

- [Search for observations](#) – Query the database for observations and download FITS files containing all the objects observed in each configuration.
- [Search for objects](#) – Query the database for particular objects and retrieve their spectrum.
- [Processing status](#) – Follow the status of processing, link to log files and weather monitoring.
- [Help](#) – How to use the archive.
- [Spectra Data and Formats Document](#) – Description of available data, naming convention and formats.

My requests Logout

CASU provide access to data both internally to survey teams and externally to consortia. Database access via IRIS will provide enhanced resilience.

Gaia-ESO Survey Archive

Home Overview Observations Objects Processing Help

My requests Logout

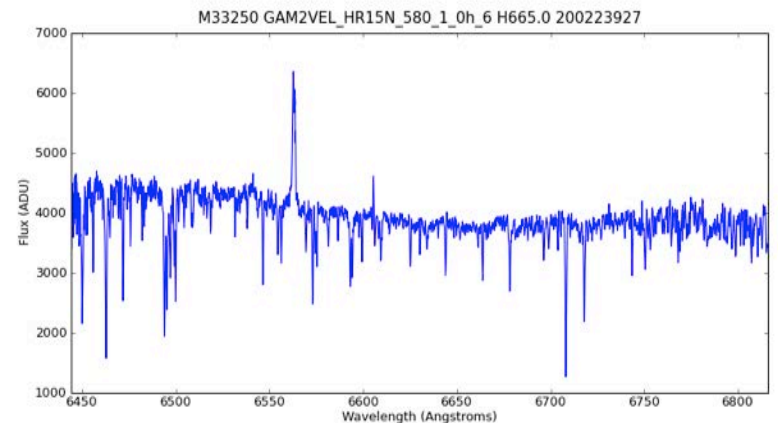
GES_MW_1528035-504234HR10

Coordinates: 15:28:03.5 -50:42:34
 File name: C20120312_00034_fn.fits
 ESO name: GIRAF_2012-03-13T08:46:21.579.fits
 OB ID: 200226309
 Date of observation: 2012-03-13T08:46:21.579
 Exposure Time: 1800.0034
 Airmass: 1.12
 Seeing: 1.32

Grating: H548.8
 Filter name: HR10
 Grating resolution: 40000.0
 Grating order: 10
 Allocated objects: 118 (1)
 Allocated sky: 21 (0)
 Unallocated objects: 1
 Last modified: 2012-09-04 08:49:42

Targets

Name	Coordinates	Magnitude	Spectrum
6012201	15:27:02.31 -50:38:39.4	16.92	Spectrum
6000182	15:27:04.62 -50:49:58.3	16.61	Spectrum
6015115	15:27:09.21 -50:36:07.3	16.87	Spectrum
6007202	15:27:14.85 -50:42:56.6	16.89	Spectrum
6007485	15:27:18.47 -50:42:40.3	16.60	Spectrum
6001370	15:27:21.43 -50:48:34.6	16.92	Spectrum
6010092	15:27:21.99 -50:40:26.0	16.88	Spectrum
6014454	15:27:27.63 -50:36:44.1	16.74	Spectrum



Coordinates: 08:08:52.14 -47:22:57.9
 Image name: C20120101_00015_st

Cambridge Astronomy Survey Unit – Institute of Astronomy

A

18 Nov 2020

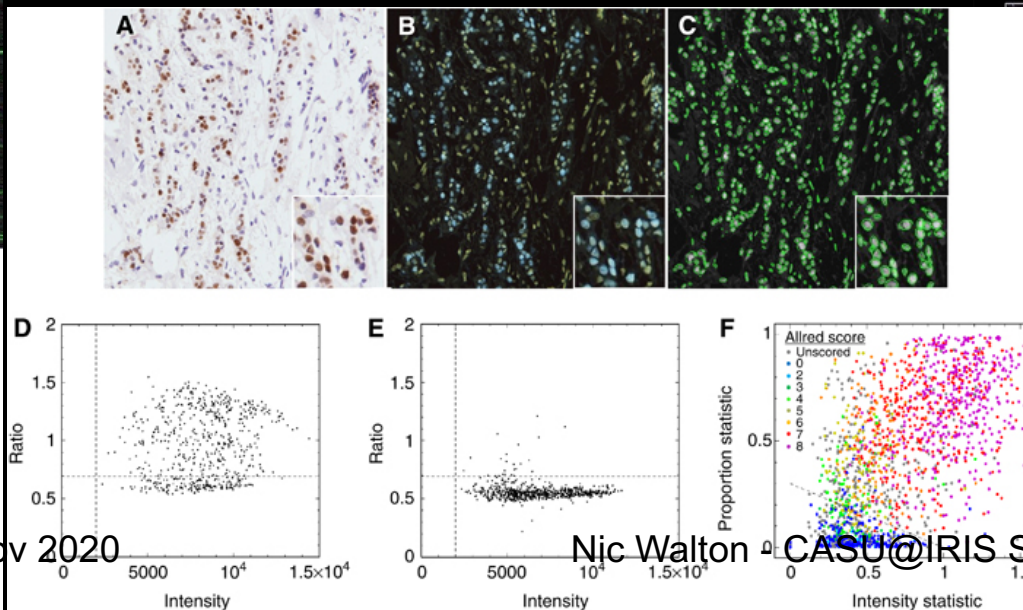
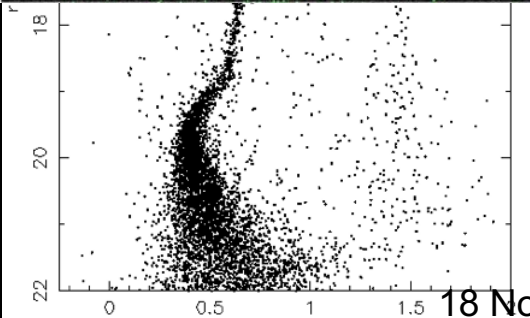
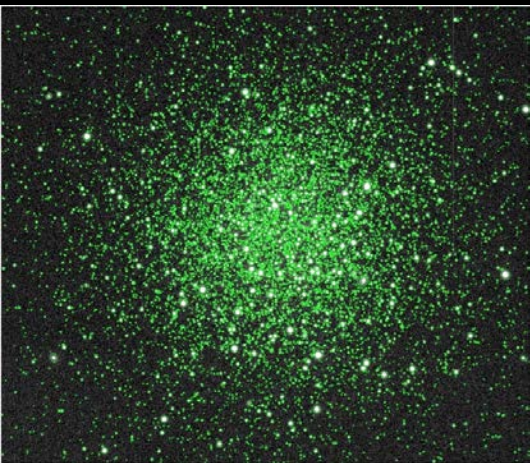


Applying 'astro' techniques to the medical domain

Image: ESO



astronomical image analysis and data analysis developed at CASU, IoA, Cambridge applied to medical imaging data



IMAXT Data Analysis and Infrastructure

Cancer Research UK Grand Challenge

Interactive data visualisation

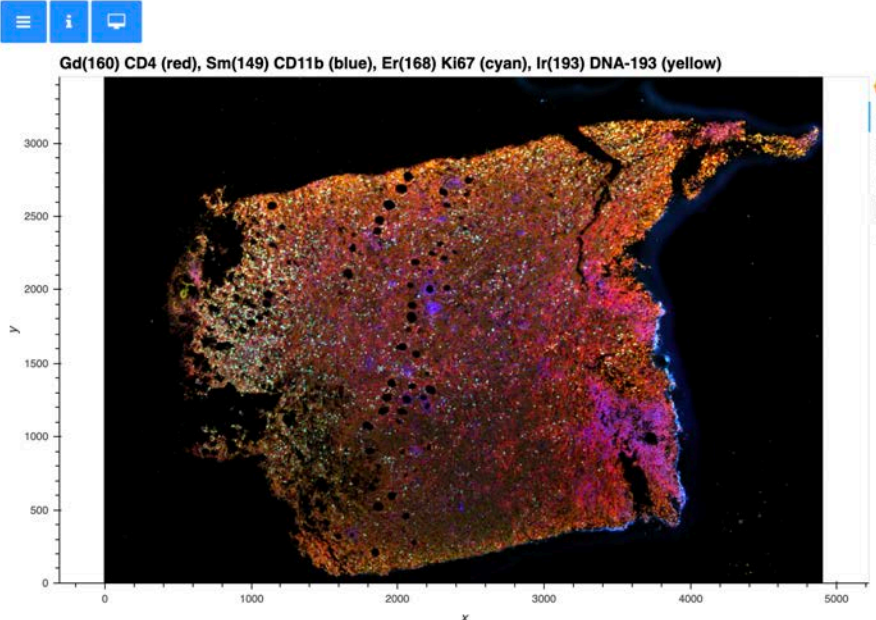
CREATING A VIRTUAL REALITY TUMOUR

- 1 A detailed reference picture of a tumour is taken.
- 2 Wafer thin slices are cut from the tumour.
- 3 The slices are deeply analysed, right down to their genetic information.
- 4 The information is processed and the tumour is rebuilt in virtual reality.

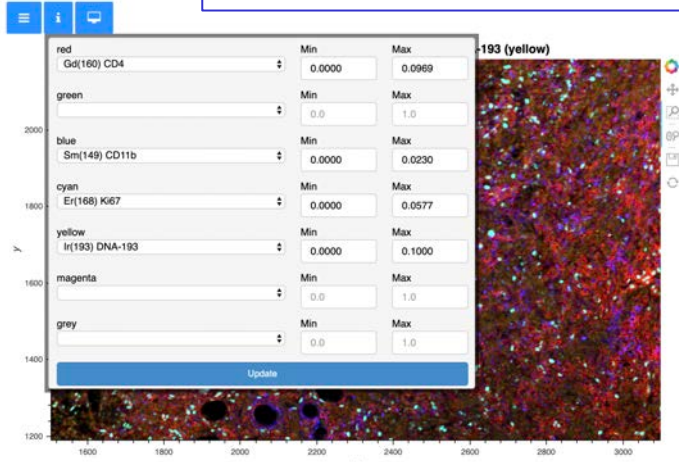


LET'S BEAT CANCER SOONER
cruk.org

CANCER RESEARCH UK
GRAND CHALLENGE

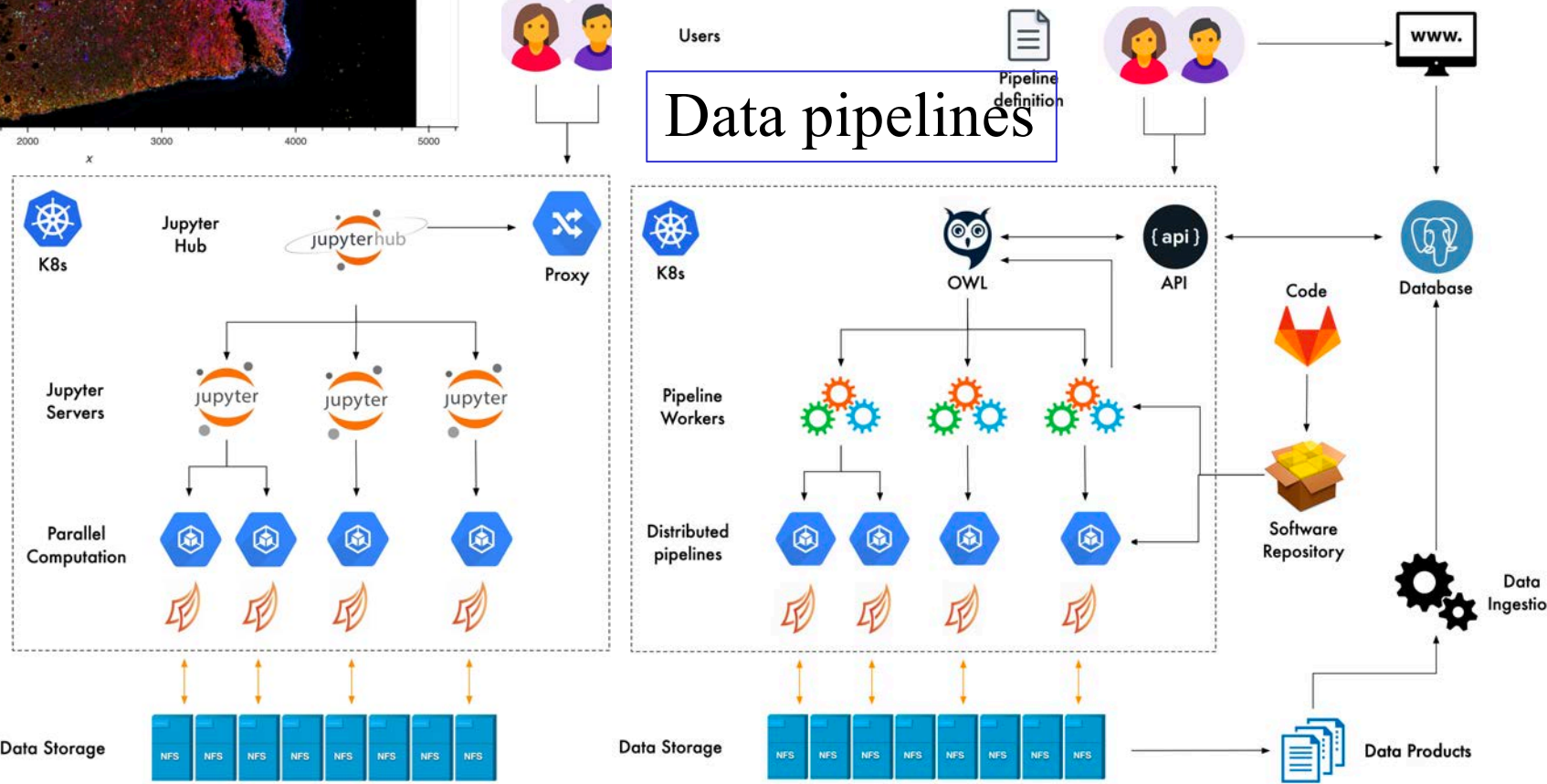


Deployment could be made on Openstack@IRIS



Interactive data access

Data pipelines



CASU responsible for primary image analysis and data infrastructure – transfer of CASU technology from astro to medical domain

JupyterHub Deployment: CASU@IRIS

The screenshot shows the CASU Interactive Analysis Platform Launcher. The interface has a teal header with the CASU logo and 'ioa' branding. Below the header is a menu bar with 'File', 'Edit', 'View', 'Run', 'Kernel', 'Git', 'Snippets', 'Tabs', 'Settings', 'Help', and 'CASU'. The version is 'imaxt/casu-notebook:2020.07.21'. The main area is titled 'Launcher' and features a file browser on the left showing a directory structure with 'cutouts.ipynb' and 'cutouts.py'. The central area is divided into three sections: 'Notebook', 'Console', and 'Other'. Each section contains icons for various languages and tools: Python 3, Go, Javascript (Node.js), Octave, Python 3.8, R, Terminal, Diagram, Python File, Text File, Markdown File, and Tensorboard. A 'Show Contextual Help' button is at the bottom. The status bar at the bottom indicates 'Saving completed' and 'Launcher'.

The screenshot shows the CASU Interactive Analysis Platform with a Jupyter Notebook open. The notebook is titled 'cutouts.ipynb' and contains the following code:

```
[7]: fut = []
for item in res.head(n=12).itertuples():
    fut.append(delayed(extract_cutout)(item, ra, dec, size))

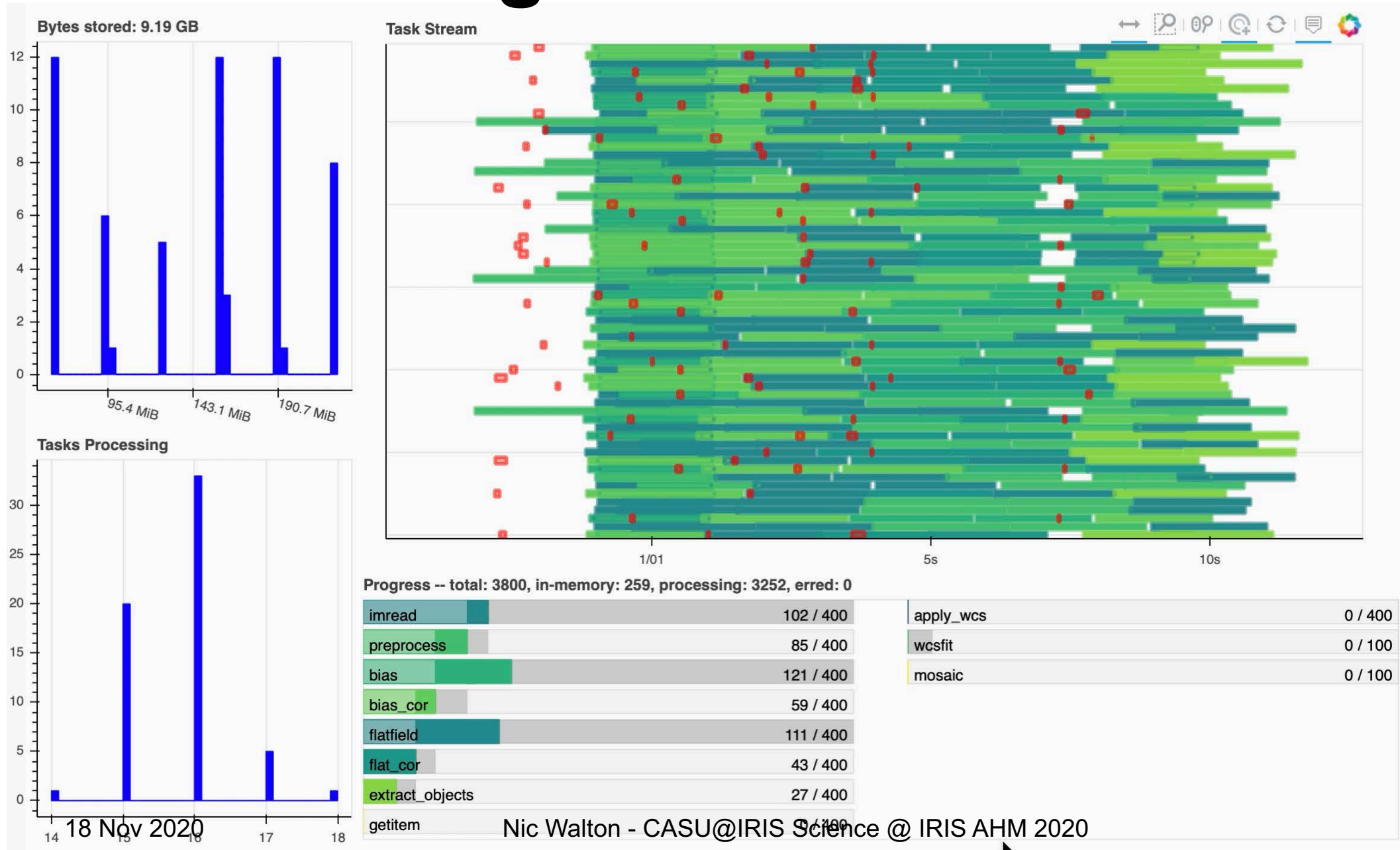
images = dask.compute(*fut)

fig, ax = plt.subplots(nrows=3, ncols=4, figsize=(16,12))
ax = ax.ravel()

for j, img in enumerate(images):
    cutout_imshow(ax[j], img)
```

The notebook output displays a 3x4 grid of 12 galaxy cutouts, each showing a different view of a galaxy with dark spots and a bright central region. The status bar at the bottom indicates 'Saving completed', 'Mode: Edit', and 'Ln 10, Col 33 cutouts.ipynb'.

Processing 100 images in distributed containers using 60 cores



Query VISTA database around position

```

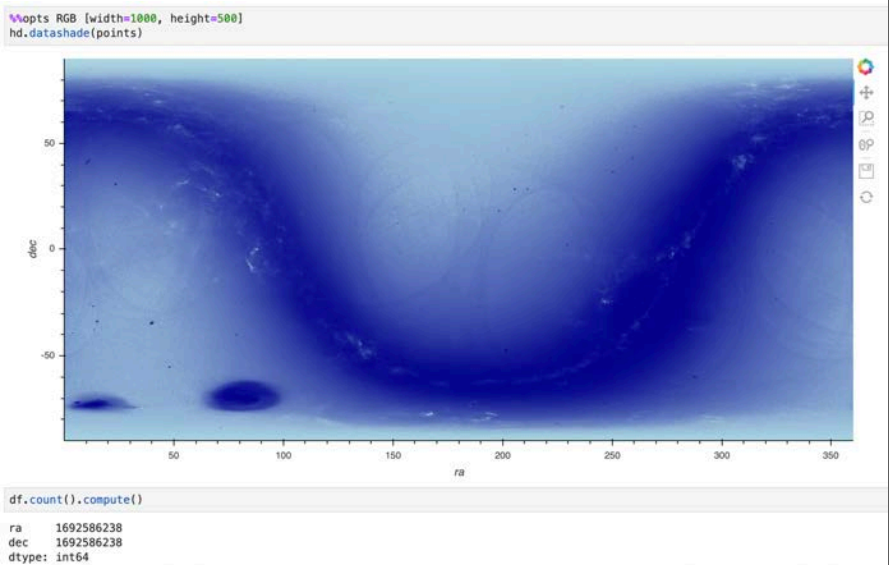
vistadb = VISTADB()

# Coordinates to search for
ra, dec = 194.30, -64.75

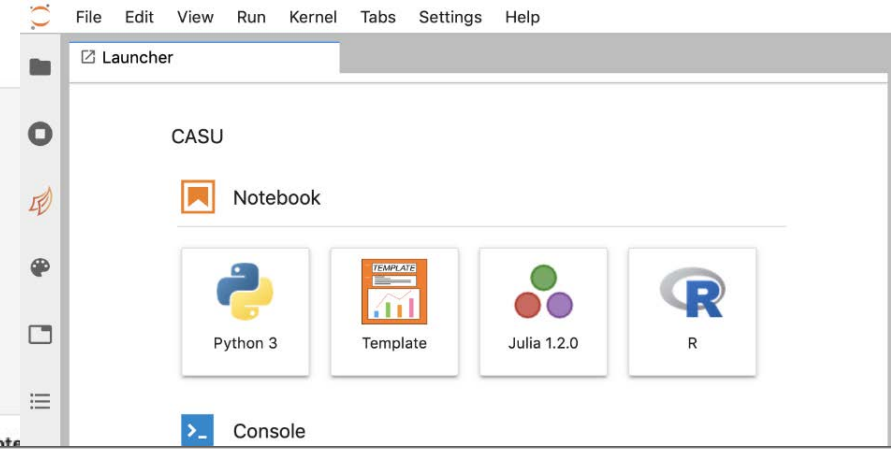
# Columns to print
columns = ['filename', 'coords', 'filtername', 'surveyname', 'nightobs',
          'totexptime', 'obsfwhm', 'obstatus', 'qcstatus']

# Execute query and display first 10 results
res = vistadb.query_radec(ra, dec)
res.columns.head(n=10)
    
```

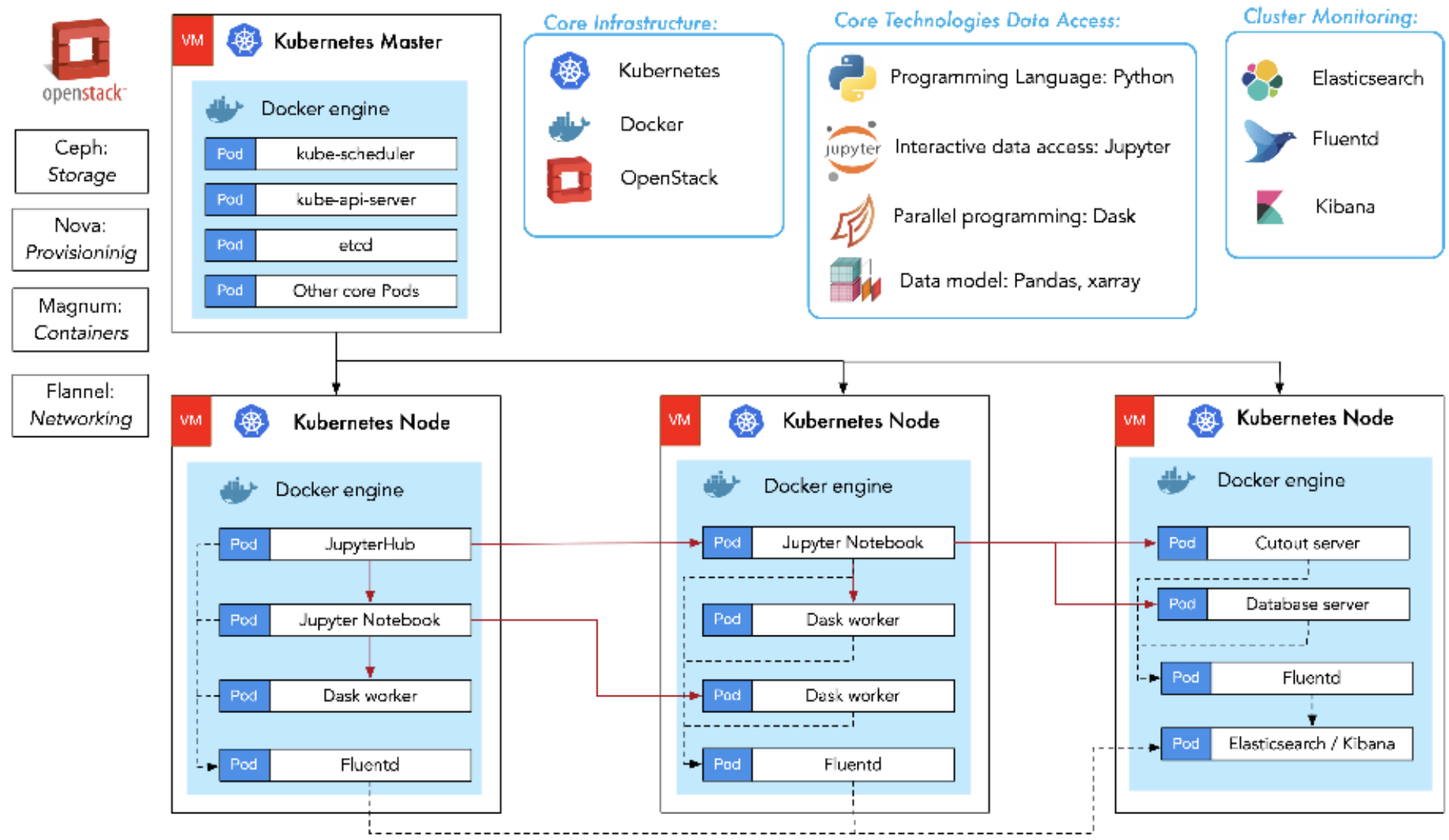
id	filename	coords	filtername	surveyname	nightobs	totexptime
62124	v20100218_00330_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201002	
68702	v20100228_00495_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201002	
70725	v20100304_00444_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201003	
81752	v20100315_00327_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201003	
83062	v20100316_00361_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201003	
83235	v20100316_00499_st_tl.fit	13:02:18.02 -64:35:28.7	H	VVV	201003	
83254	v20100316_00511_st_tl.fit	13:02:18.22 -64:35:29.9	Ks	VVV	201003	
83273	v20100316_00523_st_tl.fit	13:02:18.22 -64:35:29.9	J	VVV	201003	
120735	v20100422_00425_st_tl.fit	13:02:18.02 -64:35:28.7	Y	VVV	201004	
120754	v20100422_00437_st_tl.fit	13:02:18.22 -64:35:29.9	Z	VVV	201004	



18 Nov 2020



Architecture diagram for the IRIS-CASU deployment. OpenStack provides the infrastructure as a set of virtual machines (VM) where a Kubernetes (k8s) cluster is installed.



PLATO (ESA M3)

<http://sci.esa.int/plato>

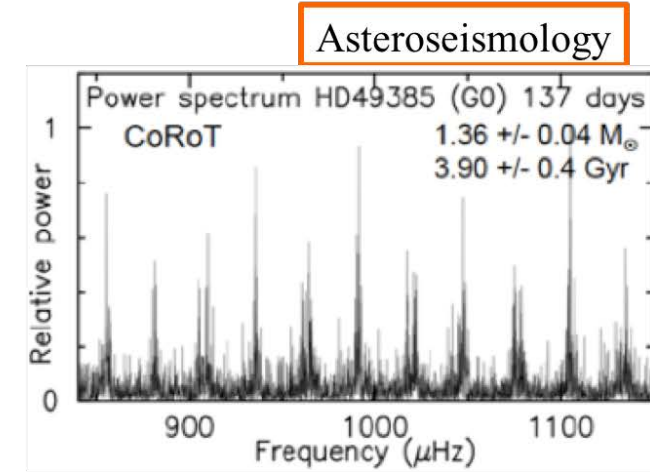
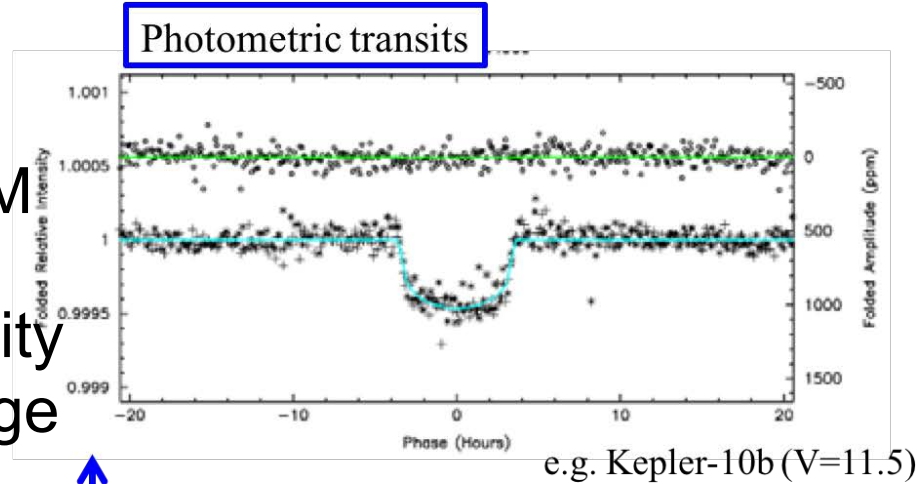


Goals: will detect and characterize planets down to Earth-size by high precision photometric transits around ~1M bright stars.

Launch end 2026. 4 to 8 years operations plus post operations

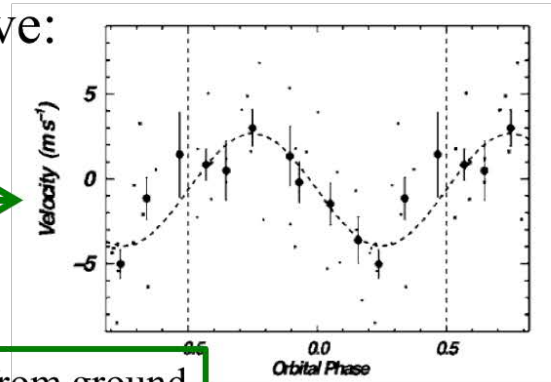
PLAnetary Transits and Oscillation of stars

Goals: will detect and characterize planets down to Earth-size by high precision photometric transits around ~1M bright stars. Planetary masses will be determined by ground-based radial velocity measurements. Stellar parameters like age and mass will be obtained by asteroseismology

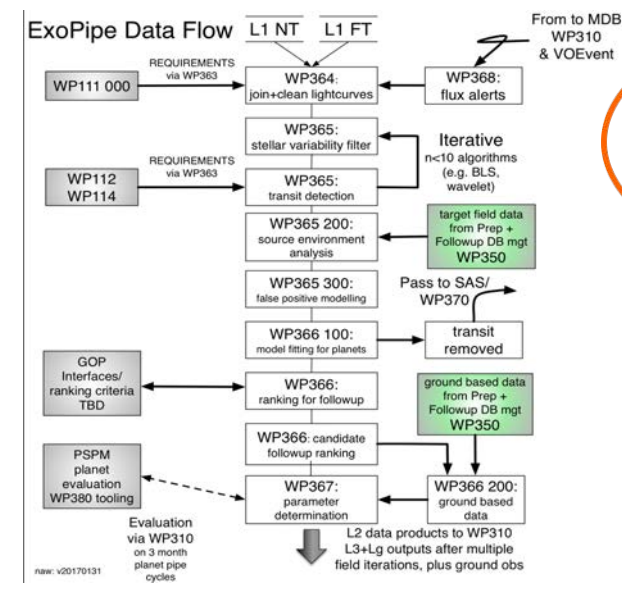


For solar like stars PLATO will give:

- Radius to ~3%
- Mass to ~10%
- Age to ~10%



What? IoA development and operations of Exoplanet Analysis System pipelines.



When? Launch end 2026. 4 to 8 years operations plus post operations



Gaia and DPAC @ Cambridge

Credit: ESA Gaia/ DPAC



Mapping the Milky Way: 2 Billion Stars
Gaia Data Release 2 (Apr 2018):

<https://www.cosmos.esa.int/web/gaia/dr2>

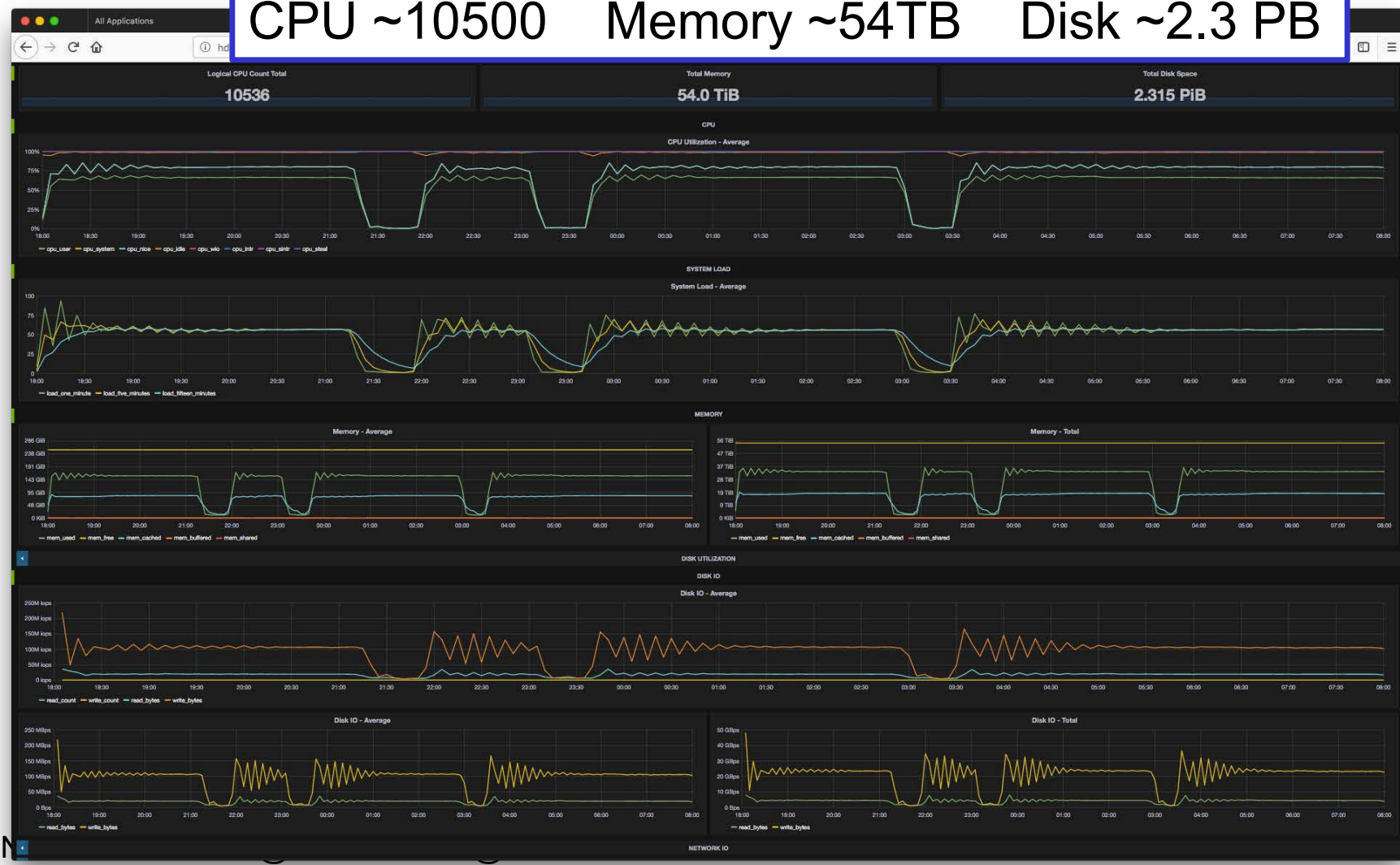
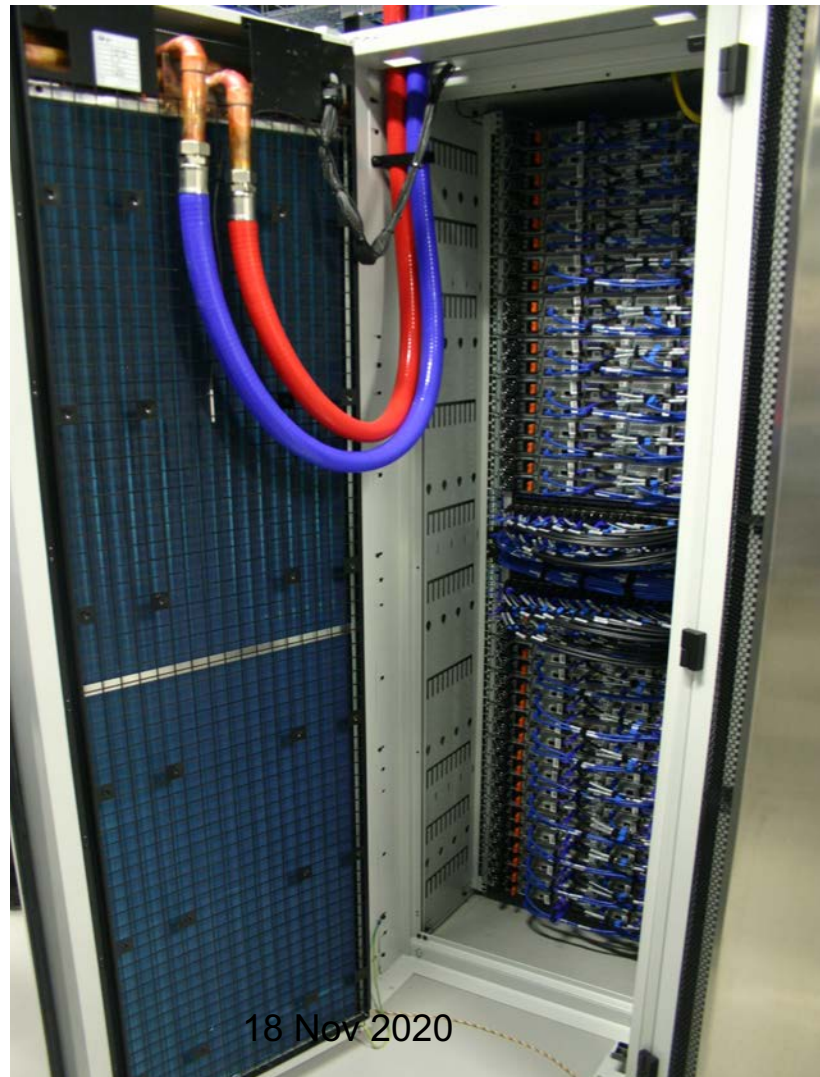
Gaia Early Data Release 3 (3rd Dec 2020)

Cambridge DPCI Cluster @ West Cambridge Data Centre

220 nodes provide combined
compute+storage



CPU ~10500 Memory ~54TB Disk ~2.3 PB

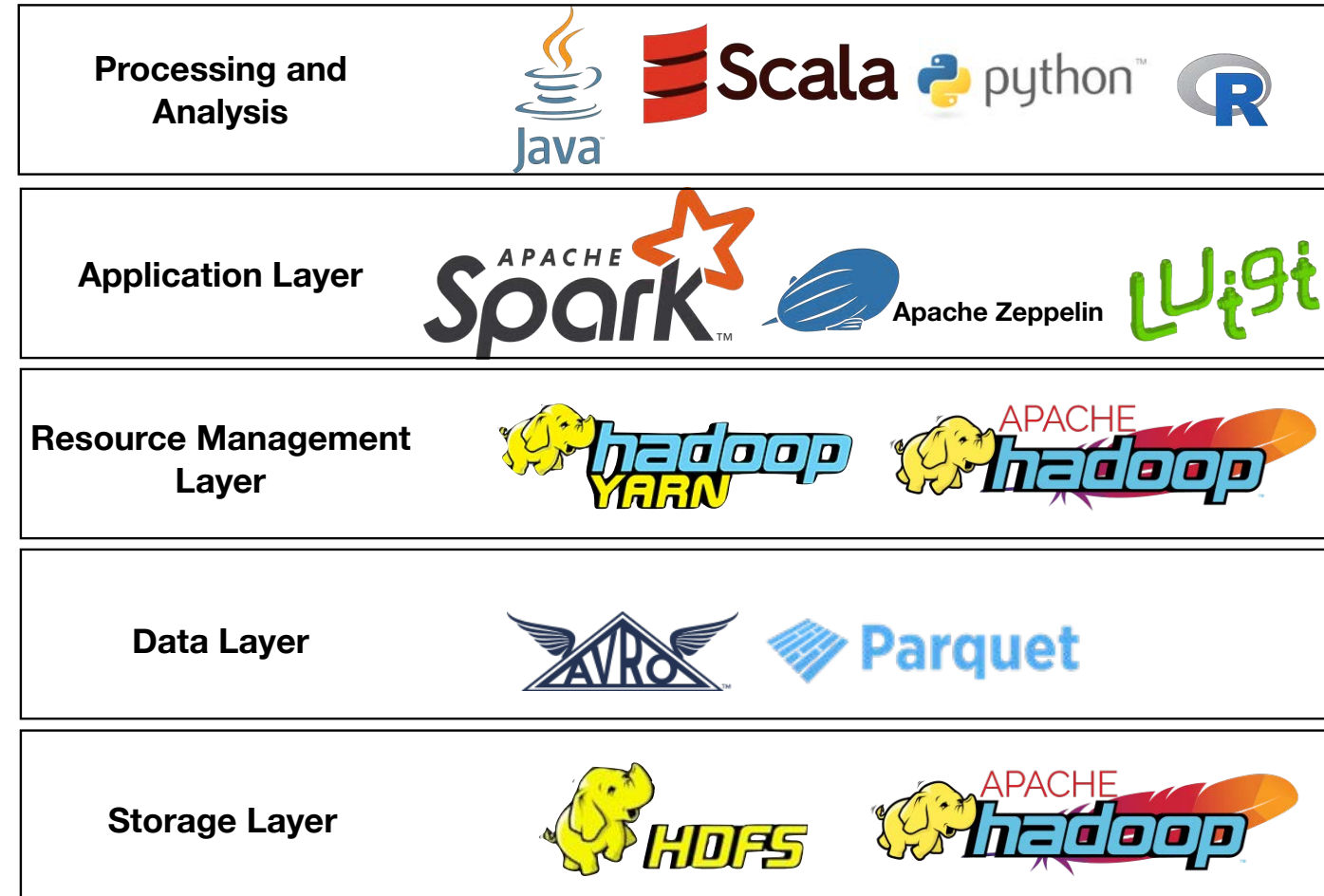


18 Nov 2020



Cambridge DPCI Software Stack

- The processing stack has transitioned from MapReduce to Spark
- Provides flexibility as Spark processing can be deployed on non Hadoop Cluster architectures
- Opens the possibility to deploy on IRIS type infrastructures



Will investigate use of IRIS for future Gaia Core processing



CASU actively participating in the IRIS initiative

- Current activity involves deployment of data analysis and database access to VISTA imaging data
- Expand capability to provide user access to analysis chains linked to all CASU science data products
 - Improved science consortium access to internal data releases
- Investigate deployment of core Gaia processing to IRIS in the 2023/24 timeframe (assumes IRIS or similar provision longer term and UKSA/STFC agreement)
 - Experience informs deployment of PLATO processing centre post 2025