CASU@IRIS: enabling wide area survey science

Nicholas Walton et al
Institute of Astronomy,
University of Cambridge

Carina: Credit: ESO/ VPHAS/ CASU

Areas of the LMC: Credit: ESO

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

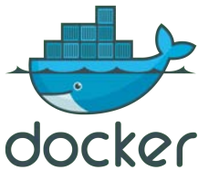
CASU: Data for Data Intensive Science

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 (K8s) allows for deployment, scaling and management of containers in a cluster

CASU

- OpenStack manages large pools of compute resources

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

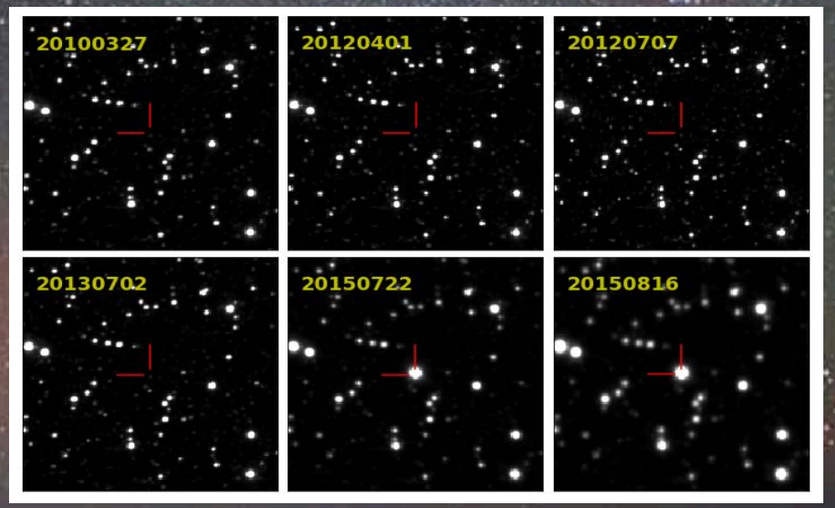
Monthly delivery of data products to Public Survey PIs

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

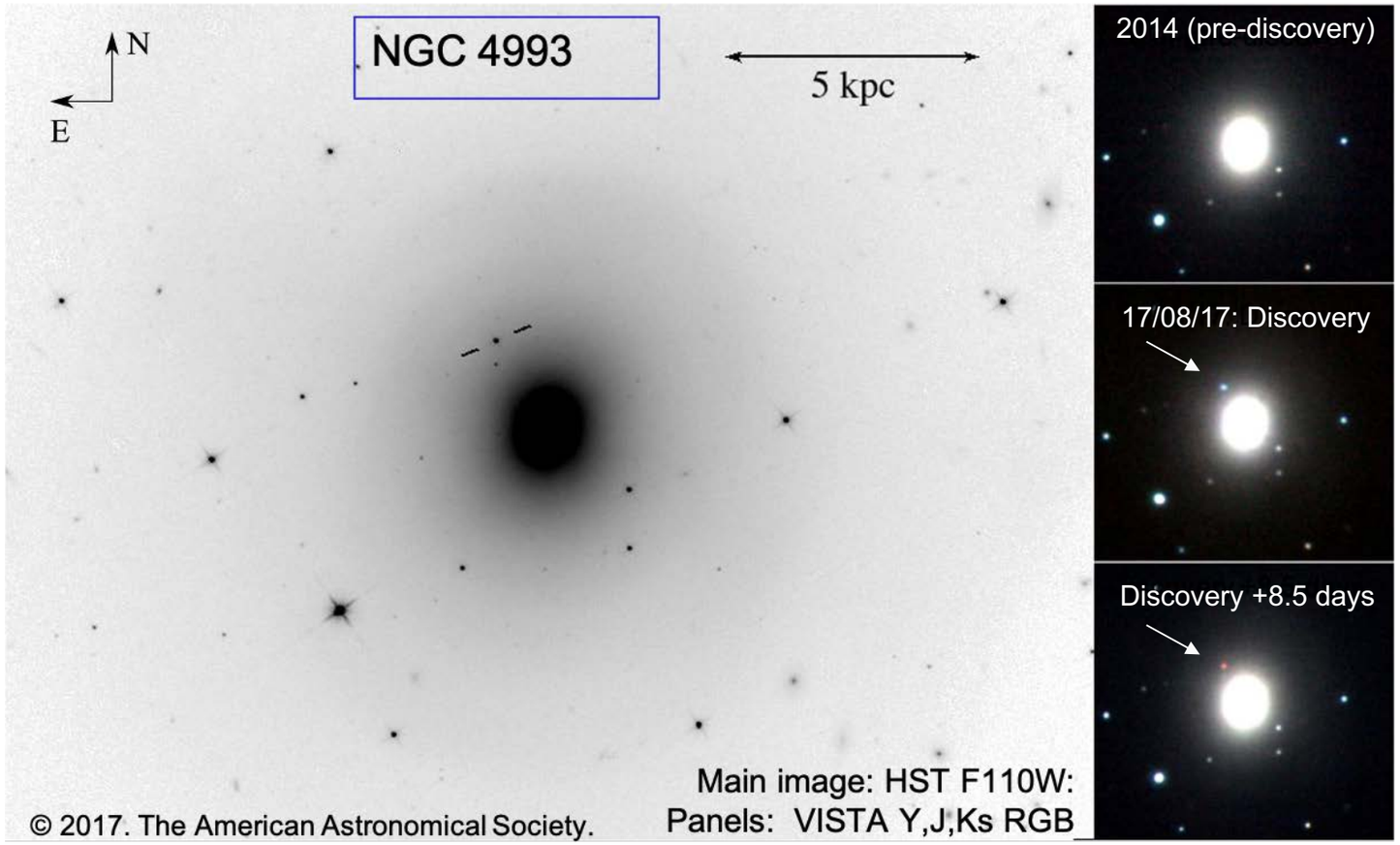
VISTA + VIRCAM



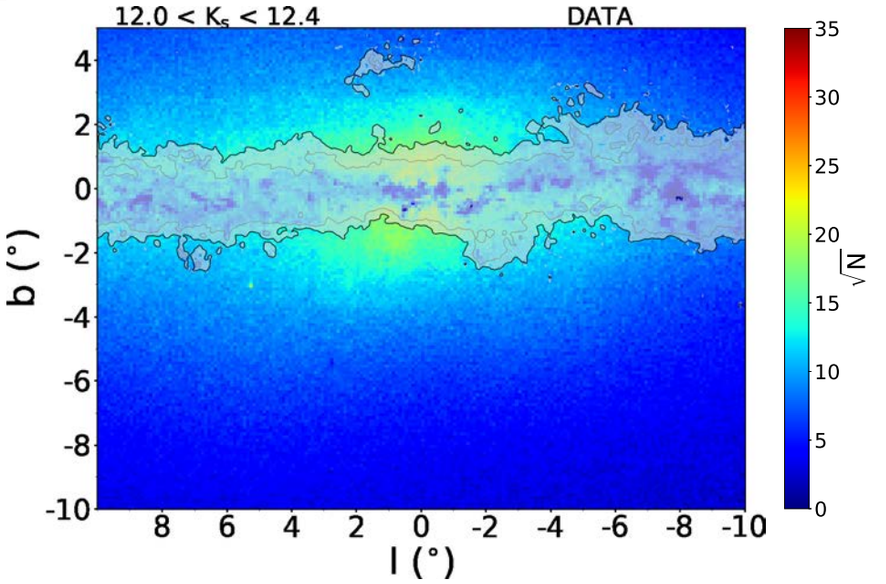
VISTA Alerts: rapid processing

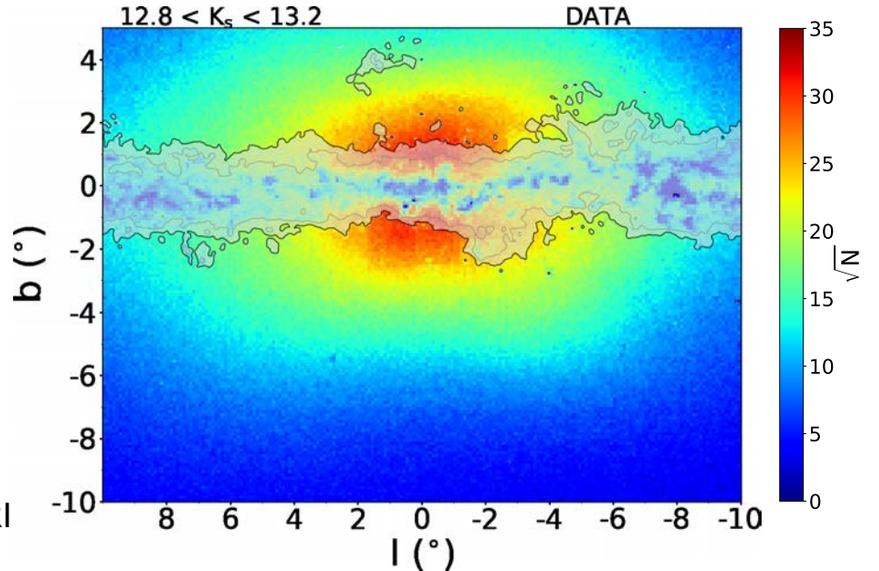
Imaging Surveys: VISTA

IoA developed and operate pipeline processing for VISTA IR surveys



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





Simion et al, 2017

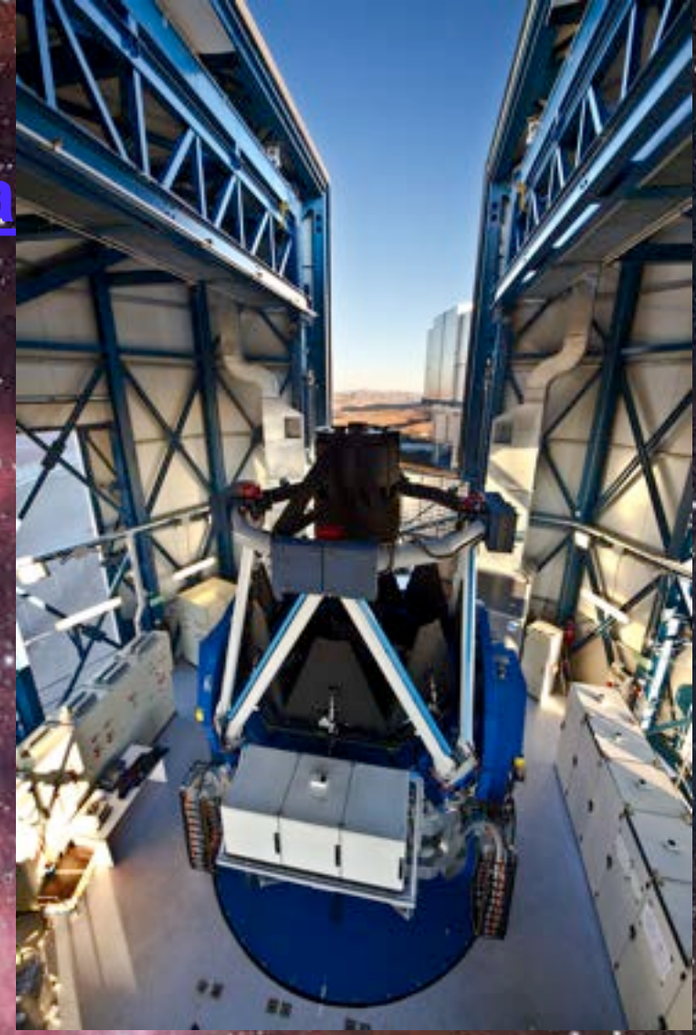
Imaging Surveys: VST

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

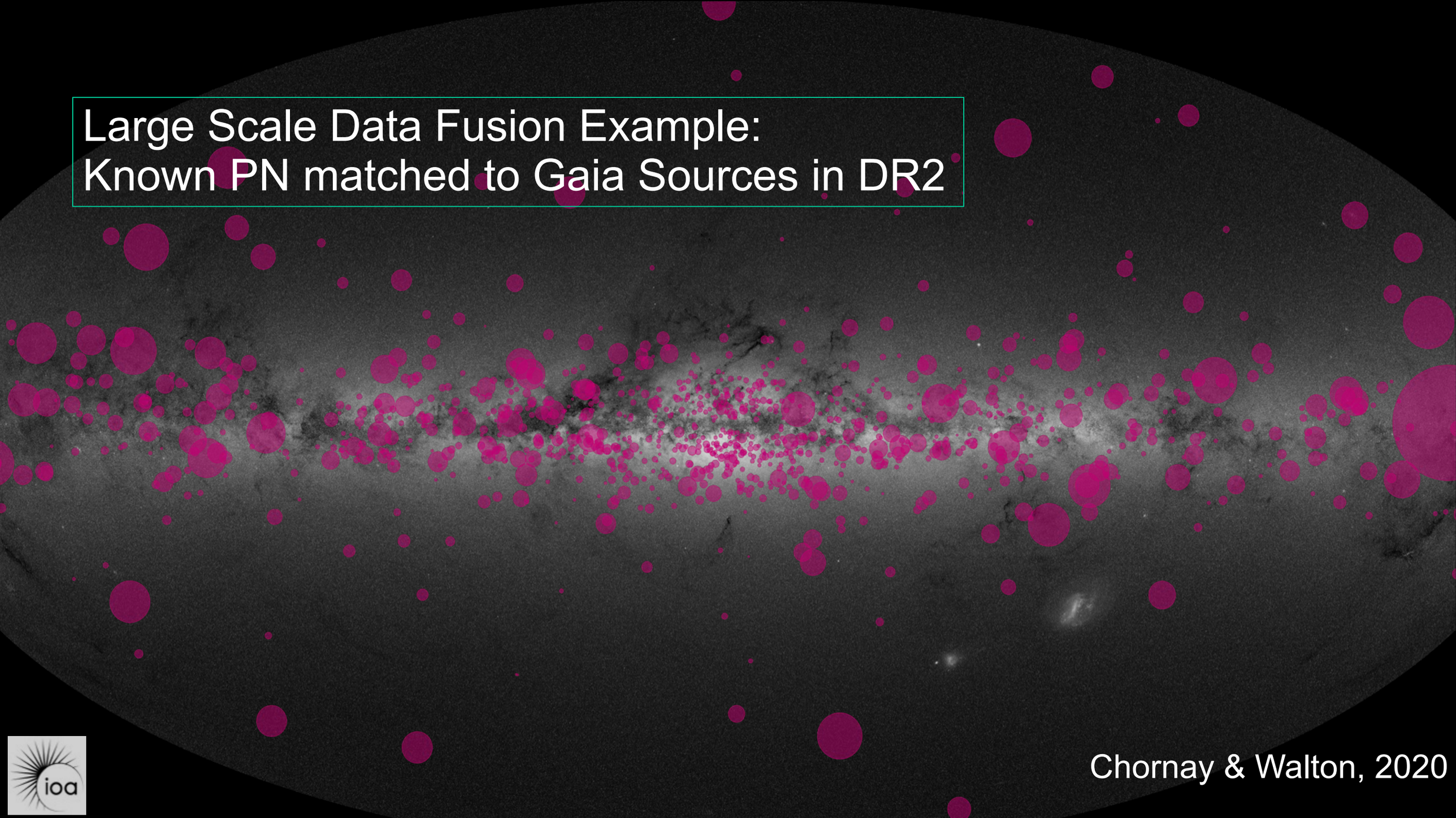
Why? Range of ESO public surveys, e.g. ATLAS, VPHAS, etc

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

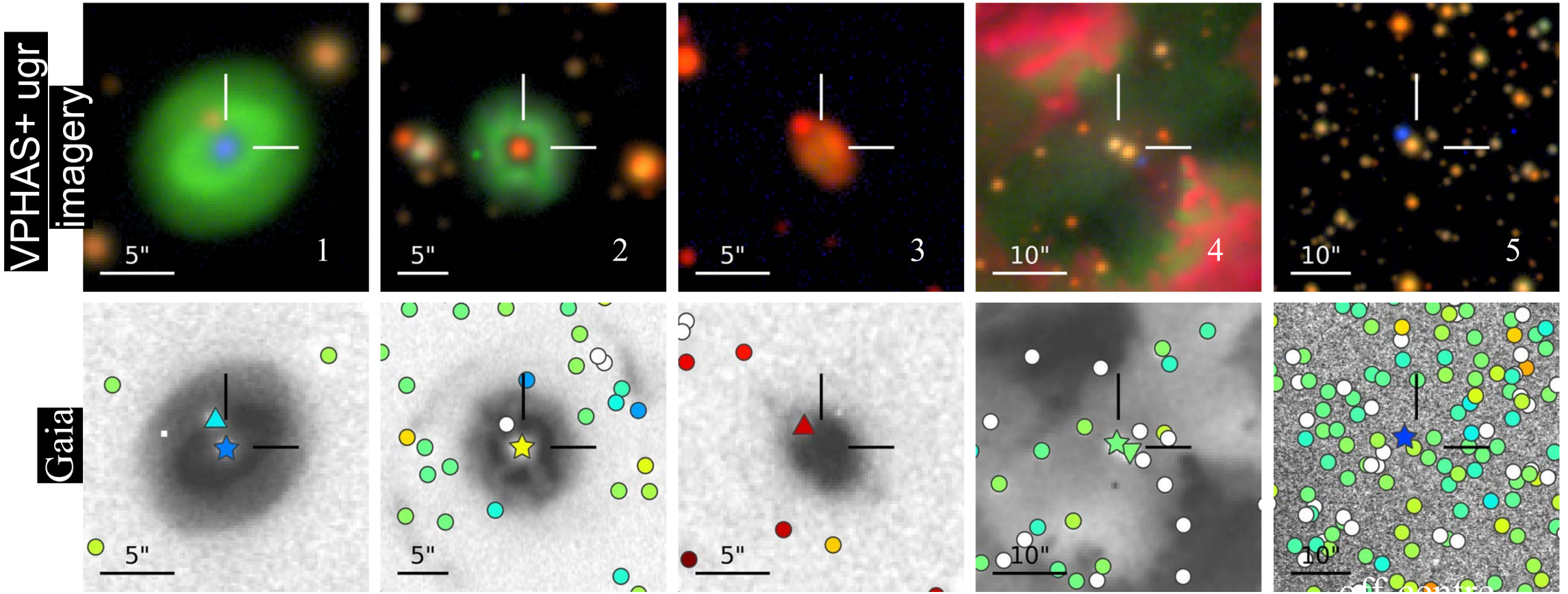
When? Surveys ongoing, 5+ years



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



Chornay & Walton, 2020



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

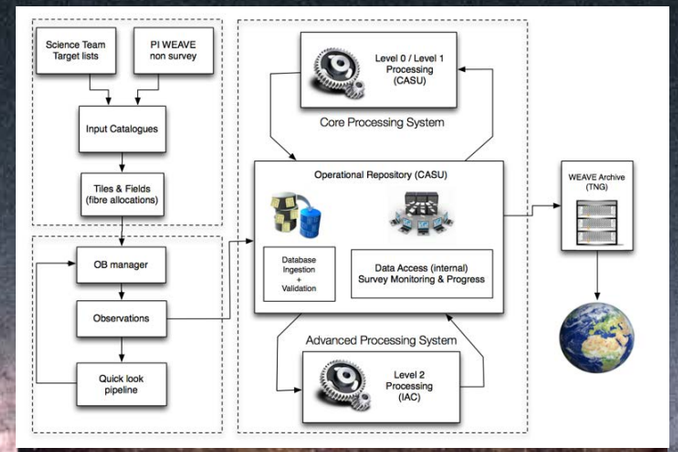
Spectroscopic Pipelines

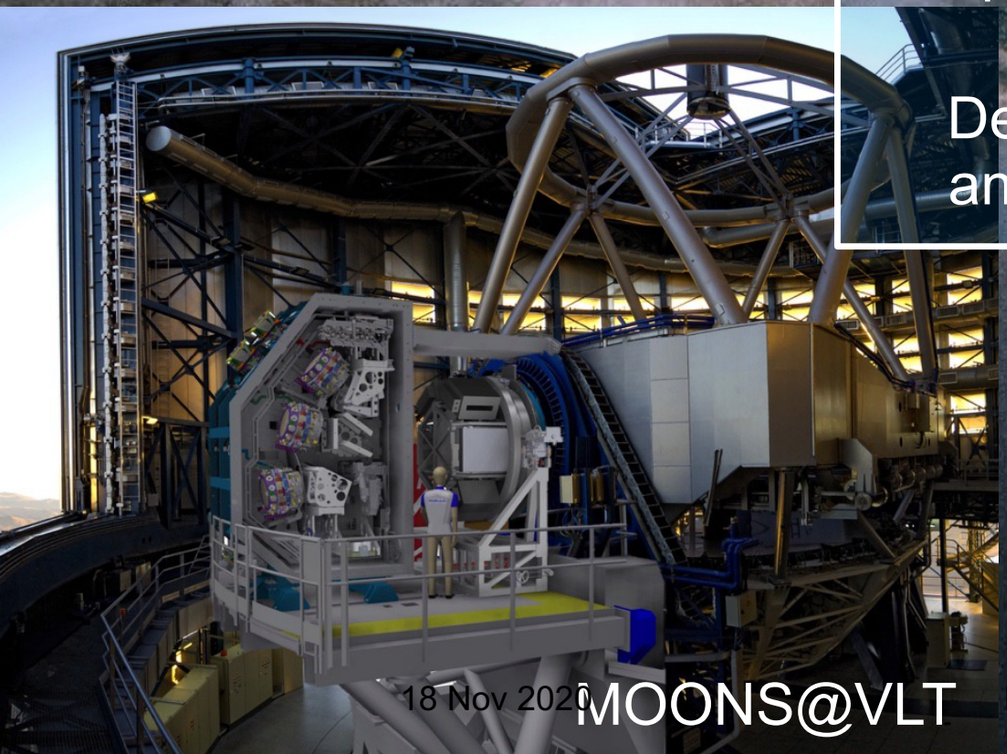
Design of complete data system architecture

Development of QC pipeline

Development of Operational Repository and spectral extraction pipelines

Design of interfaces to spectral analysis pipelines and Archive





MOONS@VLT

4MOST@VISTA

WEAVE@WHT

4MOST

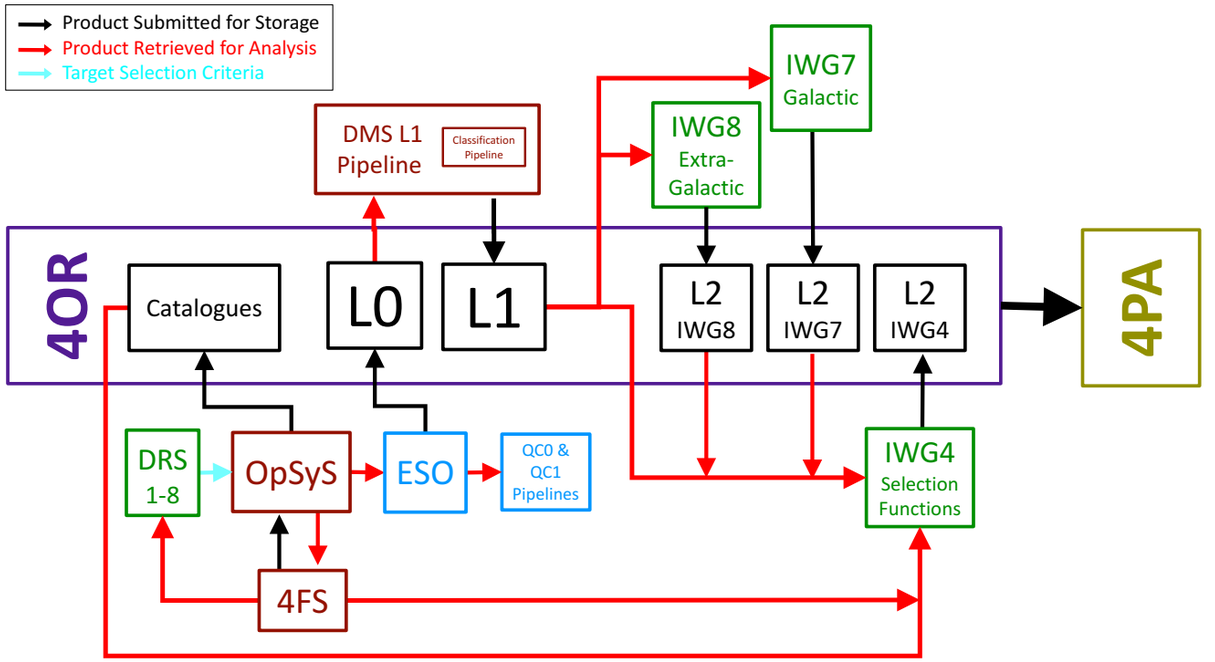
https://www.4most.eu

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

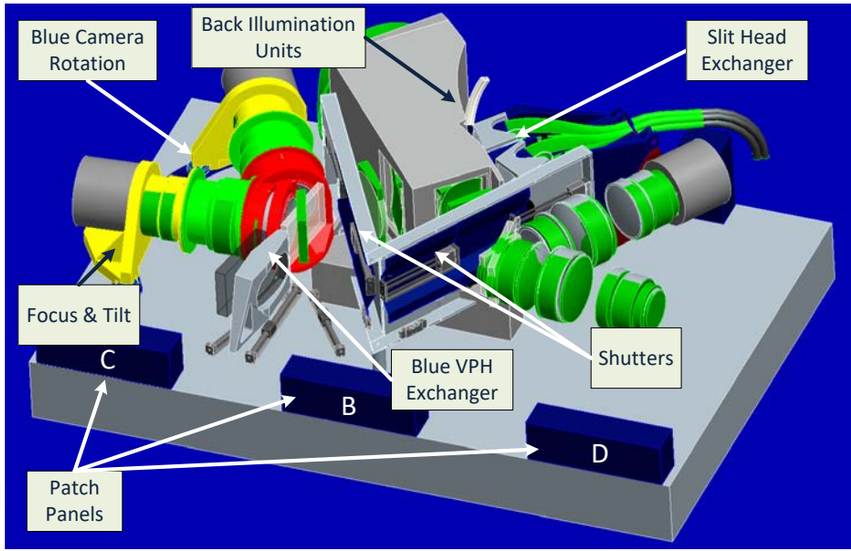
Science case	S/N / Å	r_{AB}-mags	Targets (Millions)
S1 Milky Way Halo LR Survey	10	16–20.0	1.5
S2 Milky Way Halo HR Survey	140	12–15.5	0.08
S3 Milky Way Disk and Bulge LR Survey	10–30	14–18.5	10.7
S4 Milky Way Disk and Bulge HR Survey	140	14–15.5	1.8
S5 Galaxy Clusters Survey	4	18–22.0	1.1
S6 AGN Survey	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



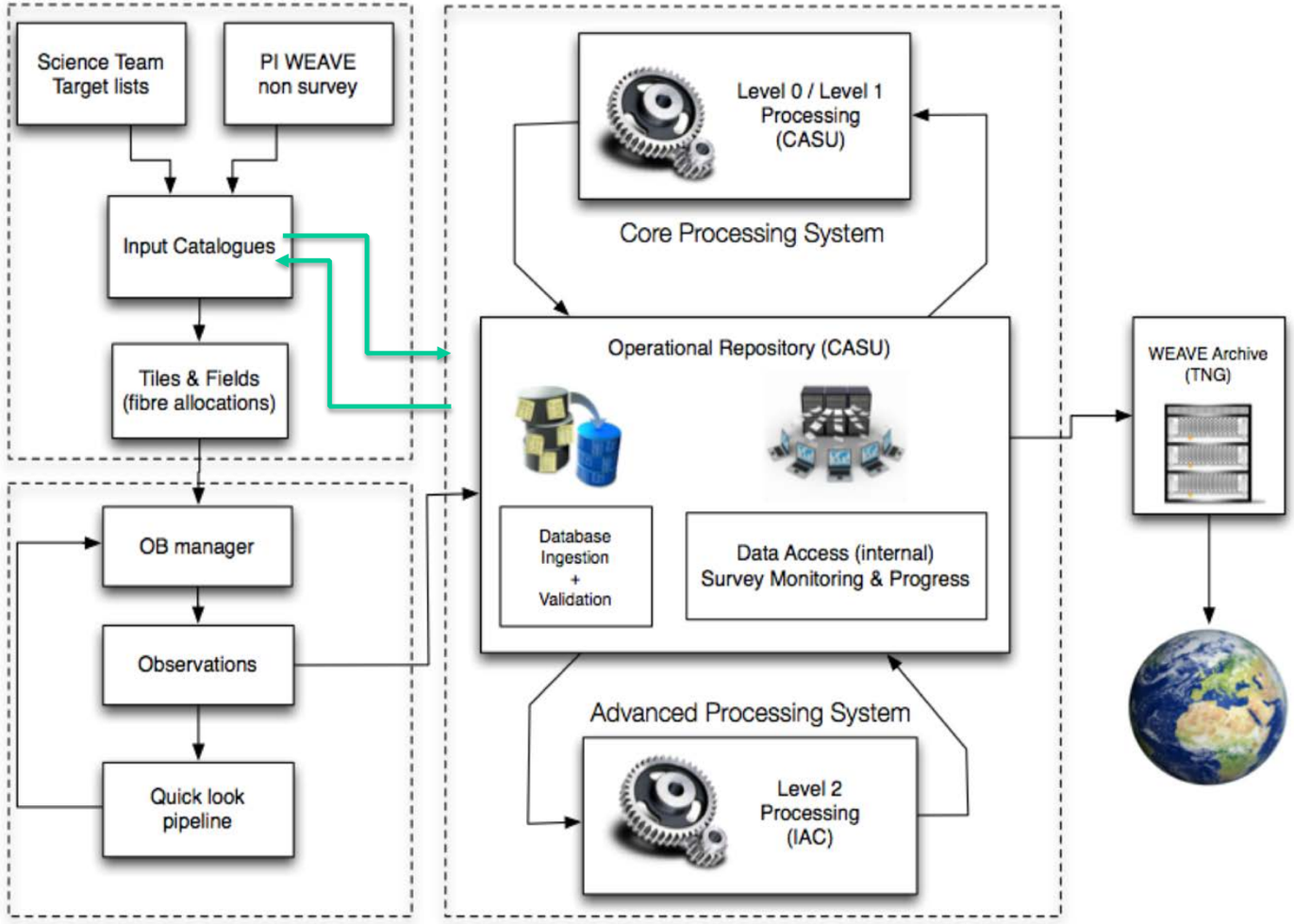
WEAVE:
http://www.ing.iac.es/weave

~1000 fibres (+mIFU and IFU)
over ~π deg^2
at R up to 20,000
for λ ~366-959nm

WEAVE 7 year surveys commence early 2020

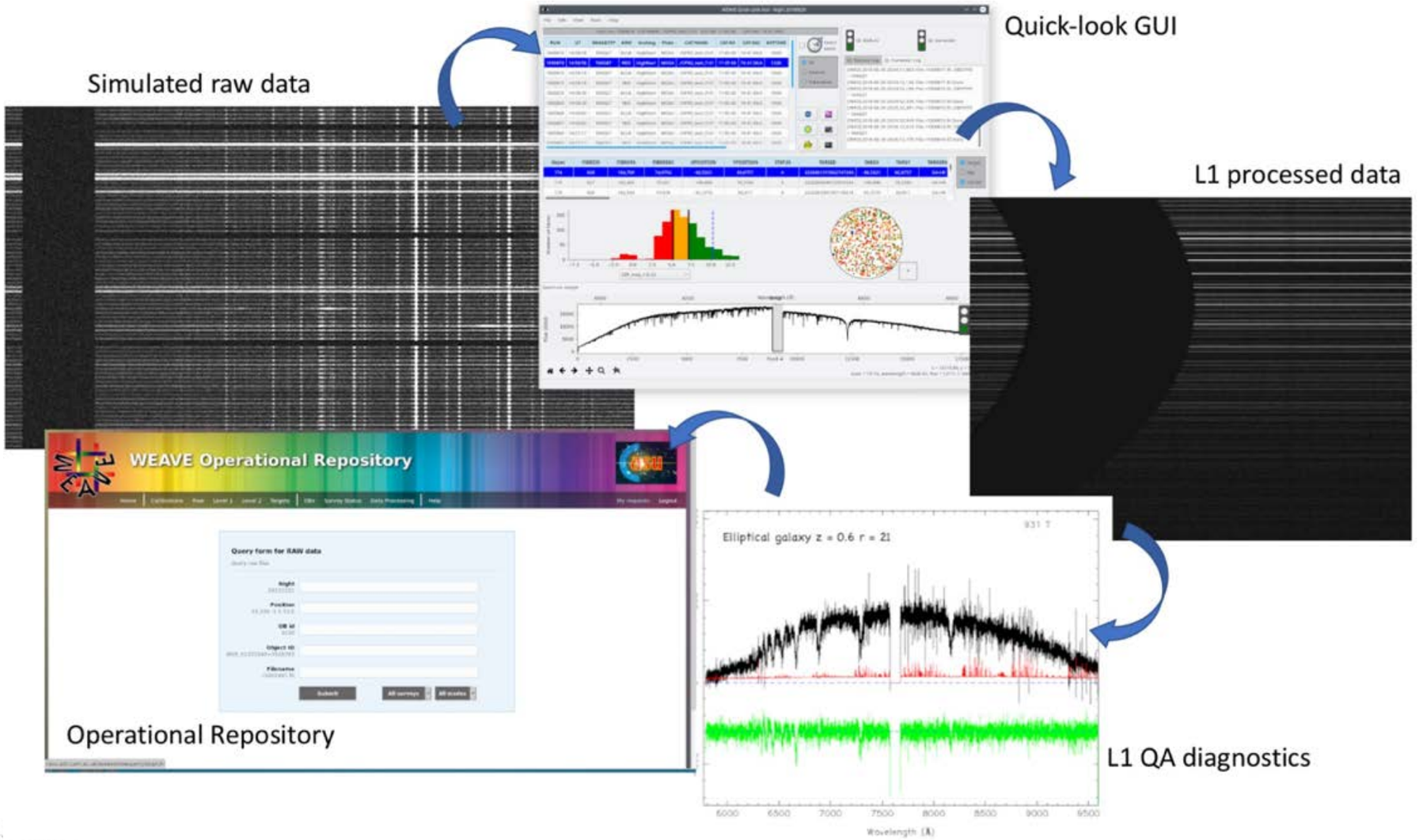


WEAVE Data Flow (simplified)



Goals: Galactic structure (Gaia complement), Galaxy evolution (LOFAR complement)
IoA 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

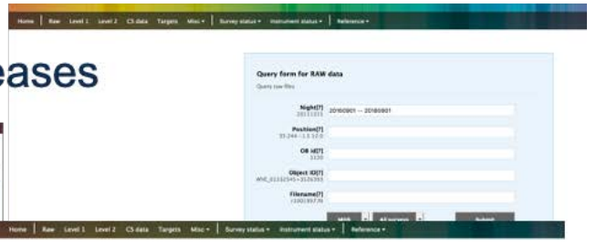


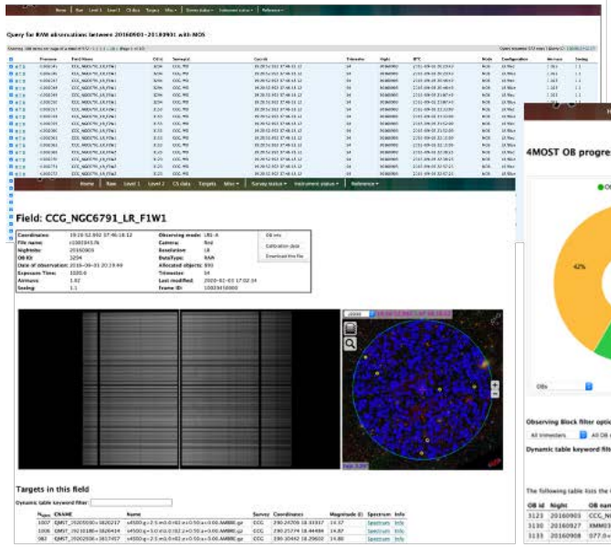
Output from OpR3 allowed full assessment of:

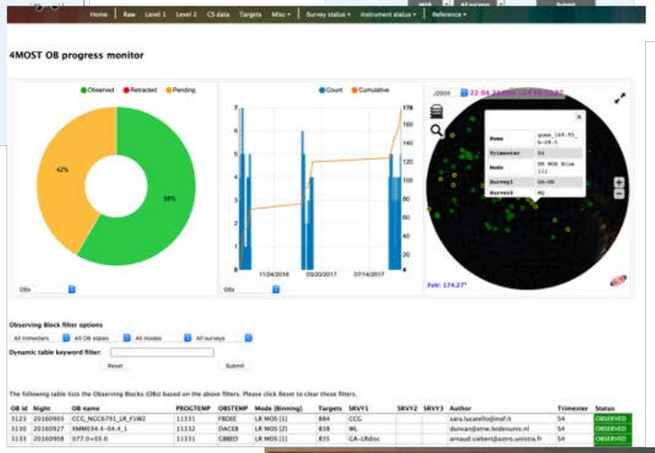
- Survey scheduling
- Quality of L1 and L2 data
- WEAVE archive
- Overall system performance
- Operations both WEAVE and Science teams

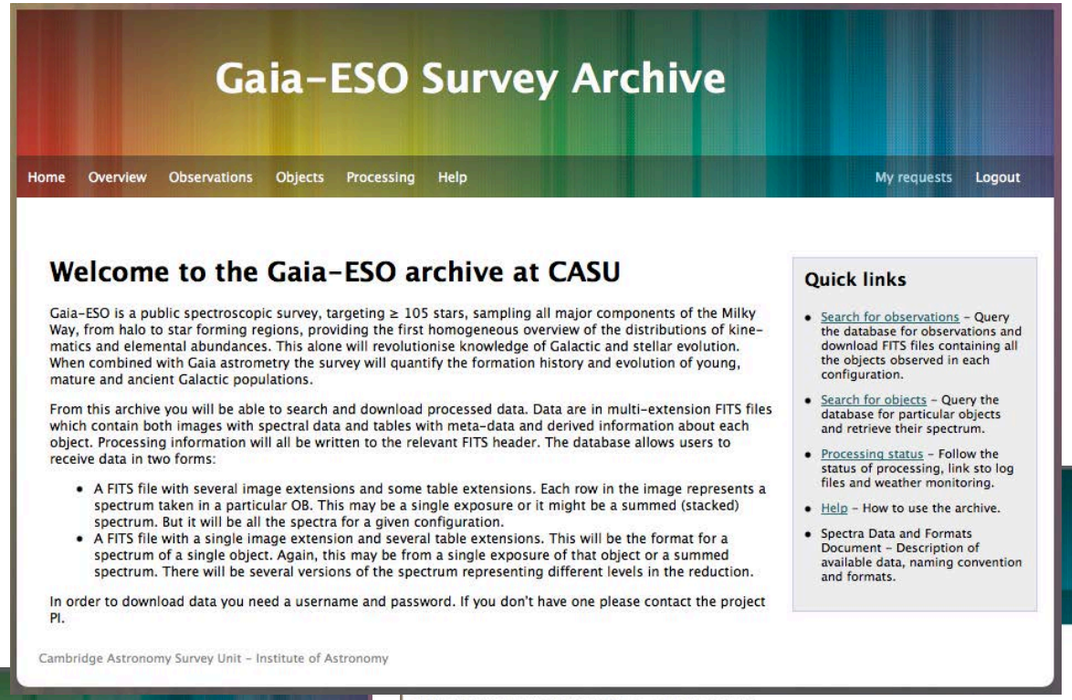
CASU Archives and Operational Repositories

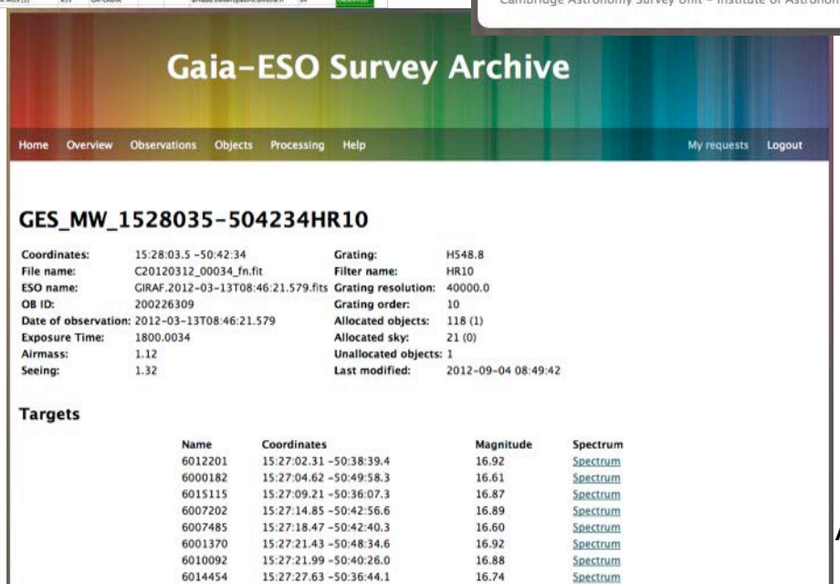
4OR: supporting internal releases

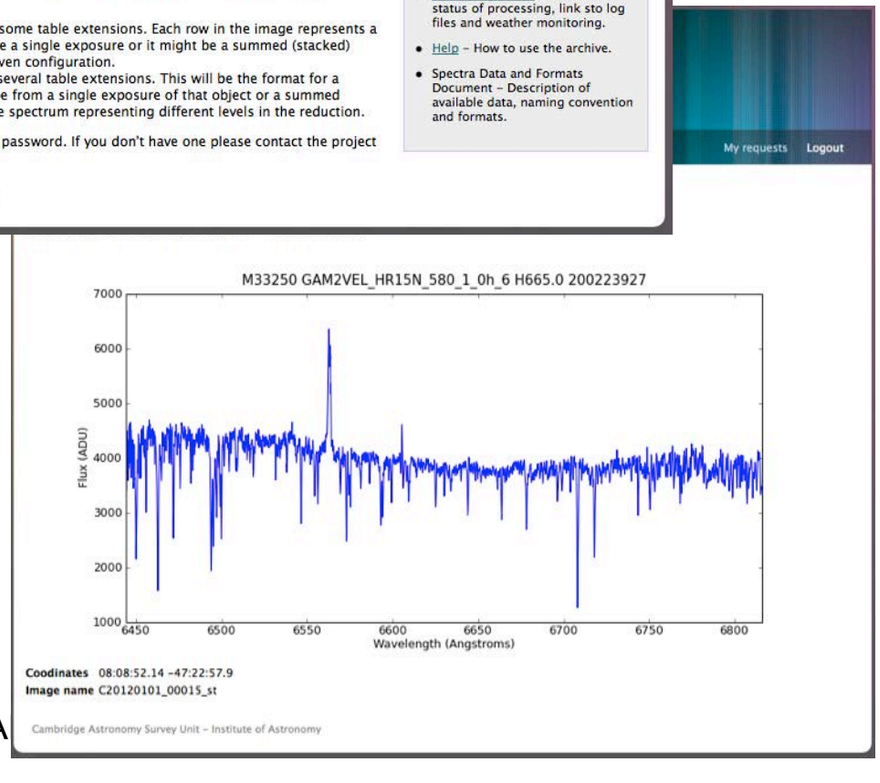












CASU provide access to data both internally to survey teams and externally to consortia. Database access via IRIS will provide enhanced resilience.

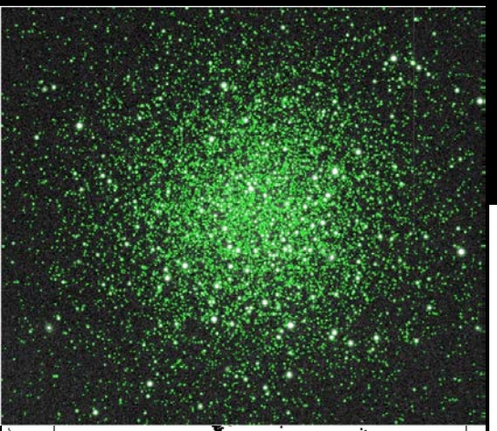
Applying ‘astro’ techniques to the medical domain

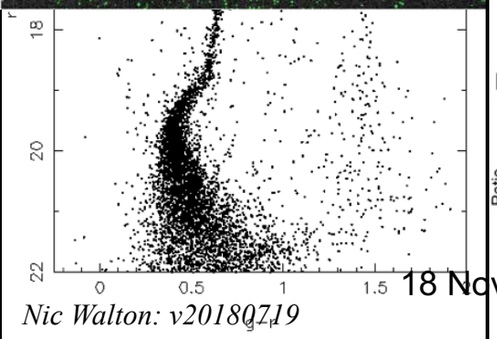
Image: ESO

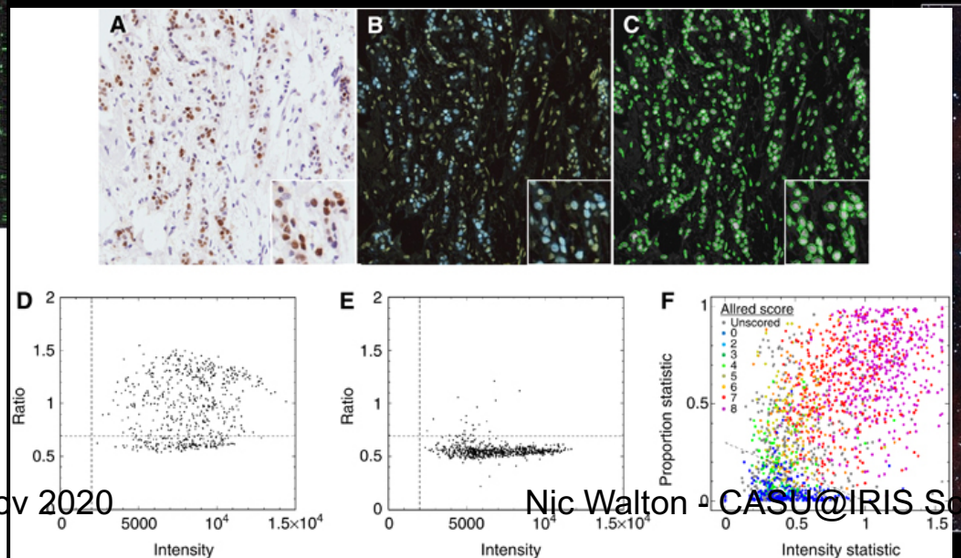
Science & Technology Facilities Council
10 Years of Impact and Inspiration

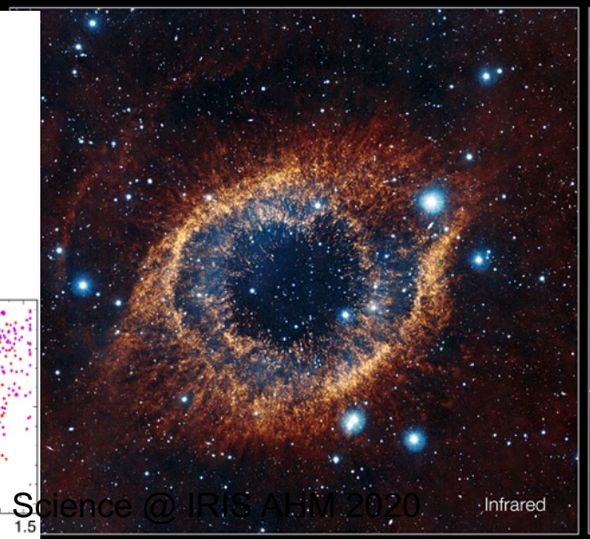
UK SPACE AGENCY

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







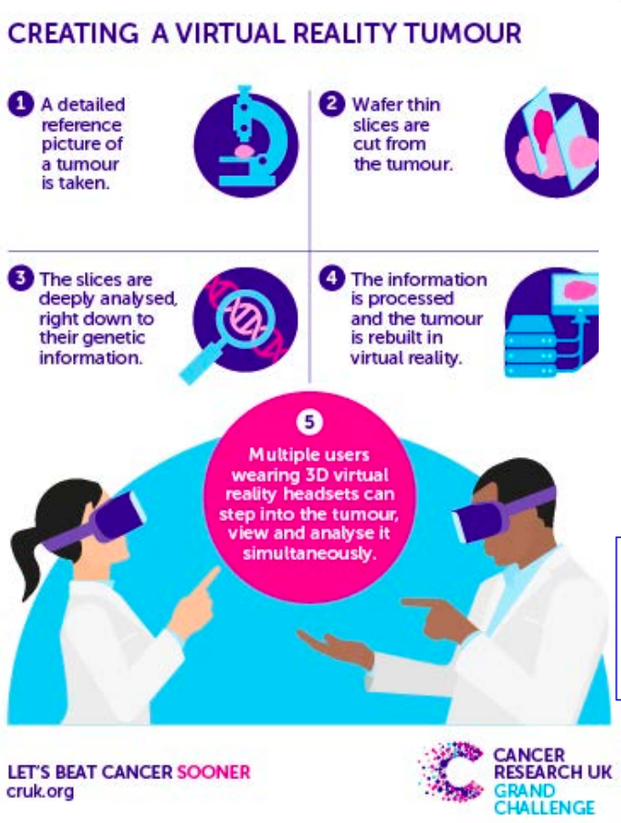


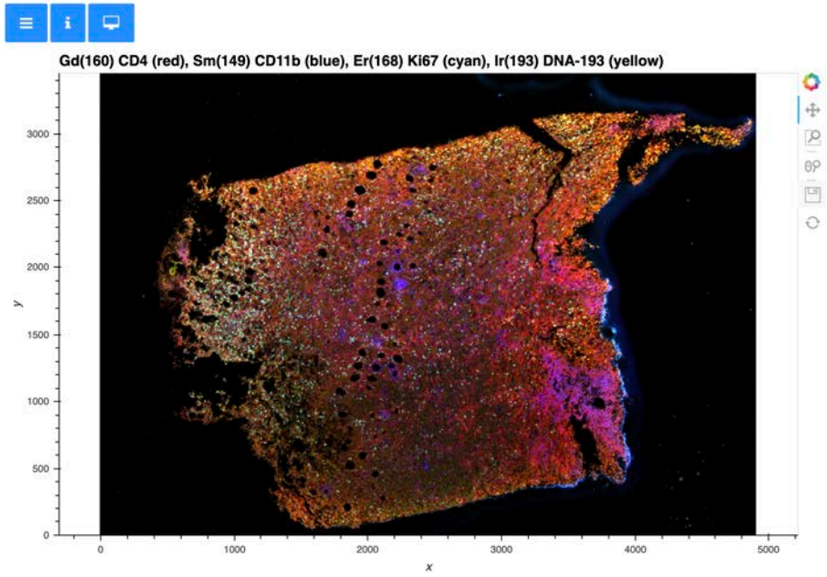
Nic Walton: v20180719

IMAXT Data Analysis and Infrastructure

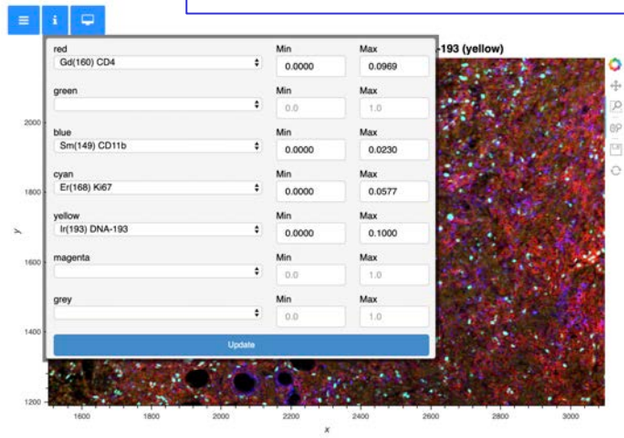
Cancer Research UK Grand Challenge

Interactive data visualisation



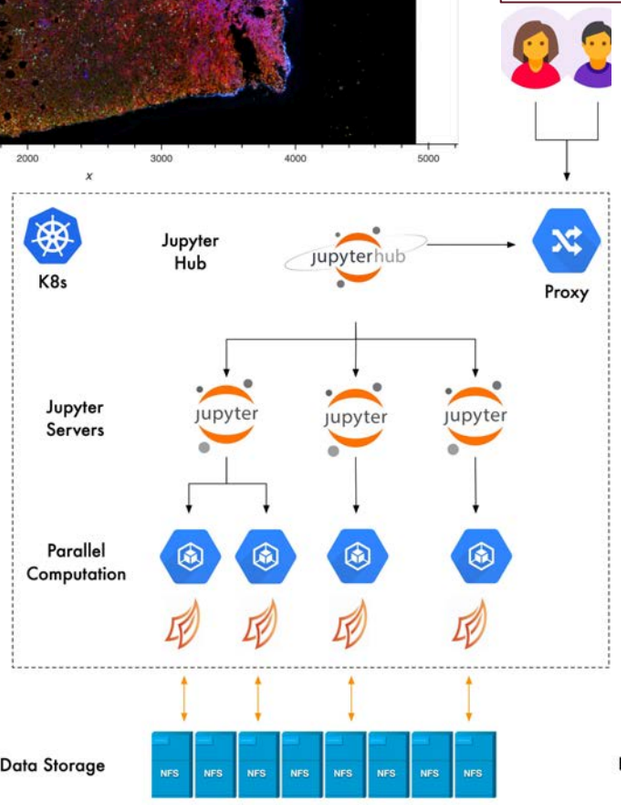


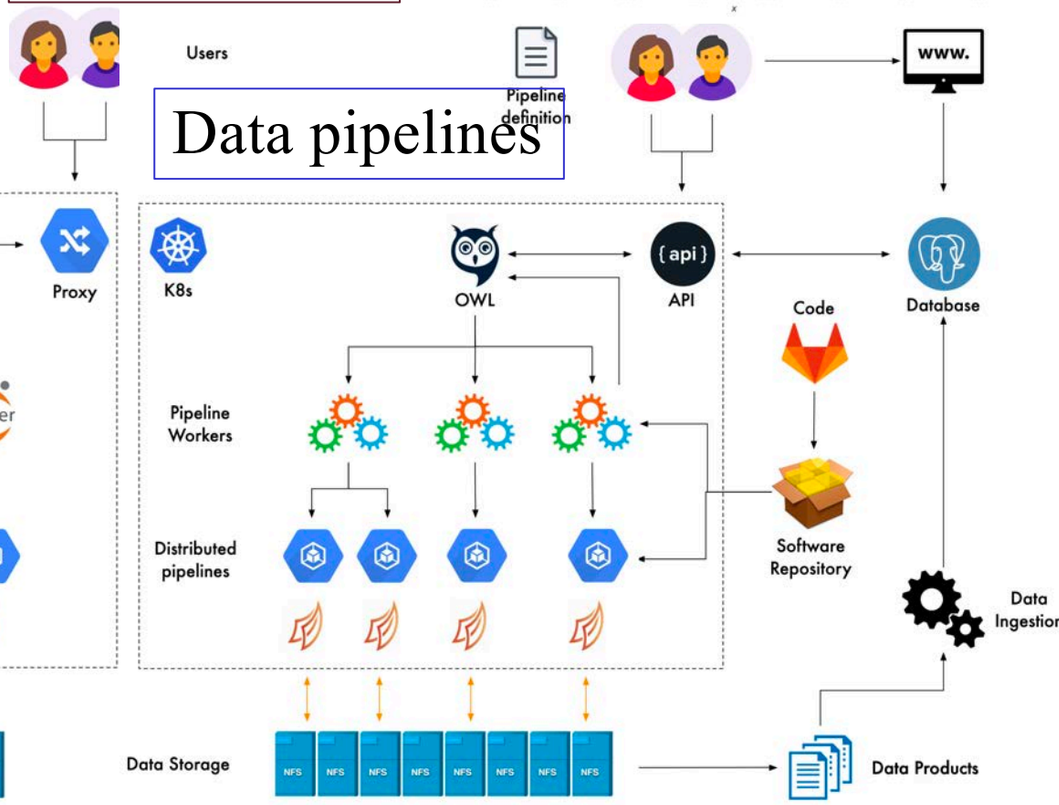
Deployment could be made on Openstack@ IRIS



Interactive data access

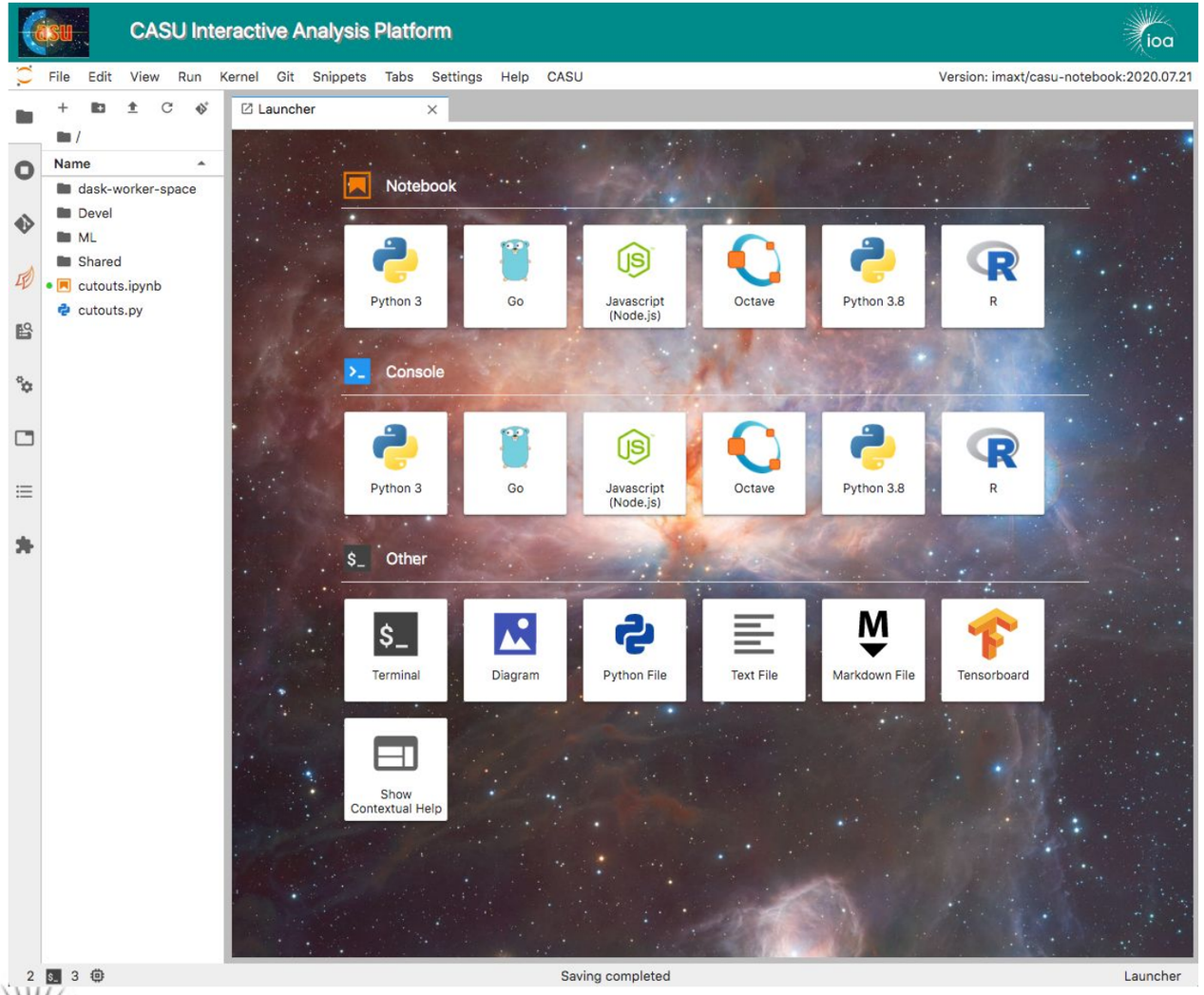
Data pipelines

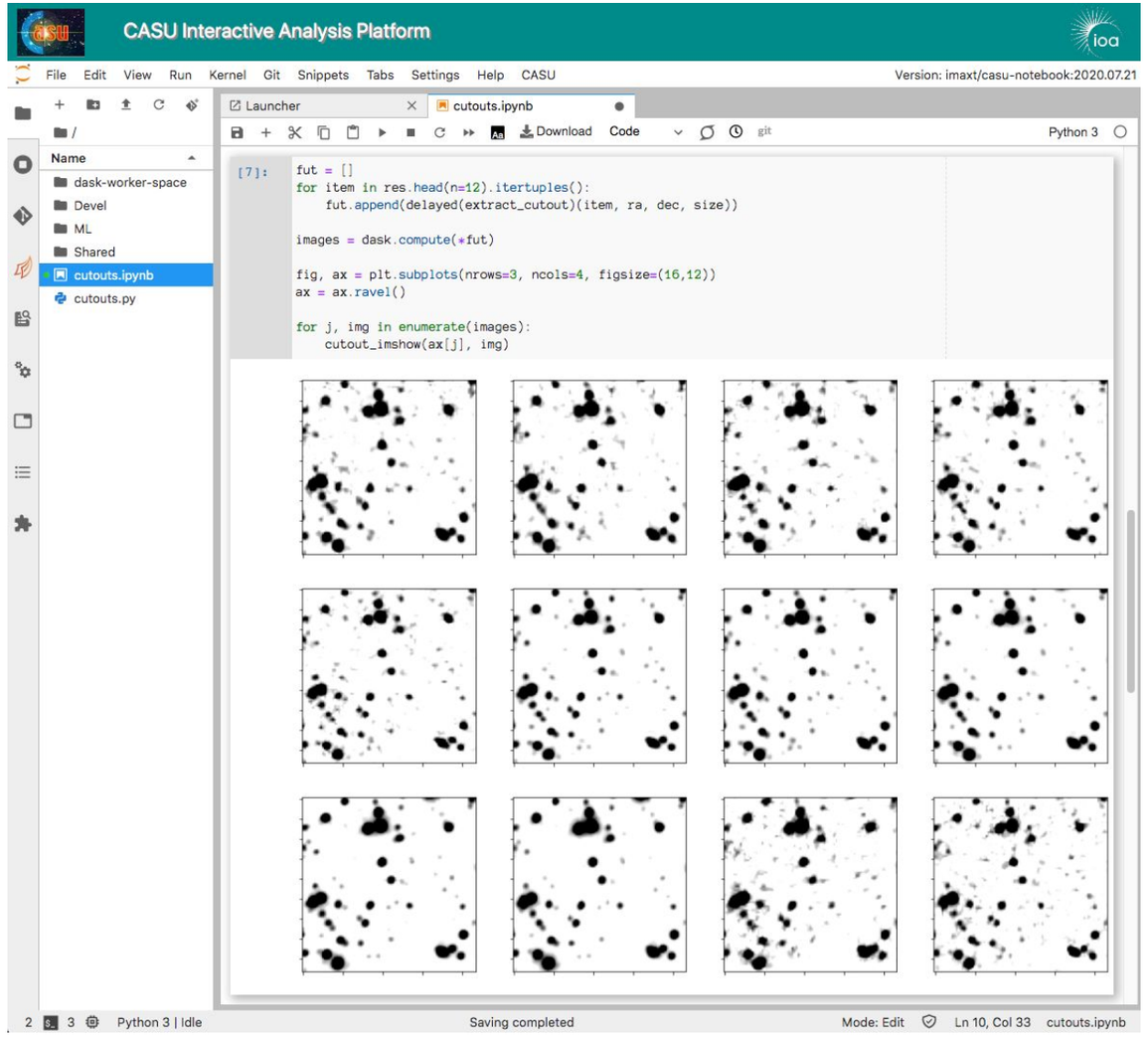




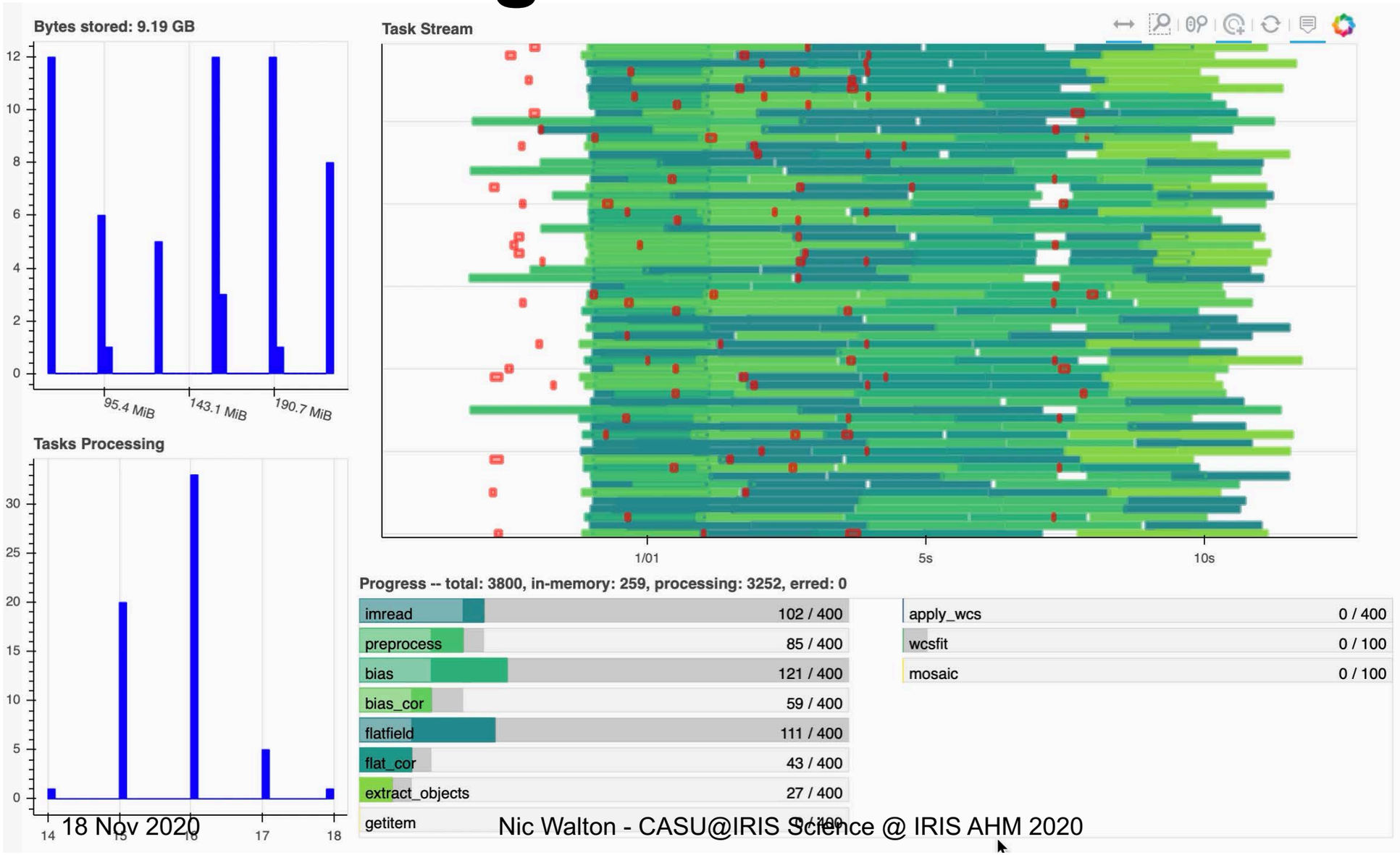
CASU responsible for primary image analysis and data infrastructure – transfer of CASU technology from astro to medical domain

JupyterHub Deployment: CASU@IRIS





Processing 100 images in distributed containers using 60 cores

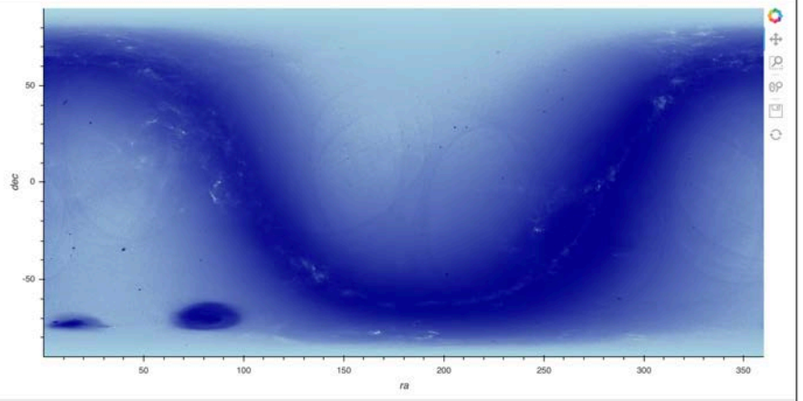


Query VISTA database around position

```
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 dissplay first 10 results
res = vistadb.query_radec(ra, dec)
res[columns].head(n=10)
```

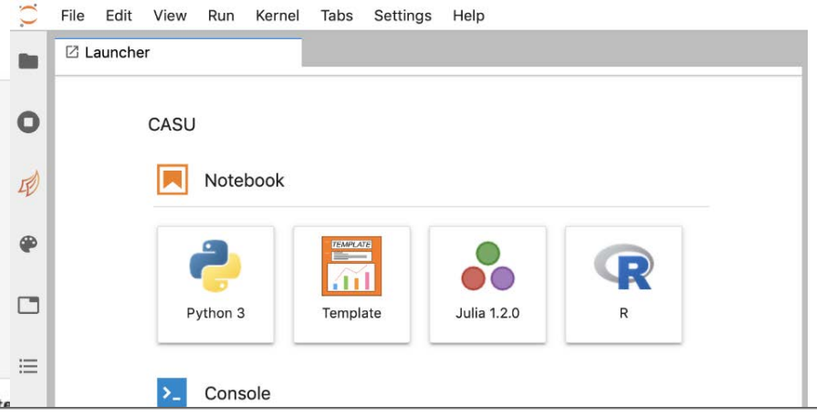
id	filename	coords	filtername	surveyname	nightobs	tote
62124	v20100218_00330_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201002	
68702	v20100228_00495_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201002	
70725	v20100304_00444_st_tl.fit	13:02:18.02 -64:35:28.7	Ks	VVV	201003	
81752	v20100315_00327_st_tl.fit	13:02:18.02 -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	

```
%%opts RGB [width=1000, height=500]
hd.datashade(points)
```

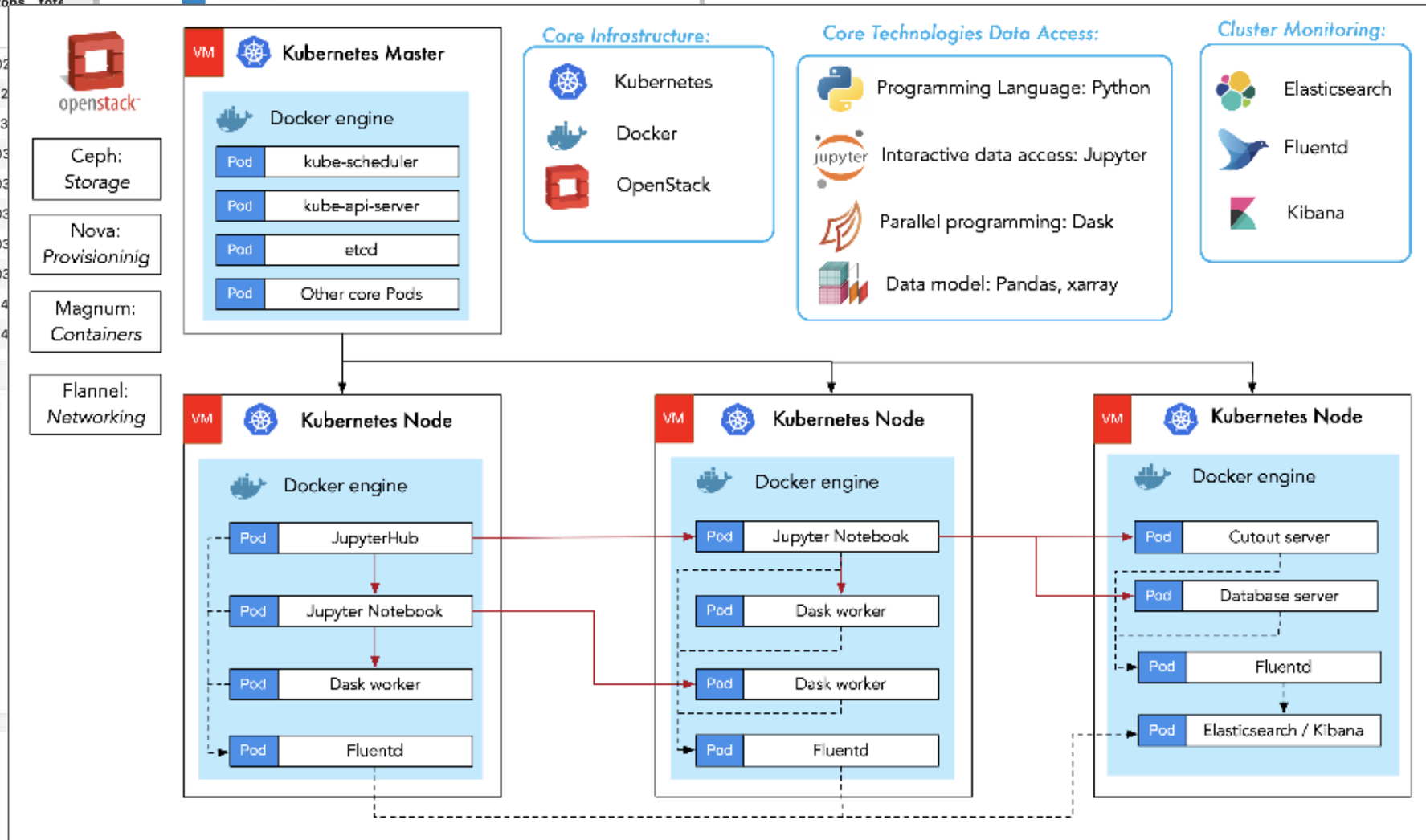



```
df.count().compute()
```

```
ra      1692586238
dec     1692586238
dtype: int64
```



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.

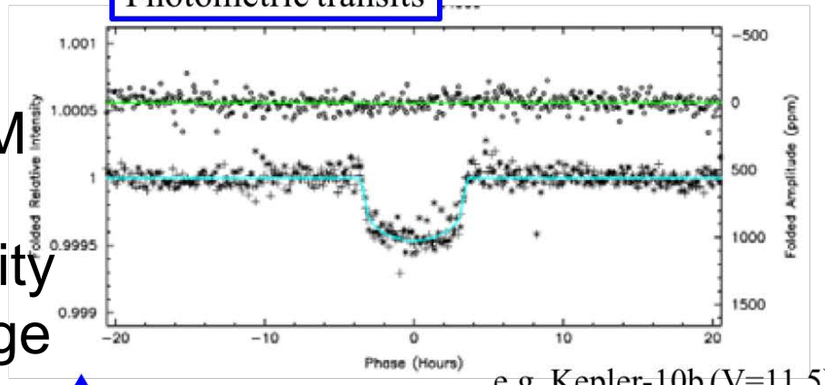
Launch end 2026. 4 to 8 years operations plus post operations

PLATO
PLAnetary Transits and Oscillation of stars

PLATO 2.0: Rauer et al 2013

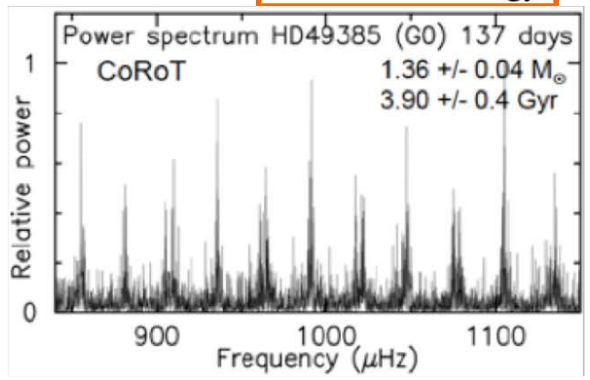
Goals: will detect and characterize planets down to Earth-size by high precision photometric transits around ~1M bright stars. Planetary masses will be determined by ground-based radial velocity measurements. Stellar parameters like age and mass will be obtained by asteroseismology

Photometric transits



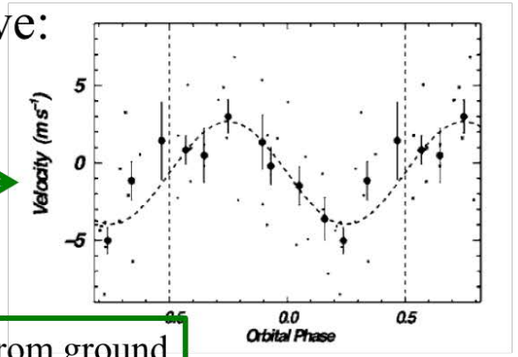
e.g. Kepler-10b (V=11.5)

Asteroseismology



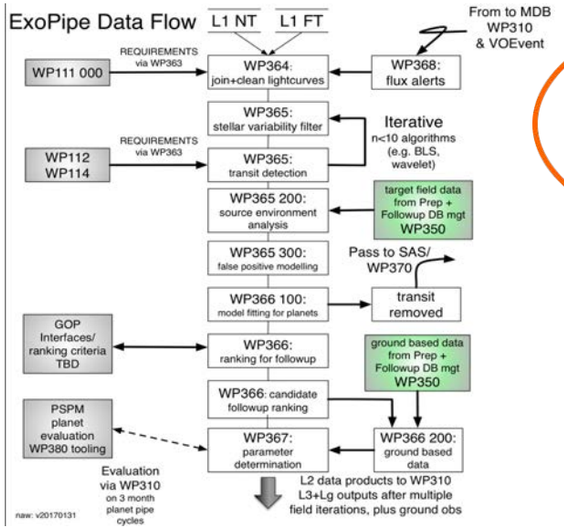
For solar like stars PLATO will give:

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



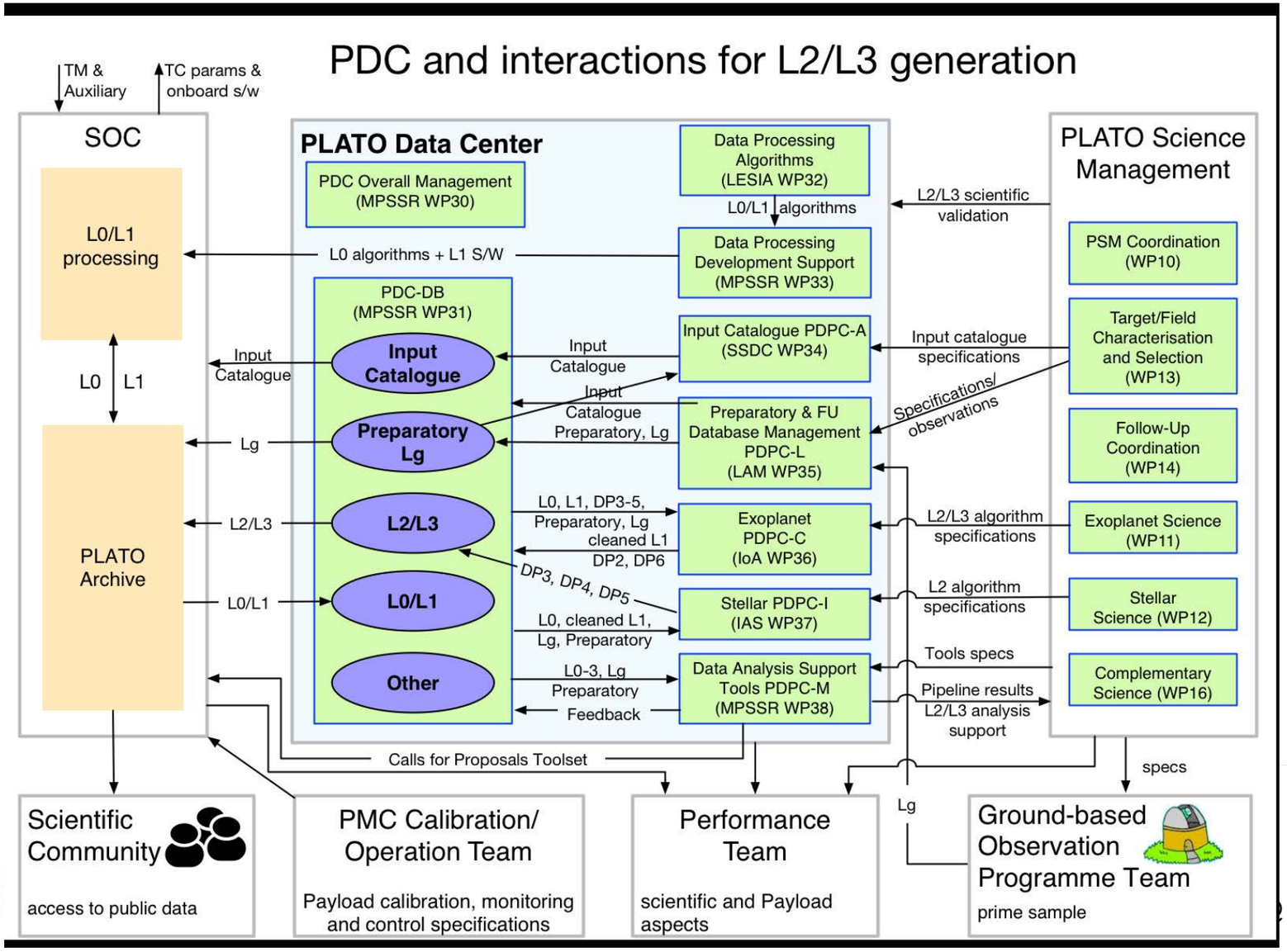
RV follow-up from ground

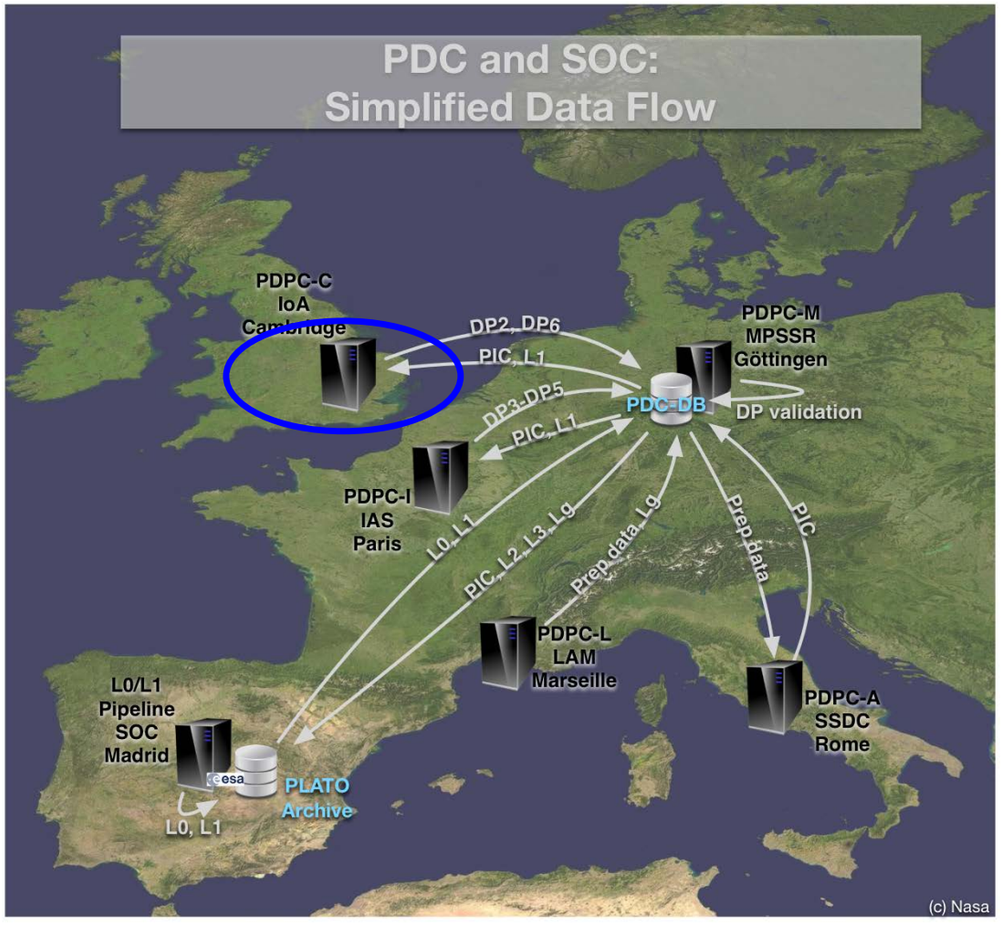
What? IoA development and operations of Exoplanet Analysis System pipelines.



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

PLATO Data Processing Centre
Exoplanet Analysis at @ Cambridge





Large data flow: ~110 GB/ day
Complex data

Gaia and DPAC @ Cambridge

Credit: ESA Gaia/ DPAC

Mapping the Milky Way: 2 Billion Stars

Gaia Data Release 2 (Apr 2018):

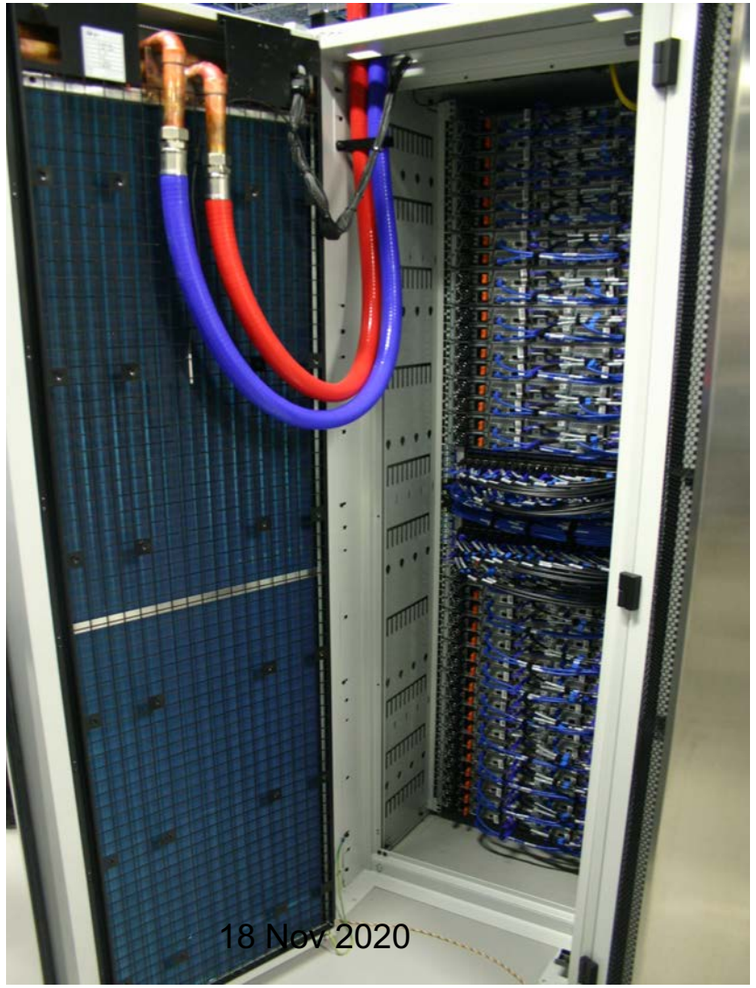
https://www.cosmos.esa.int/web/gaia/dr2

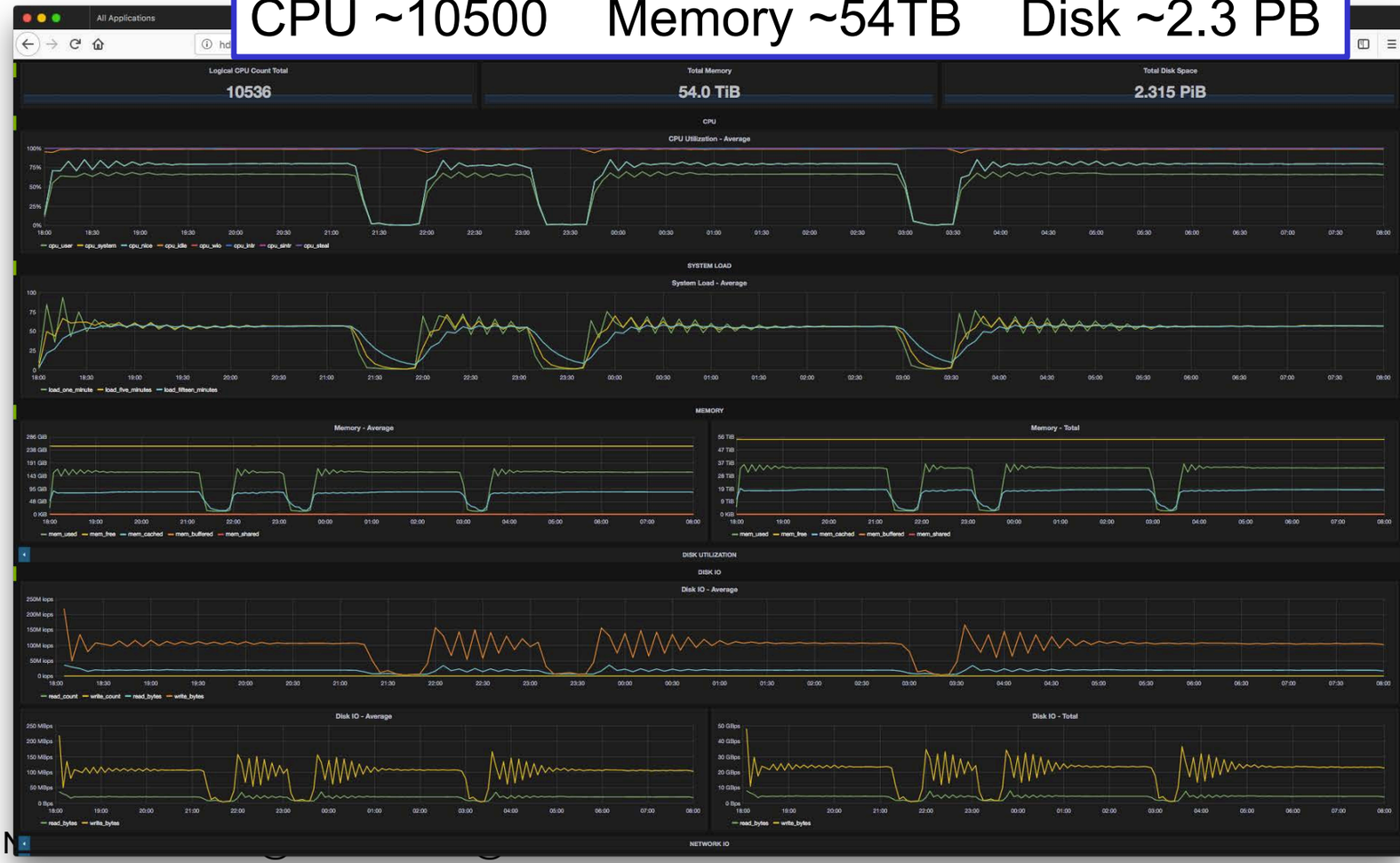
Gaia Early Data Release 3 (3rd Dec 2020)

Cambridge DPCI Cluster
@ West Cambridge Data Centre

220 nodes provide combined compute+storage

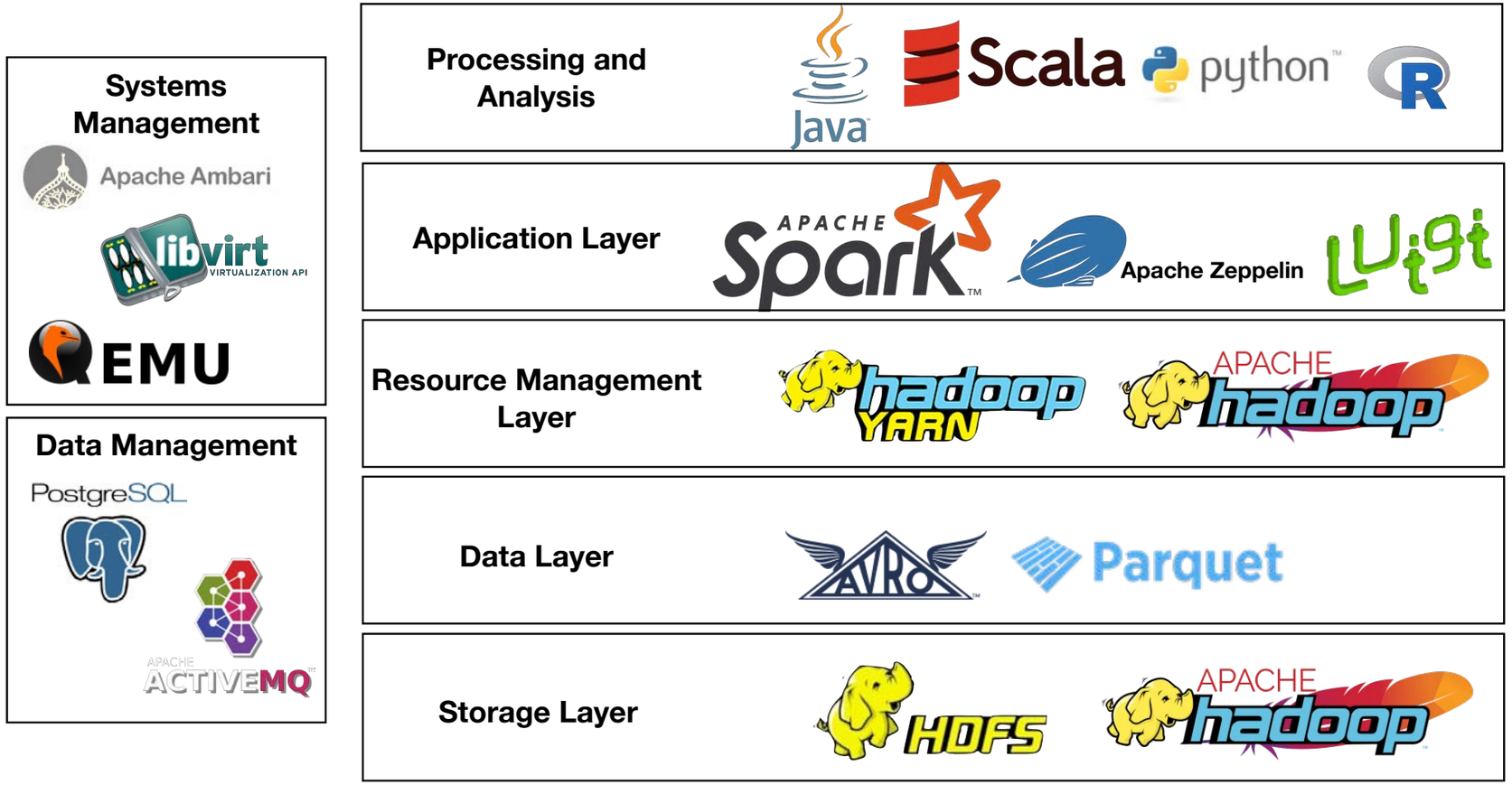
CPU ~10500 Memory ~54TB Disk ~2.3 PB





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