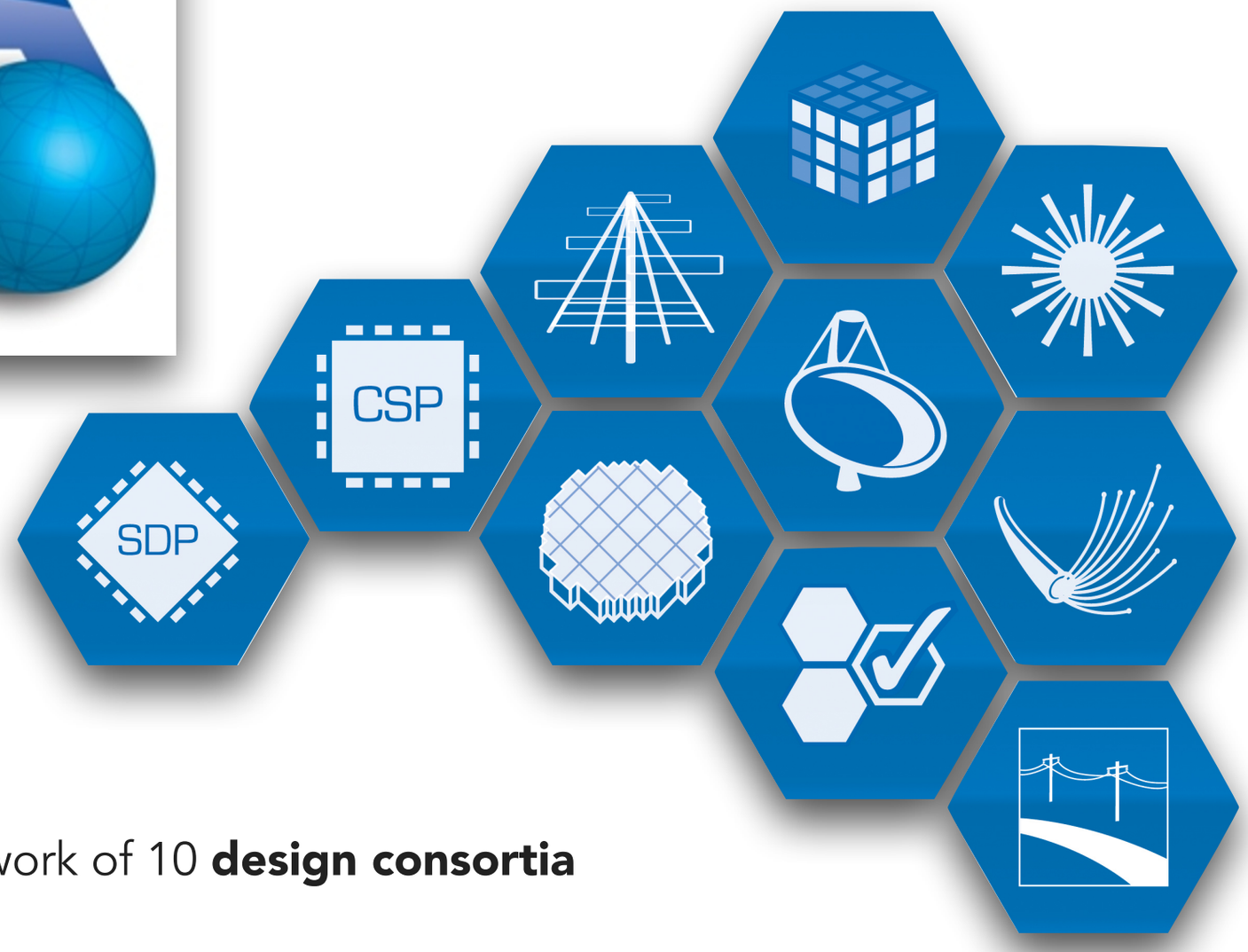
Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

SKA

Anna Scaife
Jodrell Bank Centre for Astrophysics

Interferometry Centre of Excellence

The Square Kilometre Array International Organisation (**SKAO**)



The SKAO oversees the work of 10 **design consortia**

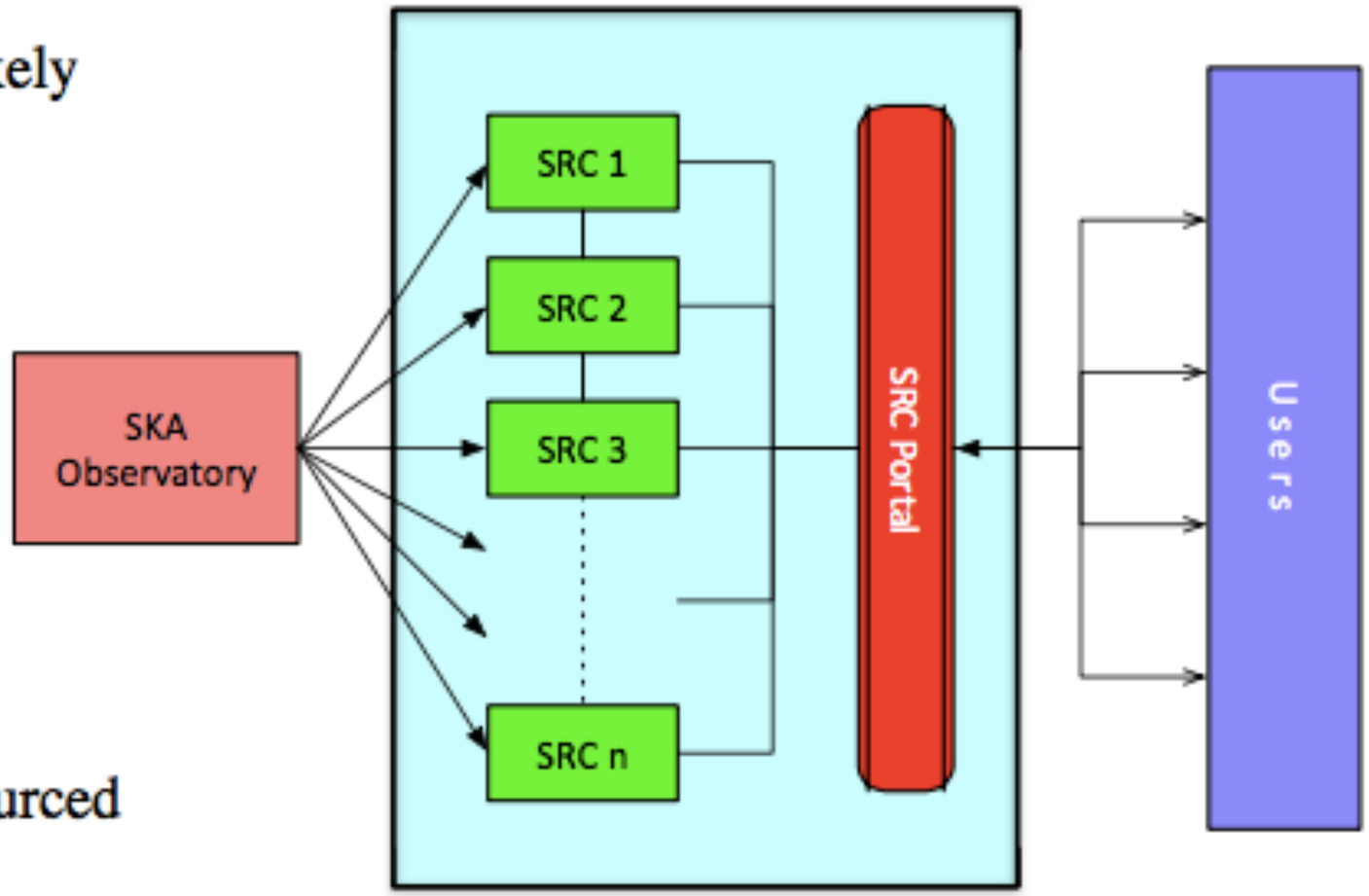

SKA
SQUARE KILOMETRE ARRAY

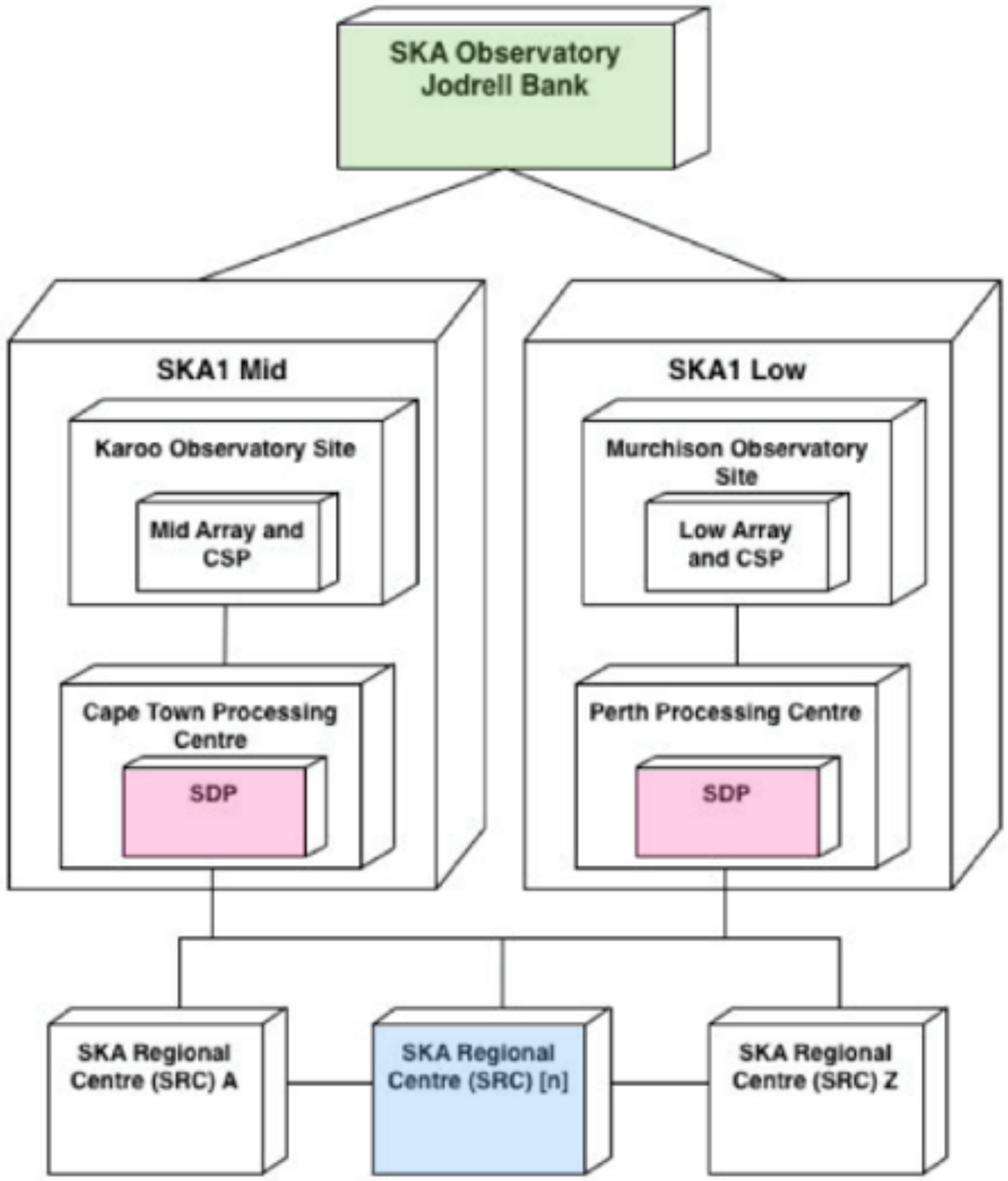

SKA
SQUARE KILOMETRE ARRAY
SDP
SCIENCE DATA PROCESSOR


Aeneas
Advanced European Network of E-infrastructures
for Astronomy with the SKA

SKA Regional Centres

- Science Data Centres (SDCs) will likely host the SKA science archive
- Provide access and distribute data products to users
- Provide access to compute and storage resources for users
- Provide analysis capabilities
- Provide user support
- Multiple regional SRCs, locally resourced

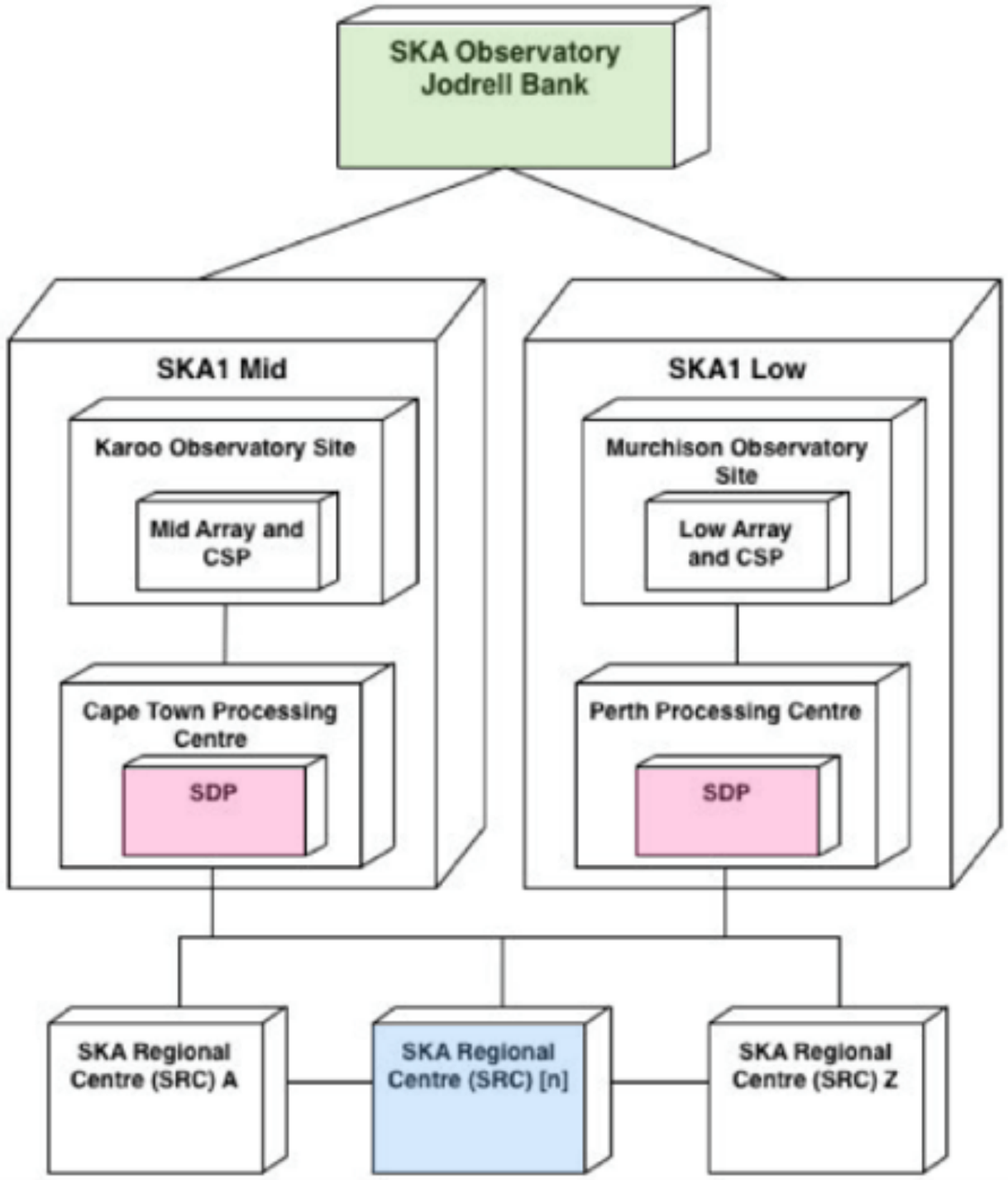




CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

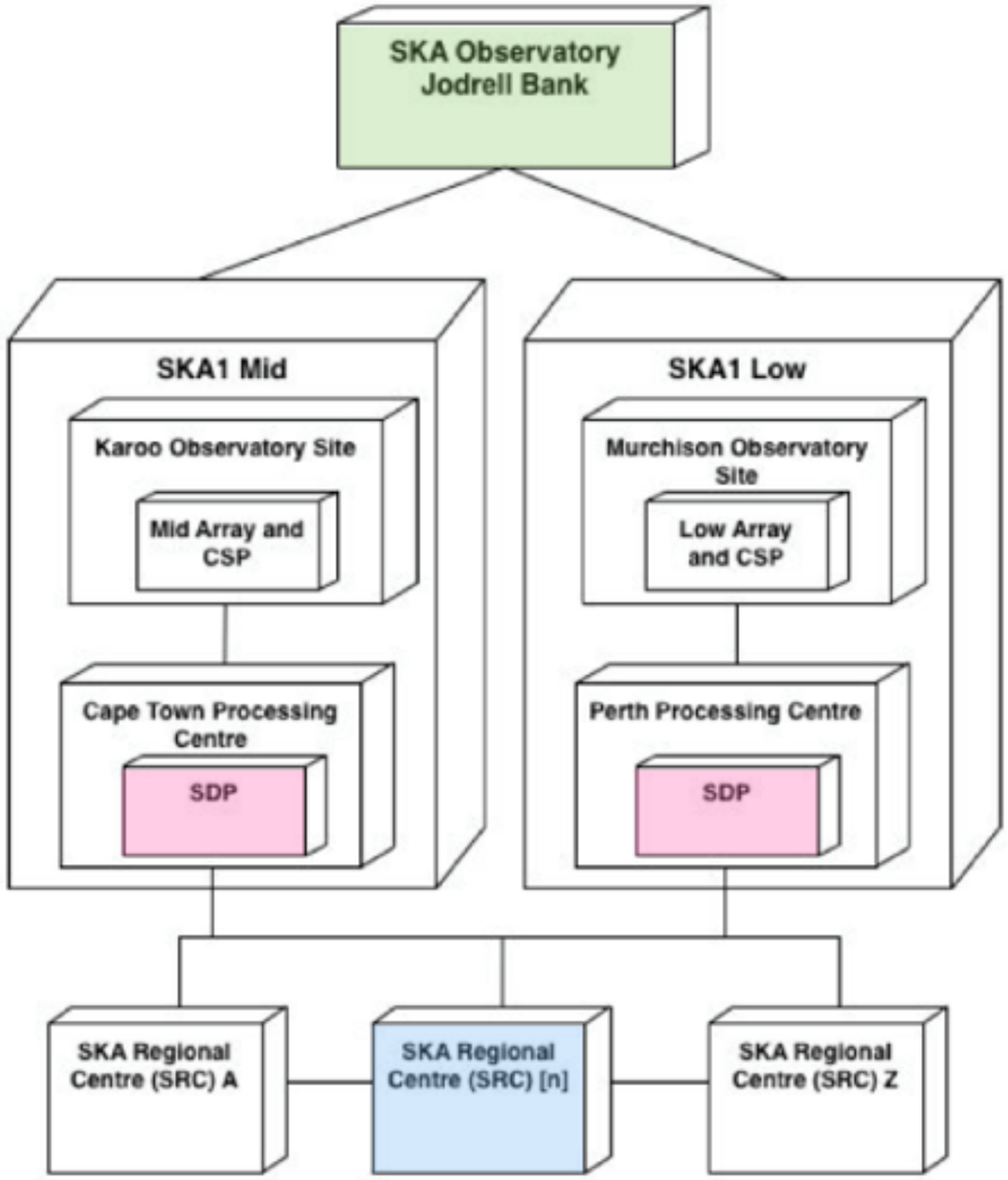
REGIONAL DATA CENTRE



CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE



SQUARE KILOMETRE ARRAY

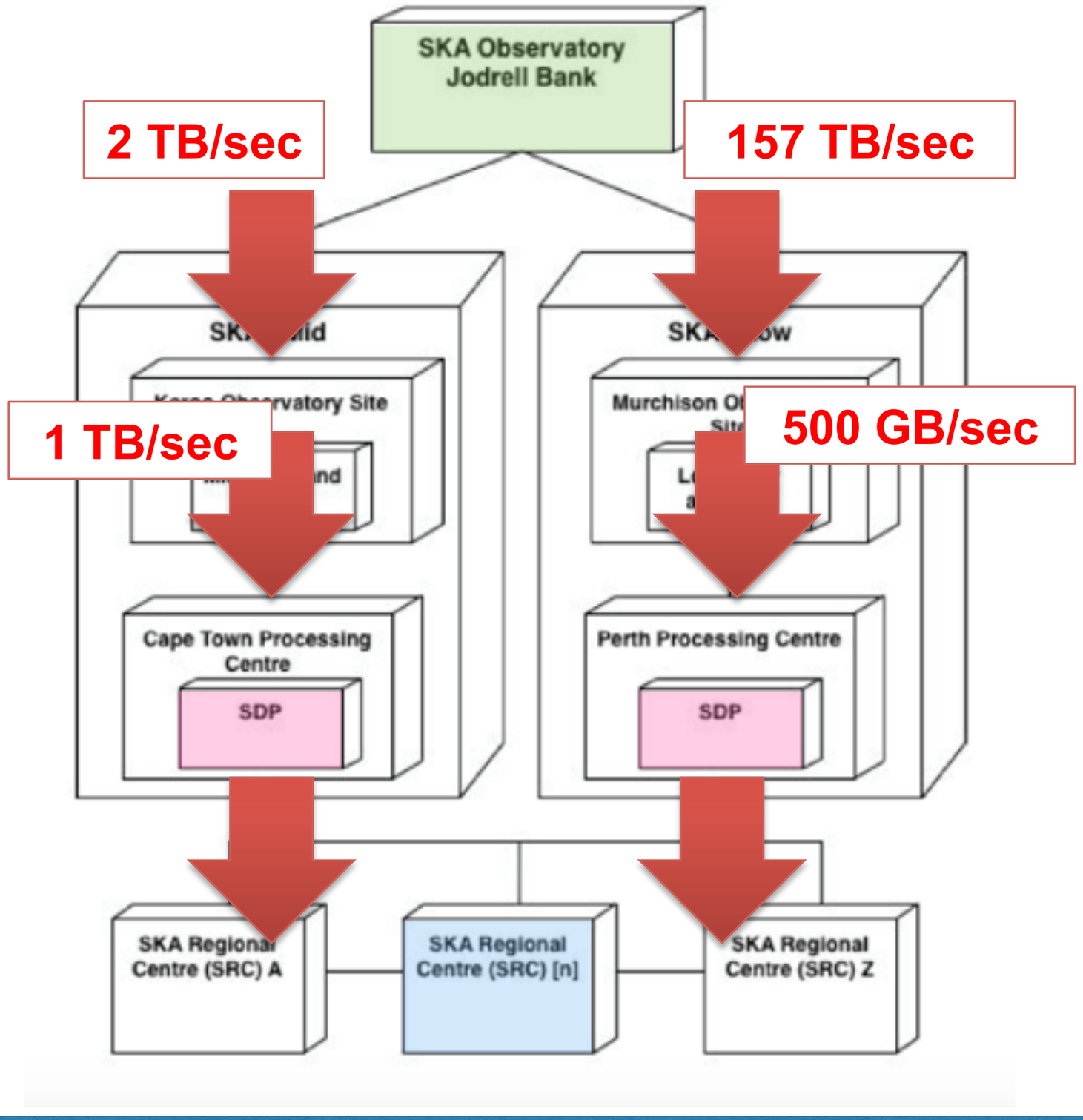
CENTRAL SIGNAL PROCESSOR

CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSOR

SCIENCE DATA PROCESSING

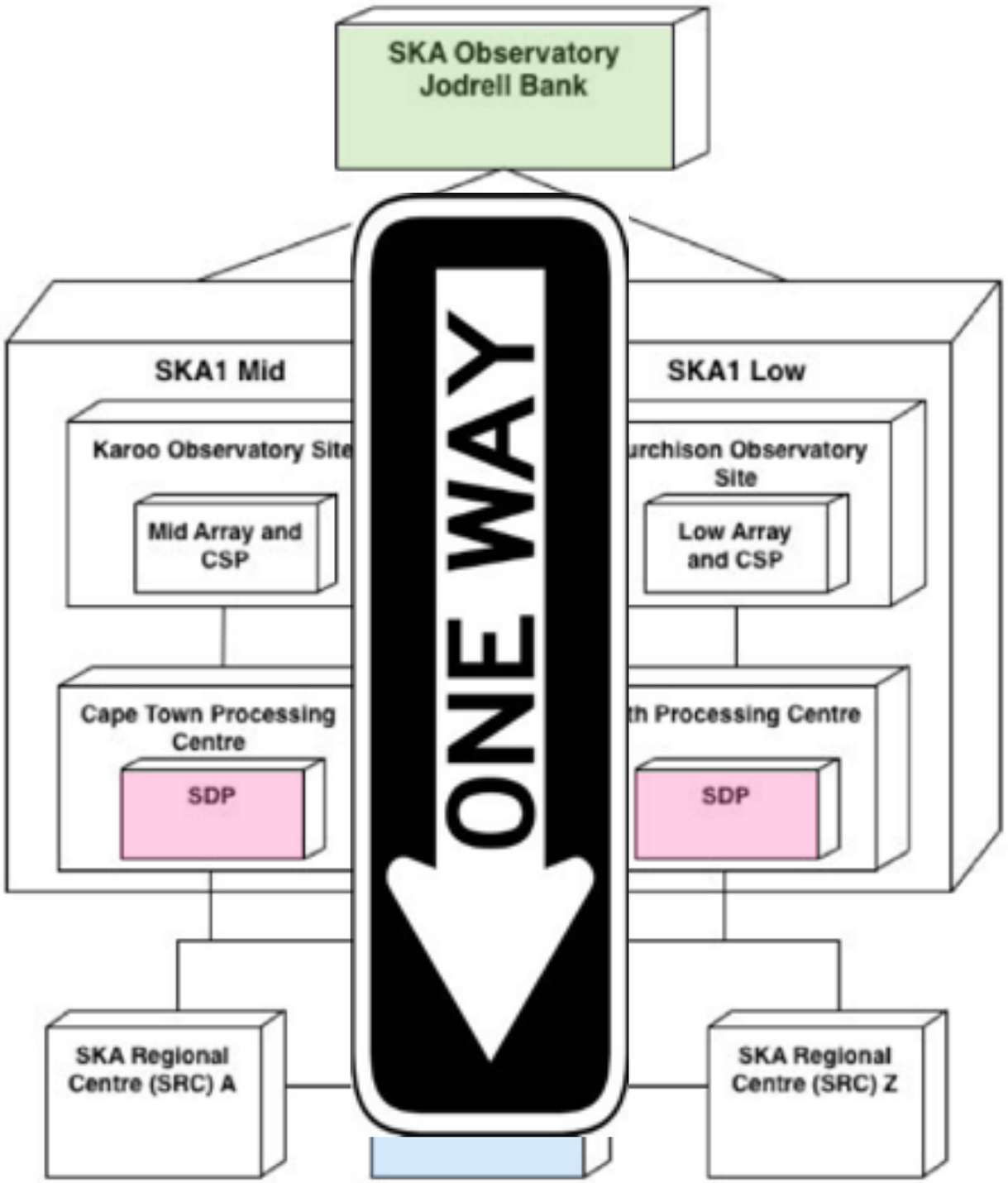
REGIONAL DATA CENTRE



CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE



Standardized data products

A standard SKA1-MID image data product has
30k x 30k pixels

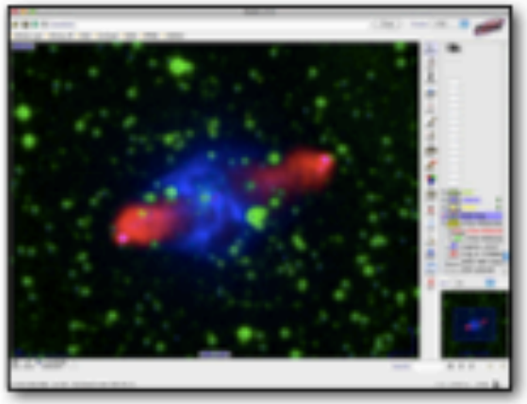
SKA1 will have up to **65k frequency channels**
and **4 polarisations**

At 4 Bytes per voxel that equates to
30k x 30k x 65k x 4 x 4
= **936 TeraBytes**

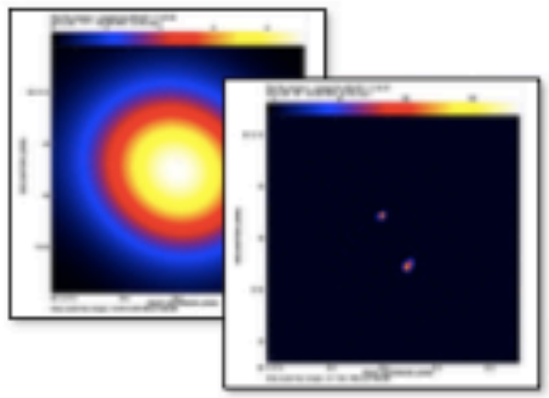
Regional Centre Functionality

Data Discovery

- Observation database
- Quick-look data products
- Flexible catalog queries
- Integration with VO tools
- Publish data to VO



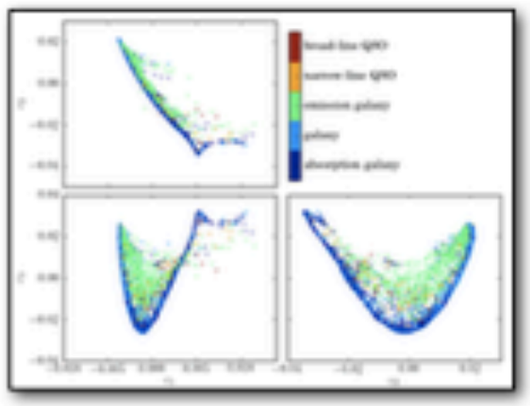
Data Processing



- Reprocessing
- Calibration and imaging
- Source extraction
- Catalog (re-)creation
- DM searches

Data Mining

- Multi-wavelength studies
- Catalog cross-matching
- Transient classification
- Feature detection
- Visualization



What do we know about data and its consumption

- Hierarchical data structure within each experiment, key science based namespaces – further extended to be PI-based
- Granularity at which data is managed varies from experiment to experiment
- Pre-determined ”push” mechanism from the SDP to Regional centres
- Data storage will be at various locations possibly under different administrative domains
- Number of users would be a few thousand
- “Passive users” consuming data will also be generating secondary data which may not be smaller than raw data
- Likely X.509 certificates for authentication
- Commonly used tools are CASA, AIPS, Miriad, PRESTO, SIGPROC

Data Products at the Regional Centre

- Image type data products
  - Image cubes
    - Continuum Survey, Magnetism, Hi Kinematics, ISM
    - Data archive for these experiments would range from a fraction of a PB to 120 PB
    - Since hours of telescope time differ, it is useful to look at data generated per 6 hour observation. This will range from 0.1 to 100 GB
  - U-V Grid – calibrated visibilities
    - EoR experiments on SKA1 LOW
    - Data archive of almost 220 PB
    - Per Observation ~270 GB
- Non-image data products
  - Pulsar search and timing experiments
    - Data archive of 250 GB to a few PB, per observation less than 3 GB
    - LSM Catalogue, Transient catalogue, Pulsar timing solutions, Transient buffer data, Sieved pulsar and transient candidates

Experience with existing e-Infrastructure

- SKA.GridPP meeting in Manchester 2016
- skatelescope.eu VO established
- Testing on GridPP started in 2017 as part of AENEAS
- Lots of help from Andrew & Alessandra

Three initial programs:

1. Interferometric imaging compute model
2. Object detection & classification (image based)
3. Synergistic science incorporating multi-wavelength surveys (catalogue based)

Alessandra Forti

Andrew McNab

Rohini Joshi

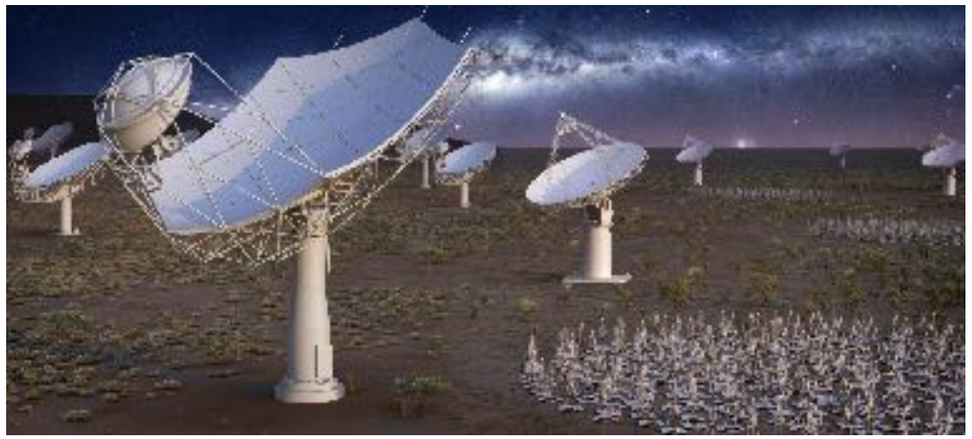
Thérèse Cantwell

Alex Clarke

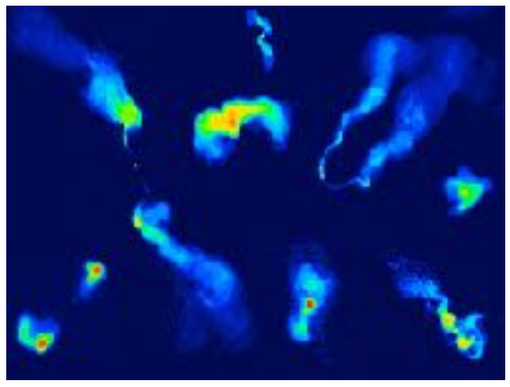
(1) Interferometric Imaging

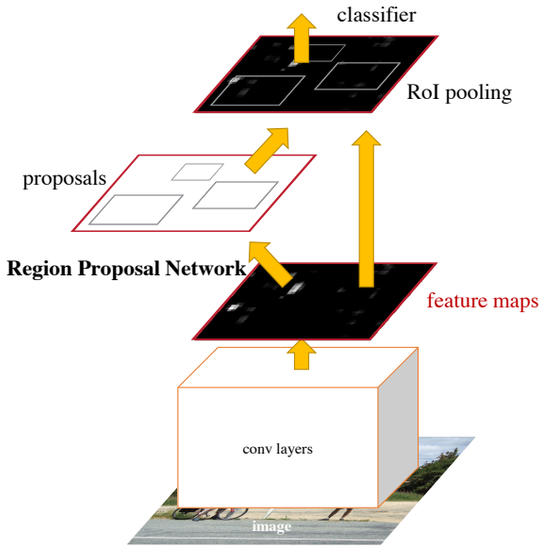
- LOFAR data – GOODS-N survey. One observation is 3.5 TB
- Uploaded to the Manchester storage nodes using LTA's HTTP interface and a parametric set of jobs
- Accessed using Logical File Names (LFNs)
- Time-slice sub-bands into manageable chunks to process them
- Calibration using LOFAR software on CVMFS being tested on GridPP – looking into singularity as an alternative
- Using DIRAC's DMS and WMS
- GridPP liaisons - Andrew McNab and Alessandra Forti

(2) Object detection & classification (image based)



Classifying sources by eye takes too much time!





- **Using region-based CNN to detect objects in a field and classify them**
- **Specifically using a python based implementation of faster R-CNN for CPU**

Automatically generate source catalog including

- **https://github.com/rbgirshick/py-faster-rcnn**
- **Plan to containerise using Singularity for testing on GridPP**

```
[vagrant@localhost 8237808]$ ls
cow_file.txt  StdOut
[vagrant@localhost 8237808]$ cat cow_file.txt
 ---------------------------
< The cow is on the grid! >
 ---------------------------
        \   ^__^
         \  (oo)\_______
            (__)\       )\/\
                ||----w |
                ||     ||
[vagrant@localhost 8237808]$
```

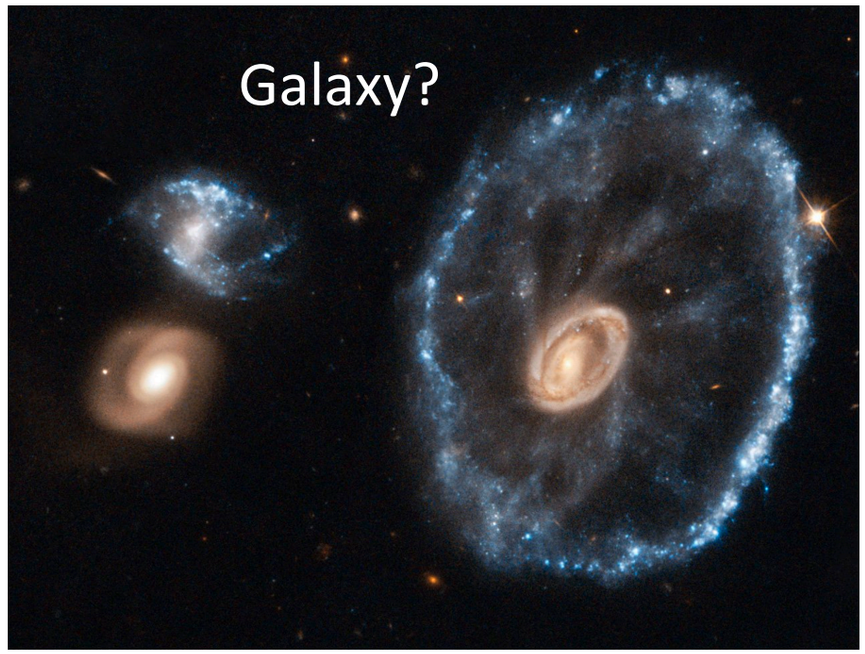
(3) Synergistic science incorporating multi-wavelength surveys (catalogue based)

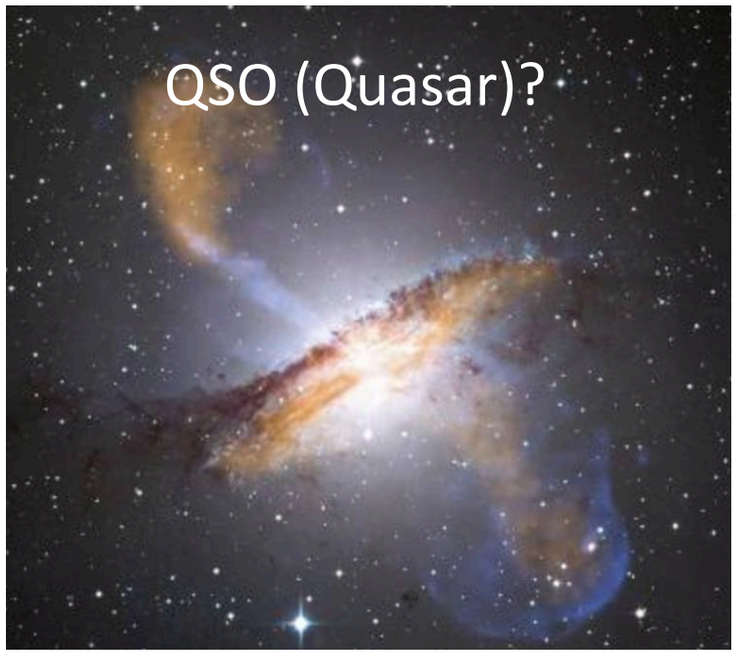
SciServer is an online system for accessing & analysing scientific big data projects in astronomy, and other areas.

Identify source position and morphology from SKA image → Cross-match with archival multi-wavelength survey data from SciServer → Apply Machine Learning algorithms to inform about SKA sources

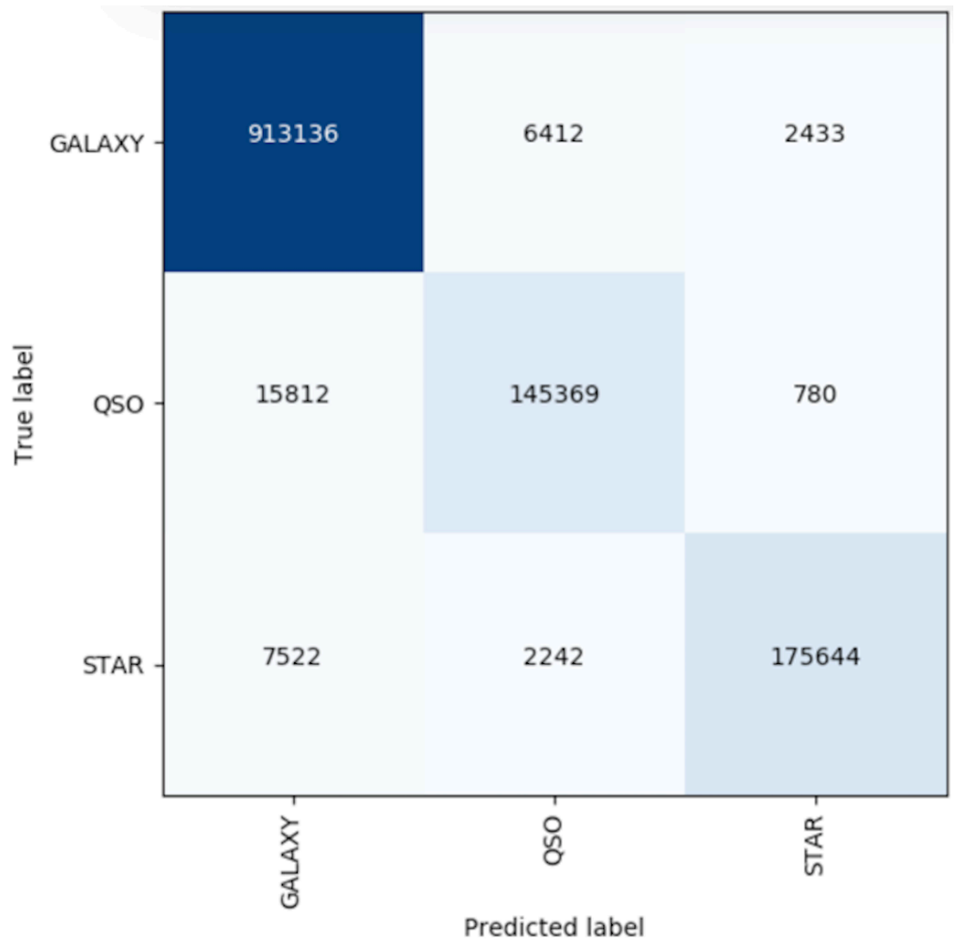
Apply this schematic in a distributed computing set up to run analysis on millions of objects, each with hundreds of data points describing them

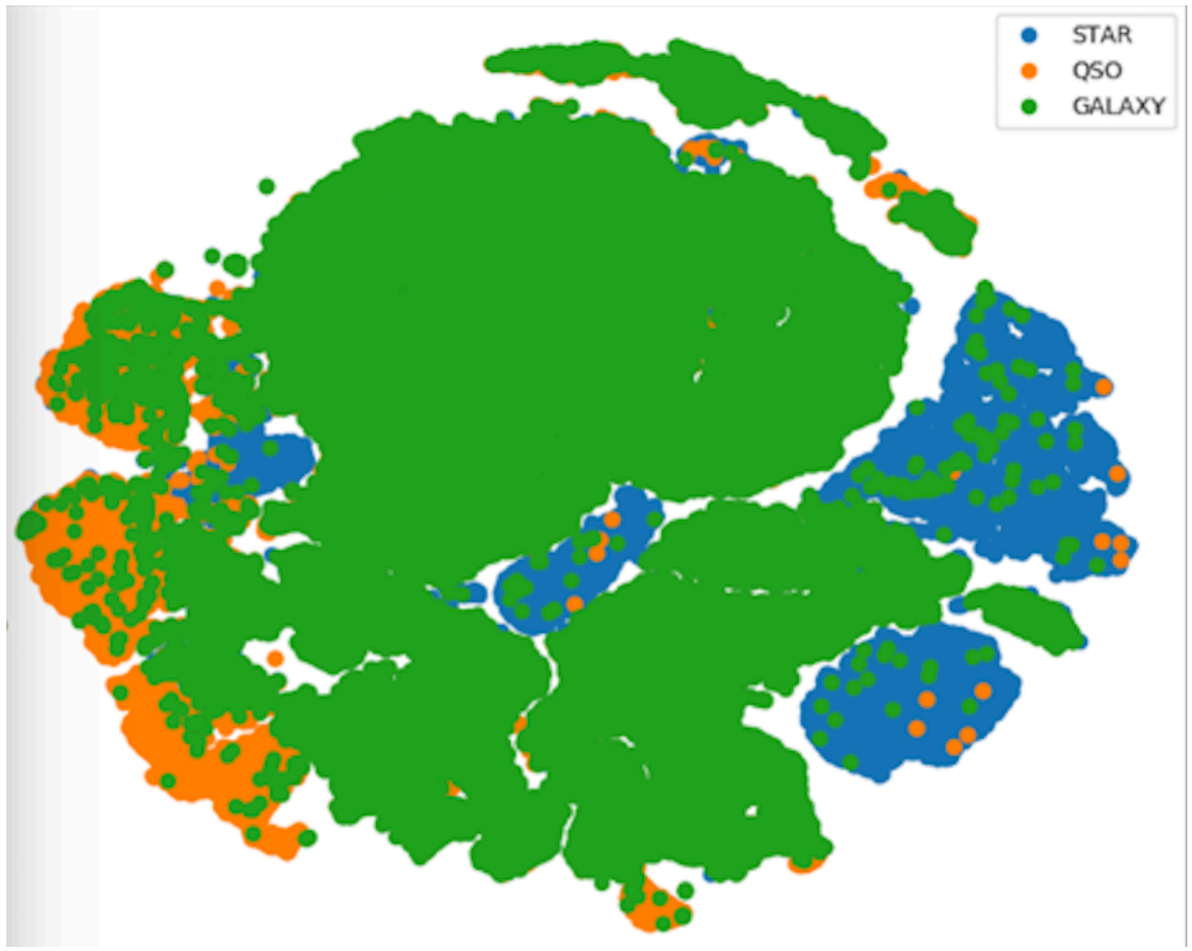
The more data available, the more complex and unique sources a computer can learn to recognise and describe





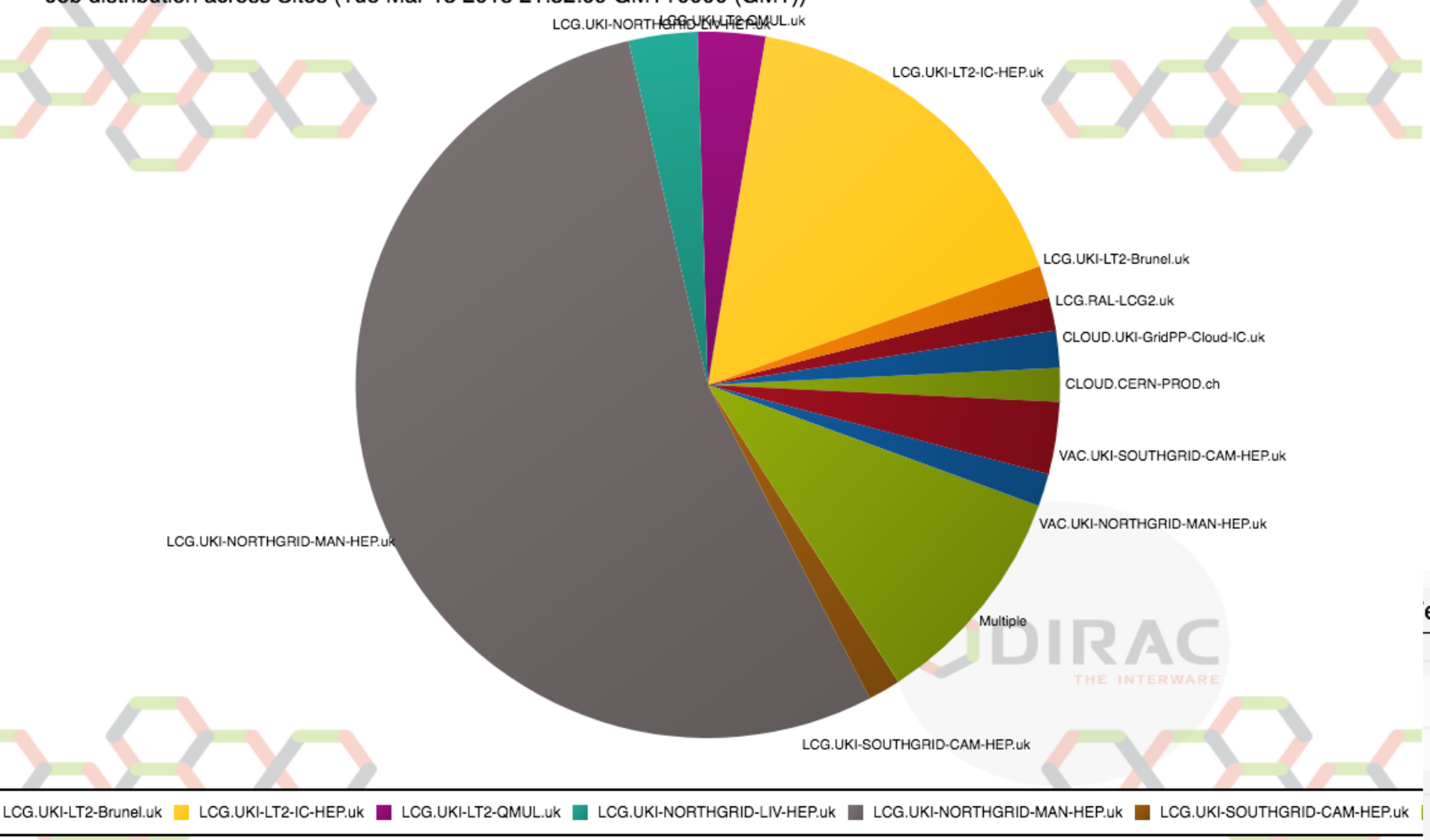
Clustering algorithms and decision trees learn what objects are from the data





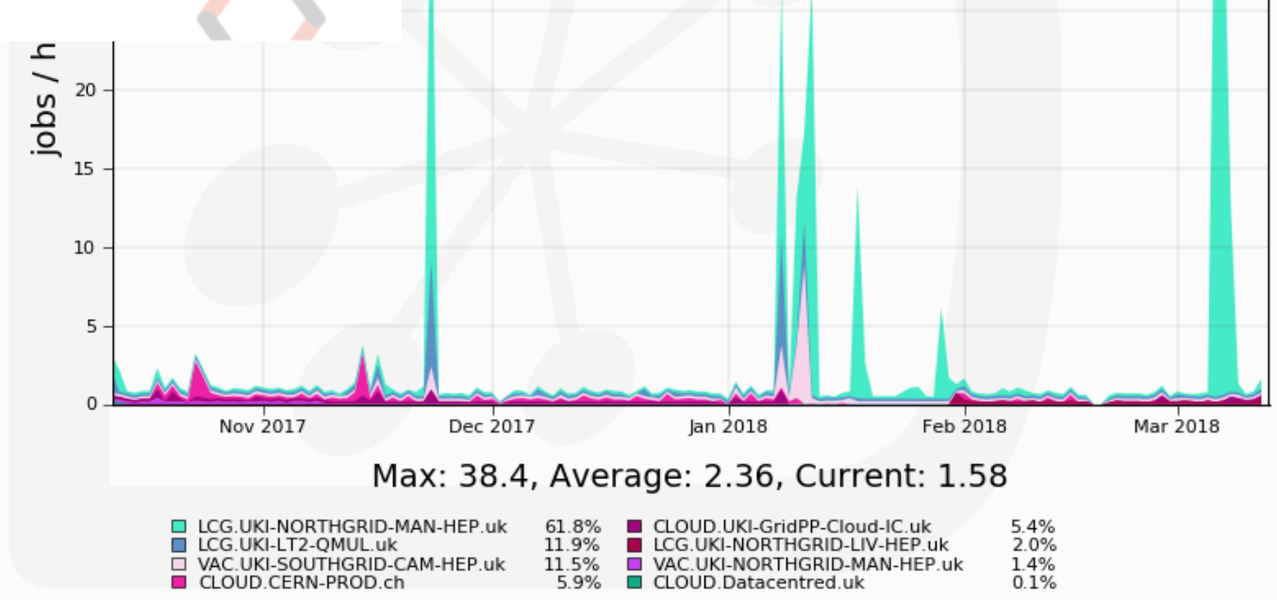
>1e6 objects - 16 photometry data points describing each from optical, IR, radio

Job distribution across Sites (Tue Mar 13 2018 21:32:09 GMT+0000 (GMT))



Jobs by Site

eeks from Week 41 of 2017 to Week 10 of 2018



Max: 38.4, Average: 2.36, Current: 1.58

LCG.UKI-NORTHGRID-MAN-HEP.uk	61.8%	CLOUD.UKI-GridPP-Cloud-IC.uk	5.4%
LCG.UKI-LT2-QMUL.uk	11.9%	LCG.UKI-NORTHGRID-LIV-HEP.uk	2.0%
VAC.UKI-SOUTHGRID-CAM-HEP.uk	11.5%	VAC.UKI-NORTHGRID-MAN-HEP.uk	1.4%
CLOUD.CERN-PROD.ch	5.9%	CLOUD.Datacentred.uk	0.1%

Generated on 2018-03-13 21:07:35 UTC

Next steps…

- Time domain compute model: identify critical elements & prototype;
- Memory restrictions - options;
- skatelescope DIRAC instance? Rucio+PANDA?