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

AENEAS: An SKA Regional Centre for Europe

Rohini Joshi
University of Manchester

rohini.joshi@manchester.ac.uk

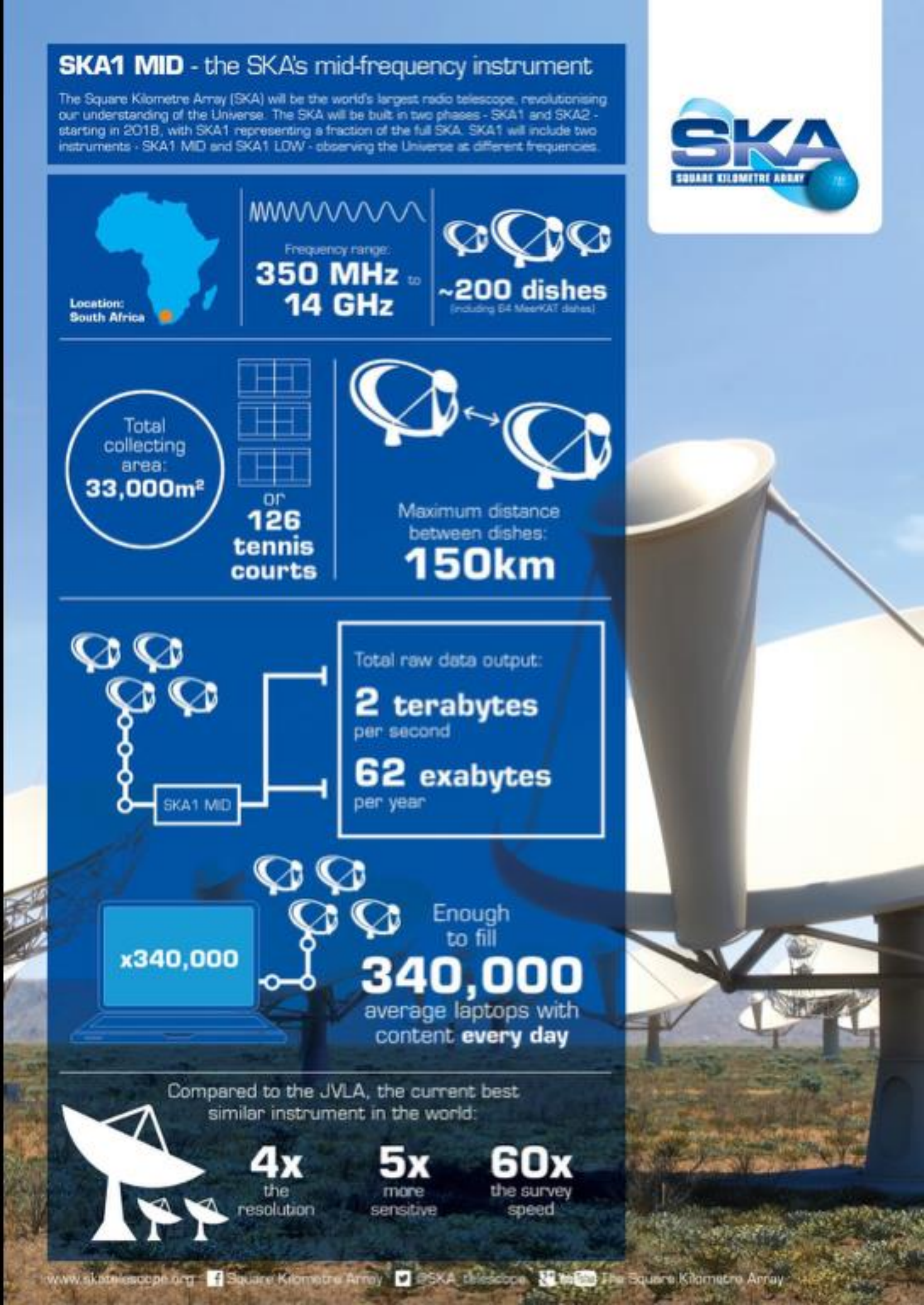
Develop a concept and design for a distributed, federated European Science Data Centre (ESDC) to support the astronomical community in achieving the scientific goals of the Square Kilometre Array

Outline

- What is the SKA?
- SKA Regional Centres
- AENEAS goals and objectives
- Data products at a RC
- Prototyping known use cases on GridPP
  - Calibration and Imaging
  - Image based object detection and classification
  - Classification using external archives
  - Pulsar timing

DISH ARRAY

SKA1 MID

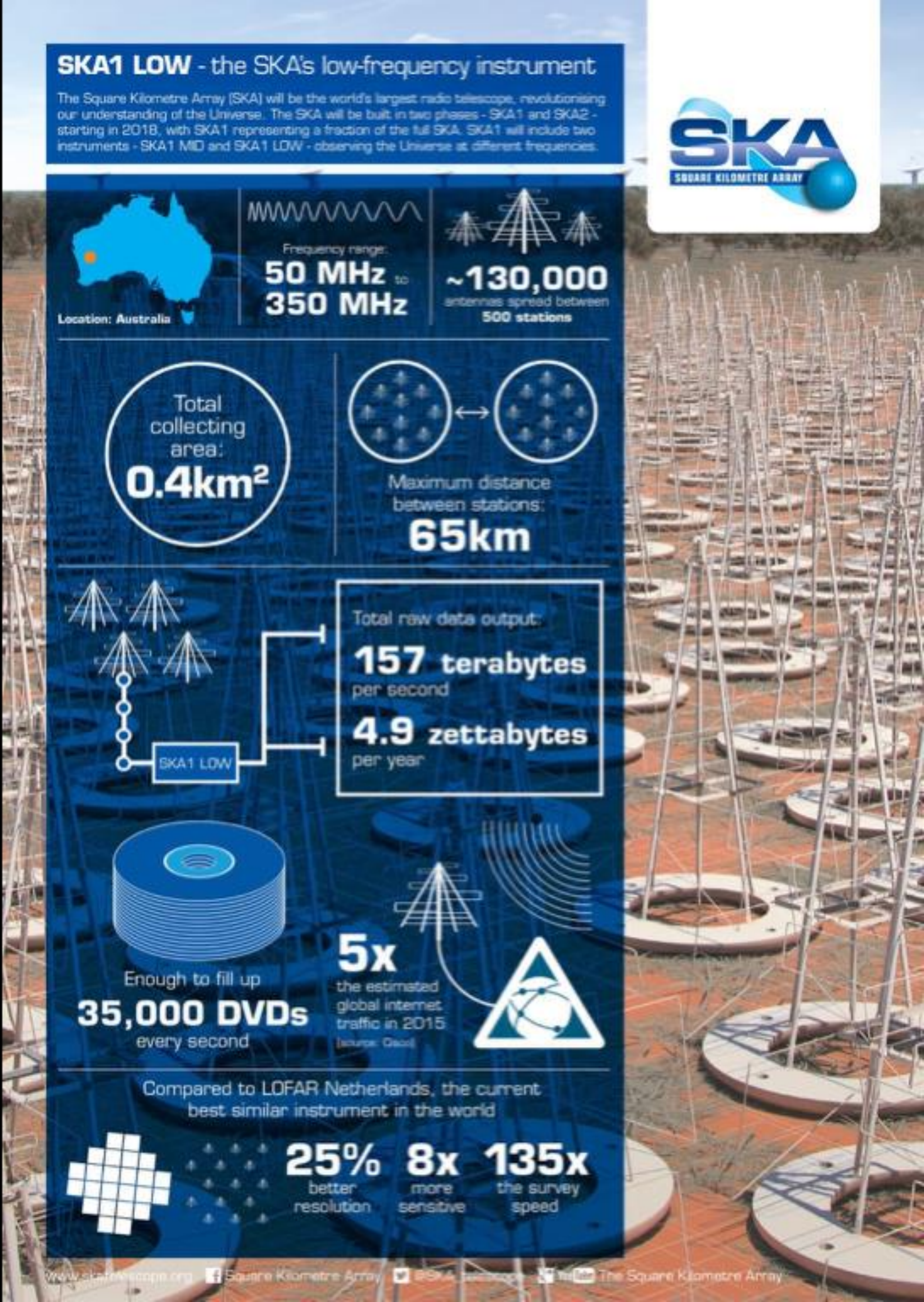


APERTURE ARRAY

SKA1 LOW

TERABYTE = 10^{12} BYTES

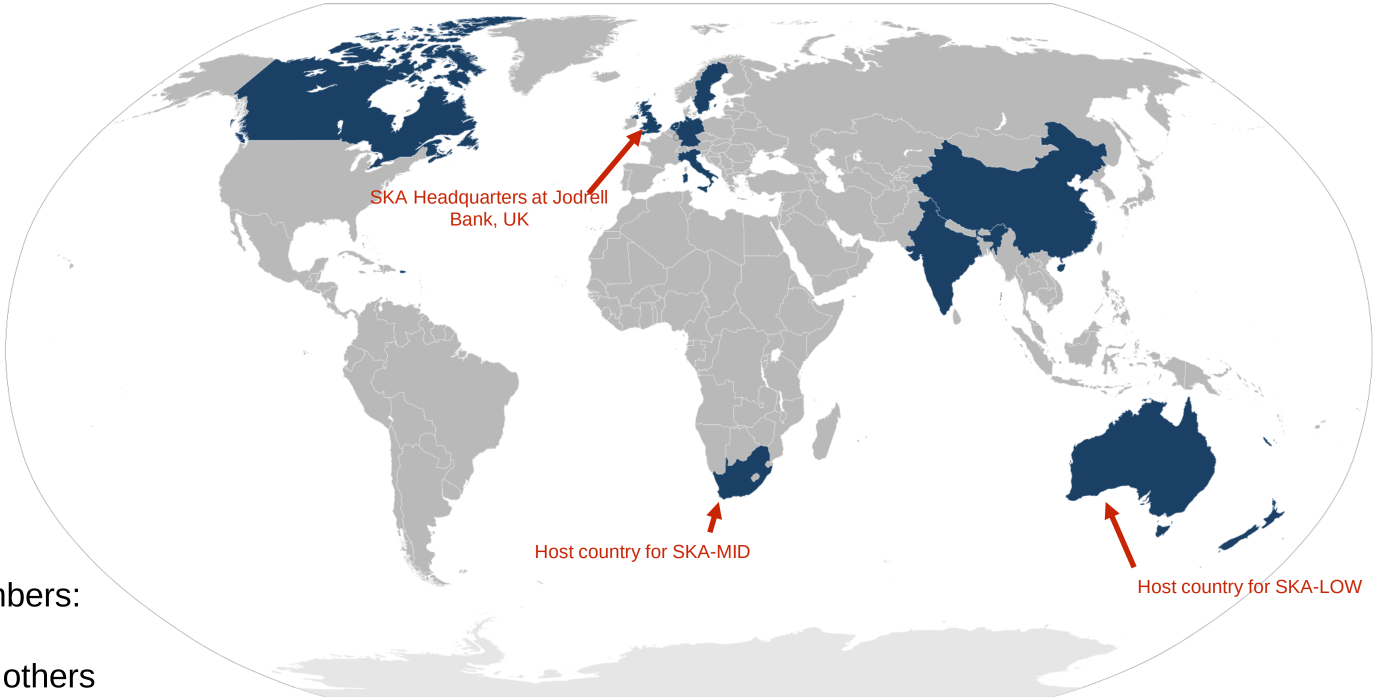
ZETTABYTE = 10^{21} BYTES



The Square Kilometre Array

- Australia
- Canada
- China
- India
- Italy
- Netherlands
- New Zealand
- South Africa
- Sweden
- UK

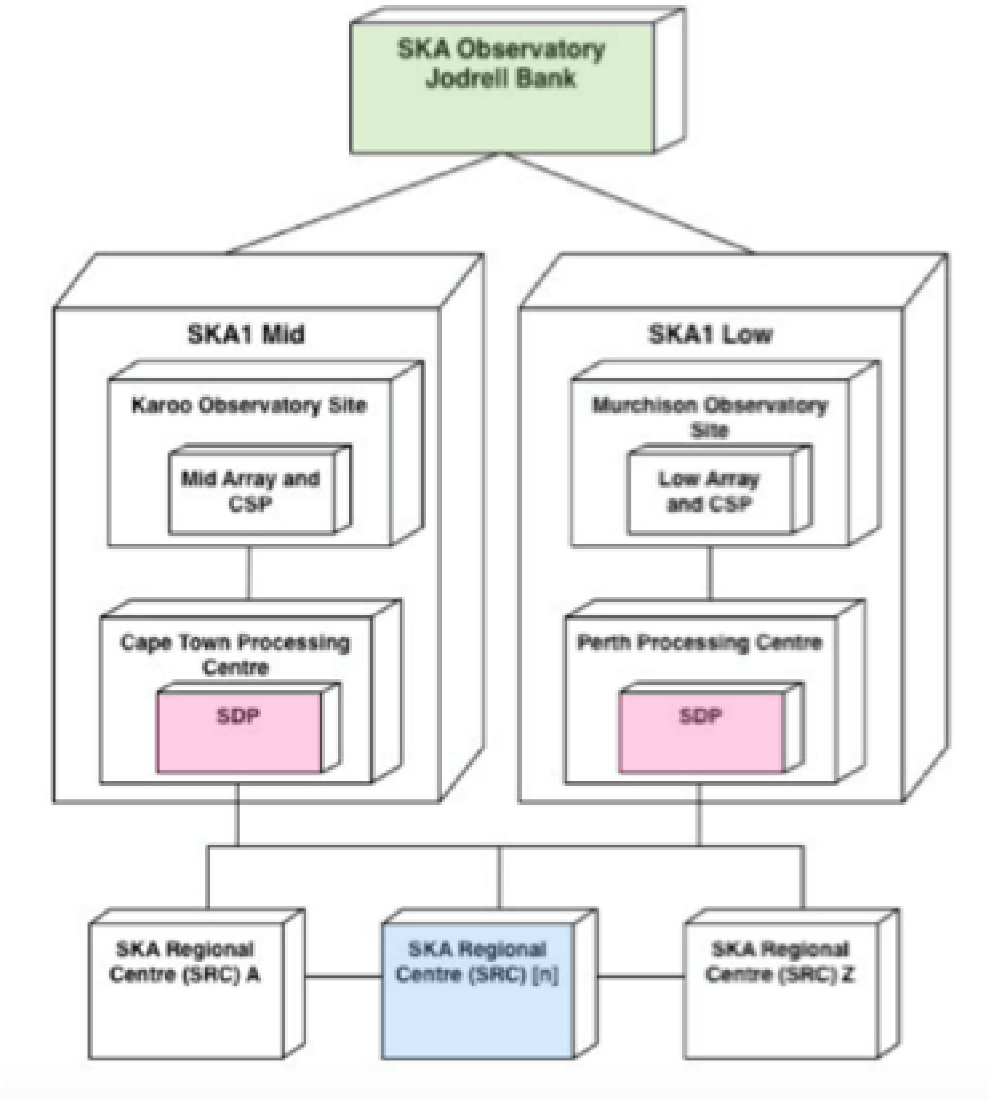
Potential new members:
Spain, Portugal,
Germany, France, others



SKA
SQUARE KILOMETRE ARRAY

SKA
SQUARE KILOMETRE ARRAY
SDP
SCIENCE DATA PROCESSOR

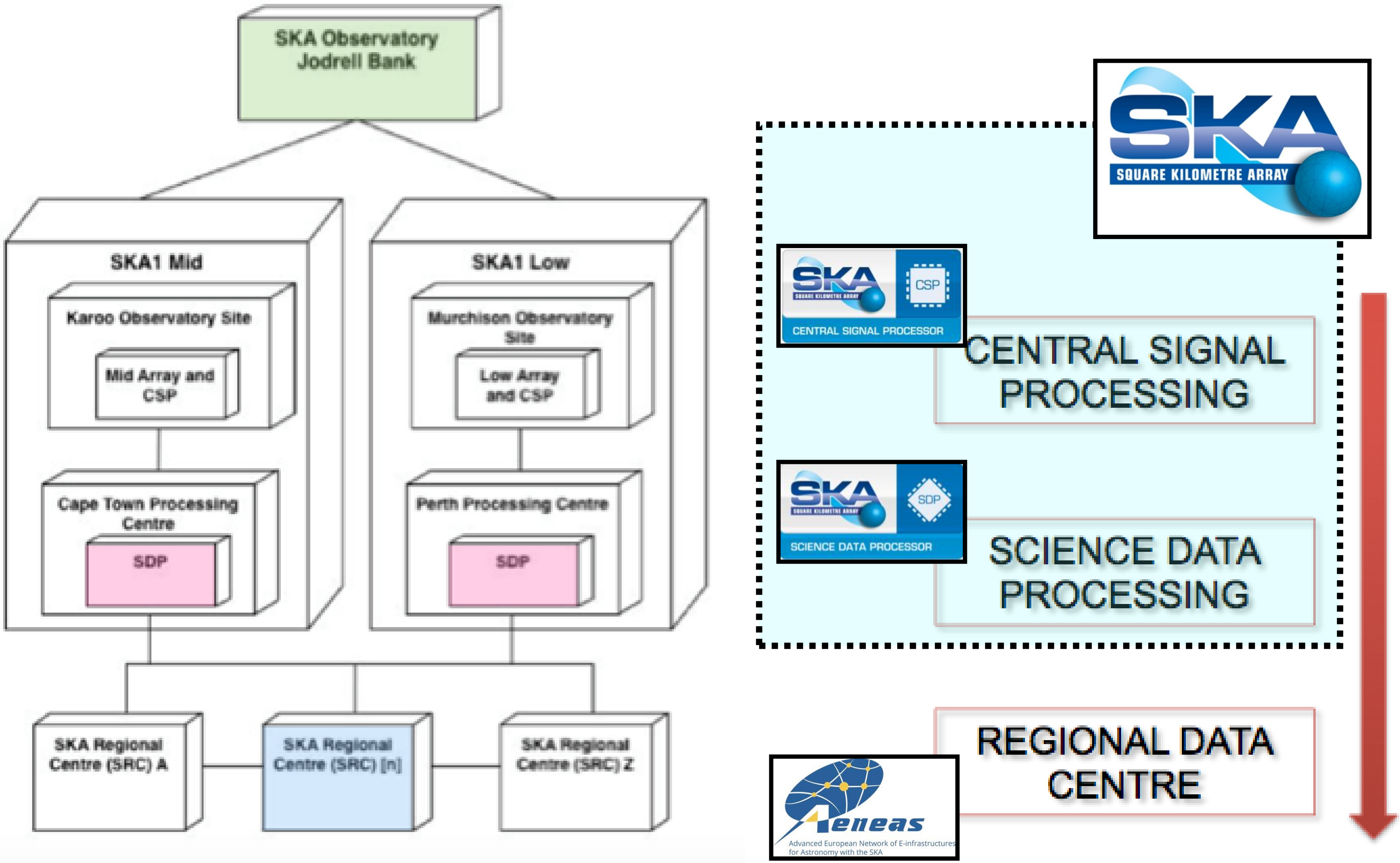
Aeneas
Advanced European Network of E-infrastructures
for Astronomy with the SKA

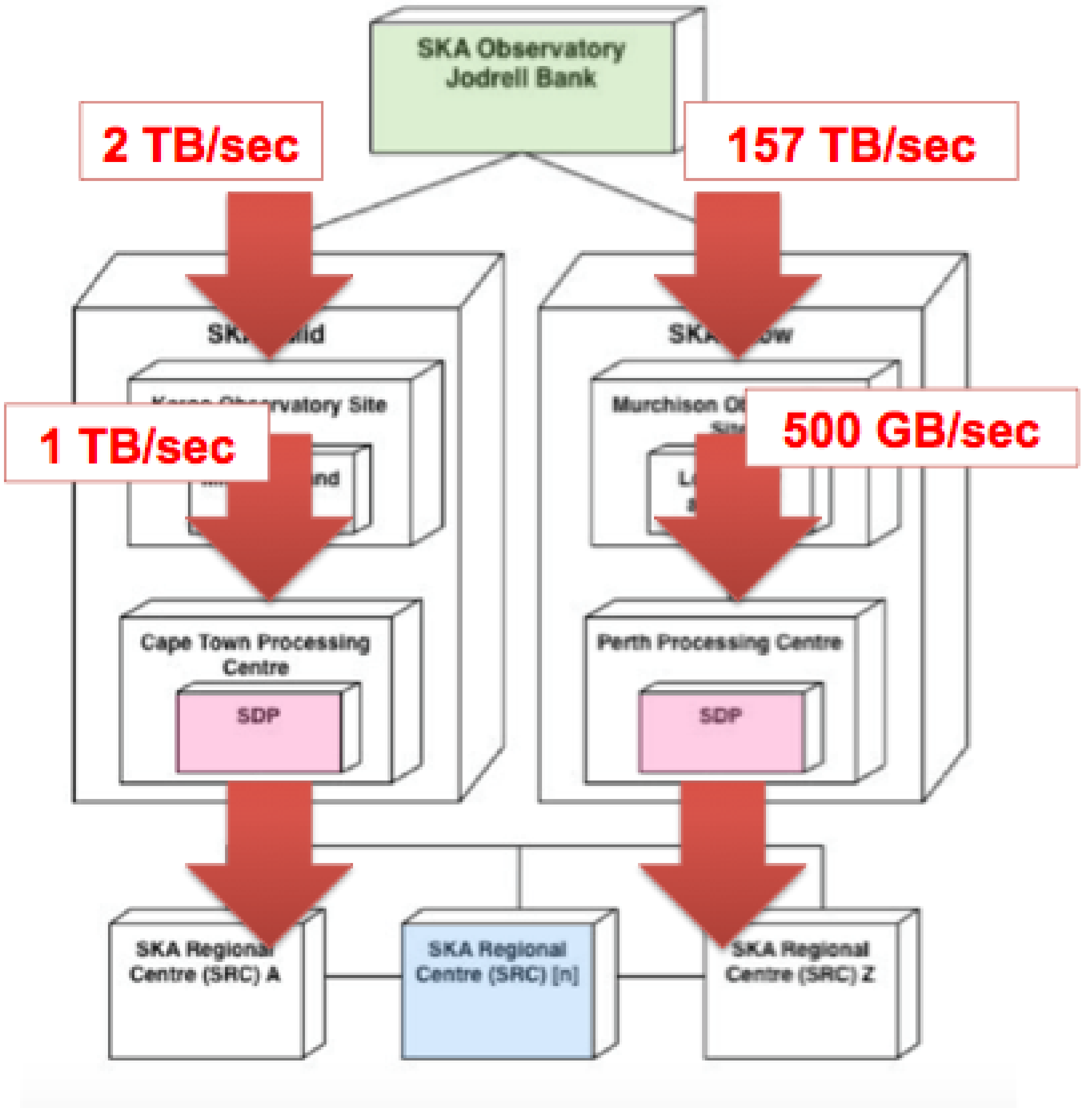


CENTRAL SIGNAL PROCESSING

SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE


SKA Observatory
Jodrell Bank
SKA1 Mid
Karoo Observatory Site
Mid Array and CSP
Cape Town Processing Centre
SDP
SKA1 Low
Murchison Observatory Site
Low Array and CSP
Perth Processing Centre
SDP
SKA Regional Centre (SRC) A
SKA Regional Centre (SRC) [n]
SKA Regional Centre (SRC) Z
SKA
SQUARE KILOMETRE ARRAY
CSP
CENTRAL SIGNAL PROCESSOR
CENTRAL SIGNAL PROCESSING
SDP
SCIENCE DATA PROCESSOR
SCIENCE DATA PROCESSING
REGIONAL DATA CENTRE
Advanced European Network of E-infrastructures
for Astronomy with the SKA



CENTRAL SIGNAL PROCESSING

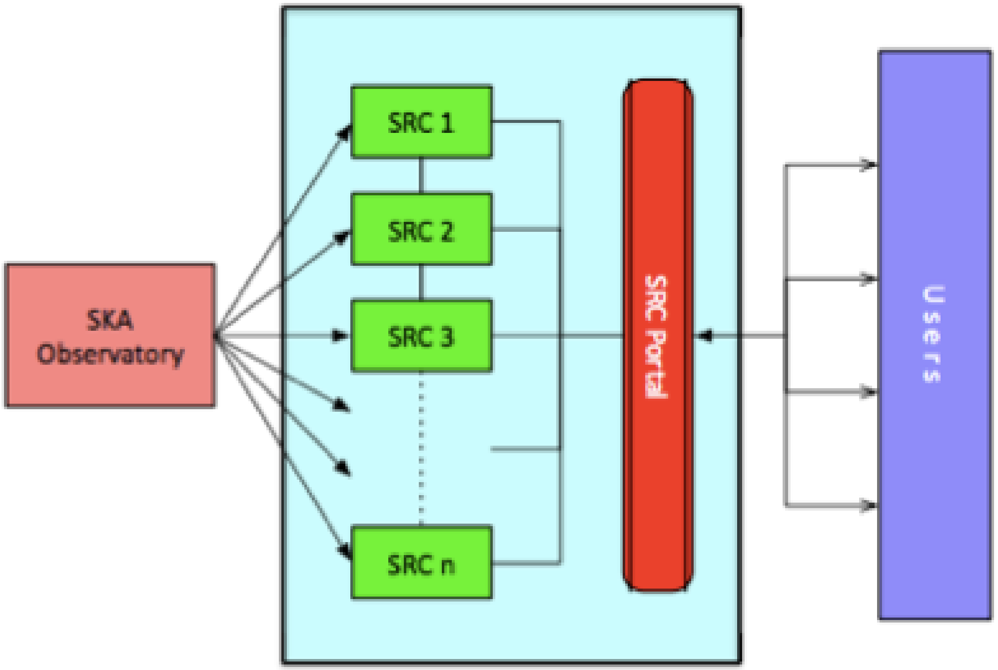
SCIENCE DATA PROCESSING

REGIONAL DATA CENTRE

Future SKA Science Archive

SKA Regional Centres

- Provision long-term SKA Science Archive
- Provide access and distribute data products to users (key science projects and PIs)
- Provision and Management of computational resources for post-processing
- Provide platform for continued development of software
- Multiple regional SRCs, locally resourced but interoperable with common core functionality



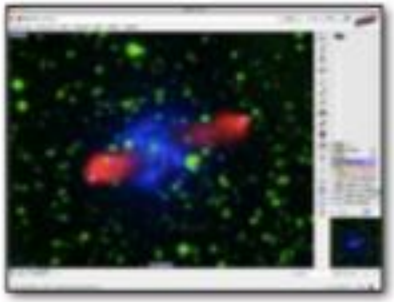
Joint SKAO/SRC functions

- User support for SKAO data products
- User support for SKAO provided software and tools
- Distribution of SKA data packs to users (SDP or SRC)

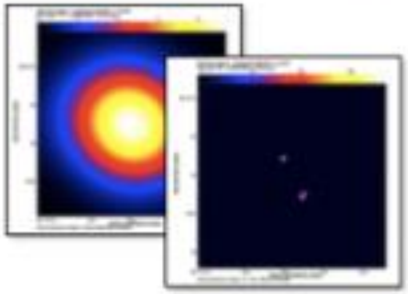
Regional Centre Functionality

Data Discovery

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



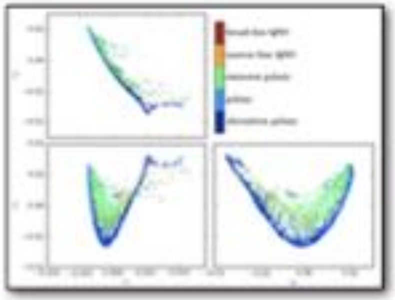
Data Processing



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

Data Mining

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



RUCIO

DIRAC
THE INTERWARE

Objectives

- Determine use cases, estimate compute and storage requirements with input from the Science Working Groups

  Solar, Heliospheric & Ionospheric Physics
  Cosmology
  Epoch of Reionization
  Extragalactic Continuum (galaxies/AGN, galaxy clusters)
  Cradle of Life
  HI galaxy science
  Magnetism
  Pulsars
  Transients

- Prototyping regional centre activities by mapping pilot compute models onto available resources
- Identify potential bottlenecks due to resource mapping or existing pipelines that do not scale well
- Commonly used tools are CASA, AIPS, Miriad, PRESTO, SIGPROC
  However we are now leaning towards containerized approach to maintain reproducibility
- Finding the balance between keeping stakeholders happy and keeping resource allocation simple

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
- Users consuming data will also be generating secondary data which may not be smaller than raw data

Prototyping known use cases: Calibration and Imaging

- Using LOFAR data since LOFAR is a pathfinder instrument
  GOODS-North survey data on DFC, 3.5 TB per observation
- CVMFS LOFAR software installation used, but soon moved to a singularity image
- Ran into memory limitations very quickly, ~ 2 GB per job slot is not enough

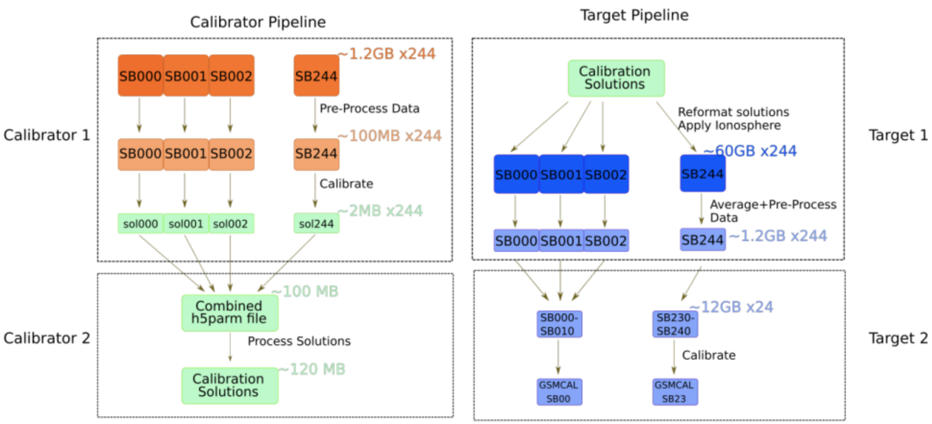
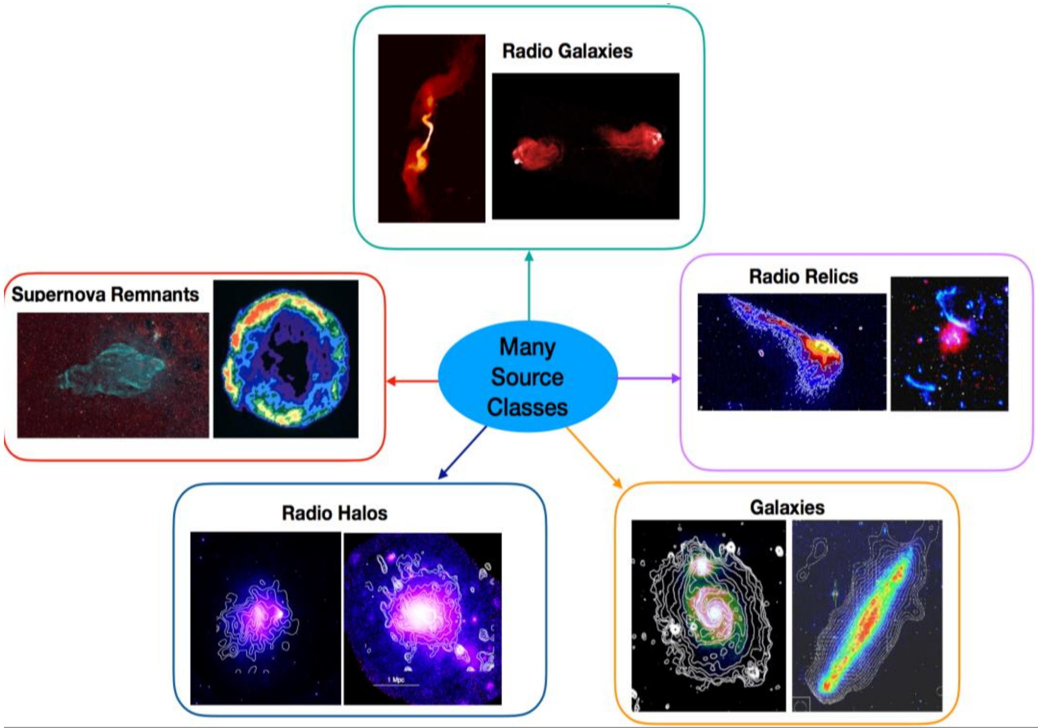


Image credit to Mechev et al. 2018

This approach will not only relax the memory requirements of the pipeline as a whole but also lend it self very well to the Transformation system

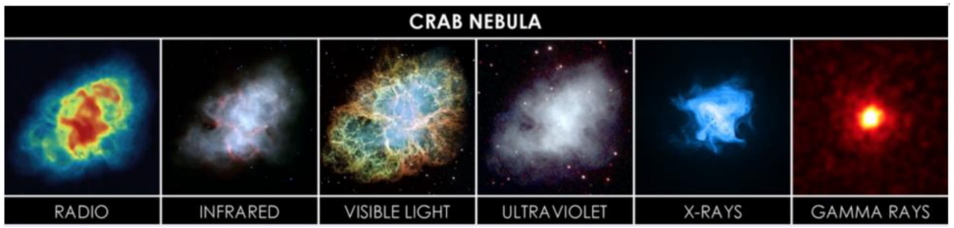
Prototyping known use cases: Image-based Object Detection & Classification

- Convolutional Neural Networks
- ~ 2 GB per job slot is not enough, few 10s of GB needed
- Once region of interests are identified, classification can be largely parallelized



Prototyping known use cases: Classification using external archives

- Distribute SKA source catalogues and cross-match with external multi-wavelength archives
- Gather results into one master catalogue with all available information, ready for machine learning



Supervised Learning:

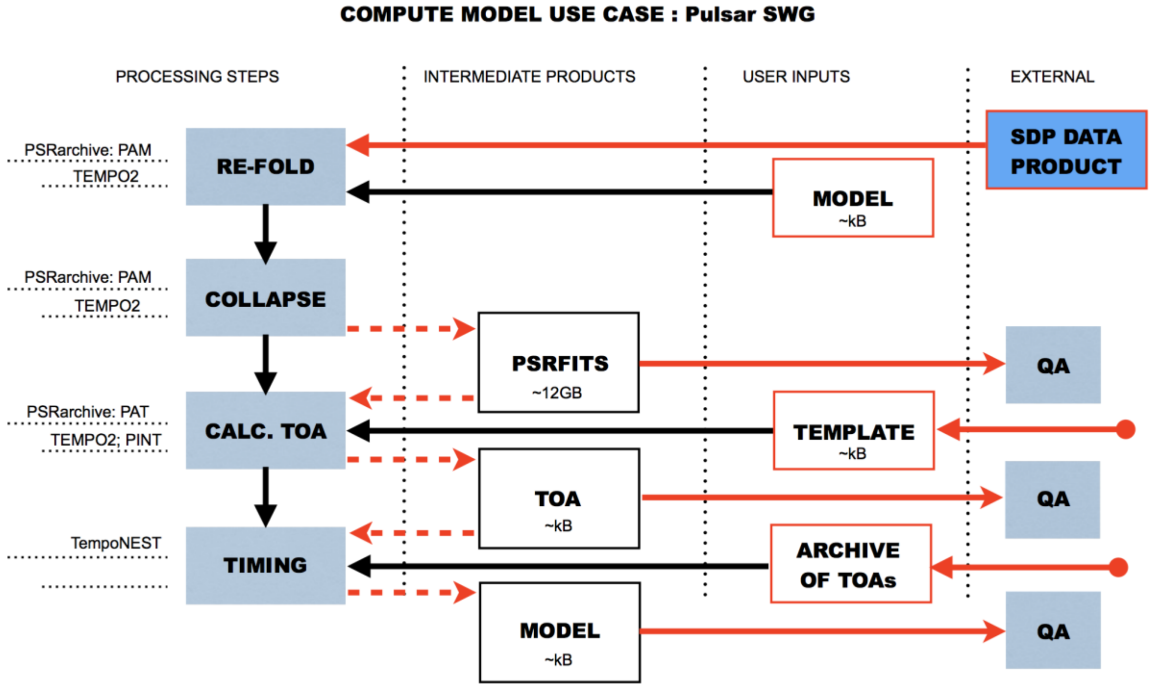
- Train machine learning algorithms on sub-sets of data with excellent external archive coverage
  (typically sources already detected before at multiple wavelengths)
- Gather and combine models from distributed runs to optimise them
  (model accuracy is typically higher when data is not distributed, but can be worth the compute trade-off)
- Distribute the trained model to classify source types on sub-sets of data with poor external archive data
  (typically brand new sources, or old sources with poor multi-wavelength data coverage)
- Opportunity to classify many sources that would otherwise go unnoticed

Unsupervised Learning:

- Clustering algorithms like T-SNE’s can automatically classify source types with no prior labeled info
- But can be very computationally intensive – Needs to be run on sub-sets of data in a distributed way.

Prototyping known use cases: Pulsar timing

- Time domain re-folding
  Lower memory and compute requirements
  Finding appropriate test data




1824