South Africa:
SKA Regional Centre Activity

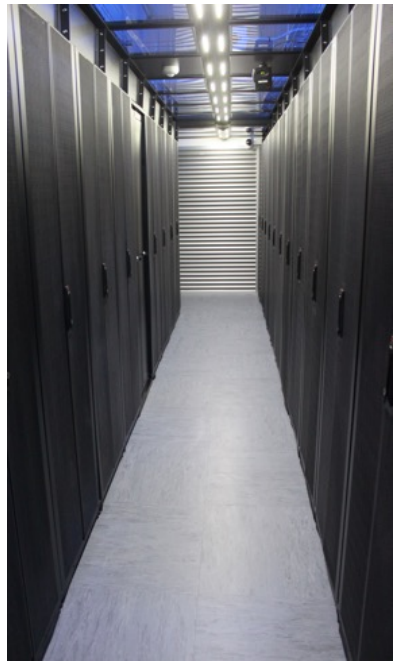
Dr. Rob Simmonds
Associate Director
IDIA

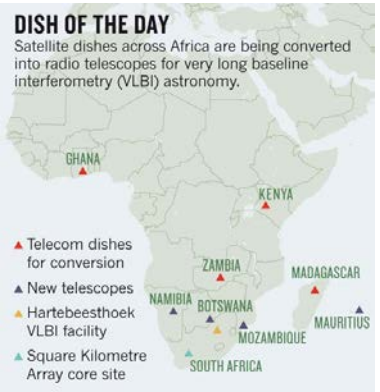
Overview

- MeerKAT RC / Tier 2 facilities
- SDP DELIV
  - Data delivery architecture for the SKA
- Data distribution and data processing for MeerKAT data
  - Collaboration between IDIA, ASTRON and SKA-SA
- SKA-SA activities
  - MeerKAT Commissioning
  - Archiving MeerKAT data
  - Pipelines / Machine learning
- CyberSKA portal and visualization activities

ARC / IDIA / Tier 2 systems

- Hardware deployed at UCT and NWU to act as MeerKAT Tier 2 system
- Currently 48 compute nodes (1500 cores) / 1.5PB storage
  - Includes 8 P100 GPUs for experimentation
  - 50Gb/s Open Ethernet core network
  - Managed using OpenStack IaaS framework
- RFP for additional compute and storage about to be issued
- Additional systems being deployed at UWC, UP and Wits will be used for distributed prototyping
- Exploring expansion of ARC to AVN / SKA partner countries





IDIA storage

- Using CEPH for cloud storage
  - Ephemeral storage to VMs
  - Block storage for creating small / mid-sized file system volumes
  - Object storage for scale out applications
- CEPH 3x replication but could reduced duplication using software erasure coding on current systems (on SAS backed volumes)
- Have BeeGFS for large data volumes
  - Higher performance (+)
  - Less data replication (using RAID 6) (+)
  - System level authorization (-)
  - Provide home, data and scratch BeeGFS volumes to trusted OpenStack hosted VMs

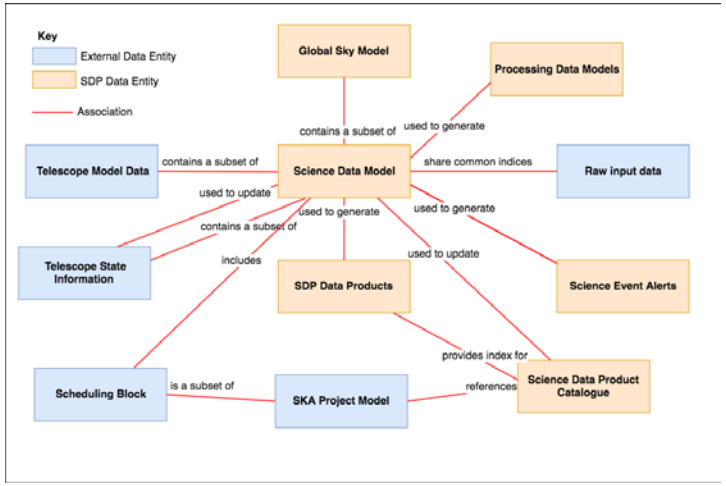
IDIA POSIX storage access

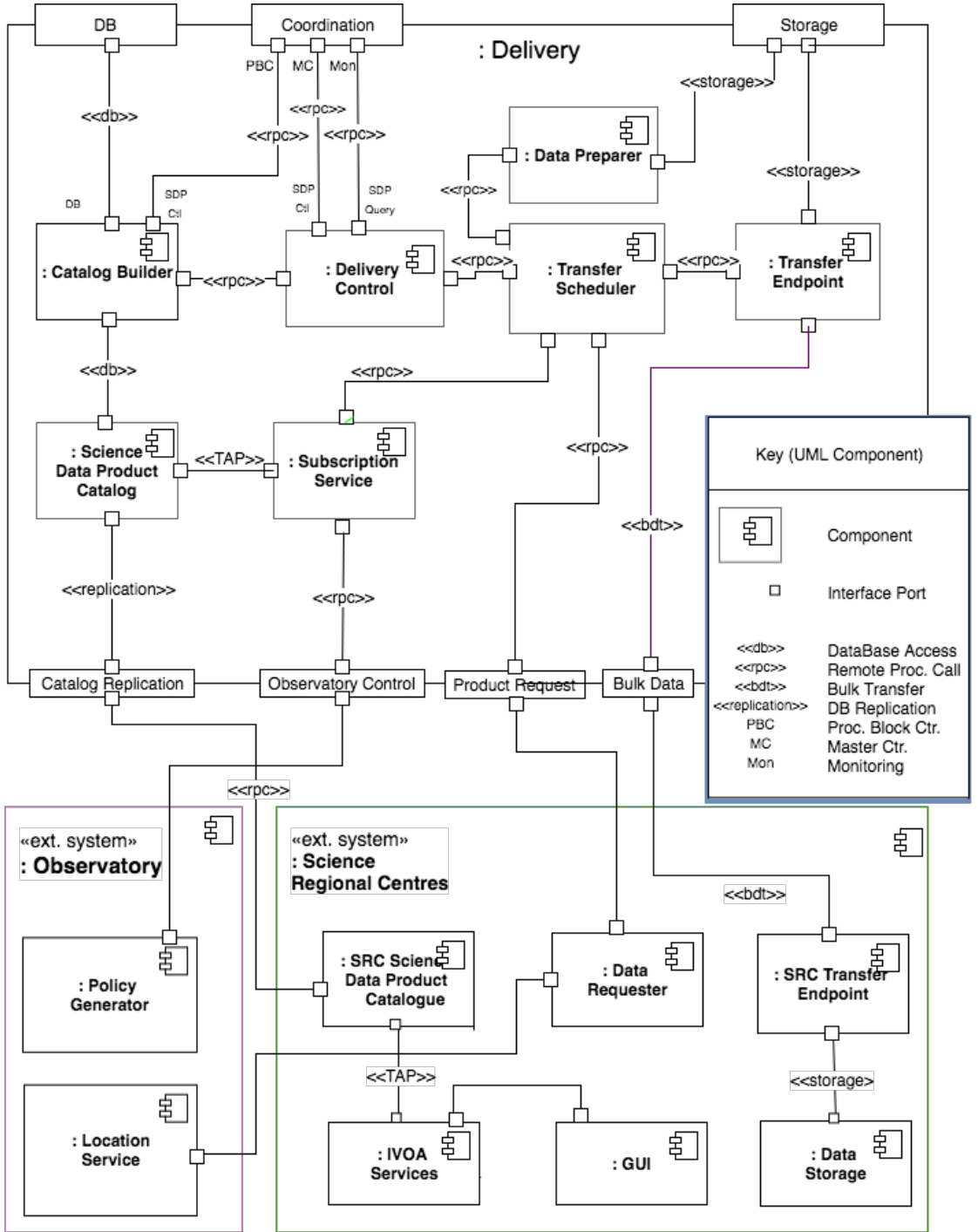
- Access to shared POSIX volumes from trusted machines
  - I.e., users don't have sudo access on these machines
- Users execute applications Singularity containers
  - Repository of application containers being created
- Users can create their own containers on non-trusted VMs
  - These can access shared POSIX volumes using SSH-FS

- Note: moved from Docker to Singularity containers due to concerns about privilege escalation

SDP Delivery Architecture

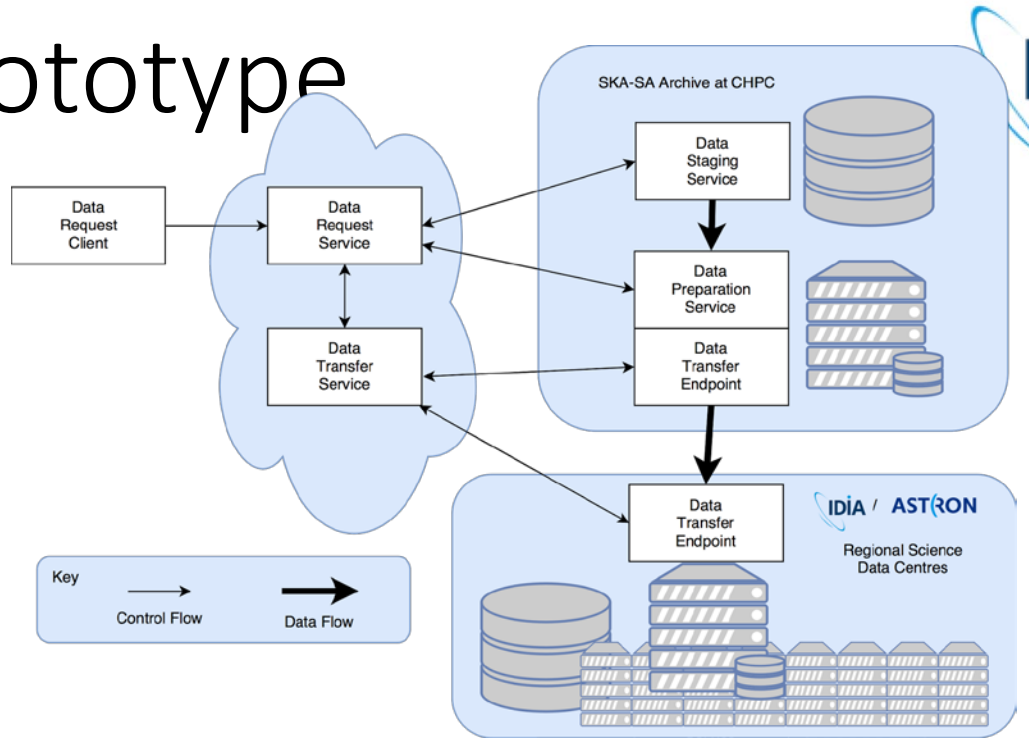
- Refining the delivery architecture to fit with revised SDP architecture
- Includes the SDP SRC interface
- Partners from UCT, CADC, SKA-SA/SAC, ASTRON, IAA and Oxford
- Still needs some thought about the what is needed in Science Data Product Catalog at SRCs

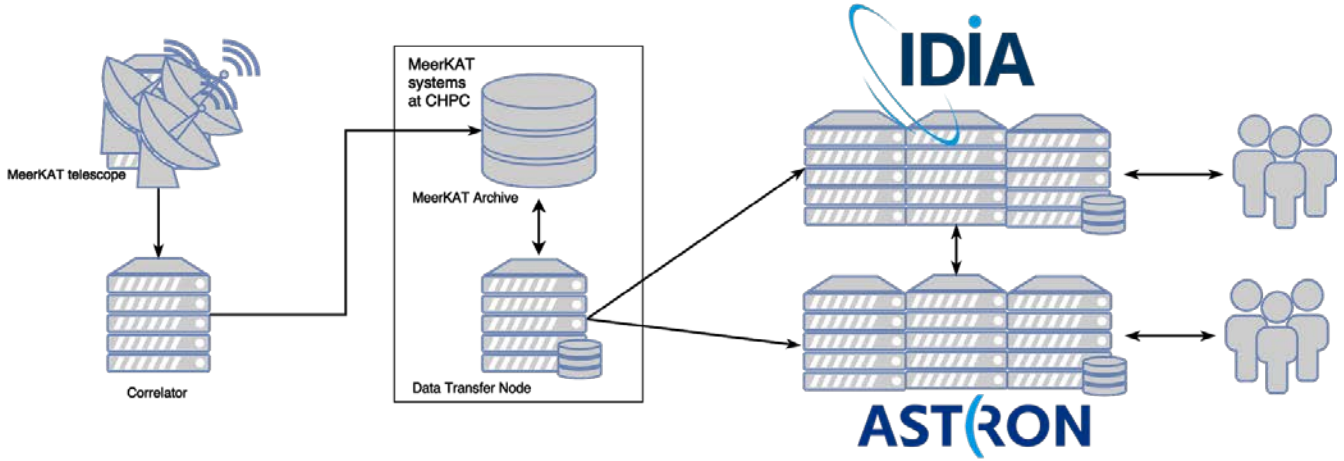




Data transport prototype

- Deployed as part of DOME project
- Utilizes CERN's FTS to control network transfers
- Queuing system used to stage data from Archive and to keep network loaded with high priority transfers
- Implements Request, Prepare and Deliver components of SDP architecture
- Deployed at SKA-SA/CHPC, IDIA and ASTRON





IVOA service prototyping

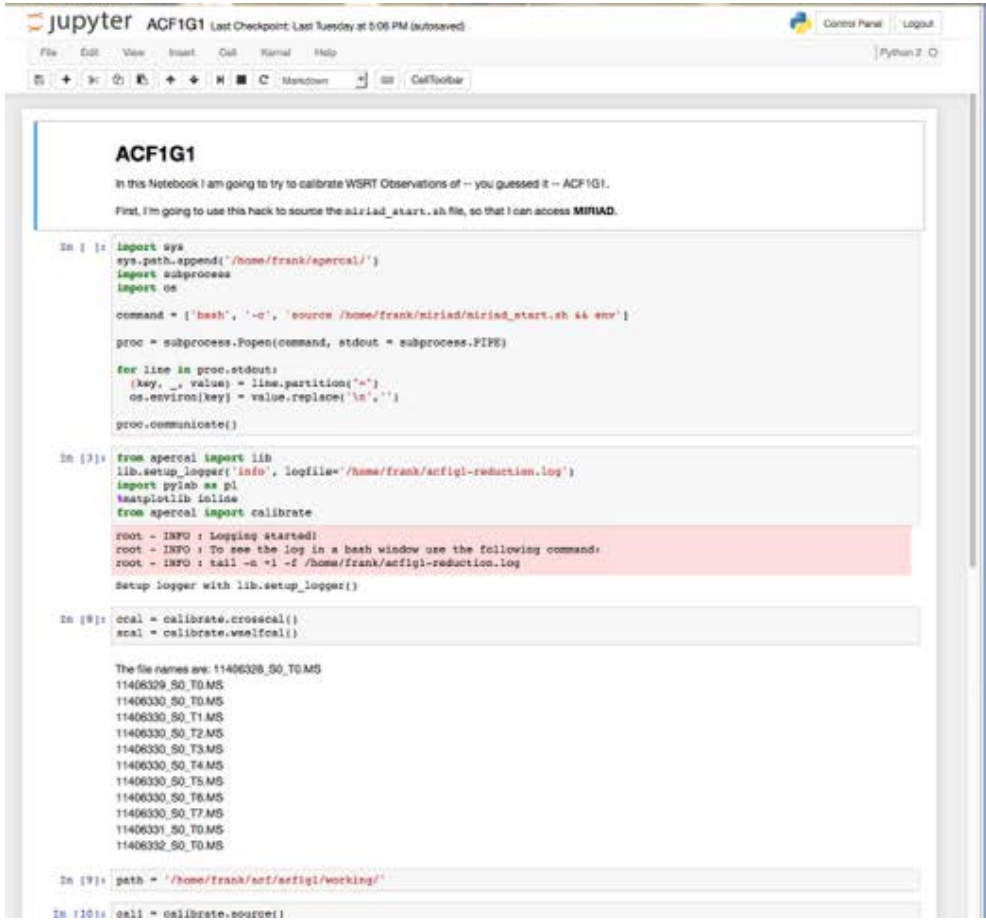
- Have access to sandbox services running at CADC
- SKA-SA providing access to MeerKAT metadata
- Mapping this metadata to COAM2 data model for use with CADC services
- Will provide prototype for SDP Query service once data modelling is complete

Pipelines

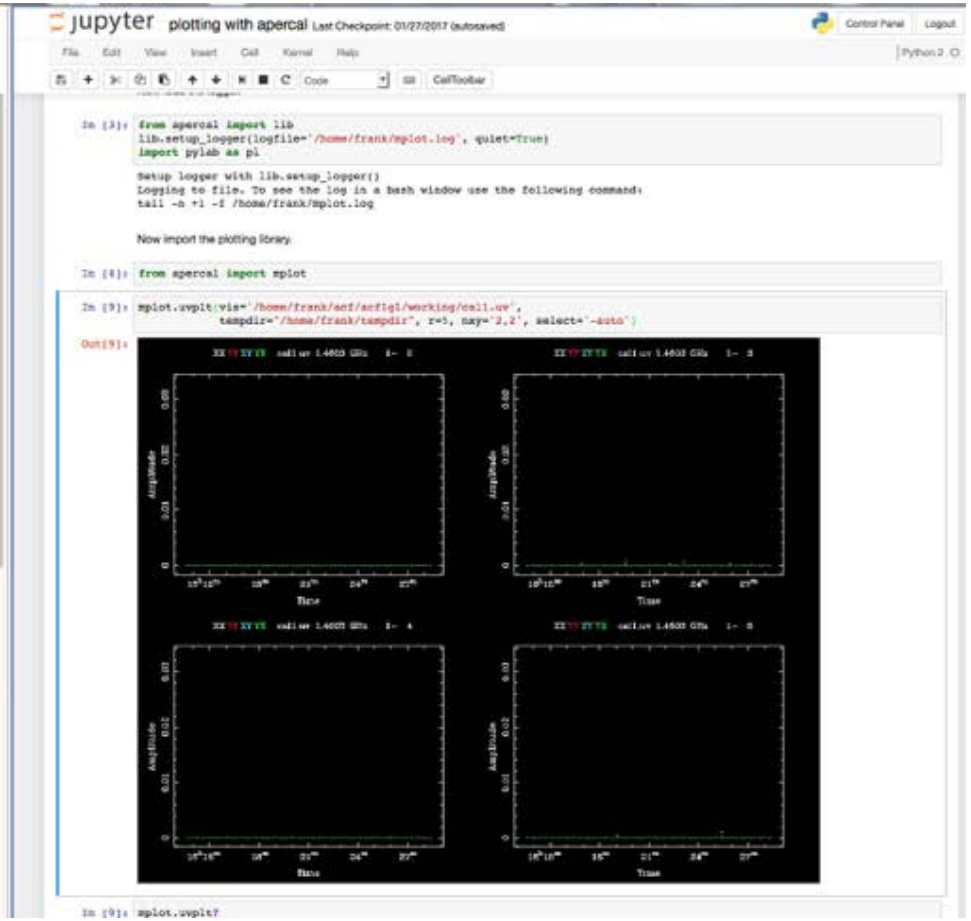
- Singularity container with:
  - CASA, drive-casa, python libraries, Jupyter Notebook & Hub.
- Continuing to develop software framework for running pipelines in Notebook
- Adopted code from JIVE project that was more robust than our initial prototype
- Being combined with data transport system to provide automated archive to product execution
- Adding DAGman/HTCondor and automated interaction with Git to produce reproducible workflows

Current Notebook Dashboard on ARC

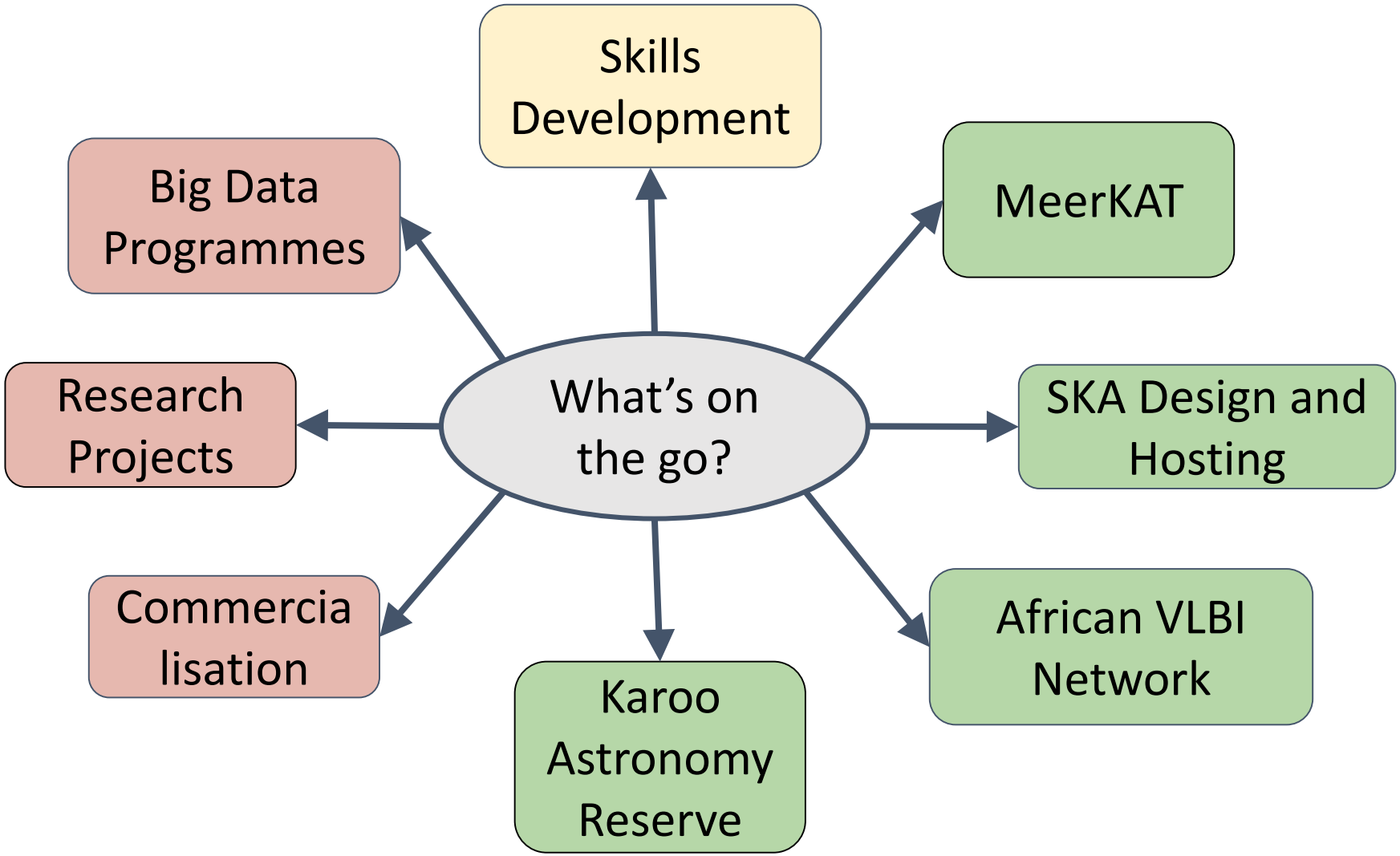
Pipeline Control &
Data Exploration



Visualization

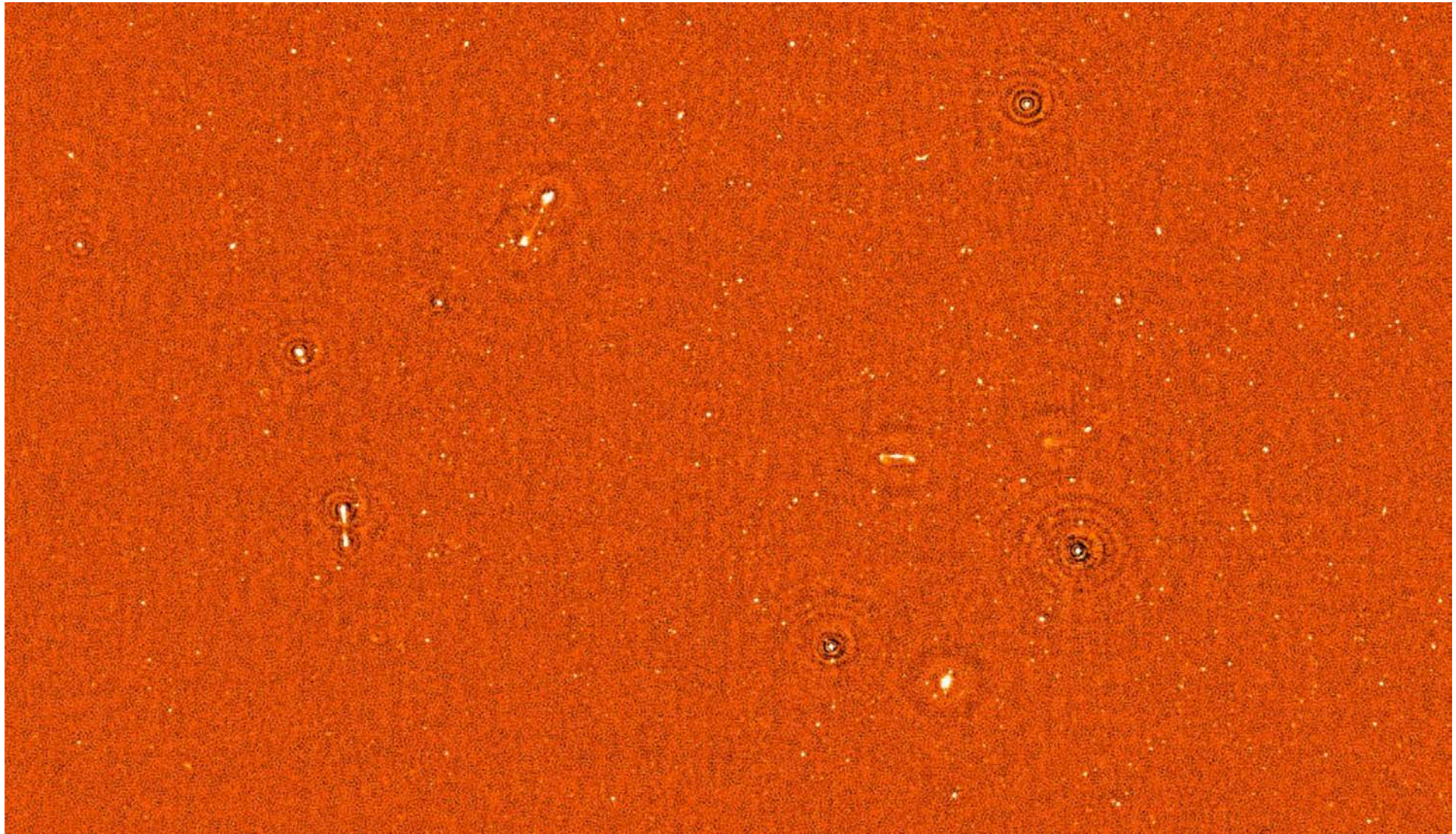


SKA SA - Programmes



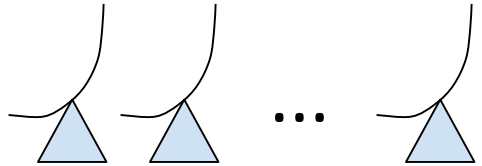
SKA AFRICA
SQUARE KILOMETRE ARRAY

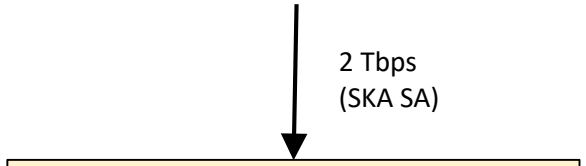
MeerKAT data processed at IDIA

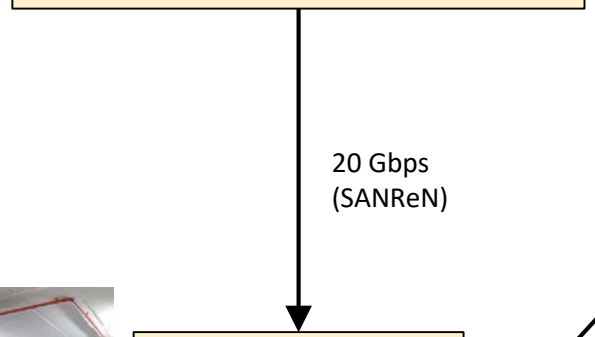


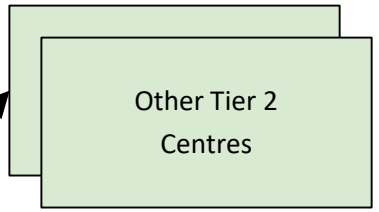
MeerKAT Processing and Archive

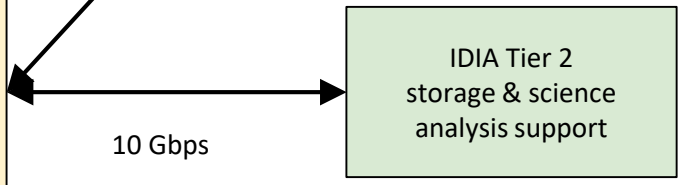
- Currently have storage in place in Carnarvon and CHPC
- Arriving today:
  - 200 TFLOPS compute
  - 10PB tape library
- Will have 20PB raw disk storage by Feb 2018 (~10TPB now)











SKA1 Era Regional Centres

197 dishes

20 Tbps
(SKAO)

Karoo Array Processor Building
(SKA CSP - SKAO controlled)

Local Tier 2's
(SA controlled)
e.g. IDIA, Sol Plaatje, etc.

2-10 Tbps
(TBD)

International SKA Regional Centres
(foreign controlled)
e.g. NL, China, Italy, etc.

100 Gbps
(TBD)

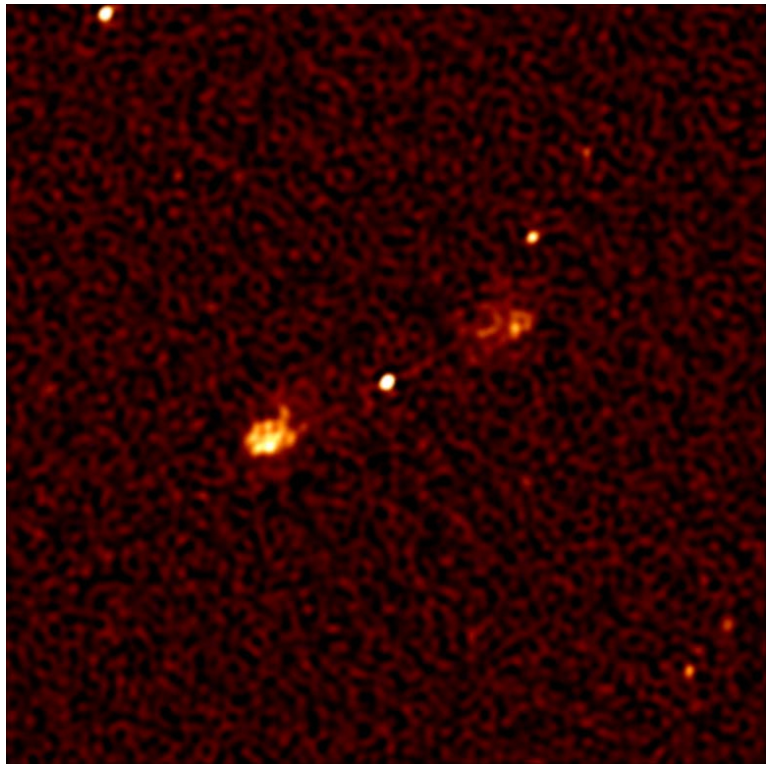
SKA Science Data Processor and Archive
(SKAO controlled)
~60 racks
(Cape Town)

SA SKA Regional Centre
(SA controlled)
(compute and archive)
~ 20 racks

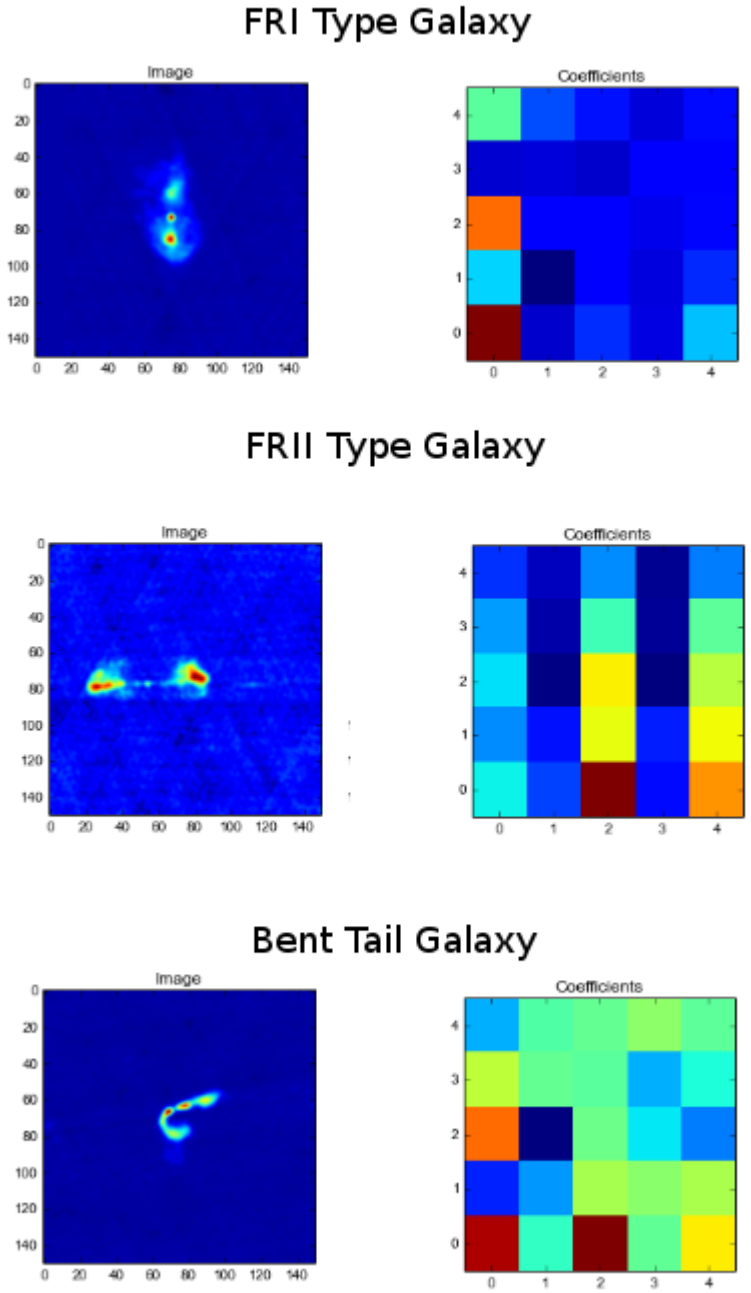
SKA AFRICA
SQUARE KILOMETRE ARRAY

Galaxy Morphology

Arun Aniyan



Shapelet and Deep Learning approaches under development



SKA AFRICA
SQUARE KILOMETRE ARRAY

Pulsar Candidate Selection

Image: NASA

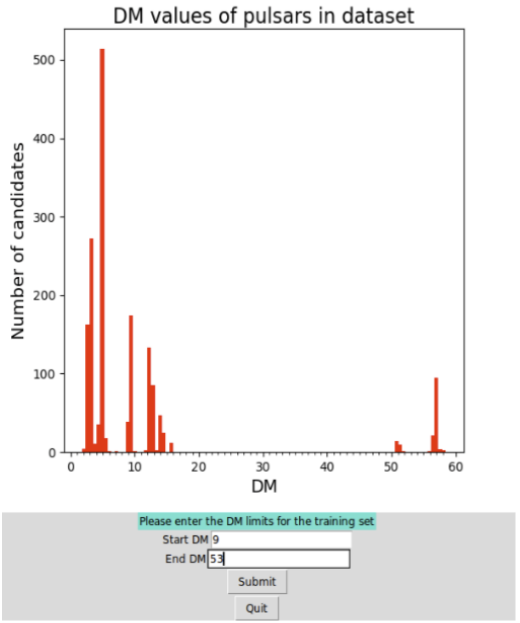
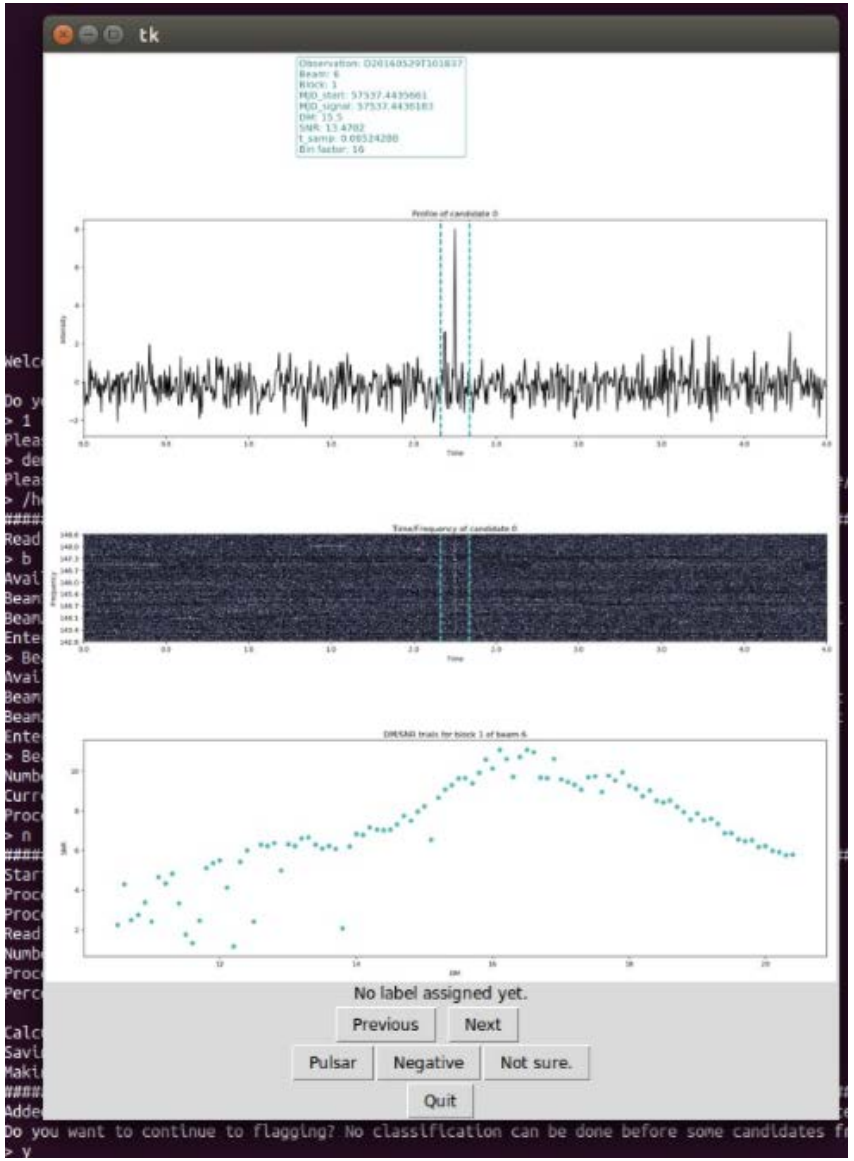
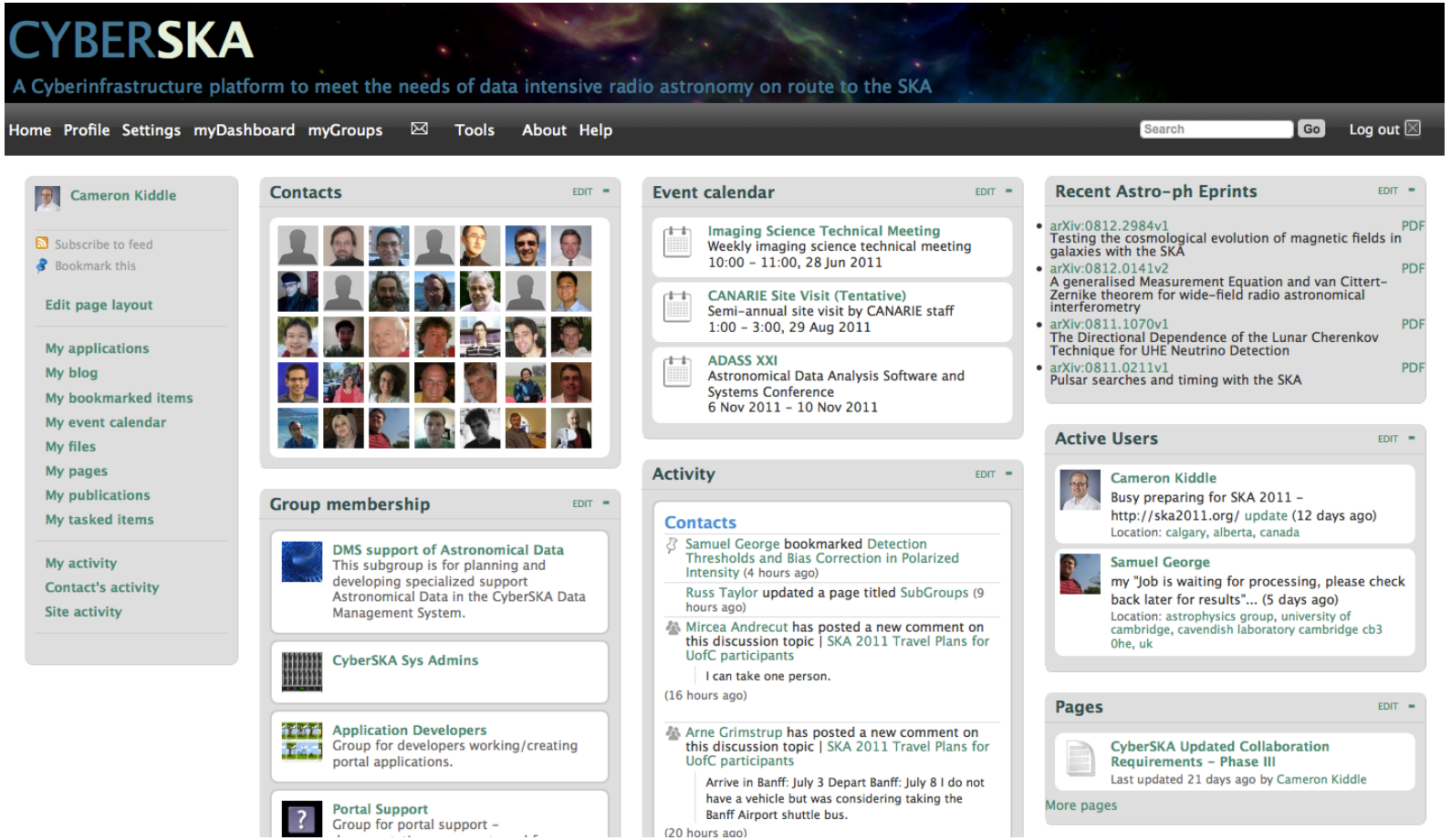


Illustration 5: Training / testing split interface;



SKA AFRICA
SQUARE KILOMETRE ARRAY

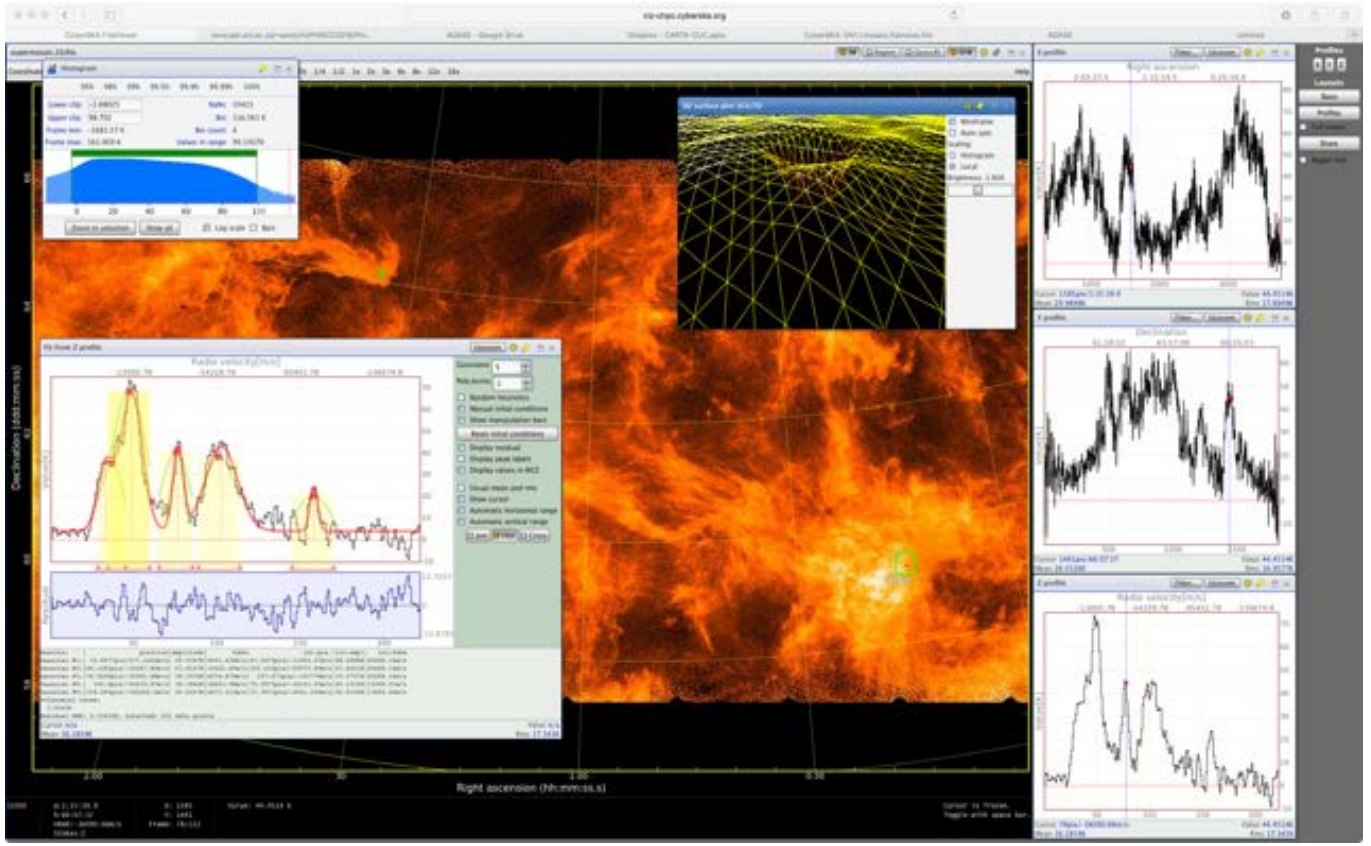
CyberSKA Portal / Gateway



- Over 700 users
- Latest version enables federation of multiple portals
- iRods used for data management
- Provides access to data sharing, collaboration, visualisation and data search tools

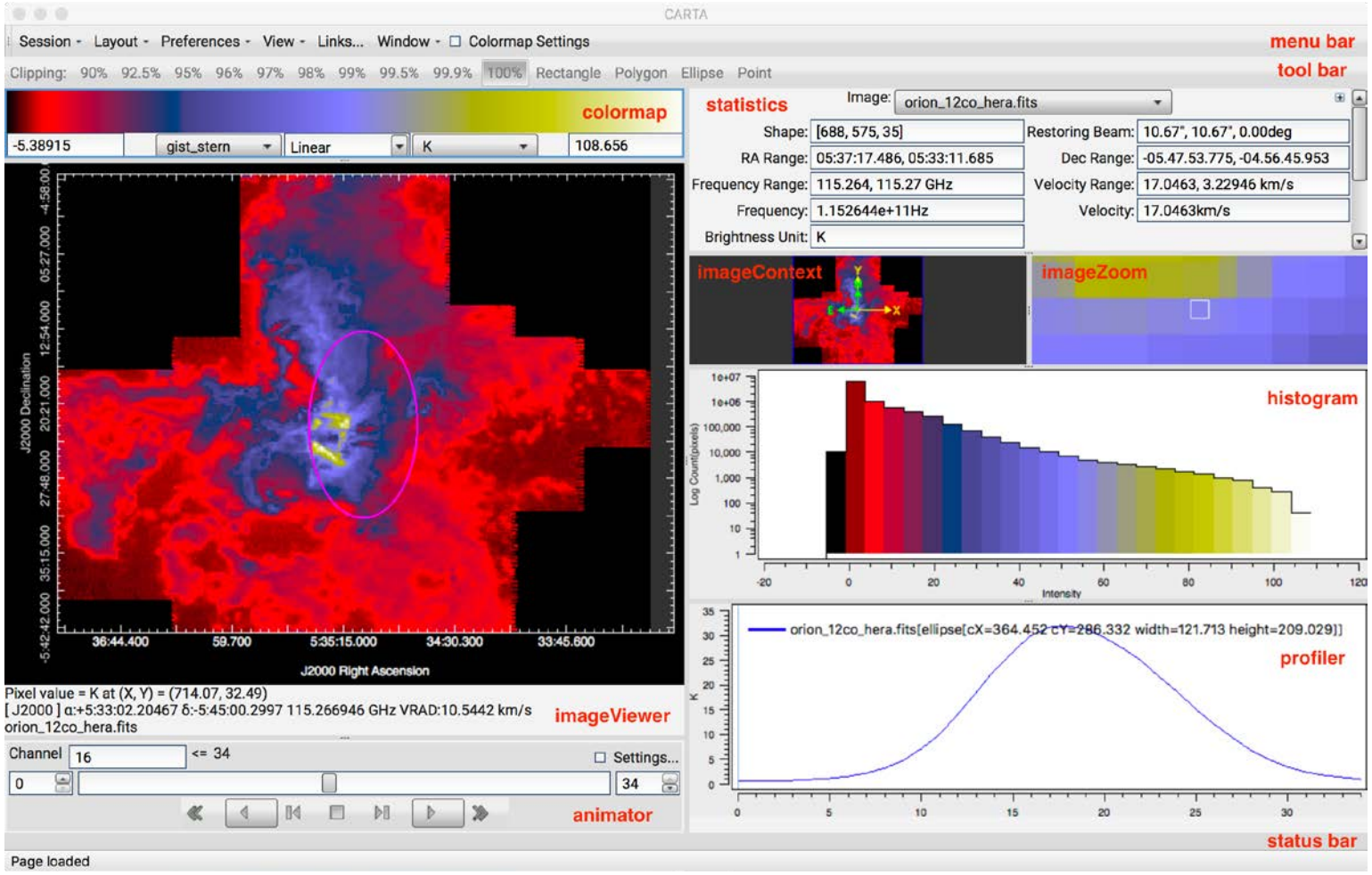
Visualisation

- CyberSKA remote radio astronomy viewer is primary visualization tool in CyberSKA portal
- Currently used with up to 360 GB data cubes
- Provides range of visual analytics algorithms
- Enables sharing of visualization sessions between distributed participants

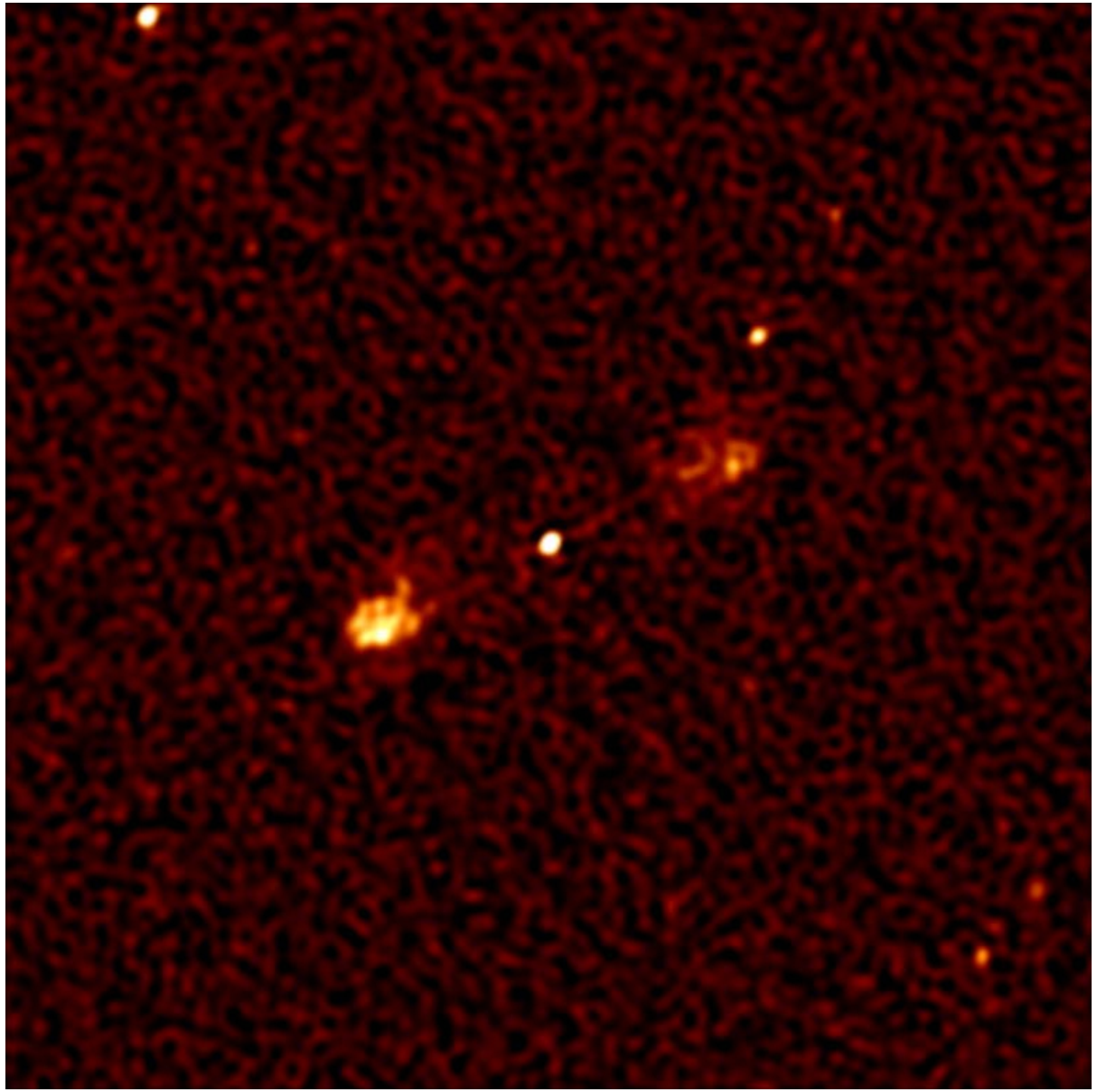


CARTA viewer

- Developing new viewer
  - NRAO collaboration
  - Will replace CASA viewer
- Aim to scale visualisation and analytics to multi-terabyte cubes
- Exploring HDF5 formats supporting parallel I/O
- Currently reworking architecture to provide common GUI across workstation and remote access platforms



Summary

- SKA SA aim to support activities for SRC development
- SDP has developed baseline architecture for delivering data to SRCs. SRC interface definition document to be reviewed soon
- MeerKAT RC framework being developed in multi-partner collaboration aand initial data starting to be distributed
- CyberSKA portal and CARTA viewer development is ongoing with CARTA architecture update to unify GUI across platforms
- South Africa planning to have SRC in addition to SKA1 Mid Processing Centre

MeerKAT First Light (FR II galaxy)