

SKA Global Partnership



SKA Organization incorporated in the UK
Headquarters at Jodrell Bank

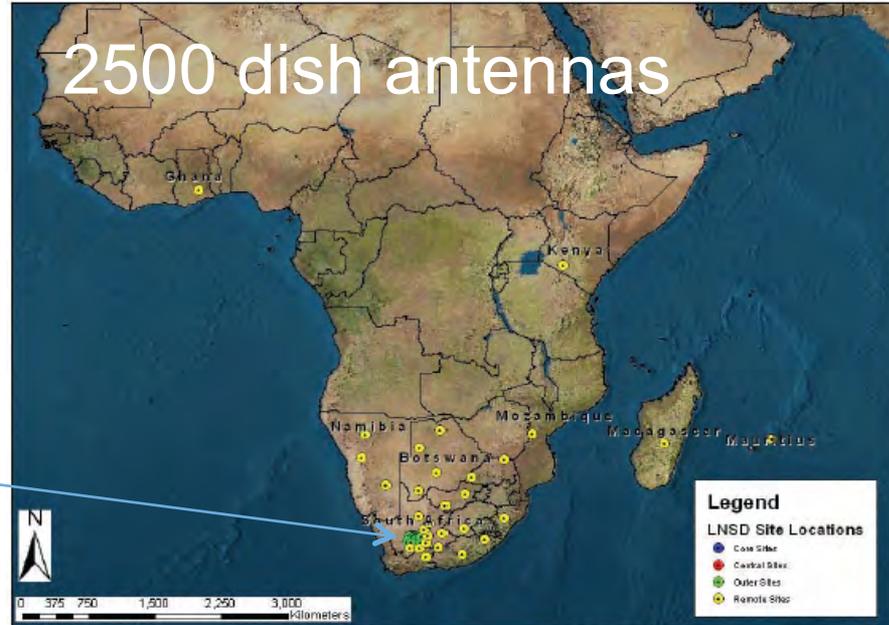
-
- Full members
 - Associate members
 - Member SKA Phase 1 and Phase 2 host countries
 - Non-member SKA Phase 2 host countries
 - SKA Headquarters host country

Southern Africa: SKA-mid frequency dish array

SKA Central Region

Dishes

2500 dish antennas



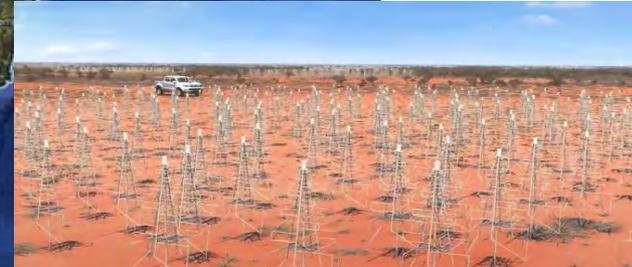
Rob Millman



Australia – SKA Low

10^6 dipole antennas

Sparse
Aperture Arrays



SKA Timeline

2010

2015

2020

2025

South Africa



SKA1
Pre-construction

MeerKAT



SKA1 Construction



SKA

1%

10%

100%

Australia



ASKAP

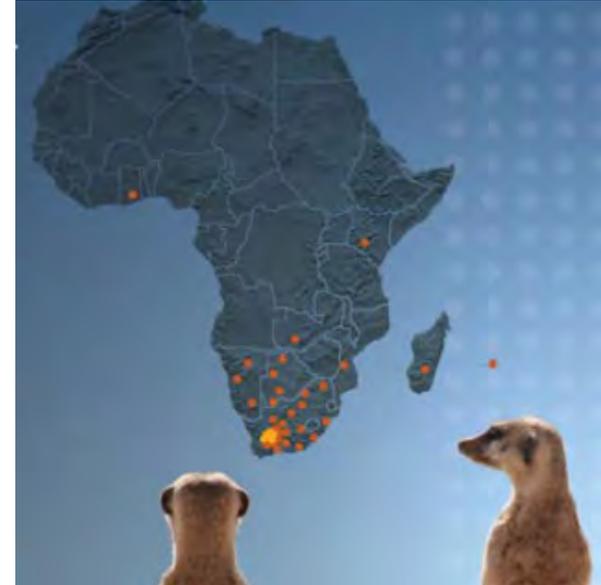


MWA



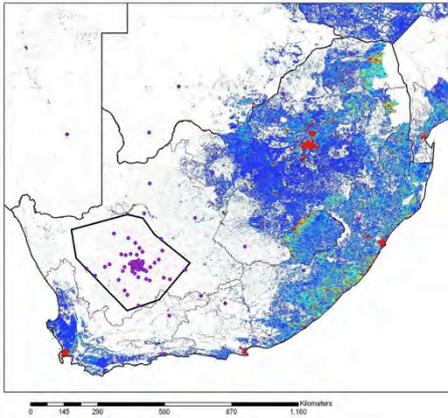
MeerKAT - phase 0 of SKA-mid

- 64 13.5m offset parabolic dishes with single-pixel feeds
- Constructed at SA SKA Site for incorporation into SKA1



MeerKAT under construction at the SA SKA Site

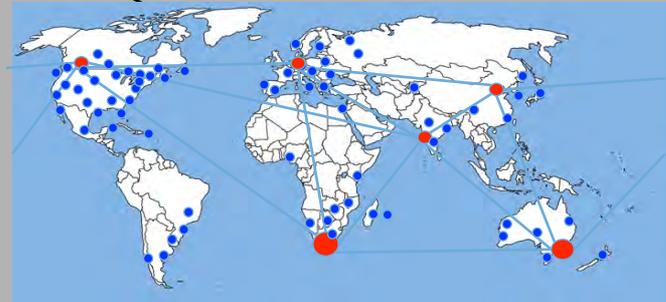
- 16 antennas mid 2016
- 32 antennas early 2017
- 64 antennas late 2017



Sociology of Radio Astronomy



- Much of the key science en route to the SKA will be achieved via large-scale survey mode observing programs executed by globally distributed teams of researchers

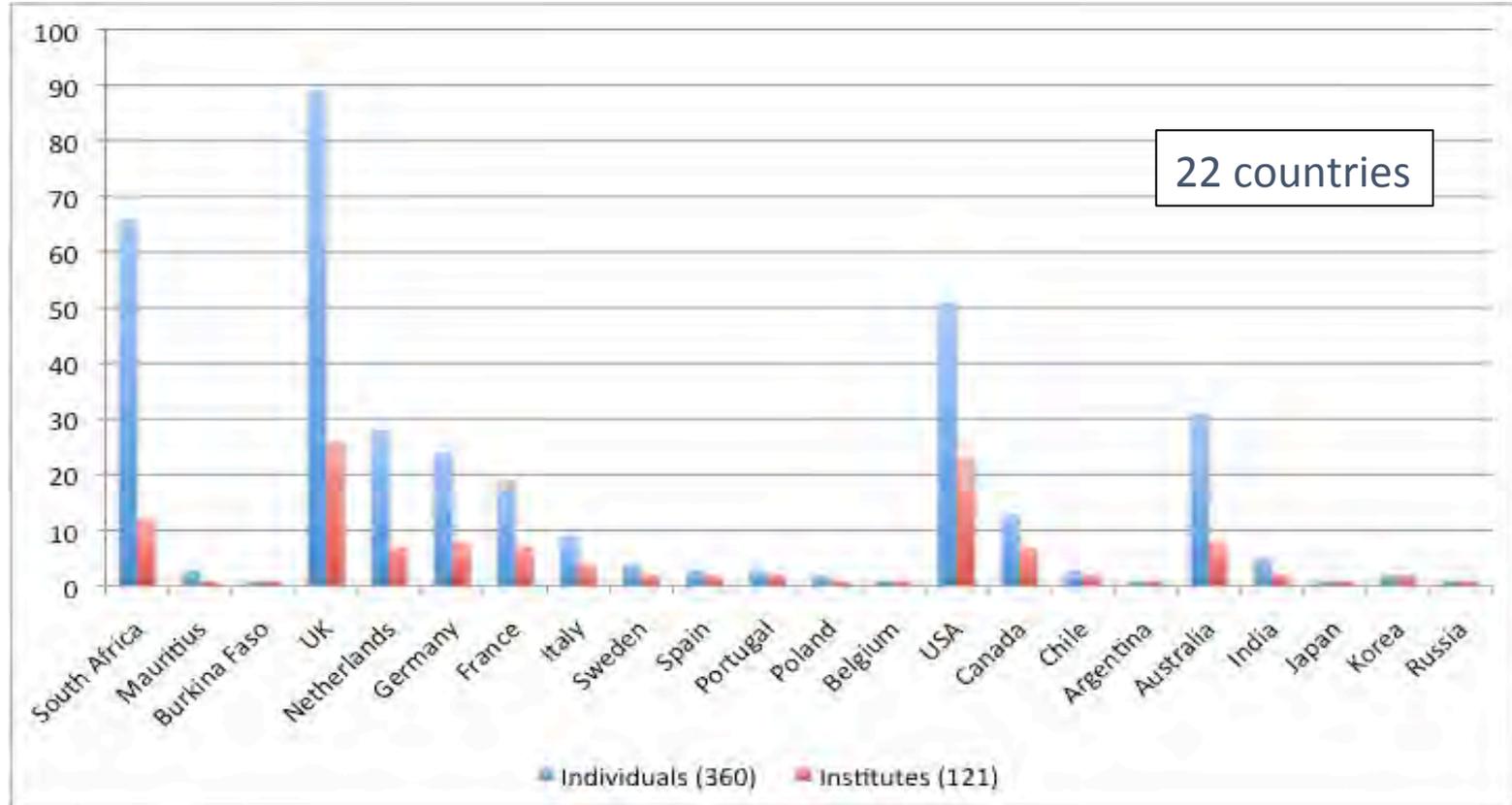


MeerKAT Large Survey Projects

- LADUMA (Deep atomic hydrogen)
- MIGHTEE (Deep radio imaging of the early universe)
- ThunderKAT (exotic phenomena, variables and transients)
- TRAPUM (pulsar search)
- Pulsar Timing
- MESMER (High-z CO)
- MeerKAT Absorption Line Survey
- MHONGOOSE (Nearby HI)
- Fornax MeerKAT HI Survey of Fornax
- MeerGAL (Galactic Plane Survey)



MeerKAT Large Surveys (43,000 hours allocated)



Time Domain Imaging: ThunderKAT and MeerLICHT

ThunderKAT

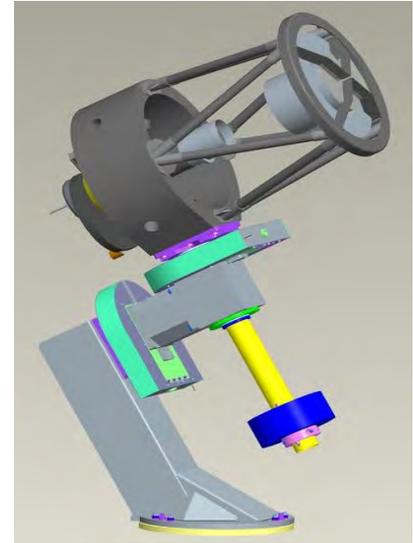
- Real-time imaging of MeerKAT Data Stream
- Targetted and commensal

MeerLICHT

- 60cm fully robotic optical telescope
- Field of view 1.6 x 1.6 degrees, slaved to MeerKAT

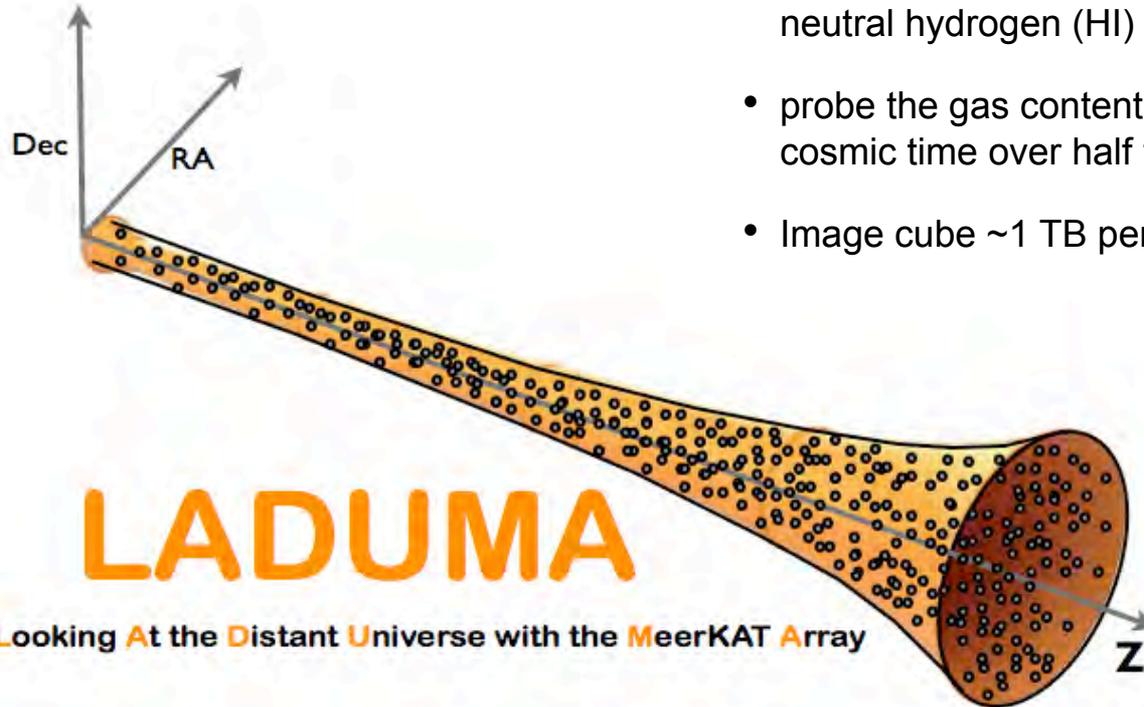
Two time-domain data streams

- Rapid robotic triggers for follow up – VoEvent tools



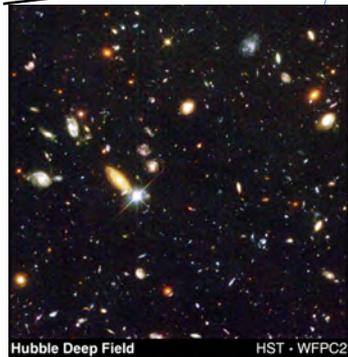
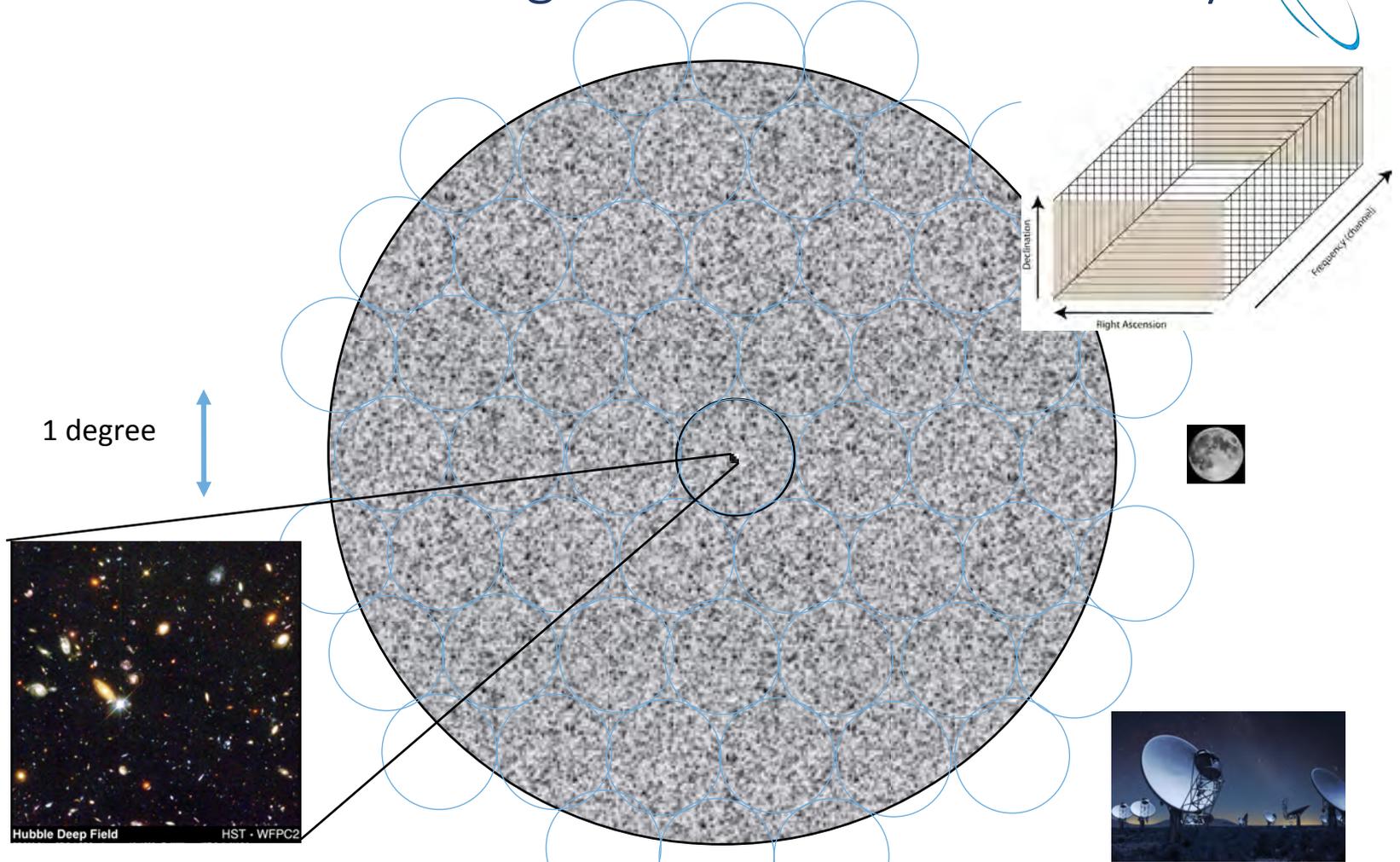
LADUMA

- 5000 h to make the deepest observations of neutral hydrogen (HI) in galaxies to date
- probe the gas content of galaxies back in cosmic time over half the age of the universe.
- Image cube ~ 1 TB per sq deg.



(Blyth, Holwerda & Baker)

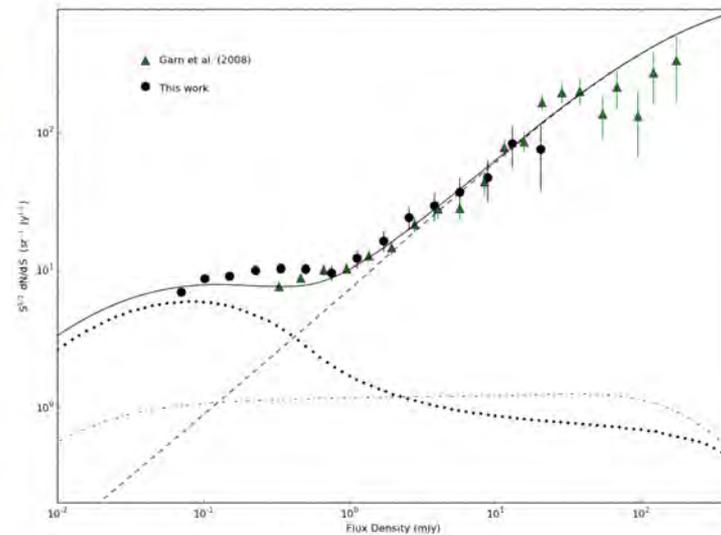
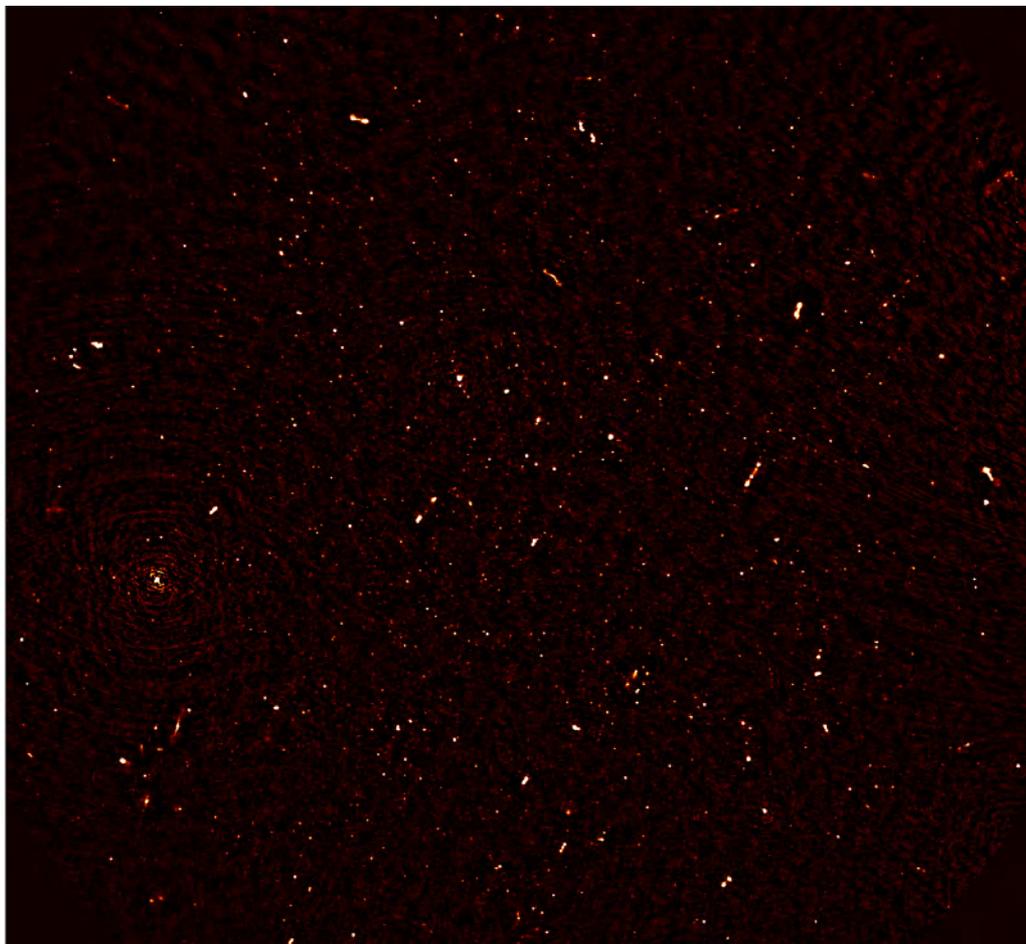
MIGHTEE – Tiered Extragalactic Continuum Survey



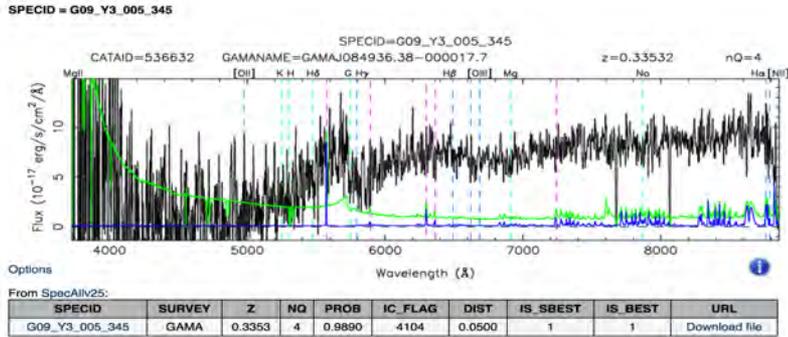
MeerKAT Survey project Image Data Sizes

- MIGHTEE – 20 sq. deg
 - Channel width ~ 1 MHz, 1000 channels per band, 2" pixels
 - four polarizations (α , δ , ν , S) ~ 1 TB per band
 - Three bands ~ 3 TB
- LADUMA – 1 sq. deg.
 - Channel width ~ 2 KHz, 32,000 channels for 770 MHz
 - (α , δ , ν) ~ 0.5 TB per band per sq deg.
 - Two bands ~ 1 TB
- A wide area survey 100 sq. deg $\rightarrow \sim 5 - 100$ TB
- All sky survey ~ 1 PB

The Loneliness of Radio Data



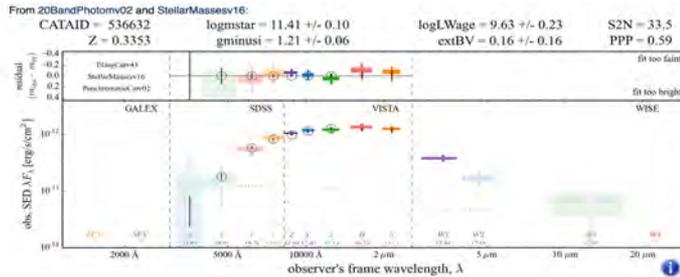
Large Multi-wavelength Data Sets



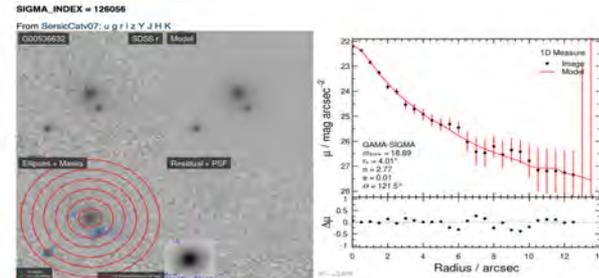
Currnet spectra of this object:

SPECID	SURVEY	Z	NQ	PROB	DIST	IS_SBEST	IS_BEST	URL	URL_IMG
429214988947140096	SDSS	0.3358	4	0	0.1800	1	0	Download file	View image
525797212534388256	SDSS	0.5810	1	0	0.1100	0	0	Download file	View image
G09_Y2_014_355	GAMA	0.3356	2	0.5950	0.0500	0	0	Download file	View image

[top]



[top]



No group information available for this object!

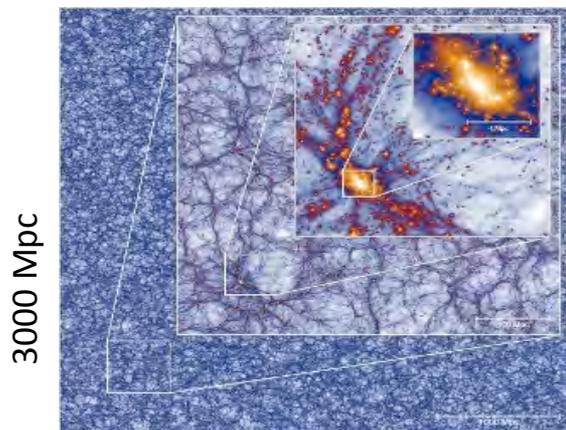
[top]

Michelle Cluver, Mattia Vaccari (UWC), Tom Jarret (UCT)

Multi-Scale Universe Simulations

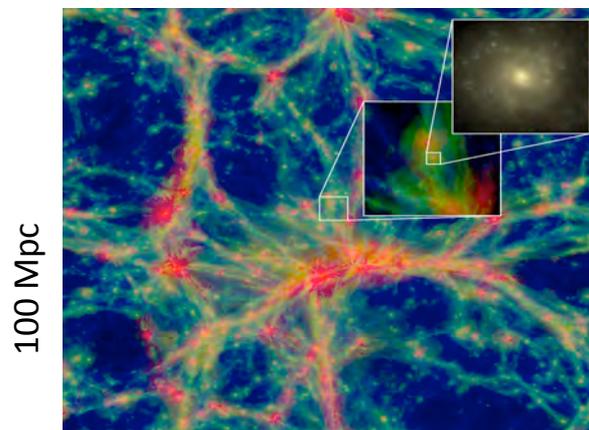
Big Data meets big simulations

- “Hubble volume” simulations (\sim Gpc):
- “Cosmological” simulations (\sim 100Mpc):
- “Zoom” simulations (\sim 1 Mpc)

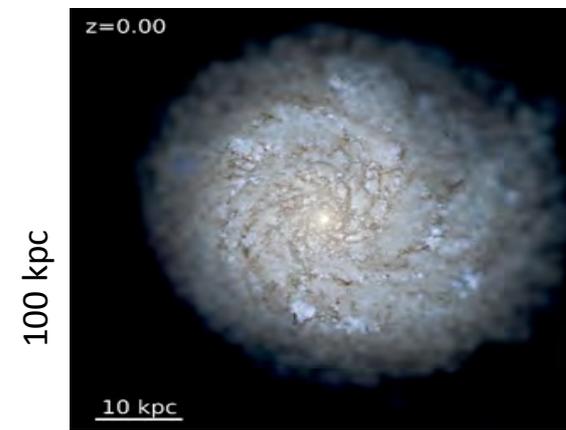


Millennium XXL simulation

Romeel Dave (UWC)

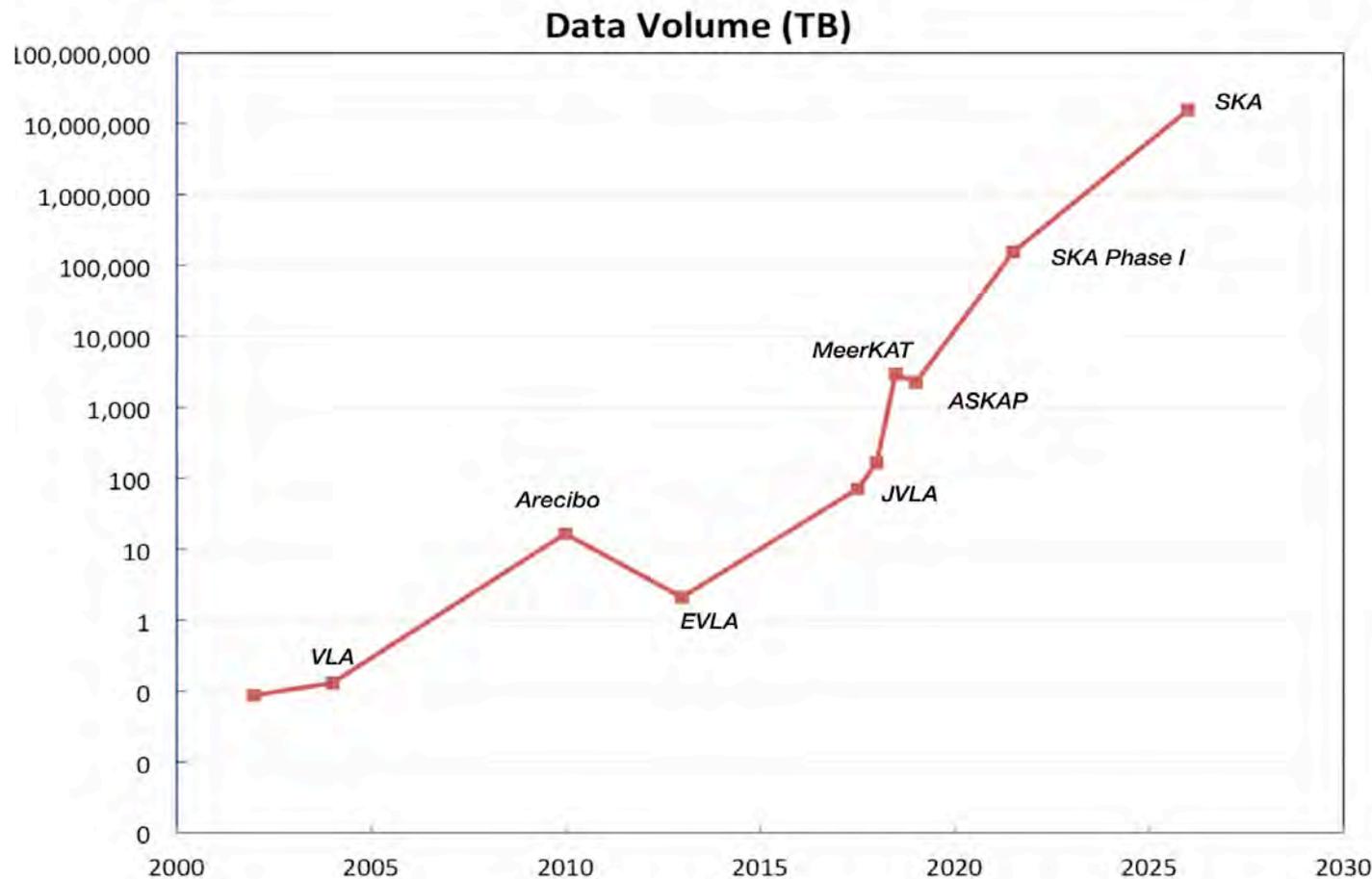


Eagle simulation

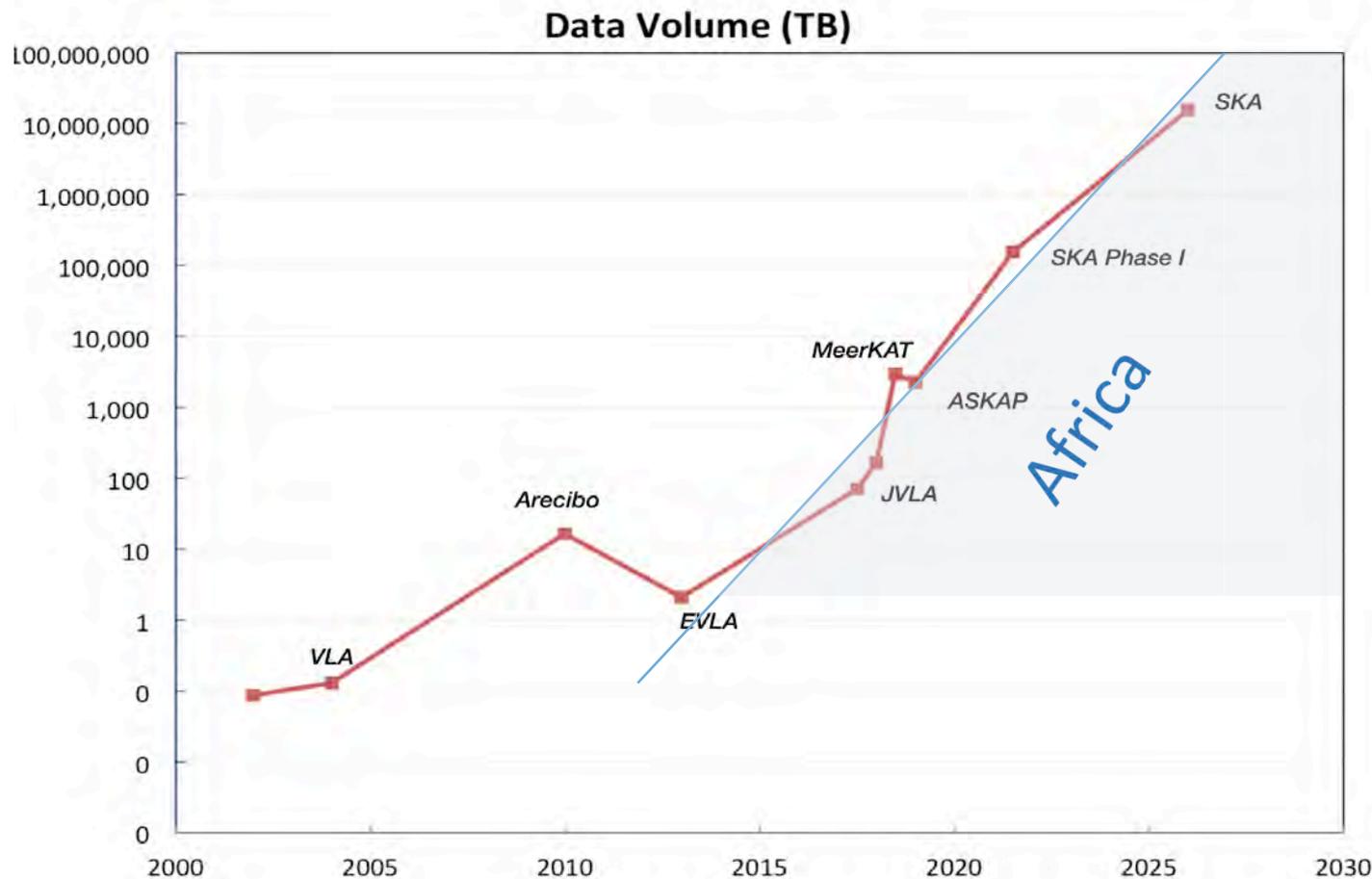


FIRE simulation

Growth of Data Volumes to Radio Astronomers

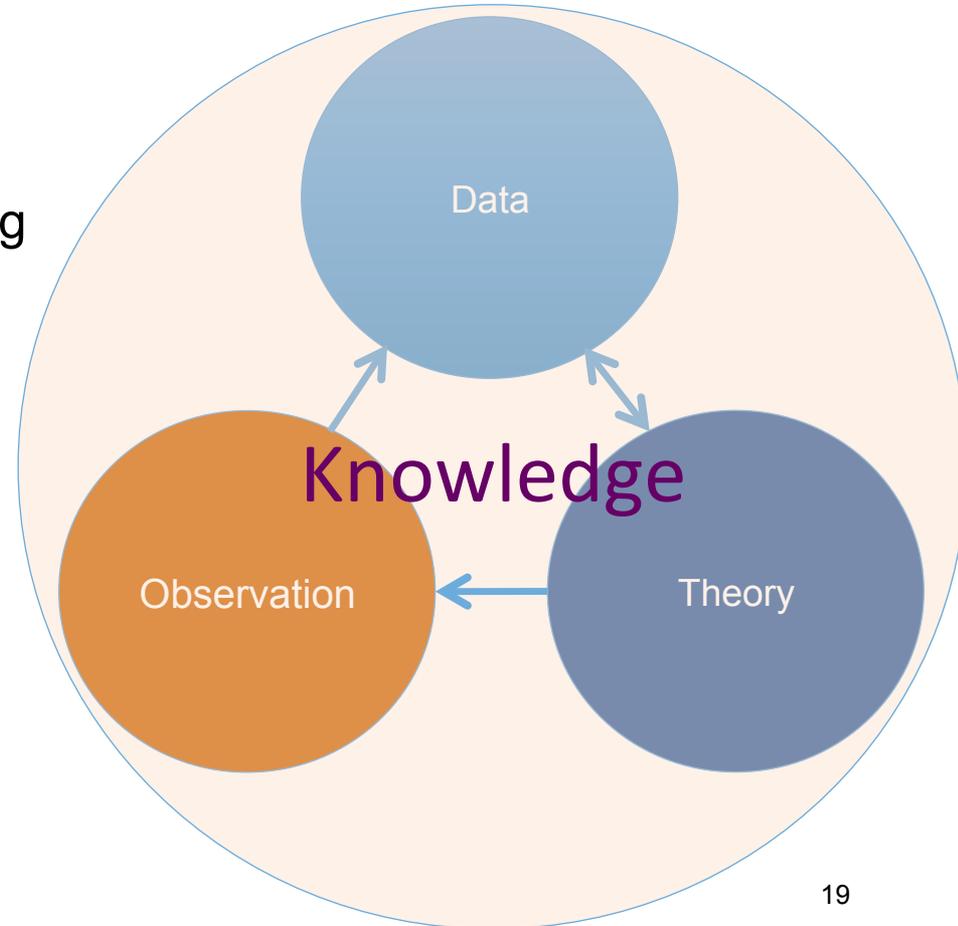


Growth of Data Volumes to Radio Astronomers

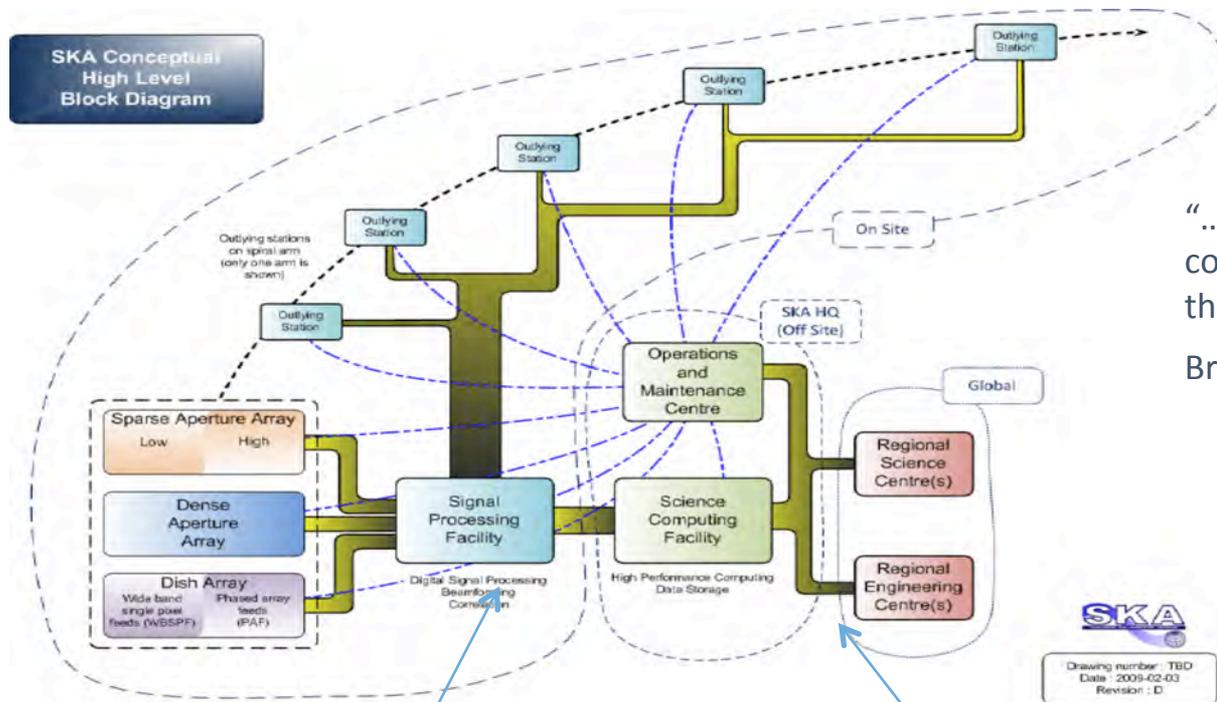


The Challenges: data to knowledge

- Exponential increase in rates and volumes
- Complex, multi-purpose, processing and analysis for a key science questions, and data mining and exploration
- Interface of big data with big simulations
- Collaborative execution of big data science projects by globally distributed teams of researchers



Data Flow....



“... the data rate in the SKA will be comparable to that circulating around the globe by the rest of humanity.”

Bruce Elmegeen (IBM Research)

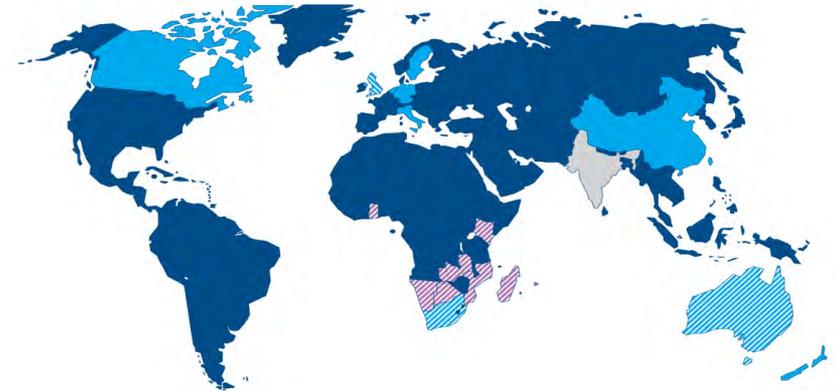
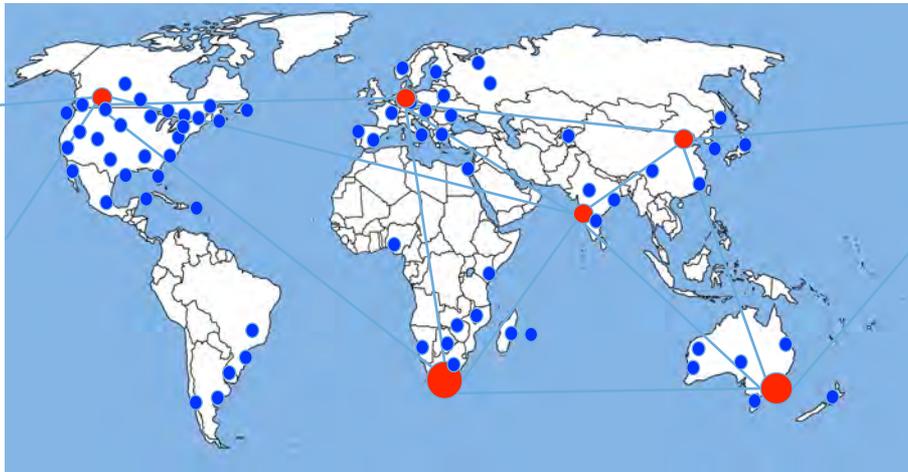
Very Large but well defined BIG data problem – a streaming data engineering challenge

Somewhat smaller but much more complex BIG data problem – astronomy, data science, engineering, sociology, governance, politics, business

SKA Global Partnership in Data

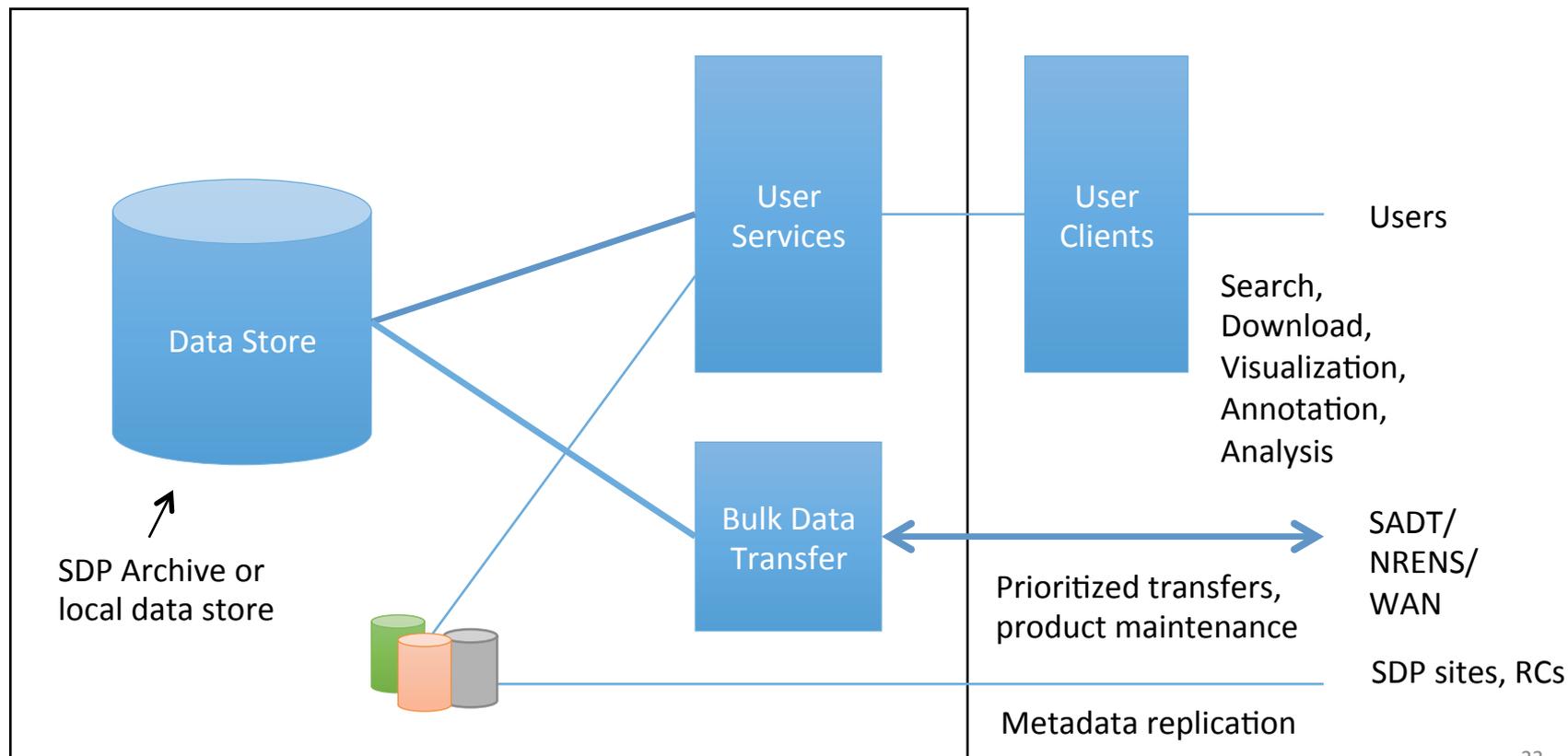


- SKAO Board established SKA Data Flow Panel with rep from each country
- Data Flow Steering Group
 - Russ Taylor (Inter-University Institute for Data Intensive Astronomy)
 - Paul Alexander (Cambridge Institute for Astronomy)
 - Michael Weiss (Netherlands Institute for Radio Astronomy)
 - Peter Quinn (International Centre for Radio Astronomy Research)
- Develop concept document for SKA Data Flow and SKA Regional Science Data Centres



- Full members
- Associate members
- Member SKA Phase 1 and Phase 2 host countries
- Non-member SKA Phase 2 host countries
- SKA Headquarters host country

SKA Data Delivery Design Architecture



SKA Data Delivery IVOA services

- Currently proposing to host the following services in for the SKA Data Delivery Systems
 - TAP Discovery (Table Access Protocol) with ObsCore Data Model
 - SIA Discovery Service (Simple Image Access)
 - SSA Discovery service (Simple Spectral Access)
 - DataLink Service (used to link related data and metadata)
- Additional service supported for Observatory support
 - SODA service (to provides cutouts)
- VOEvent (used to send transient alerts) also supported by SKA Telescope Manager

SKA Precursor Regional Science Data Centres

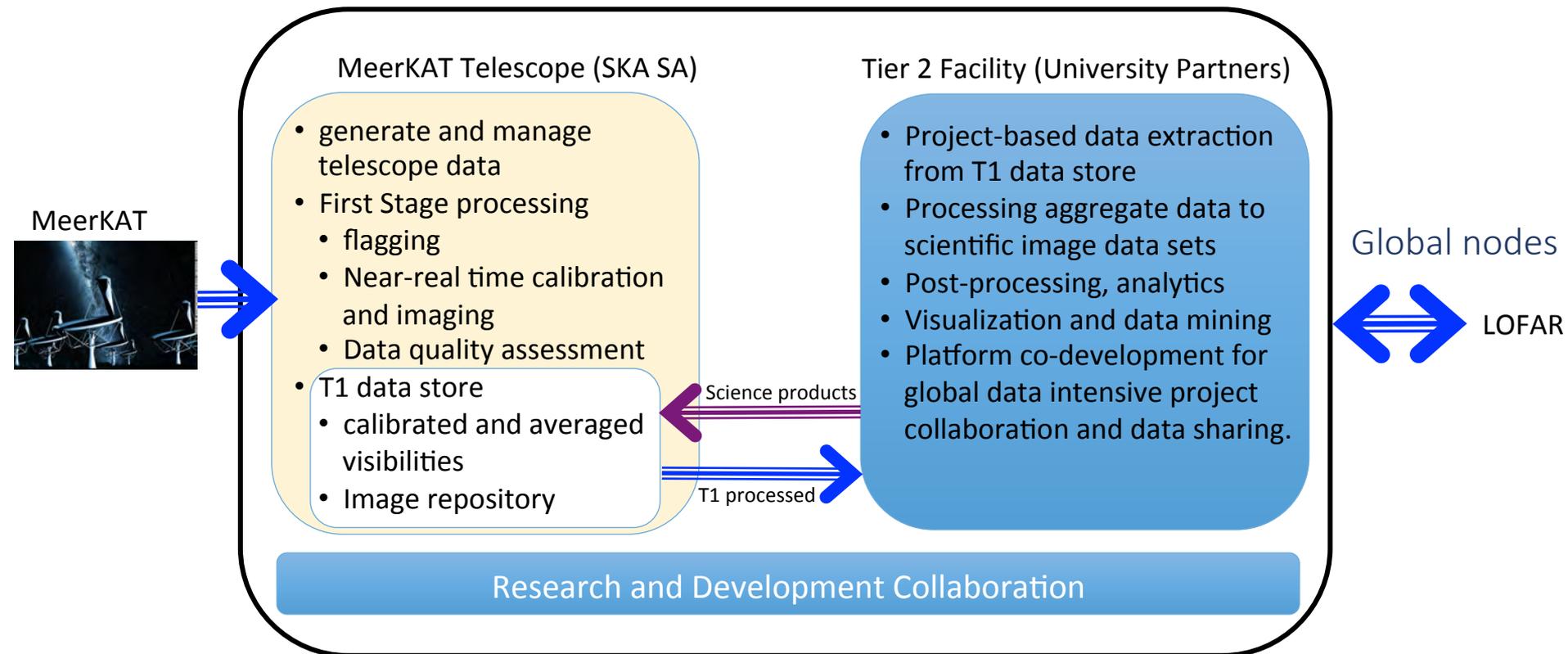
- MoU to collaborate on development of Precursor SKA RSDC
 - ASTRON (Netherlands Institute for Radio Astronomy)
 - NWO (Netherlands Organization for Scientific Research)
 - IBM-NL
 - **IDIA**
 - **SKA – SA**

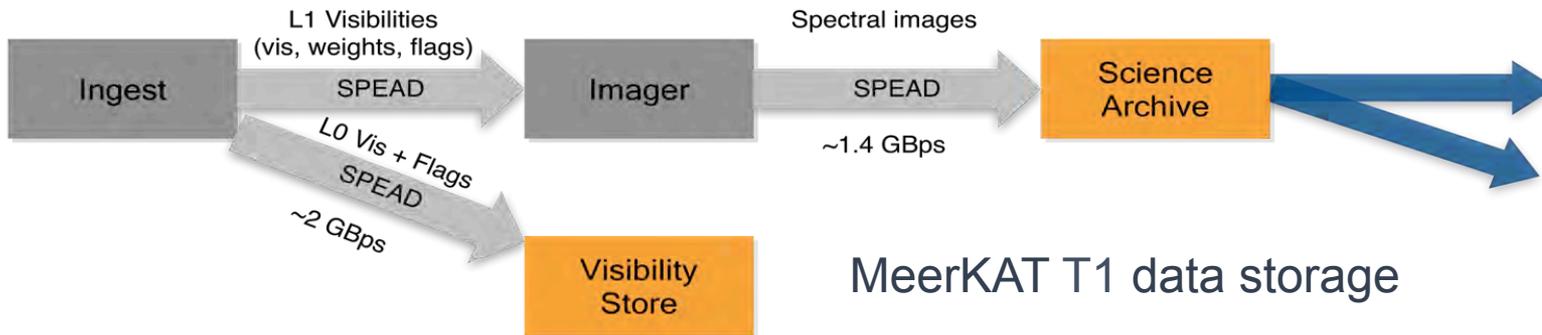
Signed 17 November 2015



SKA Precursor Regional Science Data Centres

MeerKAT and LOFAR data and use cases





Visibility Store

- Telescope visibility data
- 18 PB capacity; at CHPC Cape Town.
- Data stored on LT07 tape.
- Data access: No direct access; staged from tape.

Science Archive

- Telescope image cubes
- 10 PB capacity; site redundant.
- Data stored on HDDs.
- Data stored in CEPH object store:
 - 130000 objects per image cube
 - Object metadata stored in Apache Solr or Elastic.co
- Data access:
 - Direct via RADOS; Amazon S3 gateway
 - VO gateway

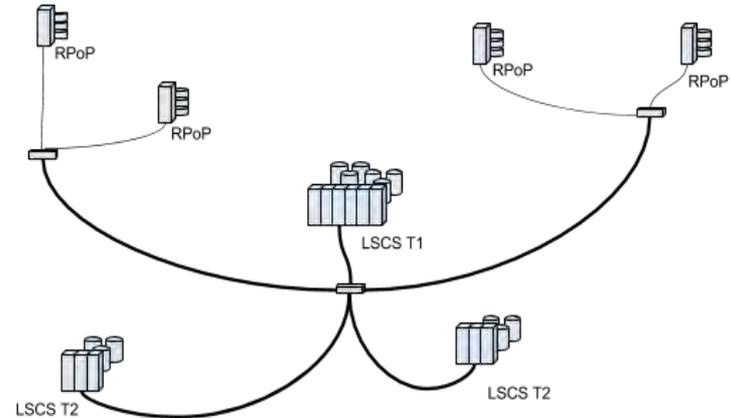
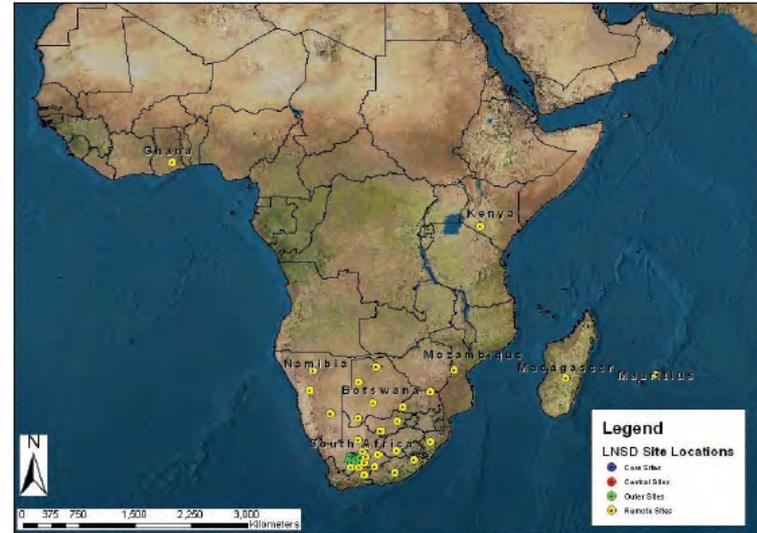
Tier 2 Data Processing Facility

- Funded and managed by IDIA University partners
- 3-4 PB persistent spinning disk storage to aggregate project visibilities from T1 data store
- Data-centric, high performance computing cluster
- On-going development platform for data processing and post-processing algorithms and catalogue outputs
- Cloud enabled services and distributed platform for project-based data access, apps to data, workflows, analytics, visualization,...
- Part of a national, distributed, tiered, data-intensive research infrastructure – African Data Intensive Research Cloud

African Data Intensive Research Cloud

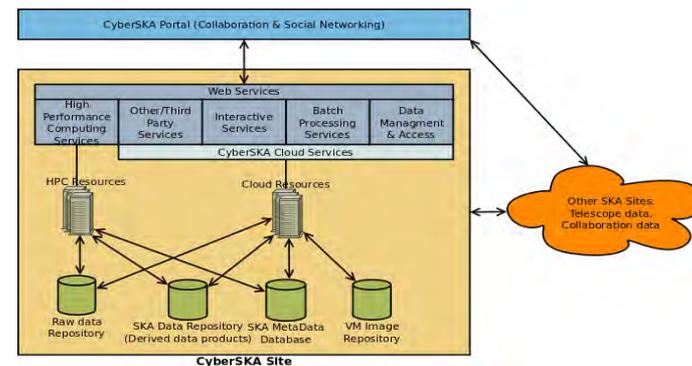
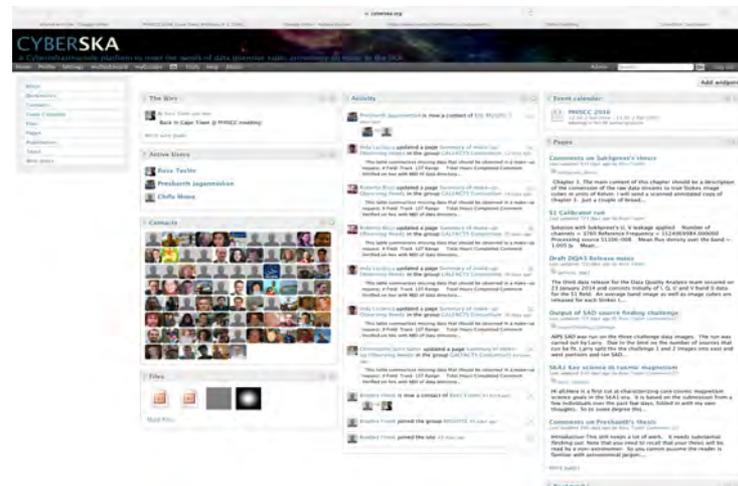


- Platform for data access and research on astronomy big data toward the SKA era
- Bring together distributed infrastructure and resources into a tiered federated cloud around data, software, analytics, visualization, collaboration,..
- Proto-type under development between IDIA partners (UCT, UWC, UP, NWU), and SKA-SA, CHPC, Canonical, Dell,...
- Initial focus on astronomy but expand to other data challenged domains, e.g. bioinformatics



CyberSKA: A cloud-enabled Big Data Research Platform

- Collaboration
 - Portal built on social networking and sharing technologies
- Data Management
 - Scalable collaborative access, sharing and searching of distributed (BIG) data sets
- Data Visualization and Visual analytics
 - On-line interactive visualization of remote Big Data
- Third Party Applications
 - Community driven site with common API



CyberSKA: A cloud-enabled Big Data Research Platform

- 620 members from around the world

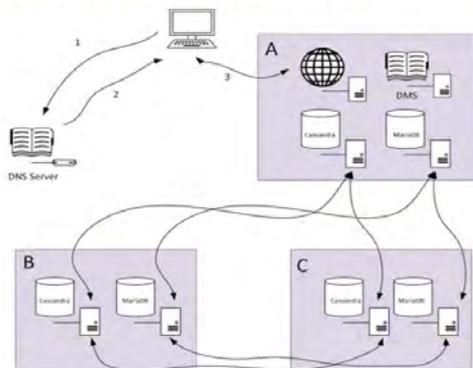


Federated CyberSKA Platform

Integrating Globally Distributed Resources into the CyberSKA Platform

David Aikema, Rob Simmonds and Russ Taylor – info@cyberska.org

Restructuring for a globally-distributed system



- DNS based geolocation used to connect to the nearest CyberSKA instance – Region A in this case.
- Metadata describing files in the DMS instances in each region is stored in Apache Cassandra database that provides eventually consistent replication of this metadata between regions.
- MariaDB using multi-master replication with global transaction ID support is used to distribute portal configuration and account information.

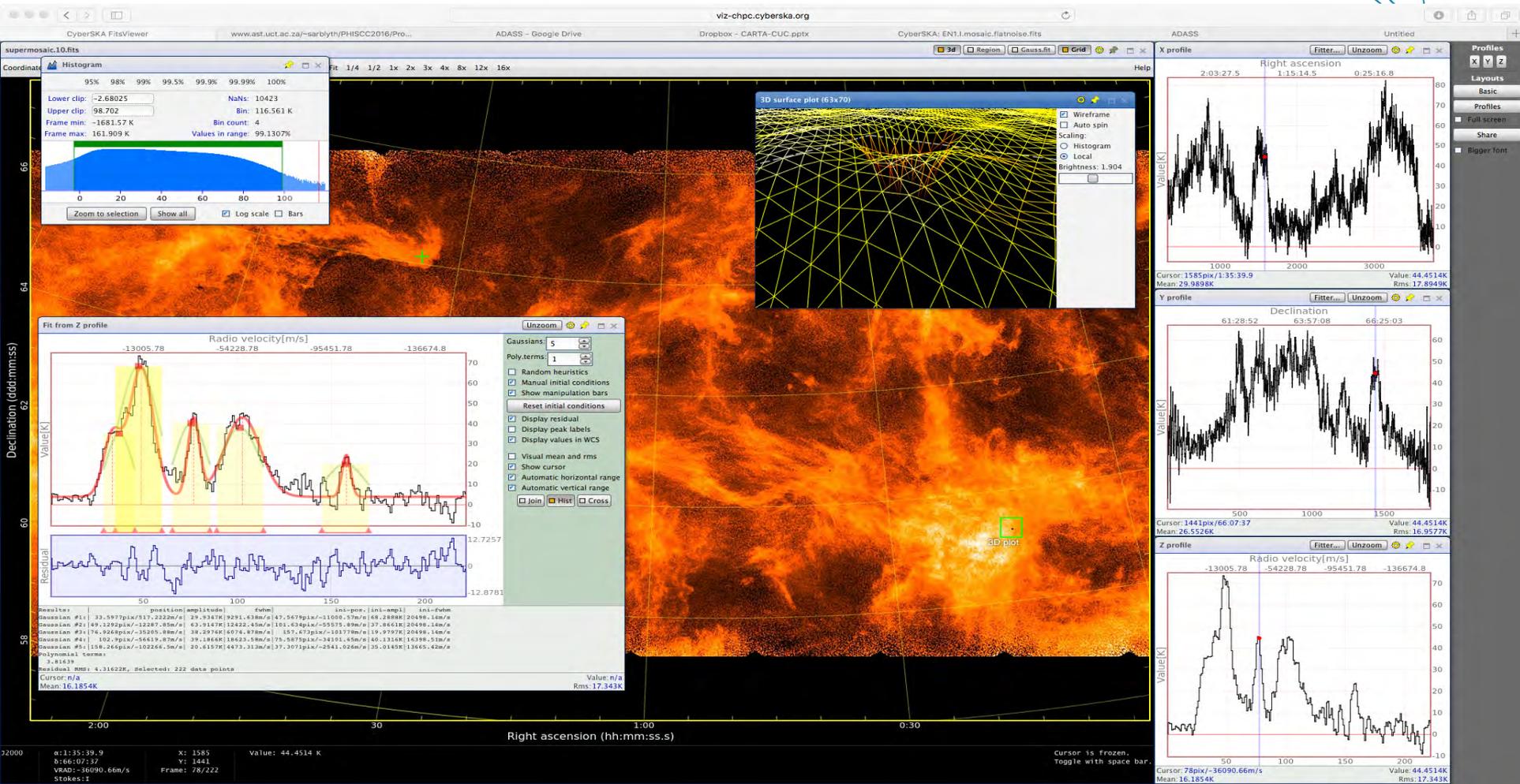
CyberSKA operates two platforms - one for production use and another for experimental purposes. This poster outlines some of the developments in our experimental testbed.

CyberSKA Testbed Resource Locations



Red indicates currently active resources in the testbed whereas yellow indicates resources awaiting integration

CyberSKA Visual Analytics Tool



File Formats for scaling to Big Data

- **FITS**
 - First released in 1981
 - Widely used in astronomy
 - Sequential access only

- **HDF5**
 - Generic container format, can be used with any data model
 - Supports parallel read/write access and streaming

SA³ IVOA Collaboration

- Setup SA IVOA PostgreSQL database with CAOM data model
- Run IVOA tools provided by CADC
 - CADC a partner in the SKA SDP Data Delivery design work package
- Perform data engineering required to support MeerKAT data
 - Work with HDF5 data files
 - Support MeerKAT radio data
- Host MeerKAT and SALT/SAAO metadata
- Time series data model?
- explore IVOA approaches to multi-wavelength data fusion, analytics and mining for MeerKAT large survey projects