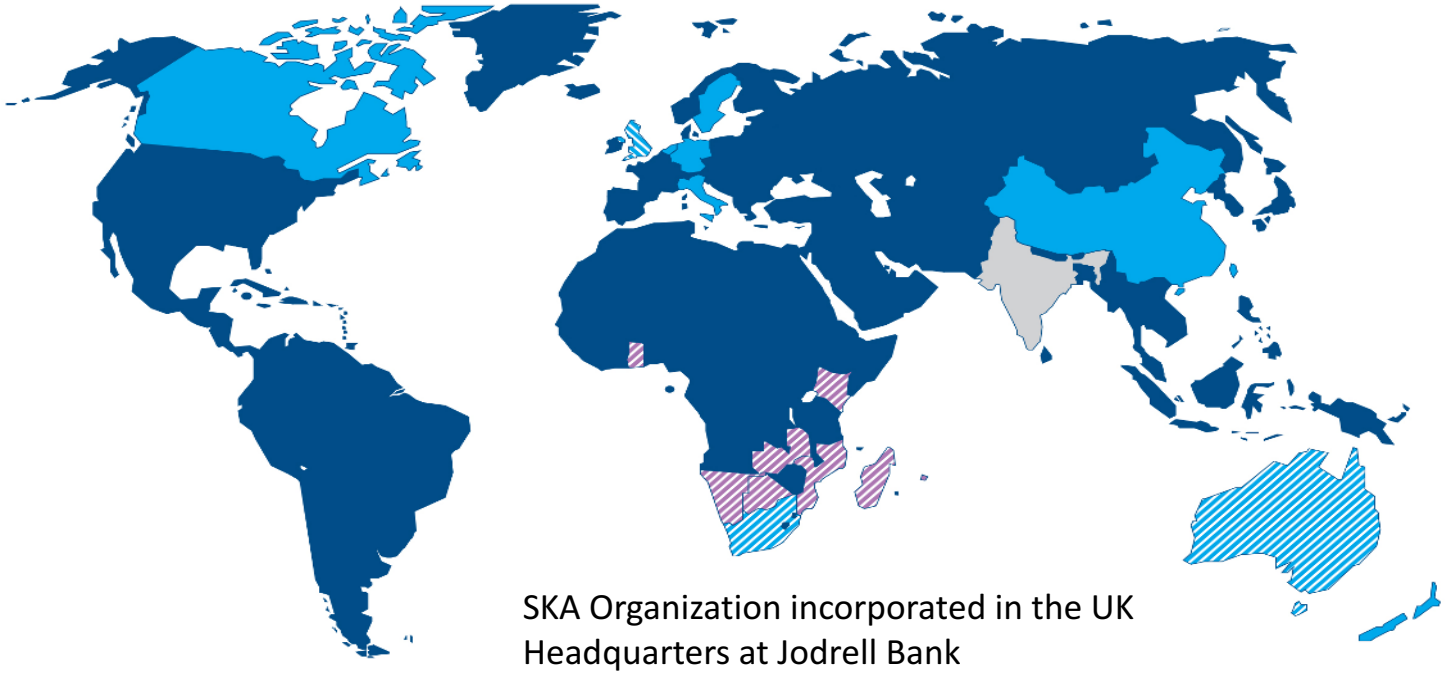
IDIA and the Big Data Challenge in South African Astronomy

Russ Taylor
SKA Research Chair
University of Cape Town and University of the Western Cape
Director
Inter-University Institute for Data Intensive Astronomy

BRICS Astronomy Workshop 2016

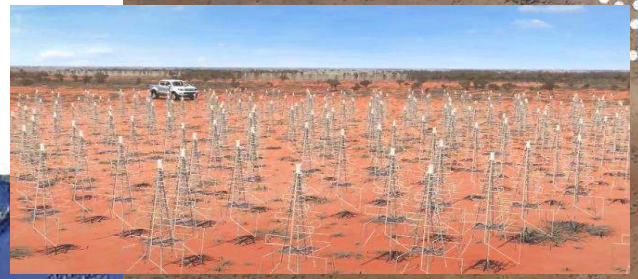
SKA Global Partnership



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

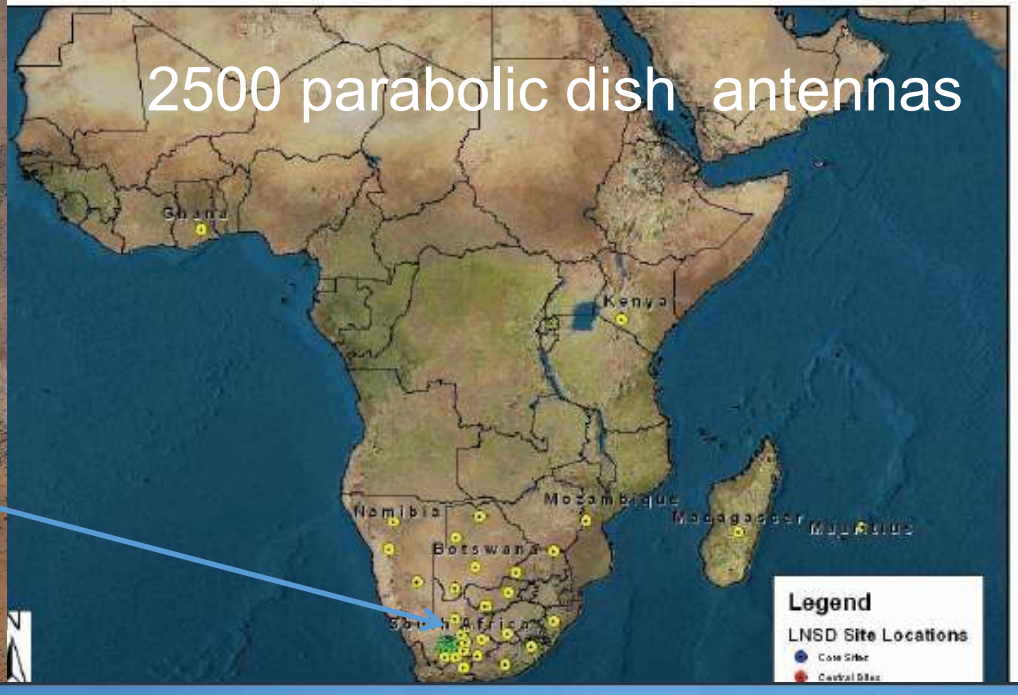
Australia: SKA-LOW



10^6 dipole antennas

Sparse
Aperture Arrays

Southern Africa: SKA-MID



SKA Timeline

2010 | 2015 | 2020 | 2025

South Africa

SKA1
Pre-construction

SKA1 Construction

SKA

MeerKAT

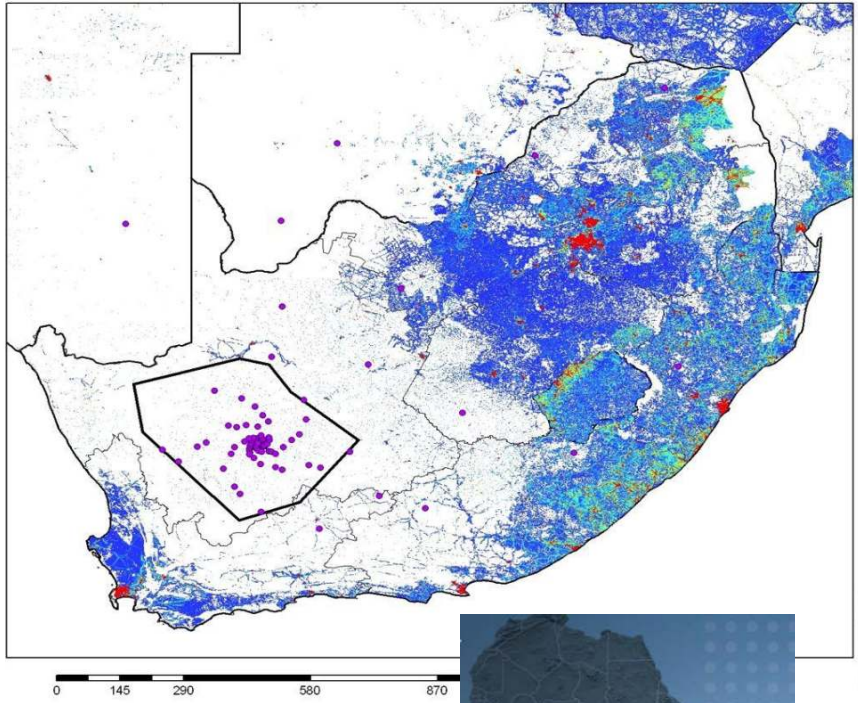
1% | 10% | 100%

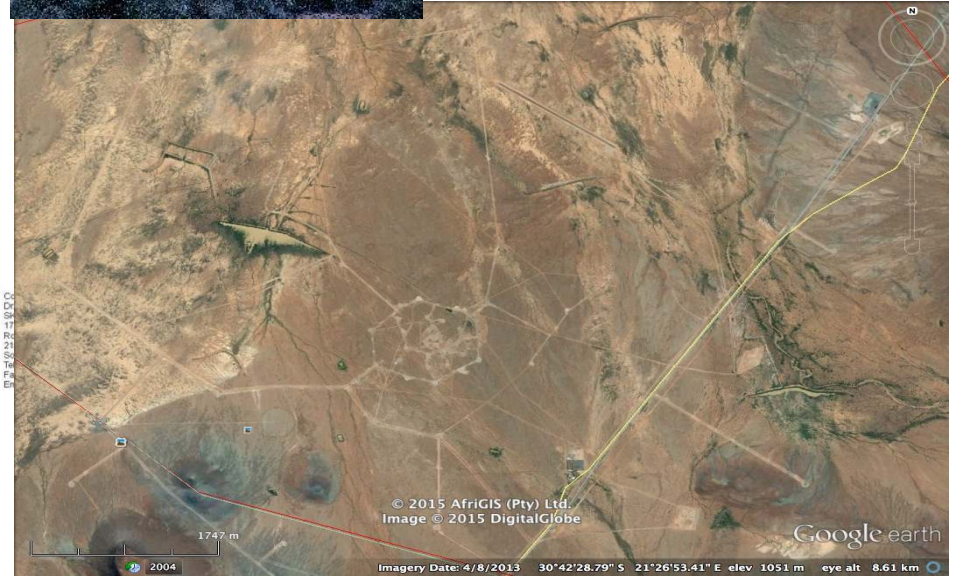
Australia

ASKAP

MWA

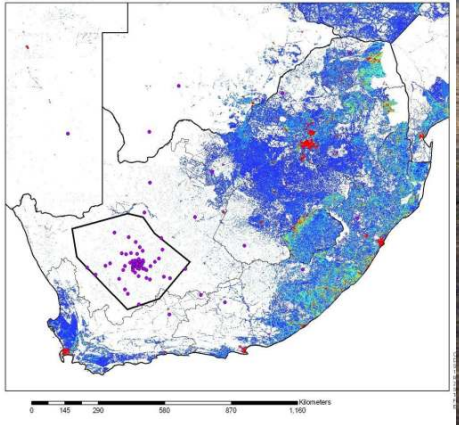
MeerKAT - phase 0 of SKA-mid



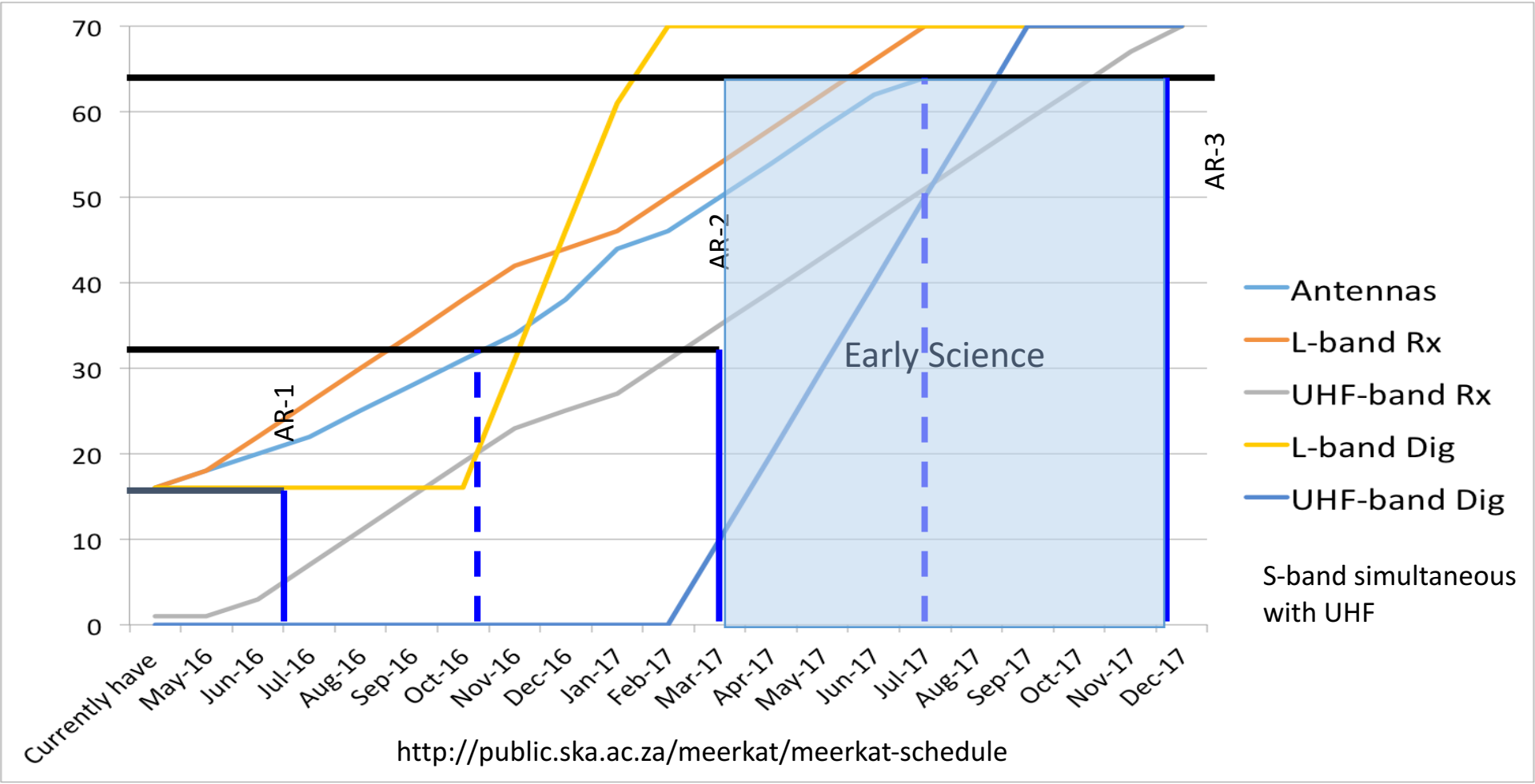


MeerKAT - phase 0 of SKA-mid

Operational end of 2017

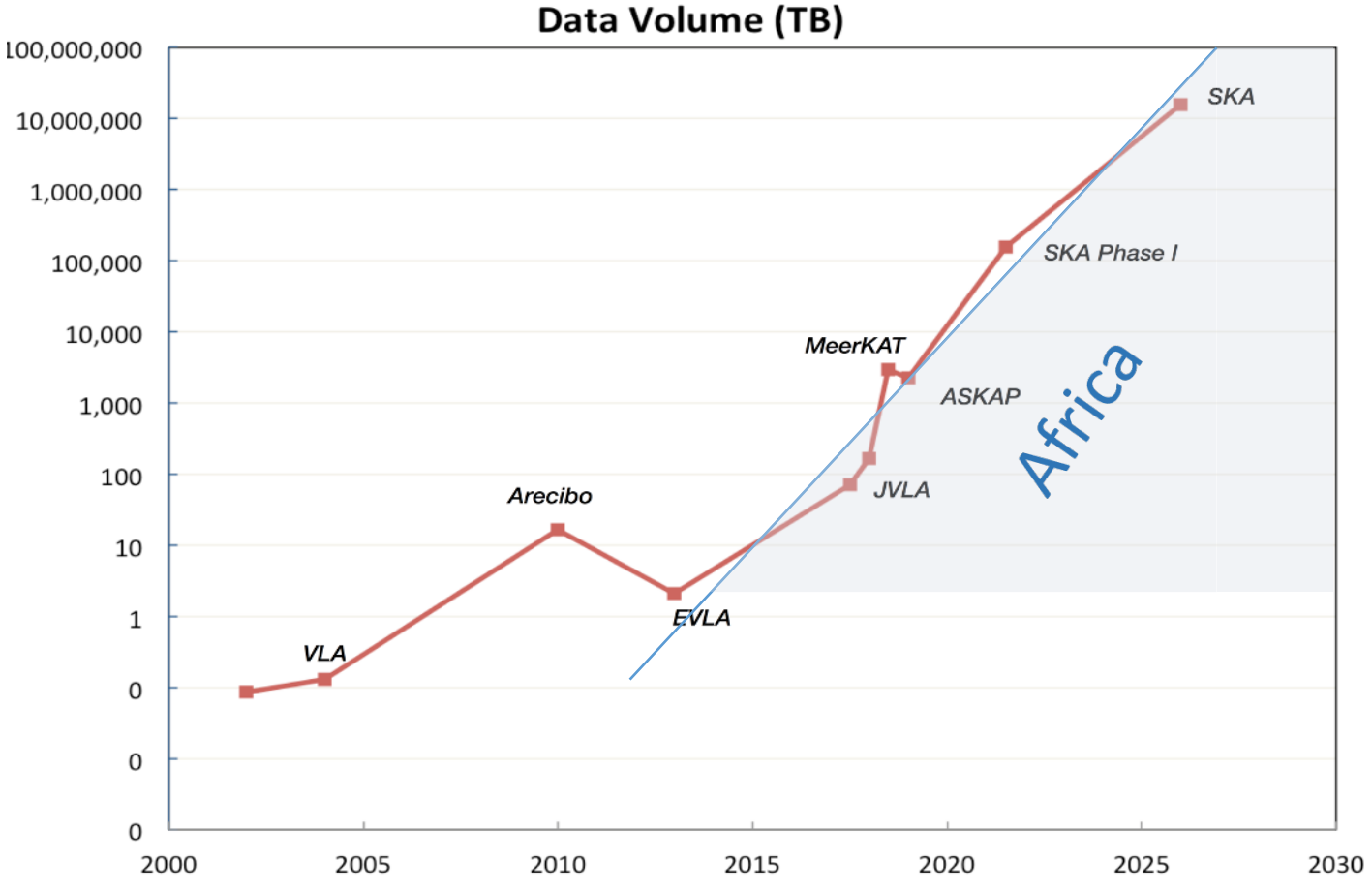


MeerKAT Schedule



http://public.ska.ac.za/meerkat/meerkat-schedule

Growth of Data Volumes to Radio Astronomers



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

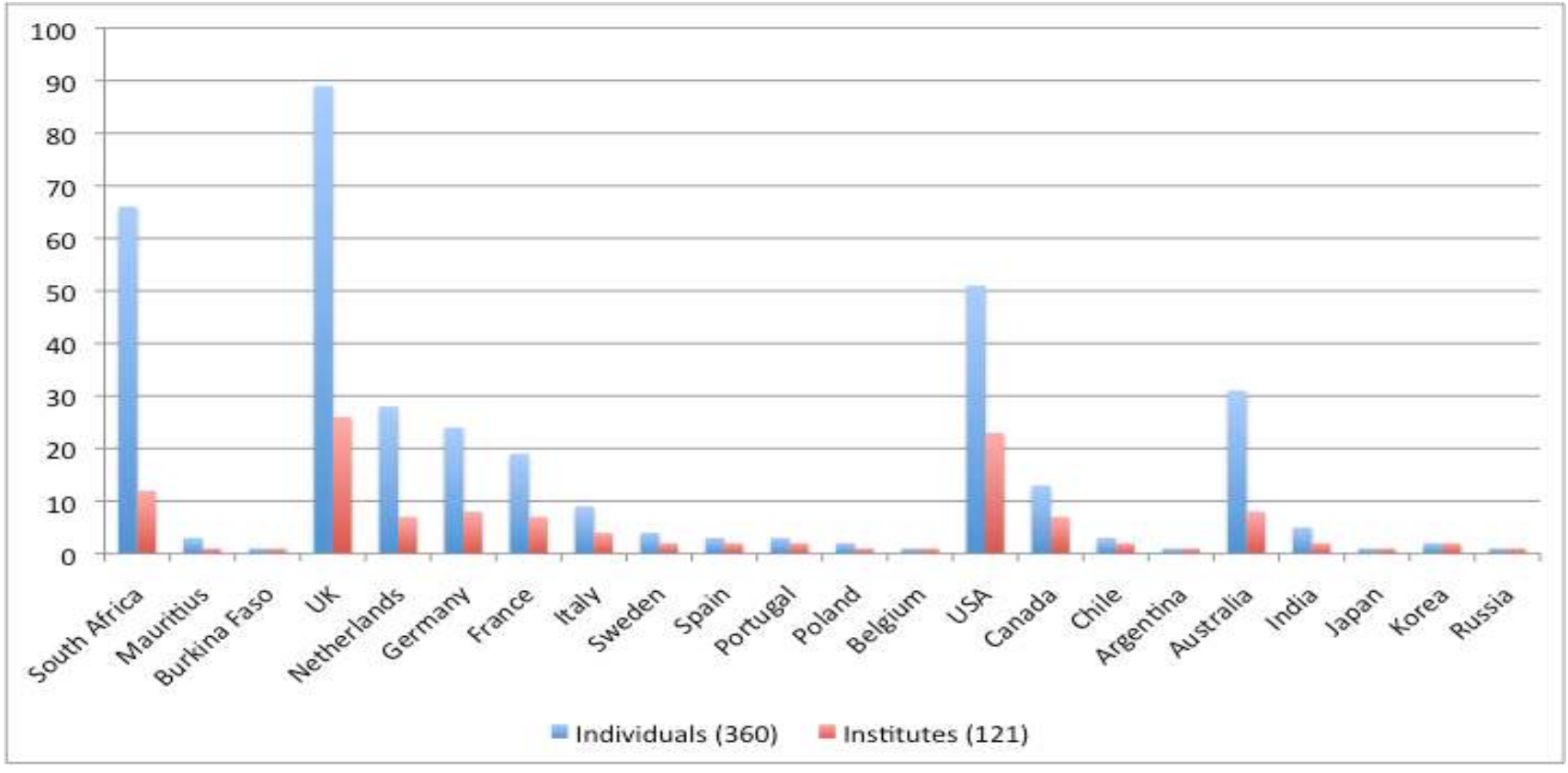
imaging

Time domain

- LADUMA (Deep atomic hydrogen)
- MIGHTEE (Deep continuum imaging of the early universe)
- Fornax (Deep HI Survey of the Fornax cluster)
- MHONGOOSE (targeted nearby galaxies HI)
- MeerKAT Absorption Line Survey (extagalactic HI absorption)
- ThunderKAT (exotic phenomena, variables and transients)
- TRAPUM (pulsar search)
- Pulsar Timing (no acronym)
- MESMER (High-z CO)
- MeerGAL (Galactic Plane Survey)

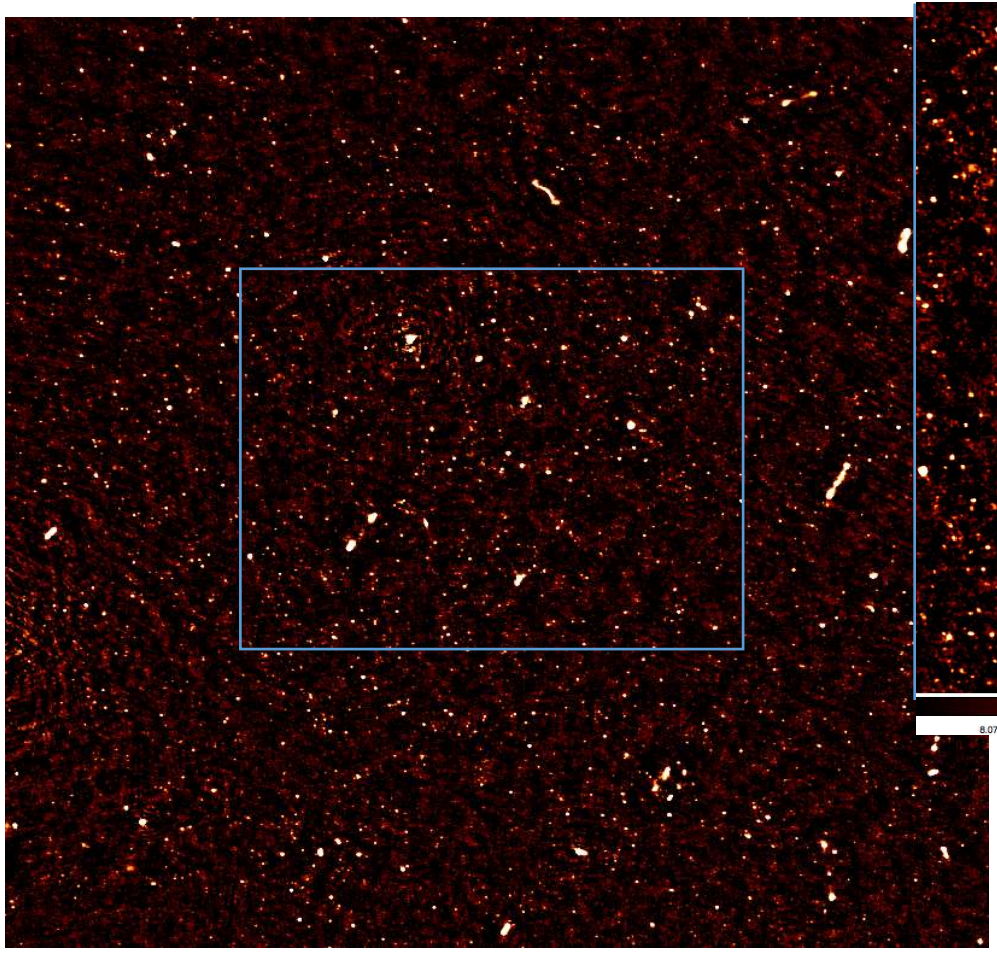
http://public.ska.ac.za/meerkat/meerkat-large-survey-projects

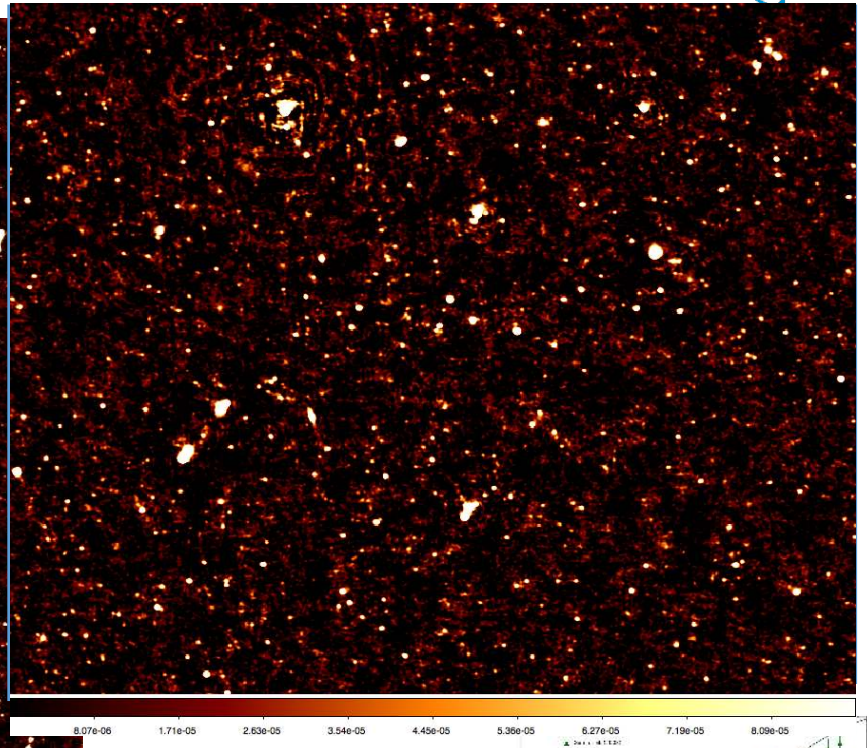
MeerKAT Large Surveys (43,000 hours allocated)



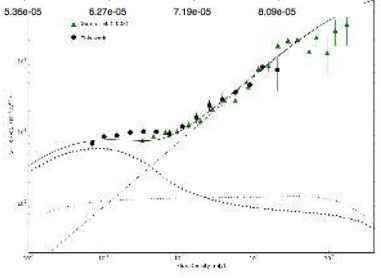
22 countries

What a MeerKAT Continuum image will look like

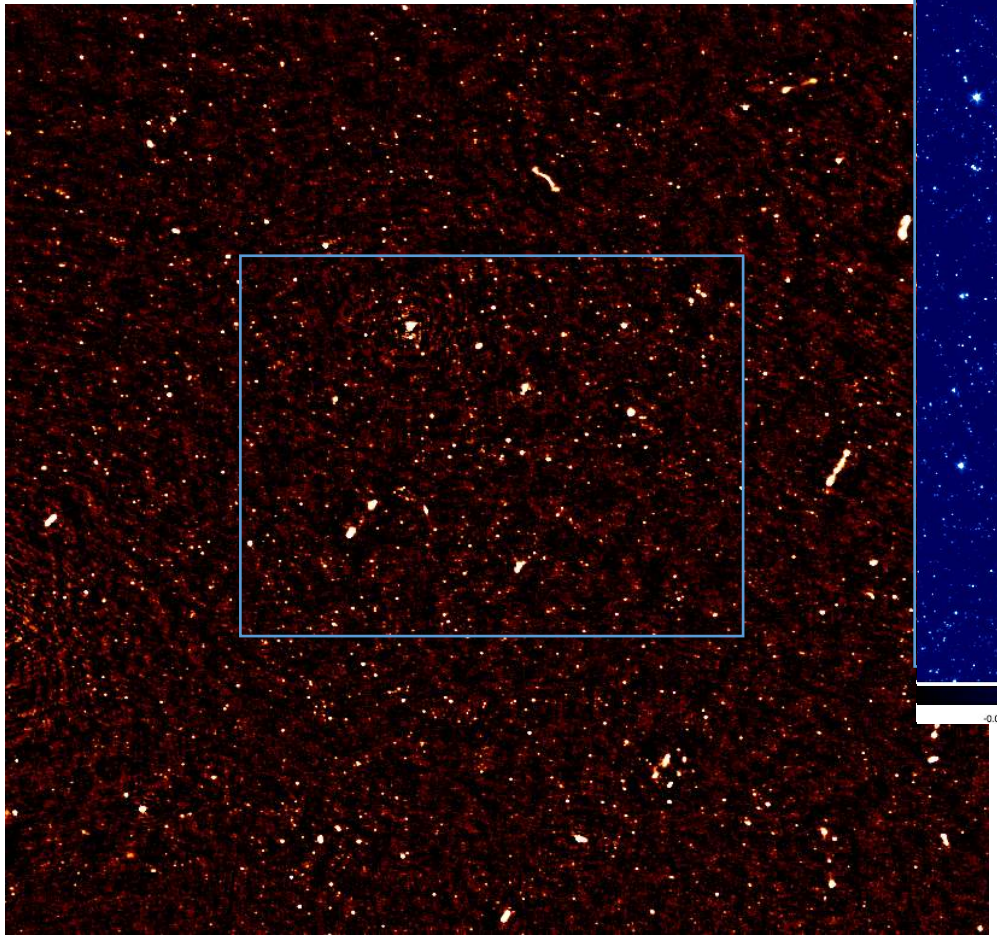


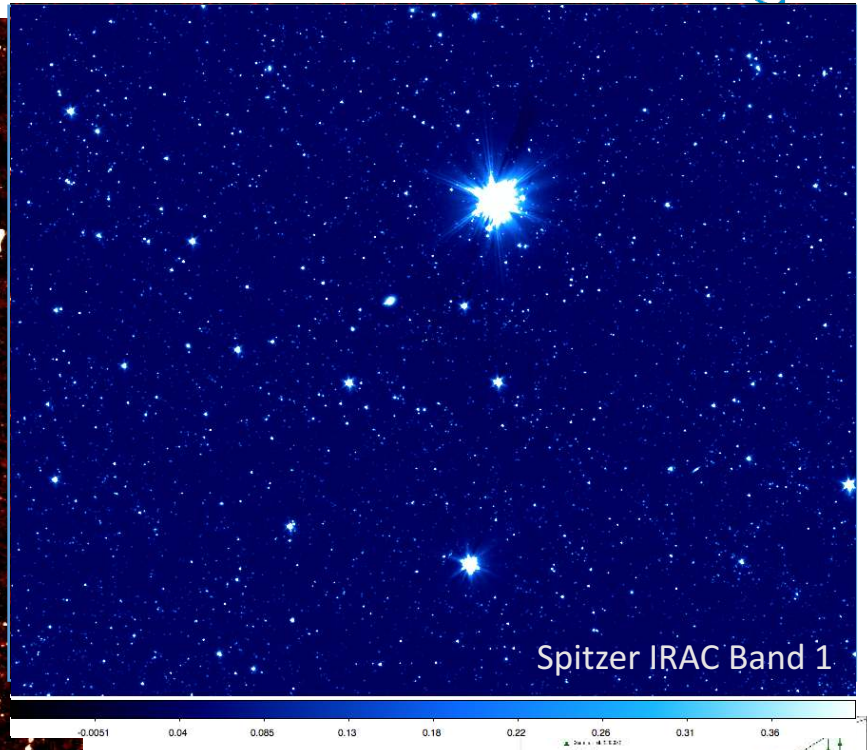


GMRT 610MHz
Resolution 5”
RMS = 10.3 microJy

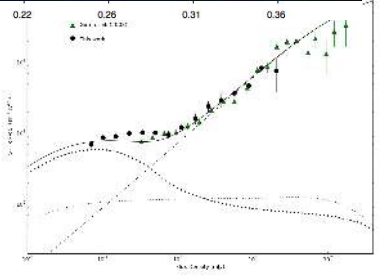


What a MeerKAT Continuum image will look like

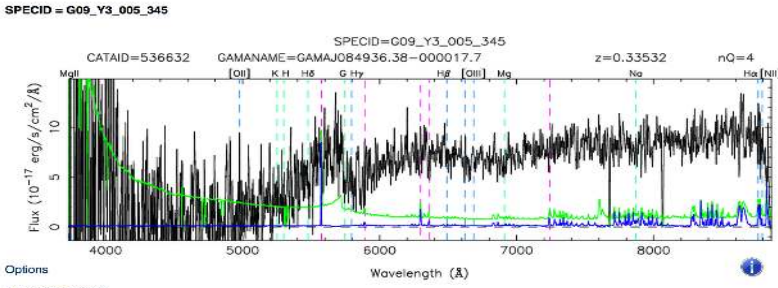




GMRT 610MHz
Resolution 5”
RMS = 10.3 microJy



Fusion Large Multi-wavelength Data Sets



From SpecAllv25:

SPECID	SURVEY	Z	NQ	PROB	IC_FLAG	DIST	IS_SBEST	IS_BEST	URL
G09_Y3_005_345	GAMA	0.3353	4	0.9890	4104	0.0500	1	1	Download file

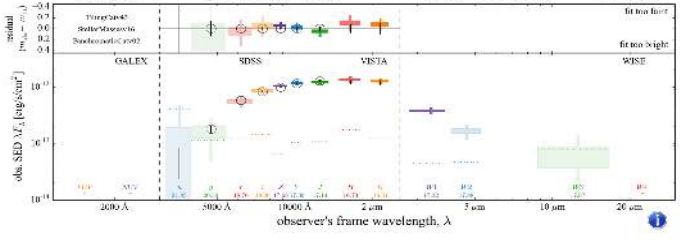
Other spectra of this object:

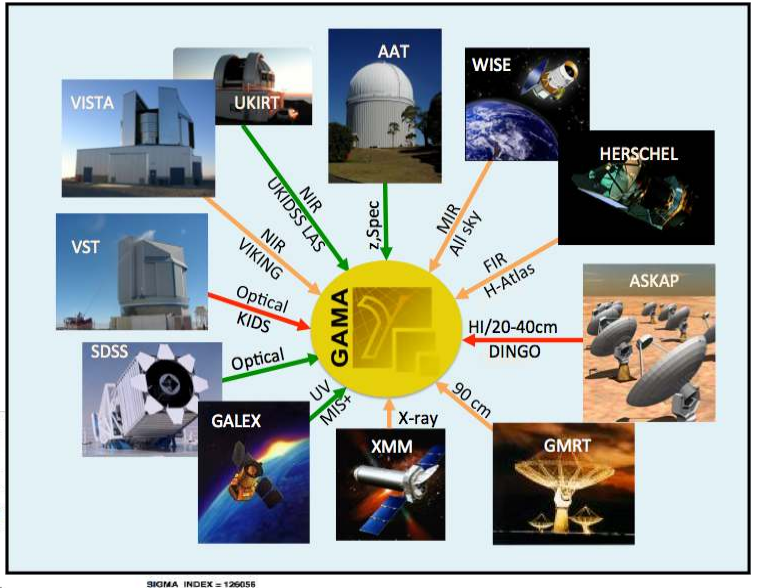
SPECID	SURVEY	Z	NQ	PROB	DIST	IS_SBEST	IS_BEST	URL	URL_IMG
429214988994714009 6	SDSS	0.3358	4	0	0.1800	1	0	Download file	View image
525797212534368256	SDSS	0.5810	1	0	0.1100	0	0	Download file	View image
G09_Y2_014_355	GAMA	0.3356	2	0.5950	0.0500	0	0	Download file	View image

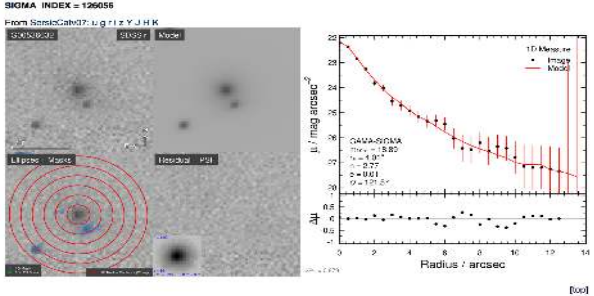
[top]

From 20BandPhotomv02 and StellarMassesv16:

CATAID = 536632 logmstar = 11.41 +/- 0.10 logLWage = 9.63 +/- 0.23 S2N = 33.5
Z = 0.3353 gminusi = 1.21 +/- 0.06 extBV = 0.16 +/- 0.16 PPP = 0.59







No group information available for this object

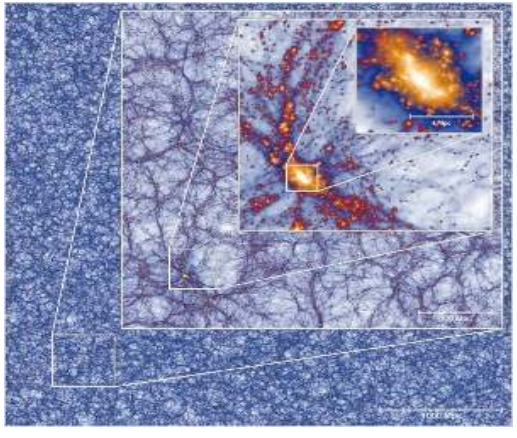
Michelle Cluver, Mattia Vaccari (UWC), Tom Jarret (UCT)

Comparison to Multi-Scale Universe Simulations

Big Data meets big simulations

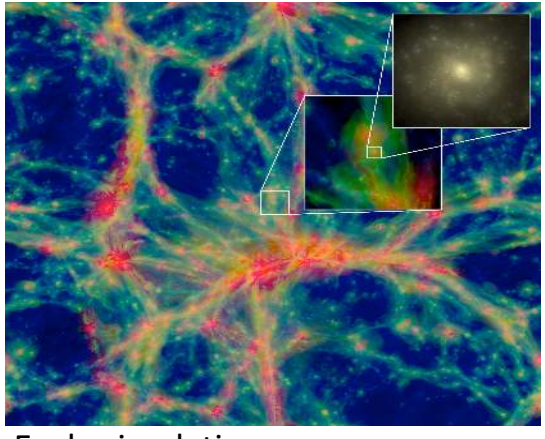
- “Hubble volume” simulations (~Gpc):
- “Cosmological” simulations (~100Mpc):
- “Zoom” simulations (~1 Mpc)

3000 Mpc



Millennium XXL simulation

100 Mpc



Eagle simulation

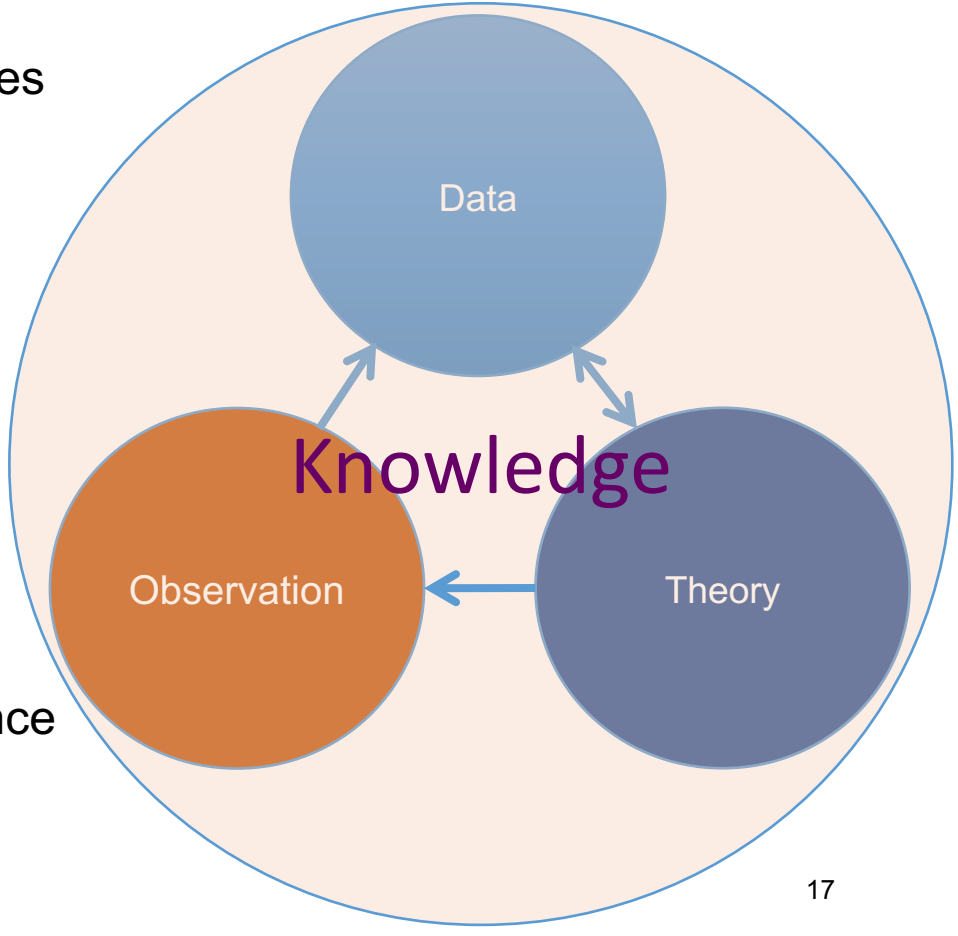
100 kpc

FIRE simulation

Romeel Dave (UWC)

The Challenges: data to knowledge

- Exponential increase in rates and volumes
- Complex, multi-purpose, processing and analysis for science information, data mining and exploration
- Fusion of big multi-wavelength data
- Fusion of big observational data with big simulations
- Collaborative execution of big data science projects by globally distributed teams of researchers



IDIA

Inter-University Institute for Data Intensive Astronomy

from big data to big ideas

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

UNIVERSITY of the WESTERN CAPE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

IDIA Launched September 3, 2015

- Three founding universities – NWU, UCT, UWC
- University of Pretoria joined 14 February 2016
- SAP Associate Member July 2016
- Open to additional partners

IDIA Goals

- **Build data science capacity and expertise at SA partner universities** for leverage the SA investment in MeerKAT as a precursor to SKA science
- **Develop SKA-driven multi-disciplinary, multi-university, data-intensive research and training programs** bringing together astronomy, statistics, computer science, eResearch,…
- **Develop data-centric computational platforms** to enable data-intensive research in the new research paradigm
- **Build strategic relationships to transfer and exchange knowledge and training** between sectors and domains.

Major Research Themes

- **Design SKA data access and delivery systems and architectures**
  - Task Leadership for SKA SDP “Data Delivery” work package
  - Prototyping of Precursor SKA Regional Science and Data Centres
  - Federated African Data Intensive Research Cloud (ADIRC)
- **Middleware cyber-platforms** for collaborative research with remote and distributed Big Data
- **Data processing algorithms and software** in support of MeerKAT large survey projects .
- **Data Science Research** for data mining and knowledge extraction for MeerKAT and SKA Key Science

SKA Precursor Regional Science Data Centres

- MoU to collaborate on development of Precursor SKA RSDC
  - **NL, SKA-SA, IDIA**
- Bring together MeerKAT and LOFAR key science

MoU Signed 17 November 2015

EU Horizon 2020 Project Approved

Lead by ASTRON in the Netherlands

28 participants, including 3 in South Africa

- IDIA
- CSIR
- NRF (SKA-SA)

A proposal in response to H2020 INFRASUPP-3-2016-2017 (Part A)

Design and specification of a distributed, European Science Data Centre (ESDC) to support the pan-European astronomical community in achieving the scientific goals of the SKA.

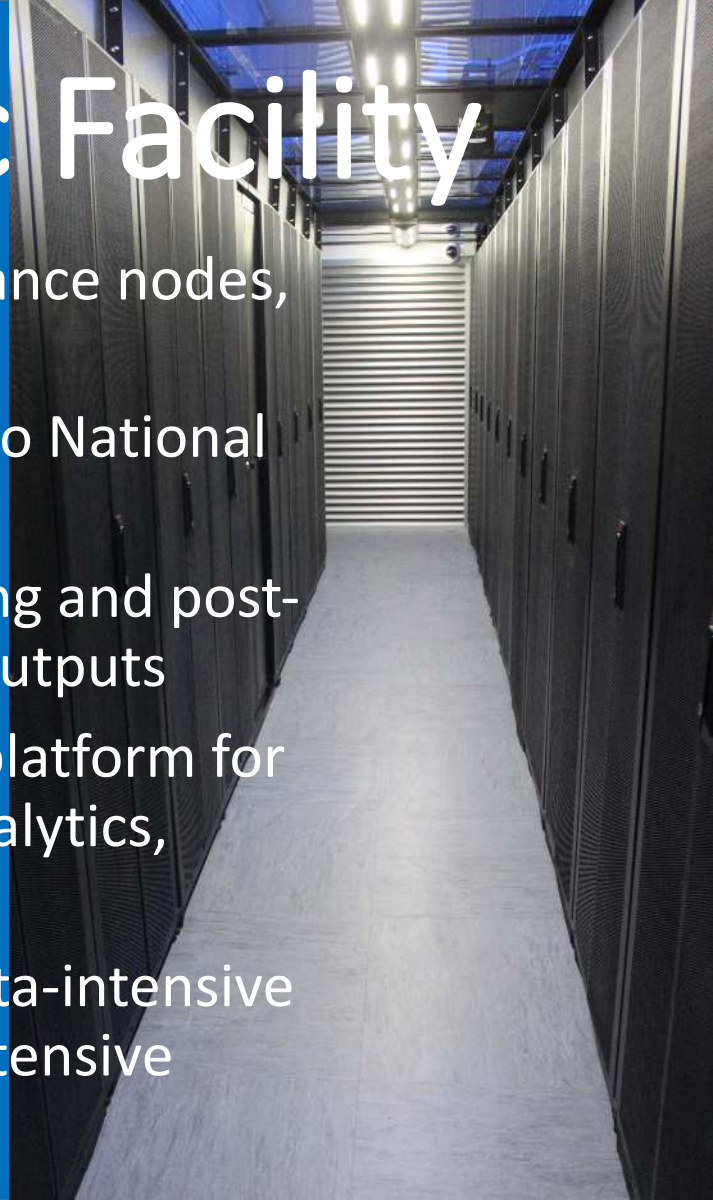
US National Radio Astronomy Observatory

- Collaboration of development for data-intensive radio astronomy projects and visualization for large radio astronomy data cubes
- Software system used for processing for Jansky Very Large Array and Atacama Large Millimetre Array and MeerKAT

Signed 17 January 2016

Tier2 Data-Centric Facility

- 8-10 PB tiered storage, 80 high-performance nodes, GPU systems
- housed at UCT with 10GB/s connection to National Data transport ring
- development platform for data processing and post-processing algorithms and data mining outputs
- Cloud enabled services and distributed platform for data access, apps to data, workflows, analytics, visualization,…
- Part of a national, distributed, tiered, data-intensive research infrastructure – African Data Intensive Research Cloud



SKA Precursor Regional Science Data Centres

MeerKAT and LOFAR data and use cases

MeerKAT

MeerKAT Telescope (SKA SA)

- generate and manage telescope data
- First Stage processing
  - flagging
  - Near-real time calibration and imaging
  - Data validation

T1 data store
- calibrated and averaged visibilities
- Image repository

Science products

T1 processed

Tier 2 Facility (University Partners)

- Project-based data extraction from T1 data store
- Processing aggregate data to scientific image data sets
- Post-processing, analytics
- Visualization and data mining
- Platform co-development for global data intensive project collaboration and data sharing.

Global nodes

LOFAR

Research and Development Collaboration

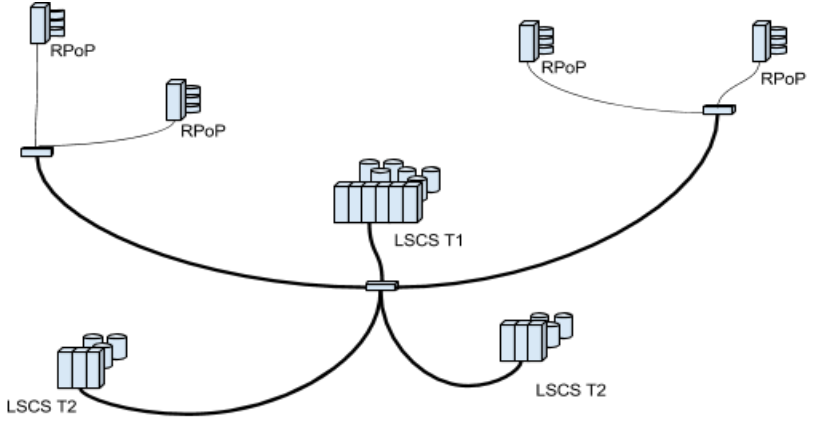
African Data Intensive Research Cloud

- Cloud-based, distributed platform for data access and data intensive research.
- tiered, federated cloud around data, software, analytics, visualization, collaboration
- Proto-type and test bed under development among IDIA partner universities

IST-Africa 2016 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2016
ISBN: 978-1-905824-54-0

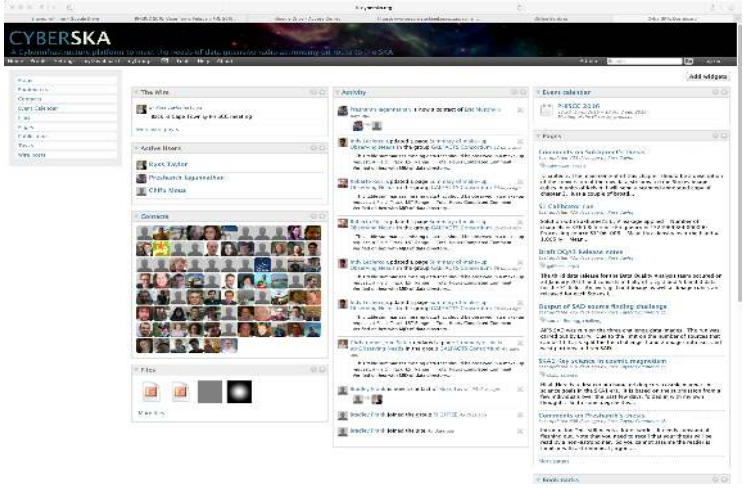
The African Data Intensive Research Cloud

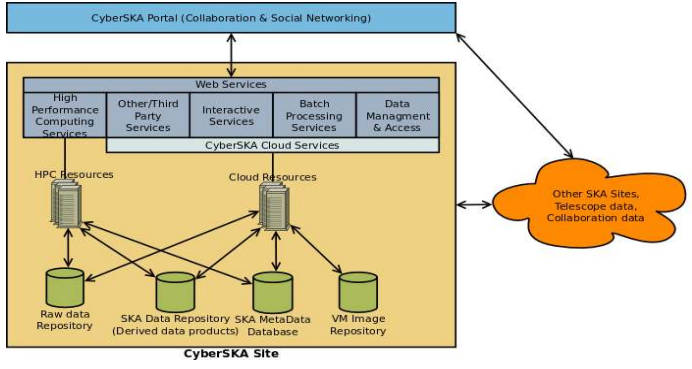
Rob SIMMONDS[1], Russ TAYLOR[2,3], Jasper HORRELL[4], Bernie FANAROFF[5], Happy SITHOLE[6], Sakkie JANSE VAN RENSBURG[7], Boeta PRETORIUS[8]



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



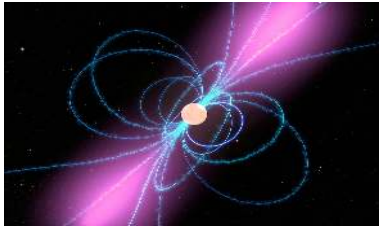


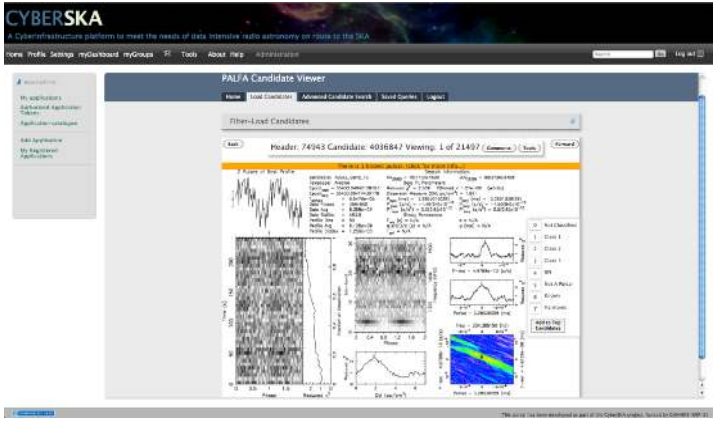
CyberSKA: A cloud-enabled Big Data Research Platform

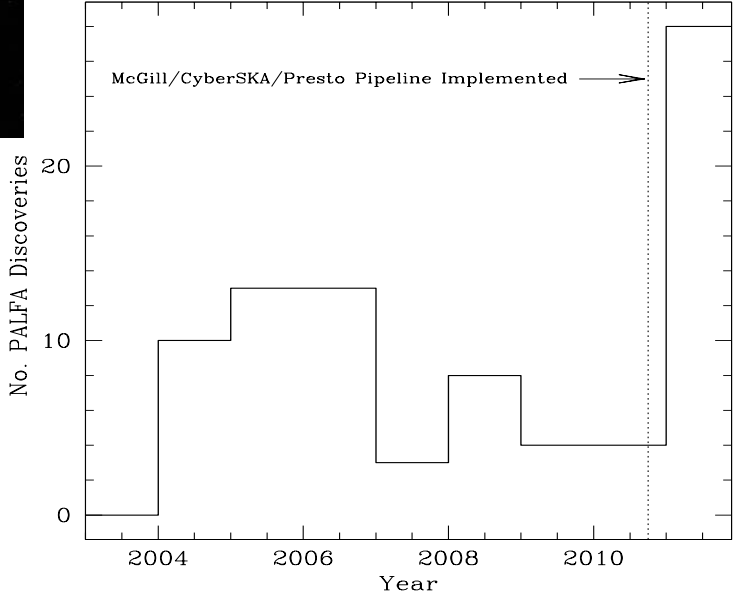
- 680 members from around the world

CyberSKA: Enabling Discovery

- PALFA: Millisecond Pulsar Search – 117 global collaborators





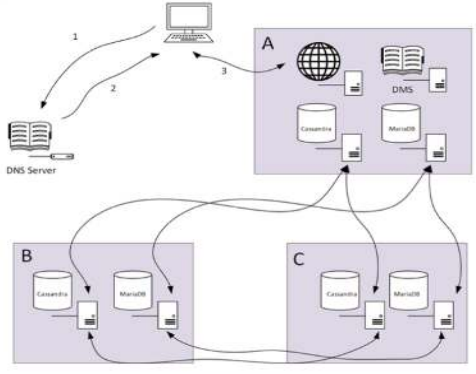


Federated CyberSKA Platform

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

IDIA Data Science Workshop, 12-13 April 2016

- Image and Time Domain Data Processing -> **science data products**
- Post-processing and extraction of information -> **cataloging**
- Visualization and visual analytics -> **exploration, seeing into the data**
- Data Mining and Multi-wavelength Science -> **contextualization**
- Simulations and comparison of data and theory -> **interpreting the data**
- Exploration of the Unknown -> **what the heck?**

Training Data Scientists

- New M.Sc. in Data Science and Big Data at UCT and UP
- IDIA Postgraduate and postdoctoral bursaries for multi-disciplinary “astro-informatics” research projects
- Cross-sector and multi-institution development teams for R&D in new data systems and technologies
- Partnerships with industry for sponsorship of postgraduate and postdoctoral internships
- collaboration with SA undergraduate data science programs (e.g. Sol Plaatje University) for IDIA projects in data science


The Square Kilometre Array
2021
Data to Knowledge
(People, facilities, systems, algorithms)