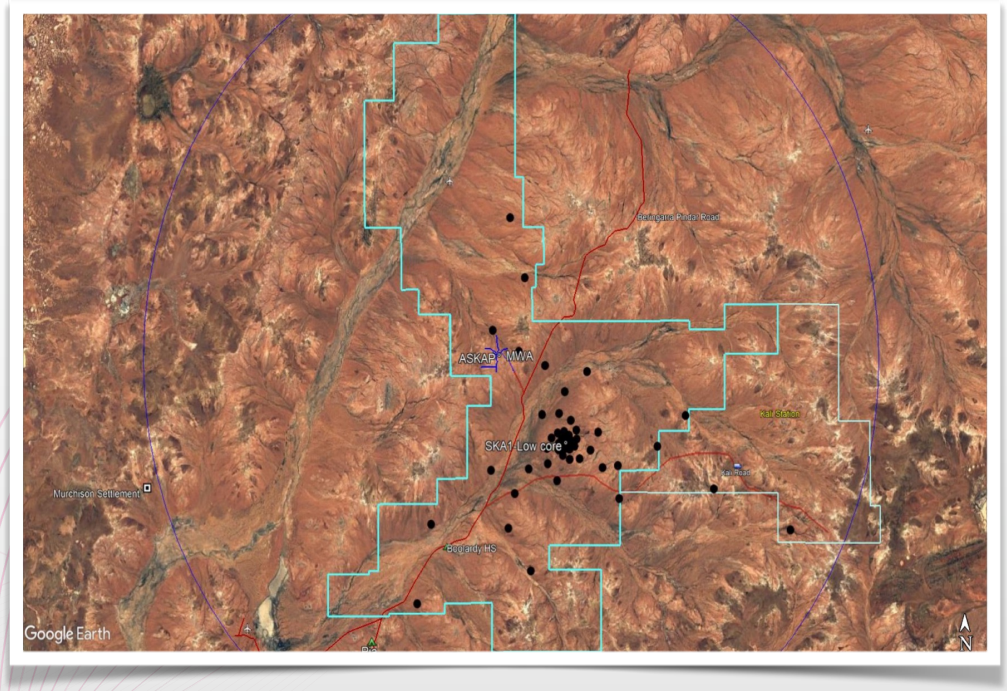
SKA and its data challenge

Dr A. Bonaldi, SKAO Project scientist

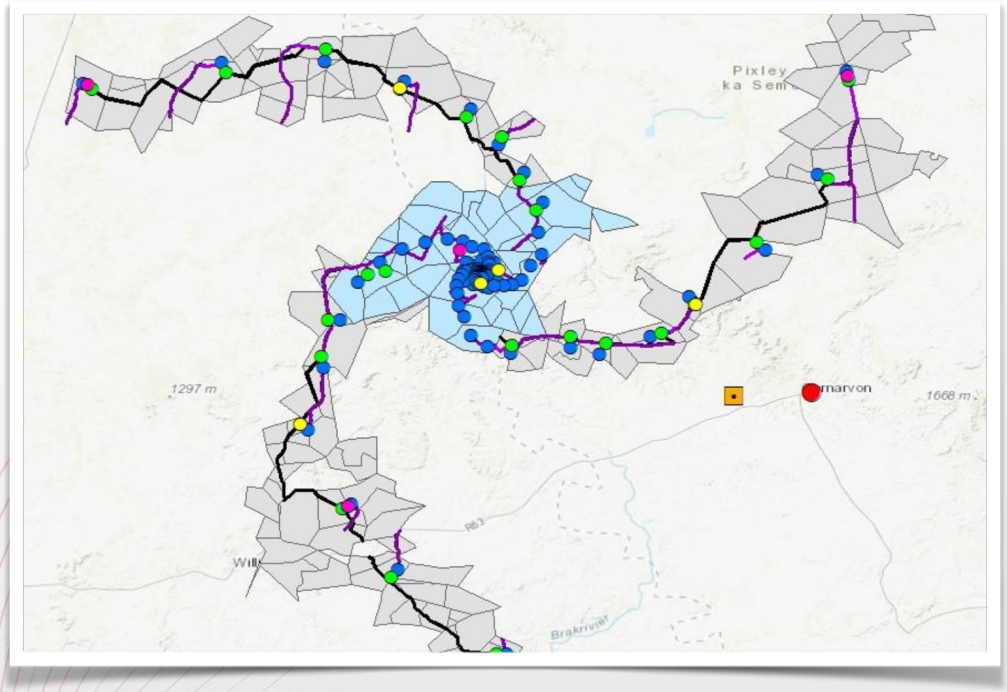
ADA X summer school, Crete, 18-22 September 2023

SKA-Low in Australia

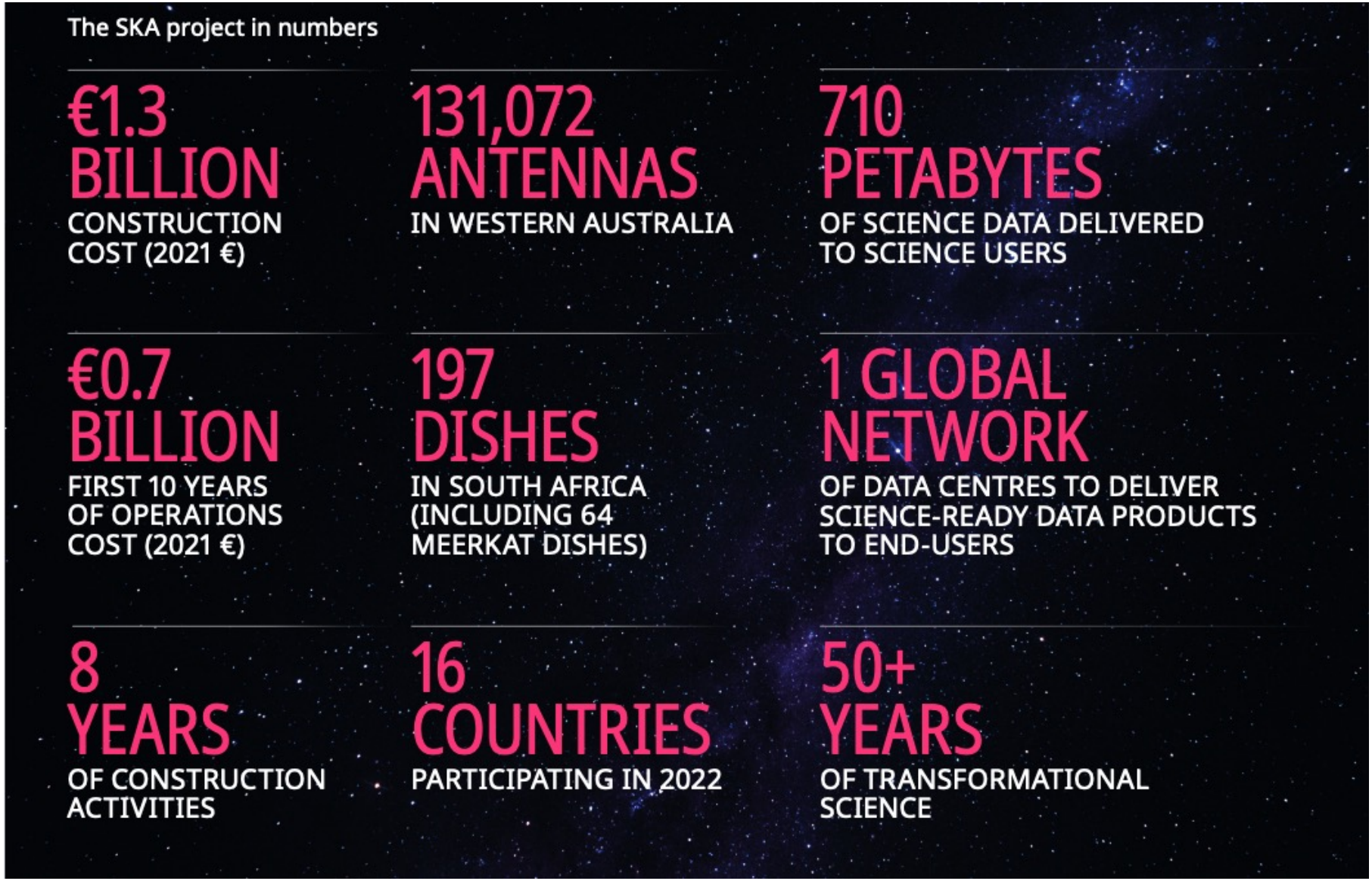


- 131,072 log-periodic antennas, spread across 512 stations
- Frequency range: 50 MHz - 350 MHz
- Wavelength range: 6 m – 0.85 m
- Maximum distance between antenna stations: 74km

SKA-Mid in South Africa



- 197 fully steerable dishes, including the existing MeerKAT dishes
- Frequency range: 350 MHz - 15.4 GHz
- Wavelength range: 0.85 m – 2 cm
- Maximum distance between dishes: 150km

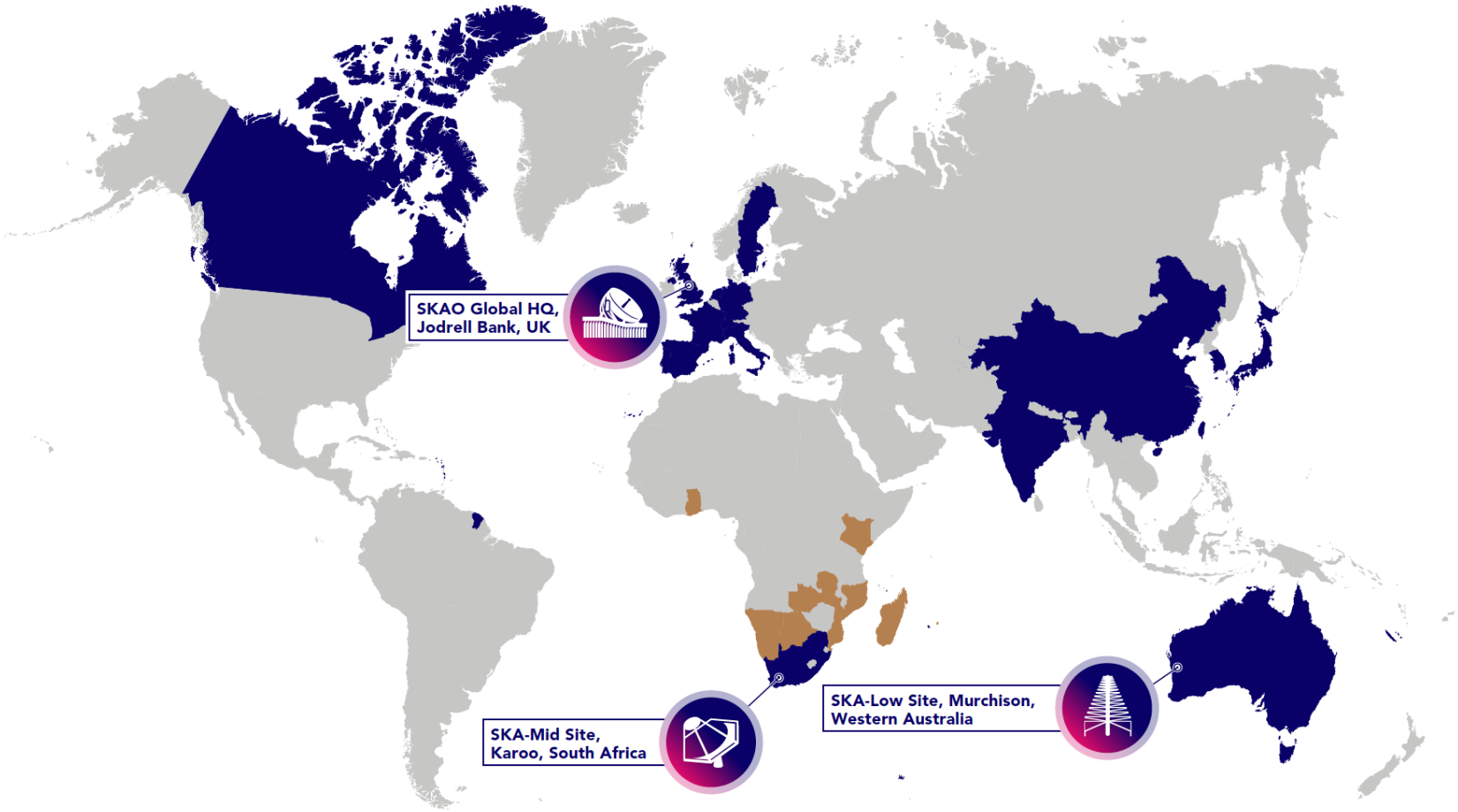

The SKA project in numbers
€1.3 BILLION
CONSTRUCTION COST (2021 €)
131,072 ANTENNAS
IN WESTERN AUSTRALIA
710 PETABYTES
OF SCIENCE DATA DELIVERED TO SCIENCE USERS
€0.7 BILLION
FIRST 10 YEARS OF OPERATIONS COST (2021 €)
197 DISHES
IN SOUTH AFRICA (INCLUDING 64 MEERKAT DISHES)
1 GLOBAL NETWORK
OF DATA CENTRES TO DELIVER SCIENCE-READY DATA PRODUCTS TO END-USERS
8 YEARS
OF CONSTRUCTION ACTIVITIES
16 COUNTRIES
PARTICIPATING IN 2022
50+ YEARS
OF TRANSFORMATIONAL SCIENCE

Construction Schedule

Construction commencement ceremonies, Dec 5-6 2022

Milestone Event		SKA-Mid	SKA-Low
AA0.5	4 dishes 6 stations	2025 Q1	2024 Q4
AA1	8 dishes 18 stations	2026 Q1	2025 Q4
AA2	64 dishes 64 stations	2027 Q1	2026 Q4
AA*	144 dishes 307 stations	2027 Q4	2028 Q1
Operations Readiness Review		2028 Q1	2028 Q2
End of staged delivery programme		2028 Q3	2028 Q3
Full SKA	197 dishes 512 stations	TBD	TBD

First science verification expected in 2026/27

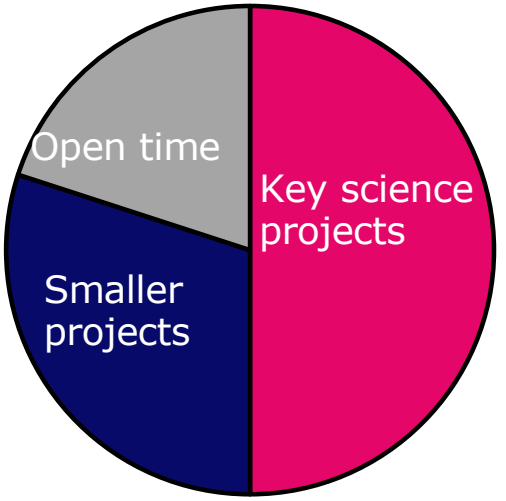


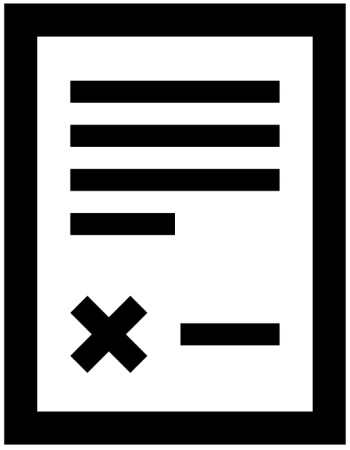
SKAO Partnership - includes SKAO Member States* and SKAO Observers (as of June 2022)

African Partner Countries

Funding model

- SKAO member countries contribute to the SKAO construction and operations cost at an agreed level
- Telescope access is based on contribution level
- Construction contracts awarded to member countries whenever possible, to guarantee fair return of investment





Examples of impact of investment in radio astronomy

THE INVENTION OF WIFI

MAGNETIC RESONANCE IMAGING (MRI)

REFERENCE SYSTEMS FOR SPACE NAVIGATION AND GP

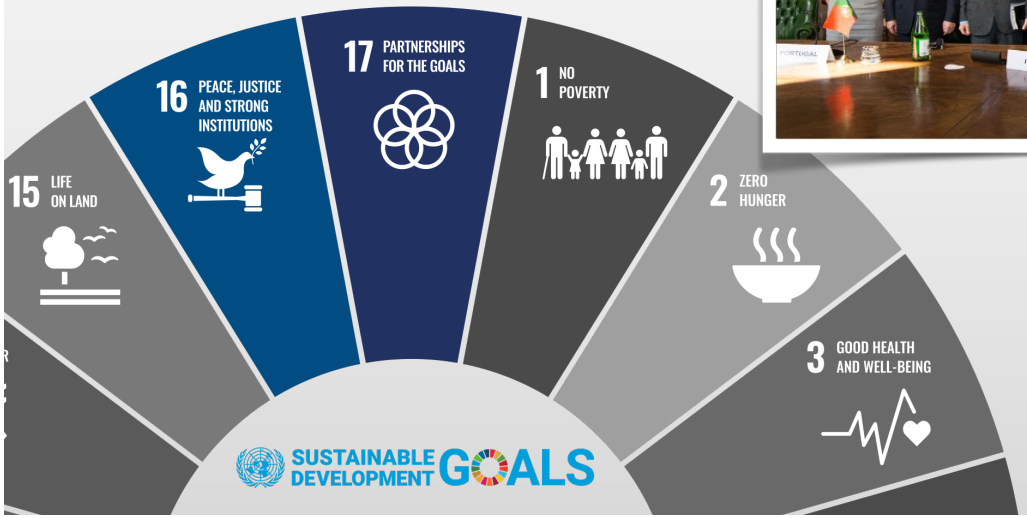
HIGH-PRECISION MONITORING OF TECTONIC PLATE MOVEMENTS

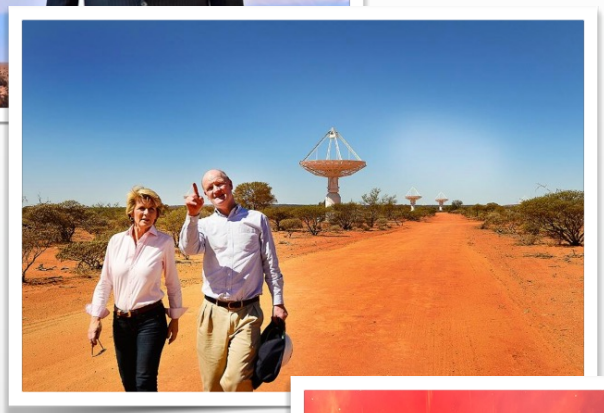
SPACE TRACKING


Extragalactic Continuum
Science Working G
Cosmology
Science Working G
Cradle of Life
Science Working G
Pulsars
Science Working
Epoch of Reionization
Science Working G
HI Galaxy Science
Science Working G
Extragalactic Spectral Lines
Science Working G
Transients
Science Working G
Cosmic Magnetism
Science Working G
Solar and Heliospheric Physics
Science Working G
Our Galaxy
Science Working G
High Energy Cosmic Particles
Focus Group
SQUARE KILOMETRE ARRAY
ADVANCING ASTROPHYSICS with the SQUARE KILOMETRE ARRAY
VOLUME 1
SKA ORGANISATION
ADVANCING ASTROPHYSICS with the SKA
VOLUME 2

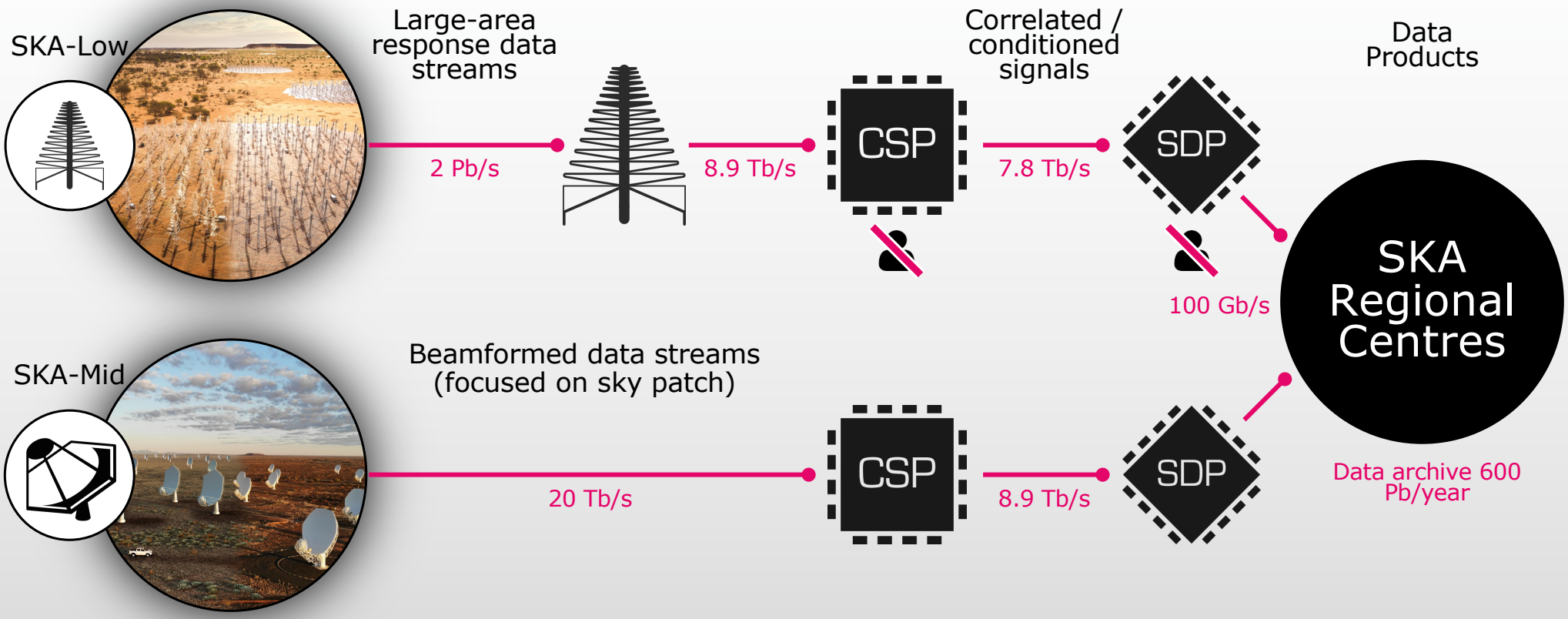
SKAO partnership as a science diplomacy tool

- Build international connections
- Encouraging government-level interaction
- A vehicle for collaboration

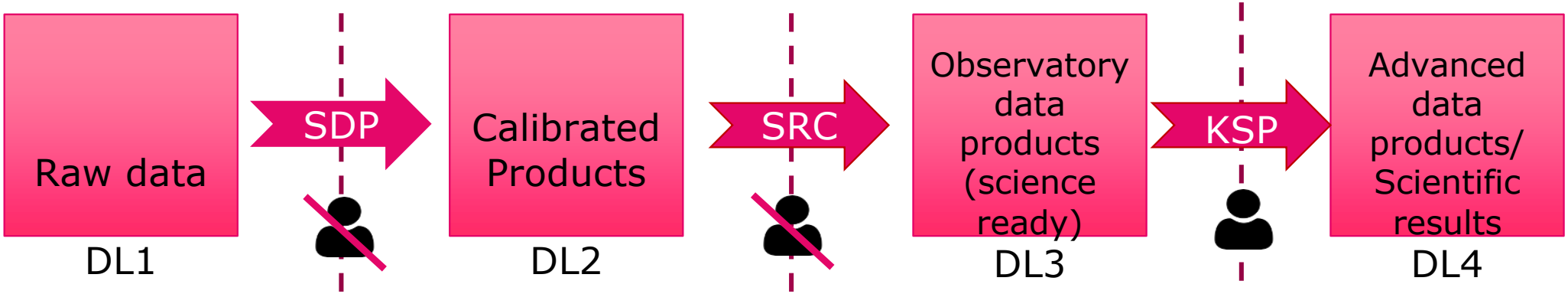




SKAO data processing stages



SKA data journey: Data Layers (DL)



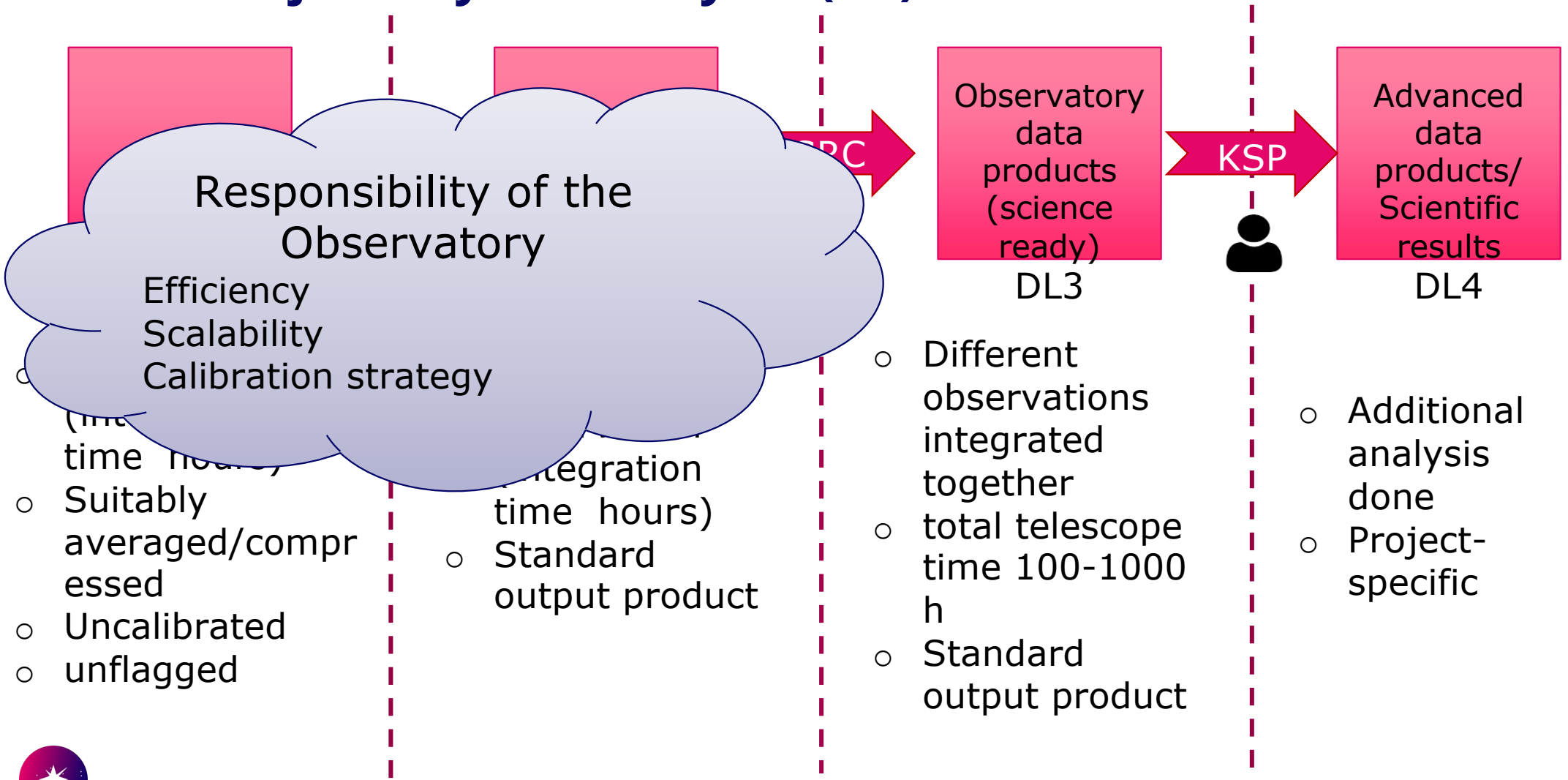
- per observation (integration time hours)
- Suitably averaged/compressed
- Uncalibrated
- unflagged

- per observation (integration time hours)
- Standard output product

- Different observations integrated together
- total telescope time 100-1000 h
- Standard output product

- Additional analysis done
- Project-specific

SKA data journey: Data Layers (DL)



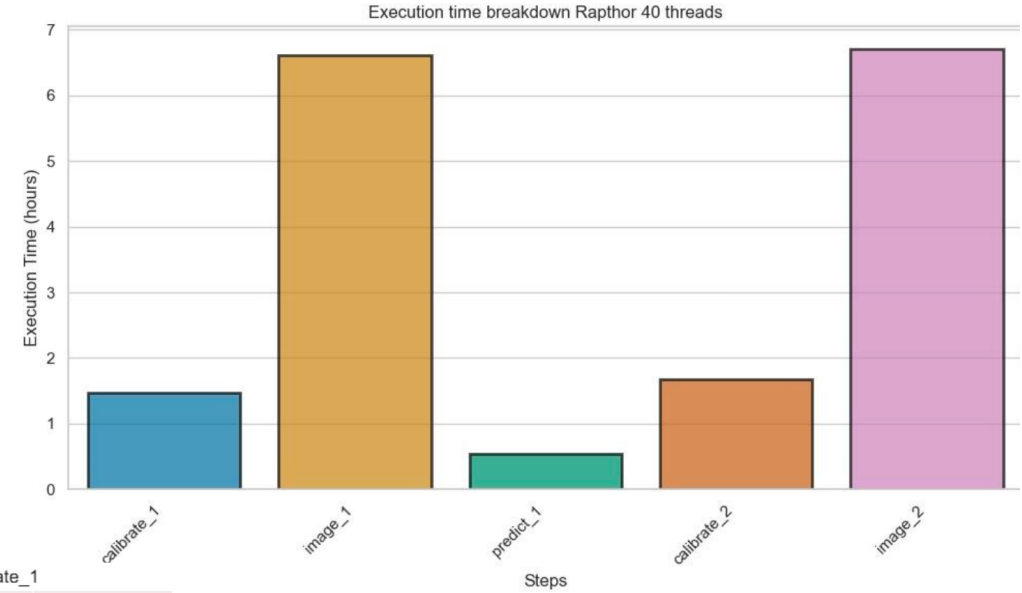
Paving the way

SDP: scalability of calibration and imaging

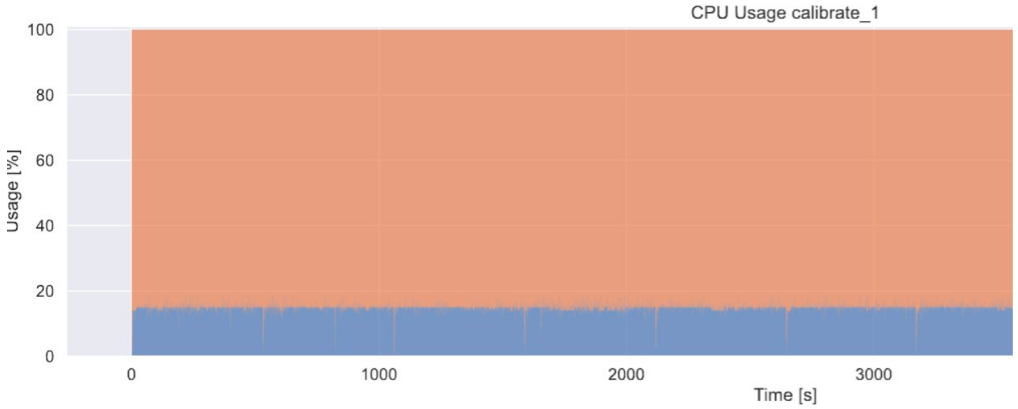
Slide credit: Shan Mignot

Execution metrics

- Low CPU usage
  - < 20% for DP3
  - < 60% for WSClean



execution time for the different steps of the pipeline (iterations 1 & 2)

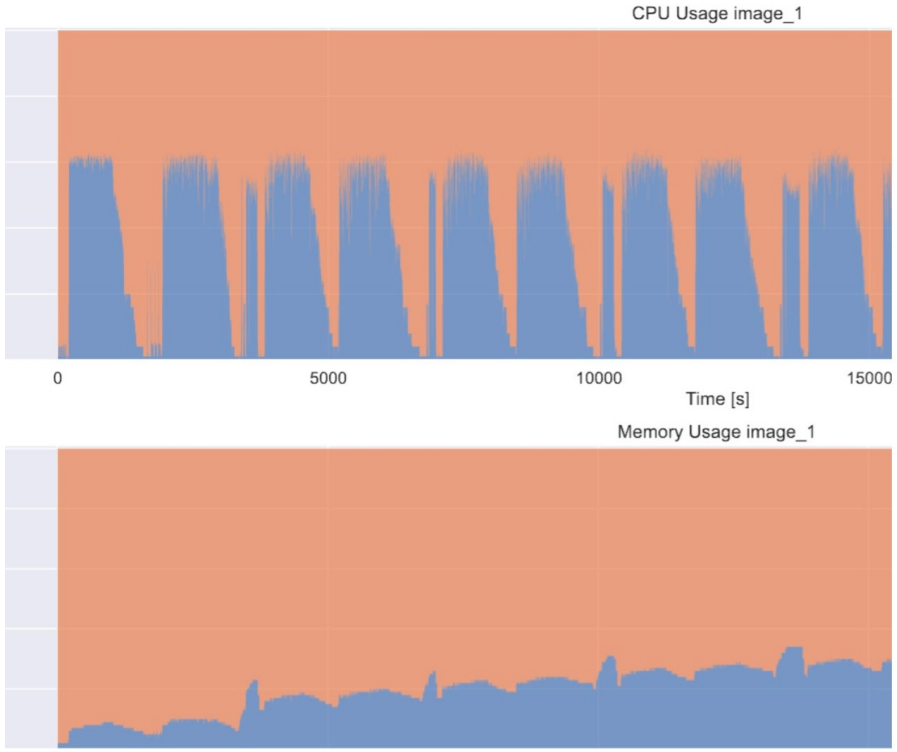


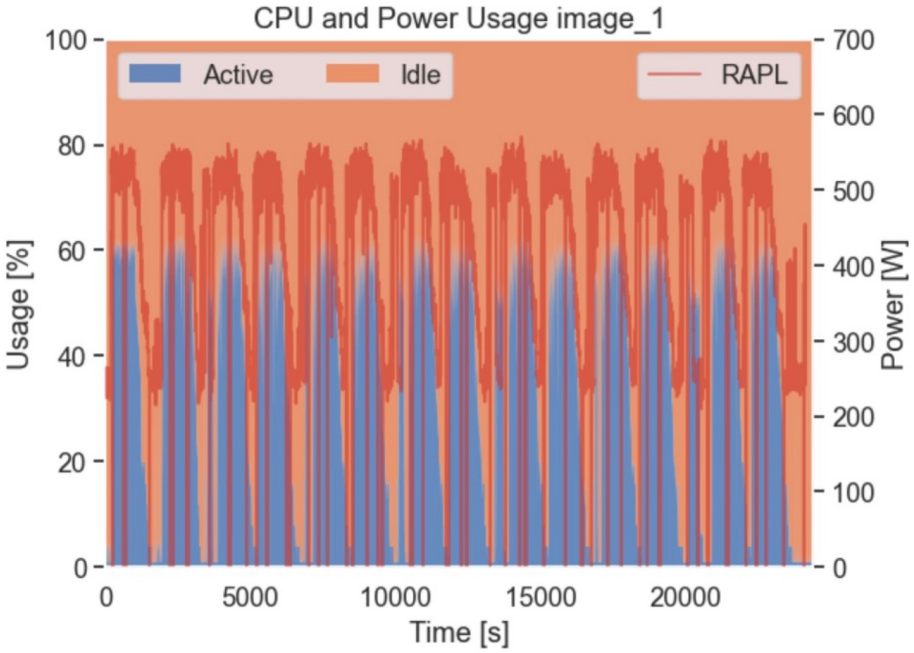
core usage for calibrate_1 (DP3)

SDP: scalability of calibration and imaging

Slide credit: Shan Mignot

Execution metrics (II)



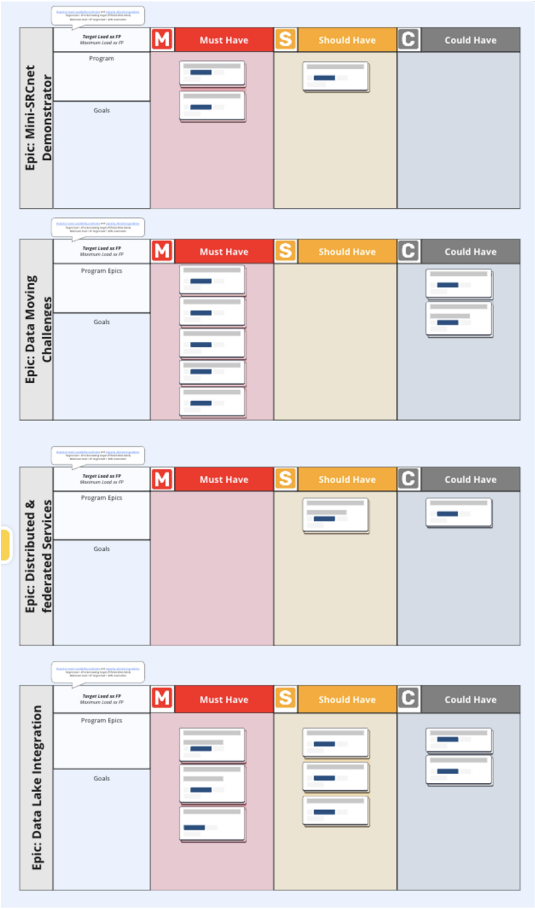


resource usage recordings for the image_1 step (WSClean)

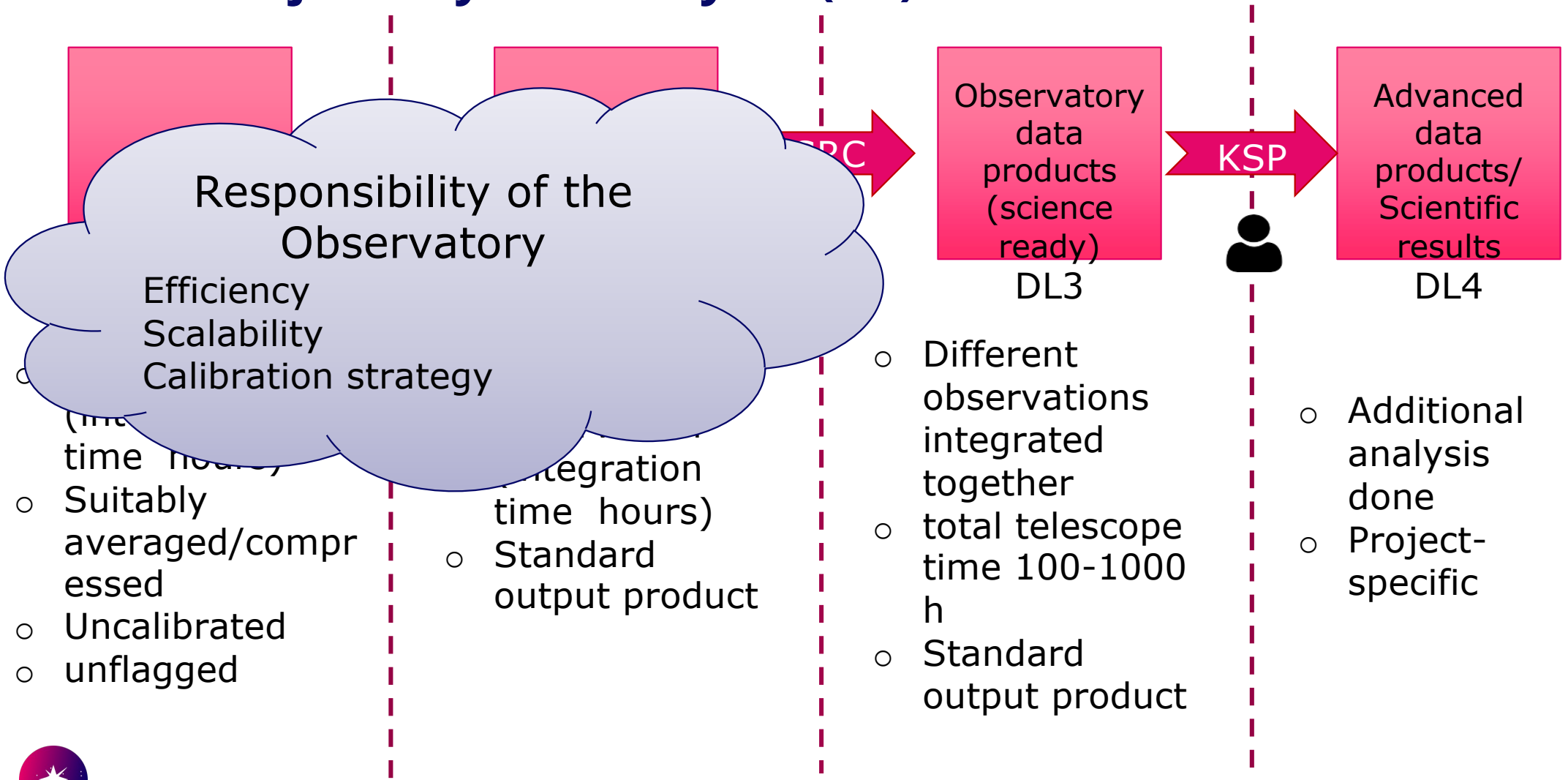
SRCNet prototyping efforts

Some of the current work areas:

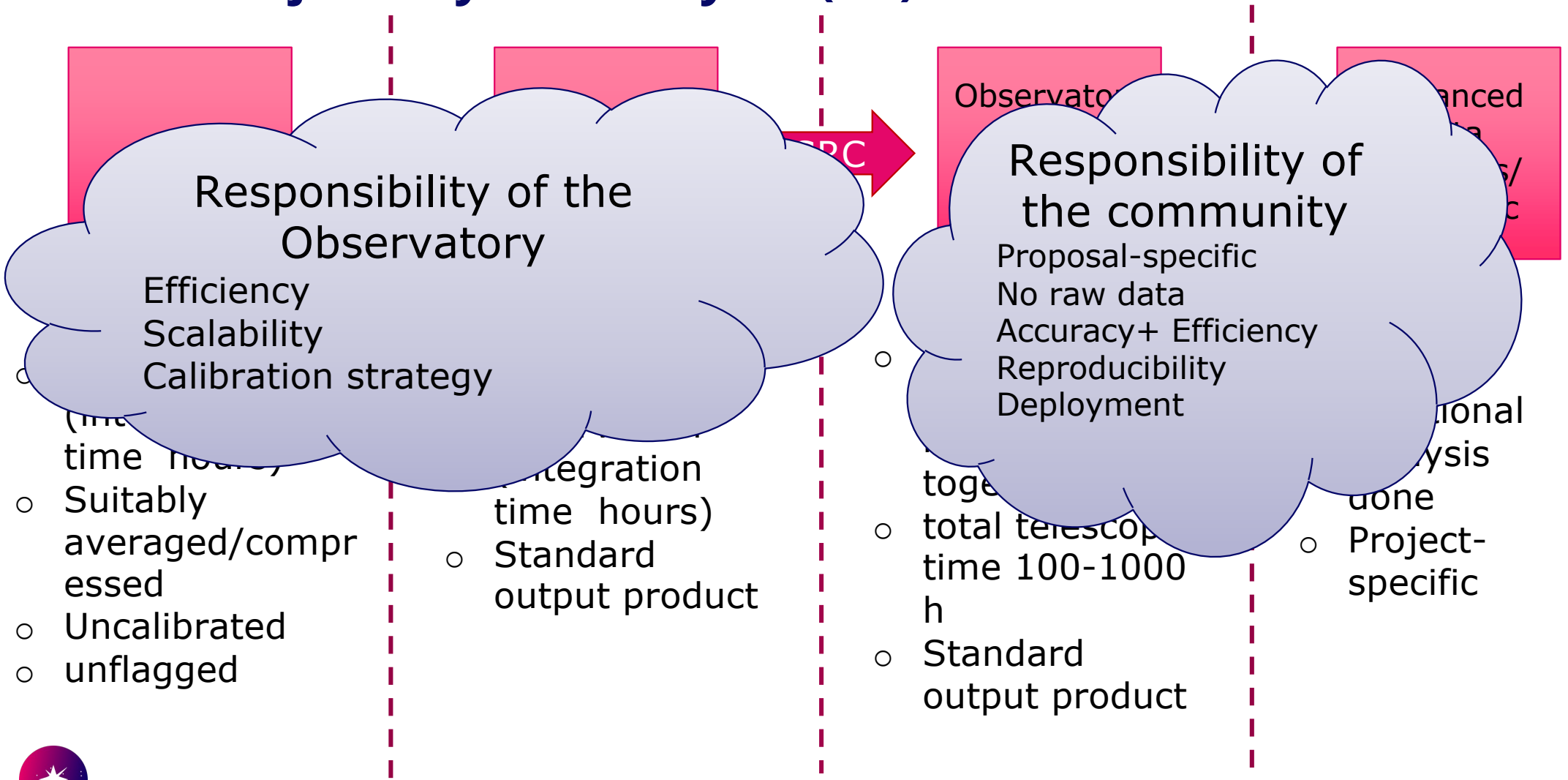
- Distributed and federated services
- Data lake integration
- AAI
- Mini-SRC demonstrator
- Software distribution
- Example workflows
- SRC Science Analysis platform
- Data moving challenges



SKA data journey: Data Layers (DL)



SKA data journey: Data Layers (DL)



Community training

- SKA regional centre training event (27/2-14/2 2022)
  - Hands-on Containerization
  - Lessons and tutorials on Github, Gitlab, Containers (docker, singularity)
- IAA-CSIC Severo Ochoa SKA Open Science School (Granada 8-10/5/23)
  - Endorsed by SRC science-user engagement
  - Lessons on Open science, Open source, reproducibility, licenses, Open source package managers

Science Data Challenges

"The purpose of SDCs is to prepare the astronomical community, and SKAO itself, for the novel, yet challenging, nature of SKA data"

Science Data Challenges: What are we trying to achieve ?

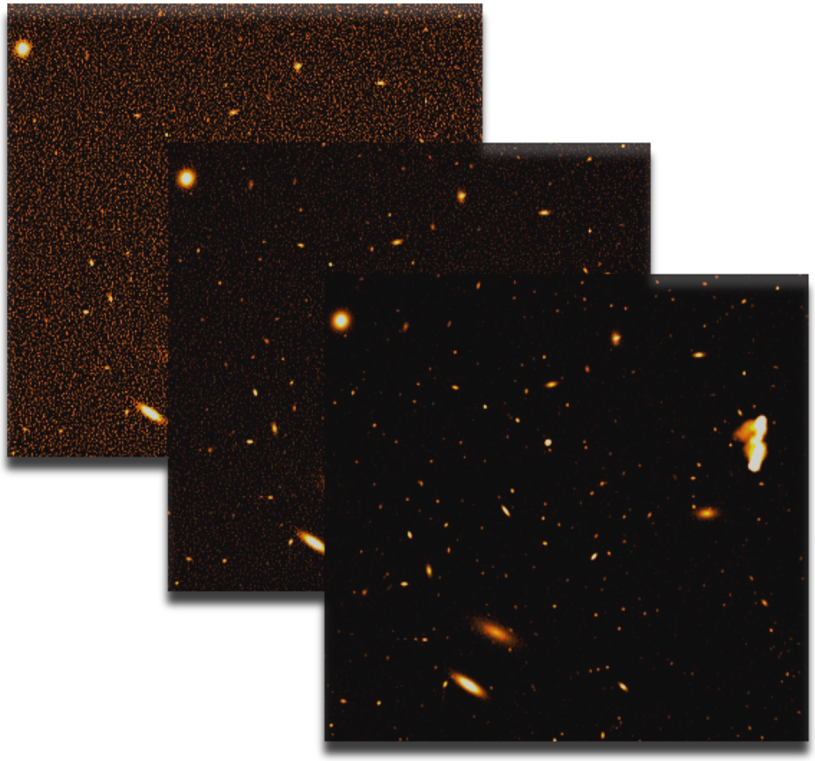
- Prepare Science Community
  - Science extraction from SKA Observatory Data Products (ODPs)
  - Stimulate advance of state-of-the-art in source finding, source characterisation and reliable inference of astrophysical parameters
  - Promote reproducibility and analysis pipeline sharing
- Develop proto-SRC Network
  - Test increasingly realistic data transfer, user access and customised user processing in proto-SRC environment
- Constrain SDP Pipeline development
  - Identify gaps in sky, telescope and error models
  - Determine necessary calibration quality and identify any other factors that might inhibit science extraction from ODPs

https://www.skao.int/en/464/ska-science-data-challenge-1

Science Data Challenge 1

Radio Continuum emission

- **Continuum emission** images, SKA MID Bands 1, 2 and 5, integrations 8, 100, 1000h
- Images populated by star forming galaxies **(SFGs)** and active galactic nuclei **(AGN)**
- High telescope sensitivity → highly **crowded** images
- **The challenge**: to **find and characterise sources**
- **Data volume = 30 GB**
- 9 competing teams



Zoom-in of the 1.4 GHz maps, showing the same region of the sky with different telescope integrations: 8, 100, 1000 h from left.

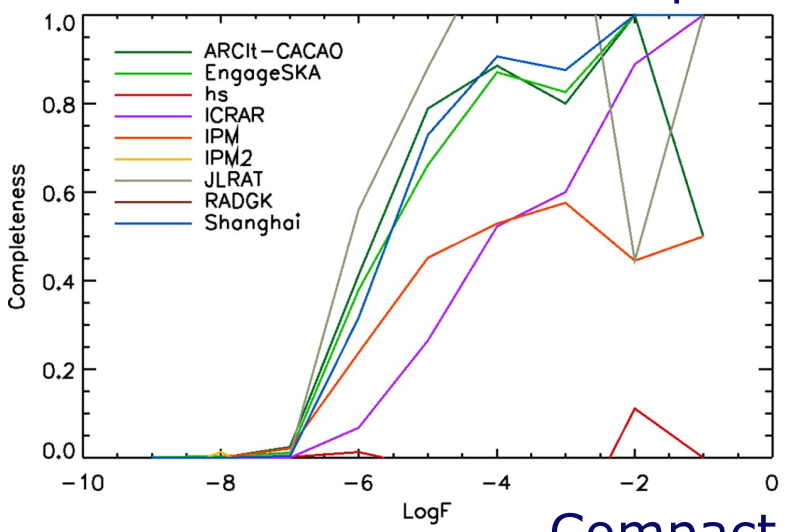
Square Kilometre Array Science Data Challenge 1: analysis and results FREE

A Bonaldi ✉, T An, M Brüggen, S Burkutean, B Coelho, H Goodarzi, P Hartley, P K Sandhu, C Wu, L Yu, M H Zhoolideh Haghighi, S Antón, Z Bagheri, D Barbosa, J P Barraca, D Bartashevich, M Bergano, M Bonato, J Brand, F de Gasperin, A Giannetti, R Dodson, P Jain, S Jaiswal, B Lao, B Liu, E Liuzzo, Y Lu, V Lukic, D Maia, N Marchili, M Massardi, P Mohan, J B Morgado, M Panwar, P Prabhakar, V A R M Ribeiro, K L J Rygl, V Sabz Ali, E Saremi, E Schisano, S Sheikhnezami, A Vafaei Sadr, A Wong, O I Wong

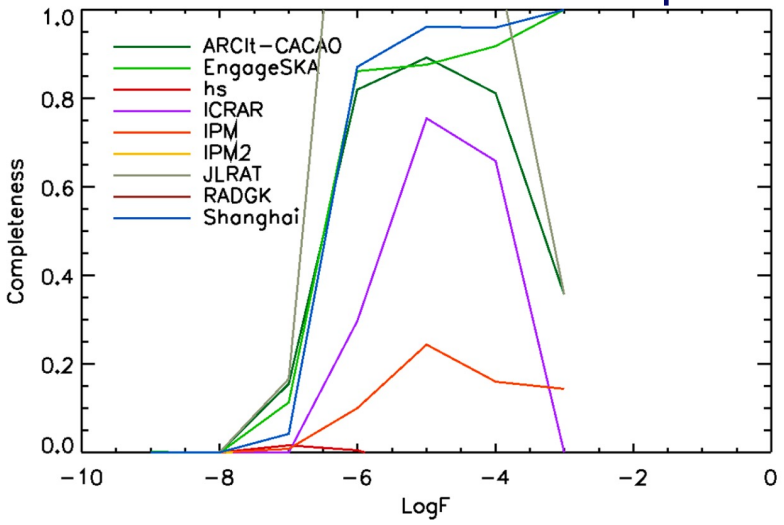
Monthly Notices of the Royal Astronomical Society, Volume 500, Issue 3, January 2021, Pages 3821–3837, https://doi.org/10.1093/mnras/staa3023

- Source crowding proved major challenge for correct source IDs and spectral classification
- Complementarity of different detection and characterisation approaches apparent
- Data tessellation to tackle data size needs to be redundant / scale-dependent
- Resolved sources with complex morphology a challenge for most methods

Complex



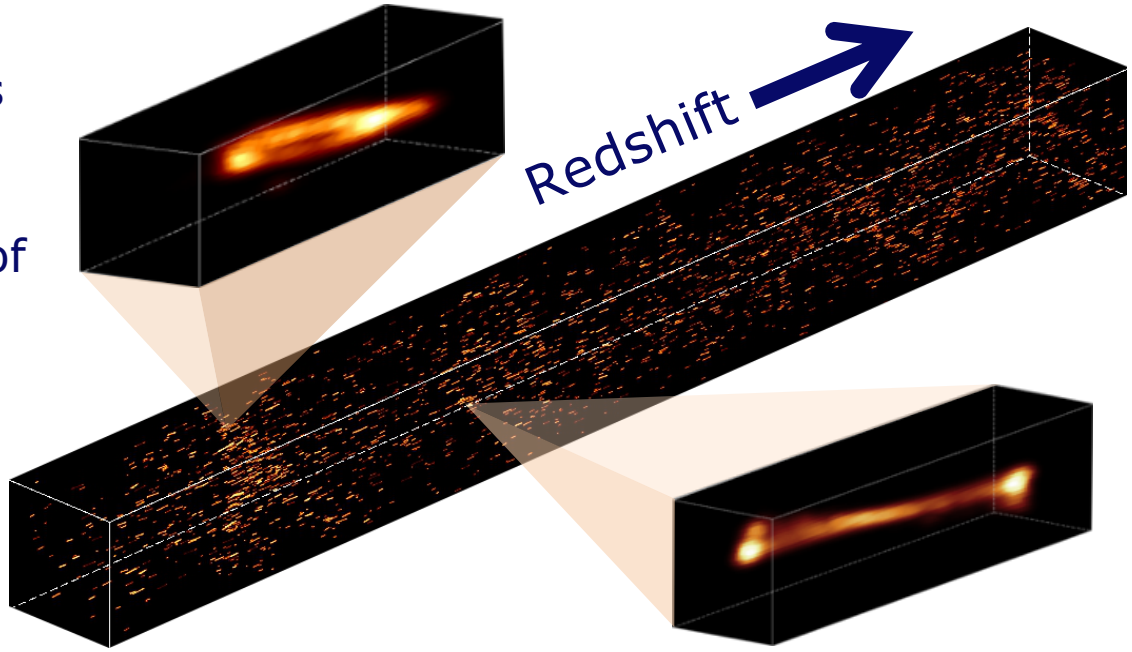
Compact



Science Data Challenge 2

Neutral hydrogen (HI) spectral line emission

- **21cm spectral line** image cube, simulating deep SKA MID observations (**redshift 0.25 to 0.5**)
- Image cube populated by **HI** content of **galaxies**
- **2000 h** integration time across **20 sq deg** field of view
- The challenge**:** to **find and characterise HI sources**
- **Data volume = 1 TB**
- 12 competing teams



Sample noise-free simulated HI image cube

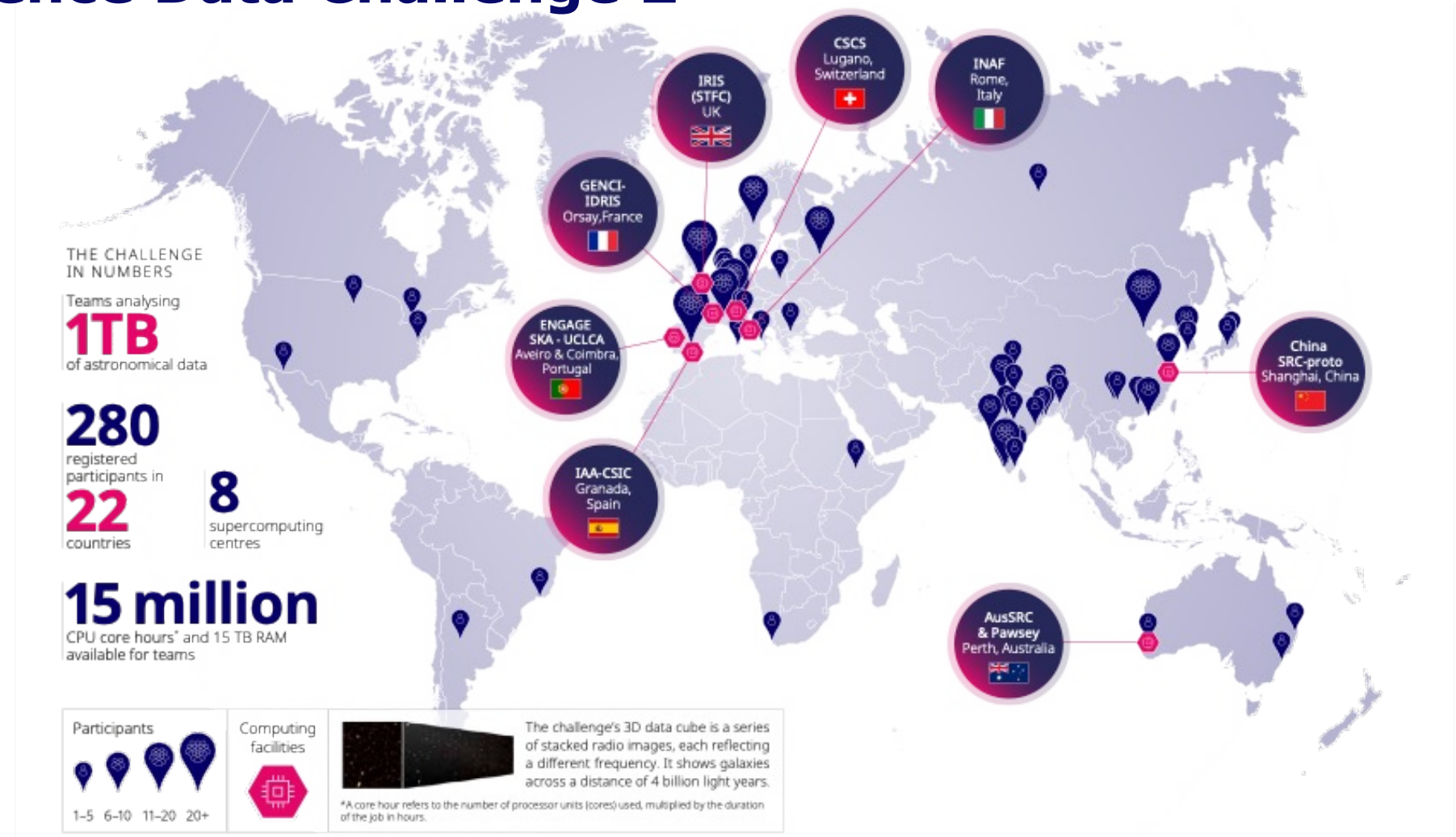
HPC Facility Partners – Why?

- Store the dataset in multiple locations, where teams will be able to access
- Provide computational resources to inspect and analyse the dataset without transferring

HPC Facility Partners – How?

- Teams state their computational needs as part of the SDC registration
- The SDC team collaborate with the facility partners to identify the best matches with teams
- Teams access the data through the chosen facility
- The data is made available at multiple facilities at the same time to ensure a fair challenge
- Teams are able to process the data there

Science Data Challenge 2

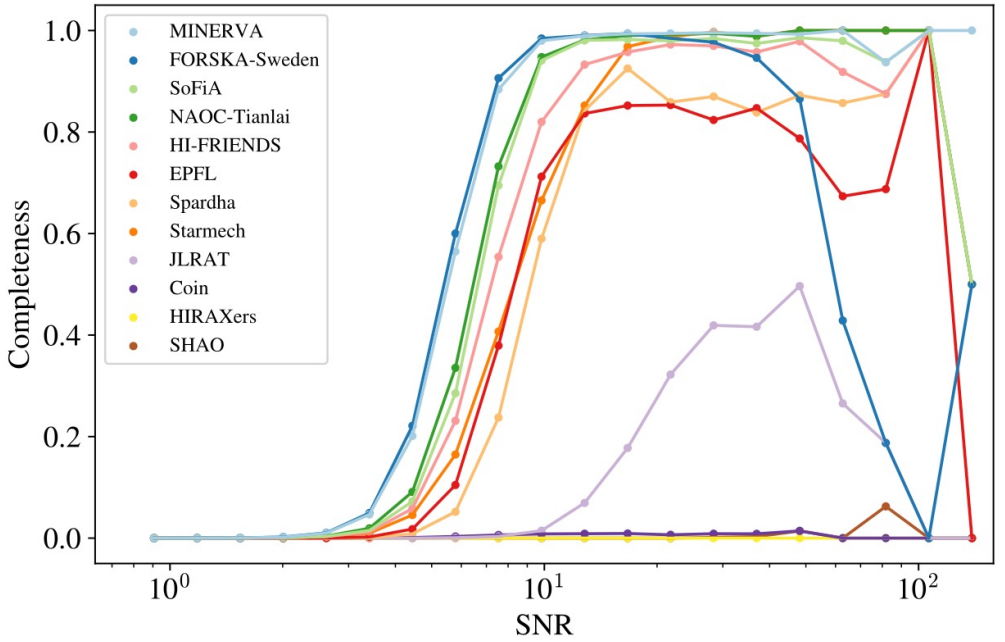


SKA Science Data Challenge 2: analysis and results

Get access >

P Hartley ✉, A Bonaldi, R Braun, J N H S Aditya, S Aicardi, L Alegre, A Chakraborty, X Chen, S Choudhuri, A O Clarke, J Coles, J S Collinson, D Cornu, L Darriba, M Delli Veneri, J Forbrich, B Fraga, A Galan, J Garrido, F Gubanov, H Håkansson, M J Hardcastle, C Heneka, D Herranz, K M Hess, M Jagannath, S Jaiswal, R J Jurek, D Korber, S Kitaeff, D Kleiner, B Lao, X Lu, A Mazumder, J Moldón, R Mondal, S Ni, M Önnheim, M Parra, N Patra, A Peel, P Salomé, S Sánchez-Expósito, M Sargent, B Semelin, P Serra, A K Shaw, A X Shen, A Sjöberg, L Smith, A Soroka, V Stolyarov, E Tolley, M C Toribio, J M van der Hulst, A Vafaei Sadr, L Verdes-Montenegro, T Westmeier, K Yu, L Yu, L Zhang, X Zhang, Y Zhang, A Alberdi, M Ashdown, C R Bom, M Brüggen, J Cannon, R Chen, F Combes, J Conway, F Courbin, J Ding, G Fourestey, J Freundlich, L Gao, C Gheller, Q Guo, E Gustavsson, M Jirstrand, M G Jones, G Józsa, P Kamphuis, J-P Kneib, M Lindqvist, B Liu, Y Liu, Y Mao, A Marchal, I Márquez, A Meshcheryakov, M Olberg, N Oozeer, M Pandey-Pommier, W Pei, B Peng, J Sabater, A Sorgho, J L Starck, C Tasse, A Wang, Y Wang, H Xi, X Yang, H Zhang, J Zhang, M Zhao, S Zuo

Monthly Notices of the Royal Astronomical Society, Volume 523, Issue 2, August 2023, Pages 1967–1993, https://doi.org/10.1093/mnras/stad1375



- Detailed understanding of noise properties within data products vital to optimising source detection
  - Need to account for actual **variability of RMS level as function of angular scale** (that reflects array configuration)
- Best results obtained with hybrid methods based on “agreement” between traditional and machine learning approaches (but reliant on high quality training data for ML!)
- Within caveats of only limited simulation realism:
  - Improved understanding of biases in HI surveys
  - Improved understanding of likely survey completeness versus SNR, ~50% at 5σ

Reproducibility awards

SDC2

Software Sustainability Institute

www.software.ac.uk

Reproducibility:

Is the software:

- Well-documented
- Easy to install
- Easy to use

Reusability:

Does the software:

- Use an open licence
- Have findable code
- Use code standards
- Use built-in tests

	Reproducibility of the solution Can the software pipeline be re-run easily to produce the same results? Is it: • Well-documented *Research software documentation best practice* • Easy to install *Top tips for packaging software* • Easy to use *Top tips for documentation*	
Well-documented	High-level description of what/who the software is for is available	
	High-level description of what the software does is available	
	High-level description of how the software works is available	
	Documentation consists of clear, step-by-step instructions	
	Documentation gives examples of what the user can see at each step e.g. screenshots or command-line excerpt	
	Documentation uses `monospace` fonts for command-line inputs and outputs, source code fragments, function names, class names etc	
	Documentation is held under version control alongside the code	
Easy to install	Full instructions provided for building and installing any software	
	All dependencies are listed, along with web addresses, suitable versions, licences and whether they are mandatory or optional	
	All dependencies are available	
	Tests are provided to verify that the installation has succeeded	
	A containerised package is available, containing the code together with all of the related configuration files, libraries, and dependencies required. *Using .e.g. Docker/Singularity*	
Easy to use	A getting started guide is provided outlining a basic example of using the software *e.g. a README file*	
	Instructions are provided for many basic use cases	
	Reference guides are provided for all command-line, GUI and configuration options	

	Reusability of the pipeline Can the code be reused easily by other people to develop new projects? Does it: • Have an open licence *Choosing an open source licence* • Have easily accessible source code *Choosing a repository for your project* • Adhere to coding standards *Writing readable source code* • Utilise tests *Testing your software*	
Open licence	Software has an open source licence *e.g. GNU General Public License (GPL), BSD 3-Clause*	
	Licence is stated in source code repository	
	Each source code file has a licence header	
Accessible code	Access to source code repository is available online	
	Repository is hosted externally in a sustainable third-party repository *e.g. SourceForge, LaunchPad, GitHub: Introduction to GitHub*	
	Documentation is provided for developers	
Code standards	Source code is laid out and indented well	
	Source code is commented	
	There is no commented out code	
	Source code is structured into modules or packages	
	Source code uses sensible class, package and variable names	
	Source code structure relates clearly to the architecture or design	
Testing	Source code has unit tests	
	Software recommends tools to check conformance to coding standards *e.g. A 'linter' such as PyLint for Python*	

NOW ON:

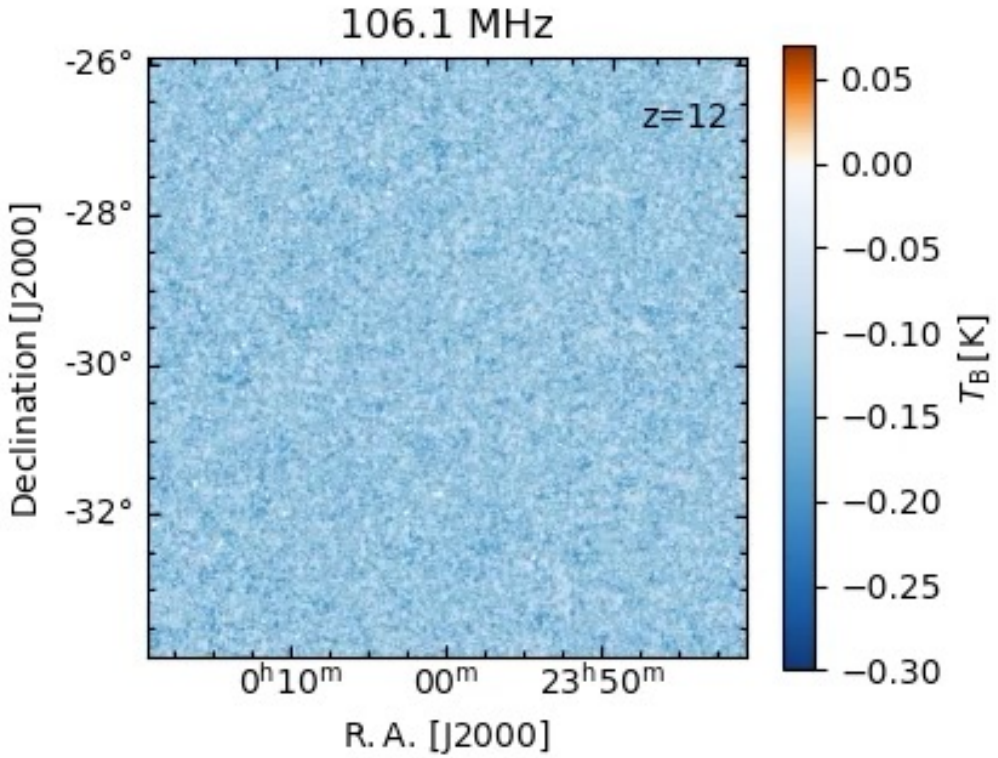
Science Data Challenge 3

Epoch of Reionisation

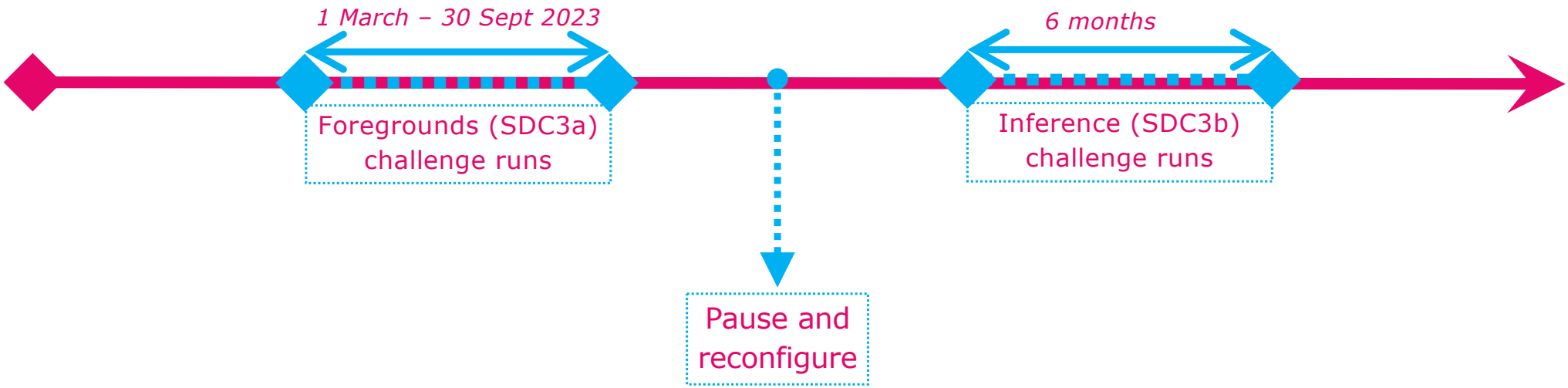
Science Data Challenge 3

Developed in collaboration with SKA EoR SWG members

- SDC3a "**Foregrounds**" (SDC3a; SWG Coordinators: C. Trott, V. Jelic)
  - **Foreground removal** exercise
  - SDC3a registration ran from 10th October 2022 – 15th November 2022
- SDC3b "**Inference**" (SDC3b; SWG Coordinators: A. Mesinger, G. Melema)
  - Extraction of **cosmological parameters**
  - SDC3b launching Q1 2024



SDC3 Timeline



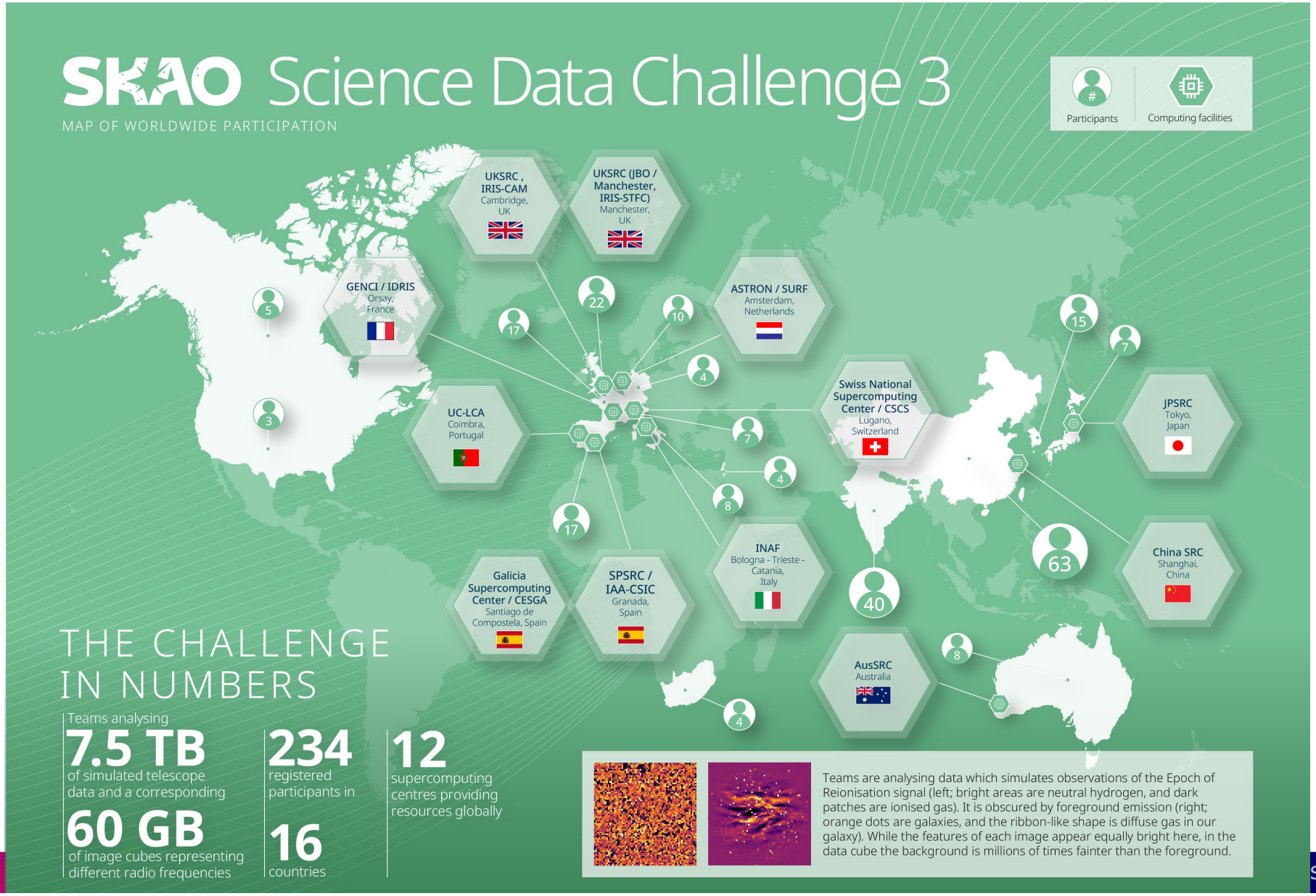
*Not to scale

Science Data Challenge 3a – Dataset(s)

- General
  - Observation track length HA = -2 to +2 hours
  - Thermal noise equivalent 1000 h
  - Field of View: one SKA1-Low pointing at RA, Dec = 0h, -30deg
- Visibilities
  - Size 5.4 TB
  - Integration time 10 s
  - Channel width 100 kHz
  - Frequency coverage 106 - 196 MHz
- Image cube -> 2048 x 2048, 16 arcsec pixels, natural weighting

SKAO Science Data Challenge 3

MAP OF WORLDWIDE PARTICIPATION



THE CHALLENGE IN NUMBERS

Teams analysing

7.5 TB of simulated telescope data and a corresponding

60 GB of image cubes representing different radio frequencies

234 registered participants in

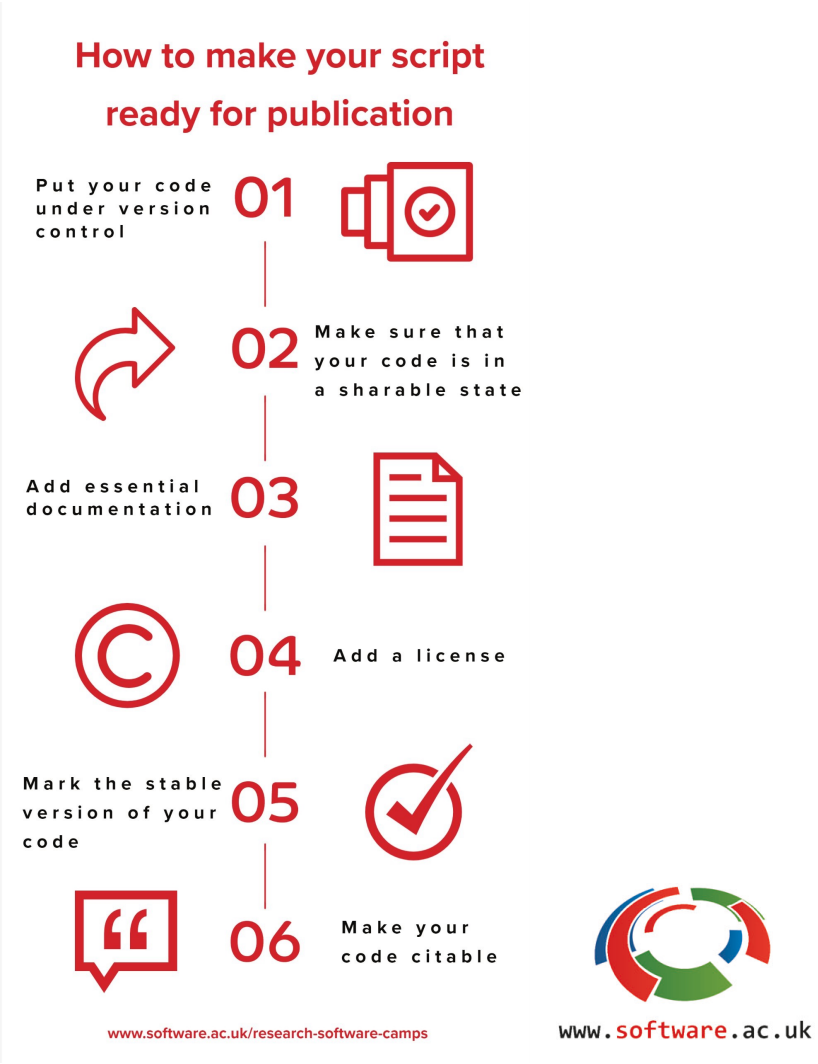
16 countries

12 supercomputing centres providing resources globally

Teams are analysing data which simulates observations of the Epoch of Reionisation signal (left; bright areas are neutral hydrogen, and dark patches are ionised gas). It is obscured by foreground emission (right; orange dots are galaxies, and the ribbon-like shape is diffuse gas in our galaxy). While the features of each image appear equally bright here, in the data cube the background is millions of times fainter than the foreground.

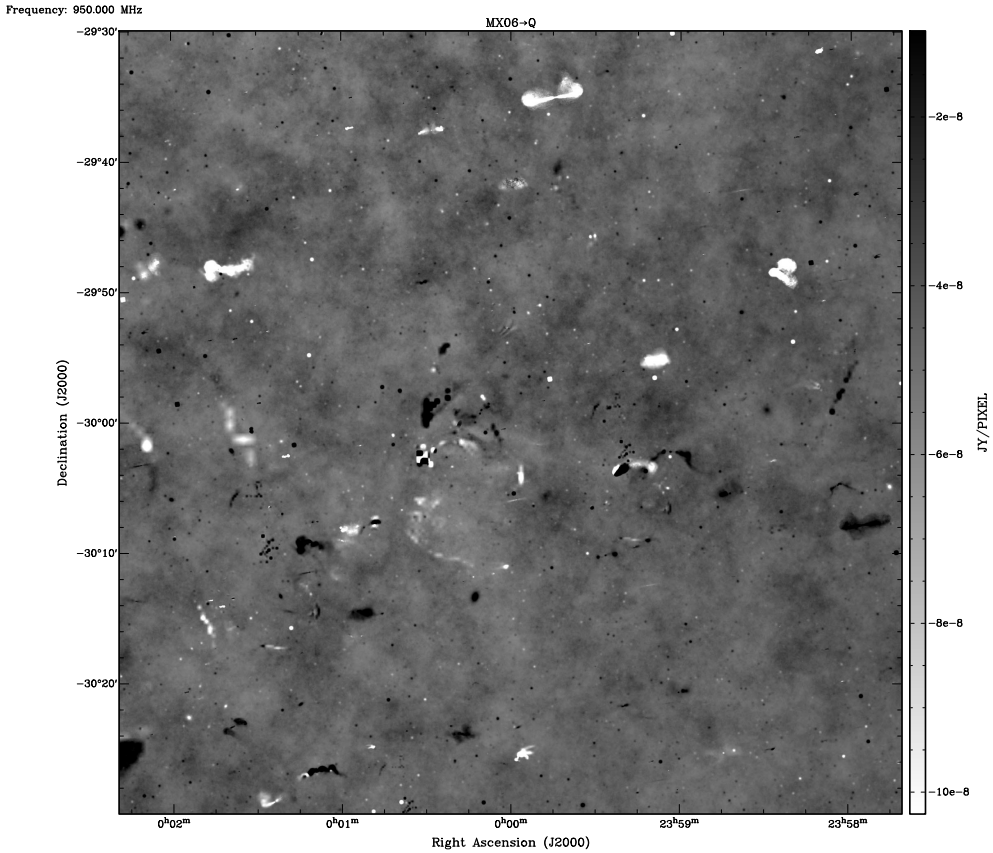
Reproducibility awards
SDC3

- Revised award system
- Reproducibility ‘badges’
- Based on Software Sustainability Institute’s six steps to reproducibility
- Simpler for teams to follow and achieve



Science Data Challenge 4 – Magnetism

- Developed in collaboration with Magnetism SWG (Akahori, Vernstrom, Vacca, …)
  - Scope still being refined, but full Stokes compact plus diffuse sky model with IGM, ISM, and ionosphere propagation
  - 10 square deg, 950 – 1760 MHz, 3 arcsec beam, source finding and characterisation
  - 100 square deg, 100 – 350 MHz, 350 – 1760 MHz, 10 arcsec beam, source finding and characterisation
  - Thermal noise equivalent few 1000 h
- Sky and Propagation Models nearing completion and looking good
- Telescope and Error Models
  - OSKAR for LOW
  - RASCIL for MID



Propagated Stokes Q Sky Model at 950 MHz

Science Data Challenges: What are the emerging benefits ?

- SDC1:
  - Identified telescope number limitations/inconsistencies within major packages: *casa, aips, miriad* that have now been remedied by the developers
  - Developed better understanding of crowded field and complex source finding
- SDC2:
  - Proto-SRC data product hosting, user interaction and analysis
  - Identified SNR versus scale as vital component of optimised source detection
  - Better understanding of HI survey biases and completeness
  - Reproducibility awards have stimulated code sharing and reuse
- SDC3:
  - Proto-SRC data distribution (eg. rucio), data hosting and user analysis
  - Expect to constrain (DI, DD, and bandpass) calibration precision requirements for SDP
- SDC4:
  - Identified the (storage and computing) benefits of non-linear frequency sampling for broad-band Stokes data (optimum sampling for given RM_{Max} scales as υ^3 and can save factor 5 for SKA bands)
  - Working with package developers to provide support (eg. FREQ-LOG axis support within carta)
  - Use of SRC-prototype JupyterHub service as analysis interface for participants
- All:
  - Growing repository of Sky Models and simulation code for community (re-)use

For more information on the SKA science data challenges:

- https://www.skao.int/en/science-users/160/skao-data-challenges
- Links to all past and current data challenges
- Datasets and scoring codes available once challenge is concluded

- Send a request via email to the relevant SWG co-chairs
- Contact details of the co-chairs available on the website

Questions?

We recognise and acknowledge the Indigenous peoples and cultures that have traditionally lived on the lands on which our facilities are located.

SKAO

www.skao.int