
SKAO
SKA SWG Update
Robert Braun, SKAO Science Director
21 June 2022

SKA Science Update

- New SWG co-chairs
- Science Data Challenge progress (Philippa, Anna, Simon)
- SWG Banners (Tyler)
- AOB

New SWG Co-Chairs

- Cradle of Life:
  - Welcome to Cherry Ng (University of Toronto)
  - Thanks to Laurent Lamy!
- HI Galaxies:
  - Welcome to Betsy Adams (ASTRON)
  - Thanks to Paolo Serra!

SDC2 results paper

- Results and analysis from SDC2 now in preparation for submission to MNRAS
- 12 finalist teams from over 40 institutions
- High level findings:
  - Complementary methods,
  - Mix of new and existing techniques; machine learning and non-machine learning,
  - SoFiA package very popular thanks to excellent documentation and ease of use,
  - Analysis of biases and HI mass recovery with redshift

SKA Science Data Challenge 2: analysis and results

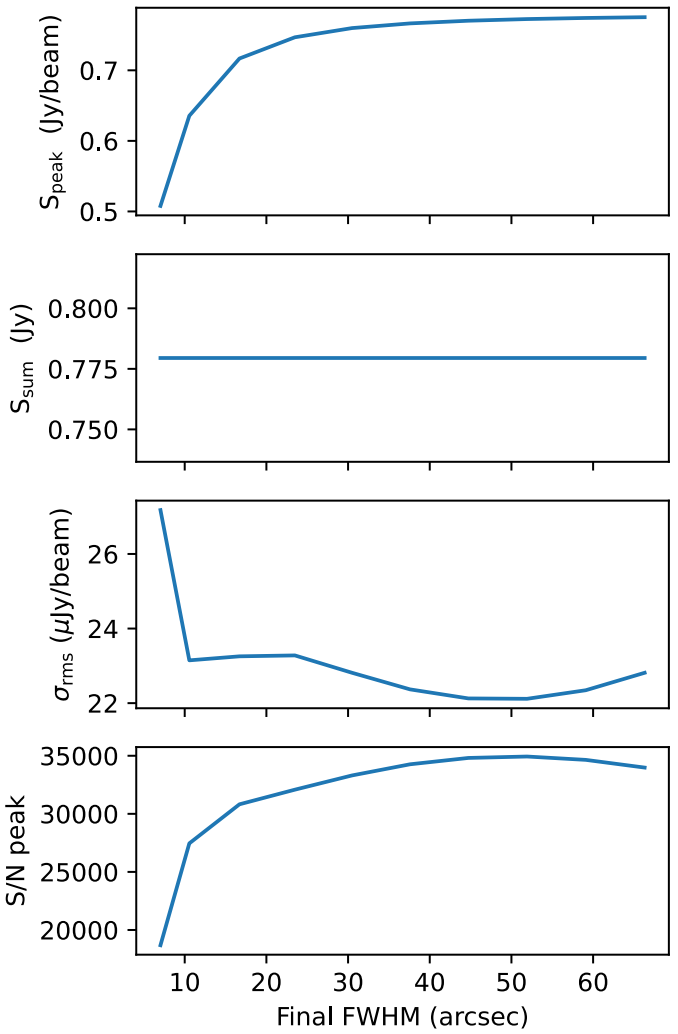
P. Hartley, A. Bonaldi, R. Braun, [Order TBC:] D. Cornu[1], B. Semelin[1], X. Lu[1], S. Aicardi[2], P. Salomé[1], A. Marchal[3], J. Freundlich[4], F. Combes[1,5], C. Tasse[6,7], C. Heneka[8], M. Delli Veneri[9], A. Soroka[10], F. Gubanov[10], A. Meshcheryakov[11], B. Fraga[12], C.R. Bom[12], M. Brüggen[8], A. K. Shaw[13], N. Patra[14], A. Chakraborty[15], R. Mondal[16], S. Choudhuri[17], A. Mazumder[15], M. Jagannath[18], M. J. Hardcastle[19], J. Forbrich[19], L. Smith[20], V. Stolyarov[20,21], M. Ashdown[20], J. Coles[20], H. Håkansson[22], A. Sjöberg[22], M. C. Toribio[23], M. Önnheim[22], M. Olberg[23], E. Gustavsson[22], M. Lindqvist[23], M. Jirstrand[22], J. Conway[23], K. M. Hess[24,25,26], R. J. Jurek[27], S. Kitaeff[28], P. Serra[29], A. X. Shen[30,31], J. M. van der Hulst[25], T. Westmeier[28], A. Alberdi[33], J. Cannon[34], L. Darriba[33], J. Garrido[33], J. Gósza[35], D. Herranz[36], M. G. Jones[37], P. Kamphuis[38], D. Kleiner[29], I. Márquez[33], J. Moldón[33], M. Pandey-Pommier[39], M. Parra[33], J. Sabater[40], S. Sánchez[33], A. Sorgho[33], L. Verdes-Montenegro[33], G. Fourestey[41], A. Galan[41], C. Gheller[29], D. Korber[41], A. Peel[41], M. Sargent[41], E. Tolley[41], B. Liu[42], R. Chen[42], B. Peng[42], L. Yu[42], H. Xi[42], K. Yu[43], Q. Guo[43], W. Pei[43], Y. Liu[43], Y. Wang[43], X. Chen[43], X. Zhang[44], S. Ni[44], J. Zhang[44], L. Gao[44], M. Zhao[44], L. Zhang[45], H. Zhang[45], X. Wang[45], J. Ding[45], S. Zuo[46], Y. Mao[46], A. Vafaei Sadr[47], M. Kunz[47], B. Bassett[48], D. Crichton[49], V. Nistane[47], N. Oozeer[35], S. Jaiswal[50], B. Lao[50], J. N. H. S. Aditya [50], Y. Zhang[50], A. Wang[50], and X. Yang[50]
Affiliations can be found after the references

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
The Square Kilometre Array Observatory (SKAO) will explore the radio sky to unrivalled depths in order to conduct transformational science. SKAO data products made available to astronomers will be correspondingly large and complex, requiring the application of advanced analysis techniques in order to extract key science findings. To this end, SKAO is conducting a series of Science Data Challenges, each designed to familiarise the scientific community with SKAO data and to drive the development of new analysis techniques. We present the results from Science Data Challenge 2 (SDC2), which invited participants to find and characterise neutral hydrogen (HI) sources in a simulated data product representing a 2000 h SKA MID spectral line observation. Through the generous support of eight international supercomputing facilities, participants were able to undertake the Challenge using dedicated computational resources. This model not only supported the accessible provision of a realistically large dataset, but also provided the opportunity to test several aspects of the future SKA Regional Centre network. Sitting alongside the main challenge, 'reproducibility awards' were made in recognition of those pipelines which demonstrated Open Science best practice. The Challenge saw over 100 finalists develop a range of new and existing techniques, in results which highlight the strengths of multidisciplinary and collaborative effort. The winning strategy – combining predictions from two independent machine learning techniques – underscores one of the main Challenge outcomes: that of method complementarity. It is likely that the combination of methods in a so-called ensemble approach will be key to exploiting very large astronomical datasets.

SDC2 signal-to-noise analysis

- Expressing SDC2 outcomes in terms of source signal-to-noise values
- Challenging to define a meaningful measure of integrated signal-to-noise, since it is intimately tied to the quality of a "matched filter"
- SKA noise properties unlike current telescopes
  - RMS noise remains ~constant when *spatially* smoothing between range of about 0.4 to 100 arcsec FWHM (at 1.2 GHz)
- Possible implications for source finding approaches
  - Need to include accurate noise models to optimize detection strategy
- See SDC2 paper for discussion



Science Data Challenge 3 (SDC3)

- SDC3 will consist of two tiers:
  - **SDC3 "foregrounds" (SDC3a)**
    - Simulations for the challenge are maturing, and we are aiming for a challenge **start date of October 2022,** for a duration of six months
    - Computational costs are likely to vary significantly according to approaches (see later slides)
  - **SDC3 "inference" (SDC3b)**
    - SDC3 inference will be run after a short break following the conclusion of "foregrounds", so it will take place **during 2023.**
    - Again, computational costs are likely to vary between teams
- Working with members of EoR SWG to develop both tiers
- **SKAO** is developing a website dedicated to SDC3, with Challenge registration due to open soon

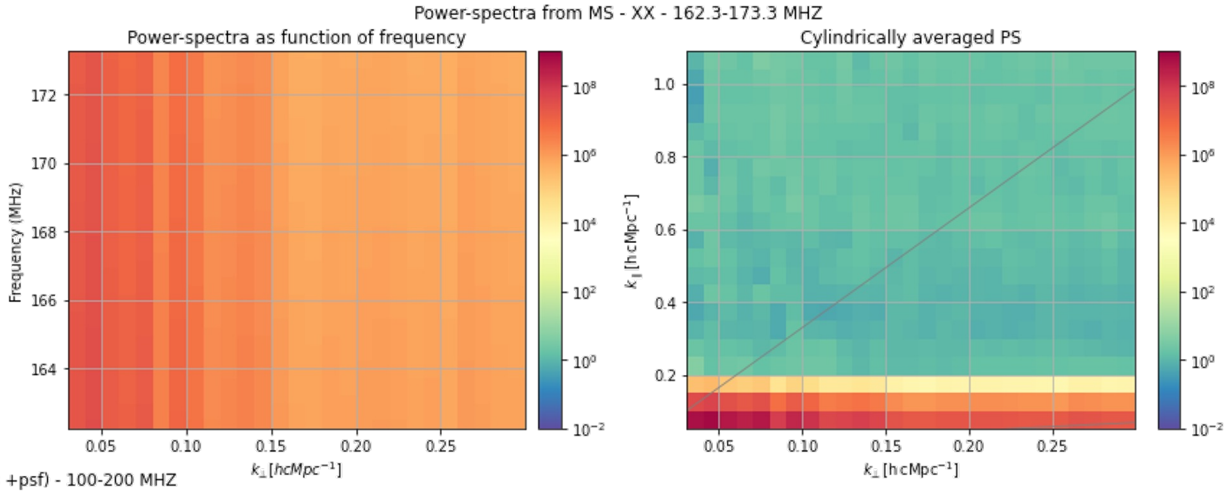
Status of SDC3 Foregrounds data products

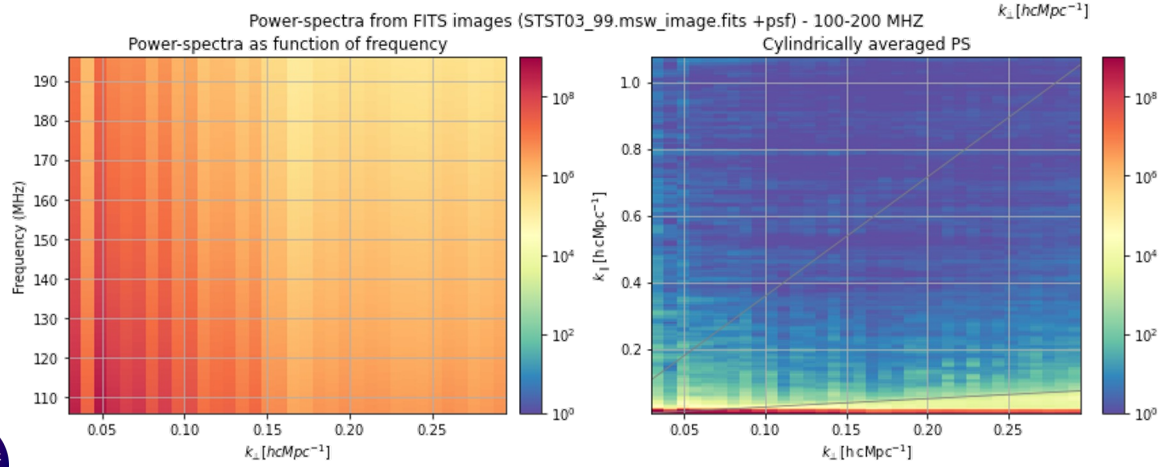
- The datasets that will be provided to participants are reaching maturity
- Currently under review by EoR specialists
- Positive feedback from review will allow us to confirm target start date of October 2022
- Data products that will be available to participants:
  - 3 GB image cube
  - 1 TB visibility set

SDC3 foregrounds dataset feedback

- Preliminary dataset provided to a few EoR SWG experts for feedback
- Several issues identified
- Improved version underway

Validation of the measurement set
(Courtesy F. Mertens)





Validation of the image cube
(Courtesy F. Mertens)

Foreground All-scale Radio Modeller (FARM)

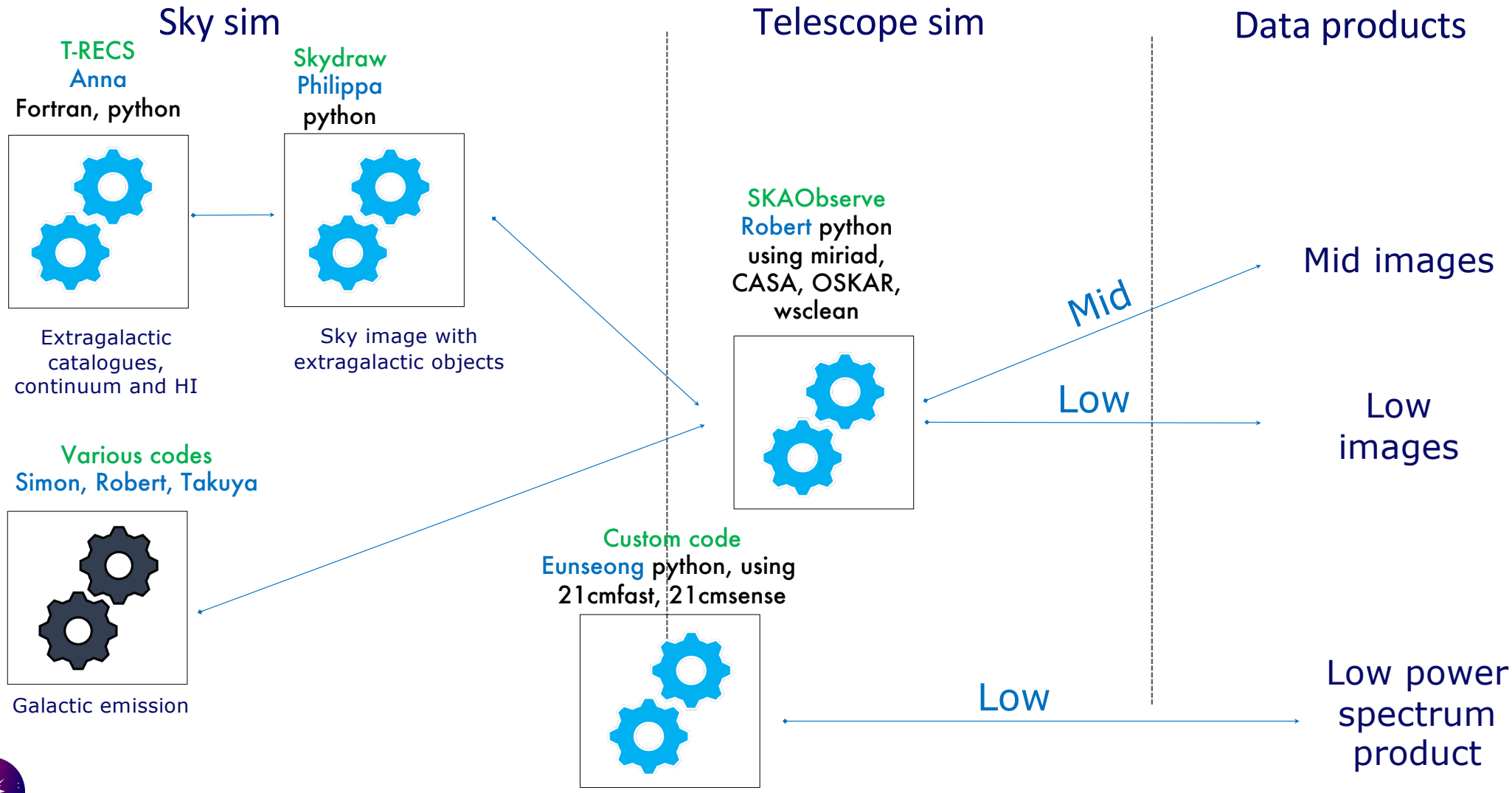
"FARM adapts SKAObserve script into a python-based, modular, reusable, and extensible CLI- or GUI-based foreground-simulation tool"

- Can use different telescopes
- Executions to simulate the foreground use a configuration file to specify parameters
- Can 'switch on/off' different effects/foreground components
- Specify calibration errors/effects
- Can generate some components from scratch, or uses models

```
[directories]
  root_name = "example_1"  # root name for all output files (not directory)
  telescope_model = "/Users/SDC3/Data/TelescopeModels/V512/telescope.tm"
  output_dcy = "/Users/SDC3/Desktop/test_output_farm"

# observation not yet incorporated in to FARM
[observation]
  time = 2027-01-01T10:00:01  # Start time (UTC) YYYY-MM-DDTHH:MM:SS
  t_total = 600  # Total on source time [s]
  n_scan = 1  # Number of observational 'scans' of target field
  min_gap_scan = 1200  # Minimum time-gap between consecutive scans [s]
  min_elevation = 30.0  # Minimum elevation [deg]

  [observation.field]
    ra0 = "00:00:00.0"  # HH:MM:SS.S
    dec0 = "-30:00:00.00"  # DD:MM:SS.SS
    frame = "fk5"  # Astropy-compatible frame string e.g. fk4, fk5, galactic etc
    nxpix = 1024  # Number of pixels in x/right ascension
    nypix = 1024  # Number of pixels in y/declination
    cdelt = 28.125  # arcsec

  [observation.correlator]
    freq_min = 115e6  # Hz
    freq_max = 169e6  # Hz
    nchan = 217  # Number of evenly spaced channels across whole bandwidth, int
    chanwidth = 250e3  # channel bandwidth for smearing calculation, Hz
    t_int = 60.0  # visibility integration time, s
```

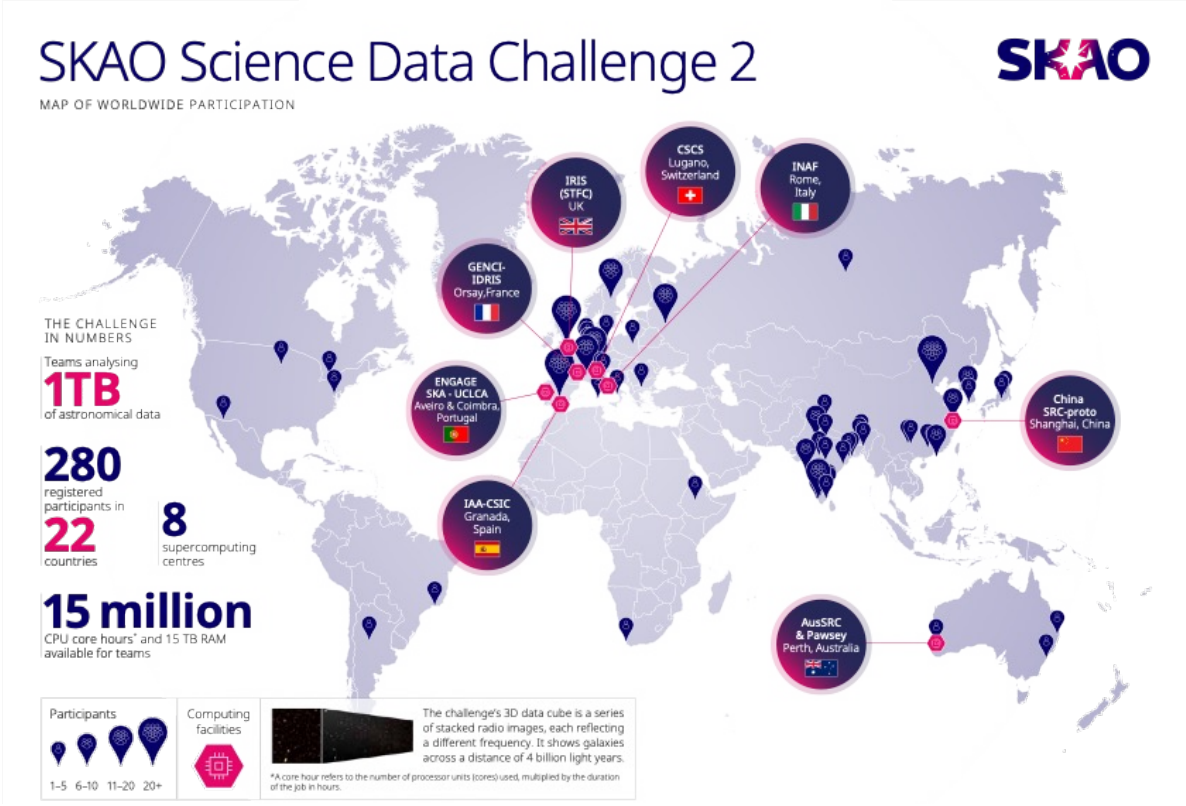
SDC simulation framework

To be migrated to Gitlab!

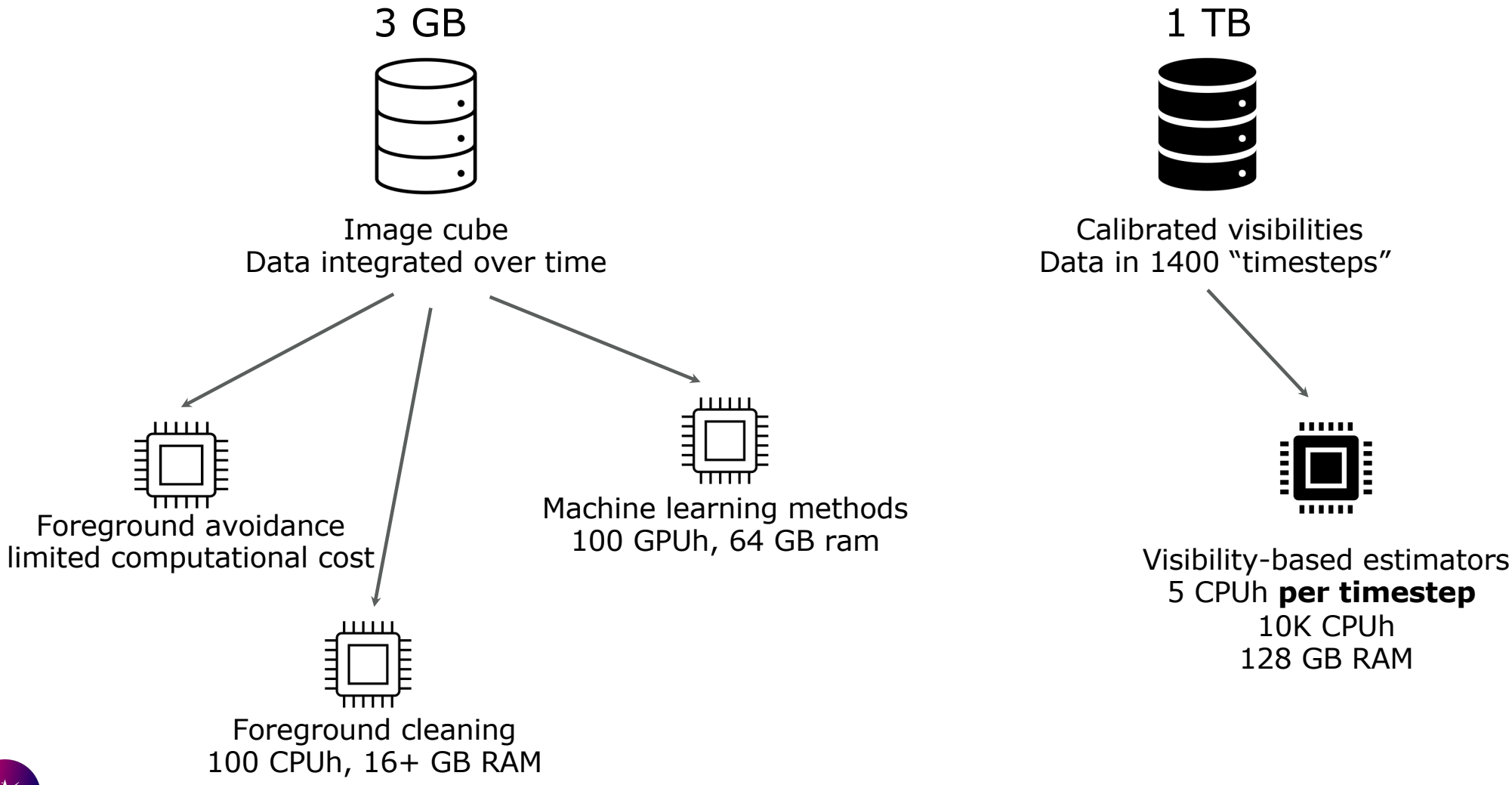


SDC Computational support model

- SDC2 received invaluable support from international computing facilities
  - Enabled SKAO to provide a 1 TB data product in an accessible way
  - Test cases for SKA Regional Centre (SRC) prototyping
- SDC3 will also receive computational support for participating teams
  - Computational cost estimates of Foregrounds analysis are currently being finalized
  - A new registration and time allocation model
    - Streamlining the computational resource allocation process and help teams to get the most from their resources



SDC3 Computational costs



SDC3 Computational costs: current estimates

with thanks to EoR SWG members for estimates

- We expect ~10 teams to take part in SDC3a
- We expect that some teams will deal only with the smaller (~3 GB) image dataset
  - Estimated compute cost: ~ <100 CPUh per full pipeline run; <64 GB RAM
  - i.e. ~3 hours on 30 core VM
- Some teams will rely on the larger (~1 TB) visibilities dataset (or both)
  - Estimated compute cost: cost dominated by one step of the processing <10k CPUh; <128 GB RAM.
  - i.e. 20 days in 30 core VM
  - This step is not likely to be run many times

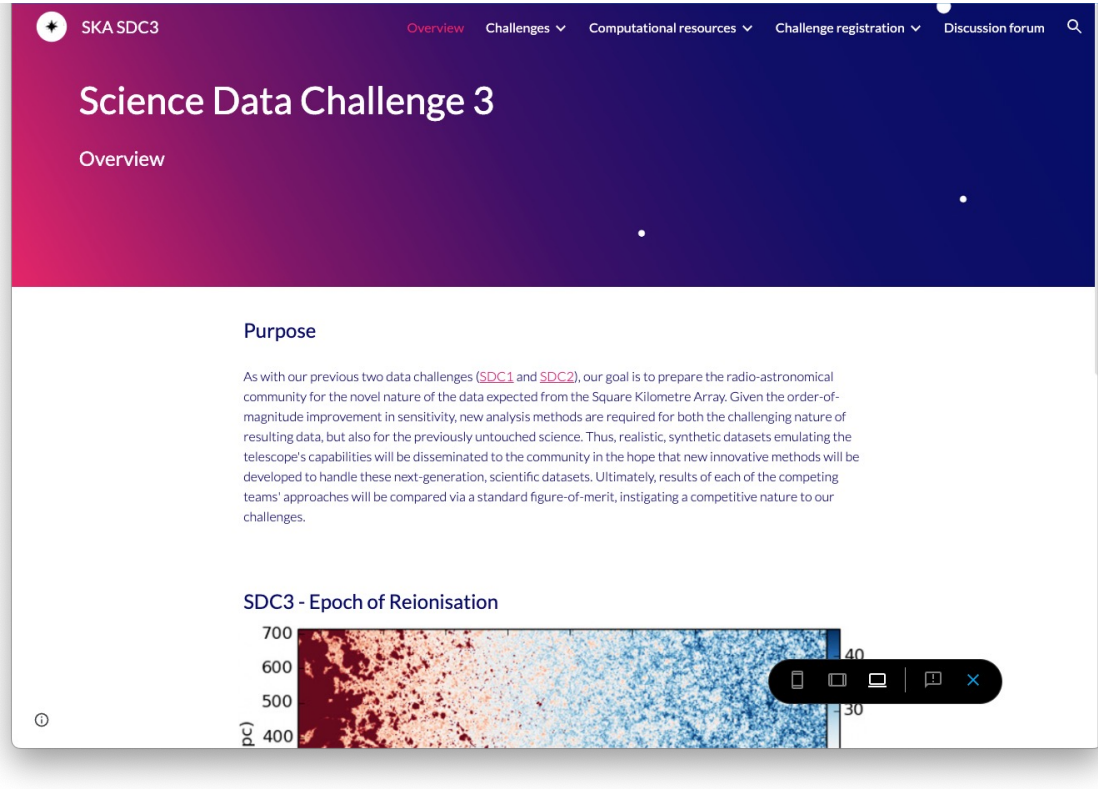
SDC3 registration

As part of SDC3 registration participants will be invited to submit a short proposal, which:

- Allows us to assess teams' software and hardware requirements
- Enables us to match teams to HPC facilities (who themselves have provided us with the month-by-month availability of their facilities)
- Facilitates more effective use of HPC facilities

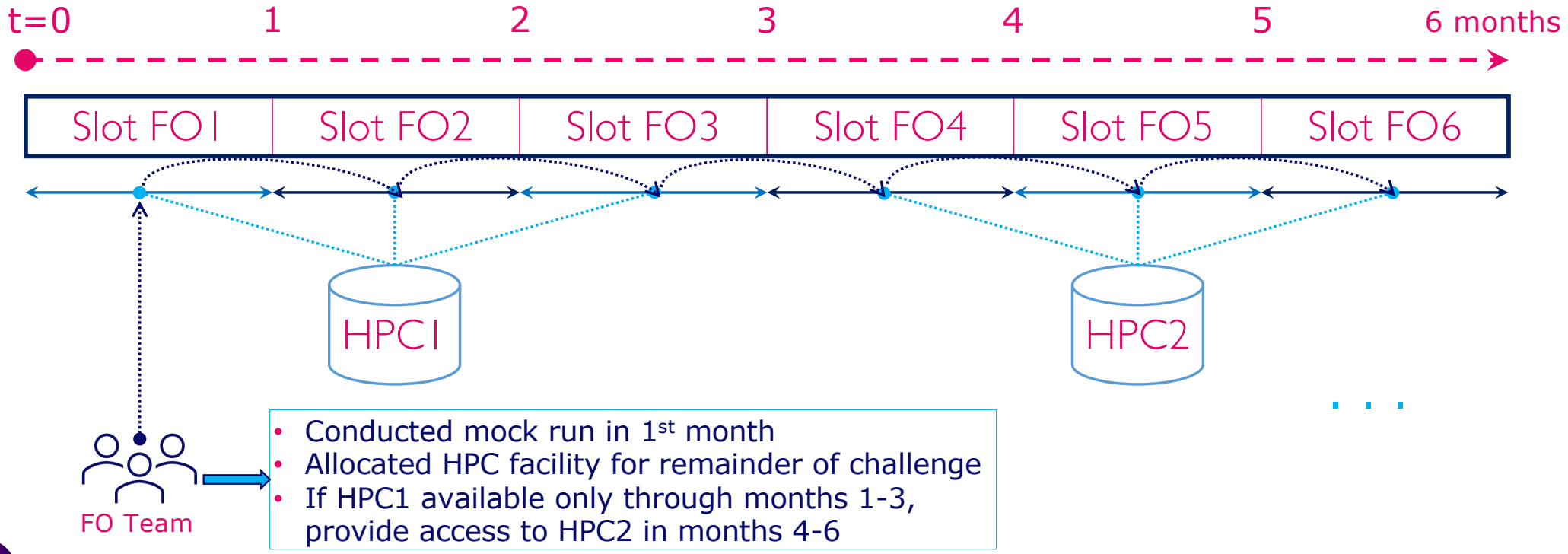
SDC3 Registration

- Registration will open soon
- Via SDC3 website (under construction)



SDC3 computing support allocation

All participants assigned to HPC-facility. Asked to conduct a 'mock' run on server in month 1, then given permanent account on that facility for remainder of the challenge



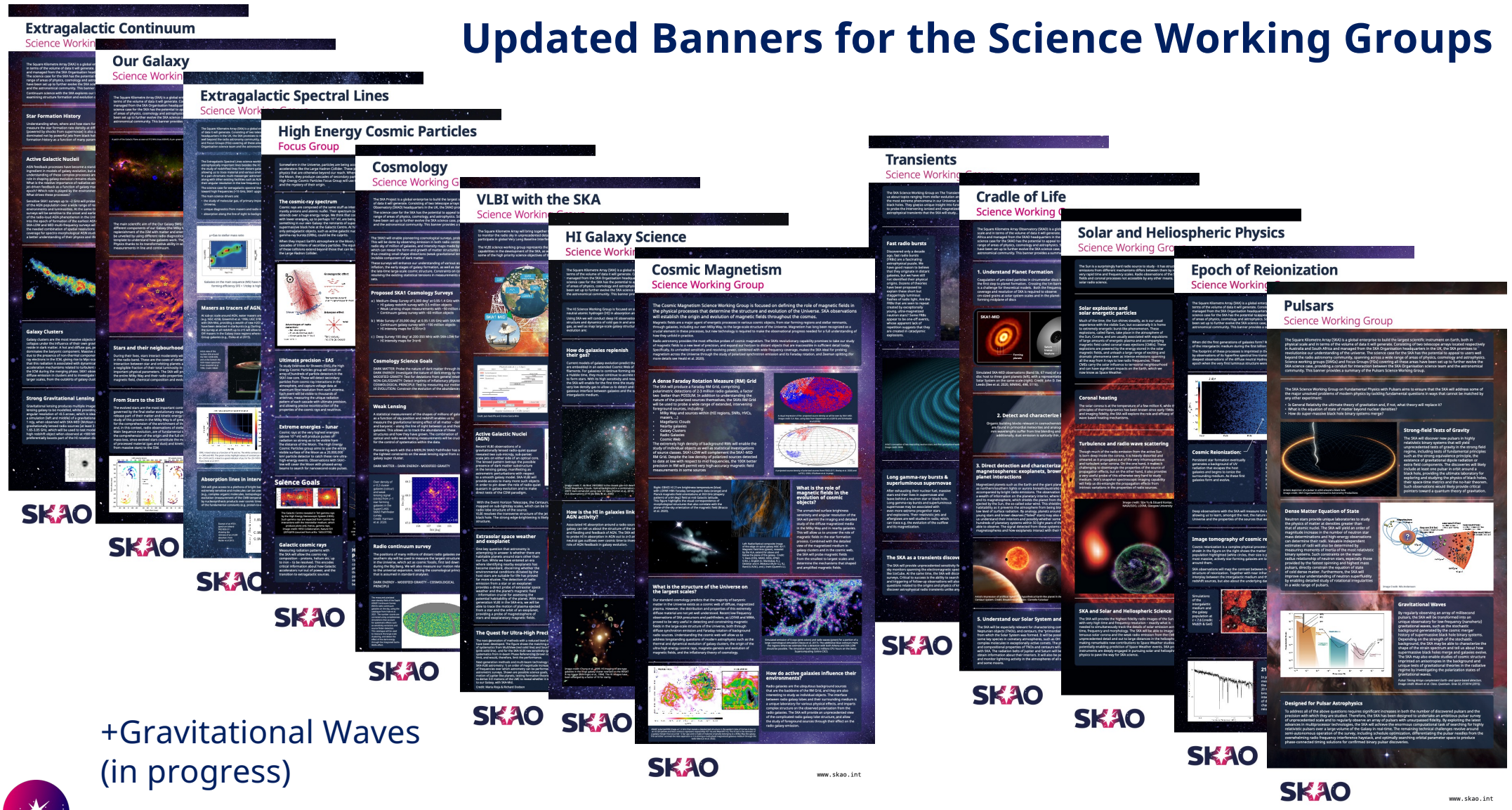
SRC prototyping activities

- The data challenges provide the opportunity to test some of techniques being prototyped for future SKA Regional Centre (SRC) Network
- The SRC development work is now being conducted within the ‘SAFe’ methodology
  - SDC development work could interface via raising features or by identifying prototype products that would be useful to test during a science data challenge
- We would be cautious to ensure that any prototype testing would happen at a suitable level of maturity of a prototyped product, and within a timescale that suits all supporting facilities.
- **No major prototyping planned for SDC3a**
  - Possibly testing by participants of e.g. image viewing software (CARTA)

Reproducibility, Open Science and best practice

- The data challenges provide the opportunity to familiarise the science community with some of the best practices concerning software and data processing
- Option to again include 'reproducibility awards' alongside main challenge:
  - Possibly revise format of award
  - For example: can SKAO team re-run a team's pipeline?
- Idea of an environmental award to run alongside the main challenge

Updated Banners for the Science Working Groups



+Gravitational Waves
(in progress)

Any Other Business

- Upcoming meetings
  - EAS2022 "S7: Building bridges: The lifecycle of dust and gas in the Milky Way with ALMA and SKA", 27 June – 1 July (https://eas.unige.ch/EAS_meeting/session.jsp?id=S7)
  - EAS2022 "SS23: Towards the SKA Observatory: Artificial Intelligence in radio astronomy", 27 June – 1 July, (https://eas.unige.ch/EAS_meeting/session.jsp?id=SS23)
  - EAS2022 "SS5: Neutral hydrogen: the next generation of science and simulations", 27 June – 1 July, (https://eas.unige.ch/EAS2022/session.jsp?id=SS5)
  - "Coordinated Surveys of the Southern Sky", 10 – 14 October, (https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/meetings/getMeetings.html?number=6792)
  - …
- Other Points?

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