SKA SWG Update

Robert Braun, SKAO Science Director

15 November 2022

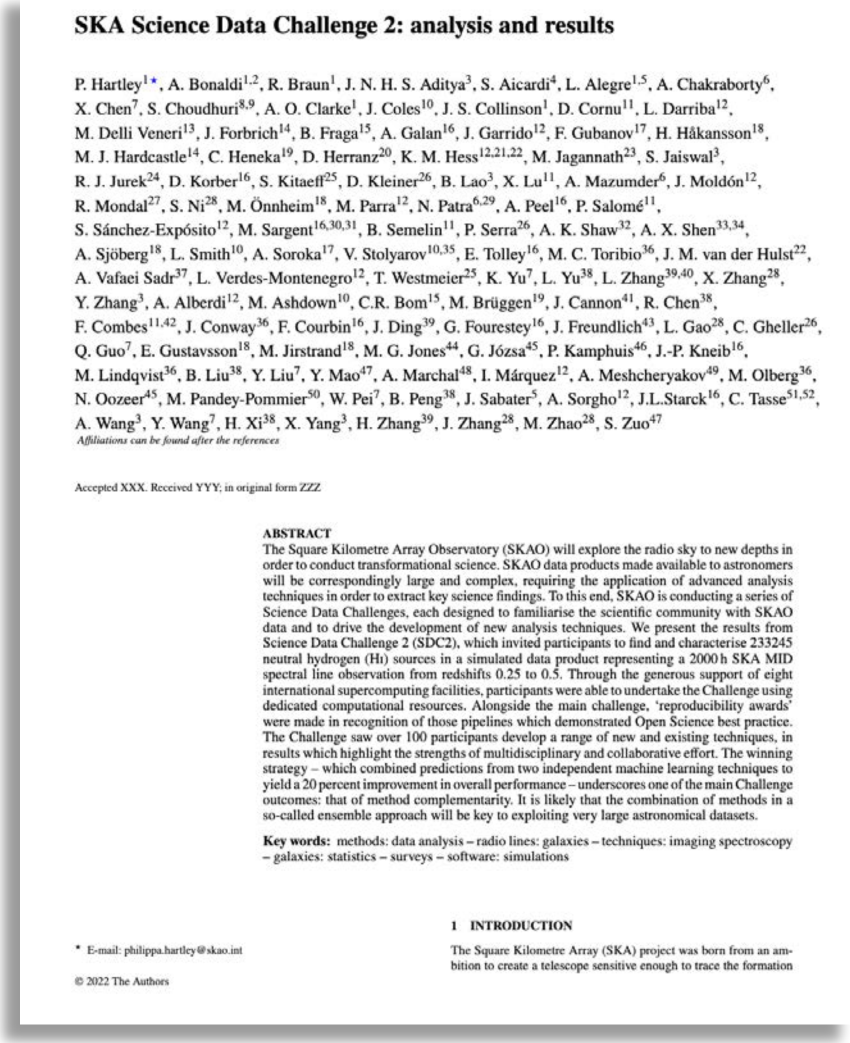
SKA Science Update

- Data Challenges Update
- Science Meetings
- AOB

Science Data Challenge 2 results paper

In review

- Describing the Challenge, the simulations, teams’ methods, results and analysis
- Submitted to MNRAS
- Over 100 challenge participants
- Over 50 worldwide institutions



SKA Science Data Challenge 2: analysis and results

P. Hartley[1⋆], A. Bonaldi[1,2], R. Braun[1], J. N. H. S. Aditya[3], S. Aicardi[4], L. Alegre[1,5], A. Chakraborty[6], X. Chen[7], S. Choudhuri[8,9], A. O. Clarke[1], J. Coles[10], J. S. Collinson[1], D. Cornu[11], L. Darriba[12], M. Delli Veneri[13], J. Forbrich[14], B. Fraga[15], A. Galan[16], J. Garrido[12], F. Gubanov[17], H. Håkansson[18], M. J. Hardcastle[14], C. Heneka[19], D. Herranz[20], K. M. Hess[12,21,22], M. Jagannath[23], S. Jaiswal[3], R. J. Jurek[24], D. Korber[16], S. Kitaeff[25], D. Kleiner[26], B. Lao[3], X. Lu[11], A. Mazumder[6], J. Moldón[12], R. Mondal[27], S. Ni[28], M. Önnheim[18], M. Parra[12], N. Patra[6,29], A. Peel[16], P. Salomé[11], S. Sánchez-Expósito[12], M. Sargent[16,30,31], B. Semelin[11], P. Serra[26], A. K. Shaw[32], A. X. Shen[33,34], A. Sjöberg[18], L. Smith[10], A. Soroka[17], V. Stolyarov[10,35], E. Tolley[16], M. C. Toribio[36], J. M. van der Hulst[22], A. Vafaei Sadr[37], L. Verdes-Montenegro[12], T. Westmeier[25], K. Yu[7], L. Yu[38], L. Zhang[39,40], X. Zhang[28], Y. Zhang[3], A. Alberdi[12], M. Ashdown[10], C.R. Bom[15], M. Brüggen[19], J. Cannon[41], R. Chen[38], F. Combes[11,42], J. Conway[36], F. Courbin[16], J. Ding[39], G. Fourestey[16], J. Freundlich[43], L. Gao[28], C. Gheller[26], Q. Guo[7], E. Gustavsson[18], M. Jirstrand[18], M. G. Jones[44], G. Józsa[45], P. Kamphuis[46], J.-P. Kneib[16], M. Lindqvist[36], B. Liu[38], Y. Liu[7], Y. Mao[47], A. Marchal[48], I. Márquez[12], A. Meshcheryakov[49], M. Olberg[36], N. Oozeer[45], M. Pandey-Pommier[50], W. Pei[7], B. Peng[38], J. Sabater[5], A. Sorgho[12], J.L.Starck[16], C. Tasse[51,52], A. Wang[3], Y. Wang[7], H. Xi[38], X. Yang[3], H. Zhang[39], J. Zhang[28], M. Zhao[28], S. Zuo[47]

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 new 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 233245 neutral hydrogen (Hı) sources in a simulated data product representing a 2000 h SKA MID spectral line observation from redshifts 0.25 to 0.5. Through the generous support of eight international supercomputing facilities, participants were able to undertake the Challenge using dedicated computational resources. Alongside the main challenge, ‘reproducibility awards’ were made in recognition of those pipelines which demonstrated Open Science best practice. The Challenge saw over 100 participants develop a range of new and existing techniques, in results which highlight the strengths of multidisciplinary and collaborative effort. The winning strategy – which combined predictions from two independent machine learning techniques to yield a 20 percent improvement in overall performance – 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.

Key words: methods: data analysis – radio lines: galaxies – techniques: imaging spectroscopy – galaxies: statistics – surveys – software: simulations

1 INTRODUCTION

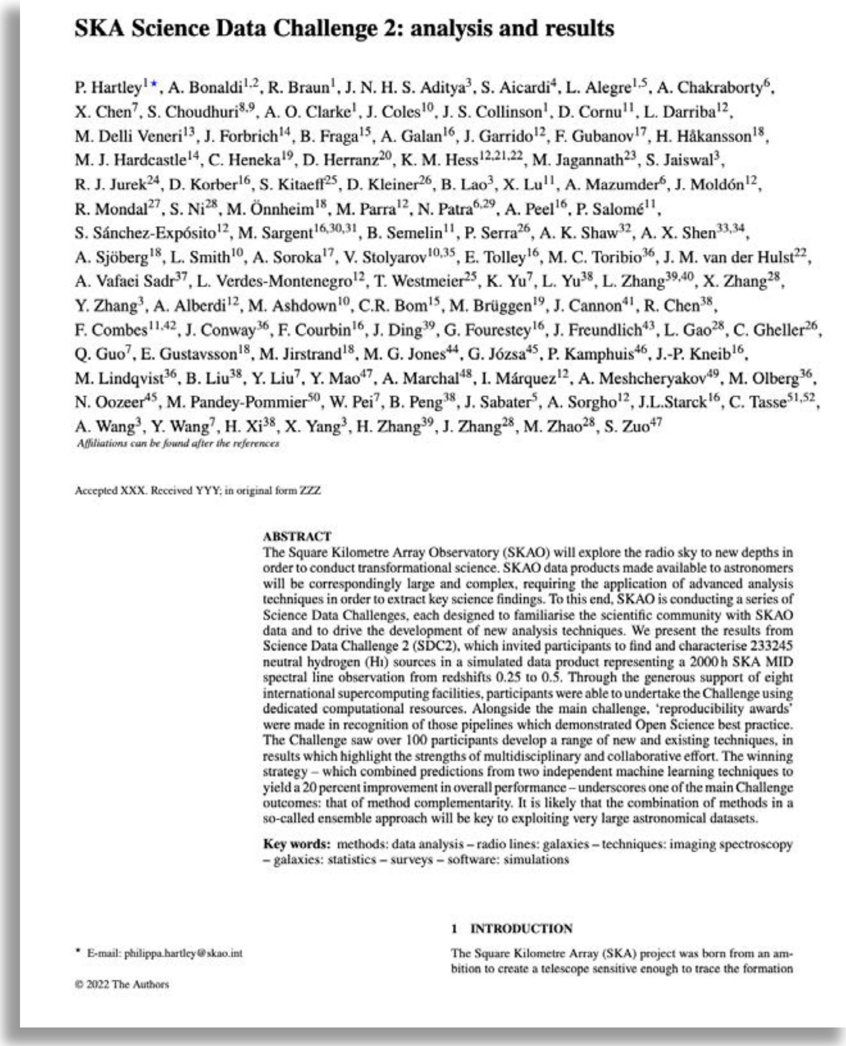
The Square Kilometre Array (SKA) project was born from an ambition to create a telescope sensitive enough to trace the formation

⋆ E-mail: philippa.hartley@skao.int

Science Data Challenge 2 results paper

In review

- 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[1*], A. Bonaldi[1,2], R. Braun[1], J. N. H. S. Aditya[3], S. Aicardi[4], L. Alegre[1,5], A. Chakraborty[6], X. Chen[7], S. Choudhuri[8,9], A. O. Clarke[1], J. Coles[10], J. S. Collinson[1], D. Cornu[11], L. Darriba[12], M. Delli Veneri[13], J. Forbrich[14], B. Fraga[15], A. Galan[16], J. Garrido[12], F. Gubanov[17], H. Håkansson[18], M. J. Hardcastle[14], C. Heneka[19], D. Herranz[20], K. M. Hess[12,21,22], M. Jagannath[23], S. Jaiswal[3], R. J. Jurek[24], D. Korber[16], S. Kitaeff[25], D. Kleiner[26], B. Lao[3], X. Lu[11], A. Mazumder[6], J. Moldón[12], R. Mondal[27], S. Ni[28], M. Önnheim[18], M. Parra[12], N. Patra[6,29], A. Peel[16], P. Salomé[11], S. Sánchez-Expósito[12], M. Sargent[16,30,31], B. Semelin[11], P. Serra[26], A. K. Shaw[32], A. X. Shen[33,34], A. Sjöberg[18], L. Smith[10], A. Soroka[17], V. Stolyarov[10,35], E. Tolley[16], M. C. Toribio[36], J. M. van der Hulst[22], A. Vafaei Sadr[37], L. Verdes-Montenegro[12], T. Westmeier[25], K. Yu[7], L. Yu[38], L. Zhang[39,40], X. Zhang[28], Y. Zhang[3], A. Alberdi[12], M. Ashdown[10], C.R. Bom[15], M. Brüggen[19], J. Cannon[41], R. Chen[38], F. Combes[11,42], J. Conway[36], F. Courbin[16], J. Ding[39], G. Fourestey[16], J. Freundlich[43], L. Gao[28], C. Gheller[26], Q. Guo[7], E. Gustavsson[18], M. Jirstrand[18], M. G. Jones[44], G. Józsa[45], P. Kamphuis[46], J.-P. Kneib[16], M. Lindqvist[36], B. Liu[38], Y. Liu[7], Y. Mao[47], A. Marchal[48], I. Márquez[12], A. Meshcheryakov[49], M. Olberg[36], N. Oozeer[45], M. Pandey-Pommier[50], W. Pei[7], B. Peng[38], J. Sabater[5], A. Sorgho[12], J.L.Starck[16], C. Tasse[51,52], A. Wang[3], Y. Wang[7], H. Xi[38], X. Yang[3], H. Zhang[39], J. Zhang[28], M. Zhao[28], S. Zuo[47]

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 new 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 233245 neutral hydrogen (HI) sources in a simulated data product representing a 2000 h SKA MID spectral line observation from redshifts 0.25 to 0.5. Through the generous support of eight international supercomputing facilities, participants were able to undertake the Challenge using dedicated computational resources. Alongside the main challenge, 'reproducibility awards' were made in recognition of those pipelines which demonstrated Open Science best practice. The Challenge saw over 100 participants develop a range of new and existing techniques, in results which highlight the strengths of multidisciplinary and collaborative effort. The winning strategy – which combined predictions from two independent machine learning techniques to yield a 20 percent improvement in overall performance – 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.

Key words: methods: data analysis – radio lines: galaxies – techniques: imaging spectroscopy – galaxies: statistics – surveys – software: simulations

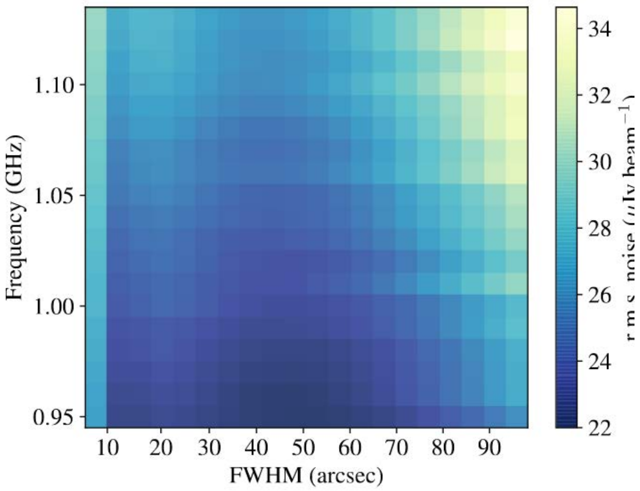
1 INTRODUCTION

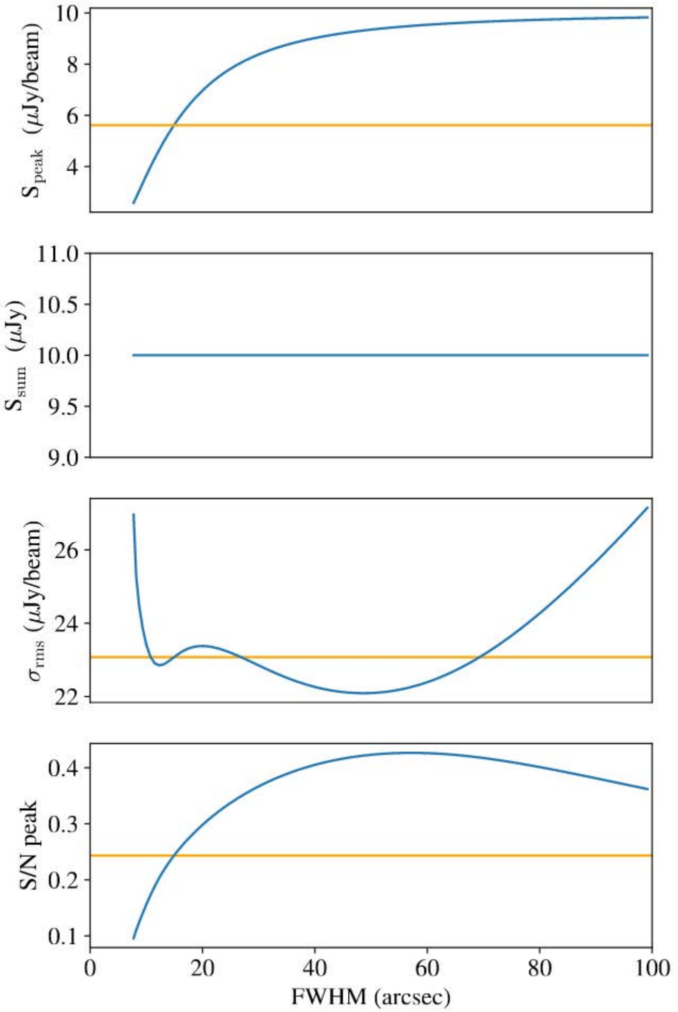
The Square Kilometre Array (SKA) project was born from an ambition to create a telescope sensitive enough to trace the formation

[*] E-mail: philippa.hartley@skao.int

Science Data Challenge 2 results paper

- Expressing SDC2 scores in terms of source signal-to-noise values
- Meaningful measure of signal-to-noise
- Use SKA MID noise properties:
  - RMS noise remains ~constant when *spatially* smoothing up to ~70 arcsec FWHM
- Possible implications for source finding approaches





Science Data Challenge 2 results paper

Reproducibility awards

In partnership
with the Software
Sustainability
Institute

www.software.ac.uk

Reproducibility:

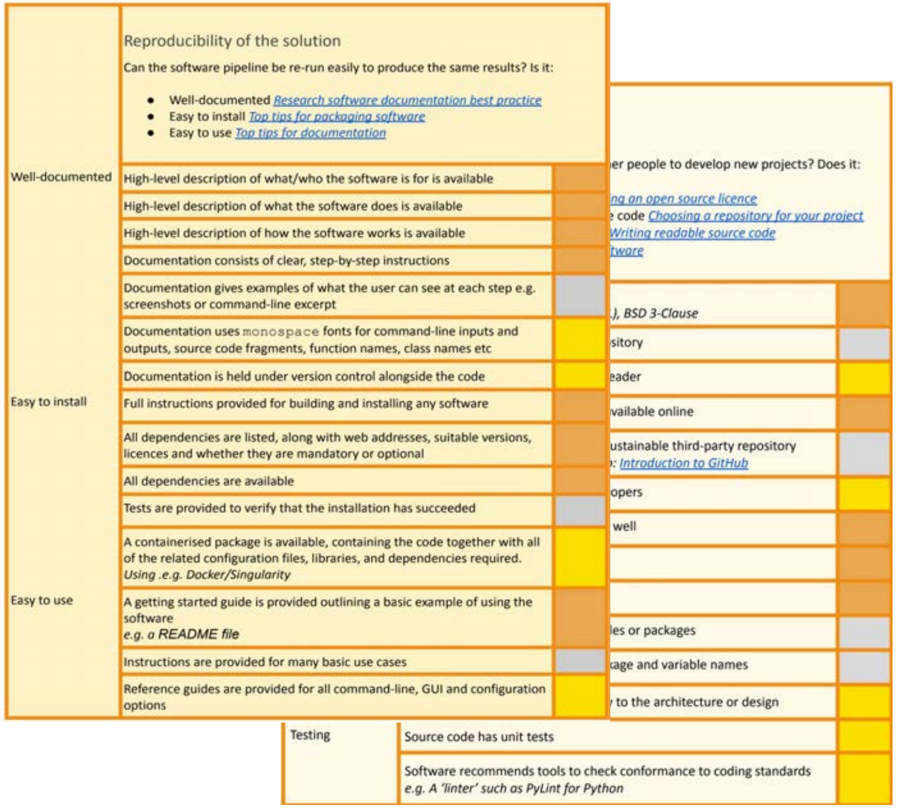
Is the software:

- 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	🟨

ier people to develop new projects? Does it:

ng an open source licence
e code *Choosing a repository for your project*
Writing readable source code
tware

	.), BSD 3-Clause	🟧
	sitory	⬜
	eader	🟨
	vailable online	🟧
	ustainable third-party repository : *Introduction to GitHub*	⬜
	opers	🟨
	well	🟧
		🟧
		🟧
	les or packages	⬜
	kage and variable names	⬜
	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*	🟨

Science Data Challenge 2 results paper

Reproducibility awards

In partnership
with the Software
Sustainability
Institute

www.software.ac.uk

Reproducibility:

Is the software:

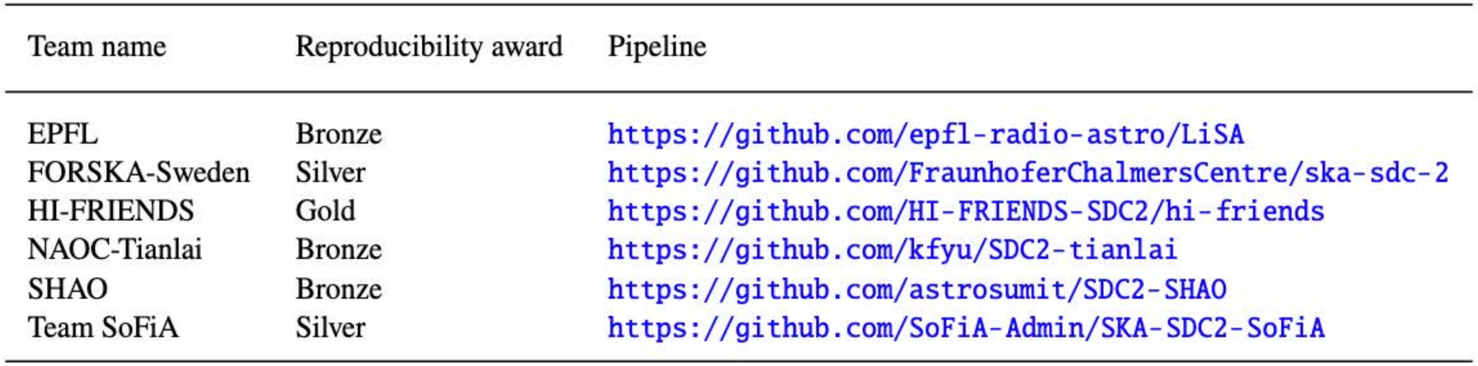
- Well-documented
- Easy to install
- Easy to use

Reusability:

Does the software:

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

Results



Team name	Reproducibility award	Pipeline
EPFL	Bronze	https://github.com/epfl-radio-astro/LiSA
FORSKA-Sweden	Silver	https://github.com/FraunhoferChalmersCentre/ska-sdc-2
HI-FRIENDS	Gold	https://github.com/HI-FRIENDS-SDC2/hi-friends
NAOC-Tianlai	Bronze	https://github.com/kfyu/SDC2-tianlai
SHAO	Bronze	https://github.com/astrosumit/SDC2-SHAO
Team SoFiA	Silver	https://github.com/SoFiA-Admin/SKA-SDC2-SoFiA

Award announcement to be featured in next edition of Contact

Tiered EoR Data Challenge

- **SDC3a Foregrounds**
- Foreground Subtraction + 21cm Power Spectrum Extraction (SWG contacts: Trott & Jelic)
- Target Participants: SWGs like CD/EoR, Cosmology, Continuum, etc.
  - Input Data: Calibrated Visibilities and High Fidelity Image
- Challenge will be based on:

a) Ability to remove the point source + diffuse foregrounds from the data-set

b) Ability to extract the cylindrical power spectrum

- Verification of the results from participants

c) Comparison with the original input signal power spectrum

Tiered EoR Data Challenge

- **SDC3b Inference**
- Extraction of reionization parameters (SWG contacts: Mesinger & Mellema)
- Target Participants: SWGs like CD/EoR
  - Input Data: EoR PS + noise and residual foreground contamination
- Challenge will be based on:

a) Ability to extract the IGM and source properties

- Verification of the results from participants

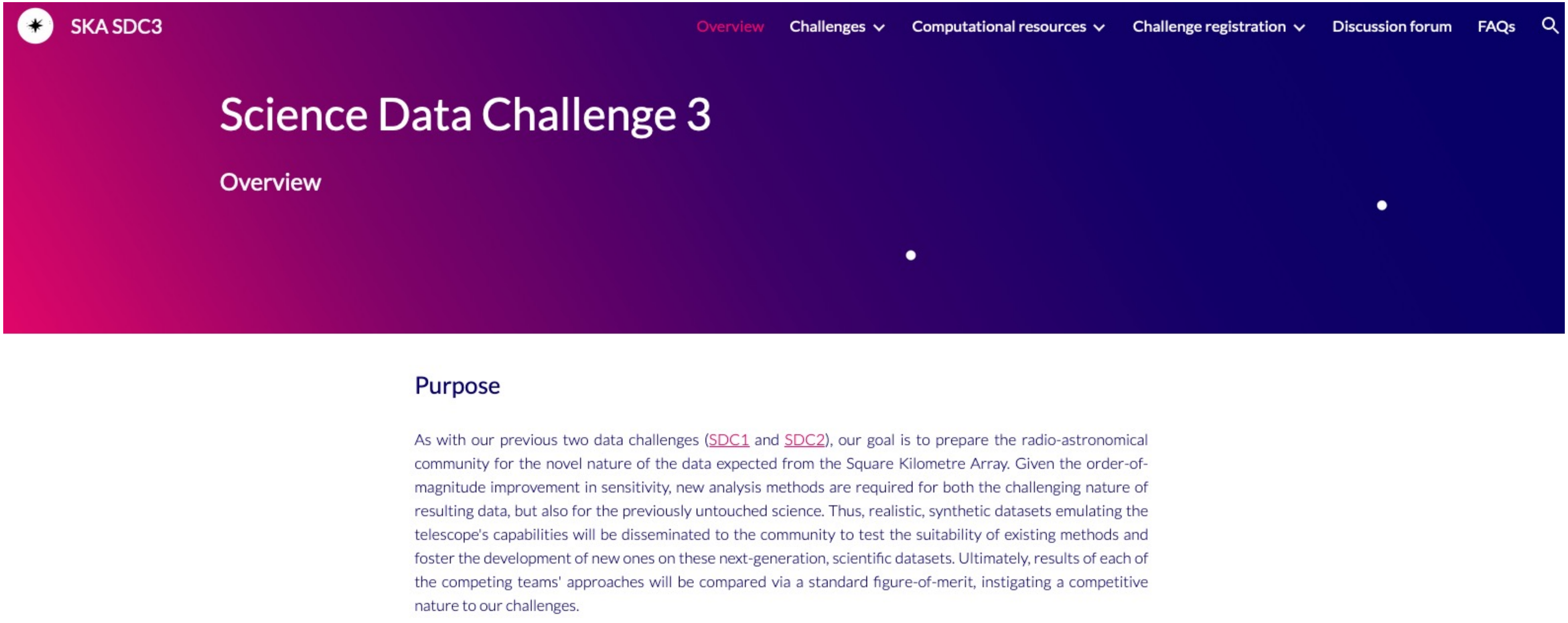
b) Comparison with the input ionisation history

Tiered EoR Data Challenge: Timeline

- SDC3a foregrounds: end of 2022, 6 months duration
- SDC3b inference: after SDC3 foregrounds, 6 months duration
  - Two independent datasets, different EoR model
  - Teams will be able to complete them individually
  - SDC3 foregrounds results will be propagated to the SDC3 inference simulation by adding foreground residuals to the input EoR PS and/or filtering some modes

Tiered EoR Data Challenge

sdc3.skao.int



Science Data Challenge 3

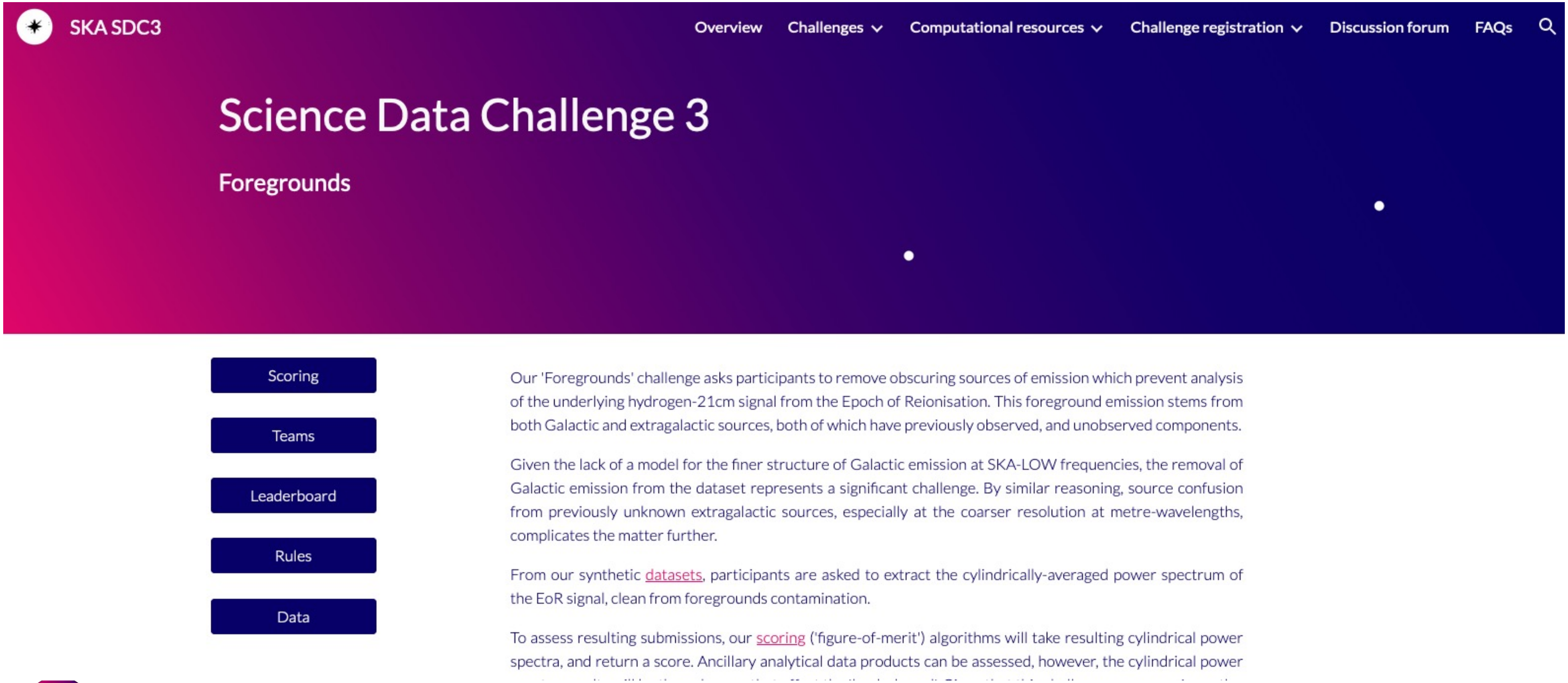
Overview

Purpose

As with our previous two data challenges (SDC1 and SDC2), our goal is to prepare the radio-astronomical community for the novel nature of the data expected from the Square Kilometre Array. Given the order-of-magnitude improvement in sensitivity, new analysis methods are required for both the challenging nature of resulting data, but also for the previously untouched science. Thus, realistic, synthetic datasets emulating the telescope's capabilities will be disseminated to the community to test the suitability of existing methods and foster the development of new ones on these next-generation, scientific datasets. Ultimately, results of each of the competing teams' approaches will be compared via a standard figure-of-merit, instigating a competitive nature to our challenges.

Tiered EoR Data Challenge

sdc3.skao.int



Science Data Challenge 3

Foregrounds

Our 'Foregrounds' challenge asks participants to remove obscuring sources of emission which prevent analysis of the underlying hydrogen-21cm signal from the Epoch of Reionisation. This foreground emission stems from both Galactic and extragalactic sources, both of which have previously observed, and unobserved components.

Given the lack of a model for the finer structure of Galactic emission at SKA-LOW frequencies, the removal of Galactic emission from the dataset represents a significant challenge. By similar reasoning, source confusion from previously unknown extragalactic sources, especially at the coarser resolution at metre-wavelengths, complicates the matter further.

From our synthetic datasets, participants are asked to extract the cylindrically-averaged power spectrum of the EoR signal, clean from foregrounds contamination.

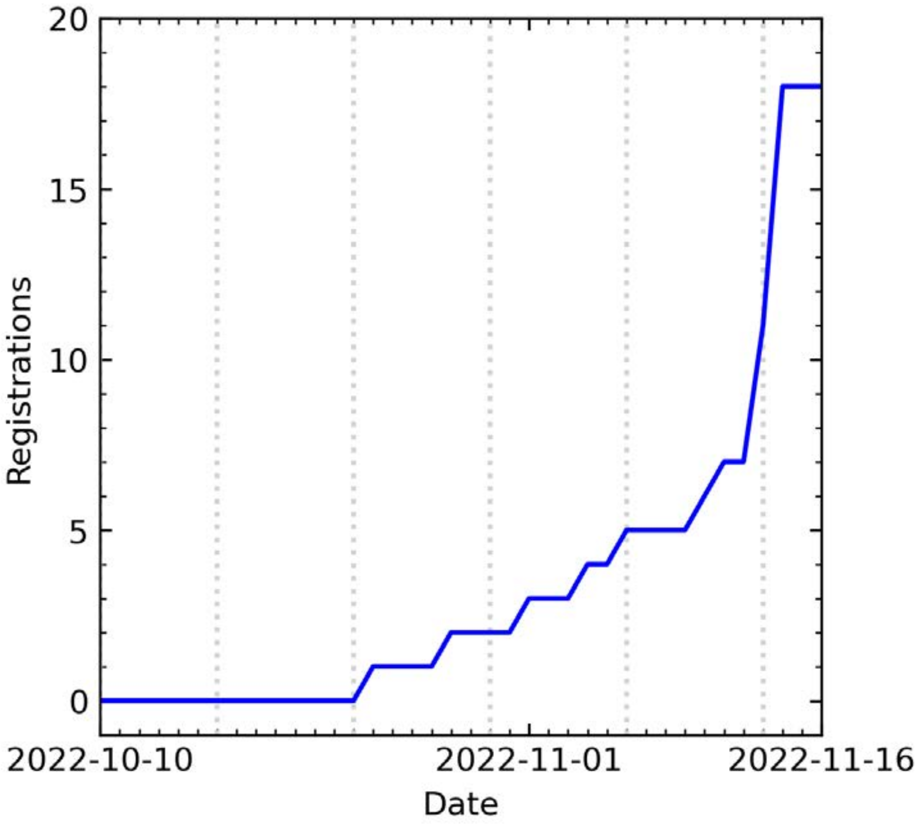
To assess resulting submissions, our scoring ('figure-of-merit') algorithms will take resulting cylindrical power spectra, and return a score. Ancillary analytical data products can be assessed, however, the cylindrical power

EoR Data Challenge: Computational Facility Partners

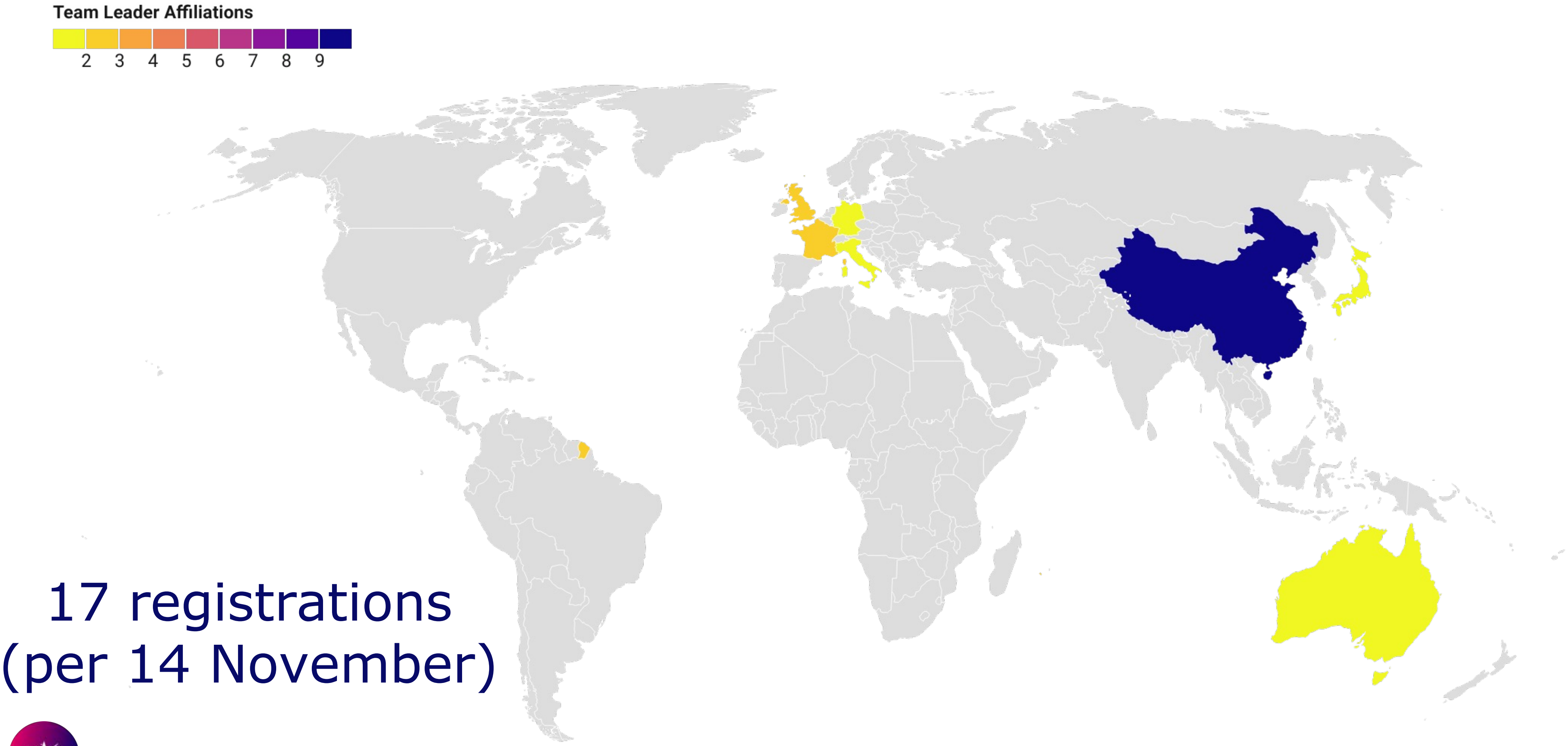
- Why computational facility partners?
  - Store the dataset in multiple locations, where teams will be able to access
  - Provide computational resources to inspect and analyse the dataset without transferring
- How will it work?
  - Teams will state their computational needs as part of the SDC3 registration
  - The SDC team will collaborate with the facility partners to identify the best matches with teams
  - Teams will access the data through the chosen facility
  - The data will be made available at multiple facilities at the same time to ensure a fair challenge
  - Teams will be able to process the data there
- Which facilities for SDC3?
  - IRIS
  - INAF ICT facility
  - SPSRC
  - GENCI-IDRIS
  - EngageSKA - UCLCA
  - Swiss SRC
  - ChinaSRC
  - ASTRON/SURF
  - AUS SRC
  - JPSRC

SDC3a Registrations

- Registration started on 10th October 2022
- To-date, there have been 18 registrations from 7 countries
- Expect more by end of registration today (15th November 2022)

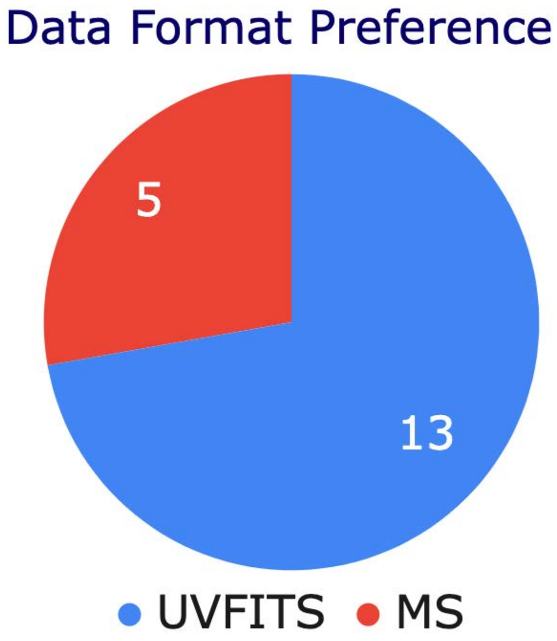


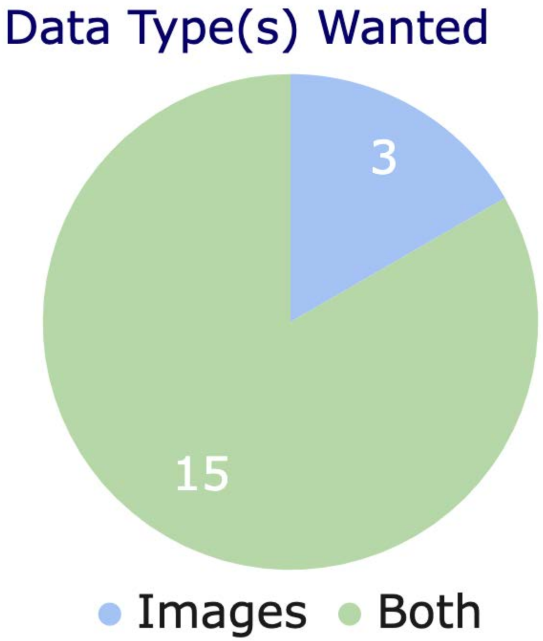
SDC3a Registrations

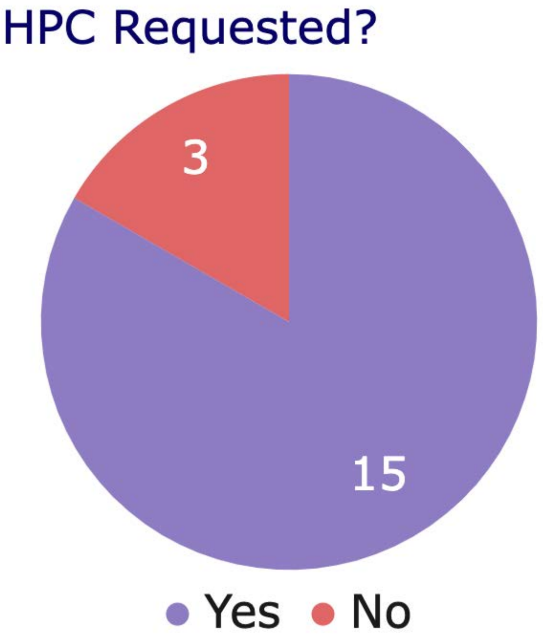


17 registrations
(per 14 November)

SDC3a Registrations







Science Meetings

- Joint ESO/SKAO Conference and Workshop was planned for week of 14 November 2022 “Coordinated Surveys of the Southern Sky”, in Garching: week of 27 February 2023
- Joint SKAO/ngVLA Science Conference week of 30 April 2023, in Vancouver
  - Web site in development, SOC formed
- EAS 2023, SKAO Lunch Session (1.5 hour) proposed
- IAU GA 2024 in Cape Town, several Letters of Intent for SKAO related Symposia have been submitted, including in EoR and HI areas, any news here?

Any Other Business

- New SWG mailing lists are now available (including core sub-lists) with same conventions throughout
  - e.g. swg-transients@skao.int, swg-col-core@skao.int, swg-vlbi@skao.int, swg-particles@skao.int
  - Old list names will still work
- Is there general SWG interest in central hosting of WG notes?
- News from SWG Chairs?
- …

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