one observatory
two telescopes
three continents

SKA Science Data Challenges

Philippa Hartley
SKAO Scientist

Swiss SKA Days 2022

The SKA data journey

SKA LOW

SKA MID

5 + 9 Tb/s
Approx 300 high definition movies per second!

SDP: Science Data Processor

SDP prototype, Cambridge, UK

600 PB/yr

SRC: SKA Regional Centre

SRC

SRC

SRC

SRC

SRC

SKA portal

Distributed facilities

User data products up to **TB**s in size

Key Science Projects

Scientific analysis

SKA Science Data Challenges

Primary goals:

- Familiarise the science community with **size and complexity of SKA data**
- Support the **design** of future SKA observations
- Drive the development of **data analysis techniques**

Additional benefits:

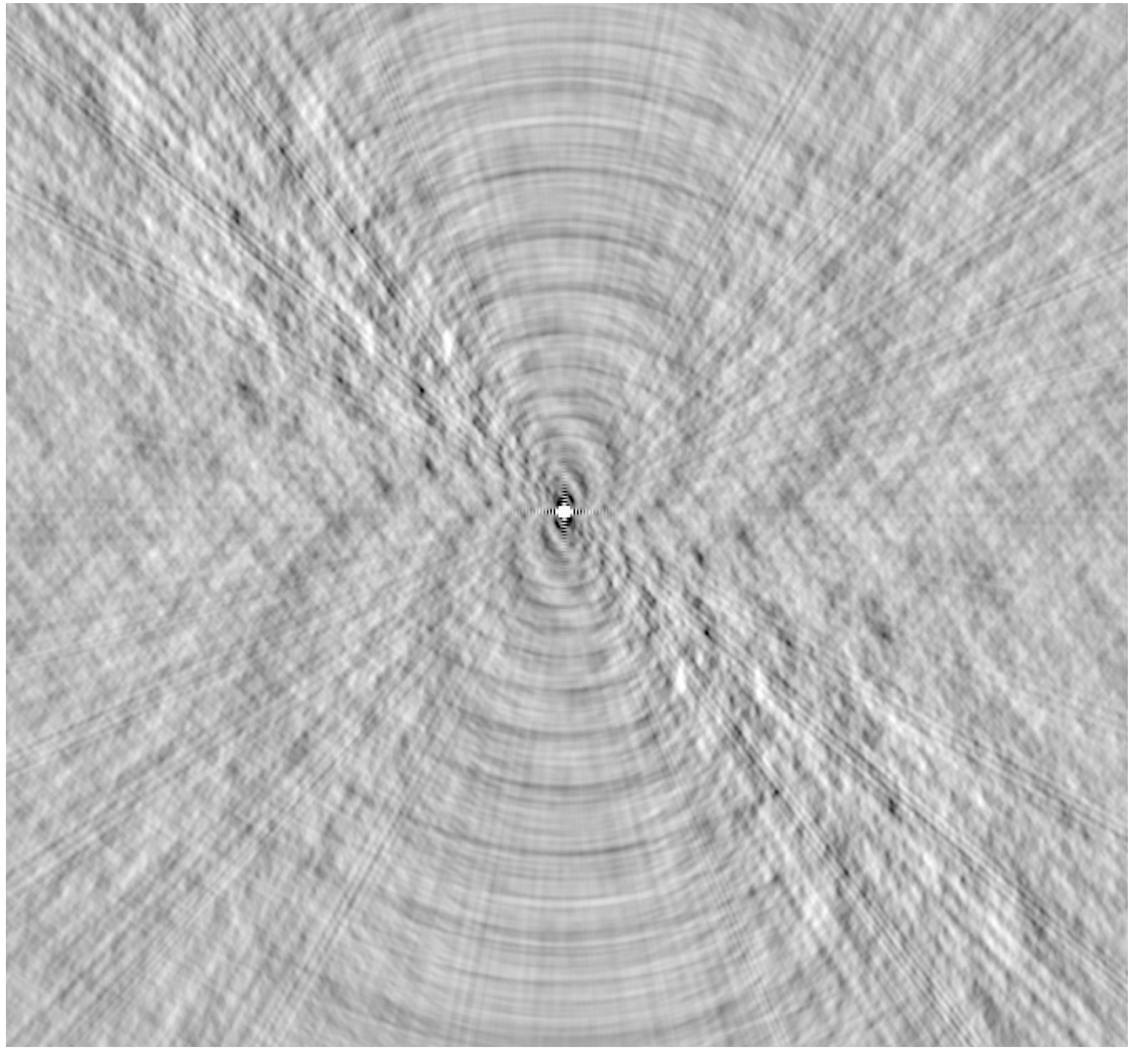
- Familiarise the science community with **data access models**
- Test SKA Regional Centre **prototyping**
- Encourage best practices for **Open Science** and **reproducibility**

SDC data products are made publicly available for the long term

SKA characteristics

SKA-unique features of the data products:

- In the **image plane**, not visibilities
- “**Benign**” dirty beam
- Deconvolved down to **8h exposures**
- Very deep -> **towards confusion limit**
- Very **large number** of sources to detect and classify

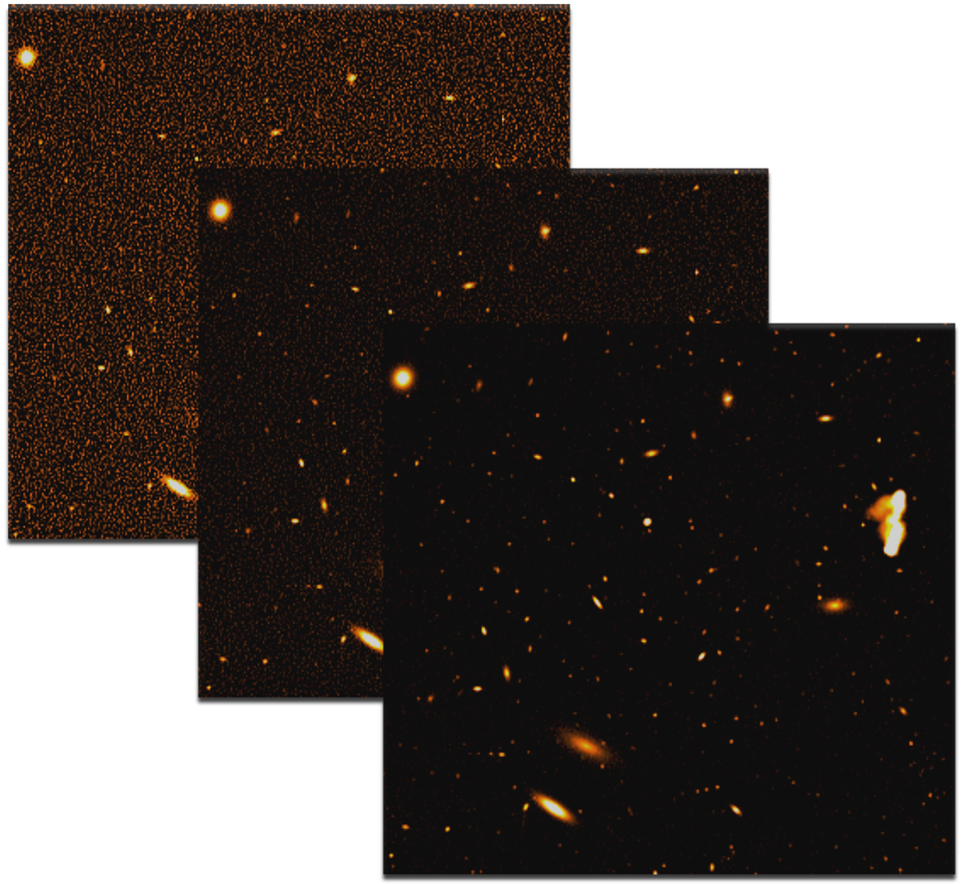


SKA MID 1.4 GHz beam

Science Data Challenge 1

Continuum emission

- **Continuum emission** images, simulating observations for SKA MID Bands 1, 2 and 5
- Images populated by star forming galaxies **(SFGs)** and active galactic nuclei **(AGN)**
- 3 telescope **integrations** each: 8, 100 and 1000h
- High telescope sensitivity → highly **crowded** images
- **The challenge**: to **find and characterise sources**
- **SDC1 website**



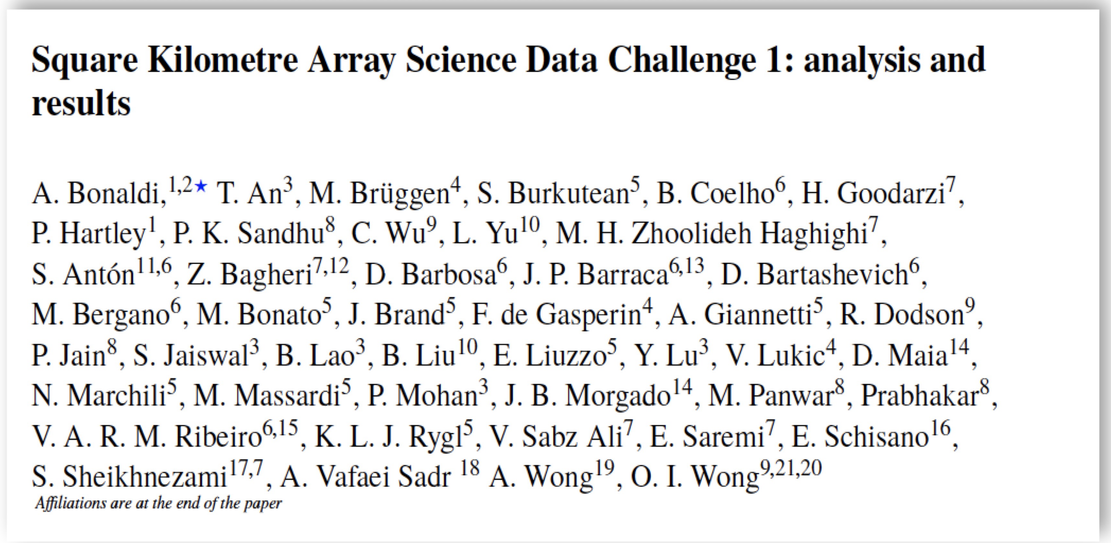
Zoom-in of the 1.4 GHz maps, showing the same region of the sky with different telescope integrations: 8, 100, 1000 h from left.

Science Data Challenge 1

Continuum observations

Main findings:

- Very **crowded** skies demand new approaches
- Variety of methods including **latest machine learning** techniques
- **Complementarity** of methods: tendency to score well either on finding galaxies *or* measuring them



Square Kilometre Array Science Data Challenge 1: analysis and results

A. Bonaldi,[1,2⋆] T. An[3], M. Brüggen[4], S. Burkutean[5], B. Coelho[6], H. Goodarzi[7], P. Hartley[1], P. K. Sandhu[8], C. Wu[9], L. Yu[10], M. H. Zhoolideh Haghighi[7], S. Antón[11,6], Z. Bagheri[7,12], D. Barbosa[6], J. P. Barraca[6,13], D. Bartashevich[6], M. Bergano[6], M. Bonato[5], J. Brand[5], F. de Gasperin[4], A. Giannetti[5], R. Dodson[9], P. Jain[8], S. Jaiswal[3], B. Lao[3], B. Liu[10], E. Liuzzo[5], Y. Lu[3], V. Lukic[4], D. Maia[14], N. Marchili[5], M. Massardi[5], P. Mohan[3], J. B. Morgado[14], M. Panwar[8], Prabhakar[8], V. A. R. M. Ribeiro[6,15], K. L. J. Rygl[5], V. Sabz Ali[7], E. Saremi[7], E. Schisano[16], S. Sheikhnezami[17,7], A. Vafaei Sadr [18] A. Wong[19], O. I. Wong[9,21,20]

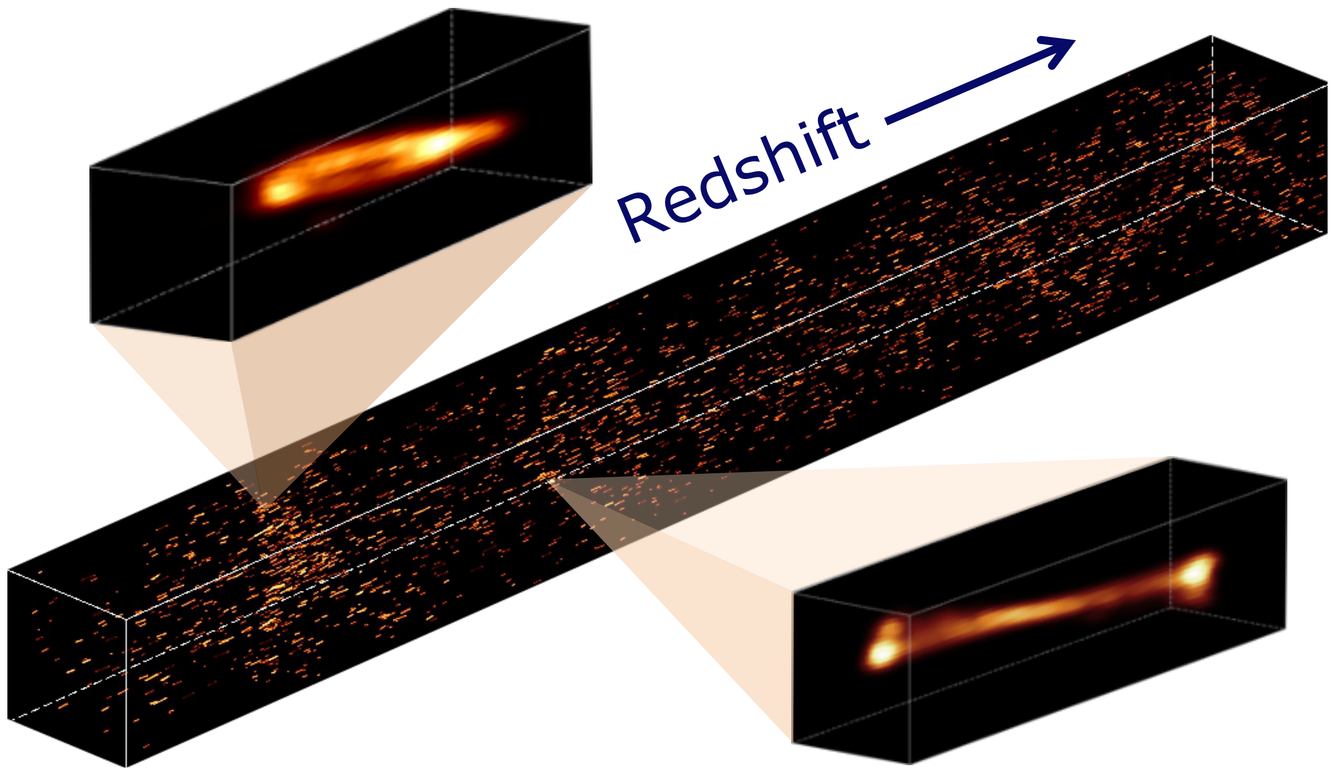
Affiliations are at the end of the paper

Monthly Notices of the Royal Astronomical Society, Volume 500, Issue 3, January 2021, Pages 3821–3837

Science Data Challenge 2

Neutral hydrogen (HI)

- **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**
- SDC2 website



Sample noise-free simulated HI image cube

SDC2 results paper

- 12 finalist teams from over **50** 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
- Results and analysis from SDC2 prepared for **submission to MNRAS**

SKA Science Data Challenge 2: analysis and results

P. Hartley[0], A. Bonaldi[0], R. Braun[0], J. N. H. S. Aditya[50], S. Aicardi[2], L. Alegre[40], A. Chakraborty[15], X. Chen[43], S. Choudhuri[17], A. O. Clarke[0], J. S. Collinson[0], D. Cornu[1], L. Darriba[33], M. Delli Veneri[9], J. Forbrich[19], G. Fourestey[41], B. Fraga[12], A. Galan[41], J. Garrido[33], C. Gheller[29], F. Gubanov[10], H. Håkansson[22], M. J. Hardcastle[19], C. Heneka[8], D. Herranz[36], K. M. Hess[24,25,26], M. Jagannath[18], S. Jaiswal[50], R. J. Jurek[27], D. Korber[41], S. Kitaeff[28], D. Kleiner[29], B. Lao[50], X. Liu[43], A. Mazumder[15], J. Moldón[33], R. Monda[41], S. Ni[44], M. Önnheim[22], M. Parra[33], N. Patra[14], A. Peel[41], P. Salomé[1], S. Sánchez-Expósito[33], M. Sargent[41,51,52], B. Semelin[1], P. Serra[29], A. K. Shaw[13], A. X. Shen[30,31], A. Sjöberg[22], S. Smith[20], A. Soroka[10], V. Stolyarov[20,21], E. Tolley[41], M. C. Toribio[23], J. M. van der Hulst[25], A. Vafaei Sadr[47], L. Verdes-Montenegro[33], T. Westmeier[28], K. Xu[43], L. Yu[42], L. Zhang[45], X. Zhang[44], Y. Zhang[50], A. Alberdi[33], M. Ashdown[20], C.R. Bom[12], M. Brüggen[8], J. Cannon[34], R. Chen[42], J. Coles[20], F. Combes[1,5], J. Conway[23], J. Ding[45], J. Freundlich[4], L. Gao[44], Q. Guo[43], E. Gustavsson[22], M. Jirstrand[22], M. G. Jones[37], G. Józsa[35], P. Kamphuis[38], M. Lindqvist[23], B. Liu[42], Y. Liu[43], Y. Mao[46], A. Marchal[3], I. Márquez[33], A. Meshcheryakov[11], M. Olberg[23], N. Oozeer[35], M. Pandey-Pommier[39], W. Pei[43], B. Peng[42], J. Sabater[40], A. Sorgho[33], C. Tasse[6,7], A. Wang[50], Y. Wang[43], H. Xi[42], X. Yang[50], H. Zhang[45], J. Zhang[44], M. Zhao[44], S. Zuo[46]

Affiliations can be found after the references

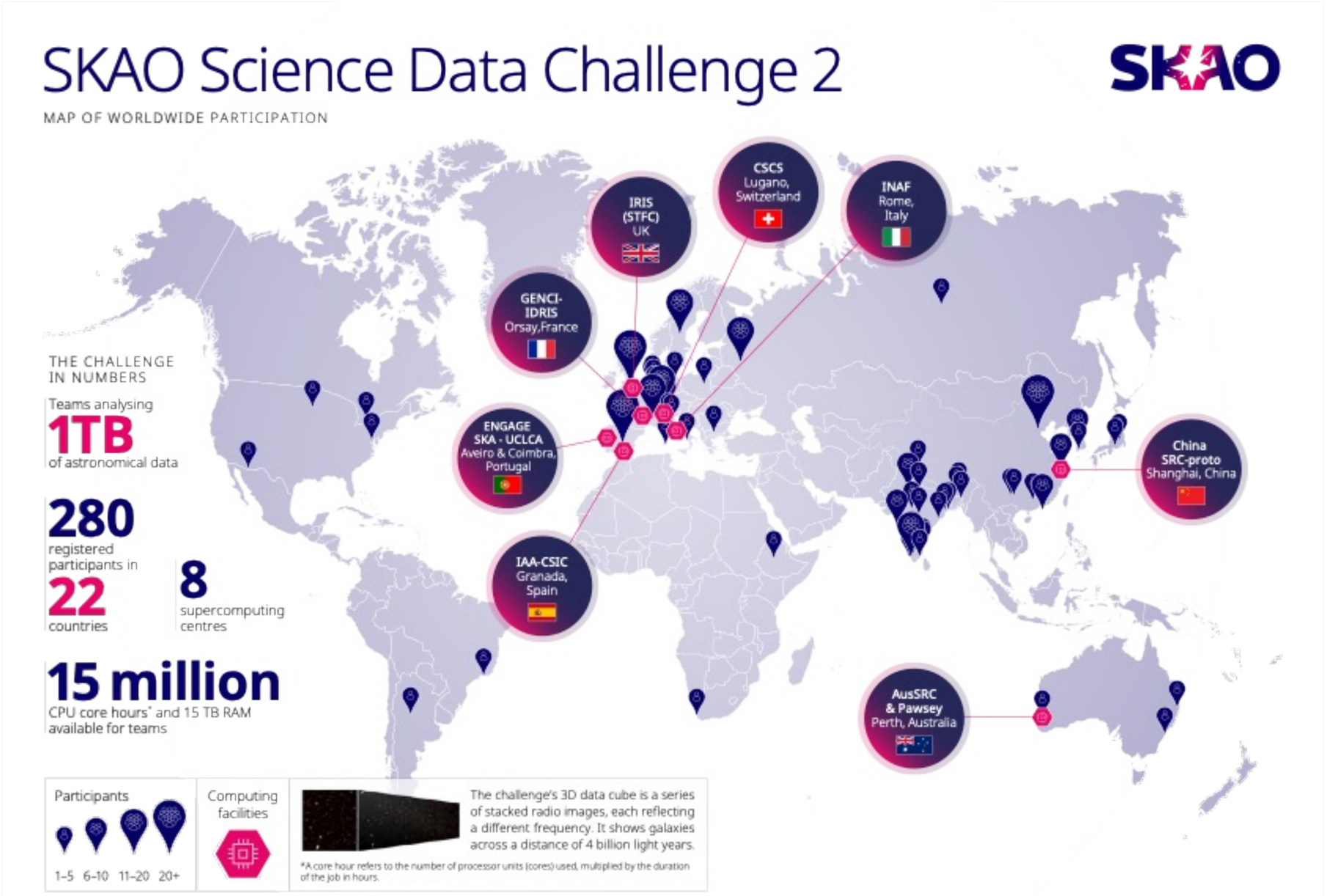
Accepted XXX. Received YYY; in original form ZZZ

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.

SDC computational facility partners

- **Support from eight** international computing facilities essential to success of SDC2
- Enabled accessible provision of **realistically large dataset**
- Test aspects of the future **SKA Regional Centre** model, e.g.:
  - Community data access
  - New technologies for distributed platform
  - SKA is committed to Open Science best practice

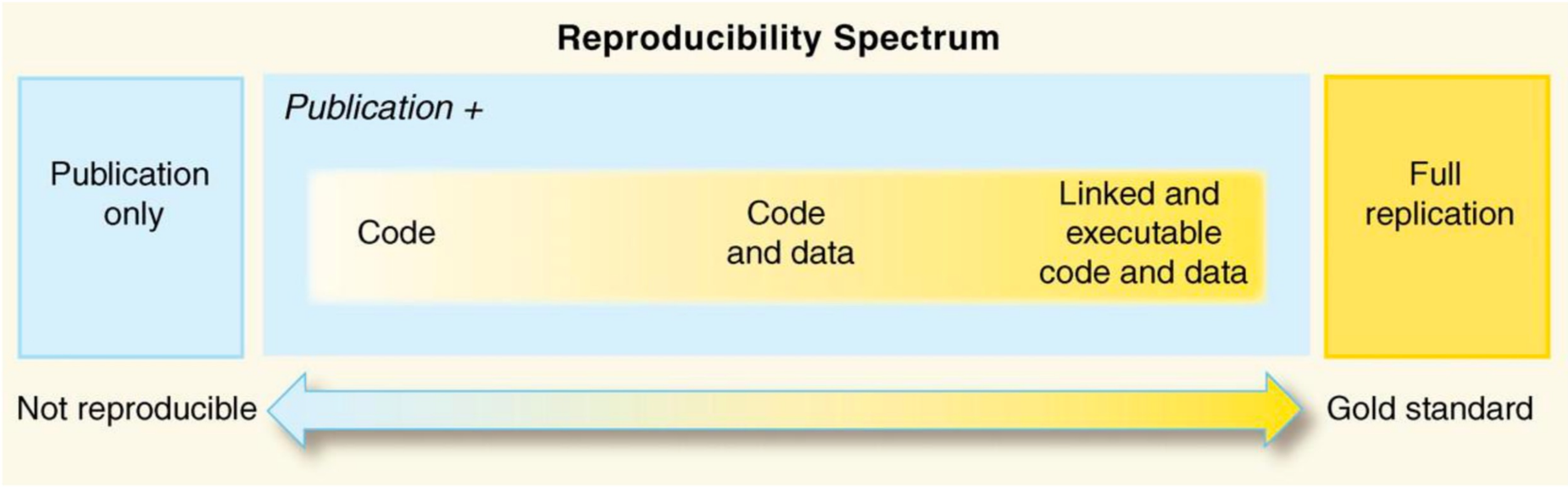


Reproducibility awards

In partnership with the Software Sustainability Institute

www.software.ac.uk

- An essential part of the scientific method, **reproducibility** leads to better, more efficient science.
- **Reusability** generalises this principle to create software that can be adapted by others, allowing previous work to be built upon for the future: a key feature of Open Science
- SKA is committed to delivering on the **FAIR** principles for scientific data management



Credit: Rachael Ainsworth

Reproducibility awards

In partnership with the Software Sustainability Institute

www.software.ac.uk

Reproducibility:

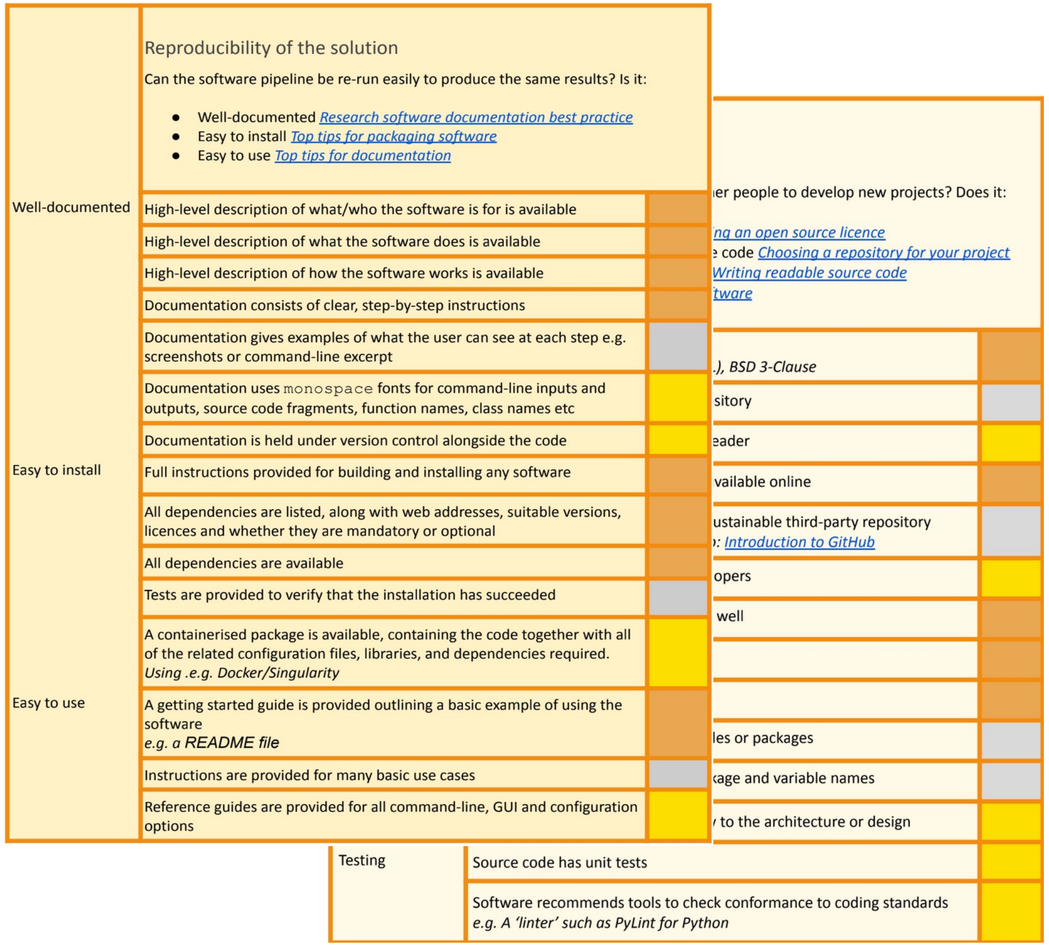
Is the software:

- 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

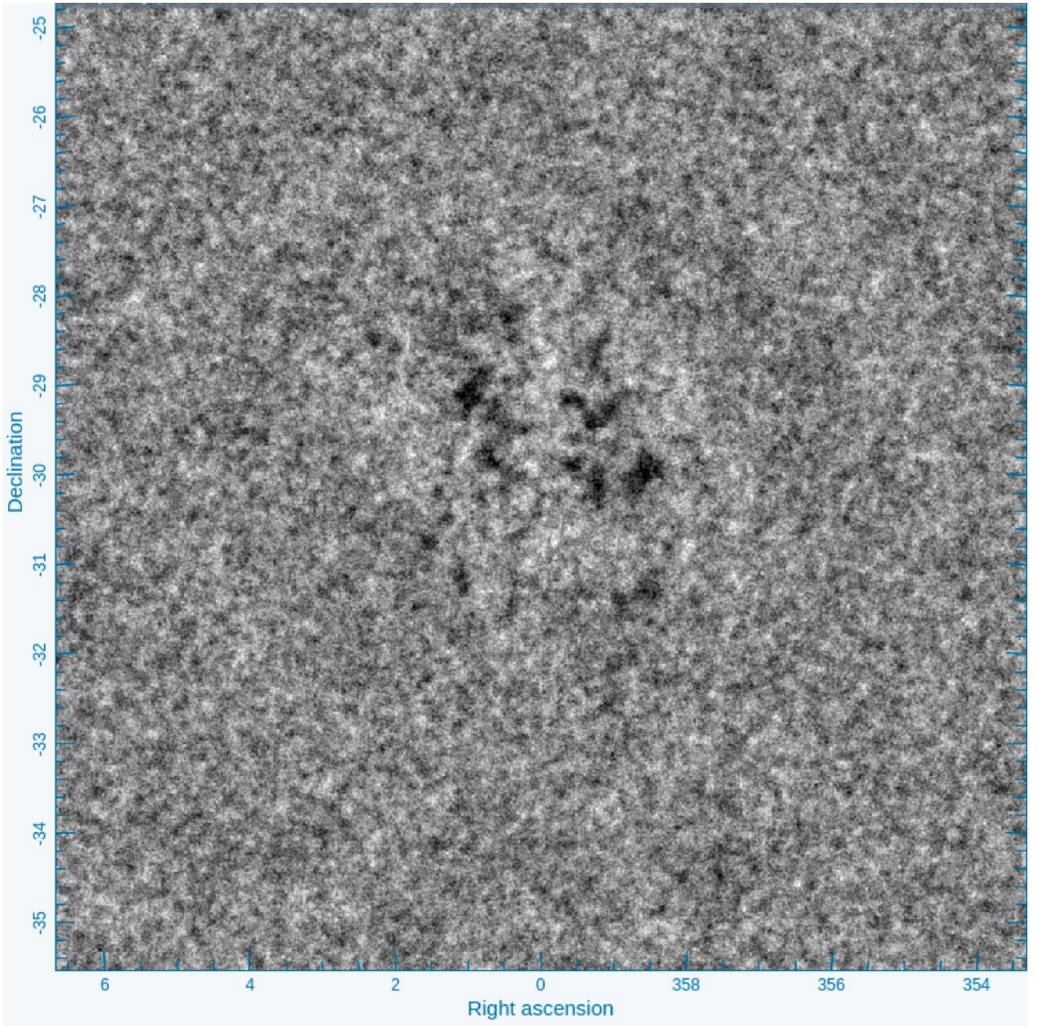


Science Data Challenge 3
Epoch of Reionisation (EoR)

Developing in collaboration with SKA EoR SWG members

Two parts:

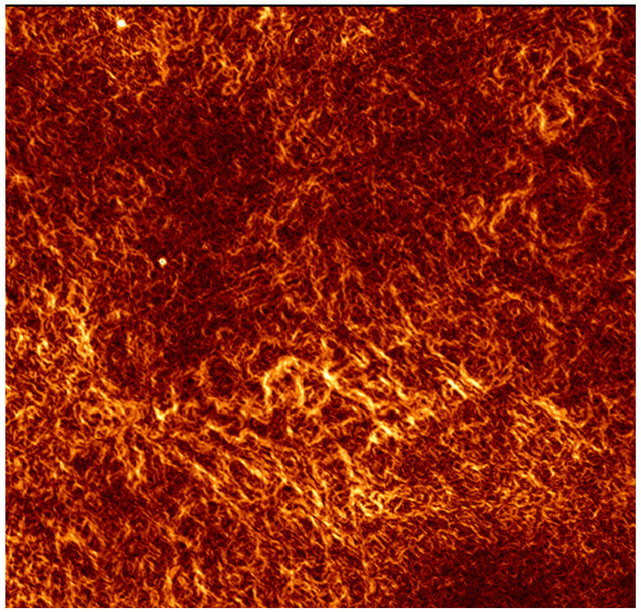
- SDC3 ”**Foregrounds**” (SDC3a; SWG Coordinators: C. Trott, V. Jelic)
  - **Foreground removal** exercise
  - SDC3a registration **will open soon**: SDC3 website
- SDC3 “**Inference**” (SDC3b; SWG Coordinators: A. Mesinger, G. Melema)
  - Extraction of **cosmological parameters**
  - SDC3b launching 2023



Sample EoR signal with noise added

Future Science Data Challenges

- **Cosmic magnetism** SDC (T. Akahori+), **Transients** SDC, and more
- Stay tuned for news and updates!



‘Snakes’ of cosmic magnetism. Credit: B. Gaensler et al.

Quasar schematic. Credit: NASA