one observatory
two telescopes
three continents

SKAO Science Data Challenges

Philippa Hartley
SKAO Scientist

SKA-China Workshop on SKA Science and Operations
Wednesday 28th September 2022

The SKAO 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

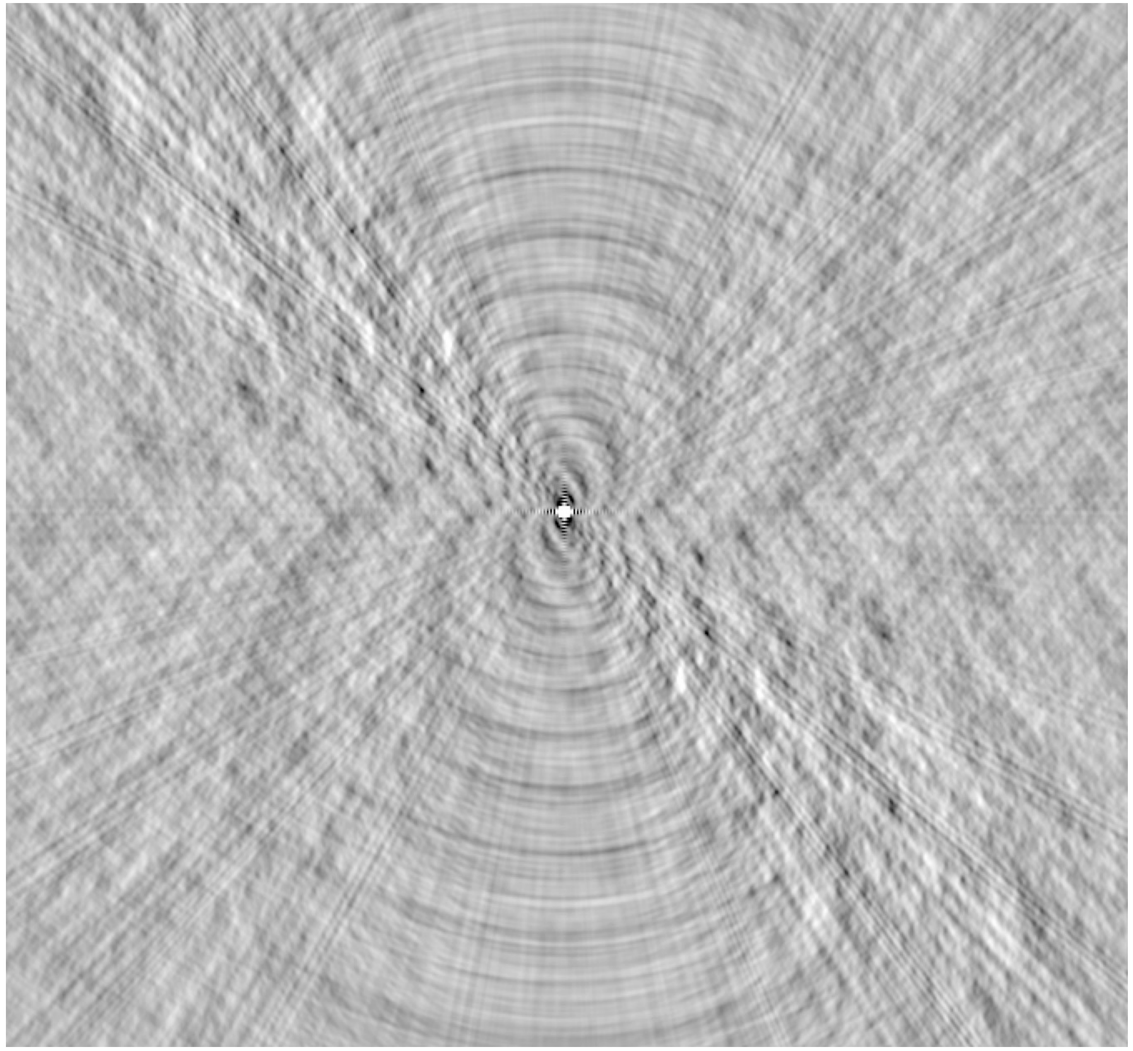
Key Science Projects

Scientific analysis

SKA telescope 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

SKAO 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

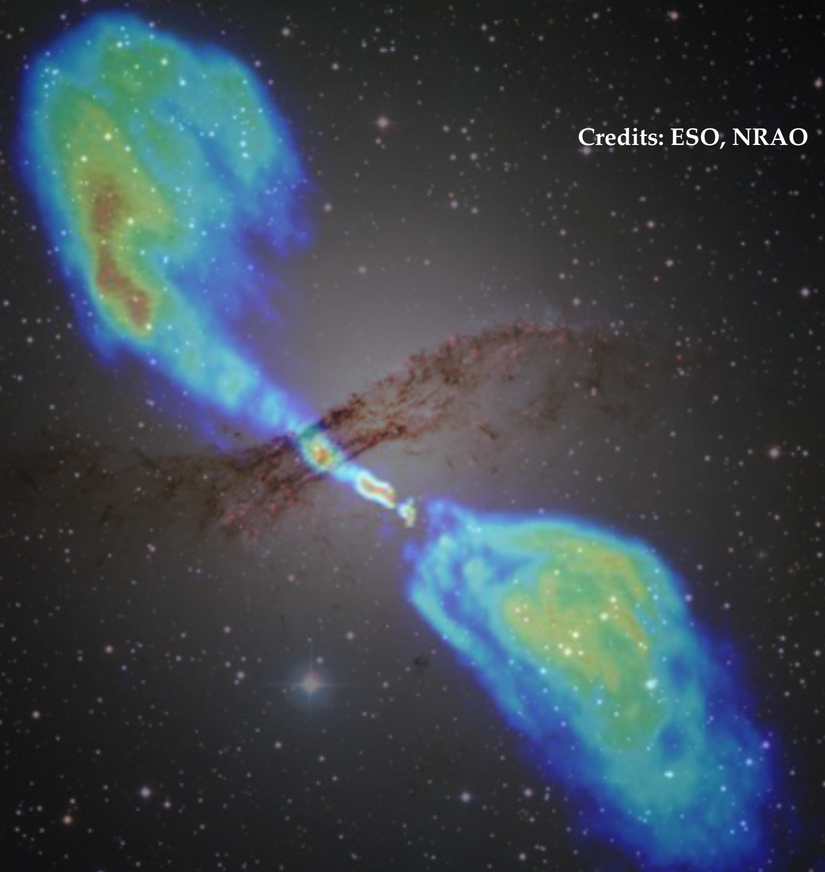
Science Data Challenge 1 (SDC1)

***Continuum* emission**

Continuum science with the SKAO:

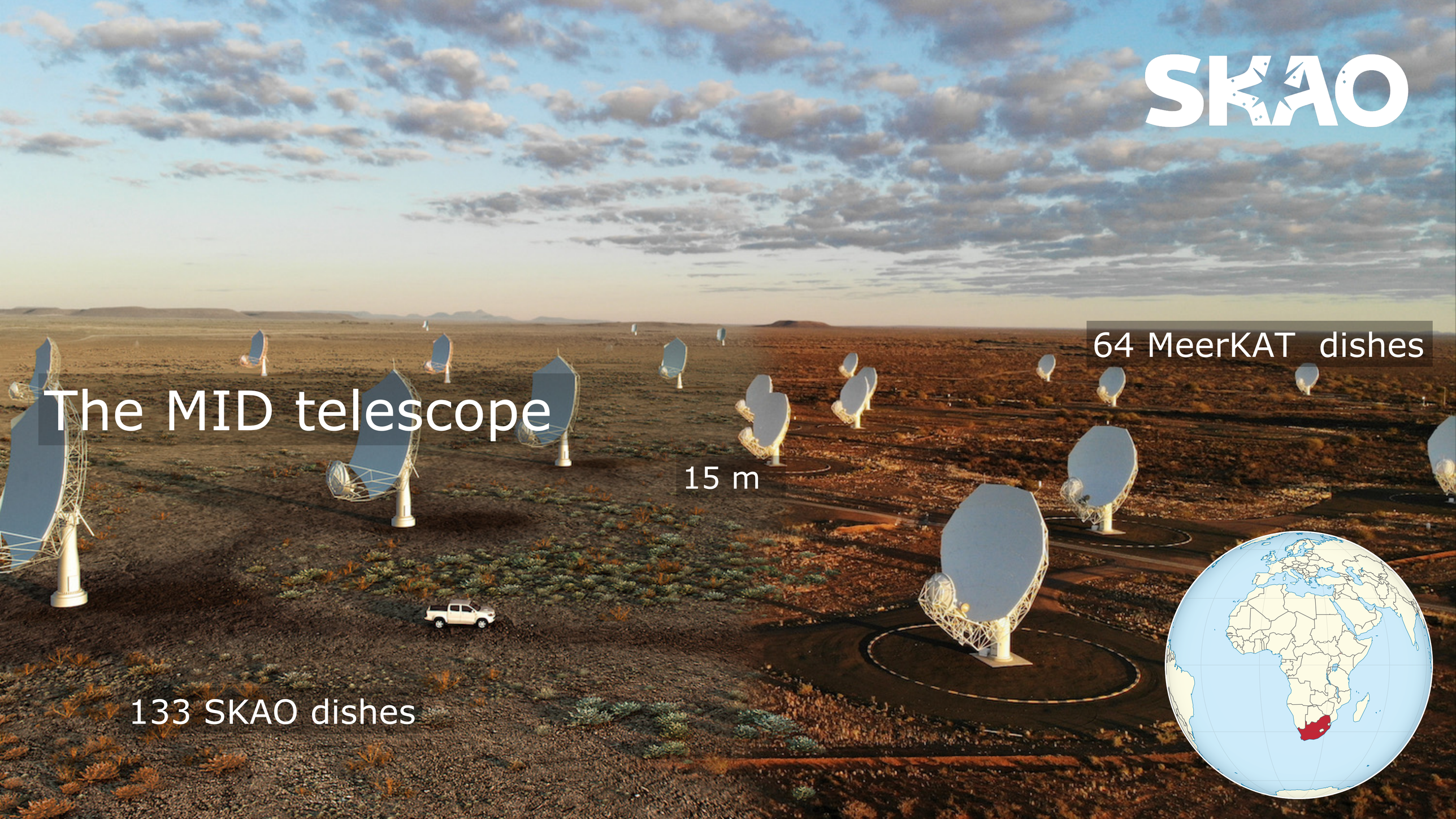
- explore our Universe from the nearest star-forming galaxies to high-redshift clusters of galaxies
- examine structure formation and evolution on all scales.

Extragalactic continuum Science Working Group:
https://www.skao.int/en/science-users/science-working-groups-focus-groups/107/extragalactic-continuum



Credits: ESO, NRAO

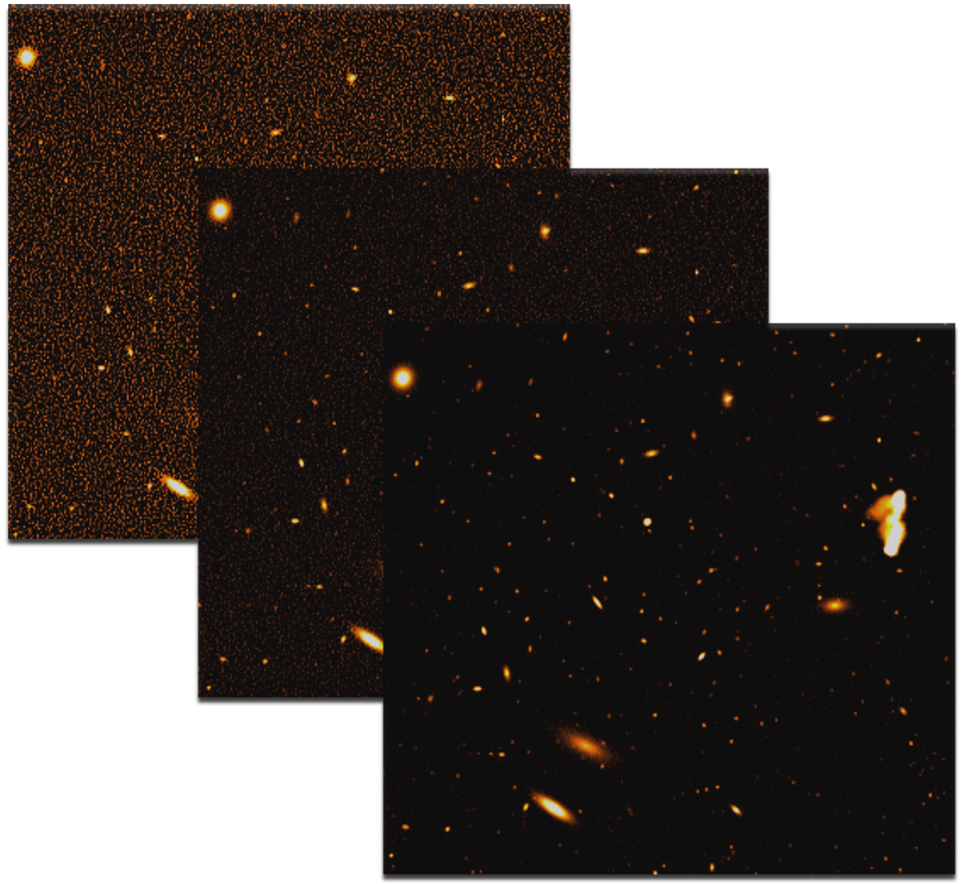
Centaurus A


SKAO
The MID telescope
64 MeerKAT dishes
15 m
133 SKAO dishes

Science Data Challenge 1 (SDC1)

***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 (SDC1)

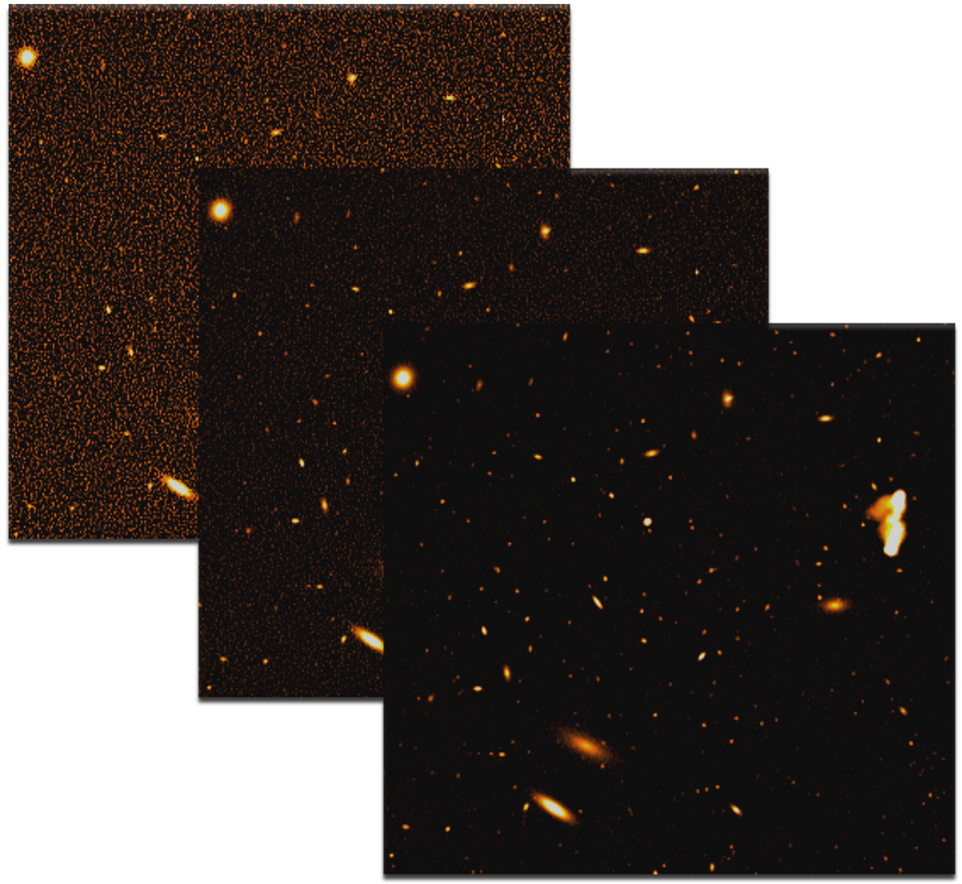
Continuum emission

- **25** square degrees
- **1.5, 0.60 and 0.0913** arcsec FWHM
- Dirty beam sidelobes **$4x10^{-4}$**

- No simulated calibration errors
- No simulated pointing errors

	560 MHz	1400 MHz	9200 MHz
8 h	2880	710	430
100 h	810	200	120
1000 h	255	73	38
confusion	15	0.36	0.0002

Noise RMS, nJy/beam

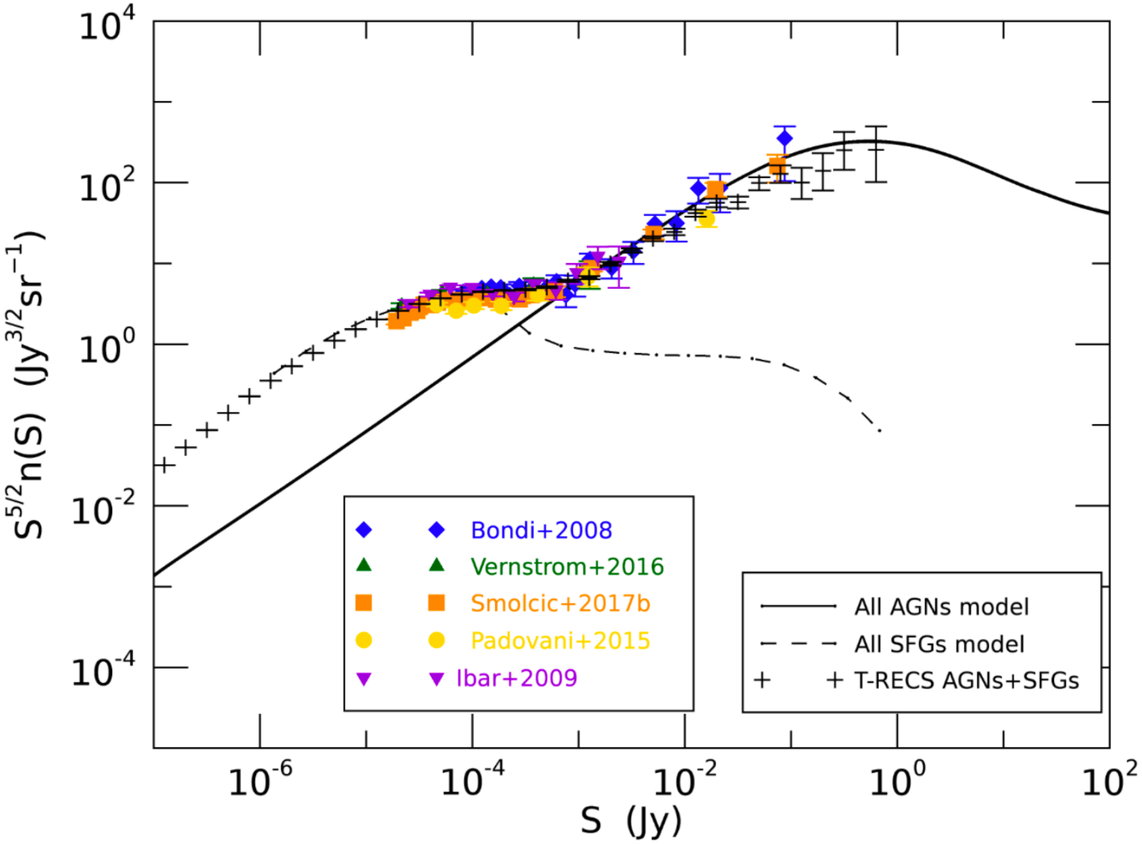


Zoom-in of the 1.4 GHz maps, showing the same region of the sky with different telescope integrations: 8, 100, 1000 h from left.

SDC1 simulations: continuum catalogue

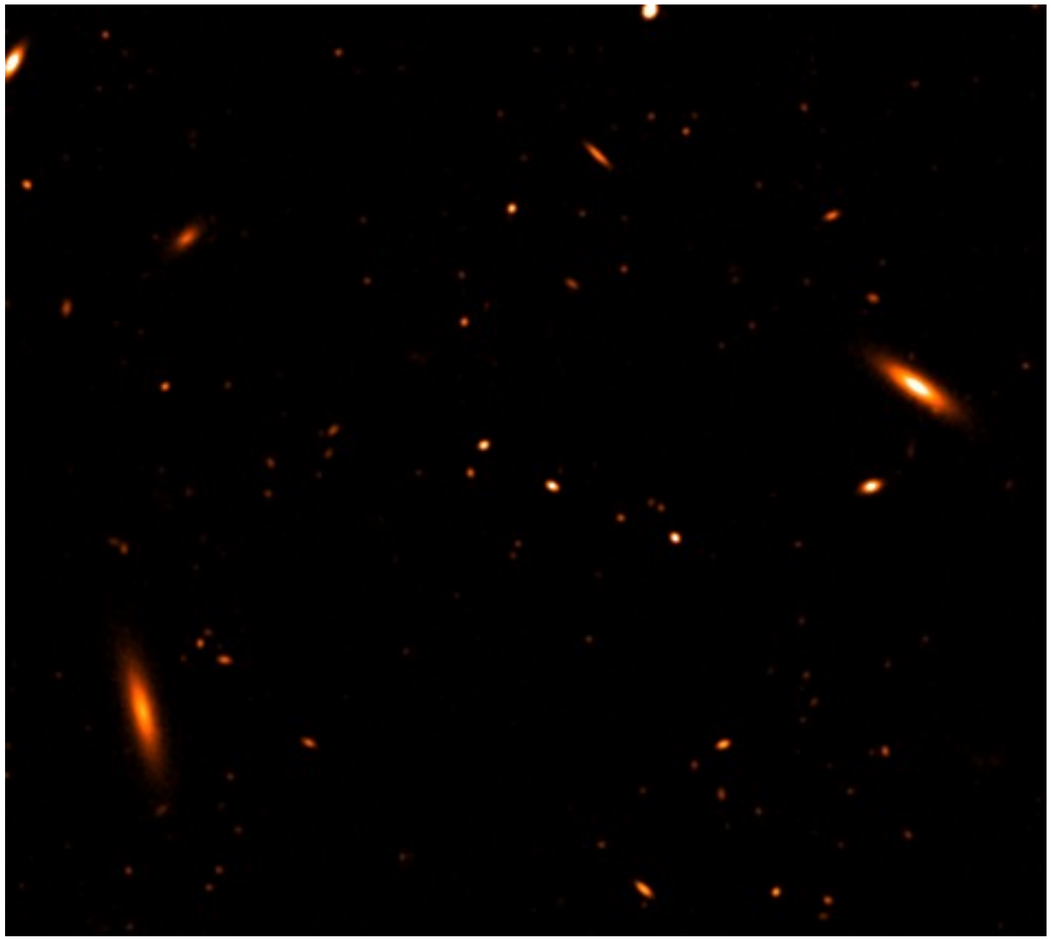
The Tiered Radio Extragalactic Continuum Simulation (T-RECS) *Bonaldi+ 2019*

- Good agreement with source counts at the **sub-mJy level**
- **Polarisation**
- Realistic **clustering** via P-millennium simulation (Baugh et al. 2018)



1.4 GHz differential source counts

SDC1 simulations: source morphology

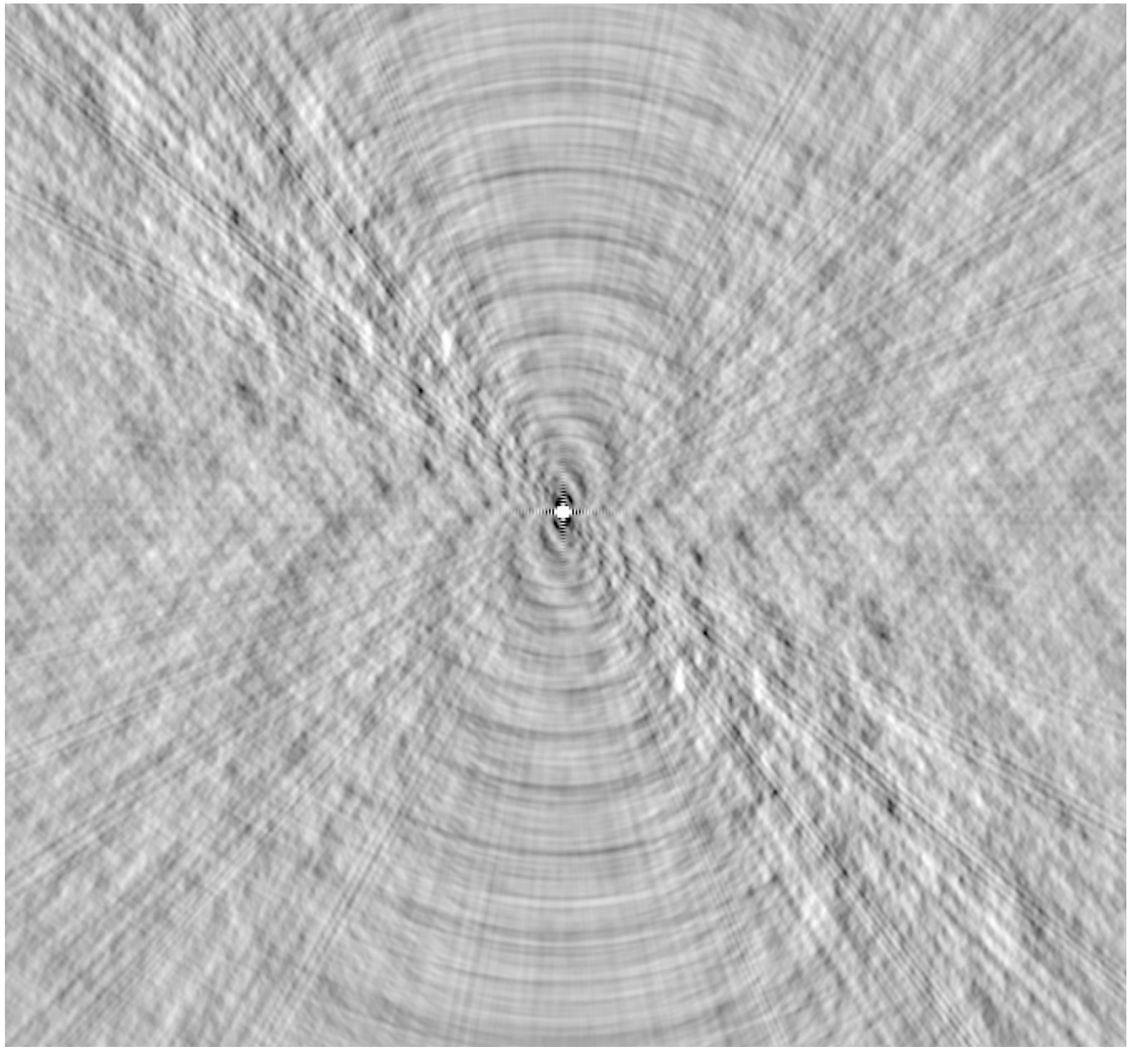


SFGs: Exponential Sersic profiles
Flat spectrum AGN: Gaussians and points

Steep spectrum AGN: DRAGNs
database, University of Manchester

SDC1 simulations: observed continuum

- Addition of **noise**
- **‘dirty’ beam** simulated
- Imperfect **deconvolution** signatures



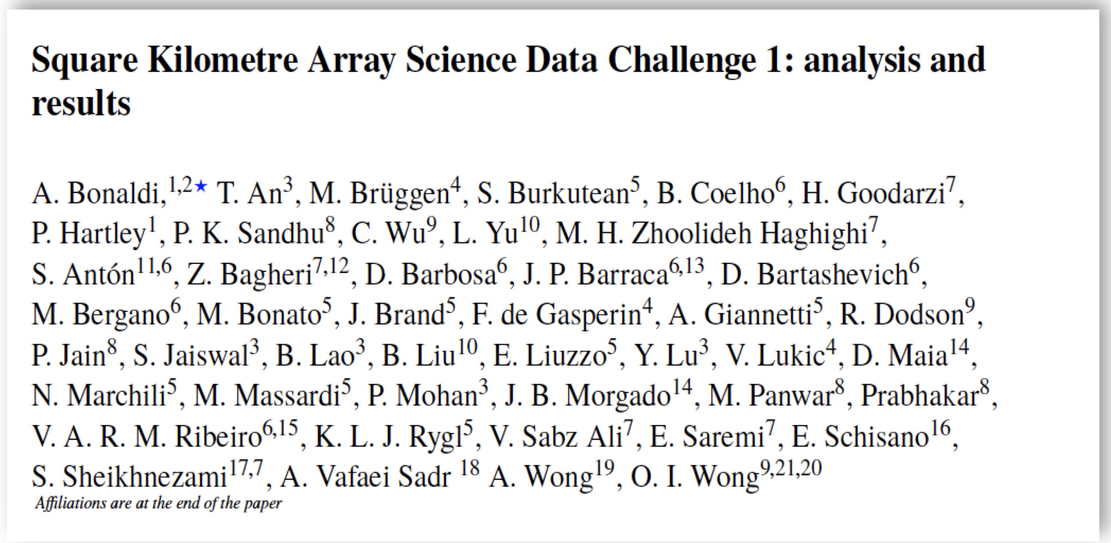
SKA MID 1.4 GHz beam

SDC1 results

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 (SDC2)

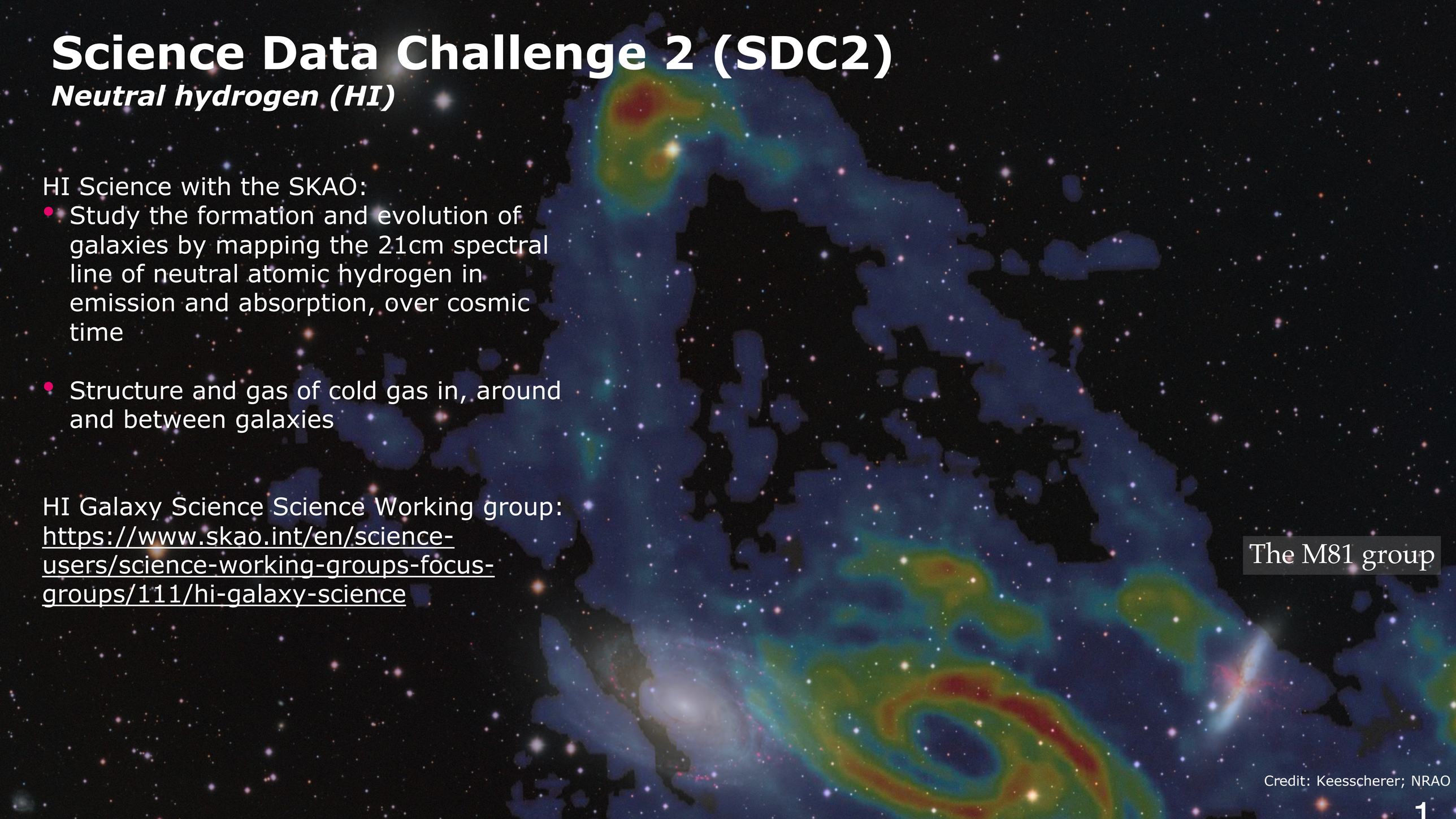
Neutral hydrogen (HI)

HI Science with the SKAO:

- Study the formation and evolution of galaxies by mapping the 21cm spectral line of neutral atomic hydrogen in emission and absorption, over cosmic time

- Structure and gas of cold gas in, around and between galaxies

HI Galaxy Science Science Working group: https://www.skao.int/en/science-users/science-working-groups-focus-groups/111/hi-galaxy-science



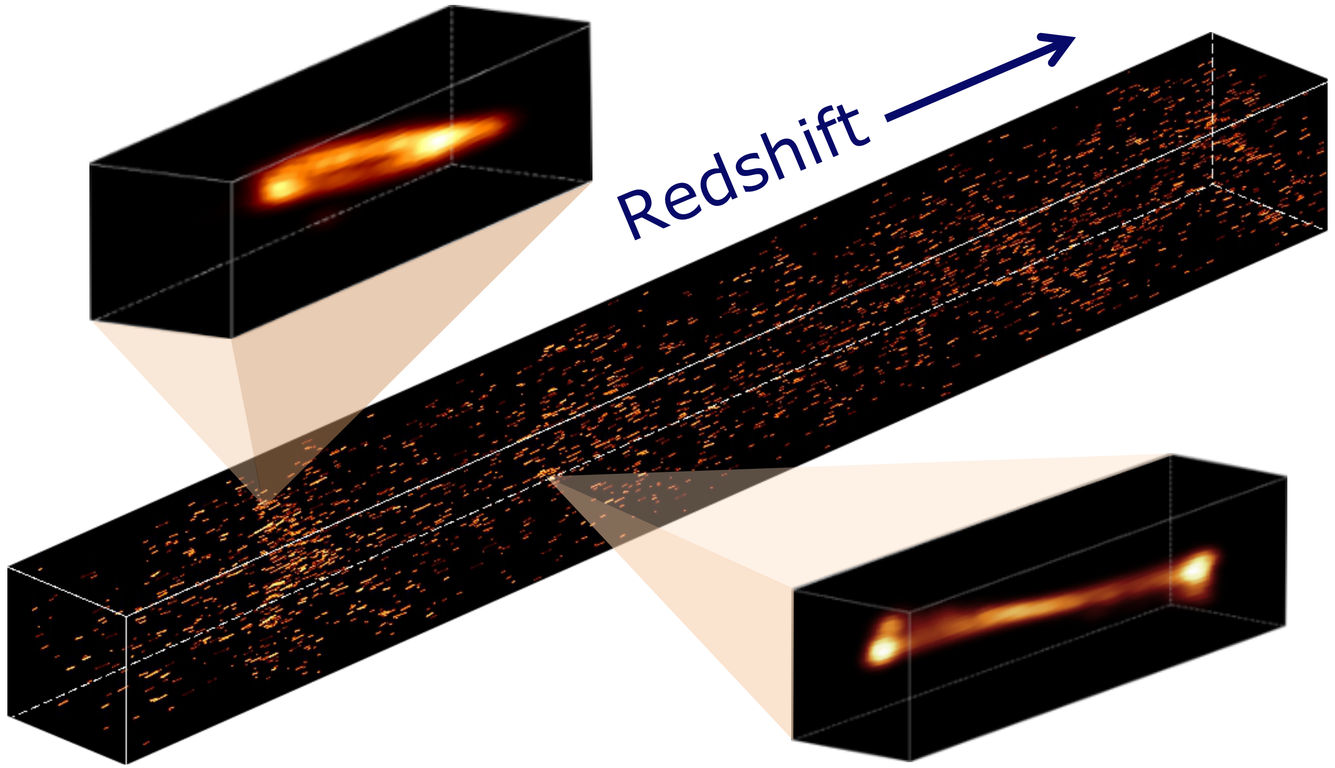
The M81 group

Credit: Keesscherer; NRAO

Science Data Challenge 2

Neutral hydrogen (HI)

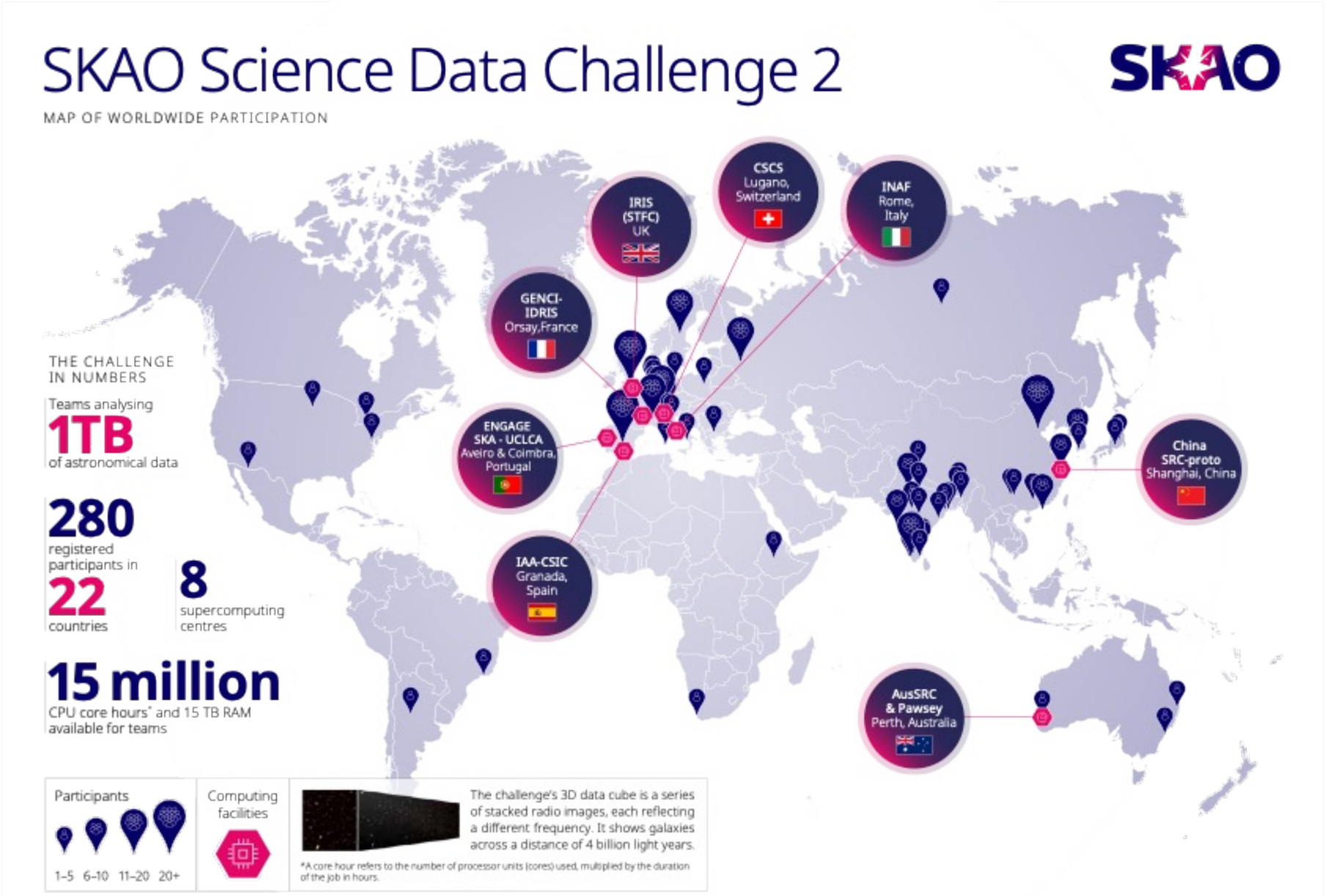
- **21cm spectral line** image cube, simulating deep **SKA MID** observations (redshift 0.25 to 0.5)
- Image cube populated by **HI** content of **galaxies**
- **2000 h** integration time across **20 sq deg** field of view
- The challenge**:** to **find and characterise HI sources**
- **Data volume = 1 TB**
- SDC2 website



Sample noise-free simulated HI image cube

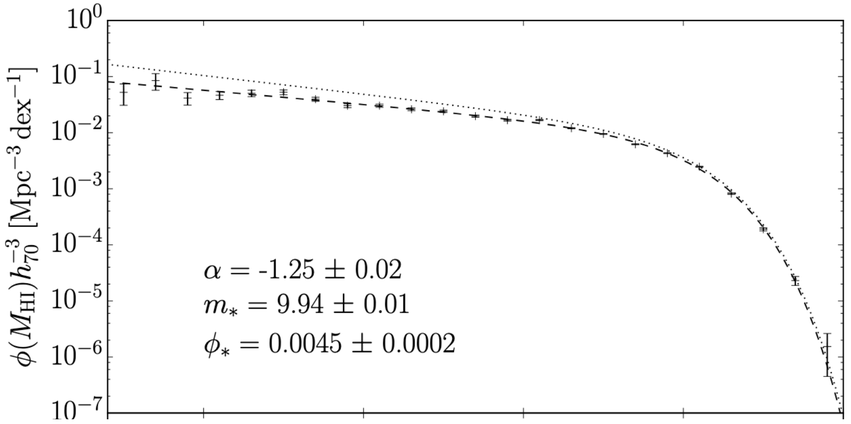
SDC computational facility partners

- **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

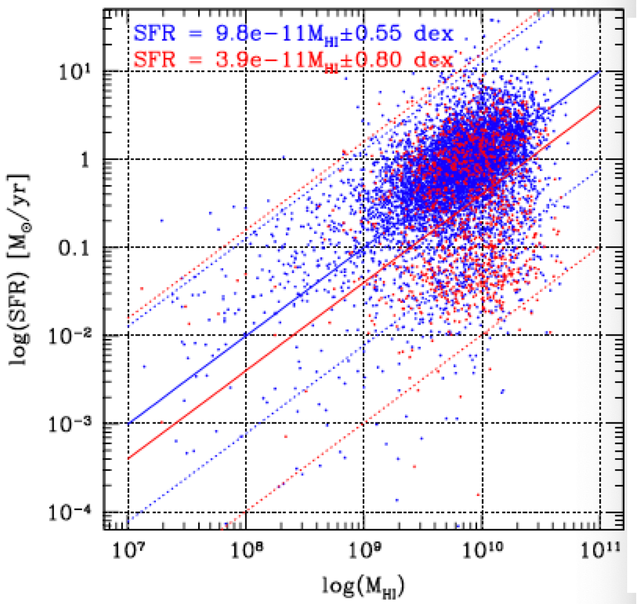


SDC2 simulations: HI and continuum catalogues

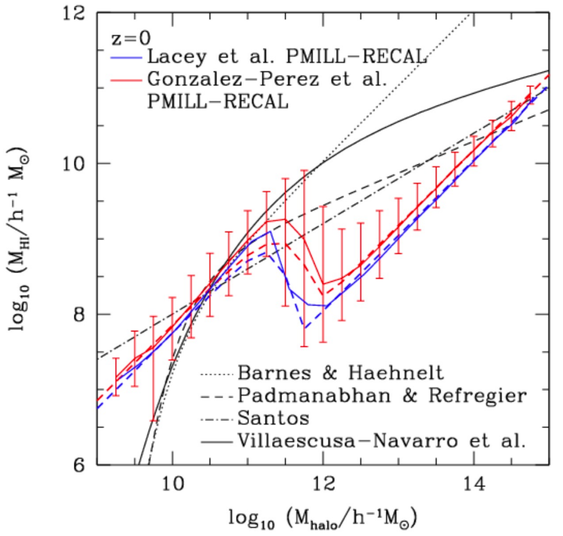
- HI and continuum **catalogues** produced using T-RECS modules
  - **HI catalogue**: HI masses drawn from ALFALFA survey HI mass function
  - **Continuum catalogue**: HI masses *predicted* from ALFAFLA M_{HI}-SFR (for SFGs) and P-Millennium M_{HI}-M_{halo} (for AGN)
- Large scale structure via P-millennium simulation
- Catalogues **cross-matched** by HI mass



Hi mass function, Jones+ 2018



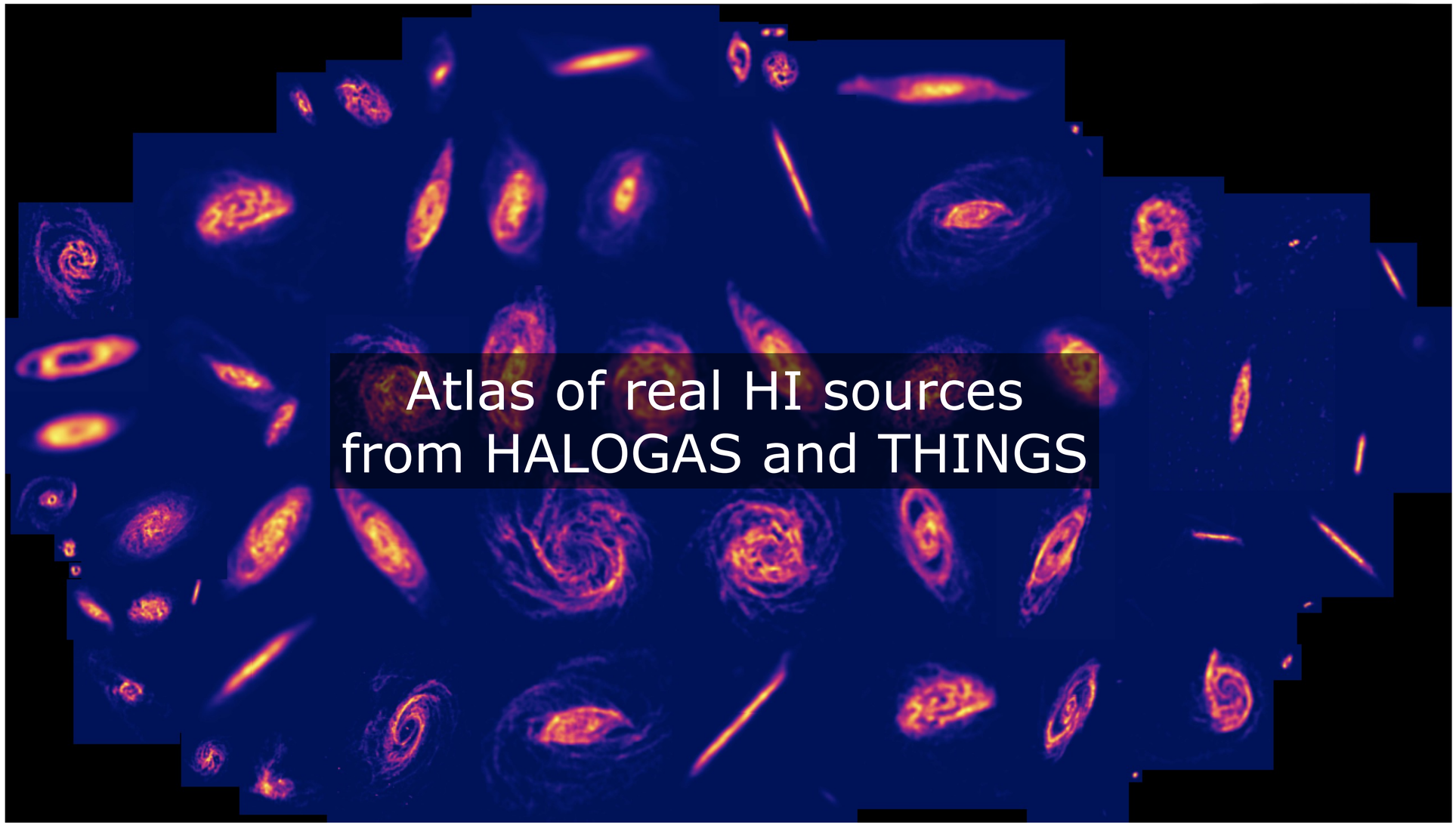
Data from ALFALFA, Giovanelli+ 2005



P-Millennium, Baugh+2018

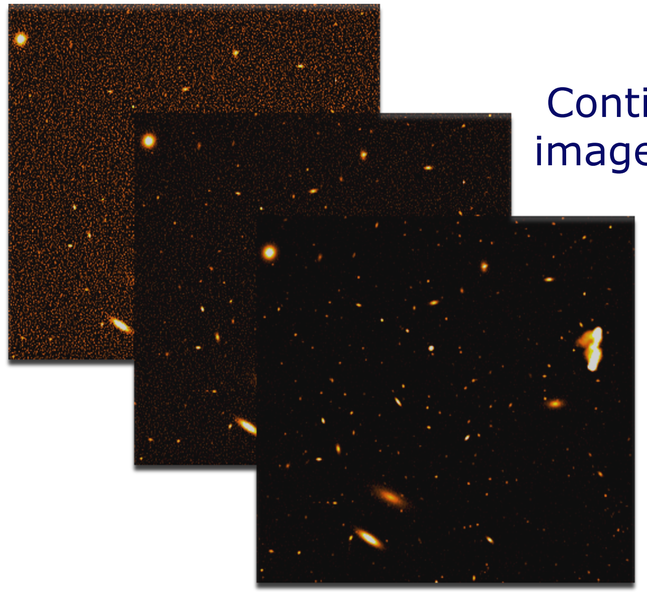
SDC2 simulations: source morphology

Heald et al. 2011;
Walter et al. 2008



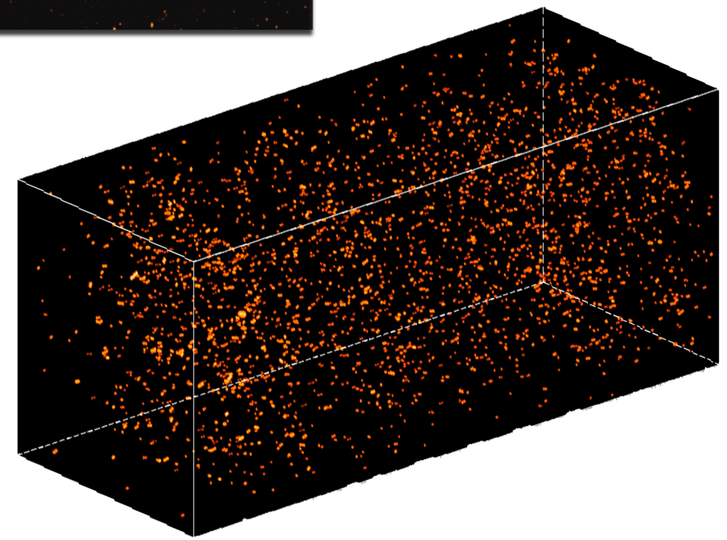
SDC2 simulations: observed HI

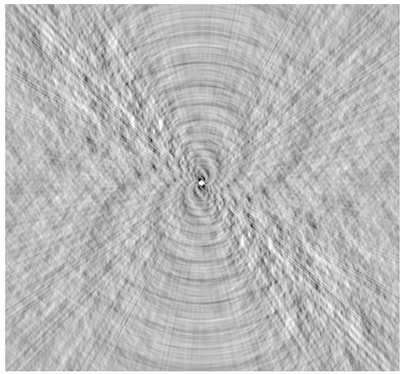
- HI **absorption** signatures calculated and net absorption/emission cube produced
- Imperfect **continuum subtraction**
- Residual **deconvolution** signatures using a simulated ‘dirty’ beam
- **Noise** added, with **RFI** flagging simulated



Continuum image cube

HI image cube





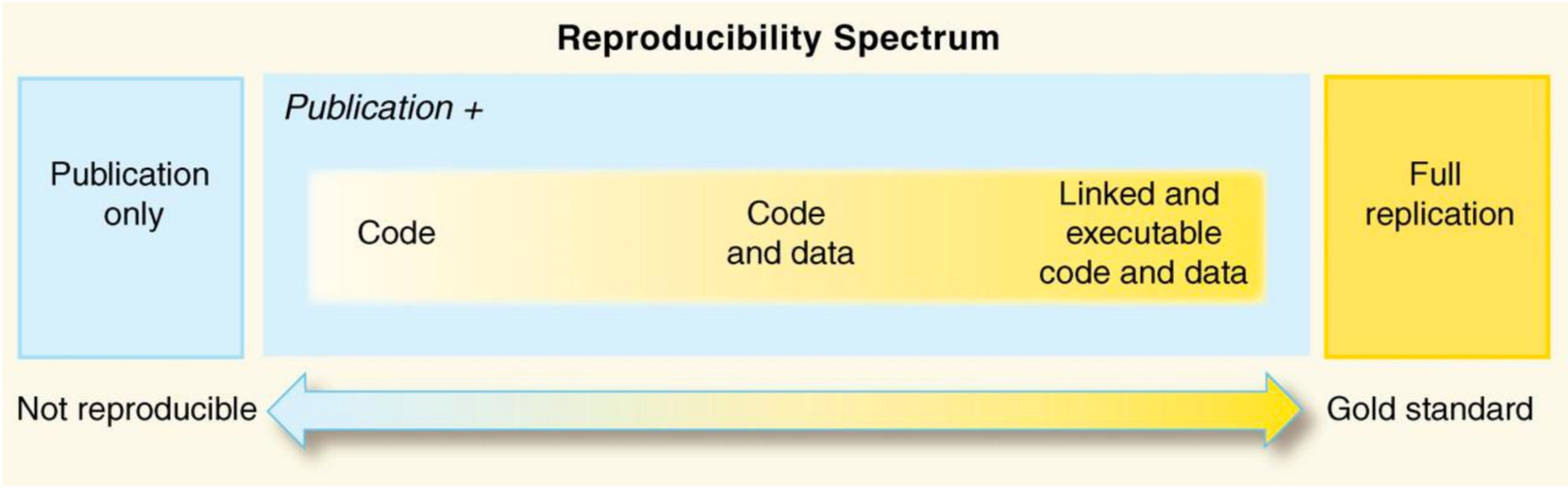
SKA MID 1.4 GHz beam

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:

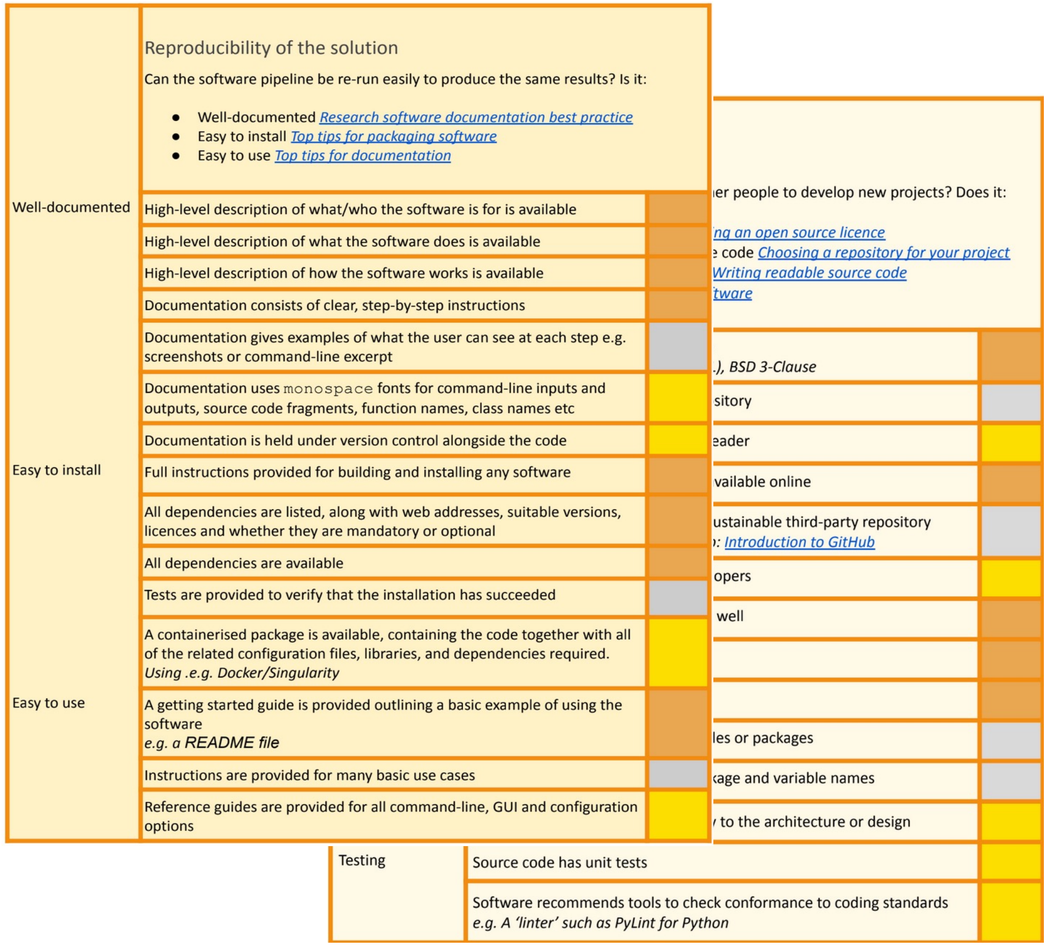
Is the software:

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

Reusability:

Does the software:

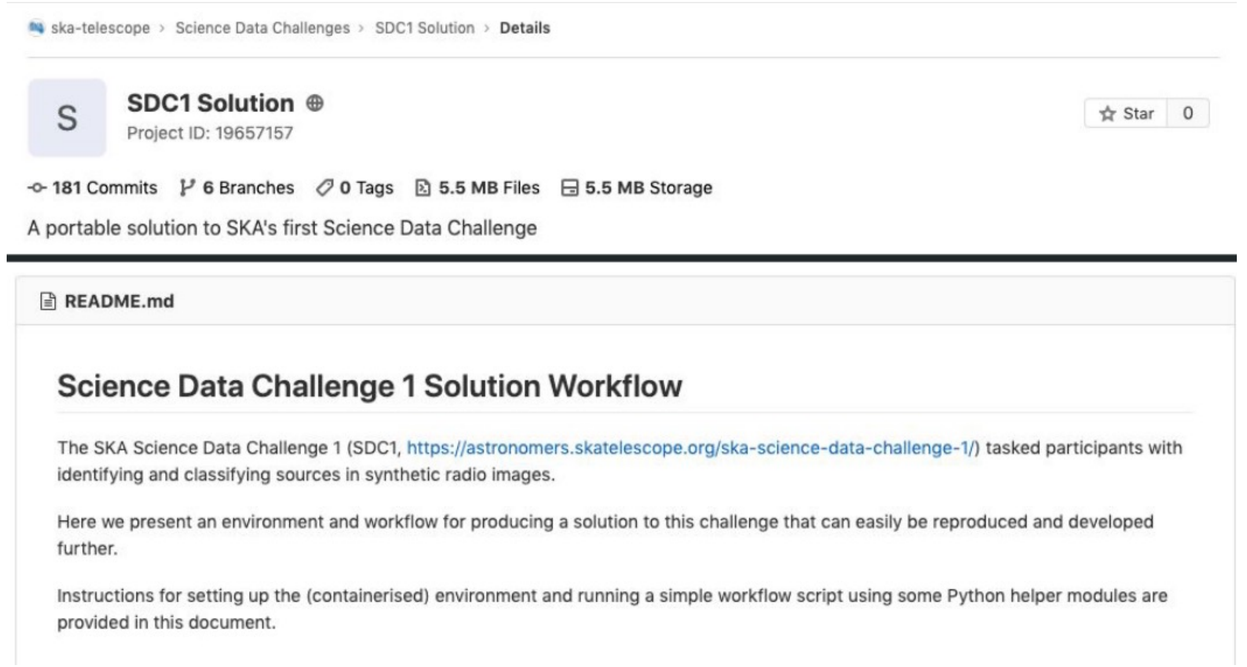
- Use an open licence
- Have findable code
- Use code standards
- Use built-in tests



SDC1 solution

- **Fully reproducible** solution by Alex Clarke (SKAO)
- Utilises **containerisation** in order to package up all software and environment
- Demonstrates **best practices** in research and software development

https://gitlab.com/ska-telescope/sdc/sdc1-solution



SDC2 results

Neutral hydrogen (HI)

- **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. Kleiber[41], S. Kitaeff[28], D. Kleiner[29], B. Lao[50], X. Lu[1], A. Mazumder[15], J. Moldón[33], R. Mondal[15], 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], L. 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[23], L. Verdes-Montenegro[33], T. Westmeier[28], K. Yu[43], L. Yu[42], L. Zhang[45], X. Zhang[44], Y. Zhang[46], 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.

In preparation

In preparation

Science Data Challenge 3
Epoch of Reionisation (EoR)

- Studies of the first generations of galaxies
- Footprints in diffuse neutral hydrogen gas, tracing the evolution of cosmic structure in the $6 < z < 30$ range

EoR Science Working Group:

https://www.skao.int/en/science-users/science-working-groups-focus-groups/106/epoch-reionization

The LOW telescope

131,000 x 2 m antennas

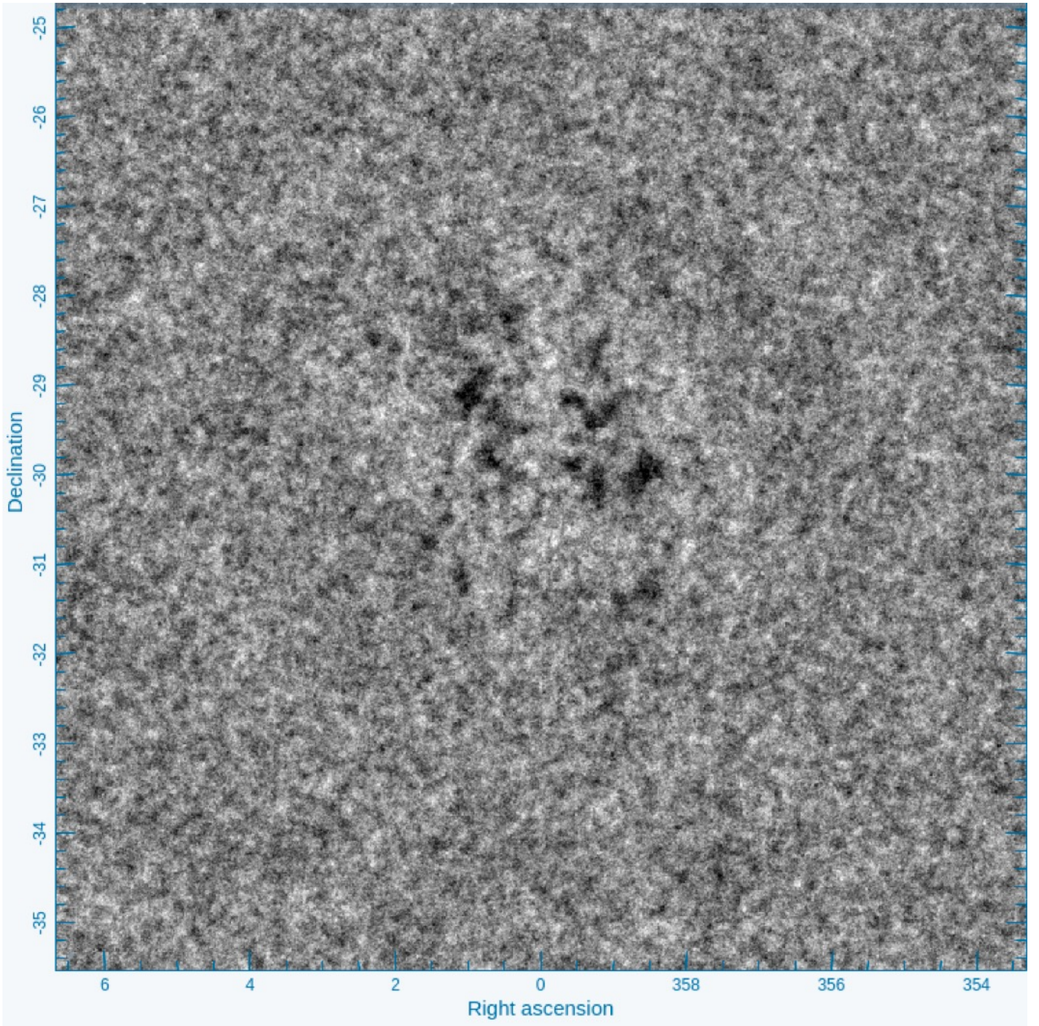
Science Data Challenge 3 (SDC3)

Epoch of Reionisation (EoR)

Developing in collaboration with SKA EoR SWG members

Two parts:

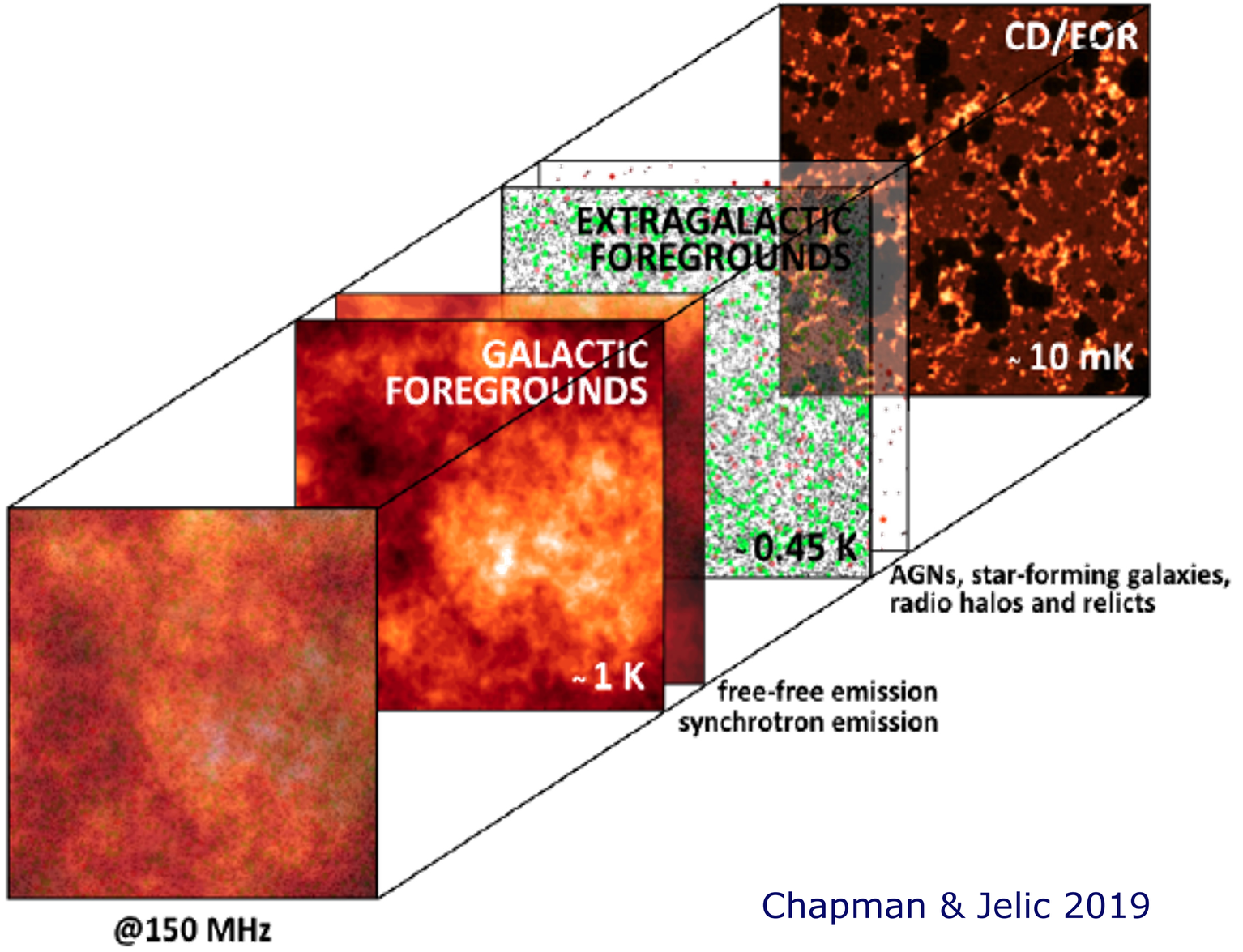
- SDC3 ”**Foregrounds**” (SDC3a; SWG Coordinators: C. Trott, V. Jelic)
  - **Foreground removal** exercise
  - SDC3a registration **will open soon**: SDC3 website
- SDC3 “**Inference**” (SDC3b; SWG Coordinators: A. Mesinger, G. Melema)
  - Extraction of **cosmological parameters**
  - SDC3b launching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



Sample EoR signal with noise added

Science Data Challenge 3a (SDC3a)
"Foregrounds"

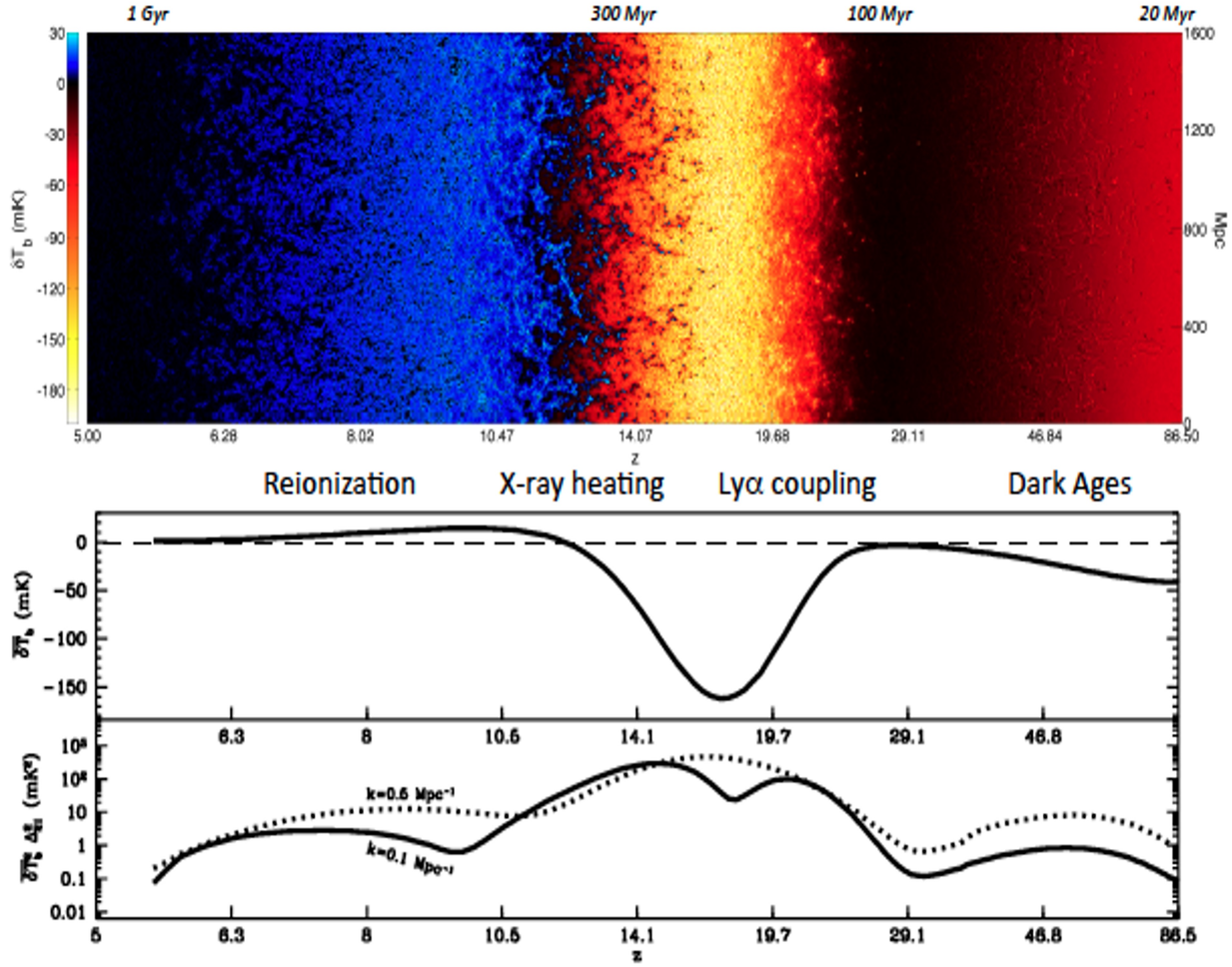
- Target participants: SWGs including EoR, Cosmology, Continuum
- Input data: calibrated visibilities and high fidelity image cube
- Challenge will be based on:
  - a) Ability to remove the point source + diffuse foregrounds from the data-set
  - b) Ability to extract the spherical and cylindrical power spectrum



Chapman & Jelic 2019

Science Data Challenge 3a (SDC3b)
"Inference"

- Target Participants: SWGs including EoR
- Input data: EoR power spectrum + noise and residual foreground contamination
- Challenge will be based on:
  - a) ability to extract the input EoR history (ionisation fraction)



Mesinger+ (2016)

Science Data Challenge 3 (SDC3)

Website [SDC3 website]

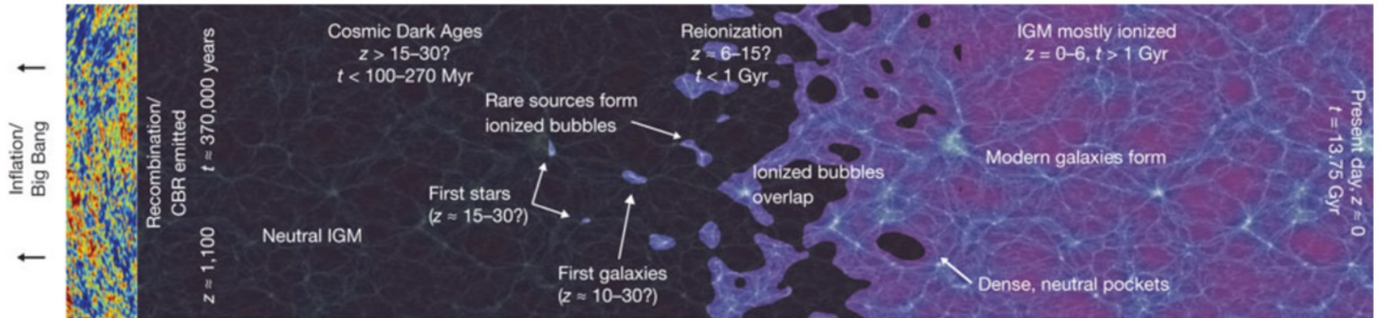
SKA SDC3 | Overview | Challenges | Computational resources | Challenge registration | Discussion forum | FAQs

Epoch of Reionisation

Given our current understanding, cosmological history prior to the current state of the universe can be divided in to several distinct epochs:

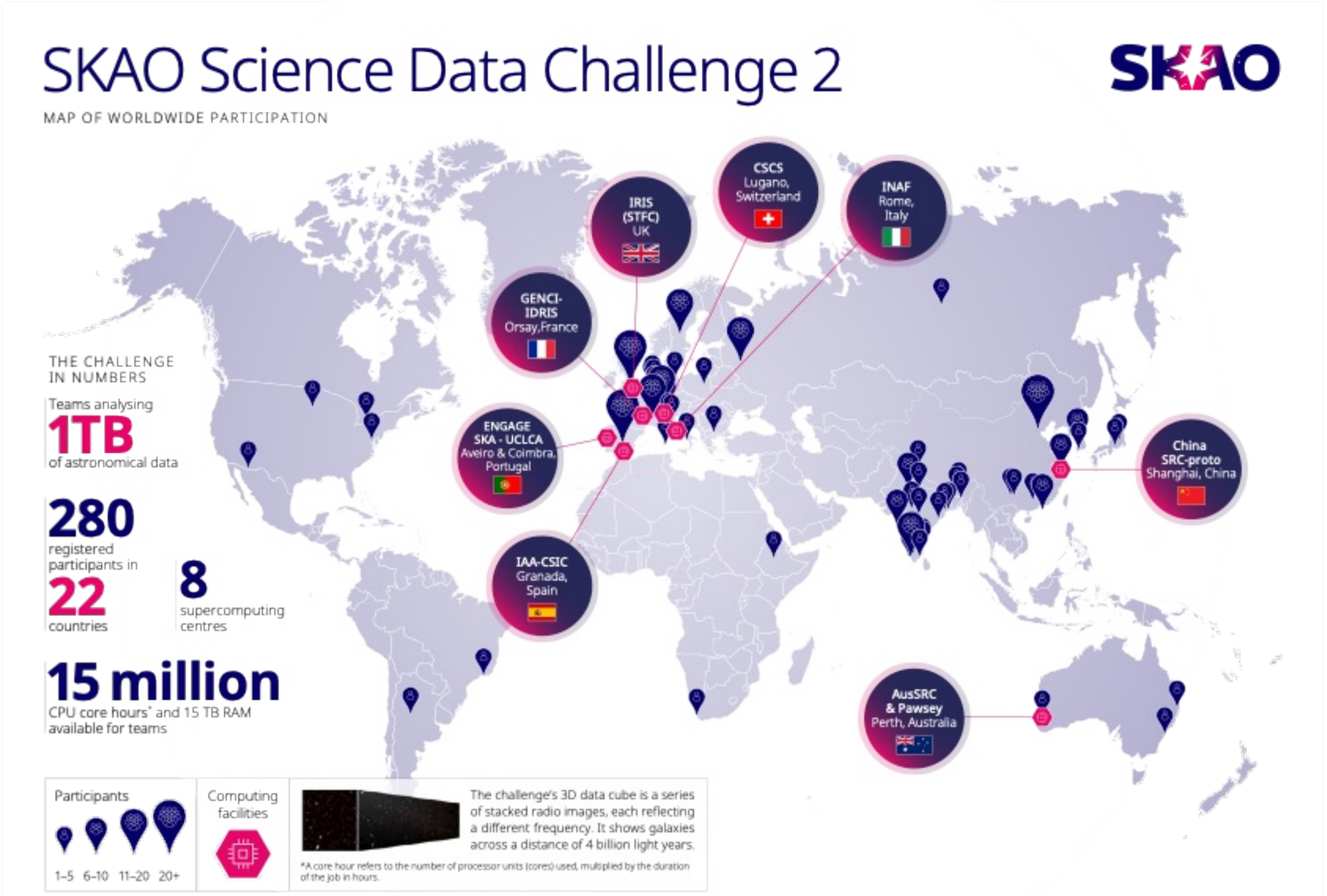
1. Cosmic Dark Ages (1100 < z < 30)
2. Cosmic Dawn (30 < z < 15)
3. Epoch of Reionisation (15 < z < 6)

After the Big Bang, our universe expanded and cooled to the point at which ionised hydrogen recombined to its atomic state, with 21cm emission subsequently being observed in absorption against the background. With the appearance of a significant population of galaxies, the first stars formed and began to heat their surroundings. This heating shifted the absorption profile of the 21 cm Hydrogen line into emission. However, as this heating continued, more of the surrounding gas became ionised to the point at which this emission ceased. Having only roughly constrained the time periods at which these important events occurred, a primary science goal of the SKA is to constrain that timeline, which is where Science Data Challenge 3 comes in.



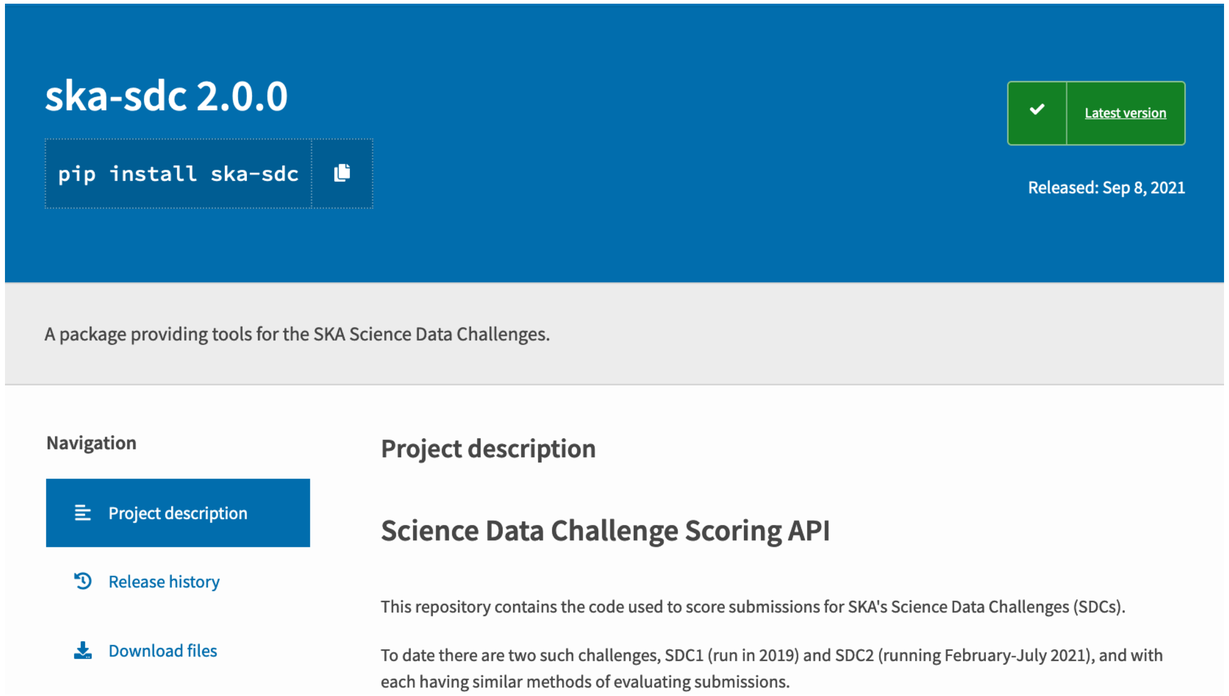
SDC computational facility partners

- **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



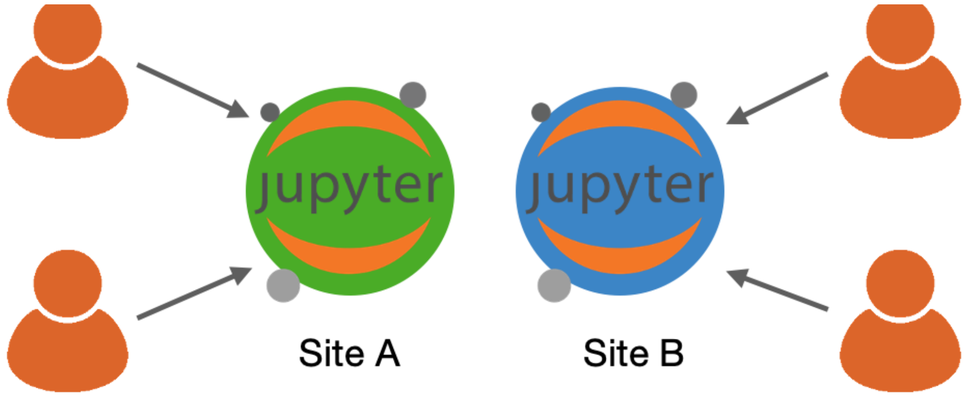
Prototyping for the SKA Regional Centres (SRCs)

- SDC2 **scoring service** made use of **web technologies** for remote scoring by participants
- SDC3 **data transfer** as a test case for **Rucio** solution: “Store, manage, and process data in a heterogeneous distributed environment”



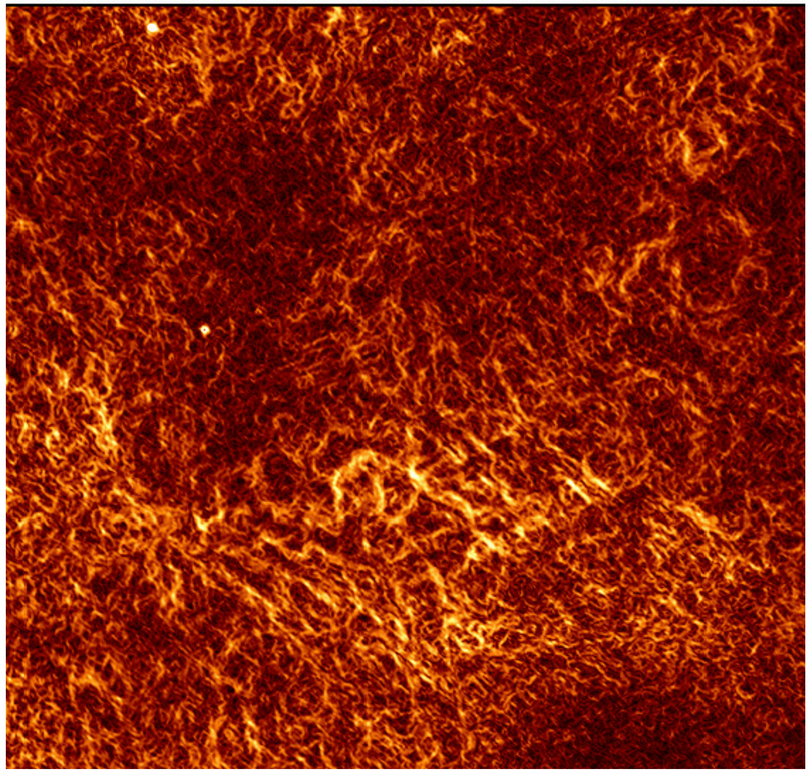
Prototyping for the SKA Regional Centres (SRCs)

- Potential to support JupyterHub environment (to work with **containers**)
  - Further enable reproducibility
  - Would support science community to be able to deploy pipelines
- Deploy **image viewers**, e.g. CARTA



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

Summary

- **Science Data Challenges** goal to prepare the community for the **size and complexity** of SKAO data
- Two data challenges completed: **collaboration and method complementarity** a highlight of both
- **Generous support** from international computing facilities
  - Enables accessible provision of **realistically large dataset**
    - Test aspects of the future **SKA Regional Centre** model
- **EoR Challenge** launching soon
- **Thank you!**