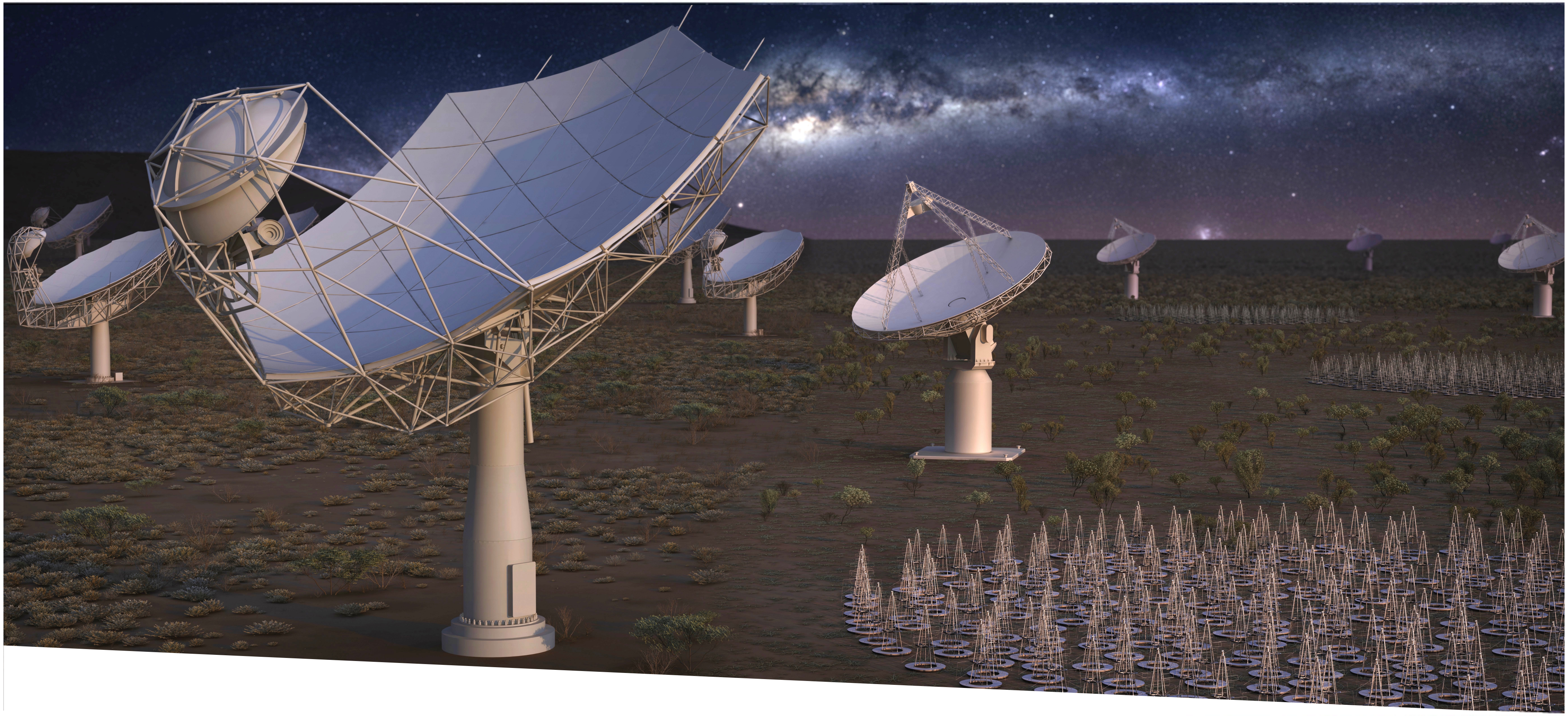
SKA data challenges for SRCs



SQUARE KILOMETRE ARRAY

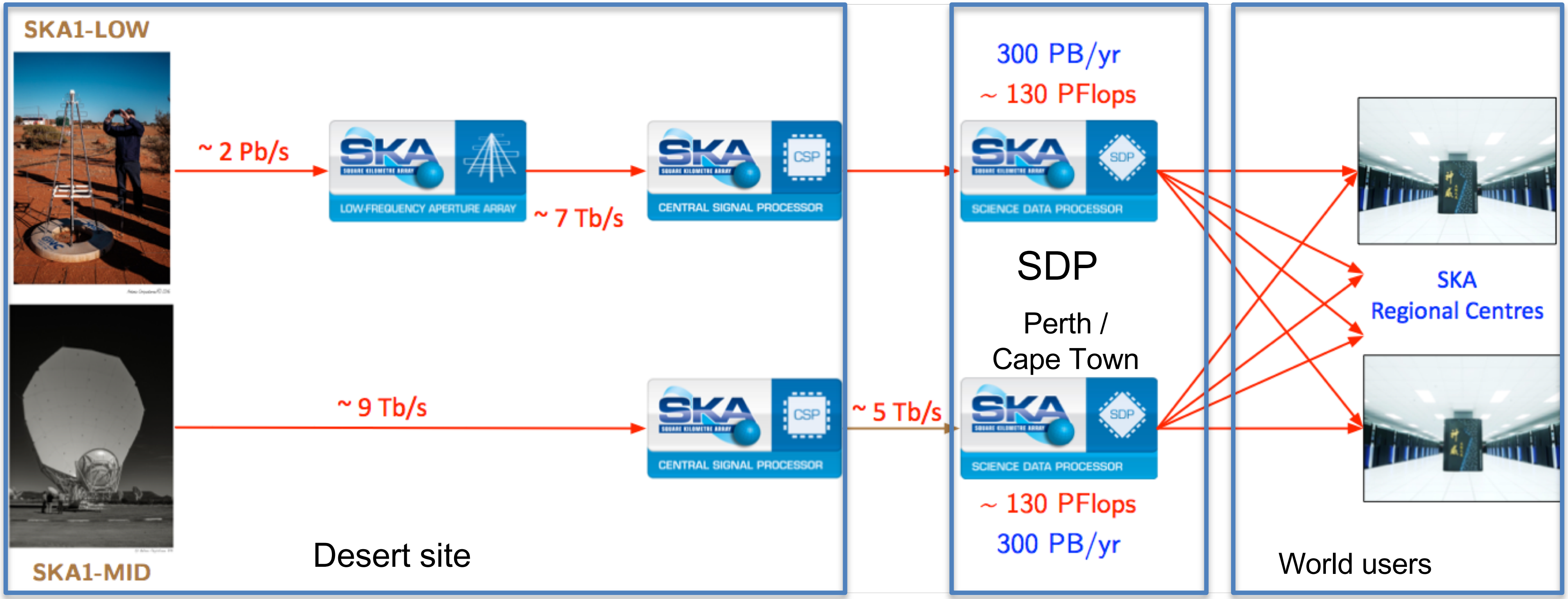
Exploring the Universe with the world's largest radio telescope

A. Bonaldi
Project scientist

The SKA's data journey

- Data flow challenges

50 x data rate reduction by Science Data Processors

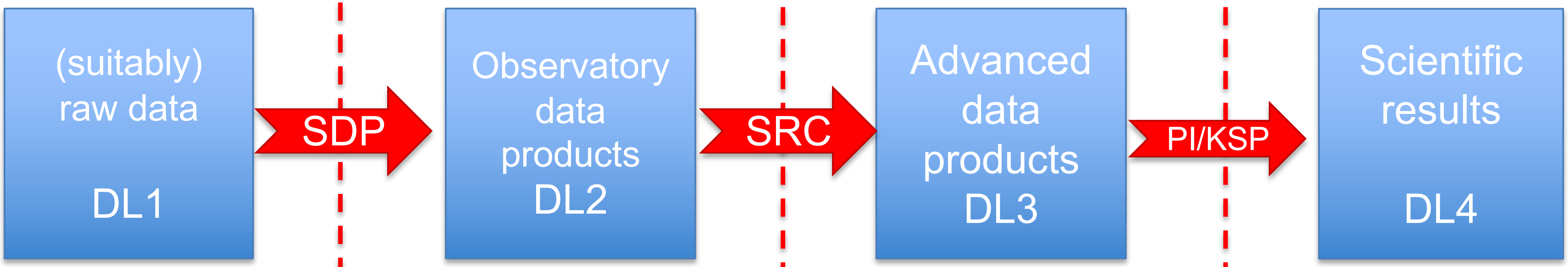

SKA1-LOW
~ 2 Pb/s
SKA LOW-FREQUENCY APERTURE ARRAY
~ 7 Tb/s
SKA CSP CENTRAL SIGNAL PROCESSOR
300 PB/yr
~ 130 PFlops
SKA SDP SCIENCE DATA PROCESSOR
SDP
Perth / Cape Town
~ 9 Tb/s
SKA CSP CENTRAL SIGNAL PROCESSOR
~ 5 Tb/s
SKA SDP SCIENCE DATA PROCESSOR
~ 130 PFlops
300 PB/yr
SKA Regional Centres
SKA1-MID
Desert site
World users

- But, have a series of buffers
- *Raw voltage data can be stored for about 2 minutes*

- *Raw visibility data can be stored for about 2 weeks*

- *Final data products will be stored forever*

SKA data challenges: Data Layers (DL)



(suitably) raw data DL1

SDP

Observatory data products DL2

SRC

Advanced data products DL3

PI/KSP

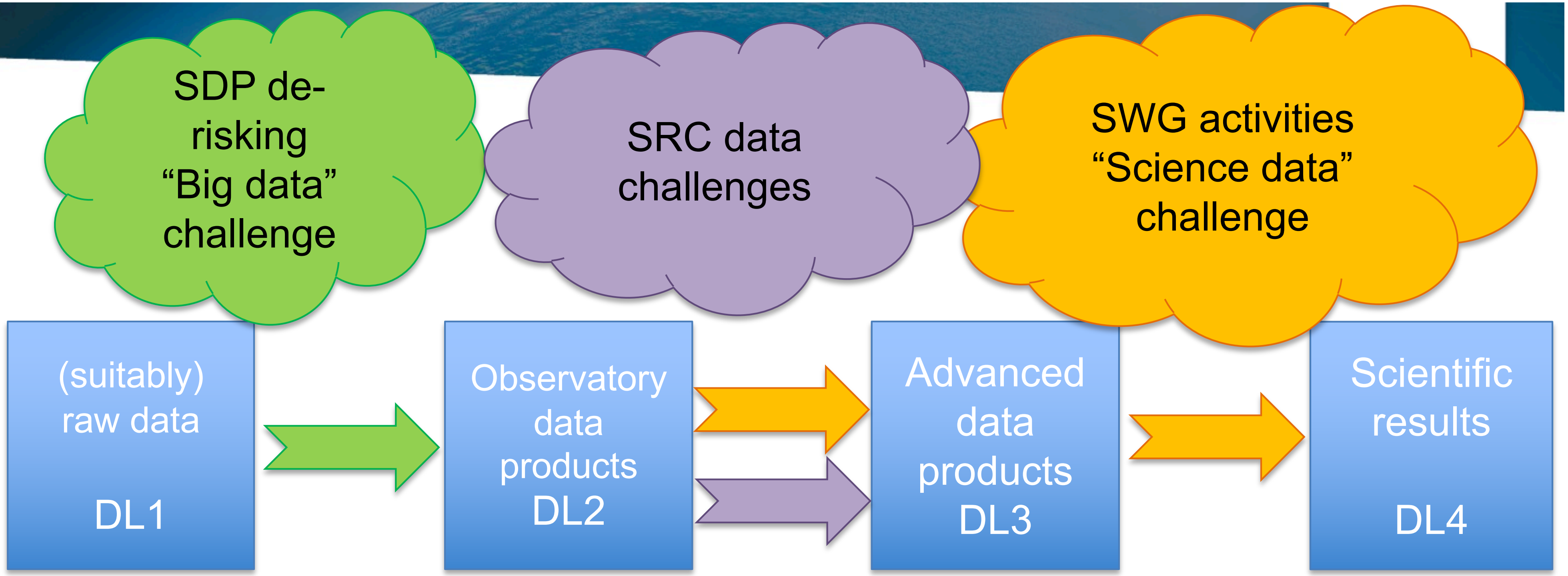
Scientific results DL4

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

- Calibrated products
- Observation-level: per observation (a few h)
- Project-level: total awarded time (100-1000 h)

- Further analysis of DL2 products
- User-generated at the SRCs
- Following best practices (reproducibility/open science)

- Additional analysis done



SDP de-risking “Big data” challenge

SRC data challenges

SWG activities “Science data” challenge

(suitably) raw data

DL1

Observatory data products

DL2

Advanced data products

DL3

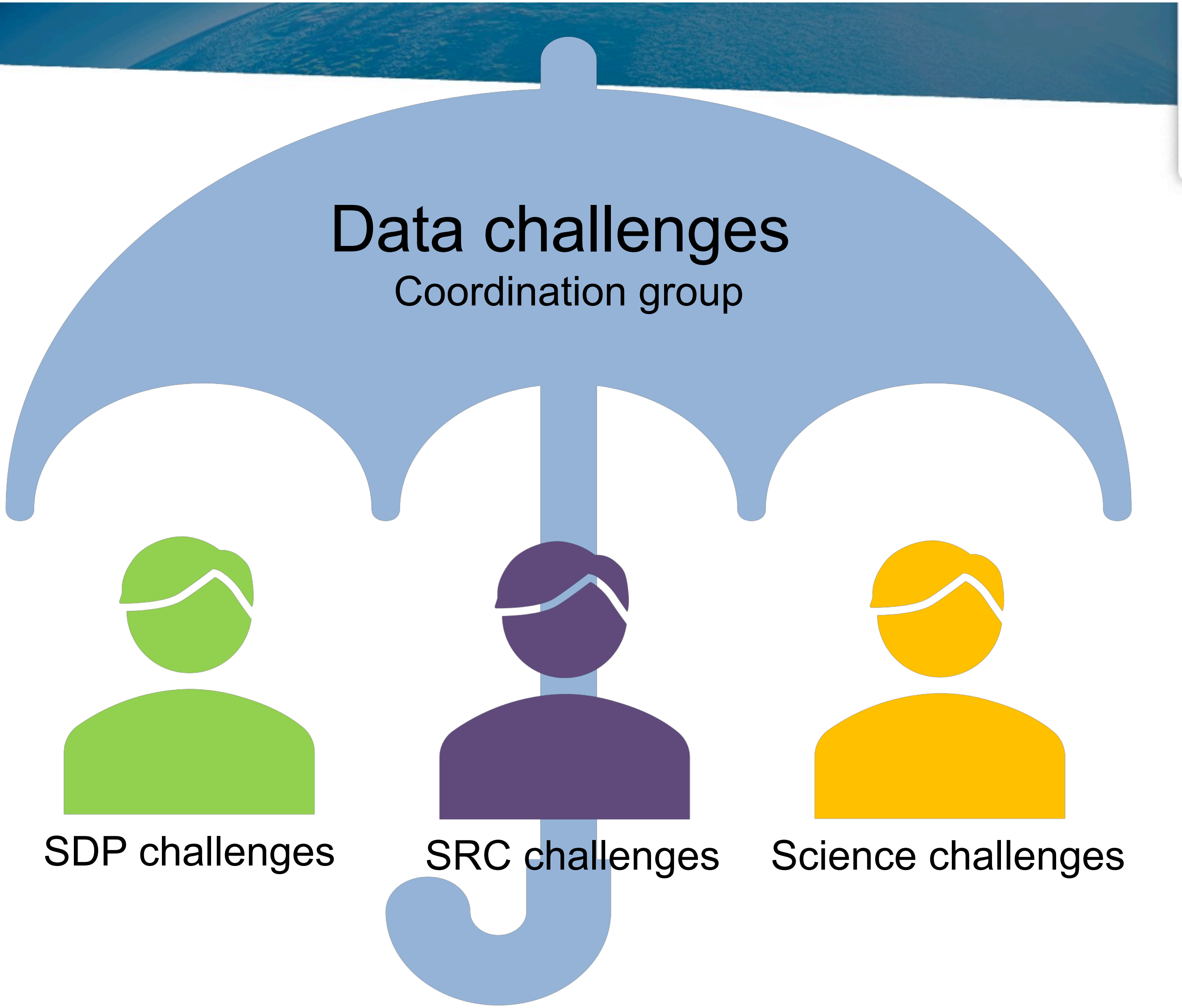
Scientific results

DL4

- Focus on efficiency/scalability
- Calibration strategy and implementation
- Data size 10-100 TBs

- Focus on algorithm development
- Proposal-Specific processing
- Data size few TBs

- Technical: data movement, format, protocols, security, databases
- Algorithms - Best practices



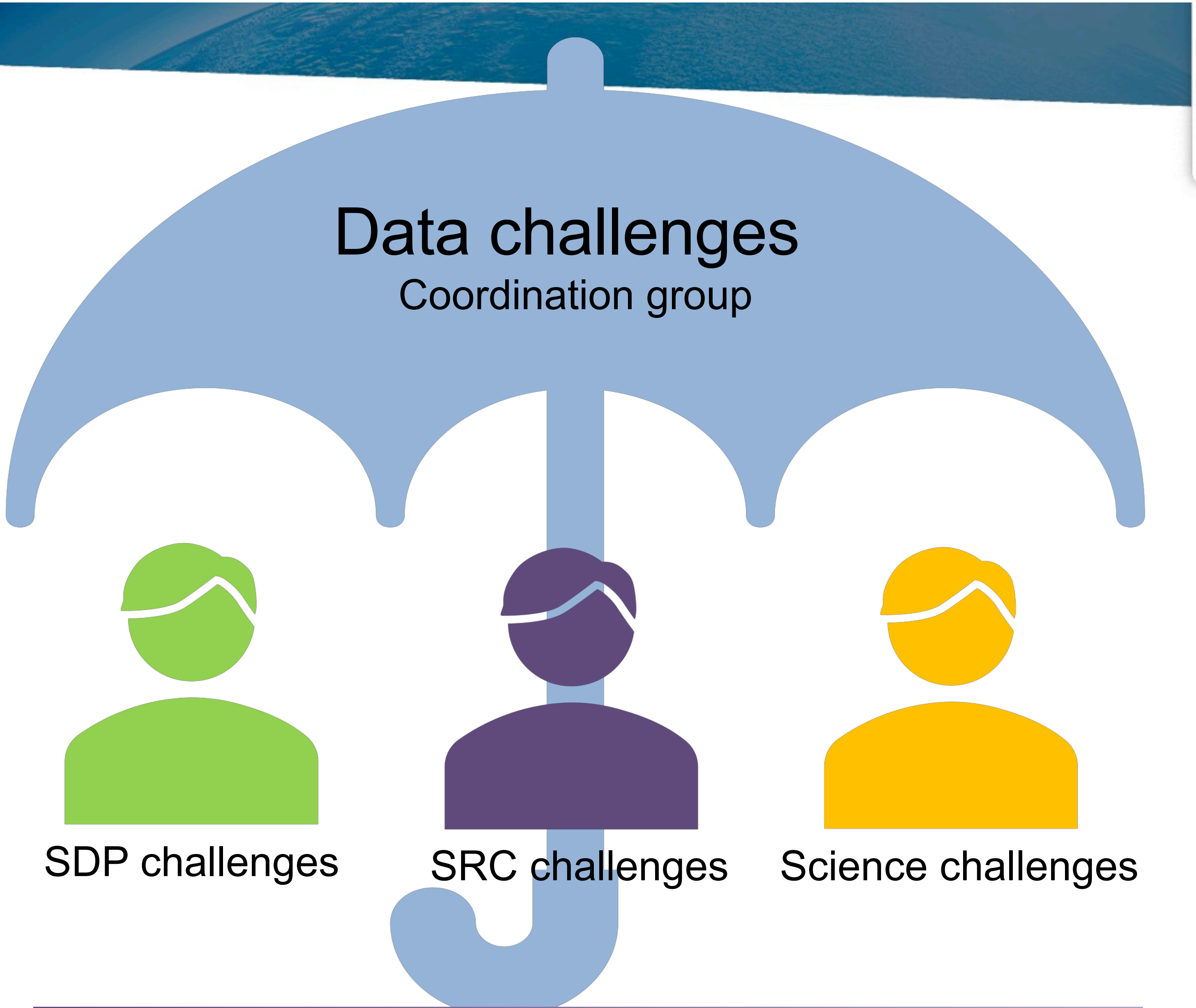
Data challenges

Coordination group

SDP challenges

SRC challenges

Science challenges



Data challenges

Coordination group

SDP challenges

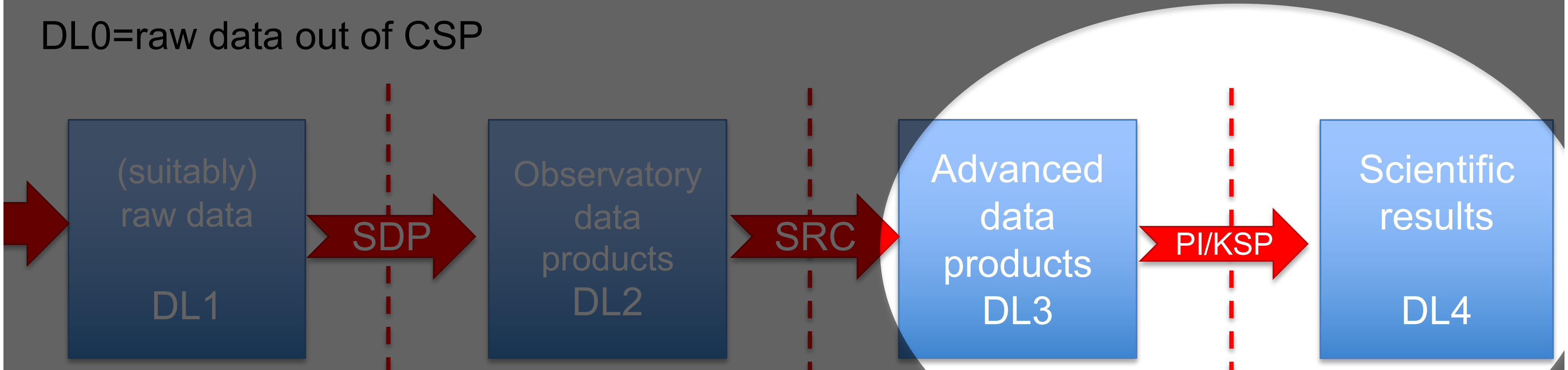
SRC challenges

Science challenges

- Chair: R. Bolton
- Priorities and strategies from the SRC community
- Coordination through SRCSC

SKA data challenges: Data Layers (DL)

DL0=raw data out of CSP



- ○ per observations (integration time hours)
- ○ Suitably averaged/compressed
- ○ Uncalibrated
- ○ unflagged
- ○ …

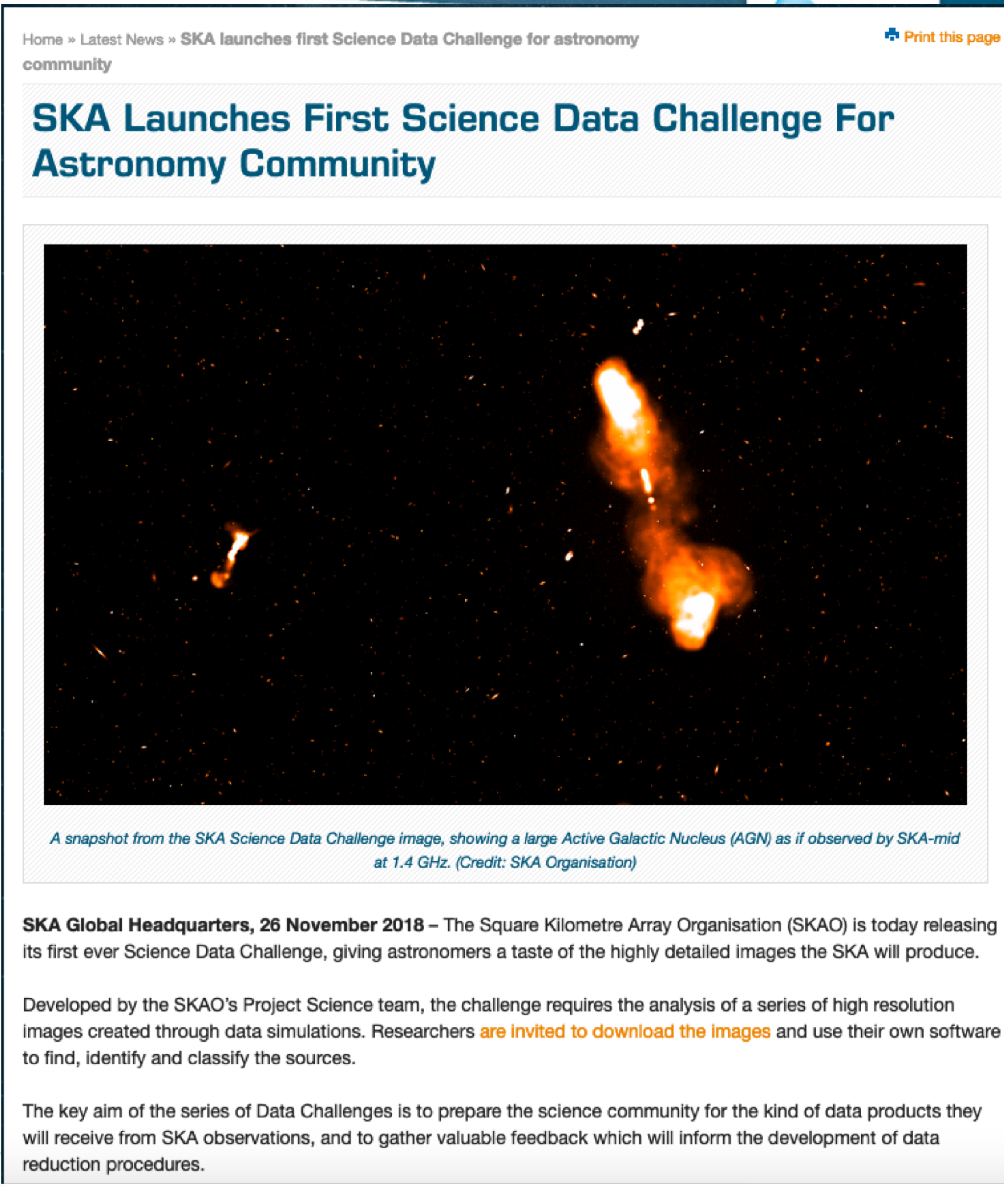
- ○ Calibrated products
- ○ Observation-level: per observation (a few h)
- ○ Project-level: total awarded time (100-1000 h)

- ○ Derived from further analysis of DL2 products
- ○ User-generated at the SRCs
- ○ Needs to follow best practice (reproduceability/ open science)

- ○ Additional analysis done

Science data challenge 1 (SDC1)

- Science-ready (SRC) imaging product
- Radio continuum, SKA Mid
- Not too challenging data sizes
- 1 pointing, 3 freqs, 3 depths
- Source finding
- Source identification, classification & characterization



Home » Latest News » **SKA launches first Science Data Challenge for astronomy community**

SKA Launches First Science Data Challenge For Astronomy Community

A snapshot from the SKA Science Data Challenge image, showing a large Active Galactic Nucleus (AGN) as if observed by SKA-mid at 1.4 GHz. (Credit: SKA Organisation)

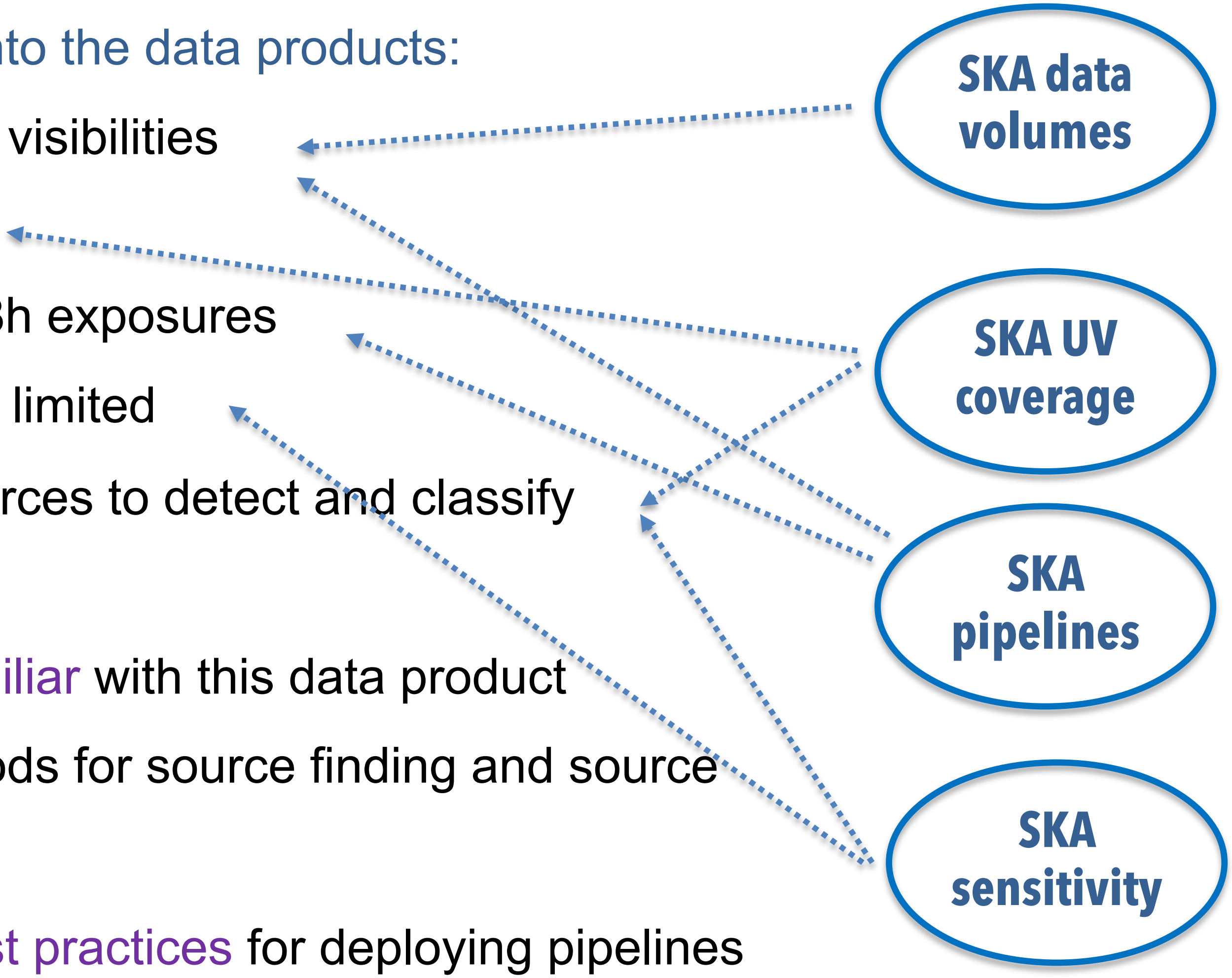
SKA Global Headquarters, 26 November 2018 – The Square Kilometre Array Organisation (SKAO) is today releasing its first ever Science Data Challenge, giving astronomers a taste of the highly detailed images the SKA will produce.

Developed by the SKAO's Project Science team, the challenge requires the analysis of a series of high resolution images created through data simulations. Researchers are invited to download the images and use their own software to find, identify and classify the sources.

The key aim of the series of Data Challenges is to prepare the science community for the kind of data products they will receive from SKA observations, and to gather valuable feedback which will inform the development of data reduction procedures.

Science data challenge 1 (SDC1)

- SKA unique features map into the data products:
  - ✓ In the image plane, not visibilities
  - ✓ "Benign" dirty beam
  - ✓ Deconvolved down to 8h exposures
  - ✓ Very deep -> confusion limited
  - ✓ Very big number of sources to detect and classify
- SDC1 goals:
  - ✓ Get the community familiar with this data product
  - ✓ Develop efficient methods for source finding and source characterization
  - ✓ Begin talking about best practices for deploying pipelines @ SRCs

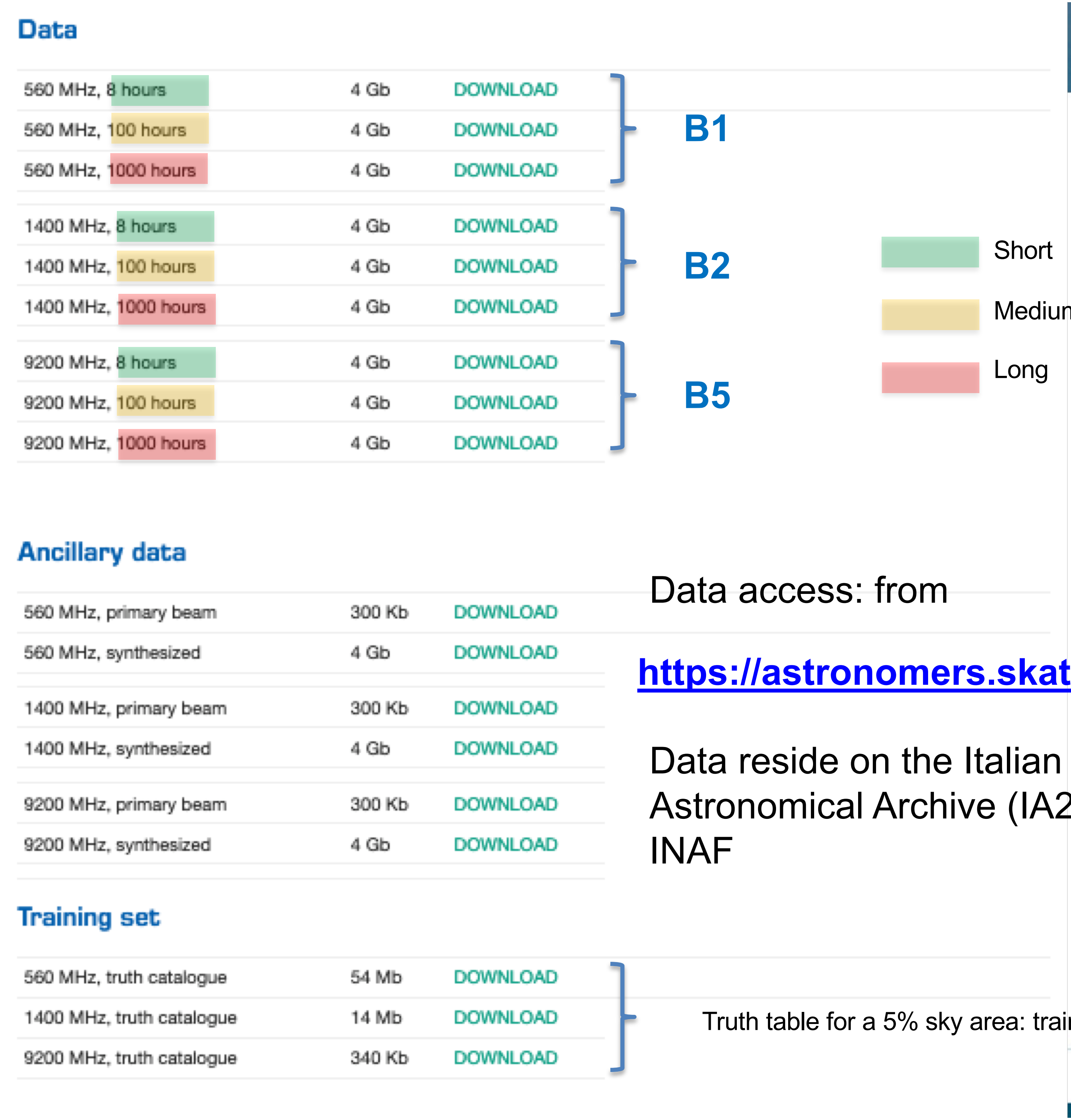


SKA data volumes

SKA UV coverage

SKA pipelines

SKA sensitivity



Data

560 MHz, 8 hours	4 Gb	DOWNLOAD
560 MHz, 100 hours	4 Gb	DOWNLOAD
560 MHz, 1000 hours	4 Gb	DOWNLOAD
1400 MHz, 8 hours	4 Gb	DOWNLOAD
1400 MHz, 100 hours	4 Gb	DOWNLOAD
1400 MHz, 1000 hours	4 Gb	DOWNLOAD
9200 MHz, 8 hours	4 Gb	DOWNLOAD
9200 MHz, 100 hours	4 Gb	DOWNLOAD
9200 MHz, 1000 hours	4 Gb	DOWNLOAD

B1

B2

B5

🟩 Short

🟨 Medium

🟥 Long

Ancillary data

560 MHz, primary beam	300 Kb	DOWNLOAD
560 MHz, synthesized	4 Gb	DOWNLOAD
1400 MHz, primary beam	300 Kb	DOWNLOAD
1400 MHz, synthesized	4 Gb	DOWNLOAD
9200 MHz, primary beam	300 Kb	DOWNLOAD
9200 MHz, synthesized	4 Gb	DOWNLOAD

Data access: from

https://astronomers.skatelescope.org/

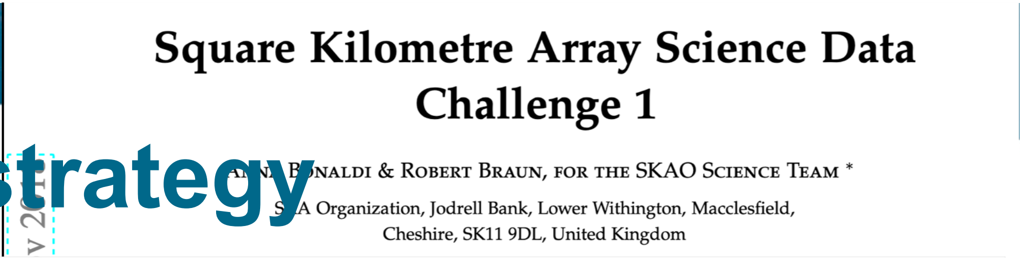
Data reside on the Italian Center for Astronomical Archive (IA2) operated by INAF

Training set

560 MHz, truth catalogue	54 Mb	DOWNLOAD
1400 MHz, truth catalogue	14 Mb	DOWNLOAD
9200 MHz, truth catalogue	340 Kb	DOWNLOAD

Truth table for a 5% sky area: training set

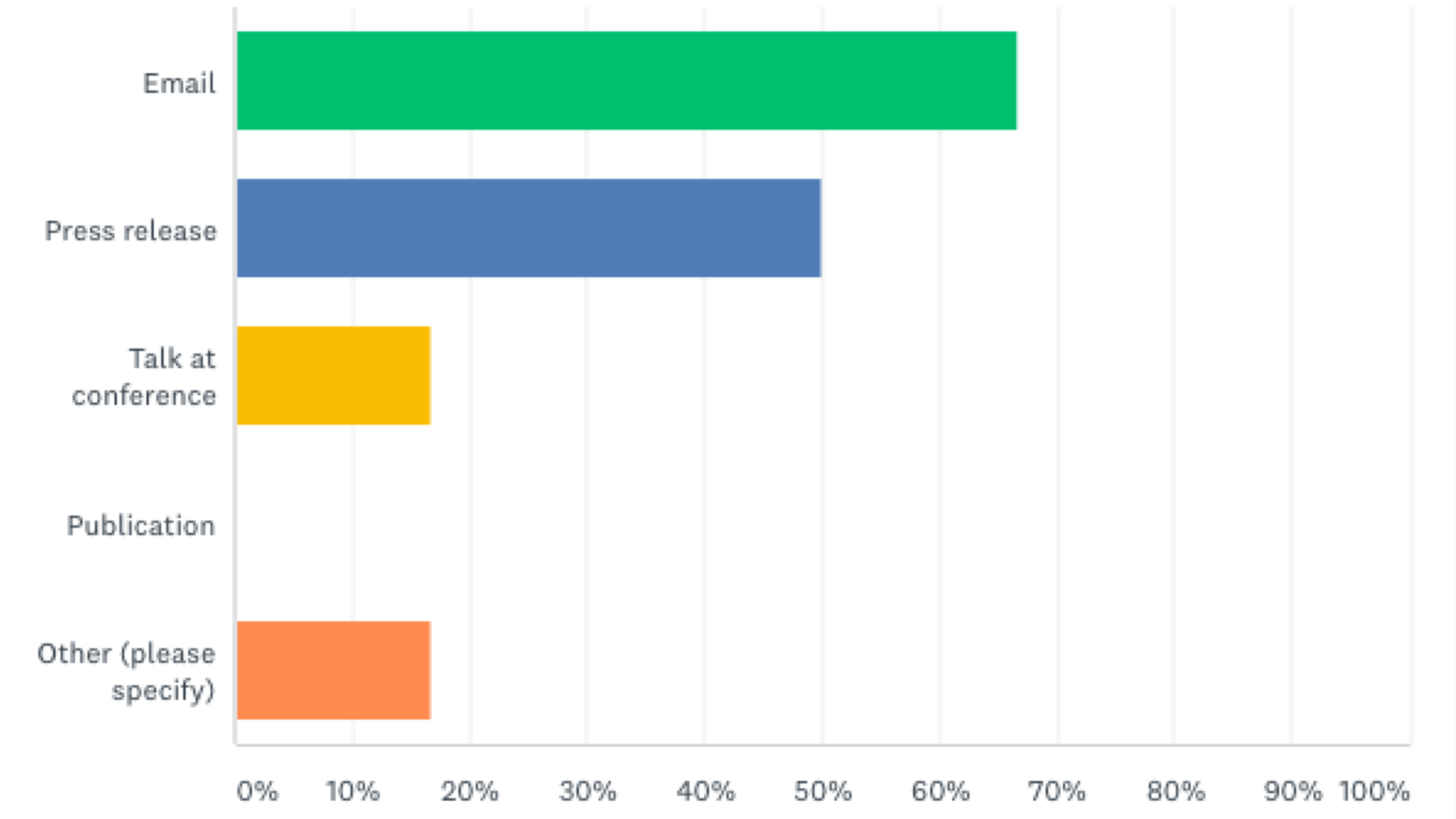
SDC1 communication strategy



Square Kilometre Array Science Data Challenge 1

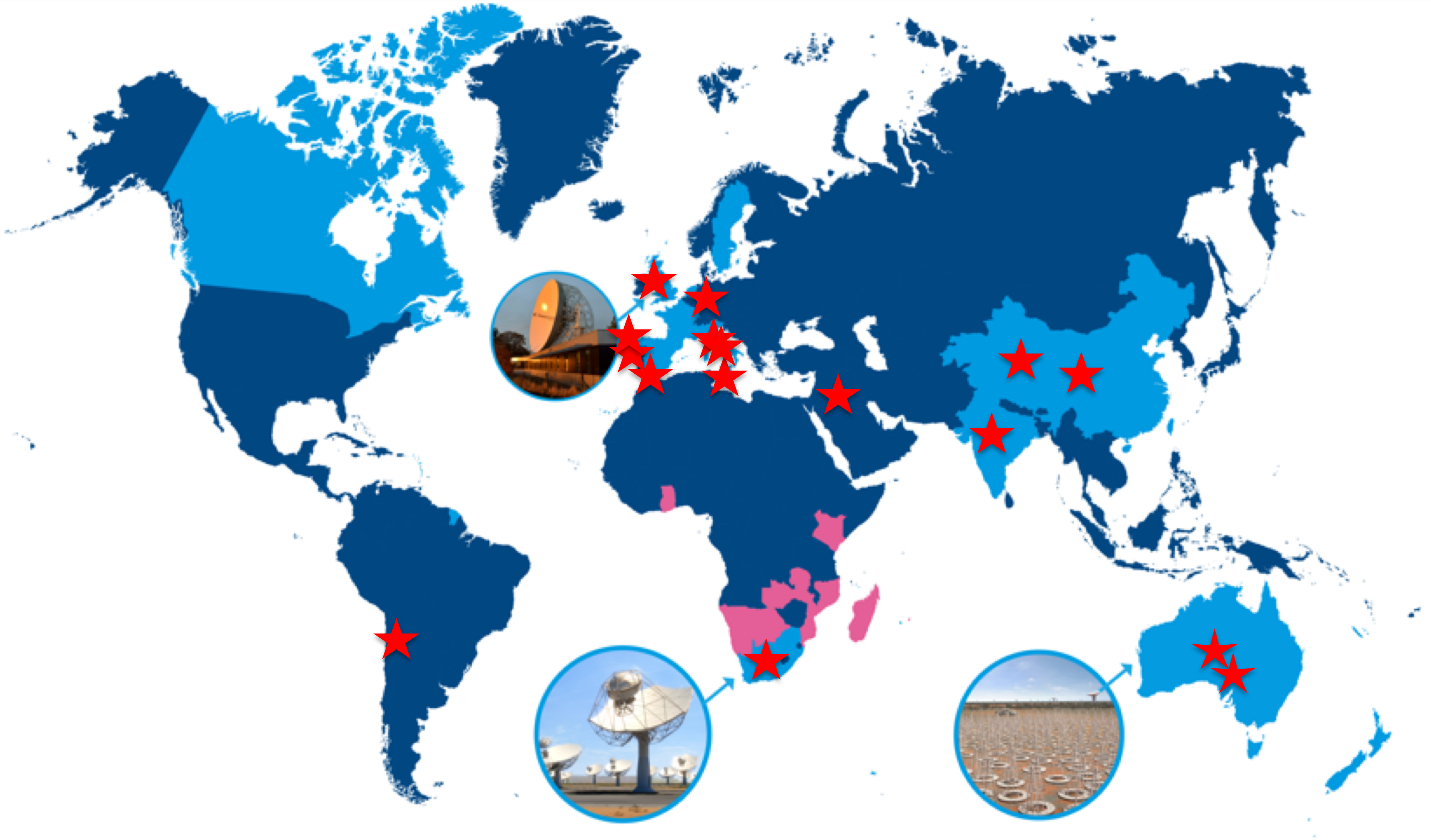
Anna Bonaldi & Robert Braun, for the SKAO Science Team *

SKA Organization, Jodrell Bank, Lower Withington, Macclesfield,
Cheshire, SK11 9DL, United Kingdom



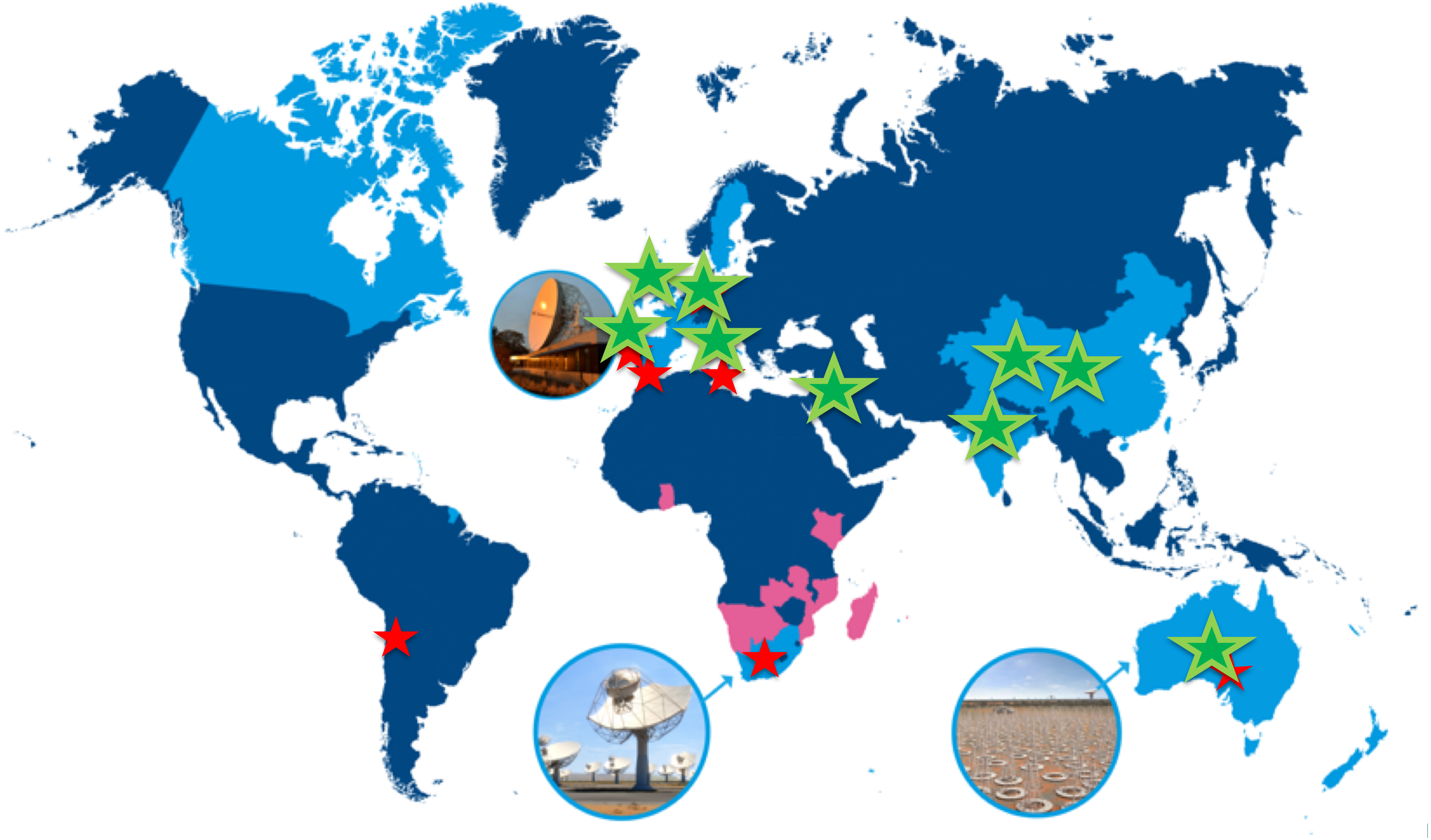
The SDC1 teams!

17 teams registered to SDC1

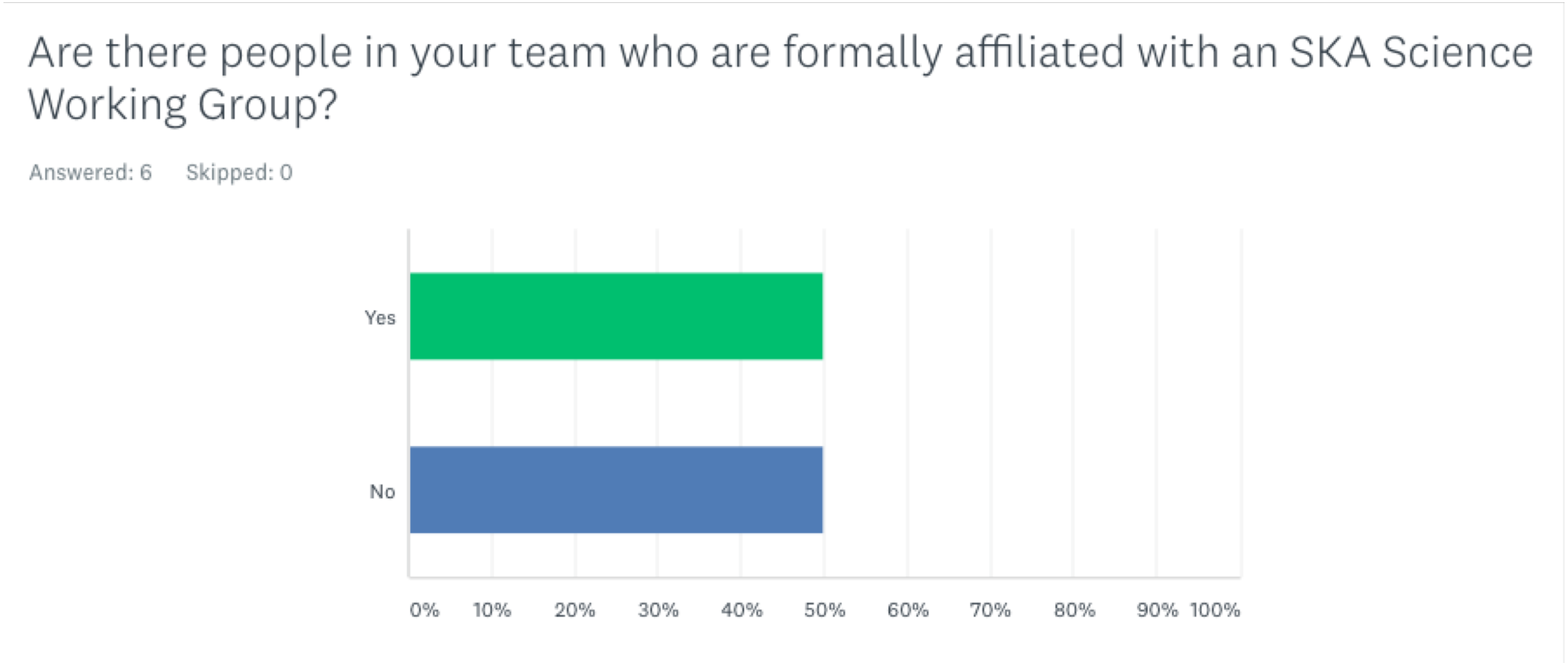


The SDC1 teams!

9 teams submitted results by the deadline of 30th April



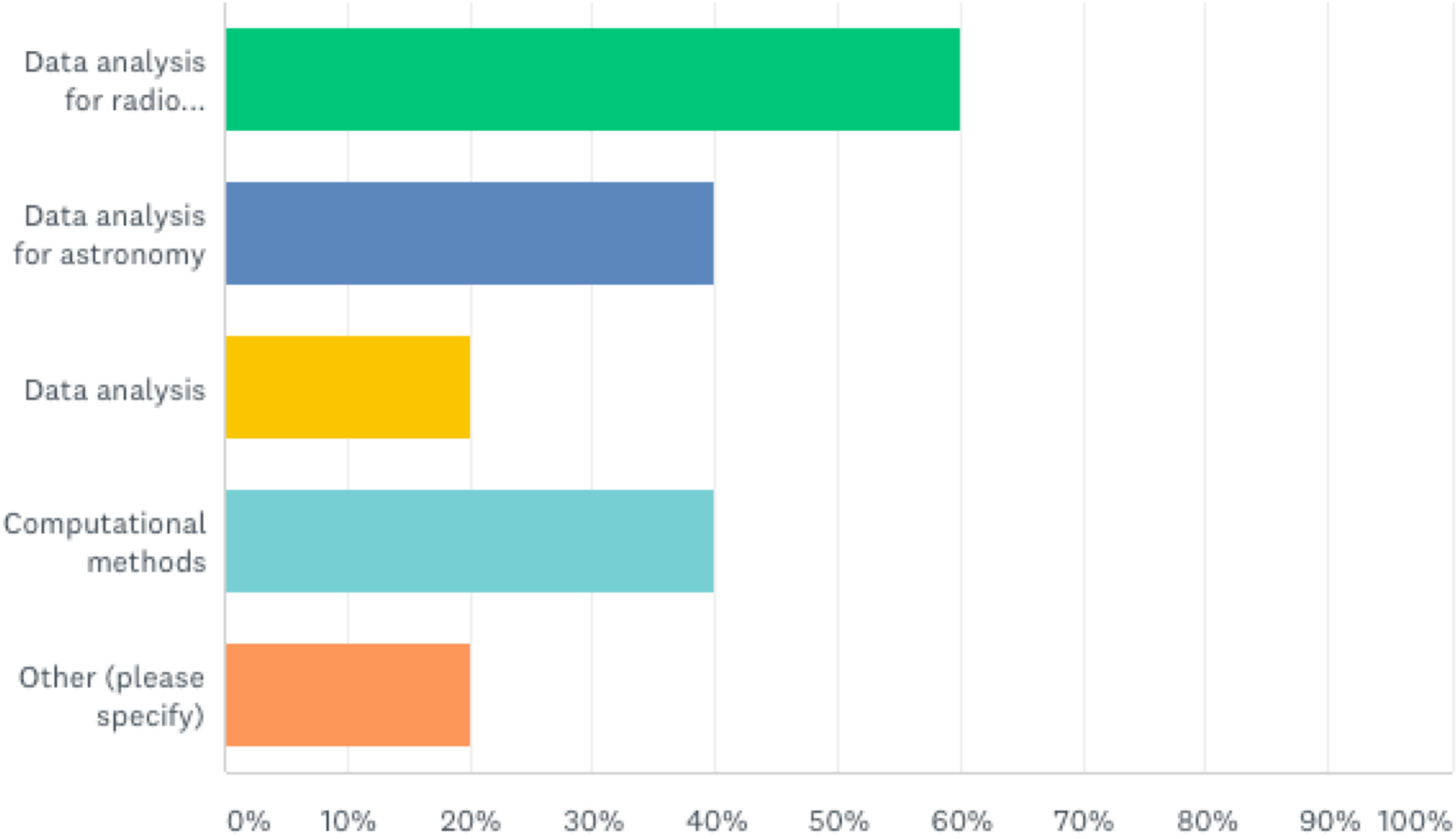
Team's provenance



Are there people in your team who are formally affiliated with an SKA Science Working Group?

Answered: 6   Skipped: 0

Team’s expertise



SDC1 Leaderboard

Ox-ICRAR
UK / Australia

Catherine Hale, Aaron Robotham, Matt Jarvis

The team acknowledges STFC and Oxford Hintze Centre for Astrophysical Surveys (funded through the Hintze Family Charitable Foundation), ICRAR (for ASGR)

ICRAR
Australia

Chen Wu, Ivy Wong

Richard Dodson, Elle Gholami

The team acknowledges CSIRO/Bracewell, ICRAR/Pleiades

EngageSKA
Portugal

Bruno Coelho

Domingos Barbosa, Sonia Antón, Jorge Bruno Morgado, Valério A. R. M. Ribeiro, Dzianis Bartashevych, João Paulo Barraca, Miguel Bergano, Dalmiro Maia

B.C. acknowledges support from the Advanced EU Network of E-infrastructures for Astronomy with SKA (AENEAS) [illegible] by COMPETE 2020 and FCT, Portugal.

ARCIt-CACAO
Italy

Sandra Burkutean, Marcella Massardi, Jan Brand, Elisabetta Liuzzo, Matteo Bonato

Eugenio Schisano, Andrea Giannetti, Kazi Rygl, Nicola Marchili

hs
Germany

Vesna Lukic, Marcus Brüggen

the team acknowledges the use of the LOC2 cluster at University of Hamburg

IITK
India

Pankaj Jain, Pritpal Kaur, Mohit Panwar Prabhakar

IPM
Iran

Zahra Bagheri, Hadis Goodarzi, Elham Saremi, Somayeh Sheikhnezami, Vajiheh Sabz Ali

Arash Danesh, Habib G. Khosroshahi, Armin Rasouli, Fatemeh Tabatabaei, Alireza Vafaei Sadr, Mohammad Hossein Zhoolideh Haghigh

The team acknowledges funding by Institute for Research in Fundamental Sciences (IPM)

JLRAT
China

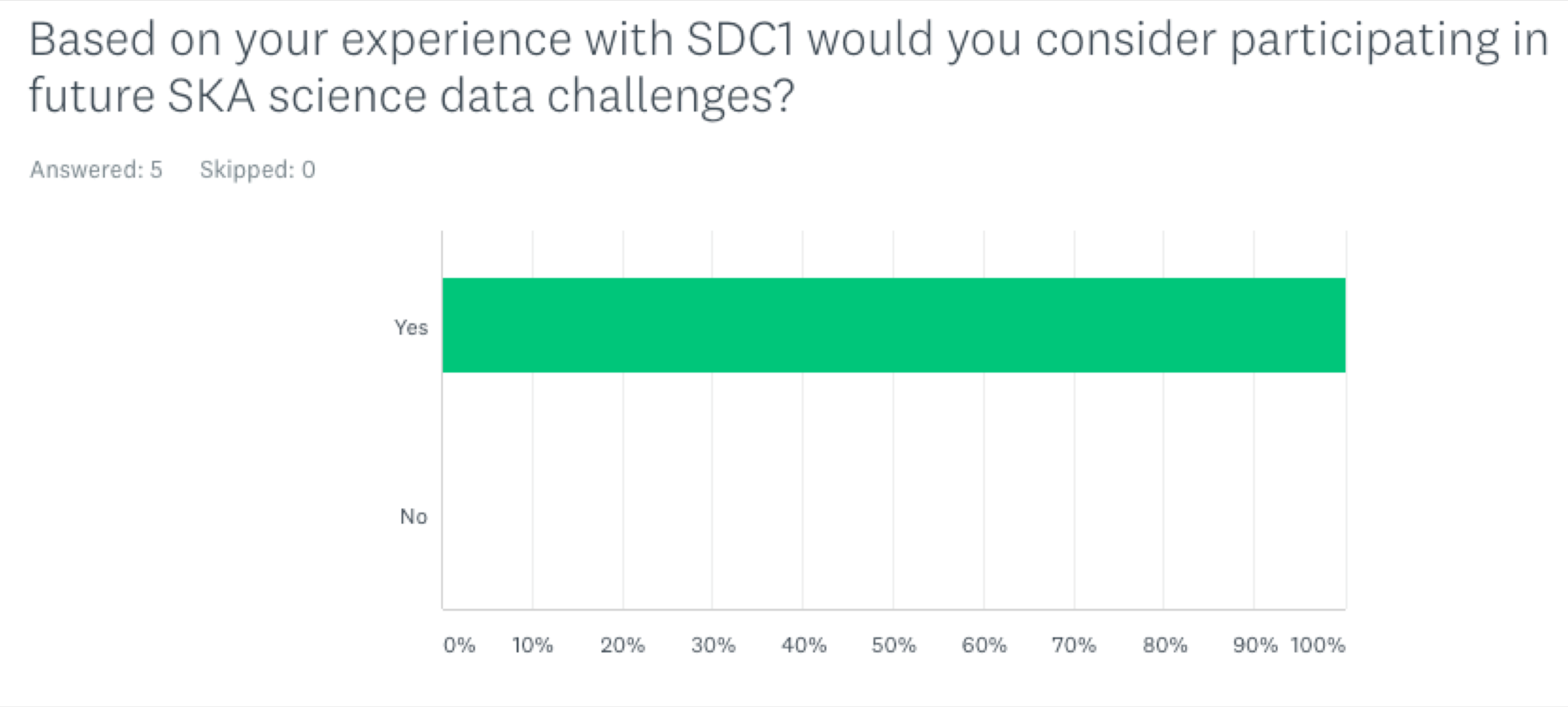
Lei Yu, Bin Liu

Shanghai
China

Tao An, Sumit Jaiswal, Yang Lu, Prashanth Mohan, Baoqiang Lao

National Key R&D Programme of China under grant number 2018YFA0404603

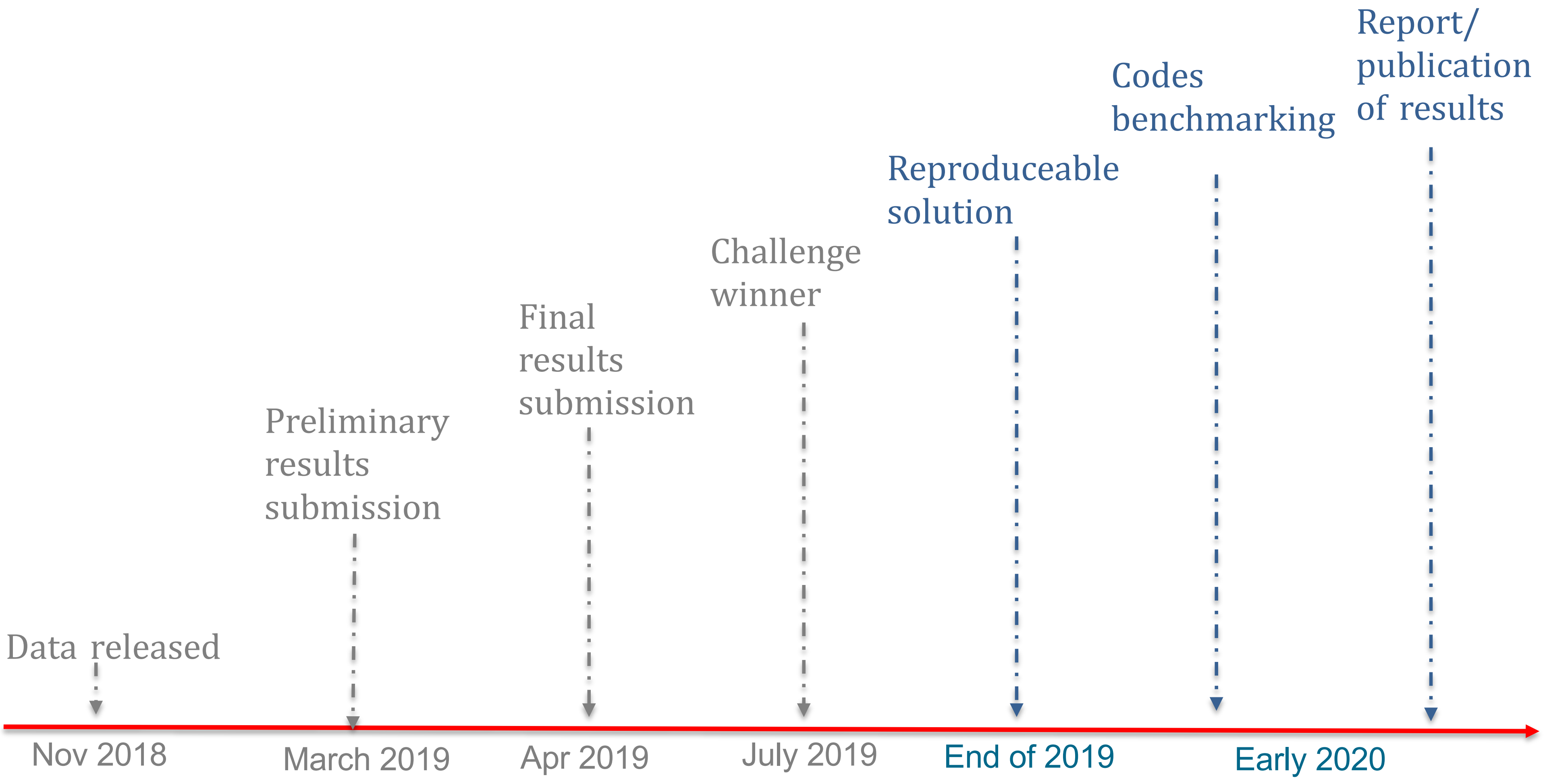
Positive feedback!



Based on your experience with SDC1 would you consider participating in future SKA science data challenges?

Answered: 5 Skipped: 0

SDC1 timeline and progress



Data released

Nov 2018

Preliminary results submission

March 2019

Final results submission

Apr 2019

Challenge winner

July 2019

Reproduceable solution

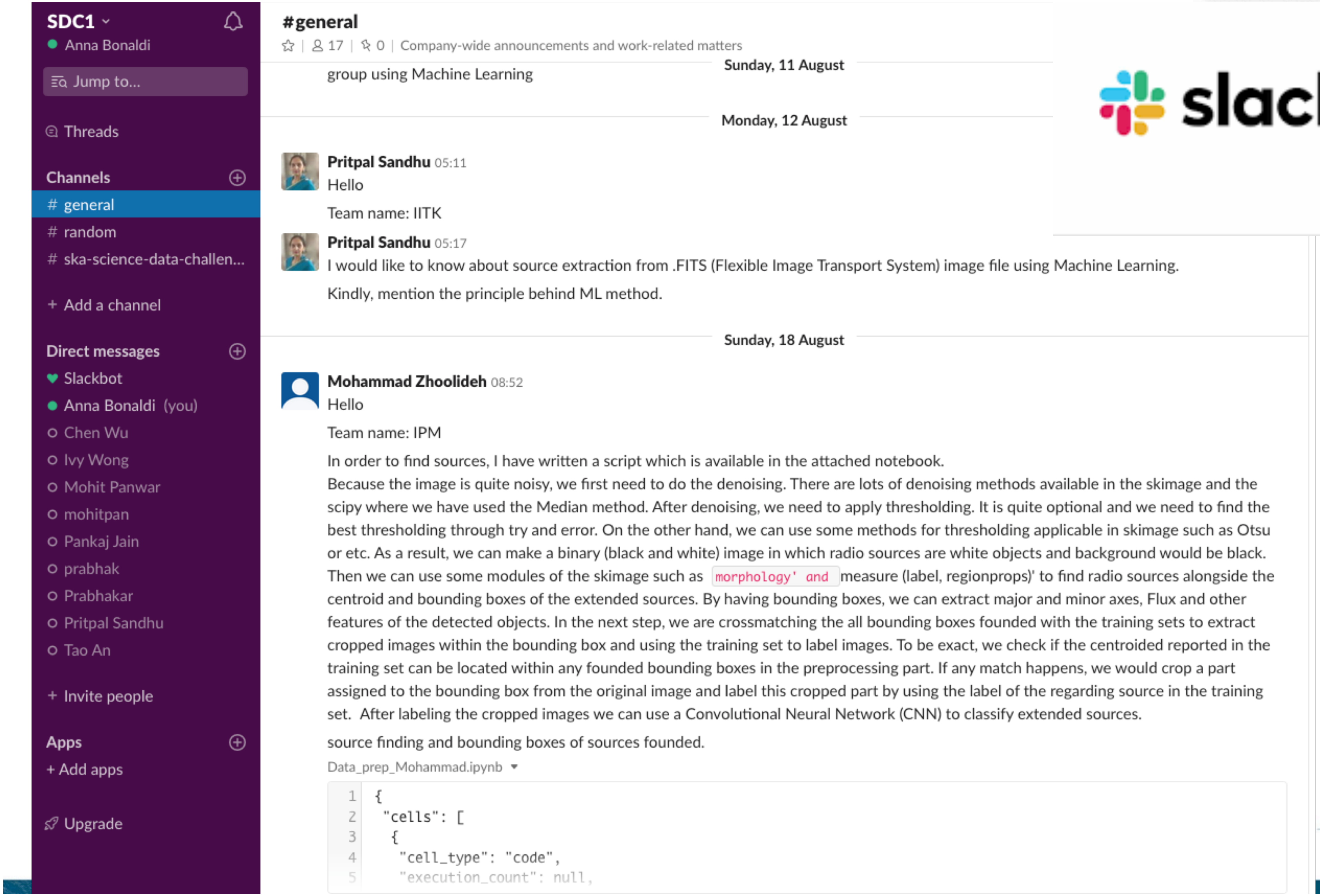
End of 2019

Codes benchmarking

Report/ publication of results

Early 2020

Collaborative phase



#general

☆ | 17 | 0 | Company-wide announcements and work-related matters

Sunday, 11 August

group using Machine Learning

Monday, 12 August

Pritpal Sandhu 05:11

Hello

Team name: IITK

Pritpal Sandhu 05:17

I would like to know about source extraction from .FITS (Flexible Image Transport System) image file using Machine Learning.

Kindly, mention the principle behind ML method.

Sunday, 18 August

Mohammad Zhoolideh 08:52

Hello

Team name: IPM

In order to find sources, I have written a script which is available in the attached notebook.
Because the image is quite noisy, we first need to do the denoising. There are lots of denoising methods available in the skimage and the scipy where we have used the Median method. After denoising, we need to apply thresholding. It is quite optional and we need to find the best thresholding through try and error. On the other hand, we can use some methods for thresholding applicable in skimage such as Otsu or etc. As a result, we can make a binary (black and white) image in which radio sources are white objects and background would be black. Then we can use some modules of the skimage such as `morphology' and` measure (label, regionprops)' to find radio sources alongside the centroid and bounding boxes of the extended sources. By having bounding boxes, we can extract major and minor axes, Flux and other features of the detected objects. In the next step, we are crossmatching the all bounding boxes founded with the training sets to extract cropped images within the bounding box and using the training set to label images. To be exact, we check if the centroided reported in the training set can be located within any founded bounding boxes in the preprocessing part. If any match happens, we would crop a part assigned to the bounding box from the original image and label this cropped part by using the label of the regarding source in the training set. After labeling the cropped images we can use a Convolutional Neural Network (CNN) to classify extended sources.

source finding and bounding boxes of sources founded.

Data_prep_Mohammad.ipynb ▾

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": null,
```

Best practices for SDC1

- Open
  - Is your pipeline publicly available (e.g. Github?)
  - How is it licensed?
  - Which licenses/dependencies does it need?
  - Is there a documentation?
- Reproduceable
  - Can you containerize your pipeline?
    - Info on containers circulated
    - Support offered to move towards reproducibility

Thanks!

SDP list of observation-level products

From SDP CDR documentation

- Science Alert Catalogue
- Transient Source Catalogue
- Science Product Catalogue Data Product
- Image Products 1: Image Cubes
- Image Products 2: Uvgrids
- Calibrated Visibilities
- Sieved Pulsar and Transient Candidates
- Pulsar Timing Solutions
- Dynamic spectrum data
- Transient Buffer Data
- Science Data Model

Observation-level products will be combined into project-level products.

Added-value products will be derived from the Observatory data products