SKA data challenge 1 solution

Alex Clarke, James Collinson, Rohini Joshi, Rob Barnsley, Rosie Bolton

SKA Observatory

E-OSSR Onboarding Presentation

Date

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Introduction/Instructions

Aim:
tech reports (~20 min talk and 2-3 page summary documents) on community software for OSSR

- Content:
  - science case and ”user story” (two sides: data analyst side and OSSR side)
  - added value of OSSR
  - update on questions from OSSR's first questionnaire and software registration survey
    - Both replies will be provided before the talk by FG1 lead
  - Discussion on on-boarding: open points, requirements…

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Introduction – 2-3 slides

- Science case: Source finding from images, followed by source classification from catalogued information.

  This is an example solution to the SKA data challenge 1, which we hope will inspire astronomers to provide their code and software pipeline in future SKA data challenges and in their own research (since there is a lack of software pipelines provided with published results).

  As a secondary goal, of course we would be pleased if all or parts of our software are used by others in their research, though we expect it to be adapted in such scenarios rather than used out of the box.

- Software or Service Name: SKA data challenge 1 solution
  - Purpose: A packaged solution to the first SKA data challenge.
  - Use case: Source finding from images to obtain a catalogue containing e.g: position, size, flux. Plus a machine learning classification workflow to classify star-forming galaxies and active galactic nuclei from the resulting catalogue.
  - Integration in processing or analysis workflow: Image analysis, followed by catalogue analysis. Modular, so can pick parts required.

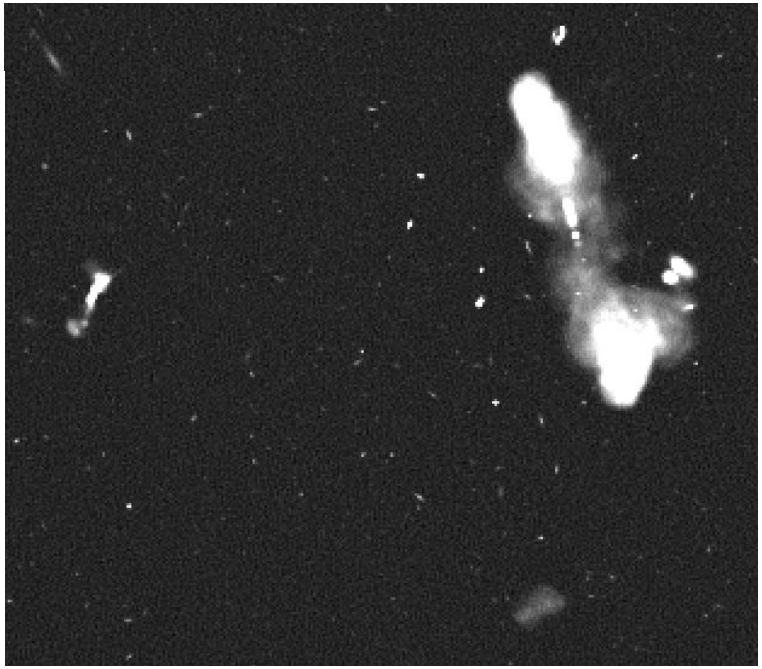
Funded by the European Union's Horizon 2020 - Grant N° 824064

Introduction – 2-3 slides

The SKA data challenge 1

https://astronomers.skatelescope.org/ska-science-data-challenge-1/

Images are provided at 3 frequencies (560, 1400, 9200 MHz). They are 32000 by 32000 pixels, at resolutions of 2.4, 1.2 and 0.6 arcseconds, respectively. They have the same pointing centre, but cover different sky areas. Primary beam files are also provided.



A truth catalogue is provided for a small area of sky, but when the challenge ran in 2019 the full truth catalogue was not provided. We have access to it now.

The challenge was to find sources in each of these images, and classify them as either a star-forming galaxy (SFG), or steep/flat spectrum active galactic nuclei (SS-AGN/FS-AGN).

Funded by the European Union's Horizon 2020 - Grant N° 824064

Software/Service Development – 2 slides

- Software Development Lifecycle Strategies

  Earlier this year we set aside approximately 3 months (May-July) to create a solution with a software pipeline. No more time has been set aside for development, but we plan to enhance our Jupyter notebook version in the near future with some interactive aspects such as the Aladin-lite viewer.

- Development: coding styles, versioning, maintenance, documentation, software quality standards

  We followed PEP 8 style guide, aiming for the code to be modular with classes and functions. There is only one version and we are not setting aside any time for maintenance. Documentation is provided automatically through Gitlab.

  When the challenge ran in 2019 an issue was that participants did not provide their code, only the solution catalogue. Our aim was to provide an example of a research software pipeline, which people can use or refer to, in particular for subsequent SKA data challenges where we hoped to inspire people to provide their software solutions.

- Testing and efficiency optimization strategies

  Testing was exclusive to the data challenge 1 strategy and data products. We explored strategies to expand its use (coping with larger images, calculating spectral indices from an image cube), but these extensions were not directly included.

- platform integration and metadata

  The software is provided in a container, so can be deployed on any platform that has Docker or Singularity.

- software licenses

  We don't provide or require any licenses, it is hosted publicly on Gitlab/Github. We hope to publish the results in the Open Journal of Astrophysics soon.

- General guidelines that are followed

Funded by the European Union's Horizon 2020 - Grant N° 824064

Software/Service Requirements – 2 slides

- Operating System, compilation environment
- Hardware requirements: With such large images, a large amount of RAM is required to run optimally. We had access to 90 GB of RAM, but tested it on a 40 GB of RAM. In theory it can run on machines with even less RAM but we have not tested this yet. There will certainly be a performance trade off, but ultimately to process such large images you will need a sufficiently powerful machine.
- Containerisation and portability requirements: Docker/singularity
- Workflow / interface requirements to other software/services: None anticipated outside of the container

Funded by the European Union's Horizon 2020 - Grant N° 824064

OSSR Integration – 2 slide

- What is available?

The code: https://gitlab.com/ska-telescope/sdc/sdc1-solution/-/tree/master/

The data: https://astronomers.skatelescope.org/ska-science-data-challenge-1/

- What will be onboarded: source code, container, test workflow and data
- Are there open points and requirements: <40GB of RAM, Docker

- What is the ”user story” of a EOSC user taking on the software/service?
  - From the data side (what data can be analysed and how): Very small sample images are provided inside the container for the purpose of unit tests. The full images and catalogues are available via the SKA data challenge website and are downloaded as part of the workflow.
  - From the OSSR side (how to find data and easy use demos, tutorials, documentation, …): The code includes a single executable script that runs the whole workflow. Documentation is hosted on Gitlab. We are creating a jupyter notebook for a more interactive version, but the full workflow is not designed to run in a notebook.

Funded by the European Union’s Horizon 2020 - Grant N° 824064

Time for a short demo

- Show how the software is used and what is the outcome

The whole solutions takes 3 hours to run on 32 core machine with 90GB RAM.

The steps are:

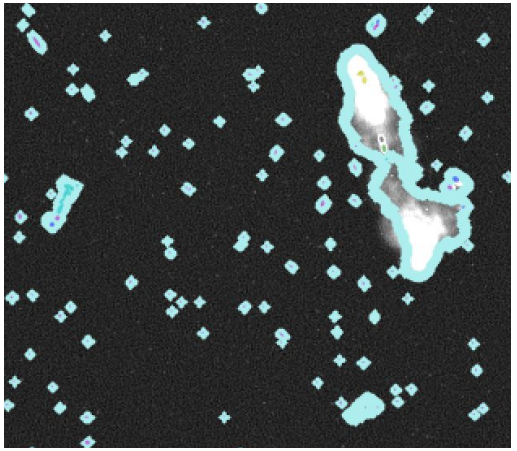
1. Apply primary beams to the images.
2. Run source finder and output a source catalogue for each image.
3. Gather data from the catalogues and implement a machine learning workflow for source classification.

   This involves training a model using the truth catalogue from the training area (or the full truth catalogue from the whole image), and evaluating its performance.

   A user may want to try using different features, or model hyper-parameters, or even different source-finding hyper-parameters to maximise the performance of their model.
4. An output catalogue of sources is provided to the community with properties derived from the source-finder and also the class labels derived from the machine learning model. Metrics are also provided per source and overall that evaluate the accuracy of the classifications. Further algorithms can be investigated, such as dimension reduction sing neighbour-graphs like the plot to the right.

- What should and can a EOSC user do with the software?

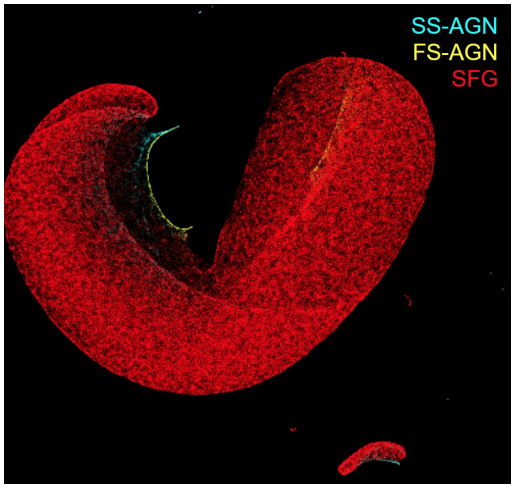
  A user can reproduce the workflow to get the expected result. They can also tune parameters to expand on the workflow. They may also take parts of the software for their own use, such as the primary beam function, source-finding function, classification function.



```
Image: data/images/560mhz_1000h.fits, frequency: 560 MHz
Score was 231518.67518287958
Number of detections 343272
Number of matches 340602
Number of matches too far from truth 1452
Number of false detections 2670
Score for all matches 234188.67518287958
Accuracy percentage 68.75728127928772
Classification report:
              precision    recall  f1-score   support

  1 (SS-AGN)     0.5109    0.0208    0.0401      6716
  2 (FS-AGN)     0.8448    0.2089    0.3350      5342
     3 (SFG)     0.9684    0.9992    0.9836    328544

    accuracy                         0.9675    340602
   macro avg     0.7747    0.4097    0.4529    340602
weighted avg     0.9574    0.9675    0.9548    340602
```



Funded by the European Union's Horizon 2020 - Grant N° 824064

A more in-depth demo of the results

1. Find sources using PYBDSF

2. Classify sources using a Random Forest
   - Take the training area, use these labelled sources to train the Random Forest
   - Applying this model to the rest of the field
   - Compare with truth catalogue

Two aspects for quantifying performance when comparing to truth catalogue:

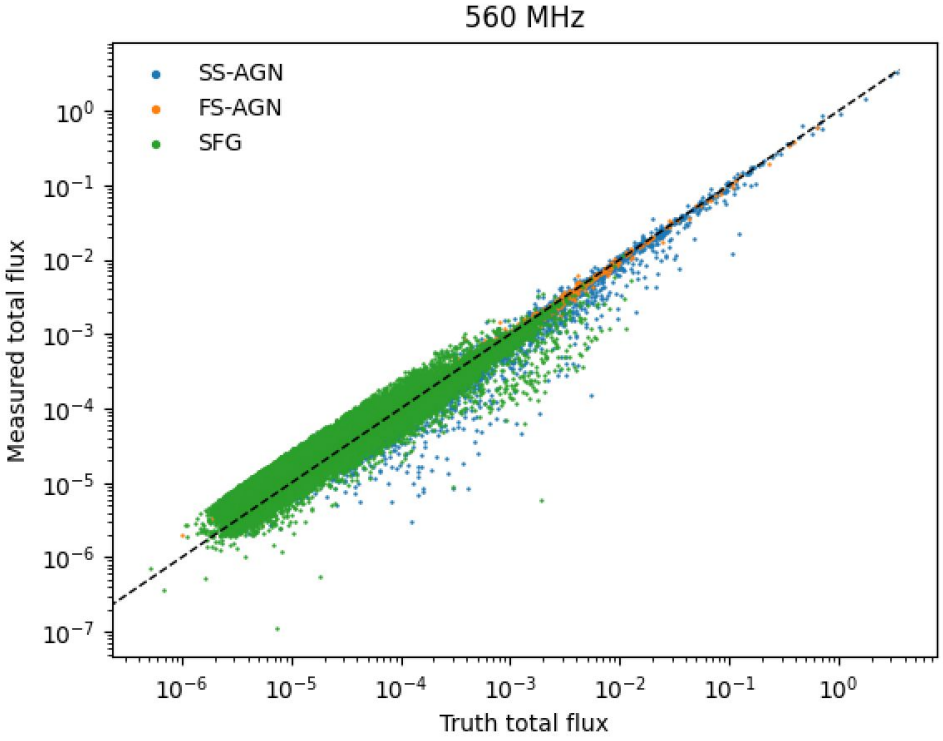
How many sources are found? How many true/false detections?

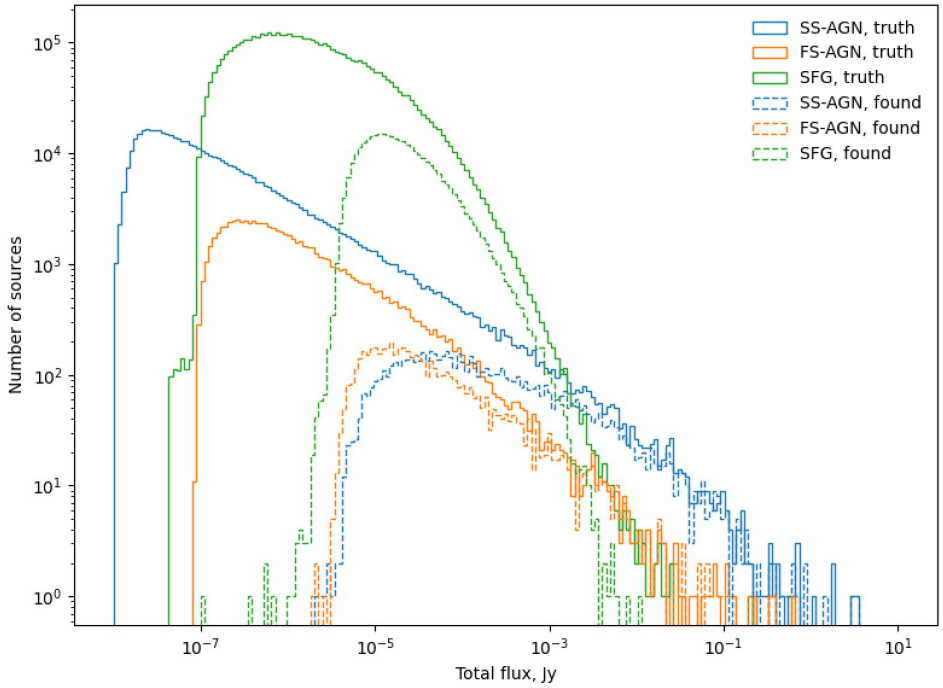
Of the true detections, how many of them are classified correctly?

The scoring package takes class label into account when calculating a score based on number of true/false source detections

For quantifying the performance of a classifier, we can only evaluate precision, recall and F1 score for true detections

Source counts, and comparison to true flux





Precision, recall, F1 score for the Random Forest (RF) - 560 MHz

Fit RF to half of the sources in the training area, test on other half:

	precision	recall	f1-score	support
FSAGN	0.8250	0.2426	0.3750	136
SFG	0.9801	0.9984	0.9892	10424
SSAGN	0.5926	0.1280	0.2105	125

Apply this model to classify the rest of the sources in the field:

	precision	recall	f1-score	support
FSAGN	0.8453	0.2383	0.3718	5342
SFG	0.9714	0.9964	0.9838	328561
SSAGN	0.4993	0.1575	0.2395	6736

Run the scoring code with these class labels:

```
Score was 231607.00169231306
Number of detections 343272
Number of matches 340639
Number of matches too far from truth 1433
Number of false detections 2633
Score for all matches 234240.00169231306
Accuracy percentage 68.7648806191637
```

The F1 score is the harmonic mean of precision and recall, interpreting true positives (TP), false positives (FP) and false negatives (FN):

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Precision, recall, F1 score for the Random Forest (RF) - 560 MHz

Fit RF to half of the sources in the training area, test on other half:

	precision	recall	f1-score	support
FSAGN	0.8250	0.2426	0.3750	136
SFG	0.9801	0.9984	0.9892	10424
SSAGN	0.5926	0.1280	0.2105	125

Apply this model to classify the rest of the sources in the field:

	precision	recall	f1-score	support
FSAGN	0.8453	0.2383	0.3718	5342
SFG	0.9714	0.9964	0.9838	328561
SSAGN	0.4993	0.1575	0.2395	6736

Run the scoring code with these class labels:

```
Score was 231607.00169231306
Number of detections 343272
Number of matches 340639
Number of matches too far from truth 1433
Number of false detections 2633
Score for all matches 234240.00169231306
Accuracy percentage 68.7648806191637
```

This is a hugely imbalanced classification problem. Let's compare scores for guessing against the RF model:

Model	SDC Score	Average F1-score
RF:	231 607	0.53
Guess all SFG:	231 475	0.33
Guess all SS-AGN:	188 275	0.01

Guessing they are all SFG will result in much worse precision, recall and F1 scores:

	precision	recall	f1-score	support
FSAGN	0.0000	0.0000	0.0000	5342
SFG	0.9645	1.0000	0.9820	328561
SSAGN	0.0000	0.0000	0.0000	6736

Conclusion: the inclusion of the class label in the scoring package gives a score biased towards finding star-forming galaxies. Thus the score from the scoring package is not a fair representation of overall performance, it is only fair at assessing the identification of sources, not their class.

F1 scores are a fair representation of the performance of a multiclass, imbalanced classification scheme.

Open Points and Discussion Time

- Which of your questions have not been covered so far?
- What do you want to discuss?

Thanks for listening

Alex Clarke (SKAO)

Funded by the European Union's Horizon 2020 - Grant N° 824064

TOC of Tech Report

- Introduction
  - ESFRI/RI and Partner, Science Case
  - Software and Service Name
- Software/Service Development Strategy
- Software/Service Requirements
- OSSR Integration
  - Status
  - Content
  - User Story

Funded by the European Union's
Horizon 2020 - Grant N° 824064