Deep Learning for galaxy detection on radio-astronomical surveys

Adrien ANTHORE, M2 student
Observatoire de Paris, PSL

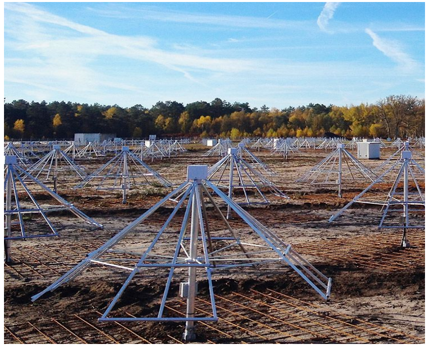
supervised by **David Cornu**

Big data from interferometers

Example of giant interferometers:

LOFAR

source: https://www.obs-nancay.fr/



NenuFAR

source: https://www.obs-nancay.fr/

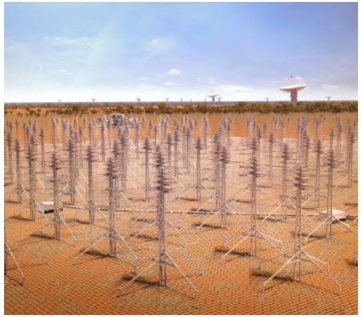
ALMA

Credit : ALMA (ESO/NAOJ/NRAO)/W. Garnier (ALMA)

<u>New scale for radio-astronomical datasets</u>

Upcoming interferometer:
Square Kilometer Array

→ SKA: **700 Po** archived data **per year**



SKA-low

SKA-mid

Credit: SKAO

Very Large amount of data

Classical methods VS **Machine Learning**

Classical methods

Pros:

- Physics driven
- Easy to interpret

Cons:

- Bad scaling with data size/dimensionality

Machine Learning

Pros:

- Data driven
- Can find complex pattern
- Efficient at low SNR

Cons:

- Require training

Problem: **How to construct the training?**

How to construct the training?

Simulated data

Pros:

- Generate examples as much as we want
- Compensate data imbalance (Specific objects or contexts)

Cons:

- Simulated object biased
- Instrumental properties biased

Observational data

Pros:

- Contains all instrumental effects and limitations

Cons:

→ How to define training sample?

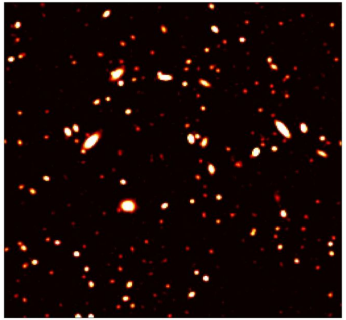
- Deeper survey/other instruments
- Other methods
- Precursors:
  - Are they representative?
  - Require a very high confidence labellisation

Simulated data: SKA data challenges

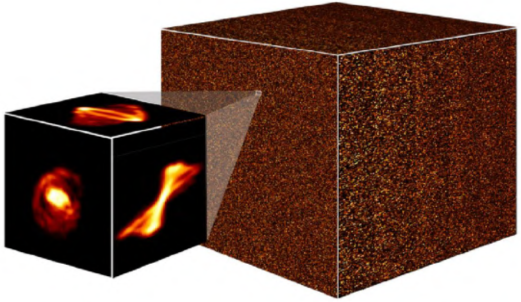
↪Objective: develop **innovative detection methods**

Challenge task:
- Sources detection (RA, DEC, freq)
- Characterization (flux, size, angles)

SDC1: simulated image



SDC2: simulated cube



Minerva team developed a method to address these challenges

SDC1: **Best** a posteriori score

SDC2: **1st place** in the challenge

My work: builds on the continuity of these results

Methodology of the MINERVA team for source detection

Classical computer vision object detection:
Deep Learning yields the **best performances**

⬇

Is it applicable with astrophysical data?

MINERVA team adapted the method

Challenge task:

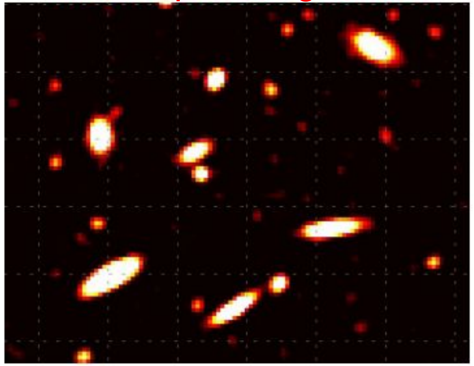
- Sources detection → **Boxes**
- Characterization → **Regression**

↪state of the art results with simulated data

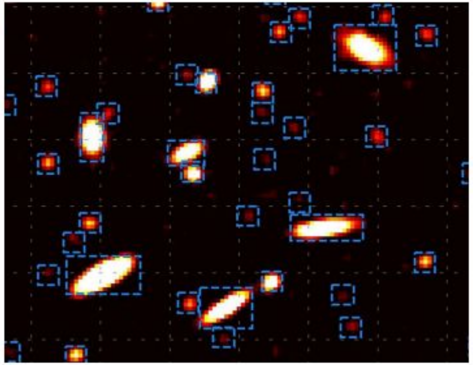
New objective: apply the method to **observational data**

1st step: **direct application** to precursors instruments

Input image



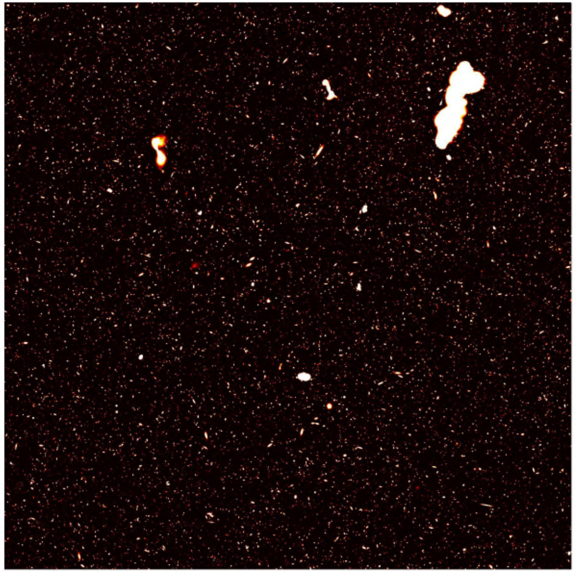
Output: list of bounding boxes



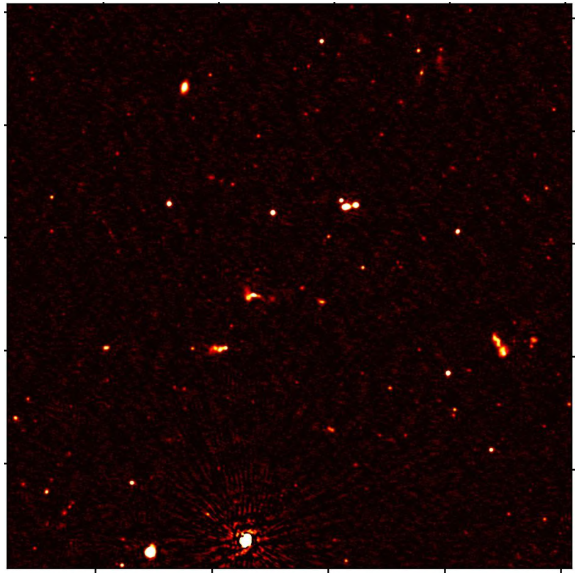
Comparable surveys

<u>My work</u>: focus on **2D continuum emission** (SDC1)

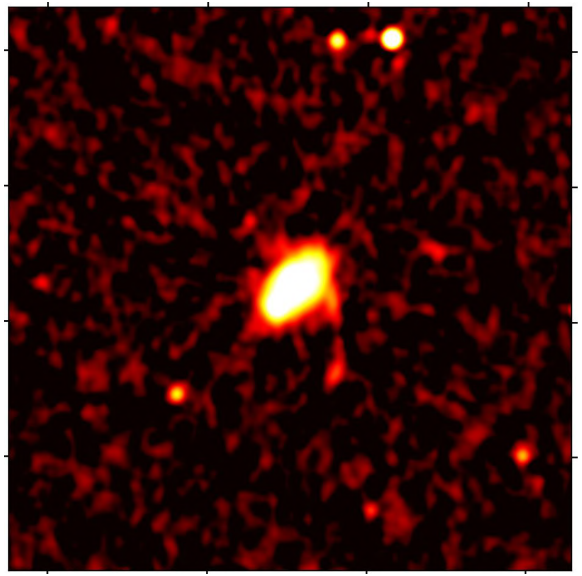
<u>SDC1</u>



<u>LoTSS DR2</u>



<u>RACS-low DR1</u>

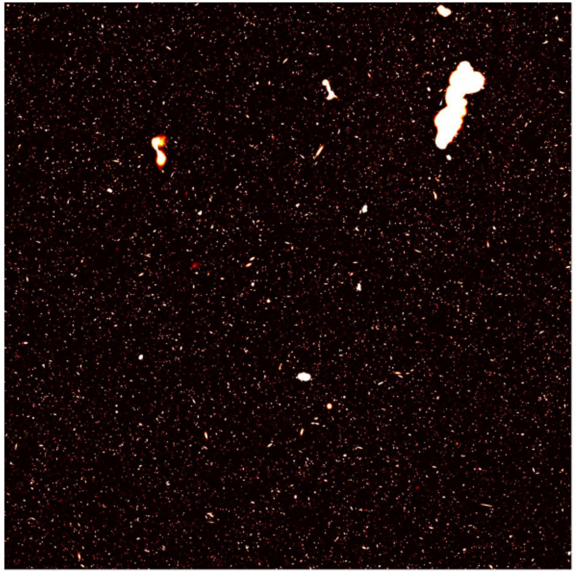


color = same renormalization approach
Subcuts of 25x25 arcmim²

Are they comparable?

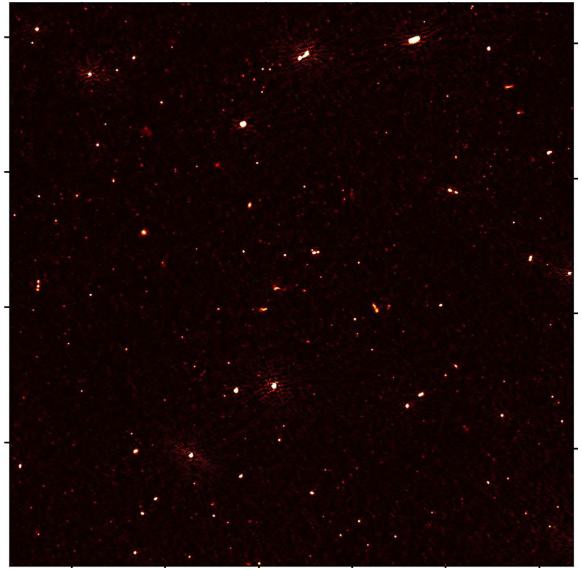
Comparable surveys

SDC1



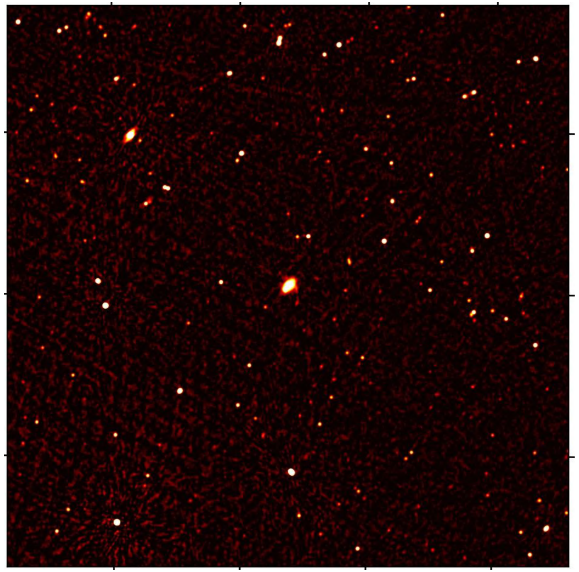
25x25 arcmim²

LoTSS DR2



63x63 arcmim²

RACS-low DR1



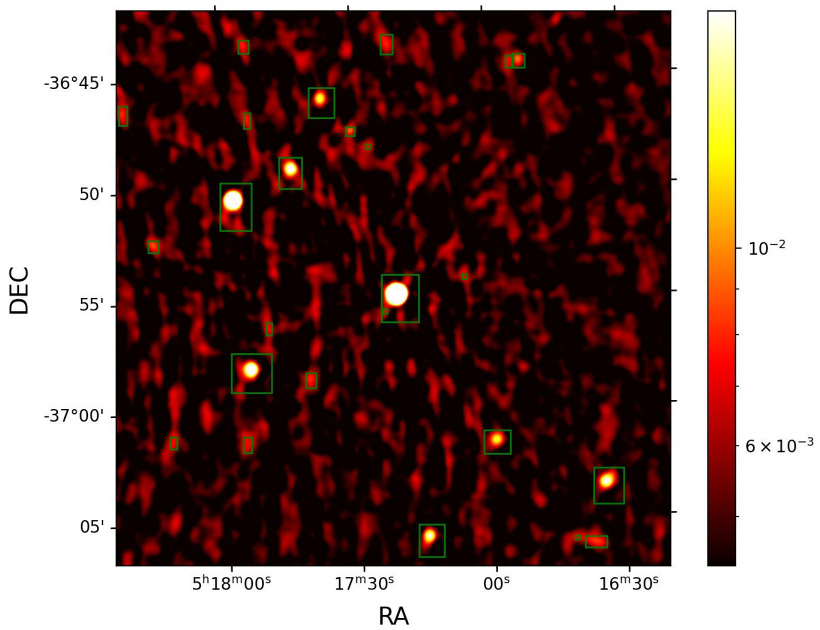
125x125 arcmim²

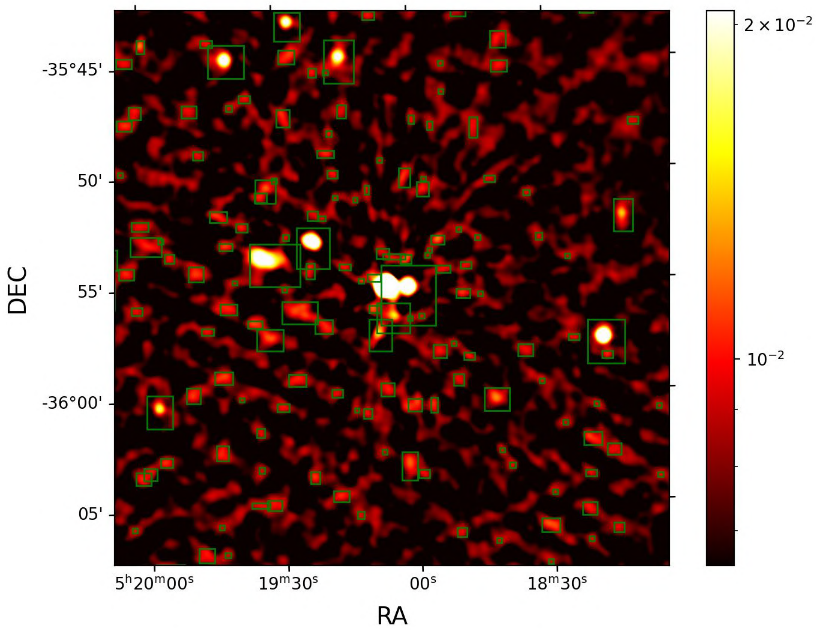
Problem similar regardless of the instrument's resolution

→**Only the scale changes**

Detection on example fields

Mosaic: RACS-DR1_0506-37A



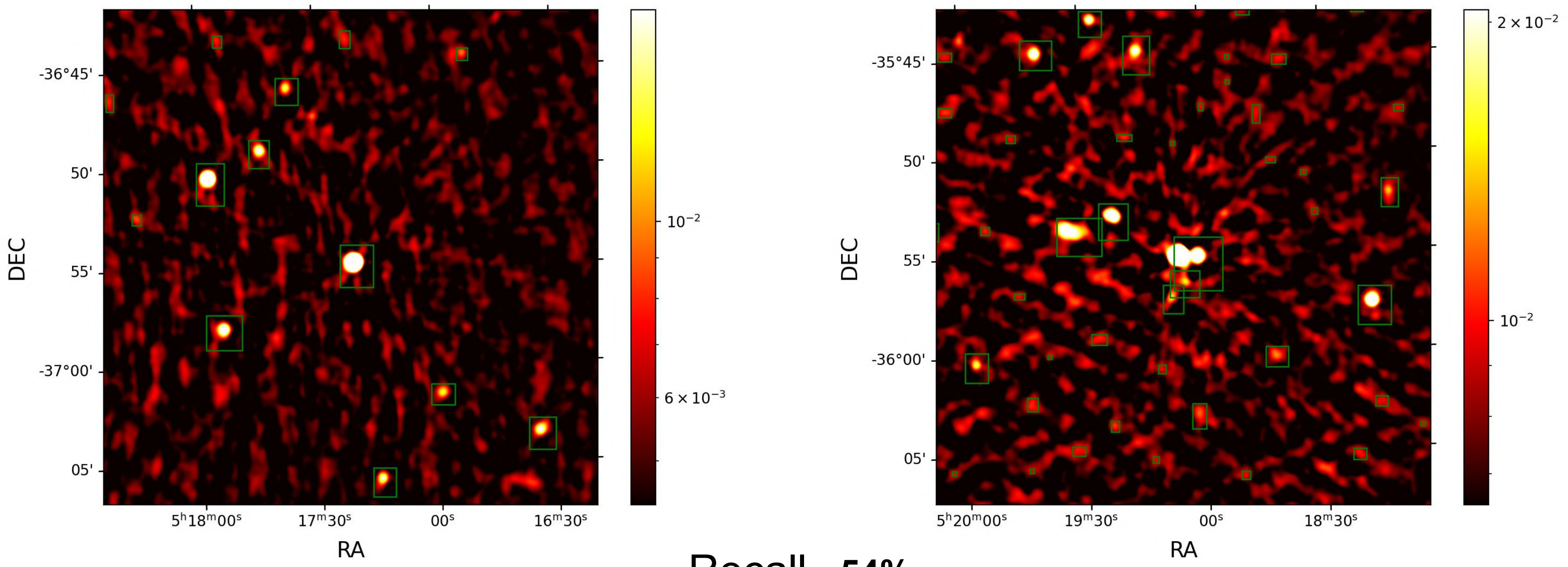


Visible false detections:

→<u>Filter the detection based on the observed flaws</u>

(Artifacts, specific morphologies, …)

Enhanced detection



Recall ~**54%**

Precision ~**65%**

→Detection still limited by **specific instrumental effects**

Limits of this pipeline

→Satisfactory results on both surveys

Limits:

- Artifacts
- Specific morphologies
- SNR limit at which a detection is relevant

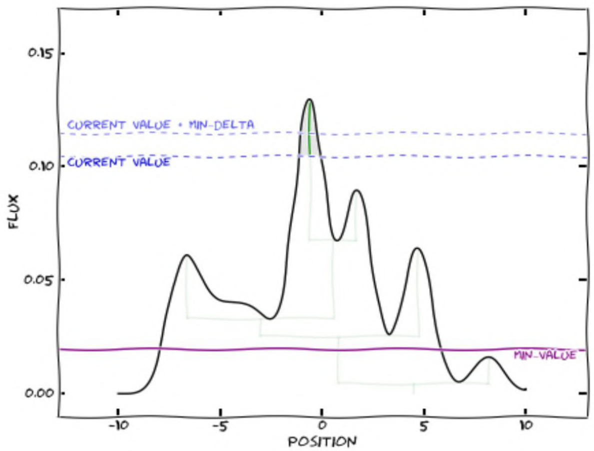
→SDC1 data **don't represents all specificities** from each instrument

To further enhance the detection:
We need a **specific training**

Methods for a specific training

- Deeper Survey
- Observational follow ups
- Create a classical catalog

Catalog **optimized for ML detection** ≠ today's existing catalogs
Our method: Astrodendro



Source: https://dendrograms.readthedocs.io

Based on 3 parameters derived by the observed fields

min_npix →Spatial resolution of the instrument

min_delta →Energy sensitivity of the instrument

min_value →Pixel distribution in a mosaic

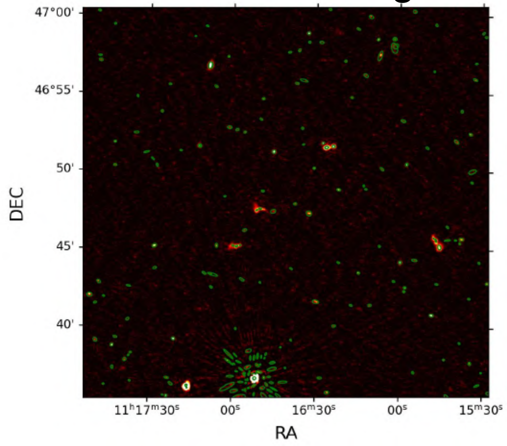
→ Application on example fields

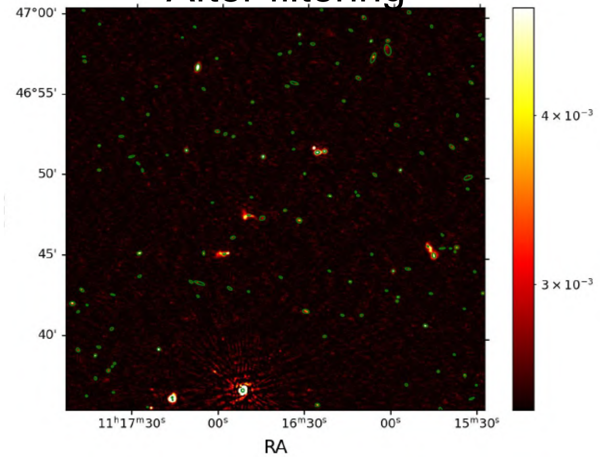
Classical detection on example fields

Before filtering | After filtering

LoTSS:

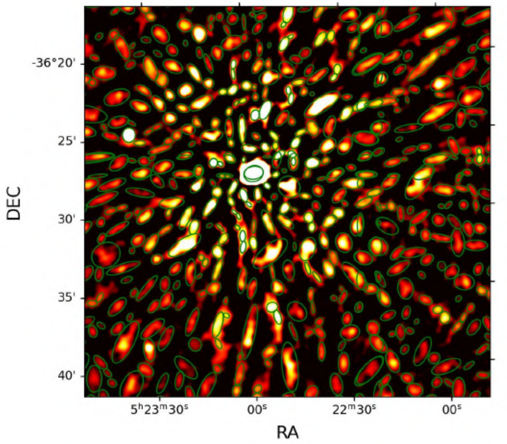
P7Hetdex11

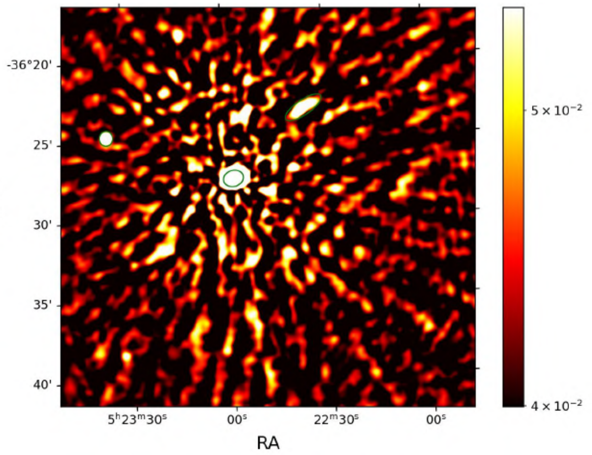




RACS:

RACS-DR1_0506-37A





Limits of our classical method

RACS:

- Recall ~**50%**
- Precision ~**90%**

LoTSS:

- Recall ~**70%**
- Precision ~**50%**

Best performances (Best fields):

- Recall ~**70%**
- Precision ~**90%**

We are reaching the **limit of information** that we can access in **radio data**

To further enhance the results:
→**IR/Optic identification** (Hardcastle et al. 2023)

↪Objective: reintroduce true detections rejected by our criteria

Conclusion

Direct application of MINERVA's method:

- LoTSS data Done!
- RACS data Done!

Construction of high-confidence radio catalog:

- LoTSS data Done!
- RACS data Done!

Construction of enhanced multi-wavelength catalogs:

- LoTSS data In progress…
- RACS data In progress…

Fine training & application of the network: To be done.

↪M2 Internship