Detection and characterization of galaxies using Machine Learning on a massive radio-astronomical dataset

Internship performed at the **LERMA**
supervised by **David Cornu**

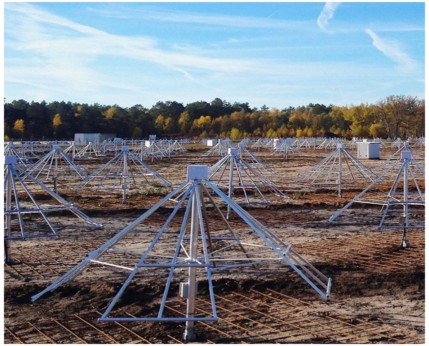
l'Observatoire de Paris | PSL

Big data from interferometers

Example of giant interferometers:

LOFAR

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



NenuFAR

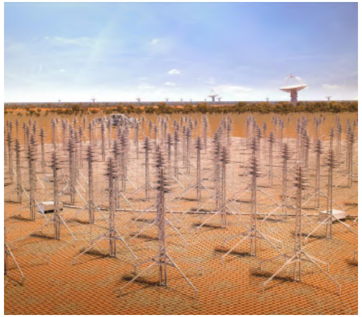
source: https://www.obs-nancay.fr/

ALMA

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

Upcoming interferometer:
Square Kilometer Array

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



SKA-low

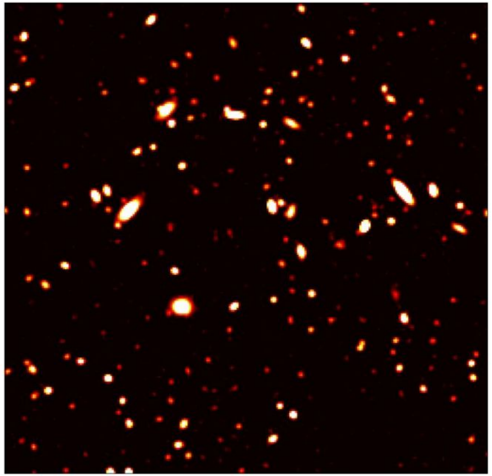
SKA-mid

Credit: SKAO

How to develop innovative detection methods in preparation for SKA ?

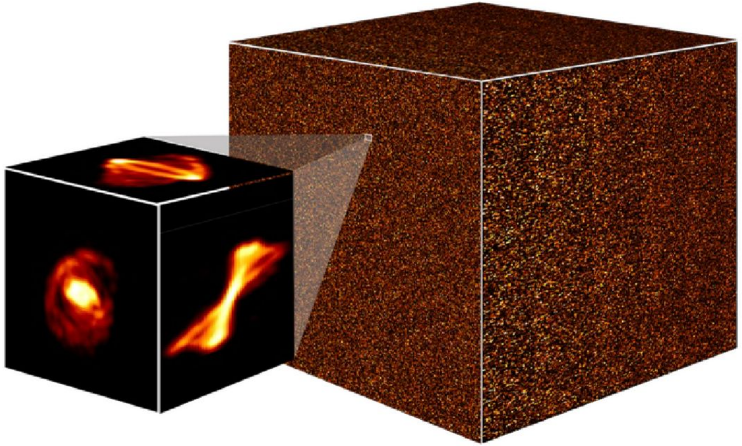
↪ SKA Science Data Challenges (SDC): **simulated data**

SDC1: simulated image



Best a posteriori score

SDC2: simulated cube



First place in the challenge (2021)

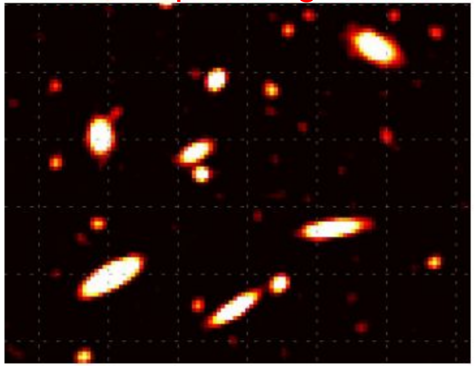
Which method for large datasets ?

↪ Large datasets require **high-performance** statistical approaches: **Machine Learning**

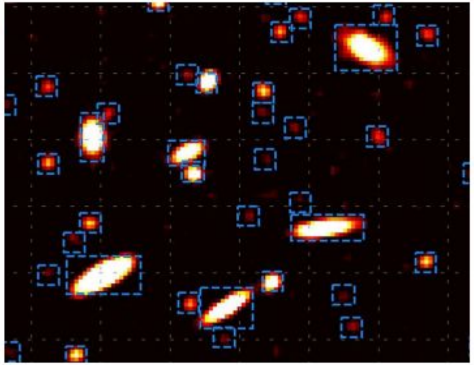
MINERVA SDC1 detection pipeline

Training on SDC data (simulated data) → **Run on SDC data (simulated data)** → **Catalog of detections**

Input image



Output: list of bounding boxes



State of the art results with simulated data

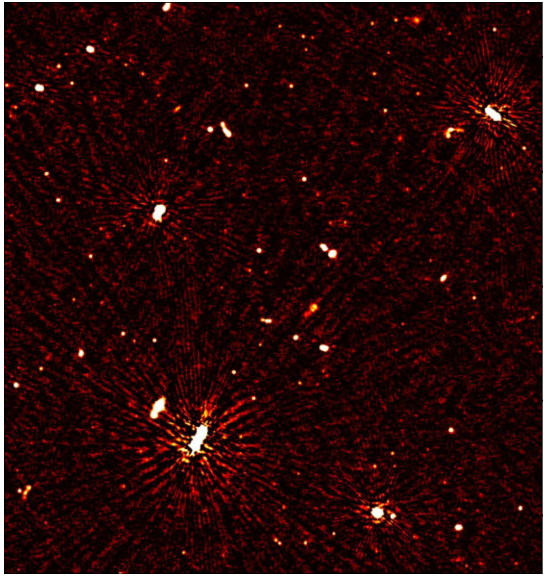
What about real data ?

LOFAR Two-Metre Survey (LoTSS)
A massive radio-astronomical dataset

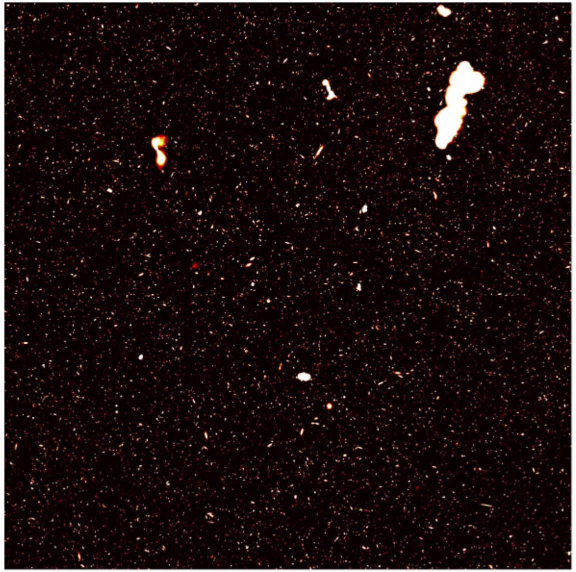
<u>LoTSS DR2: Shimwell et al. 2022</u>

- Low frequency (120-168 MHz)
- 27% of the Northern sky
- 4,396,228 radio sources
- Classical detection method (PyBDSF)
- 800 mosaics

Subcuts of 25 arcmim²



LoTSS



SDC1 (560 Mhz)

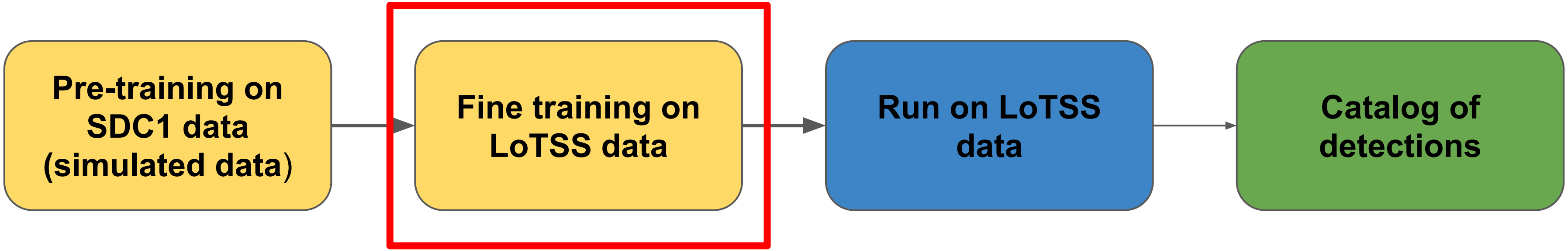
→ LoTSS catalog used as a high quality source **catalog reference** to **evaluate detection performances**

Source detection pipeline update for LoTSS

Naive SDC1 to LoTSS pipeline:

→ Satisfactory results, but difficult to deal with artifacts…

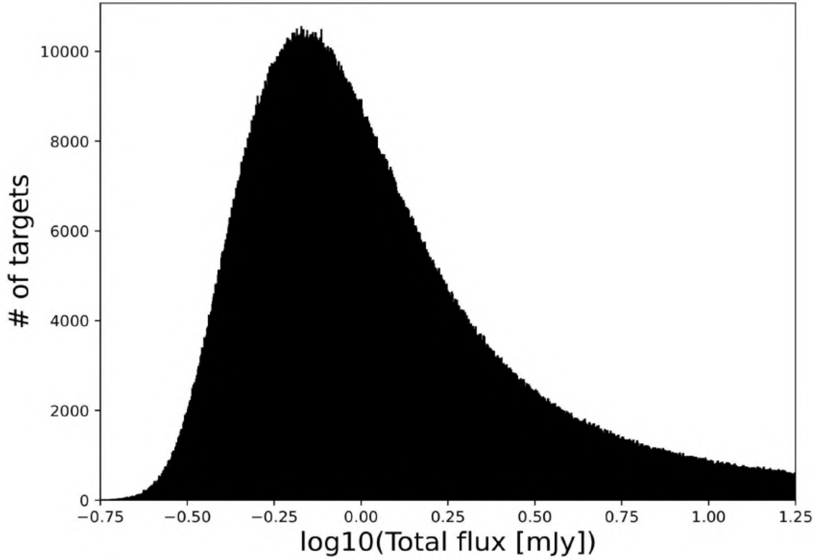
LoTSS dedicated pipeline: allows fine training of the network on LoTSS



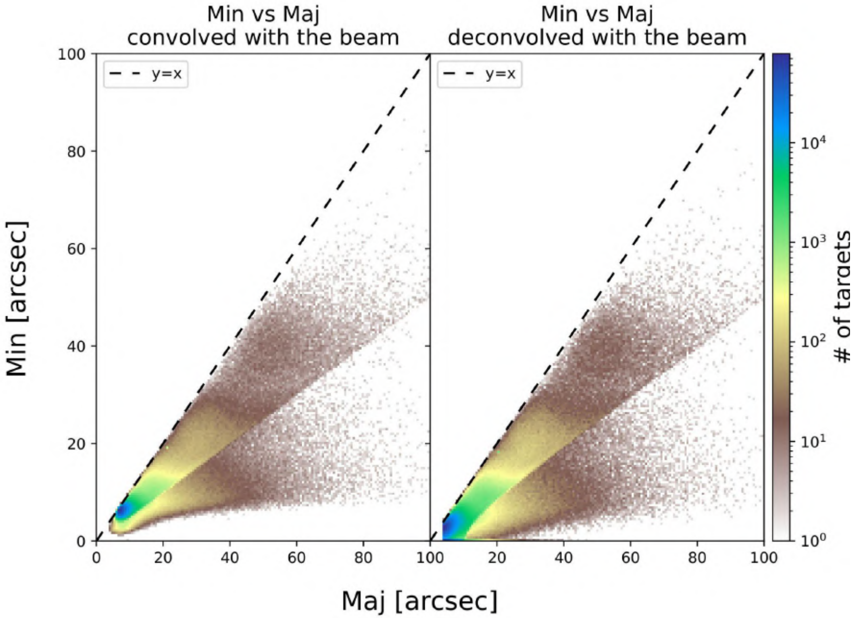
→ **My work**: construct a **high-confidence catalog**

Pipeline informed LoTSS catalog analysis

↪ Summary statistics



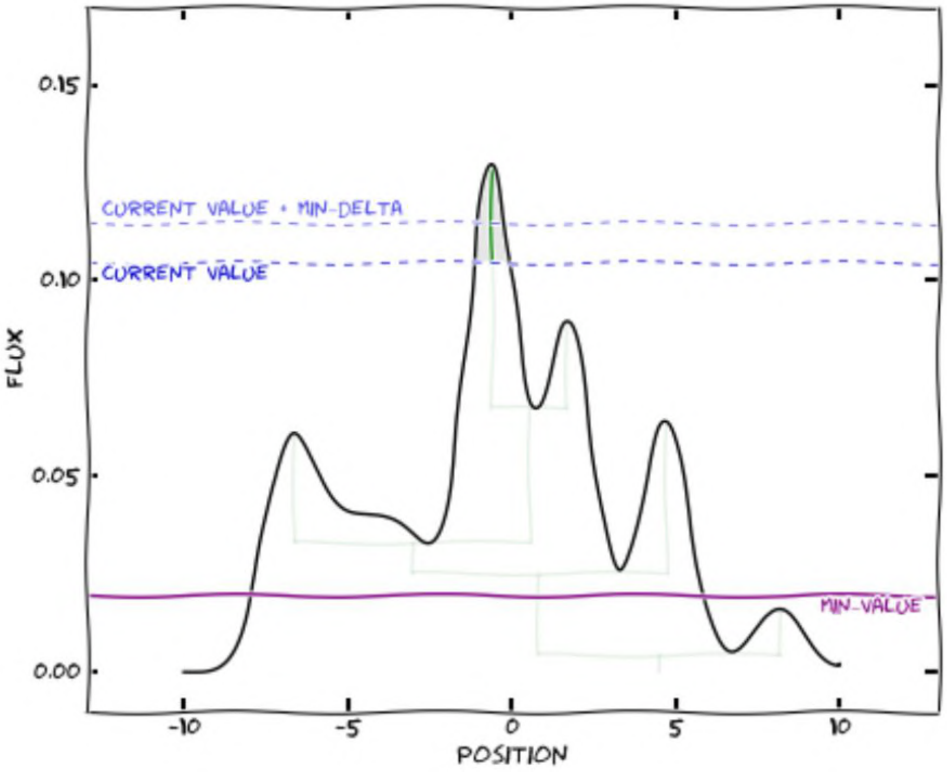
Flux distribution of the sources in LoTSS catalog
(cleaned through citizen science (RGZ))



Size distribution of the sources in LoTSS catalog

We expect these distributions to represent specificities of the LoTSS DR2 underlying statistics

Alternative non-ML method for object detection: Astrodendro



Source: https://dendrograms.readthedocs.io

Astrodendro parameters:
min_npix ; *min_value* ; *min_delta*

<u>For a single field (P7Hetdex11)</u>

With default parameters:

- **Large computation time** (hours)
- Lots of **false detections**
- **Too many detections**

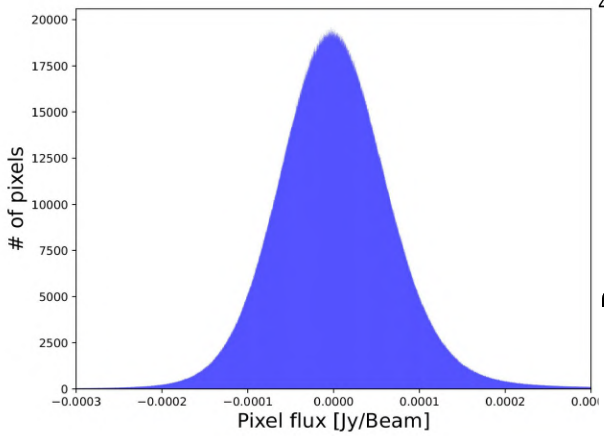
  (~18,000 sources/degree^2,

  LoTSS ~1,200 sources/degree^2)

Requires an **automated** selection of these parameters for **each field**

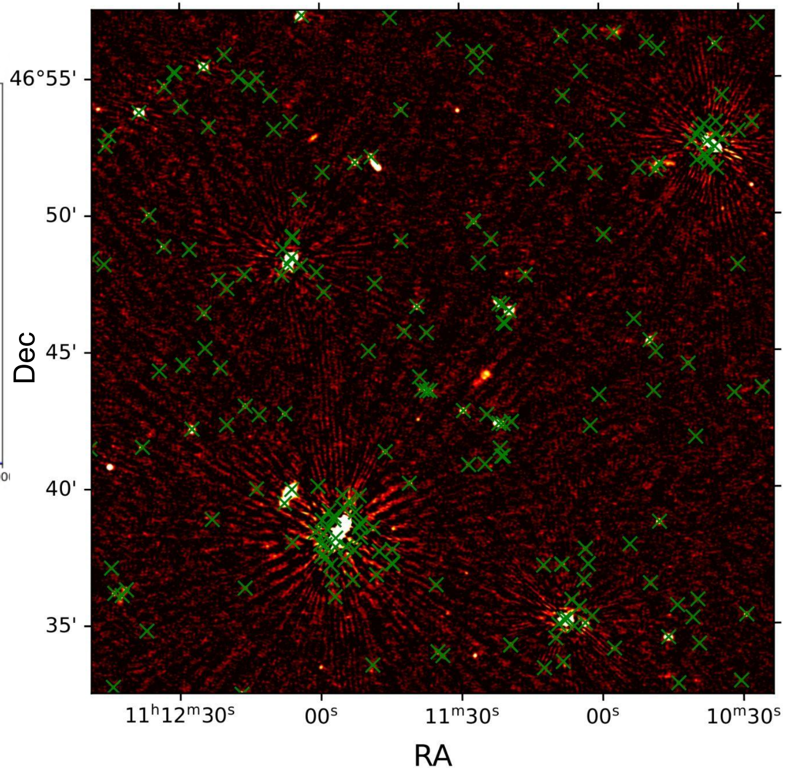
How to select optimal parameter for each field automatically ?

↪ Parameters derived from the observed field



Pixel distribution in a mosaic

→fit *min_value & min_delta*



With PyBDSF as reference:

Recall = *N_match / N_PyBDSF*

Precision = *N_match / N_dendro*

Using adapted parameters:

- 1500 sources/degree²
- Recall **100%** → **~60%**
- Precision **<10%** → **~70%**

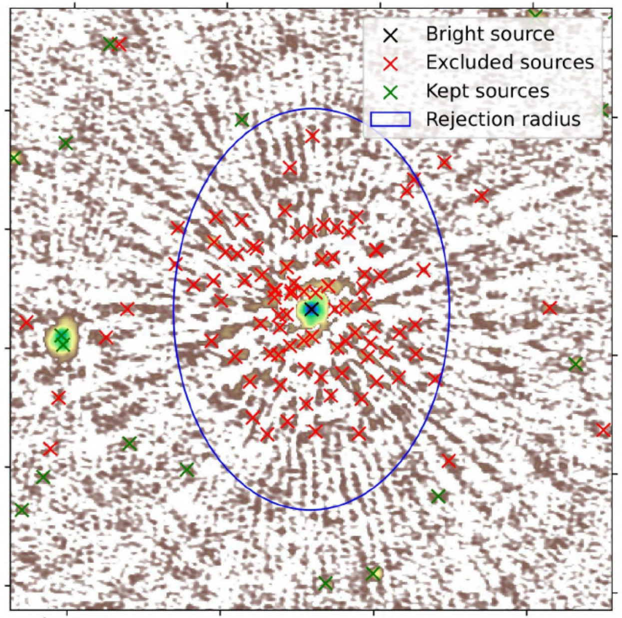
Many artifacts cataloged

→ Need more filtering

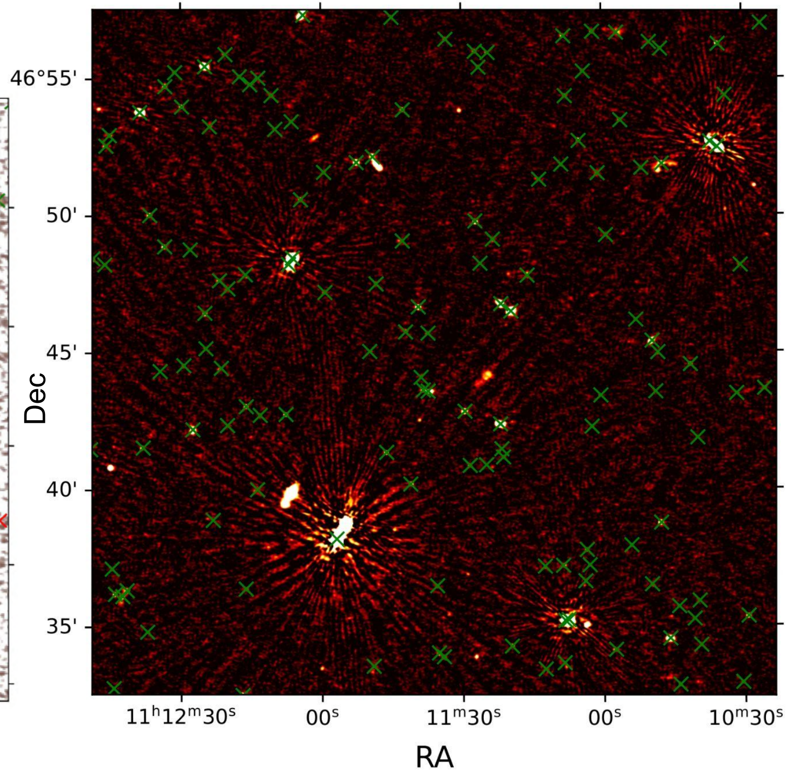
How to enhance the detection ?

↪ Post-detection filtering

- **Eccentricity limit**
- **Distance to bright sources**



LoTSS sub-field
centered on a bright source
(10’x10’)



Using adapted parameters & filtering:

- 800 sources/degree²
- Recall **~60%** → **70%**
- Precision **~70%** → **90%**

Fewer artifacts cataloged
→ Apply on full LoTSS field

Final catalog results
Method applied on all LoTSS mosaics

Results **averaged** on each field:

Recall	$59.9 \pm 15.5\%$
Precision	$61.9 \pm 7.5\%$
Flux relative difference	0.10 ± 0.17
Major axis relative difference	-0.46 ± 0.21
Minor axis relative difference	-0.54 ± 0.14

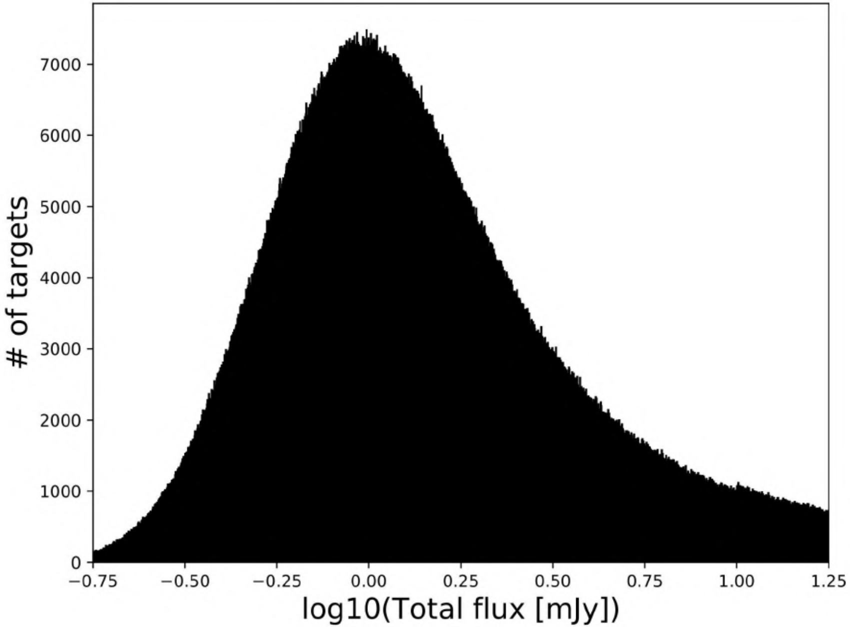
Best field

Recall~70% ; Precision~90% (~7000 detections)

Worst fields

Recall~0.02% ; Precision~100% (2 detections)

Recall~70% ; Precision~10% (~13000 detections)



Flux distribution of the sources in Astrodendro catalogs

→ For **high purity and recall**, Astrodendro begins to be **comparable** to PyBDSF

Although flaws are still identifiable, this method begins to challenge it.

Conclusion

Main results:

- Avg Recall = 59.9 ± 15.5 %
- Avg Precision = 61.9 ± 7.5 %

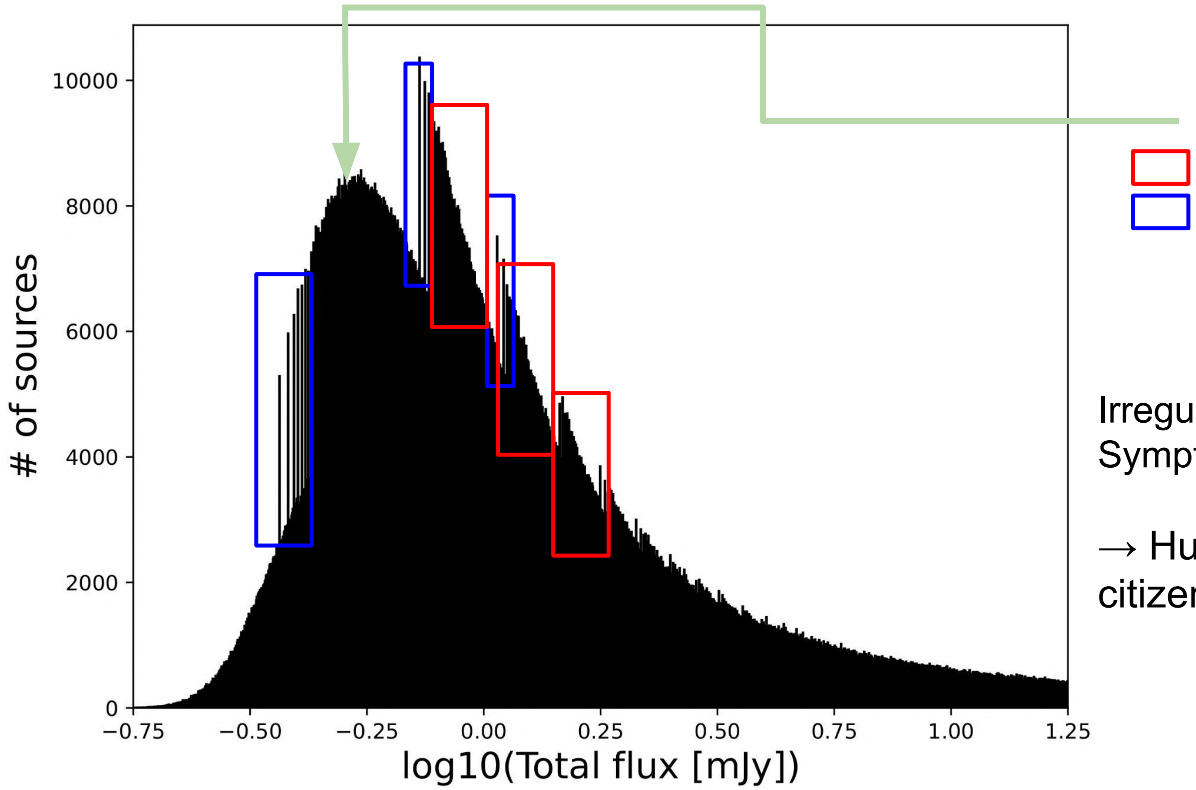
Overall pipeline update progress:

- Construction of a high-confidence catalog
- Construction of the training dataset
- Training of the network and detection on the LoTSS fields To be done.

Perspectives:

- Unification of all the individual field catalogs
- Enhancement using cross-matches (visible & infrared)

LoTSS flux distribution without human inspection



Irregular distribution:
Symptomatic of **questionable detection**

→ Human inspection made through citizen science: Radio Galaxy Zoo (RGZ)