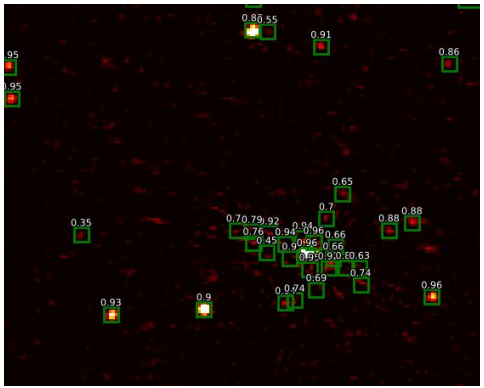
Deep Learning for galaxy detection on radio-astronomical surveys

Adrien ANTHORE, M2 student
Observatoire de Paris, PSL

supervised by **David Cornu**



Radio-interferometric observations

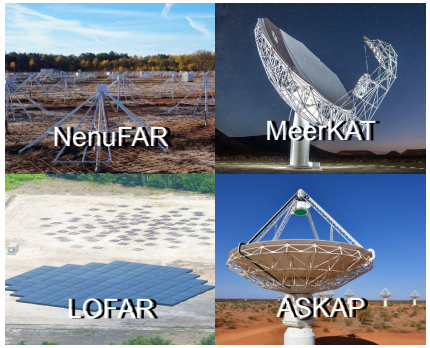
Radio observations:

- Continuum: **Synchrotron effect**
- 21 cm line (HI): **Neutral hydrogen**

Complex data:

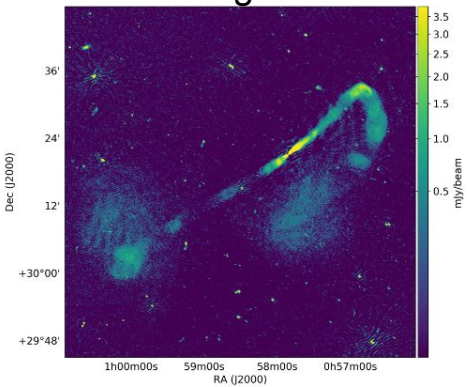
- Create images from visibilities
- Variety of features (sources, artifacts, etc.)

Today's instruments:



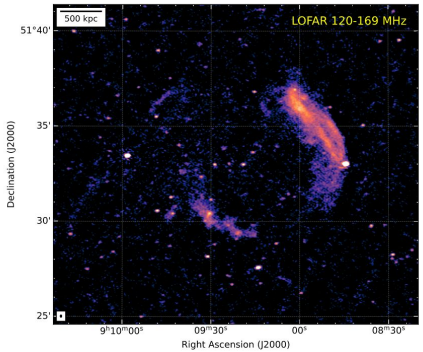
Examples of applications:

Active galaxies



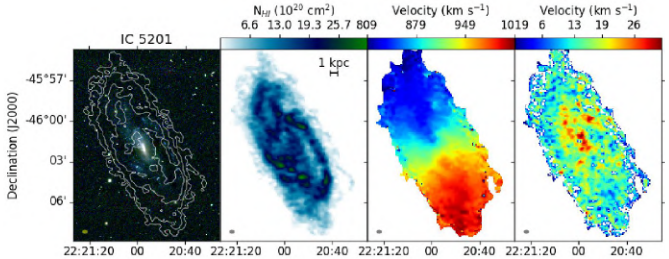
Credit: Shimwell et al. (2022)

Galaxy clusters



Credit: Rajpurohit et al. (2024)

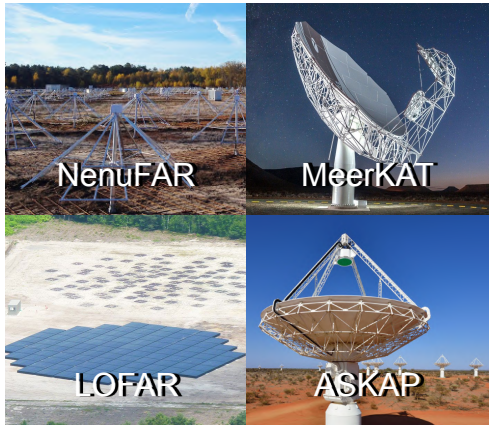
Galactic dynamics



WALLABY, Kleiner et al. 2019

Radio-interferometric observations

Today's instruments:



Forthcoming:

Current generation: **PB scale**
e.g. LoTSS DR2: 7.6 PB of data (3451 h)

Classical analysis methods (widely used today) **does not scale well**

Square Kilometer Array (SKA)
Forecast: **700 PB** of archived data **per year**

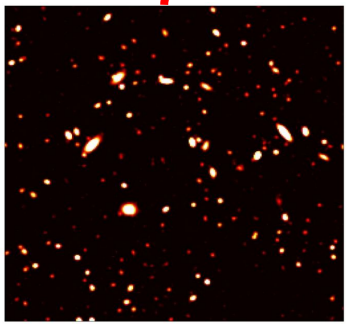
Classical methods are even more challenged

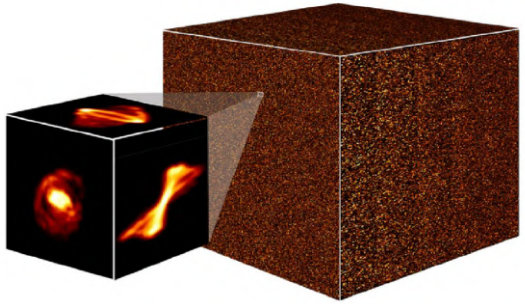
Data Challenges (SDC)

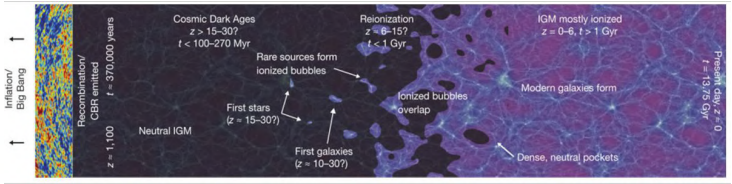
Simulated datasets resembling typical expected SKA data

Subset of labeled data were provided to train **statistical methods**

Source detection and characterization







Crédit : Robertson et al. (2010)

SDC1
Continuum 2D
Each image: **4 Go**

SDC2
HI emission cube
Full cube: **1 To**

SDC3
Visibilities and images
Full size: **17 To**

→These challenges provide a good framework for developing methods

Source detection and characterization

To perform any task (classification, measurement, etc.): →**First: identify the objects of interest**.

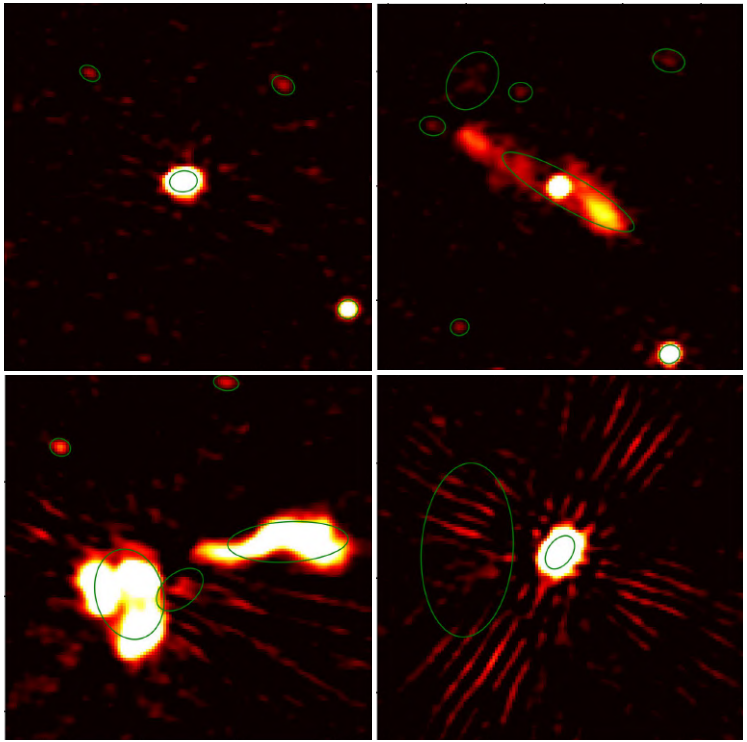
Classical methods: **widely used**
Examples: SExtractor, PyBDSF, …

→**Common to find false detections**

Pros:
- Physics driven
- Easy to interpret
- High reproducibility

Cons:
- **Bad scaling with data size/dimensionality**

Typical cases:



Methodology of the MINERVA team

MINERVA: MachINe lEarning for Radioastronomy at obserVatoire de PAris

Deep learning is state-of-the-art method for object detection in everyday life images

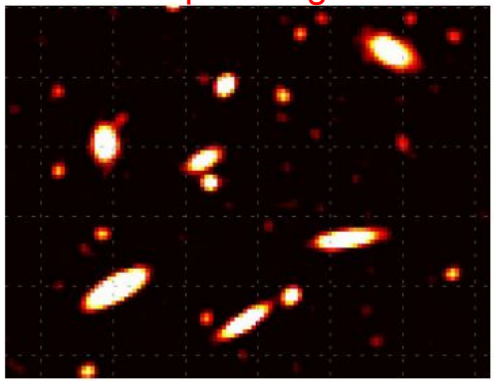
- Detection → **Boxes**
- Characterization → **Regression**

MINERVA's approach: Do **state-of-the-art computer vision** methods adapt to radio data?

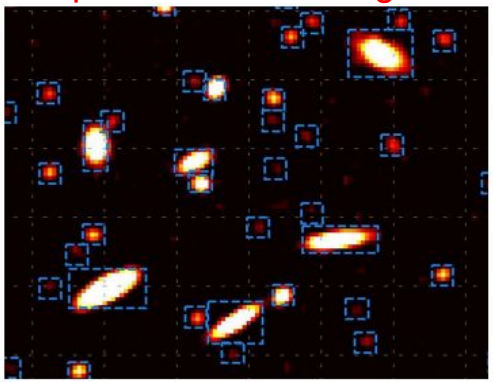
Solution: build a **supervised deep learning method** dedicated to astronomical data (YOLO-CIANNA, Cornu+ 2024)

Can find complex patterns automatically
Good scaling with data size/dimensionality

Input image

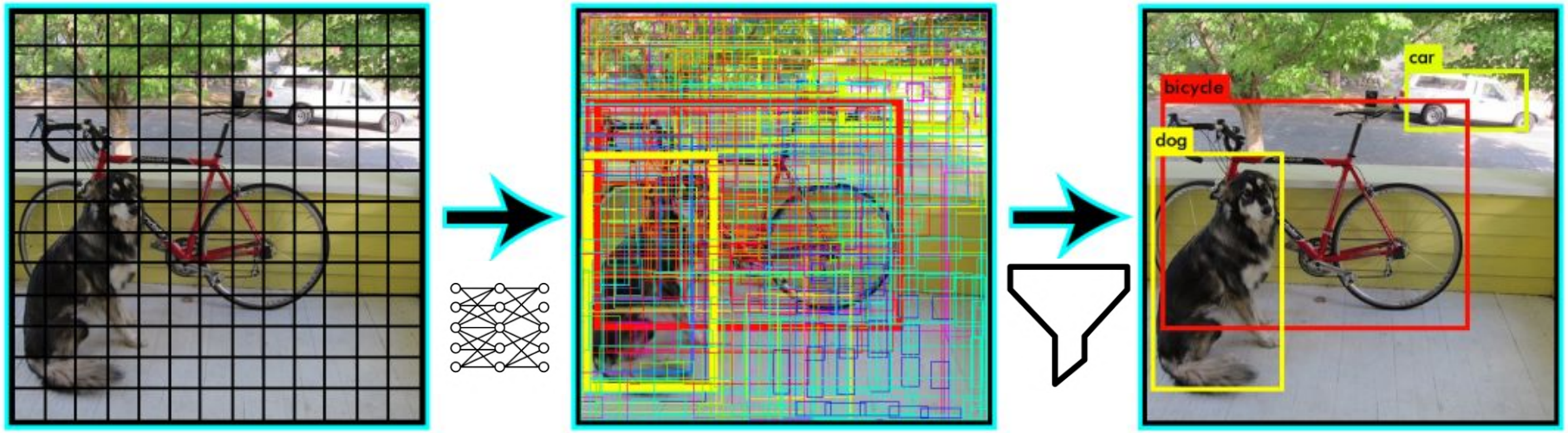


Output: list of bounding boxes



You Only Look Once: YOLO

Redmon et al. 2015, 2016, 2018



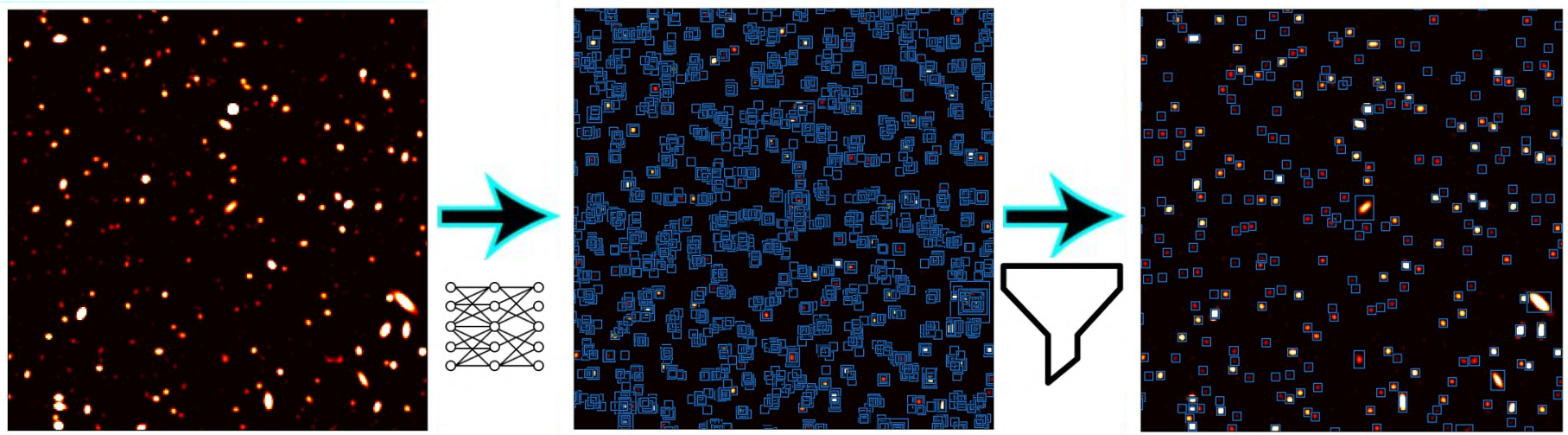
Raw network detection

- Grid the image
- Detect a list of bounding box
- Associate a score to each box

Detections filtering

1) Most probable boxes are kept
2) We take the most probable box and removes overlapping ones

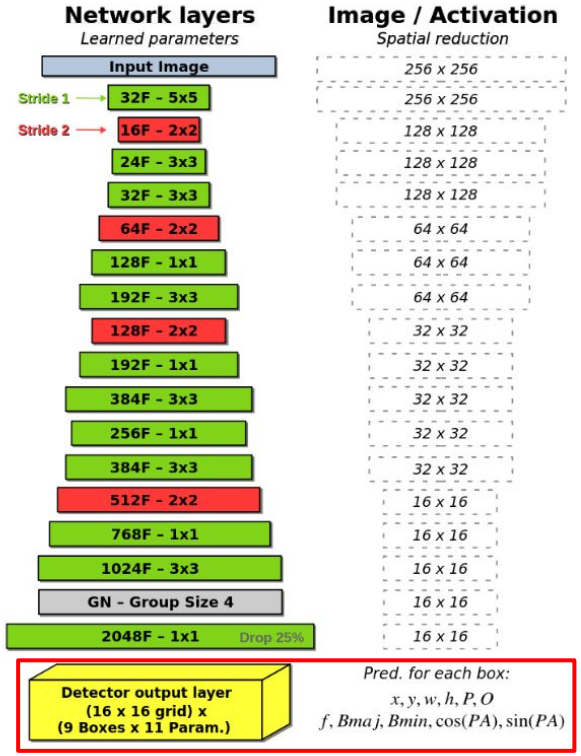
You Only Look Once: YOLO



Same process overall with astronomical images

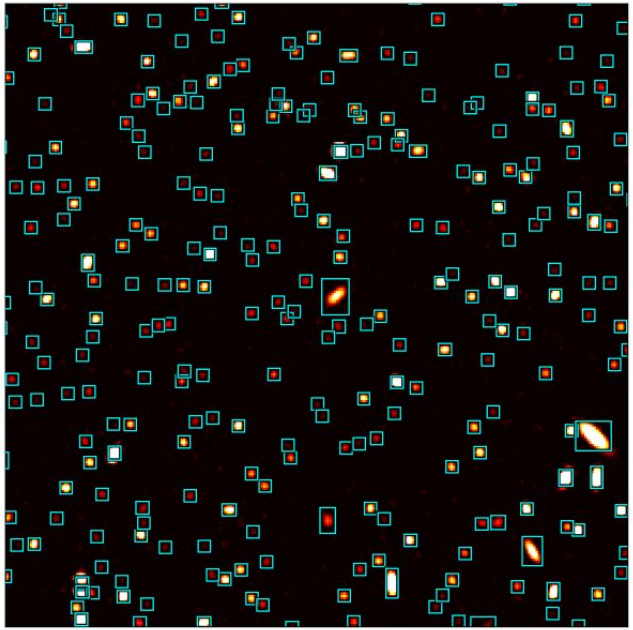
Each bounding box is associated with **parameters**:
RA, DEC, Flux, Bmin, Bmaj, PA

Performances on the SDC1



Cornu et al. 2024

SDC1 data (simulated)



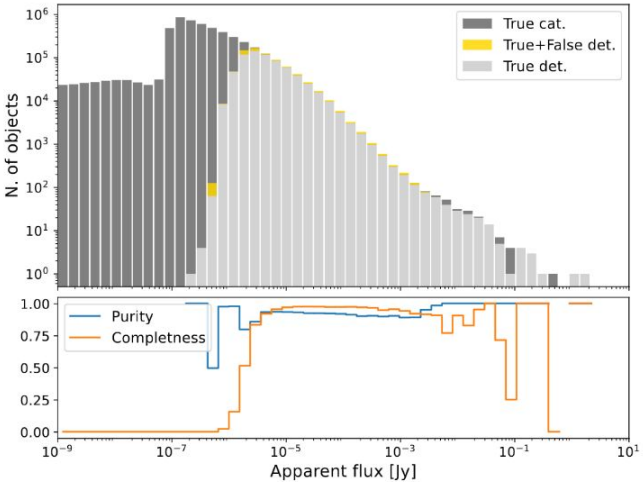
Matched detections

Works on simulated data

Sources predicted parameters (Position, Flux, Size)

Performances on the SDC1

Cornu et al. 2024



Matching criteria

$$E_{tot} = \sqrt{E_{pos}^2 + E_{size}^2 + E_{flux}^2}, \text{ with}$$

$$E_{pos} \propto \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2} / \hat{S}',$$

$$E_{size} \propto \left| S - \hat{S} \right| / \hat{S}', \text{ and}$$

$$E_{flux} \propto \left| f - \hat{f} \right| / \hat{f},$$

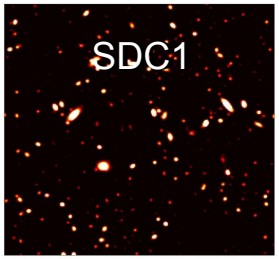
Average per-source score

	Team (method)	M_s (Score)	N_{det}	N_{match}	N_{false}	$N_{bad} \in N_{false}$	Purity	$\bar{s}$	
	Post-challenge results								
ML (CNN)	MINERVA (YOLO-CIANNA-ref)	**479758**	718760	677025	41735	15787	94.19%	0.7703	*After challenge ending*
	MINERVA (YOLO-CIANNA-alt)	414937	538649	533659	4990	2402	**99.07%**	0.7869	
ML (CNN)	JLRAT2 (JSFM2)	**298201**	502146	484212	17934	2274	96.43%	0.6529	
	Original challenge results								
Classical	Engage-SKA (PROFOUND)	**200939**	421992	418384	3608	2677	99.15%	0.4889	*Original leaderboard*
Classical	Shanghai (multiple methods)	**158841**	292646	291553	1093	698	99.63%	0.5486	
ML (CNN)	ICRAR (CLARAN)	**142784**	279898	259806	20092	6875	92.82%	0.6269	
	...	...	...	...	...	...	...	...	

From simulation to observation

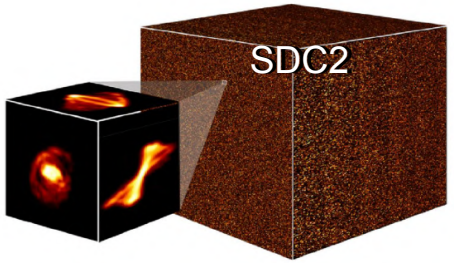
New objective: apply the method on **observational data**

2D continum



best score *a posteriori*

3D HI emission cube



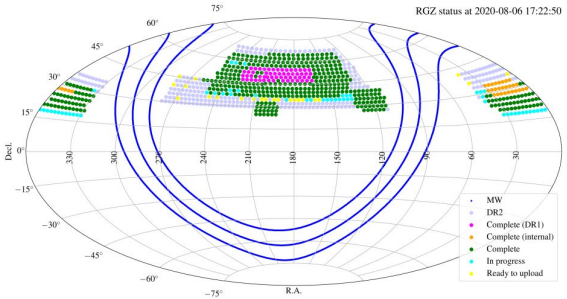
1st place in the challenge

Simulated

Analogue observational

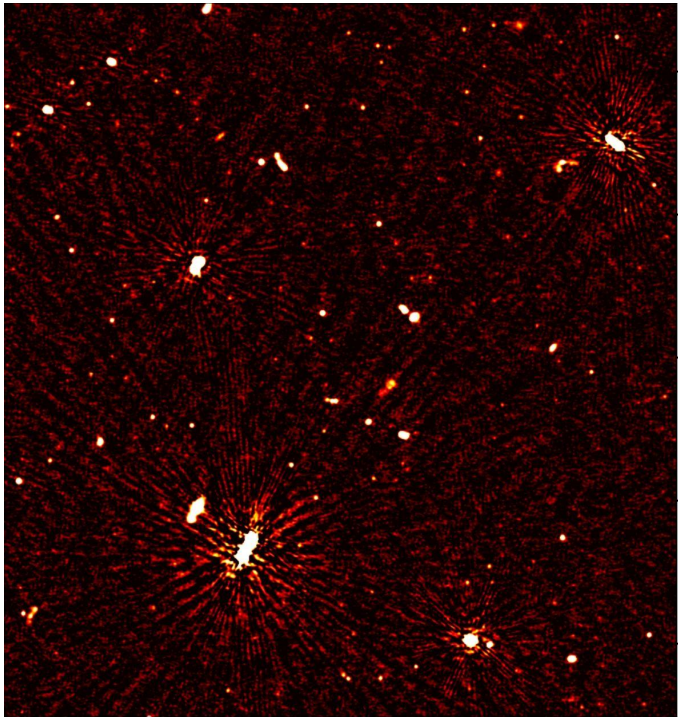
My work: deploy the method to LOFAR data

LOw Frequency ARray (LOFAR)



LoTSS DR2 coverage

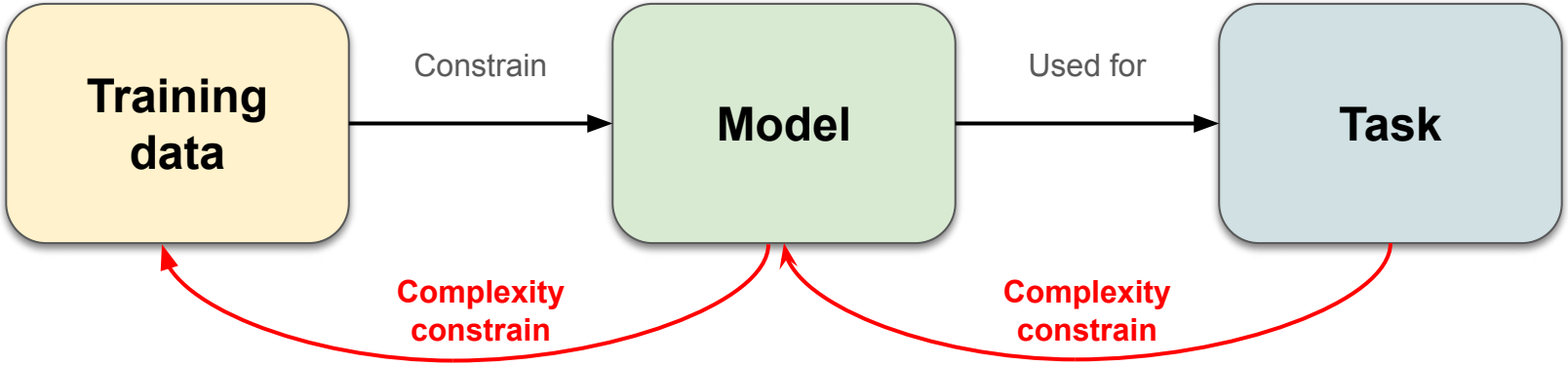
subfield of 25x25 arcmim²



LOFAR Two meter Sky Survey (LoTSS) DR2:
Shimwell et al. 2022

- Frequency: 120-168 MHz
- 27% of the Nothern hemisphere
- **4,396,228 cataloged sources** from **PyBDSF**
- 841 mosaics

Training dataset for supervised method



This training dataset must:

- Be **complete**: all specificities of the data must be represented (objects, effects, contexts, …)
- Have **pure labels**: labels should be as close as possible to the expectancy.

Not/wrongly labeled target:

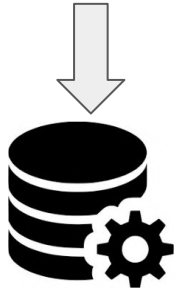
→If the network detects it, will **lower the probability of all the same kind of sources**

Labeled questionable target:

→If the network detects it, may **increase the probability of detecting noise**

How to train the model for application on observational data?

Observational data



Training dataset

Training

1st option: **using observational data**

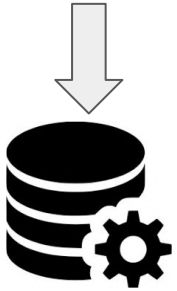
Pros:
- Contain all instrumental effects and observational limitations

Cons:
- Limitation in examples
- Scarcity effects
- Difficult to label data

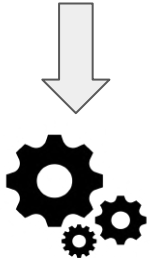
Because it requires a lot of exemple: labeling a **large portion of the survey makes ML useless**

How to train the model for application on observational data?

Simulated data



Training dataset



Training

2nd option: **using available simulated data**

Pros:

- As many examples as necessary
- Compensating for scarcity effects

Cons:

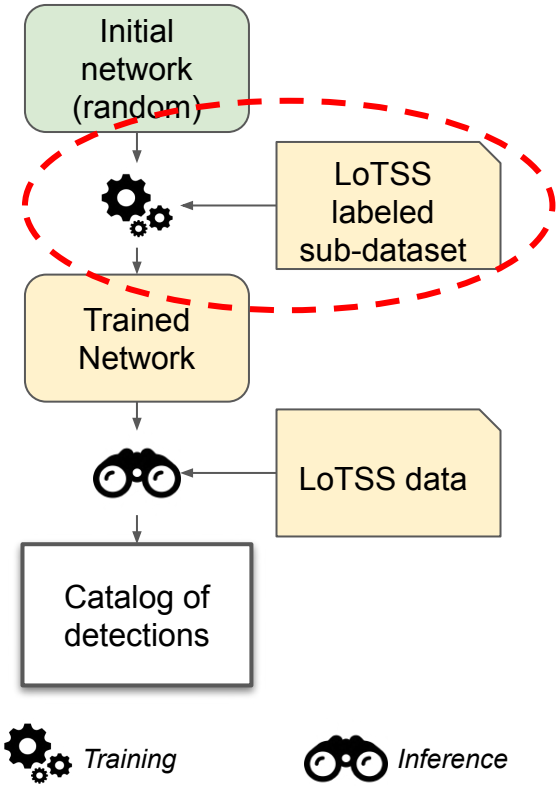
- Potentially biased or simplistic:

Physical model	Instrumental Model

In practice the two approaches can be combined

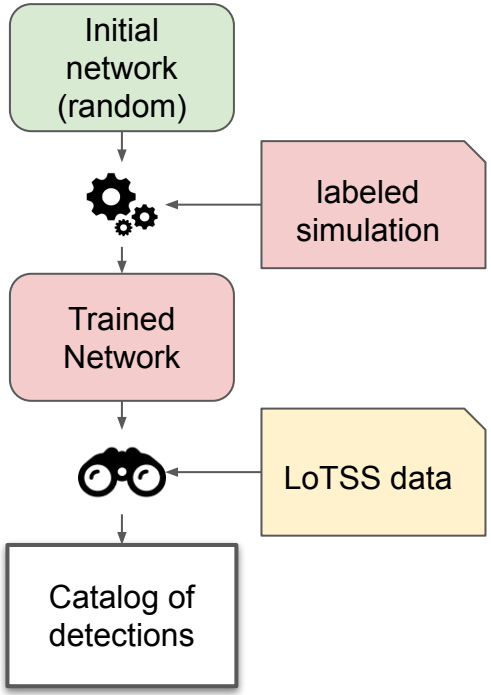
Approaches

LoTSS only

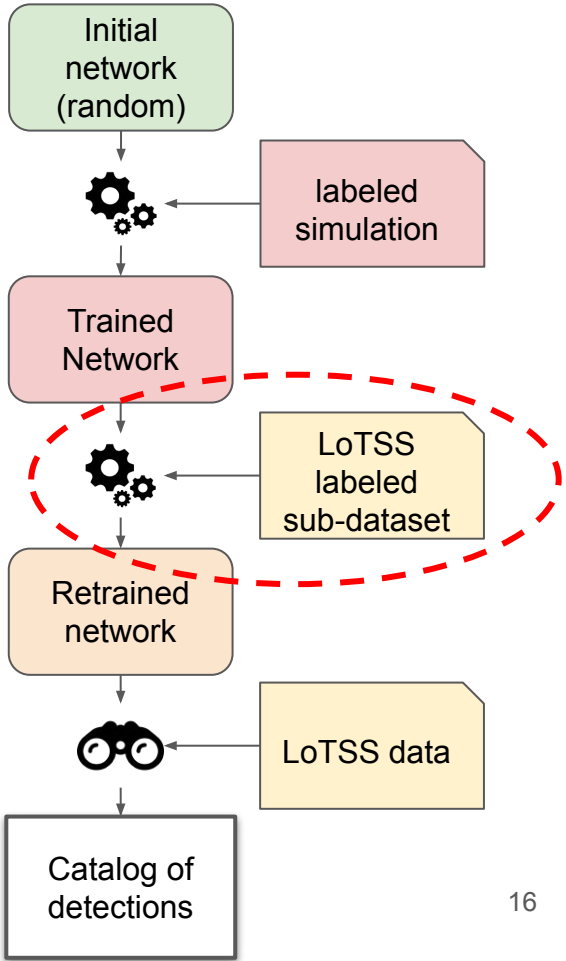


Training *Inference*

Direct application

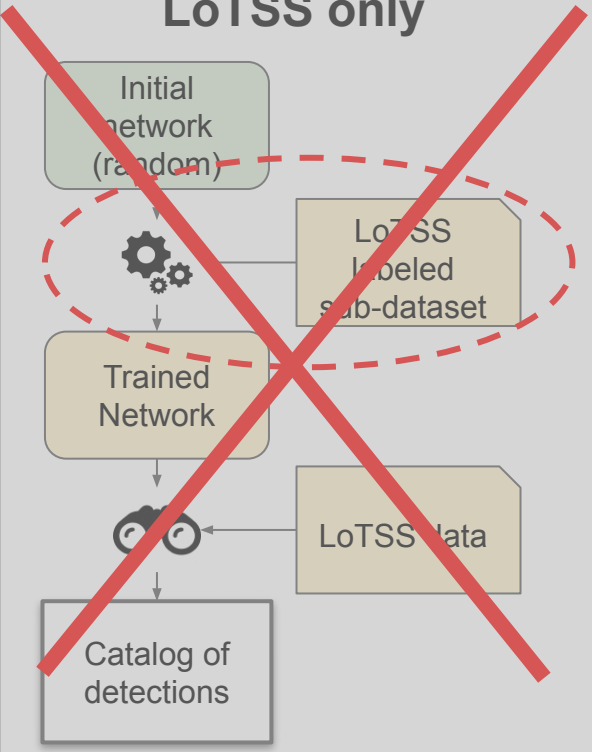


Transfer-Learning



Approaches

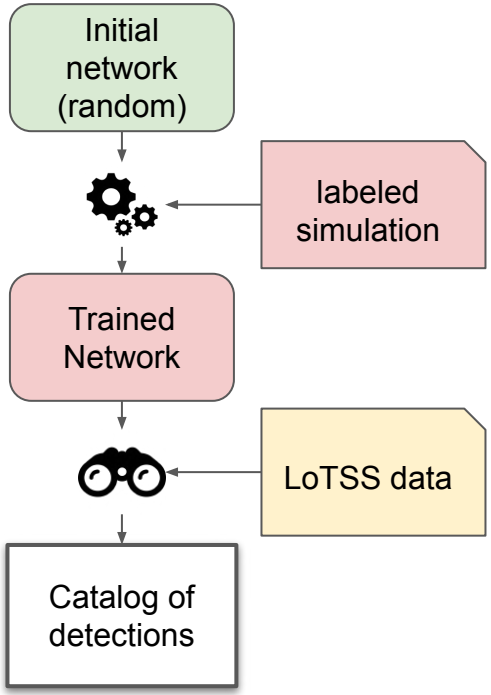
LoTSS only



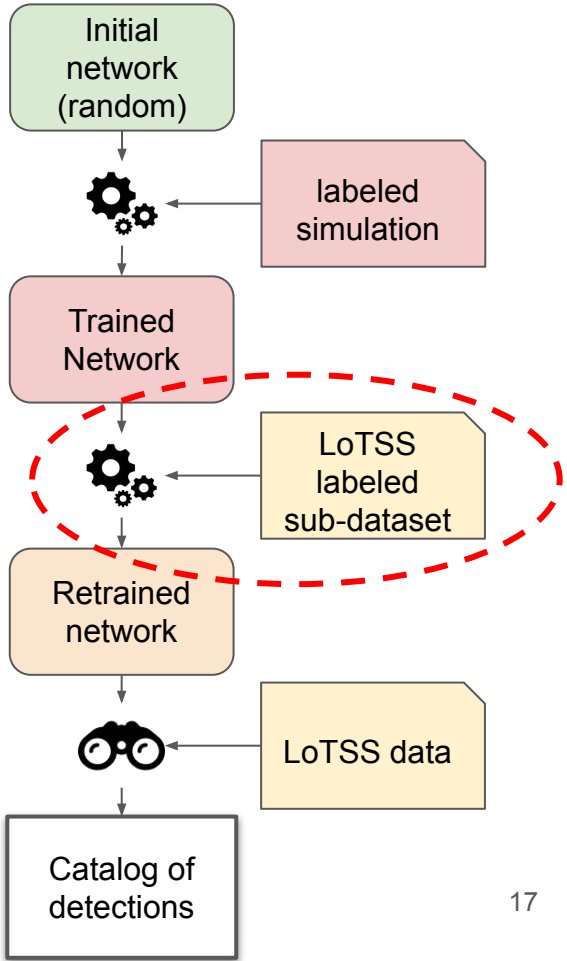
Training Inference

I)

Direct application



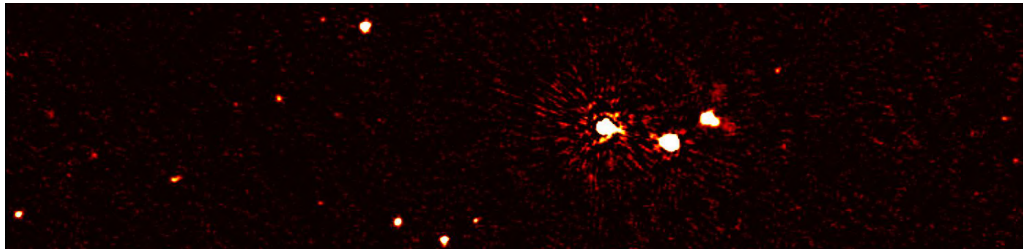
II) Transfer-Learning



Application of the method to LoTSS data

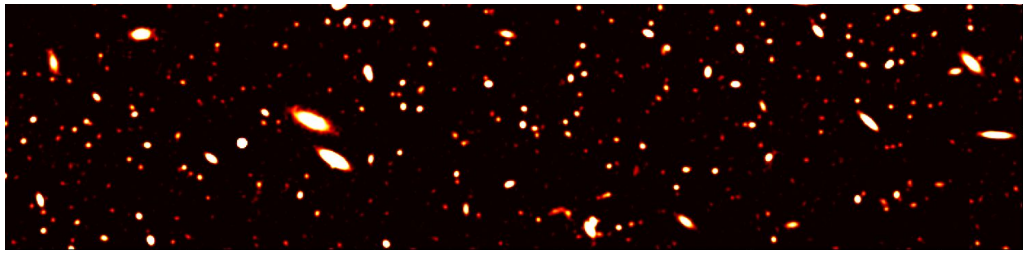
The inference data must match the training data: is it the case?

Similar?



LoTSS DR2
(Observational)
144 MHz

subfields of 150x700 pix



Simulated data
(SKAO SDC1)
560 MHz

Similarities:

- Same point-like sources
- Luminosity profiles
- Blending

Dissimilarity:

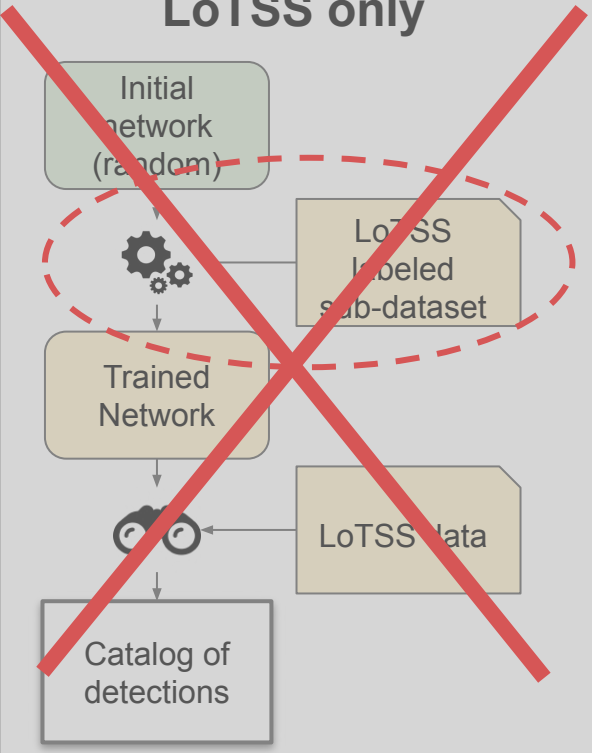
- Resolution
- Pixels dynamics/Sensitivity
- Morphological diversity
- Instrumental specificities

Similar enough, **but requiere**:

- Match **the pixel dynamics**
- Match **the sampling**

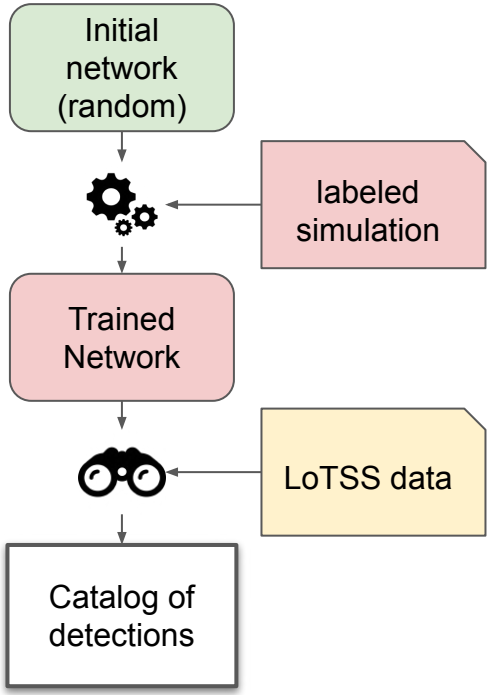
Approaches

LoTSS only

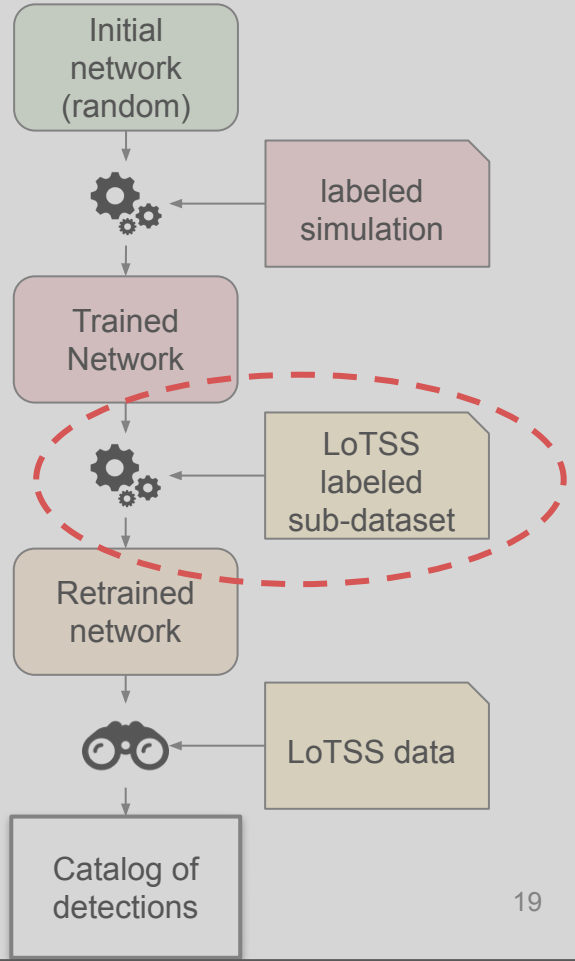


Training Inference

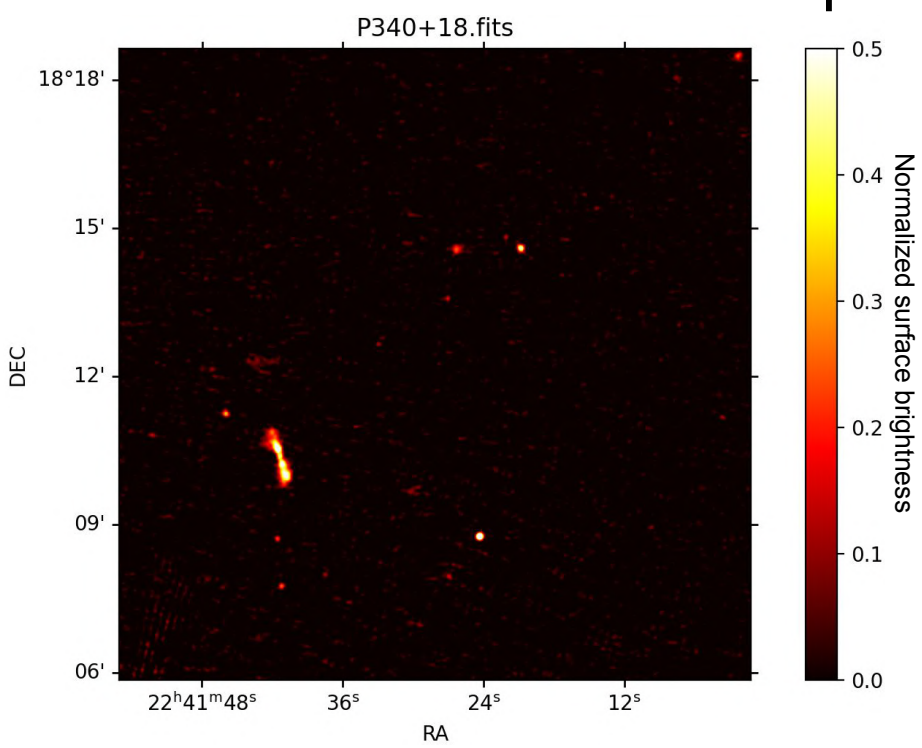
I) Direct application



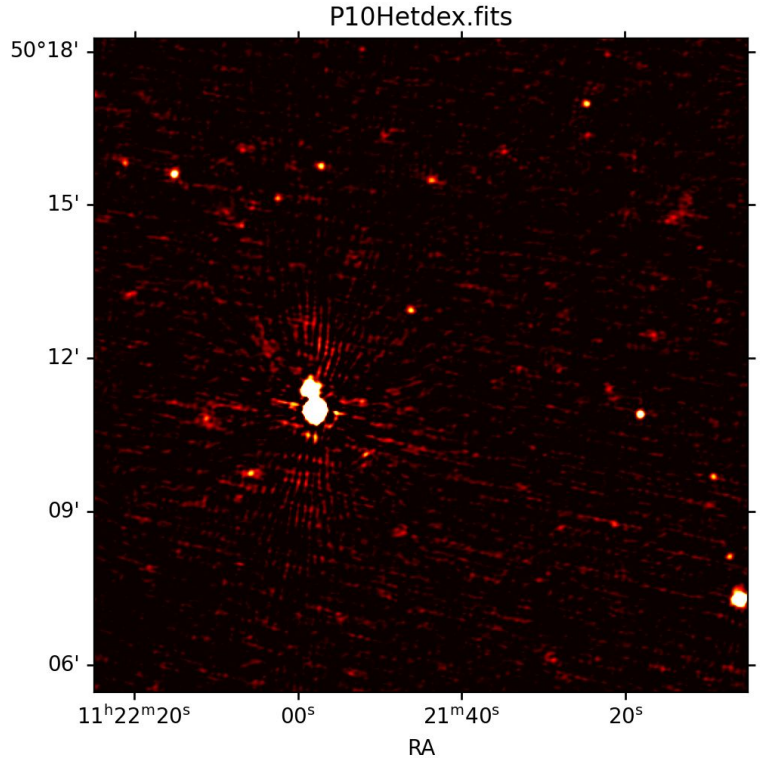
II) Transfer-Learning



Exemple fields



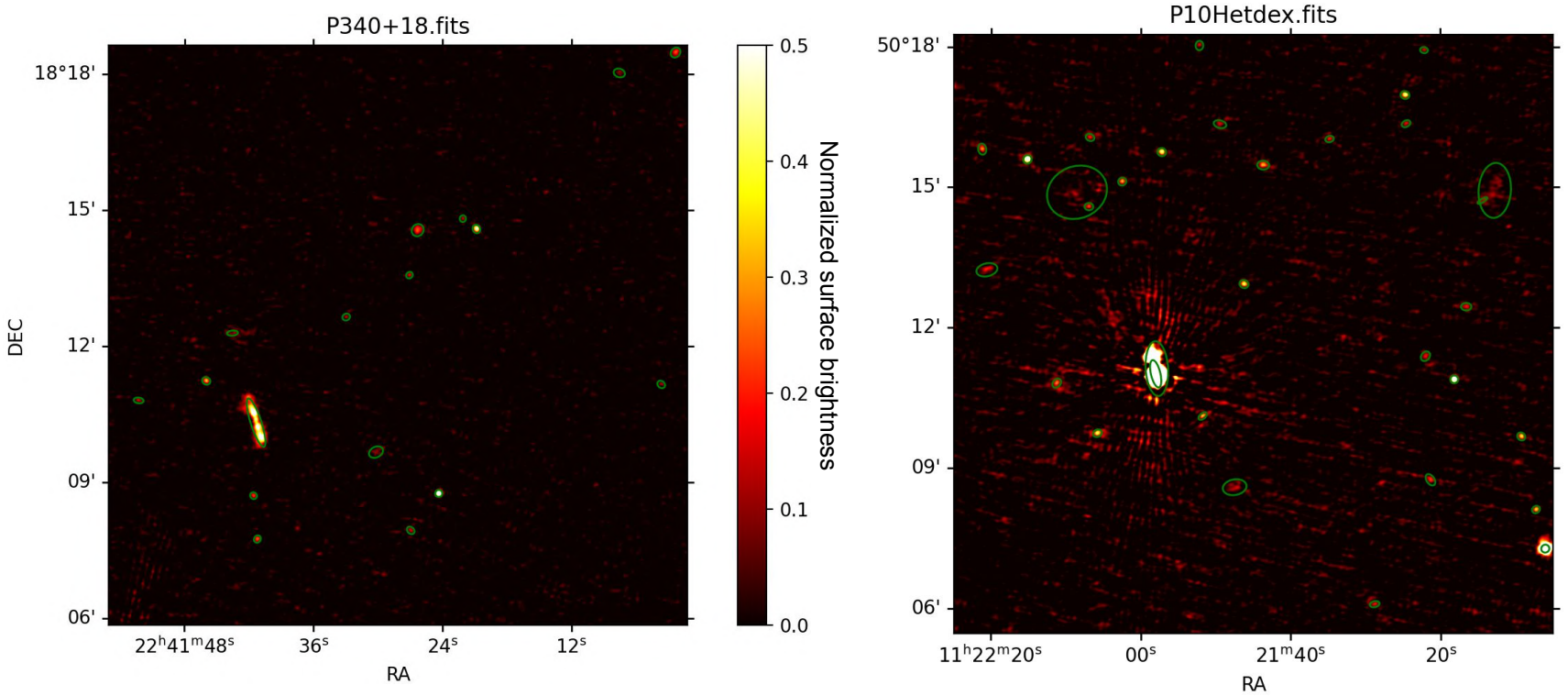
Field **without artifact**



Field **with artifacts**

Object of interest: Point-like sources; Extended sources; Artifacts around bright sources; Blending; Other artifacts

Evaluation on reference



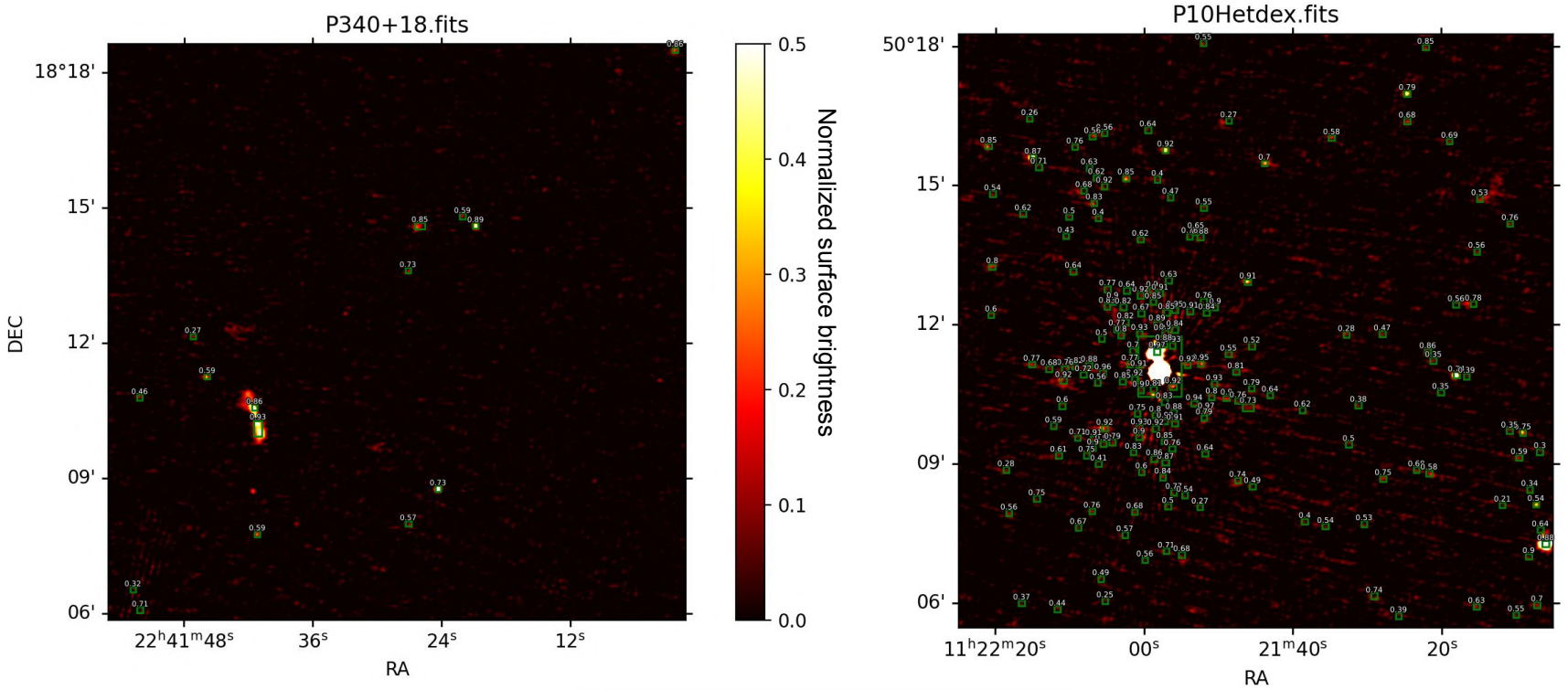
Reference: LoTSS DR2 (Shimwell et al. 2022)

Methode: PyBDSF

Recall = *N_match* / *N_ref*

Precision = *N_match* / *N_test*

Results: “Direct” approach



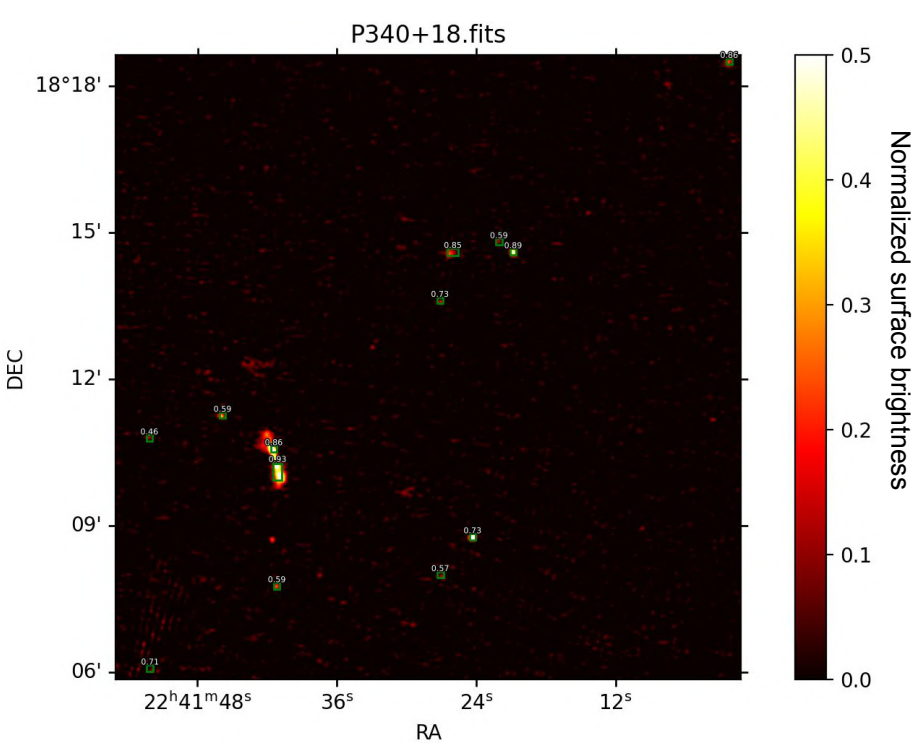
~50% Recall ; ~20% Precision

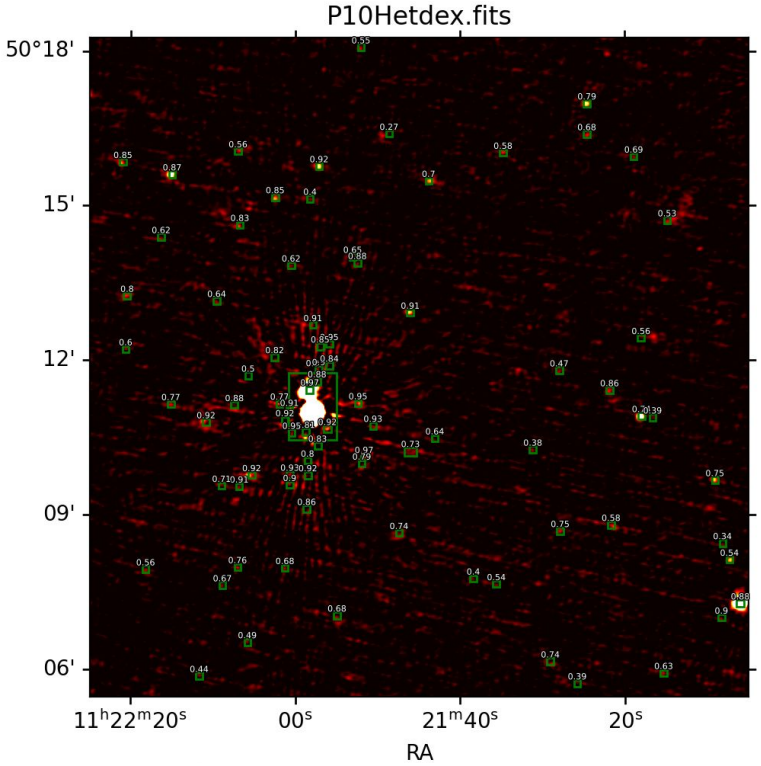
Too many false detections

XX

Detection and associated probability

Results: “Direct” approach + post-process





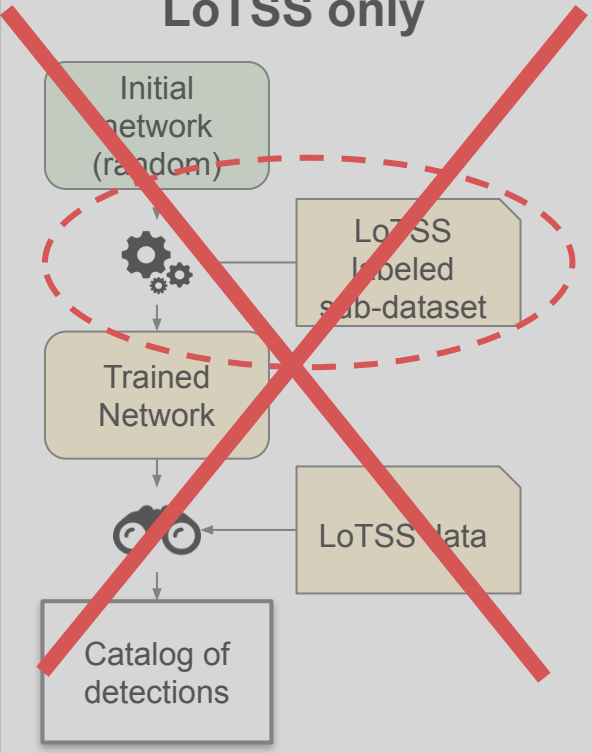
~45% Recall ; ~30% Precision

Less false detection
We reached the limit of improvement

XX
Detection and associated probability

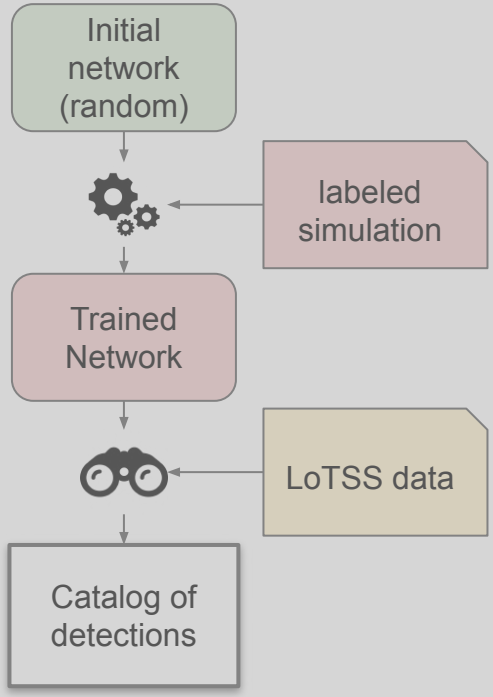
Approaches

LoTSS only

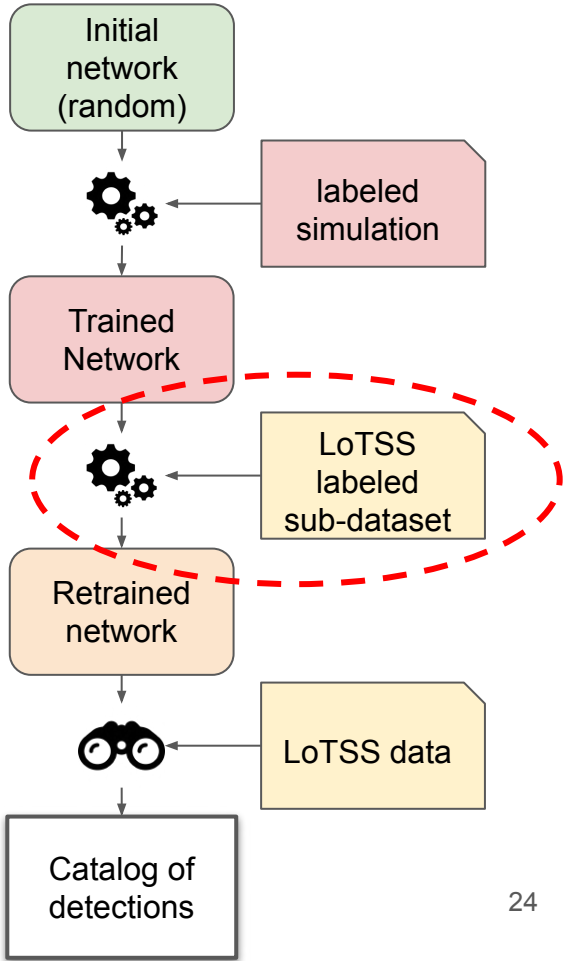


Training Inference

I) Direct application



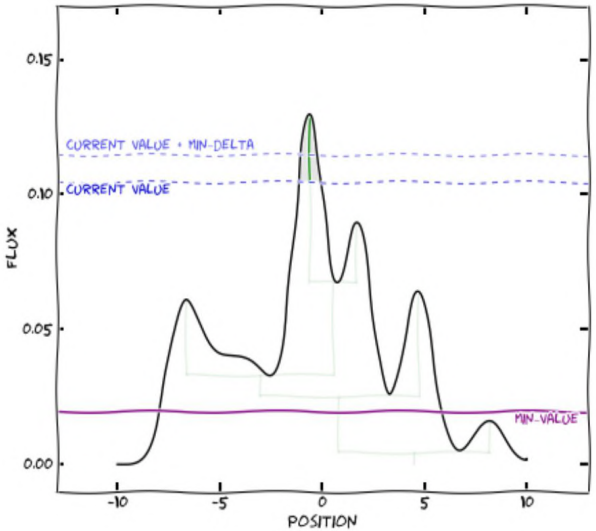
II) Transfer-Learning



Method for building the training set

Catalog optimized for ML detection ≠ existing catalogs today

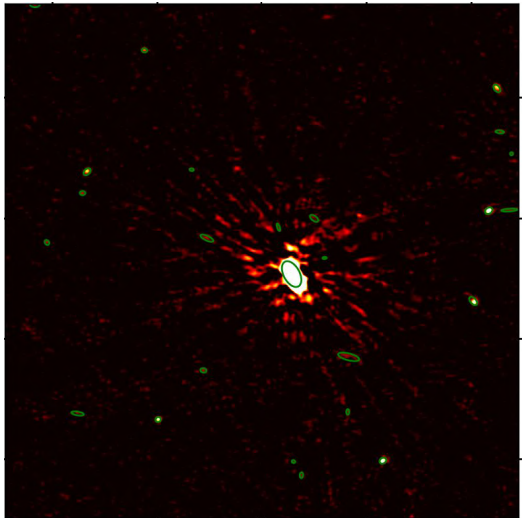
Choice of method: **Astrodendro**



Credit: https://dendrograms.readthedocs.io

Labeling process

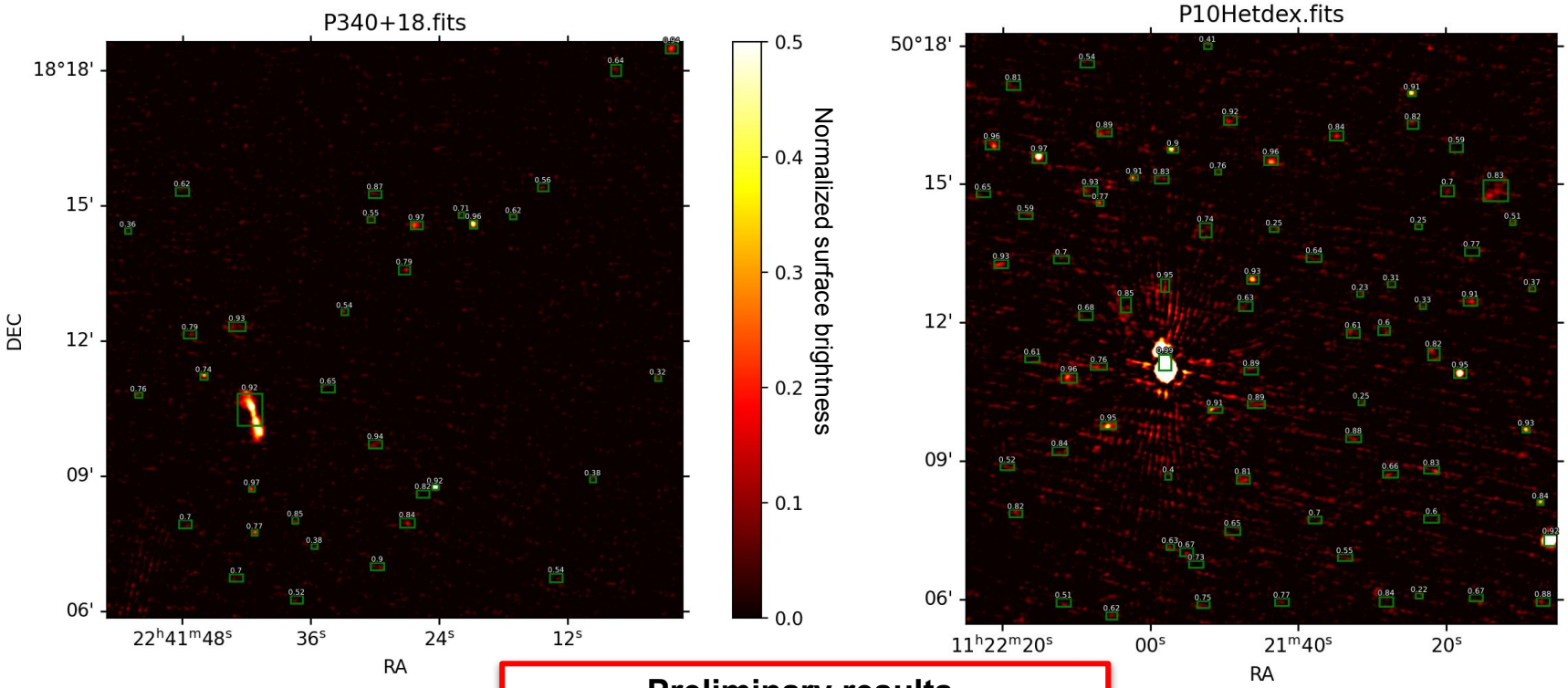
1) Detection in a subfield
2) Filtering after detection
3) IR / Optical counterparts



example of the training area
13x13 arcmin²

By combining classical detection method with counterparts:
We manage to construct a reliable training dataset

Preliminary results: “Transfer-Learning” approach



~90% Recall ; ~50% Precision

Preliminary results, possible enhancement:

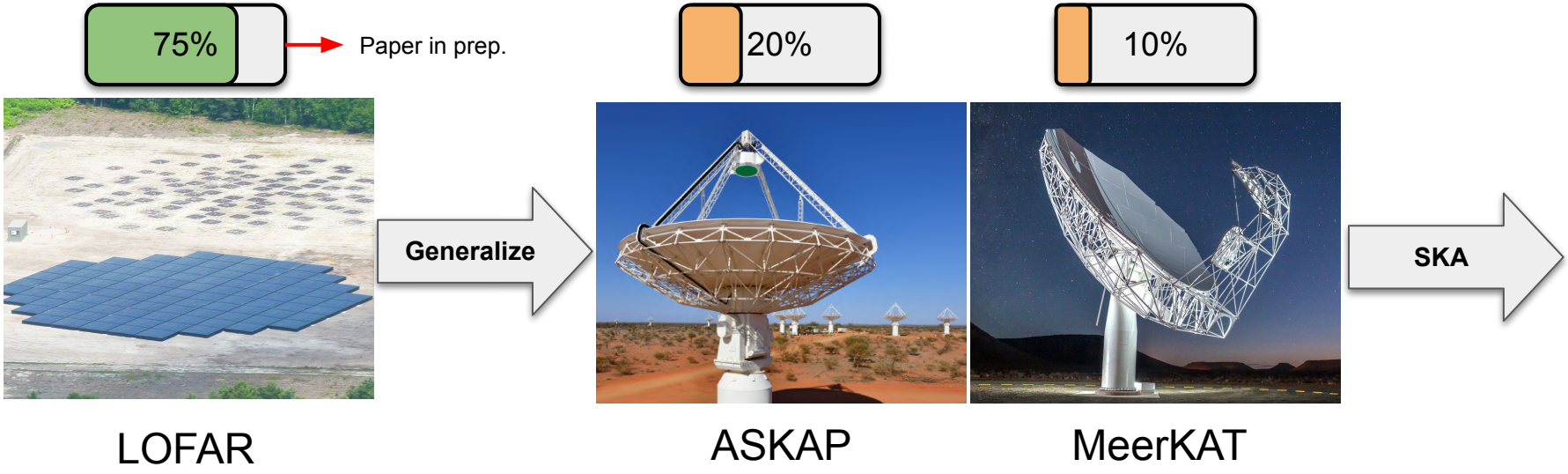
- Improving training dataset
- Post-processing

XX

Detection and associated probability

Prospects

1. Generalize the approach to a set of continuum and HI surveys in preparation for SKA



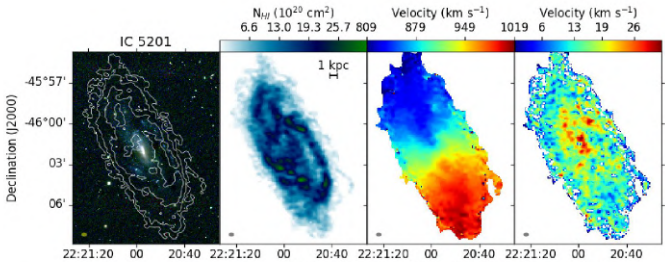
Auxiliary projects

- Improvement of the simulations (T-RECS, ALMA-sim, …)
- Exploration of new network architectures
- Application to non-radio surveys (JWST, Euclid, etc.) and combine the information

Prospects

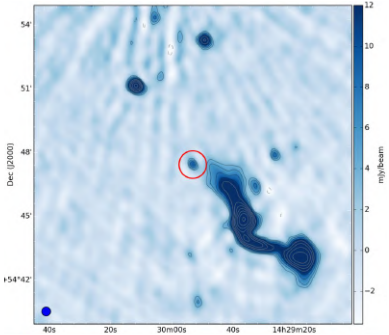
2. Establish statistics on detections

Evolution and property of galaxies



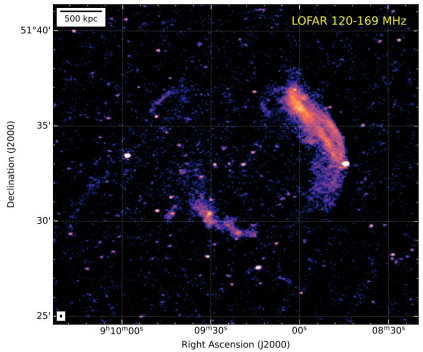
WALLABY, Kleiner et al. 2019

Radio galaxies at z > 6



LoTSS, Shimwell et al. 2017

Galaxy clusters



Rajpurohit et al. 2024

Rare object detection

Lochner and Bassett 2020