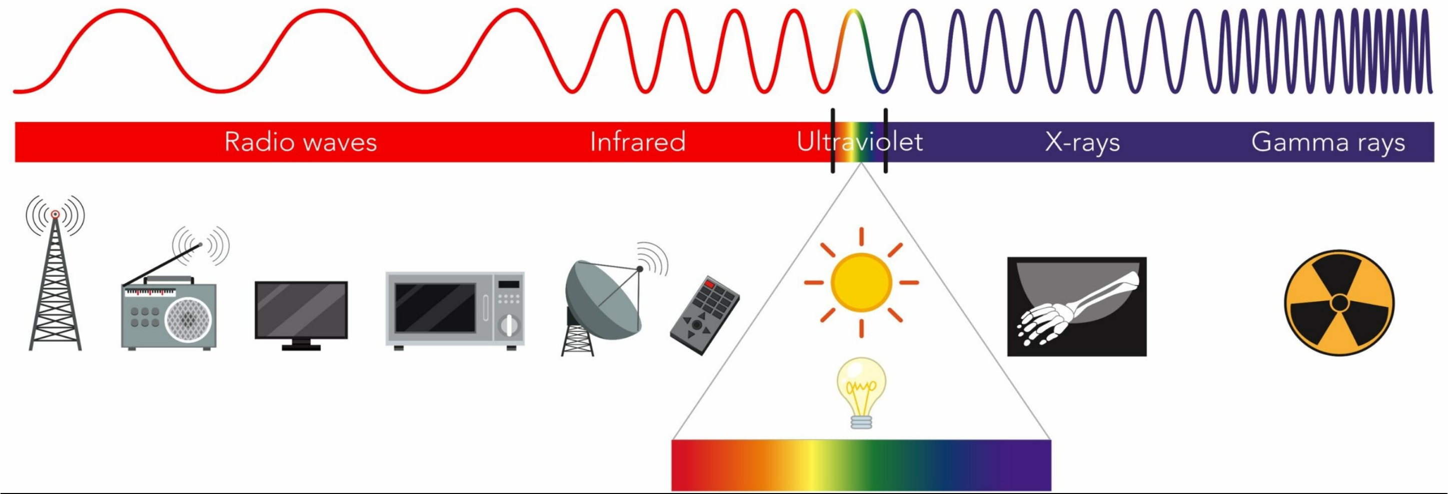
Galaxy detection with deep learning in radio-astronomical datasets

David Cornu*, P. Salomé, B. Semelin,
X. Lu, S. Aicardi, A. Marchal, J. Freundlich,
G. Sainton, F. Mertens, F. Combes, C. Tasse

LERMA, Observatoire de Paris, PSL

AISSAI, Toulouse, 2024

Galaxy observation across the EM spectrum



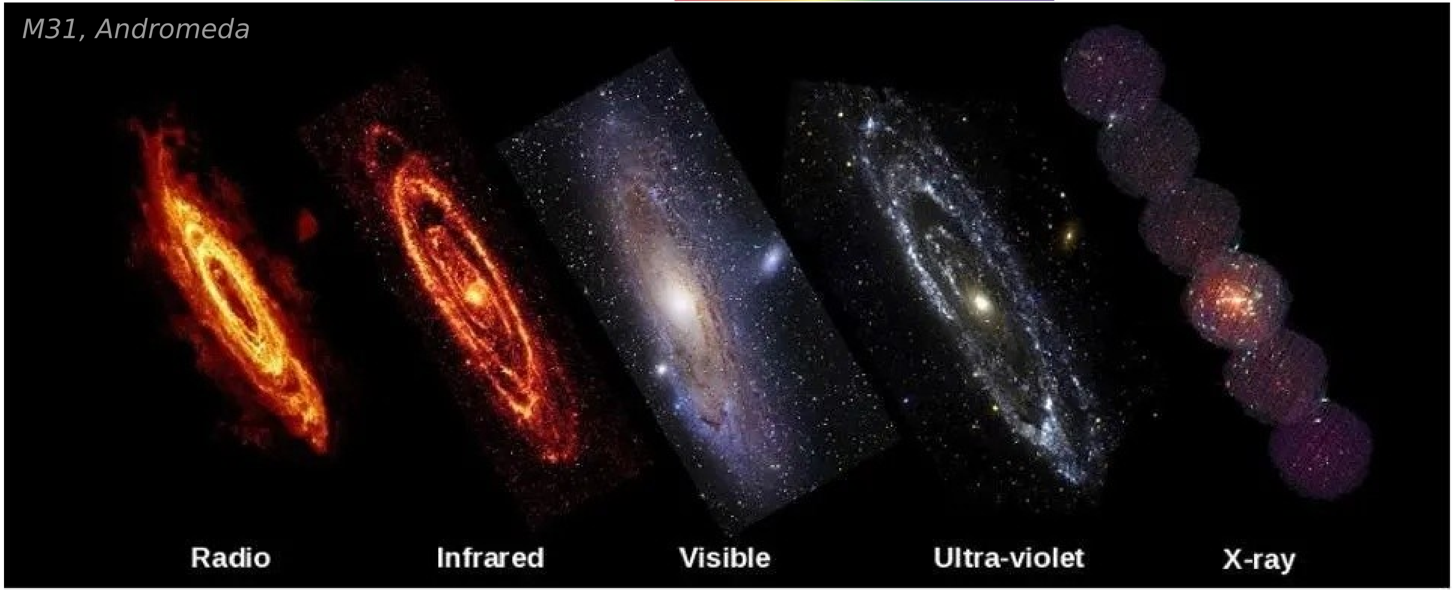
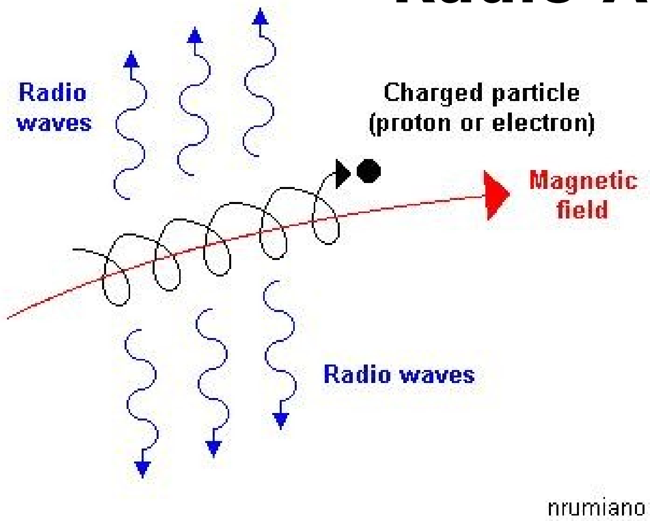


Image from Planck mission team/NASA/ESA

Radio-Astronomy: radio emission

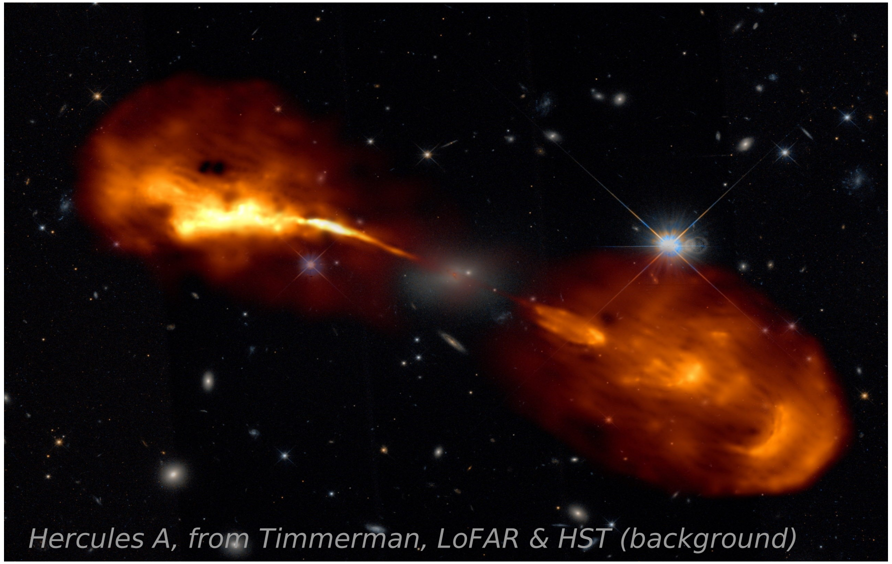


Continuum emission, mainly synchrotron radiation
Induced by the acceleration of charged particles in a magnetic field.
Continuous over a relatively large wavelength window.

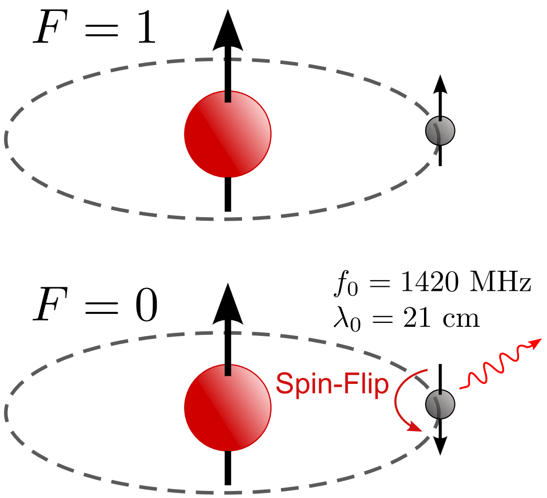
Two main types of continuum-emitting galaxies:

1) Star-forming galaxies with emission from the interstellar medium
2) Active Galactic Nuclei (AGN) with emission from relativistic jets


M51 VLA, NRAO/AUI/NSF; HST composite by B. Saxton

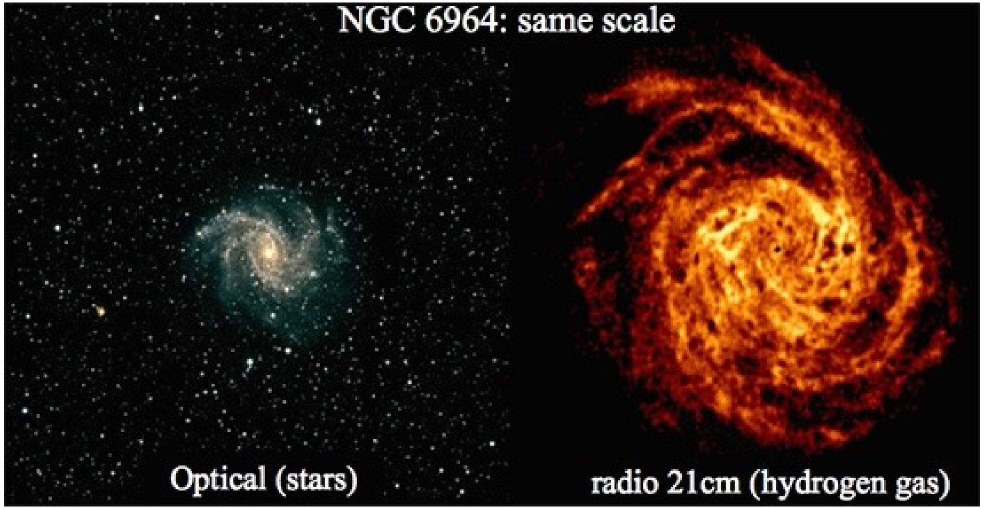

Hercules A, from Timmerman, LoFAR & HST (background)

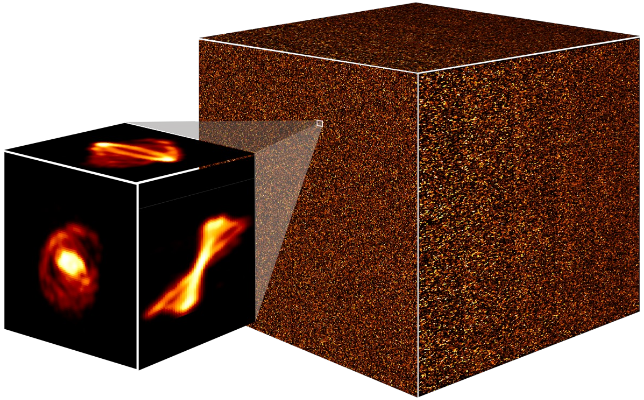
Radio-Astronomy: radio emission



"21cm" or "HI" emission
Spectral line created by the s1 hyperfine transition (spin-flip) of the hydrogen, with a characteristic emission at a wavelength of 21 cm.

HI observations **are often hyper-spectral,** allowing to reconstruct rotation curves of galaxies, and to create **large HI emission cubes**.





3D HI emission cube
(from SDC2, Hartley et al, 2023)

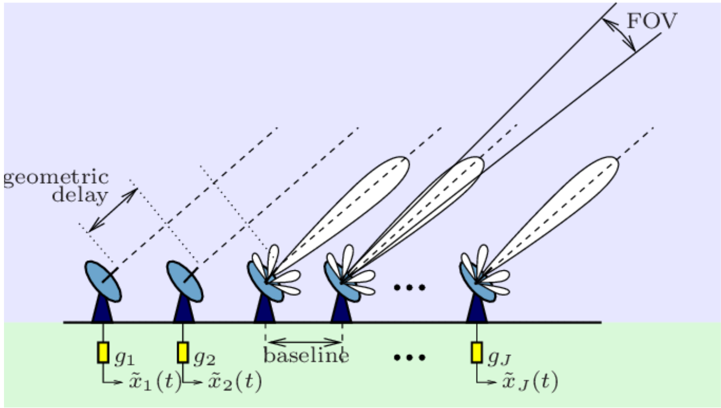
Radio-Astronomy: observation

Astronomical radio-emissions are observed with antennas or arrays of various kinds

Large single dish

Dish array

Multi-pole arrays



Radio Interferometry:

→ Using multiple antennas to emulate a single receiver
virtual telescope size = maximum distance between antennas.

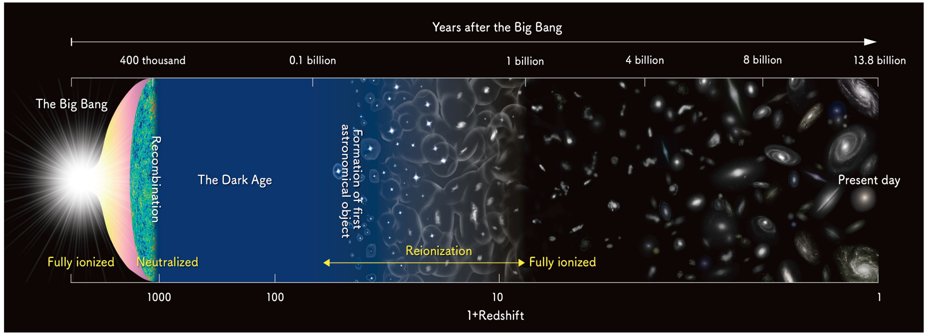
Allow for a strong increase in resolution and sensitivity at the cost of **complex computing for image reconstruction**.

→ In addition, current radio astronomical facilities already produce PB scale databases.

The Square Kilometer Array

Future largest radio-telescope

Evolution of the universe and astrophysical objects

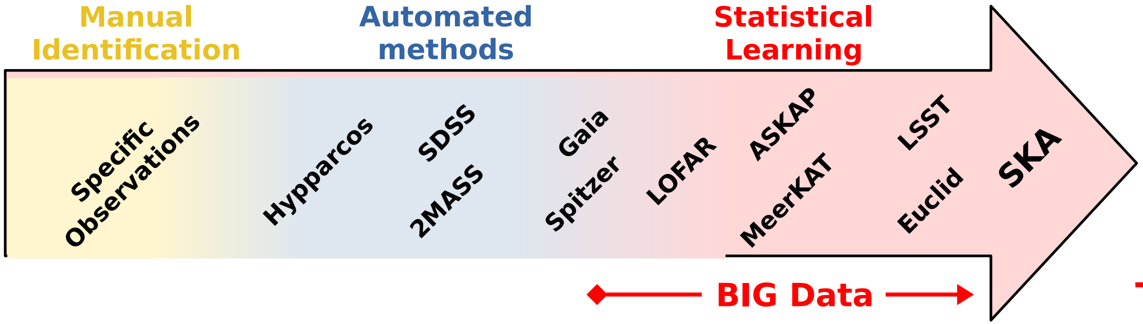


Construction phase started ! ETA : ~2028

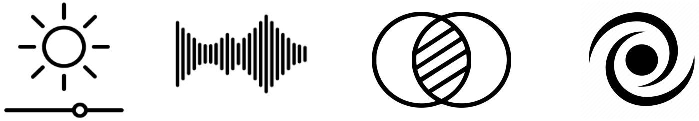
Big Data

~700 PB / year (stored)

~1.5 millions 500 GB HDD / year



Complex data



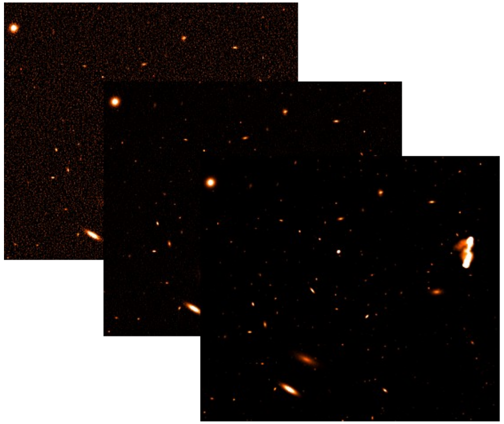
→ Requires new innovative methods

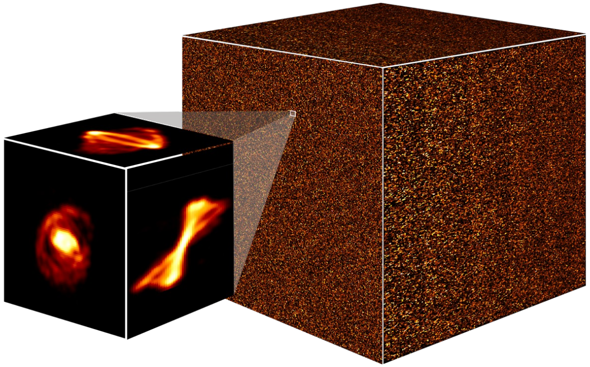
SKAO Science Data Challenges (SDCs)

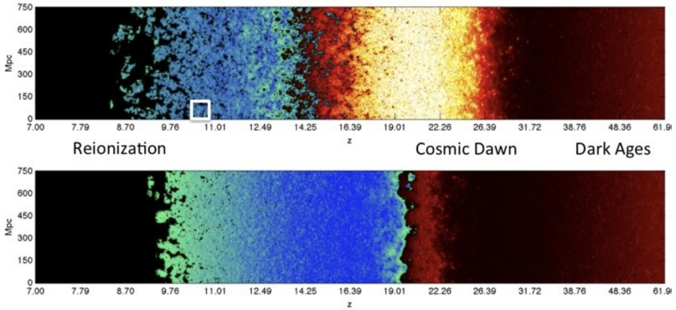
Simulated dataset that should resemble typical SKA data products

Objective: prepare astronomers, stimulate the creation of new data analysis pipeline

Source detection and characterization







SDC1: Continuum 2D images
3 integration times x 3 bands
Each image = 4 GB
From Dec 2018 to April 2019

SDC2: Hyperspectral cube of HI emission
Full cube = 1 TB
From Feb 2021 to July 2021

SDC3: 21 cm emission
Visibility and Image
Full size ~ 17 TB
EoR Focused, 2023-2025

MINERVA - MachINe lEarning for Radioastronomy at obserVatoire de Paris

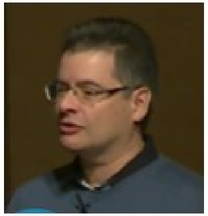
Officially ended on December 2023

Main research fields

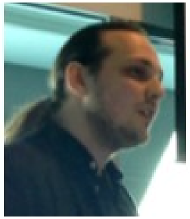
- **Cosmic Dawn / EoR**
- **Transients phenomenon**
- **Large Radio-survey mining**

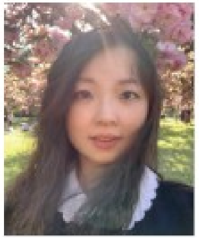
SKA Science Data Challenge 2 (SDC2) team using ML methods

From the MINERVA group

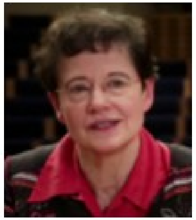

B. Semelin

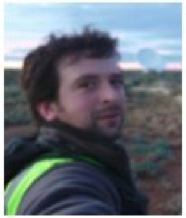

P. Salomé


D.Cornu


X. Lu


S. Aicardi


F. Combes


C. Tasse

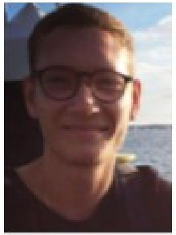

F. Mertens


G. Sainton

\+

External collaborators


J. Freundlich

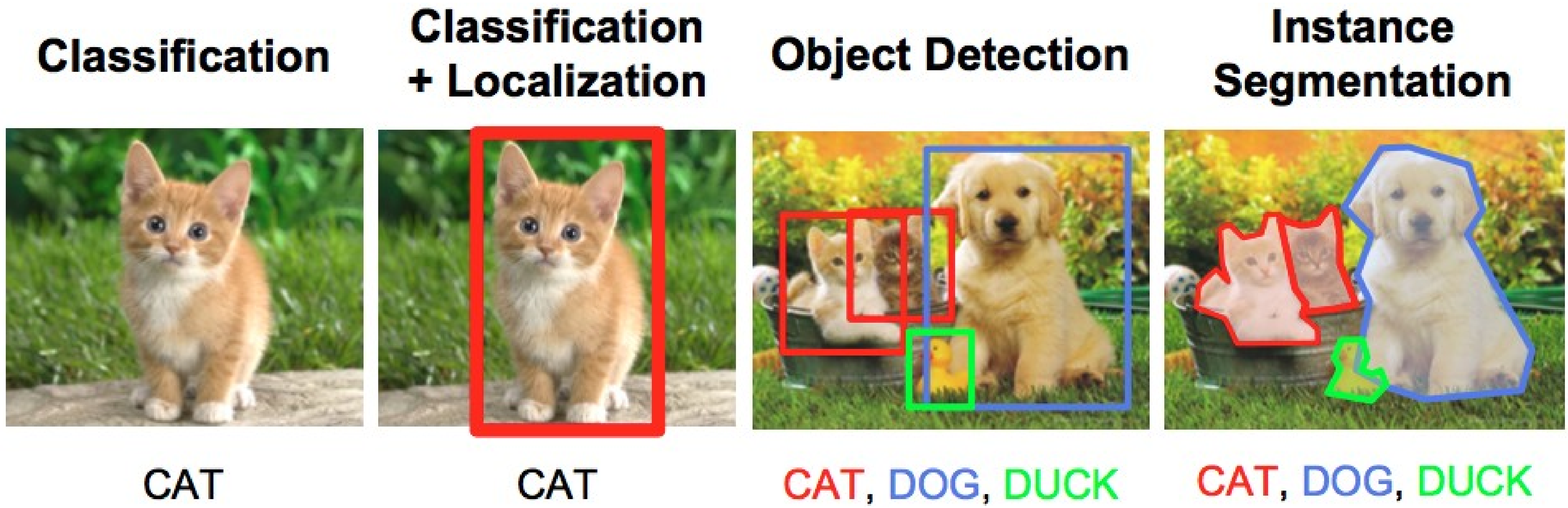

A. Marchal

Students


A. Anthore

Types of object detection in images

**Image from Stanford Deep Learning course cs224*

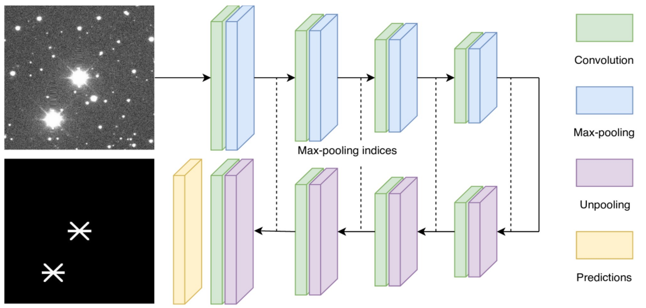


Deep Learning methods for object detection

Segmentation-based

Methods: U-net, mask R-CNN, ...

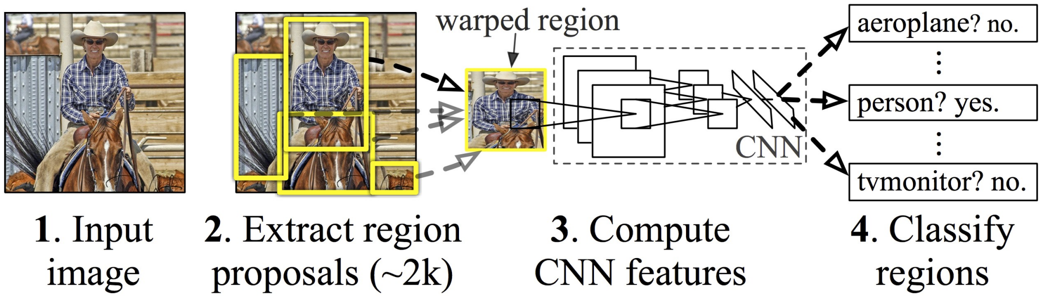
Pros: segmentation maps, shallow latent space, ...



Region-based

Methods: R-CNN (Fast and Faster), SPP-net, ...

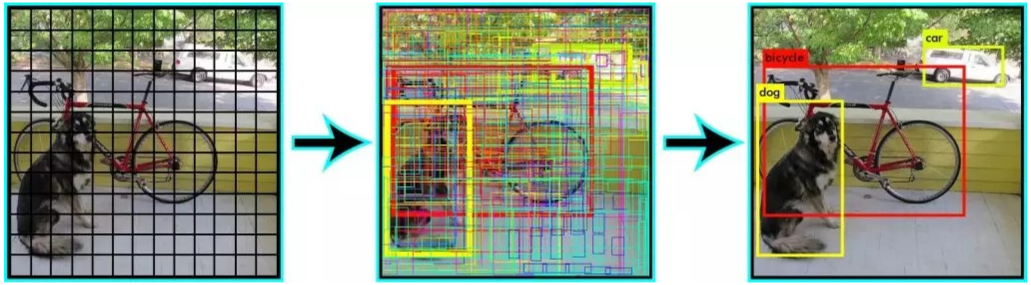
Pros: Best accuracy, ...



Regression-based

Methods: SSD (Single Shot Detector), **YOLO (You Only Look Once)**, ...

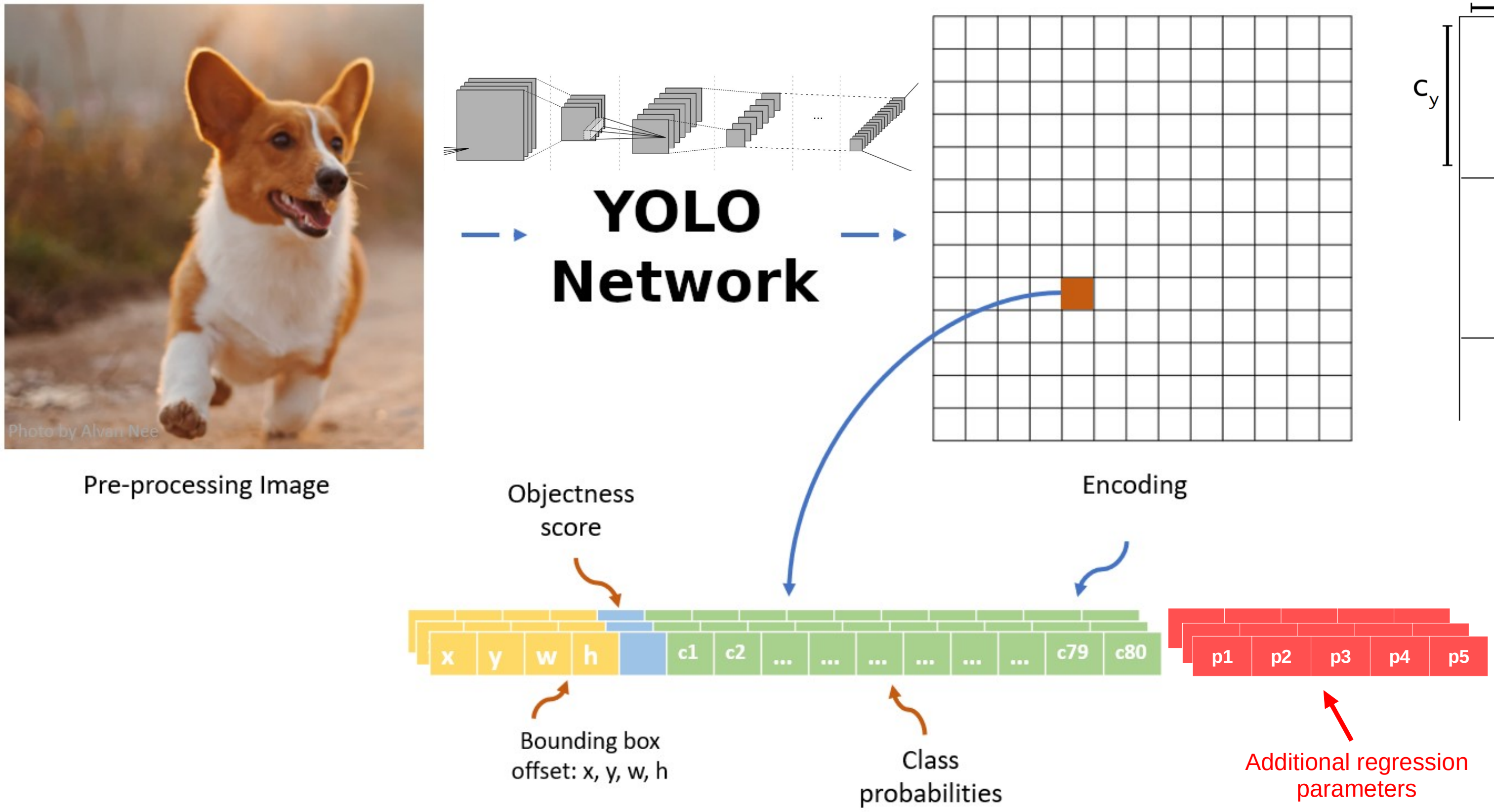
Pros: Very Fast, straightforward architecture,...

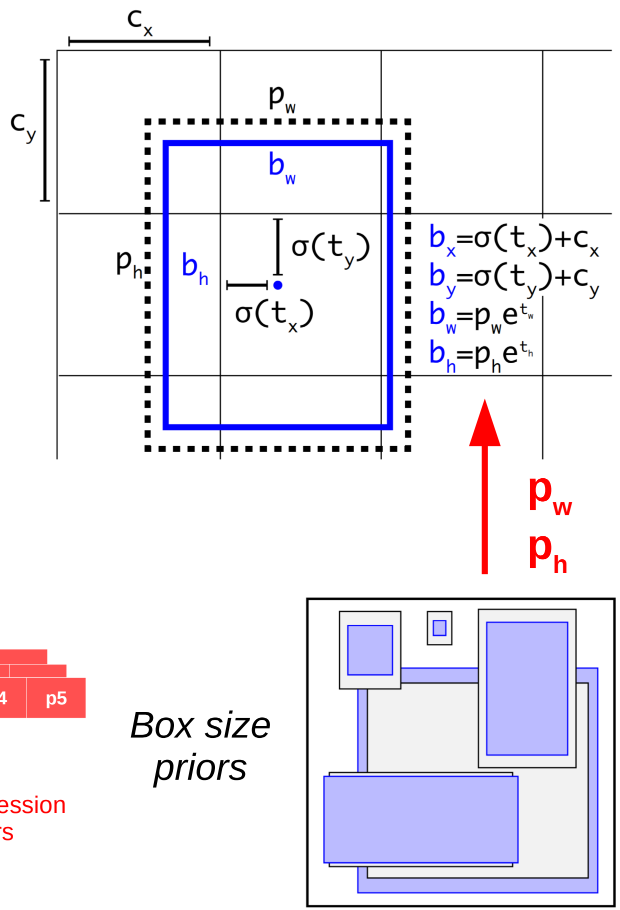


You Only Look Once – YOLO !

Originally introduced in Redmon et al. 2015 (V1), 2016 (V2), 2018 (V3)

**Images from blog post and Redmon papers*

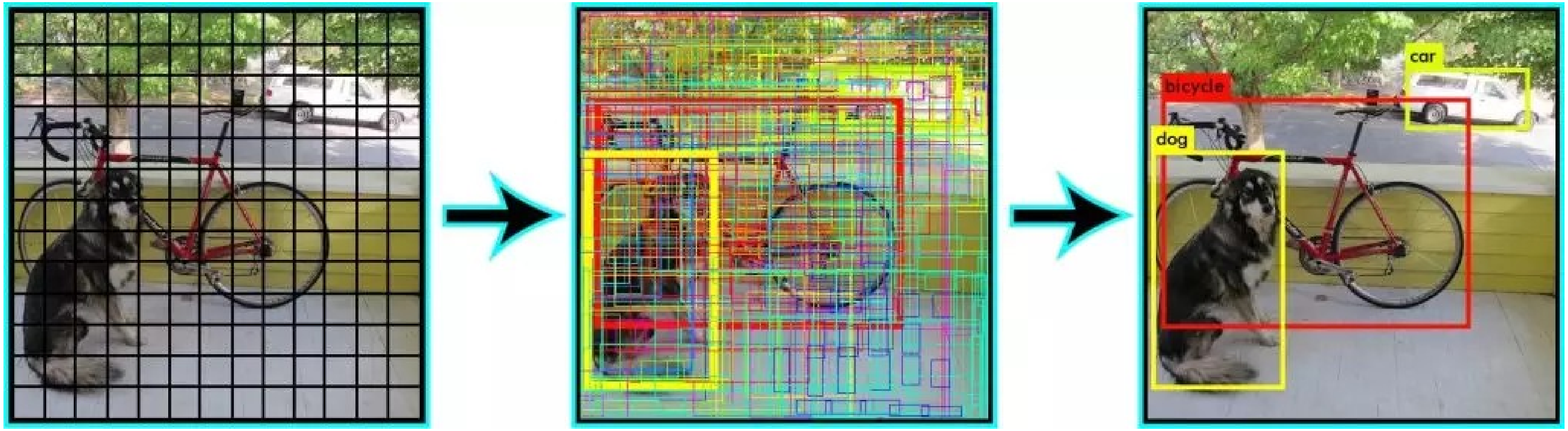




The last layer is conv. → boxes « share » weights spatially.

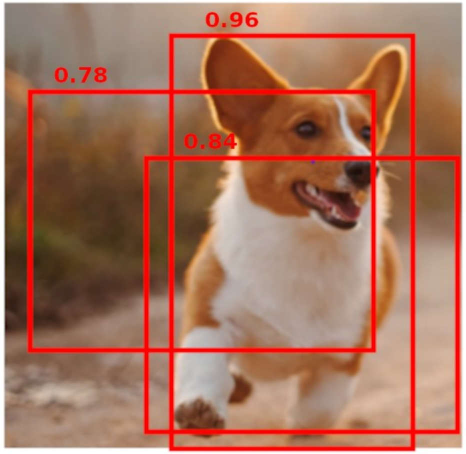
The output is a 3D cube encoding all possible boxes on the output grid.

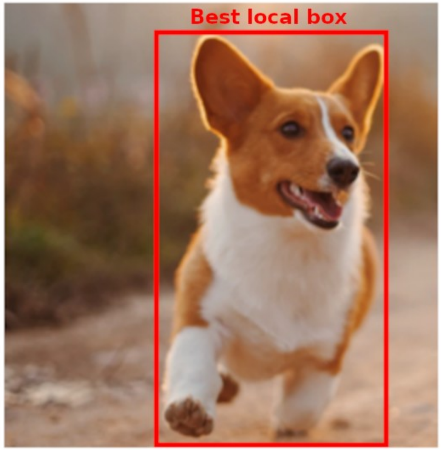
Non Max Suppression



1) Most probable boxes are kept using a threshold in objectness

2) NMS takes the most probable box and removes overlapping ones based on IoU





CIANNA

Convolutional Interactive Artificial Neural Networks by/for Astrophysicists

General purpose framework (Keras, PyTorch, ...)
BUT developed for **astronomical applications**

CPU OpenMP OpenBLAS / GPU NVIDIA CUDA

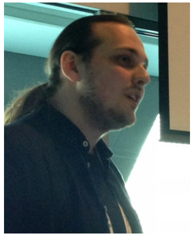
x86 Intel/AMD & ARM **Supports modern TC acceleration**

Full user interface

Successfully deployed on

- **Laptops / Workstation**
- **Local compute servers**
- **Mesocenters**
- **Large computing facilities**

CIANNA dev. team


D. Cornu


G. Sainton

github.com/Deyht/CIANNA
Open source – Apache 2 license

Custom YOLO implementation
(detailed in Cornu et al. 2024)

Activation **Cost** **Association**

- **Supplementary parameters per box**
- **Cascading loss**
- **Custom association process**

→ Matches YOLO V2 accuracy on classical VOC datasets

Application to SKAO SDC1

SKA SDC1 summary paper, Bonaldi et al. 2021

Data:

Large continuum images of the same field

- 3 frequencies: 560 MHz, 1.4 GHz, and 9.2 GHz)
- 3 integration times: 8, 100, and 1000h

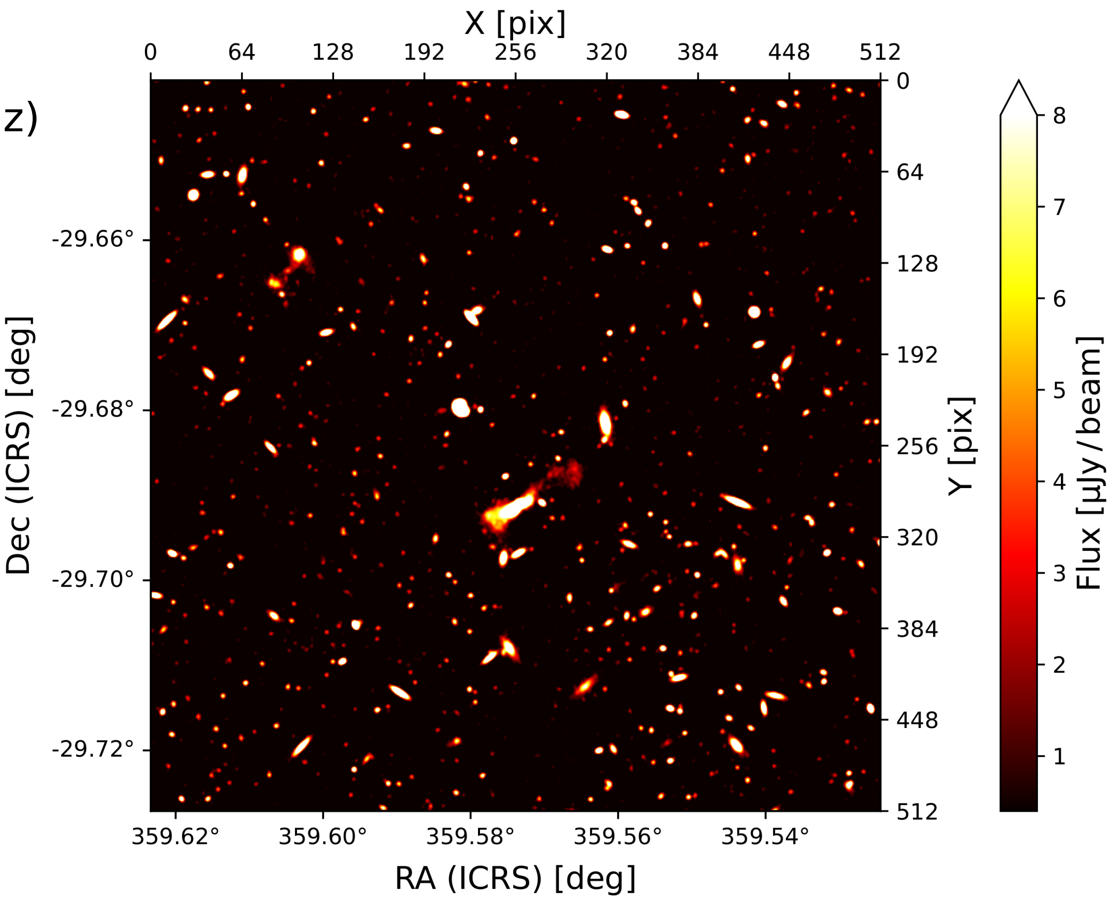
Each image is 32,768 pixel square = 4GB.

A labeled 5% surface fraction is provided for ML methods training !

The challenge:

1. Find the sources (RA, Dec)
2. Characterize each source:
   → (Flux, Bmaj, Bmin, PA, ...)

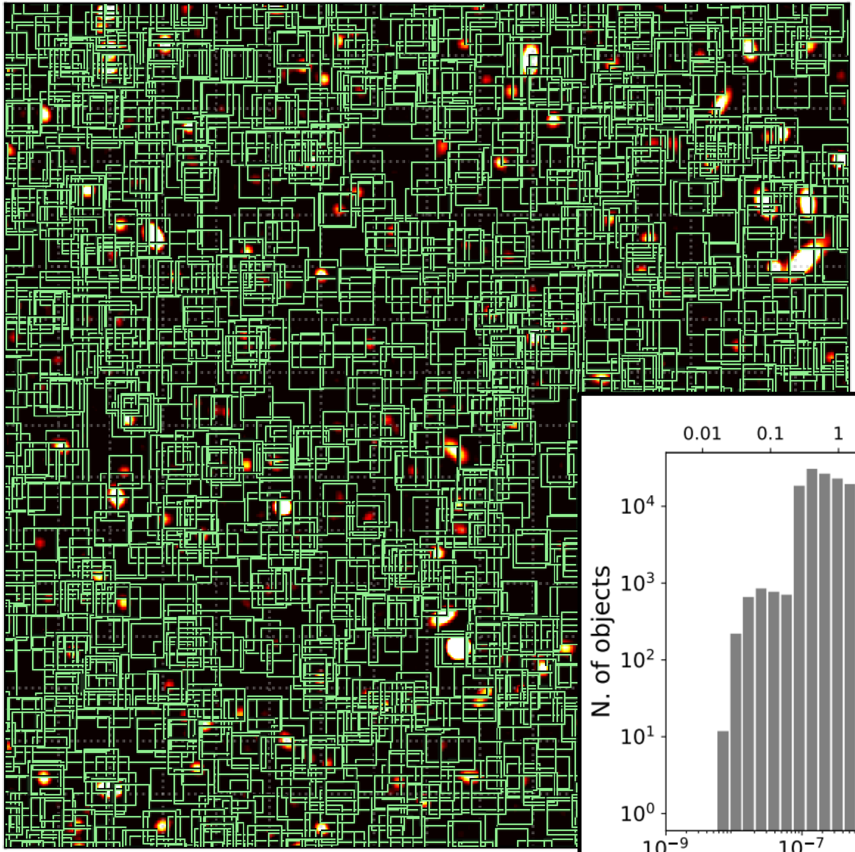
Example sub-field

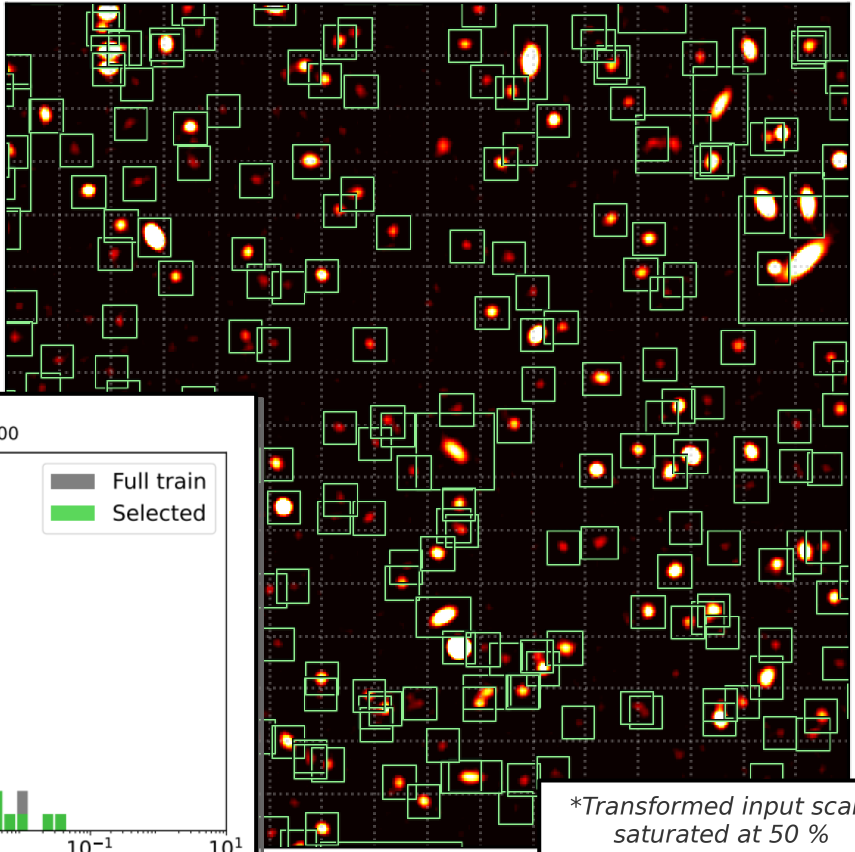


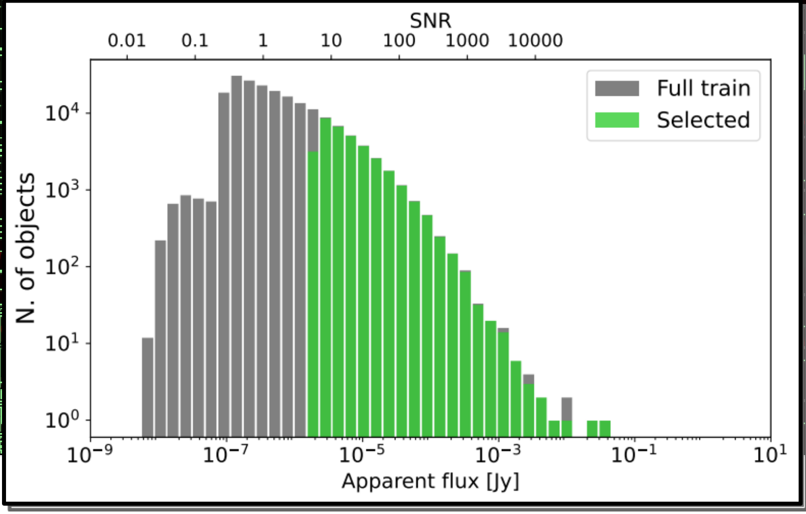
SKA SDC1 took place early 2020. Data from the challenge are still freely accessible on the dedicated web-page.

Training data selection function

- CNN must not be given the task to detect “impossible / invisible” sources!
- Selection based on surface brightness → only ~10% of the labeled catalog remains

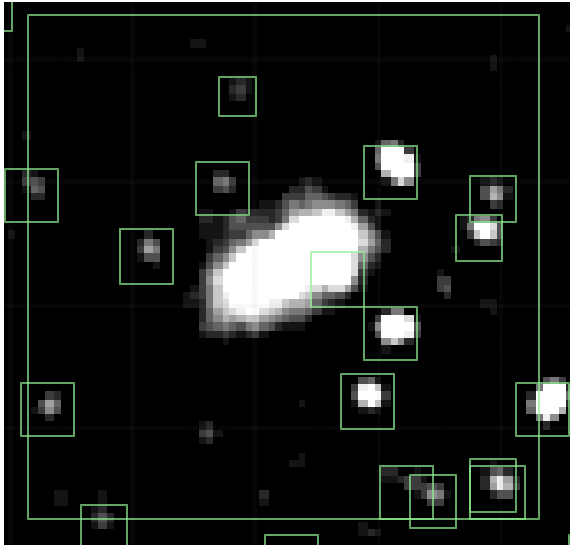






*Transformed input scale saturated at 50 %

Other difficulties and method modifications



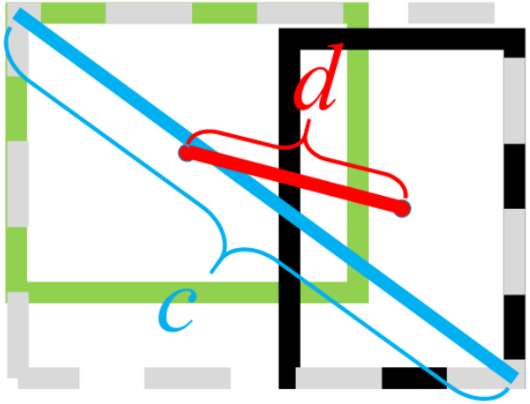
Images are crowded with small sources than can blend

- Smaller architectural reduction factor and adjusted NMS
- The minimum box size is clipped to ∼ beam size
- Multiple identical small size priors are used simultaneously
- Change the YOLO association process to be "prediction aware"

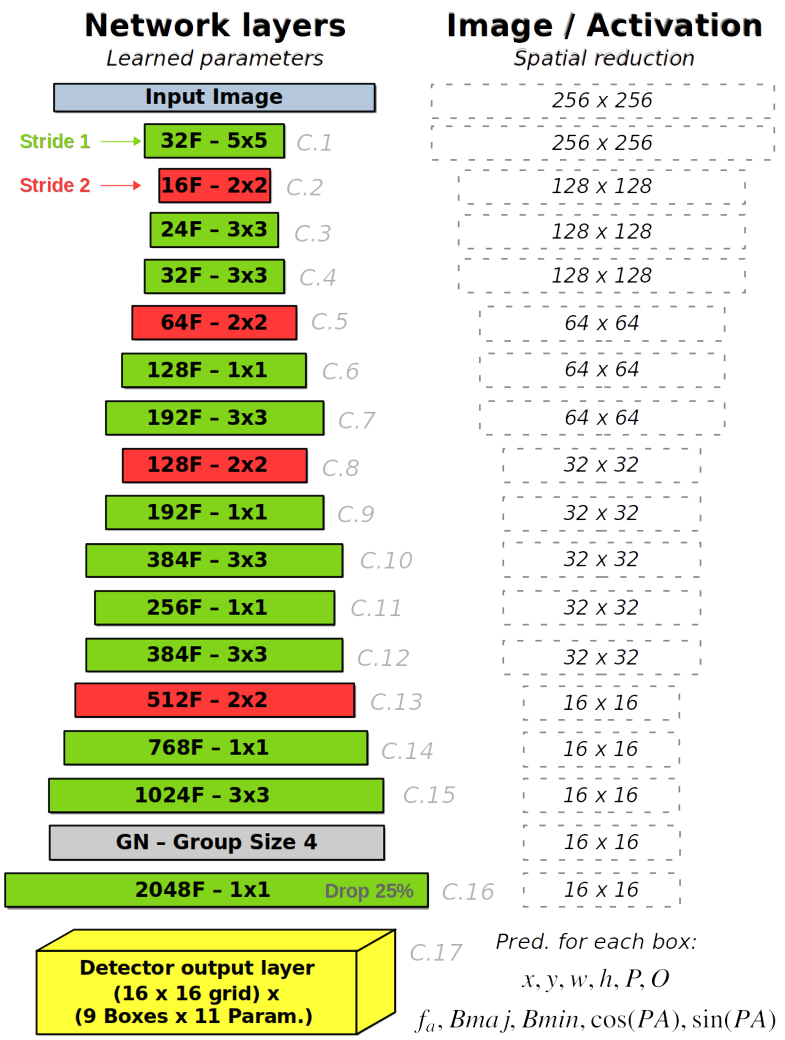
Require extreme positioning accuracy

- The loss function is manually biased for position accuracy
- Change the association metric to a distance aware DIoU

$$DIoU = IoU - \frac{d^2}{c^2}$$



Custom SDC1 model



Architecture:

- **17 conv. layers** → **~13 Million parameters (~50MB)**
  → +8% in score compared to the classical darknet19 backbone
- **9 box priors** ranging from 10 to 32 pixels
- **Modified YOLO** → For each box **5 additional parameters** are predicted: **Flux, Bmaj, Bmin, cos(PA), sin(PA)**
- No class prediction

Training the network using IANNA

- **256x256** cutouts are randomly selected in the training area (54 MB)
- ~ **34000 sources** in the selected training catalog
- Data are augmented based on cutout position and flips

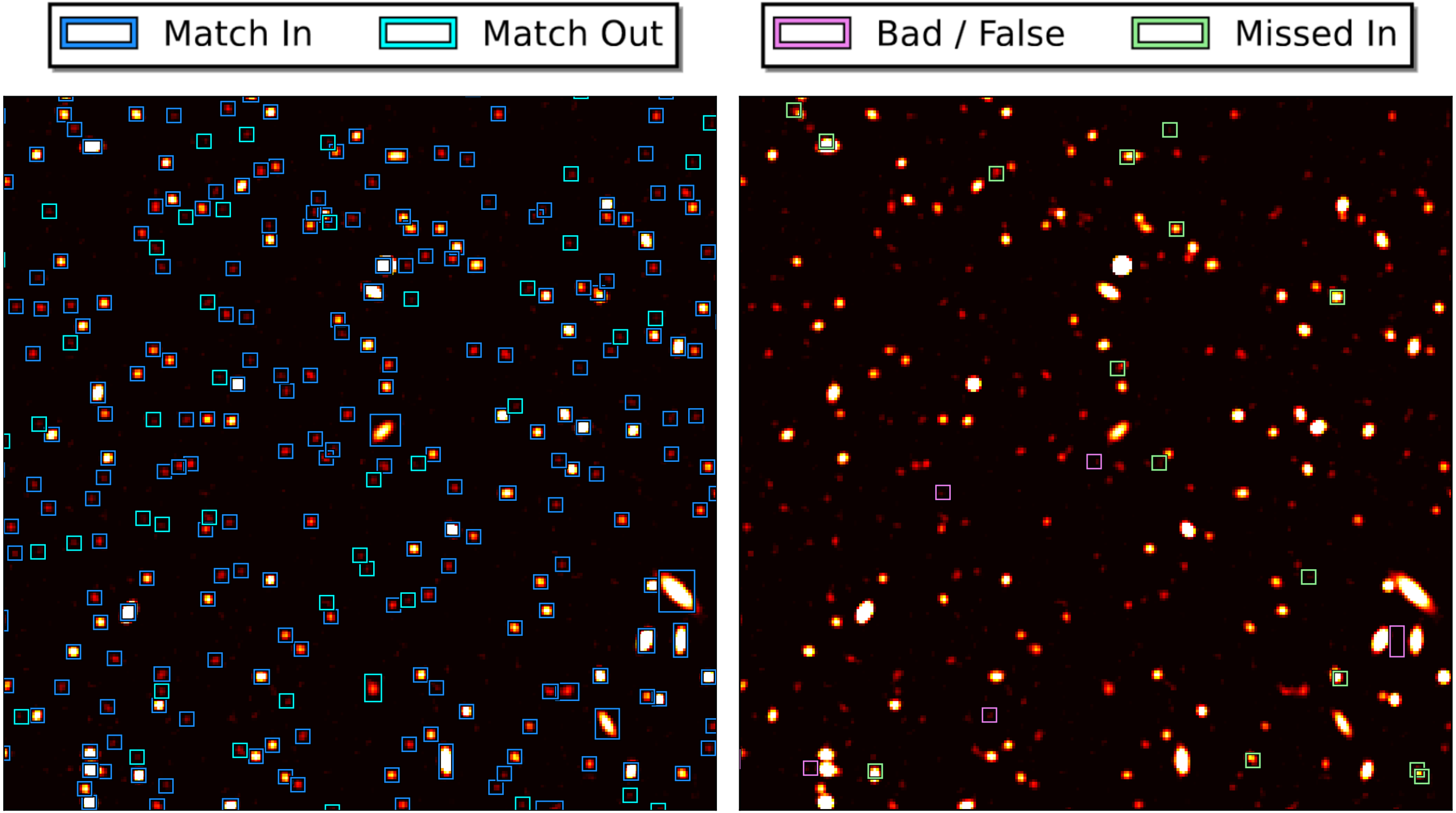
Using a single RTX 4090 GPU, training time is **~ 4 hours**

Inference:

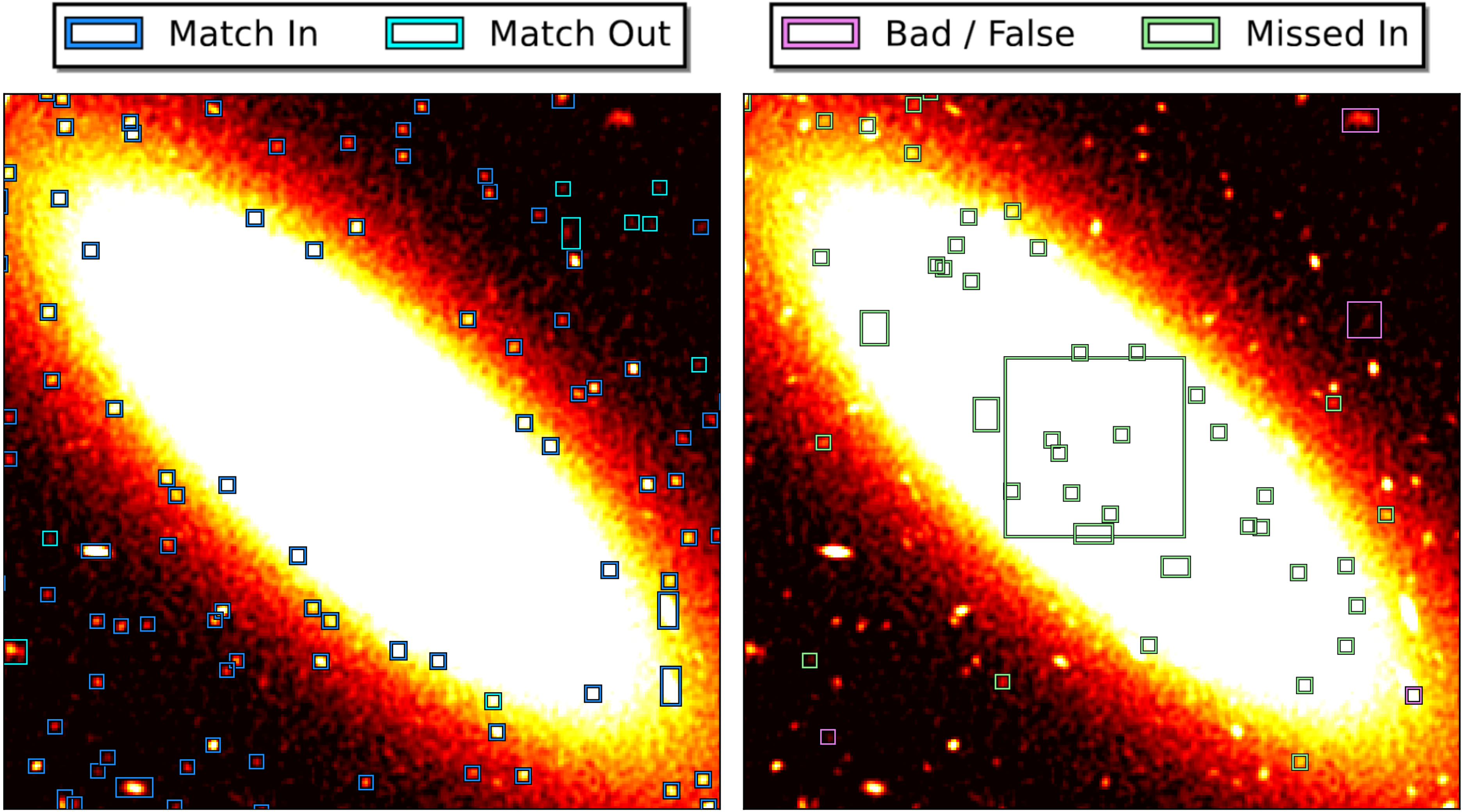
- The full SKA SDC1 image is split in 512x512 regions with an overlap of 32 pixels, → **~4500 images**
- Overlapping regions are filtered with a dedicated secondary NMS

The full inference in FP16-TC takes ~8 sec → 130 Mpix/s

Detection example fields



Detection example fields



Results comparison

Based on Bonaldi et al. 2021 + submitted catalogs

$$s_i = \frac{1}{7}\sum_{j}^{7} s_i^j \qquad M_s = \sum_{i}^{N_{match}} s_i - N_{false}$$

Average per-source score

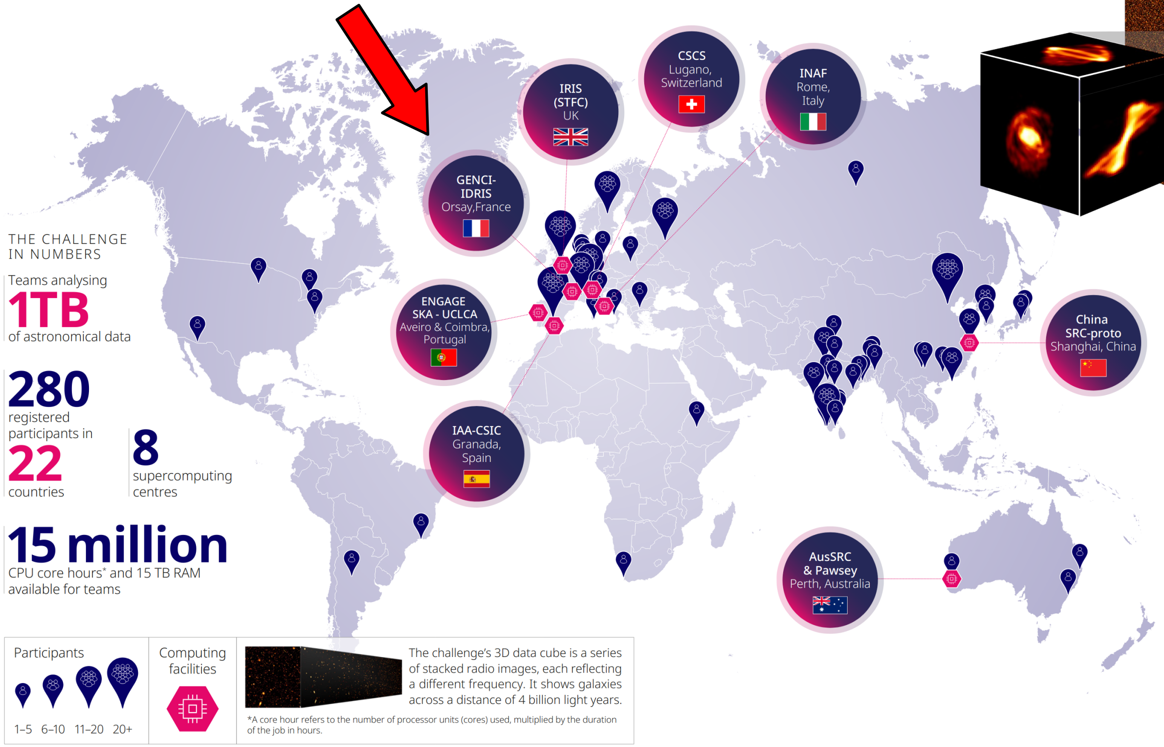
	Team (method)	M_s (Score)	N_{det}	N_{match}	N_{false}	$N_{bad} \in N_{false}$	Purity	$\bar{s}$	
	Post-challenge results								
ML (CNN)	MINERVA (YOLO-CIANNA)	**480450**	724480	680000	44480	16839	93.86%	0.7719	*After challenge ending*
	↪ *purity-focus thresholds*	418434	541542	536412	5130	2506	**99.06%**	0.7896	
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	
	7 other teams, one other ML attempt		...	...	...	...	...	...	

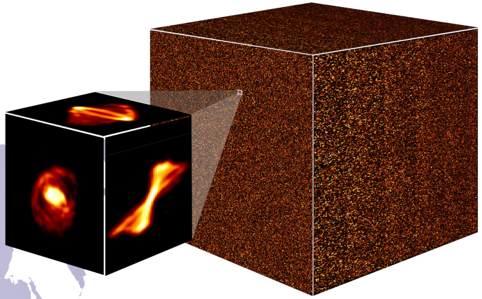
→ SDC1 is still a very interesting dataset for source detection pipeline development !

MINERVA team paper, YOLO-CIANNA → Cornu et al. 2024

SCIENCE DATA CHALLENGE 2

GENCI – Jean Zay supercomputer, GPU partition





Data: a 3D cube of simulated HI emission

- 20 square deg area
- 950 to 1150 MHz frequency (30KHz res; z = 0.235–0.495)
- 2000h integration time
- **Size of 1 TB !**
- 40GB cube for training

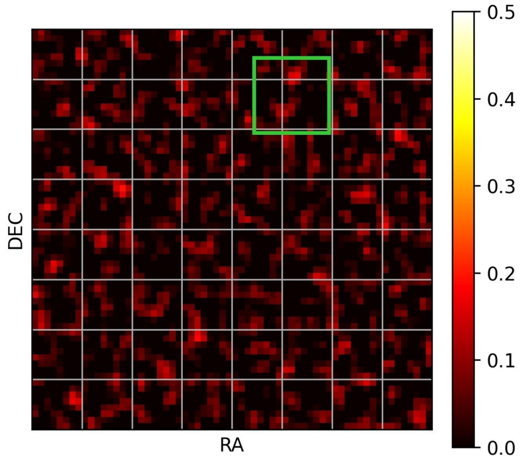
The challenge:

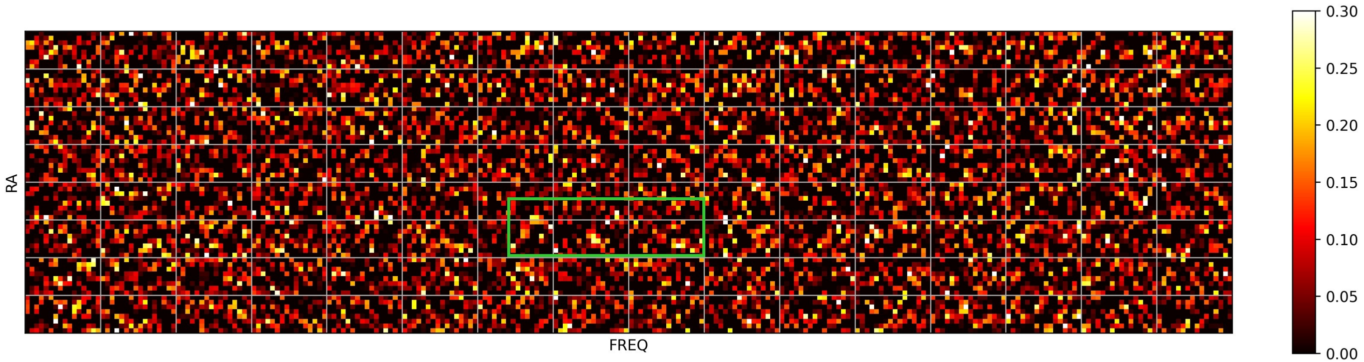
1. Find the sources (RA, Dec, Freq)
2. Characterize each source:
   → Flux, HI size, line width, PA, Inclination

Compute facilities: teams were dispatched on 8 compute facilities to prepare the model of data access through the future SKA Regional centers

**Challenge data are accessible on the dedicated web-page.*

Selection function difficulties



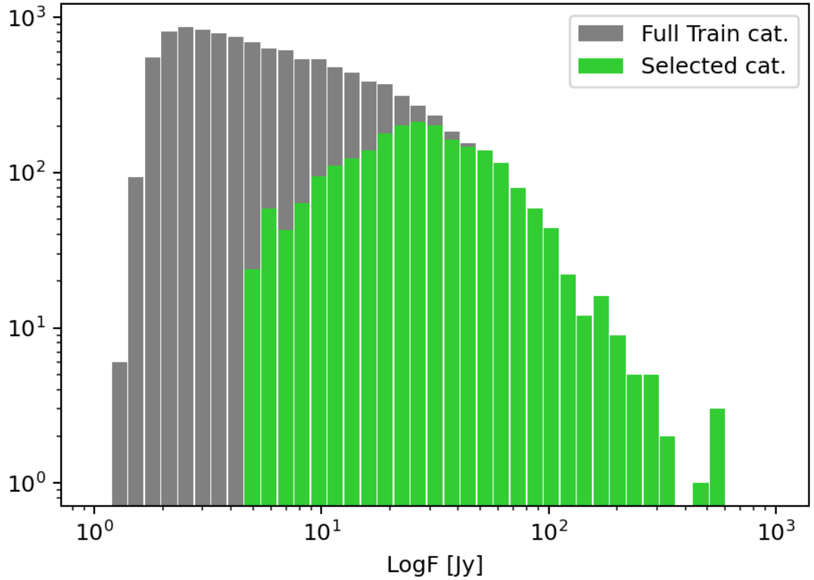


Selection function based on brightness or SNR are not sufficient to fully represent the noisy 3D information.

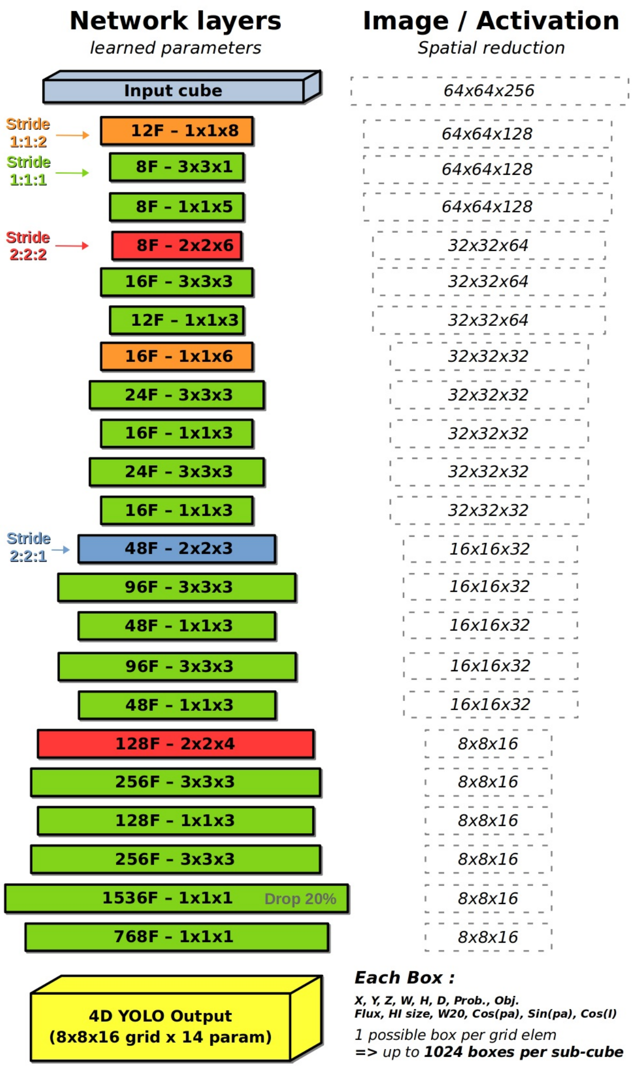
1st order combined selection :

- SNR & volume brightness
- Classical detection (FoF)

"Self learning" (~active learning):
After a first training, **un-selected true sources** with **high predicted objectness** can be re-injected in the training sample.



Model modifications for SDC2



YOLO parameters:

- **Generalized to 3D detection**
- **23 layers ~ 4 Million parameters**
- 1 single box prior per grid element! (prior 10x10x40)
- Predict 6 additional source parameters
- No class prediction

Training the network using IANNA

- **64x64x256** cubes are randomly selected in training area (40 GB)
- Around **2000 sources** in the selected train catalog
- Data are augmented using shifting and flips

Using a single RTX 4090 GPU
→ Training time up to **12 hours** (already good results after 6-8h).

Inference:

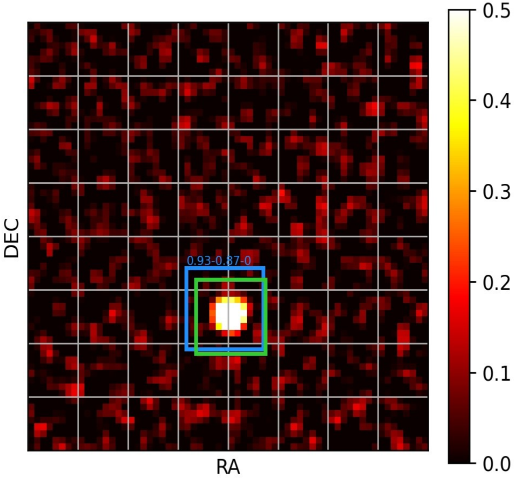
- The full SKA SDC2 1 TB cube is split in regions with large overlaps
- Box in overlapping regions are filtered with a dedicated NMS

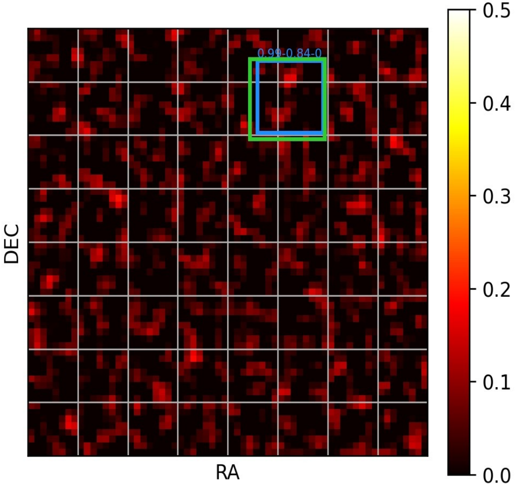
The full cube prediction takes ~1 hour (vastly dominated by data loading time) **using a single RTX 4090** ***(raw 260 ips)***

D. Cornu

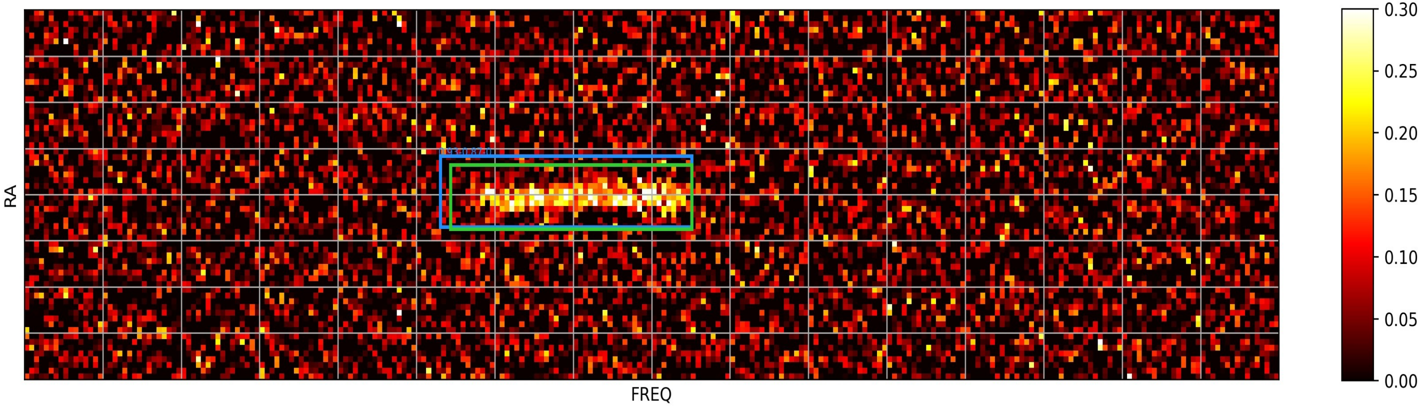
Source boxes detection

True boxes
vs Predicted boxes

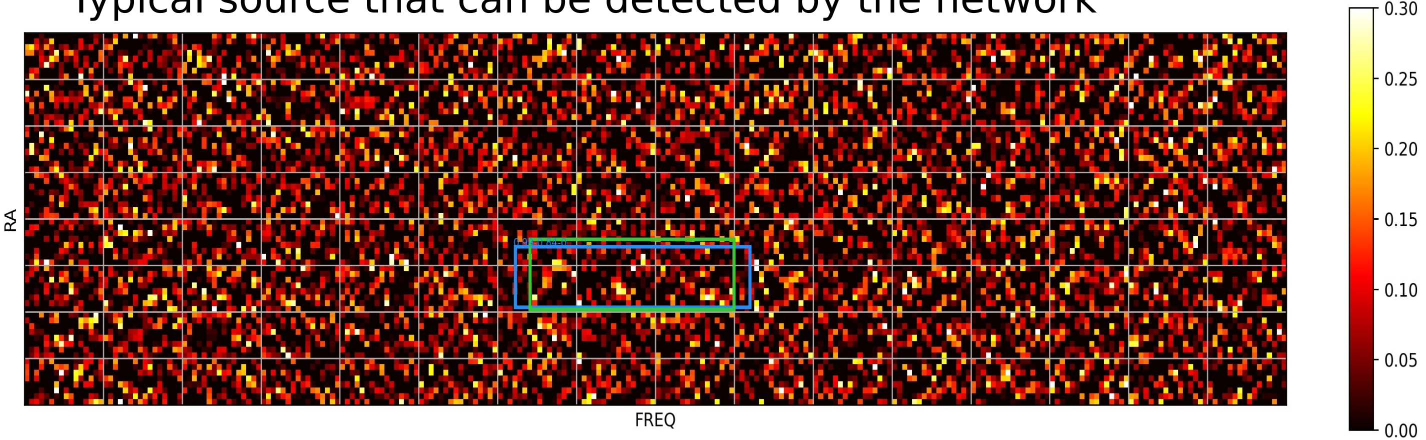




Brightest source (not typical!)



Typical source that can be detected by the network



*averaged over 20 channels in FREQ and 20 pixels in DEC respectively

LEADERBOARD

Results from Hartley et al. 2023

	Team name	Score	N_d	N_m	Accuracy
ML	MINERVA New*	23482	34441	31709	83
ML	MINERVA	23254	32652	30841	81
ML + SoFiA	FORSKA-Sweden	22489	33294	31507	77
SoFiA	Team SoFiA	16822	24923	23486	78
SoFiA	NAOC-Tianlai	14416	29151	26020	67
SoFiA	HI-FRIENDS	13903	21903	20828	72
Wavelets + ML	EPFL	8515	19116	16742	65
SoFiA	Spardha	5615	18000	13513	75
SoFiA	Starmech	2096	27799	17560	70
ML	JLRAT	1080	2100	1918	66
Wavelets + ML	Coin	-2	29	17	60
ML	HIRAXers	-2	2	0	-
Other	SHAO	-471	471	0	-

Minerva: **YOLO-only score, obtained after the challenge end*

MINERVA: *YOLO and CHADHOC combination*

FORSKA*: U-Net segmentation, parameters using SOFIA (Håkansson et al. 2023)*

EPFL: *Denoising with 3D wavelet filtering, identification with jointed likelihood, Parameters with several CNNs*

JLRAT: *Region proposal CNN detection, classical for parameters*

Coin: *Multiple CNNs for detection and dedicated CNNs for parameters*

HIRAXers: *Multiple CNNs for both detection and for parameters*

Key insight from SDC2: better scores when combining pipelines of different nature

How to transition to SKA precursors?

Multiple groups are already at work, developing ML pipelines for several instruments

LOFAR

ASKAP

MeerKAT

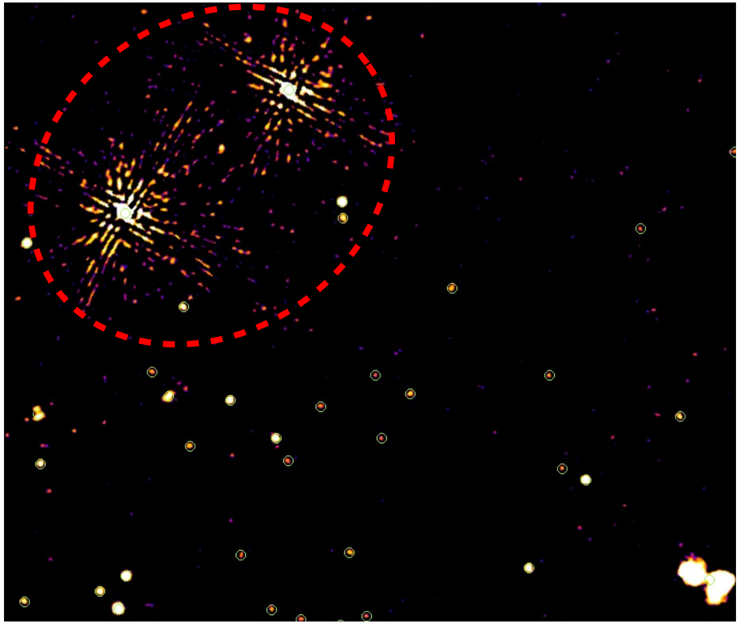
A. Anthore

On going work to generalize YOLO-CIANNA to the LoTSS and RACS surveys

Exploratory work to generalize YOLO-CIANNA to WALLABY and LADUMA

Challenges of working with real data

Example on the LoTSS survey (LOFAR)

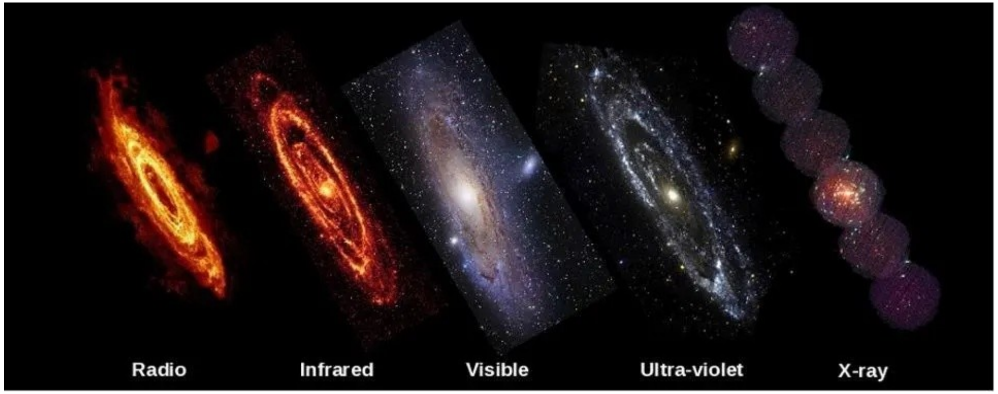


Difficulties : Artifacts / Noise / Resolution / Sizes / Morphology

How to define the training sample?

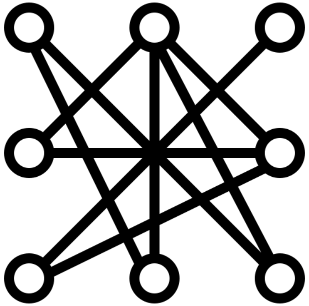
- **Use costly observations on few sources**
  Pros: Very robust labels **Cons:** few examples & imbalance
- **Use classical detection methods!**
  Pros: Easy to use, large samples **Cons:** possible bias
- **Use simulations (e.g SKA SDCs models)**
  Pros: infinite examples **Cons:** bias, instrument model required
- **Use Citizen Science (e.g Radio Galaxy Zoo)**
  Pros: "Easy" **Cons:** bias / errors, limited to human capability
- **Combine all of the above!**
  Pros: Very complete / diverse **Cons:** difficult to balance
- **Self / Active - Learning or Unsupervised**
  Train with one sample, then use one of the above to refine « new candidates », or try various flavor of unsupervised methods
  Pros: limits defined by the method and the data themselves, less human bias.

Toward multimodal astronomical analysis



Multi-wavelength or multi-messenger observations

Instrumental setups



$$F = G\frac{m_1 m_2}{r^2}$$

$$\upsilon = H_0 D$$

$$\frac{R(t_0)}{R(t_e)} = (1+z)$$

...

Known physical laws