A distributed data-mining software platform for extreme data across the compute continuum

Real-time detection of Solar and Jovian radio bursts with NenuFAR: a part of the EXTRACT project.

An overview of TASKA – A use cases

Emilie Mauduit, Cedric Viou, Louis Bondonneau, Baptiste Cecconi, Julien Girard, Fadi Nammour

The EXTRACT Project has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110

The EXTRACT project

EXTreme data Across the Compute conTinuum

- Develop novel data-driven orchestration mechanisms to efficiently deploy and execute data mining workflows.

- Personal Evacuation Route (PER) :

Transient Astrophysics with a Square Kilometer Array (TASKA)

- Radio observations with higher resolution and sensitivity → next generation radio telescopes like NenuFAR or SKA produce massive data volumes
- ⟹ Need for efficient tools to manage, analyse and store data.

- **TASKA** uses **EXTRACT** technologies to handle these massive data fluxes (~100GB/hr in beamformed for NenuFAR)
- Several projects:
  - **A1-A2** : dedicated to real-time detection of solar spikes and jovian millisecond bursts (C.Viou, E.Mauduit, L.Bondonneau),
  - **C** : creating an interactive workflow that connects multiple platforms
  - **D** : developing an imager with deep learning to reconstruct transient radio sources

NenuFAR radiotelescope

- $\sim$ 2000 antennas (96 MR of 19 antennas + 6 distant MR
- Beamforming mode: 4.3 TB/hr (~ 35 PB/yr)
- Imaging mode : 8.6 TB/hr (~ 74 PB/yr)

➢ Local storage: 3.5 PB

➢ Data reduction by scientists → diminution of data volume by a factor 100

➢ Reduced data are transferred towards data centres (3 PB in Meudon and 5 PB in Orléans)

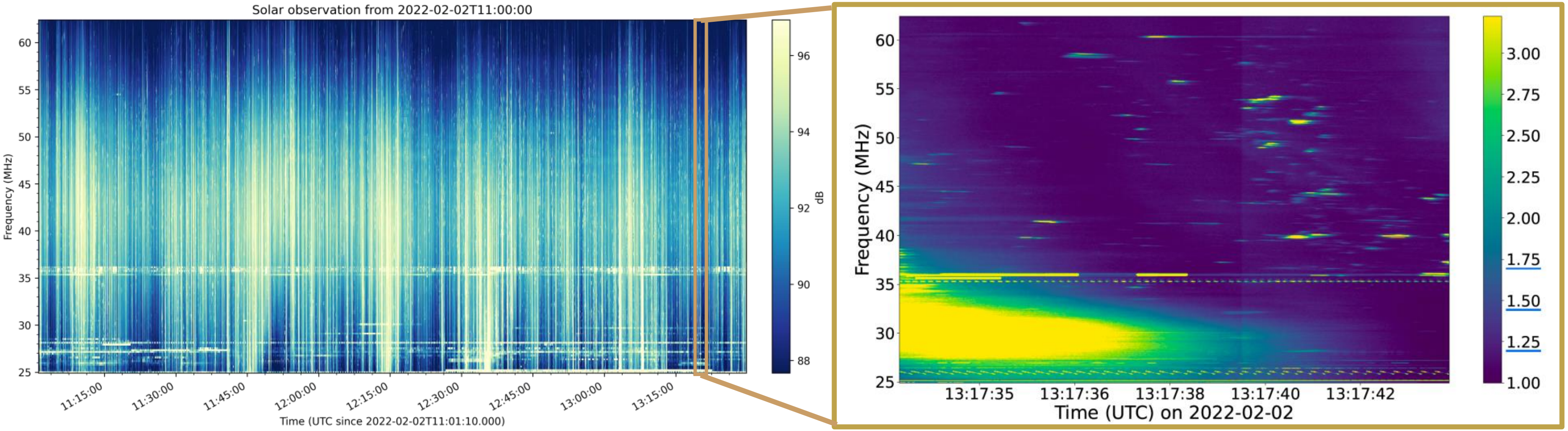
⟹ Data must be processed where it is stored !

- Range 10-80 MHz
- resolution :
  - dt = 0.3-84 ms
  - df = 0.1-12 kHz

TASKA – A1 : Solar spikes

Solar radio spikes

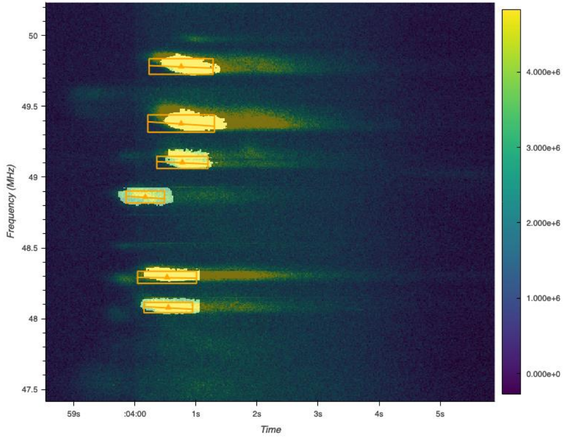
- Dedicated to automatically detect solar radio spikes using SpikeNet

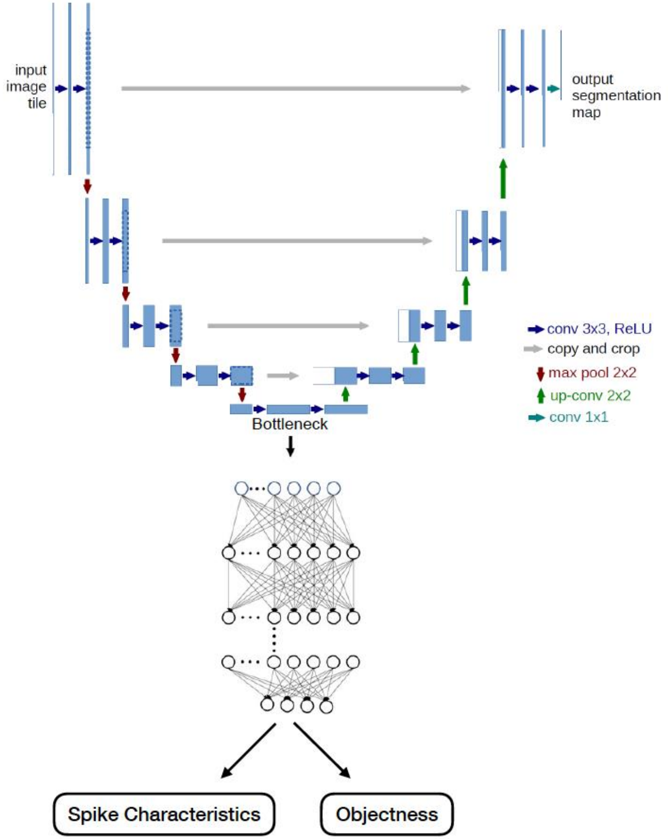


[Murphy et al, OJA, 2024]

Detection method : SpikeNet

- ML model trained on ~100 000 samples of solar spikes (64 by 64 pixels)
- Produced segmentation masks for radio spikes in the validation set and predicted the spike characteristics, i.e. location in time and frequency, duration, spectral width and drift rate
- Bursts that are not fully in one tile can not be detected



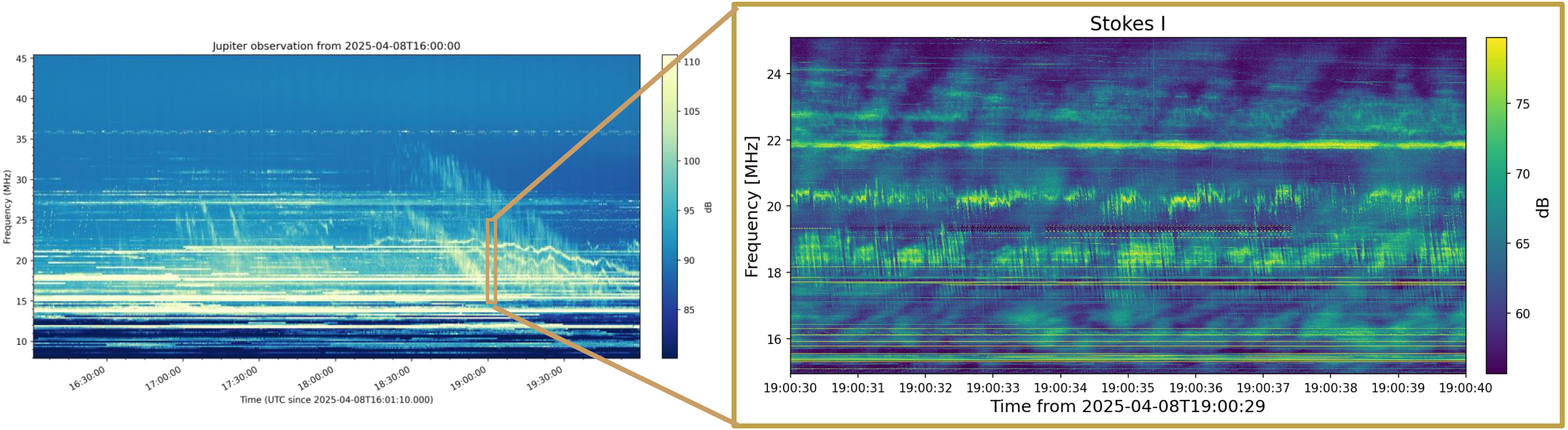


[Murphy et al, OJA, 2024]

TASKA – A2 : Jovian S-bursts

Jovian S-bursts

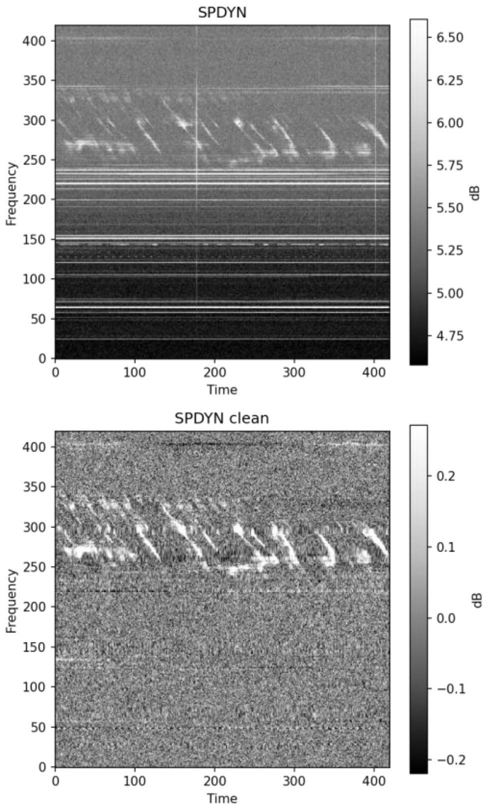
- Developing an algorithm for real-time detection of fine drifting structures during radio observations with NenuFAR
- Use a method originally developed for Jupiter observations with NDA-JunoN data *[Mauduit et al, 2023, Nat Coms]*



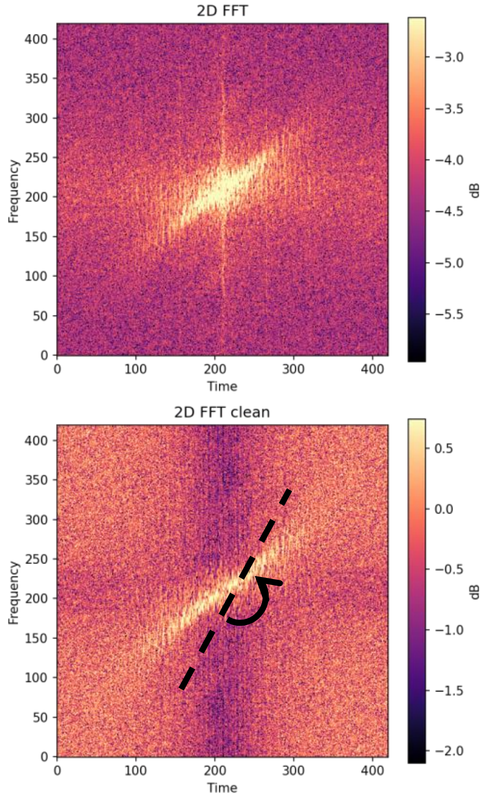
Detection algorithm : spyburst

spyburst

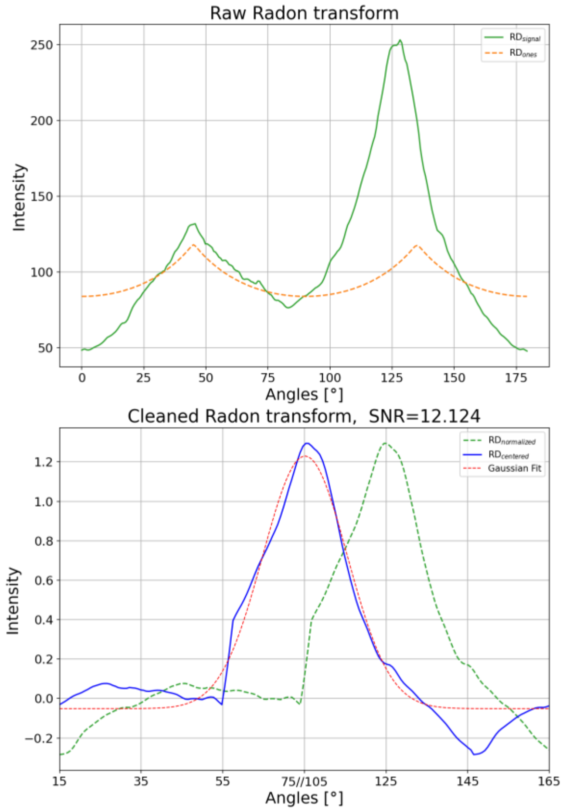
1 - RFI mitigation

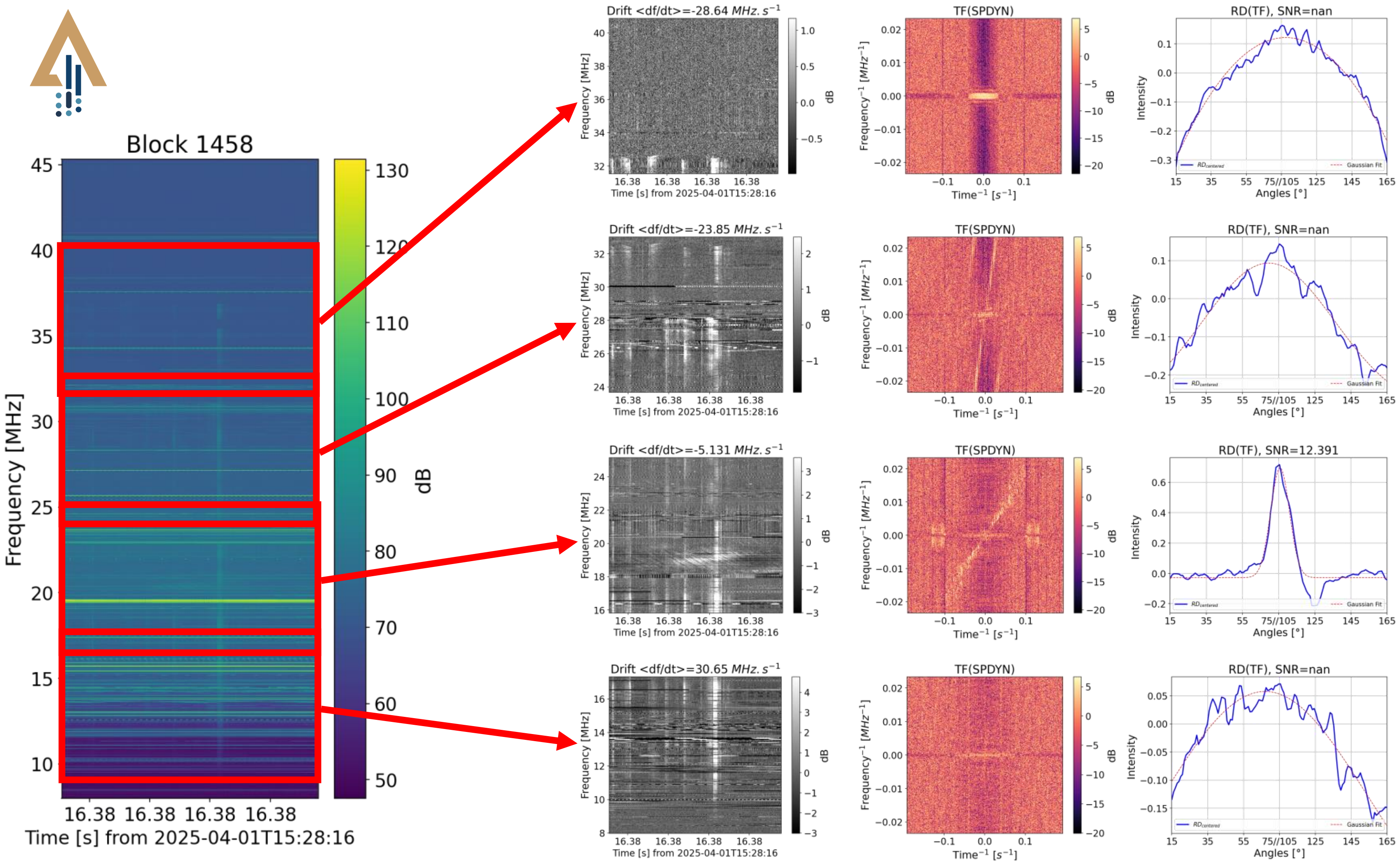


2 – FFT2D



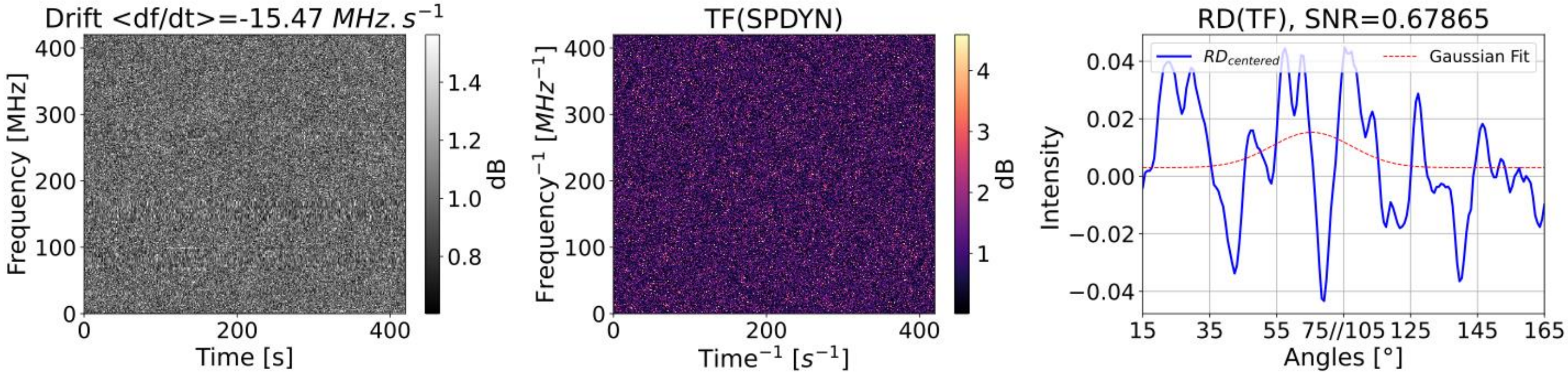
3 - Radon transform



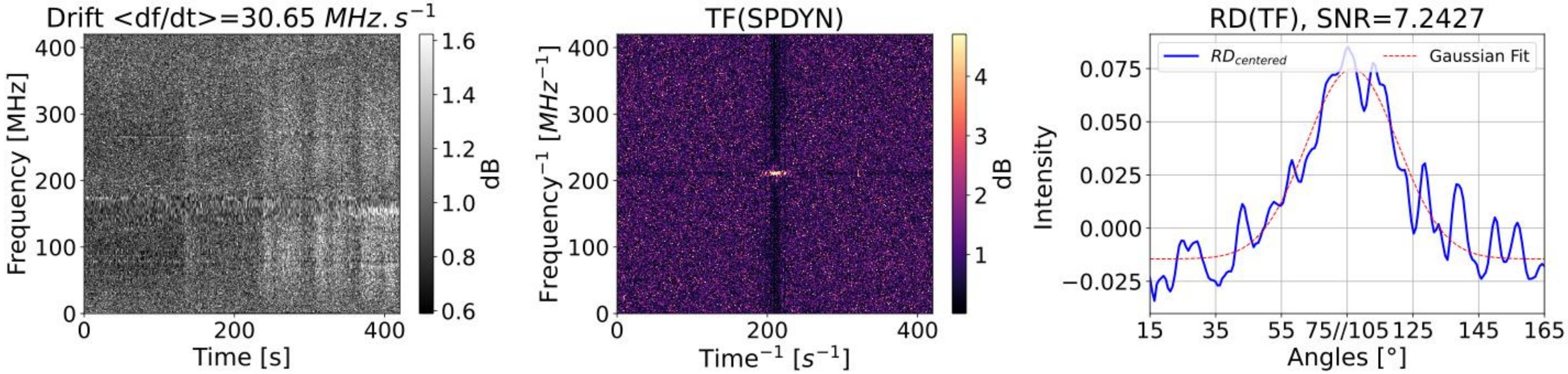

Block 1458
Frequency [MHz]
Time [s] from 2025-04-01T15:28:16
dB
Drift <df/dt>=-28.64 MHz. s⁻¹
TF(SPDYN)
RD(TF), SNR=nan
Drift <df/dt>=-23.85 MHz. s⁻¹
TF(SPDYN)
RD(TF), SNR=nan
Drift <df/dt>=-5.131 MHz. s⁻¹
TF(SPDYN)
RD(TF), SNR=12.391
Drift <df/dt>=30.65 MHz. s⁻¹
TF(SPDYN)
RD(TF), SNR=nan
Frequency⁻¹ [MHz⁻¹]
Time⁻¹ [s⁻¹]
Intensity
Angles [°]
RD_centered
Gaussian Fit

Exemples de non-détection

Pas de détection



Pente positive



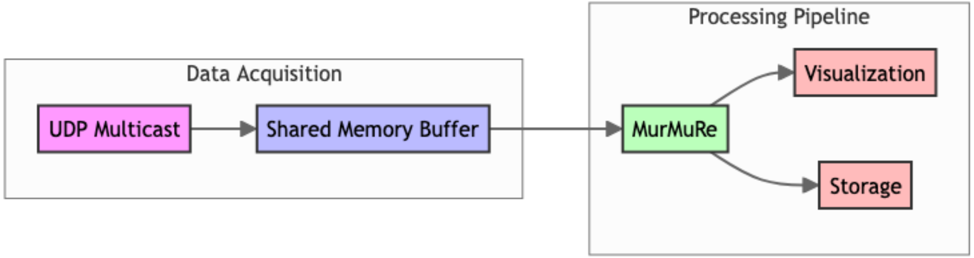
MurMuRe pipeline and EXTRACT interfaces

Modular Multicast Receiver

MurMuRe

L.Bondonneau (ORN)

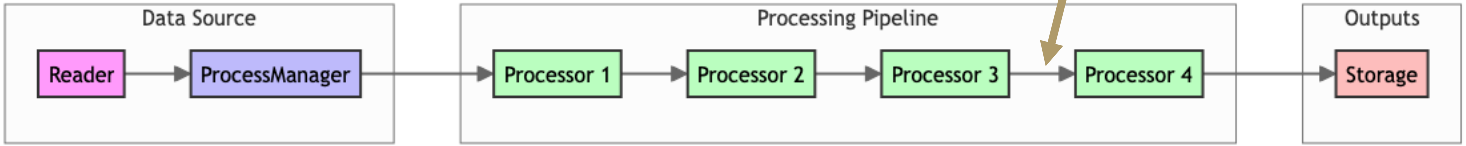
Pipeline Architecture



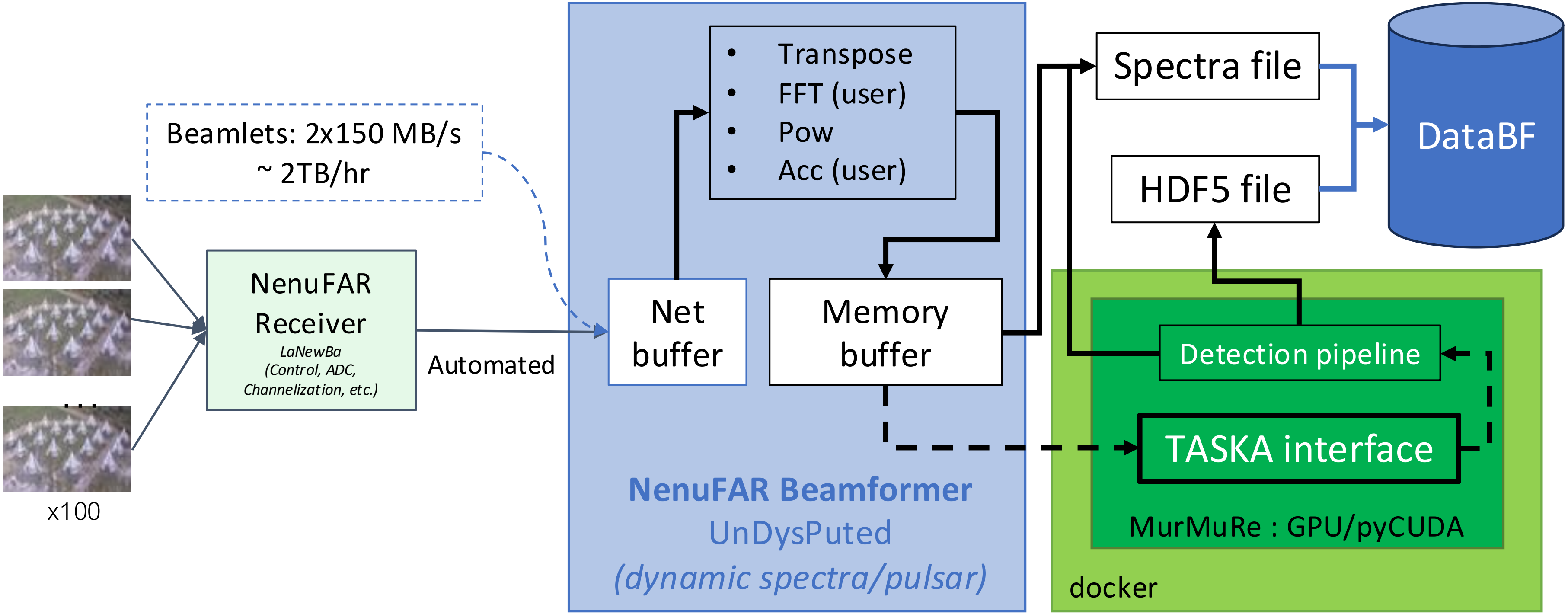
A1 or A2 processors

- MurMuRe :

Simple Example

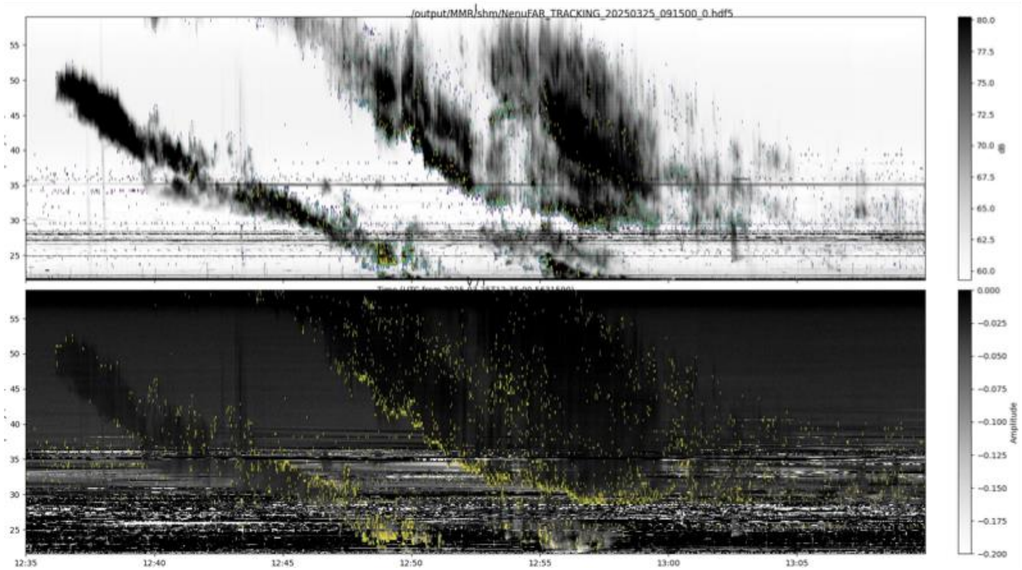


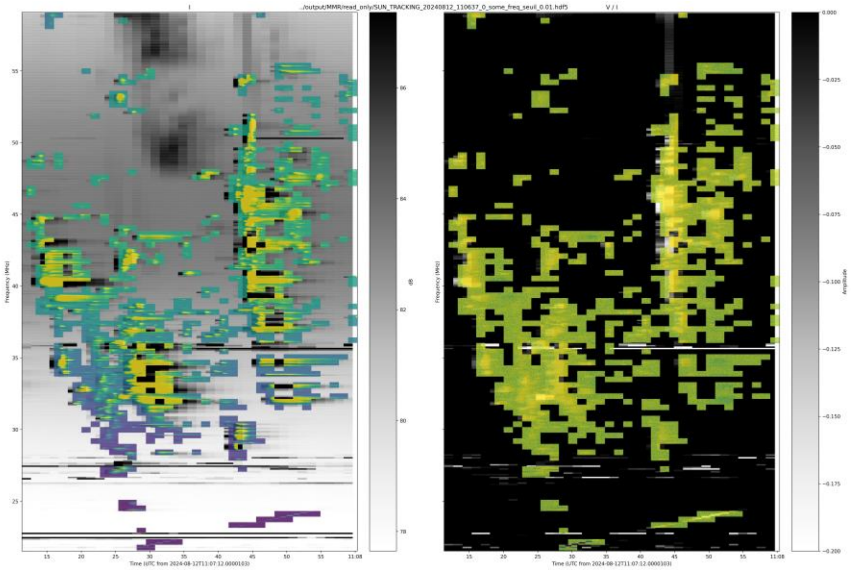
Undysputed-tf with MurMuRe and EXTRACT



Real-time detection with NenuFAR

SpikeNet successfully applied to real-time observations





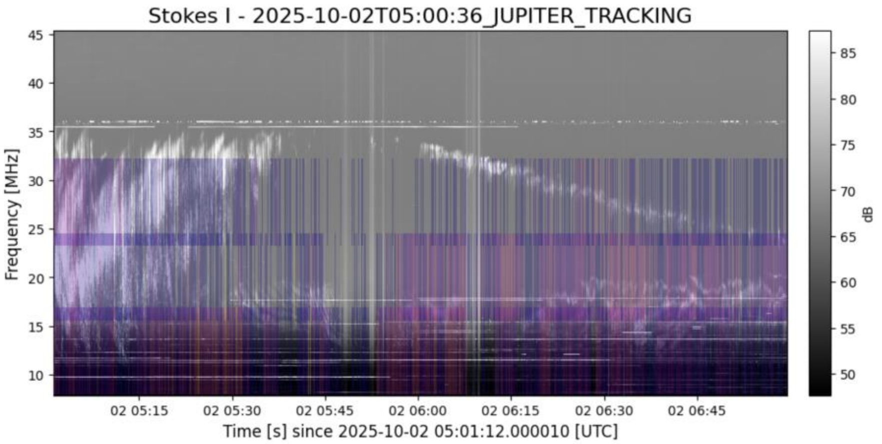
Observation campaign : from 22 to 29/03/2025, 09h15-14h50, 21-84 MHz

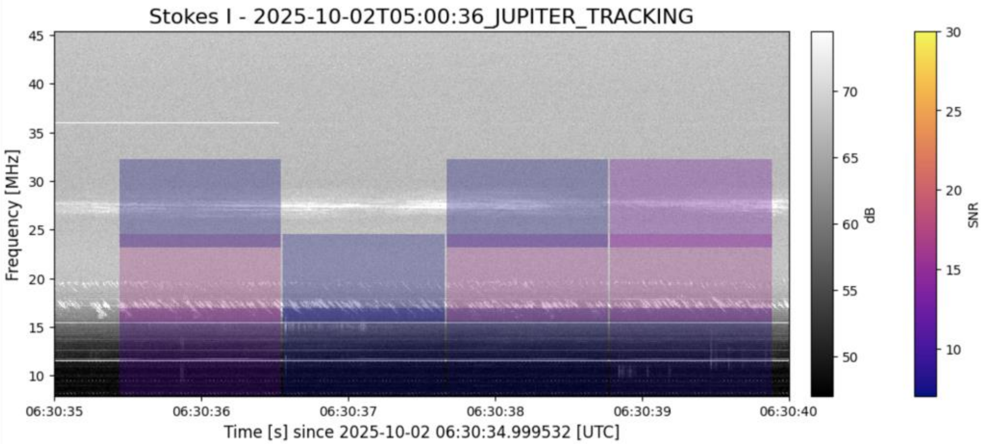
→ 10x data volume reduction !

GitLab repository

	Original pipeline	TASKA-A2 pipeline	
df, dt	6.1 kHz, 21 ms	98 kHz, 1.34 s	6.1 kHz, 21 ms
.spectra file size	27 Gb/hr	0.037 Gb/hr	-
.hdf5 file size	-	-	2.5 Gb/hr

Spyburst successfully applied to real-time observations





→ Réduction du volume de données d'un facteur 10 environ !

[GitLab repository]

	Original pipeline	TASKA-A2 pipeline	
df, dt	3.05 kHz, 2.5 ms	24 kHz, 1.06 s	3.05 kHz, 2.5 ms
.spectra file size	250 Gb/hr	0.082 Gb/hr	-
.hdf5 file size	-	-	40 Gb/hr

Conclusions

Conclusions

- Important step toward smart data filtering for next generation radiotelescopes such as NenuFAR or SKA.
- It also paves the way for “analog to information” processing, which would drastically reduce the storage needs.
- The type of emissions studied in this work require a high time-frequency resolution, but they are embedded within larger slowly-varying emissions that can be studied at a lower resolution.
- This approach helps optimizing data storage while maintaining its value for scientific analysis, thus preparing for scalable solutions in the SKA era.
- Code available on GitLab, easy to clone and use with documentation and notebook tutorials provided.

[https://gitlab.obspm.fr/extract]

Next steps and perspectives

- Develop a CNN for S-bursts based on anomaly detection.
- Optimize existing algorithms to have better detections → see V. Harff poster
- spyburst is being used with data from other instruments (LWA, Juno, ..)
- Find an even more efficient way to store the outputs.
- Include multi-scale processing

SpikeNet

MurMuRe

spyburst

Thank you !

emilie.mauduit@obspm.fr

A distributed data-mining software platform for extreme data across the compute continuum

Follow us on social media:

www.extract-project.eu

The EXTRACT Project has received funding from the European Union's Horizon Europe programmeunder grant agreement number 101093110