EXTRACT

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

TASKA

Towards a digital computing platform for modern radio astronomy

Baptiste Cecconi, Julien N Girard, Emilie Mauduit, Louis Bondonneau, Cédric Viou, Stéphane Aicardi, Fadi Nammour, Victor Landeau and the EXTRACT collaboration

See deliverables & demos extract-project.eu

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

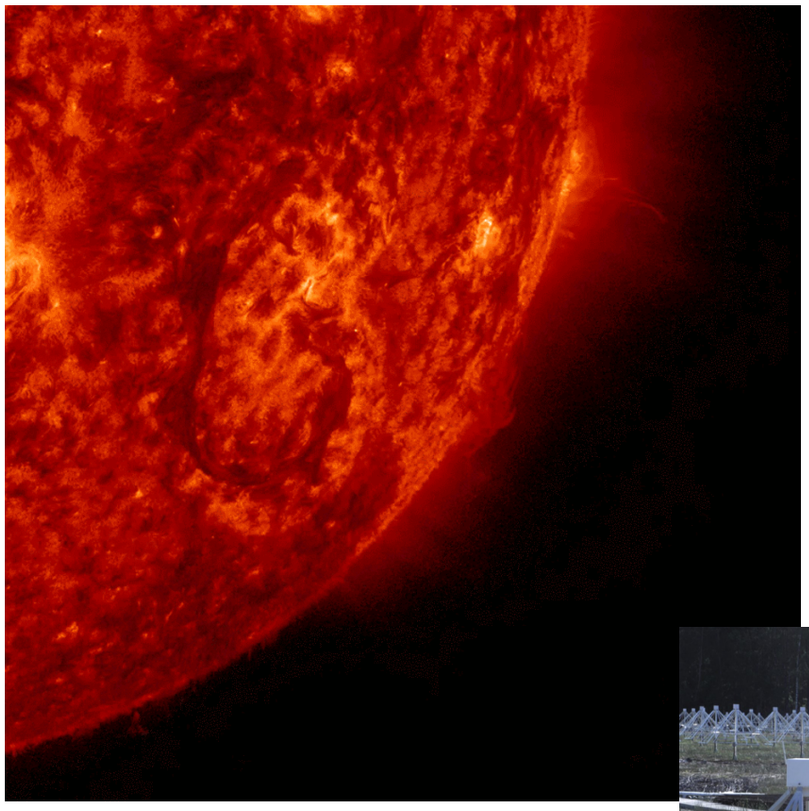
EXTRACT Use-Cases, sharing the same platform

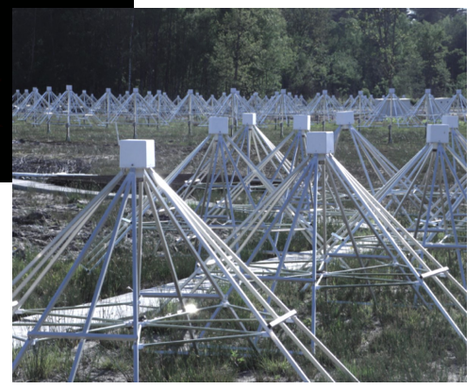
PER

Personalised Evacuation Route (PER)
in the City of Venice based on an
Urban Digital Twin and an AI engine

TASKA

Transient Astrophysics with the Square Kilometre Array pathfinder (TASKA)
NenuFAR generating high-volume and high-velocity data





NenuFAR

New extension in Nançay Upgrading loFAR

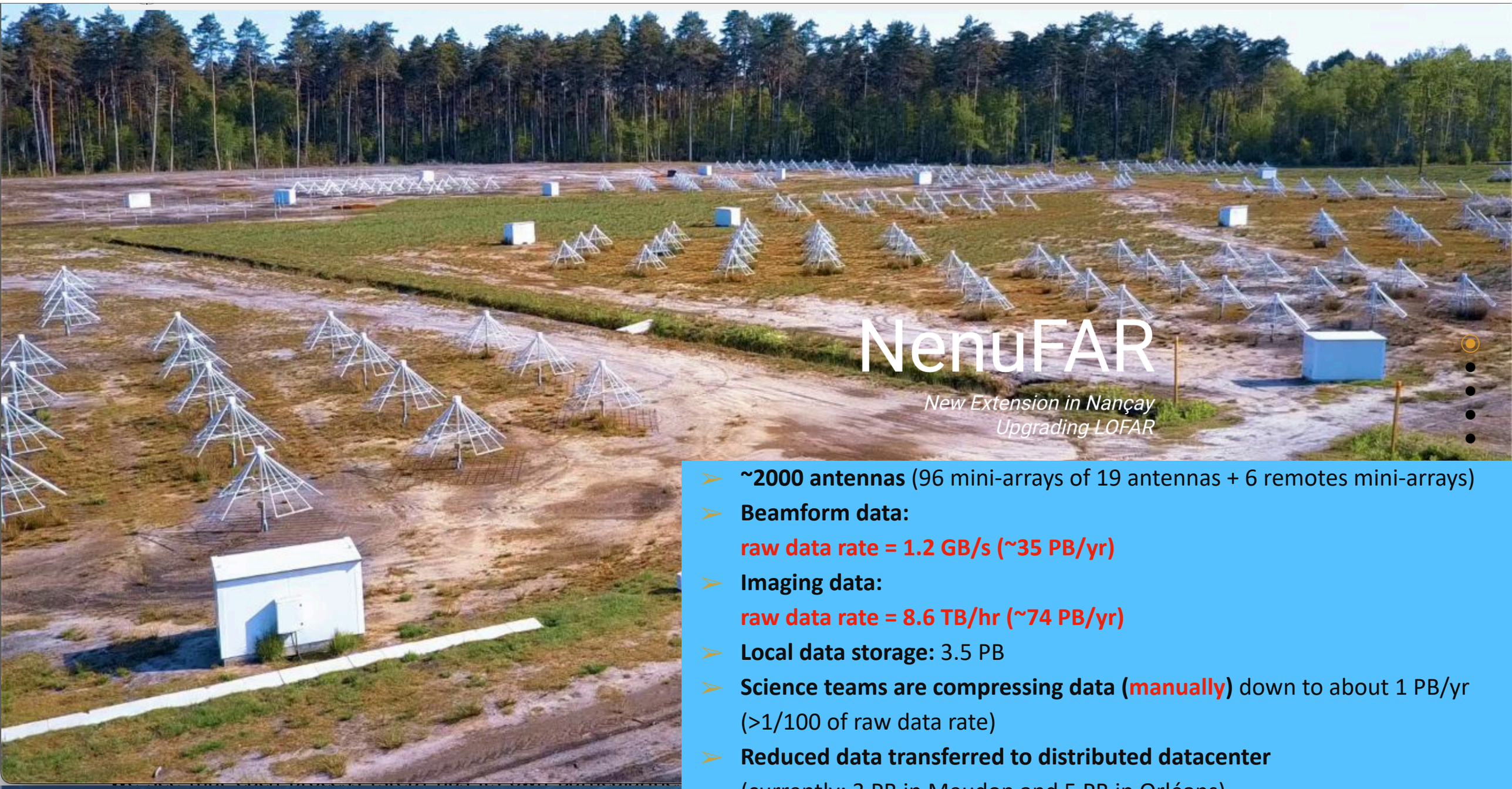
Pathfinder de SKA-LOW

Measures raw voltages

F = 20-80 MHz

N_A~2000 antennas

Operated as a phased array or an interferometer



- **~2000 antennas** (96 mini-arrays of 19 antennas + 6 remotes mini-arrays)
- **Beamform data:**
  raw data rate = 1.2 GB/s (~35 PB/yr)
- **Imaging data:**
  raw data rate = 8.6 TB/hr (~74 PB/yr)
- **Local data storage:** 3.5 PB
- **Science teams are compressing data (manually)** down to about 1 PB/yr (>1/100 of raw data rate)
- **Reduced data transferred to distributed datacenter** (currently: 3 PB in Meudon and 5 PB in Orléans)

TASKA

Astronomical signal quality

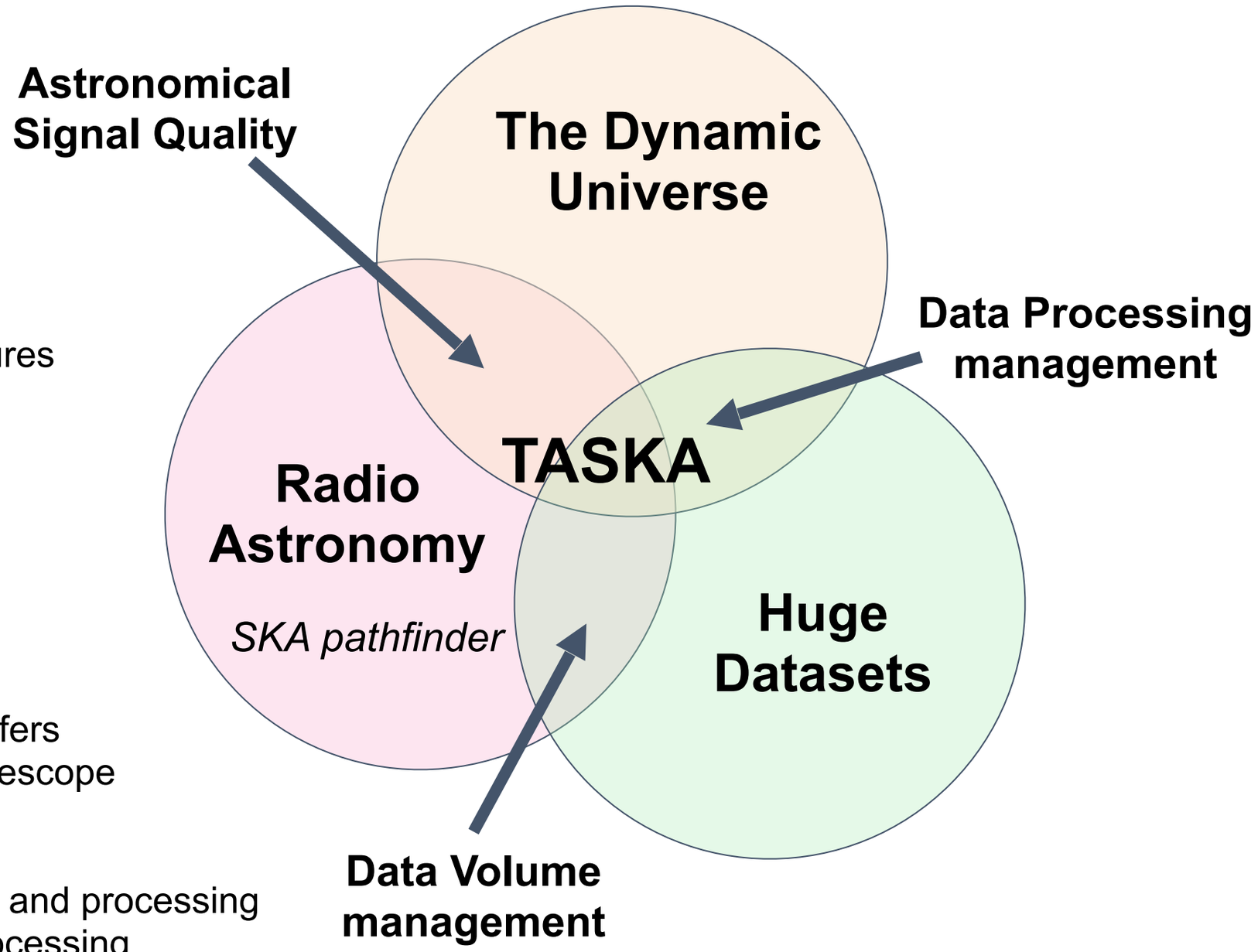
- High resolution & sensitivity
- Instrumental configuration
- serendipitous & complex time-freq structures
- Calibration

Data Volume Management

- Raw data:
  - unmovable data set (~10s TB)
  - Very demanding storage and transfers
  - “In-place” pre-processing at the telescope
- Intermediate data:
  - Ingestion of incoming data
  - Orchestration: distribution, storage and processing
  - Automation for multiple parallel processing

Data Processing Management

- Knowledge of the tools for analytics
- Knowledge of the post-processing scenarii
- Optimize the time and load of the post-processing steps
- Source restoration & classification
- Ensure the creation & verification of scientific products



TASKA Relevance: to gather Radio astronomy, HPC, Orchestration / Distribution experts together

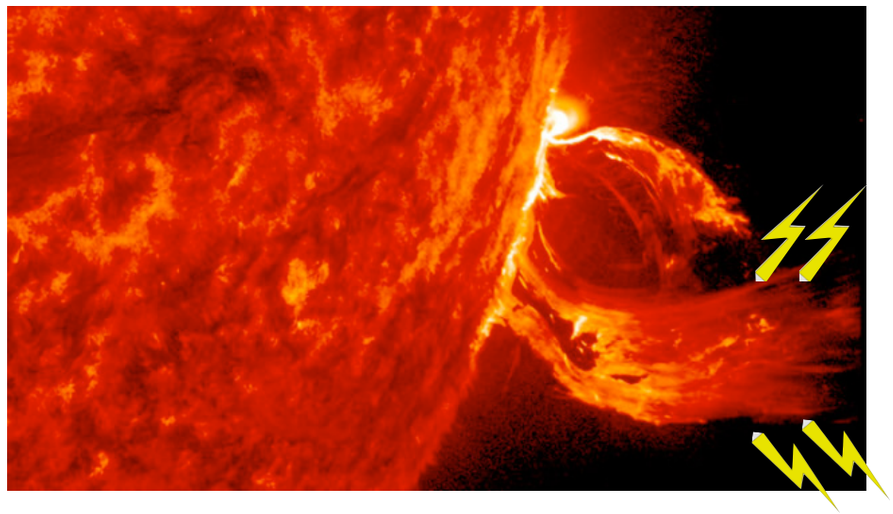
TASKA Overview

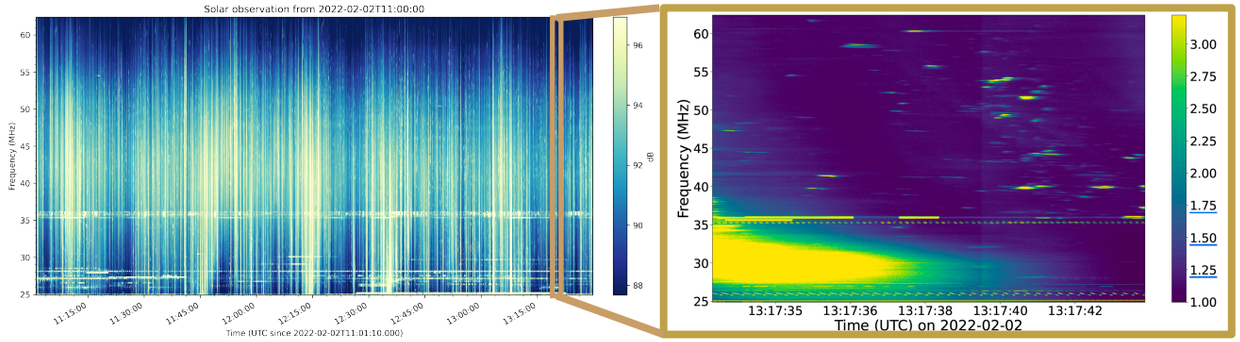
- **Use Case A:** Early detection and selective resolution data recording (space optimality)
  Emilie Mauduit **Edge**
- **Use Case B:** Adaptive scheduling (automatic decision making)
- **Use Case C:** Workflow orchestration of interferometric data processing with a focus on improving the processing speed, accuracy and automation on large datasets
  Julien Girard **Cloud**
- **Use Case D:** Prototype development for “dynamic” imaging of the variable Universe
  Fadi Nammour **IA**
- **Use Case E:** Advanced data reduction workflows for multi-dimensional real-time analysis and inference (joining A and C together)

Use case A: Early detection & selective recording

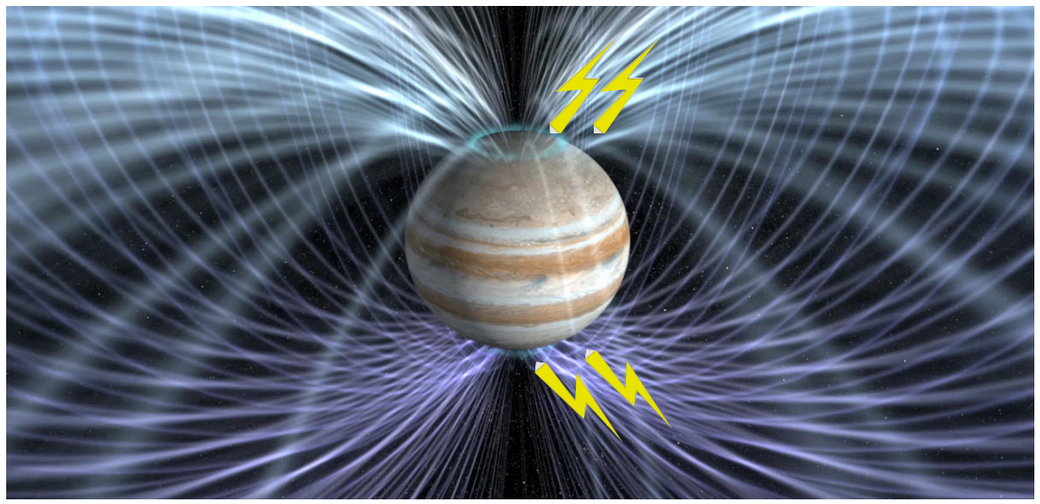
- Detection of solar radio spikes and Jupiter S-bursts

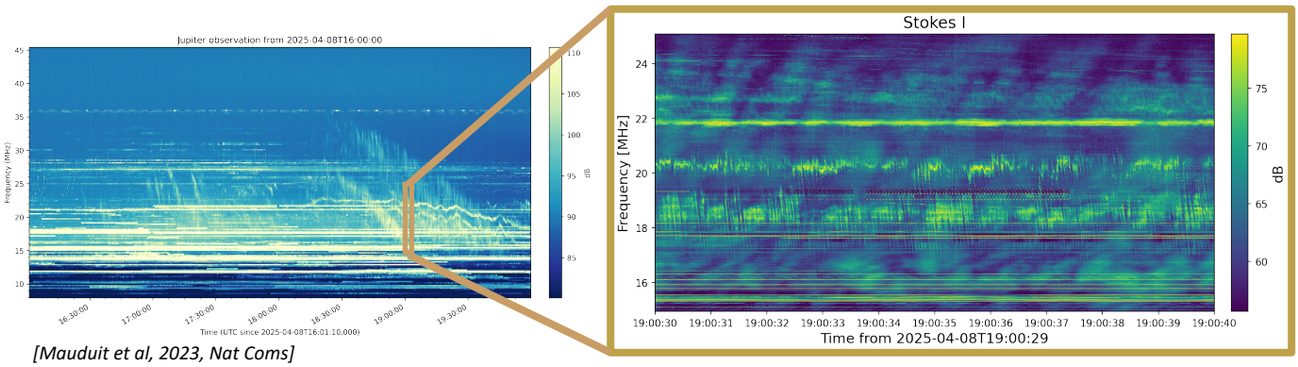
Edge





[Murphy et al, OJA, 2024]





[Mauduit et al, 2023, Nat Coms]

	Original pipeline	TASKA-A1 pipeline
(df,dt)	(6.1 kHz, 21 ms)	(98 kHz, 1.34 s)
Spectra	27 GB/hr.	0.037 GB/hr
HDF5	--	2.5 GB/hr

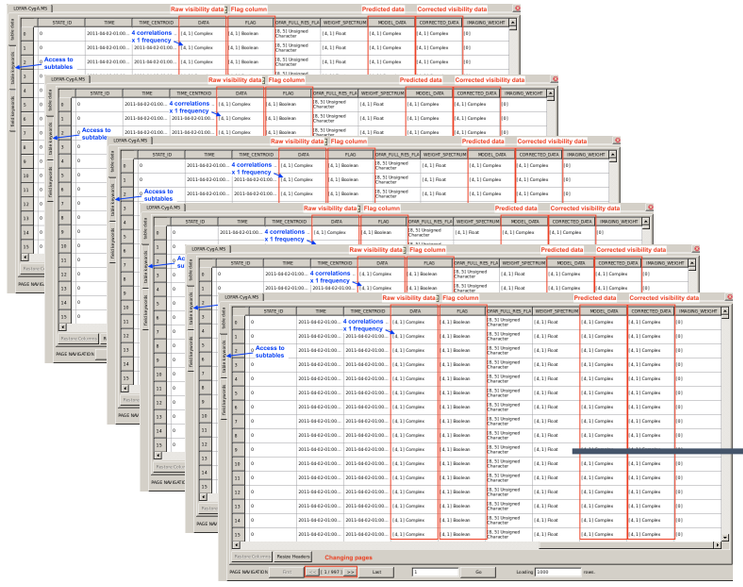
	Original pipeline	TASKA-A1 pipeline
(df,dt)	(3.05 kHz, 2.5 ms)	(21 kHz, 1.1 s)
Spectra	250 GB/hr.	0.084 GB/hr
HDF5	--	19 GB/hr

→ 10x data volume reduction !

TASKA Overview

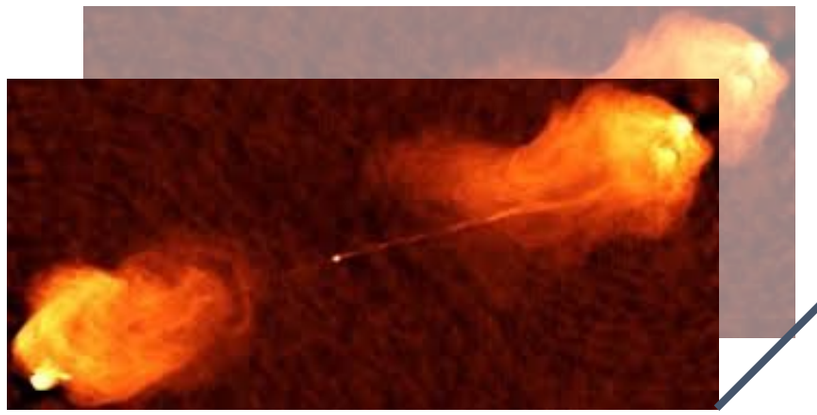
- **Use Case A:** Early detection and selective resolution data recording (space optimality) **Edge**
- **Use Case C:** Workflow orchestration of interferometric data processing with a focus on improving the processing speed, accuracy and automation on large datasets **Cloud**
- **Use Case D:** Prototype development for “dynamic” imaging of the variable Universe (DL transient imaging) **IA**
- **Use Case E:** Advanced data reduction workflows for multi-dimensional real-time analysis and inference (joining A and C together)

Use-Case C: Workflow orchestration for radiointerferometry



Data transfer
data processing

(Flagging, Rebining,
Calibration, Imaging)



Final product: time/freq
Image cubes

Starting dataset: Visibilities
(Measurement Sets (MS) Format)

Optimal dataset distribution ?
(Multiple sites)

Composable reduction?

(Multiple tools)
from combination of
known analytic “bricks”

Analytics

Multiple tools
(scientific quality, fidelity)

On-going experiments: 1) multi-site data & 2) distributed computing

Stockage objet (S3)

Obs. de Paris

OVH (Gravelines)

BRGM (Orléans)

URV

Cloud computing

Obs. de Paris

OVH (Gravelines)

URV

BRGM (Rancher)

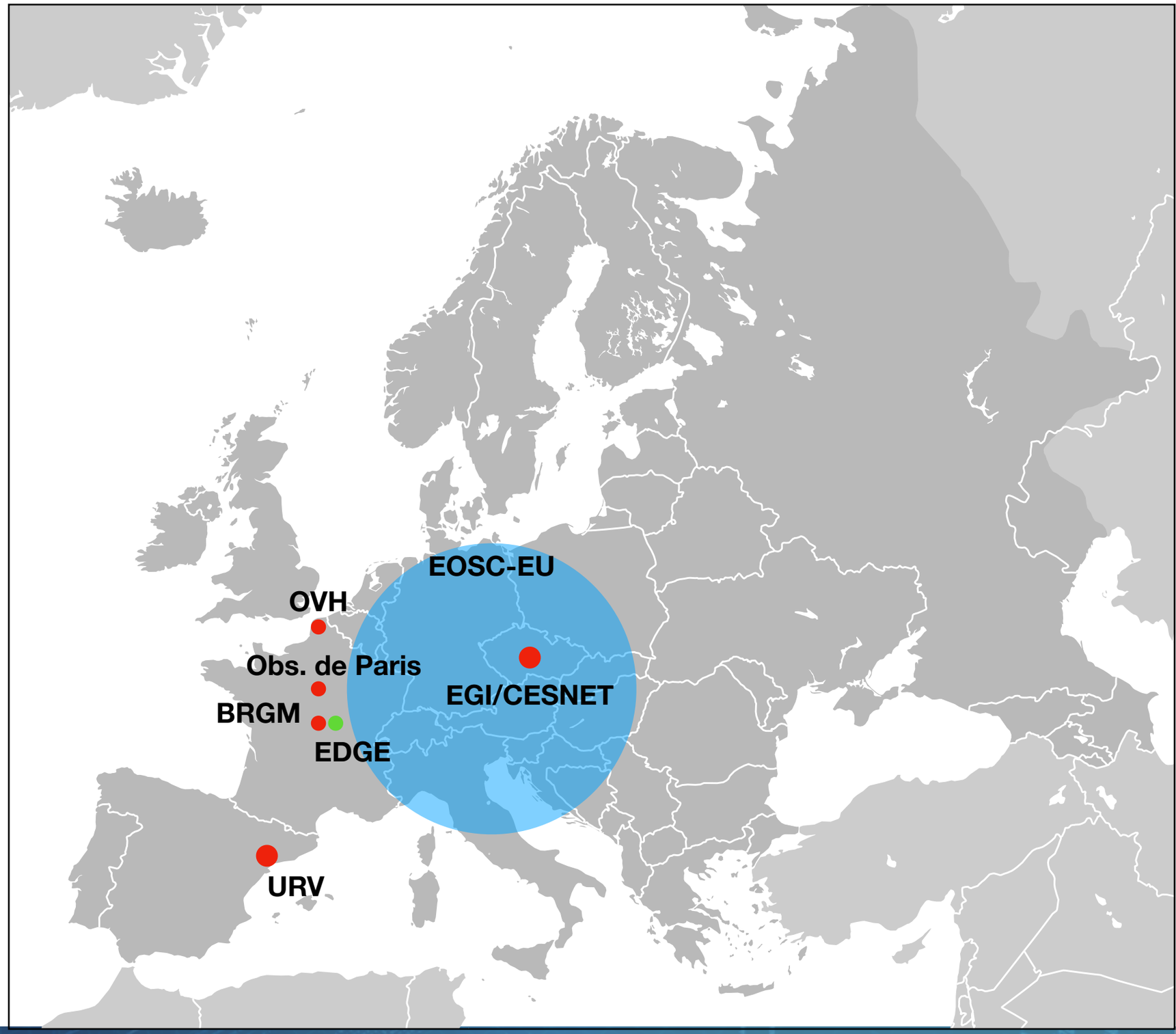
EGI/CESNET (Rancher)

EOSC-EU-Node

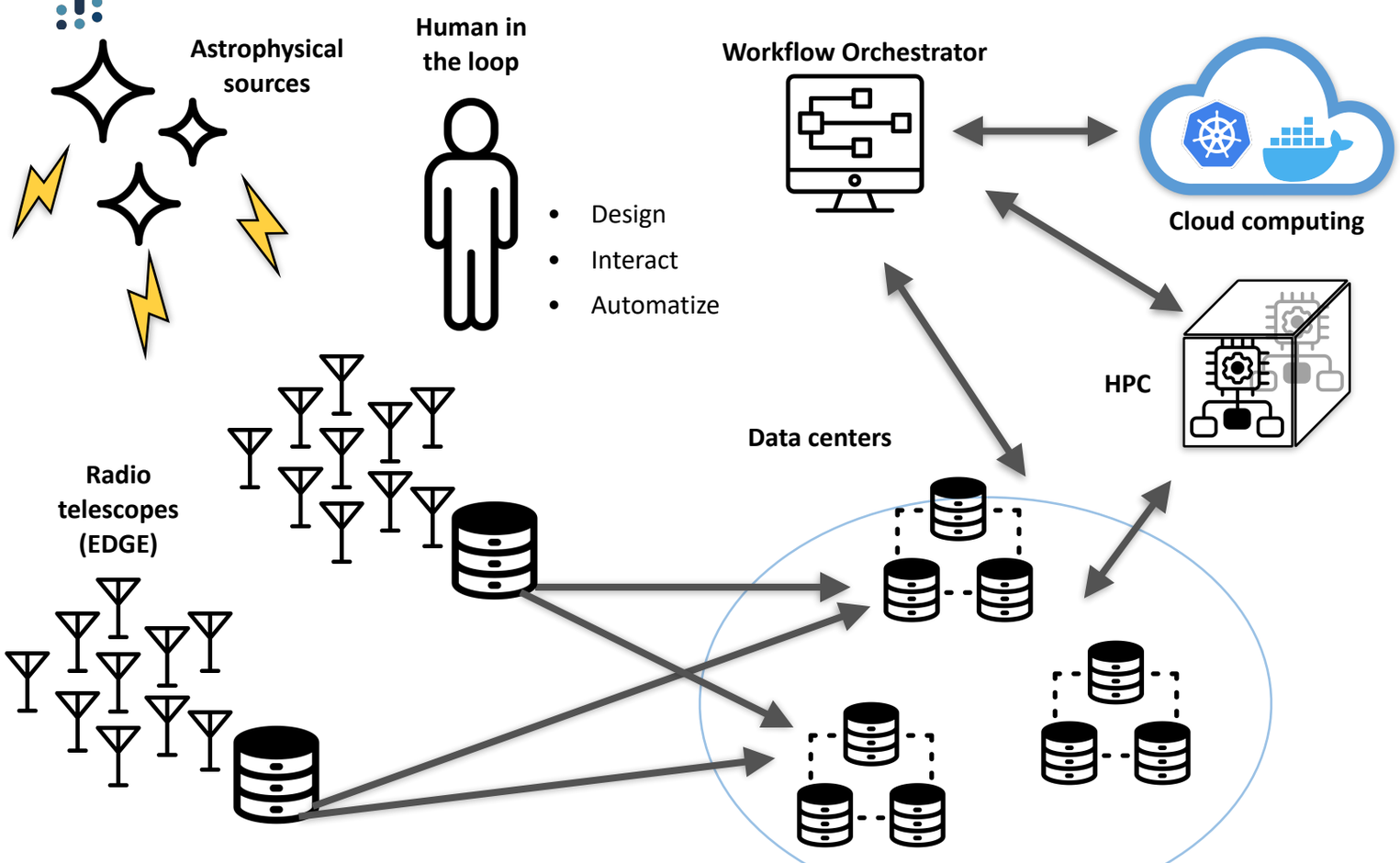
Data catalog

EDGE

Observatoire Radioastronomique de Nançay



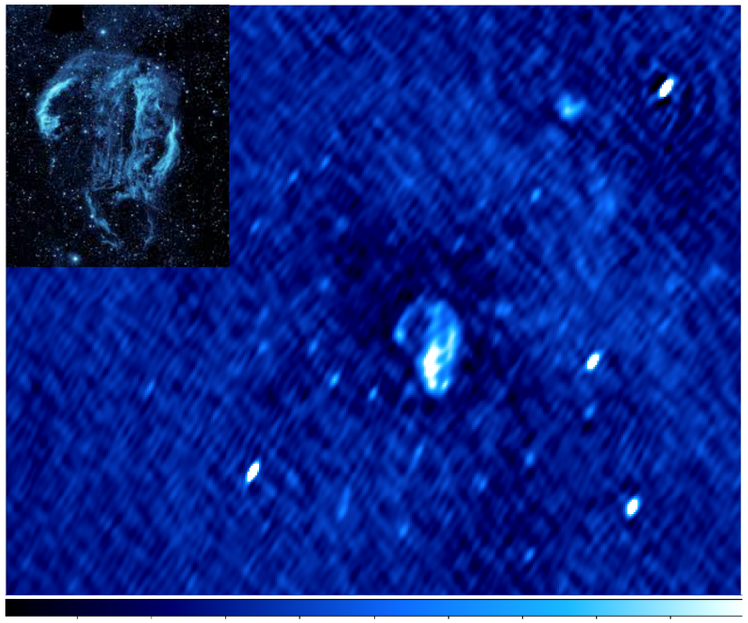
Workflow concept



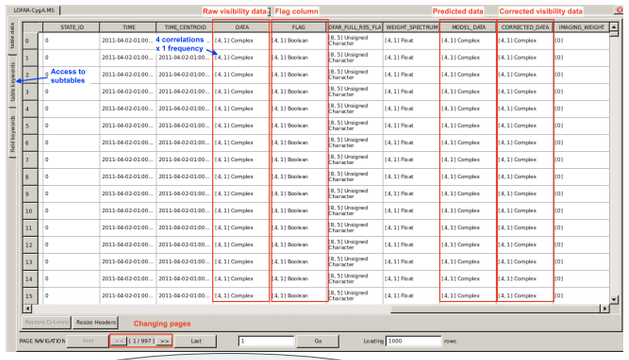
- **Each step needs to be optimized on computing inhomogeneous facilities (Cloud, HPC, etc.)**

  partitionning, parallelization, compression…

- **Scientific products are retrievable locally**



Distributed data



Raw Visibility Data

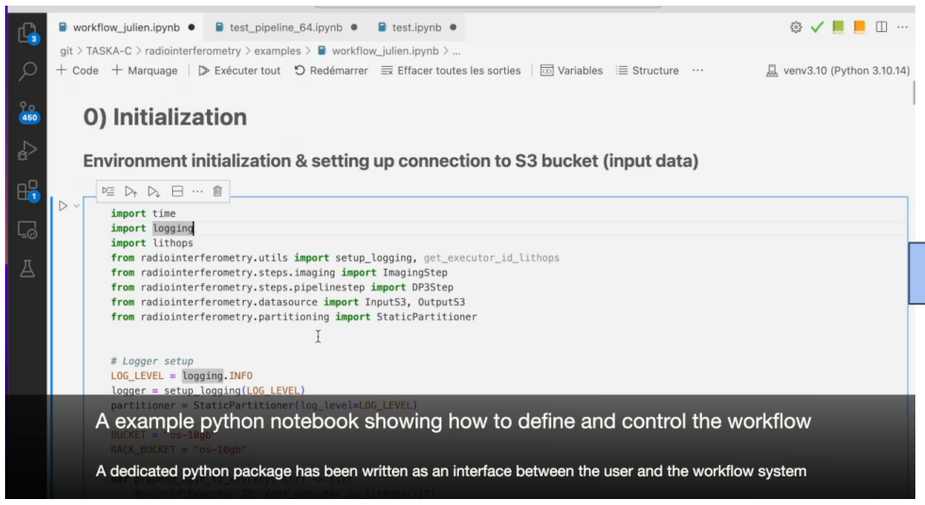
Complex direction-dependent effects

The final products are then retrieved on the scientist computer

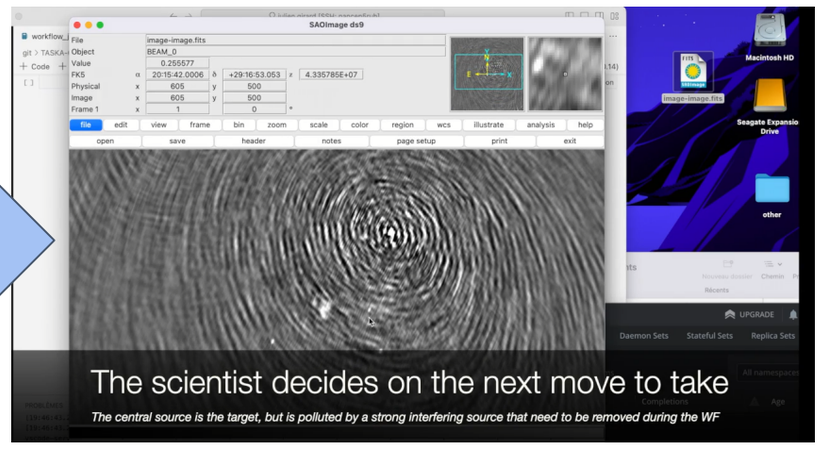
TASKA - “Interactive” Workflow

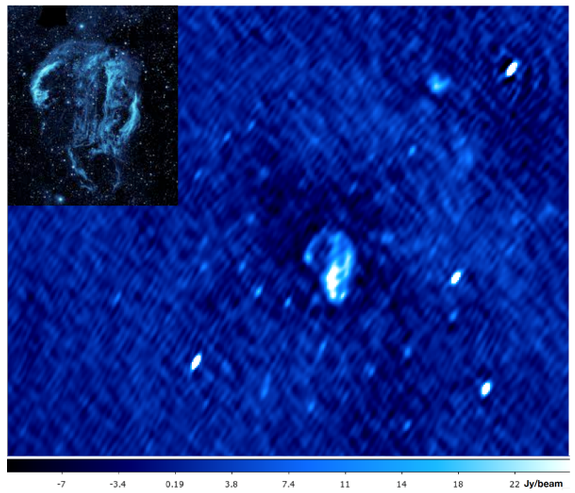
- Built as a “wrapper” that interacts with the astronomy community tools
  High potential impact because of the platform deployment in other communities (security, medical, resource management, etc.)
- Easy to invoke, easy to code, easy to customize, easy to “chain”: *natively made for workflows*
- Each task has a “definition” block and a “run” block: *separating the workflow building from its running*
- Run as a python script or in a python notebook (cf. DEMO video)

```
import time
import logging
import lithops
from radiointerferometry.utils import setup_logging, get_executor_id_lithops
from radiointerferometry.steps.imaging import ImagingStep
from radiointerferometry.steps.pipelinestep import DP3Step
from radiointerferometry.datasource import InputS3, OutputS3
from radiointerferometry.partitioning import StaticPartitioner
```



Controlled through a python notebook (S3, data partitioning, worker management, …)





The final products are then retrieved on the scientist computer

…as if the process and data were **local**

TASKA - Designing, failproofing a workflow

COMPSS User-oriented monitoring

When designing/debugging a workflow

We want:

- Ability to loop

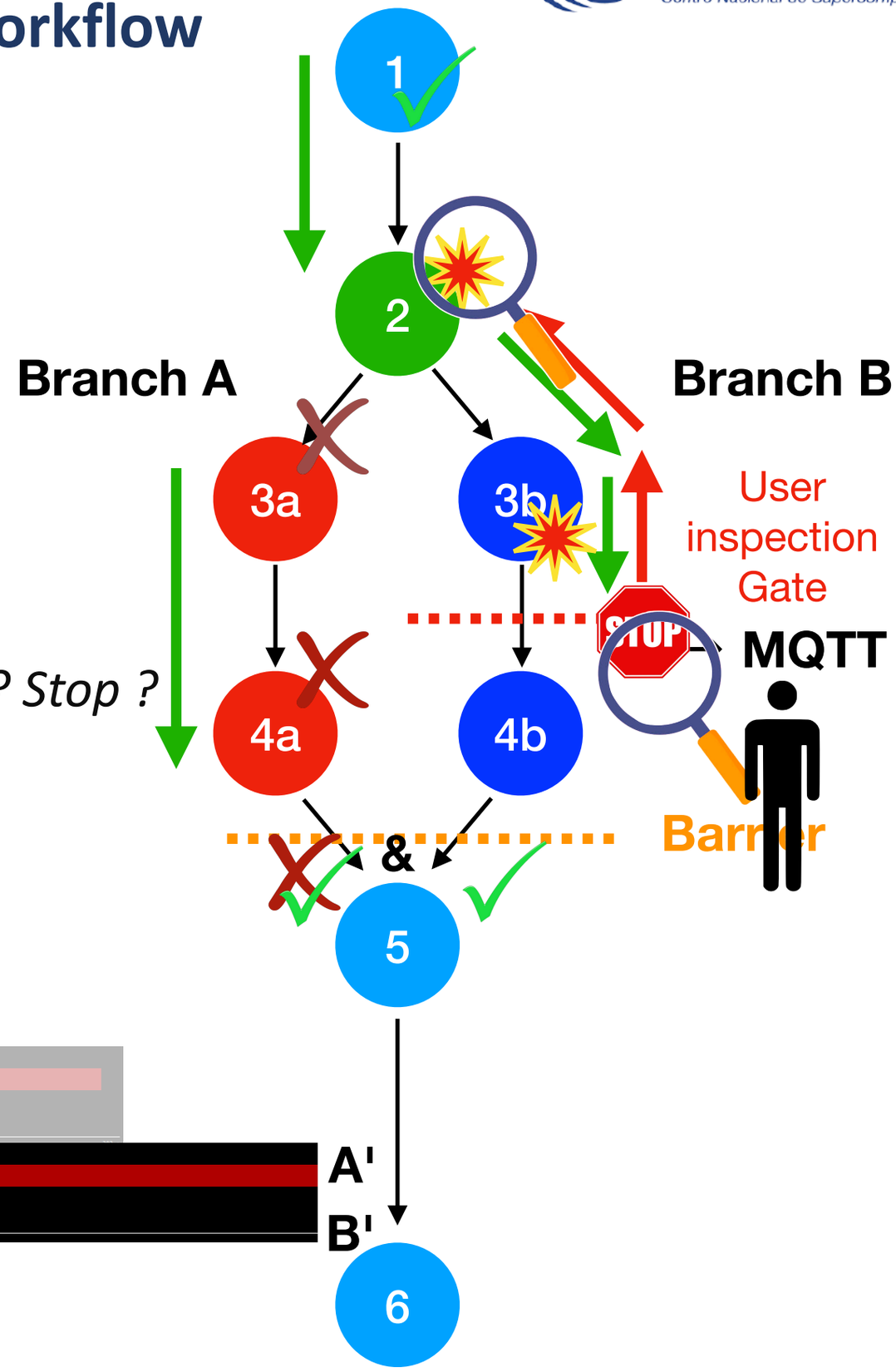
E.g: self-cal loop + stopping criterion

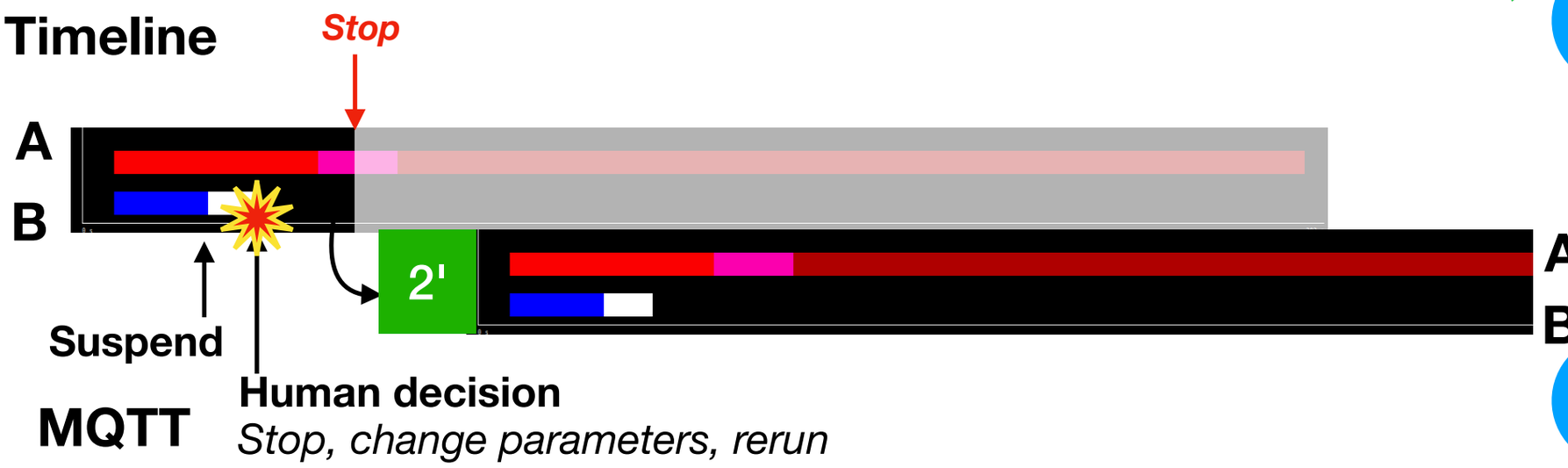
- Ability to suspend/resume workflow on logic

Criterion met: Continue? End after current task? Stop ?

- Ability to rerun branches

"Goto"





What's next for use case C?

- **Deployment on multiple sites (storage & computing)**

- **Implement more complex workflows & larger datasets**

- **Still need to test including DDFacet as a task**

- **Project extension** (→March 2026)

  Refine data catalog and interaction with the workflow

  Scaling up to real ~100s GB dataset

- **Implementation of Pulsar use case** (Cherry Ng)

  Not imaging data (Pulsar data format)

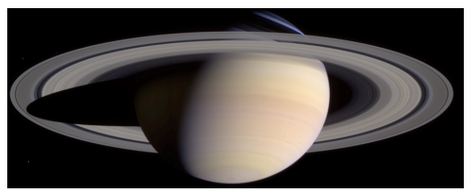
  Heavy GPU needs (pulsar dedispersion)

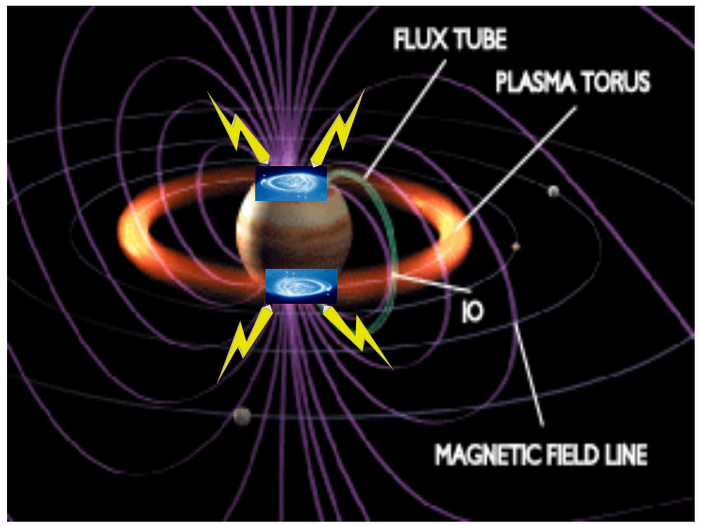
  Output products: Pulsars parameters

TASKA Overview

- **Use Case A:** Early detection and selective resolution data recording (space optimality)
- **Use Case C:** Workflow orchestration of interferometric data processing with a focus on improving the processing speed, accuracy and automation on large datasets
- **Use Case D:** Prototype development for “dynamic” imaging of the variable Universe (DL transient imaging)
- **Use Case E:** Advanced data reduction workflows for multi-dimensional real-time analysis and inference (joining A and C together)

Diversity of transient sources





Radio signatures

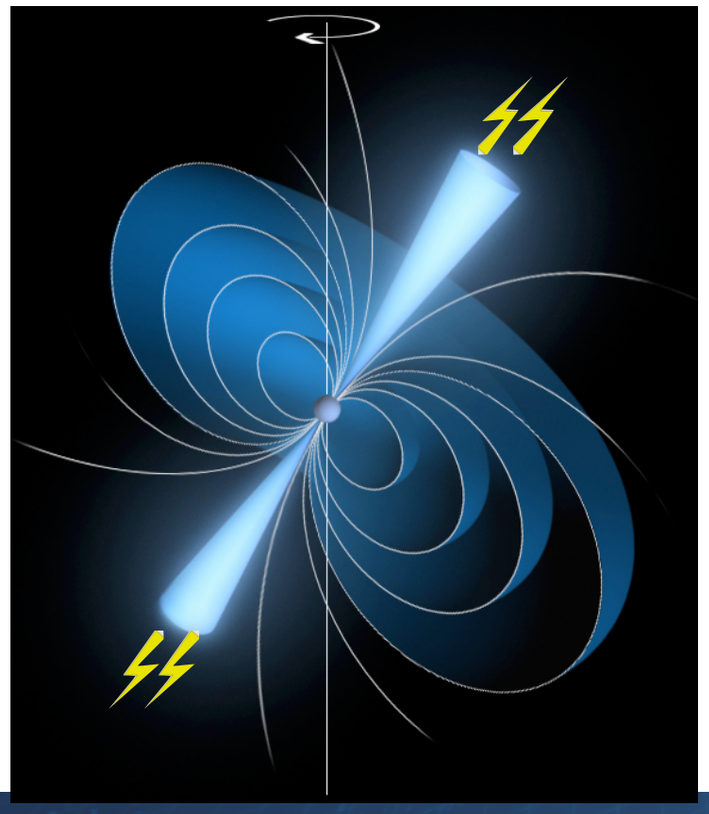
- mark the presence of magnetic fields

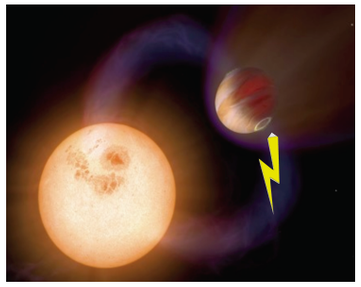
- have a rich spectral and temporal features

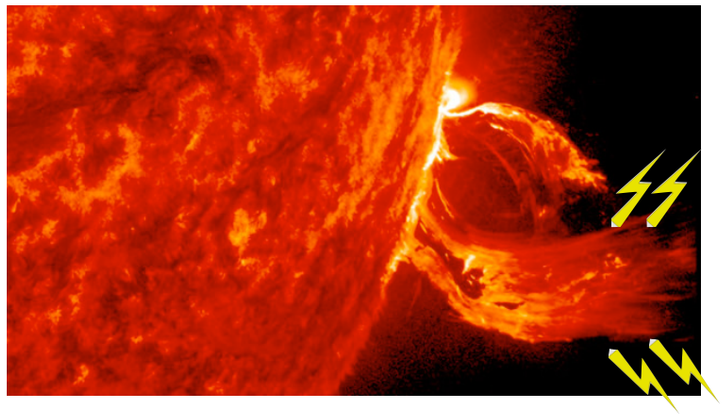
- associated to high energy events

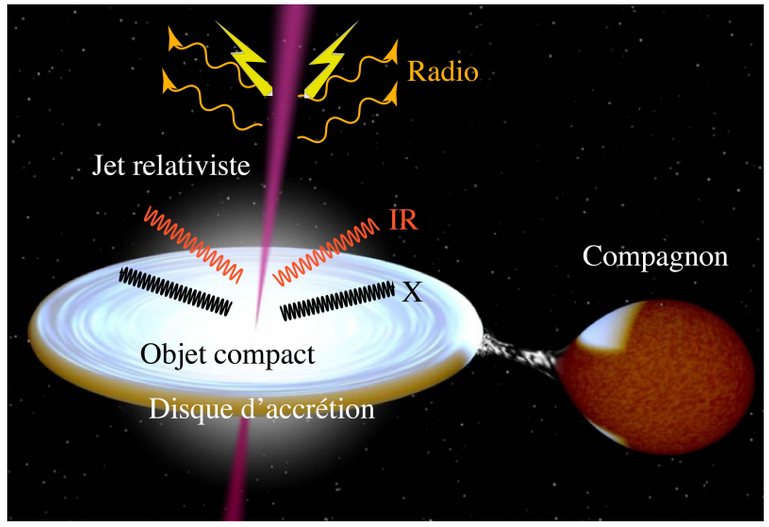
at all scales in energy, distances, durations…

<< 1ms 1s 10s 1m 10m 1h 1d 1 w 1 mo t





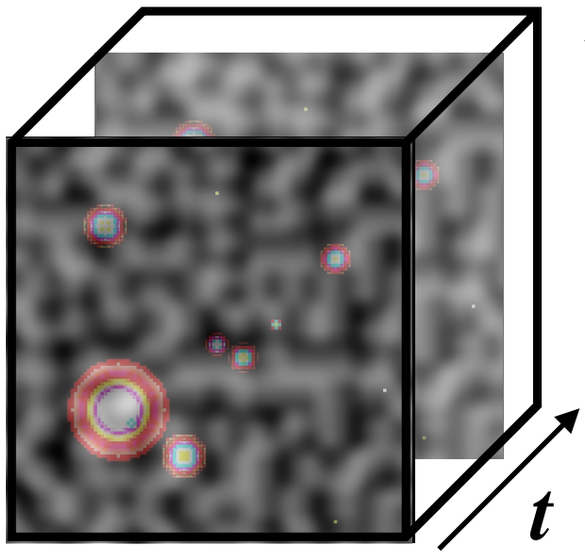




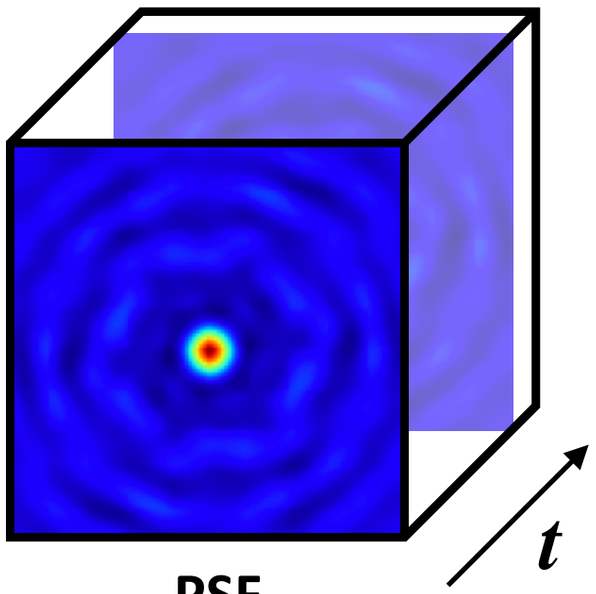
Data Model

$$Y_t = H_t * X_t + N_t$$

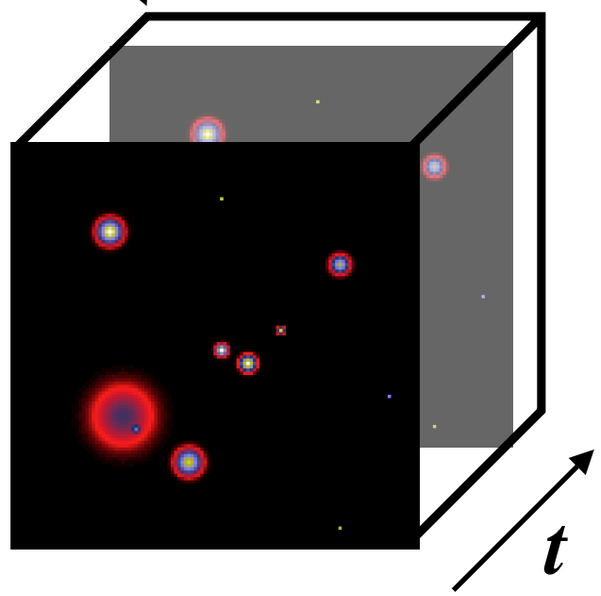
Correlated



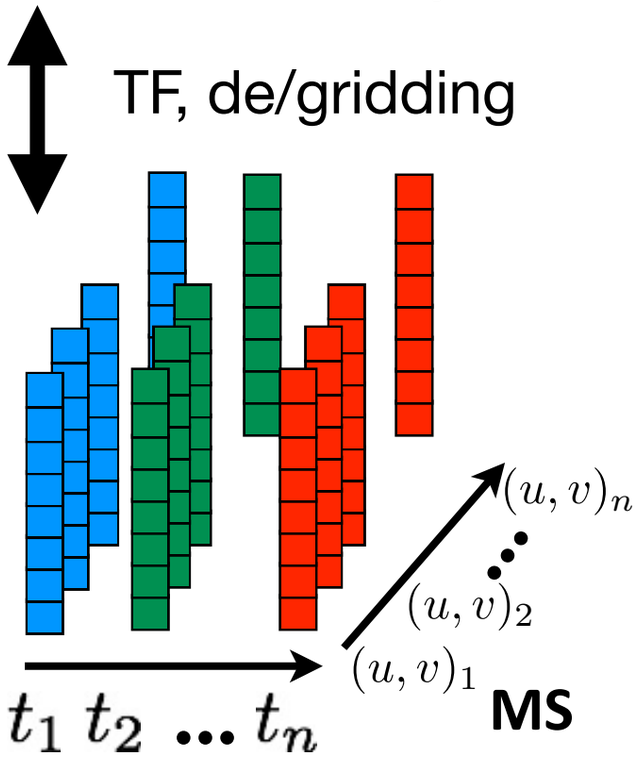
Dirty



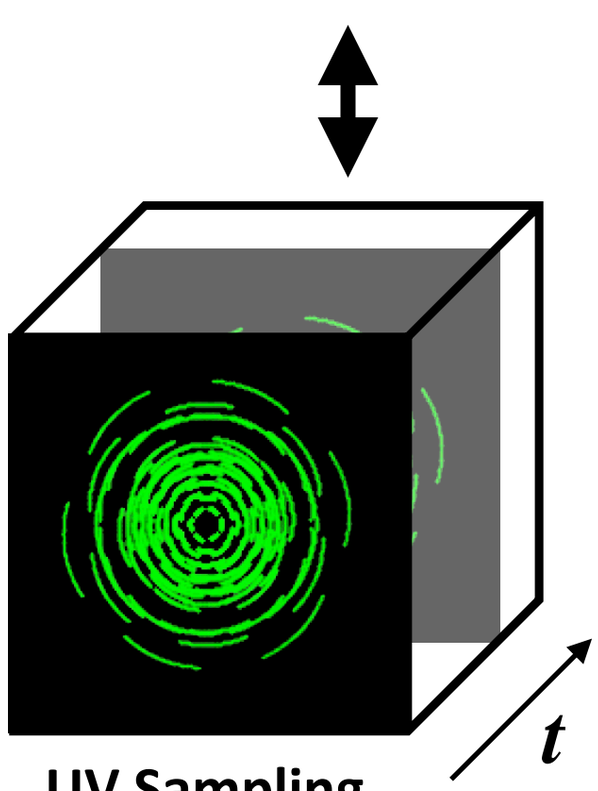
PSF



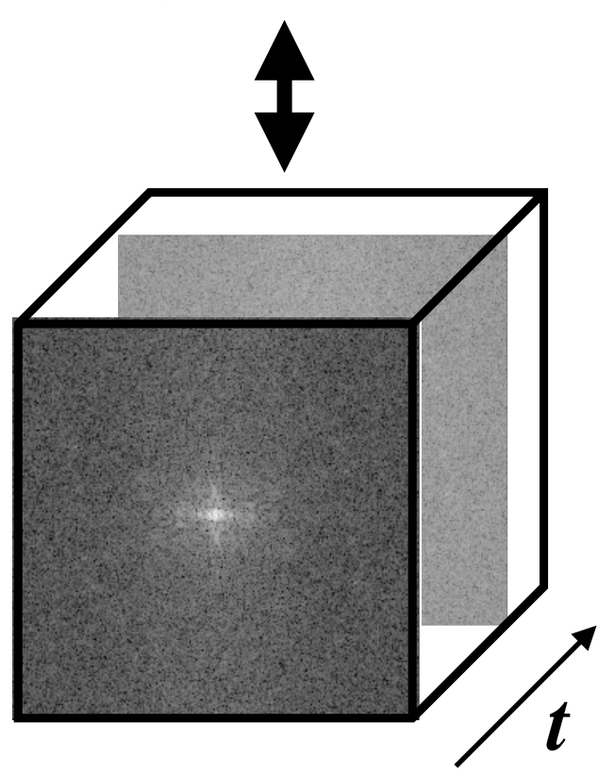
Sky model



Visibility data

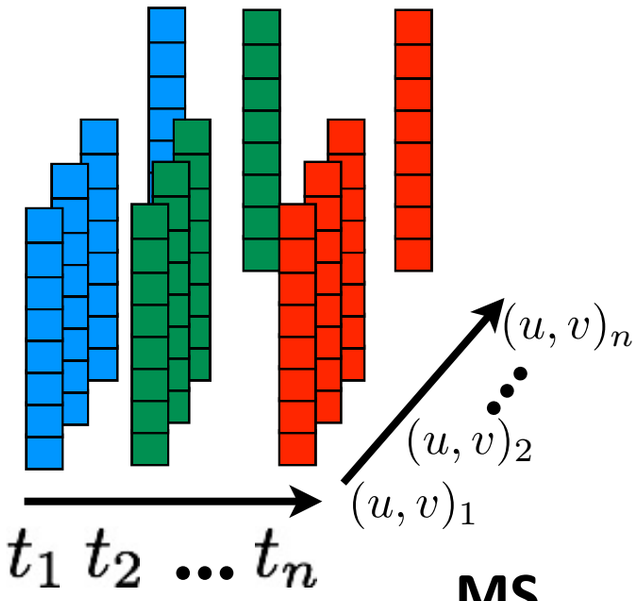


UV Sampling



Visibility function

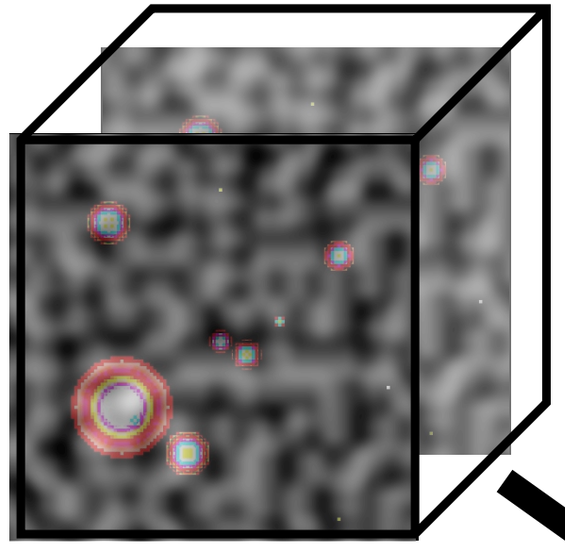
Use Case D



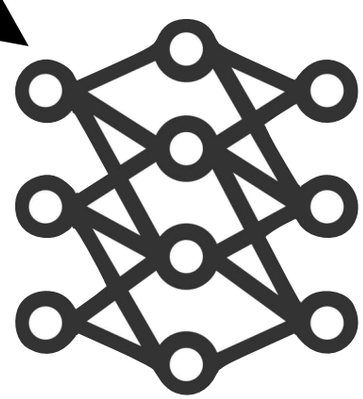
Raw data

MS

Data Gridder



Dirty cubes



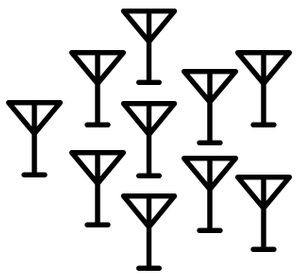
DeflationNet

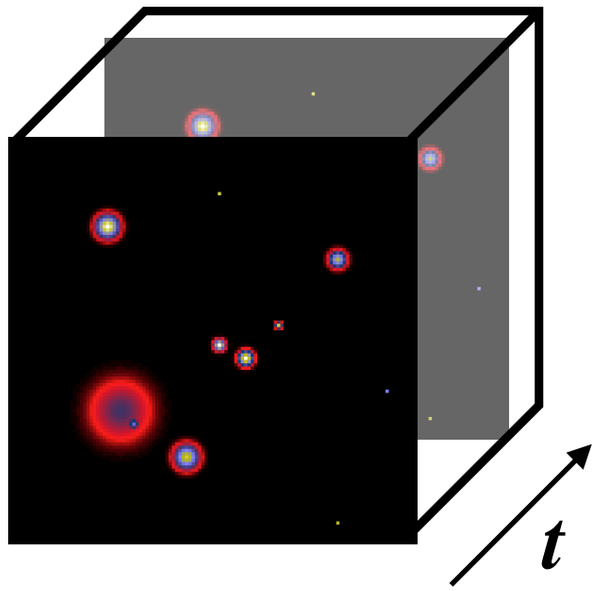
Chiche, Girard, et al., 2023

Nammour, Girard, in prep

Trained to reconstruct a dynamic sky

Observation simulator





Sky model

Preliminary training

Training time : ~27 hours

Learning rate : 1E-4

Loss : mean MSE

Number of epoch : 100

Optimiser : ADAM

Precision : Mixed (FP16-FP8)

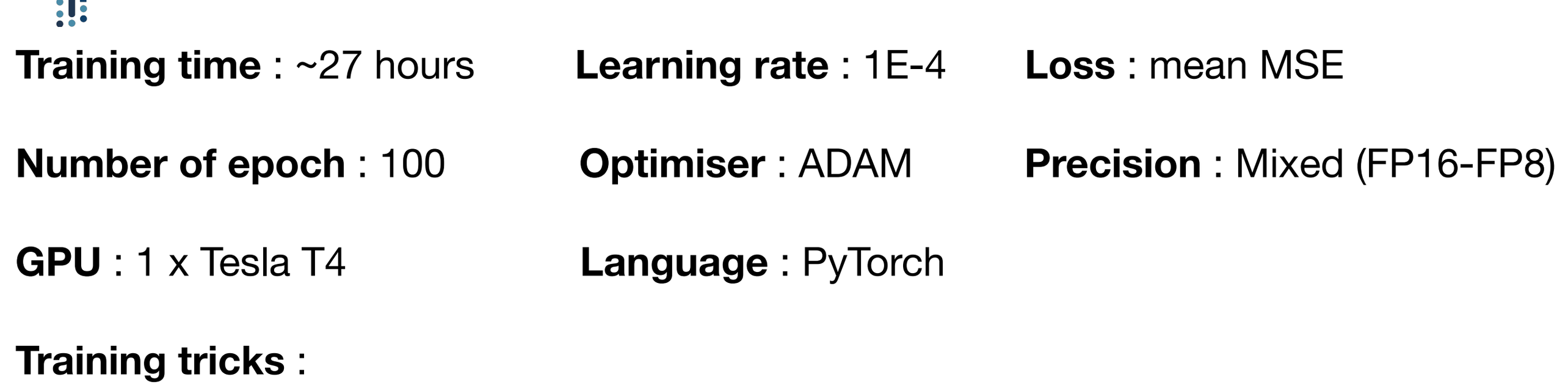
GPU : 1 x Tesla T4

Language : PyTorch

Training tricks :

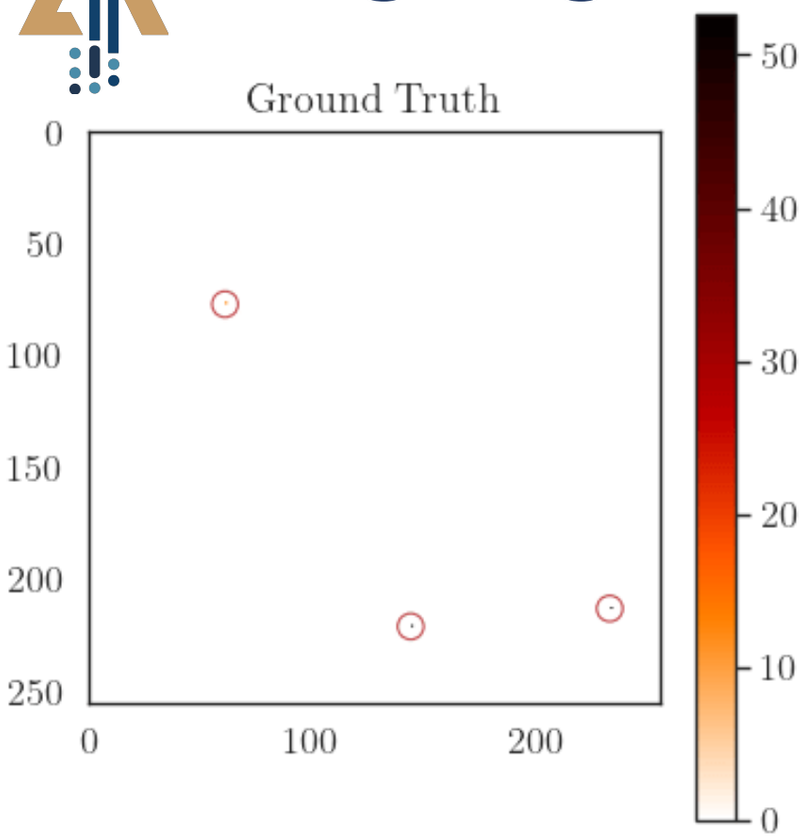
- Input flux normalisation factor : 100
- [Optional] Gaussian kernel convolution (7x7)

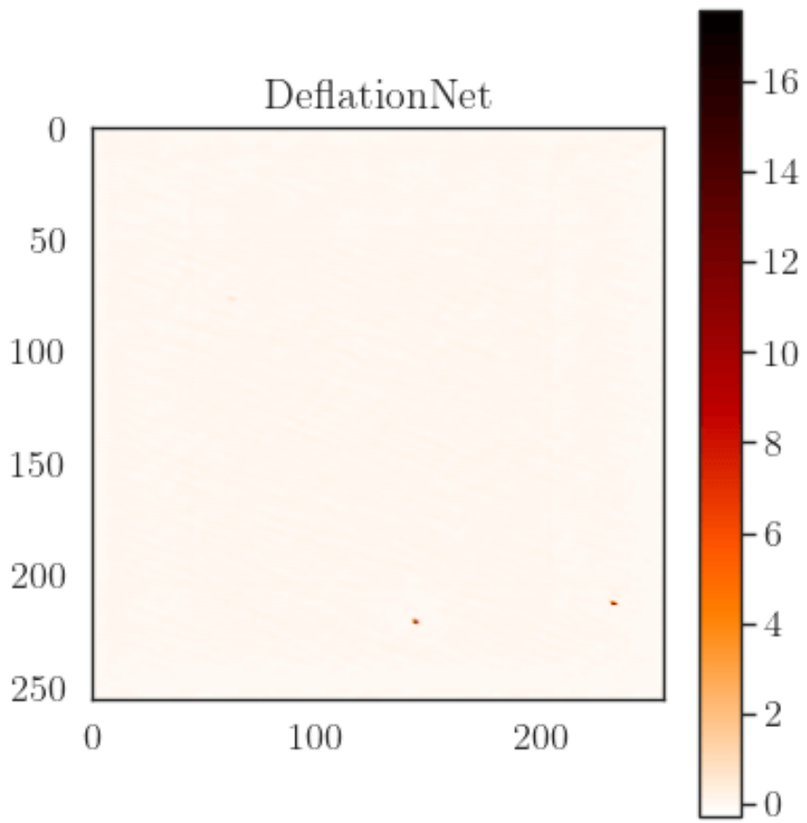
Start gentle training and increasing the complexity/generality

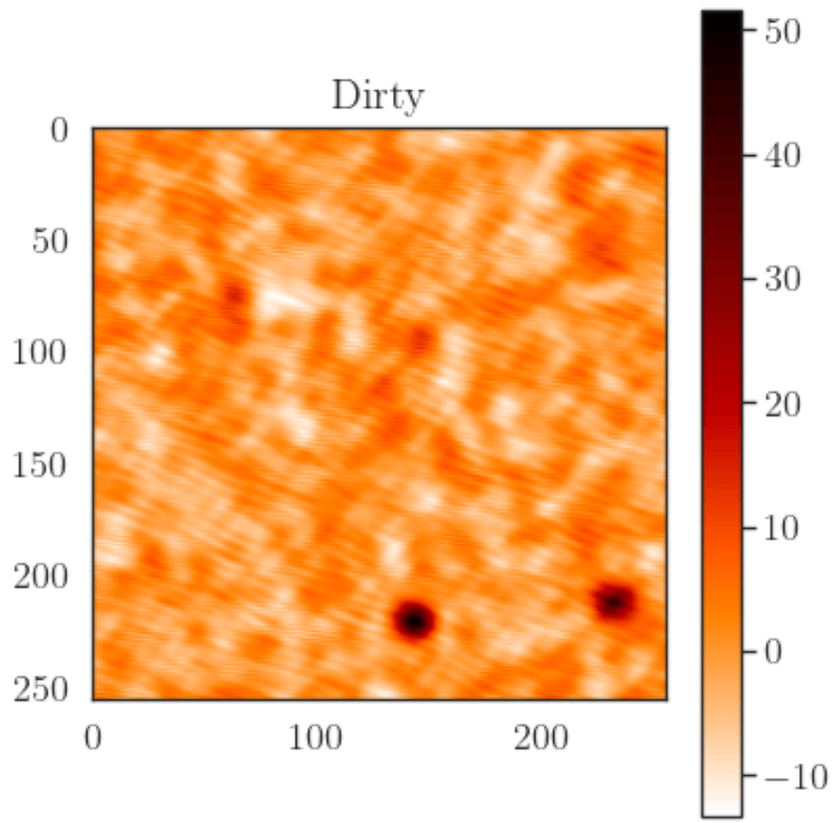


On-going results

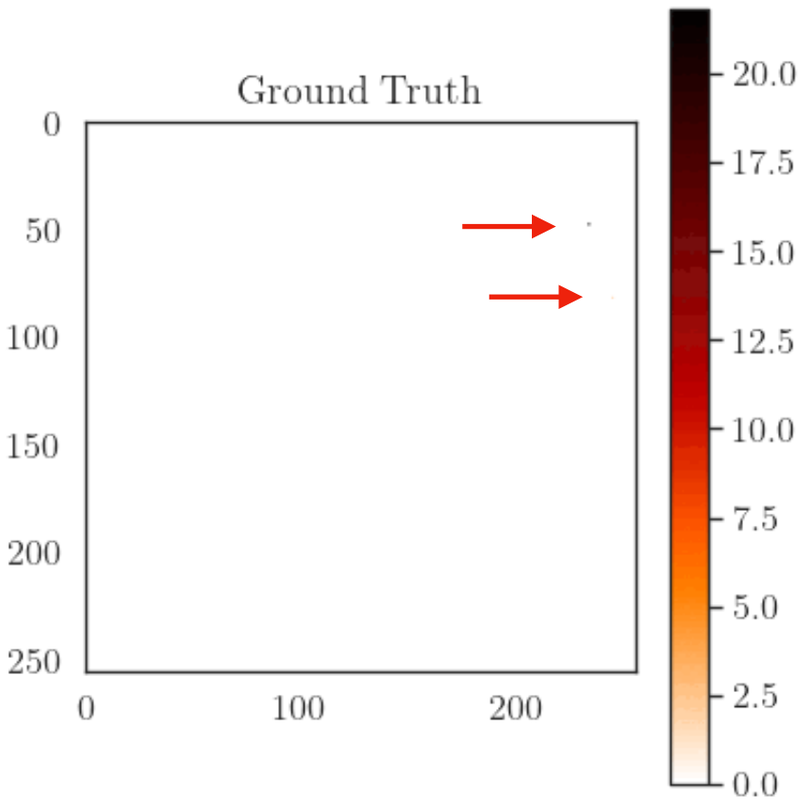
e.g. 2 constant sources 1 variable source

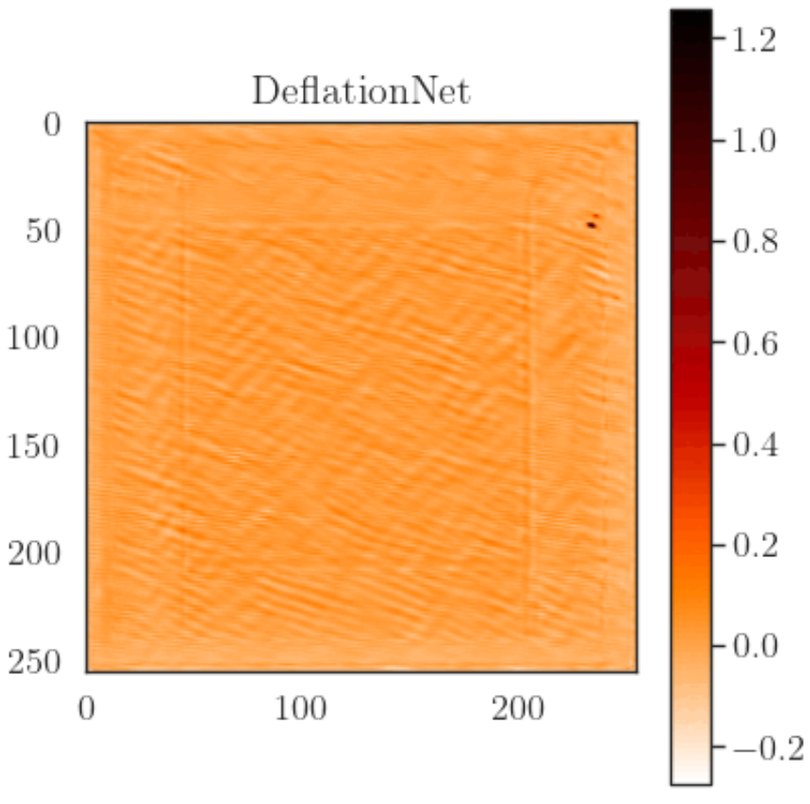


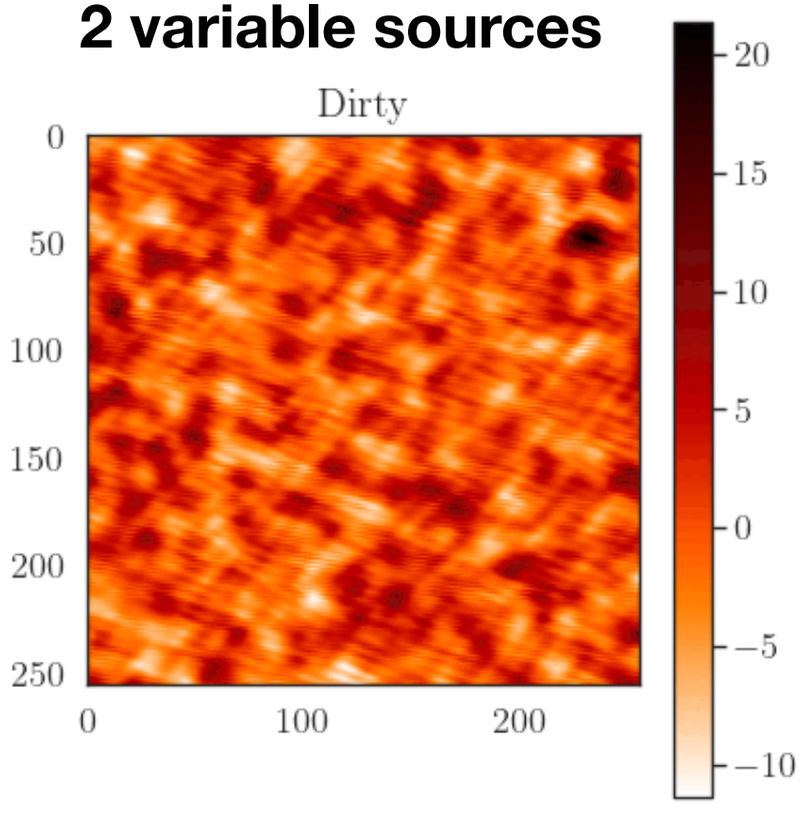




1 constant source
2 variable sources



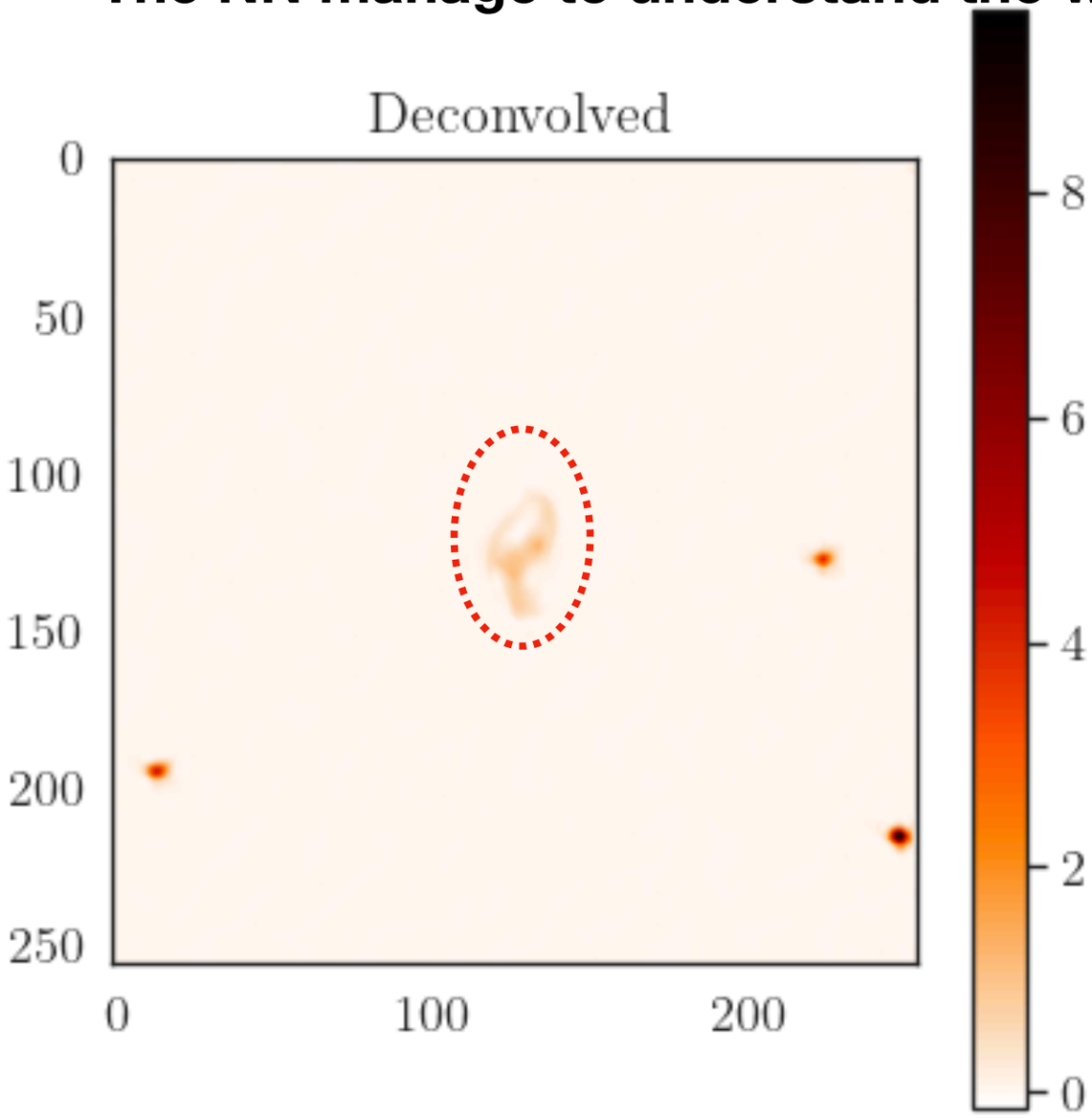


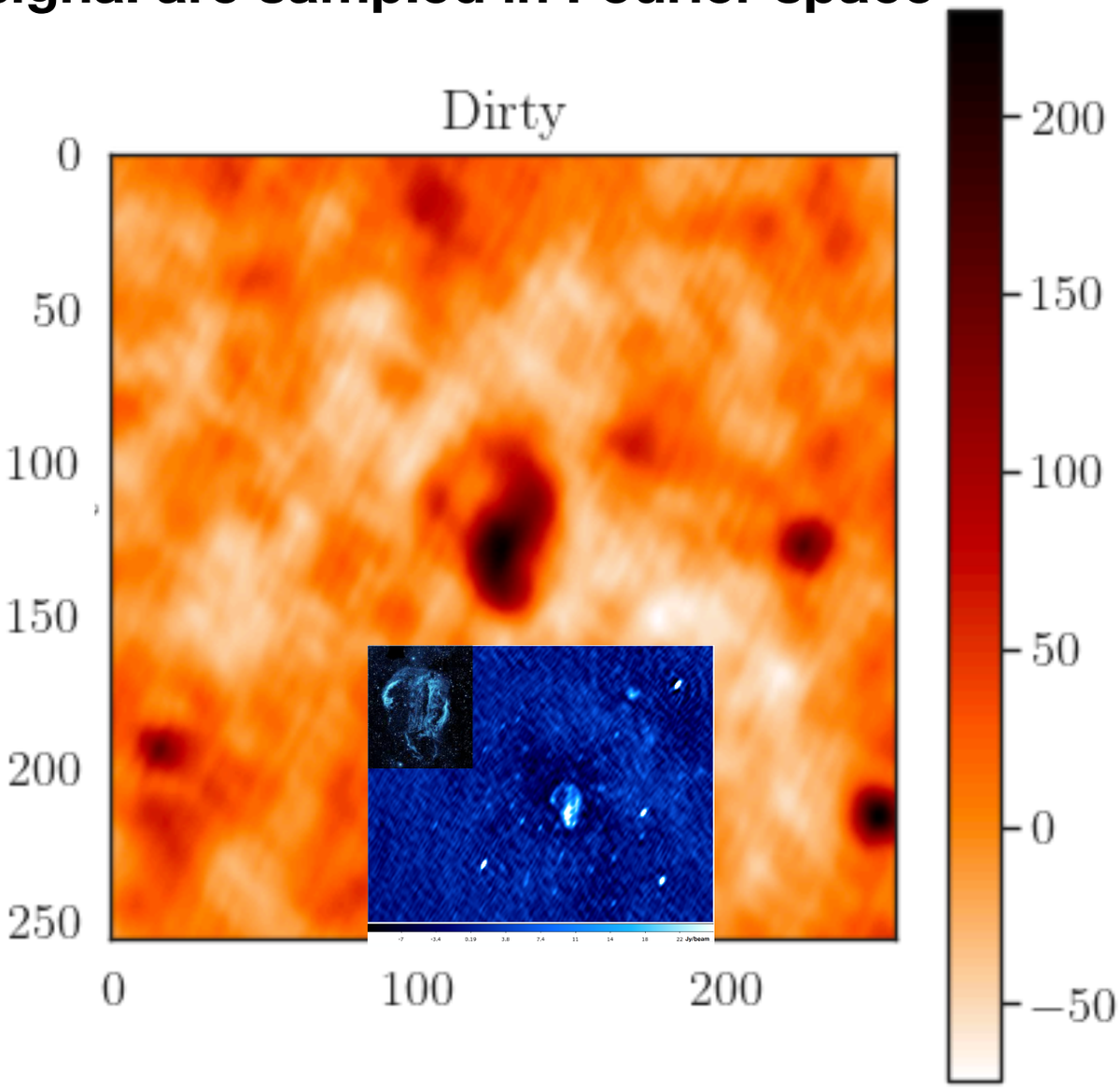


Crash test on Cygnus Loop

On-going

- The NN has been trained with constant and variables (gaussian) sources only
- Clues indicate the capability to deconvolve static and resolved sources
- The NN manage to understand the way signal are sampled in Fourier space





What's next for use case D?

- Complete code development for result analysis

- Optimise the solution (training, dataset, network)

- Adapt solution to real data

- Test different network architecture

- Test new use cases

- Generalize solution (different observation parameters, telescopes,…)

- Submit article

EXTRACT - TASKA - Summary

TASKA-A

- Real time detection (possibly with AI) on high resolution data stream (dynamic spectra): implemented on NenuFAR beamformer backend

TASKA-C

- We have developed a **framework for distributed data computing** on cloud clusters
- Currently validating
  - unsupervised/automated workflow
  - running a step on an HPC resource
  - running on a multi-cluster scale (data distributed in several data centers)
- Application on NenuFAR (SKA pathfinder)
- Clear huge potential for SRCNet

TASKA-D

- On going work on new imager for dynamical sources with AI-based video reconstruction

Questions?

fadi.nammour@obspm.fr

EXTRACT

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