

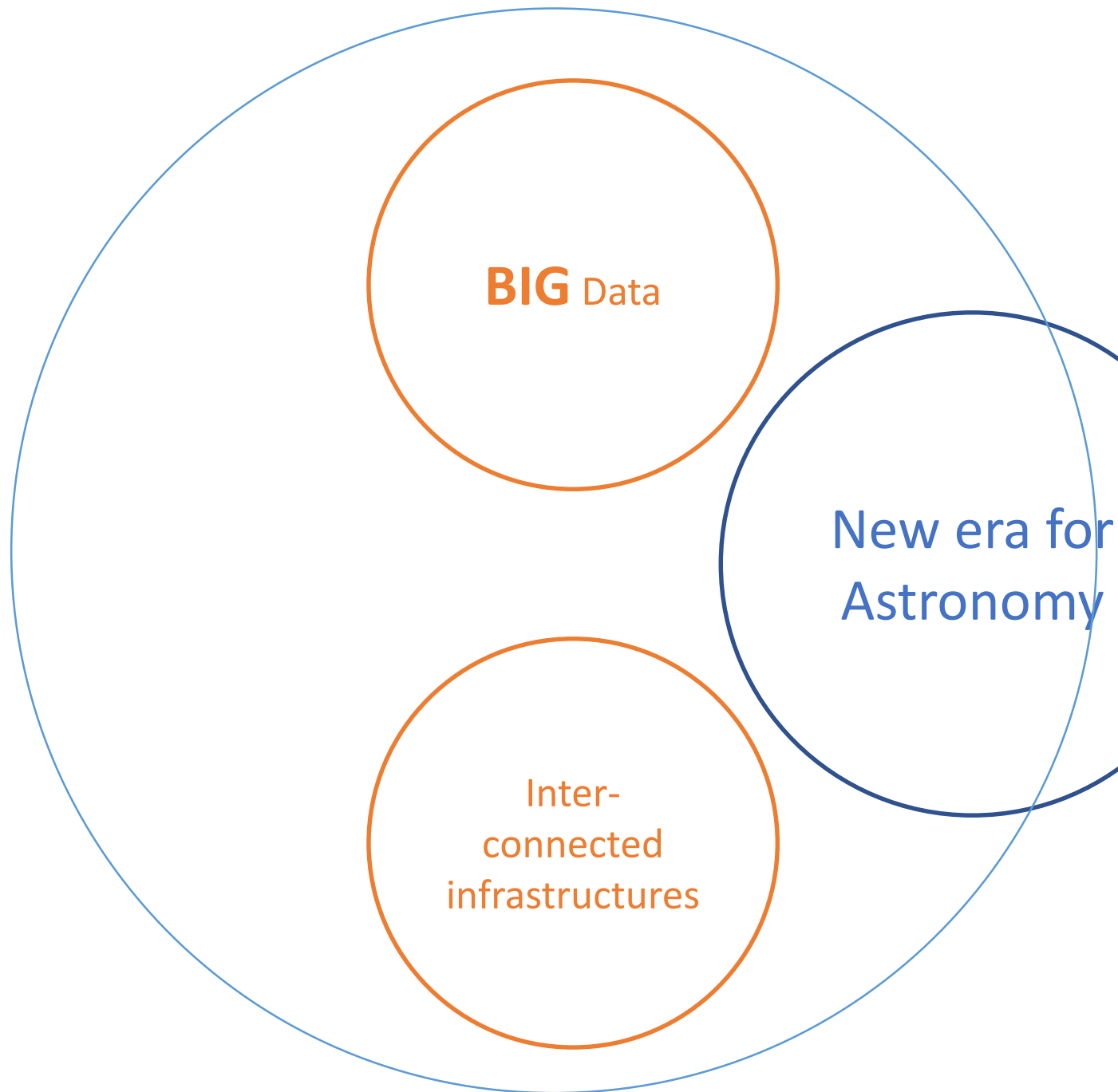
OBELICS

Observatory e-environments linked by common challenges

- towards a multi-messenger era

Thomas Vuillaume,
The new era of multi-messenger astrophysics,
Groningen, 28 March 2019





New data
management
Observatory E-environments Linked by common Challenges

- enabling **interoperability** and **software re-use** for the data generation, integration and analysis of the ASTERICS ESFRI and pathfinder facilities
- creating **new analysis methods** an open **innovation environment** for establishing **open standards** and **software libraries** for **multi-wavelength** and **multi-messenger data**
- developing **common solutions**
- **new skills** studying **advanced analysis** algorithms and software frameworks for data processing

The OBELICS galaxy

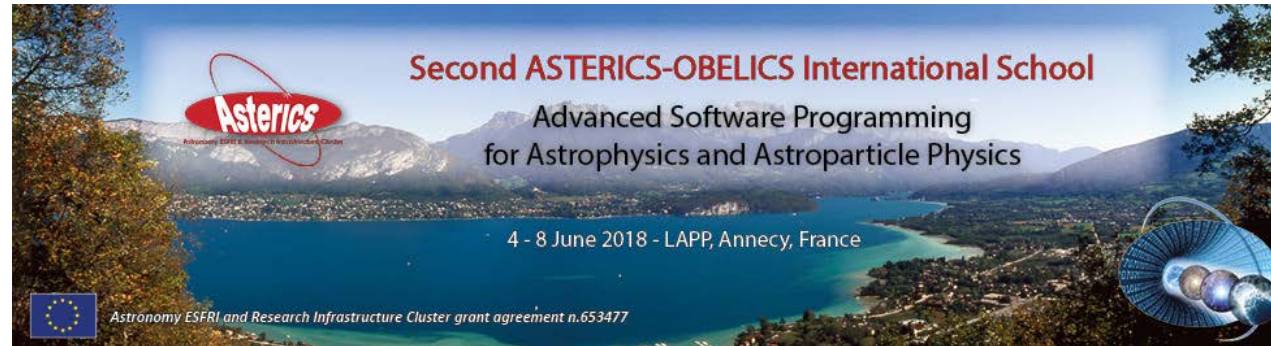


Building a community: workshops

- Three workshops bringing together representatives of the ESFRI and major industrial partners
- Creating a community to face the common data challenges
 - Invited talks
 - Tutorials and live demos
 - Panel discussions
- Creation of a machine learning community in astronomy and astroparticle physics



Training the community: schools

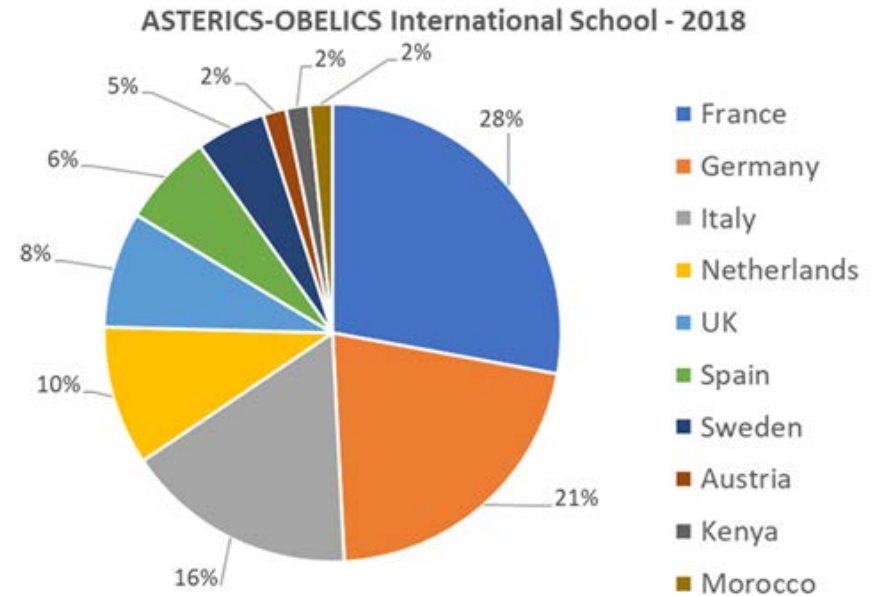


- Three summer schools “Advanced software programming for Astrophysics and Astroparticle physics”
- Python oriented
- Training young and seniors to develop good, reusable and efficient software together to tackle the data analysis challenges of the community



Training the community: schools

- > 60 students every year
 - From all over the world
 - From all communities
- Lectures
- Hands-on
- Tutors experts in all wavelengths and messengers
- Great feedback from the students
- **To be continued...**
 - **Third one in two weeks**
 - **And more with ESCAPE2020**



Covered topics:

- Good coding style
- Git
- Python data science libraries (numpy, pandas, scipy, astropy...)
- Data visualisation
- Machine Learning
- Julia

Addressing the challenges

–

OBELICS highlights, a non-exhaustive list

CORELib: A COsmic Ray Event LIbrary

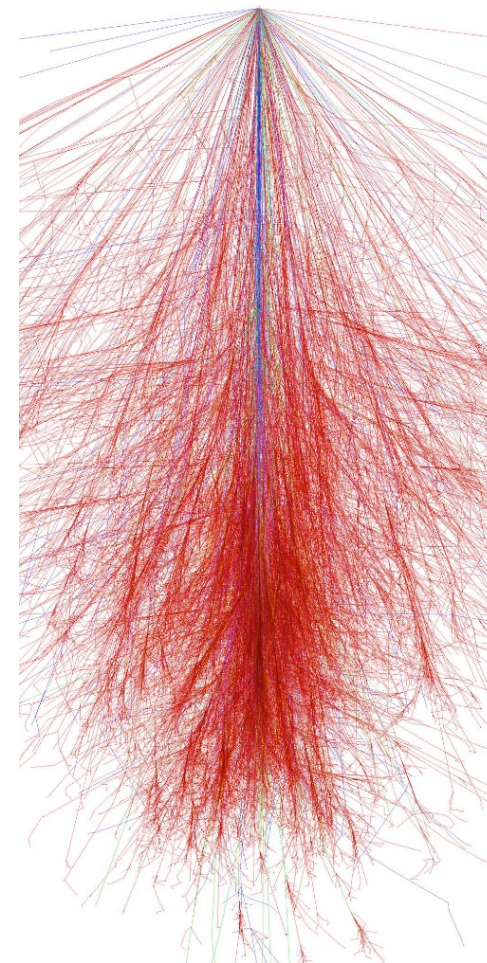
- Library of cosmic events generated by CORSIKA
- Can be used “as-is”
- Save a lot of computing resources

Pilot production
(available via SFTP)

Energy range (GeV)	Number of events
200-1000	10^7
10^3 - 10^4	10^7
10^4 - 10^5	10^6
10^5 - 10^6	10^5
10^6 - 10^7	10^4
10^7 - 10^8	10^3
10^8 - 10^9	10^2

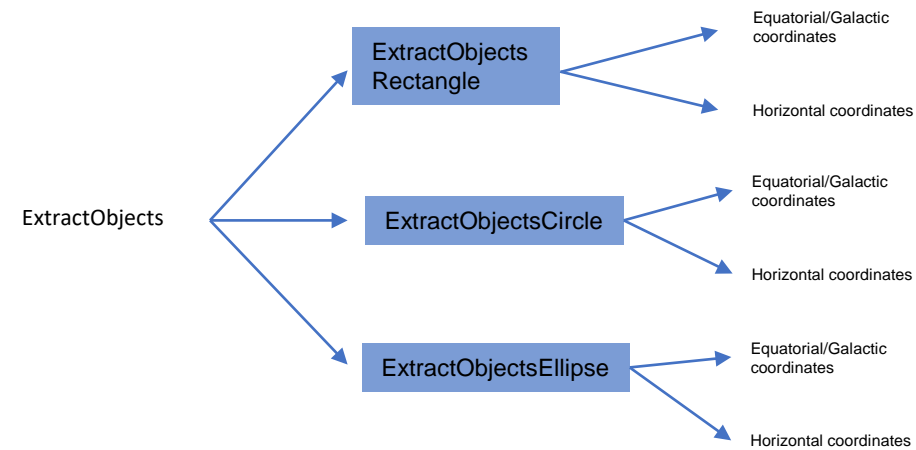
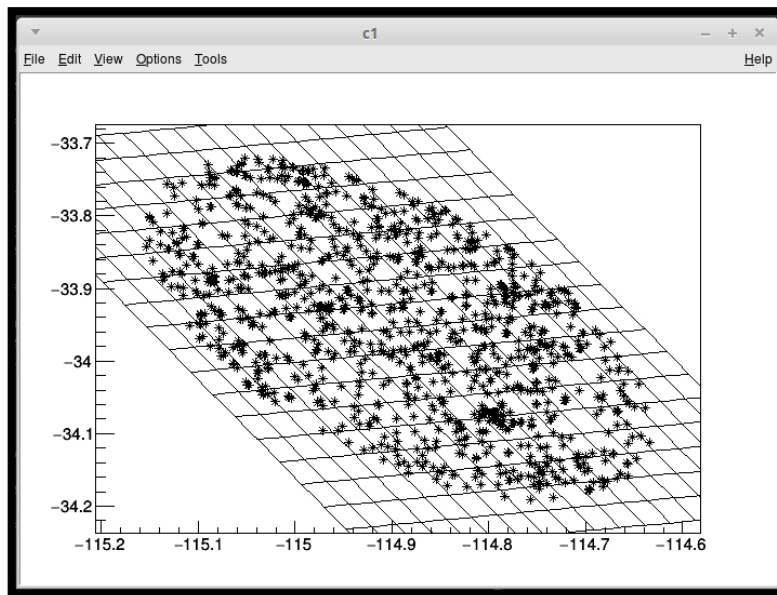
Full production
(accessible via GRID)

Energy range (GeV)	Number of events
200-1000	15×10^5
10^3 - 10^4	15×10^5
10^4 - 10^5	15×10^5
10^5 - 10^6	15×10^5
10^6 - 10^7	15×10^5
10^7 - 10^8	15×10^5
10^8 - 10^9	15×10^5



ROAst: ROOT extensions for ASTronomy

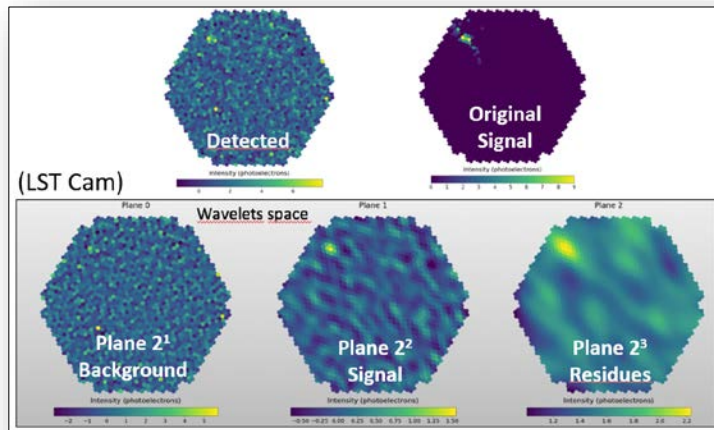
- extend the ROOT capabilities adding packages and tools for astrophysical research:
 - access to astronomical catalogues
 - coordinate conversion tools
 - high-precision Moon and Sun position models relative to the Earth
 - graphical tools to produce commonly used plots (general and partial skymaps)



PyWI - Python Wavelet Imaging

- Adapt tools from CEA Cosmostat (www.cosmostat.org) used in Euclid and SKA to CTA
 - ISAP library (2D, 3D wavelet transform and filtering)
 - Bundle in a Python library

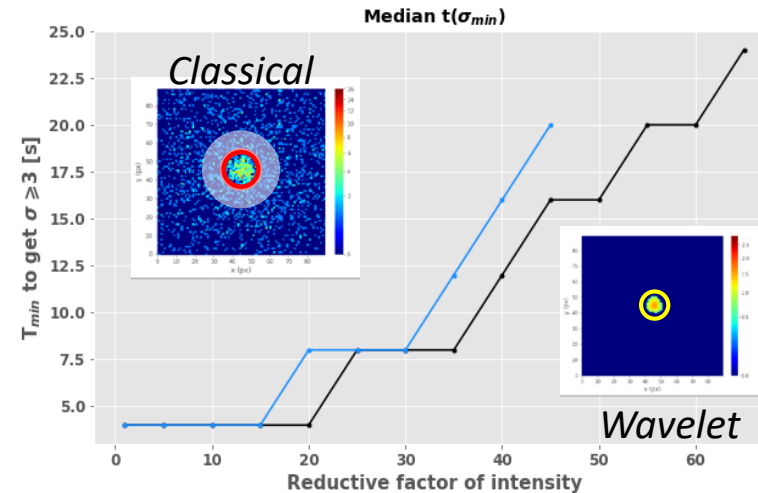
Camera images



- Night sky background filtering
- Sensitivity gain

→ Higher sensitivity and faster detection = better alert response and generation

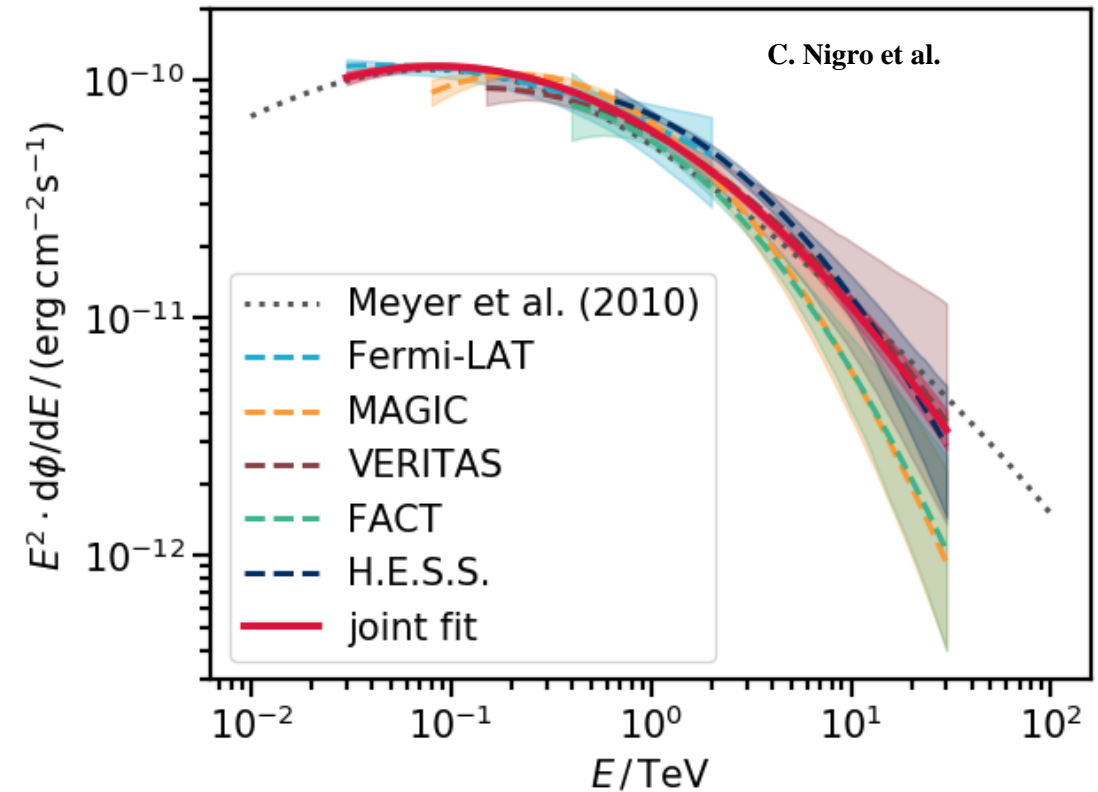
Sky images



- +50% on detection threshold
- Gain on detection time

A common data format for VHE astronomy

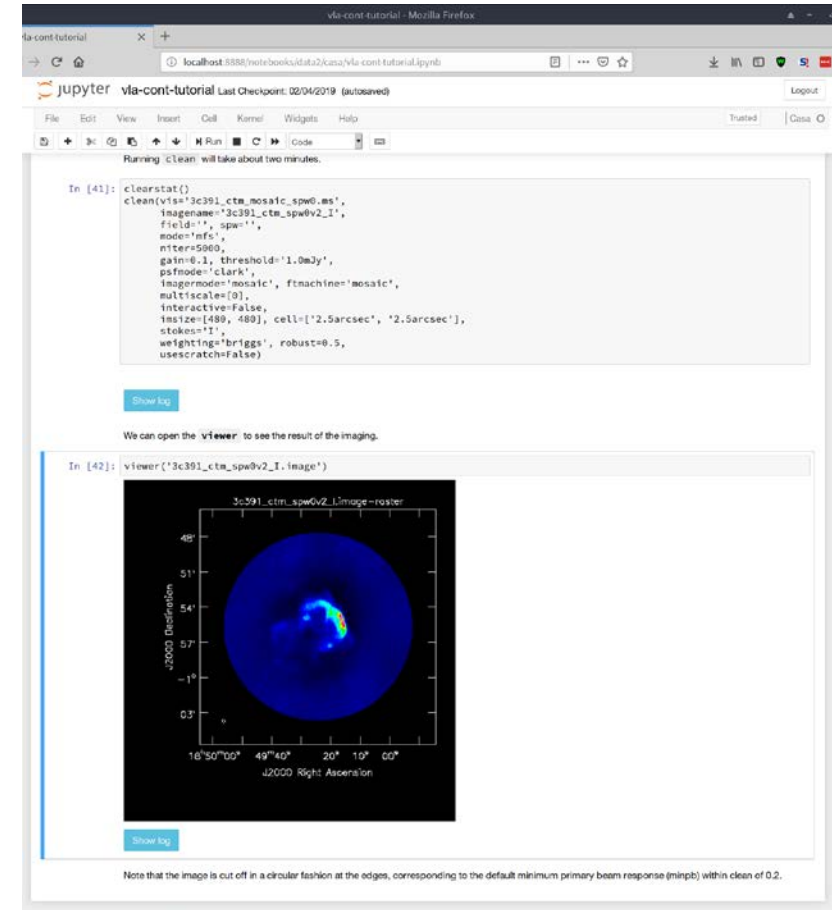
- Current IACT data and software are mostly p
- CTA will operate as an **open observatory** wit
- Development of a common data format:
 - event lists (photon-like events) + instrument re
 - in FITS
 - Discussed openly at <https://gamma-astro-data>
- CTA science tool prototypes ***Gammapy*** and
- Conversion of current IACTs data to a commo
 - CTA science tools validation
 - Legacy data that allow reproducible and multi-i
- First multi-instrument spectral analysis of the Crab Nebula:
 - Common data format + *Gammapy*



See Léa Jouvin's talk

CASA in Jupyter notebooks

- CASA: Leading data reduction package for Radio Astronomy (ALMA, VLA)
- JIVE recently added VLBI capabilities
- Jupyter notebooks allow easy remote execution where data is
- CASA data format (MSv3) also to be used for SKA.



VLA tutorial running as Jupyter notebook

CASA in Jupyter notebooks

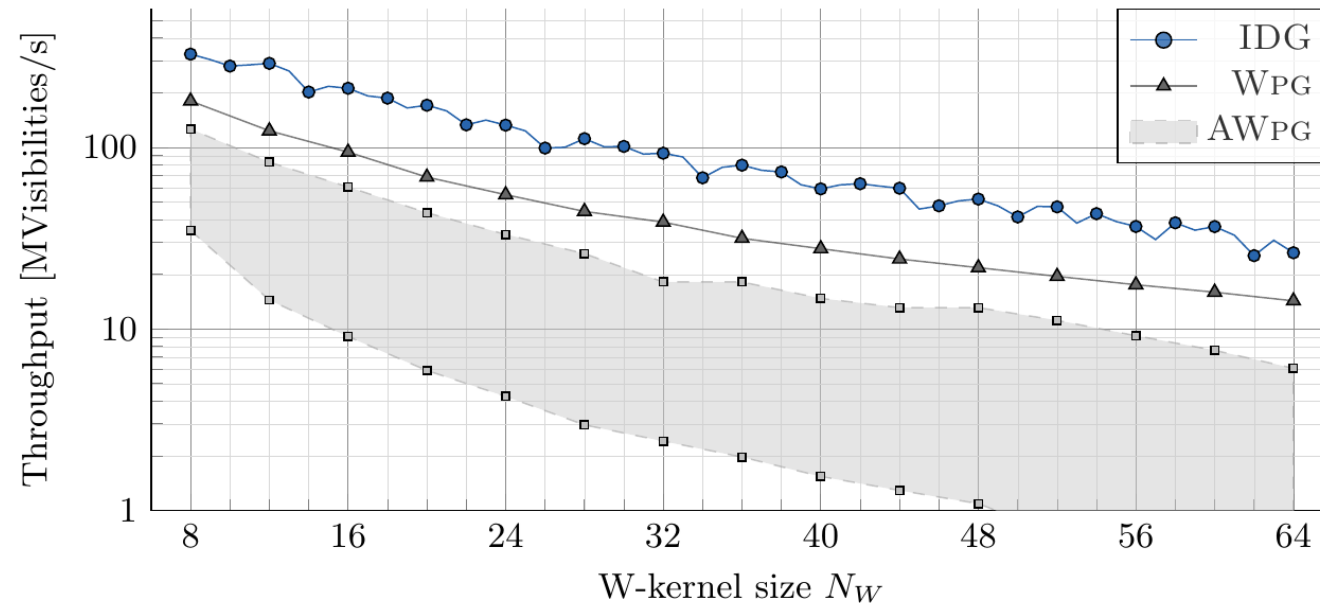
With Recipe Integration:

- Recipe: Minimal-recomputation for exploratory data analysis (Nikolic, Small and Kettenis) (also part of OBELICS)
- Allows skipping steps that have already been calculated when executing notebooks
- Available as docker and singularity containers
- <https://github.com/aardk/jupyter-casa>

Image Domain Gridding

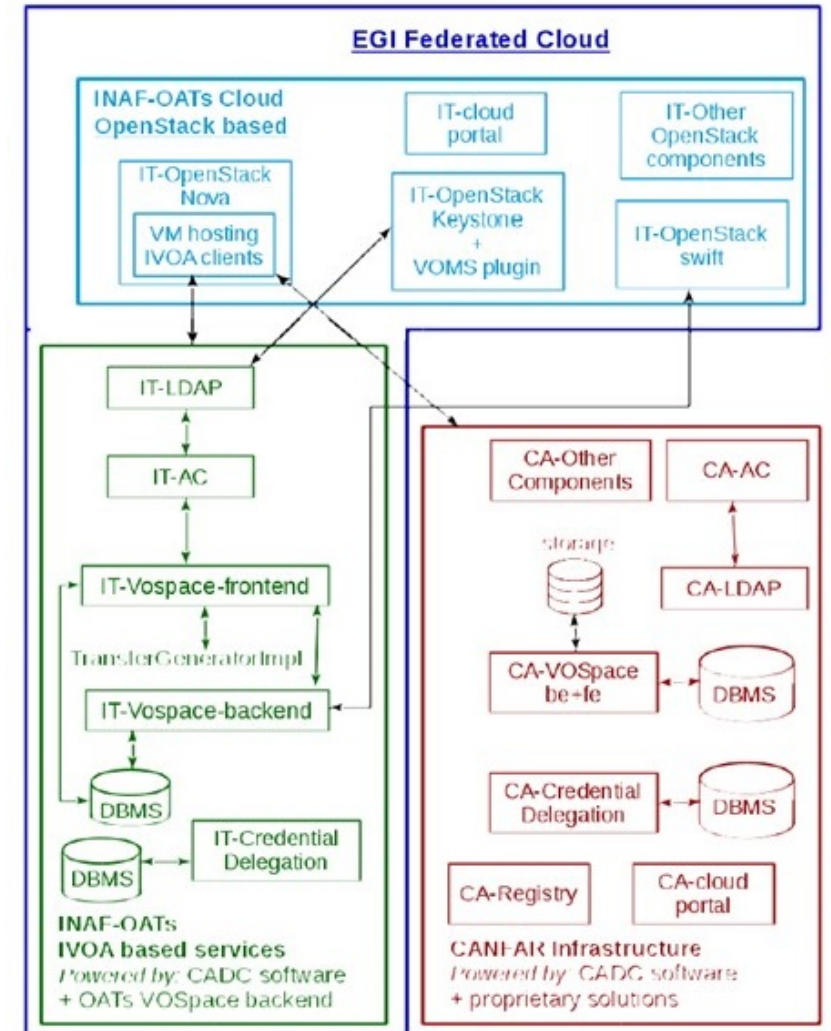
The Low Frequency Array (LOFAR) requires very computationally intensive data analysis

- Image Domain Gridding designed for massively parallel hardware
- Optimised for GPUs computations
- Integrated into WSClean (imager)
- Will be used in the preFactor pipeline currently commissioned by the Radio Observatory



Authentication & Authorisation, Workflows

- Evaluation of available workflow systems
- Evaluation of available A&A systems
- Design and implementation of several dedicated workflows embedding A&A functions:
 - Data processing workflow for CTA
 - Science gateway as web-based tool for high-energy astrophysics
 - Data processing workflows at INAF-IA2 for GIANO and HARPS-N pipelines and studies for their integration within the CWL (Common Workflow Language) framework
- Deployment of tools for federation of geographically distributed clouds
- Coordination with EGI and IVOA for standardization



Fluorescence radiation in Corsika

- Extensive Air Showers (EAS) generate Cherenkov and Fluorescence radiation. Fluorescence is normally neglected in Cherenkov observatories
 - Wanted to check this assumption and investigate its possible uses
 - Implemented fluorescence in CORSIKA, the most widely used EAS code
 - Studied Fluorescence contamination in Cherenkov-signals
 - Checked first constraints for using Cherenkov telescopes in “fluorescence mode”
 - Code potentially useful for a large number of situations
- *Detailed simulation of the effect of fluorescence radiation in Cherenkov Telescopes observations*, D. Morcuende, J. Rosado, J.L. Contreras, F.

Data and Software preservation through Containerisation

Docker Images for Software Development

- Fully reproducible environments
- Isolated workspaces
- Reusable base images
- Integrated into GitLab CI Pipelines in KM3NeT

Singularity Images in Production Pipelines

- Seamless integration into the (file) system of the host computer
- Completely independent base environment
- Perfect fit for Grid computing with heterogeneous systems
- "Software installation" is basically just copying a prebuilt Singularity image
- Already used in ANTARES production pipelines and currently tested in KM3NeT

Data Provenance

- File history is part of the data preservation
- Provides historical records of the data and its origins, like
 - container IDs (e.g. Singularity Registry)
 - specific software and library versions
 - and additional parameters (e.g. command line arguments or configuration files)
- Data and Software stored together in a single image file
- Reproducibility of results through containerised analysis chain

Data format studies for CTA

- Benchmarks of data formats for CTA low-level data
 - Historical data format
 - New custom data formats for CTA data
 - HDF5
- Benchmarks of data compression for CTA low-level data
- Development of a new compression algorithm for digitalized signals from physics experiments
 - Test on CTA MC data = speed-up of 19 compared to LZMA (keeping the best compression ratio)
 - *Polynomial data compression for large-scale physics experiments, Aubert, P. et al 2018*

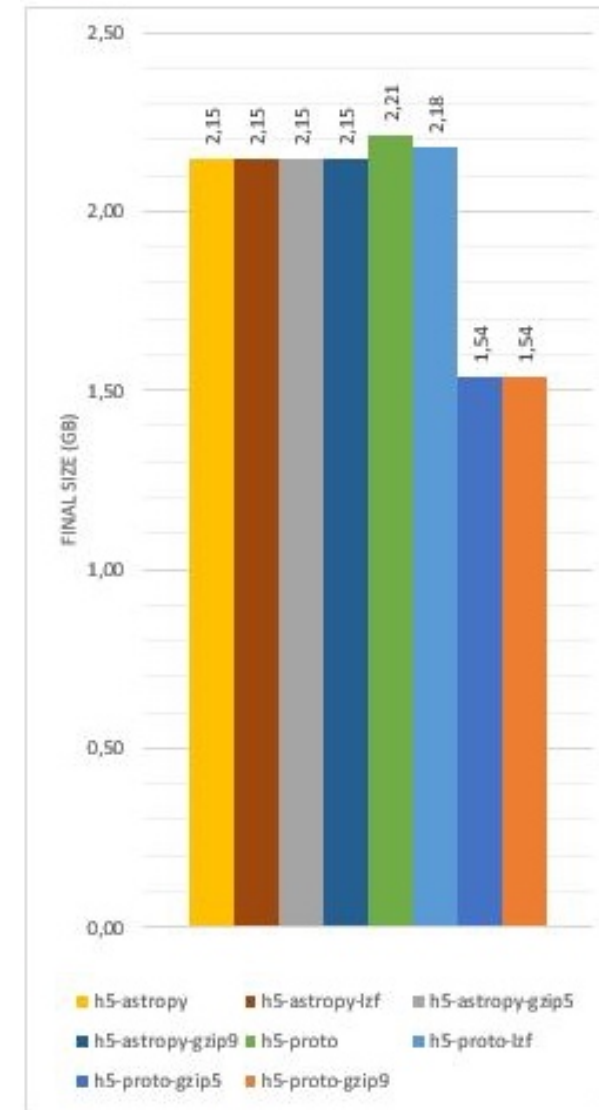
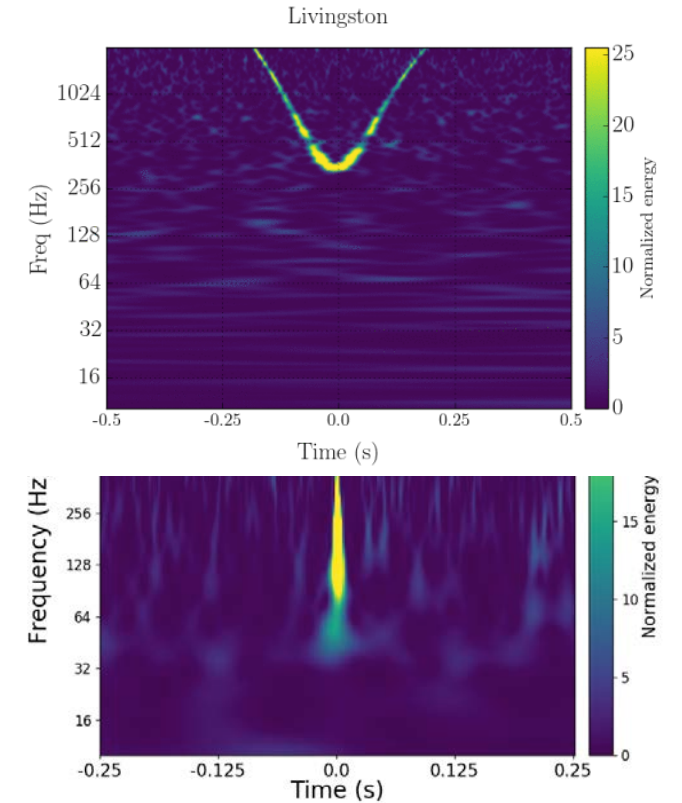
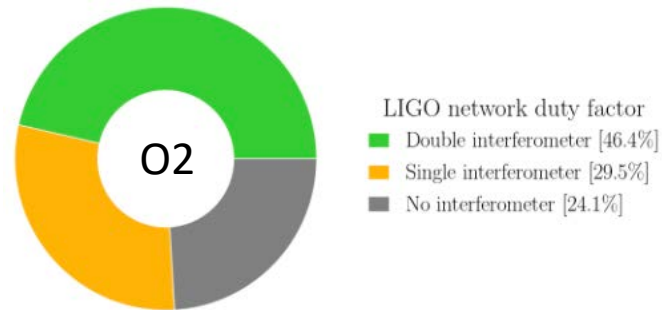
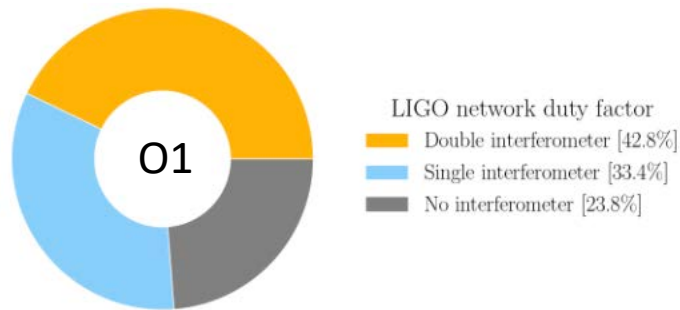


Figure 1: Final size.

Detecting Virgo's glitches with deep learning

- Study, identify and reduce the transient noise present in the gravitational wave detectors using deep learning techniques
 - Raw time-series as input instead of spectrograms
 - Both strain data and auxiliary channels
- Goal: analyze single-detector data
 - Detect more events



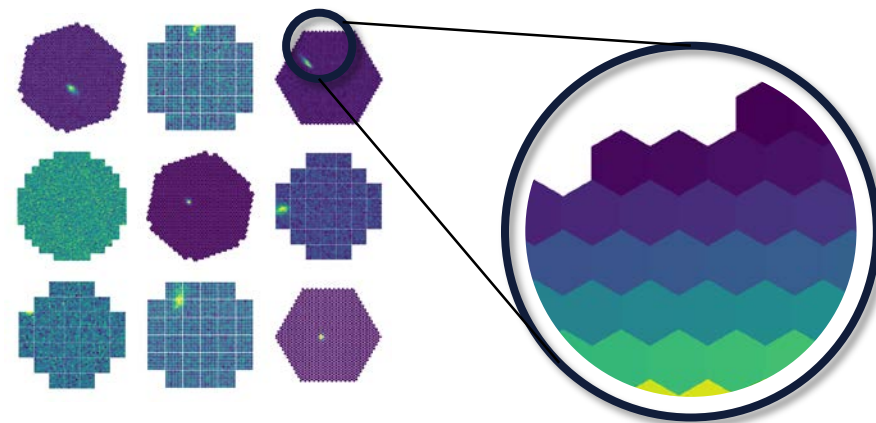
GammaLearn: Deep Learning for CTA

- Goal: Using state-of-the-art image analysis
 - Improve sensitivity and resolution
 - Fast reconstruction = fast alert generation

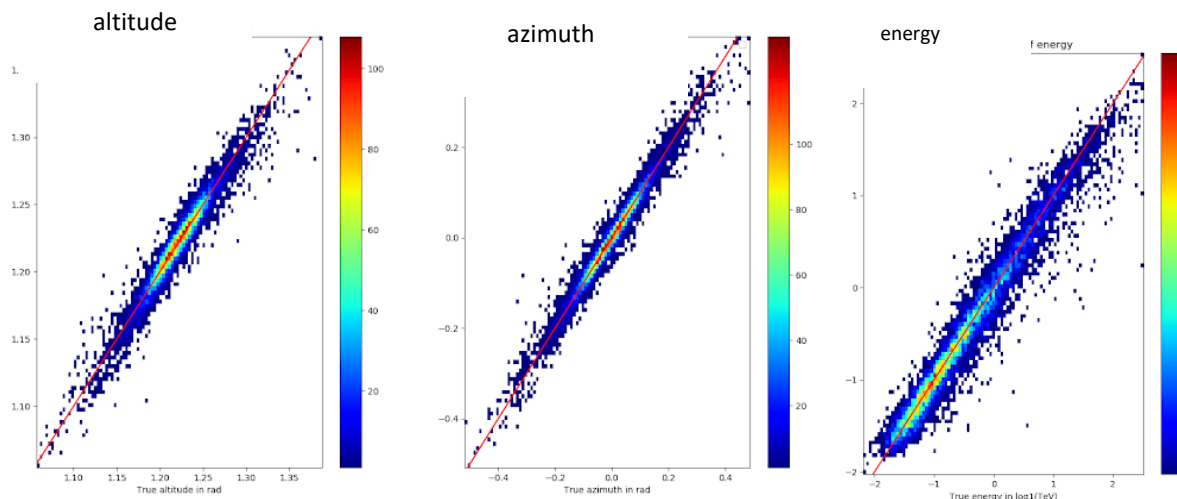
- GammaLearn framework to ease R&D and production

<https://lapp-gitlab.in2p3.fr/GammaLearn>

- The hexagonal problem



- Indexed Convolution package to deal with any pixel layout
- *Indexed Operations for Non-rectangular Lattices Applied to Convolutional Neural Networks*, Jacquemont, M. et al 2019
- <https://github.com/IndexedConv>
- Great interest for other physics detectors



And many more developments...

DEFAULT PRE-PROCESSING PIPELINE RELEASE 2016
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Default Pre-Processing Pipeline source code](#)
The LOFAR Default Pre-Processing Pipeline (DPPP) software reads and writes radio-interferometric data in the form of Measurement Set (MS) files.
AWIMAGER2 RELEASE 2016
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: AWImager](#)
AWImager is a software package for the automatic calibration and imaging of radio interferometry data.

CTA - AUTHENTICATION AND AUTHORIZATION INFRASTRUCTURE API RELEASE 2016
[Developer: INAF](#) | [Licence: Freely available](#) | [Website](#) | [Repository: CTA-AuthAPI source code](#)
The INAF CTA Authentication and Authorization Infrastructure API provides a set of tools and software widely used by the CTA community. An extended (though not exhaustive) list of tools provided by this technology embrace XANADU software package, GammaLib & ctools, Fermi Science Tools, Aladin, IRAF. The gateway is based on the Liferay platform. It provides a Workflow Management System (WMS) with a REST API.

CASA JUPYTER KERNEL RELEASE 2018
[Developer: JIVE](#) | [Licence: GNUv2](#) | [Website](#) | [Repository: CASA Jupyter kernel source code](#)
CASA Jupyter kernel is a subtask 1 of the D-PPPP project.

CASASYNTHESIS RELEASE 2016
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: CASASynthesis source code](#)
CASASynthesis is a software package for the automatic calibration and imaging of radio interferometry data.

DPPT RELEASE 2016
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: DPPT source code](#)
Extensions to DPPP, the streaming framework for radio interferometry data.

SUMMARY OF EXISTING SOFTWARE TO BE USED
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Summary of existing software to be used source code](#)
This package includes a set of machine learning tools that allow to classify those transient signals, in order to better understand their origin.

MACHINE LEARNING ALGORITHMS FOR TRANSIENT SIGNAL CLASSIFICATION FOR GRAVITATIONAL WAVE ASTRONOMY RELEASE 2016
[Developer: CNRS-APC](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Machine learning algorithms for transient signal classification for gravitational wave astronomy source code](#)
Gravitational wave observations are limited by a background of transient signals from instrumental and environmental origin. This package includes a set of machine learning tools that allow to classify those transient signals, in order to better understand their origin.

MW-INFERENCE RELEASE 2016
[Developer: UCAM](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: MW-Inference source code](#)
Software libraries for Bayesian and Neural Network, multi-wavelength and transient source detection and classification.

STOA - SCRIPT TRACKING FOR OBSERVATIONAL ASTRONOMY RELEASE 2018
[Developer: UCAM](#) | [Licence: Apache](#) | [Website](#) | [Repository: STOA - A Workflow Management System for Radio Astronomy source code](#)
STOA is a software package for the automatic calibration and imaging of radio interferometry data.

3 SOFTWARE TO BE USED
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Summary of existing software to be used source code](#)
This package includes a set of machine learning tools that allow to classify those transient signals, in order to better understand their origin.

repository.asterics2020.eu

JPP RELEASE 2016
[Developer: CNRS-CPM](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Jpp source code](#)
Jpp is a java-inspired collection of C++ classes and applications for PDF creation, multidimensional interpolation, function minimisation and plotting. Based on a much broader KM3Net software framework, developed by Maarten de Jong, it is released here in a more generic, experiment-independent format. The package uses a flexible, templated class structure, which avoids the need for a large number of specialized classes.

IRB - INSTRUMENT RESPONSE BUILDER RELEASE 2018
[Developer: CEA](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: IRB - Instrument Response Builder source code](#)
The INAF CTA science gateway aims at providing a web instrument for high energy astrophysics. It leverages open source technologies giving web access to a set of tools and software widely used by the CTA community. An extended (though not exhaustive) list of tools provided by this technology embrace XANADU software package, GammaLib & ctools, Fermi Science Tools, Aladin, IRAF. The gateway is based on the Liferay platform. It provides a Workflow Management System (WMS) with a REST API.

HIGH PERFORMANCE COMPUTING DATA FORMAT GENERATOR RELEASE 2018
[Developer: CNRS-LAPP](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: High Performance Computing data format generator source code](#)
HPC programming techniques are largely applied in many fields. New generation research projects require code programming for efficient processing of large data volumes. In this scope the optimisation of data model and data format is critical. A HPC data format is designed to help the CPU data pre-fetching. Data have to be contiguous, cache friendly and the data locality has to be preserved. The tables of data also have to be aligned on vectorial registers with respect to their types. Therefore, no peal is needed before a loop.

INAF CLOUD SCIENCE PLATFORM RELEASE 2018
[Developer: INAF](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: INAF Cloud Science Platform source code](#)
The INAF Cloud Science Platform explores a possible technological solution for large projects, to implement an integrated approach to data access, manipulation and sharing. In particular, in case of worldwide distributed collaborations, the platform provides a set of tools and software widely used by the CTA community. An extended (though not exhaustive) list of tools provided by this technology embrace XANADU software package, GammaLib & ctools, Fermi Science Tools, Aladin, IRAF. The gateway is based on the Liferay platform. It provides a Workflow Management System (WMS) with a REST API.

ROAST RELEASE 2018
[Developer: INAF](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: ROAST source code](#)
The ROOT analysis framework is one of the most used software for the analysis and indeed it is the "de facto" high-energy physics. The goal of the ROAST library (ROot extensions for Astronomy) is to extend the ROOT framework adding packages and tools for astrophysical research.

PLISA RELEASE 2018
[Developer: INAF](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: PLISA source code](#)
Parallel Library for Identification and Study of Astroparticles.

SMART RELEASE 2016
[Developer: CEA](#) | [Licence: ISAP licence - CeCILL](#) | [Website](#) | [Repository: SMART source code](#)
Sparse Methods for arrays of telescopes (name to be finalised). Library for image analysis based on sparse methods. Will be used for signal/background discrimination of atmospheric showers and then for source morphology studies on sky images.

PYWI (PYTHON WAVELET IMAGING) RELEASE 2018
[Developer: CEA](#) | [Licence: GPLv3](#) | [Website: http://www.pywi.org/](#) | [Repository: PyWi \(Python wavelet imaging\) source code](#)
PYWI is an image filtering library aimed at removing additive background noise from raster graphics images. The image filter relies on multiresolution analysis methods (Wavelet transforms) that remove some scales (frequencies) locally in space. These methods are particularly efficient when signal and noise are located at different scales (or frequencies).

SUMMARY OF EXISTING SOFTWARE TO BE USED
[Developer: ASTRON](#) | [Licence: GPLv3](#) | [Website](#) | [Repository: Summary of existing software to be used source code](#)
This package includes a set of machine learning tools that allow to classify those transient signals, in order to better understand their origin.