Identification of pulsar signals in large data streams during data acquisition

Hermann Heßling

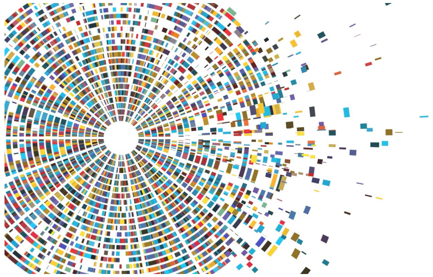
PUNCH TA3/WP3, March 17, 2025

ML-based Pipeline for Pulsar Analysis (ML-PPA)

Interdisciplinary team: TA5/WP4 + interTwin + DZA

Sagar Borra (DZA Görlitz)
Elsa Buchholz (DZA Görlitz)
Gautam Dange (FIAS Frankfurt)
Marc Drobek (DZA Görlitz)
Lars Haupt (DZA Görlitz)
Andrei Kazantsev (MPIfR Bonn)
Yurii Pidopryhora (MPIfR Bonn)
Tanumoy Saha (HTW Berlin)
Marcel Trattner (HTW Berlin)
Frank Bertoldi (U Bonn)
Hermann Heßling (DZA Görlitz)

Future Data Challenges



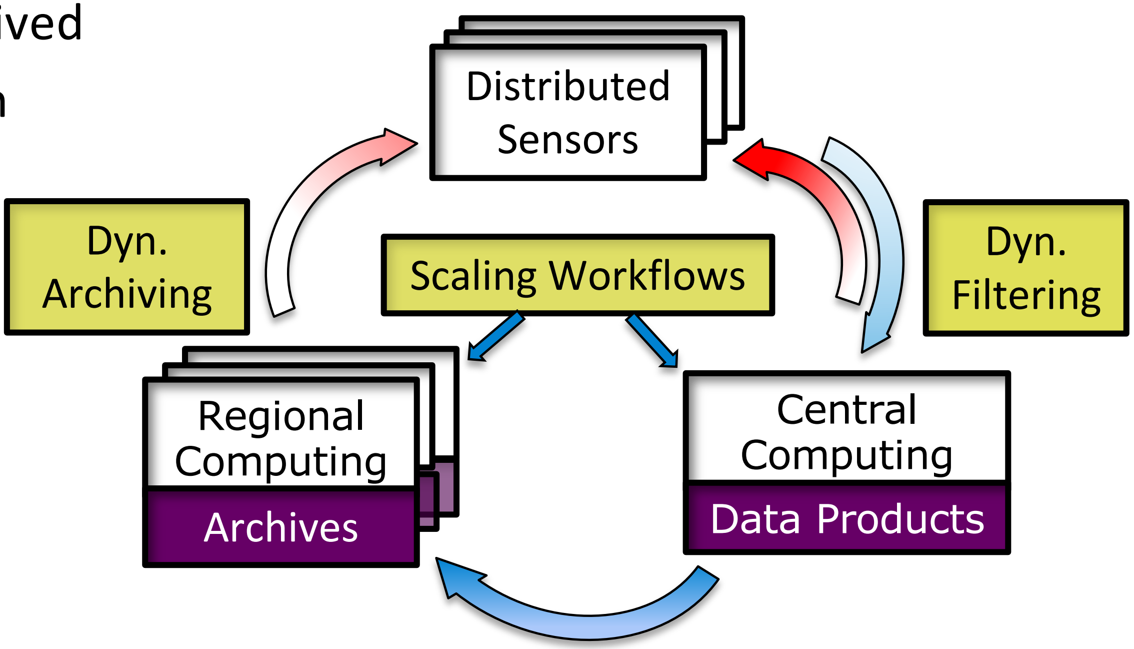
TA5: Data Irreversibility

- Only a tiny part of the data can be archived
- Decisions on what to keep are based on incomplete information

  ⇒ irreversible loss of information

Strategies

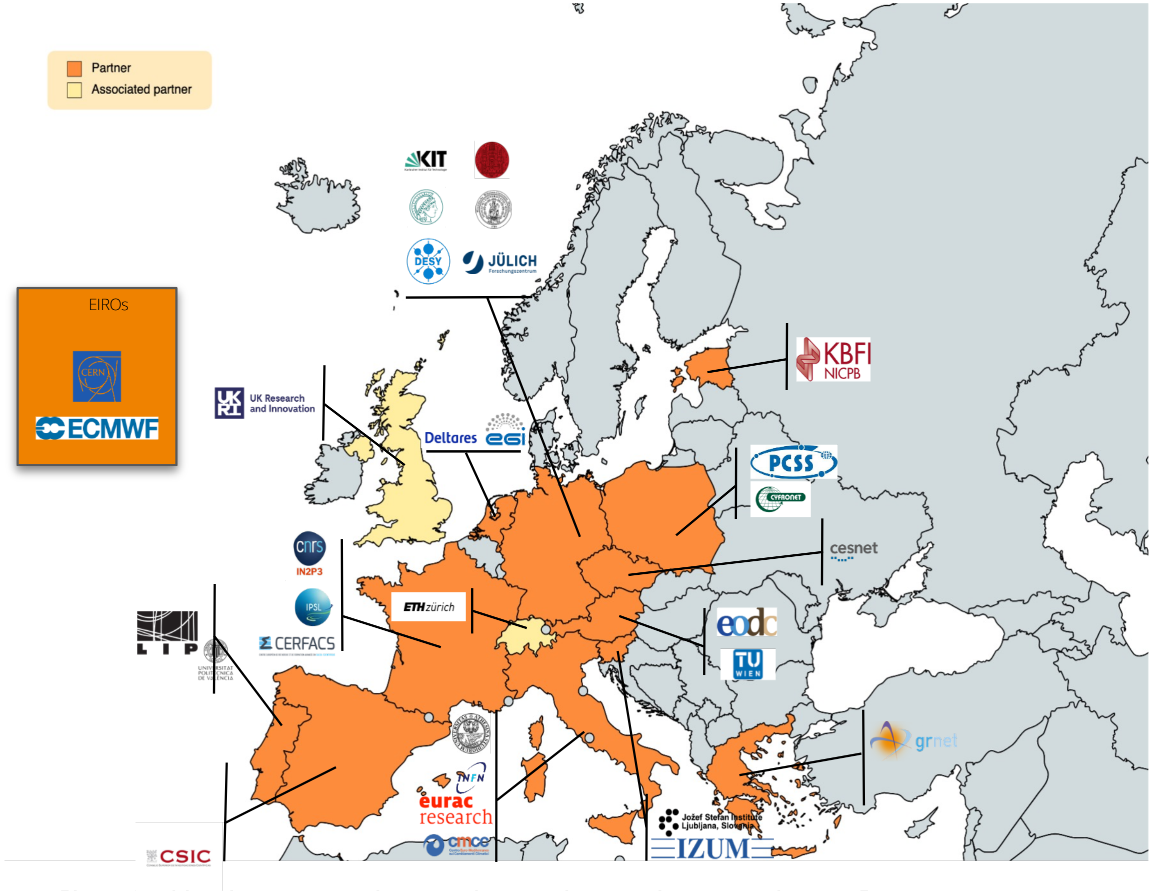
- Dynamic filtering
  - Sensor control in realtime
- Dynamic archiving
  - Feedback from offline workflows to sensor control in near-realtime
- Scaling workflows



Dynamic Life Cycle Model
[HH, M. Kramer, S. Wagner]

interTwin (2022-2025)

interTwin



[https://indico.lip.pt/event/1543/contributions/5132]

EGI Foundation as coordinator

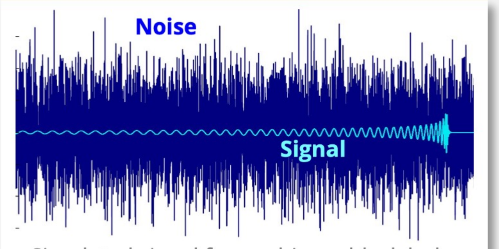
29 **Participants**, including 1 affiliated entity and 2 associated partners

Consortium at a glance

10 Providers	11 Technology providers	14 Community representants
cloud, HTC , HPC resources and access to Quantum systems	delivering the DTE infrastructure and horizontal capabilities	from 5 scientific areas; requirements and developing DT applications and thematic modules

DTs for Radio astronomy and GW astrophysics

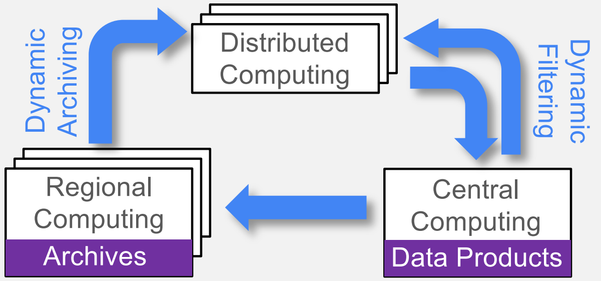
DT of the VIRGO Interferometer



Simulated signal from a binary black hole merger that has been embedded in noise.

It is meant to **realistically simulate** the noise in the detector, in order to study how it reacts to external disturbances and, in the perspective of the **Einstein Telescope**, to be able to detect noise “glitches” in **quasi-real time**, which is currently not possible. This will allow sending out **more reliable triggers** to observatories for multi-messenger astronomy.

DT for noise simulation of next-generation radio telescopes


Distributed Computing

Dynamic Filtering

Central Computing

Data Products

Regional Computing

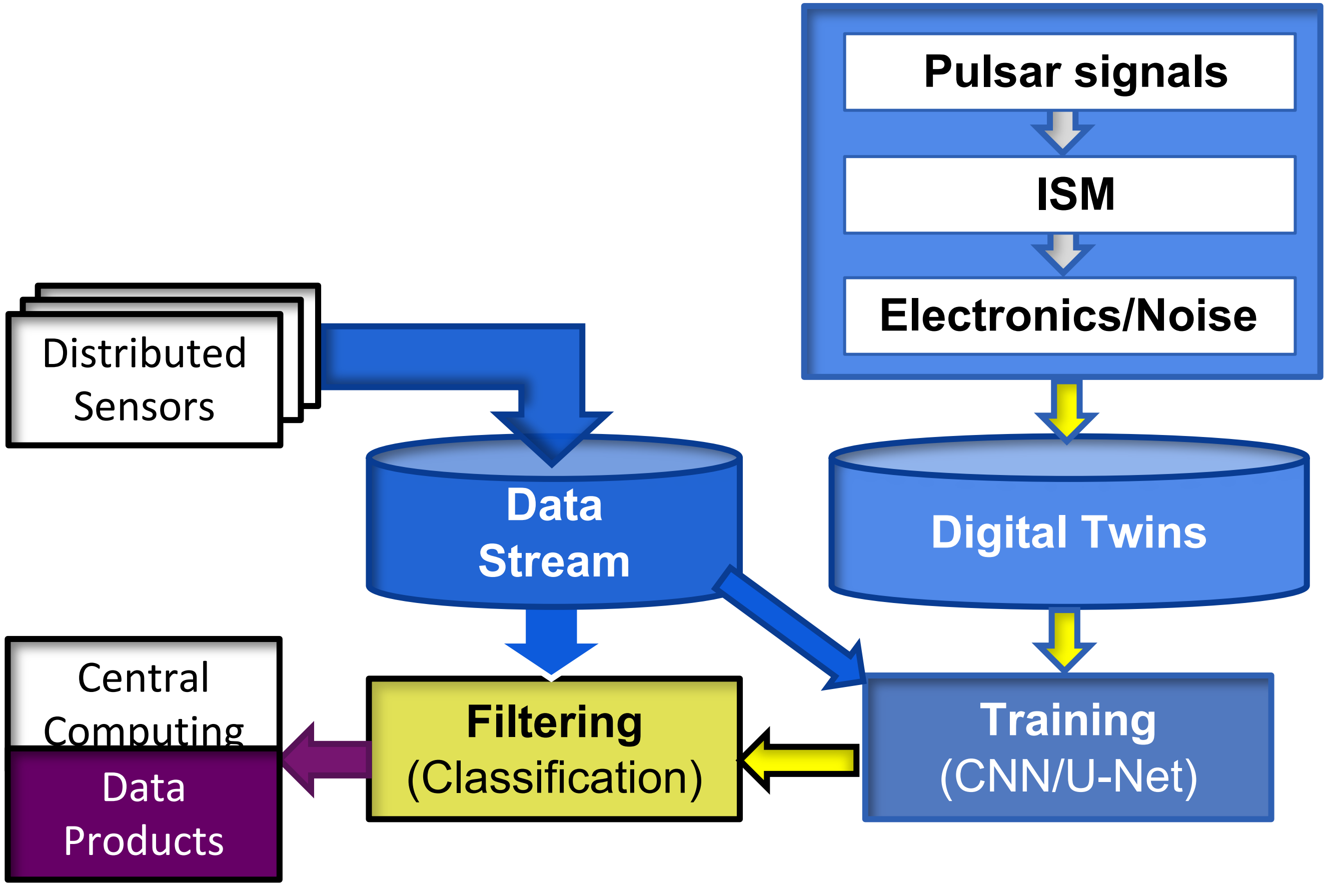
Archives

Dynamic Archiving

Meant to provide DTs to simulate the noise background of radio telescopes (**MeerKat**) will support the identification of rare astrophysical signals in (near-)real time. The result will contribute to a realisation of "**dynamic filtering**" (i.e. steering the control system of telescopes in real-time).

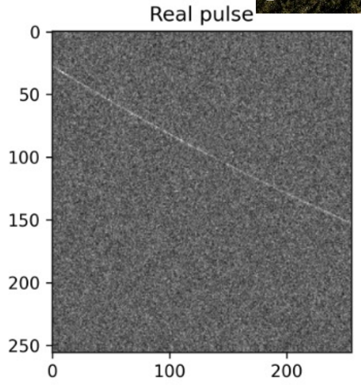
[https://indico.lip.pt/event/1543/contributions/5132]

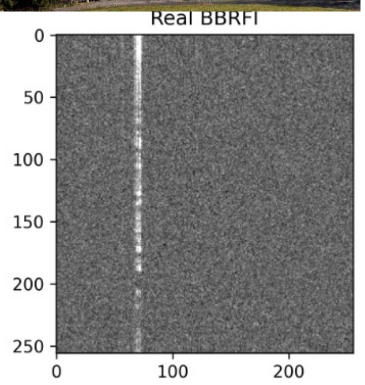
ML-PPA: Architecture



[David A. Aguilar / NASA / ESA]

[Effelsberg]





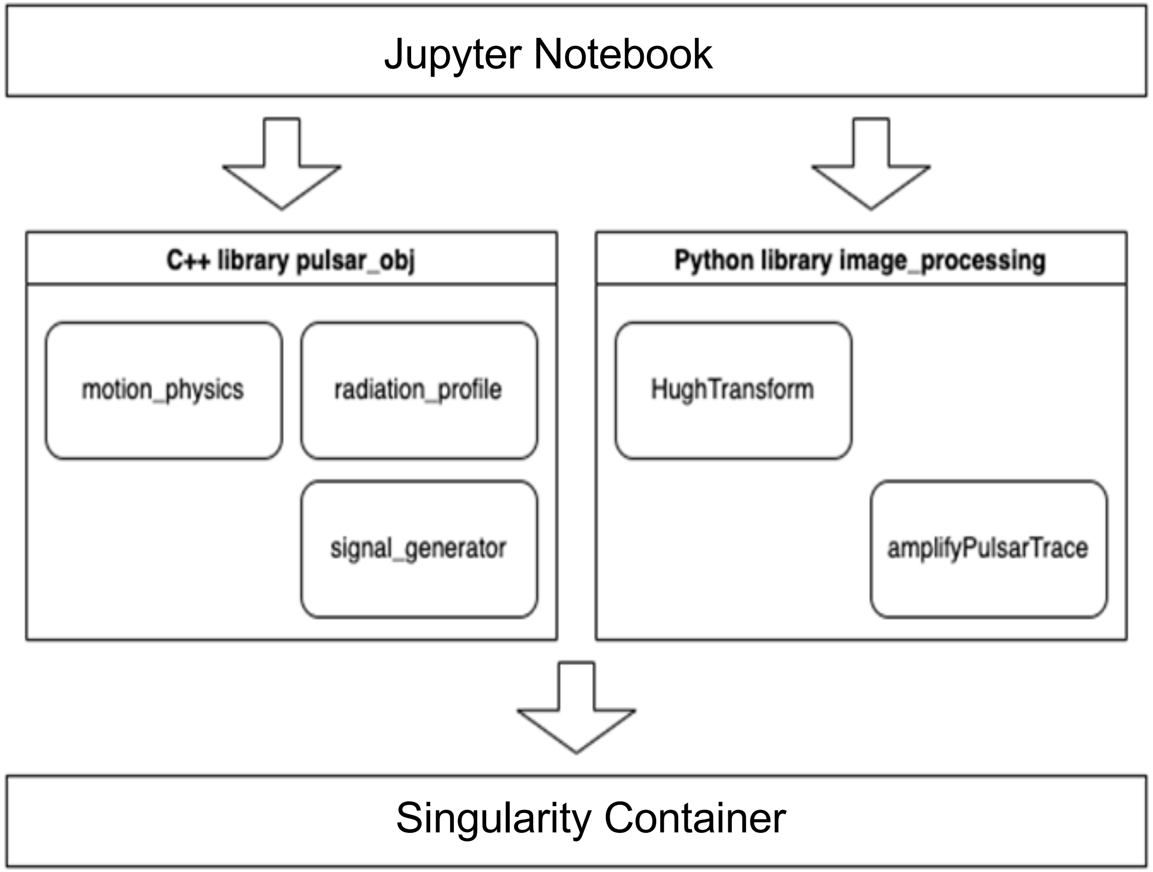
ML-PPA: Repository

Version 0.2 (https://gitlab.dzastro.de/punch/ml-ppa)

Modules

- PulsarDT physics-based DT (Python)
- PulsarDT++ C++ implementation of ML-PPA components ⇒ HPC
- PulsarRFI_Gen empirical DT
- PulsarRFI_NN ML-classifier: CNN, UNet

~ 50 page paper with detailed description


Jupyter Notebook

C++ library pulsar_obj

motion_physics

radiation_profile

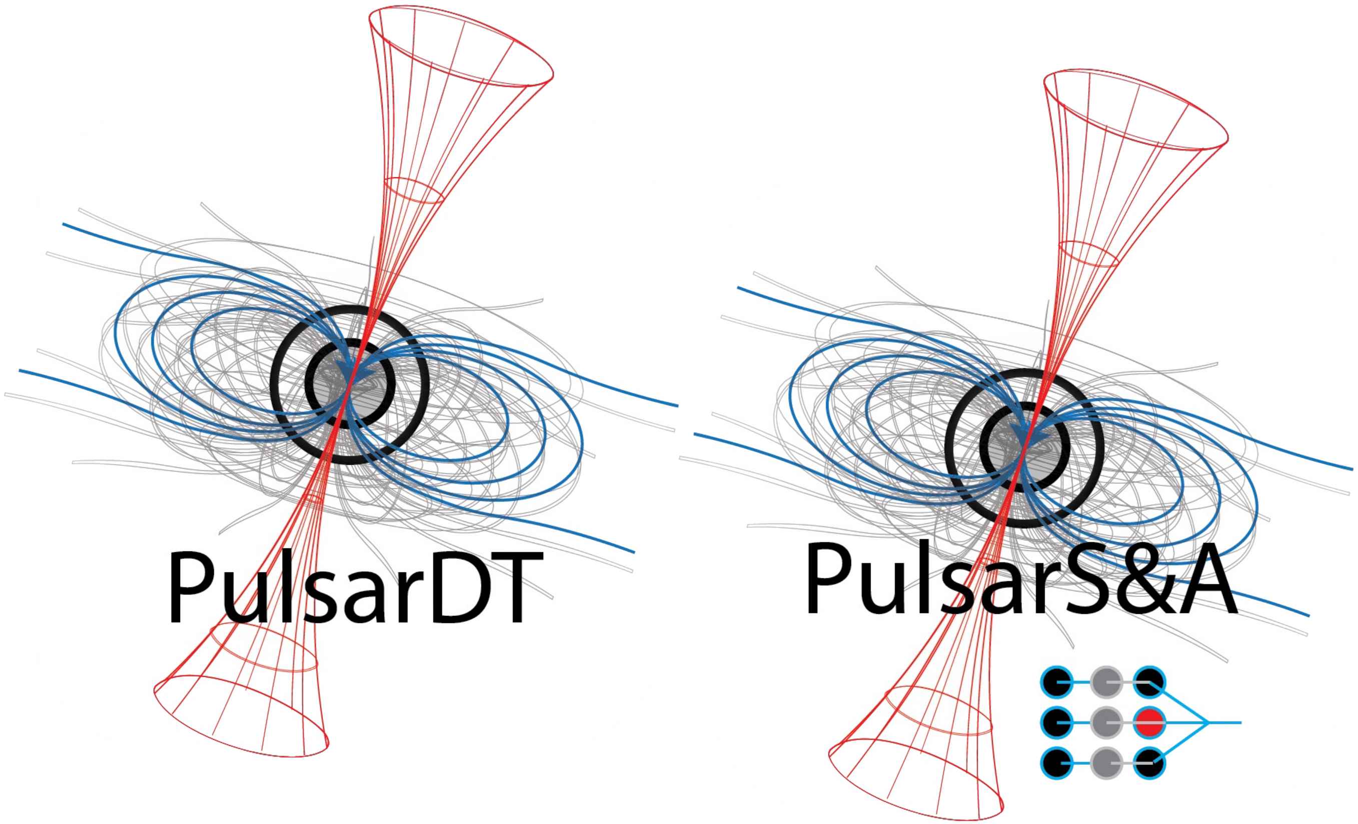
signal_generator

Python library image_processing

HughTransform

amplifyPulsarTrace

Singularity Container

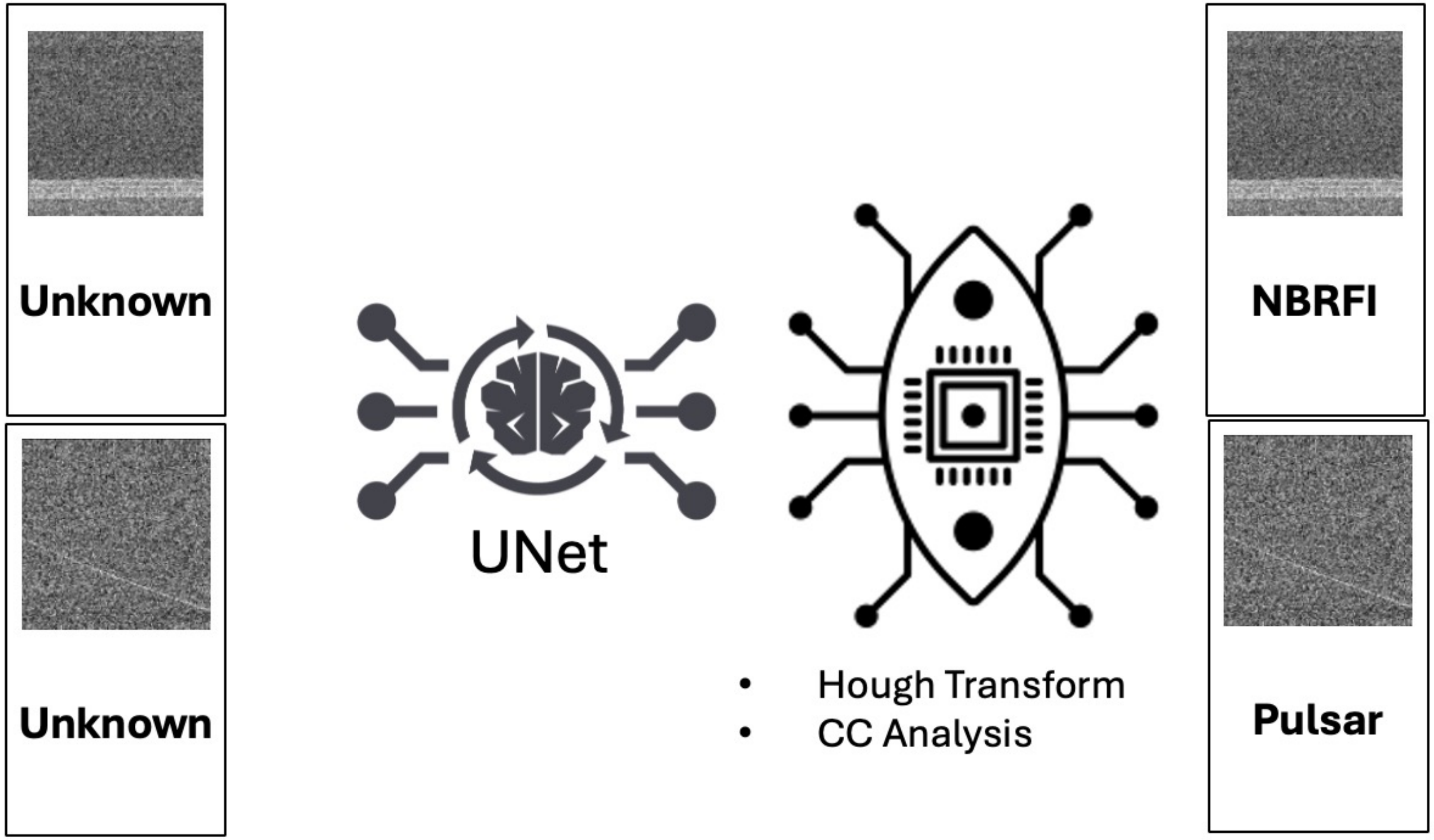


PulsarDT

PulsarS&A

Tanumoy Saha

In Version 0.1



Unknown

Unknown

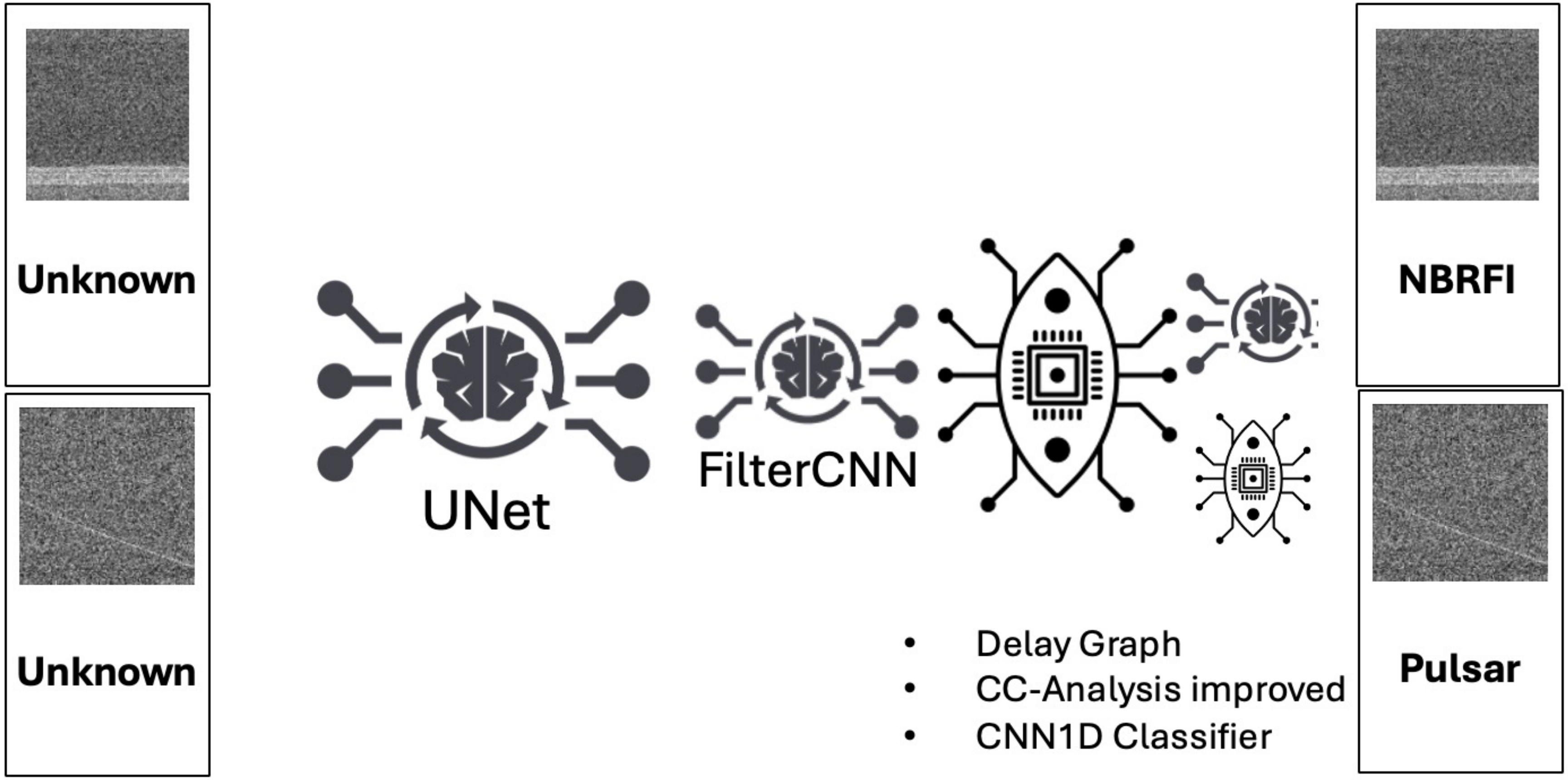
UNet

- Hough Transform
- CC Analysis

NBRFI

Pulsar

In Version 0.2

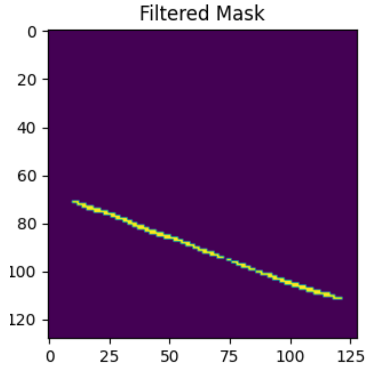


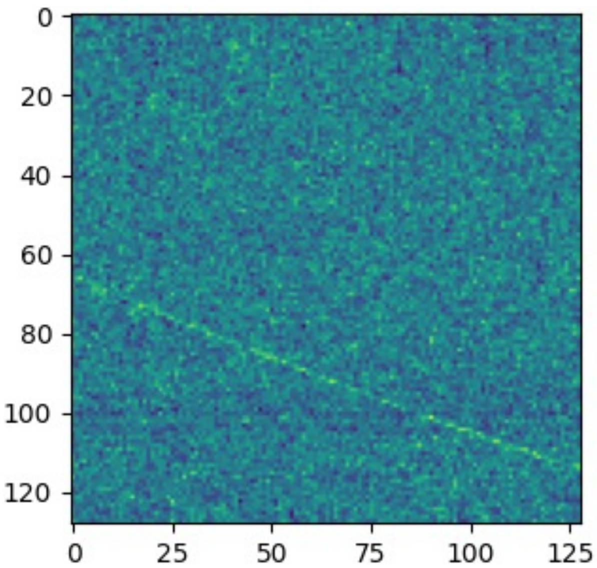
- Delay Graph
- CC-Analysis improved
- CNN1D Classifier

PulsarS&A *Version 0.2 additions*

DZA

In Version 0.2

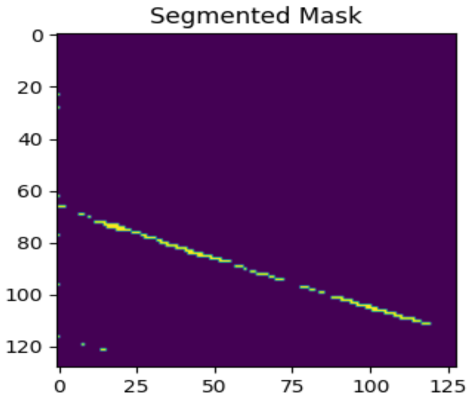




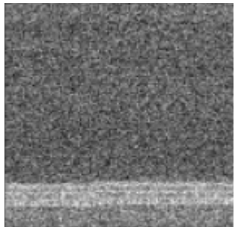
Unknown

UNet

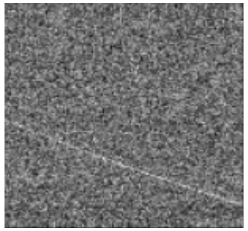
FilterCNN



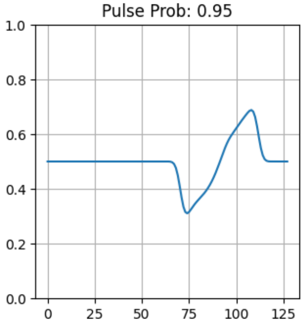
- Delay Graph
- CC-Analysis improved
- CNN1D Classifier

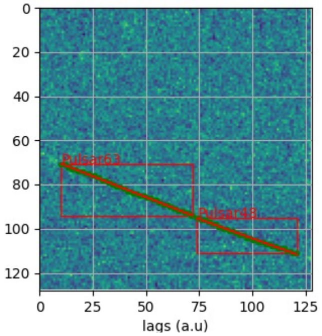


NBRFI



Pulsar





PulsarS&A *Version 0.2 additions*

DZA

Test Performance on Test Data from PulsarDT

UNet

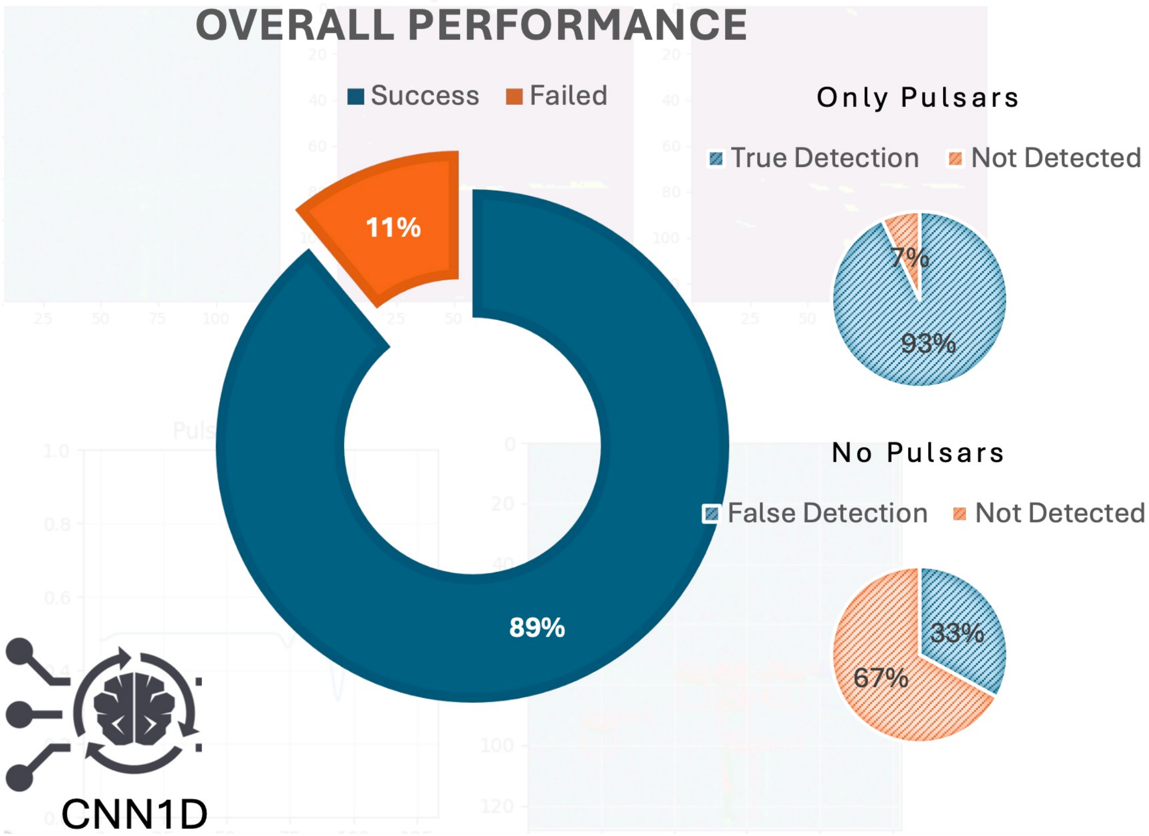
Dataset size: 500
Epochs: 10
Loss Func: WBCELoss

FilterCNN

Dataset size: 500
Epochs: 10
Loss Func: WBCELoss

CNN1D

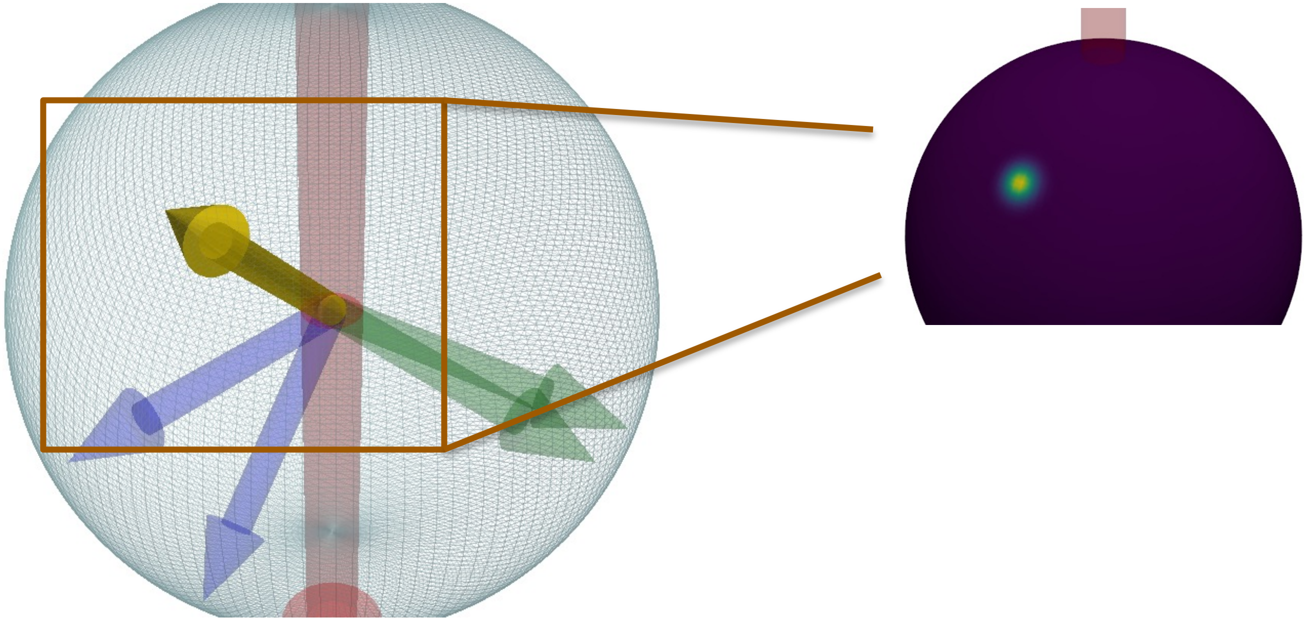
Dataset size: 500
Epochs: 15
Loss Func: WBCELoss



Version 0.2 additions

In Version 0.1

Only **Polar** spark

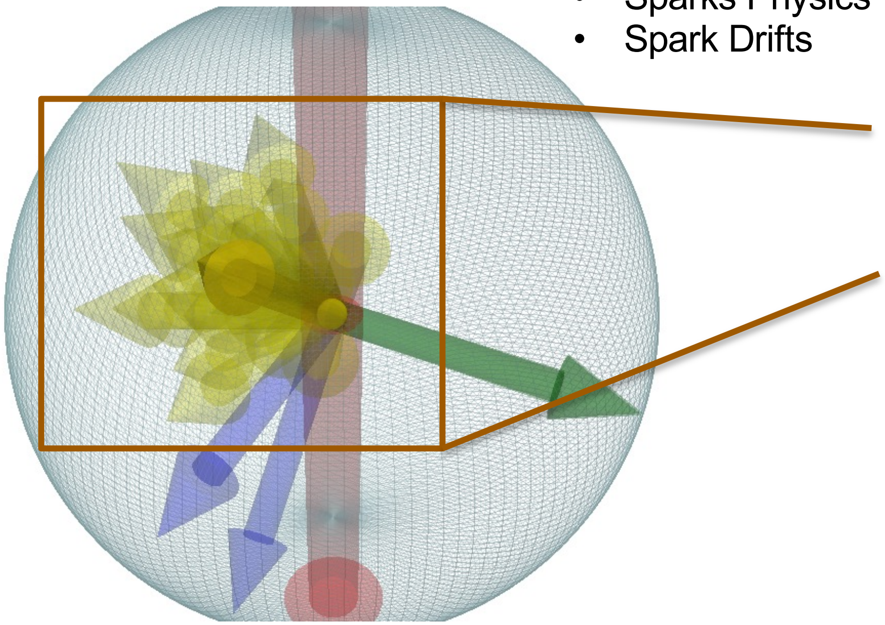


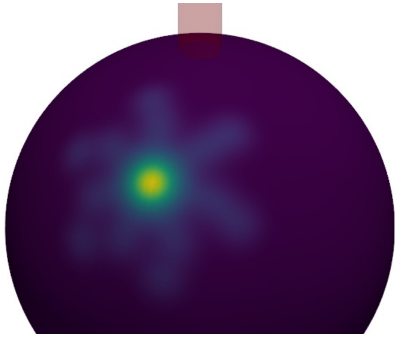
PulsarS&A *Version 0.2 additions*

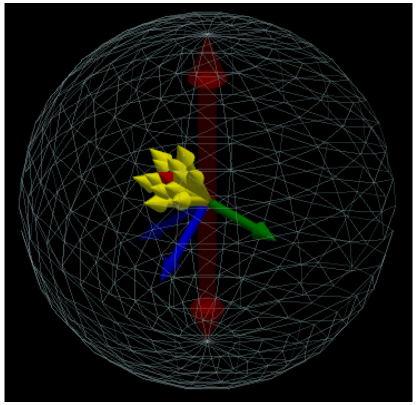
In Version 0.2

Multi sparks engineering

- Sparks Physics
- Spark Drifts







Spark model based on *Gil and Sendzk,2000*

Synthetic Pulsar and RFI Signals

Andrei Kazantsev

ML-PPA v 0.1: PulsarRFI_generator

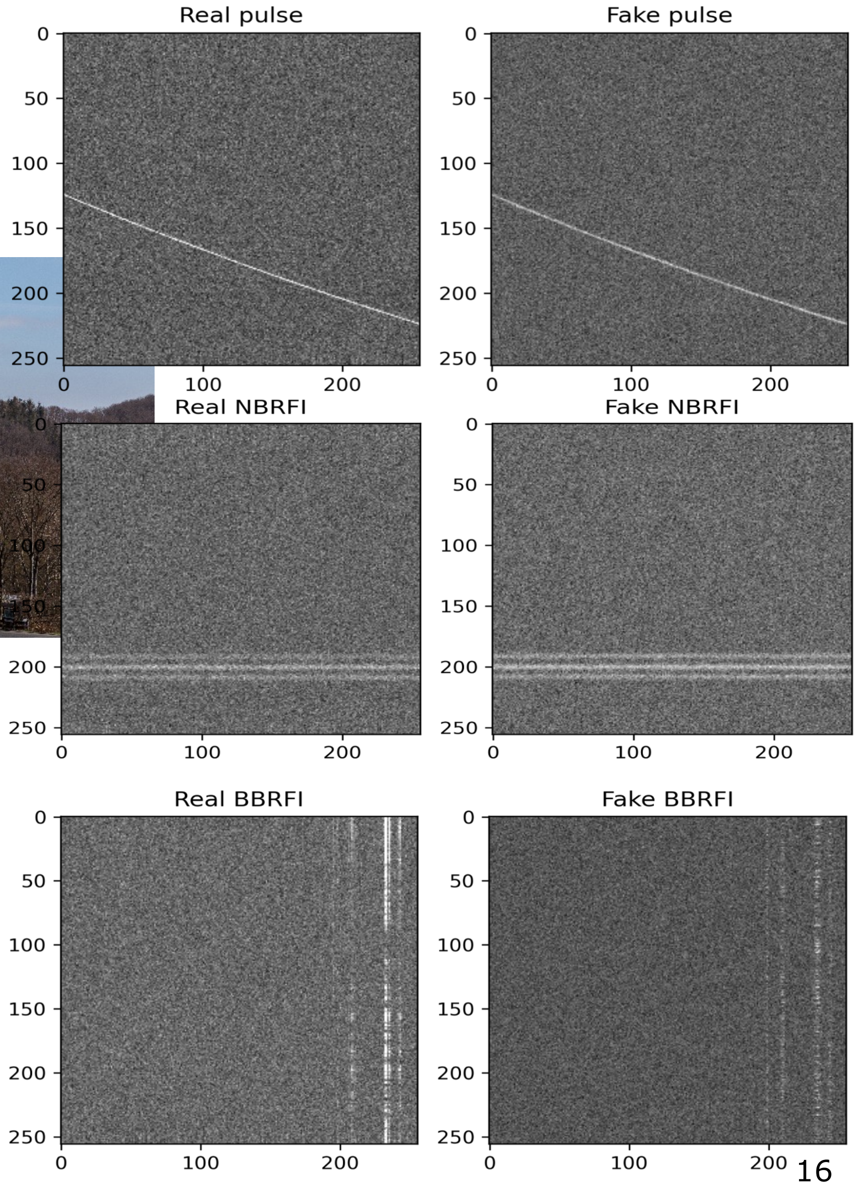
Submodule for generating synthetic pulsar pulses and radio frequency interference.

[Effelsberg]

Included effects for pulses:

- Dispersion delay,
- Scattering,
- Spectra of a pulsar,

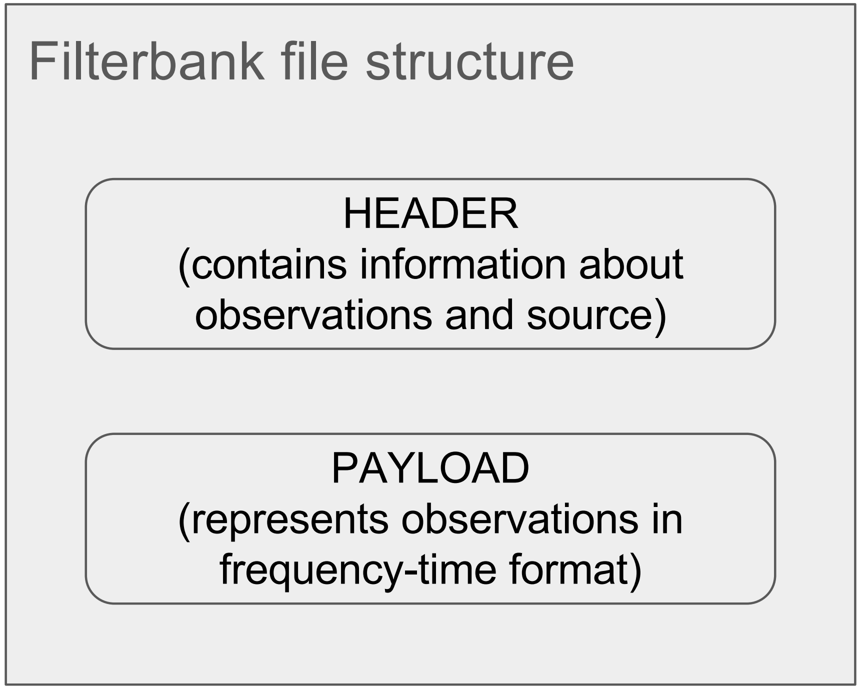
Background for synthetic data is uniform Gaussian noise.



This submodule stores the generated pulses and RFI as **NumPy arrays**. This approach is highly convenient for machine learning, as it eliminates the need for unnecessary data conversion. However, it is not canonical in radio astronomy, which limits the use of our software package by the target audience.

ML-PPA v 0.2: Filterbank file

A **filterbank file** in radio astronomy is a data format that records signals across multiple frequency channels over time. It is used to capture and store time-series data after applying a filter to isolate specific frequency ranges. This format is commonly used in the search for transient signals, such as pulsars or fast radio bursts (FRBs), by analyzing how the intensity of the signal changes across different frequencies and time intervals.



Filterbank file structure

HEADER
(contains information about observations and source)

PAYLOAD
(represents observations in frequency-time format)

Existing analogy to generate a filterbank file with synthetic pulses:

frb-faker (https://gitlab.com/houben.ljm/frb-faker/)
SIGPROC (https://sigproc.sourceforge.net/)

Existing tools for generating synthetic filterbank files lack the required functionality. For instance, **frb-faker** allows the injection of only a single pulse, whereas the **SIGPROC** program injects pulses with uniform amplitude. These programs also do not have the function of adding RFI.

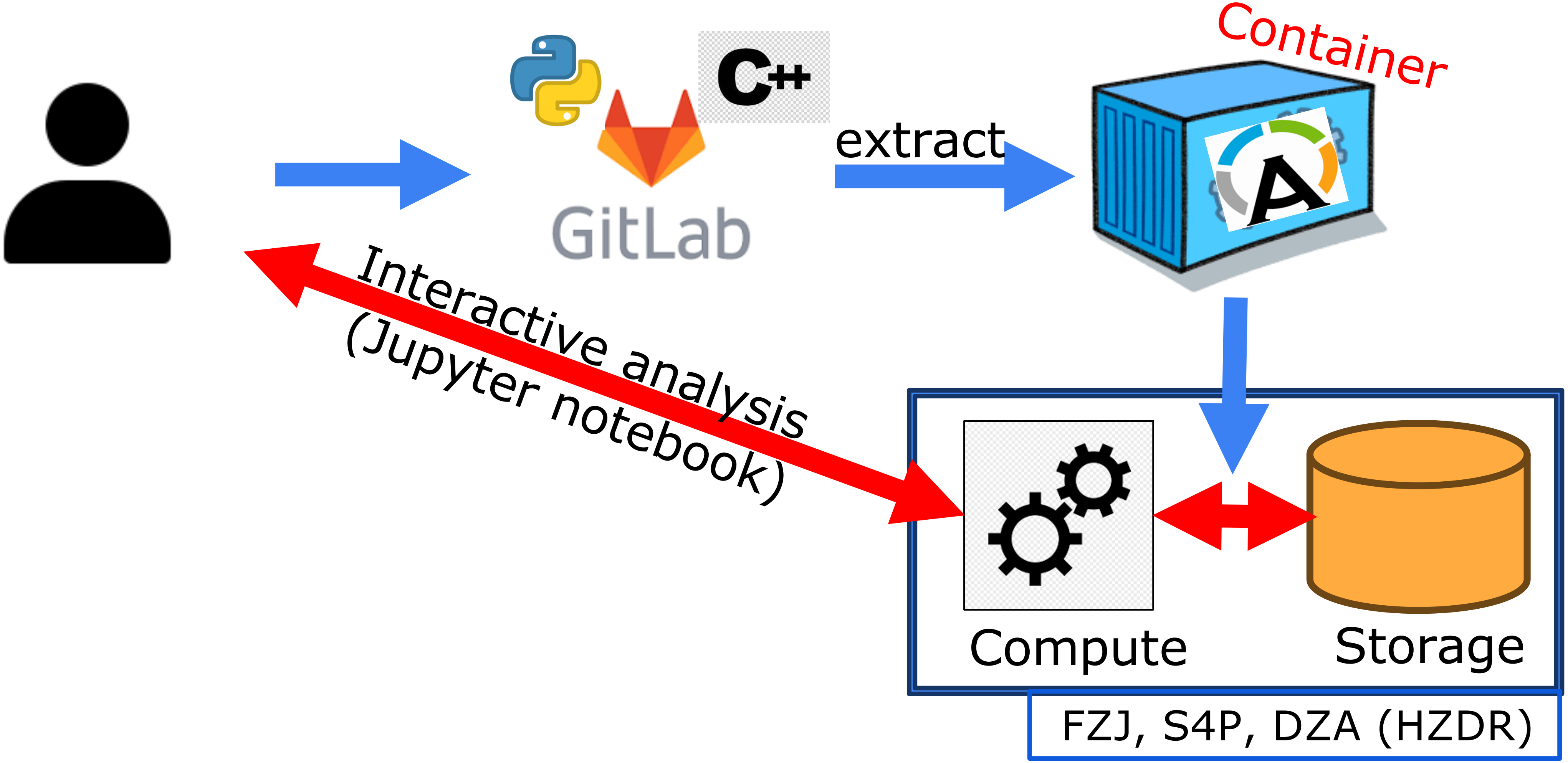
Configuration file (json format)

```
{   "basename": "test",  // The base name of the file where the data will be saved.
    "source_name": "fake_pulsar",  // The name of the source signal (pulsar) for data generation.
    "tstart": null,  // The start time of the observation in MJD (Modified Julian Date) format. If null, it will be automatically calculated.
    "f_hi": 1530,  // The upper frequency limit of the observation in MHz.
    "f_low": 1210,  // The lower frequency limit of the observation in MHz.
    "nchans": 256,  // The number of frequency channels into which the band is divided.
    "tsamp": 0.0001024,  // The sampling time, or the interval between measurements, in seconds.
    "total_duration_seconds": 30,  // The total duration of the observation in seconds.
    "dm": 57.6,  // The Dispersion Measure (DM) in pc/cm^3, which affects the delay of the signal across different frequencies.
    "period": 1.0,  // The period of the pulsar signal in seconds.
    "position_of_first_pulse": 10,  // The time in seconds at which the first pulse appears in the observation.
    "pulse_snr_range": [3, 10],  // The range of signal-to-noise ratios (SNR) for the synthetic pulsar pulses.
    "pulse_hw_range": [2, 3],  // The range of pulse half-widths (in tsamp).

    // Parameters for Broadband RFI (Radio Frequency Interference)
    "bbrfi_contamination_level": 0.5,  // The level of contamination in the data by broadband RFI (0 = clean, 1 = heavily contaminated).
    "bbrfi_amplitude_range": [3, 10],  // The amplitude range of the broadband RFI.
    "bbrfi_hw_range": [2, 3],  // The half-width range of the broadband RFI pulses.
    "num_bbrfi_per_period": [1, 3],  // The number of broadband RFI events per pulsar period.

    // Parameters for Narrowband RFI
    "nbrfi_contamination_level": 0.5,  // The level of contamination in the data by narrowband RFI.
    "nbrfi_amplitude_range": [3, 10],  // The amplitude range of the narrowband RFI.
    "nbrfi_hw_range": [3, 10],  // The half-width range of the narrowband RFI pulses.
    "nbrfi_durtion_range": [1, 5],  // The duration range of narrowband RFI events in seconds.
    "num_nbrfi_per_segment": [1, 3]  // The number of narrowband RFI events per data segment.
}
```

Software-to-the-data

Gautam Dange
Marcel Trattner

ML-PPA: workflow

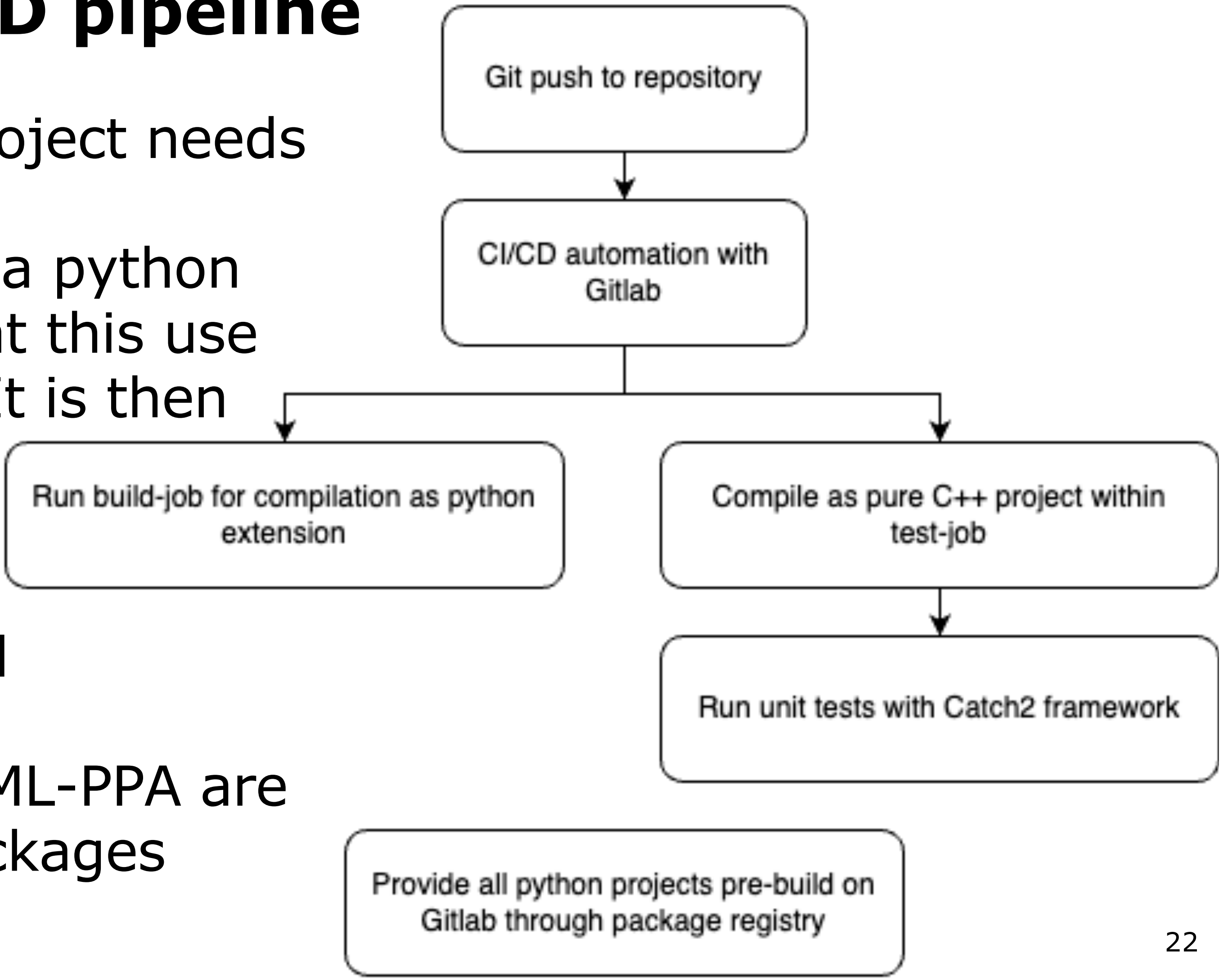


Continous Integration (CI) & Continous Deloyment (CD)

Marcel Trattner

Establishing a CI/CD pipeline

- The growing ML-PPA project needs automated testing
- PulsarDT++ is build as a python extension to ensure that this use case can be compiled. It is then build as a pure C++ project and unit tests are run to verify individual functions and calculations
- All components within ML-PPA are deployed as python packages



Git push to repository

CI/CD automation with Gitlab

Run build-job for compilation as python extension

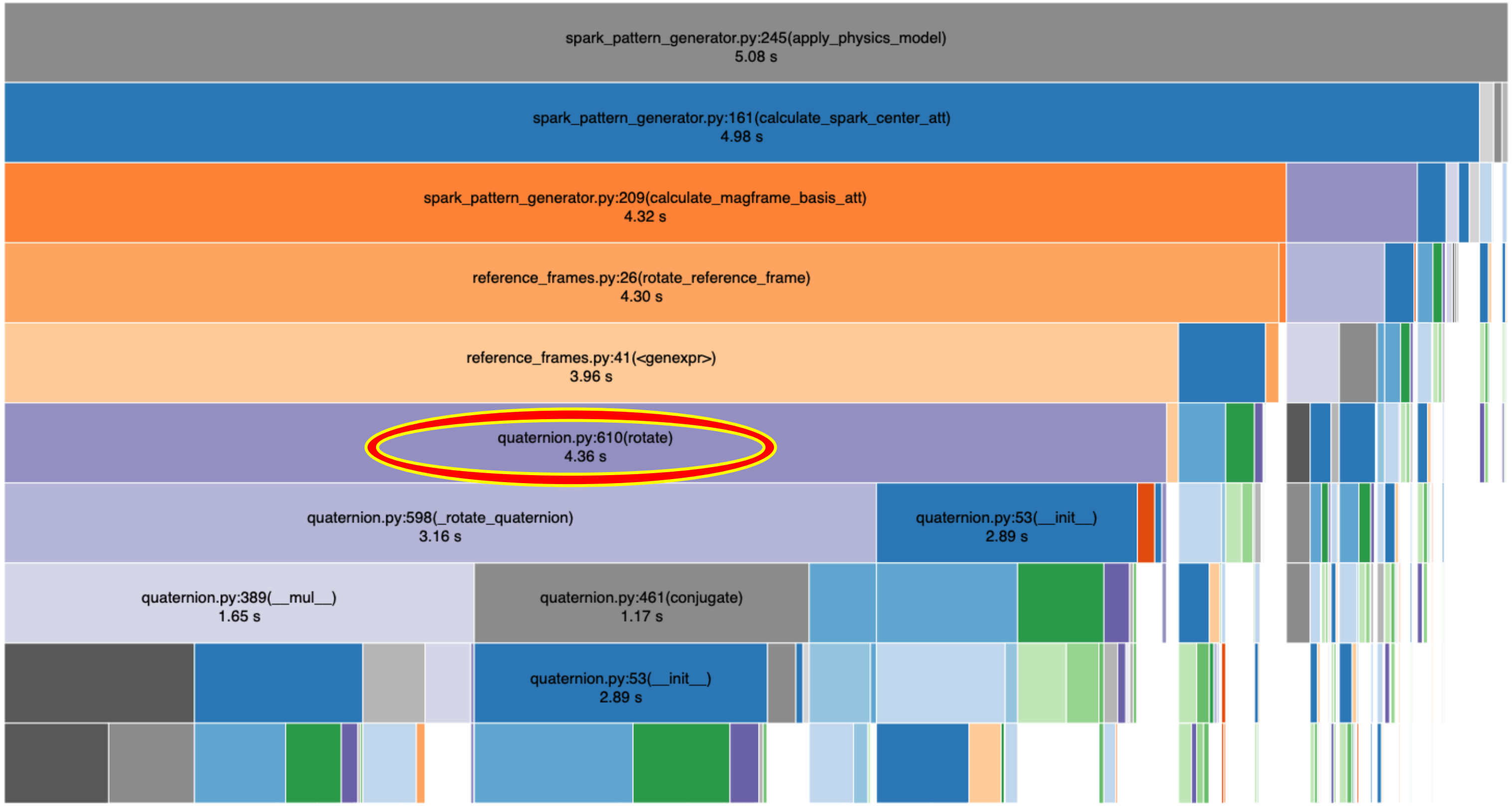
Compile as pure C++ project within test-job

Run unit tests with Catch2 framework

Provide all python projects pre-build on Gitlab through package registry

Profiling Python Code

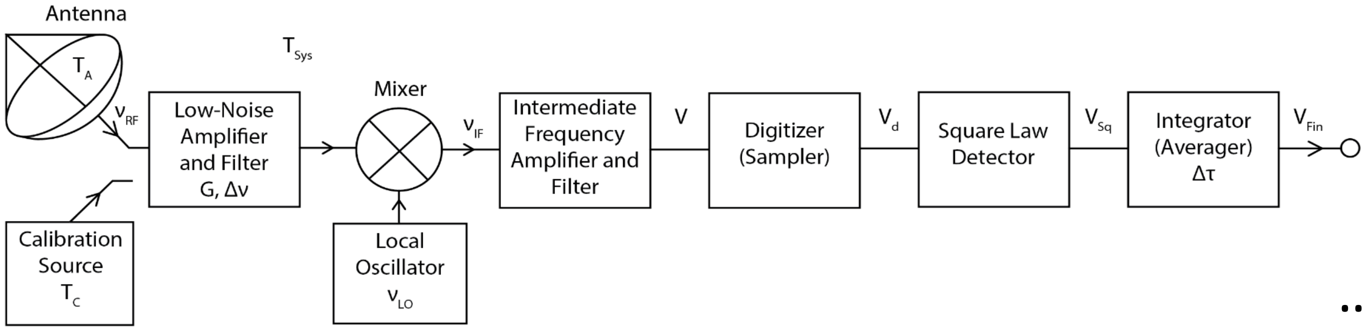
- Goal: identify the functions with the strongest impact on performance
- cProfile & SnakeViz: e.g. quaternion.py is a good candidate for **translation to C++**



Noise Background due to Electronics

Yurii Pidopryhora

Radiometer model (Superheterodine Single Sideband) **predicts white noise**



Antenna

T_A

ν_{RF}

T_{Sys}

Low-Noise Amplifier and Filter G, Δν

Mixer

ν_{IF}

Intermediate Frequency Amplifier and Filter

V

Digitizer (Sampler)

V_d

Square Law Detector

V_{Sq}

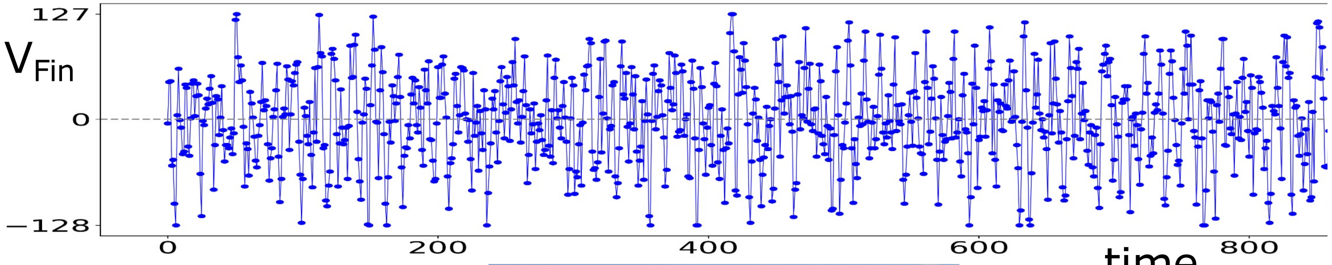
Integrator (Averager) Δτ

V_{Fin}

Calibration Source T_C

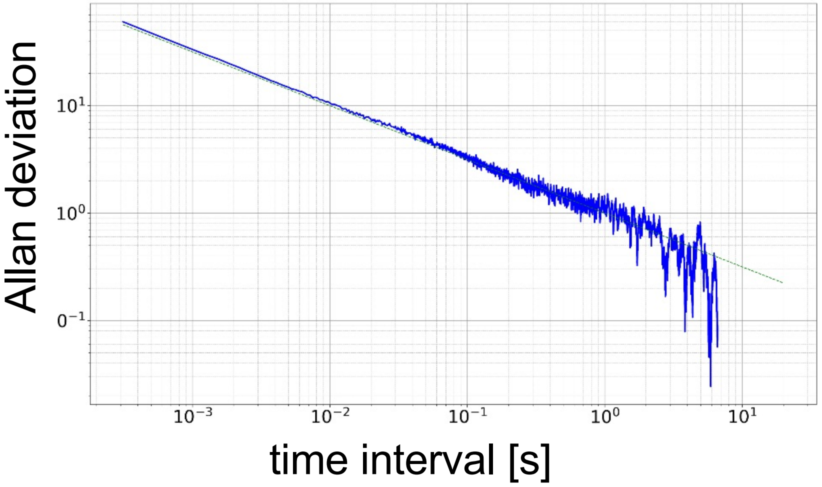
Local Oscillator ν_{LO}

A small sample of raw noisy time-series data (Effelsberg)



[Effelsberg]

... is **confirmed** by statistical analysis of **real data**:



Memory-based Computing

Elsa Buchholz
Marc Drobek
Lars Haupt
Marcel Trattner

Memory-based Computing

Julia

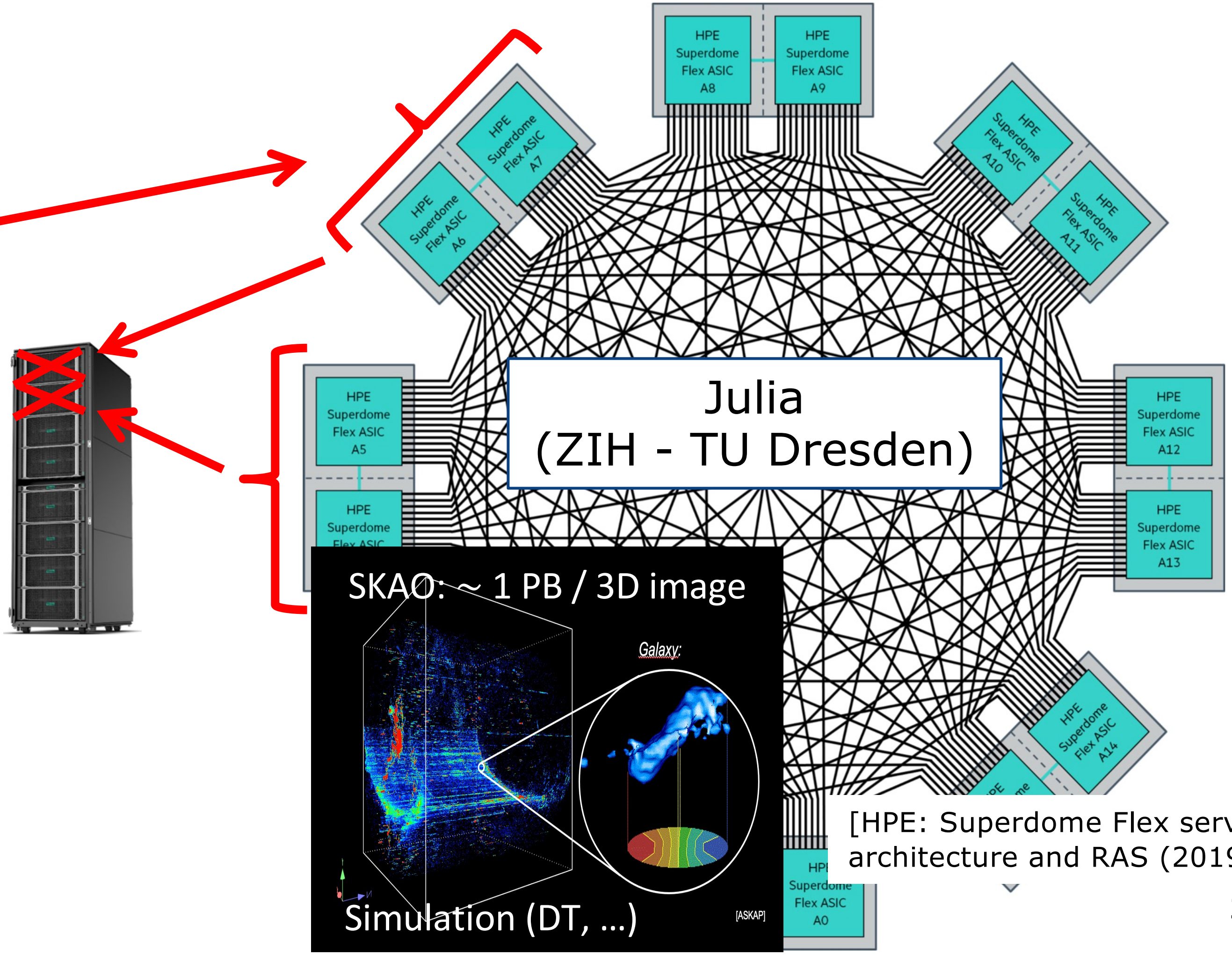
- 64 processor sockets
- 4 processors per chassis
- 18 cores per processor
- **48 TB memory**

NUMAlink cabling

- Point-to-point between each chassis
- 13.3 GB/s (Infiniband)

Subsystem for DZA

- 2 chassis
- currently being set up



Julia
(ZIH - TU Dresden)

SKAO: ~ 1 PB / 3D image

Simulation (DT, …)

HPE Superdome Flex ASIC A8
HPE Superdome Flex ASIC A9
HPE Superdome Flex ASIC A7
HPE Superdome Flex ASIC A6
HPE Superdome Flex ASIC A10
HPE Superdome Flex ASIC A11
HPE Superdome Flex ASIC A5
HPE Superdome Flex ASIC
HPE Superdome Flex ASIC A12
HPE Superdome Flex ASIC A13
HPE Superdome Flex ASIC A14
HPE Superdome Flex ASIC A0
Galaxy:
[ASKAP]

[HPE: Superdome Flex server architecture and RAS (2019)]

Summary & Outlook

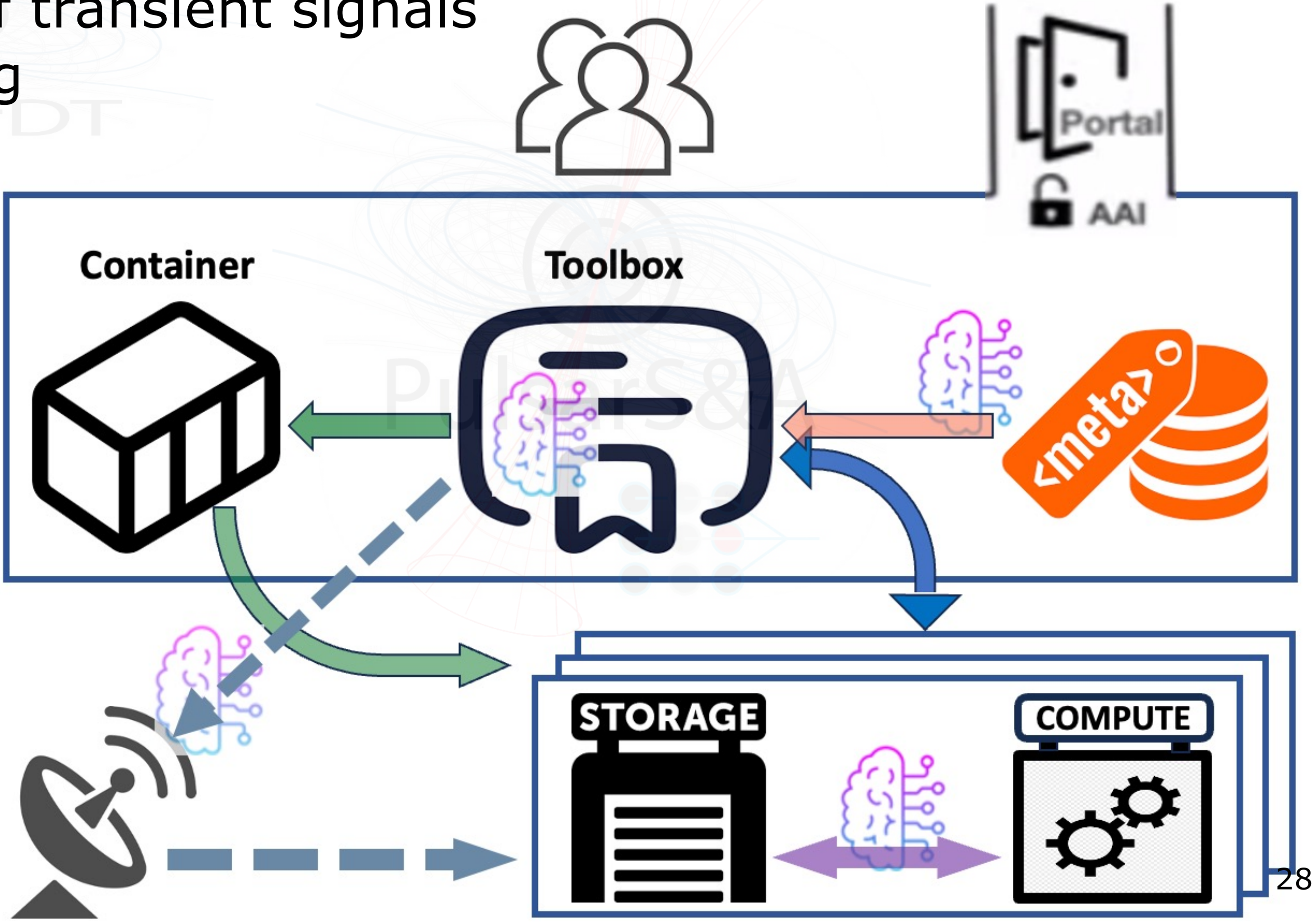
o ML-PPA v0.2 ⇒ v1.0

- Improved identification of transient signals
- Improved CI/CD + testing

o PUNCH 2.0

- MeerKAT / SKAO
- ML-PPA
  - online: small ML (FPGA)
  - offline: large ML
- Improved interactive containerization

[MeerKAT]



Thank you !