interTwin

Detecting pulsar signals in vast real-time data streams with a machine learning / digital twin-based pipeline

Yurii Pidopryhora, Max-Planck-Gesellschaft (MPG),
Max Planck Institute of Radio Astronomy, Bonn, Germany

EGI2024, Lecce, Italy, 01/10/2024

Funded by the European Union

The interTwin project is funded by the European Union - Grant Agreement Number 101058386

Our Project and Interdisciplinary Team

ML-PPA =
ML-based Pipeline for Pulsar Analysis

Gautam Dange (FIAS Frankfurt)
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 (HTW Berlin, DZA Görlitz)

External:

Radio telescope data rates

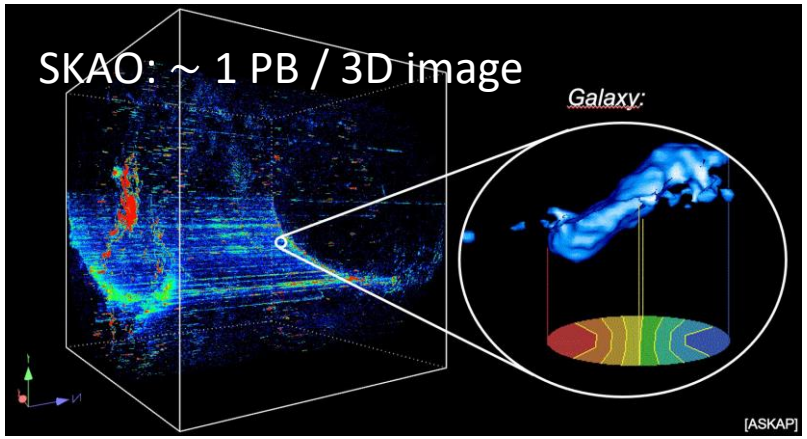
Radio telescope name	Radio telescope exterior	Bitrate per beam	Total bitrate
Effelsberg		P210-7: 11.04 Gb /s	**77 Gb / s (7)**
		UWB: 290 Gb / s	**290 Gb / s (1)**
MeerKAT		107 Mb / s	**0.1 Tb / s (~1024 beams)**
Square Kilometer Array		~ 1 Gb / s	**20 Tb / s (>2200 beams)**

Data Irreversibility

o Only a tiny part of the data can be archived

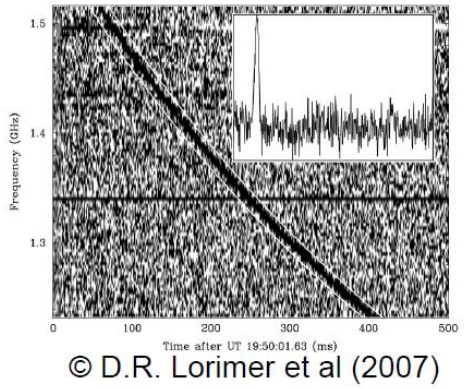
o Decisions on what to keep are based on incomplete information

⇒ irreversible loss of information



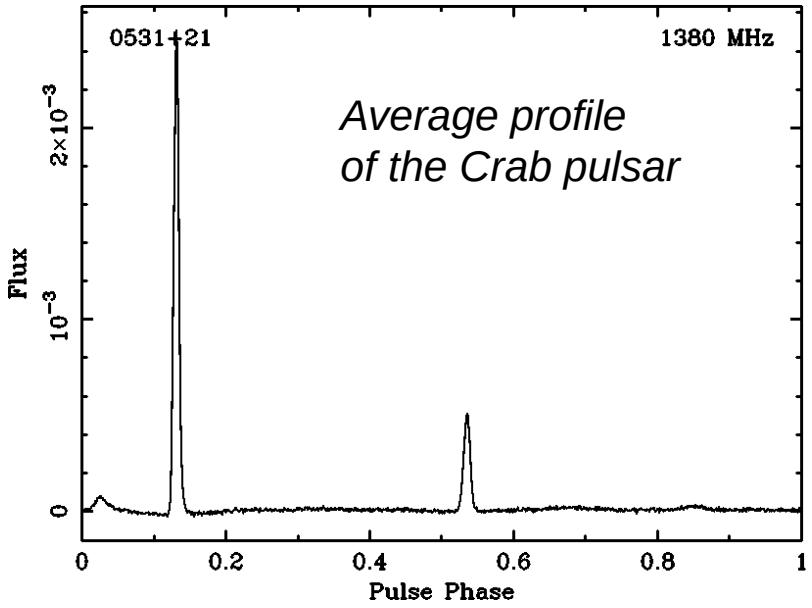
Transients & Pulsars

Crab pulsar with surroundings



© D.R. Lorimer et al (2007)

A transient



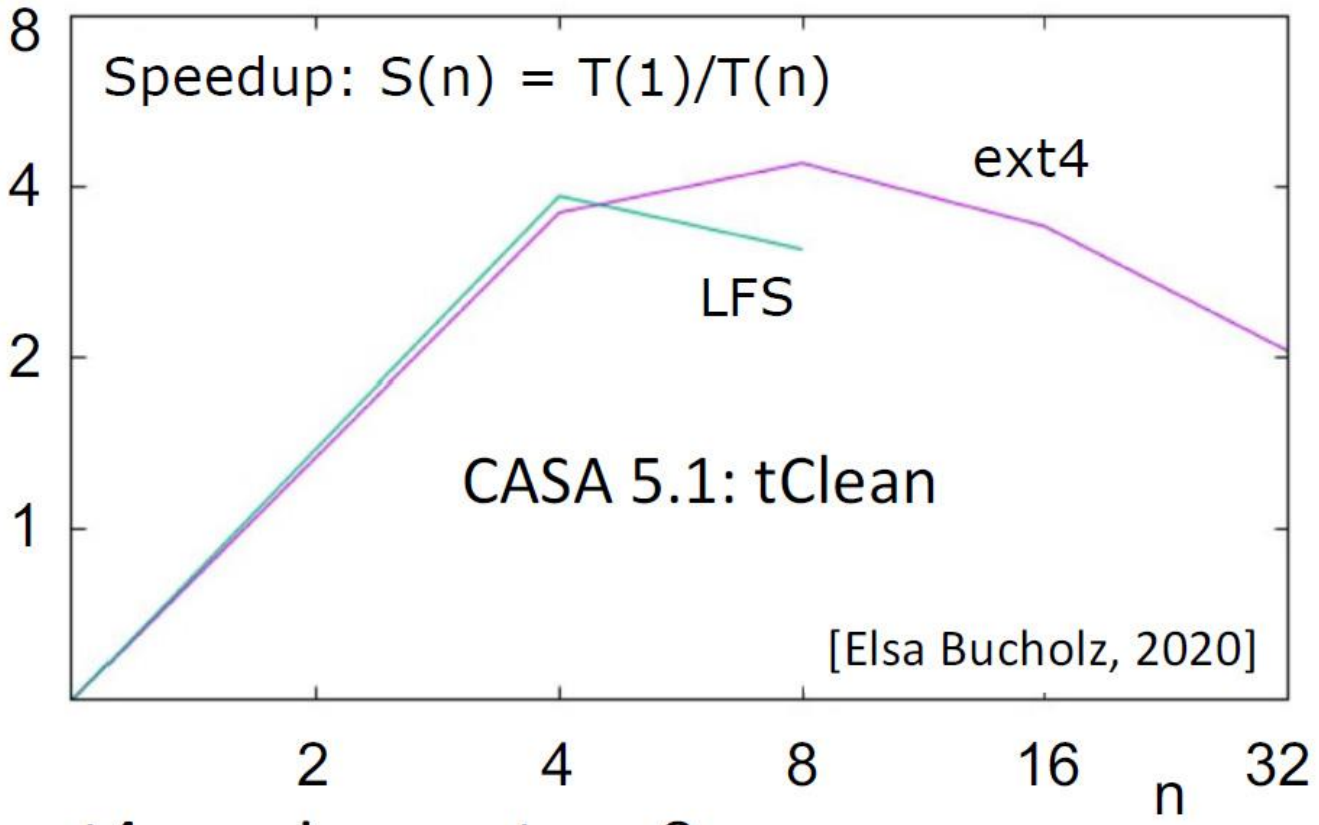
3 focus points:

- **new gen of radio telescopes threatens to flood us with data**
- **search for transients, using pulsars as proxies**
- **scalability of new data pipelines**

100m Effelsberg radio telescope

MeerKAT array (South Africa)

Common Astronomy Software Applications (CASA) testing



ext4: maximum at n = 8 cores

LFS: **CASA crashes for n > 8 cores ⇒ ML-PPA**

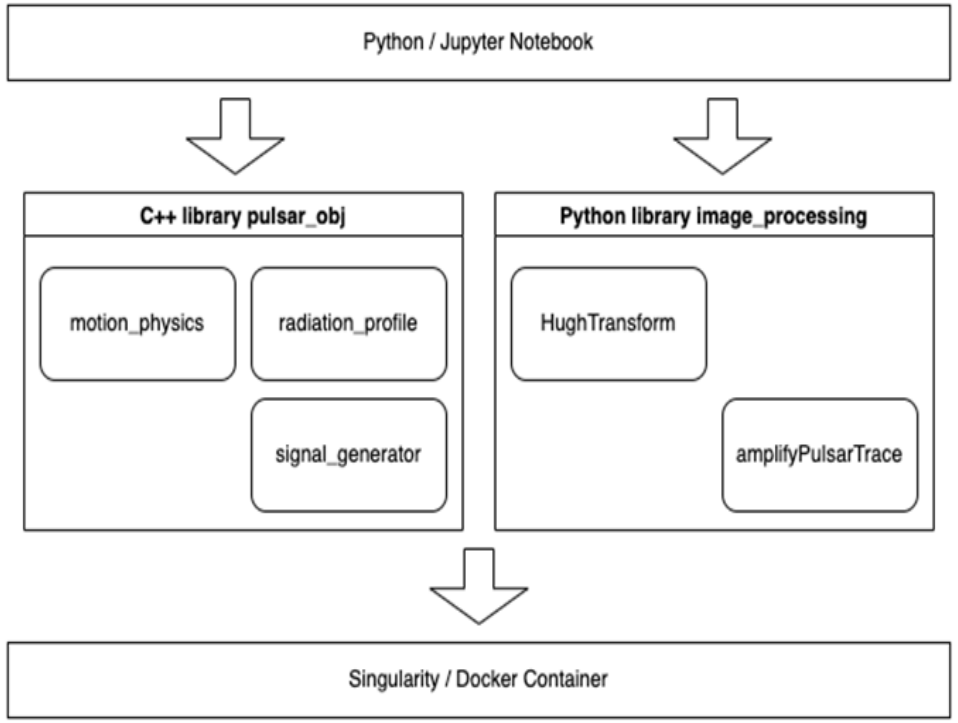
- Standard and universal data reduction software suite for radio astronomy
- Handles scaling poorly,
  need an HPC-capable alternative

ML-PPA v. 0.1

Repository: (https://gitlab-p4n.aip.de/punch_public/ml-ppa)

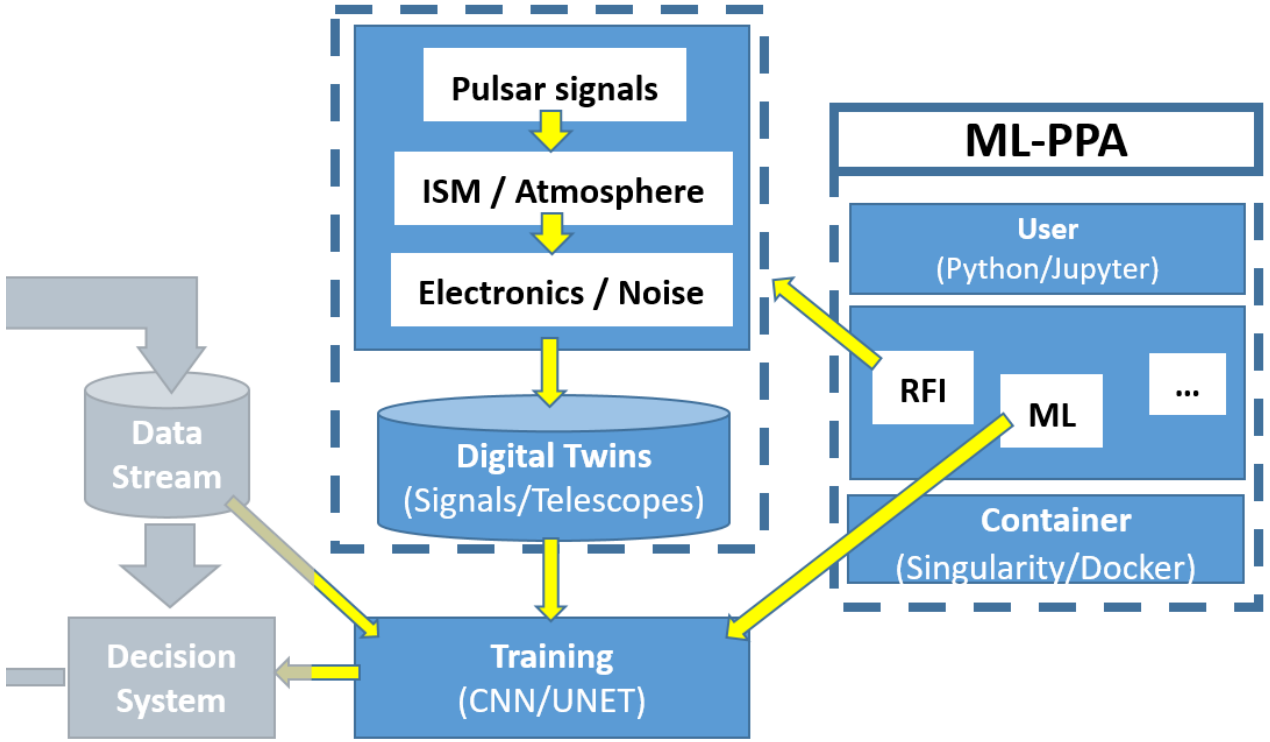
- PulsarDT: physics-based DT (Python)
- PulsarRFI_Gen: empirical DT (Python)
- PulsarRFI_NN: ML-classifier (Python)
- PulsarDT++: C++ implementation of all ML-PPA components => HPC

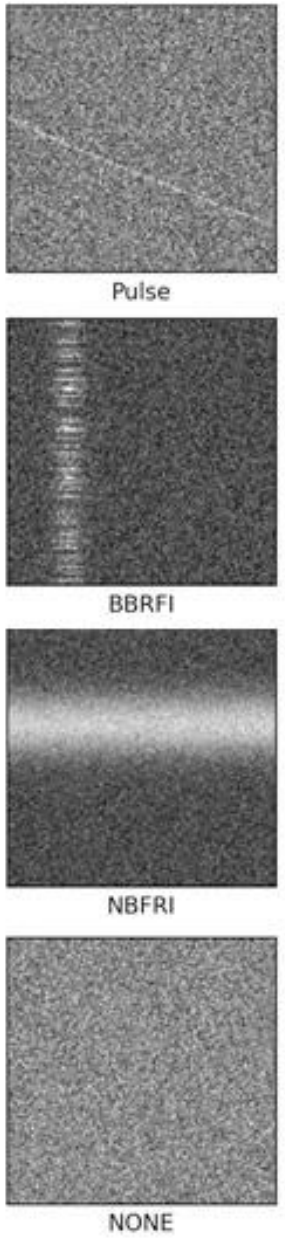
+ ~50 page **paper** with detailed description

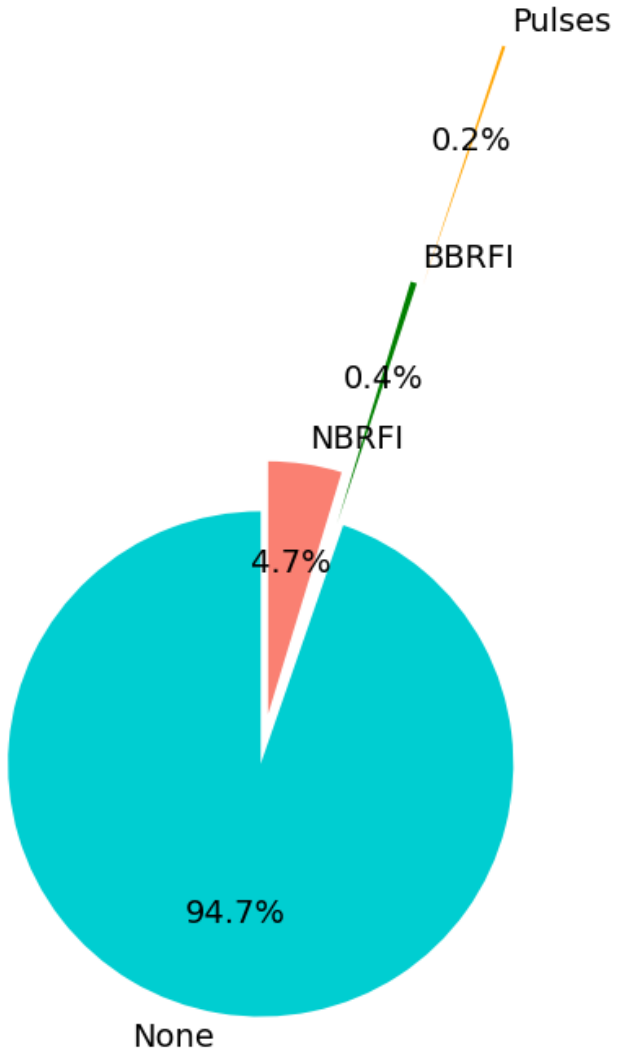


Data Classifier

ML-based **data classifier**, needs **DT** to supply **training data**.

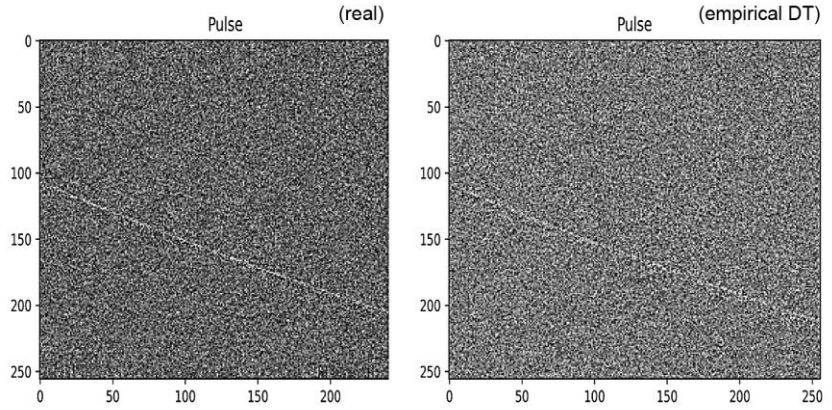


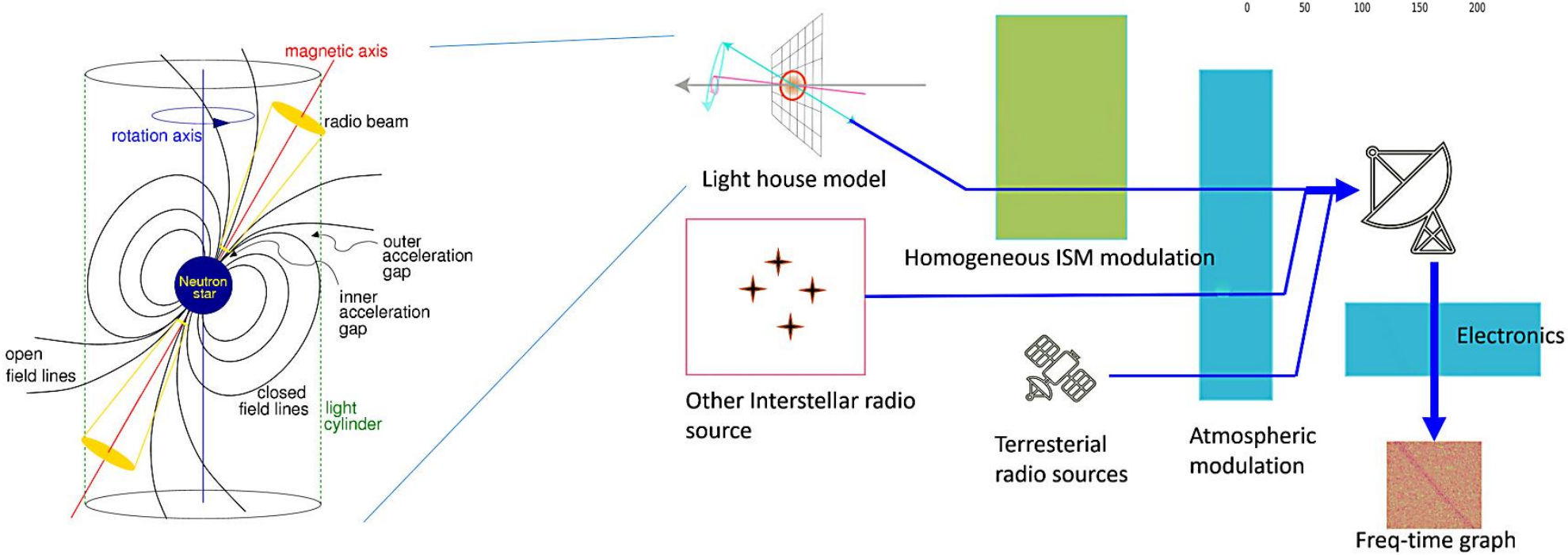




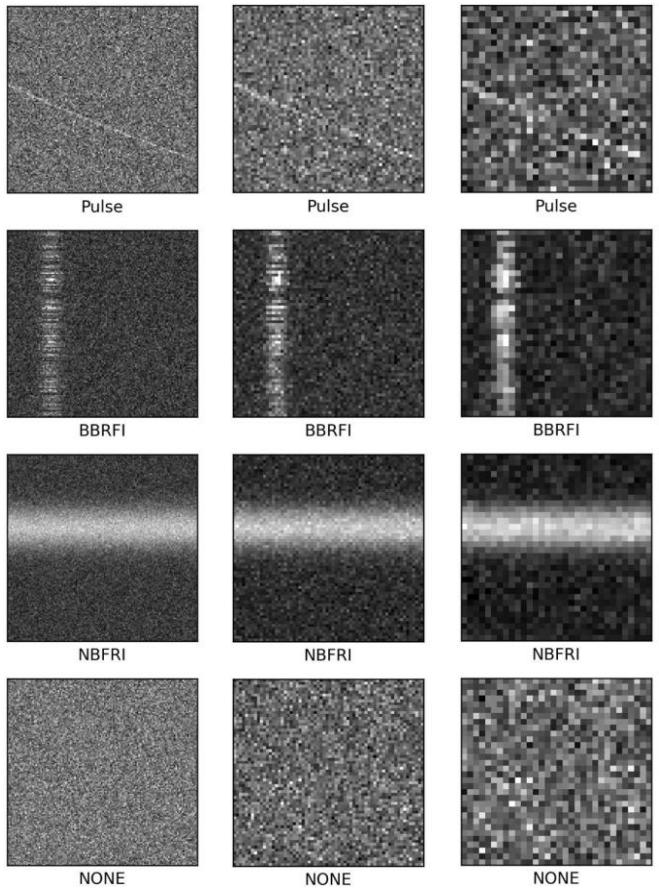
Digital Twins

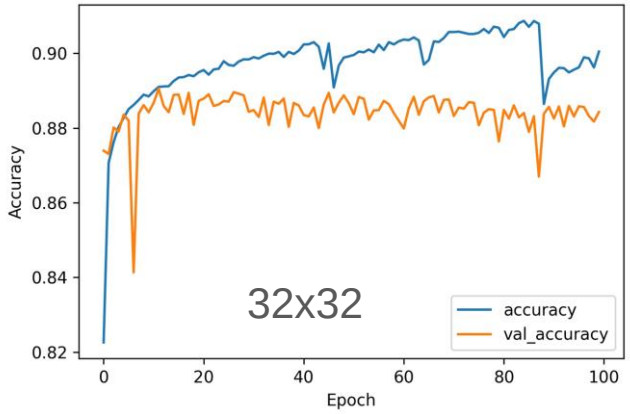
2 types of DT: **physics-driven** (main) and **empirical** (auxiliary)

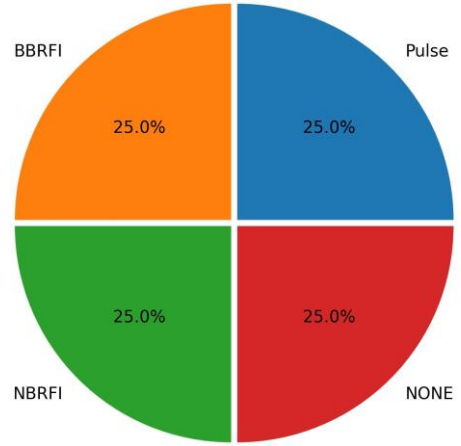


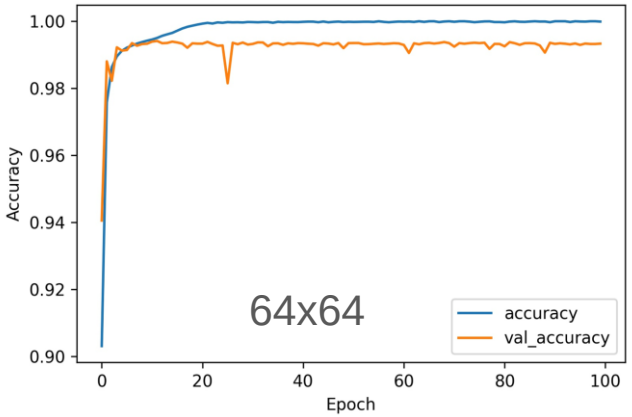


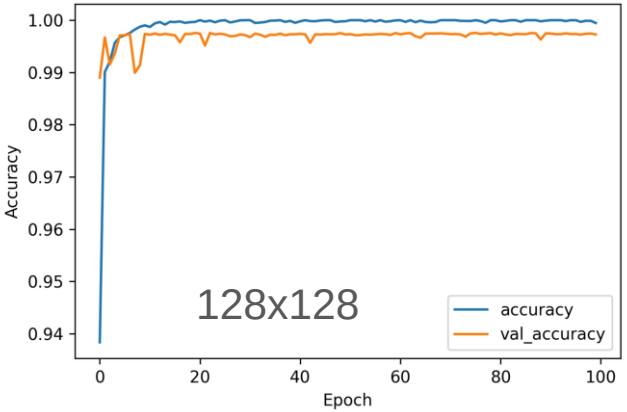
Accuracy of the model trained on synthetic data





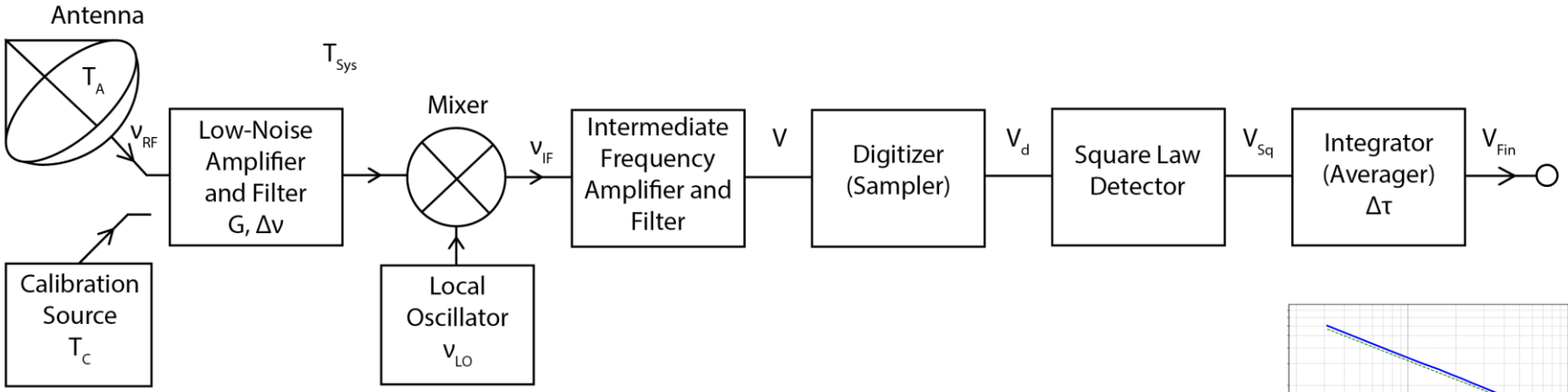






Noise Analysis

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



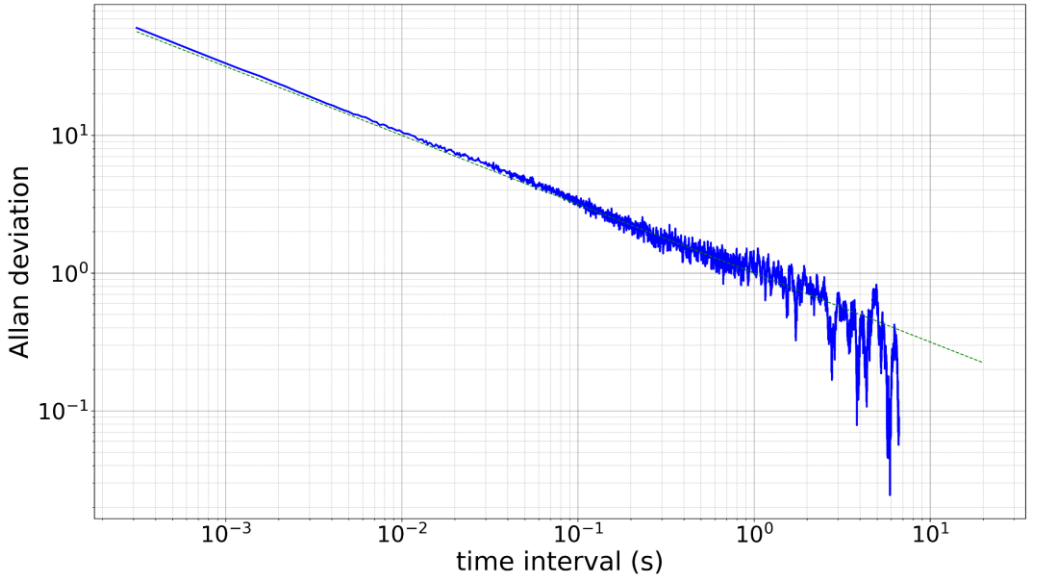
…and it is **confirmed** by statistical analysis of **real data**:

$$T_{\mathrm{Sys}} = T_{\mathrm{rad}} + T_{\mathrm{cmb}} + T_{\mathrm{bg}} + T_{\mathrm{atm}} + T_{\mathrm{spill}} + T_{\Omega}$$

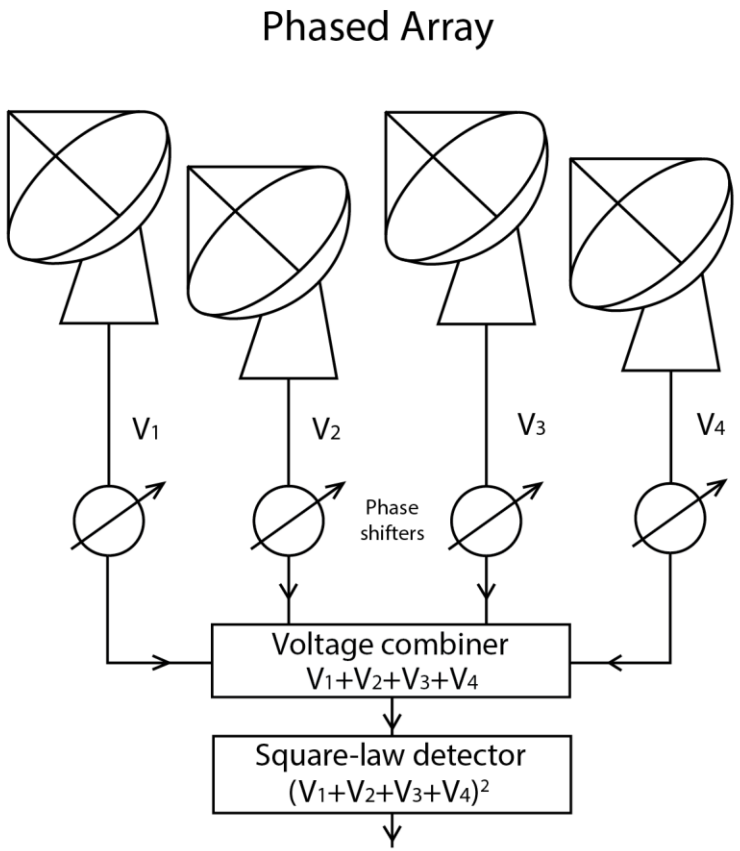
$$\langle V_{\mathrm{Fin}_T} \rangle = T_A + T_{\mathrm{Sys}}$$

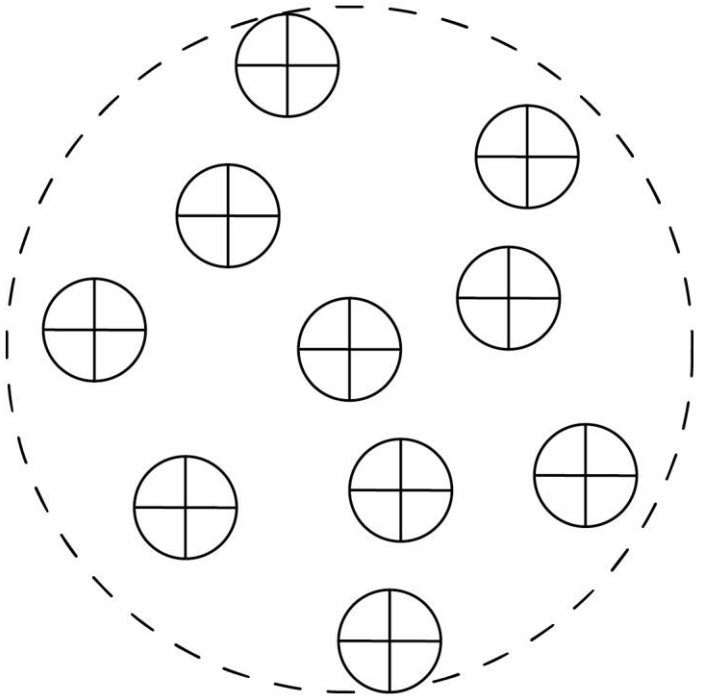
$$\sigma_{\mathrm{Fin}_T} = \frac{T_A + T_{\mathrm{Sys}}}{\sqrt{\Delta\nu\Delta\tau}}.$$

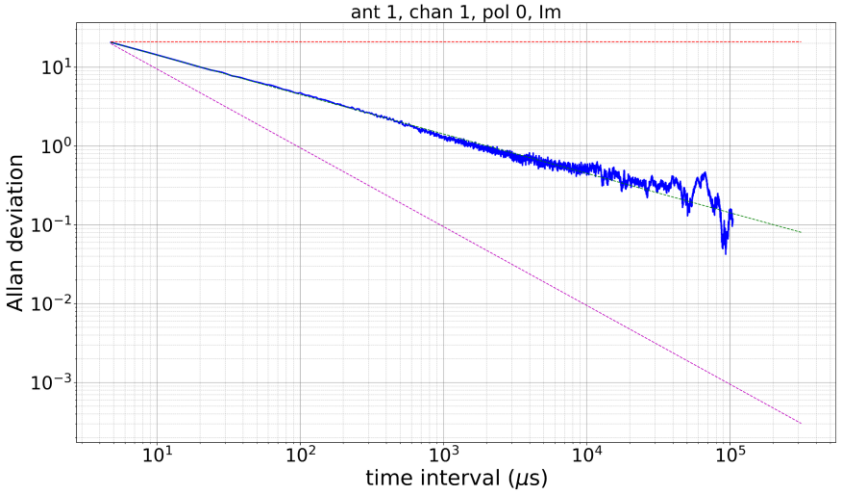
$$\mathrm{SNR} \approx \frac{T_A}{T_{\mathrm{Sys}}} \sqrt{\Delta\nu\Delta\tau}.$$

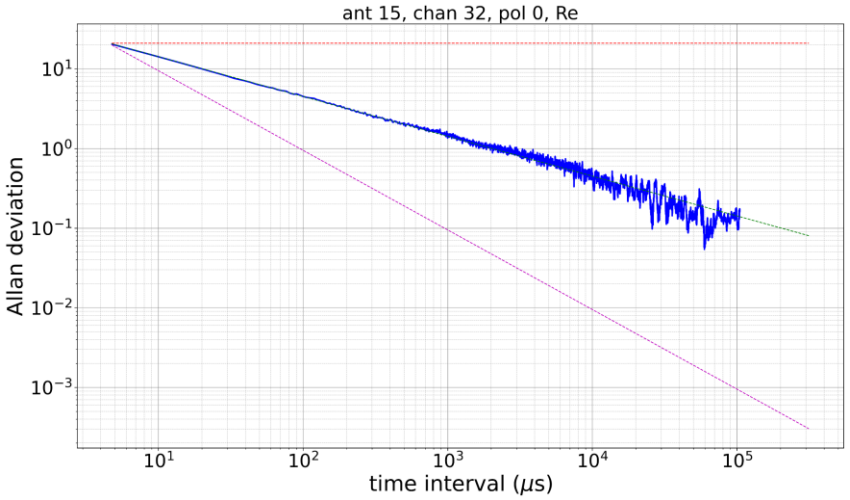


Applicability to arrays (like MeerKAT)



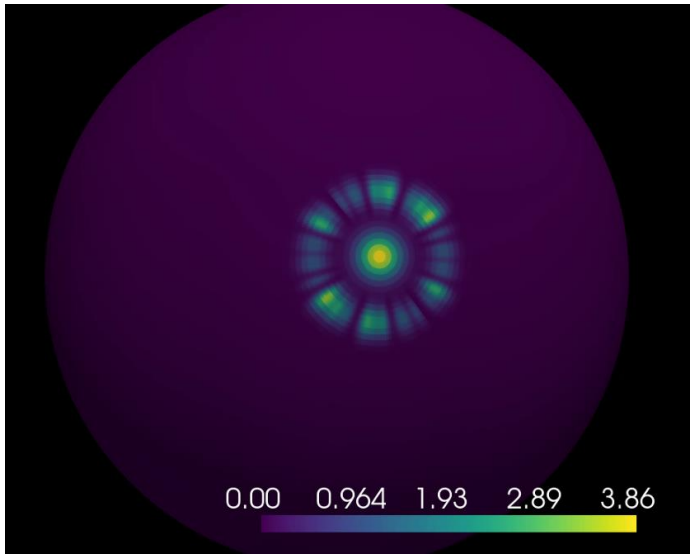


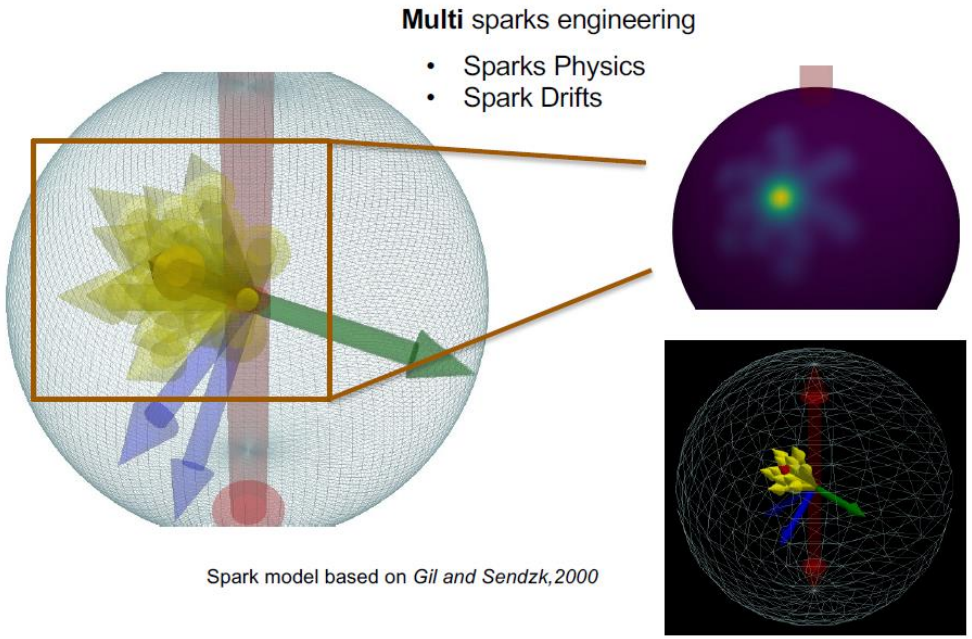




Status & plans

- Data
  - more Effelsberg data (other pulsars)
  - MeerKAT data
- Physics-driven DT
  - better pulsar model (beam properties etc.)
  - improving ISM model
  - interface
- ML-classifier
  - improving low SNR performance (exploring different ML architecture)
  - distributed training (HeAT, Horovod)
  - de-dispersion (TransientX)
- Deployment and HPC testing (different data centers)
- New release (v. 0.2) soon





Thank you!

Questions?

www.intertwin.eu

info@intertwin.eu

intertwin_eu

intertwin