Improving Pulsar Observations by Training a Machine Learning Model on the EPTA Dataset

Jacob Cardinal Tremblay[1], Jędrzej Jawor[2]
[1]Max Planck Institute for Gravitational Physics (Albert Einstein Institute)
[2]Max Planck Institute for Radio Astronomy

EPTA Data and Pipeline

- Up to 25 years of observations (EPTA Collaboration et al. 2023)
  - High-quality folded data from the Effelsberg 100-meter radio telescope
- New pipeline for Radio Frequency Interference (RFI) cleaning
  - RCVR → CLFD → MEERSurg → Manual Cleaning
- Statistical algorithms need variable thresholding
  - Manual oversight required
  - Non-standardized way of cleaning
- Ultra-Broadband Receiver (UBB)
  - ➢ Future cleaning will be much more intensive
- Improve Time of Arrival (TOA) uncertainties by:
  - Removing more RFI
  - Eliminating less "clean" data

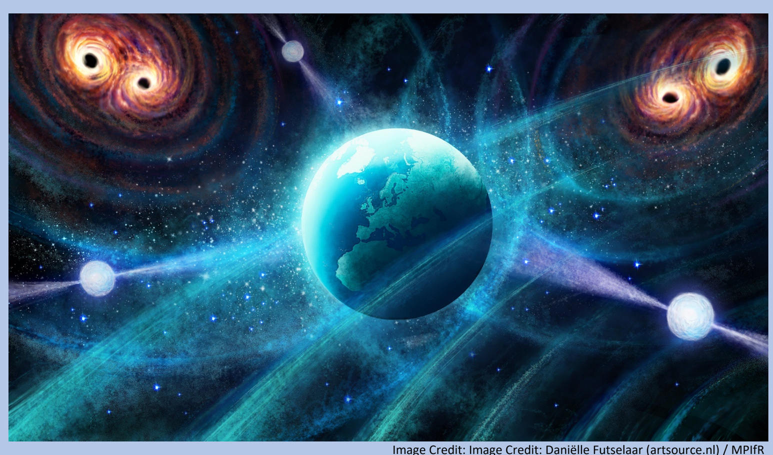
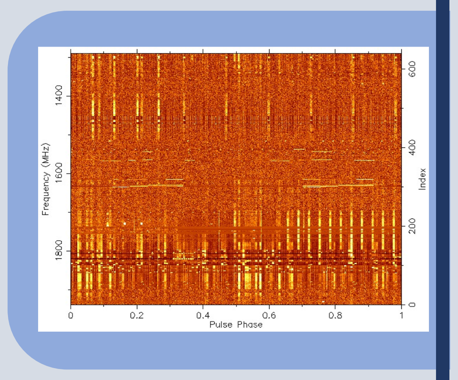


Image Credit: Image Credit: Daniëlle Futselaar (artsource.nl) / MPIfR

Radio Frequency Interference

- Radio Frequency Interference (RFI) = Unwanted human-made signals
- Large frequency bandwidth = A lot of RFI!
- RFI can reduce quality or even completely corrupt data!
- Can be narrow-band or wide-band



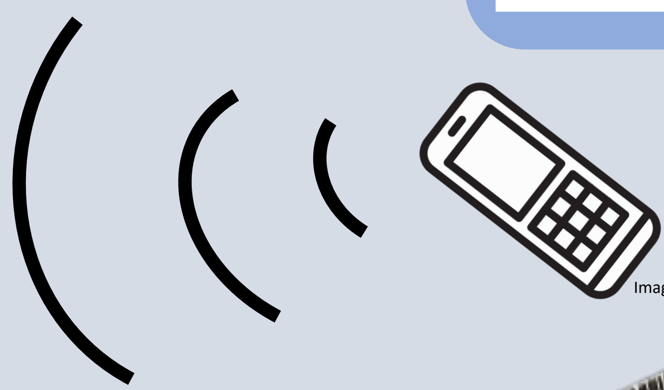


Image Credit: Selman Icons

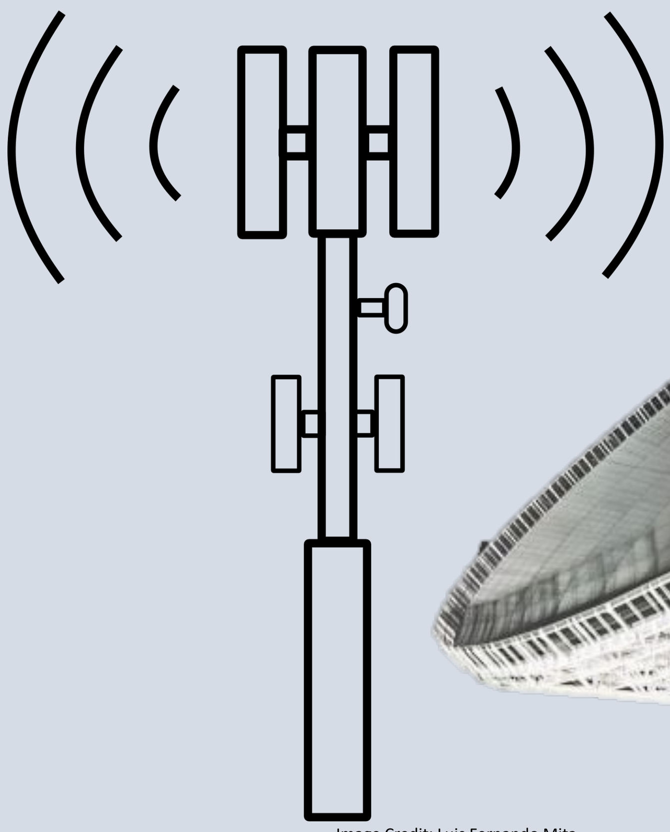


Image Credit: Luis Fernando Mita

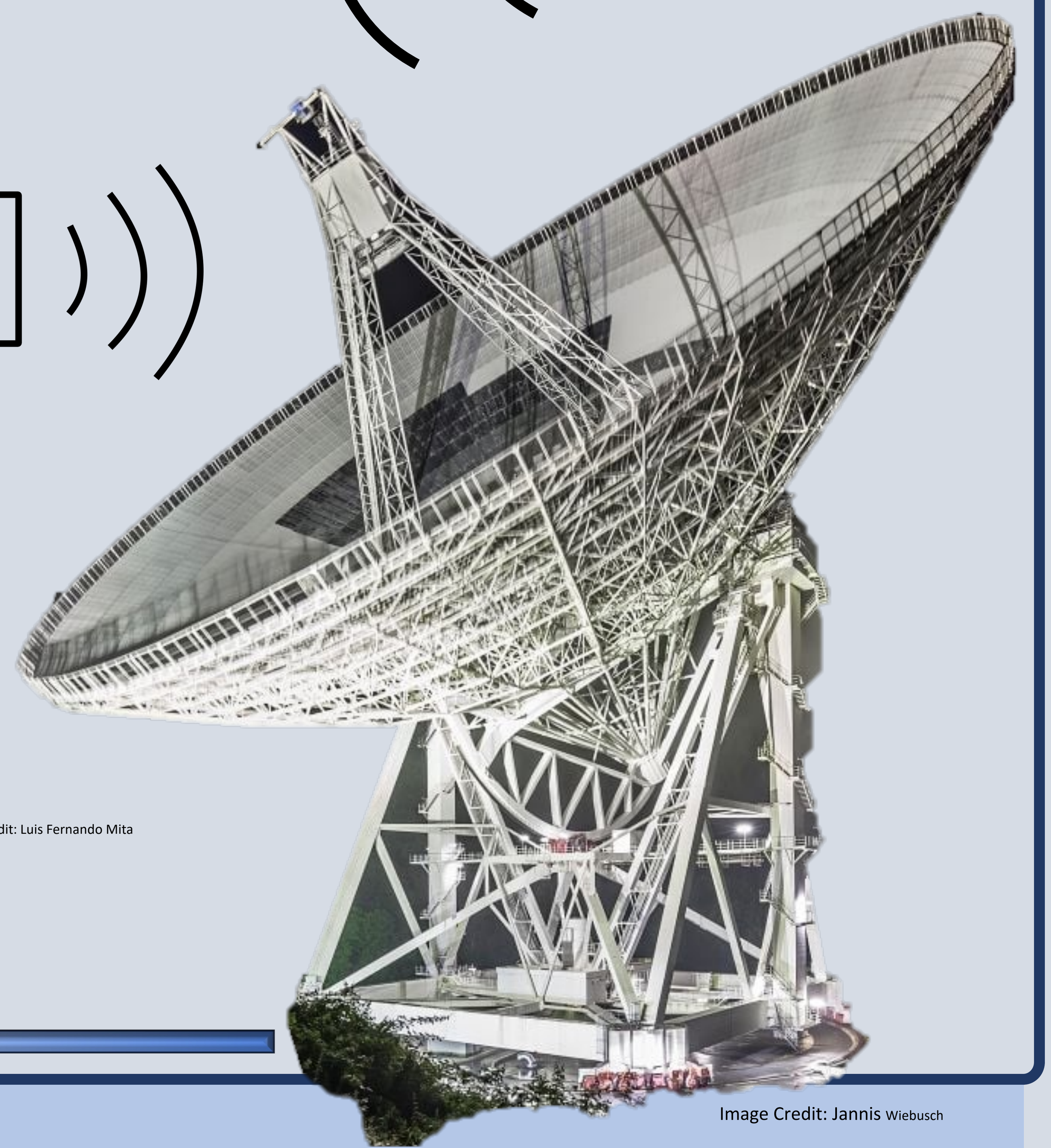
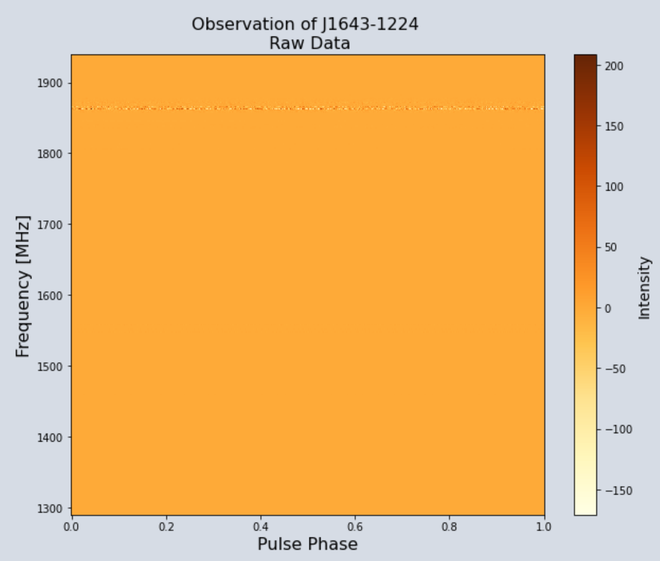


Image Credit: Jannis Wiebusch



Machine Learning

- Train a model on EPTA dataset to identify RFI
  - Random Forest/Gradient Boosting are ideal this case
  - E.g. A. Berthereau et al. 2023
- Obtain weights from previously cleaned observations
- Calculate statistics on raw data
  - E.g. median, mean, skewness, kurtosis
- Use statistics as features on which to train
  - Do not train on the actual raw data
- 80% of the data for training, 20% for testing

Observed \ Predicted	No RFI	RFI
No RFI	True Positive (TP)	False Positive (FP)
RFI	False Negative (FN)	True Negative (TN)

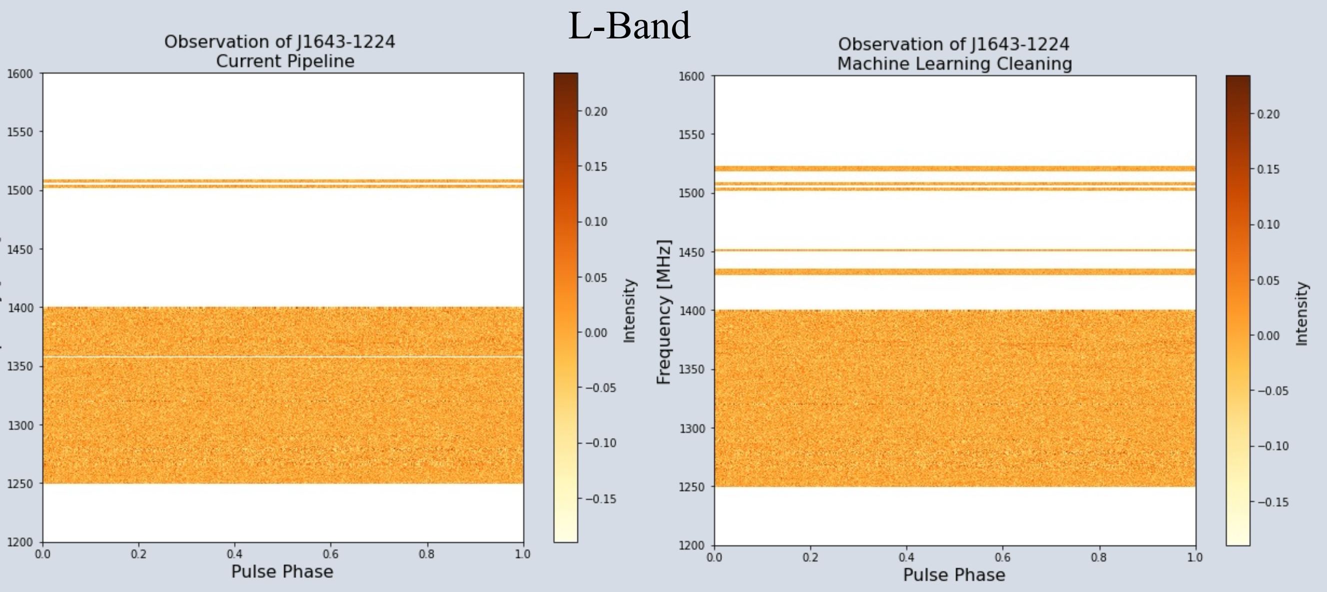
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

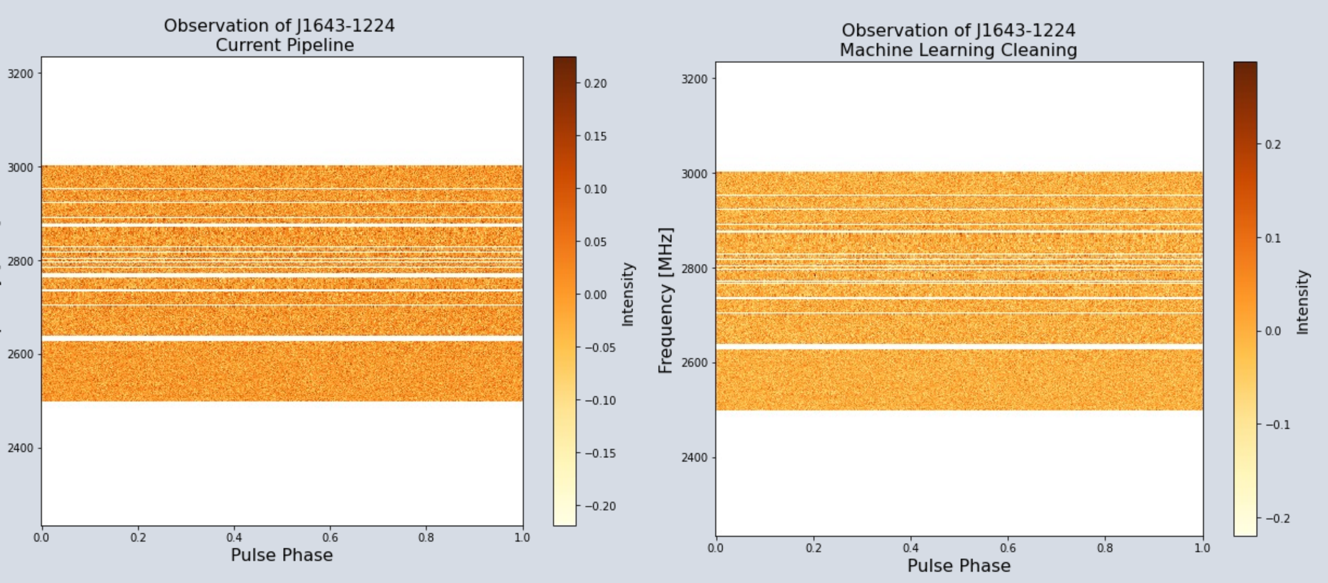
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1\text{-Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

L-Band

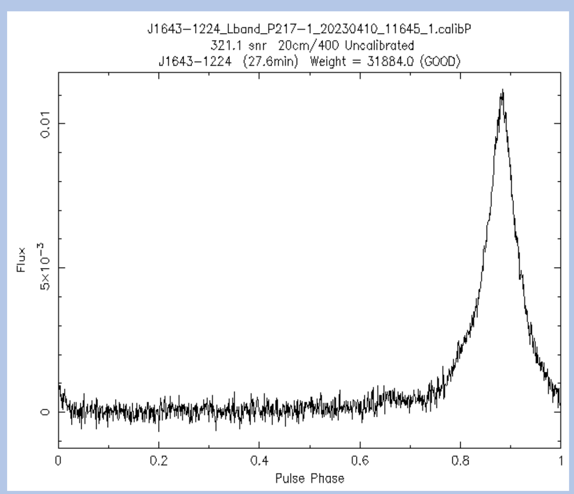


S-Band



Testing

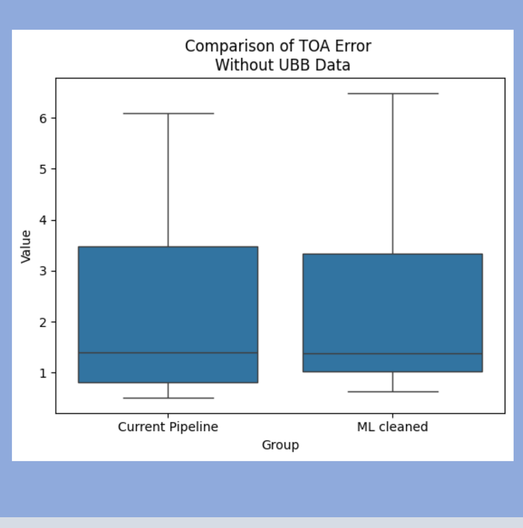
- Initial tests on pulsar J1643-1224
  - Very bright pulsar
  - "Single-peak" profile
- 74 EDD observations → ~ 6 million individual frequency channels
- Dataset is relatively balanced
  - 44% RFI, 56% no RFI

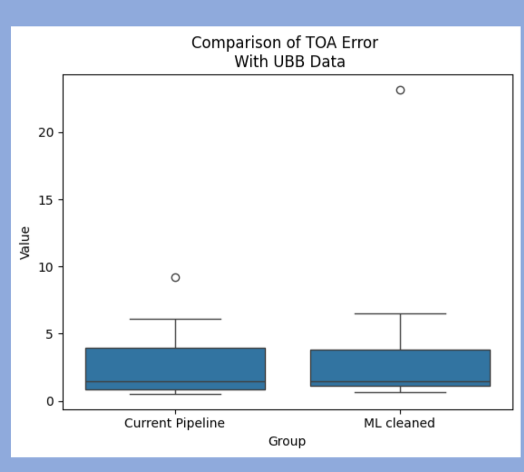


- Tested ~16,000 combinations of features to arrive at best features based on accuracy and F1-Score
  - frequency, median, mean, skewness, max

Current Results

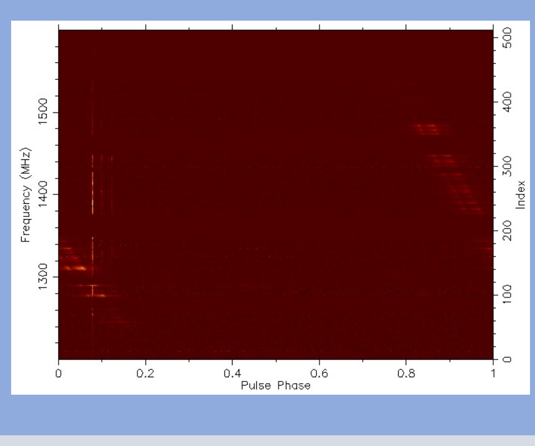
- Calculated metrics:
  - Accuracy: 94.06%
  - Precision: 90.62%
  - Recall: 99.17%
  - F1-Score: 94.70%
- TOAs calculated from cleaning
  - Uncertainties in TOAs are compared
  - ML cleaning is slightly more consistent
- However, performs worse on UBB data





Future Work

- Explore more machine learning models?
- Hyperparameter tuning
- Train on the off-pulse
- Train on multiple backends
- Train on multiple pulsars
- Train and/or test model on single pulse data
- Include wideband RFI elimination



References

EPTA collaboration, The second data release from the European Pulsar Timing Array I. The dataset and timing analysis, Astron. Astrophys. 678 (2023) A48 [2306.16224].

Anaïs Berthereau. Méthodes avancées en chronométrie des pulsars : réjection d'interférences radio par apprentissage profond et étude chronométrique du système binaire relativiste J1528-3146. Astrophysique [astro-ph]. Université d'Orléans, 2023. Français. ffNNT : 2023ORLE1043ff. fftel-04509283

Morello, V., Barr, E. D., Cooper, S., et al. 2019, MNRAS, 483, 3673

MeerGuard: https://github.com/danielreardon/MeerGuard

CoastGuard3: https://gitlab.mpcdf.mpg.de/jjawor/CG-3

Acknowledgements

Special thanks to: Marie Dumaz, Rutger van Haasteren, Boris Goncharov, Karsten Beismann, David Champion and Michael Kramer for the various roles they played in allowing this work to become a reality.