Pulsar Search Using Supervised Machine Learning

John M. Ford – Advisor: Dr. Sumitra Mukherjee

Introduction

- Pulsars are rapidly rotating neutron stars which emit a strong beam of energy. They are used to study many basic astrophysical phenomena.
- Study of these physical phenomena requires a large ensemble of pulsars to adequately sample the parameter space.
- Searching for pulsars is currently a very labor-intensive process.
- Research to date has not yielded a satisfactory automated system, with 7% of pulsars in a test data set missed by the most successful automated system to date.
- This research proposes to research, identify, and propose methods to overcome the barriers to building an improved classification system:
  - With a false positive rate of less than 5%
  - A recall of at least 99%

Research Goals

Develop an improved method of pulsar identification using supervised machine learning techniques that can achieve:

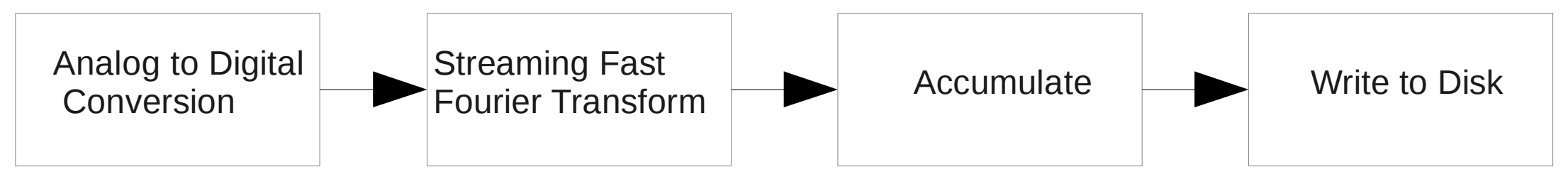
- A *false positive rate* of less than 5%
- A *Precision* of greater than 3.6%
- A *Recall* of greater than 99%

These are formidable specifications to meet, particularly the Recall of 99%

Pulsar Data Processing

Standard Pulsar Data Collection Process

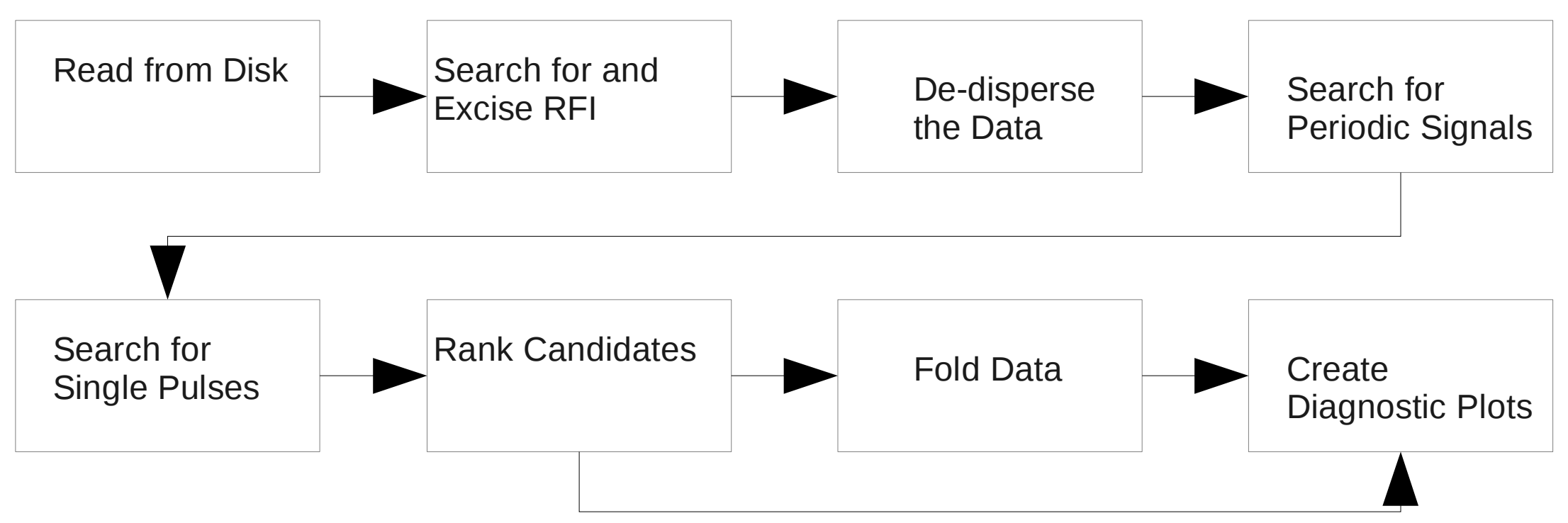
- Real-time streaming process
- 1.6 GS/sec per input sampling rate
- 400 MB/sec output data rate to disk



Analog to Digital Conversion → Streaming Fast Fourier Transform → Accumulate → Write to Disk

Standard Pulsar Signal Processing Pipeline (Off-line Processing)

- Embarrassingly parallel algorithms
- Approximately 1 CPU-day per minute of data acquisition
- Generates the diagnostic plots to the right



Read from Disk → Search for and Excise RFI → De-disperse the Data → Search for Periodic Signals → Search for Single Pulses → Rank Candidates → Fold Data → Create Diagnostic Plots

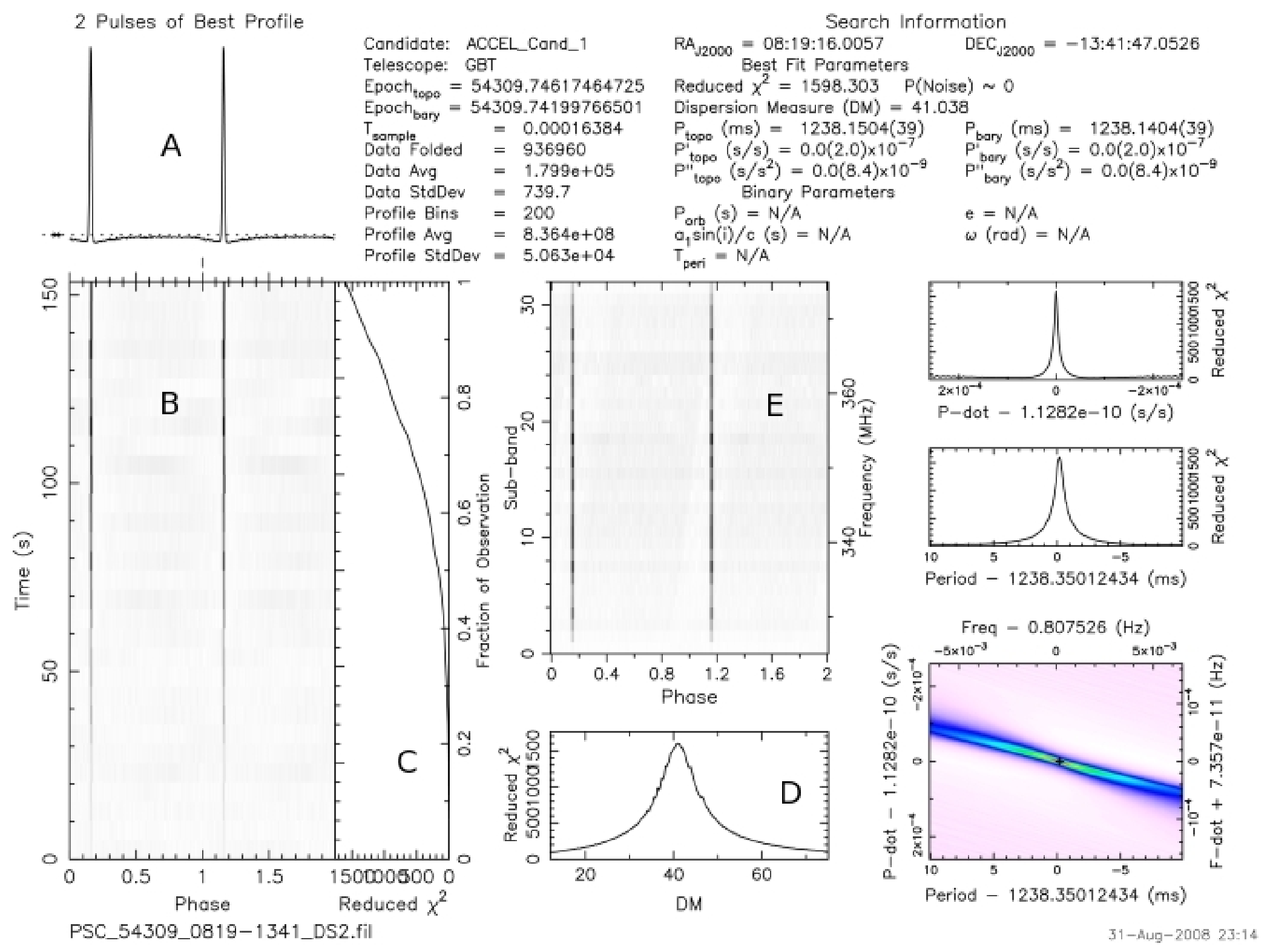
Diagnostic Plots

There are four main features in the plots that astronomers use when deciding if a candidate could be a pulsar. Refer to the plots in the next column.

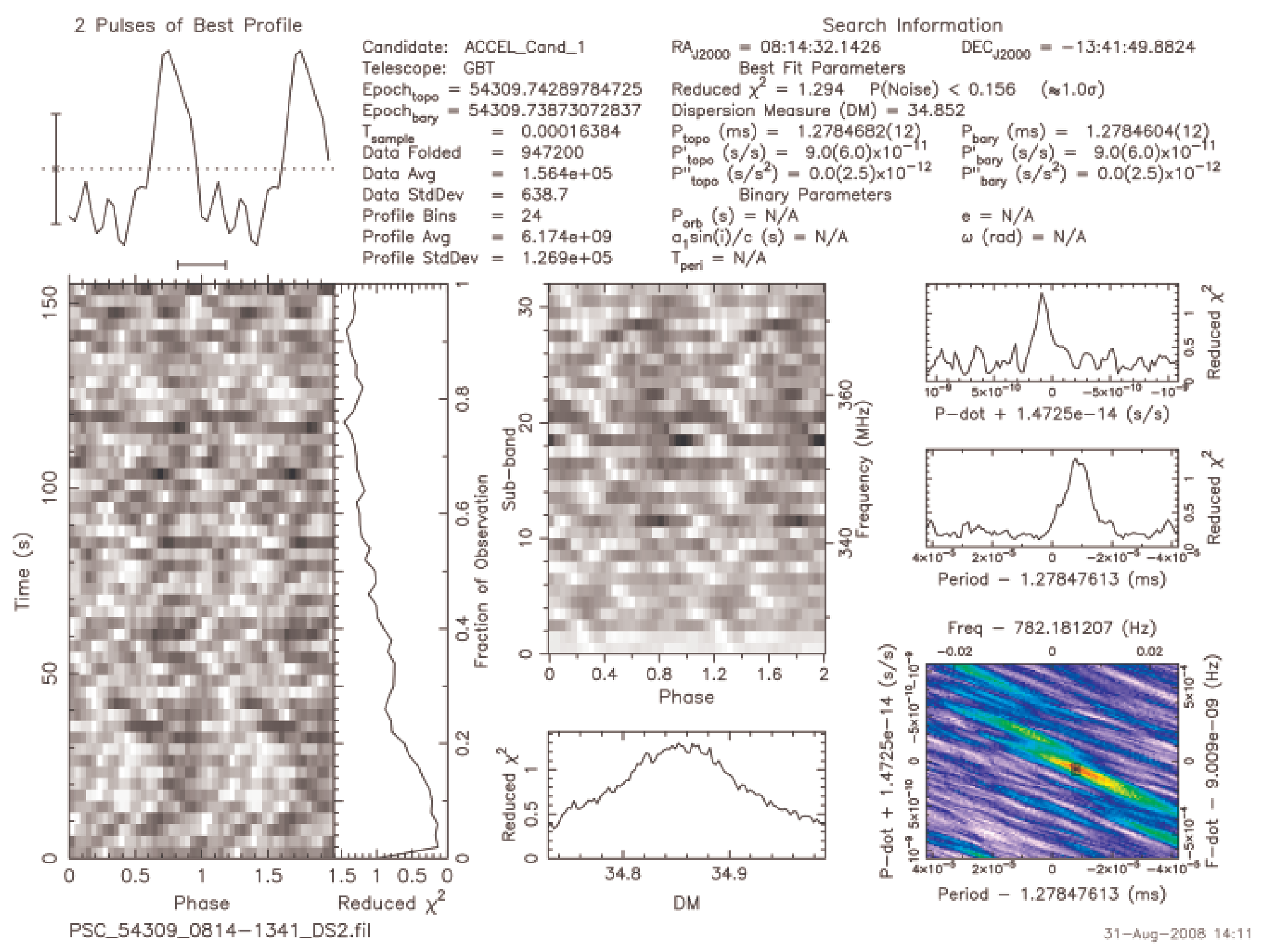
- In the *2 Pulses of Best Profile* plot in the upper left-hand corner of the figure, labeled “A”, the peaks should be significantly above the noise floor. Compare the error bars in the lower left corner of the same subplot in the other Figures.
- In the *Phase vs Time* plot, labeled “B”, vertical lines in phase with the peaks should appear throughout the entire observation time, unless the telescope beam is drifting across the sky, in which case the pulsar should smoothly come into the beam and drift out later. This indicates that the signal is continuous in time.
- In the *Phase vs Frequency* plot, labeled “E”, the vertical lines should also span most of the frequency space, indicating the signal is a broadband signal. Compare this with the signal in a plot of a man-made interference signal, where the signal is present only at a narrow band of frequencies.
- A bell-shaped curve in the *DM vs Reduced* χ^2 plot, labeled “D”, shows that the signal’s reduced χ^2 value depends strongly on DM, peaking at the trial Dispersion Measure (DM). Compare this with the plot in the other figures, where there is no strong dependence of Reduced χ^2 with DM.

Diagnostic Plots

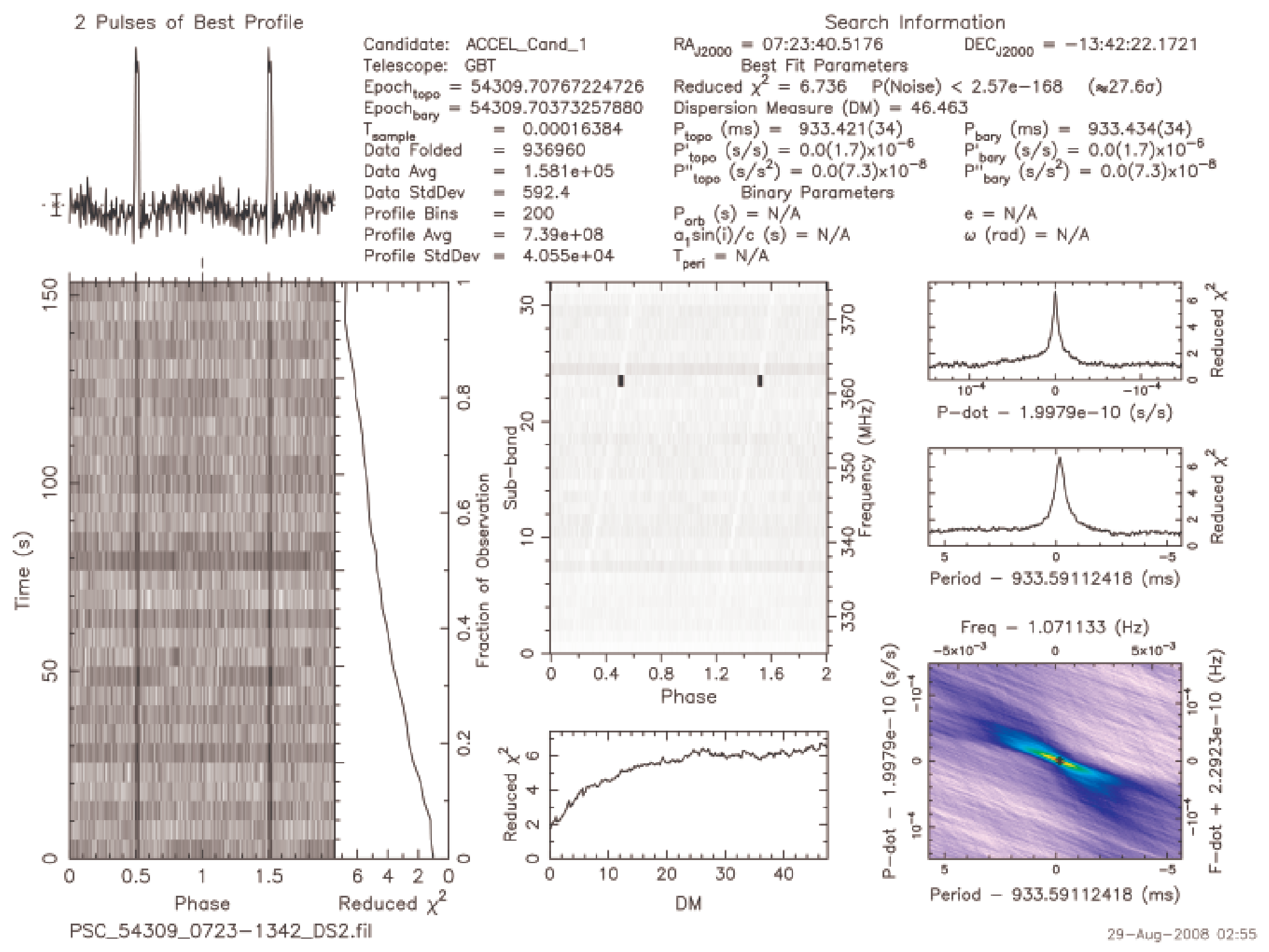
A diagnostic plot for Pulsar J0820-1350 Adapted with permission from Heatherly (2013)



A diagnostic plot containing only background noise



A diagnostic plot containing radio frequency interference



Previous Work

- Eatough et al. (2010) was the first published attempt to use a machine learning approach to examine diagnostic plots. Their approach was using an Artificial Neural Network (ANN) using the feature set described below.
- Bates et al. (2012) also attempted to use this method to find pulsars, but were not as successful.
- Keith et al. (2009) found 28 more pulsars in a data set that had already been mined by using an algorithmic approach to scoring plots, as opposed to a machine learning approach.
- A very recent system called PEACE: Pulsar Evaluation Algorithm for Candidate Extraction (Lee et al., 2013) demonstrated the utility of careful feature selection in an algorithm similar to the one in (Keith et al., 2009).

ANN Performance Statistics from Eatough et al. (2010)

	Pulsar	Non-Pulsar	Total
Selected	TP=465	FP=12535	13,000
Not selected	FN = 36	TN=2486964	2487000
Total	501	2499499	2500000

In the following statistics, the numerical values given are from Table 1

The *False Positive Rate* of the system is a measure of how well the system rejects the undesired class. It is defined as:

$$FPR = \frac{(100 * FP)}{FP + TN} = 5\% \tag{1}$$

The *Precision* of the system is a measure of how well the system discriminates between the classes. It is defined as:

$$Precision = \frac{TP}{TP + FP} = 3.6\% \tag{2}$$

The *Recall* of the system is a measure of the amount of data that is lost by the system. It is defined as:

$$Recall = \frac{TP}{TP + FN} = 92.8\% \tag{3}$$

Proposed Methodology

- Reproduce the study by Eatough et al. (2010).
- Develop a support vector machine using “R” (R Core Team, 2013) or “WEKA” (Witten, Frank, & Hall, 2011) using the same training data as Eatough et al. (2010).
- Study the characteristics of normal and millisecond pulsars.
- Study the information available for each pointing in the data archive.
- Experiment with ensemble classifiers and cascade classifiers using different randomly selected sub-sampled sets of the training data. This can yield classifiers that, together, perform better than a single classifier.
- Experiment with creating synthetic minority class members using the MSMOTE algorithm (Hu, Liang, Ma, & He, 2009). Using MSMOTE and knowledge of the physics of pulsars, additional training data can be created to assist in training the system to find rare pulsars.
- Design, prototype, train, test, and evaluate the most promising algorithms.

Summary and References

Summary

Pulsars are used in probes of fundamental physics, such as general relativity (Lorimer & Kramer, 2005). Several large-scale pulsar surveys are underway, which will generate millions of possible pulsar candidates. An automated system with a low false positive rate and high recall is needed to enable analysis of these surveys.

References

Bates, S., Bailes, M., Barsdell, B., Bhat, N., Burgay, M., Burke-Spolaor, S., . . . van Straten, W. (2012, September). The High Time Resolution Universe Survey VI: An Artificial Neural Network and Timing of 75 Pulsars. *ArXiv e-prints*. arXiv: 1209.0793 [astro-ph.SR]

Eatough, R., Molkenthin, N., Kramer, M., Noutsos, A., Keith, M., Stappers, B., & Lyne, A. (2010, October). Selection of radio pulsar candidates using artificial neural networks. *MNRAS*, *407*, 2443–2450. doi:10.1111/j.1365-2966.2010.17082.x. arXiv: 1005.5068 [astro-ph.IM]

Heatherly, S. A. (2013). Pulsar search collaboratory home page. Retrieved from https://sites.google.com/a/pulsarsearchcollaboratory.com/pulsar-search-collaboratory/Home

Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: improving classification performance when training data is imbalanced. In *Computer science and engineering, 2009. wcse ’09. second international workshop on* (Vol. 2, pp. 13–17). doi:10.1109/WCSE.2009.756

Keith, M., Eatough, R., Lyne, A., Kramer, M., Possenti, A., Camilo, F., & Manchester, R. (2009, May). Discovery of 28 pulsars using new techniques for sorting pulsar candidates. *MNRAS*, *395*, 837–846. doi:10.1111/j.1365-2966.2009.14543.x. arXiv: 0901.3570 [astro-ph.SR]

Lee, K., Stovall, K., Jenet, F., Martinez, J., Dartez, L., Mata, A., . . . Zhu, W. (2013, May). PEACE: pulsar evaluation algorithm for candidate extraction - a software package for post-analysis processing of pulsar survey candidates. *MNRAS*. doi:10.1093/mnras/stt758. arXiv: 1305.0447 [astro-ph.IM]

Lorimer, D., & Kramer, M. (2005). *Handbook of pulsar astronomy* (First). Cambridge, UK: Cambridge University Press.

R Core Team. (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques* (3rd). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Graduate School of Computer and Information Sciences - Nova Southeastern University
Mail: jf1054@nova.edu
WWW: http://scis.nova.edu