**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 6.3%
- 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 Gb/sec per input sampling rate
  - 400 MB/sec output data rate to disk

- Standard Pulsar Signal Processing Pipeline (Off-line Processing)
  - Emarrassingly parallel algorithms
  - Approximately 1 CPU-day per minute of data acquisition
  - Generates the diagnostic plots to the right

---

**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 **Phase vs Time plot**, labeled "A", 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 **Candidate** could be a pulsar. Refer to the plots in the next column.

---

**ANN Performance Statistics from Eatough et al. (2010)**

<table>
<thead>
<tr>
<th>Pulsar</th>
<th>Non-Pulsar</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected</td>
<td>TP=1454</td>
<td>FP=12985</td>
</tr>
<tr>
<td>Missed</td>
<td>TN=12672</td>
<td>FN=248700</td>
</tr>
<tr>
<td>Total</td>
<td>TP+FP=16174</td>
<td>FP+TN=1408057</td>
</tr>
</tbody>
</table>

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:

\[
PFR = \frac{100 \times FP}{TP + FP}
\]

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

\[
Precision = \frac{TP}{TP + FP}
\]

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

\[
Recall = \frac{TP}{TP + FN}
\]

---

**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**


