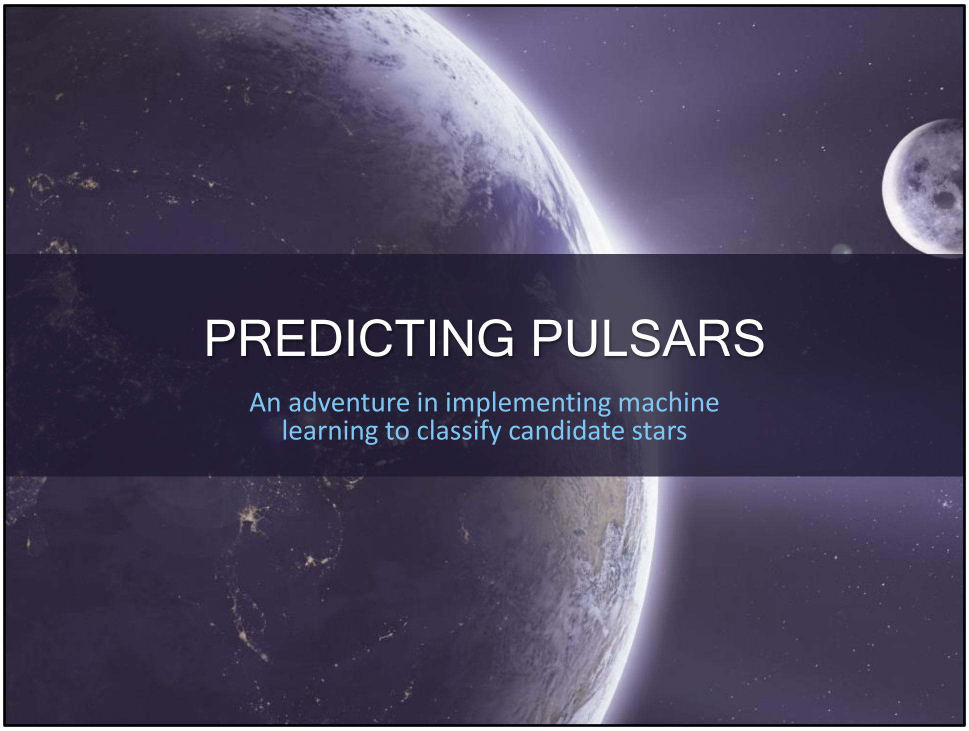
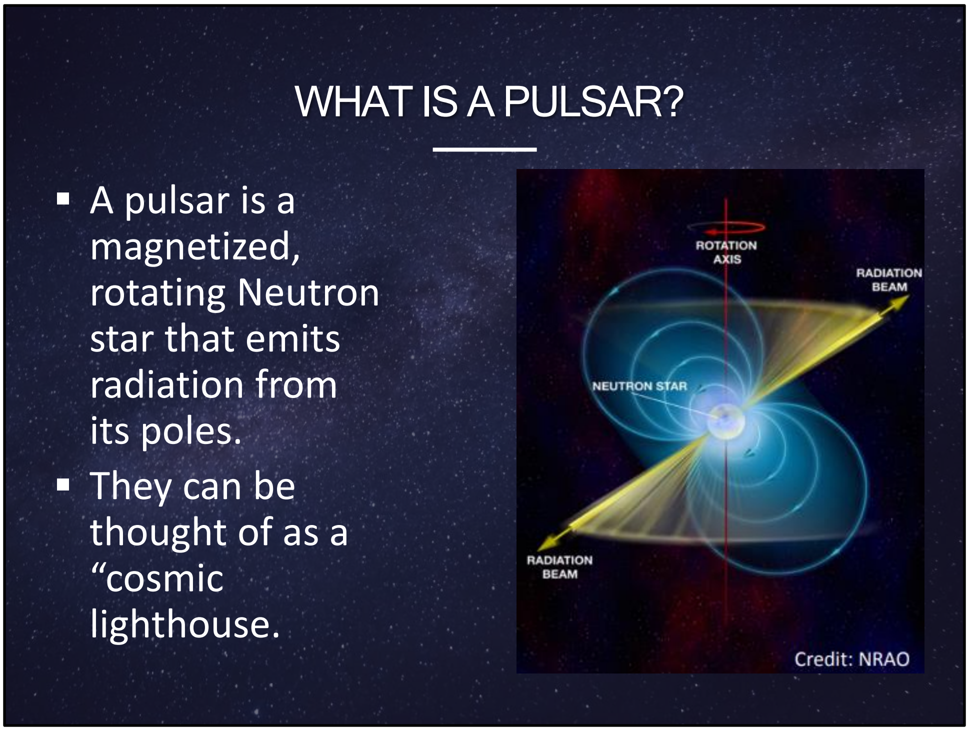


PREDICTING PULSARS

An adventure in implementing machine learning to classify candidate stars

In this project, I assumed the role of a data scientist working with an astronomical research agency to train a machine learning model to classify whether a star is a pulsar or not. To accomplish this task, I used Dr Robert Lyon’s Predicting a Pulsar Star dataset from Kaggle.

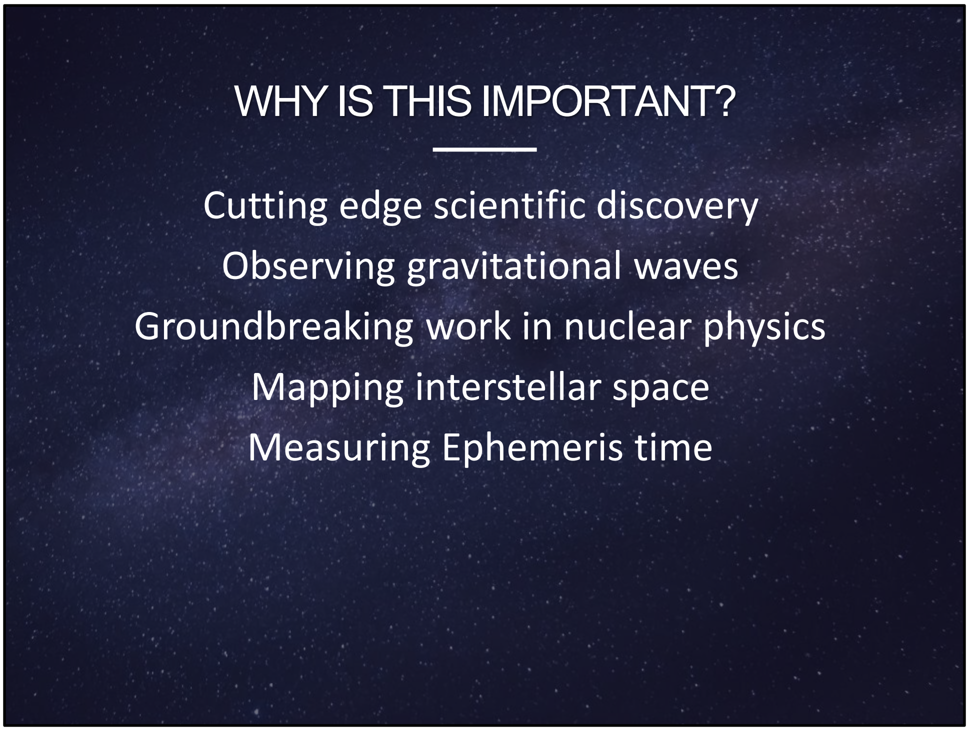


WHAT IS A PULSAR?

- A pulsar is a magnetized, rotating Neutron star that emits radiation from its poles.
- They can be thought of as a "cosmic lighthouse.

Credit: NRAO

A pulsar is a rotating, magnetized Neutron star that emits radiation(mostly radio, X-ray, and gamma wavelength) from its poles. Other than black holes, pulsars are the most dense objects in the known universe. Pulsars rotate with a very short and precise period which vary from star to star, however due to Radio Frequency Interference and background radiation noise, there are many false-positives when attempting to detect pulsars.



WHY IS THIS IMPORTANT?

Cutting edge scientific discovery

Observing gravitational waves

Groundbreaking work in nuclear physics

Mapping interstellar space

Measuring Ephemeris time

Why is pulsar research important?

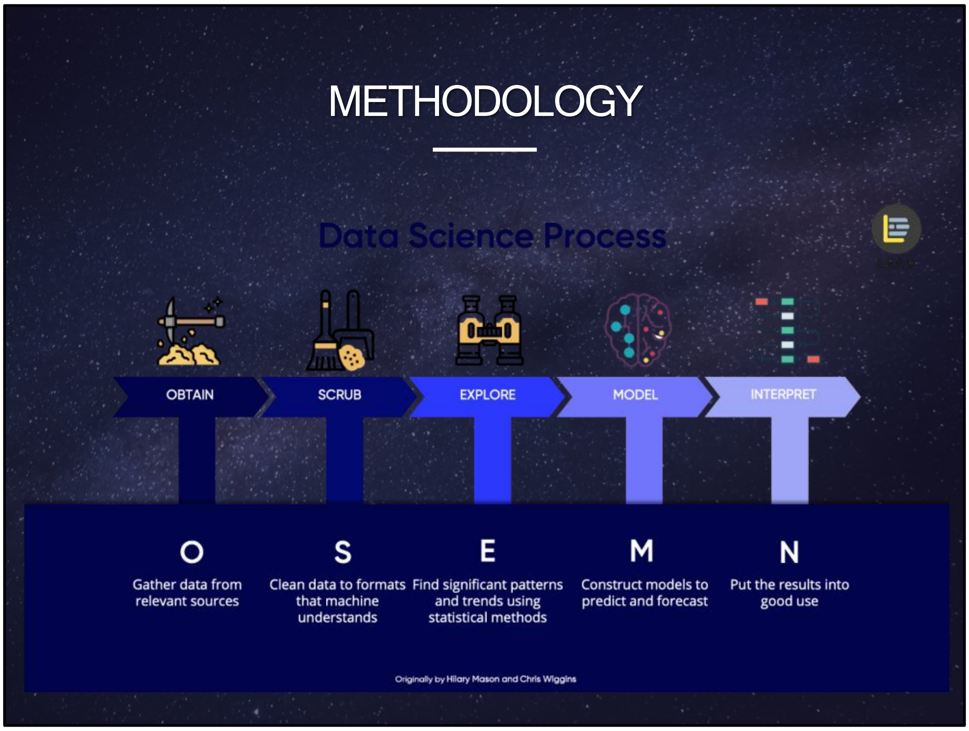
Pulsar research has already yielded two Nobel Peace Prizes, and is at the tip of the spear in Humanity’s quest to understand and explore the cosmos.

By observing the precisely timed signals from various pulsars around the galaxy, astronomers aim to measure the slight variations in these signals which are thought to be partially affected by gravitational waves. Pulsar timing arrays are being used to uncover secrets of the early universe and to test the ideas of general and special relativity.

Aside from black holes, pulsars are the most dense objects in the known universe. They’re very small (~10km radius) objects but have an average mass of ~1.4 solar masses. The density of a pulsar is several times greater than the density of atomic nuclei. There is little understood about the physics of what goes on inside a pulsar and any discovery related to this would be groundbreaking.

Due to the regularity of the signal emissions, pulsars can be reliably used to triangulate position in the galaxy. Humanity has included pulsar maps on the two Pioneer Plaques in addition to the Voyager Golden Record. As Humanity progresses to a type II civilization, this type of navigation could prove very useful.

And finally, again due to the regularity of the signals, a pulsar’s pulse can be used to measure time, particularly Ephemeris time – which is free of the irregular constraints related to fluctuating mean solar time.



METHODOLOGY

Data Science Process

O
Gather data from relevant sources

S
Clean data to formats that machine understands

E
Find significant patterns and trends using statistical methods

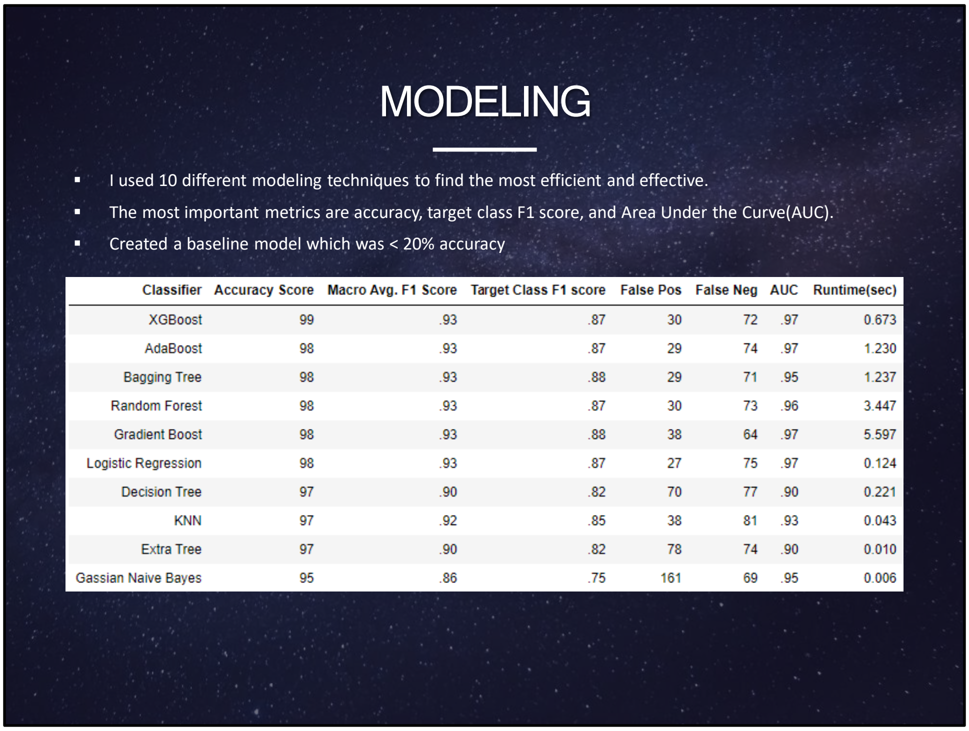
M
Construct models to predict and forecast

N
Put the results into good use

Originally by Hilary Mason and Chris Wiggins

I used the OSEMiN (oh-sem) model for data science to attack this problem.

A further look into my methodology and execution is available in my Jupyter Notebook.



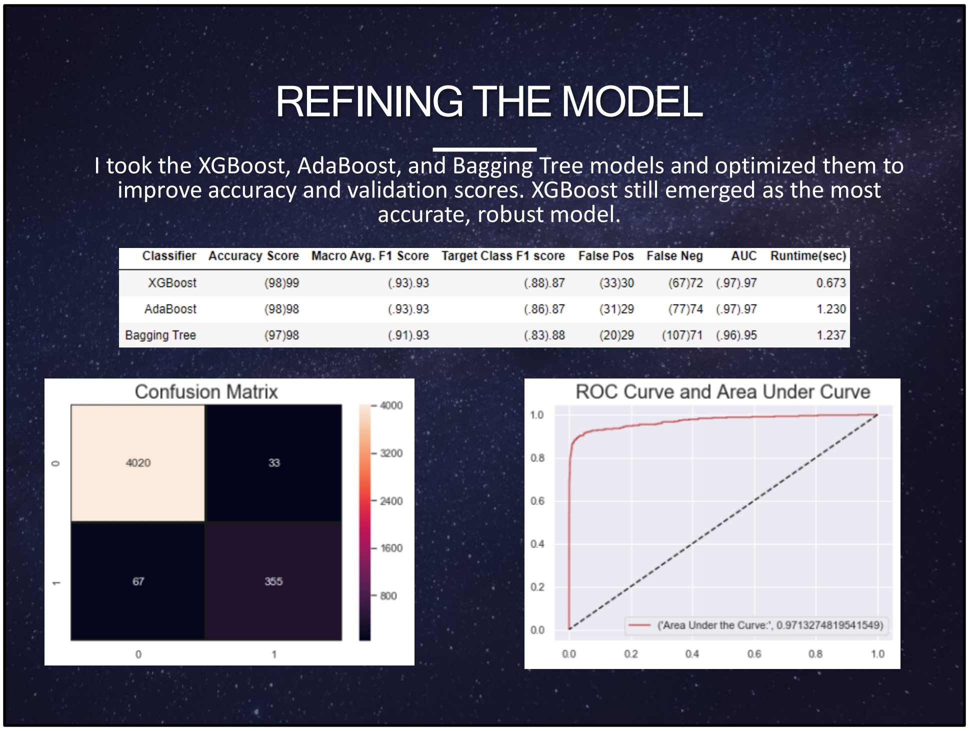
MODELING

- I used 10 different modeling techniques to find the most efficient and effective.
- The most important metrics are accuracy, target class F1 score, and Area Under the Curve(AUC).
- Created a baseline model which was < 20% accuracy

Classifier	Accuracy Score	Macro Avg. F1 Score	Target Class F1 score	False Pos	False Neg	AUC	Runtime(sec)
XGBoost	99	.93	.87	30	72	.97	0.673
AdaBoost	98	.93	.87	29	74	.97	1.230
Bagging Tree	98	.93	.88	29	71	.95	1.237
Random Forest	98	.93	.87	30	73	.96	3.447
Gradient Boost	98	.93	.88	38	64	.97	5.597
Logistic Regression	98	.93	.87	27	75	.97	0.124
Decision Tree	97	.90	.82	70	77	.90	0.221
KNN	97	.92	.85	38	81	.93	0.043
Extra Tree	97	.90	.82	78	74	.90	0.010
Gassian Naive Bayes	95	.86	.75	161	69	.95	0.006

I used 10 different modeling techniques/classifiers to find the most efficient and effective.
The most important metrics are accuracy, target class F1 score, and AUC.
Here are the results of the models prior to tuning them to increase accuracy.
These tests were performed on non-normalized data but synthetically oversampled, to account for a large class imbalance.

Prior to dealing with class imbalance, I completed a baseline Naïve Bayes model as a starting point, it achieved an accuracy score of 19% and had many false positive and false negatives.



REFINING THE MODEL

I took the XGBoost, AdaBoost, and Bagging Tree models and optimized them to improve accuracy and validation scores. XGBoost still emerged as the most accurate, robust model.

Classifier	Accuracy Score	Macro Avg. F1 Score	Target Class F1 score	False Pos	False Neg	AUC	Runtime(sec)
XGBoost	(98)99	(.93).93	(.88).87	(33)30	(67)72	(.97).97	0.673
AdaBoost	(98)98	(.93).93	(.86).87	(31)29	(77)74	(.97).97	1.230
Bagging Tree	(97)98	(.91).93	(.83).88	(20)29	(107)71	(.96).95	1.237

From the preliminary models, I took the three best performing ones and optimized their parameters.

The best model was still the XGBoost model, boasting 98% accuracy and 97.7% validation score.

RECOMMENDATIONS & FUTURE WORK

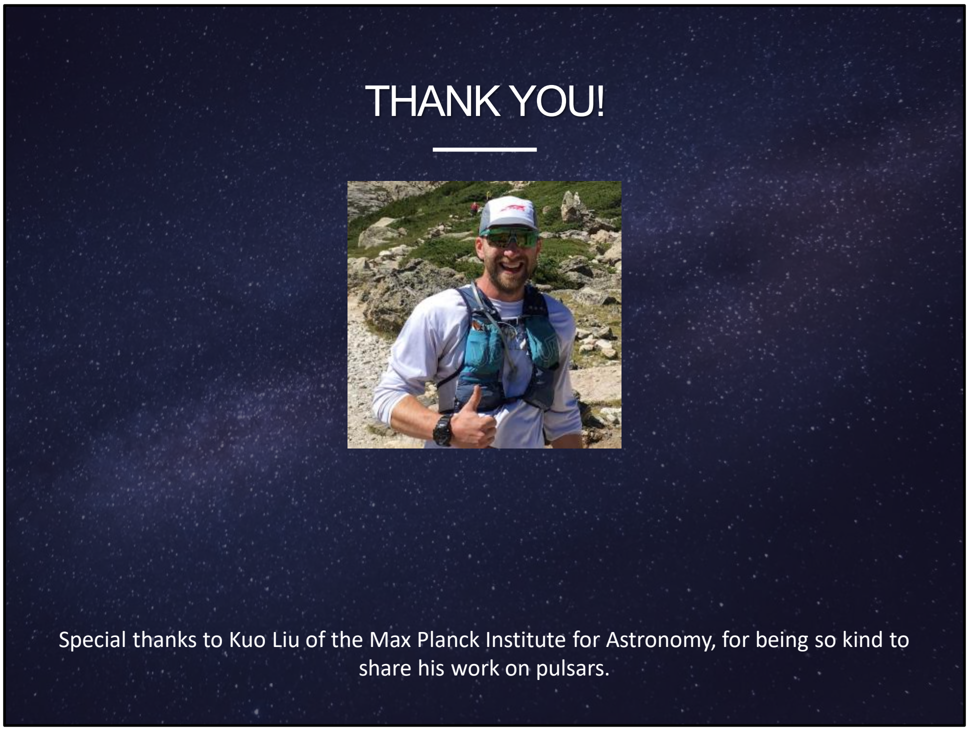
Overall, the tuned models performed well, especially XGBoost. With the ability to quickly and accurately predict pulsars from a set of candidate stars, we can identify stars for further observation and research. Here are my final thoughts on the project:

- Use the XGBoost model, it's faster, more accurate, and more resisitant to overfitting the data.
- Use a larger dataset with more anomalies.
- Use a more powerful computer or distributed computing.
- Use more complex machine learning or deep learning techniques.

With these recommendations we can work to build a fully automated classification algorithm for candidate stars and continue to push pulsar research to the next level.

Overall, the tuned models performed well, especially XGBoost. With the ability to quickly and accurately predict pulsars from a set of candidate stars, we can identify stars for further observation and research. Here are my final thoughts on the project:

Use the XGBoost model, it's the fastest, most accurate, and most resistant to overfitting of all the tested classifiers.
With a larger dataset with more anomalies and oddities, we could train a more robust model that could effectively pre-process the data by spotting false positive candidate stars earlier.
More computational power could aid this quest like using a supercomputer or a distributed system like those found in observational astronomy labs.
Using more complex machine learning or deep learning techniques could yield more accurate results.

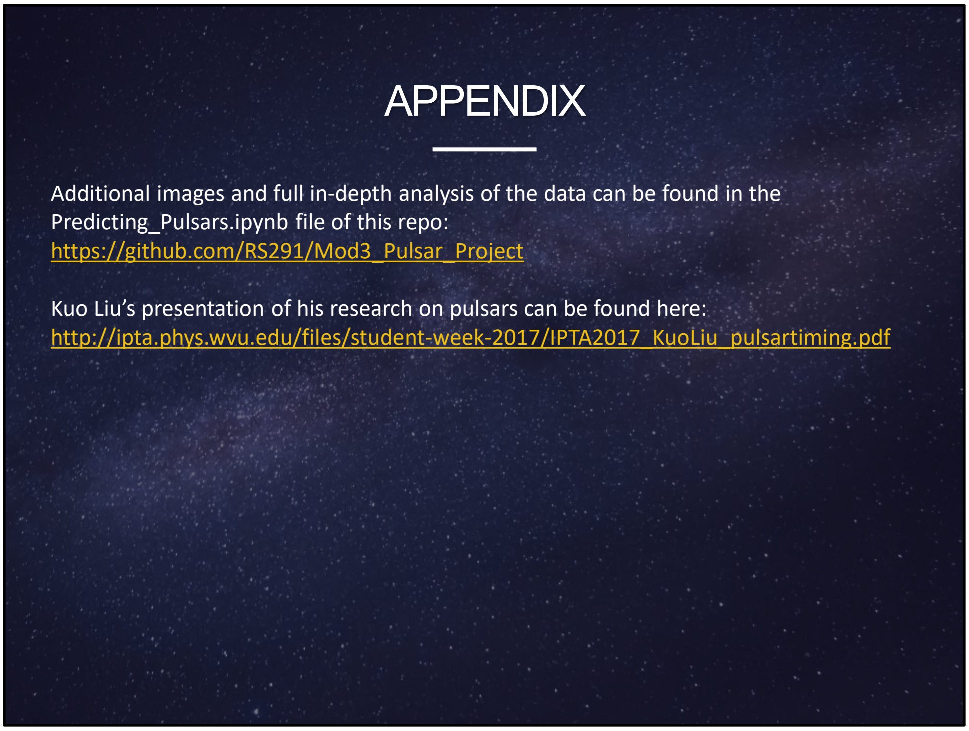


THANK YOU!

Special thanks to Kuo Liu of the Max Planck Institute for Astronomy, for being so kind to share his work on pulsars.

Thank you!

Special thanks to Kuo Liu for sharing his work on pulsars with me



APPENDIX

Additional images and full in-depth analysis of the data can be found in the Predicting_Pulsars.ipynb file of this repo:
https://github.com/RS291/Mod3_Pulsar_Project

Kuo Liu’s presentation of his research on pulsars can be found here:
http://ipta.phys.wvu.edu/files/student-week-2017/IPTA2017_KuoLiu_pulsartiming.pdf

Appendix