Machine Recognition of "Squiggles" in SETI Signal Data

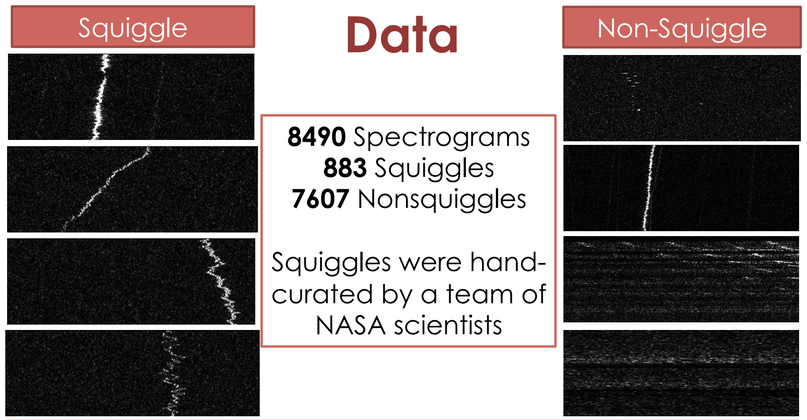
Travis Chen, Kenny Smith, Jason Wang

Motivation

The question, "*are we alone*?," has boggled scientists for centuries. The Search for Extraterrestrial Intelligence (SETI) Institute operates the Allen Telescope Array, a set of 42 antennas, to observe star systems for radio signals which may provide evidence of ET intelligence. The key problem is identifying recurring patterns in the signal stream that are not associated with known interferences, such as aircraft RFI, radio waves, etc. Recently, an unknown subset of signals inelegantly referred to as "squiggles" has become prevalent. We seek to study squiggles and their origin in two ways:

Classify Squiggle vs Nonsquiggle ⟷ Identify Squiggle Subgroups

Data

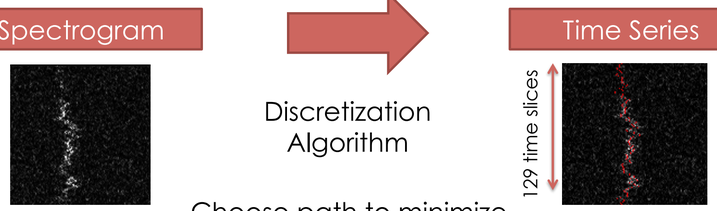


8490 Spectrograms
883 Squiggles
7607 Nonsquiggles

Squiggles were hand-curated by a team of NASA scientists

Figure 1: Each spectrogram image is 768x129 pixels, which corresponds to ~100 MHz in bandwidth (x-axis) and 93 seconds in time (y-axis).

Approach



Choose path to minimize

$L(\alpha, \beta) =$

$-\alpha*(\text{Intensity})-\beta*(\text{Neighbor Intensities})+(1-\alpha-\beta)(\text{Deviation})^2$

Figure 2: The dynamic programming algorithm traces the optimal 129-time slice vertical path. We tuned our parameters to $\alpha = 0.5,\ \beta = 0$.

Classification

Baseline

Logistic Regression on 129 scaled raw time series points

Model	ACC	AUC
Unregularized	**0.875**	**0.504**
Lasso (L1)	**0.875**	**0.5**
Ridge (L2)	**0.875**	**0.5**

Figure 3: The full dataset was split 90% training, 10% test. Using the training set, we applied 10-fold cross validation to tune model parameters. The ACC and AUC values regard the initial validation test set. Our baseline models classify the entire test set as nonsquiggles, which comprise 90% of the training set.

Intermediate

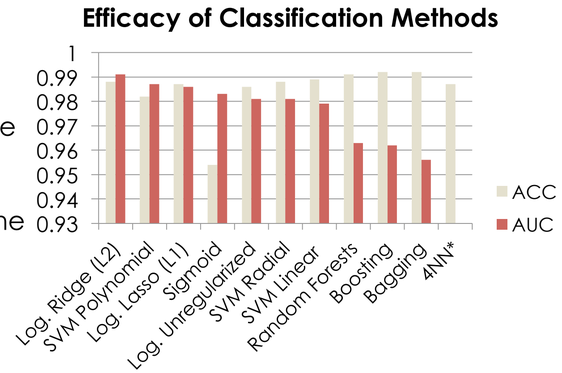
Logistic Regression on 63 discrete Fourier transform samples

Model	ACC	AUC
Unregularized	**0.955**	**0.967**
Lasso (L1)	**0.953**	**0.963**
Ridge (L2)	**0.959**	**0.962**

Figure 4: We applied a discrete Fourier transform, sampling 63 times between 0 and π. This alteration in feature space resulted in a significant improvement.

Final Model

We used a total of 72 predictors in our squiggle vs nonsquiggle classifier. Boosting and L2-regularized logistic regression resulted in the highest accuracy and AUC, respectively at >99% each.



Source/Model	Feature	Logistic (Lasso)	Logistic (Unreg. + Ridge)	SVM	Tree-Based
Image	**Signal Width**	•	•		•
Discretization Algorithm	**Loss**	•	•		
Linear Model	$\hat{\sigma}^2$	•	•	•	•
White Noise Process	$\hat{\sigma}^2$				
ARIMA (1, 1, 1) Process	$\hat{\mu}$				
	$\hat{\sigma}^2$	•	•	•	•
	$\hat{\phi}$		•	•	
	$\hat{\theta}$		•	•	
Long Memory Process	$\hat{H}$				
FFT	$X(W_{128}^n), n = 1,\ldots,63$		•		

Figure 5: The two most significant predictors across all models are highlighted above. A • denotes that a feature was deemed significant by a particular classifier.

Clustering

Results are plotted using the first two linear discriminant functions

Hierarchical Ward D2, Manhattan

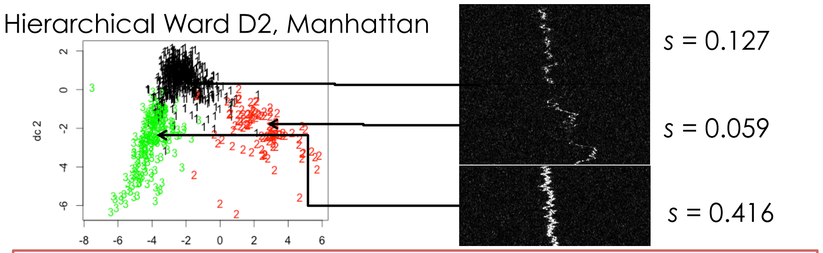


Figure 7: We can observe three clusters composed of 1. faint, quickly modulating signals, and 2. faint, slowly-modulating signals, and 3. strong low-bandwidth signals.

K-Means, Euclidean

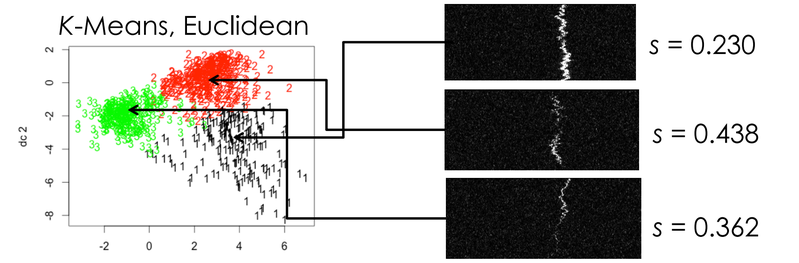


Figure 8: The clusters found using k-means correspond roughly to those we found using Ward D2 clustering and Manhattan distance.

Hierarchical Ward D2, Euclidean

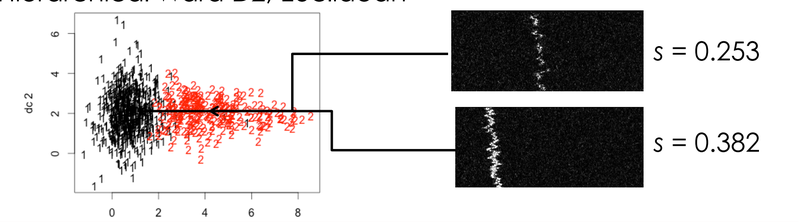


Figure 9: Here we observe two distinct clusters composed of 1. low-bandwidth signals and 2. high-bandwidth slowly modulating signals.

Insights

We performed chi-square tests to determine relationships between temporal characteristics and squiggle subgroups derived from Hierarchical Ward D2, Euclidean clustering. We found several strong correlations shown at right.

Characteristic	p-value
August	**5.75E-03**
4 AM - 8 AM	**2.31E-06**
8 AM - 12 PM	**5.79E-06**
12 PM - 4 PM	**7.32E-16**
L-Polarization	**1.24E-03**
R-Polarization	**3.14E-05**

Figure 10: In particular, we note that squiggles from the red cluster tend to occur in a 4-hour timespan.