Project SETI: Identifying the Unknown “Squiggle”

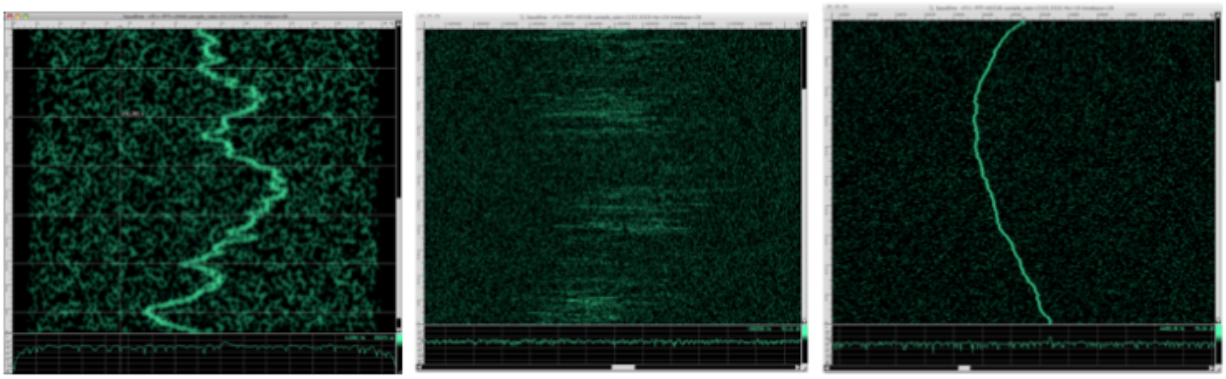
Jason Wang (jwang198), Kenny Smith (smithken), Travis Chen (travis14)

Problem Statement

The SETI Institute operates the Allen Telescope Array (ATA) to observe star systems for radio signals which may provide evidence of extraterrestrial intelligence. Signal events are recorded into a database; those events that maintain a power across all frequencies above the set threshold, are converted into spectrogram images through a fast fourier transform.

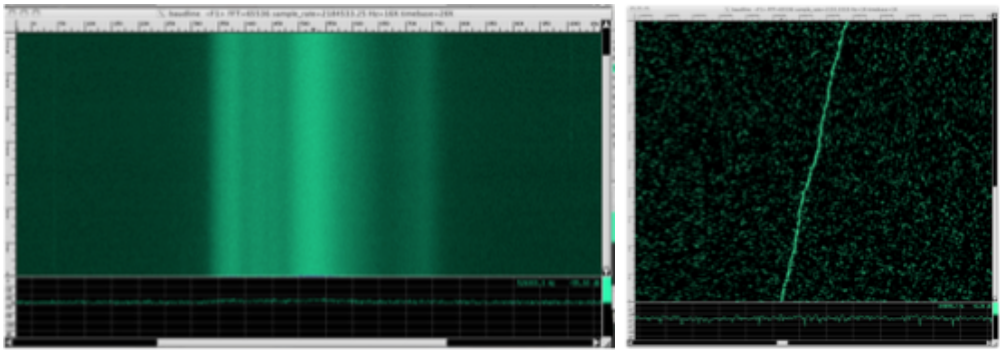
The key issue at hand is being able to pinpoint significant patterns in the signal stream that are not associated with known interferences. Using spectrogram waterfall plots curated from the ATA SETI dataset, we hope to build a classifier for an unknown subset of signals, inelegantly referred to as “squiggles.”

“Classic Squiggle” “Rapid Modulation Squiggle” “Slow Modulation Squiggle”



SETI has created classifiers for two known interferences: pulsars and linear, carrier-wave signals, but has yet to expand its toolbox.

“Pulsars” “Narrowband Signals” (with Doppler Drift)



The quest to identify “squiggles” is an open problem posed by SETI:

setiquest.org/wiki/index.php/Enhancement_of_Algorithm_to_Detect_Pulse_signals

Key Questions

We will address two key questions over the course of the quarter. We will begin with question 1 and move to question 2, time permitting.

Question 1: How can we use image processing to convert spectrogram waterfall plots into discrete time series data?

Question 2: How can we build a classifier to identify new "squiggle" signals from incoming spectrogram images?

Dataset

We will be using a dataset of spectrogram waterfall plots provided to us by the SETI Institute.

The dataset includes waterfall plots (spectrogram images) in PNG format:

- A library of >380,000 spectrogram images
- Within the library, 833 hand-identified squiggle examples

We hope to make our project useful to the larger Spark@SETI community by contributing and documenting code/findings that can be re-hashed to classify other unknown categories of signals, beyond "squiggles."

CS 229 vs. CS 341

We will be working with the same dataset and overarching project in our CS 341 (Project in Mining Massive Data Sets) class.

CS 229	CS 341
Subset of project focused on: ● Image processing: build algorithms to convert spectrogram images into discrete time series data ● Classification of "squiggle" v.s. non-"squiggle" ● Clustering among "squiggles" to identify sub-groups	Extension of project focused on: ● Build a generative model to simulate new "squiggles" ● Identify key characteristics of "squiggles" in relation to known interference types such as pulsars and linear, carrier waves

Action Plan

1) Transform spectrogram images into discrete time series datapoints
    a) Consider all 833 squiggles and ~2000 non-squiggles
    b) Dynamic Programming Algorithm:
        i) Take 1-pixel horizontal slices of the image, each slice representing a unit of time

ii) Select a set of points in the first time slice as candidates based on highest intensity
iii) Identify next time series candidate points by minimizing the loss function
(1) α Intensity + (1 - α)Deviation where $\alpha < 1$, intensity = brightness of the pixel, deviation = squared distance from previous time point/pixel
iv) Find the final time series graph denoted by path of candidate points that minimizes the loss function

2) Conduct a fast fourier transform to distill time series data in the time domain into the frequency domain
3) Extract features from the fourier-transformed and non-fourier-transformed data (e.g. using powers of each frequency, rate of modulation, etc.)
4) Classification
    a) Set aside ~200 “squiggles” for the test set
    b) Fit a classifier using logistic regression, linear discriminant analysis, etc. to separate “squiggle” v.s. Non-”squiggle”
    c) Select the model that minimizes the misclassification rate on the test set
5) Clustering
    a) Using the 833 “squiggle”, use K-means, KNN, self-organizing maps, etc. to identify characteristic groups within “squiggles” (e.g. rapid/slow/regular modulation)