Ryan Christopher
CS677

Final Project Proposal

The data science problem I am trying to solve-

Bayesian or Brahmsian?
Using Data Science to Compose Music

How can you predict the composition of music? In 2019, a partnership of Google's Magenta and PAIR (People+AI Research) celebrated Bach's 334th birthday with "the first AI-powered Google Doodle." By incorporating machine learning, orderless modeling, and Gibbs sampling (to name some of their methods), the group was able to create an interactive application that generated four part harmonies given a user's input of a one line melody.

Fast forward four months, and Magenta releases the Bach Doodle Dataset: 21.6 million harmonizations and 6 years of music submitted from the publicly available Google Doodle. Using the Bach Doodle Dataset, I plan to implement a smaller scale version of the Bach Doodle process where a user can create a melody as input and have a generated output be a multi part harmony that is modeled from the Bach Doodle Dataset.

Link to the dataset-
https://magenta.tensorflow.org/datasets/bach-doodle#download

Algorithms I plan on using-
Due to the size of the dataset, the algorithms I plan on using will be largely determined by how much or how little of the dataset I end up using. The dataset is 6.1G to 7.5G based on the format, however it is also available in a sharded format of 192 shards.

This cheat sheet from scikit-learn points me towards a Stochastic Gradient Descent Regressor if the dataset is larger than 100k samples, or the Lasso and Elastic-Net linear models if I use less than 100k samples of the dataset.

Because I want to predict values where the "values" correspond to note names in relation to the other values that signify the harmonic voices, these three models could serve as viable means to predict the output harmonies from the user's input.