Music Midi-Data Project Proposal

1. **Project Members:** Max Brown, Timothy Brown, John Ebeling, Joseph Henderson, Jake Murphy**.**

## **What the project will be:**

Music Midi data is a format of musical data in which the notes and timings of a song are recorded as data points. For an adequate project, we will attempt to use various methods to first classify a midi sequence into various categories. Our ambitious goal is to do forecasting given a sequence of midi data. More specifics are addressed in section 3 and 5.

1. **Questions you will try to answer:**

Fundamentally we seek to apply our skills to answer the following questions:

1. Given a set of MIDI data, can we predict which time period this music was composed in with 90% accuracy?
2. Given a set of MIDI data, can we predict who composed the music with 80% accuracy?
3. Given a set of MIDI data, can we generate more MIDI data to mimic a specific composer’s style?
4. Given a set of MIDI data, can we generate more MIDI data to blend the styles of two composer’s?
5. **Data sets you will use to try to answer the questions:**

Since our project will be using primarily MIDI files we are trying to gather a large amount of them online for training purposes. Right now, we will for sure be using the classical music MIDI collection found at <https://www.kaggle.com/datasets/vincentloos/classical-music-midi-as-csv>.. In addition there is a plethora of free classical music MIDI downloads online that we can pull from if we do not have sufficient data after this dataset.

Furthermore, we hope to not only use classical data for this project but also other genres such as Jazz and Pop. Possible Jazz files are found here <https://www.midiworld.com/jazz.htm>. The recency of Pop music may make it difficult to find free MIDI’s but it’s possible to scrape or otherwise acquire them, however, we will need to take care to make sure that this is done in a legal AND ethical way.

1. **Techniques you think you will use:**

We will model the songs as Markov chains. The first two questions that we will try to answer deal with classification. We plan to experiment with XGBoost and Random Forests, as well as other machine learning techniques, as classifiers using k previous time steps from our Markov Chain. We also have discussed using clustering algorithms to classify the data.

The last two questions that we have deal with predicting more data. We’ve discussed using an HMM or ARIMA to model the data. A support vector machine may be an idea for the method outside of the ACME curriculum.

1. **Metrics you expect to use to decide how good your answers are:**

For the classification problems, questions one and two listed above, we will use the metrics detailed in Section 4.10 as well as the F1 score, which is a combination of the precision and recall.

For questions three and four, we feel the techniques used for classification techniques will be insufficient because there is not a correct answer to mimicking musical style. The ultimate metric to evaluate success for questions three and four will be whether the music simply sounds good. We will strive to quantify this using music theory principles including rewarding harmonies, perfect fifths, etc.

Similar classification metrics could also be used for evaluating note prediction, although these metrics will be adjusted to account for a variety of factors. These factors include rewarding correct prediction of rhythm as well as note value, awarding correct note prediction if it is incorrect an octave, and other problems specific to music composition.

1. **How you will divide the work among the different team members:**

Now that we have found the data we have separated some of the initial work to process and visualize the data.

* Timothy and Jake will build the midi translator to easily analyze the data in Python
* Joseph and John will determine the best ways to store the data along with some of the classification classes
* Max will develop Midi plotting functions using matplotlib

As we are not completely sure how the techniques work, we plan on dividing the work in a similar fashion to our volume 4 project. We will all individually attempt to implement a machine learning method to the data. Depending on the exertion for each method, we may divide the various methods we need to try into partners and subgroups. For example, 2 members implementing a HMM with the other 3 working on clustering. We all plan to participate in the final write up.