Project Proposal
Music Genre Classification System
Goal: Classify the genre of a new, previously unseen piece of music.

Choice of datasets
http://millionsongdataset.com/
The Million Song Dataset is a collection of complementary datasets spanning various audio features (lyrics, user data, genre labels, etc.). It contains over 200,000 songs labelled with their genre, stratified and non-stratified sets with various ratios of labelled testing and training entries. The genres come from multiple sources and allow ambiguity with an optional second genre.

Methodology
First, we need to extract audio features, such as tempo, rhythm, melody, harmony, instrumentation, and timbre. Then, we will use machine learning algorithms to train the classifier on a labeled dataset of music.

Data preprocessing: Some other datasets we found, although a lot smaller, provided extracted data such as frequency spectrograms that would greatly reduce the data processing time. It is however possible to use fourier transforms fairly easily to perform this task.

Machine learning model: KNN or CNN; the goal being to reinforce the weight of edges between neurons that fire with the different patterns that map to specific genres. We want to predict genres with significant accuracy (70% seems to be doable after examining similar projects). Reducing the number of mappable genres to more general ones will also be a technique to improve accuracy.

Evaluation metric: Confusion matrix and accuracy/precision-recall/logistic loss

Application
We will be using a webapp where the input should be an audio file of a still unspecified format uploaded directly by the user (WAV seems to be the best bet for now). Linking sites that convert audio files would also make the app more user friendly. Ideally, our algorithm would predict one genre, but a more forgiving approach would be to list a few genres that have the highest probability of being correct (think of a top 5 style of listing). The genres would be outputted in a visually appealing probability chart.

Others
Ways to visualize audio data
https://www.kaggle.com/code/satoru90/music-genre-classification-xgb-deep-learning
Challenges: similarities between different genres, different approaches to music genre classification (content-based methods - audio features, collaborative filtering methods - users' listening behaviour, hybrid method - combination of both). If feasible, we could try the hybrid method.