**The presentation will be a MAXIMUM of 7 MINUTES and should give a dedicated segment/demonstration highlighting how the model performs on new data.**

**YOU MUST STICK TO YOUR ALLOTTED TIME SO THAT WE STAY WITHIN THE TIME LIMIT.**

**Sections:**

INTRO/PROBLEM (1 minute, 30 seconds): *What is the problem that you are trying to solve? Why is it important?*

**Slide 1:** Hello and welcome to the presentation of our team’s culminating project on building a neural network to automatically classify music by genre.

**Slide 2:** During today's presentation we will be taking you through first a brief introduction on music genre recognition and our project process, explain our data selection and processing methods, and outline the different models the team generated. Then we will provide a demonstration of how our final model works and explore the performance results and implications.

**Slide 3:** To start, why music genre recognition?

**Slide 4:** Given the increasing use of music streaming services, there is an ever growing number of songs being uploaded to the internet. This means that companies like Spotify and Apple need to process and sort through an even greater amount of data. Automated music genre classification could not only aid in data organization but allow for improved song recommendation to users. Machine learning models are adept at handling massive amounts of data, which make the tool ideal to use in this situation. Our project aims to develop a model that can carry out music genre recognition.

**Slide 5:** Our project was carried out in several stages: First we created a project plan and submitted a proposal to the APS360 teaching team. Then we collected and processed our data. Following the completion of data collection we created two baseline models and a CNN primary model. After the CNN model was created we decided to take steps to improve the model performance by modifying the model to a ResNet.

DATA (45 seconds): *What data did you use or collect? What does it look like? Is the amount/type of data sufficient for your project? Was there anything tricky or weird you needed to do with the data? … with the train/test/validation split?*

- Music classification

- Using raw audio data can be very large and may contain irrelevant information that can negatively affect the performance of the model.

- this sounds uncomfortable

- Preprocessing is required!

- Find an accessible dataset

- GTZAN: 1000 samples, 100 songs in 10 genres

- Take this from an audio classification problem that we are not comfortable with to a very familiar domain

- Image classification!

DATA PROCESSING (45 seconds): *What processing did you perform on your data? Why?*

- Preprocessing Roadmap

- Find an accessible dataset

- GTZAN: 1000 samples, 100 songs in 10 genres

- All in .wav file format

- Going from .wav file format to a waveform plot

- Going from Waveform to Spectrogram

- simply applying a short time fourier transform to the waveform plot

- Going from spectrogram to a mel spectrogram

- Going from mel spectrogram to log mel spectrogram

- Using GTZAN in the beginning is fine. but…

- We found that the dataset was too small and was affexting the performance of our model

- We abandoned GTZAN and went to a very famous place for music…

- Spotify

- NEW Roadmap

- Wrote a script to download 10k songs samples from spotify

- 1k per genre

- all in .mp3 file format

- write a function to convert to .wav format

- Going from .wav file format to a waveform plot

- Going from Waveform to Spectrogram

- simply applying a short time fourier transform to the waveform plot

- Going from spectrogram to a mel spectrogram

- Going from mel spectrogram to log mel spectrogram

MODELS (1 minute, 30 seconds): *What model(s) did you build? What kind of models (CNN, Transformer, GNN, etc.) are they?*

*BASELINES (30s)*

**Slide 10:** For this project we developed two baseline models to test against our primary model. These are K nearest neighbor (KNN) and artificial neural network (ANN).

The KNN model works by taking the spectrographs in vector format and classifying new data based on how close its features are to pre-classified songs. We used the training data set to populate the field, and took the closest 25 training songs to predict the class of the new songs.

In the ANN model, input data is put through hidden layers containing activation functions which transform the data. This transformed data can be used to generate a prediction for the song genre.

*FINAL PRIMARY MODELS (1 min)*

*Slide 11*

The team has implemented CNN, AlexNet, and moved forward with ResNet as the primary model. It learns low, mid, and high levels features and consists convolutional layers with skip connection. It allows ResNet to continue learning new features as well recalled what it has learned before. This also address the issue of vanishing gradient by implement multiple layers in CNN.

# as it achieved the state-of-art performance on image classification

#like edges, shapes, and objects in the image.

#such as in a particular layer, it takes input from the one layer before as well takes input from a couple layer ago.

# Its shortcut allow the gradient to be directly propagated to the earlier layer which address the issue of vanishing gradient by implement multiple layers in CNN.

#Overall, ResNet computes with higher accuracy and less error.

Slide 12

Our Reznet is based on the 18 layers and contains 4 convolutional and 2 fully connected layers. The team tested with the 101 layers which theoretically provided more depth, however the team found both have similar efficiency with our model. Therefore, the team decided to move forward with 18 layers with faster computational time. The table summarized the type of kernel, the output channel, and overall structure.

Slide 13

To overcome overfitting, other customization the team has implemented include the dropout rate, batch normalization, cross entropy loss, adam optimizer for training, epochs, learning rate, and batch size customization.

Slide 14

The team conducts normalization for data augmentation and make the pictures more vibrant.

DEMONSTRATION (45 seconds): *You will need to show a demonstration of your project evaluated on new, never before seen data. Try to be creative and engaging with the demonstration.*

Here we would like to do a demonstrate from the prof Justin Beland, favourite song, the creature by pop smoke. Mentioning by Alex, the audio will be translated to a spectrogram image and feed into the ResNet layer. From our model, The top three genres predicted by the model were Rock, pop, and blues. The prediction probabilities are 46, 25, and 14.6 respectively

The while the song is hiphop/rap, the model found it have a 5% likelihood of being such

* Now we would like to provide a demonstration of the capabilities of our model
* We asked Professor Justin Beland what his favourite song was and he said Creature by Pop Smoke \*play short audio clip\*
* We took this audio and transformed it into a spectrogram, then put it into our model
* The model predicted hip hop with X% probability

QUANTITATIVE RESULTS (45 seconds): *How are you measuring and comparing performance? How well did your model perform?*

To measure and compare performance, the team computes loss and accuracy to evaluate the model performance.

We tested our models using a larger set of data to have a fuller picture of its capabilities

The test dataset contains 1000 images, with 100 spectrograms in each genre.

Using the test set made during data processing, we found that the accuracy of our ResNet model surpasses that of the baselines we developed by 20-30%.

The highest accuracy obtained with the testing set was X%.

QUALITATIVE RESULTS (45 seconds): *What are some sample predictions generated by your model? Does it make sense? Did you “cherry-pick” the best results or are you showing a representative sample?*

Now we will explore the qualitative results we obtained, and the reason why we believe our model performs the way it does.

When observing the performance of the model by genre we can see that the genres with the greatest accuracy are X, and Y. In contrast, the genres with the lowest accuracy are A, and B.

Further, we collected the loss and accuracy of the module during each epoch of training. From the graphs, ResNet has an accuracy of almost 100 % for training, 70.5% for validation and a decreasing loss for training set with the increase of epochs. However, at X epochs, the loss and accuracy for the training and validation sets begin to diverge.

The divergence of training and validation accuracy and loss indicates that the model is overfitting to the training data.

Additionally, the greater accuracy for genres X and Y indicates that they have more distinct features that allow them to be distinguished the easiest.

These issues could potentially be solved through obtaining more data.

Thanks Slide

Thank you for watching and we hope you enjoyed this presentation!

## only add if we have extra time

TAKEAWAYS (1 minute): *What did you learn about machine learning from doing this project?*