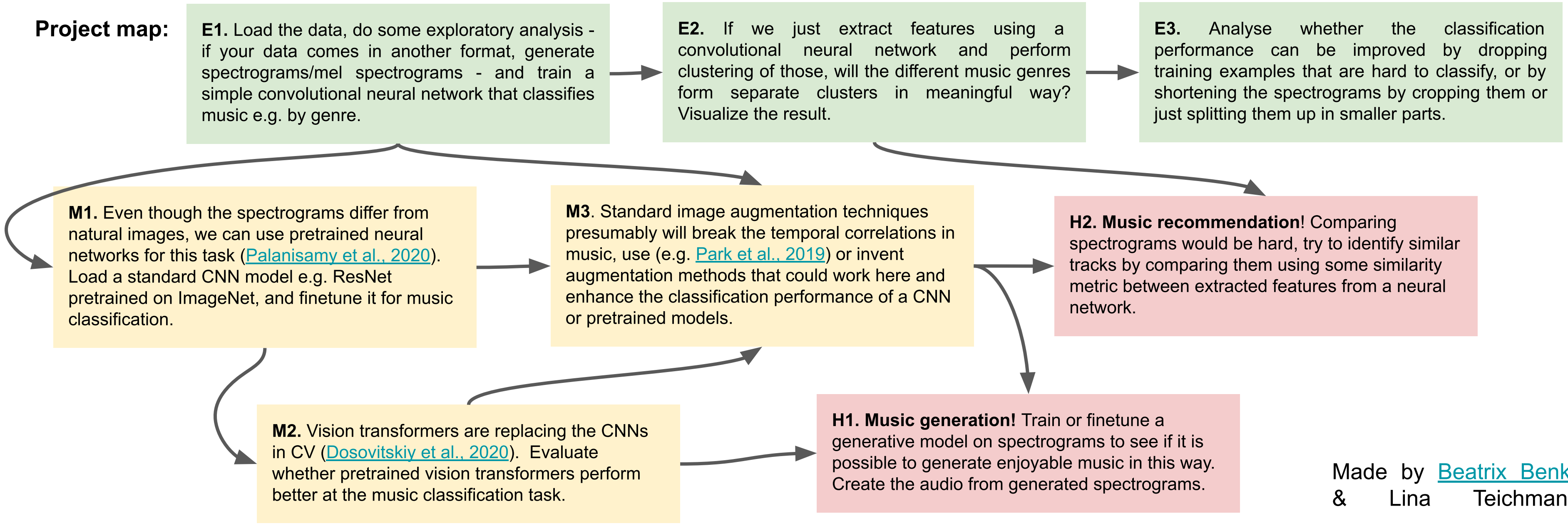
Music classification & generation with spectrograms

Background: Sound can have different digital representations: usually it is stored by encoding the shape of the waveform as it changes over time, but for analysis we often make use of visually inspectable spectrograms - obtained from a waveform by computing the Fourier transform of overlapping windows of the audio signal, and stacking the results into a 2D array. These spectrograms exhibit a lot of structure, and modelling it can enable sound - in this case music - classification, generation, even recommendation e.g. speech recognition or music recommendation.

Project setup: We provide a notebook with loading the GTZAN dataset (containing the sound files and genre labels) and the spectrograms, and training a simple CNN on them. One can use other datasets as well (e.g. FMA-small dataset).

Project map:



E1. Load the data, do some exploratory analysis - if your data comes in another format, generate spectrograms/mel spectrograms - and train a simple convolutional neural network that classifies music e.g. by genre.

E2. If we just extract features using a convolutional neural network and perform clustering of those, will the different music genres form separate clusters in meaningful way? Visualize the result.

E3. Analyse whether the classification performance can be improved by dropping training examples that are hard to classify, or by shortening the spectrograms by cropping them or just splitting them up in smaller parts.

M1. Even though the spectrograms differ from natural images, we can use pretrained neural networks for this task (Palanisamy et al., 2020). Load a standard CNN model e.g. ResNet pretrained on ImageNet, and finetune it for music classification.

M3. Standard image augmentation techniques presumably will break the temporal correlations in music, use (e.g. Park et al., 2019) or invent augmentation methods that could work here and enhance the classification performance of a CNN or pretrained models.

H2. Music recommendation! Comparing spectrograms would be hard, try to identify similar tracks by comparing them using some similarity metric between extracted features from a neural network.

M2. Vision transformers are replacing the CNNs in CV (Dosovitskiy et al., 2020). Evaluate whether pretrained vision transformers perform better at the music classification task.

H1. Music generation! Train or finetune a generative model on spectrograms to see if it is possible to generate enjoyable music in this way. Create the audio from generated spectrograms.

Made by Beatrix Benkő & Lina Teichmann