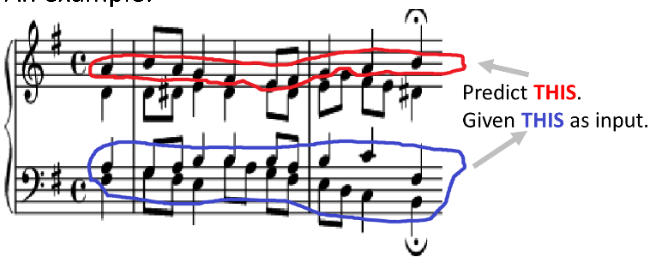
RAMT: Raw Audio Music Transformer

1. Update your project objective
   - Generate a harmony for a specific melody. The dataset is a collection of raw audio from various pieces of Bach's work. The song excerpts are split into the different voices present. By feeding a subset of these voices to our model, the aim is to predict a final voice which is in the same key and in good harmony with the input melody.
   - An example:



Predict **THIS**.
Given **THIS** as input.

2. Outline the methods you will be using
   - Using the Transformer to predict appropriate audio sequences. Our work will consider both MIDI (symbolic representation of the notes) and raw audio sequences. Since there are also multiple different sequences that could be correct, we will also investigate methods to make our models stochastic in its predictions. A motivation behind using the transformer is that they have typically less computational complexity than recurrent models, but are still able to catch global dependencies between input and output, which should hopefully make these models better suited for real-time audio systems. Finally, we also hope that we can take advantage of the attention networks in the transformer model to gain some insights into how our model behaves and what parts of the inputs it learns to focus on.
3. Outline relevant related work
   - The importance of this work is to investigate how transformer models behave on raw audio data. So far, recurrent [1] and convolutional [2] models have been predominant within the domain of raw audio data. Transformer models have gained traction within the field of music and other sequence modelling tasks. However, work such as the Music Transformer [3] and Pop Music Transformer [4], have used high-level representations of music sequences, such as MIDI, while the use of transformers on raw audio data is yet to be thoroughly explored.
     1. Wright, Alec, Eero-Pekka Damskägg, and Vesa Välimäki. "Real-time black-box modelling with recurrent neural networks." *22nd international conference on digital audio effects (DAFx-19)*. 2019.
     2. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).
     3. Huang, Cheng-Zhi Anna, et al. "Music transformer." *arXiv preprint arXiv:1809.04281* (2018).
     4. Huang, Yu-Siang, and Yi-Hsuan Yang. "Pop Music Transformer: Beat-based modeling and generation of expressive Pop piano compositions." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
4. Write at least one question/complication you would like to discuss with your advisor
   - There is not a standardised loss function for this particular application due subject preferences. Even though we have some ideas on possible approaches, it would nice to discuss this area further when we start implementing our models.
5. Write down what you expect to deliver in the final paper

- Demonstrate the efficacy of the transformer model and how it behaves when using raw audio data as input.
- Compare our models against the state-of-the-art, particularly focusing on WaveNet-style and recurrent architectures.
- Investigate various aspects of our model and how they affect both prediction performance and applicability to real-time systems.