Almost all approaches to classification problems look like the following steps:

1. Extract features from audio (Like MFCC features)
2. Put a Convolution Network/ RNN on top of the features over time
3. Train with a loss function which addresses the situation.

Recurrent Neural Network (RNN) encoder and a RNN decoder, which respectively transforms the variable-length audio sequence into a fixed-length vector and reconstructs the input sequence on the generated vector

RNN encoder to map an input sequence into a fixed-length vector, and a RNN decoder to reconstruct the input sequence from the generated vector into a sequence-to-sequence learning strategy.

he first step is to actually load the data into a machine understandable format. For this, we simply take values after every specific time steps. For example; in a 2 second audio file, we extract values at half a second. This is called **sampling of audio data,** and the rate at which it is sampled is called the **sampling rate.**