### CSE447 Proj Checkpt

### Team:

Roy Zhao, Jenny Rong, Lingji Kong

### Dataset:

We decided to use datasets that contain all kinds of languages to train our model, since we do not know which language the astronaut speaks. We will start with an initial dataset on kaggle: <https://www.kaggle.com/basilb2s/language-detection>. This dataset is a small language detection dataset, which consists of sentences in 17 languages (though this is clearly not enough for our model to predict the (i+1)-th unicode).

### Method:

We will use python as our programming language.

We will be training character-level neural network language models on each of the languages. We will first determine which of the languages the astronaut is saying. At the same time, we will also incorporate the previous words as hints to indicate which word is the astronaut trying to say.

We will be processing the data by converting strings to vectors and tokenizing the data. We will build and train with recurrent neural networks (RNN). We will be using Pytorch to build the model. The model will take the input embedding and make predictions based on the log-likelihood of the next character. Then we can convert the embedding to the next character.