Marco Lipari, Andrew Tomajian, James Wnek, Roy Elia

Choice of dataset:
https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13

We choose this dataset since it has around 20 years of Reddit comments. It has more than enough data for our purposes, although finding ways to process it all and sampling it accurately will be challenges for us to overcome.

Methodology:

Training models off of text data requires a lot of preprocessing, but there are already many widely used techniques we can draw from. Text cleaning (ex. removing links, punctuation), converting text to lowercase, removing stop words (like “a”, “the”, words which carry no meaning) and tagging parts of speech are some methods that we can use. Other techniques like stemming and lemmatization (transforming words into more basic forms by removing affixes) could also help our model, but we may want to limit their use in order to conserve some of the subtleties of the changes in language over time.

https://github.com/sharadpatell/Text_preprocessing_steps_for_NLP

The most natural choice of model for this kind of problem is Natural Language Processing (NLP). There is a spectrum of different NLP models that we can draw off of, ranging from simple techniques like Bag of Words which only look at word frequencies to more complex ones which take into account semantics and word order. In our project we can explore both coding simple models from scratch and using libraries for more complex models like BERT or GPT.

In order for our model to output a year, we can use regression. The advantage of this is that it is easy to implement error metrics like Mean Squared Error which the model can report alongside the predicted year as a sort of confidence level. A specific algorithm that is commonly used in NLP is Random Forest Regression, which takes the average of the output of many decision trees to predict a continuous variable. This algorithm suits NLP because it can better predict non-linear relationships and it is relatively computationally light for the large amount of data that we have.

We could also use logistic regression to give a categorical output (like a specific year or range of years).

https://www.datacamp.com/blog/how-to-learn-nlp
https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4
https://www.ibm.com/think/topics/logistic-regression

It is difficult to say right away how much we are expecting from our model. Ideally, it will be able to predict within a range of 2 or 3 years a significant portion of the time. We can also do tests to see how well we ourselves can predict the age of the comments and use that as the target to beat.

Application:

We will use a web application where the input is a text box. Users can copy and paste the text of a reddit comment whose age they know and see what the model will predict, or write their own comment and see what time frame it places it in.