Data

- 52,152 rows of labeled pixel data
- for each given letter, the letter that comes after it, or -1 if none

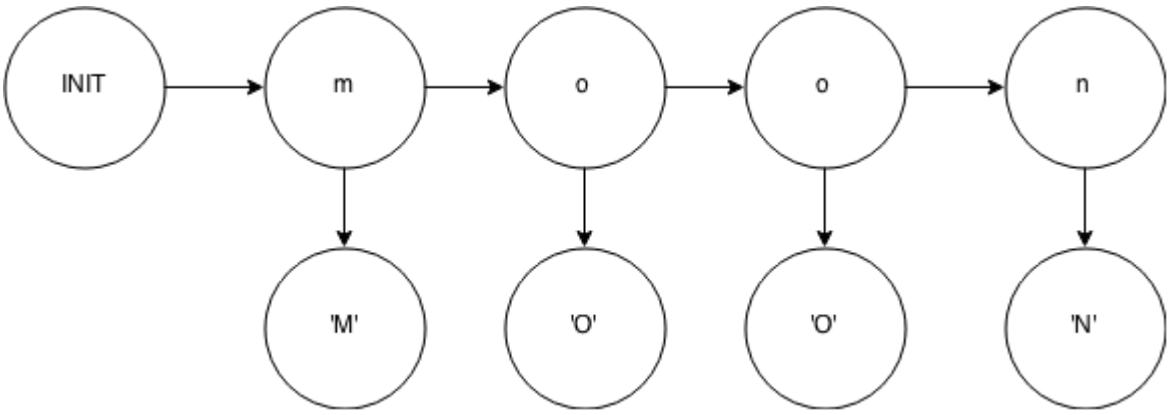
For our project, we see three levels of development: a baseline, and two extensions.

Baseline

We plan to implement two classifiers that we learned in class: Naive Bayes and Logistic Regression.

Minor Extension

Represent each word as a HMM, like below.



Each hidden state is a **letter**. Each observed state is a **pixel vector**. We need to find the following parameters:

- $\phi_{i,j}$: transition probability from state i to j
- $\theta_i(X)$: emission probability associated with state i of a pixel vector X

We plan to use MLE to find $\phi_{i,j}$. Let c_1, c_2 be consecutive letters. Then this is equivalent to finding $P(c_2|c_1)$. We can just count the frequencies of letters in our training data to obtain this.

To find $\theta_i(X)$, we find $P(X|c)$ where c is the letter that is associated with state i. Assume Naive Bayes (conditional independence) assumption on $P(X|c)$, so we have

$$P(X|c) = \prod_i P(X_i|c)$$

. This can be obtained from the training data by looking at all pixel vectors for letter c, and applying MLE to calculate $P(X_i = 0|c)$ and $P(X_i = 1|c)$ using the frequencies of the pixels that are on and off.

A word is $< X_1, ..., X_n >$ where n is the length of the word. To classify a word, feed the above parameters into the Viterbi algorithm, with the observed states as $X_1, ..., X_n$ to obtain the hidden states $c_1, ..., c_n$, which are the letters.

Question: Is this a valid extension? We are not using EM but some variation of Naive Bayes, MLE, and transition probabilities to improve the classifiers from the baseline further.

Greater Extension

Use Expectation Maximization to find out the transition and emission probabilities of the HMM above. We are not sure how EM works using HMM or where to start, but we will research this.

Question: How do we represent the observed pixel vector as proper states? Do we assume that the pixels take a multinomial distribution with 128 parameters?

Question: Is using "EM" equivalent to using the Baum-Welch algorithm?