What is language modeling?

- Assigns probabilities to sequences of tokens
  - Tokens could be words, characters; in the general case character strings

$$P(\texttt{pocket} \mid \texttt{whats in your})$$

- Why should we approach language understanding this way?
  - There are many applications for which this is useful, for example:
    - Machine translation

$$P(\texttt{pocket} \mid \texttt{from capital one. whats in your})$$
$$\textit{or}$$
$$P(\texttt{wallet} \mid \texttt{from capital one. whats in your})$$

    - Speech recognition

$$P(\texttt{yeast} \mid \texttt{i live in \_\_\_\_ village})$$
$$\textit{or}$$
$$P(\texttt{east} \mid \texttt{i live in \_\_\_\_ village})$$

  - Most importantly, we can train deep neural networks using language modeling procedures to learn features from large amounts of text.

The problem with modeling sequences

- For most interesting sequence problems (language, genes), learning the joint distribution of observed sequences is intractable.

$$P(sentence) = P(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$$

  - Consider a sequence of length $T = 10$ generated from a vocabulary containing only $N = 1000$ words, the number of possible sentences is 10^{30}. The large branching factor, N, makes estimating the probability of each possible outcome intractable.