

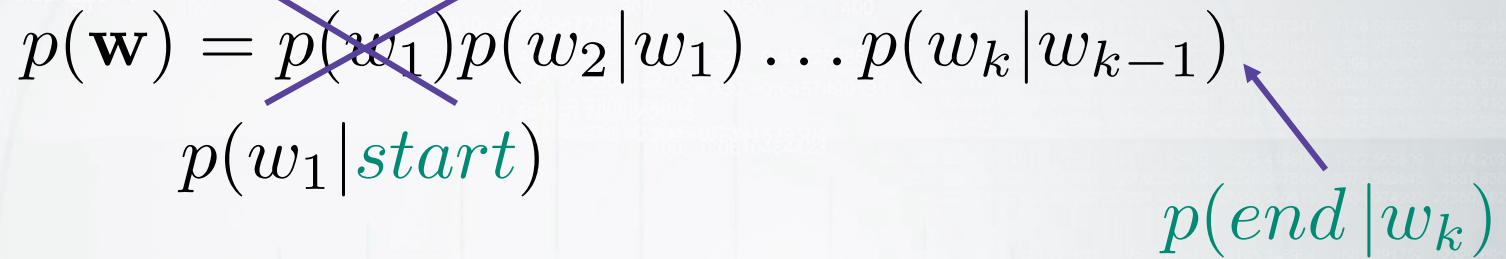
# Bigram language model

So that's what we get for  $n = 2$ :

$$p(\mathbf{w}) = \cancel{p(w_1)} p(w_2|w_1) \dots p(w_k|w_{k-1})$$

$p(w_1 | \textit{start})$

$p(\textit{end} | w_k)$



It's normalized separately for each sequence length!

$$p(\textit{this}) + p(\textit{that}) = 1.0$$

$$p(\textit{this this}) + p(\textit{this is}) + \dots + p(\textit{built built}) = 1.0$$

...