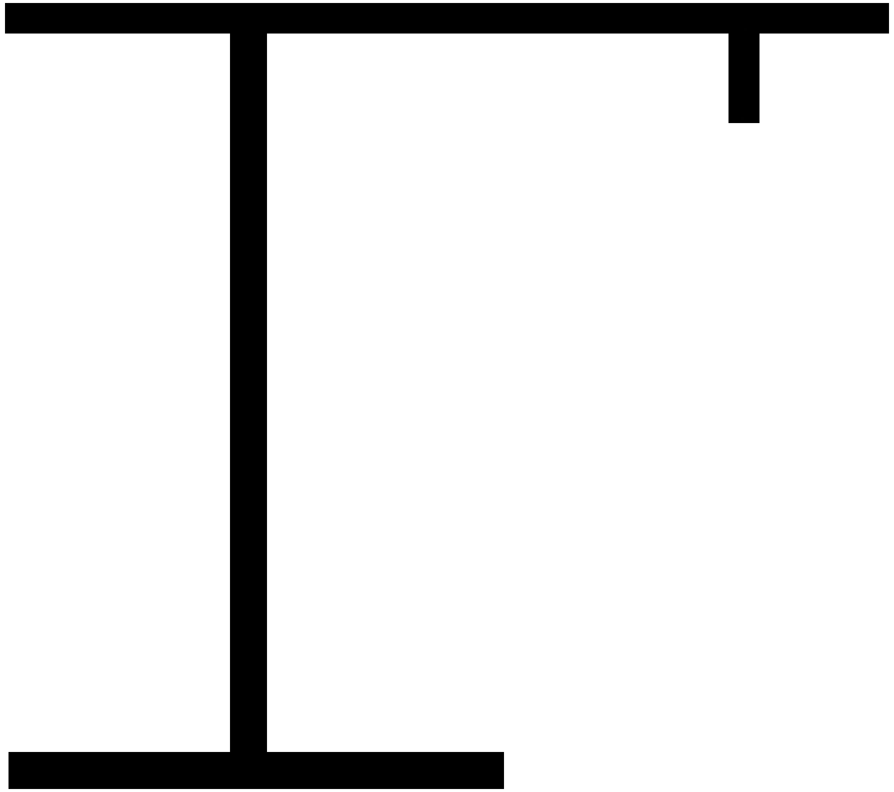
Language Models

How do we model language for machine learning?

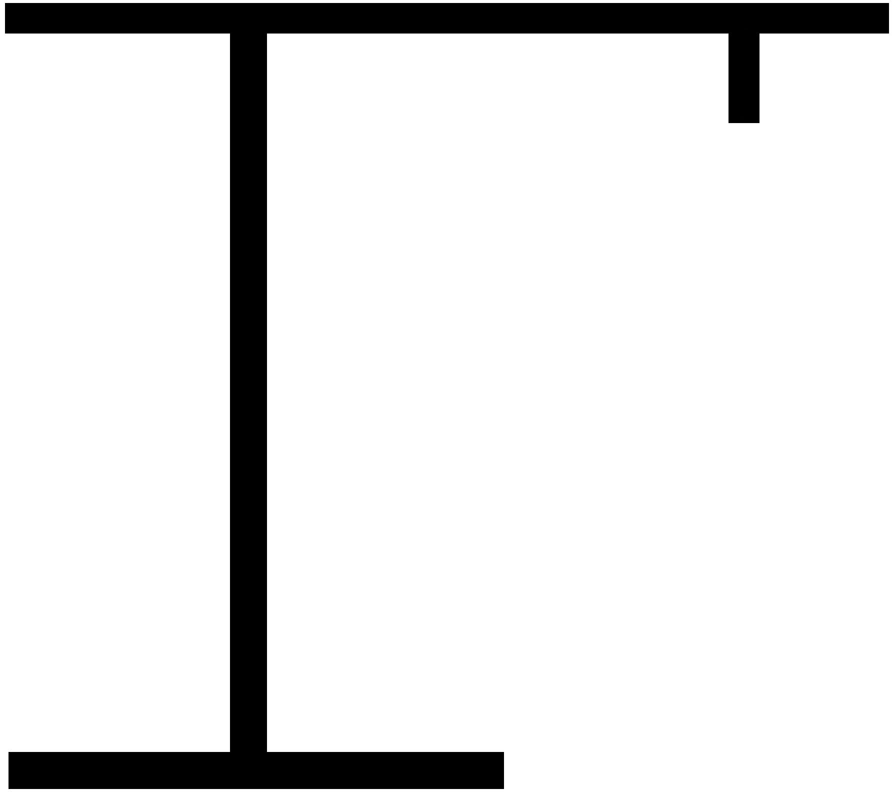
- Statistical relationships between words!

Example: Hangman



___ ___ ___ ___ ___

Example: Hangman



N G R A M

Statistical learning

Discuss:

How did you make decisions about which letter to guess when there were no other letters?

How did you make decisions about which letters to guess after that?

Statistical learning

- Decisions based on probabilities, either with context (some letters filled in) or without (the first letter)
- Remember n-grams?
    - Guessing a letter with no context is like a 0-gram
    - Guessing a letter based on the previous letter is a unigram
    - Based on the previous two is a bigram
    - Etc!
- Can be used to predicting next letter or next word

How do we get statistics for letters and words?

Discuss!

How do we get statistics for letters?

- Analyze their frequency in actual text
- Large libraries like the Corpus of Contemporary American English aggregate lots of different types of text
- We can count the frequencies of different combinations of letters
- For any n-gram, we then have statistics about what letter most often follows it

N-gram statistics example

Given a trigram of letters:

B R U __

How would we figure out the statistical probability of the next letter?

N-gram statistics example

Given a trigram of letters:

B R U __

Figure out the frequency of every possible following letter:
(From the Corpus of Contemporary American English)

BRUS: 9806
BRUT: 5193

N-gram statistics example

BRUS: 9806

BRUT: 5193

TOTAL: 14999

S: 9806 / 14999 = 0.654

T: 5193 / 14999 = 0.346

N-gram statistics for sentences

- Works the same way for sentences!
- How would you guess the probabilities compare for

“On my way” vs “On my squash”

- Many more options (26 letters vs 100,000+ words)
- Used for both!
  - Autocomplete for words and autocomplete for sentences