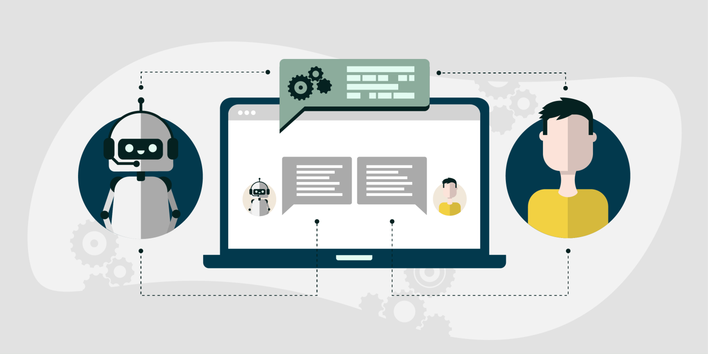
# Week 21 - Natural Language Processing

# What is Natural Language Processing?



# What is Natural Language Processing?

- Extracting meaningful information from natural language text
- Examples:
  - Sentiment Analysis
  - Chat Bots
  - Automatic Translation
  - Speech Recognition

# Steps in Natural Language Processing

1. Data Cleaning (remove punctuation, capitalization etc)
2. Tokenization (separating each word into its own entity)
3. Removing 'stop words' (such as and, or, like, ...)
4. Lemmatization (replacing words by their roots - such as mapping 'gone', 'going', 'goes', 'went' all into 'go')
5. Stemming (removing common prefixes or suffixes)

# Bag of Words

- 'Bag Of Words': text is represented as a "bag" of words without paying attention keeping word order.

- Vectorization will generate vectors which indicate the presence of tokens in different text instances.

- This looks just like a set of dummy variables - from here on, we can apply models we already know!

# Natural Language Processing

Note: this class is using a very basic algorithm in text classification. In many cases, the positions of words in a sentence have additional meaning, and there are much more sophisticated algorithms to handle this.