Machine Learning

- At a high level, machine learning is set of techniques, or *machinery*, that is designed to automatically learn complex relationships between random variables. Typically there is some output that we are interested in predicting, and there are inputs that we hypothesize affect that output. The goal is to build a model that accurately predicts the output given the input.

- In the context of natural language there are several things that make it unique:
  - The **distribution of words is highly non uniform** in the wild and follows a power law with a very long tail [1].
  - Unlike images or tabular data, **language is highly compositional** and meaning is derived from how characters, words, and phrases are composed. There are countless ways of representing the same meaning using language; this makes it hard to automate the extraction of meaning from it.

[1] Zipf, 1949

Composition

Language is **compositional**: units such as words can combine to create phrases, which can combine by the very same principles to create larger phrases. For example, a **noun phrase** can be created by combining a smaller noun phrase with a **prepositional phrase**, as in *the whiteness of the whale*. The prepositional phrase is created by combining a preposition (in this case, *of*) with another noun phrase (*the whale*). In this way, it is possible to create arbitrarily long phrases, such as,

(1.1) ...huge globular pieces of the whale of the bigness of a human head.[2]

The meaning of such a phrase must be analyzed in accord with the underlying hierarchical structure. In this case, *huge globular pieces of the whale* acts as a single noun phrase, which is conjoined with the prepositional phrase *of the bigness of a human head*. The interpretation would be different if instead, *huge globular pieces* were conjoined with the prepositional phrase *of the whale of the bigness of a human head* — implying a disappointingly small whale. Even though text appears as a sequence, machine learning methods must account for its implicit recursive structure.

- Eisenstein, 2019, Chp 1