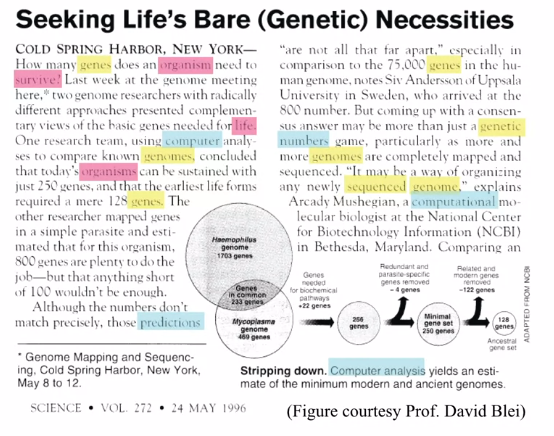
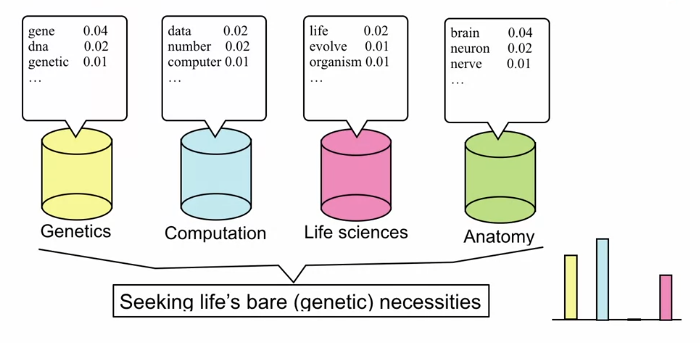
**Topic Modelling:**

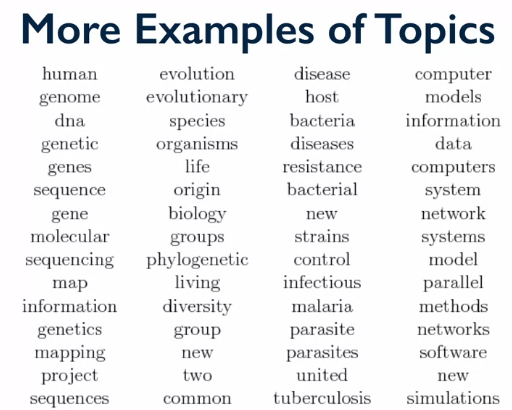
This is a high-level analysis that can be used to determine what a document is about. E.g. if you’re analysing a paper on genetics and computation then you might find that the words used in the paper come from sets of words what are commonly used in genetics, computation, life science and anatomy. E.g.:





Where the bar chart shows the total number of words or their weights seen in the particular document, form this you could tell what the paper is about!

The words in each topic (e.g. genetics) are assigned a probability that that word is in that set, this means that words like “team” could appear in multiple sets e.g. sports or science. There are many different sets of topics for these words to fall under.



The way this works is that you define the number of relevant topics you are interested in using within the document e.g. maybe 20 different topics. However, you don’t actually know the real topics, so the algorithm will return the topics it thinks are most relevant. You also don’t know the distribution of the topics in the document. This means that topic modelling is a **text clustering** problem, but you need to cluster both the documents and the words at the same time. There are many different algorithms that could be used, the most common one is **Probabilistic Latent Semantic Analysis (PLSA)** and **Latent Dirichlet Allocation (LDA).**