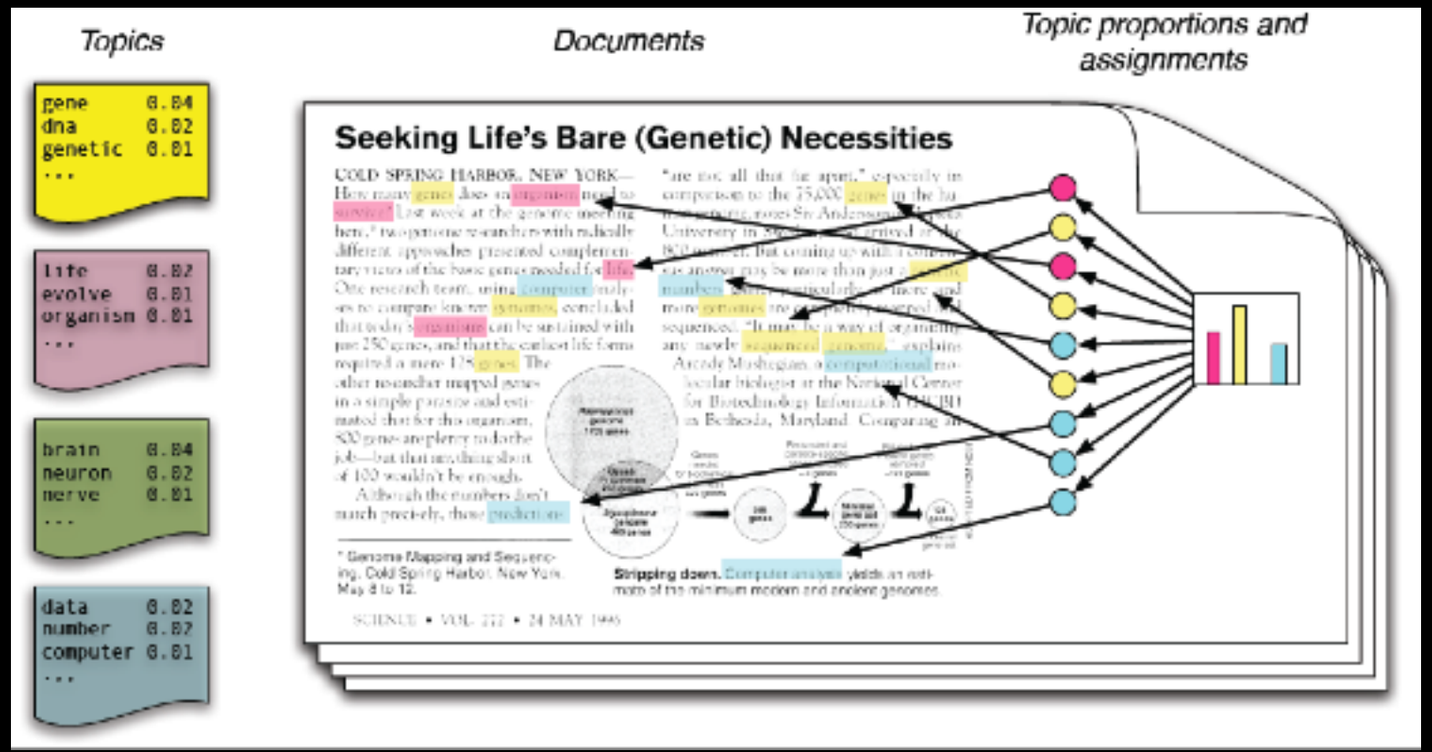
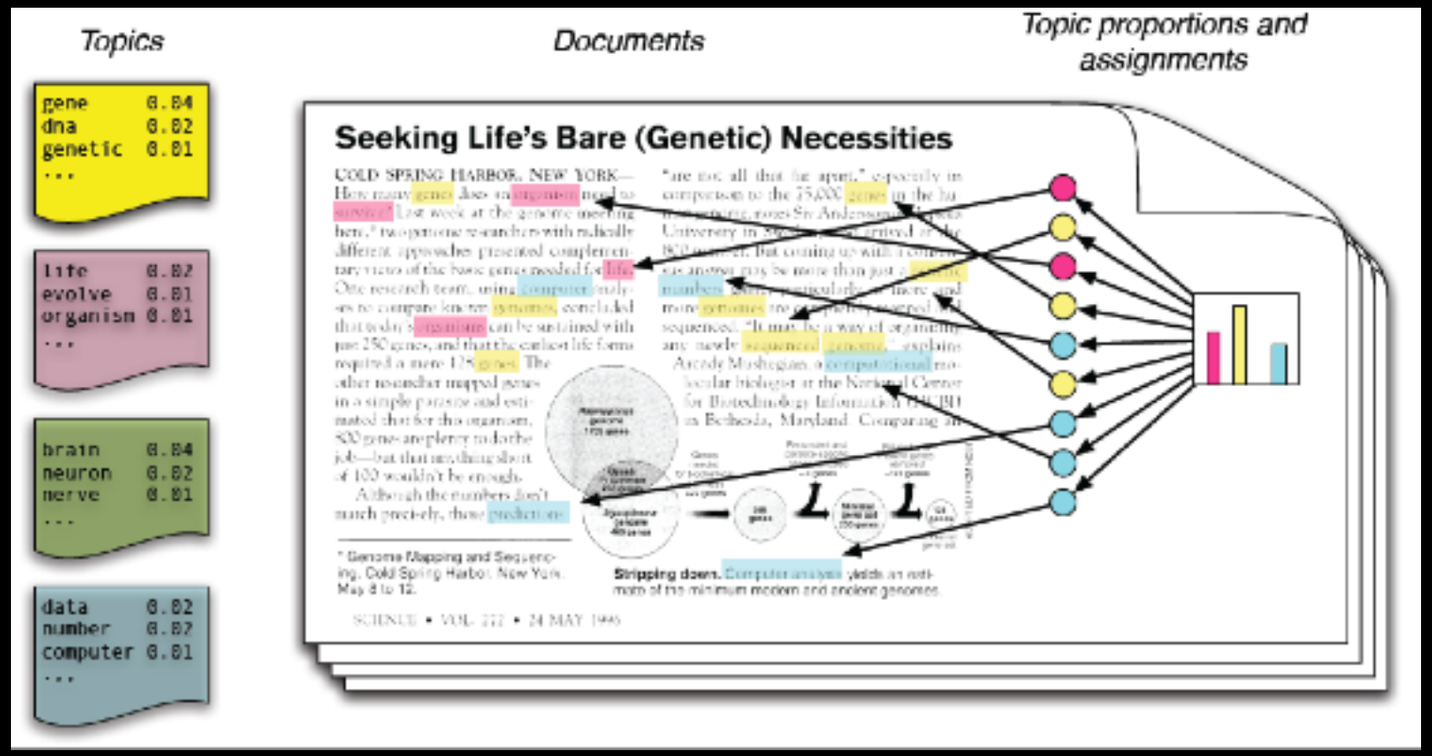


- **LDA** assumes that the way a document is written is that its author picked words (i.e., **n-grams**) from some set of topics that we don't know but want to determine.

- As such, we don't know what these topics are or even how many there are!

- So, how does the LDA determine these topics in the first place? The **LDA** simply reverse engineers this process.



- We first assume that **some K number of topics** exist across all of the documents in the corpus. There could be more topics or less. It's just a first guess!
- Next, we randomly guess how much each **document** is made of these **topics**.
- For instance, **document 1** could be made up of terms that are 30% associated to **Topic 1**, 40% associated to **Topic 2**, 10% associated to **Topic 3**, 0% associated to **Topic 4**, and 20% associated to **Topic 5**, such that they sum to 100%. Again, it's just a random guess based on nothing.
- This document-to-topic distribution is called **alpha**.

- This document-to-topic distribution is called **alpha**.