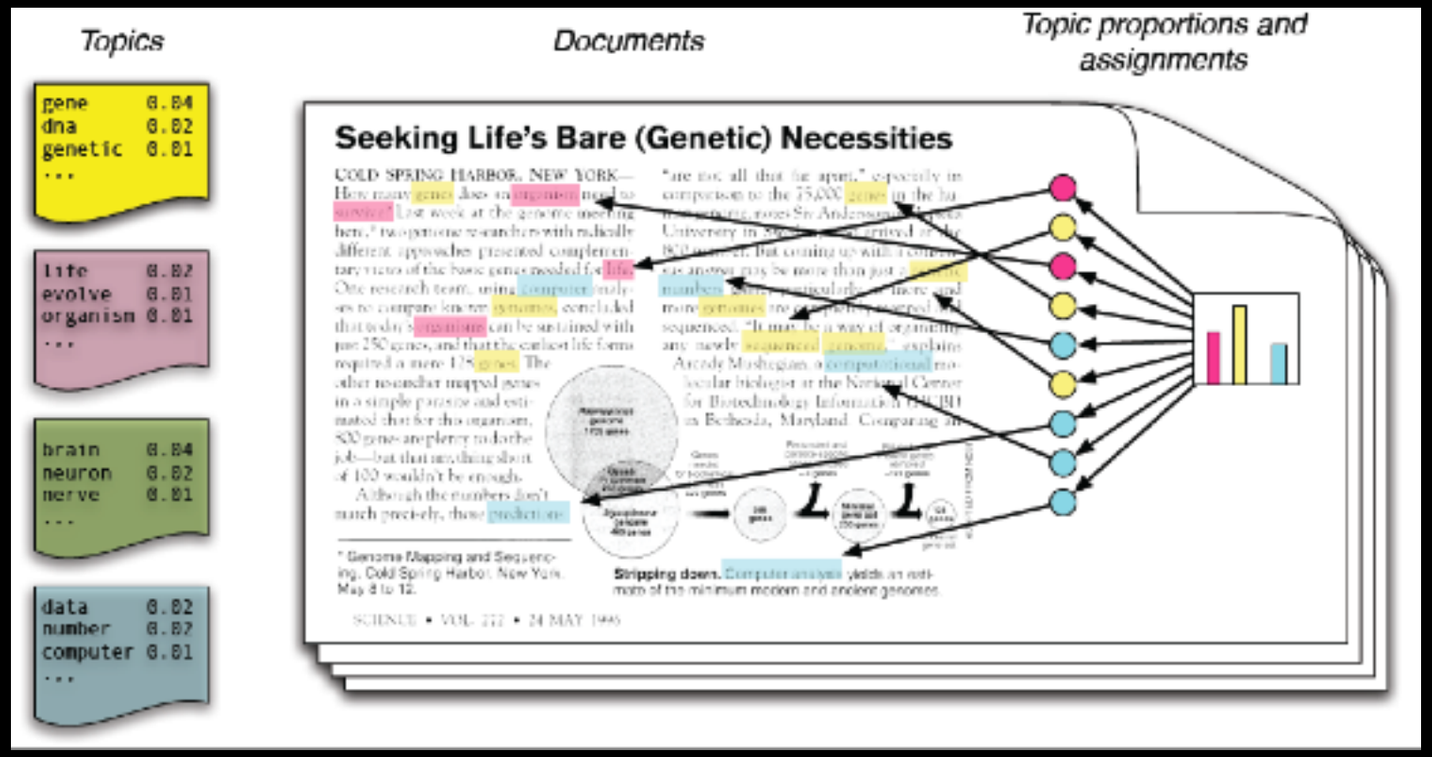
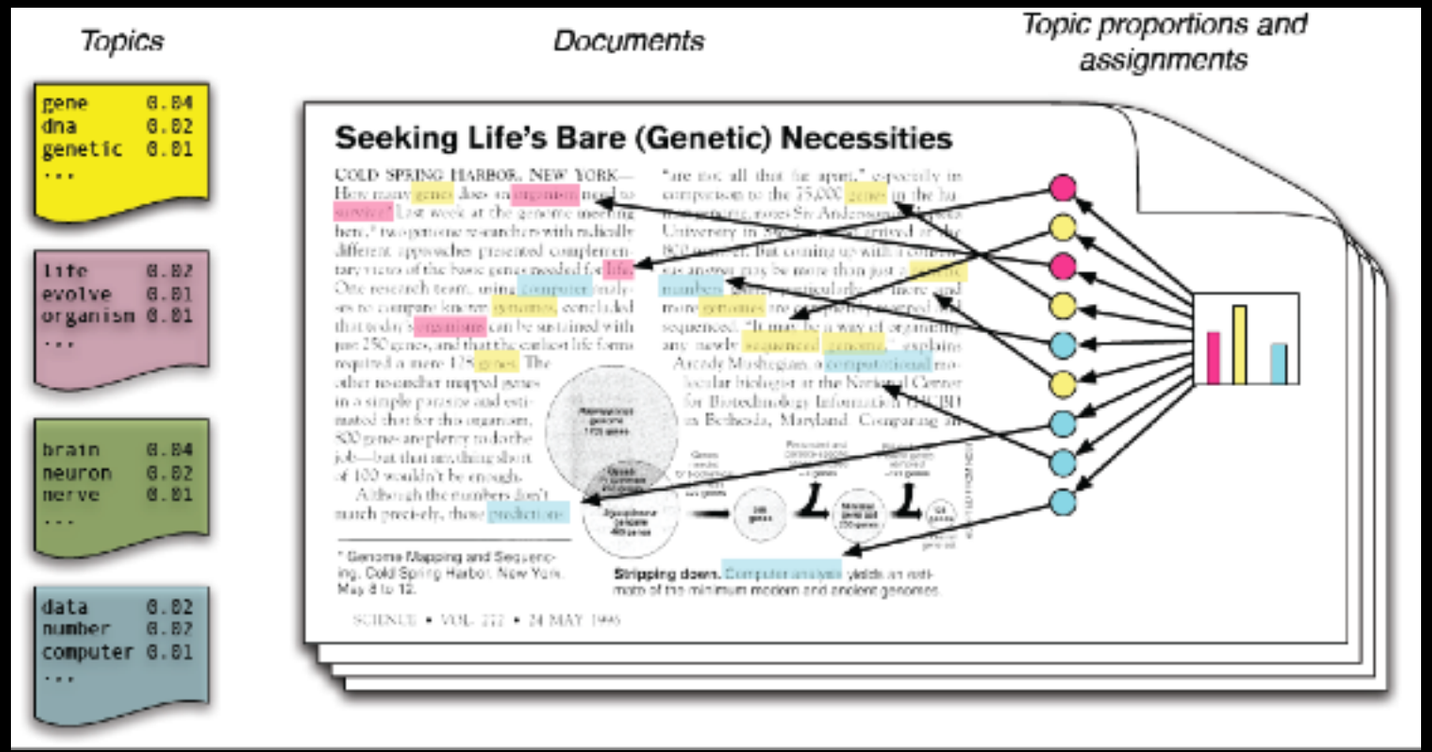


- **Topic modeling** is the process of identifying **topics** (or clusters of terms that co-occur together) in a corpus (i.e., a set of documents).

- **Latent Dirichlet Allocation** (LDA) is an example of topic model and is used to classify text in a document to a particular topic.

- **LDA** is a form of **machine learning** with text data that views documents as **bags of words**; in other words, the order of the **n-grams** does not matter.



- **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.

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