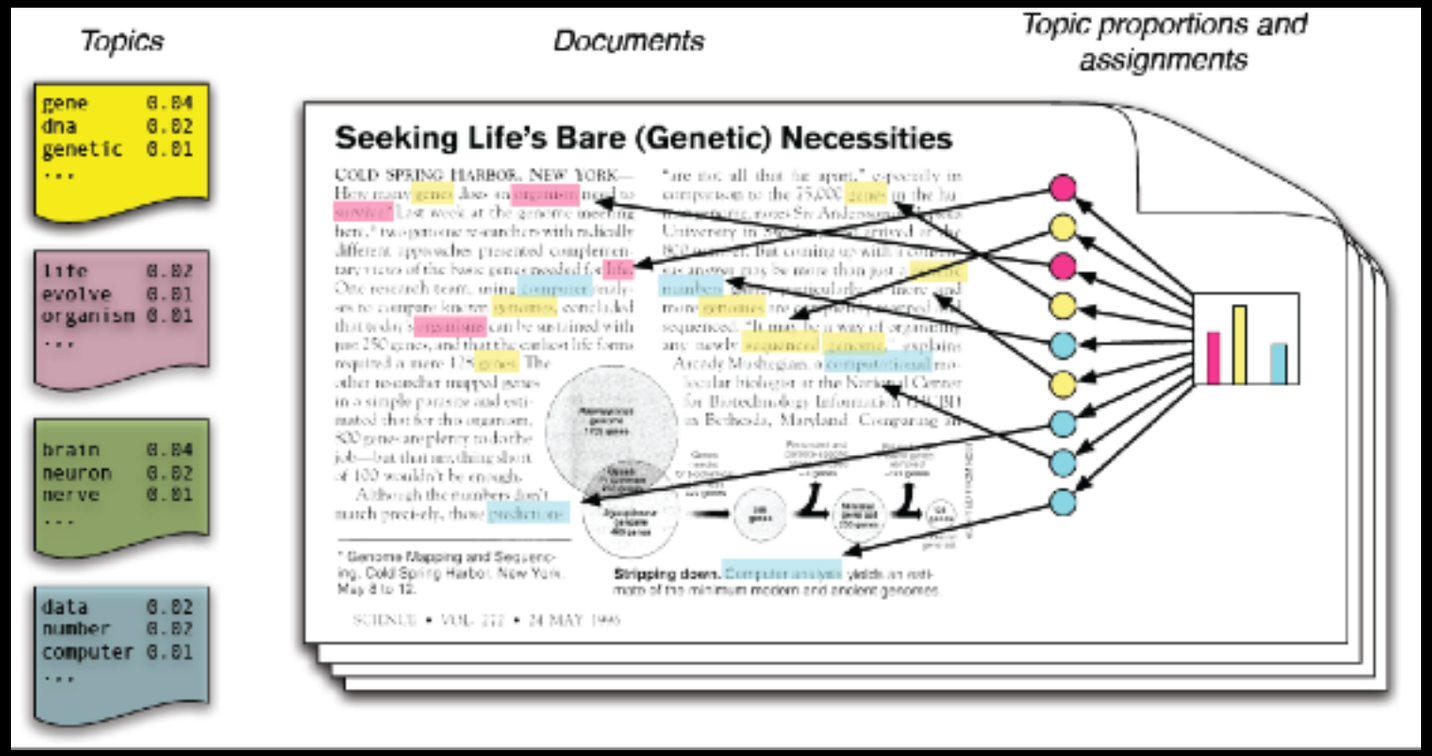
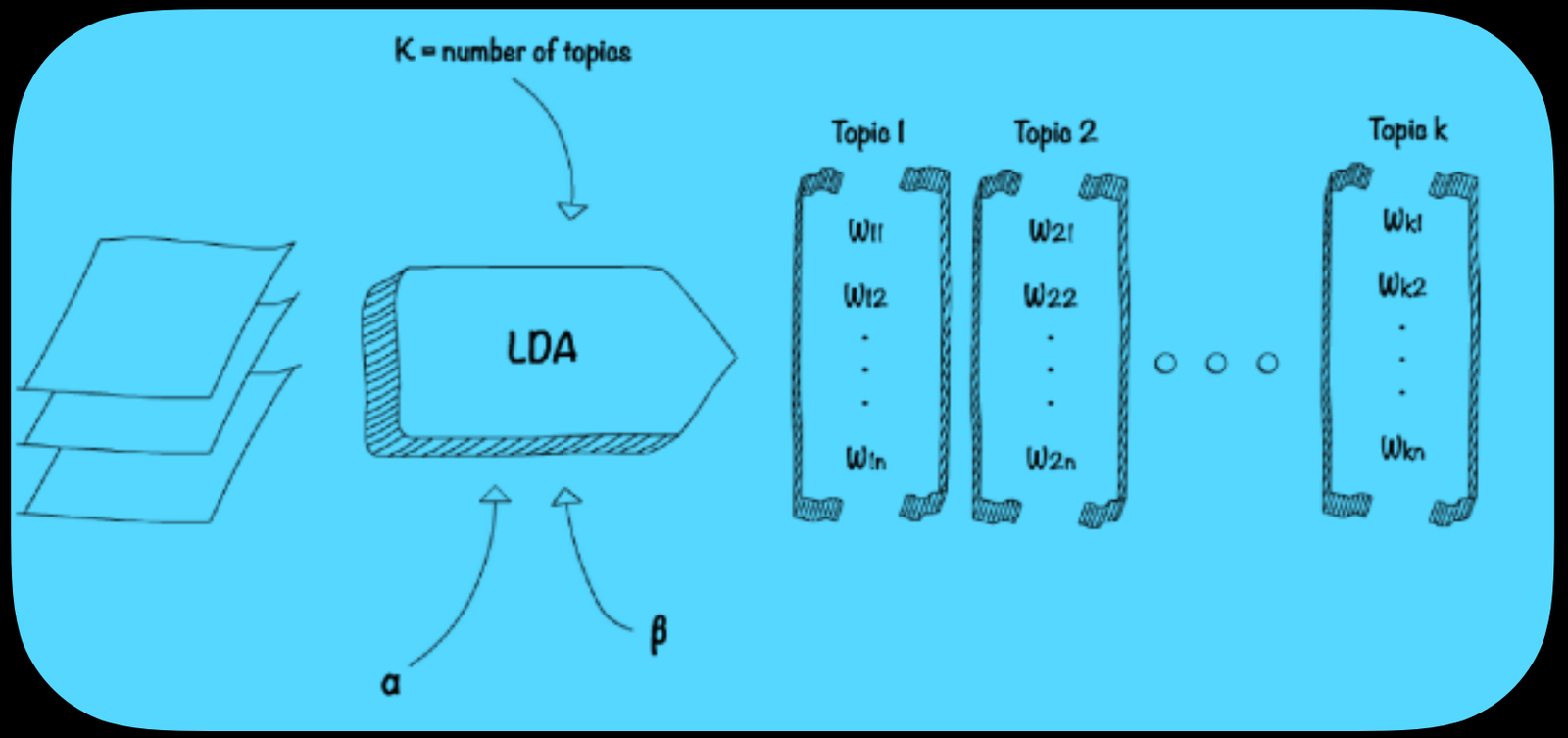


- We first assume that **some K number of topics** exist across all of the documents in the corpus. There could be more topics or less. It's just a first guess!
- Next, we randomly guess how much each **document** is made of these **topics**.
- For instance, **document 1** could be made up of terms that 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**.



- Obviously, this random distribution is probably very wrong, as it was completely assigned at random.

- So, for each **word w** in **document m**, we assign **word w** to **a new topic** based on two things:
  - What topics are already in **document m**?
  - How many times **word w** has been assigned a **particular topic** across all of the documents: This term-to-topic distribution is called **beta**.

- We then update our **alpha** and **beta** values, but it's still probably very wrong. So, what do we do?

- We then update our **alpha** and **beta** values, but it's still probably very wrong. So, what do we do?