# Text 101: tokens

We reduce the linguistic meaning of 'text', 'token' to the following:
This reduces it quite extremely, but its simplicity will suffice for us.

**Text:**

A sequence of tokens(words).

**Token**/word:

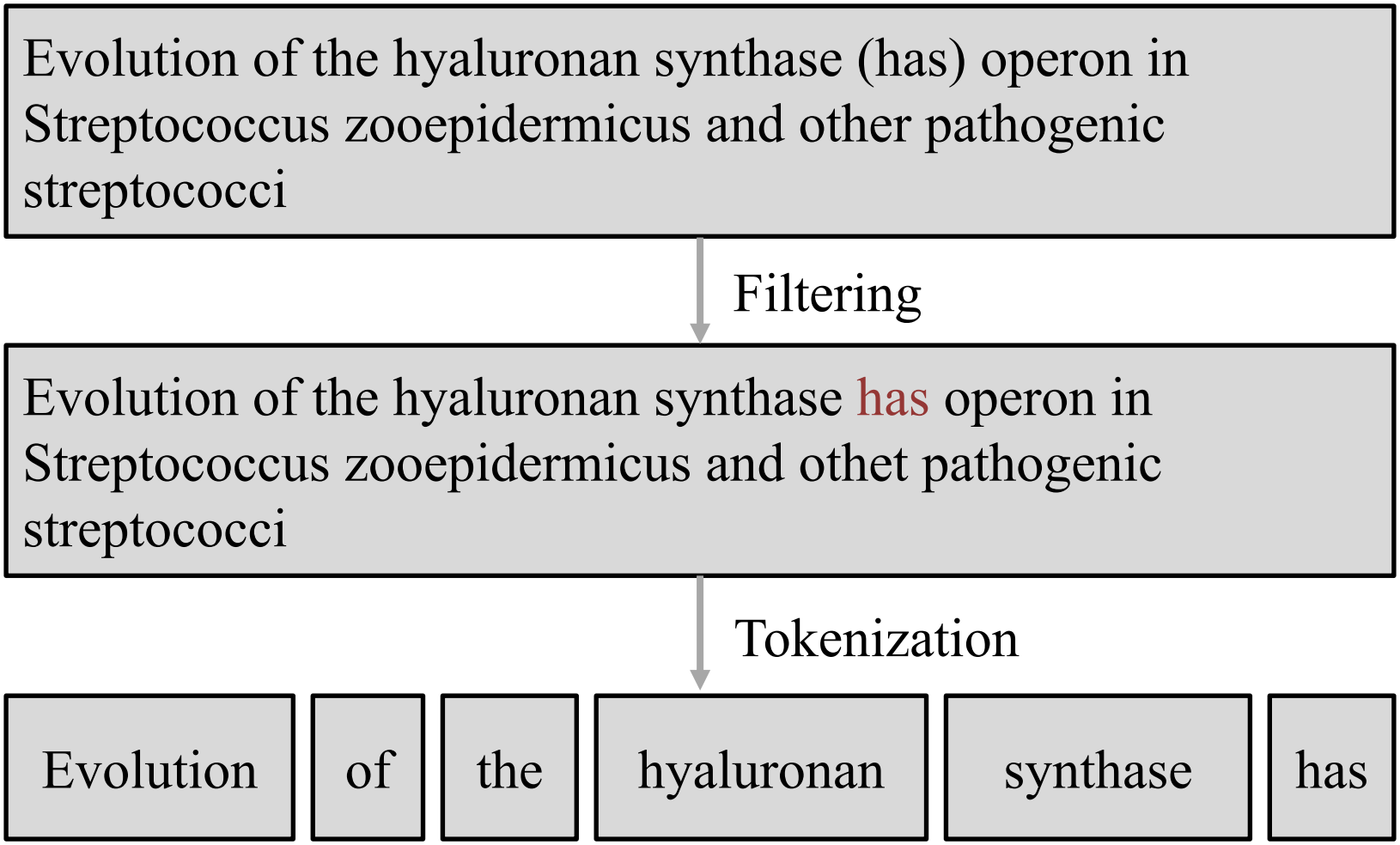
A sequence of characters.

**Character:**

An atomic element of text.

¯_(ツ)_/¯

# Text 101: tokens



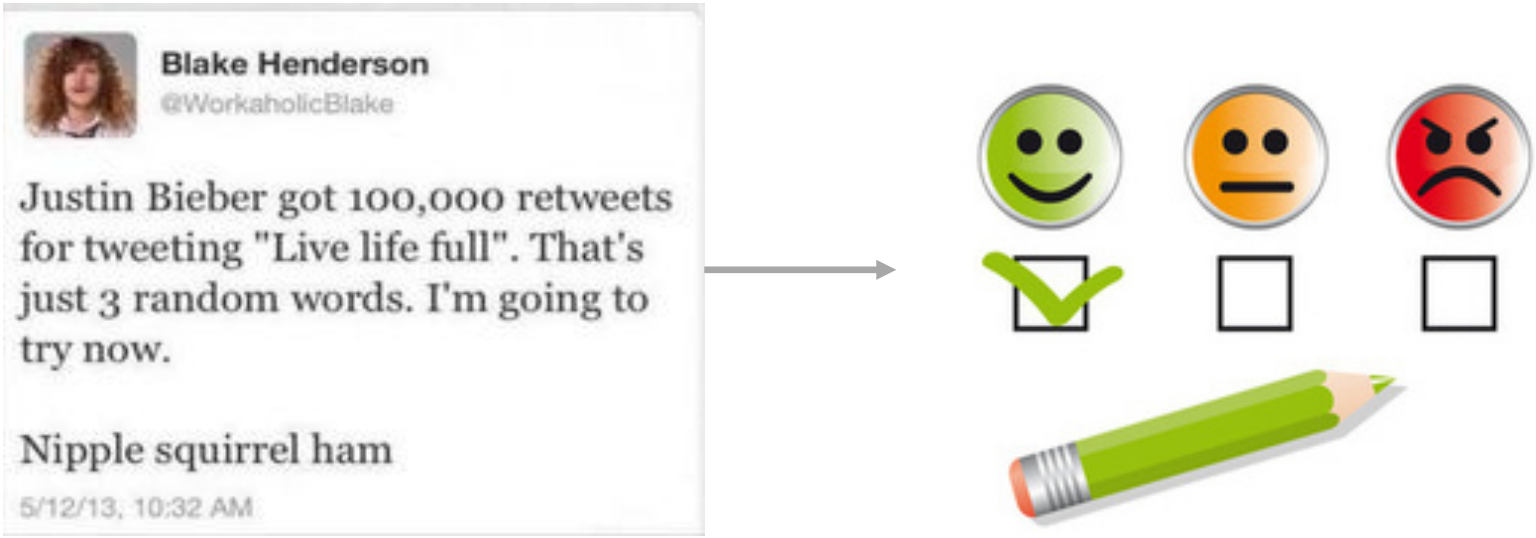
# What is a text 101: bag of words

Journal of Artificial Intelligence Research
JAIR is a refereed journal, covering the areas of Artificial intelligence, which is distributed free of charge over the Internet. Each volume of the journal is also published by Morgan Kaufmann

| | |
|---|---|
| 0 | learning |
| 3 | journal |
| 2 | intelligence |
| 0 | text |
| 1 | Internet |
| … | … |

this is the BoW approach.
We have a word count for every word in the text.

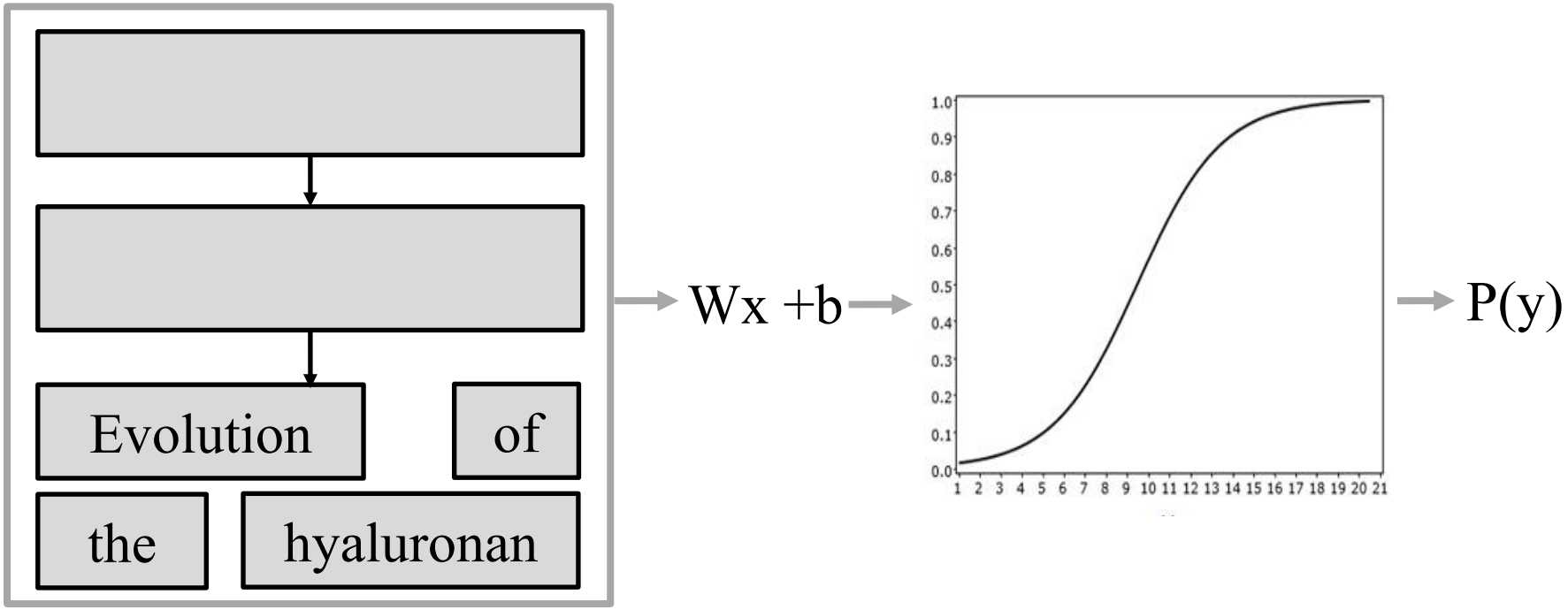
# Text classification/regression




Other usage:

- Adult content filter (safe search)
- Detect age/gender/interests by search queries
- Convert movie review into “stars”
- Survey public opinion for the new iphone Vs old one (SNA)
- ...

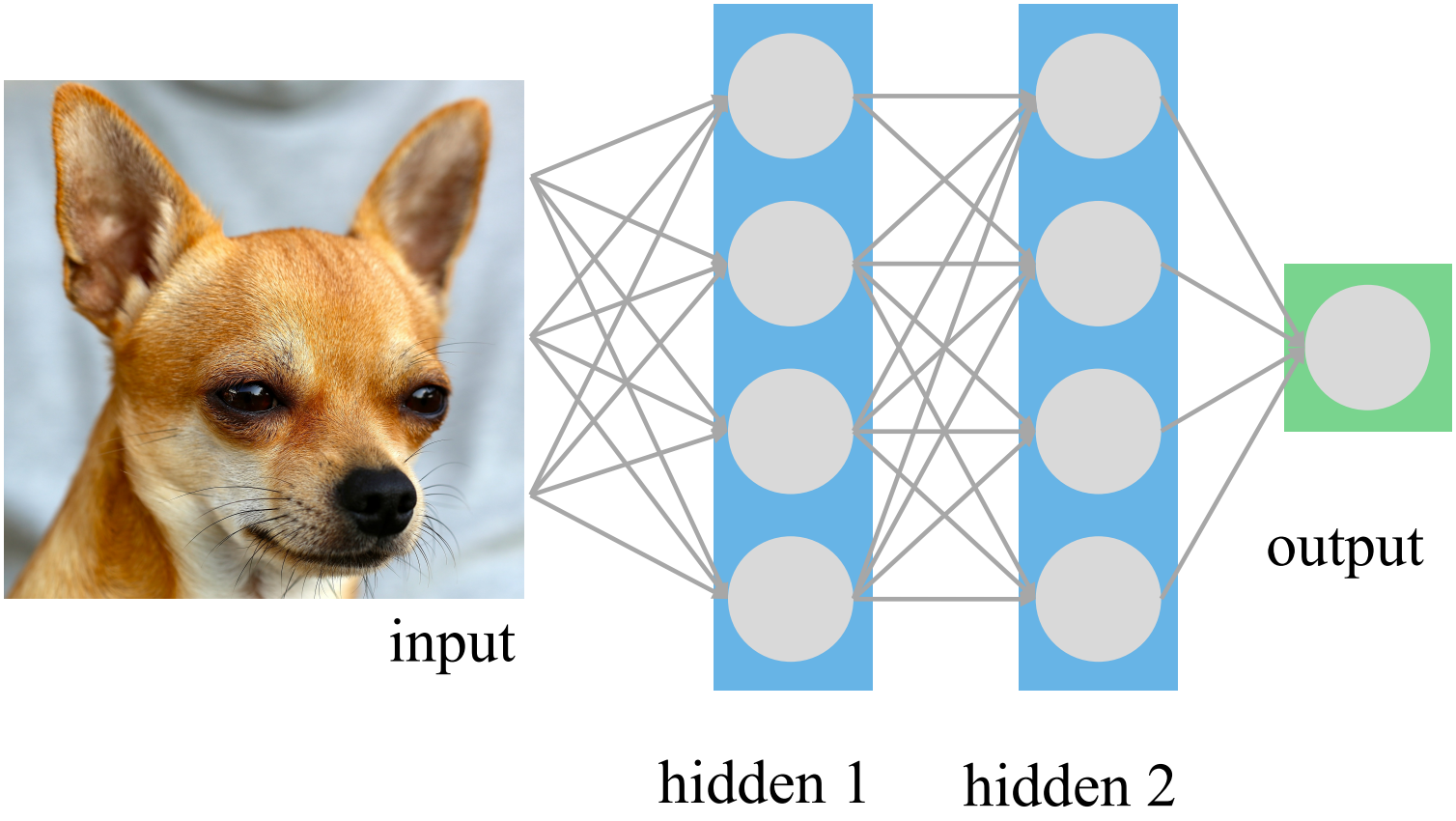
# Text classification: BoW + linear



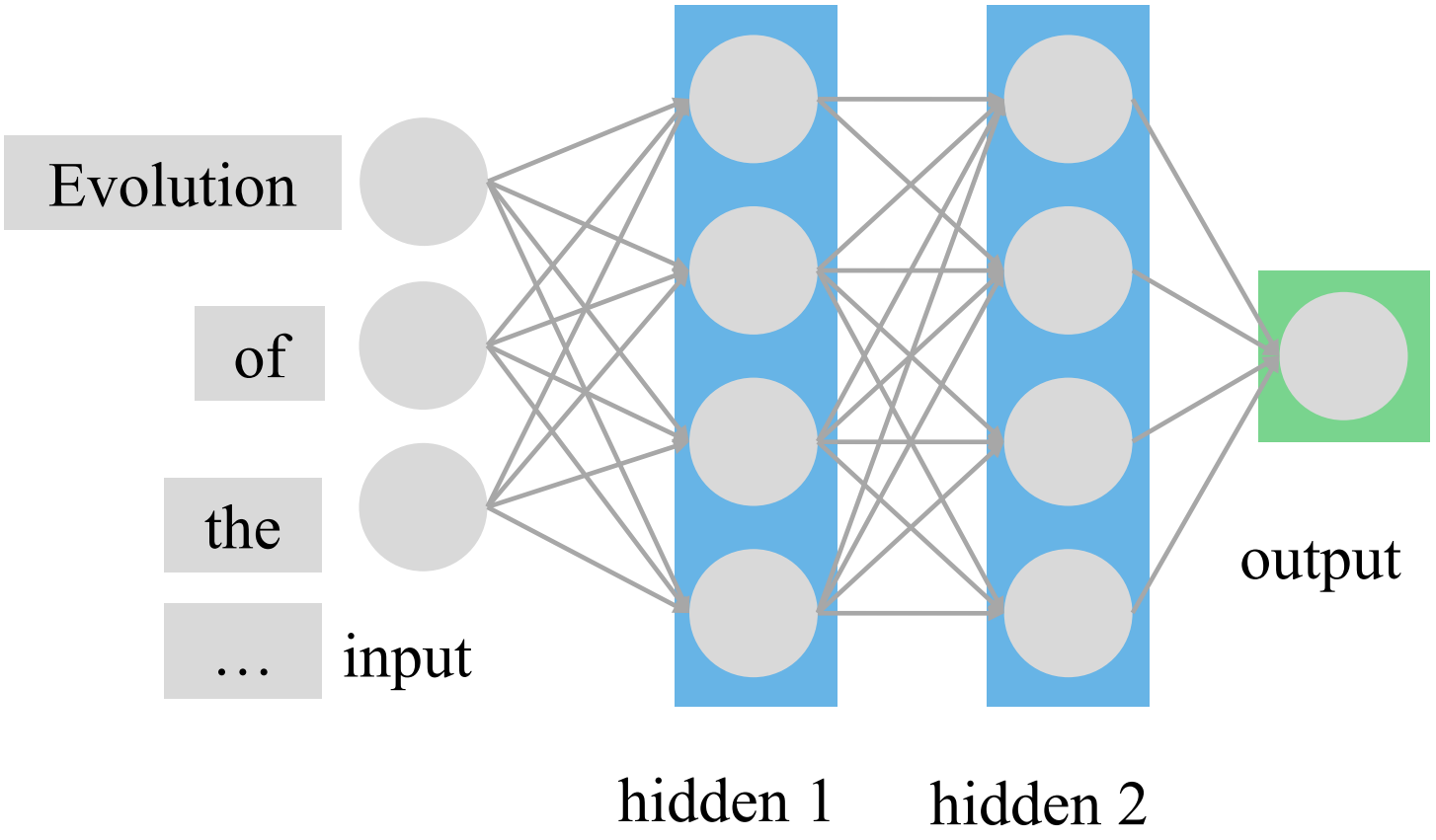

Divination:

How many features (approx.) will such model have?

# Text classification: BoW + linear

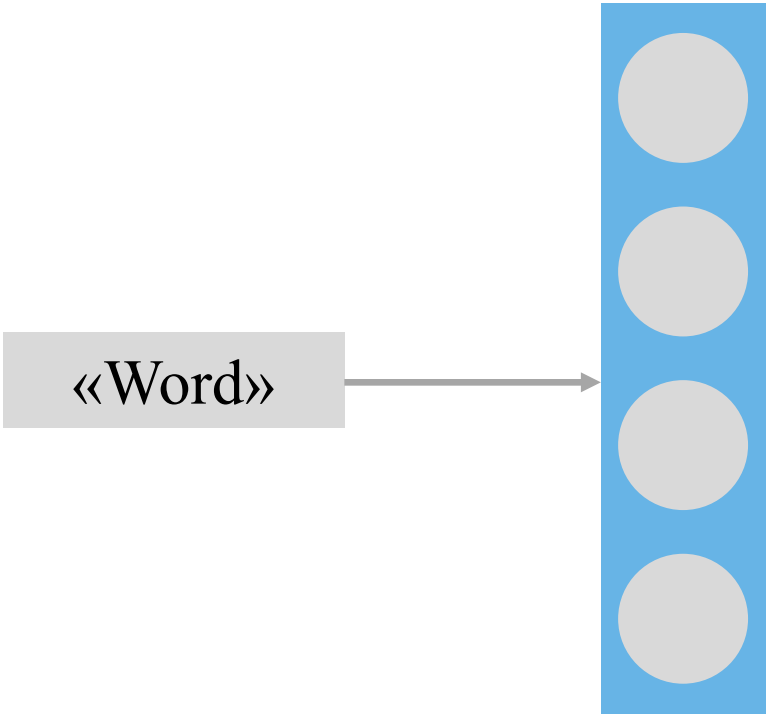


# Text classification: BoW + linear



# Word embeddings

We want a compact representation of text so that we could use it for neural nets!



# Recap: embeddings

Map data into a lower dimensional space while preserving structure. MDS, LLE, **TSNE**, etc.