

Text 101: tokens

Text:

A sequence of tokens(words).

Token/word:

A sequence of characters.

Character:

An atomic element of text.

¬_(ツ)_/

Text 101: tokens

Evolution of the hyaluronan synthase (has) operon in Streptococcus zooepidermicus and other pathogenic streptococci

Filtering

Evolution of the hyaluronan synthase **has** operon in Streptococcus zooepidermicus and othet pathogenic streptococci

Tokenization

Evolution

of

the

hyaluronan

synthase

has

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
...	...

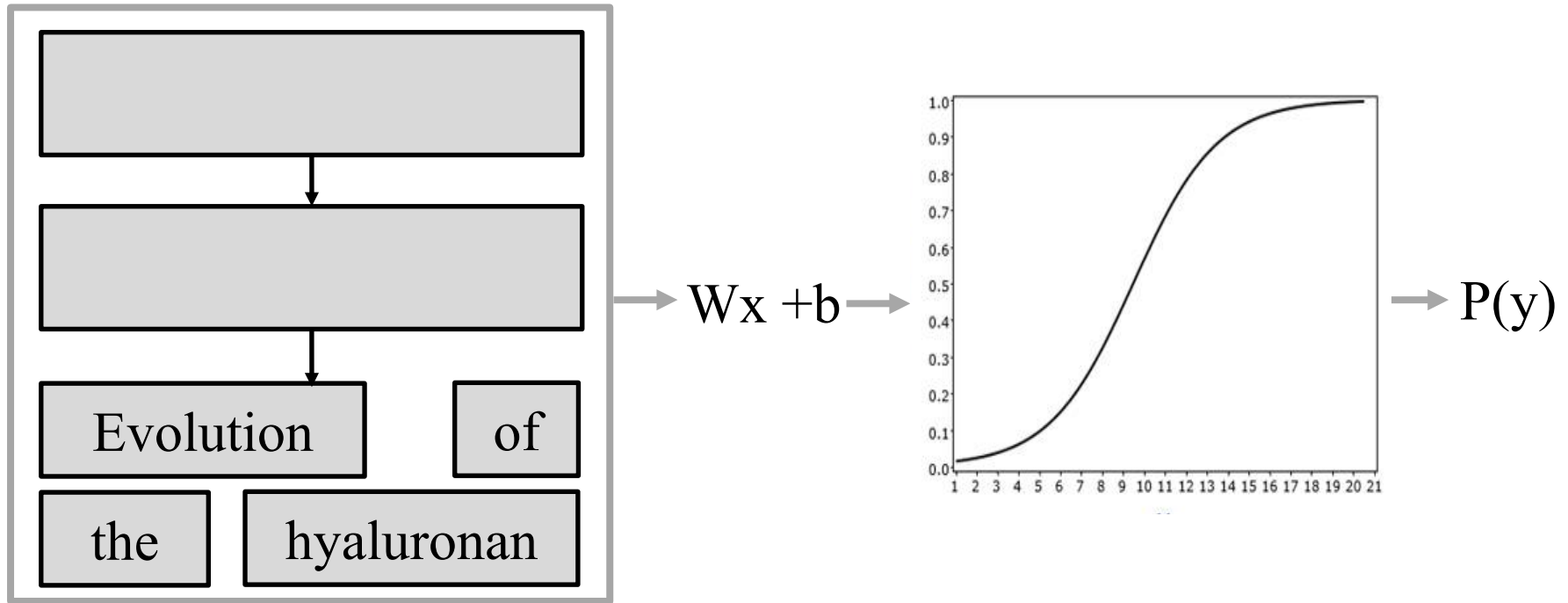
Text classification/regression



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)
- ...

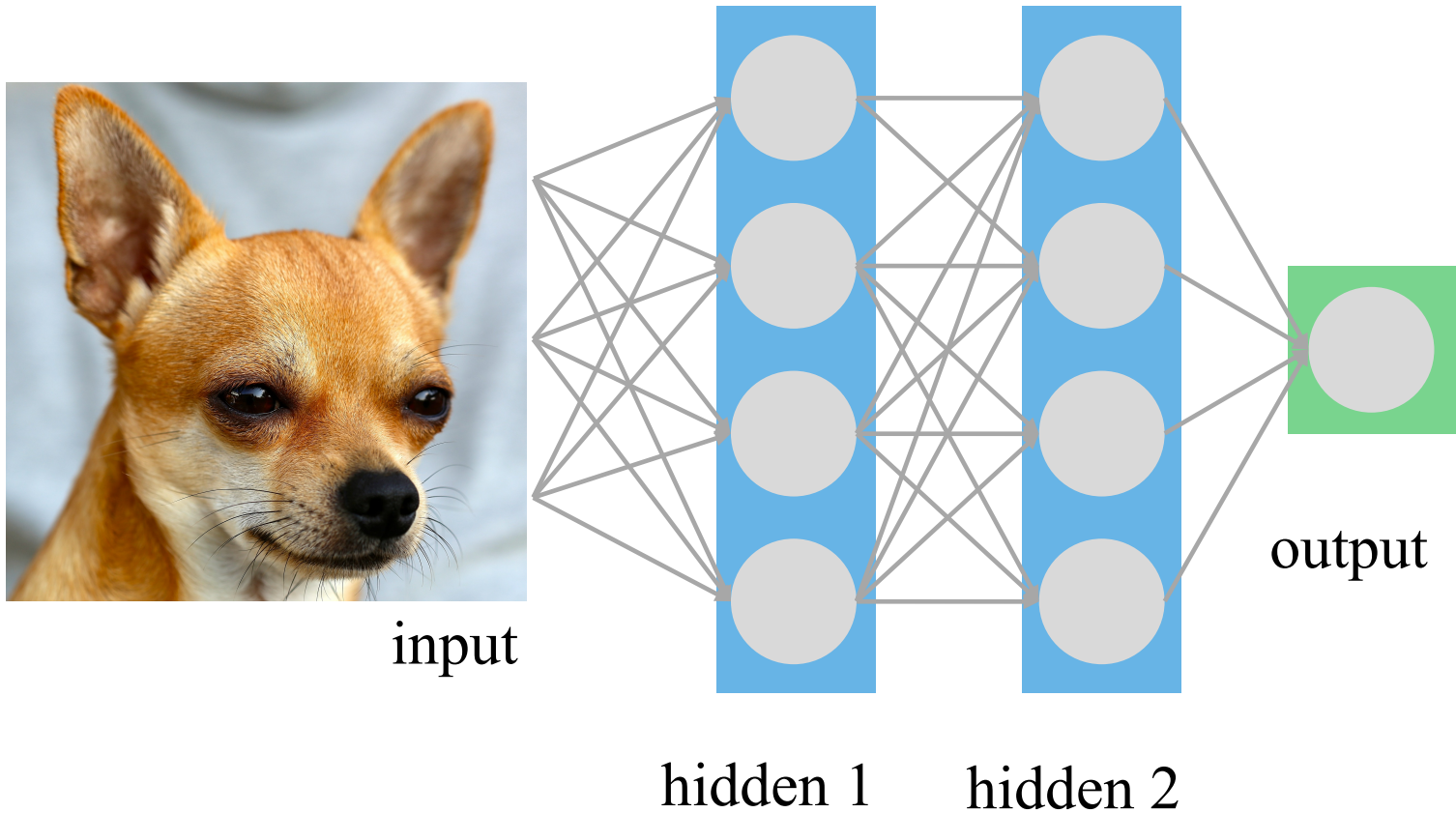
Text classification: BoW + linear



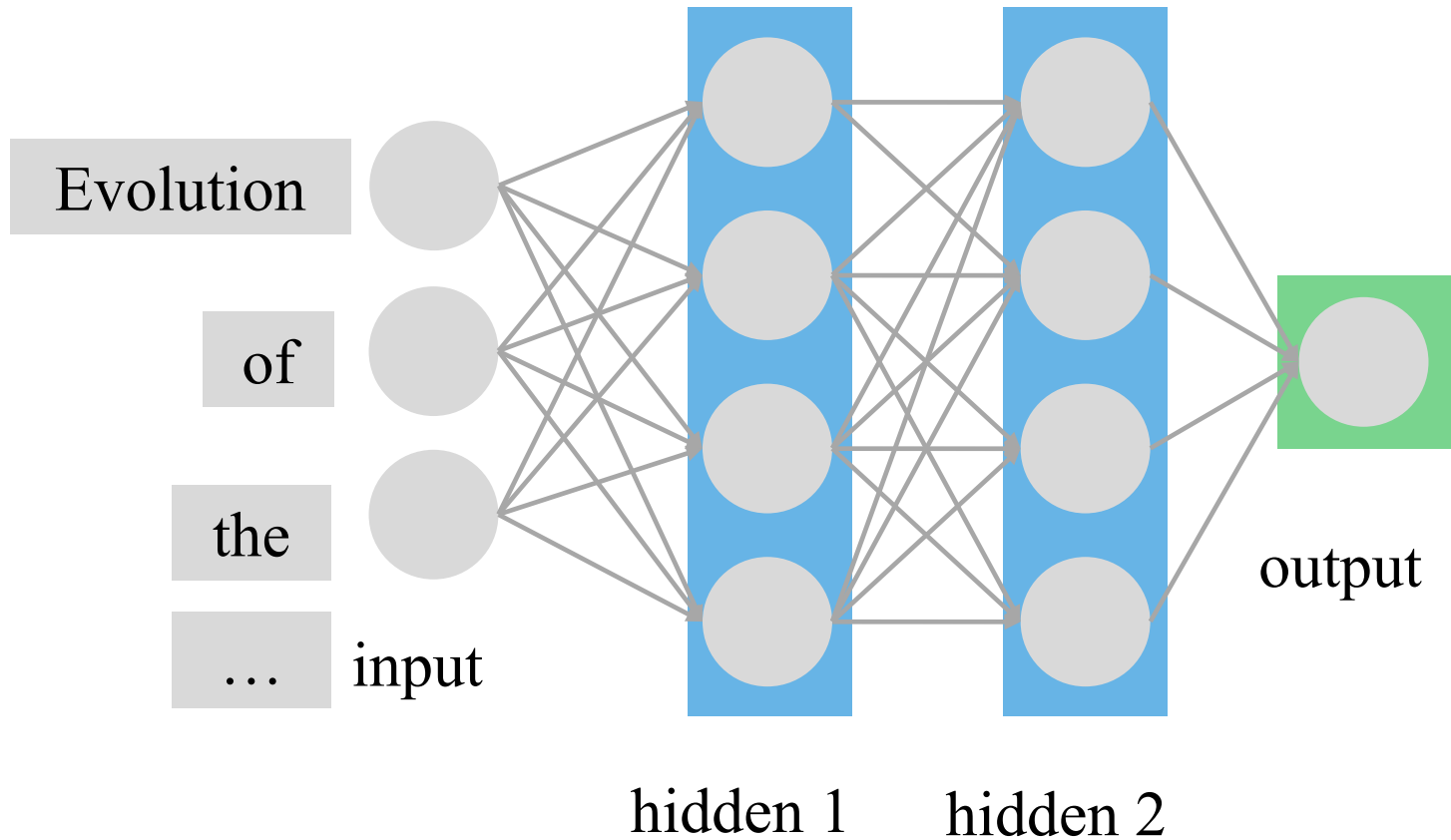
Divination:

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

Text classification: BoW + linear

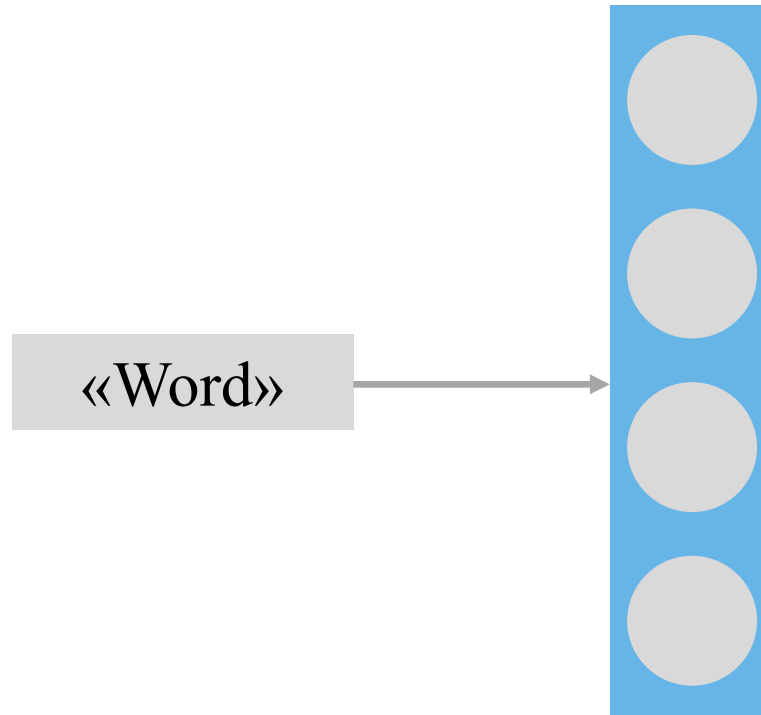


Text classification: BoW + linear



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.

