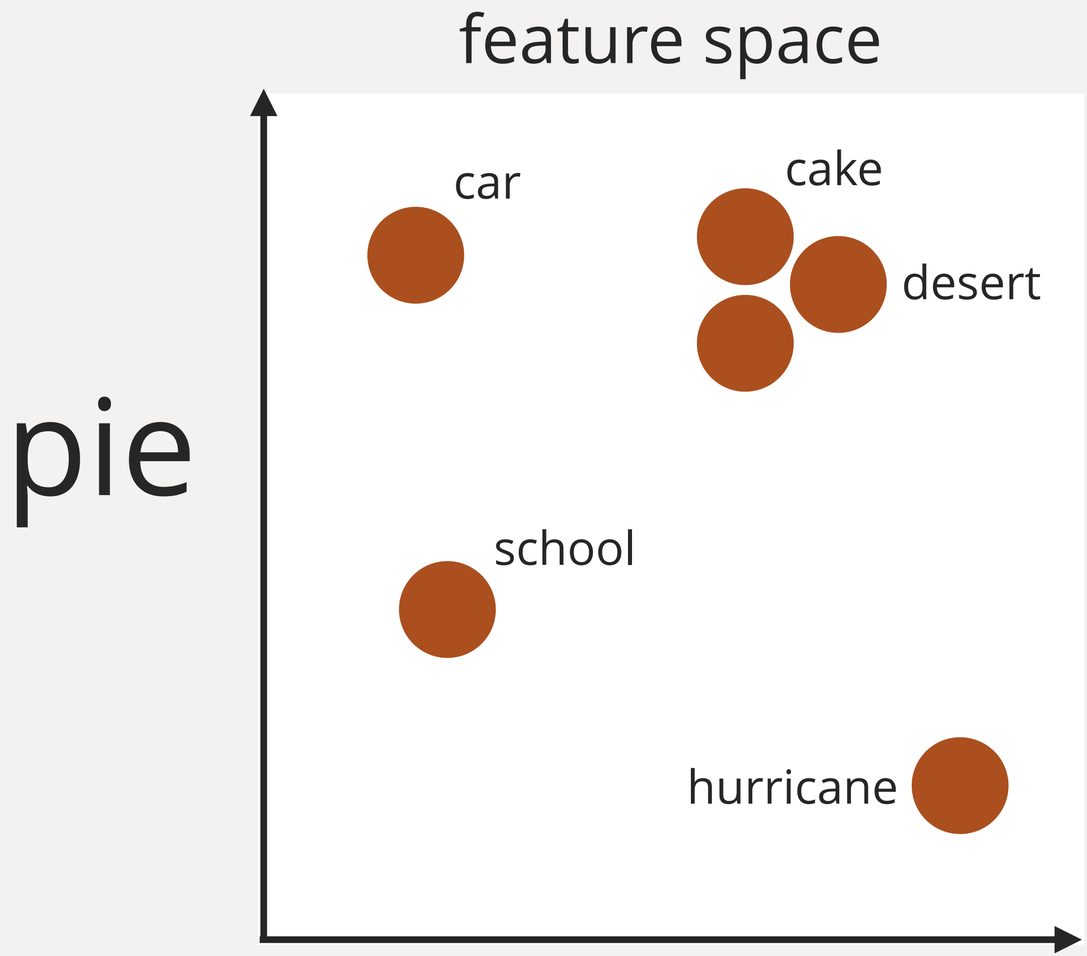
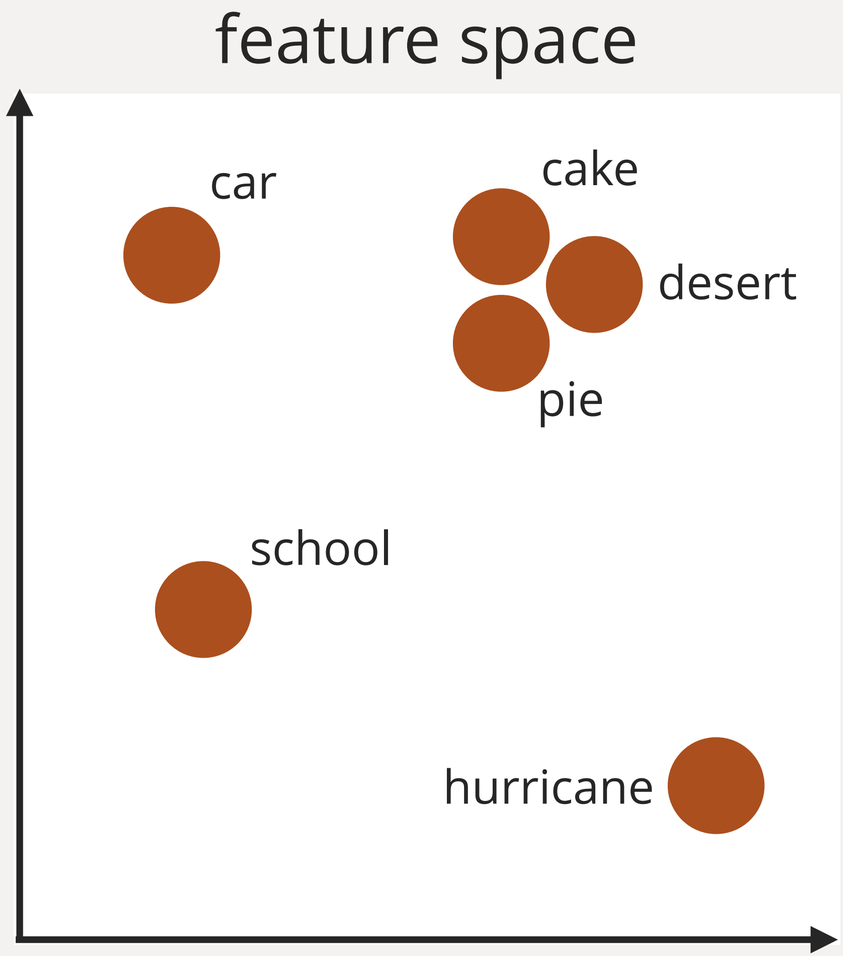
Word to Vec (Word2Vec)

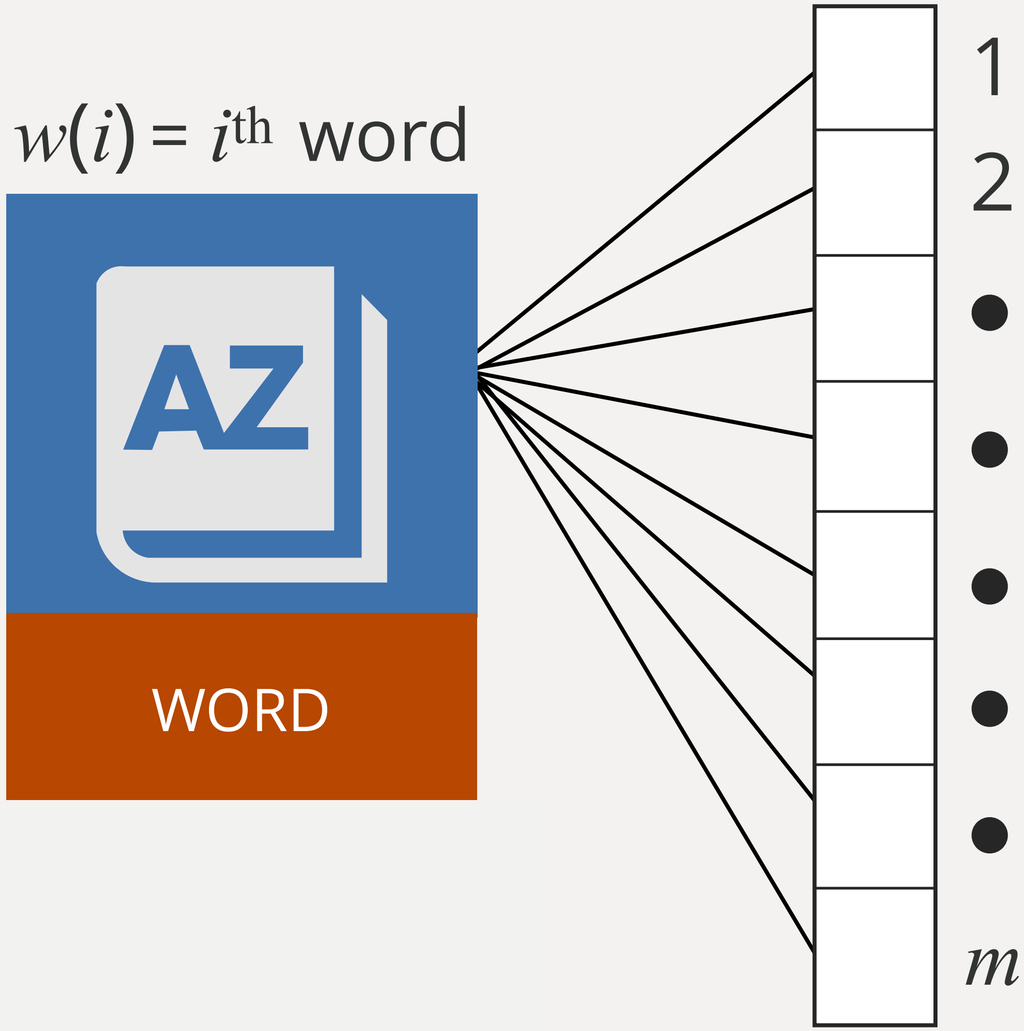
Each of the vectors that are associated with a given word is often referred to as embedding

Word Embedding

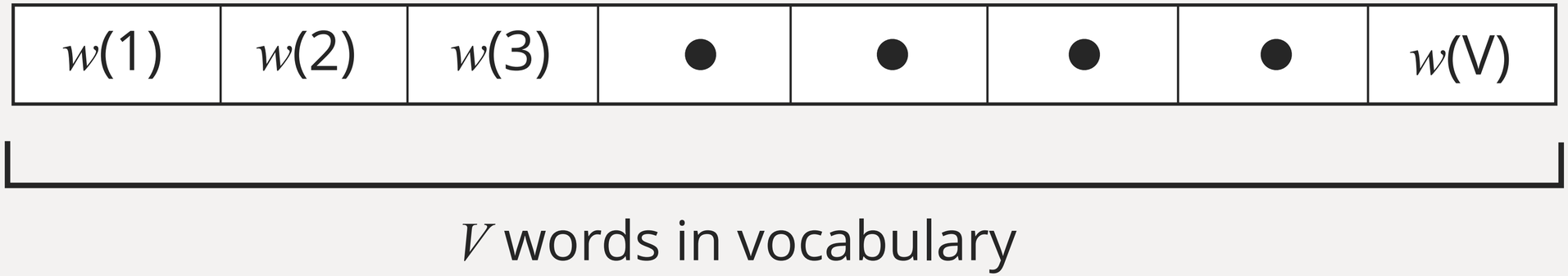


Word Embedding

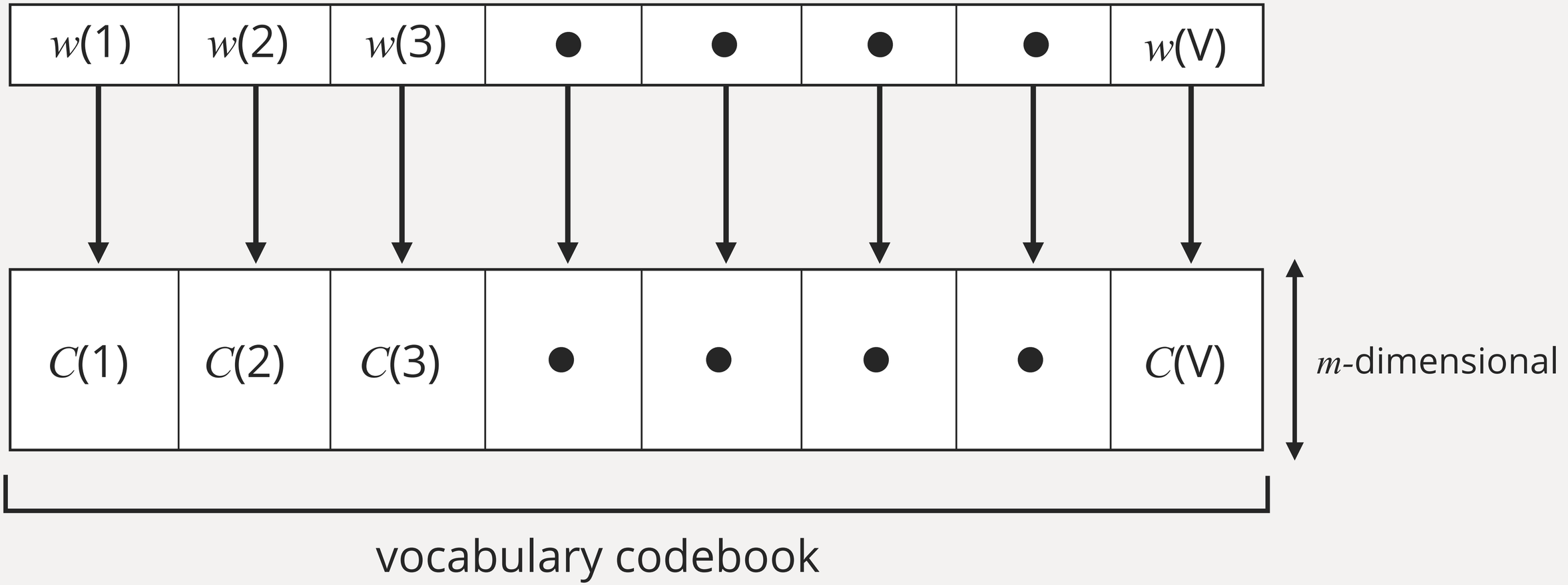


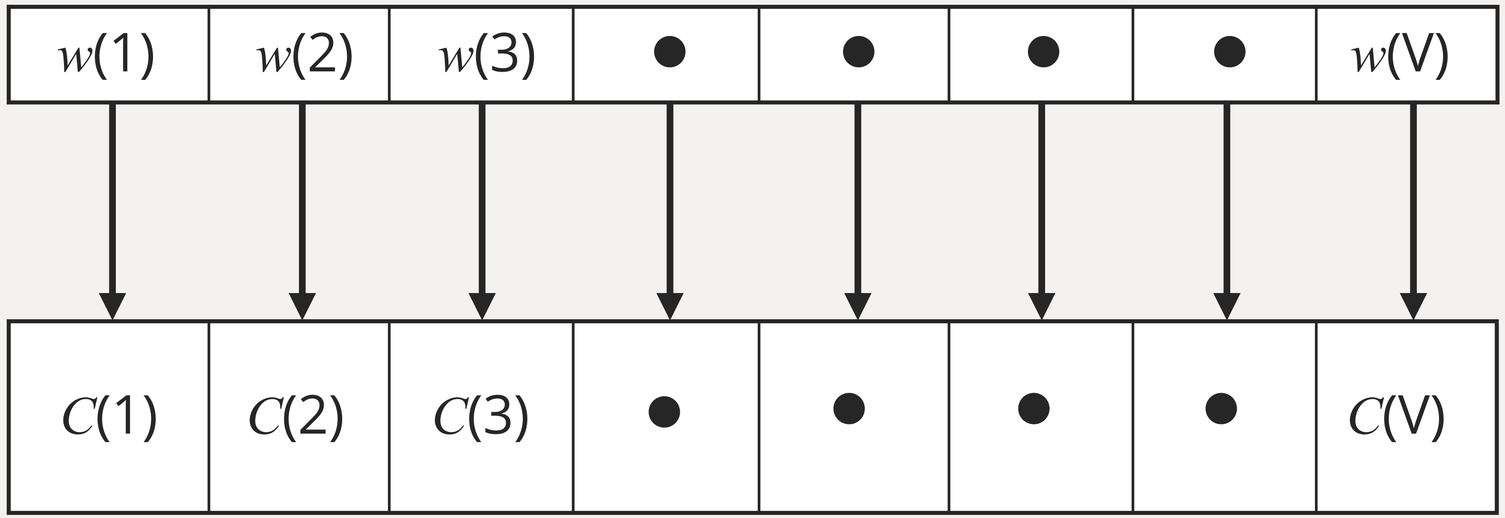

$w(i) = i^{\text{th}}$ word
AZ
WORD
1
2
m

Word2Vec



Word2Vec

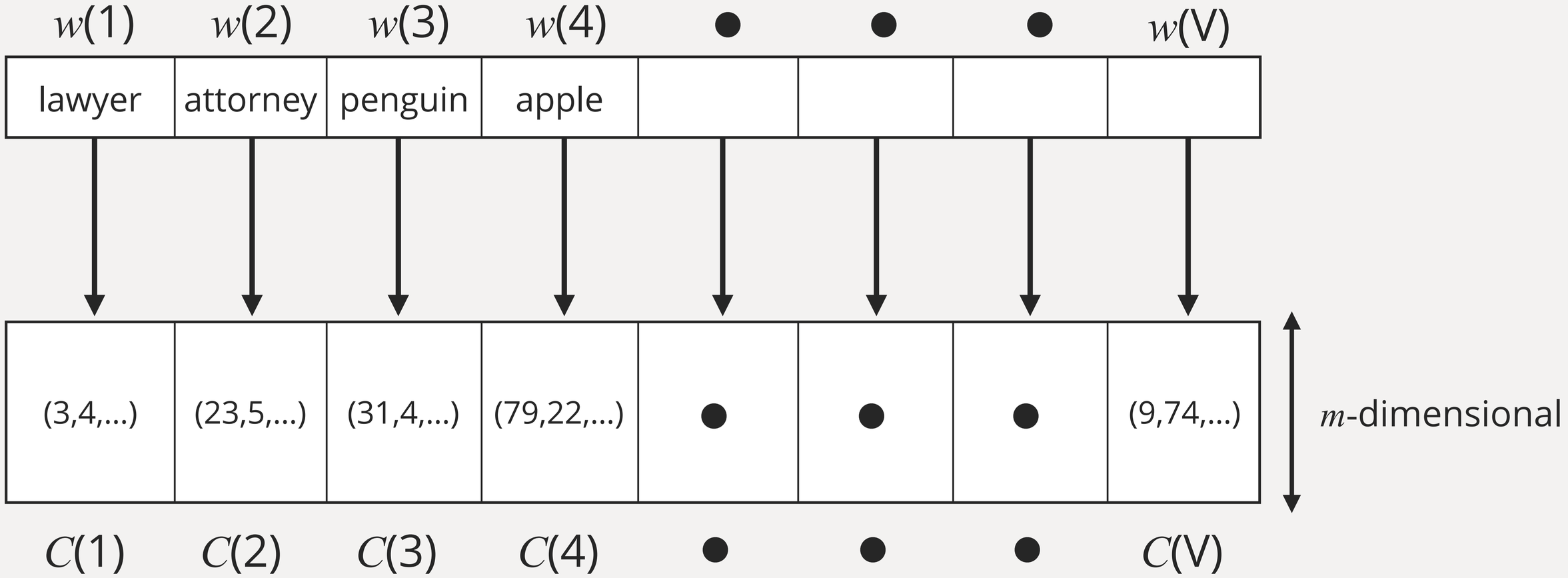




Word2Vec

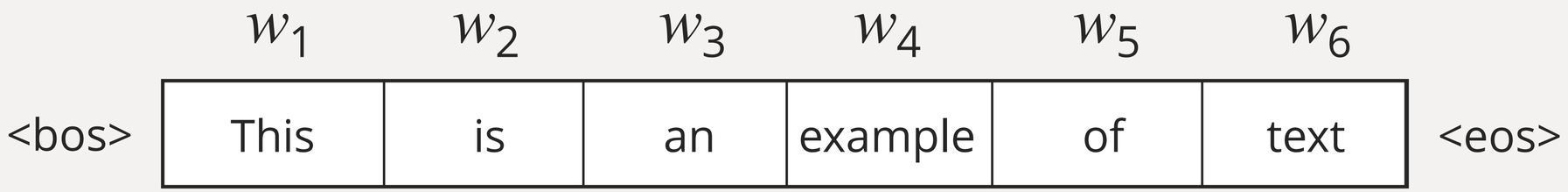
- Vector $C(i)$ is associated with i^{th} word $w(i)$
- Desire to have two similar words to in close proximity to each other
- Will learn vectors based on large corpus of text

Word2Vec


w(1)
w(2)
w(3)
w(4)
w(V)
lawyer
attorney
penguin
apple
(3,4,...)
(23,5,...)
(31,4,...)
(79,22,...)
(9,74,...)
m-dimensional
C(1)
C(2)
C(3)
C(4)
C(V)

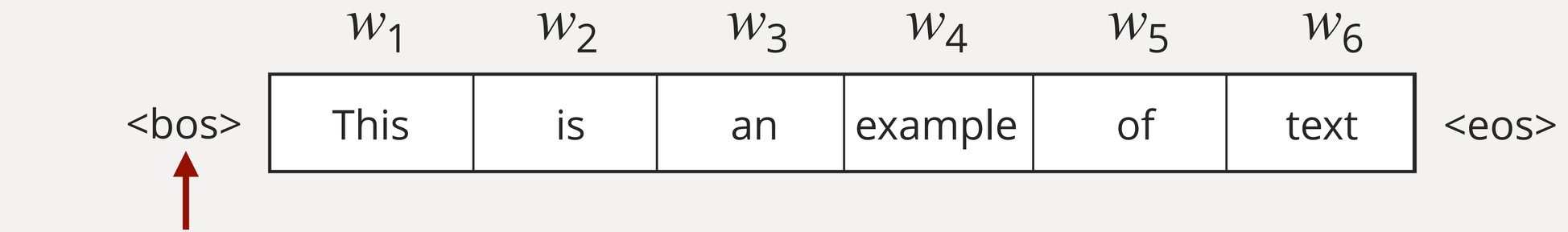
Word2Vec Representation

assume we have learned a codebook C for our vocabulary of interest



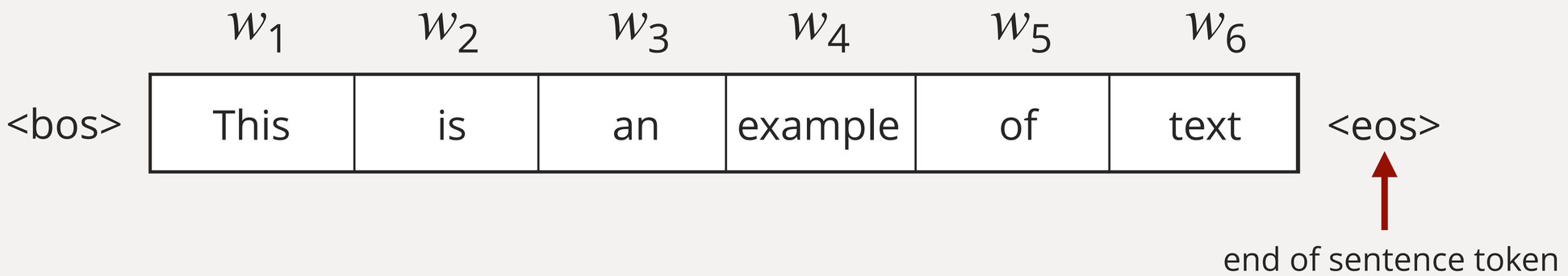
Word2Vec Representation

assume we have learned a codebook C for our vocabulary of interest



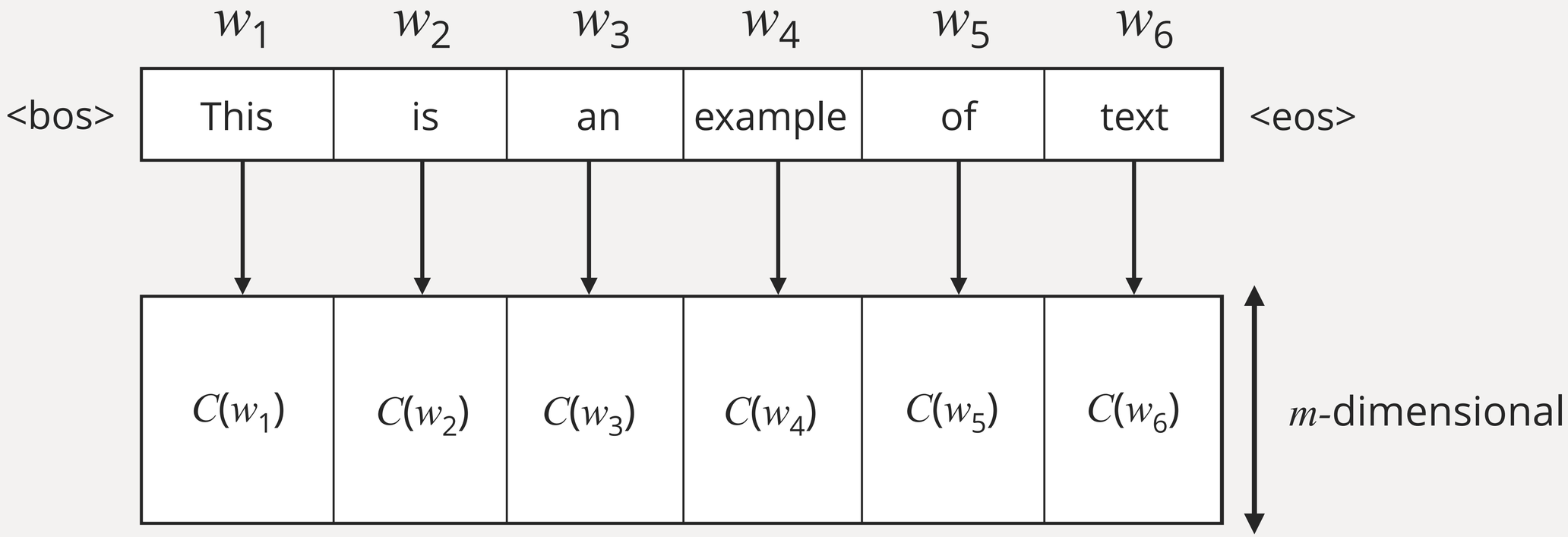
Word2Vec Representation

assume we have learned a codebook C for our vocabulary of interest



Word2Vec Representation

then any document composed of N words may be mapped to a sequence of N m-dimensional vectors



How do we learn these word vectors?