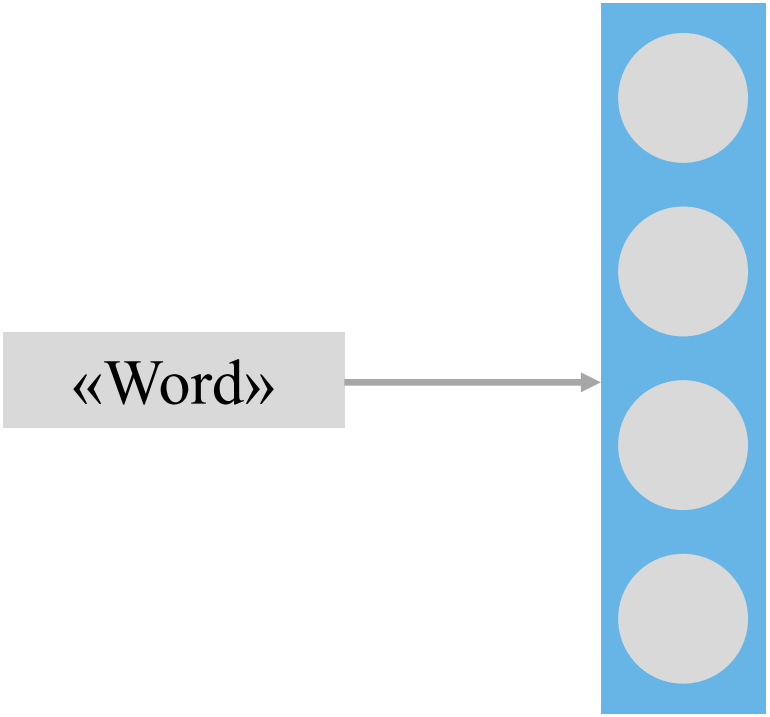
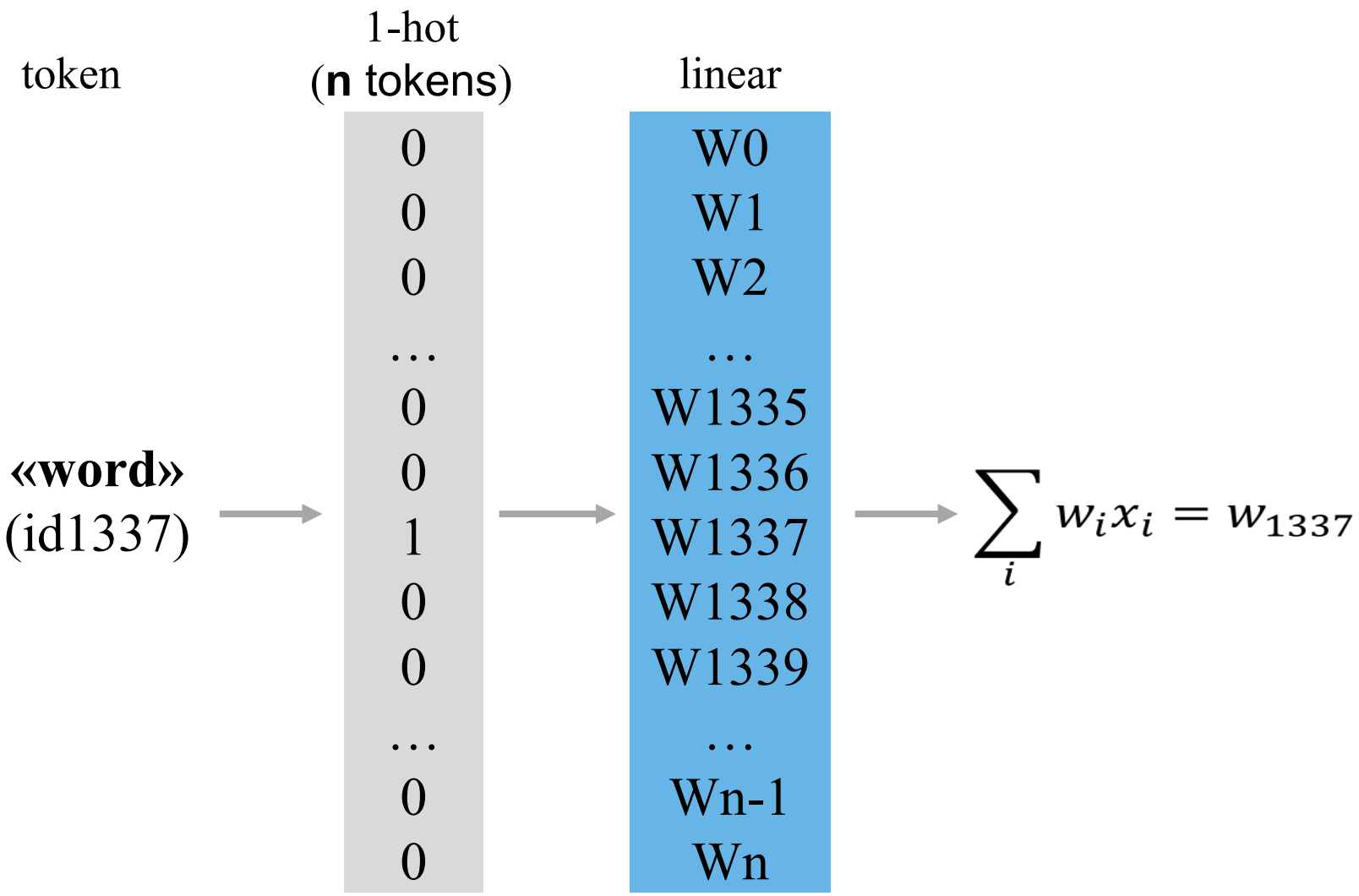
Word embeddings

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

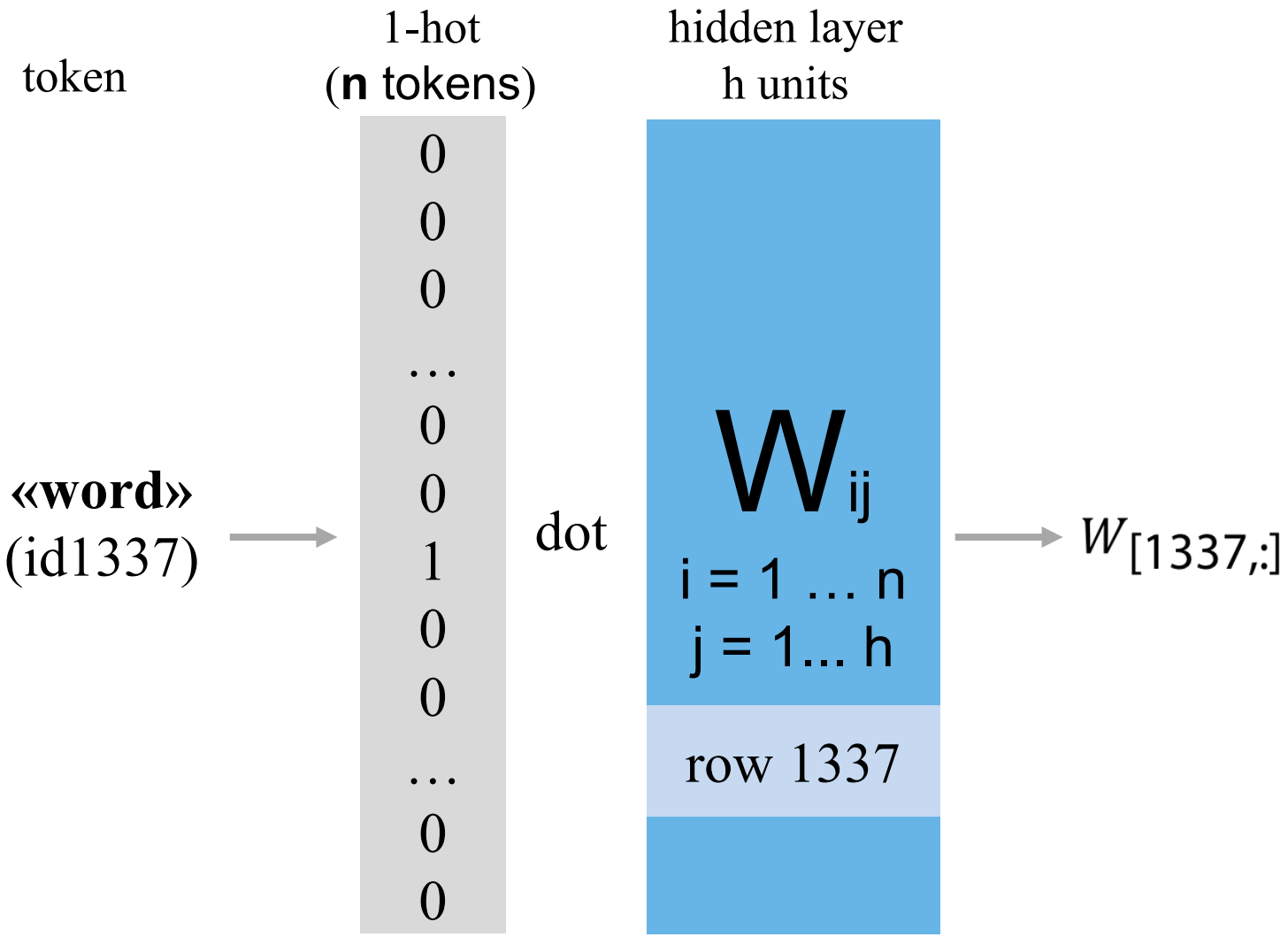


Sparse vector products



Embedding

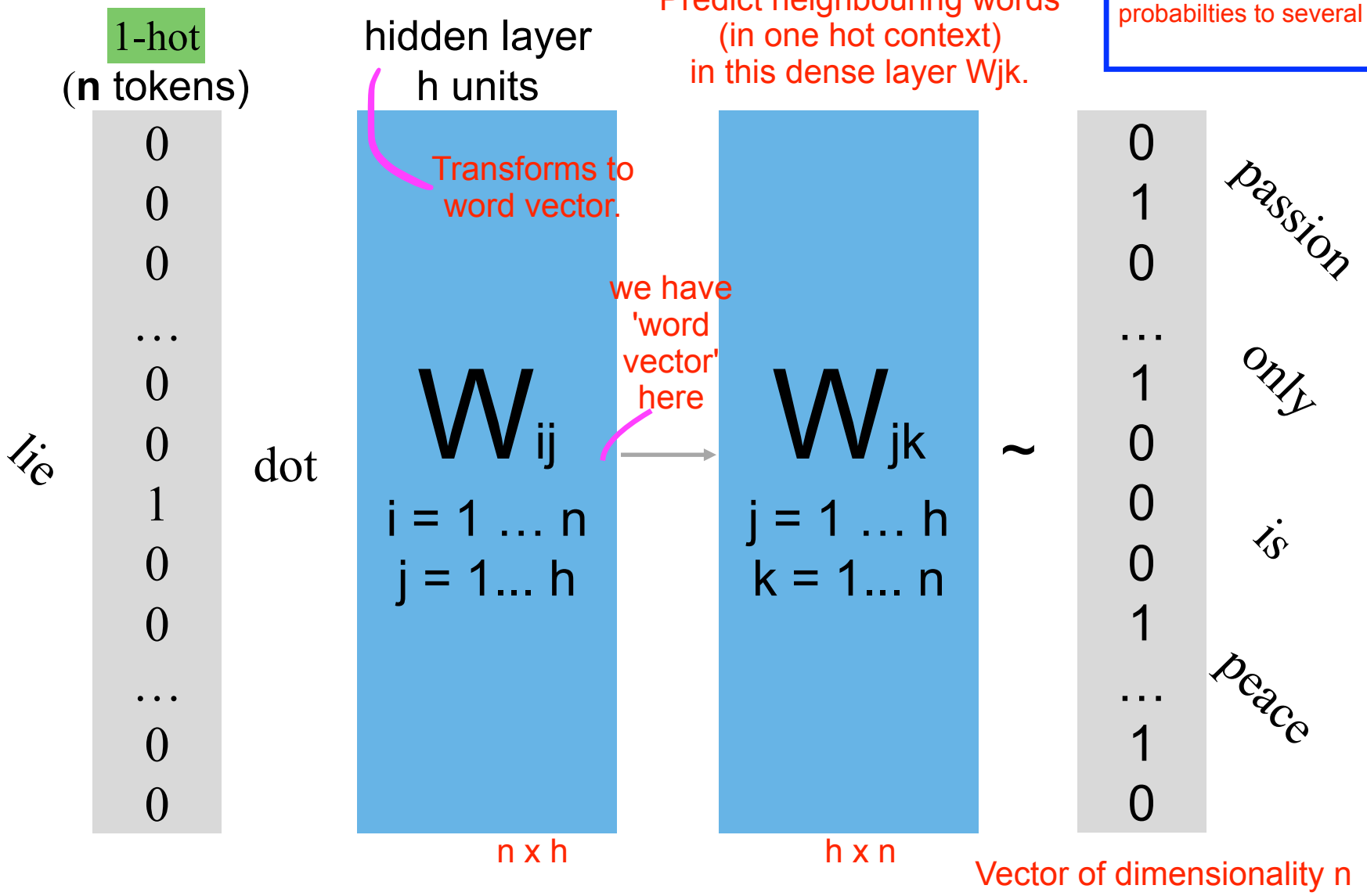
This is an 'embedding' of the word.
n would be in the millions, h would be the dimensionality of the word vector, i.e. 100, 200 or 300.



Embedding: word2vec

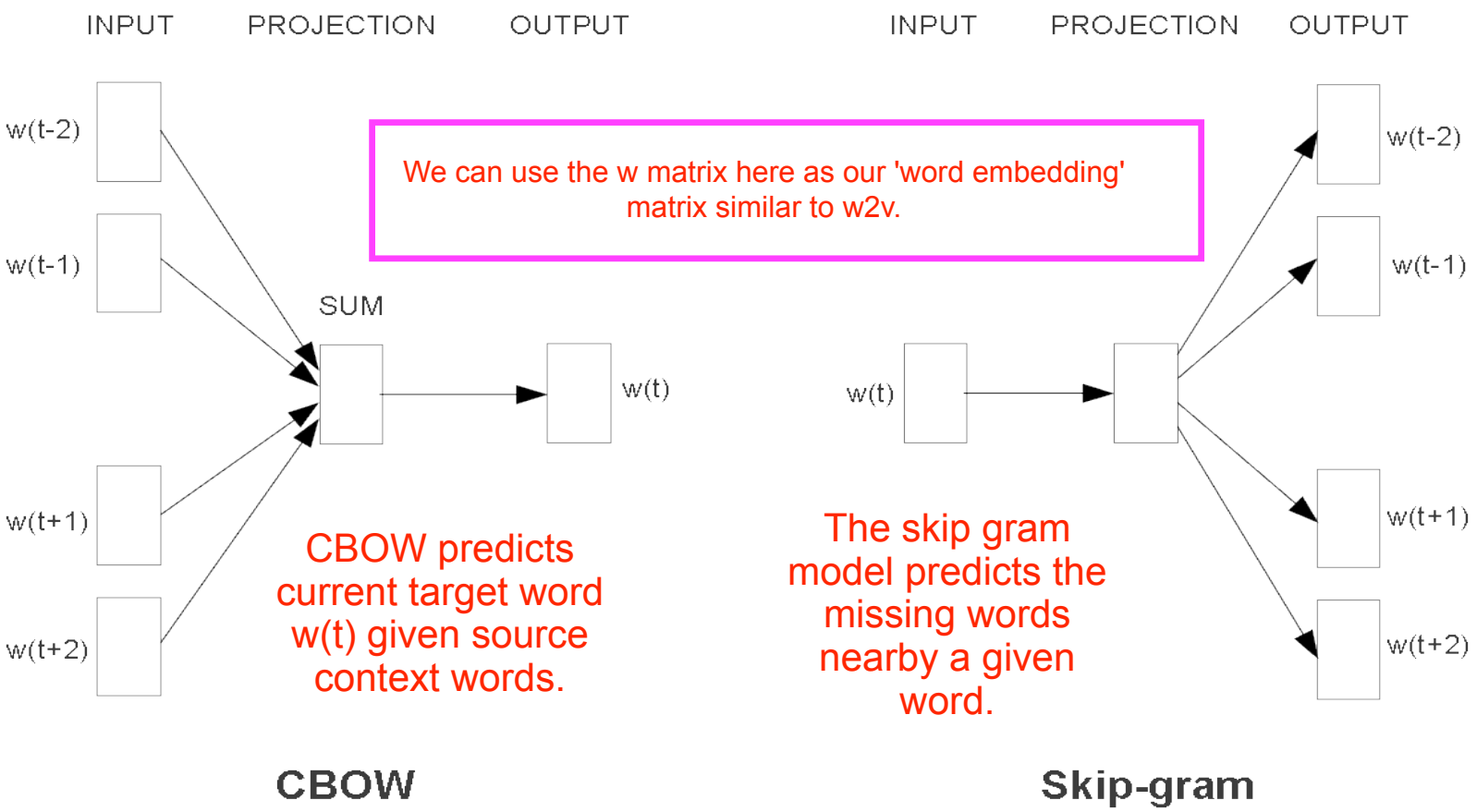
The premise is that if two words are similar, they would be assigned similar weights Wij[i1, :] and Wij[i2,:] such that their output vectors have the same 'similarity' probabilties to several words.

"Peace is a lie, there is only passion"



Embedding: word2vec

the *distributional hypothesis* : similar context = similar meaning



Yang Huijeong, http://cscp2.sogang.ac.kr/CSE4187_02/index.php/%ED%8C%8C%EC%9D%BC:8.png

Embedding: word2vec

Side effect: synonyms

“nice” ~ “beautiful”

“hard” ~ “difficult”

Side effect: word algebra

“king” - “man” + “woman” ~ “queen”

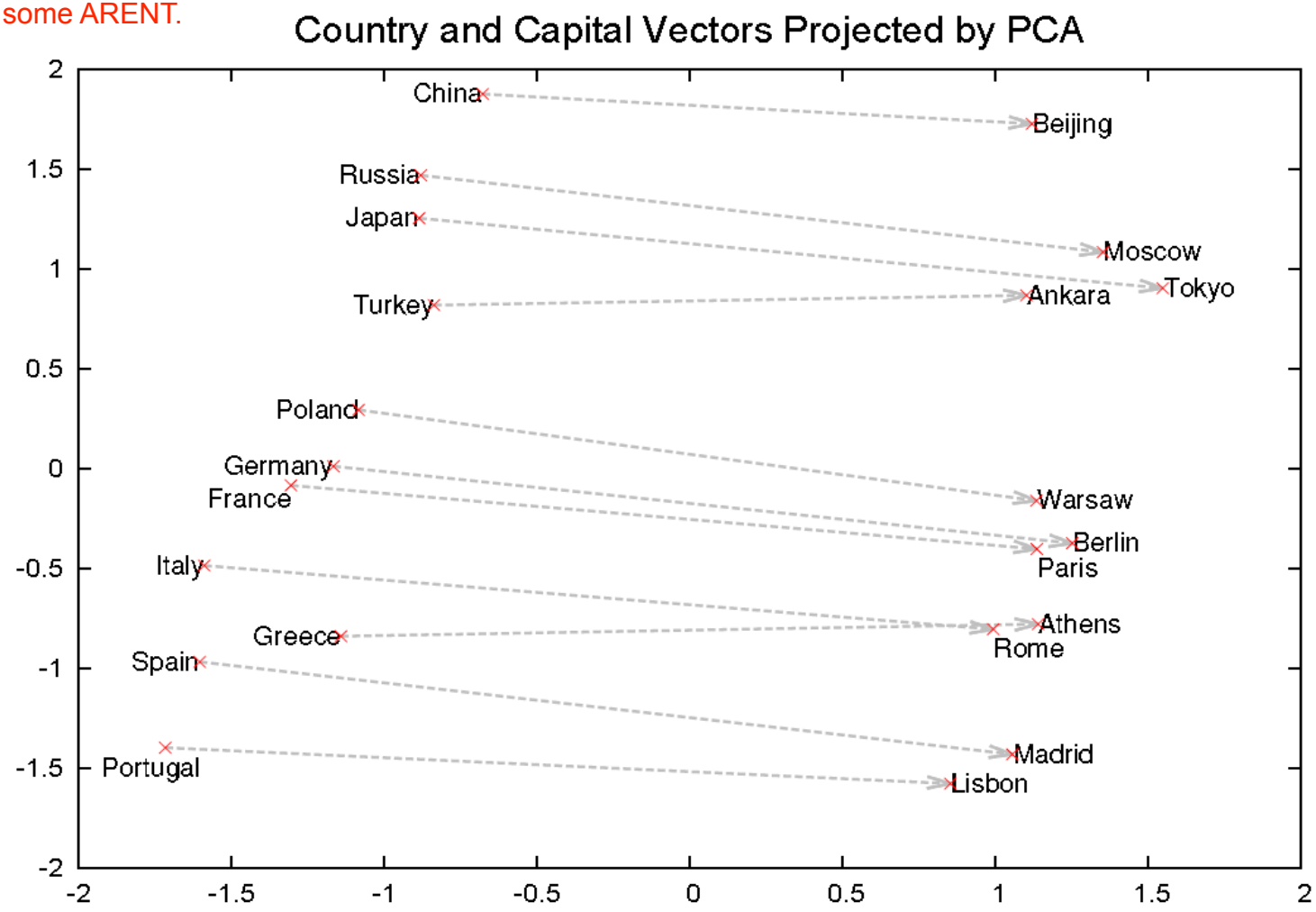
“moscow” - “russia” + “france” ~ “paris”

^^ Thsi is some cool shit

Embedding: word2vec

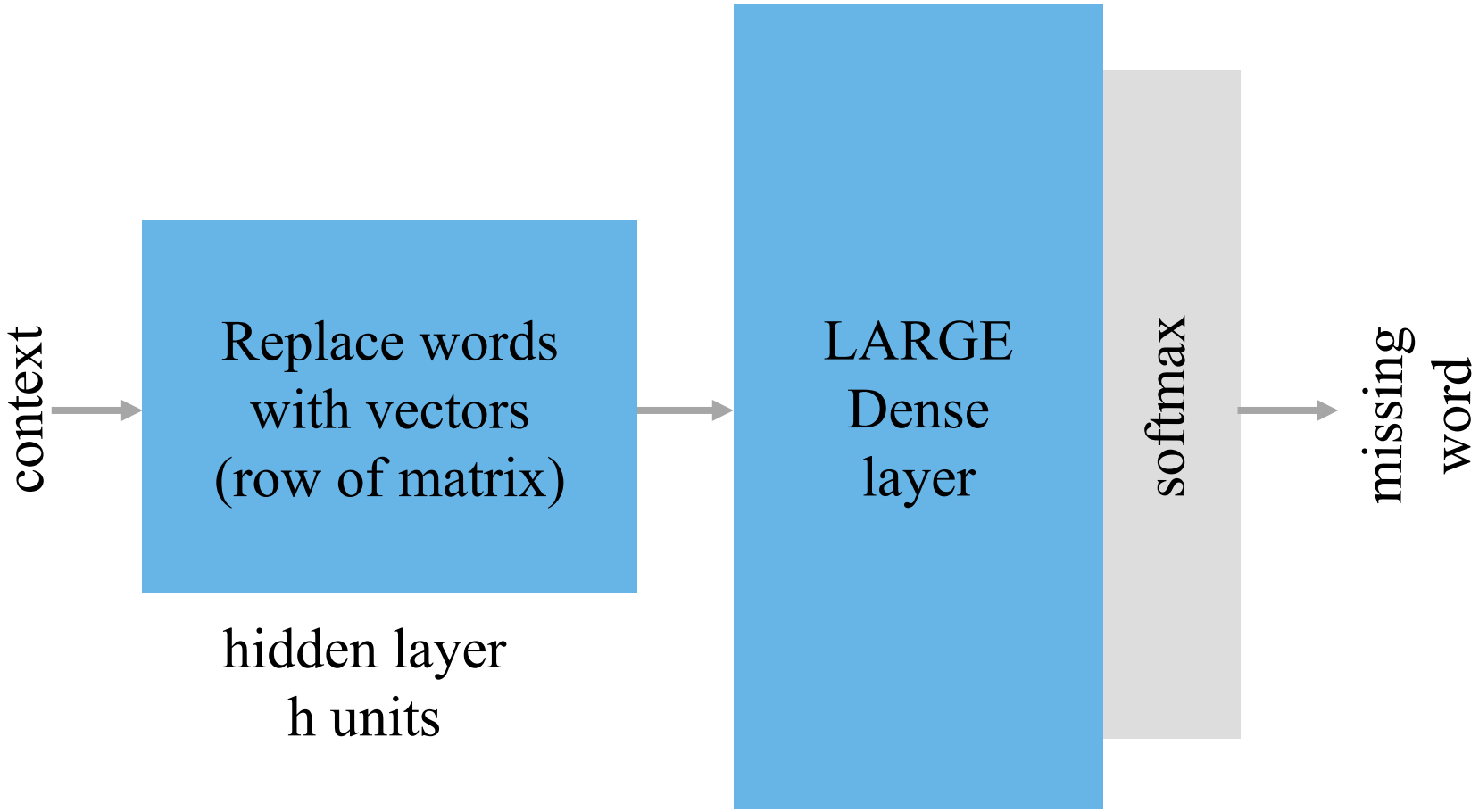
many unsupervised learning methods have interesting side effects. Some of them desirable, some ARENT.

Side effect: word algebra



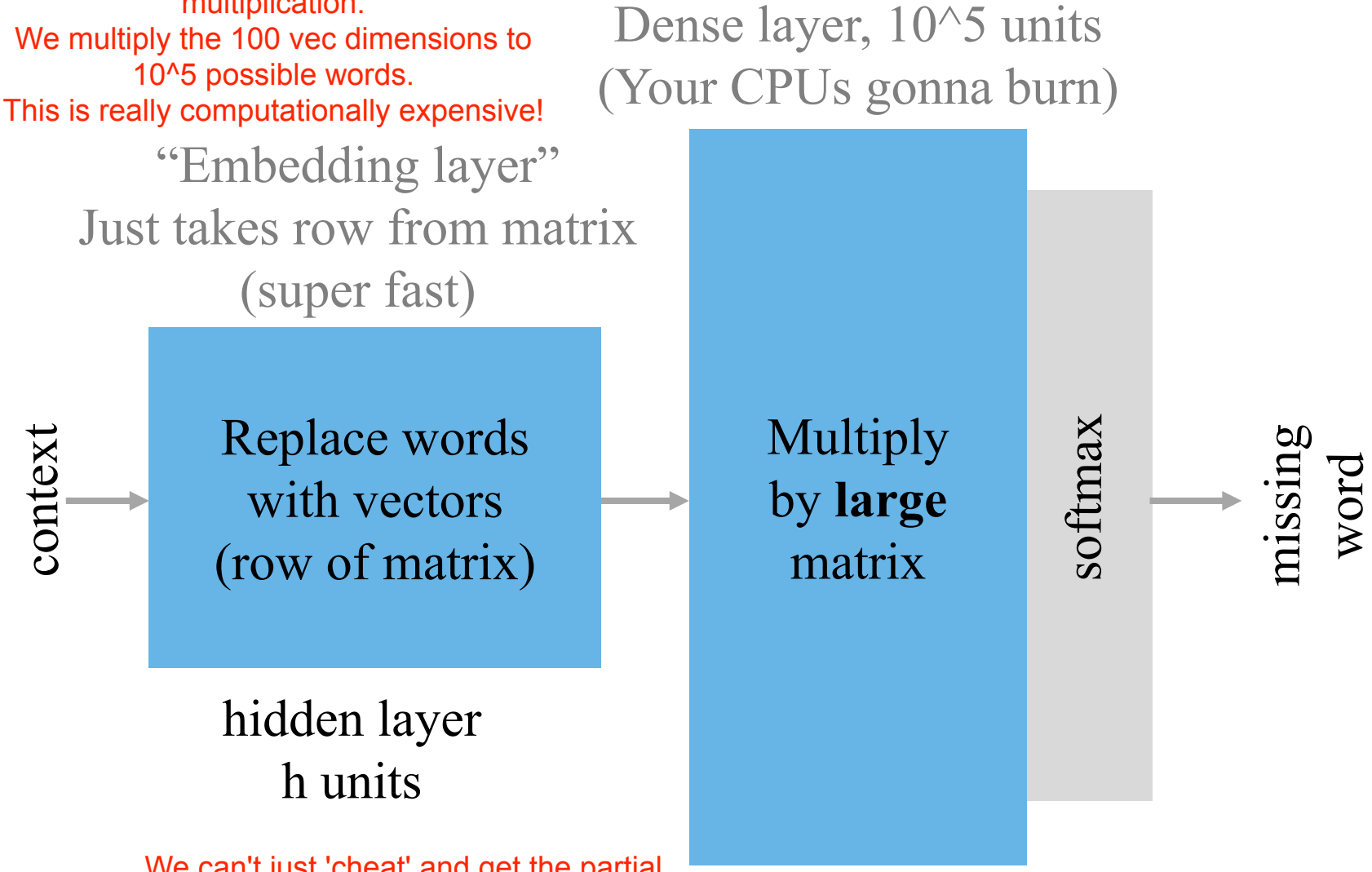
Softmax problem

Lets think about the engineering problem of training these word embedding models.
Now, for data, we can just get a corpus (available online), news, etc. and use the words there.



Softmax problem

A problem is that we have a large matrix multiplication.
We multiply the 100 vec dimensions to 10^5 possible words.
This is really computationally expensive!



We can't just 'cheat' and get the partial output just cos our context are 'one hot' inputs. We want to get the probabilities for every word being the missing word.

Also, softmax needs to know all of the logits

More word embeddings

Faster softmax:

Instead we use a 'faster' softmax,
or use other models.

- Hierarchical softmax, negative samples, …
- learn more

Alternative models: GloVe

Sentence level:

- Doc2vec, skip-thought (using rnn)

To be continued...
in the NLP course