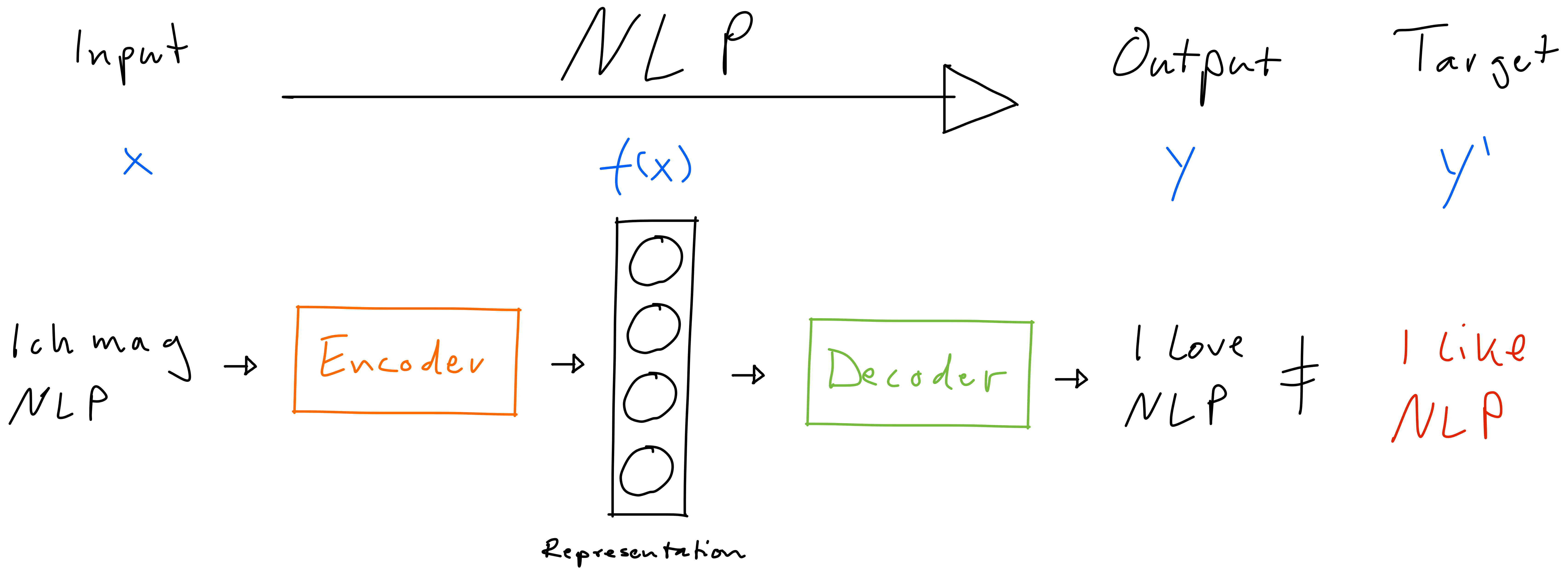
Word Embeddings and Topic Modelling

Tim Rocktäschel & Sebastian Riedel

COMP0087 Natural Language Processing

NLP in a Nutshell



- Sentences
- Documents
- Languages
- Domains
- Databases

- Manual
- Learnt
- Pretrained
- RNN
- CNN

- symbolic
- sparse
- dense
- attention

- softmax
- Sequence decoder
- tree decoder

- labels
- sentences
- trees
- graphs

- supervised
- self supervised
- weak
- unsup.
- semi-sup.

Classify This!

Train
blah blah football → A
blah blah stocks → B

Test
blah blah hockey → ?

Machine Sees this

Train

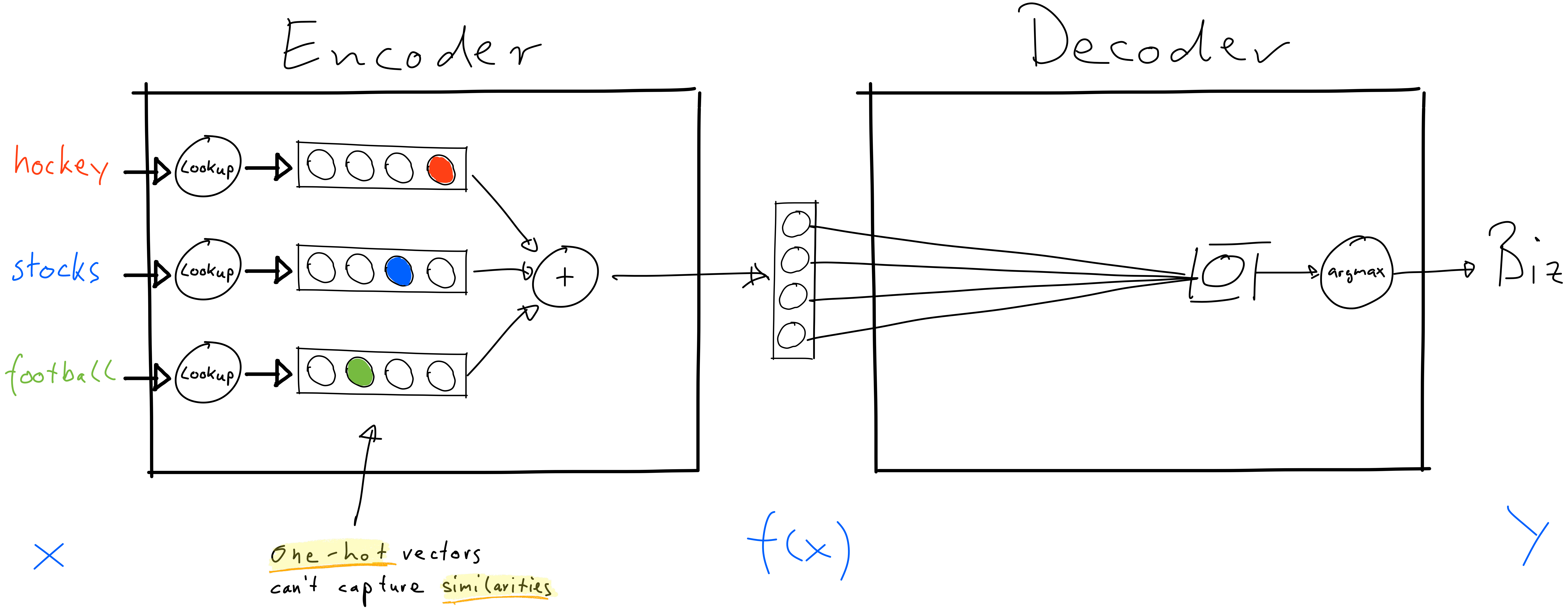
blah blah symbol1 → A

blah blah symbol23 → B

Test

blah blah symbol42 → ?

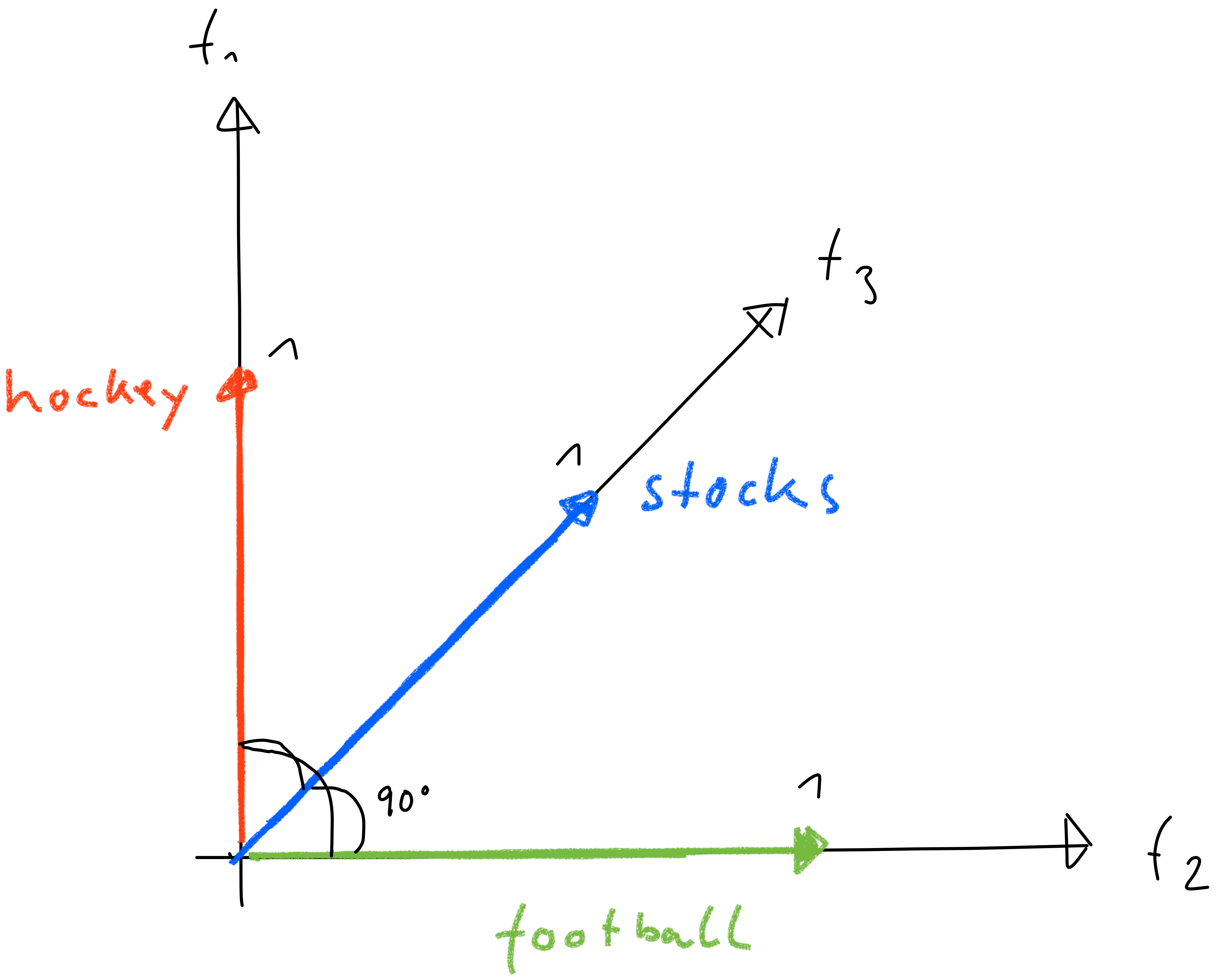
Word Representations



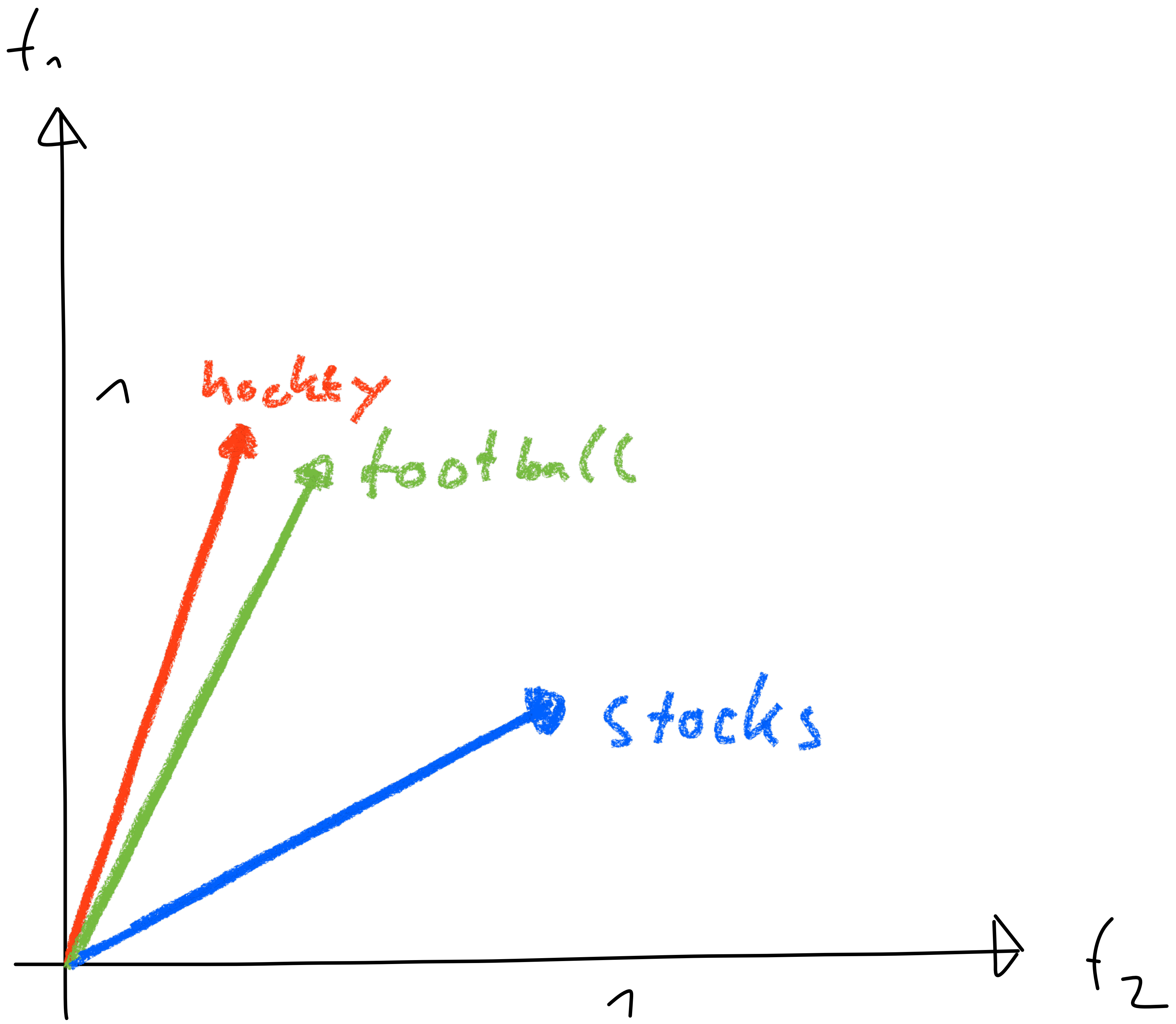
One-hot vectors can't capture similarities

Fixing one hot vectors

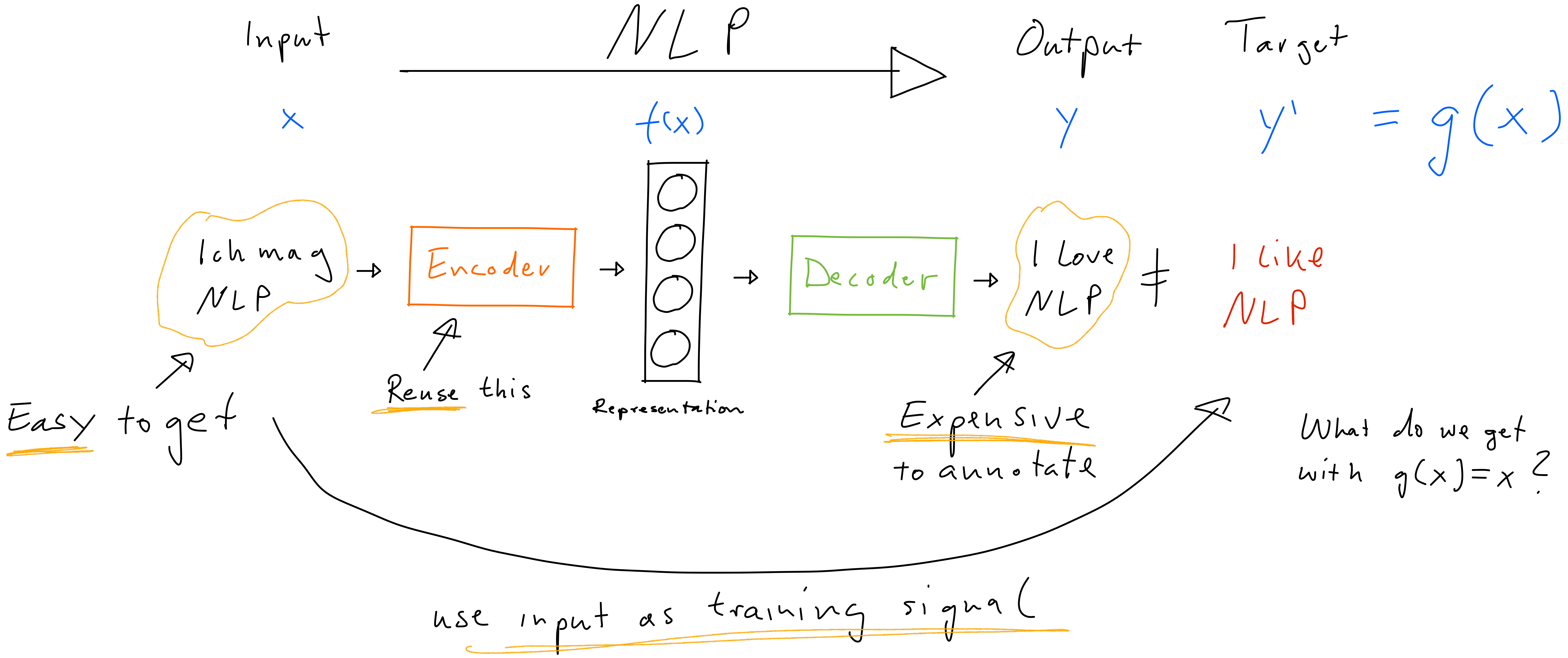
we have this



but we want this



NLP in a Nutshell

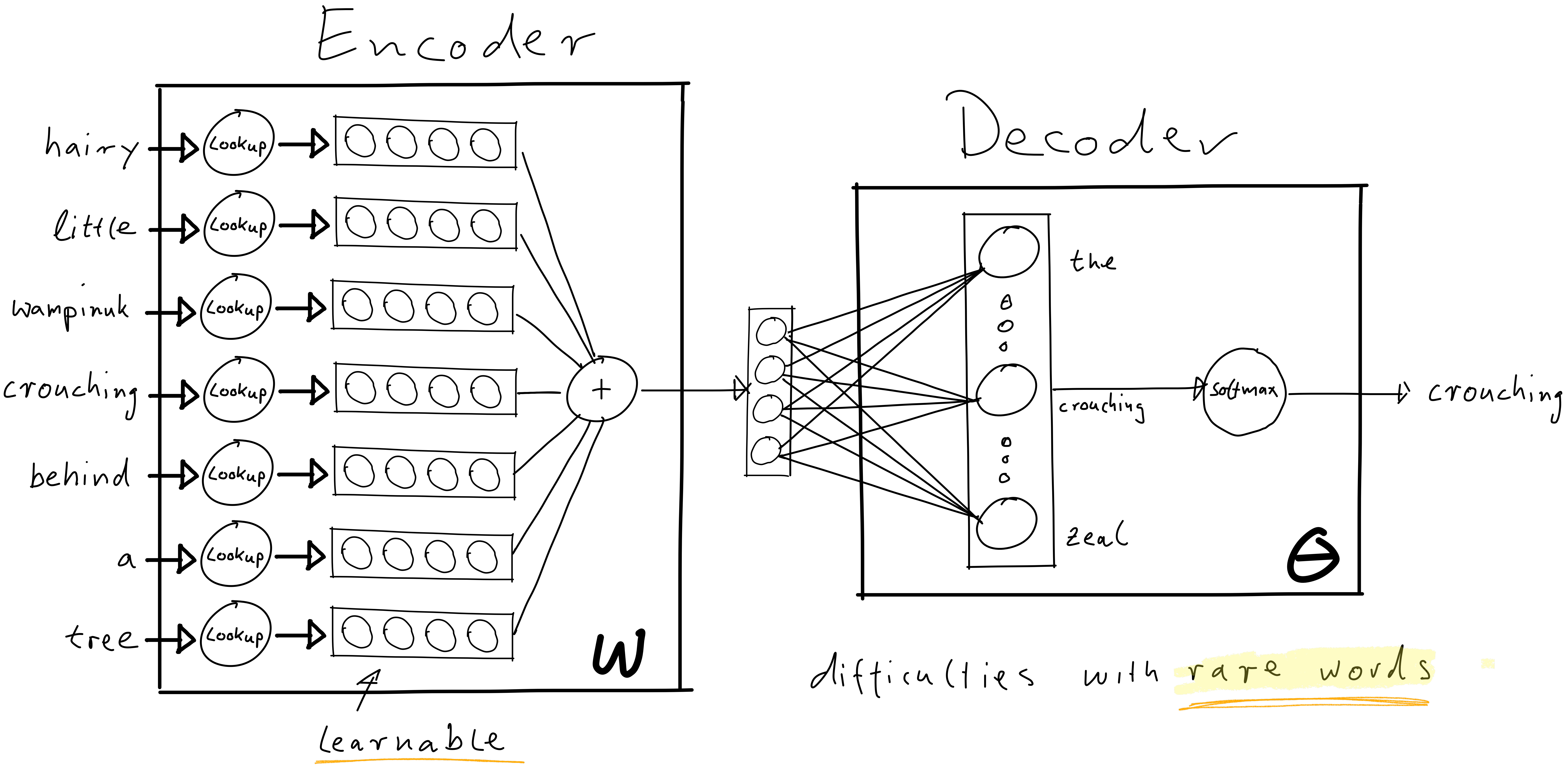


Wampinuk

Marco saw a hairy little **wampinuk** crouching behind a tree

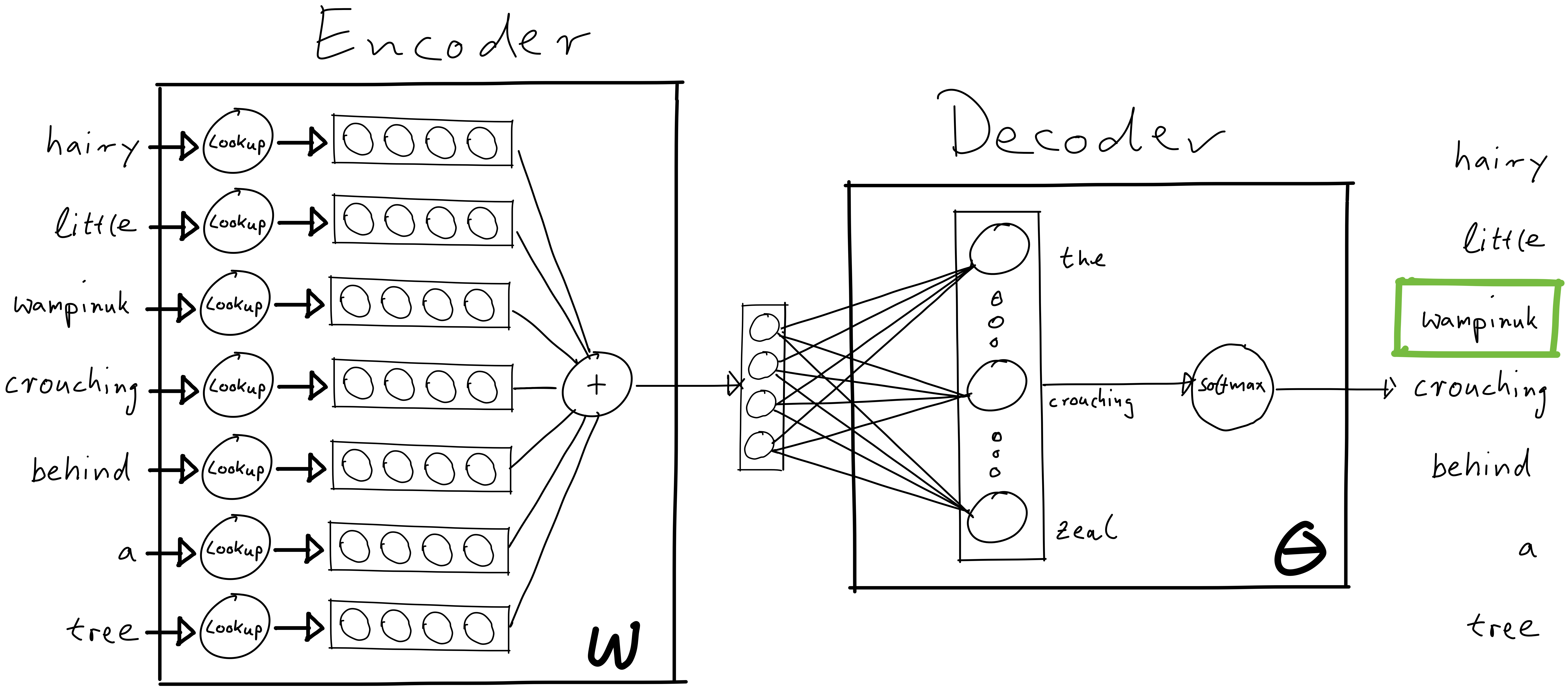
You shall know aword by the company it keeps (Firth, J. R. 1957:11)

Word2Vec: Continuous Bag of Words

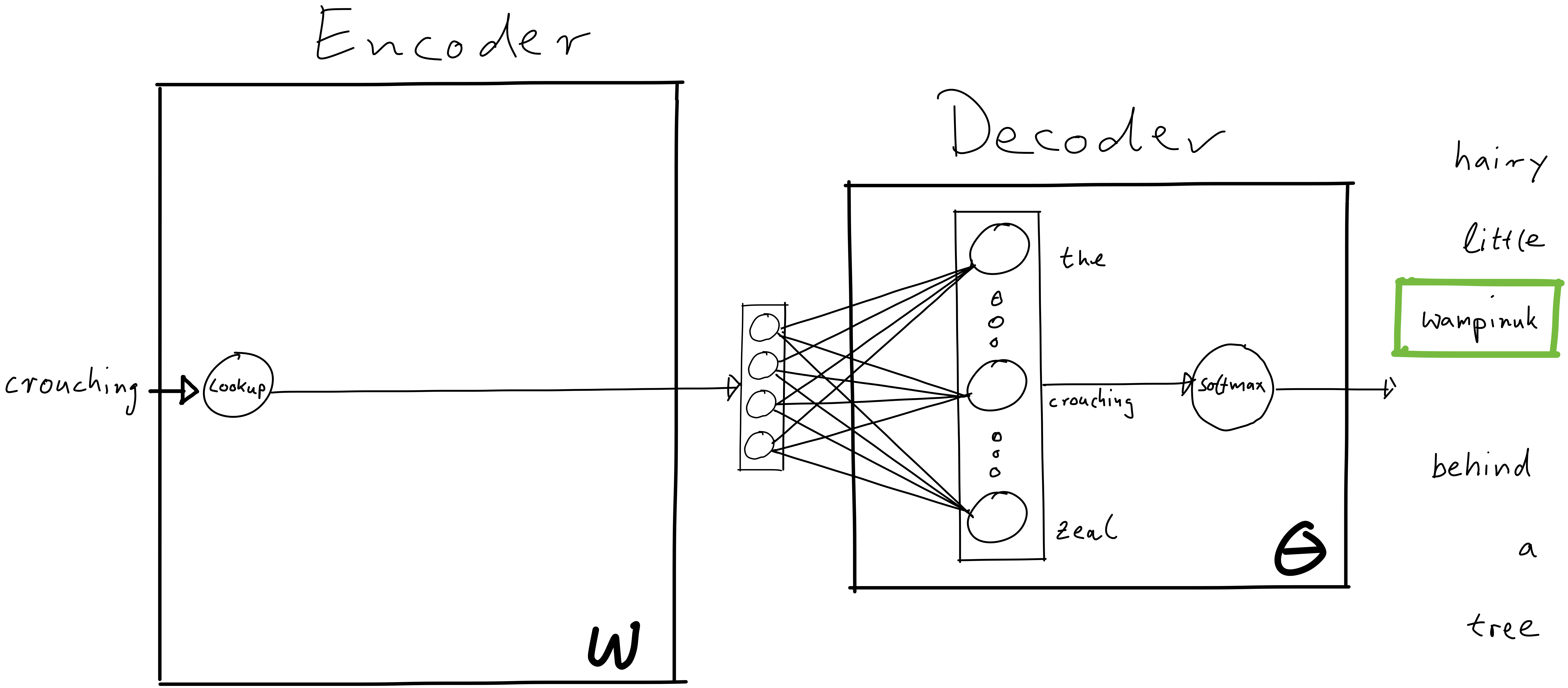


difficulties with rare words

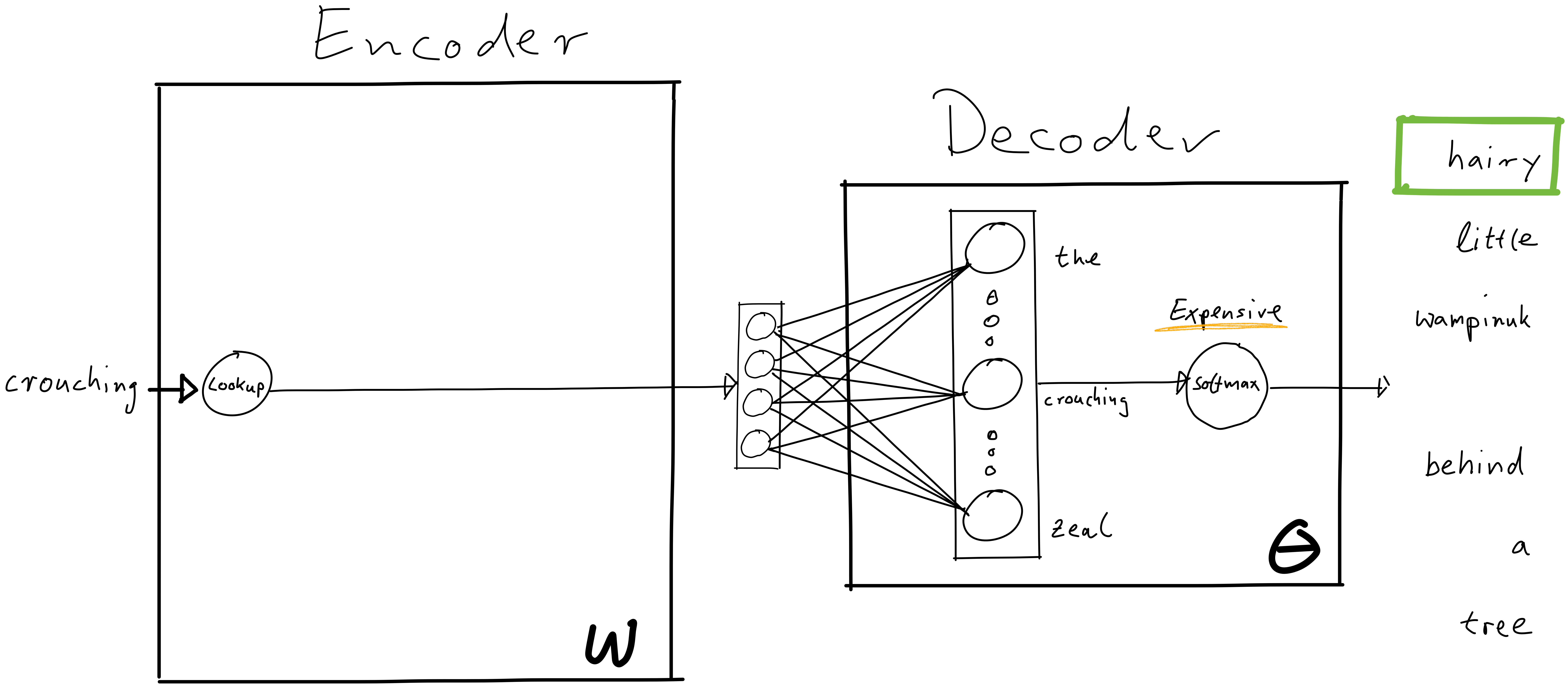
Word2Vec: Skip Gram



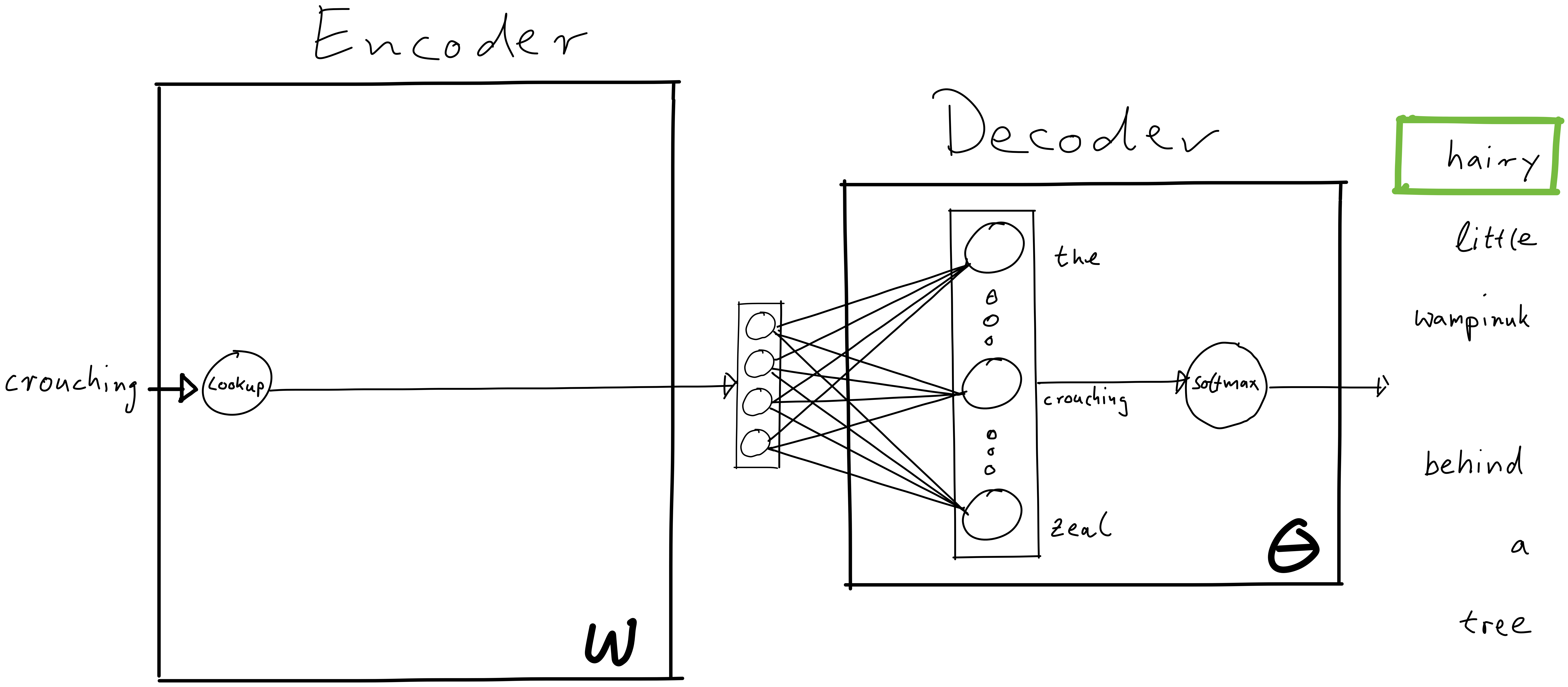
Skip Gram


Encoder
Decoder
crouching
Lookup
W
the
crouching
zeal
softmax
θ
hairy
little
wampinuk
behind
a
tree

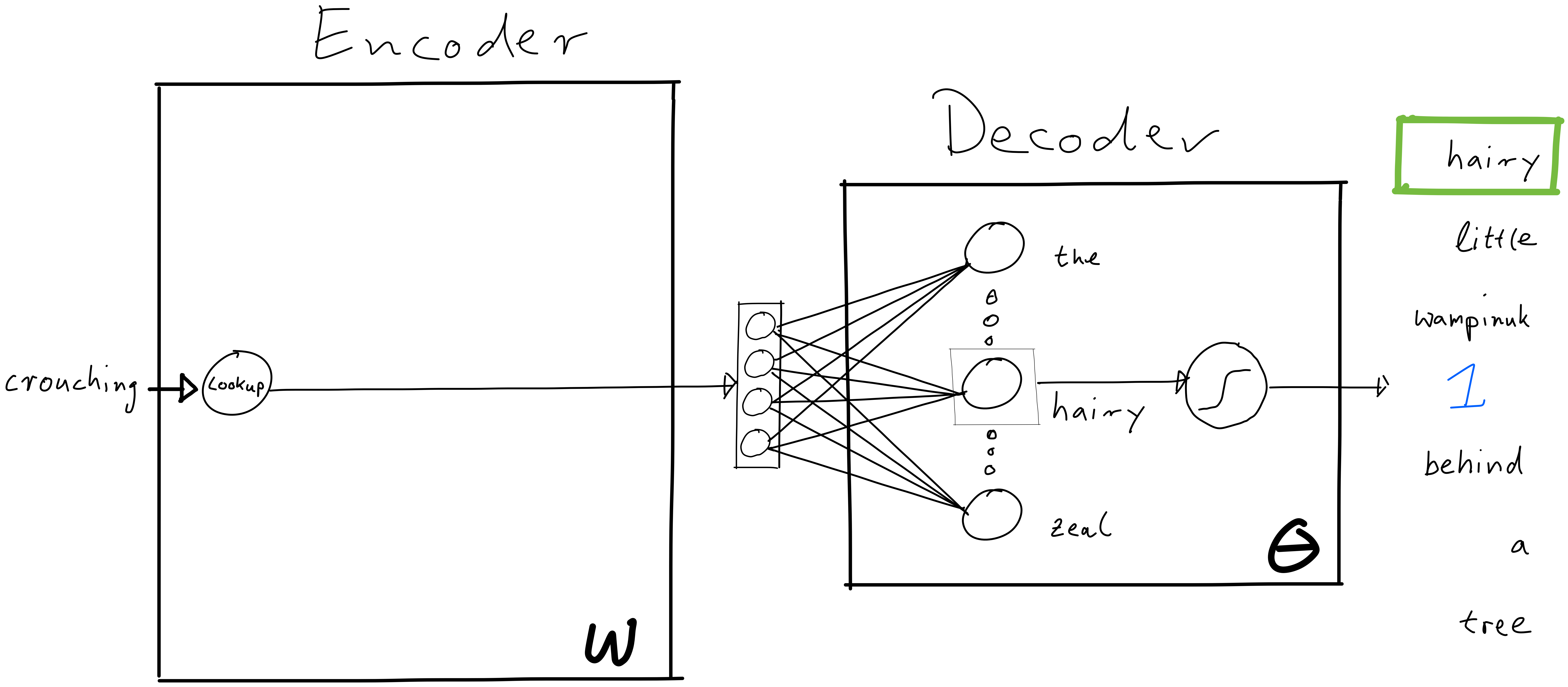
Skip Gram


Encoder
crouching
Lookup
W
Decoder
the
crouching
zeal
Expensive
softmax
Θ
hairy
little
wampinuk
behind
a
tree

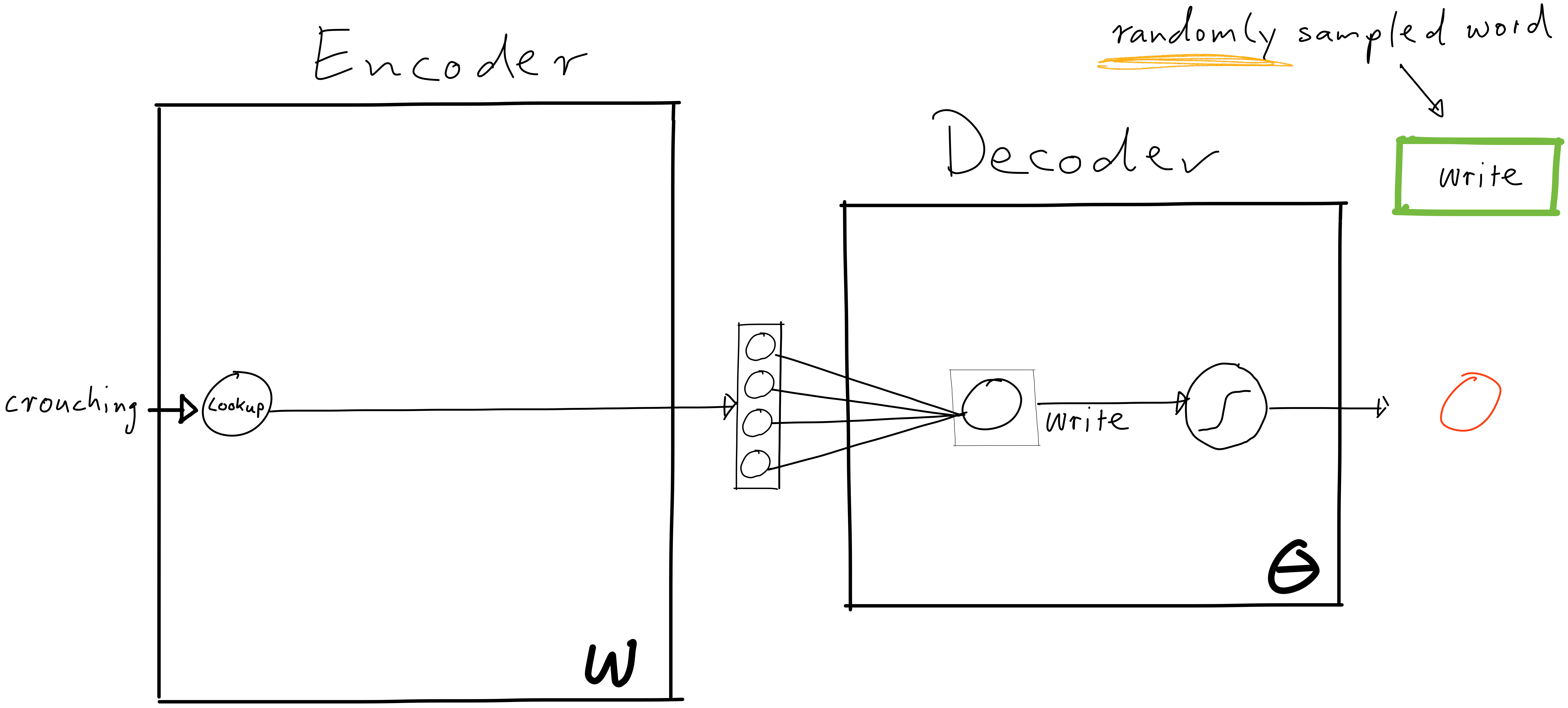
Binary Classifier



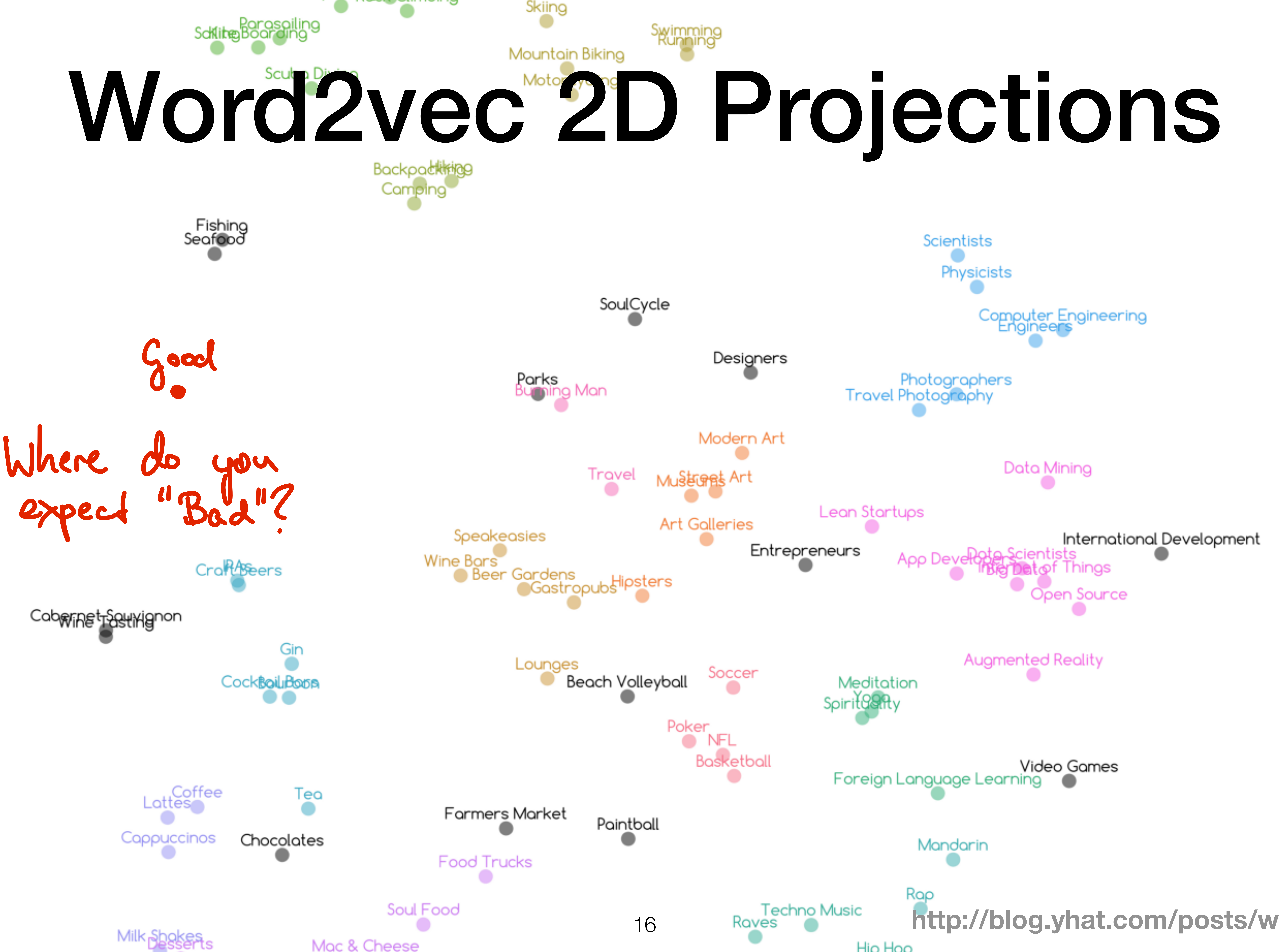
Binary Classifier



Negative Sampling



Word2vec 2D Projections

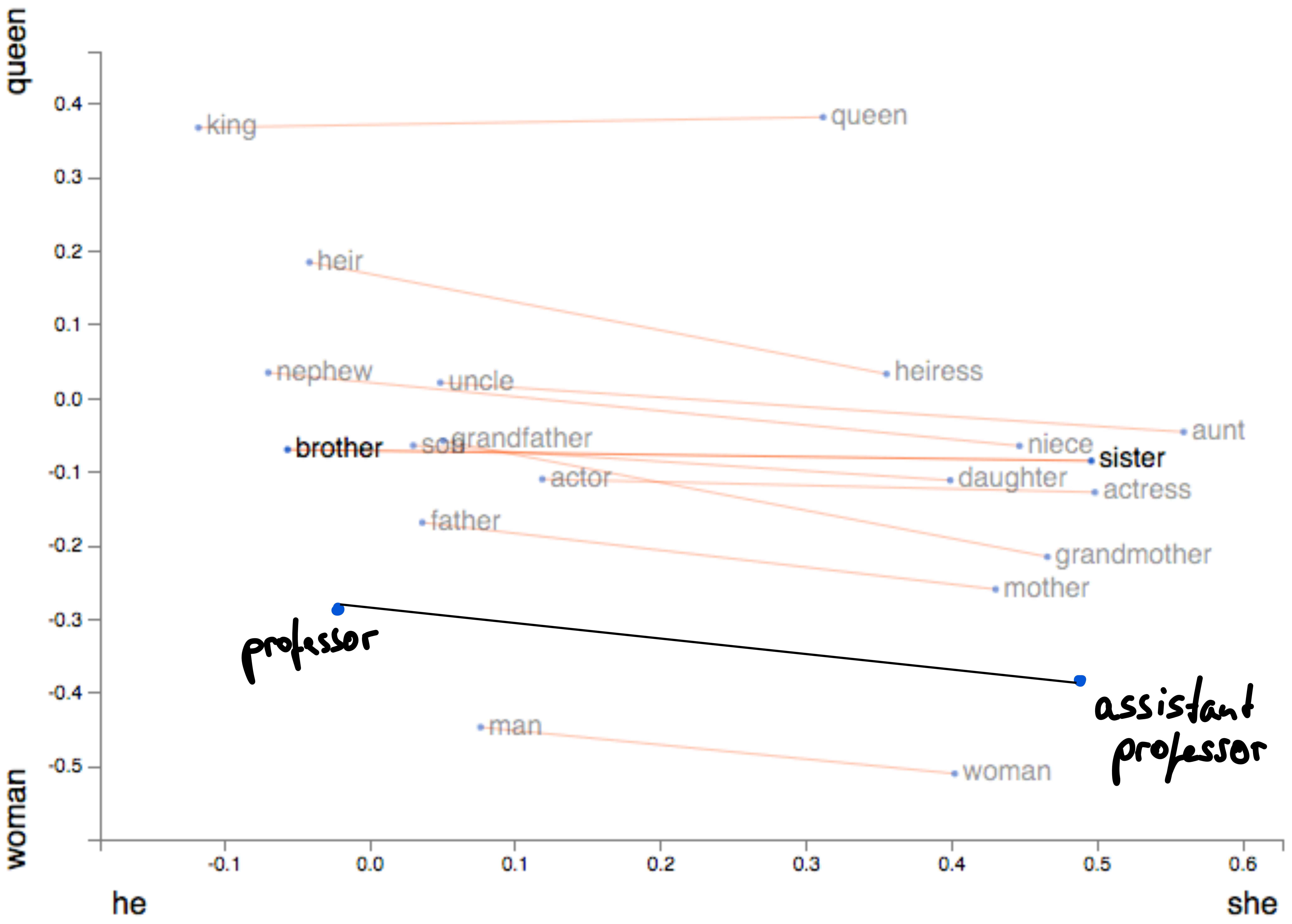


Good

Where do you expect "Bad"?

http://blog.yhat.com/posts/words2map.html

King - Man + Woman ≈ Queen



Be aware of biases in the data!

Classify This!

Train

blah blah football → A

blah blah stocks → B

Test

blah blah hockey → ?

Cluster This!

blah blah football → ?

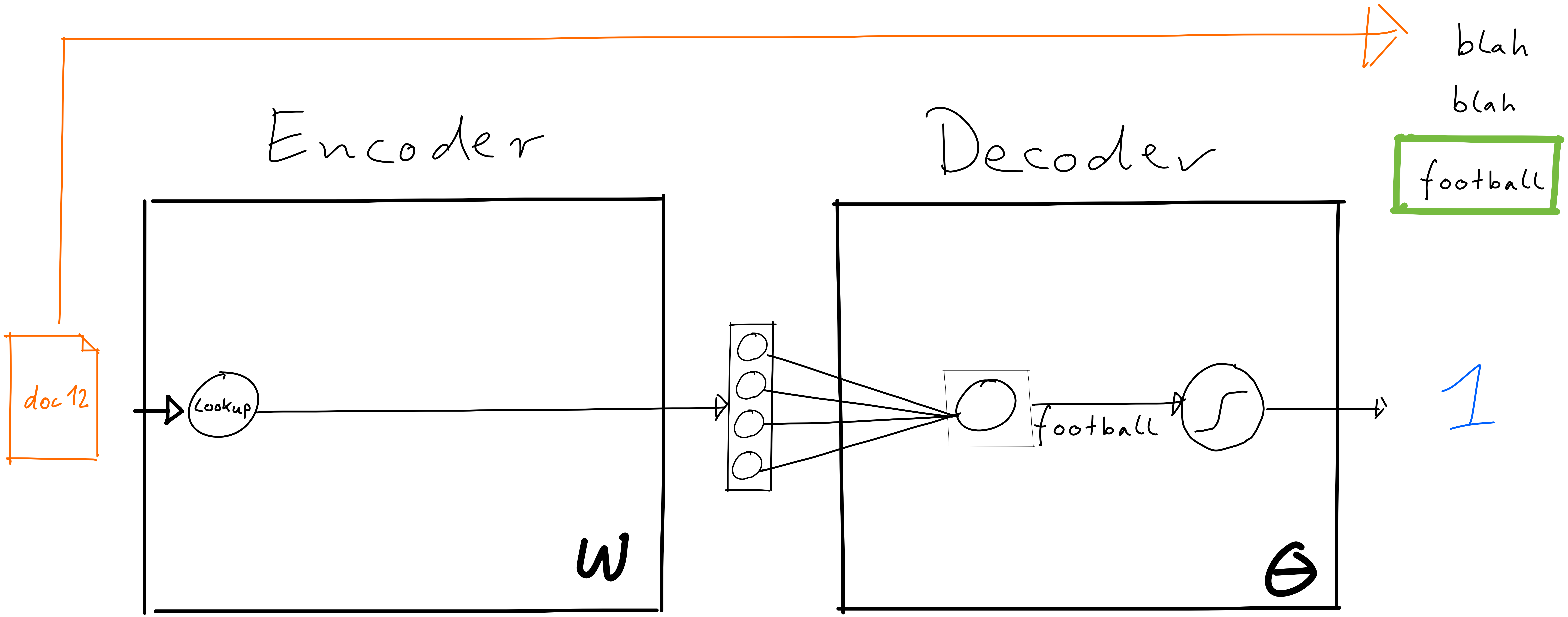
blah blah stocks → ?

Test

blah blah hockey → ?

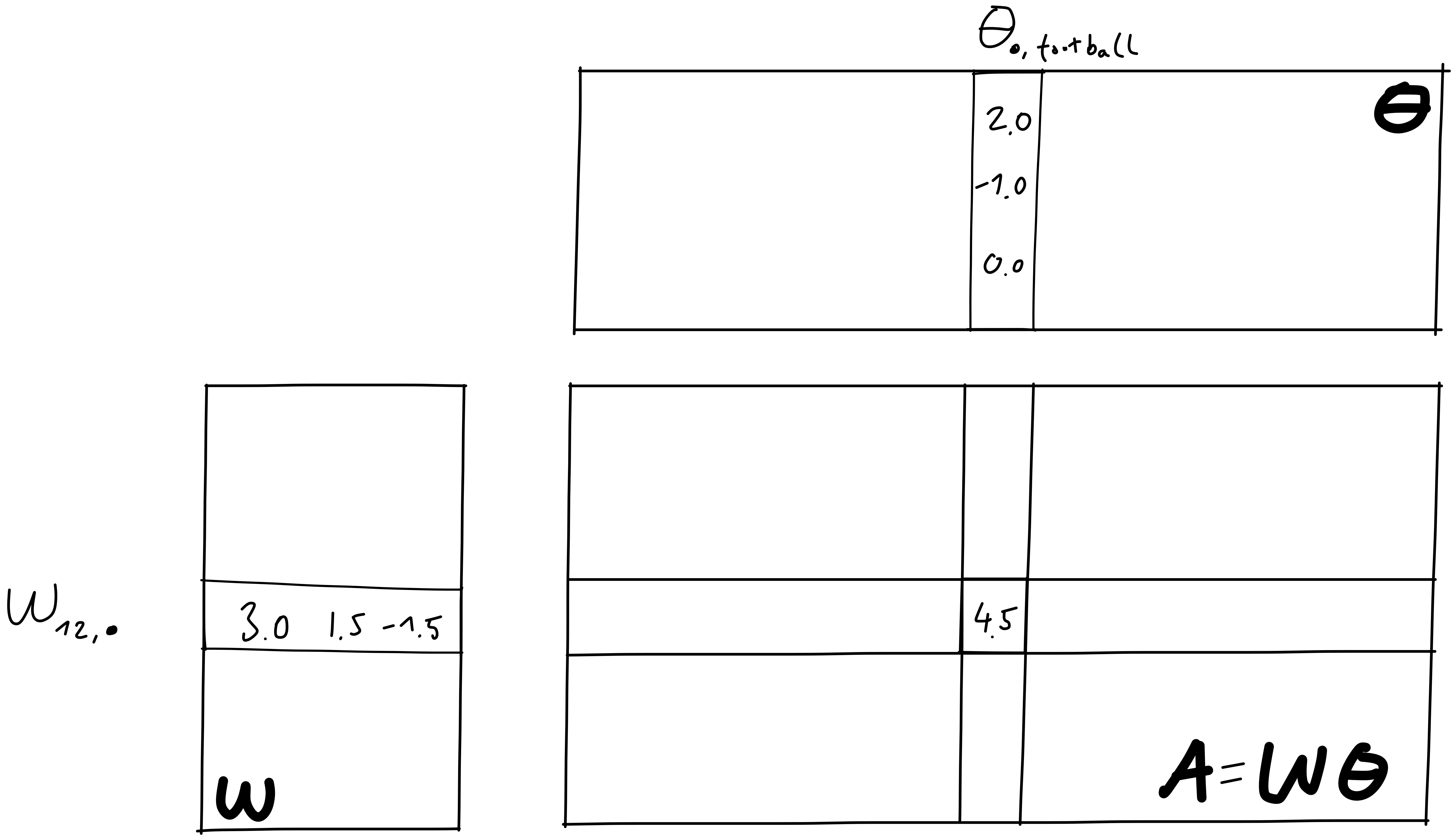
Why useful?

Topic Modelling as Self Supervision



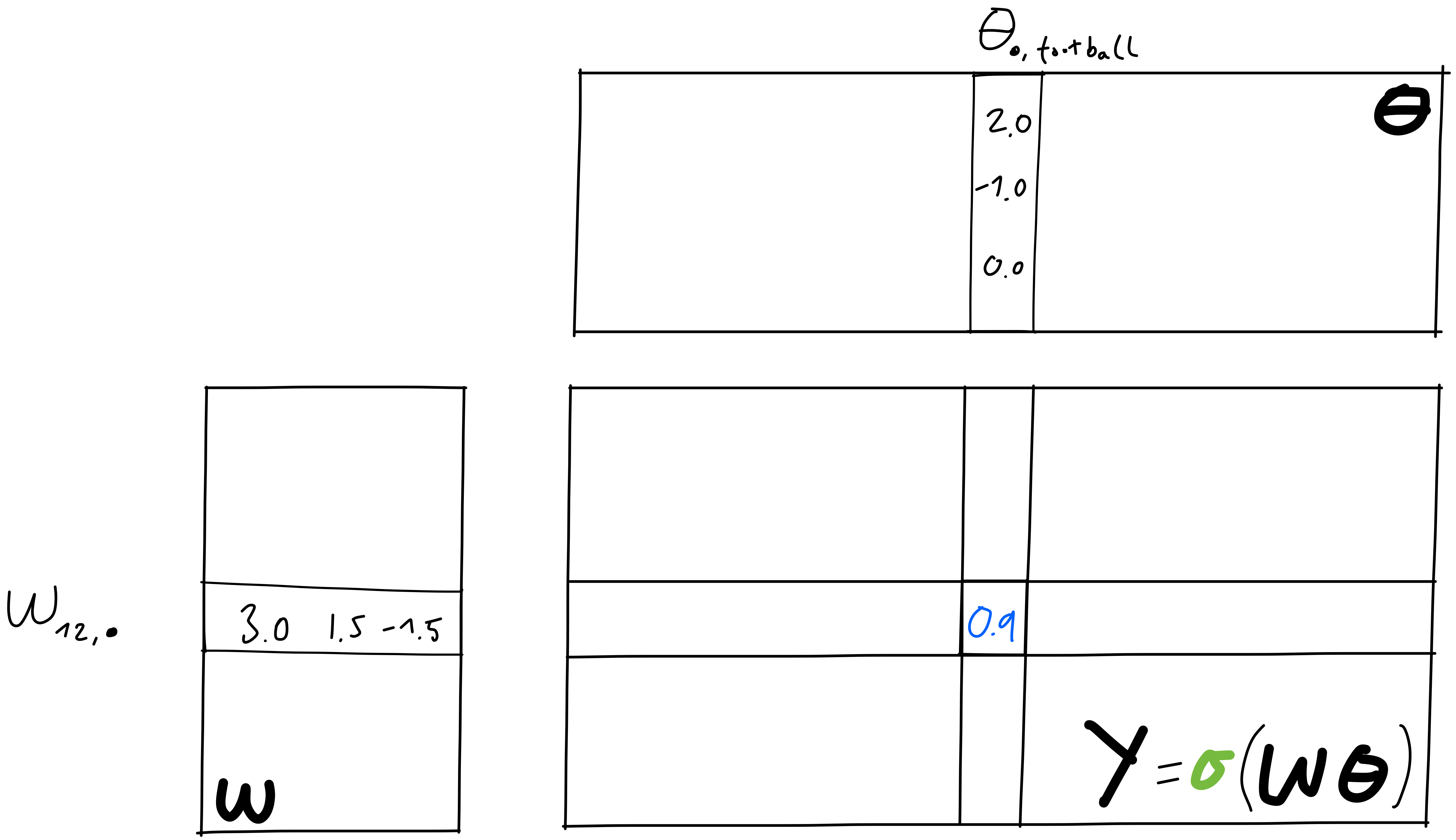
X represent the doc such that the "football" classifier says 1 and the "stocks" classifier 0 Y'

Matrix View



$$A_{12, football} = W_{12,\bullet} \Theta_{\bullet, football}$$

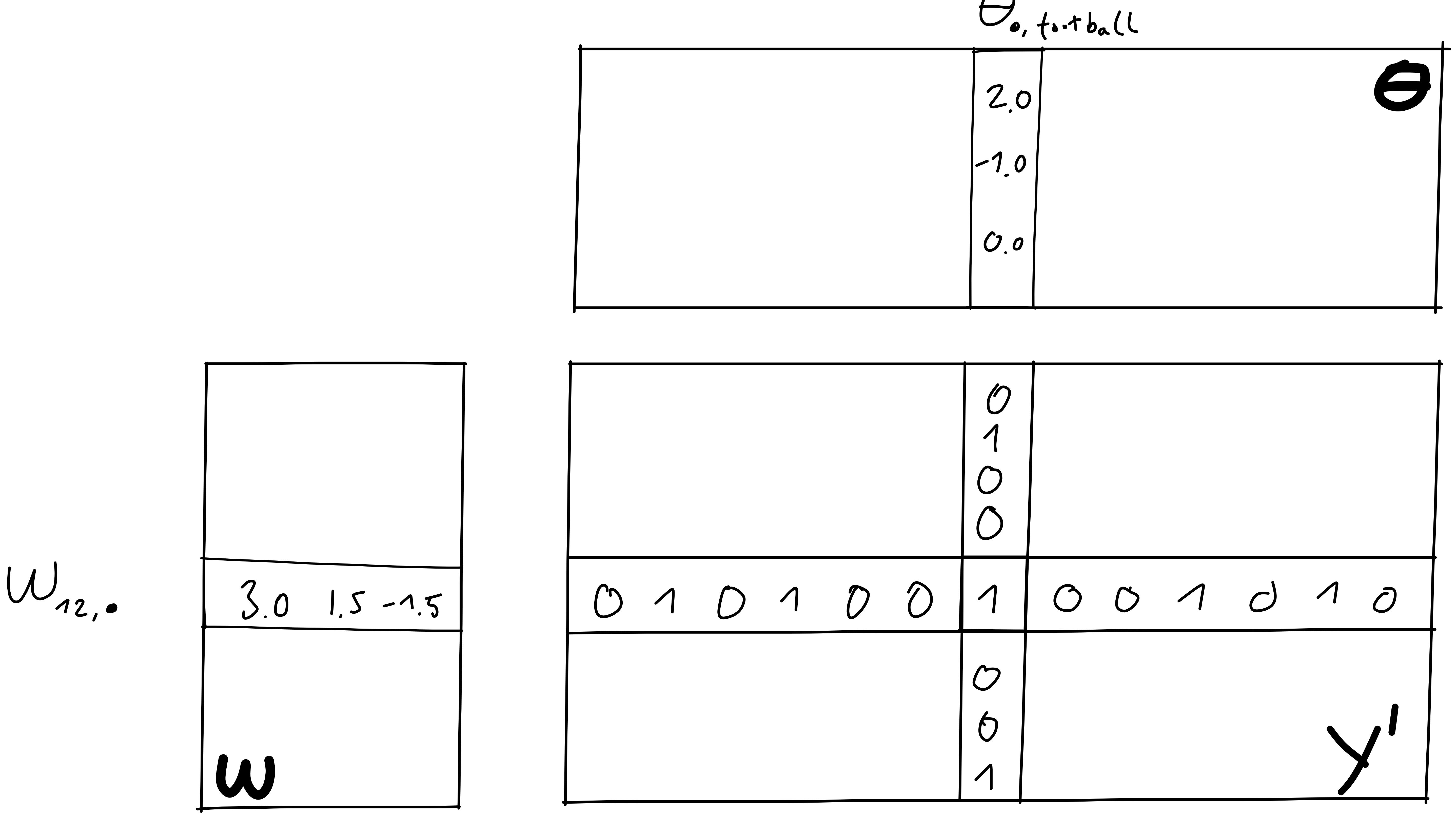
Matrix View



$$A_{12,\,\text{football}} = W_{12,\bullet}\,\theta_{\bullet,\,\text{football}}$$

$$Y_{12,\,\text{football}} = \sigma\left(A_{12,\,\text{football}}\right)$$

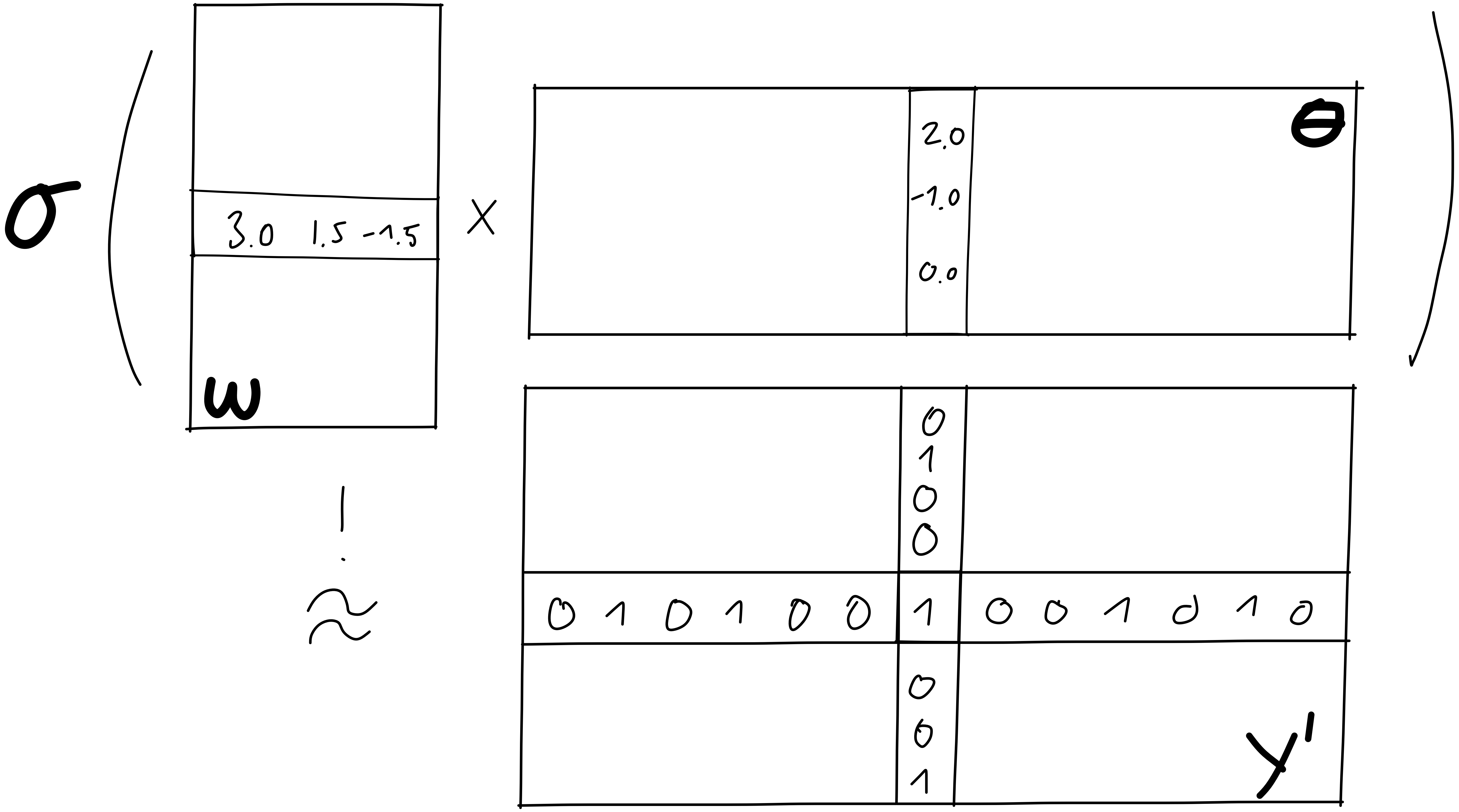
Term Document Matrix



$$A_{12,\text{football}} = W_{12,\bullet}\,\Theta_{\bullet,\text{football}}$$

$$Y_{12,\text{football}} = \sigma\left(A_{12,\text{football}}\right)$$

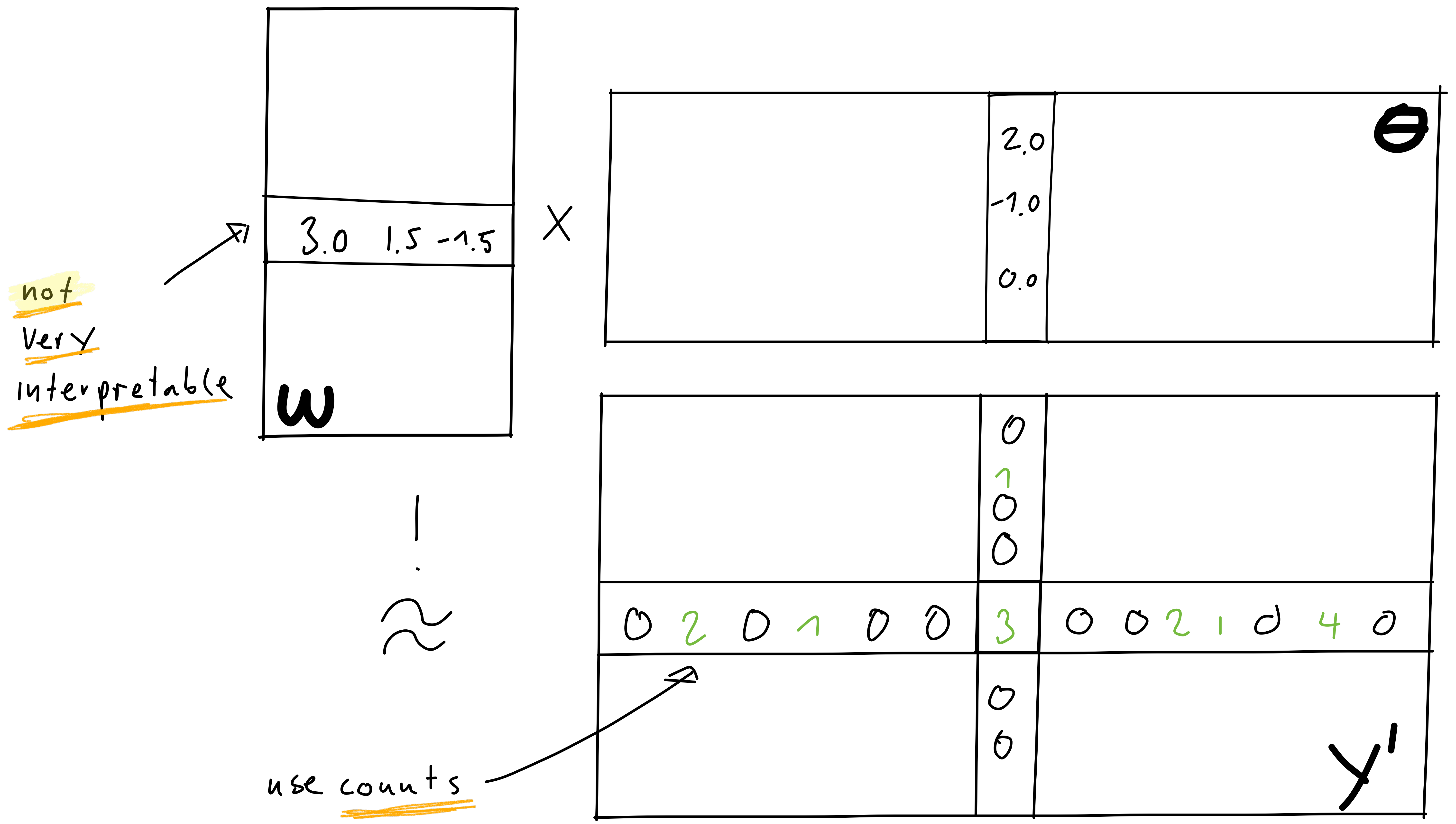
Binary Matrix Factorization



$$A_{12,\,\text{football}} = W_{12,\bullet}\, \Theta_{\bullet,\,\text{football}}$$

$$Y_{12,\,\text{football}} = \sigma(A_{12,\,\text{football}}) \overset{!}{\approx} Y'_{12,\,\text{football}}$$

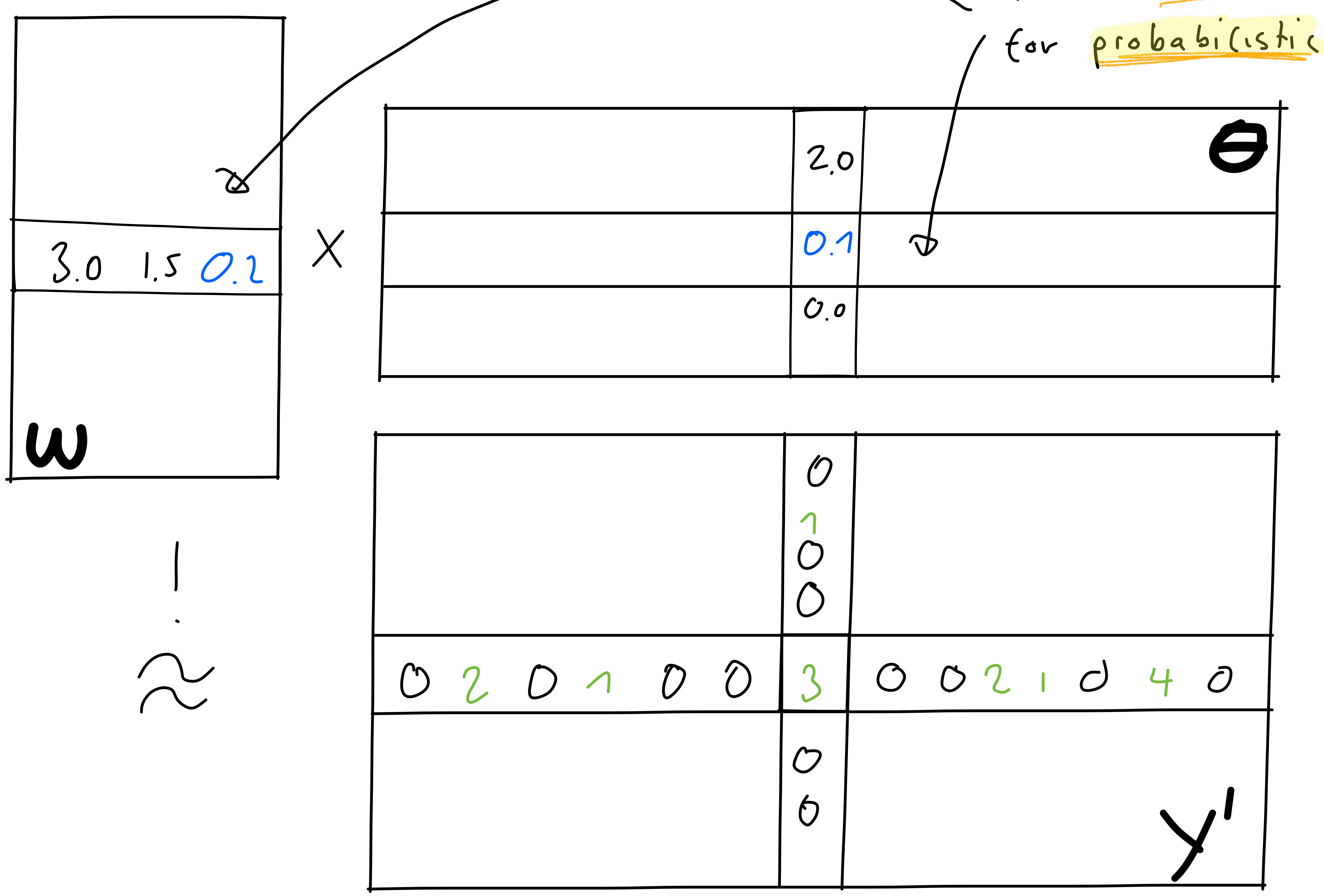
Matrix Factorization



$$A_{12,\,\text{football}} = W_{12,\bullet}\,\theta_{\text{football}}$$

$$\overset{!}{\approx} Y'_{12,\,\text{football}}$$

Nonnegative Matrix Factorization



Force normalization for probabilistic Topic Models (e.g. LDA)

$$A_{12,\text{football}} = W_{12,\bullet}\,\Theta_{\bullet,\text{football}} \overset{!}{\approx} Y'_{12,\text{football}}$$

$$\Theta \geq 0$$

$$W \geq 0$$

Enforce these constraints during training

References

- Word Embeddings
  - J&M Chapter 6
  - Goldberg Chapter 5
  - Efficient Estimation of Word Representations in Vector Space, Mikolov et al, ICLR Workshop 2013
  - GloVe: Global Vectors for Word Representation, Pennington et al., EMNLP 2014
- Topic Models
  - An Introduction to Latent Semantic Analysis, TK Landauer
  - Probabilistic Latent Semantic Analysis, T Hofmann
  - Probabilistic Topic Models, Blei
  - Exploring Topic Coherence over Many Models and Many Topics, K Stevens, EMNLP 2012