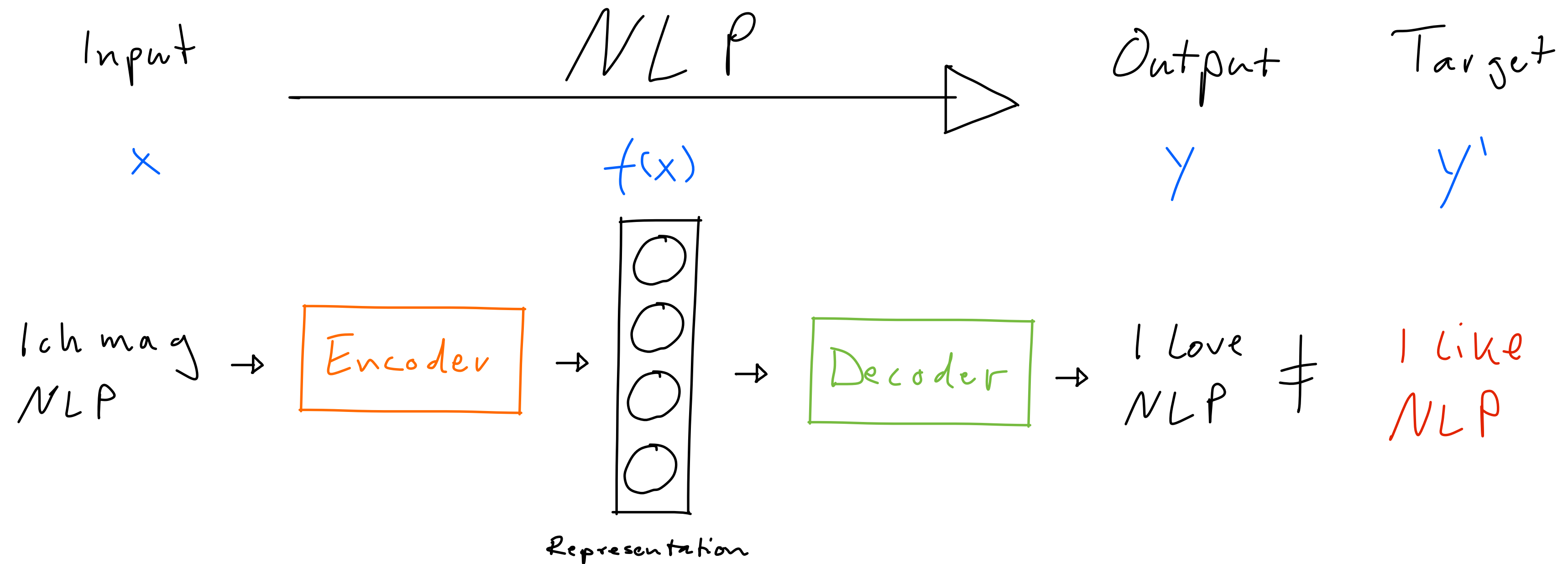


# Word Embeddings and Topic Modelling

Tim Rocktäschel & Sebastian Riedel  
COMP0087 Natural Language Processing



# NLP in a Nutshell



- |             |              |             |                    |             |                   |
|-------------|--------------|-------------|--------------------|-------------|-------------------|
| · Sentences | · Manual     | · symbolic  | · softmax          | · labels    | · supervised      |
| · Documents | · Learnt     | · sparse    | · Sequence decoder | · sentences | · self supervised |
| · Languages | · Pretrained | · dense     | · tree decoder     | · trees     | · weak            |
| · Domains   | · RNN        | · attention |                    | · graphs    | · unsup.          |
| · Databases | · CNN        |             |                    |             | · semi-sup.       |

# Classify This!

# Train

b l a h

$$b(a, h)$$

football

 $\rightarrow A$ 

b l a h

$$b(a, h)$$

stocks

$$\rightarrow \beta$$

Test

blah

blah

# hockey

$\rightarrow \mathbb{Z}$

# Machine Sees this

# Train

$b(a \cdot h)$     $b(a \cdot h)$    symbol

$\rightarrow A$

b(a h      b(a h      symbo(23

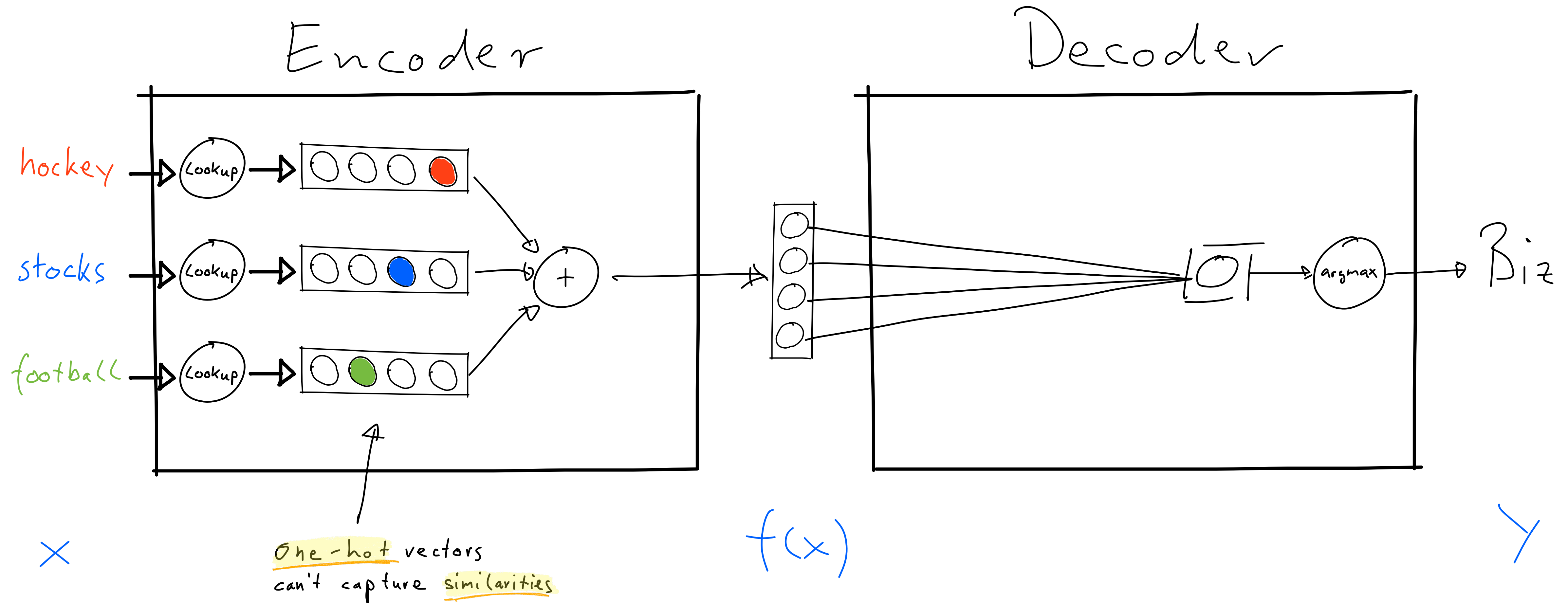
$$\rightarrow \beta$$

Test

blah blah symbol(42)

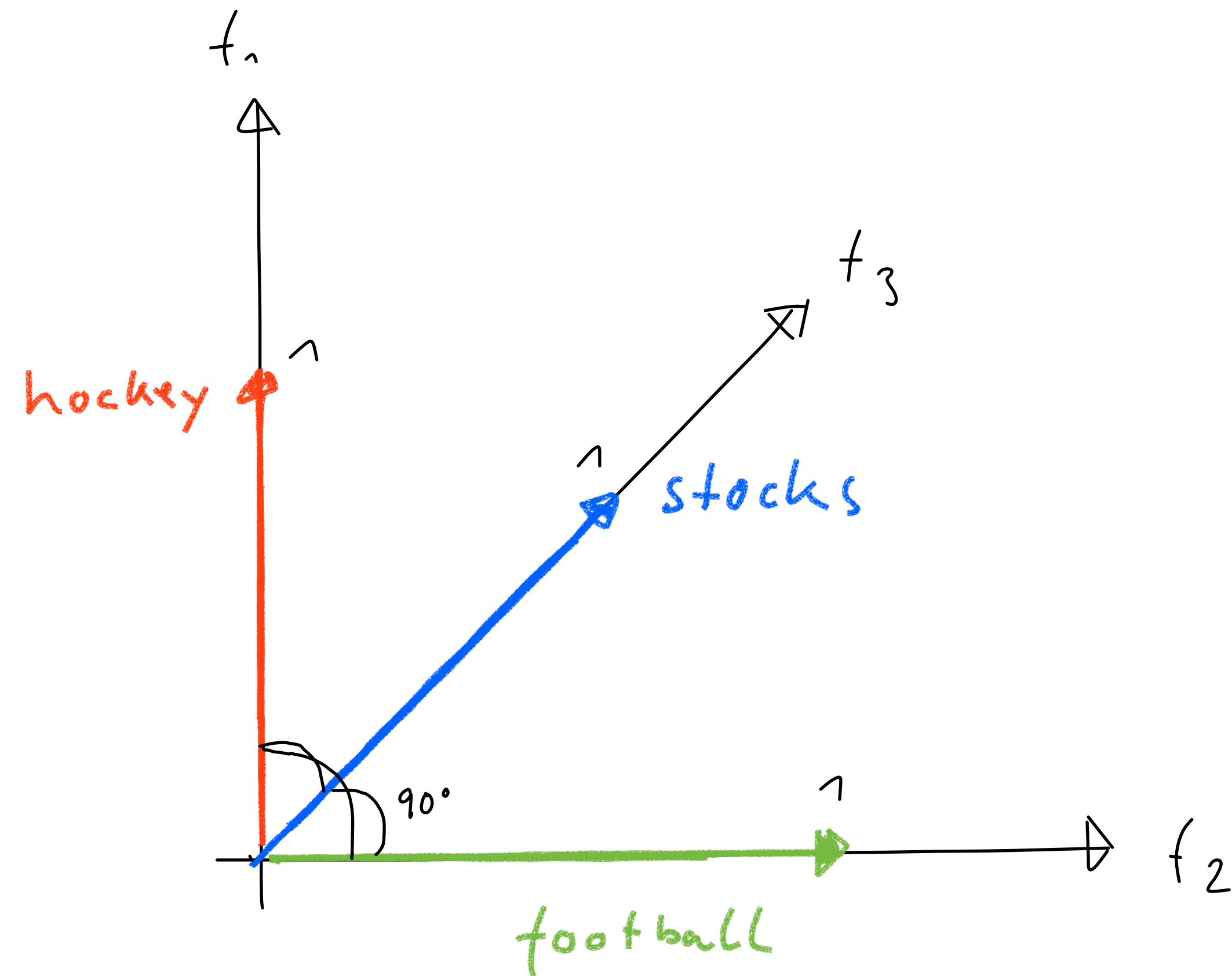
→ 2

# Word Representations

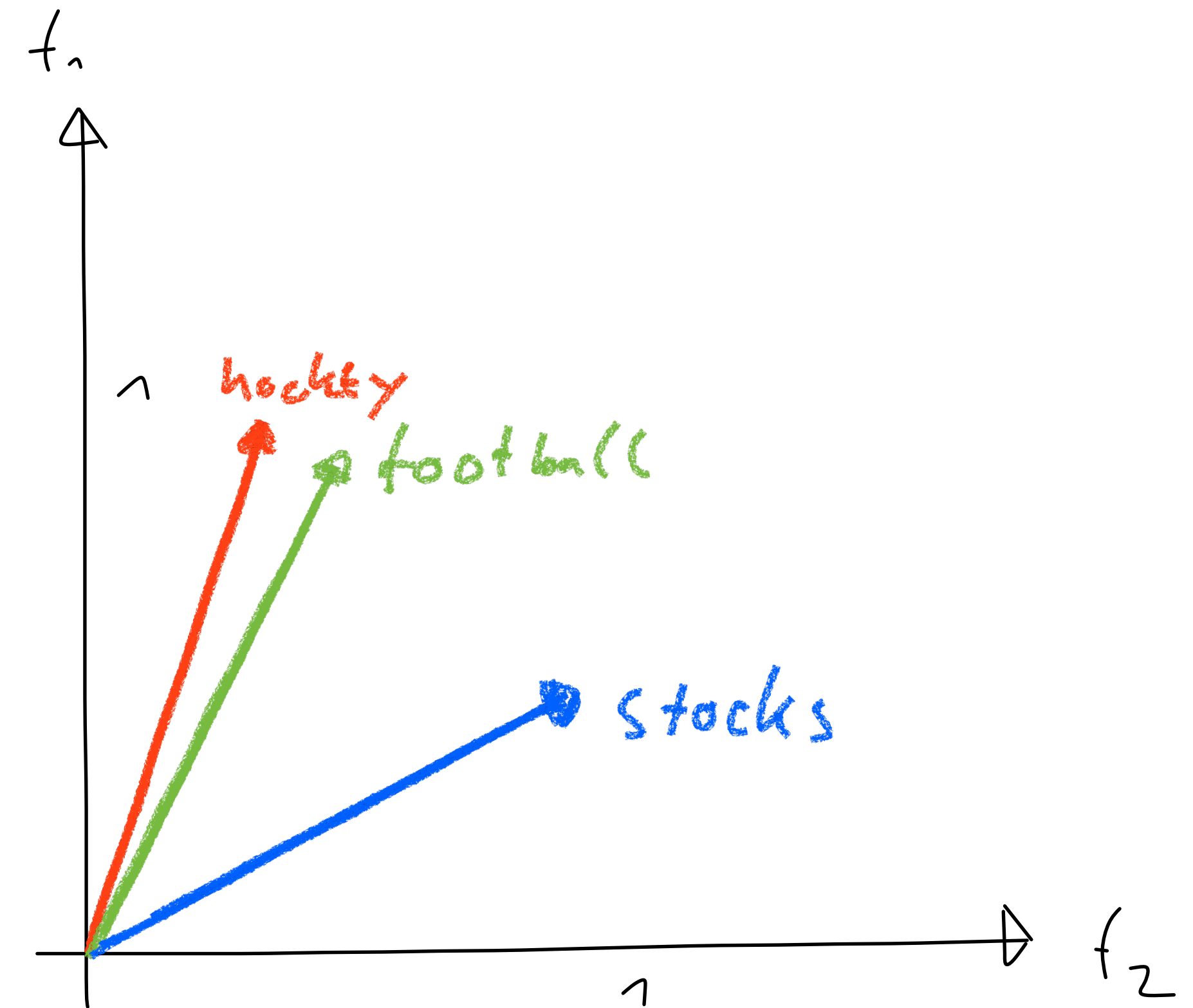


# Fixing one hot vectors

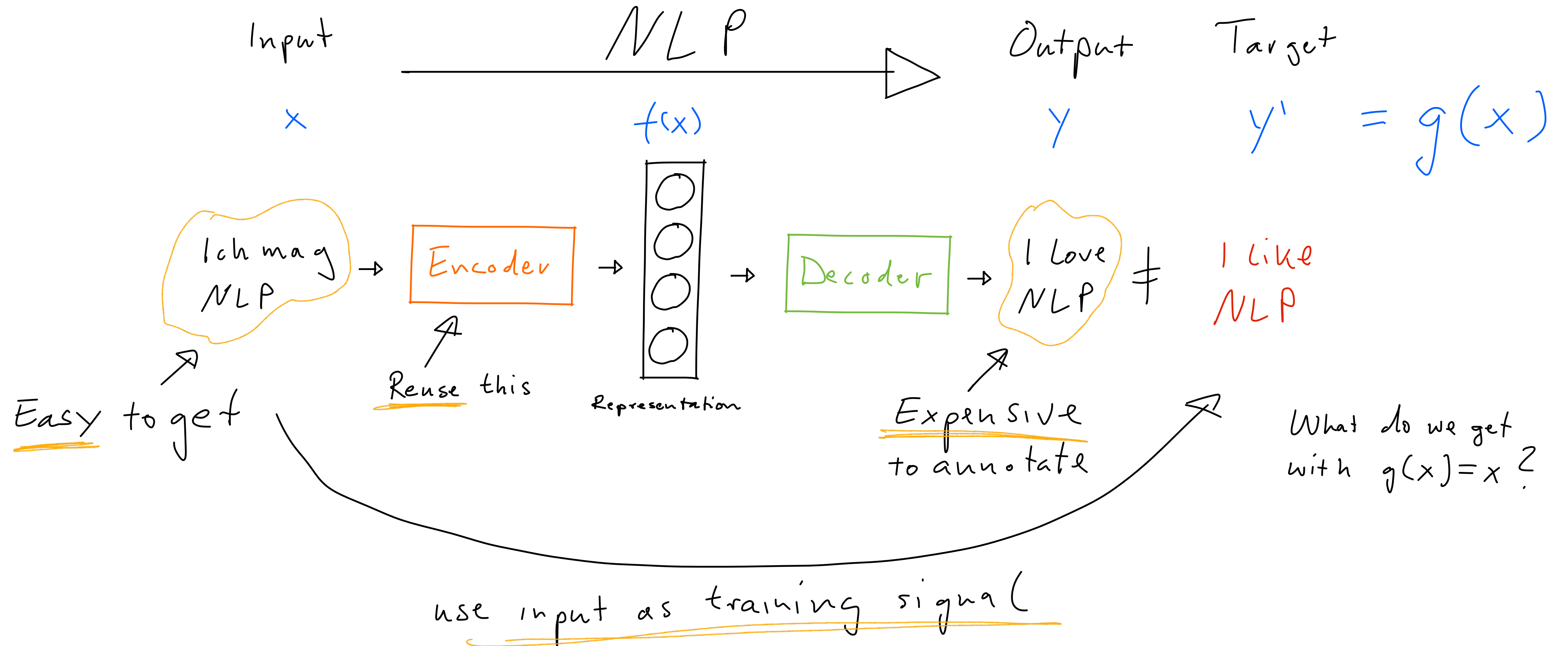
we have this



but we want this



# NLP in a Nutshell



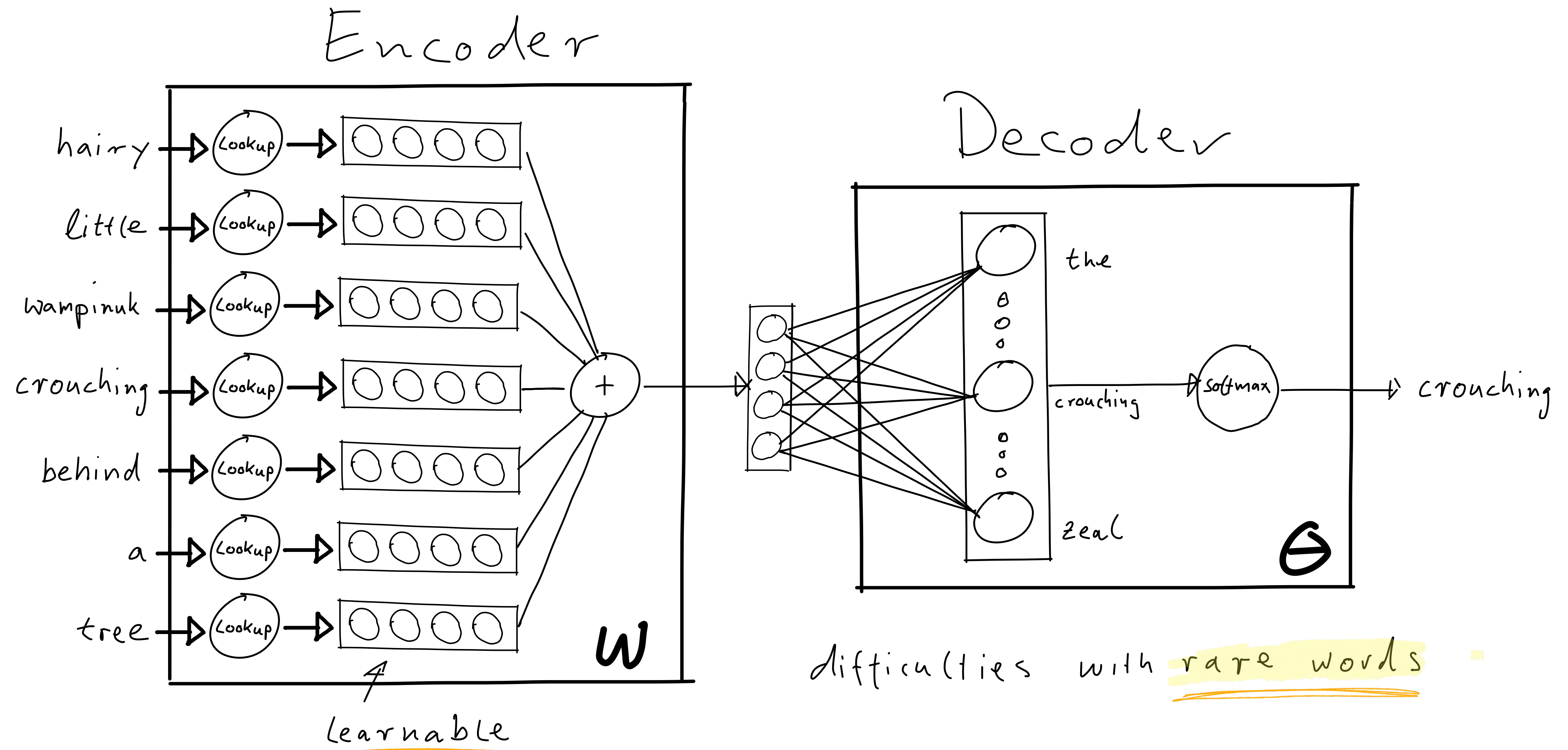
# Wampinuk

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

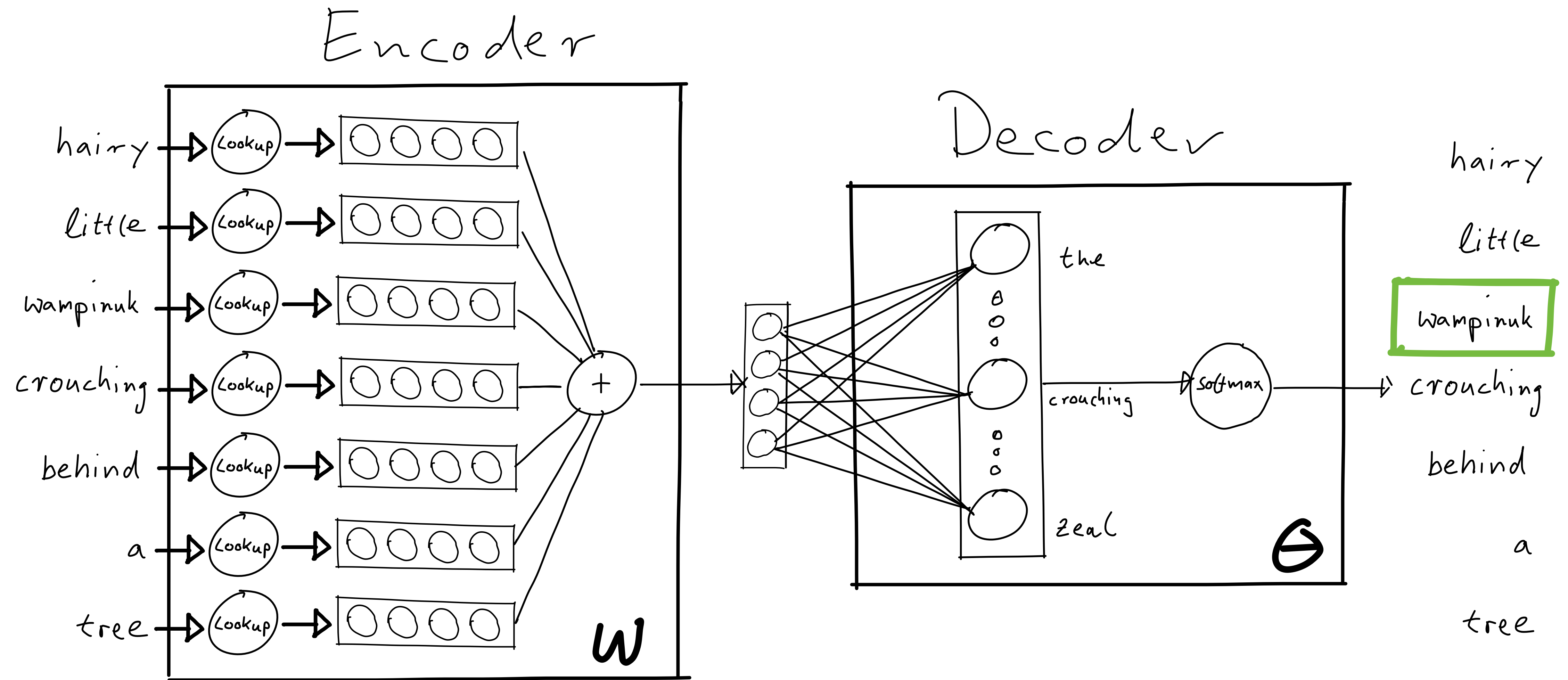
You shall know a word by the company it keeps (Firth, J. R. 1957:11)



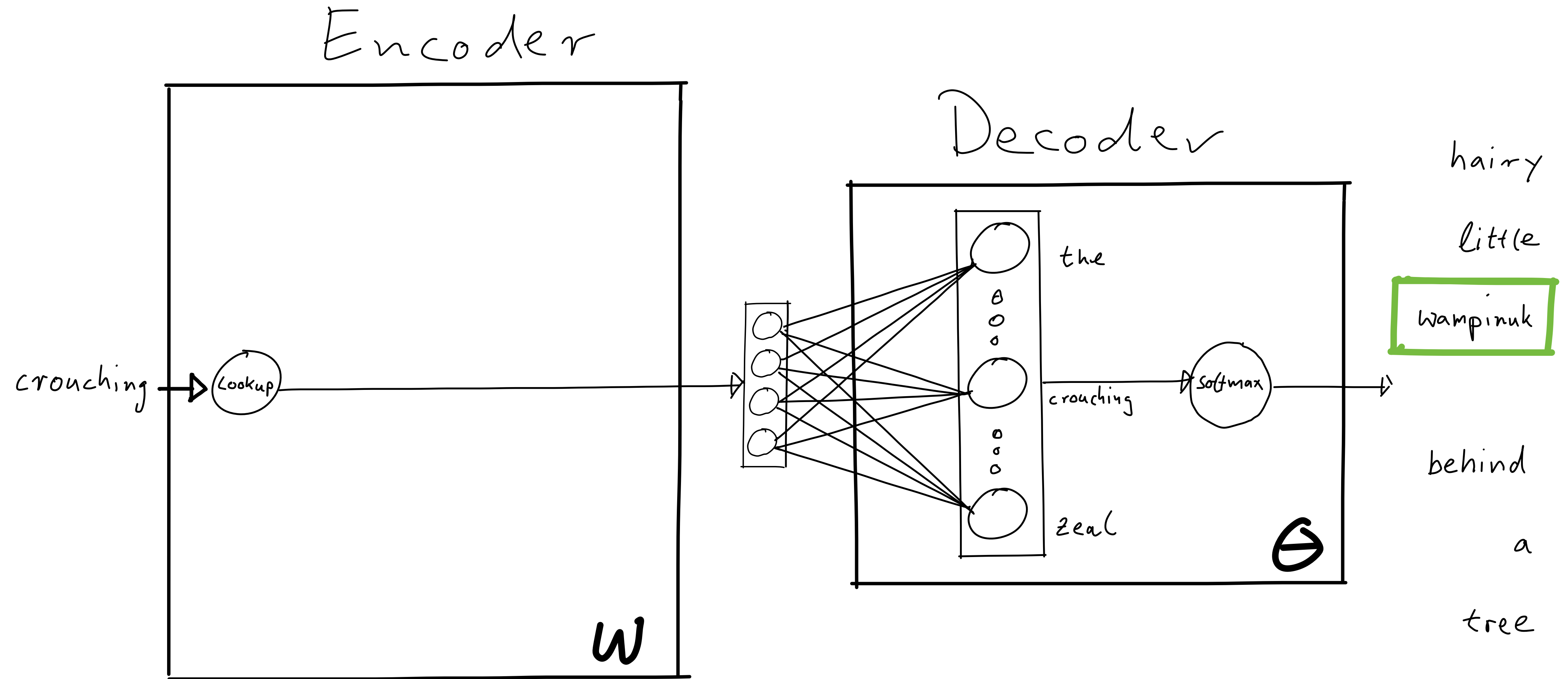
# Word2Vec: Continuous Bag of Words



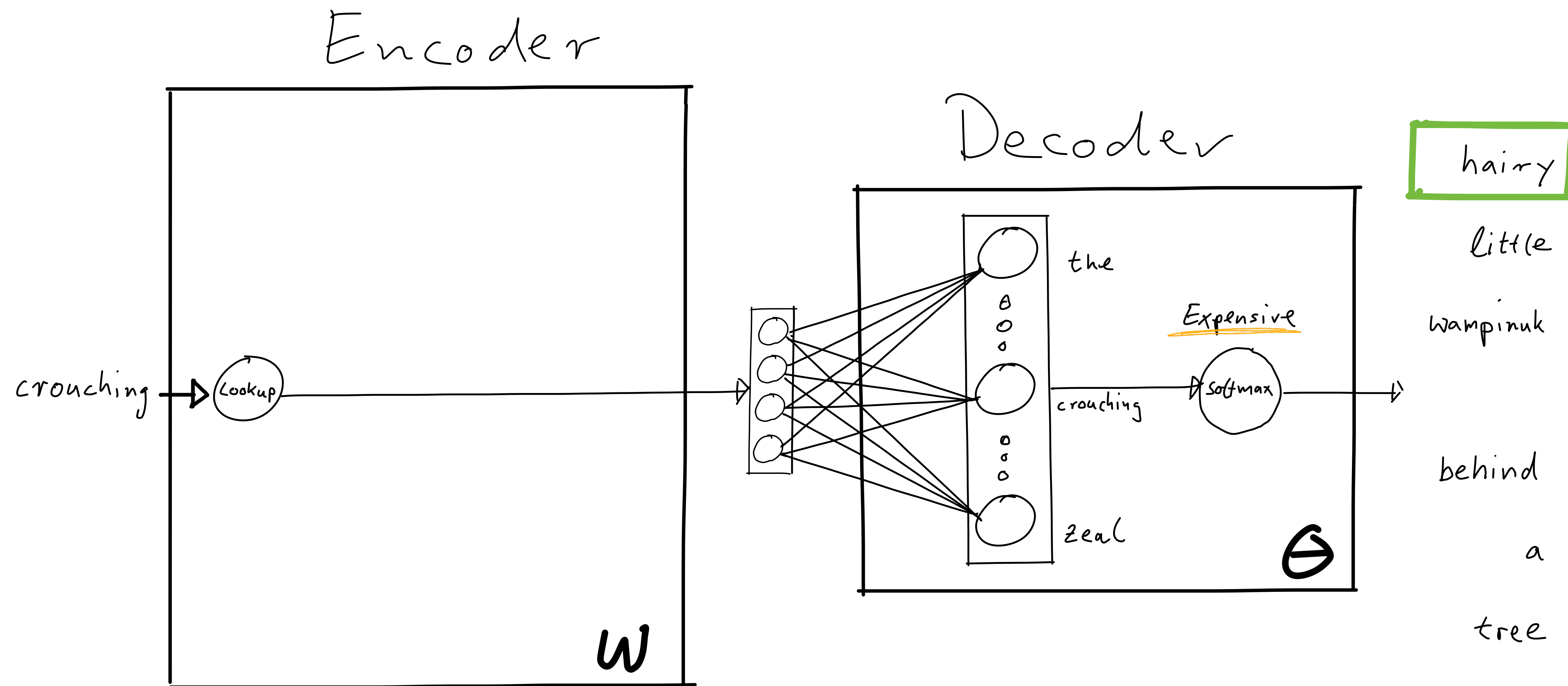
# Word2Vec: Skip Gram



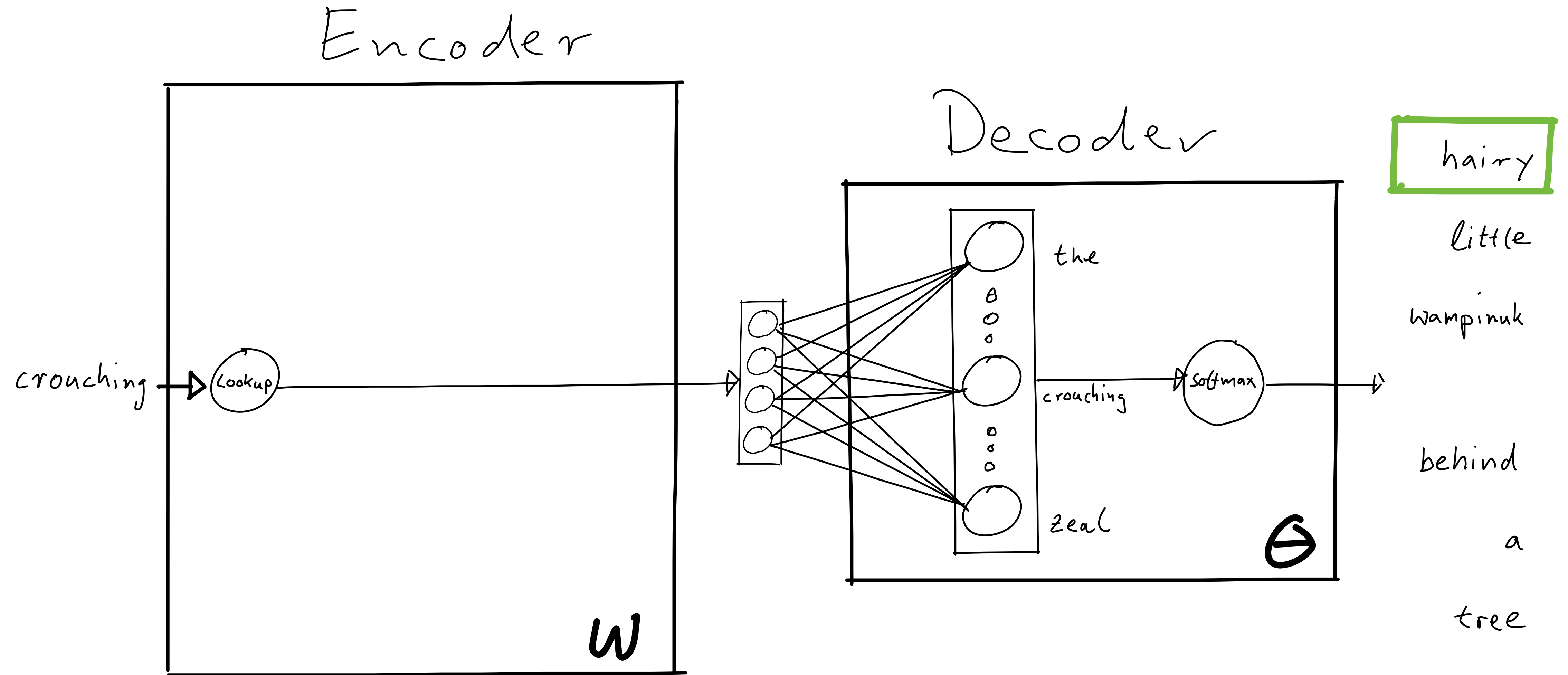
# Skip Gram



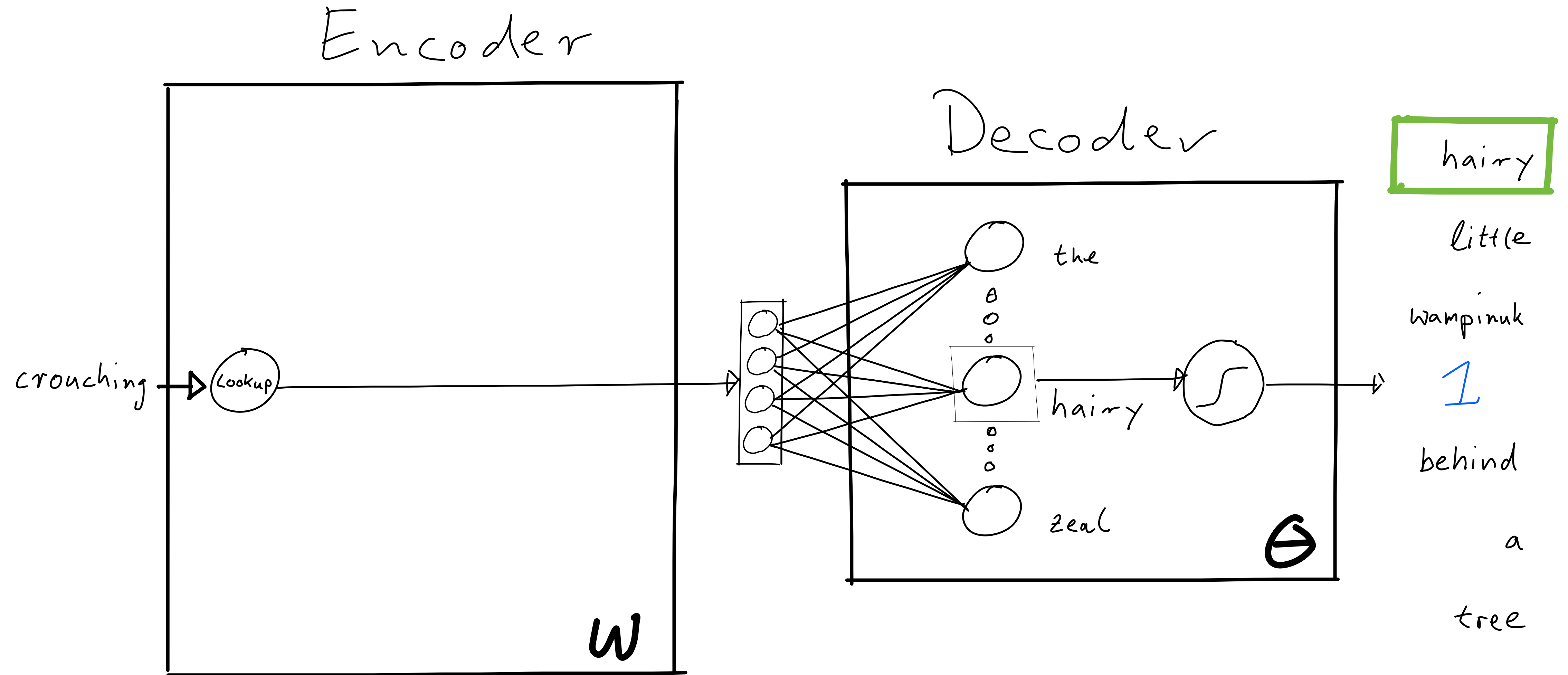
# Skip Gram



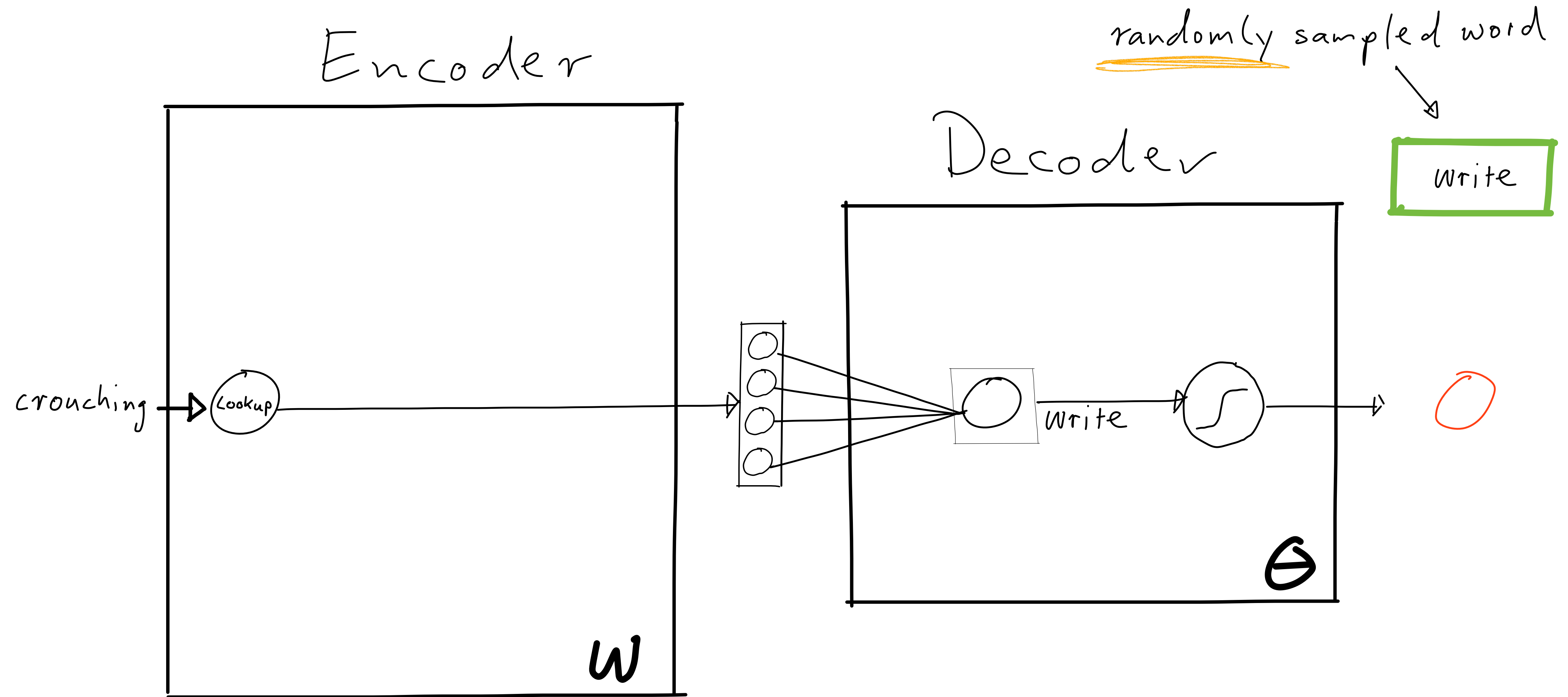
# Binary Classifier



# Binary Classifier



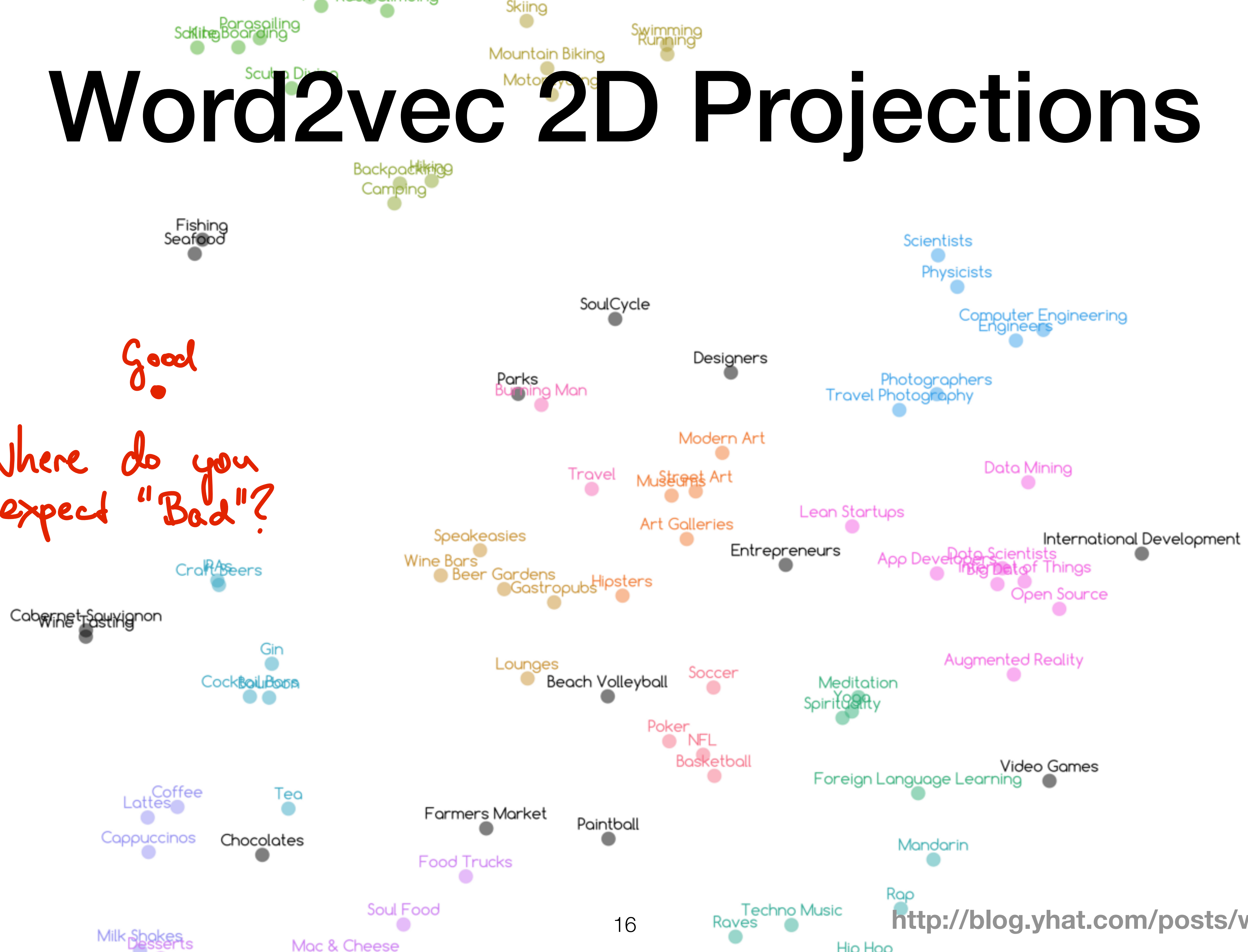
# Negative Sampling





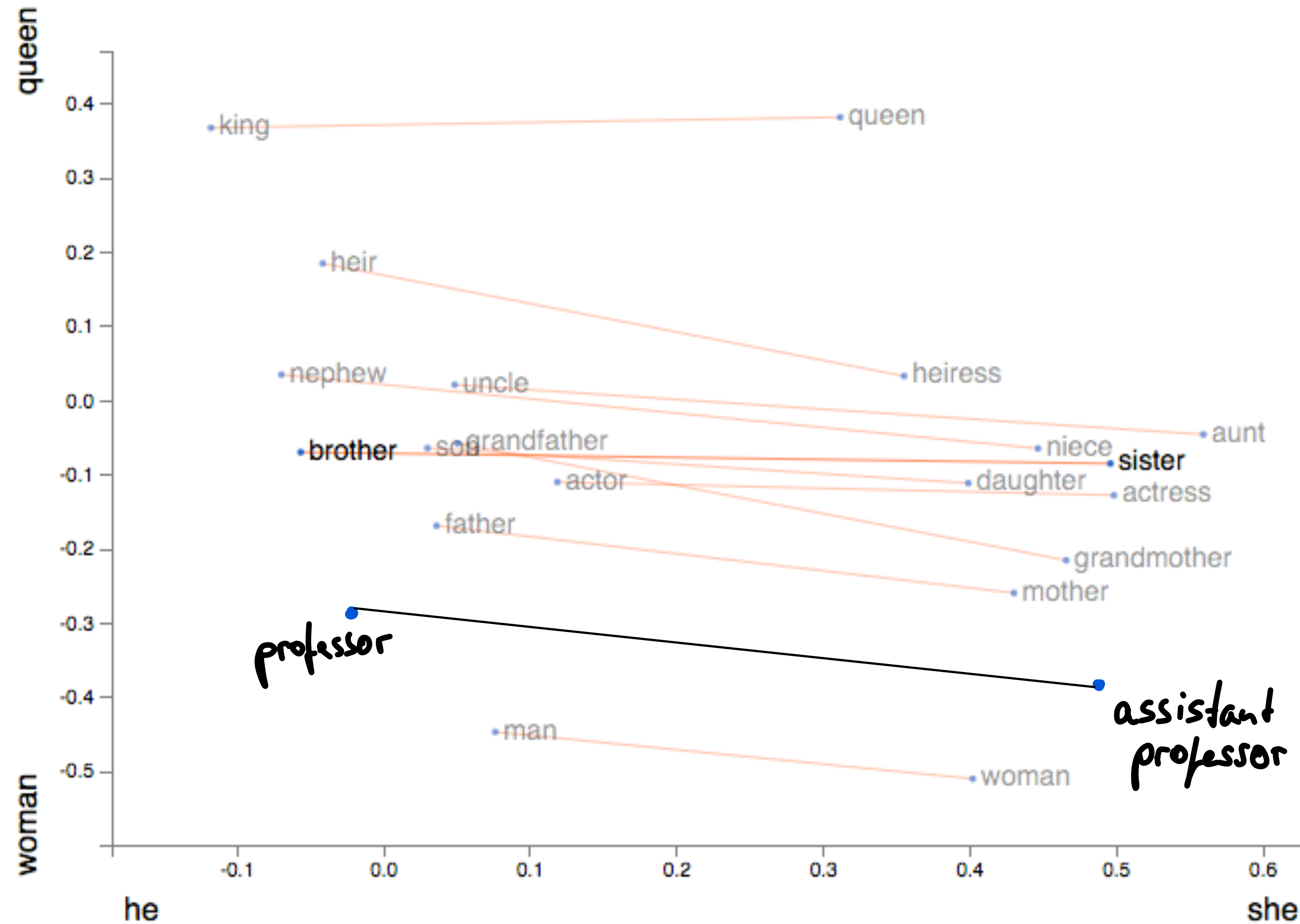
# Word2vec 2D Projections

Good  
Where do you  
expect "Bad"?





# King - Man + Woman $\approx$ Queen



Be aware  
of biases  
in the  
data!

# Classify This!

Train {

blah	blah	football	→ A
blah	blah	stocks	→ B

Test      b l a h    b l a h    hockey → ?

# Cluster This!

---

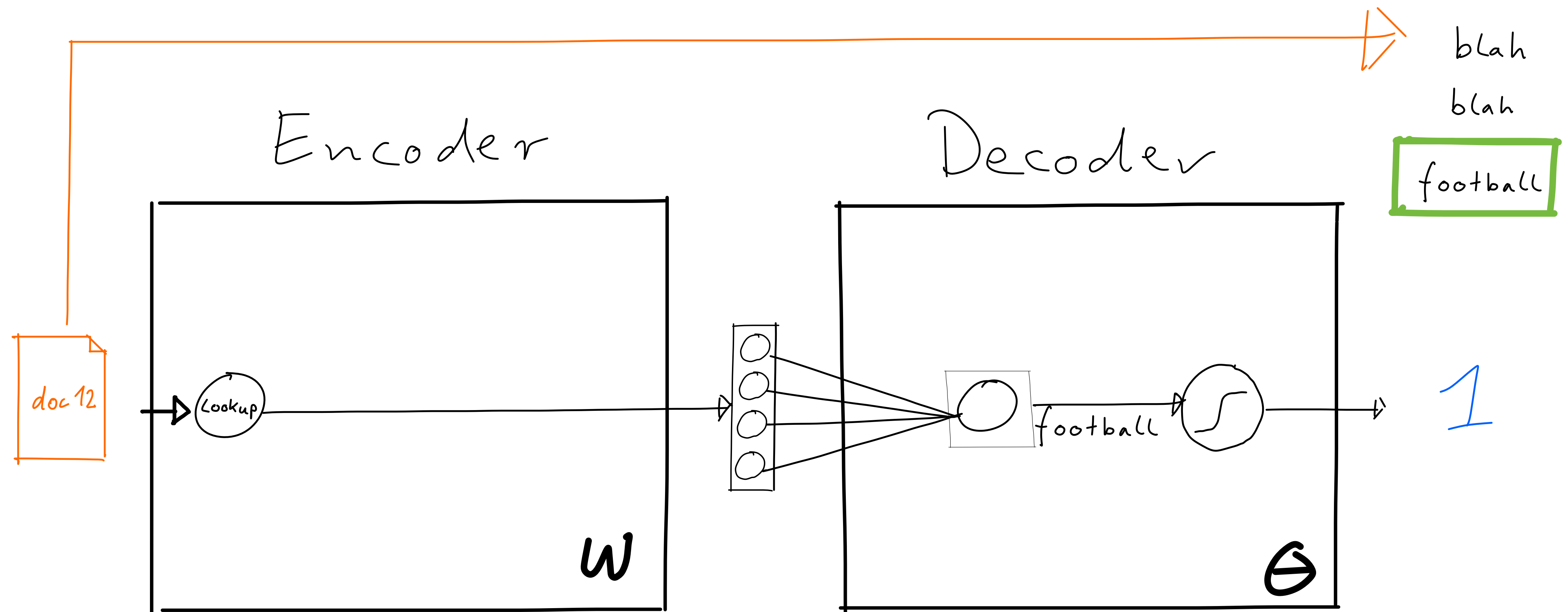
blah blah football → ?

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

$\theta_{\text{football}}$		
	2.0	$\theta$
	-1.0	
	0.0	
	4.5	
		$A = W\theta$

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

A hand-drawn diagram of a 2D coordinate system. The vertical axis is labeled  $w_{12, \bullet}$  and the horizontal axis is labeled  $w$ . The origin is marked with a small dot. The horizontal axis has tick marks labeled 3.0, 1.5, and -1.5. The vertical axis has a tick mark labeled 1.5. The axes are drawn with solid black lines, and the labels and tick marks are handwritten in black ink.

# Matrix View

The diagram illustrates the calculation of a weighted sum and its activation. It consists of three main components:

- Weight Vector  $w$ :** A vertical column with three cells. The middle cell contains the values  $3.0 \quad 1.5 \quad -1.5$ . The bottom cell is labeled  $w$ .
- Input Vector  $\theta_{\text{football}}$ :** A horizontal row with three cells. The rightmost cell contains the values  $2.0 \quad -1.0 \quad 0.0$ . The top-right cell is labeled  $\theta_{\text{football}}$ .
- Calculation and Activation:**
  - The middle cell of the weight vector  $w$  is connected to the middle cell of the input vector  $\theta_{\text{football}}$  by a horizontal line.
  - The intersection of this line and the middle cell of the third column contains the value  $0.9$ .
  - The bottom cell of the third column contains the formula  $Y = \sigma(W\theta)$ , where  $\sigma$  is highlighted in green.

$$A_{12, \text{foot+ball}} = W_{12, \bullet} \Theta_{\bullet, \text{foot+ball}}$$

$$y_{12, \text{foot+ball}} = \sigma(A_{12, \text{foot+ball}})$$

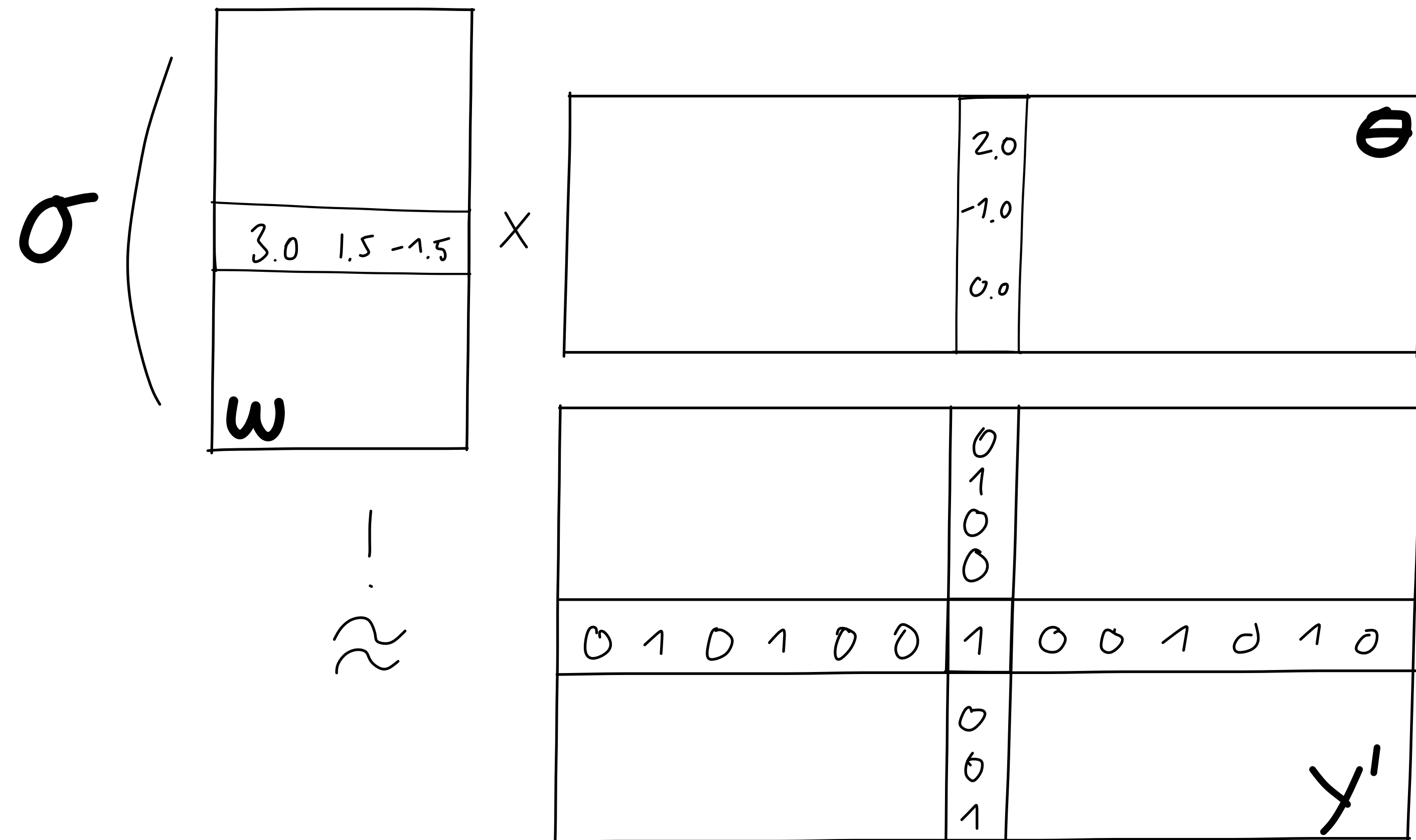
# Term Document Matrix

$w_{12,\bullet}$		$\theta_{\bullet, \text{football}}$									
		$\theta$									
		2.0 -1.0 0.0									
$w$											
		0 1 0 0									
	3.0 1.5 -1.5	0 1 0 1 0 0 1 0 0 1 0 0									
$w$											
		0 0 1									
		$y'$									

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

$$y_{12, \text{football}} = \sigma(A_{12, \text{football}})$$

# Binary Matrix Factorization



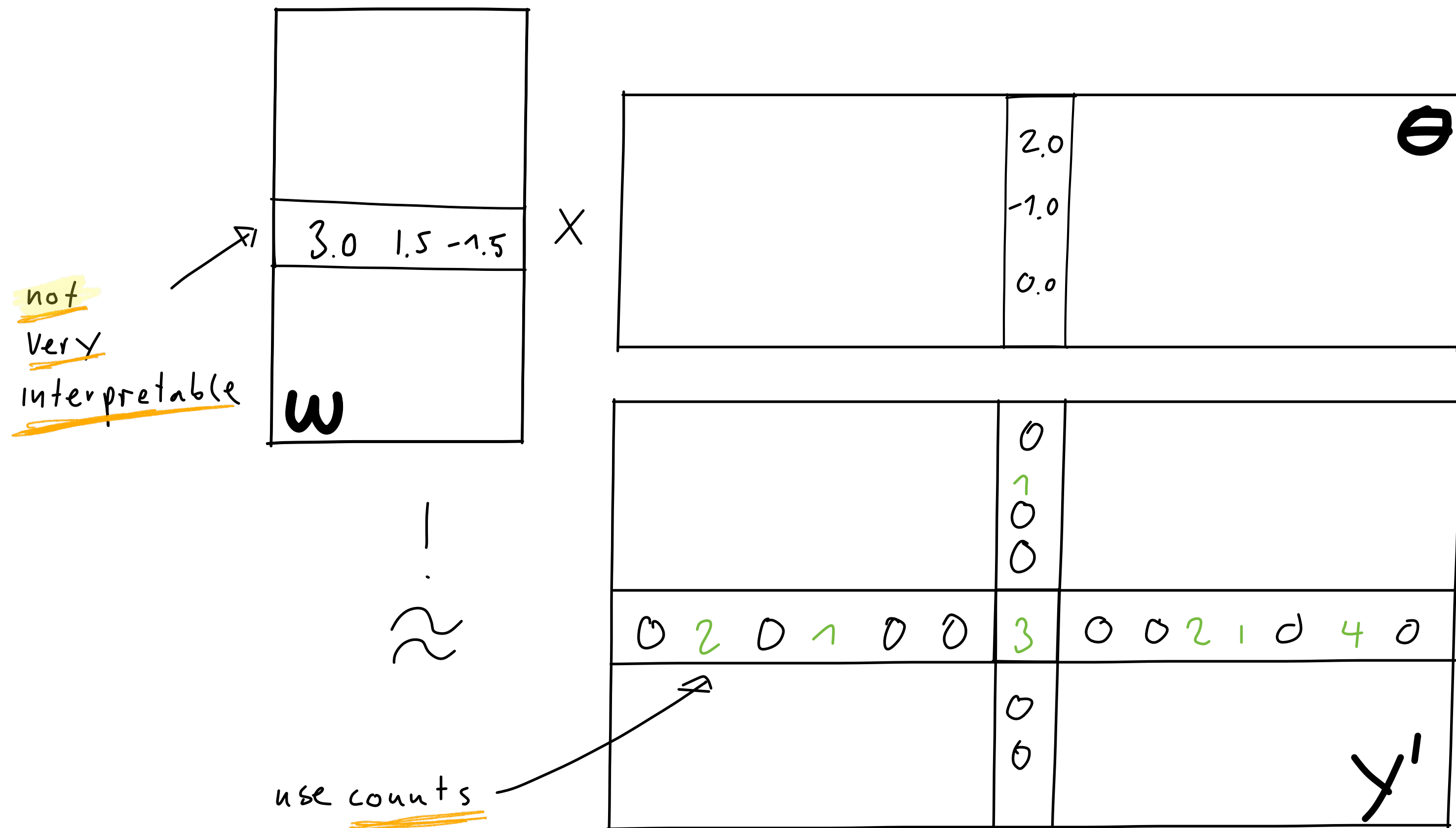
$$A_{12, \text{football}} = W_{12, \cdot} \Theta_{\cdot, \text{football}}$$

$$y_{12, \text{football}} = \sigma(A_{12, \text{football}})$$

$$\approx y'_{12, \text{football}}$$



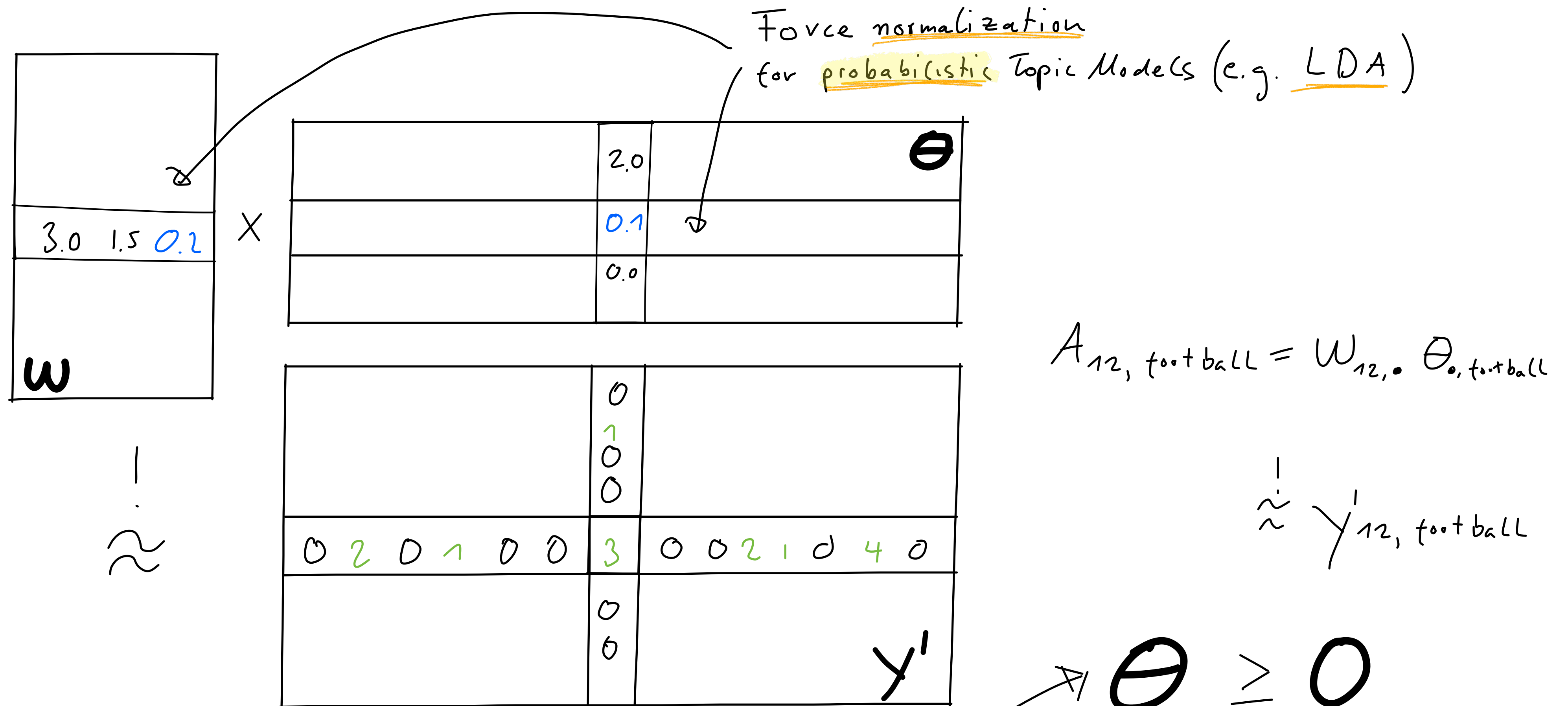
# Matrix Factorization



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

$$\hat{y}'_{12, \text{football}}$$

# Nonnegative Matrix Factorization



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