Unit 3: Learning from other people

6. Modern language models

11/24/2020

Modern language models

1. Embedding models are a general class of models for representing meaning in a vector-space

2. Embedding models can be used to understand aspects of cognition and language

3. The leading edge of models don’t represent “meaning” anymore at all

How do you know so much without being told about it?

Plato's Problem:
Even uneducated people seem to know a lot

Plato's Solution:
Knowledge is innate

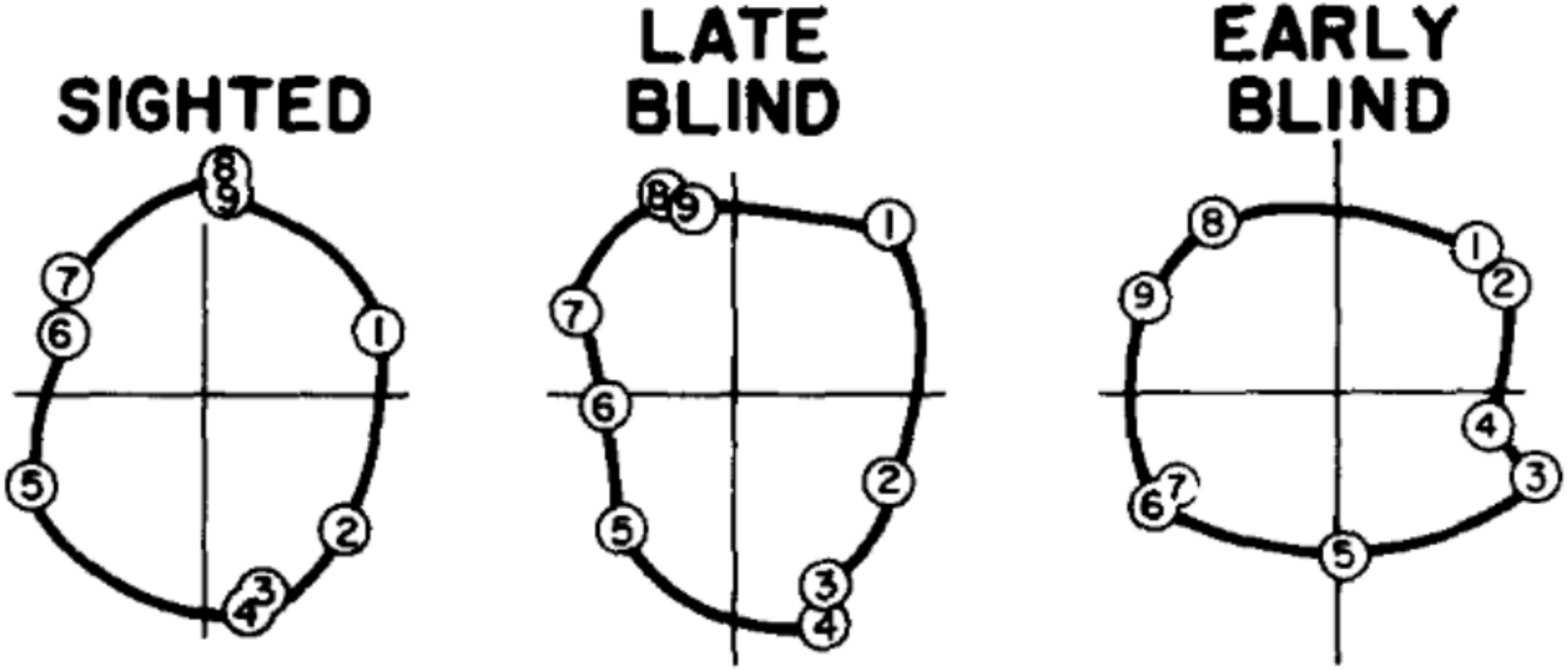
Plato (380 BC)

Chomsky's Problem:
Children seem to learn language from insufficient input

Chomsky's Solution:
Universal grammar is innate

Chomsky (1986)

Blind adults color similarities look a lot like sighted adults



COLOR LEGEND:

1. RED 2. ORANGE 3. GOLD 4. YELLOW 5. GREEN
6. TURQUISE 7. BLUE 8. PURPLE 9. VIOLET

Marmor (1978)

A solution to Plato's problem (Landauer & Dumais, 1997)

Red onions are sweeter than **white** ones

Red hair occurs naturally in one to two percent of the human population

Pittsburgh one of U.S. cities with highest number of **gray** days

Fall tips for a **green** spring lawn

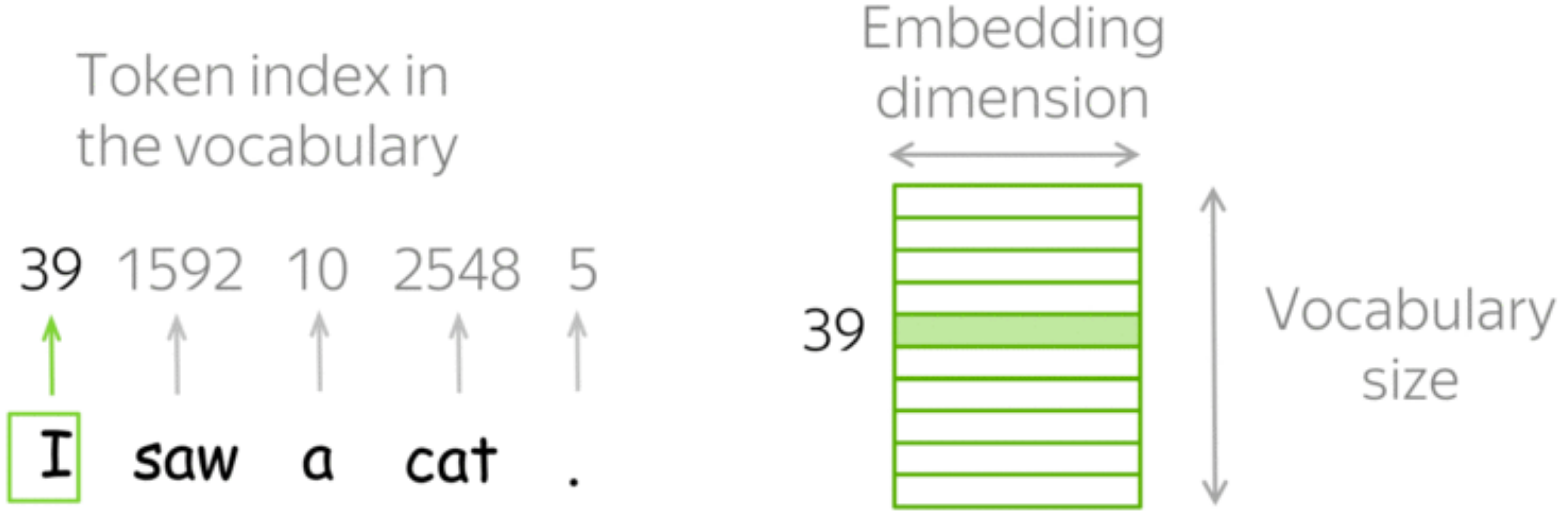
Lake Tahoe stretches 22 miles long and 12 miles wide, with clear **blue** water that's more than 99 percent pure

Direct information:
There is a relationship between e.g. red and hair

Indirect information:
Red, white, gray, green, and blue are used in *similar contexts*.

Contexts for e.g. blue and green are more similar than blue and red

Embedding models: Representing words as vectors



What goes in the embeddings?

Adapted from Lena Voita

A simple idea: embeddings as co-occurrence counts

2-sized window for cat

… I saw a cute grey cat playing in the garden …

contexts for cat

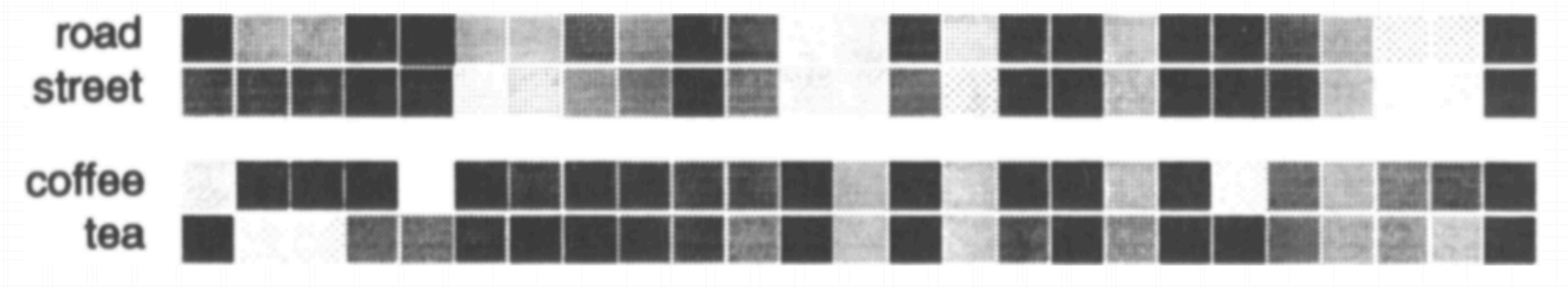
Hyperspace analogies to language (HAL) - Lund & Burgess (1996)

Example Matrix for "The Horse Raced Past the Barn Fell" (Computed for Window Width of Five Words)

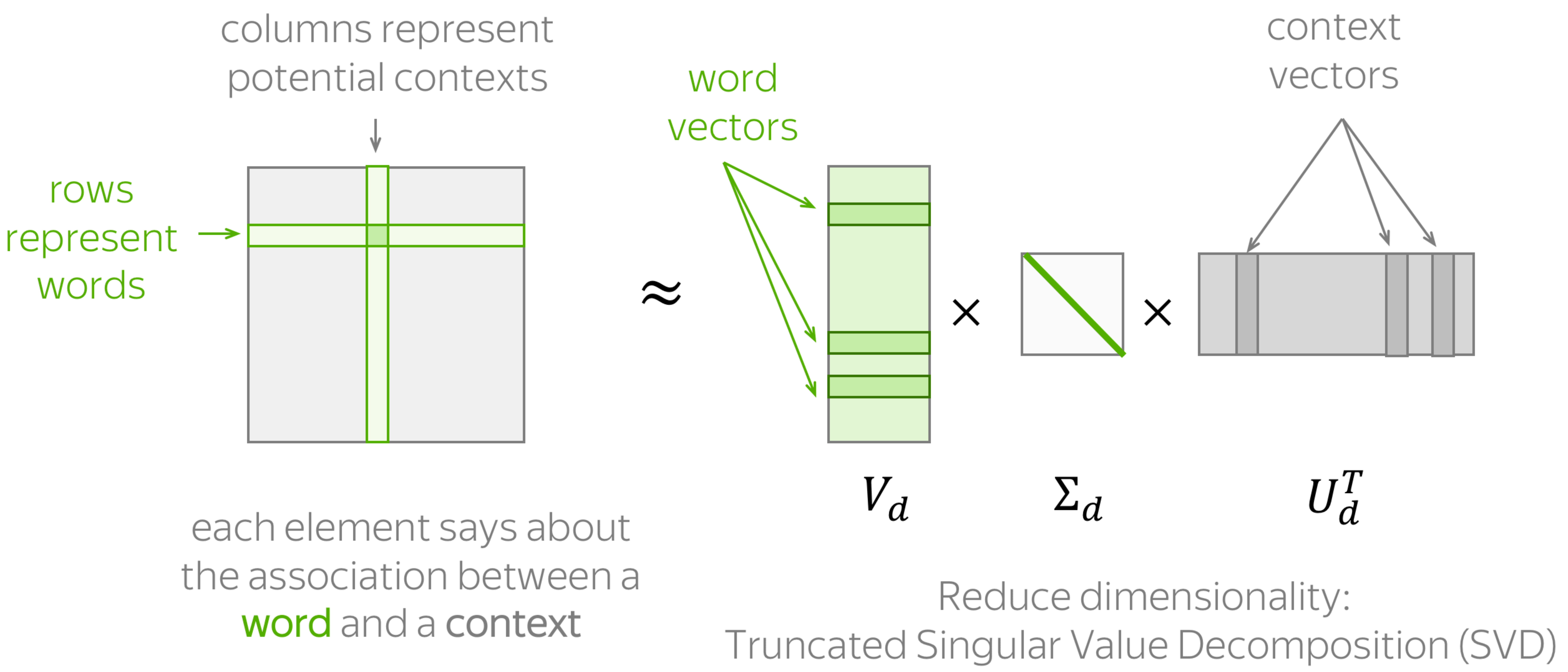
	barn	fell	horse	past	raced	the
<PERIOD>	4	5	0	2	1	3
barn	0	0	2	4	3	6
fell	5	0	1	3	2	4
horse	0	0	0	0	0	5
past	0	0	4	0	5	3
raced	0	0	5	0	0	4
the	0	0	3	5	4	2

Five Nearest Neighbors for Target Words From Experiment 1 (*n*1 ... *n*5)

Target	*n*1	*n*2	*n*3	*n*4	*n*5
jugs	juice	butter	vinegar	bottles	cans
leningrad	rome	iran	dresden	azerbaijan	tibet
lipstick	lace	pink	cream	purple	soft
triumph	beauty	prime	grand	former	rolling
cardboard	plastic	rubber	glass	thin	tiny
monopoly	threat	huge	moral	gun	large

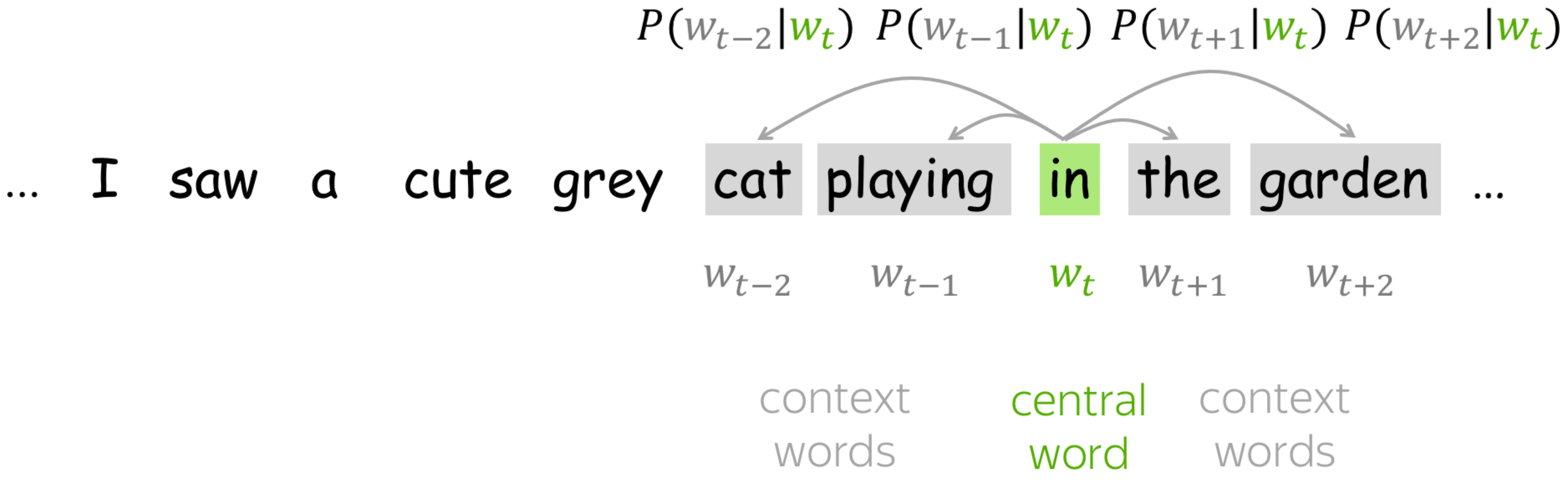


Latent semantic analysis is a smarter embedding model than HAL



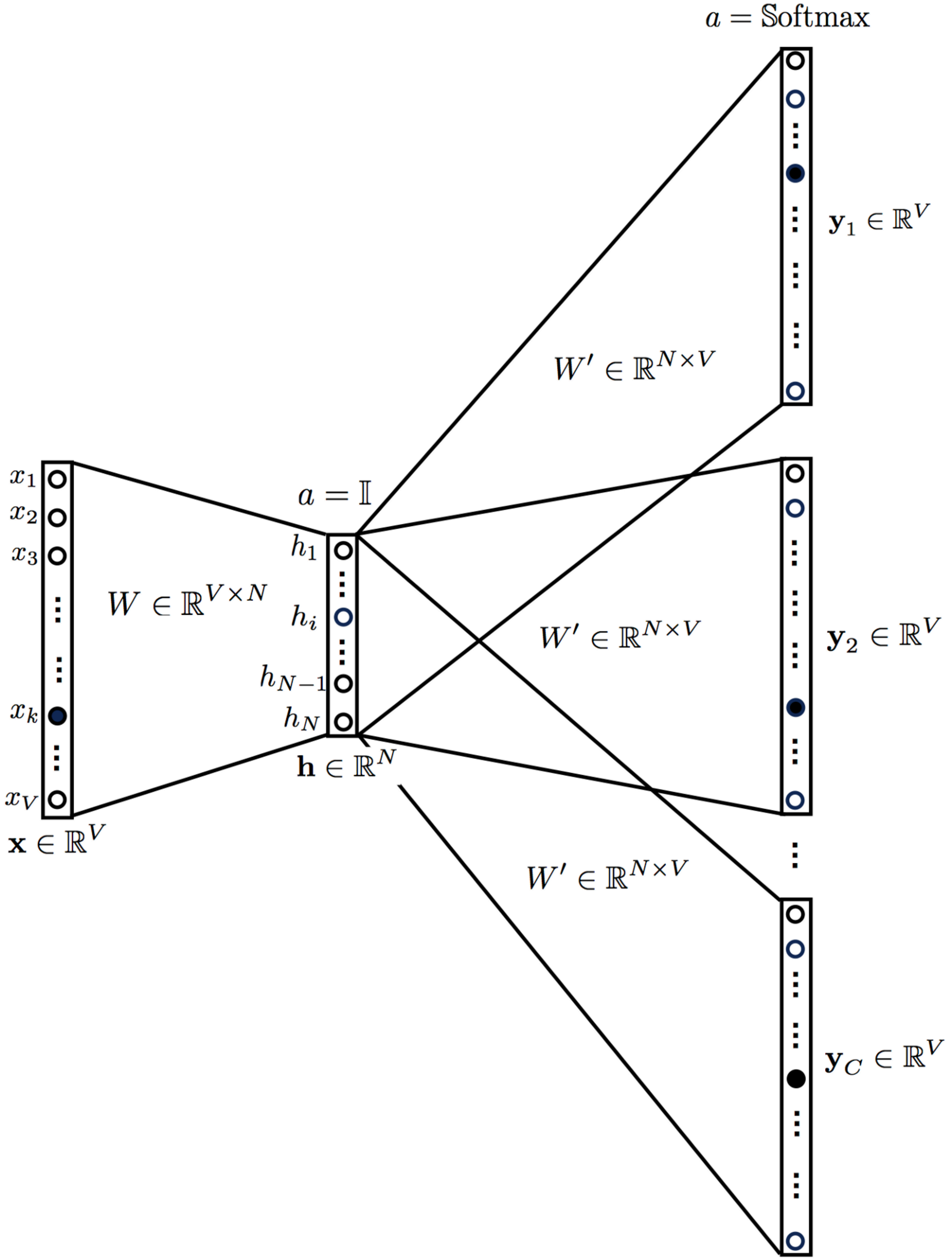
Insight: co-occurring with some words (or in some contexts) is more meaningful

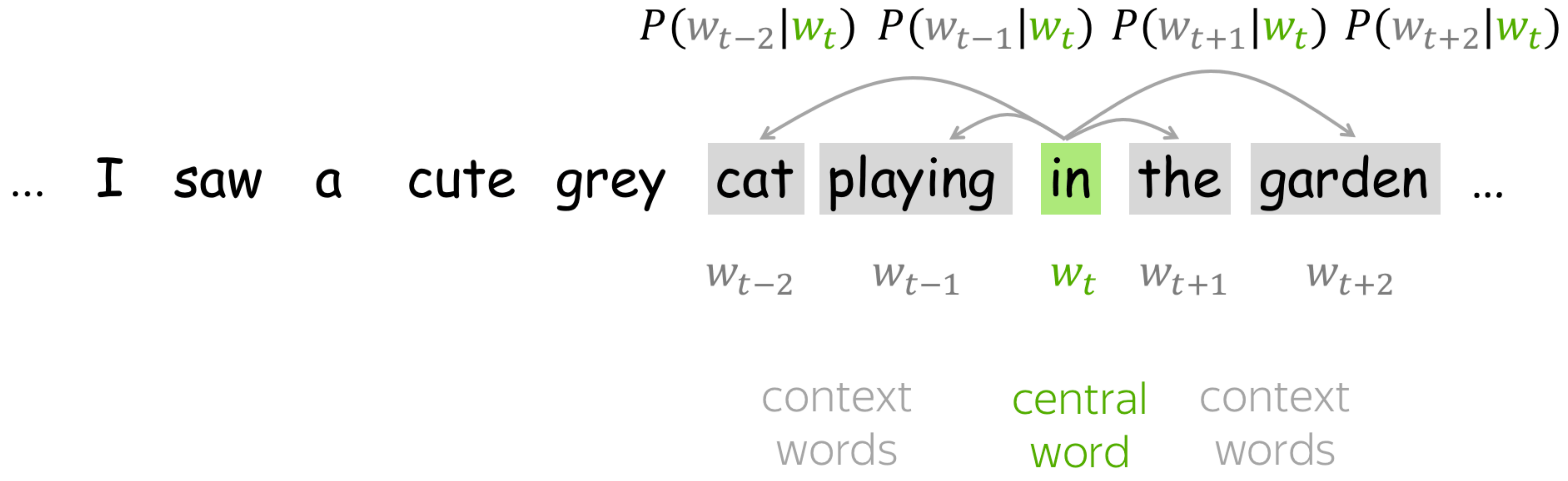
Can we do this separately for each word?



We want to predict a word's **context** from that word

Learning contexts using a skip-gram model (Word2Vec) - Mikolov et al. (2013)





Target words’ embeddings

$P(w_{t-2}|w_t)$ $P(w_{t-1}|w_t)$ $P(w_{t+1}|w_t)$ $P(w_{t+2}|w_t)$

… I saw a cute grey cat playing in the garden …

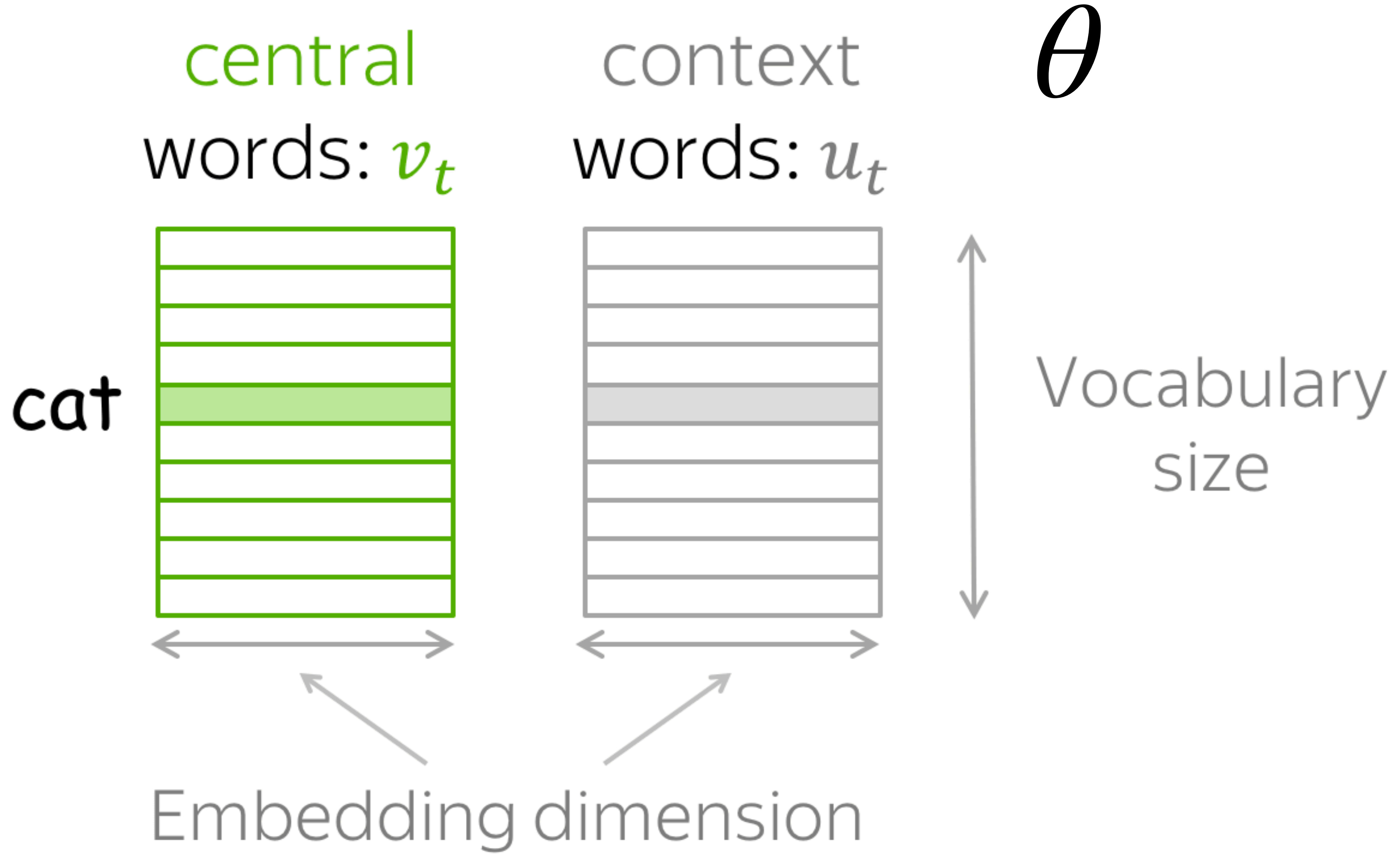
w_{t-2} w_{t-1} w_t w_{t+1} w_{t+2}

Likelihood: $$L(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m, j \neq 0} P(w_{t+j} | w_t, \theta)$$

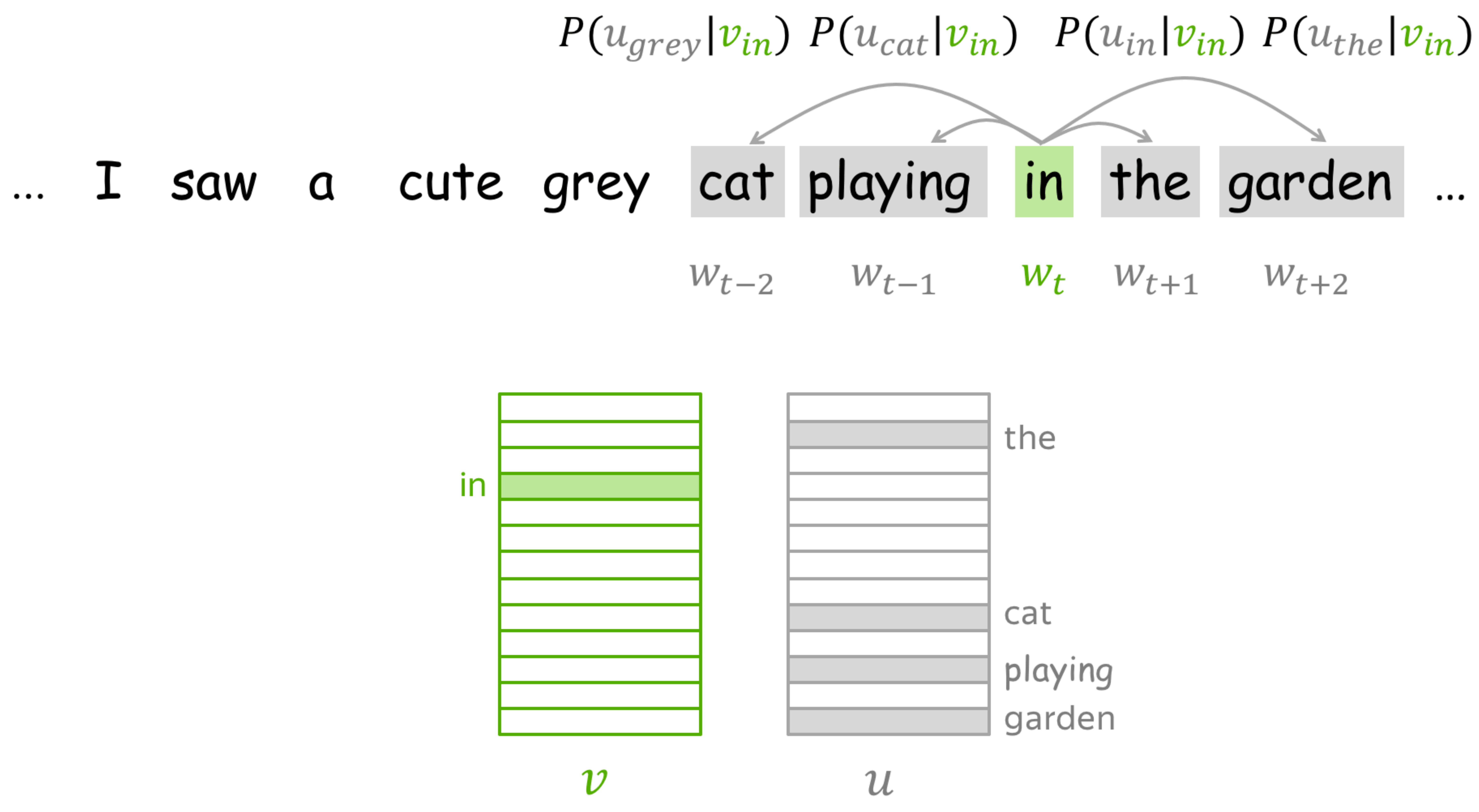
$$P(o|c) = \frac{\exp(\boxed{u_o^T v_c})}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Similarity of o and c

Normalization



Target words' embeddings



Estimating words' embeddings by gradient descent

Likelihood: $$L(\theta) = \prod_{t=1}^{T} \prod_{-m \leq j \leq m, j \neq 0} P(w_{t+j} | w_t, \theta)$$

Loss: $$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \leq j \leq m, \\ j \neq 0}} \log P(w_{t+j} | w_t, \theta)$$

$$\theta^{new} = \theta^{old} - \alpha \nabla_{\theta} J(\theta)$$

Estimating words' embeddings by gradient descent

… I saw a cute grey cat playing in the garden …

$$J_{t,j}(\theta) = -\log P(cute|cat)$$

$$= -\log \frac{\exp u_{cute}^T v_{cat}}{\sum_{w \in Voc} \exp u_w^T v_{cat}}$$

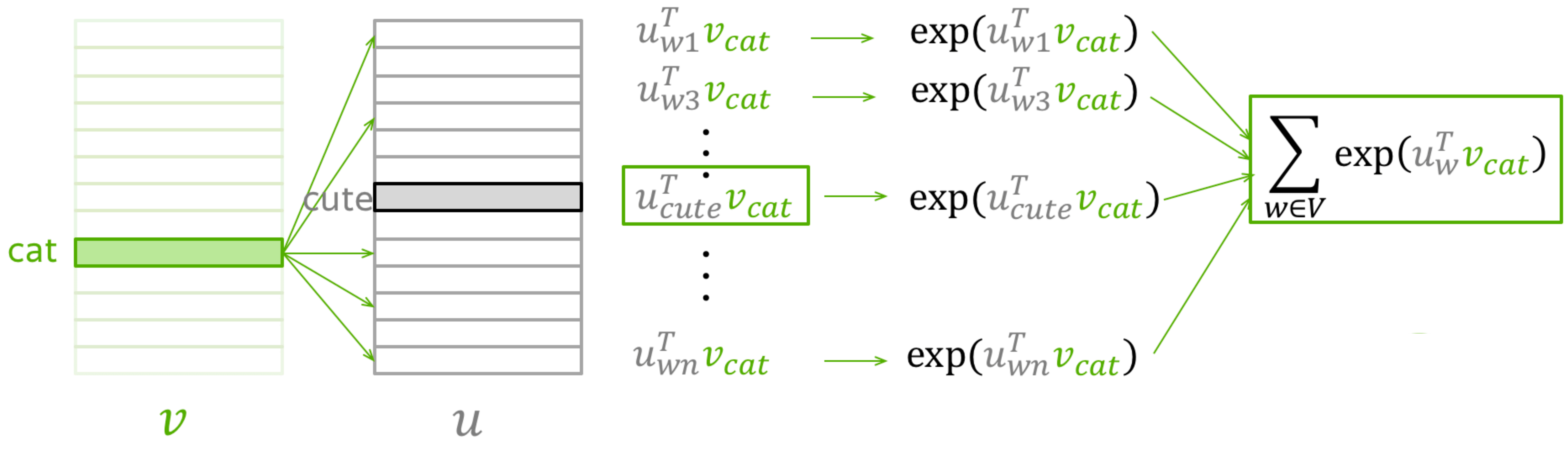
$$= -u_{cute}^T v_{cat} + \log \sum_{w \in Voc} \exp u_w^T v_{cat}$$

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Estimating words' embeddings by gradient descent

… I saw a cute grey cat playing in the garden …

$$J_{t,j}(\theta) = -\log P(cute|cat)$$

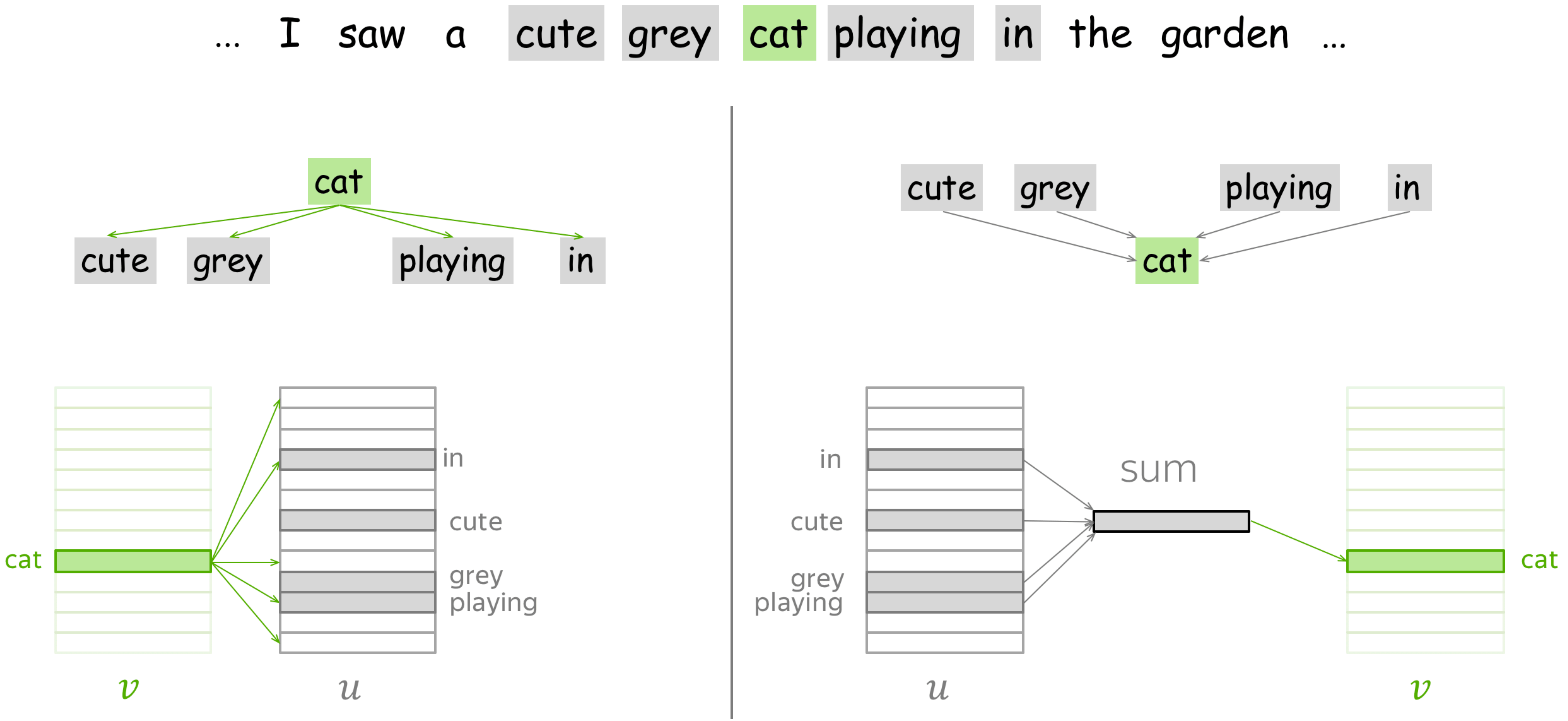


$$J_{t,j}(\theta) = \underbrace{-u_{cute}^T v_{cat}} + \underbrace{\log \sum_{w \in V} \exp(u_w^T v_{cat})}$$

$$v_{cat} := v_{cat} - \alpha \frac{\partial J_{t,j}(\theta)}{\partial v_{cat}}$$

$$u_w := u_w - \alpha \frac{\partial J_{t,j}(\theta)}{\partial u_w} \ \forall\, w \in V$$

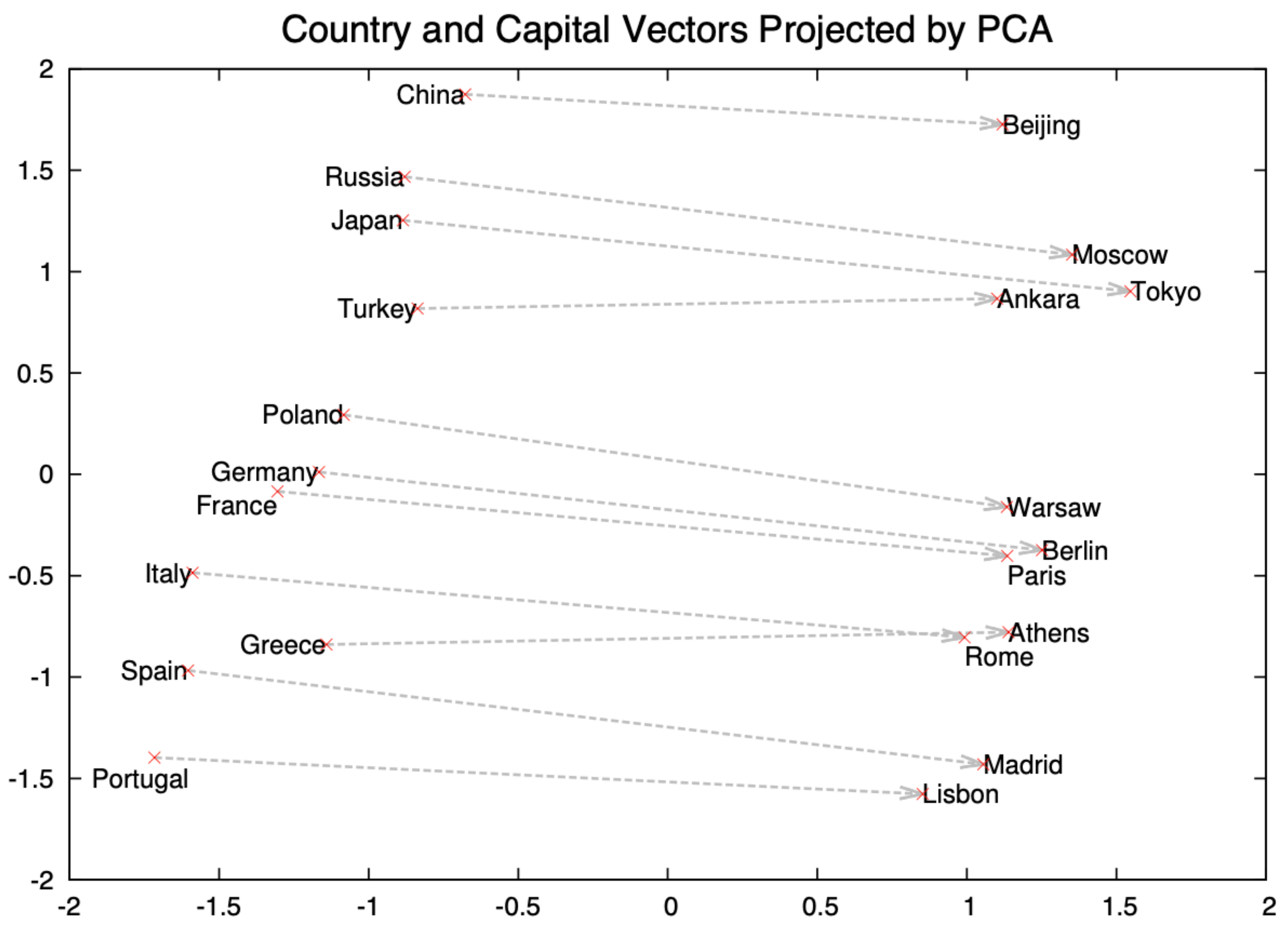
Two ways of estimating Word2Vec



Skip-gram

Continuous bag of words (CBOW)

Word2Vec geometry is surprisingly meaningful!



Mikolov et al. (2013)

Global Vectors for Word Representation (GloVe - Pennington, Socher, & Manning, 2014)

context vector

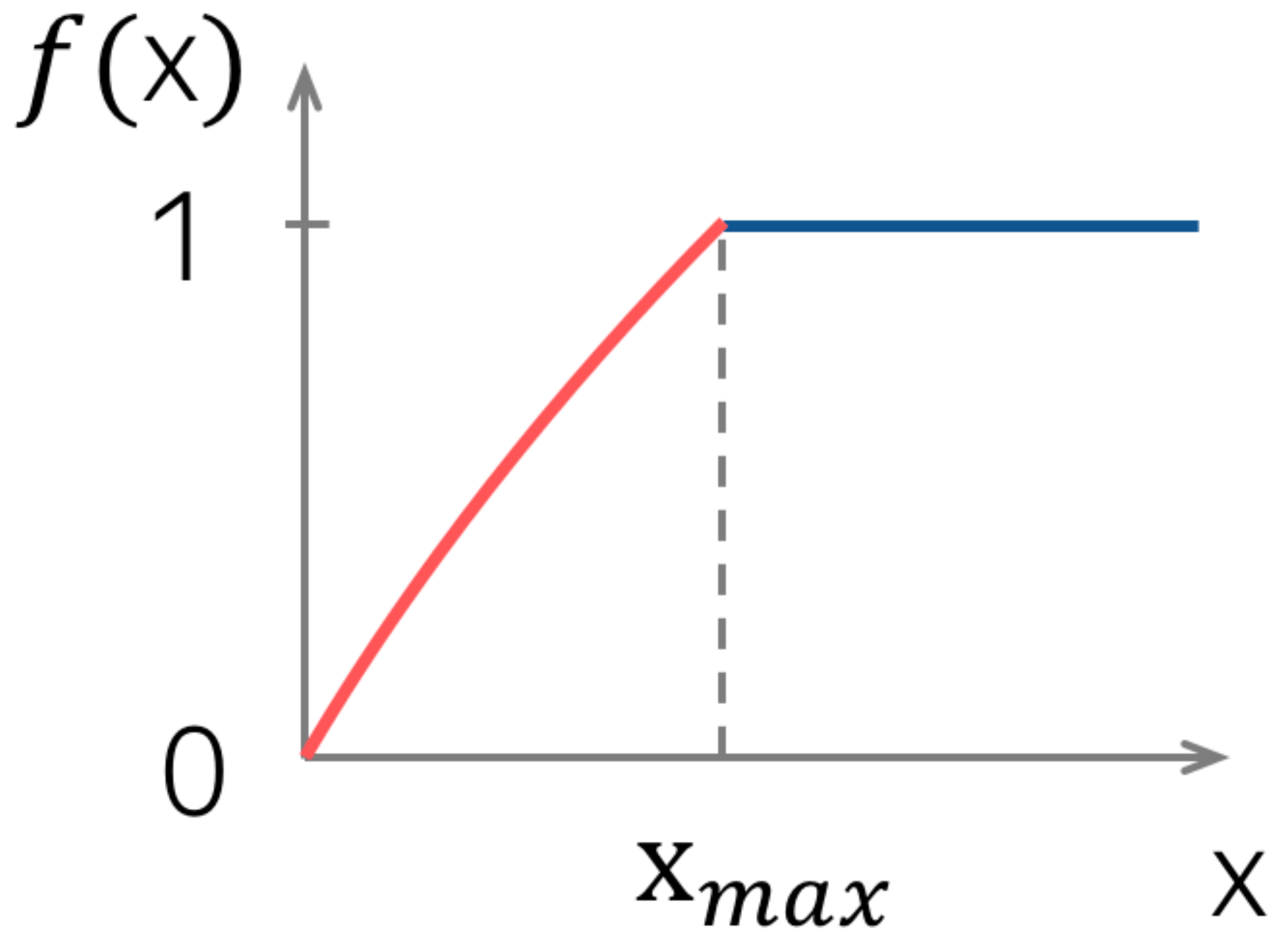
word vector

bias terms (also learned)

$$J(\theta) = \sum_{w,c \in V} \underbrace{f(\mathrm{N}(w, c))} \cdot (u_c^T v_w + b_c + \overline{b_w} - \log \mathrm{N}(w, c))$$

Weighting function to:

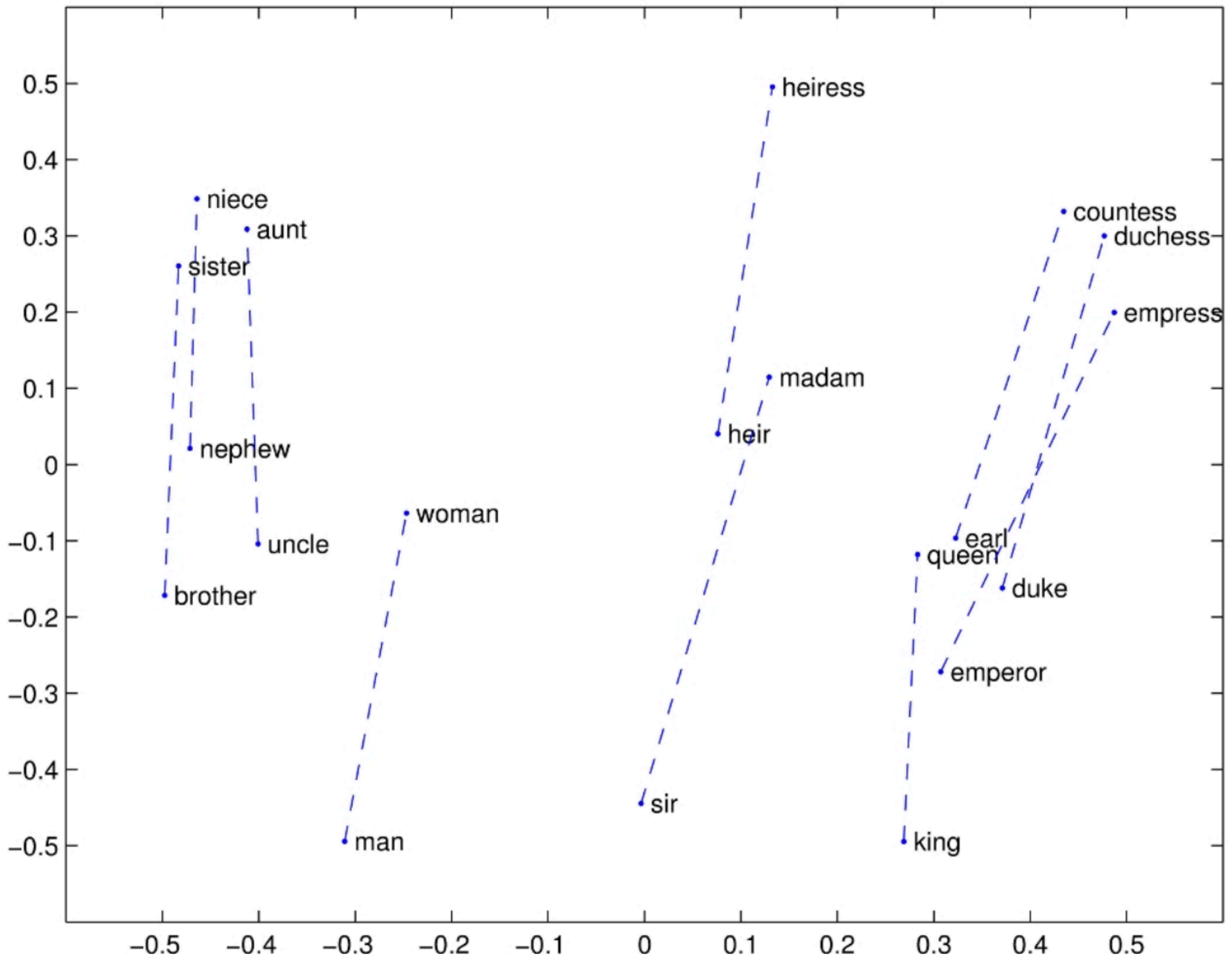
- penalize rare events
- not to over-weight frequent events

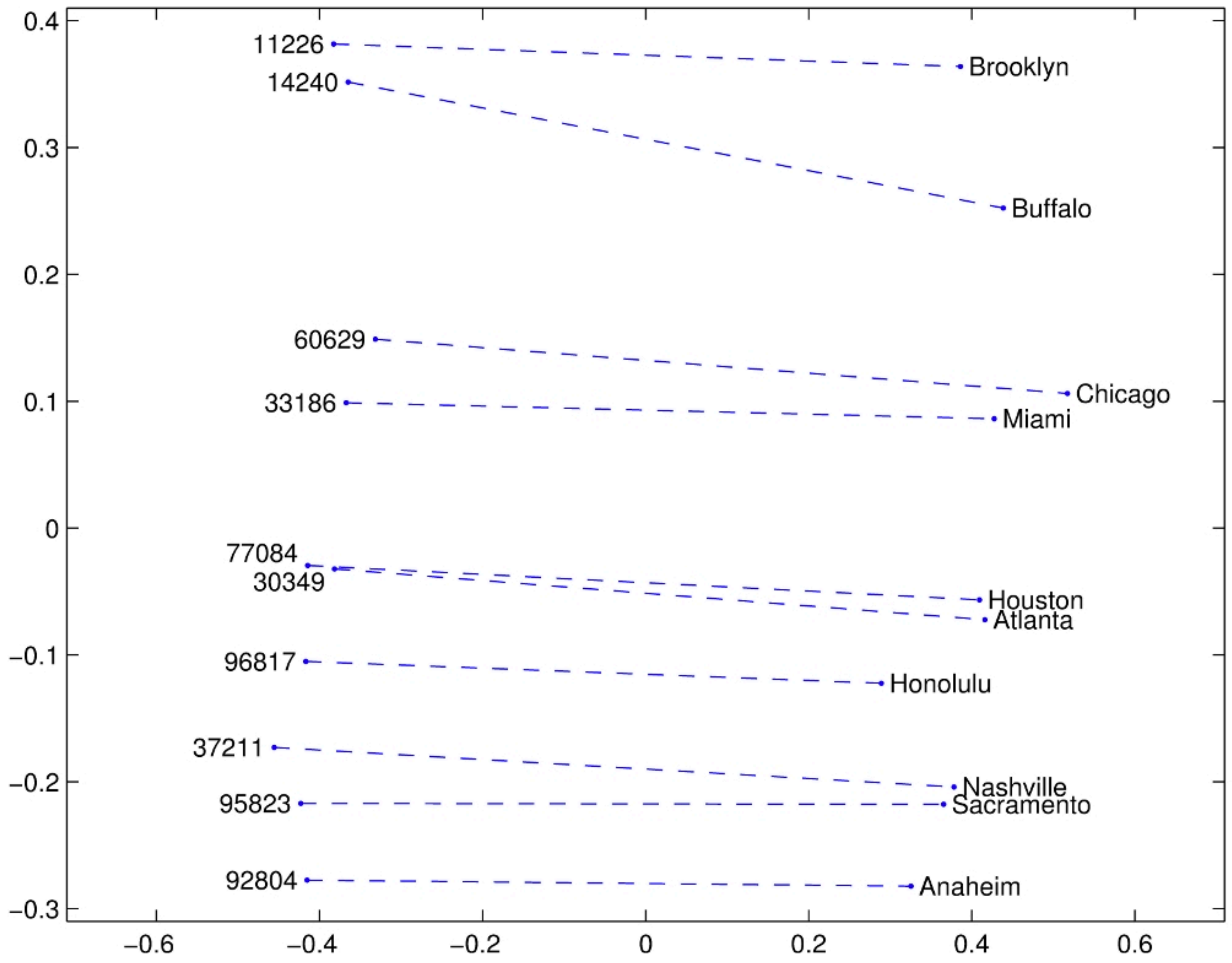


$$\begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max}, \\ 1 & \text{otherwise.} \end{cases}$$

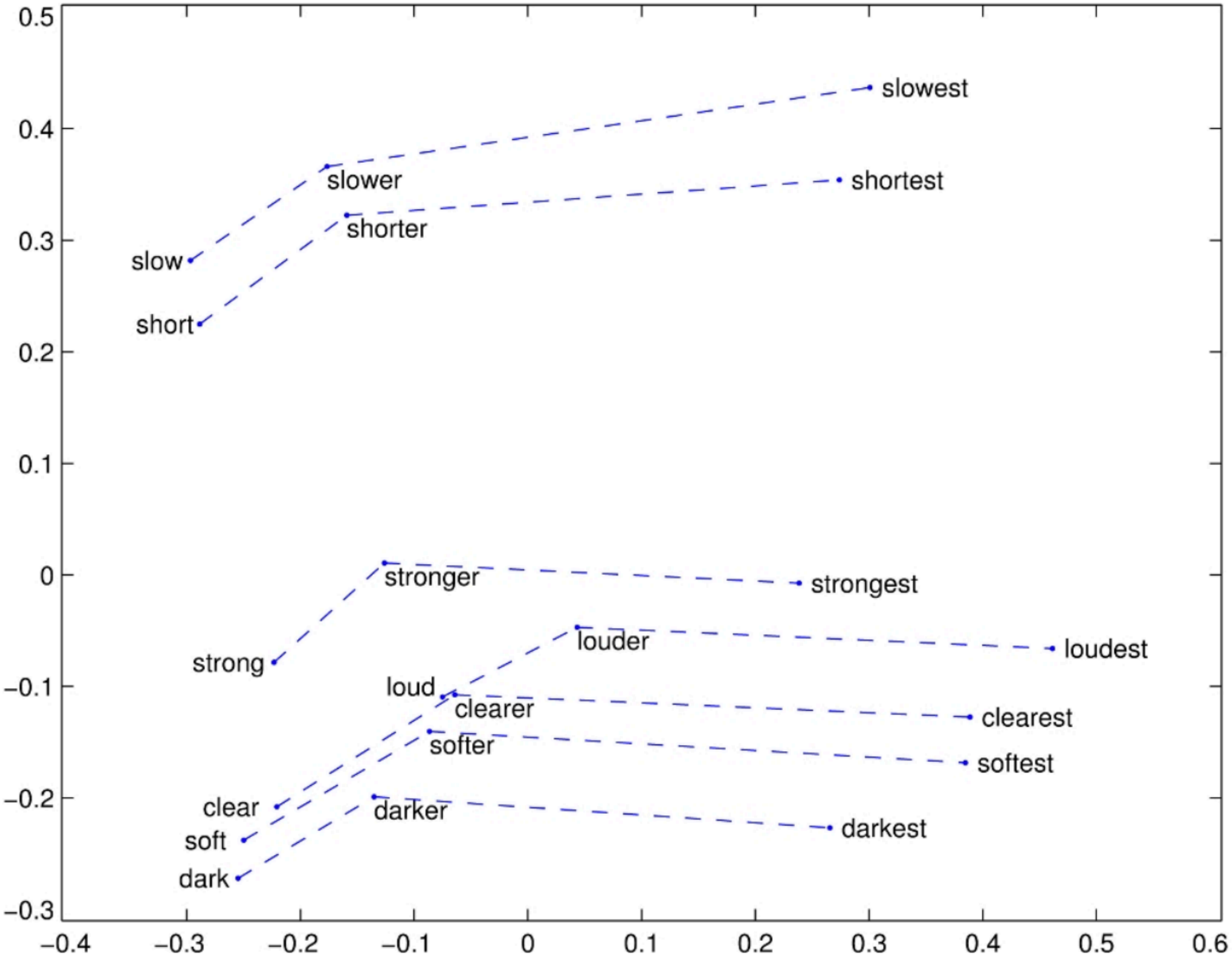
$\alpha = 0.75$, $x_{max} = 100$

The structure in embeddings





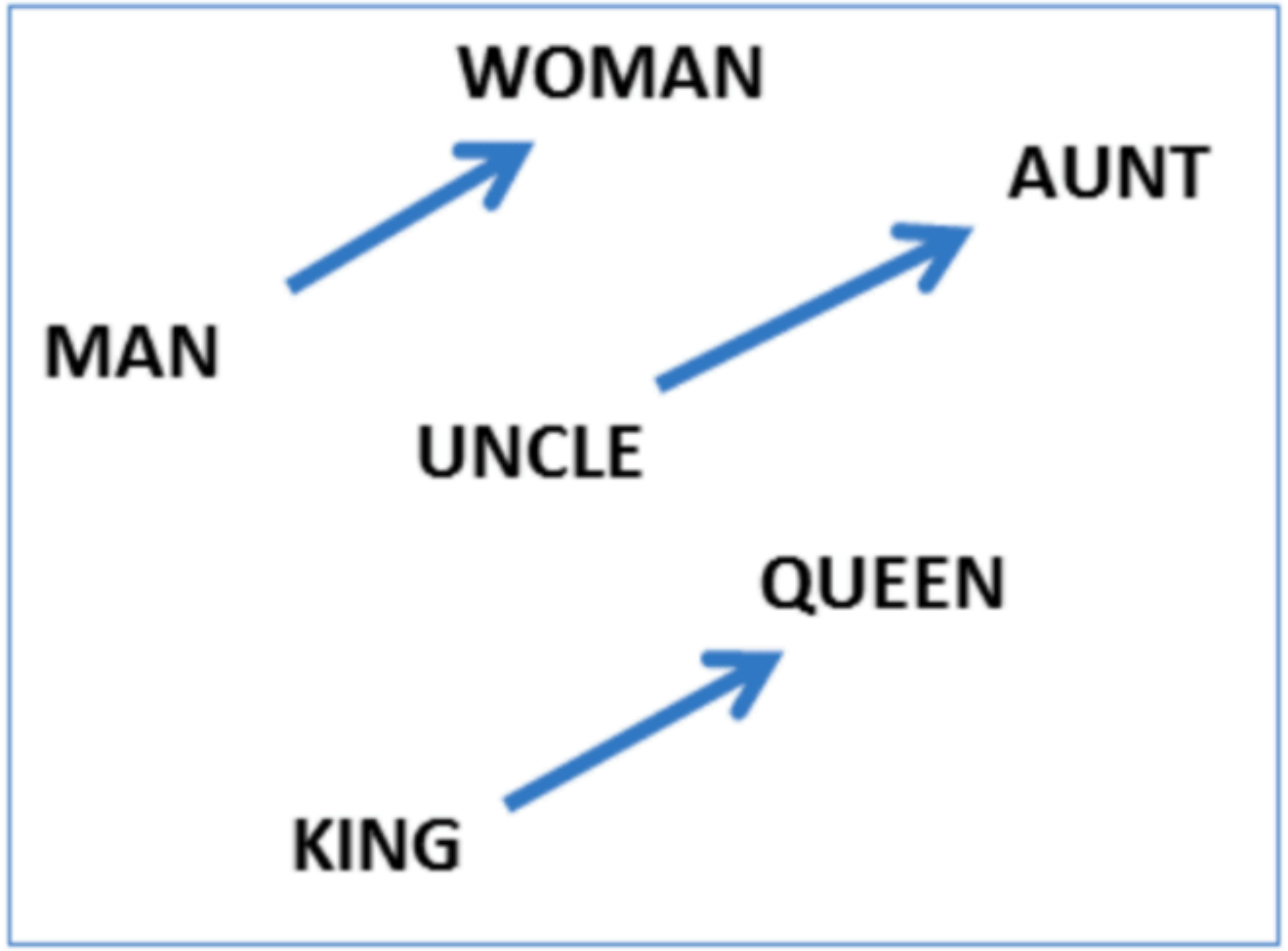
The structure in embeddings

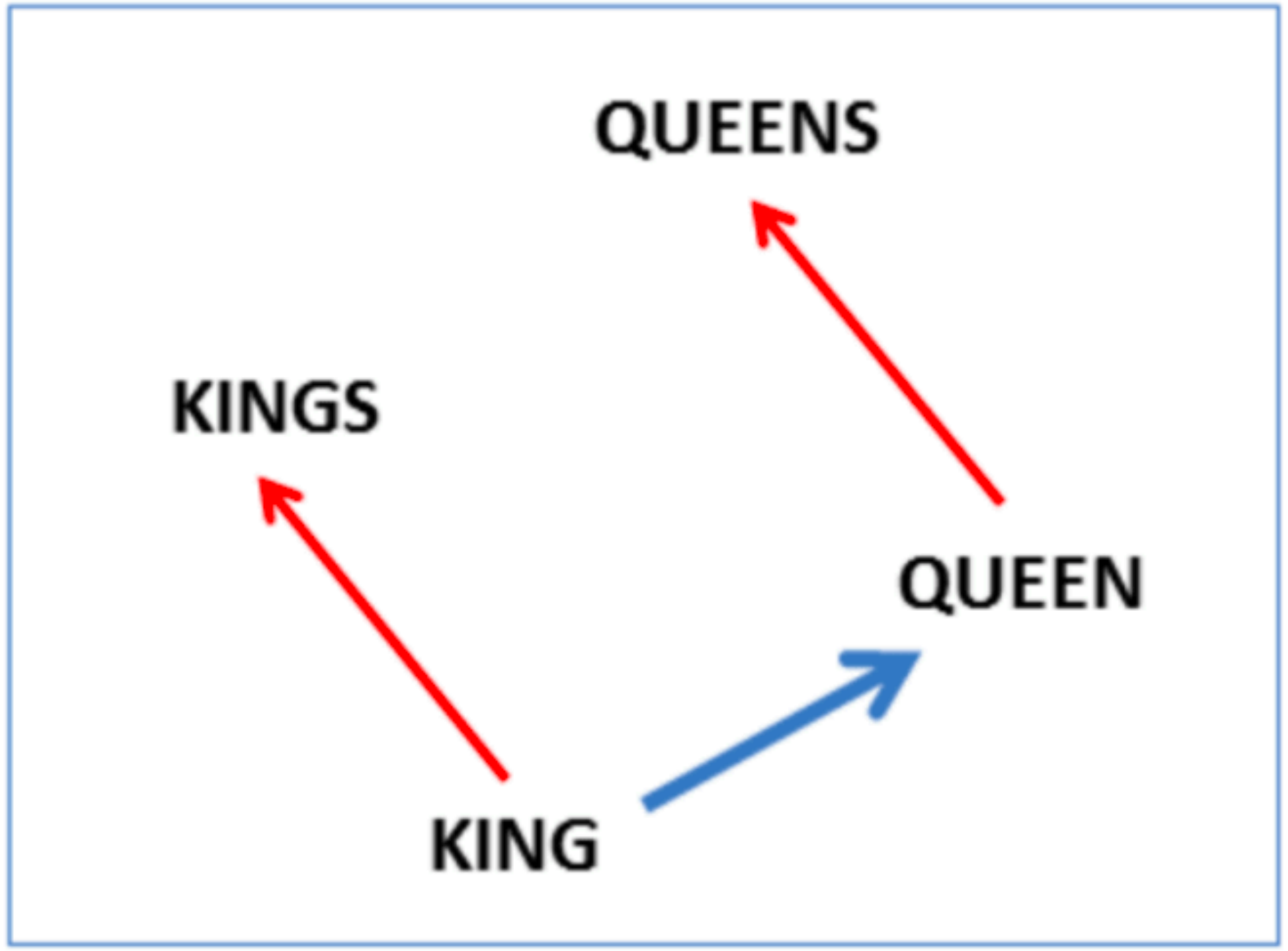


Embeddings encode both semantic and syntactic relationships

semantic: $v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen})$

syntactic: $v(\text{kings}) - v(\text{king}) + v(\text{queen}) \approx v(\text{queens})$

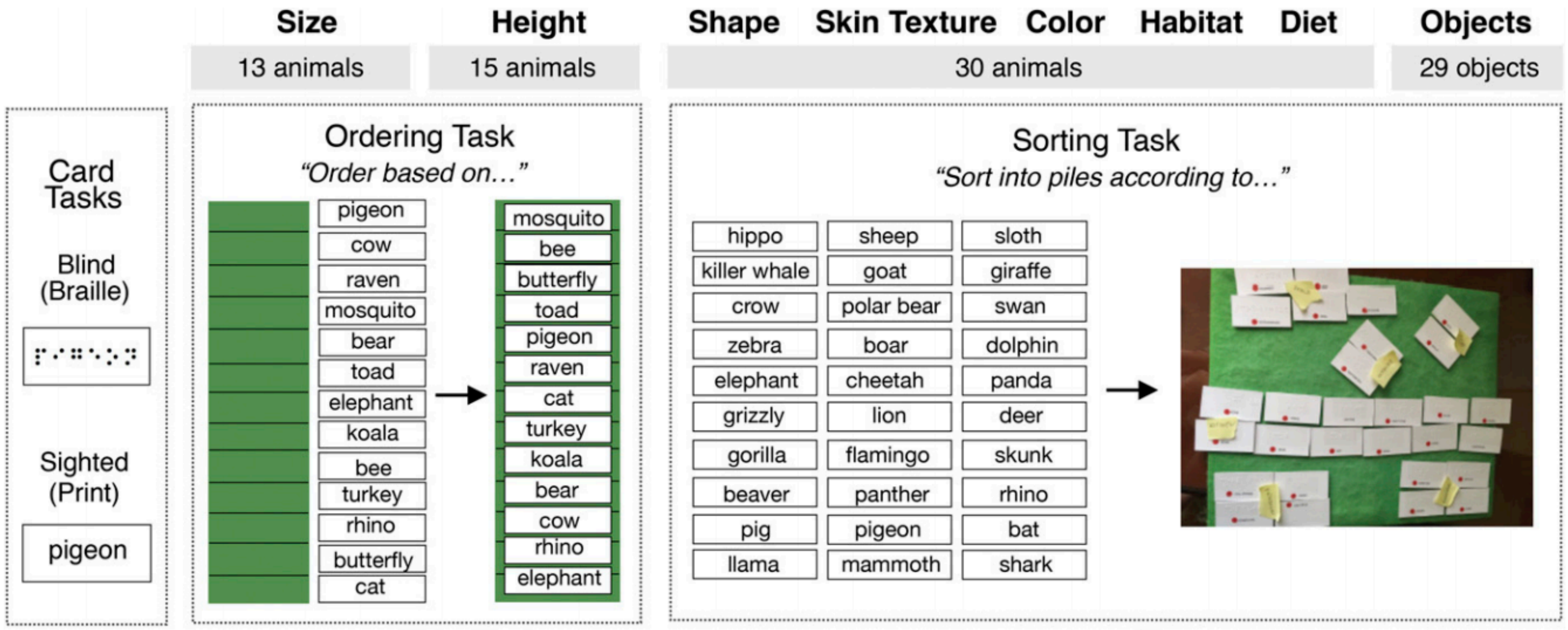




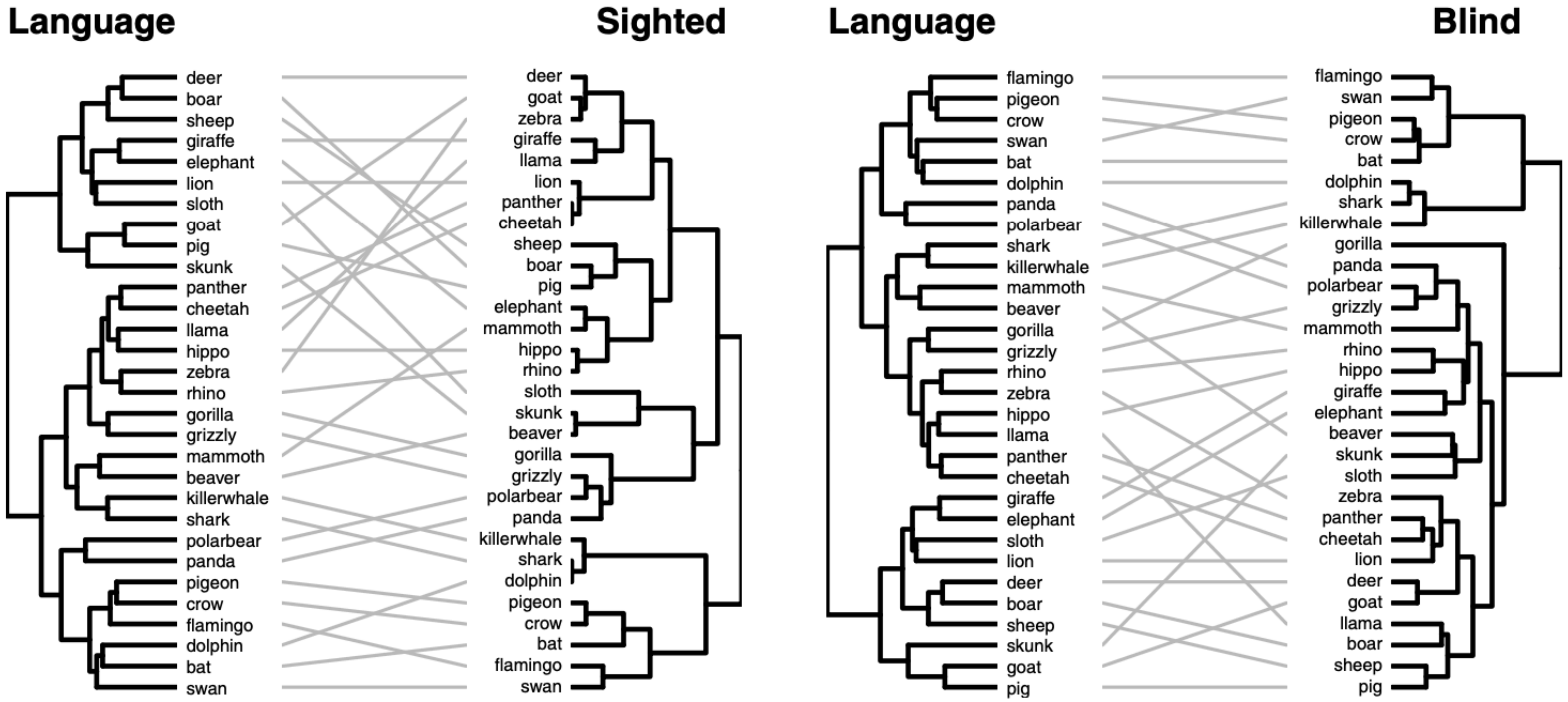
Exploring embedding models

http://vectors.nlpl.eu/explore/embeddings/en/

Embedding similarities predict human similarity judgments

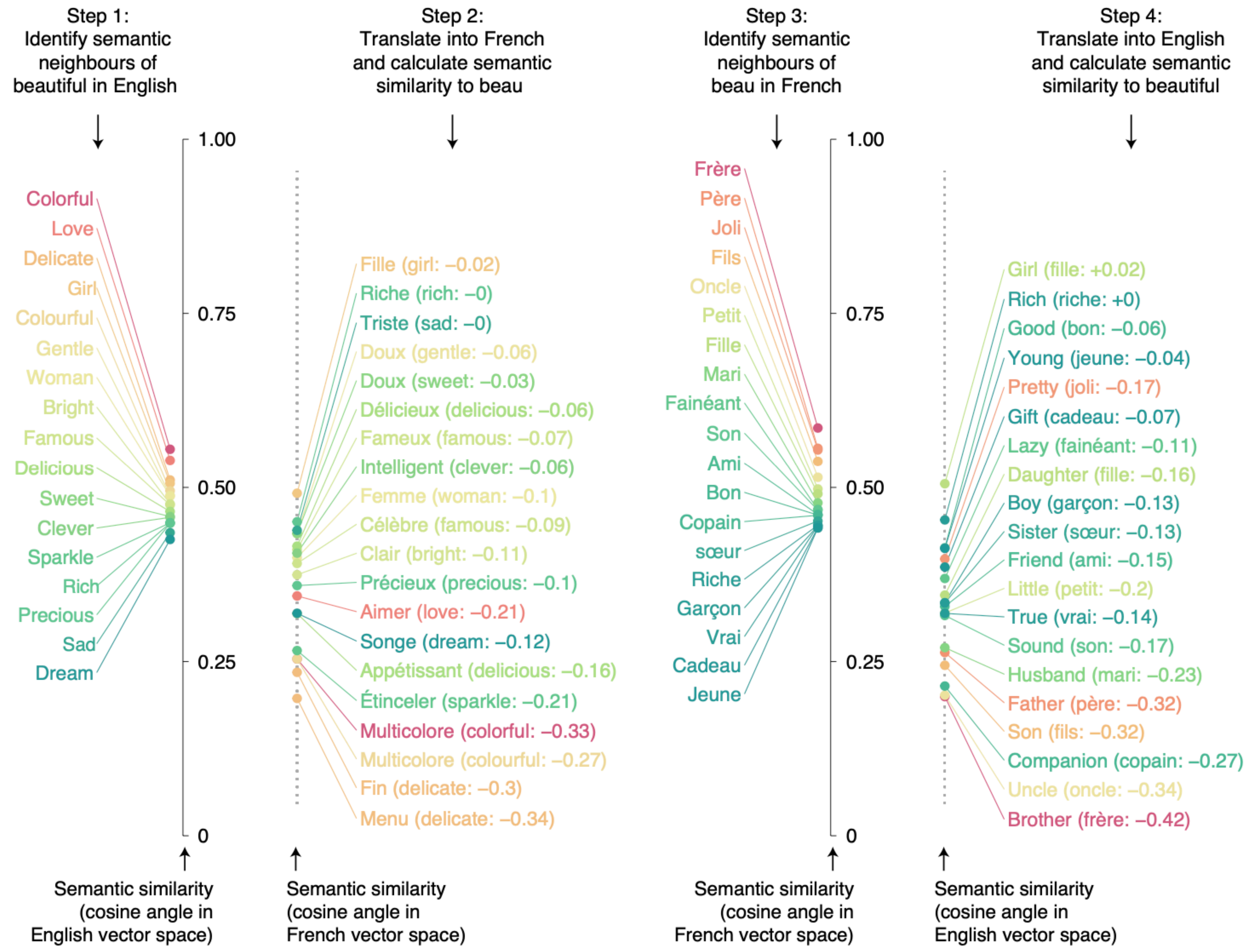


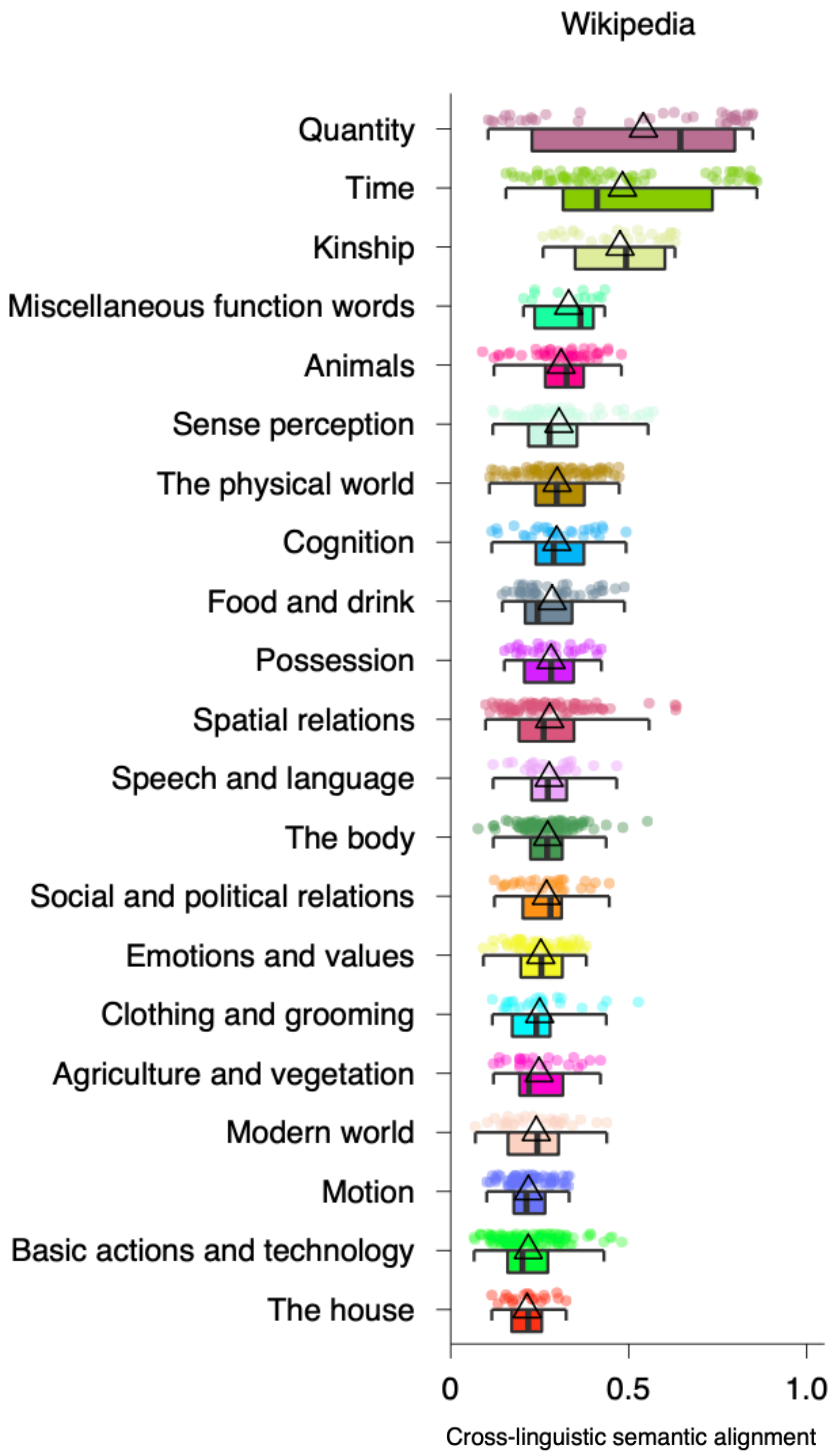
Kim, Elli, & Bedny (2019)



Lewis, Zettersten, & Lupyan (2019)

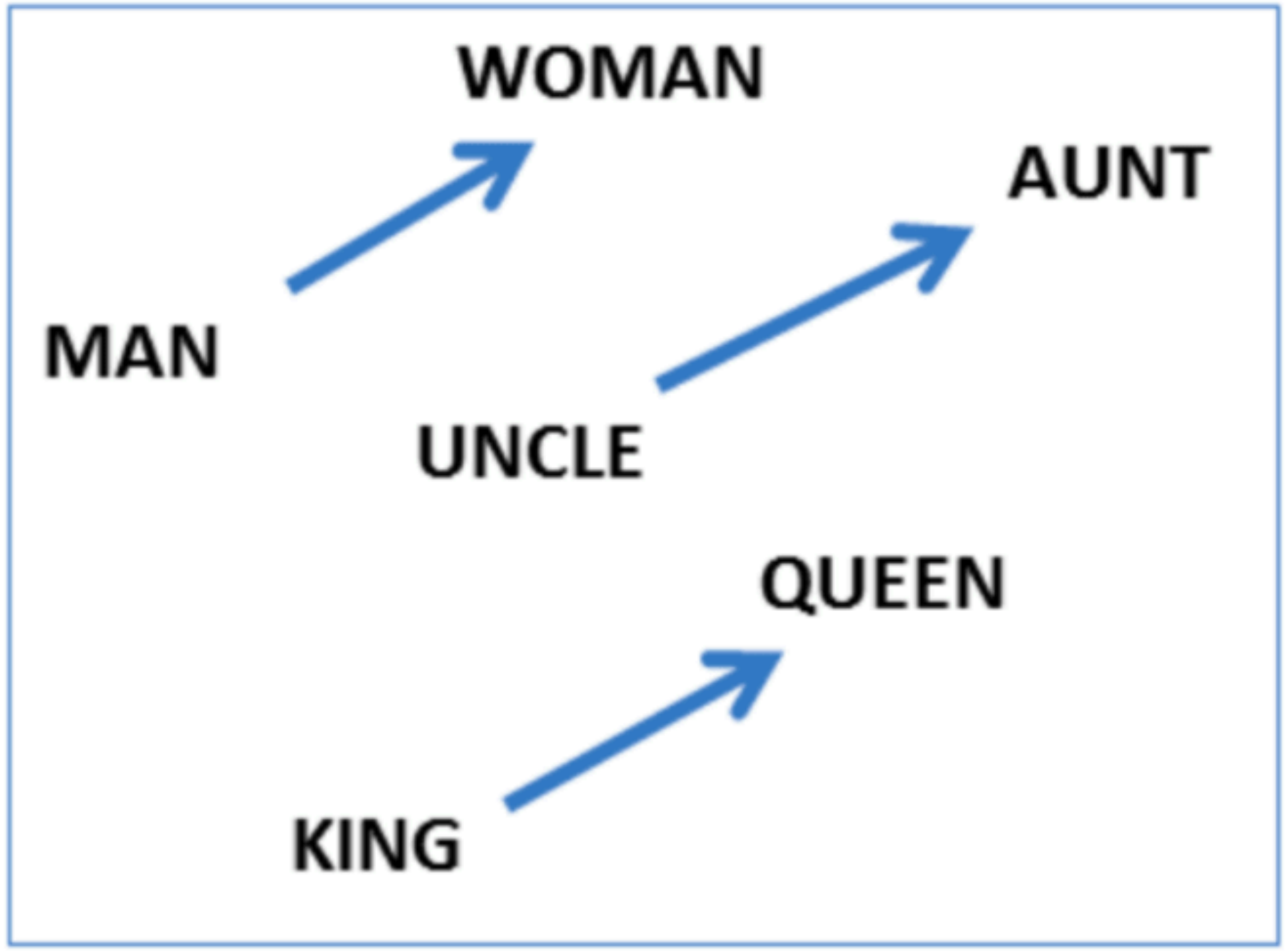
Using embeddings to estimate translatability (Thompson, Roberts, & Lupyan, 2020)

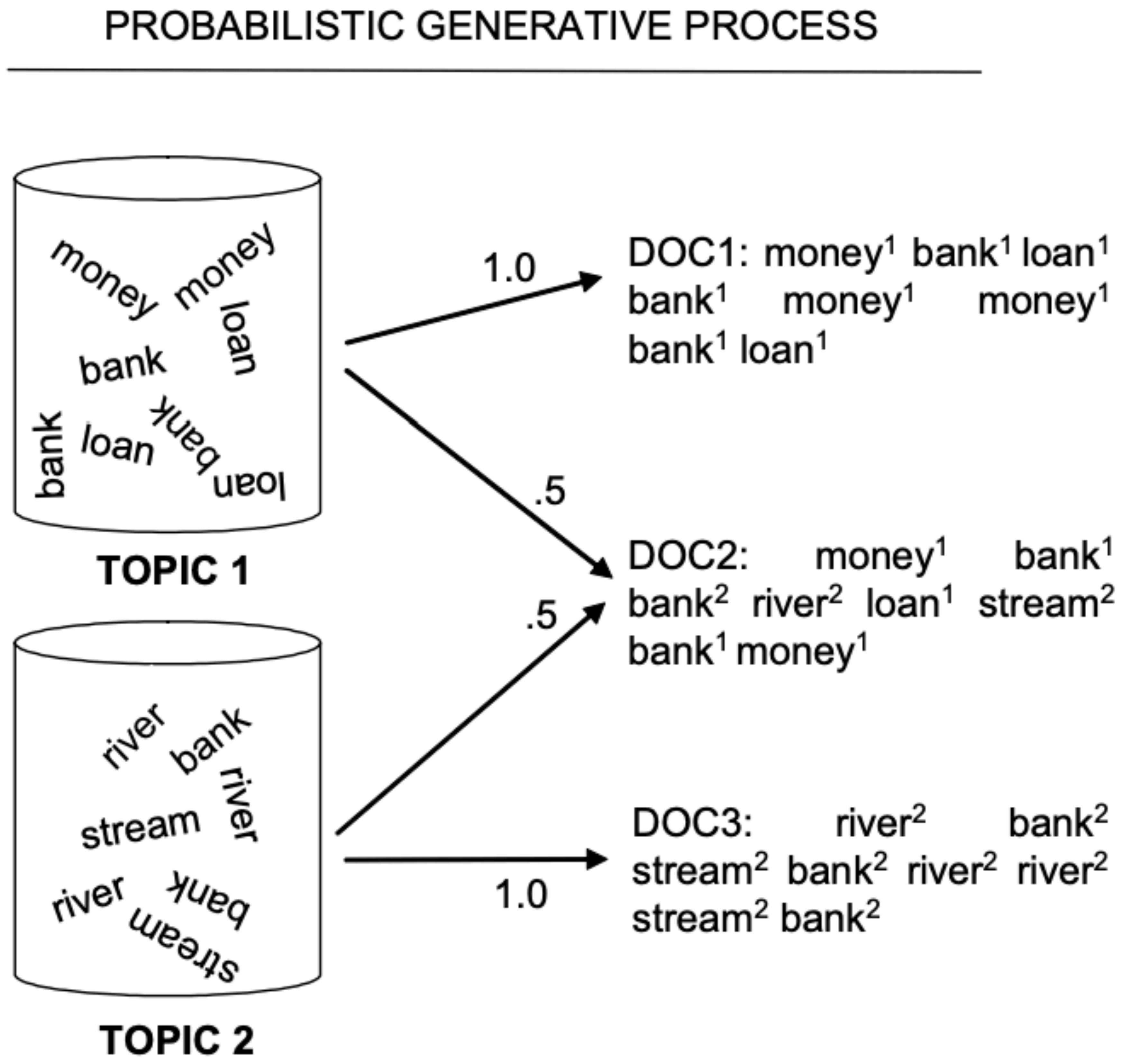




The problem with "meaning"

v(**king**) – v(**man**) + v(**woman**) ≈ v(**queen**)

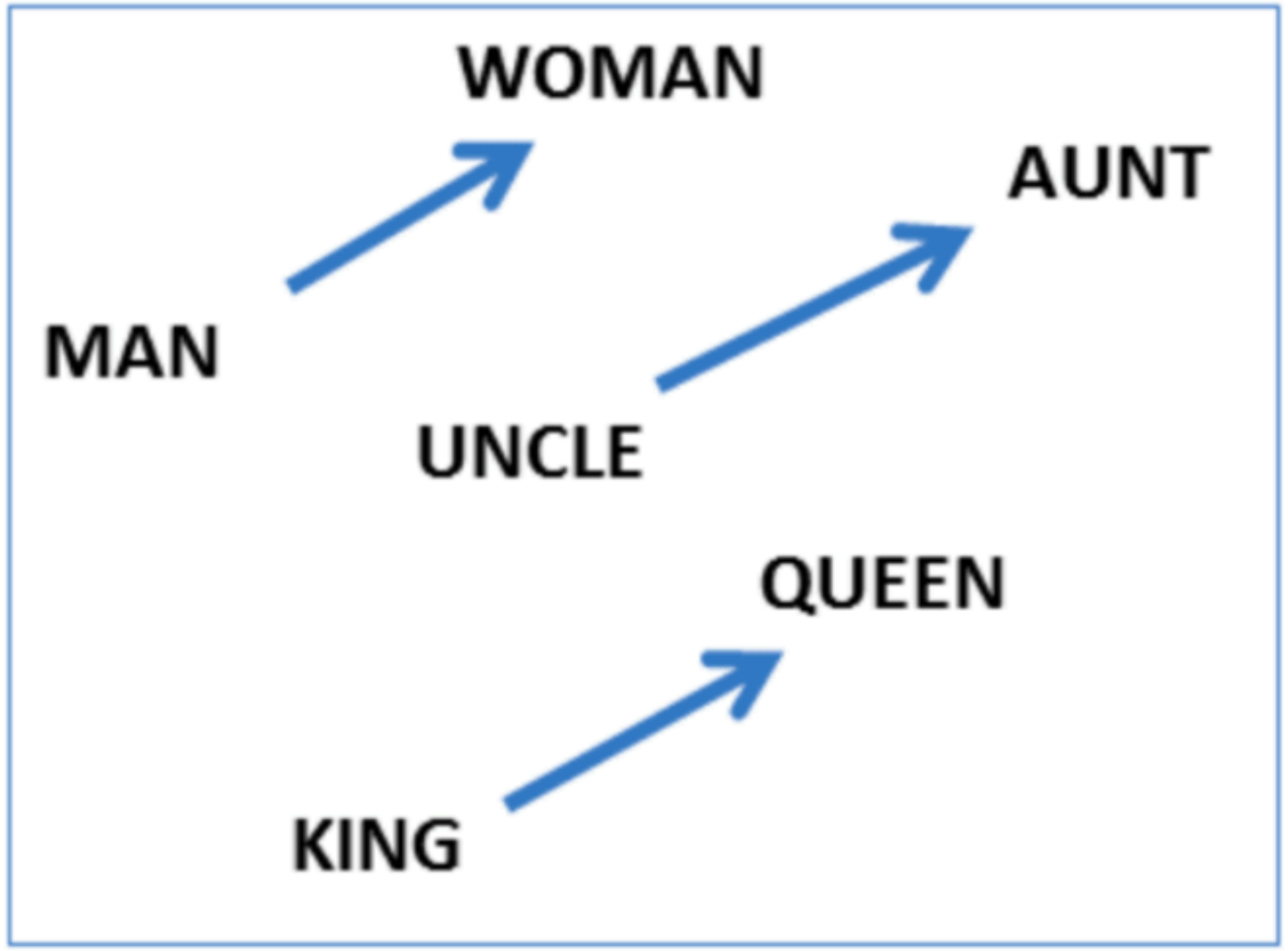


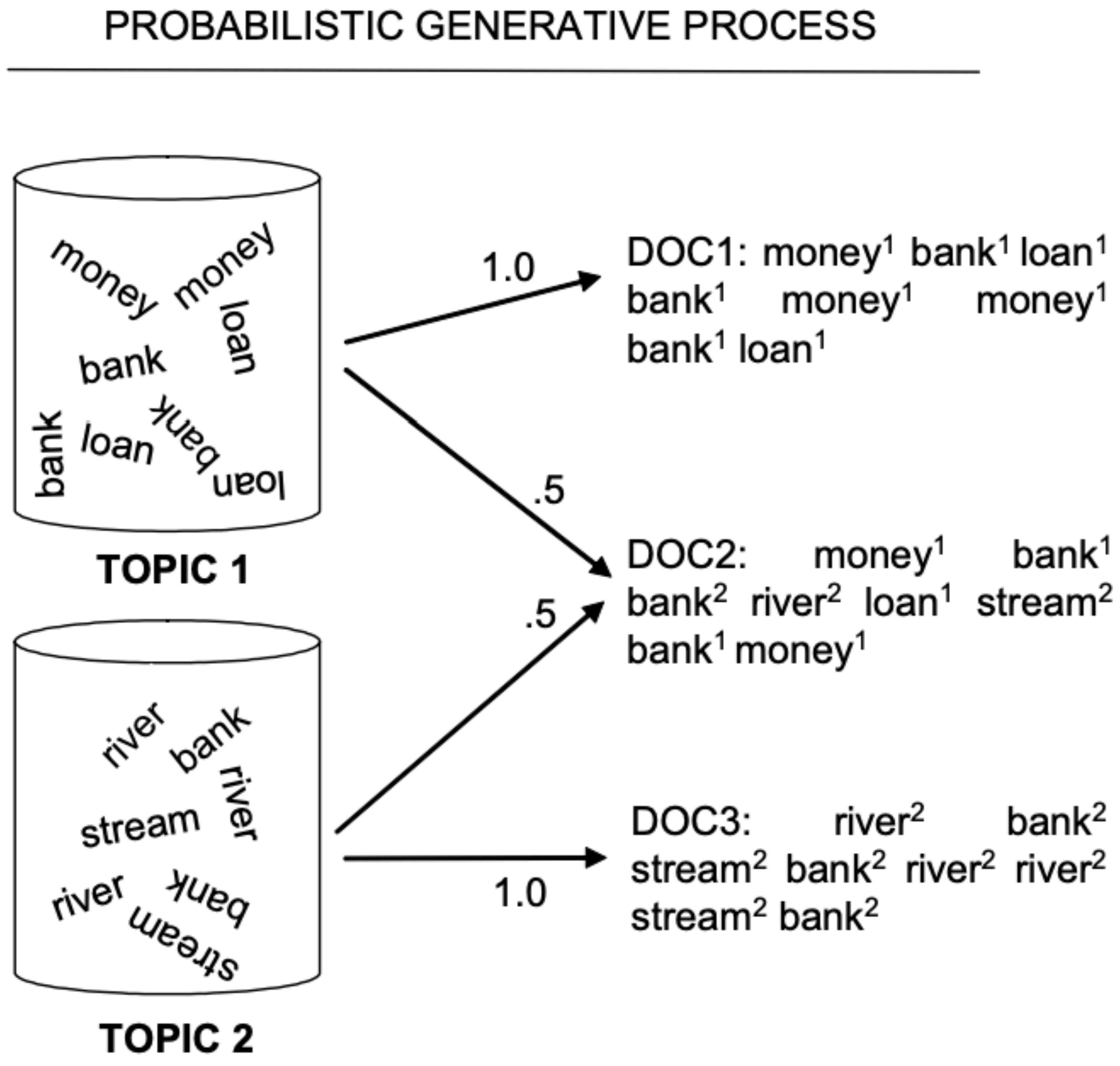


What about **big**. Or **red**. Or **monster**.

The problem with "meaning"

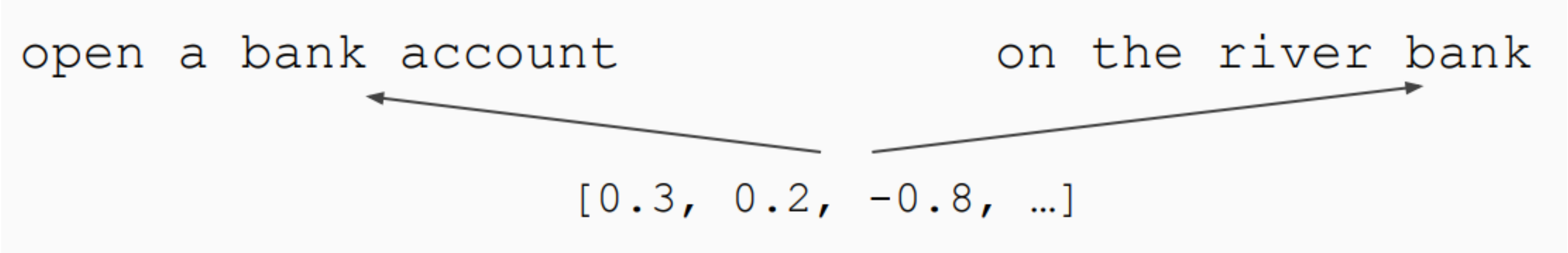
$$v(\textbf{king}) - v(\textbf{man}) + v(\textbf{woman}) \approx v(\textbf{queen})$$





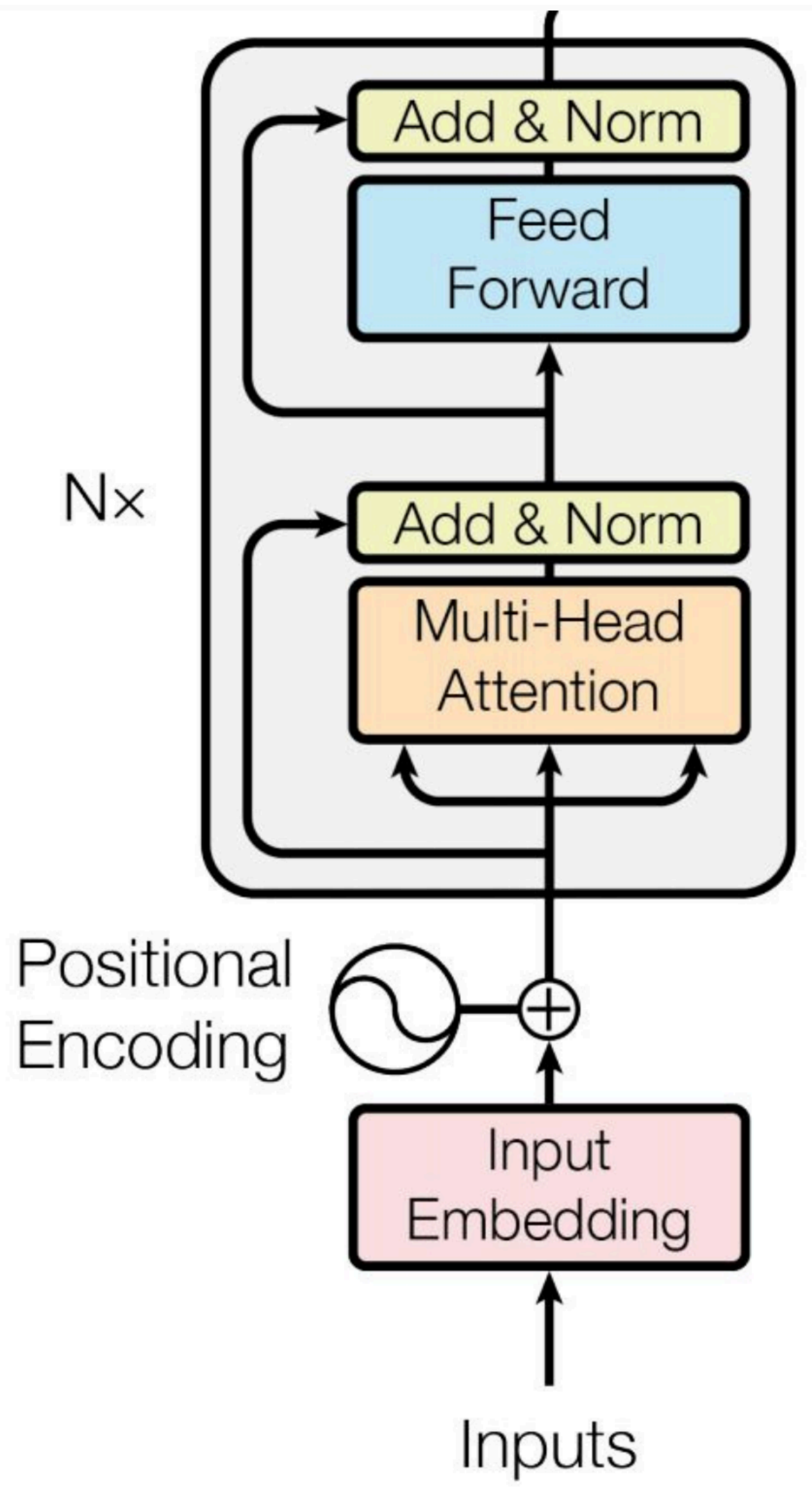
What about **big**. Or **red**. Or **monster**.

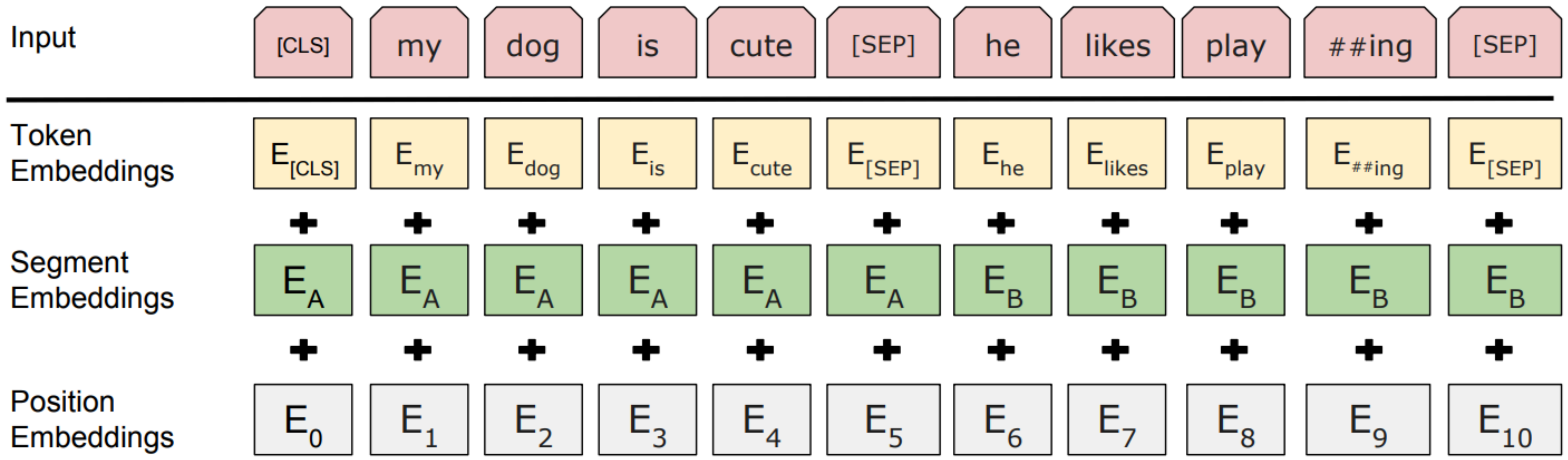
Word2Vec/Glove embeddings vs. Contextual embeddings



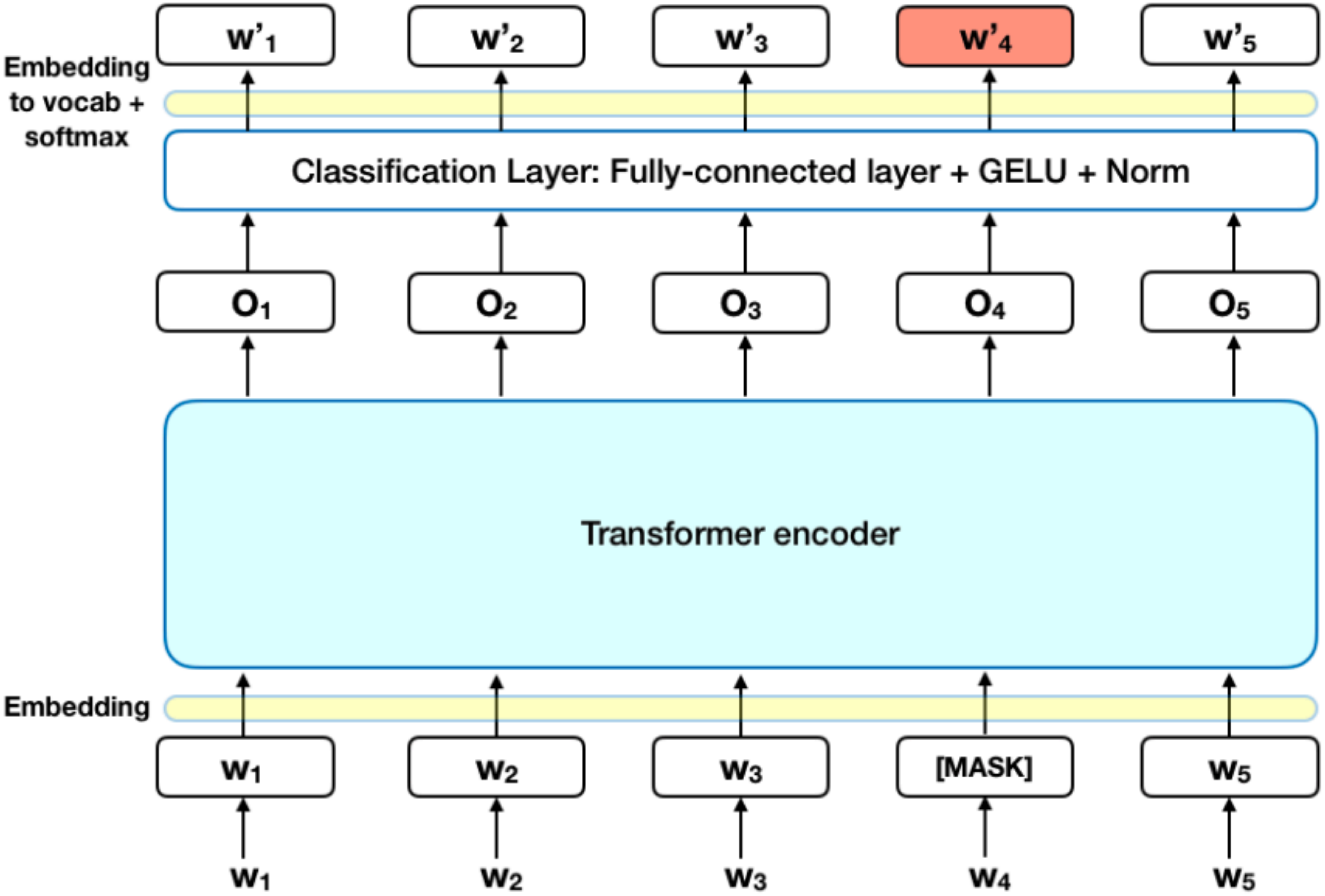
Adapted from Jacob Devlin

Bidirectional Transformers for Language Understanding (BERT - Devlin et al., 2018)

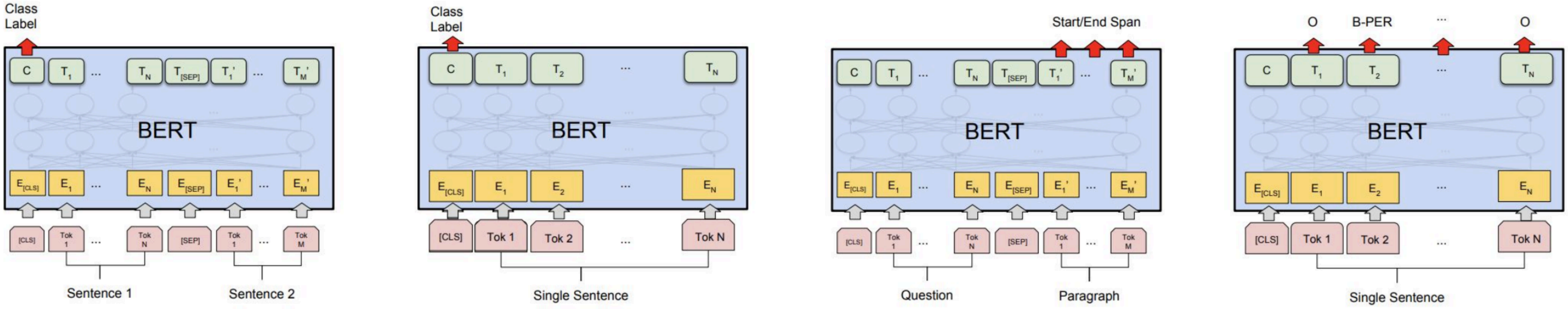




Training to predict masked words



Fine-tuning for individual tasks



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT$_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT$_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

BERT demo

https://demo.allennlp.org/masked-lm

Modern language models

1. Embedding models are a general class of models for representing meaning in a vector-space

2. Embedding models can be used to understand aspects of cognition and language

3. The leading edge of models don’t represent “meaning” anymore at all