

# Representações distribuídas de palavras

*vector space models ou embeddings*

Prof. Dr. Ricardo Primi



# Tópicos

- Visão geral do pipeline
- Matriz de dados
- Representações vetoriais de palavras
- SkipGram e word2vec
- Word embeddings

# Great power, a great many design choices

tokenization

annotation

tagging

parsing

feature selection

: cluster texts by date/author/discourse context/...



Matrix design	Reweighting	Dimensionality reduction	Vector comparison
word × document	probabilities	LSA	Euclidean
word × word	length norm.	PLSA	Cosine
word × search proximity	TF-IDF	LDA	Dice
adj. × modified noun	PMI	PCA	Jaccard
word × dependency rel.	Positive PMI	NNMF	KL
:	:	:	:

Nearly the full cross-product to explore; only a handful of the combinations are ruled out mathematically. Models like GloVe and word2vec offer packaged solutions to design/weighting/reduction and reduce the importance of the choice of comparison method. Contextual embeddings dictate many preprocessing choices.

# word x word

	:)	:/	:D	:	;p	abandon	abc	ability	able	...
:)	74	1	0	0	0	1	0	2	2	
:/	1	306	0	0	0	0	0	0	17	
:D	0	0	16	0	0	0	6	1	1	
:	0	0	0	120	0	0	0	1	9	
;p	0	0	0	0	516286	0	0	0	0	...
abandon	1	0	0	0	0	370	24	65	235	
abc	0	0	6	0	0	24	7948	77	291	
ability	2	0	1	1	0	65	77	4820	1807	
able	2	17	1	9	0	235	291	1807	14328	
:					:					
:					:					

# word x document

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

## word x discourse context

Upper left corner of an interjection x dialog-act tag matrix  
derived from the Switchboard Dialog Act Corpus:

	Reject-part	Hedge	Completion	Tag question	Hold	Quotation	Accept	...
absolutely	0	2	0	0	0	0	95	
actually	17	12	0	0	1	0	4	
anyway	23	14	0	0	0	0	0	
boy	5	3	1	0	5	2	1	
bye	0	1	0	0	0	0	0	
bye-bye	0	0	0	0	0	0	0	...
dear	0	0	0	0	1	0	0	
definitely	0	2	0	0	0	0	56	
exactly	2	6	1	0	0	0	294	
gee	0	3	0	0	2	1	1	
goodness	1	0	0	0	2	0	0	
:	:			:				

## Other designs

- adj. × modified noun
  - word × syntactic context
  - word × search query
  - person × product
  - word × person
  - word × word × pattern
  - verb × subject × object
- :

## Meaning latent in co-occurrence patterns

Class	Word
neg	awful
neg	terrible
neg	lame
neg	worst
neg	disappointing
pos	nice
pos	amazing
pos	wonderful
pos	good
pos	awesome

A hopeless learning scenario

Latent meaning      High-level goals      Guiding hypotheses      Design choices

## Meaning latent in co-occurrence patterns

Class	Word	excellent	terrible	Pr(Class=pos)	Word	excellent	terrible
neg	awful	6	113				
neg	terrible	8	309				
neg	lame	1	69				
neg	worst	9	202				
neg	disappointing	19	29				
pos	nice	118	2				
pos	amazing	91	6	≈0	$w_1$	4	82
pos	wonderful	66	7	≈0	$w_2$	5	84
pos	good	21	9	≈1	$w_3$	49	3
pos	awesome	67	2	≈1	$w_4$	41	5

A promising learning scenario

# Windows and scaling: What is a co-occurrence?

from swerve of shore **to** bend of bay , brings

4	3	2	1	0	1	2	3	4	5
---	---	---	---	---	---	---	---	---	---

from swerve of shore **to** bend of bay , brings

**Window: 3** 4 3 2 1 0 1 2 3 4 5

**Scaling: flat** 0 1 1 1 1 1 1 1 0 0

## A Model for Analogical Reasoning<sup>1</sup>

DAVID E. RUMELHART AND ADELE A. ABRAHAMSON

*University of California, San Diego*

A theory of analogical reasoning is proposed in which the elements of a set of concepts, e.g., animals, are represented as points in a multidimensional Euclidean space. Four elements A,B,C,D, are in an analogical relationship  $A:B::C:D$  if the vector distance from A to B is the same as that from C to D. Given three elements A,B,C, an ideal solution point I for  $A:B::C:?$  exists. In a problem  $A:B::C:D_1, \dots, D_i, \dots, D_n$ , the probability of choosing  $D_i$  as the best solution is a monotonic decreasing function of the absolute distance of  $D_i$  from I. A stronger decision rule incorporating a negative exponential function in Luce's choice rule is also proposed. Both the strong and weak versions of the theory were supported in two experiments where Ss rank-ordered the alternatives in problems  $A:B::C:D_1, D_2, D_3, D_4$ . In a third experiment the theory was applied and further tested in teaching new concepts by analogy.

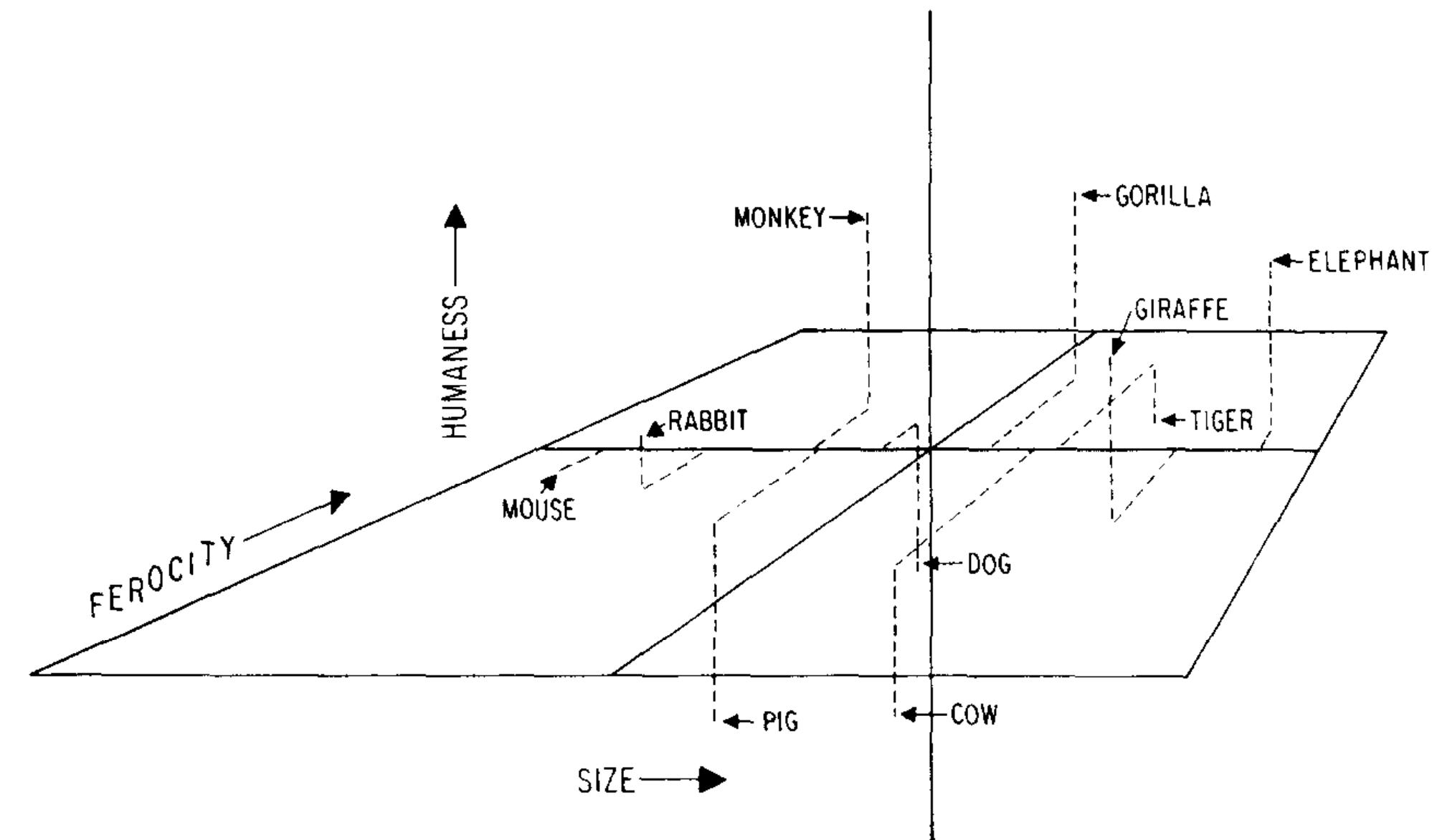


FIG. 1. The placements of a selected set of animals based on data from Henley (1969).

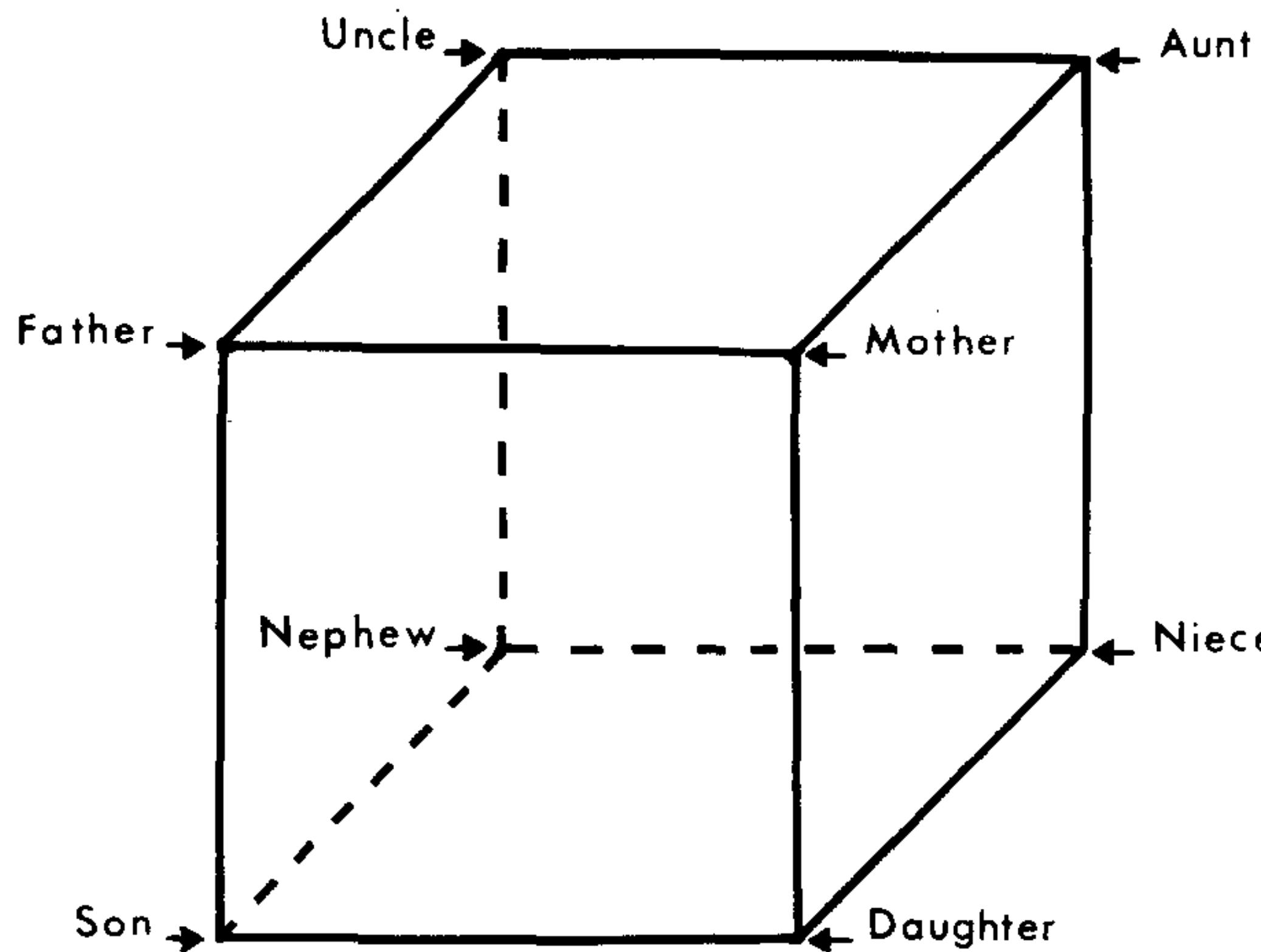


FIG. 2. Three-dimensional representation of relations among eight kinship terms.

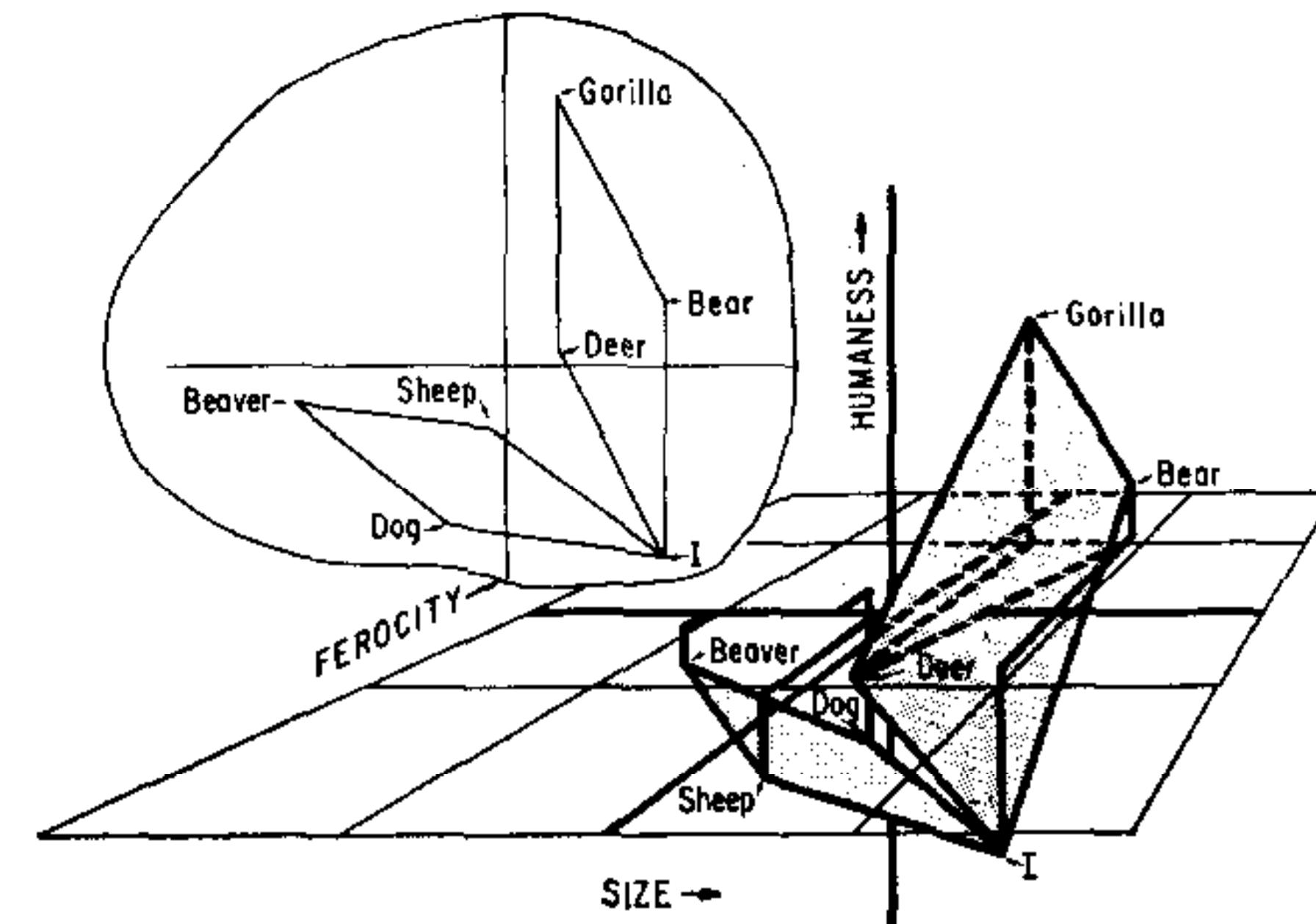
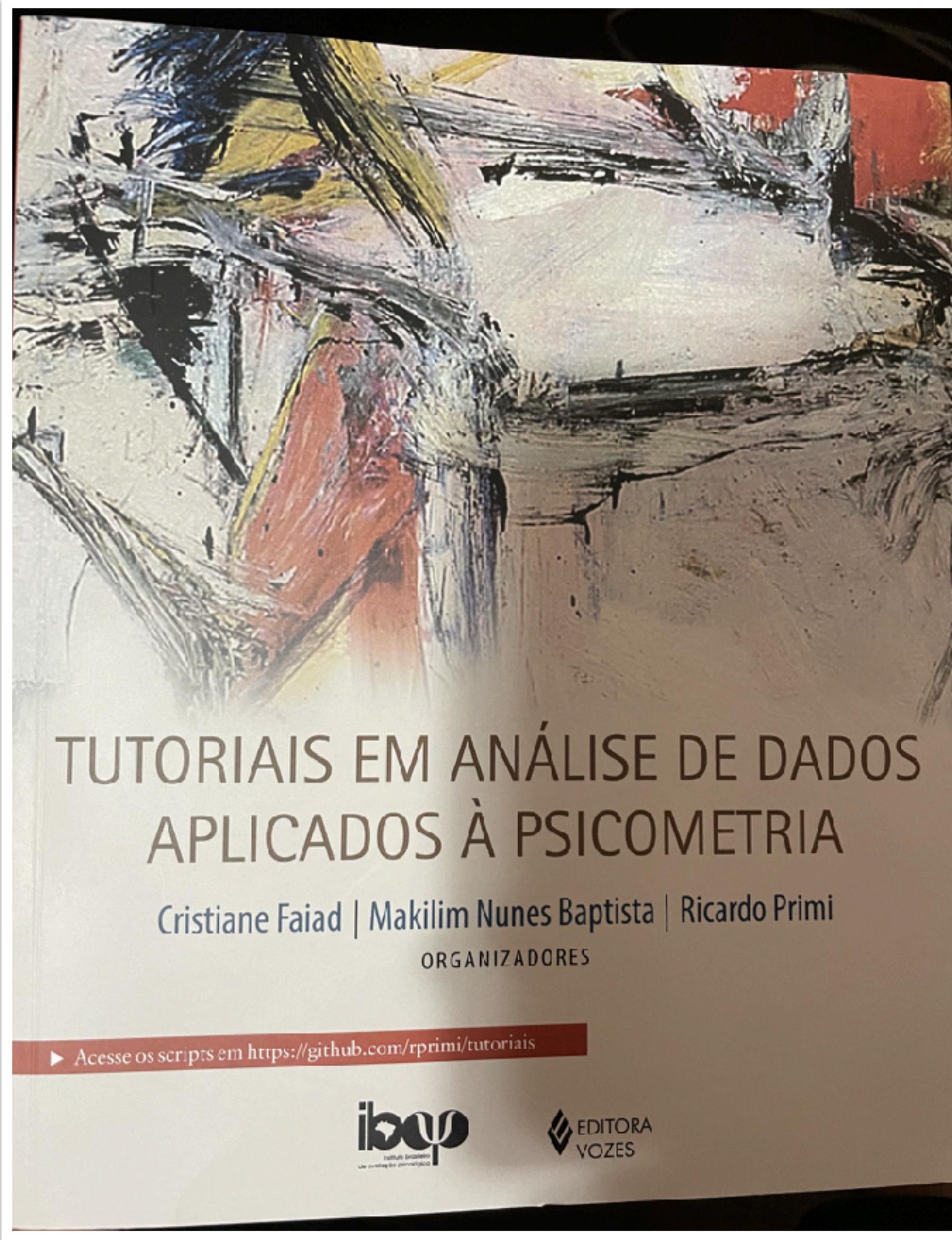


FIG. 5. Three-dimensional representation of two analogies with the same solution. The analogies are (a) GORILLA:DEER::BEAR:? and (b) BEAVER:SHEEP::DOG?:.



23  
Uso do *word-to-vec* (*word embeddings*) para análise de textos

*Ricardo Primi*  
Universidade São Francisco

O objeto de estudo da psicologia é mediado por conteúdos verbais. As teorias psicológicas foram construídas a partir de análise dos discursos de pacientes sobre suas introspecções. A teoria fatorial da personalidade, que define os traços básicos, foi construída empregando-se a análise fatorial para agrupar palavras que descreviam características pessoais. Assim a análise dos discursos verbais e a análise fatorial são métodos bastante importantes na psicologia. Mas será que existe algum método sistemático para analisar textos ou discursos diretamente?

Com o avanço recente do processamento em linguagem natural temos, à nossa disposição, um conjunto de métodos de análise de textos. Mas esses métodos são pouco explorados na psicologia. Este capítulo irá explorar o uso *word-to-vec* (vetores de palavras ou *word embeddings*) para análise fatorial de instrumentos de autorrelato. Vetores de palavras são representações numéricas que codificam as informações semânticas das palavras. Este capítulo fará uma breve definição dos vetores de palavras e, em seguida, apresentará um tutorial explicando como usar os vetores de palavras para realizar uma classificação dos questionários em que as pessoas são as próprias observadoras e relatores de seu comportamento.

do verbal, sem a necessidade de coletar dados de pessoas respondendo ao teste. Essa aplicação poderá ser estendida para uma quantidade grande de contextos na psicologia.

**Abordagem fatorial da personalidade: Fontes de dados e a hipótese léxica**

Na década de 1930, os psicólogos iniciaram estudos de identificação das características fundamentais que definem nossa individualidade. Buscaram definir uma taxonomia natural de descriptores das características de personalidade a partir das palavras do dicionário. Essa abordagem foi chamada de hipótese léxica. Ela é baseada no princípio de que "as características sociais de personalidade mais salientes foram codificadas na linguagem natural" (John, Naumann & Soto, 2008, p. 117).

Raymond Cattell (1973) argumentou que se uma determinada característica for um atributo fundamental da personalidade, ela se expressa de várias maneiras. Segundo esse raciocínio, classificou três tipos de dados: (a) respostas a questionários em que as pessoas são as próprias observadoras e relatores de seu comportamento

□ □ **O** bigode é a antena do gato  
□ **O** bigode é a antena do gato  
**O** bigode é a antena do gato  
**O** bigode é **a** antena do gato  
**O** bigode é a **antena** do gato  
**O** bigode é a antena **do** gato □  
**O** bigode é a antena do **gato** □ □

Figura 2. Exemplo do método de varredura *skip-gram* de  $k = 2$

O modelo prevê a probabilidade de observar as palavras do contexto (*outside words o*) dada a observação da palavra central (*center word c*):  $P(o|c)$ . Isto é, o modelo prevê qual palavra será mais provável de ocorrer antes ou depois de uma palavra central. O modelo é dado por:

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^W \exp(u_w^T v_c)}$$

Sendo  $u_o$  o vetor da palavra de contexto e  $v_c$  o vetor da palavra central. Note que esses vetores terão dimensão  $d$  especificada previamente à análise. Eles são vetores de variáveis latentes que irão representar cada palavra. Para simplificar, imagine que tenhamos especificado  $d = 3$ . A notação  $u_o^T v_c$  indica o produto cruzado dos vetores. Geralmente os vetores são dispostos em colunas, o símbolo  $T$  em  $u_o^T$  indica que o primeiro vetor é transposto, isso é, é transformado de um vetor-coluna para vetor-linha. Assim se  $u_o = (1, .8, 0)$  e  $u_c = (.9, .2, .4)$ , a multiplicação cruzada será:

$$u_o^T v_c = (1 \quad .8 \quad 0) \times \begin{pmatrix} .9 \\ .2 \\ .4 \end{pmatrix} = (1 \times .9 + .8 \times .2 + 0 \times .4) = 1.06$$

vetores para as palavras de nosso vocabulário (esses valores foram inventados para ilustrar o modelo):

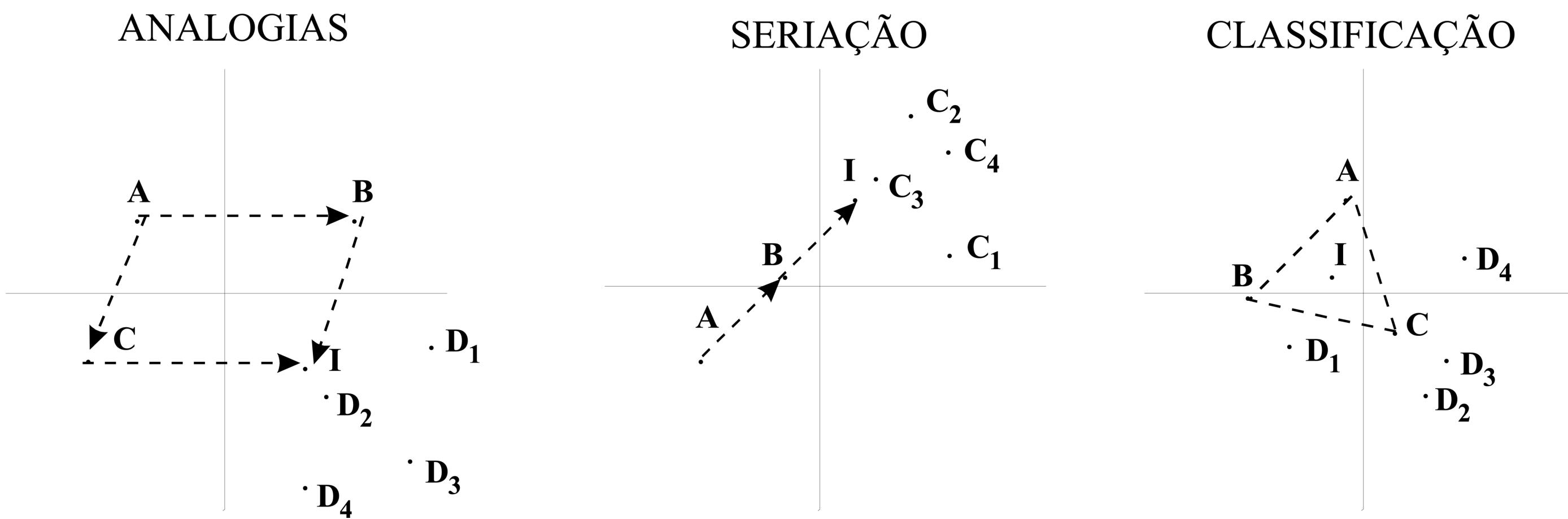
Tabela 1. *Vetores das 15 palavras do corpus de três sentenças*

vocab	d1	d2	d3
o	0.17	0.94	0.49
bigode	0.89	0.20	0.14
é	0.84	0.11	0.21
a	0.27	0.04	0.23
antena	0.94	0.18	0.12
do	0.39	0.12	0.72
gato	0.90	0.21	0.12
cachorro	0.03	0.00	0.01
um	0.63	0.71	0.17
lobo	0.63	0.19	0.59
mais	0.20	0.05	0.75
manso	0.01	0.72	0.74
cegonha	0.20	0.20	0.20
girafa	0.61	0.13	0.77
ganso	0.59	0.89	0.04

A probabilidade de observar a palavra “antena” (palavra do contexto) dado que estejamos vendo a palavra “gato” (palavra central) é função da soma dos produtos dos vetores dessas duas palavras:

$$P(\text{antena}|\text{gato}) = \frac{\exp(u_{\text{antena}}^T v_{\text{gato}})}{\sum_{w=1}^W \exp(u_w^T v_{\text{gato}})} = \frac{2.44}{26.83} = .09$$

Comparando esse resultado com probabilidade de observar a palavra “cachorro” tendo visto a palavra “gato” resultará em uma probabilidade menor  $P(\text{cachorro}|\text{gato}) = .03$  já que, nesse corpus, essas duas palavras não aparecem juntas na mesma frase.



*Figura 1.* Exemplo de representação dos termos das tarefas de raciocínio indutivo em um espaço euclidiano bidimensional.

□ □ **O** bigode é a antena do gato  
□ **O** **bigode** é a antena do gato  
**O** bigode é a **antena** do gato  
**O** bigode é **a** **antena** **do** gato  
**O** bigode é **a** **antena** **do** gato  
**O** bigode é **a** **antena** **do** gato  
**O** bigode é a **antena** **do** **gato** □ □

*Figura 2.* Exemplo do método de varredura *skip-gram* de  $k = 2$

## Problems with this discrete representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: **hotel, conference, walk**

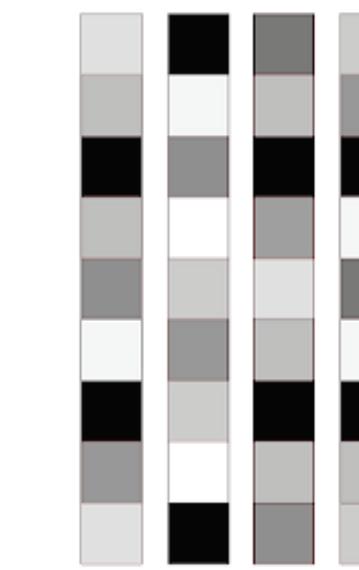
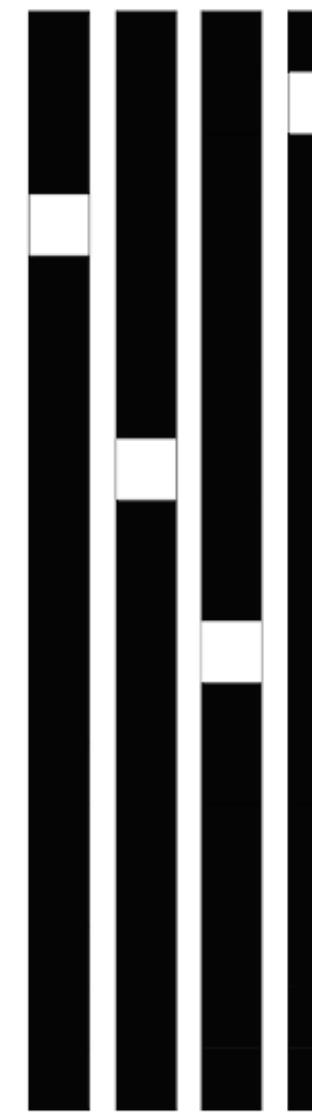
In vector space terms, this is a vector with one 1 and a lot of zeroes

**[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]**

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a “one-hot” representation. Its problem:

**motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND  
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0**

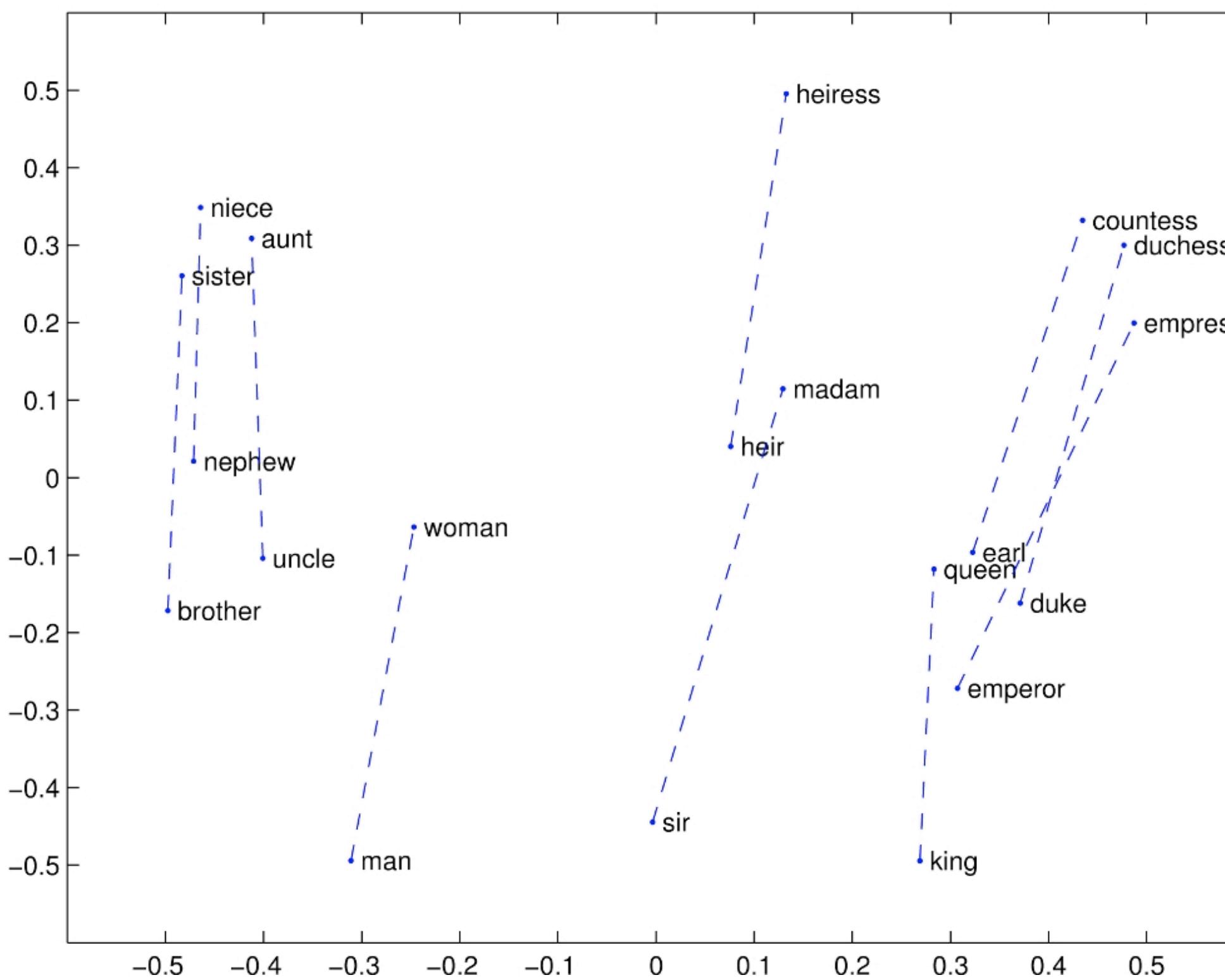


One-hot word vectors: Word embeddings:

- Sparse
- High-dimensional
- Hard-coded
- Dense
- Lower-dimensional
- Learned from data

**Figure 6.2** Whereas word representations obtained from one-hot encoding or hashing are sparse, high-dimensional, and hardcoded, word embeddings are dense, relatively low-dimensional, and learned from data.

# Glove Visualizations



<http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc#>

Início Downloads Como Utilizar

## Repositório de Word Embeddings do NILC

NILC - Núcleo Interinstitucional de Linguística Computacional

### Introdução

NILC-Embeddings é um repositório destinado ao armazenamento e compartilhamento de vetores de palavras (do inglês, word embeddings) gerados para a Língua Portuguesa. O objetivo é fomentar e tornar acessível recursos vetoriais prontos para serem utilizados nas tarefas de Processamento da Linguagem Natural e Aprendizado de Máquina. O repositório traz vetores gerados a partir de um grande córpus do português do Brasil e português europeu, ce fontes e gêneros variados. Foram utilizados dezessete córpuses diferentes, totalizando 1,395,926,282 tokens. O treinamento dos vetores ocorreu em algoritmos como Word2vec [1], FastText [2], Wang2vec [3] e Glove [4]. Mais detalhes sobre o projeto podem ser encontrados em: [Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks](#).

### Artigo produzido

Esse trabalho produziu um artigo aceito para publicação no [STIL 2017 -- Symposium in Information and Human Language Technology](#). Os anais do evento estão disponíveis [aqui](#). A versão preprint do artigo pode ser visto [aqui](#).

### Download Scripts Pré-processamento e Scripts de Avaliação

Os scripts utilizados para pré-processamento dos dados, bem como os scripts para as avaliações realizadas, estão disponíveis para [download](#).

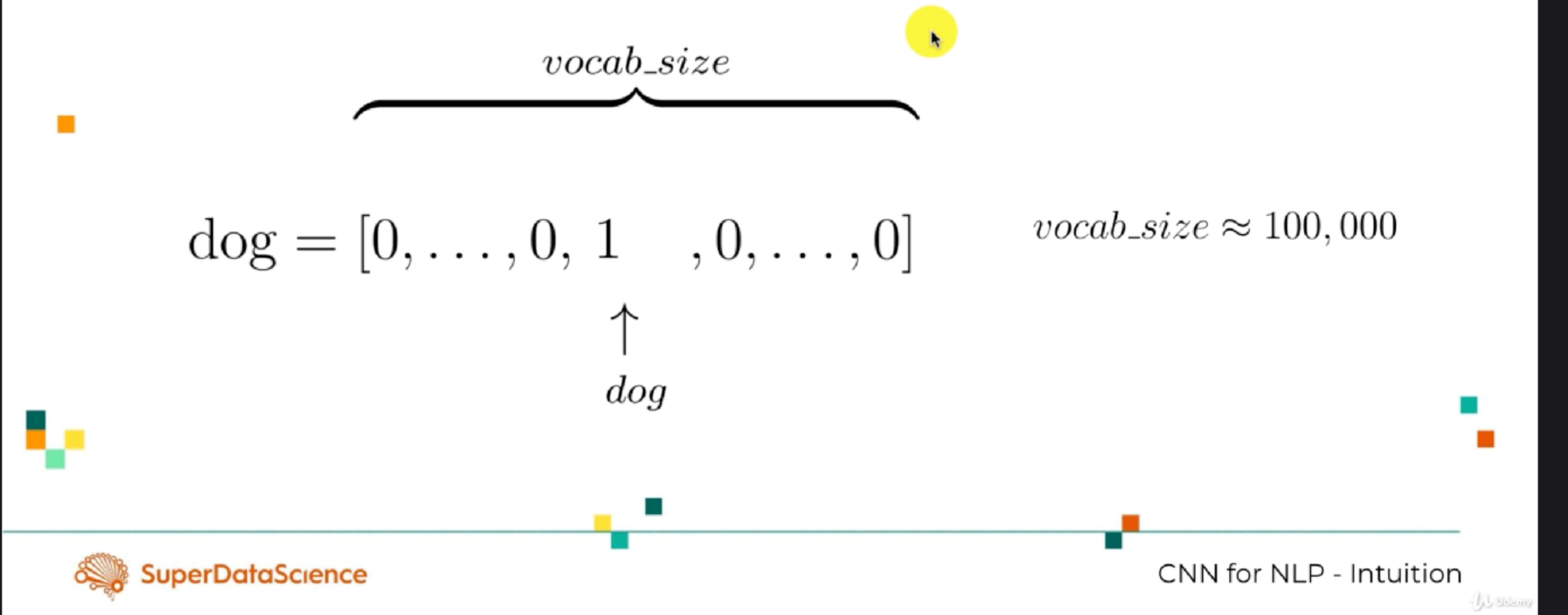
### Download Word Embeddings Pré-treinadas

Para cada modelo, foram disponibilizados vetores de palavras gerados em várias dimensões. Alguns modelos como Word2vec, FastText e Wang2vec possuem as variações CBOW e Skip-Gram, que diferenciam-se pela forma como preveem as palavras. Em "Ver Detalhes" pode-se ter acesso à rotinas de pré-processamento, limpeza e avaliação. No córpus, foram feitas tratativas de tokenização, remoção de stopwords, stemming e outras.

Word2Vec	FastText
Modelo	Corpora NILC
CBOW 50 dimensões	<a href="#">download</a>
CBOW 100 dimensões	<a href="#">download</a>
CBOW 300 dimensões	<a href="#">download</a>
CBOW 600 dimensões	<a href="#">download</a>
CBOW 1000 dimensões	<a href="#">download</a>
SKIP-GRAM 50 dimensões	<a href="#">download</a>
SKIP-GRAM 100 dimensões	<a href="#">download</a>
SKIP-GRAM 300 dimensões	<a href="#">download</a>
SKIP-GRAM 600 dimensões	<a href="#">download</a>
SKIP-GRAM 1000 dimensões	<a href="#">download</a>
Modelo	Corpora NILC
CBOW 50 dimensões	<a href="#">download</a>
CBOW 100 dimensões	<a href="#">download</a>
CBOW 300 dimensões	<a href="#">download</a>
CBOW 600 dimensões	<a href="#">download</a>
CBOW 1000 dimensões	<a href="#">download</a>
SKIP-GRAM 50 dimensões	<a href="#">download</a>
SKIP-GRAM 100 dimensões	<a href="#">download</a>
SKIP-GRAM 300 dimensões	<a href="#">download</a>
SKIP-GRAM 600 dimensões	<a href="#">download</a>
SKIP-GRAM 1000 dimensões	<a href="#">download</a>

## Word embedding

**Easy but ineffective representation of words:** one-hot encoding. No relation between words.



## Word embedding

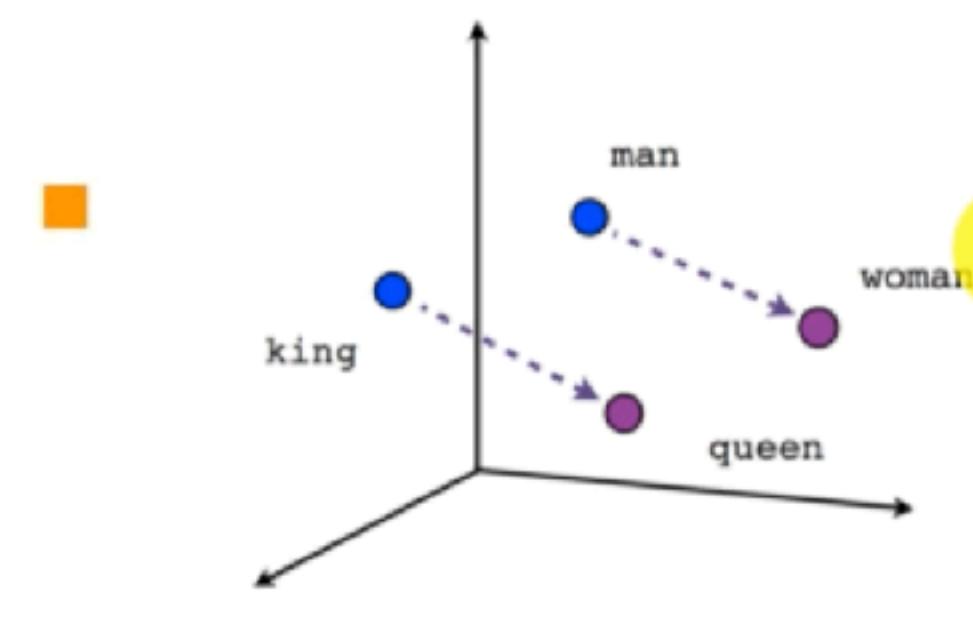
**Word embedding:** make each vector smaller -> adds relation between words.

$$\text{dog} = \overbrace{[0.194, 0.047, \dots, 0.126]}^{\textit{emb\_dim}}$$

$\textit{emb\_dim} \approx 64$

## Word embedding

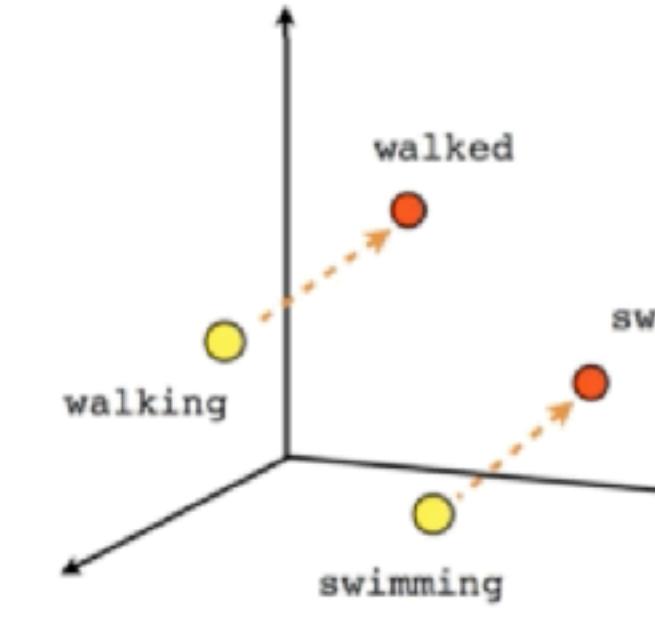
**Word embedding:** mathematical relations between words



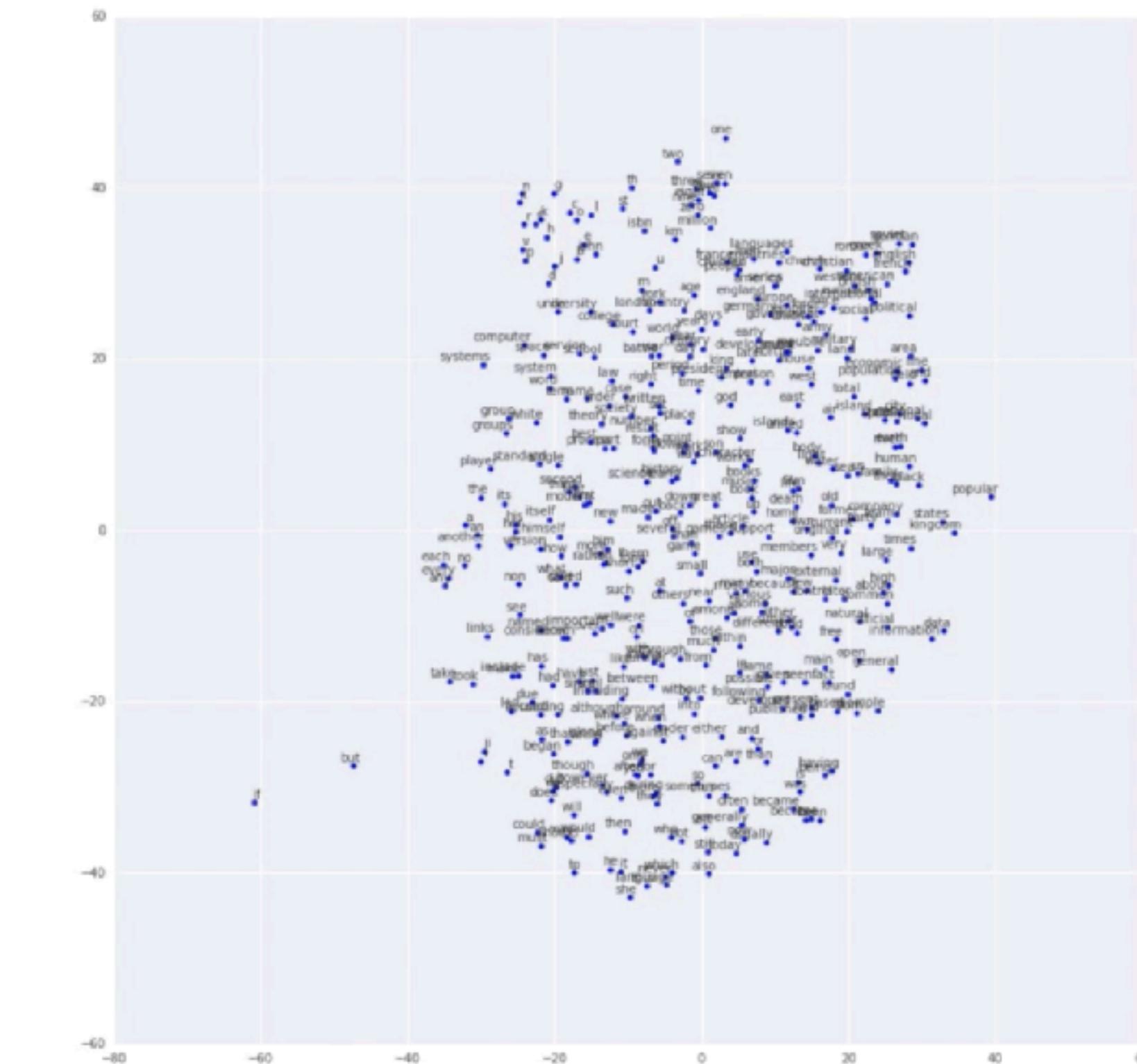
## Male-Female

[king] - [man] + [woman] = [queer]

[Paris] - [France] + [Italia] = [Rome]

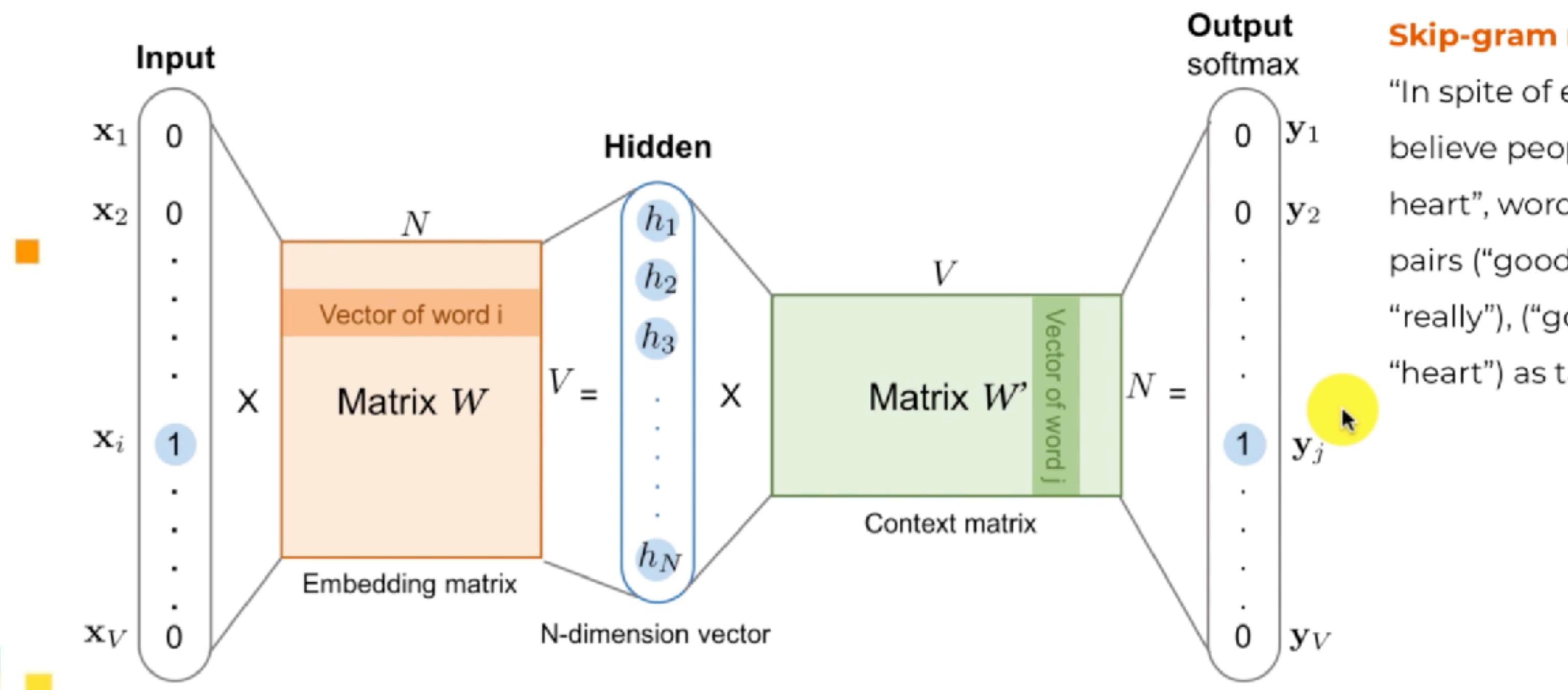


## Verb tens



CNN for NLP - Intuition

# Word embedding



**Skip-gram model:** in sentence  
“In spite of everything, I still  
believe people are really good at  
heart”, word “good” produces  
pairs (“good”, “are”), (“good”,  
“really”), (“good”, “at”), (“good”,  
“heart”) as target/context.

<https://medium.com/nlplanet/two-minutes-nlp-11-word-embeddings-models-you-should-know-a0581763b9a9>

