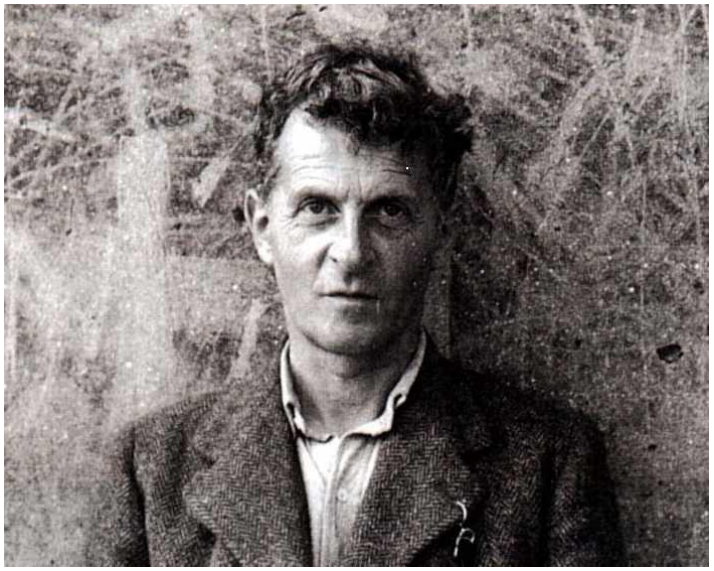
# Laboratorio de datos, clase 20

# Word embeddings con *word2vec*

Prof. Enzo Tagliazucchi

tagliazucchi.enzo@googlemail.com
www.cocuco.org



Ludwig Wittgenstein

“El significado deriva del uso de las palabras en el lenguaje”

“Las palabras se definen por su compañía”

“Dime con que palabras apareces y te diré qué significas”

“Ítems linguisticos con distribuciones similares tienen significados similares” (hipótesis distribucional)

| | Palabra 1 | Palabra 2 | Palabra 3 | Palabra 4 | Palabra 5 | |
|---|---|---|---|---|---|---|
| Relato 1 | 0 | 0.12 | 0.01 | 0 | 0 | |
| Relato 2 | 0 | 0 | 0.44 | 0.15 | 0.65 | |
| Relato 3 | 0.11 | 0.31 | 0.28 | 0 | 0 | (···) |
| Relato 4 | 0 | 0 | 0.05 | 0.21 | 0 | |
| Relato 5 | 0 | 0.13 | 0 | 0.07 | 0 | |
| | | (···) | | | | |

La correlación lineal entre filas nos da una idea de la similitud del significado entre relatos

La correlación lineal entre columnas nos da una idea de la similitud del significado entre palabras

Pero hay un problema: la mayor parte de los valores son 0

“Sobre la mesa hay un florero con margaritas y jazmines”

“El vaso lleno de flores está apoyado sobre una mesada”

Significado muy similar,
ninguna palabra en común

Solución: próximas clases
(ya tenemos todas las herramientas)

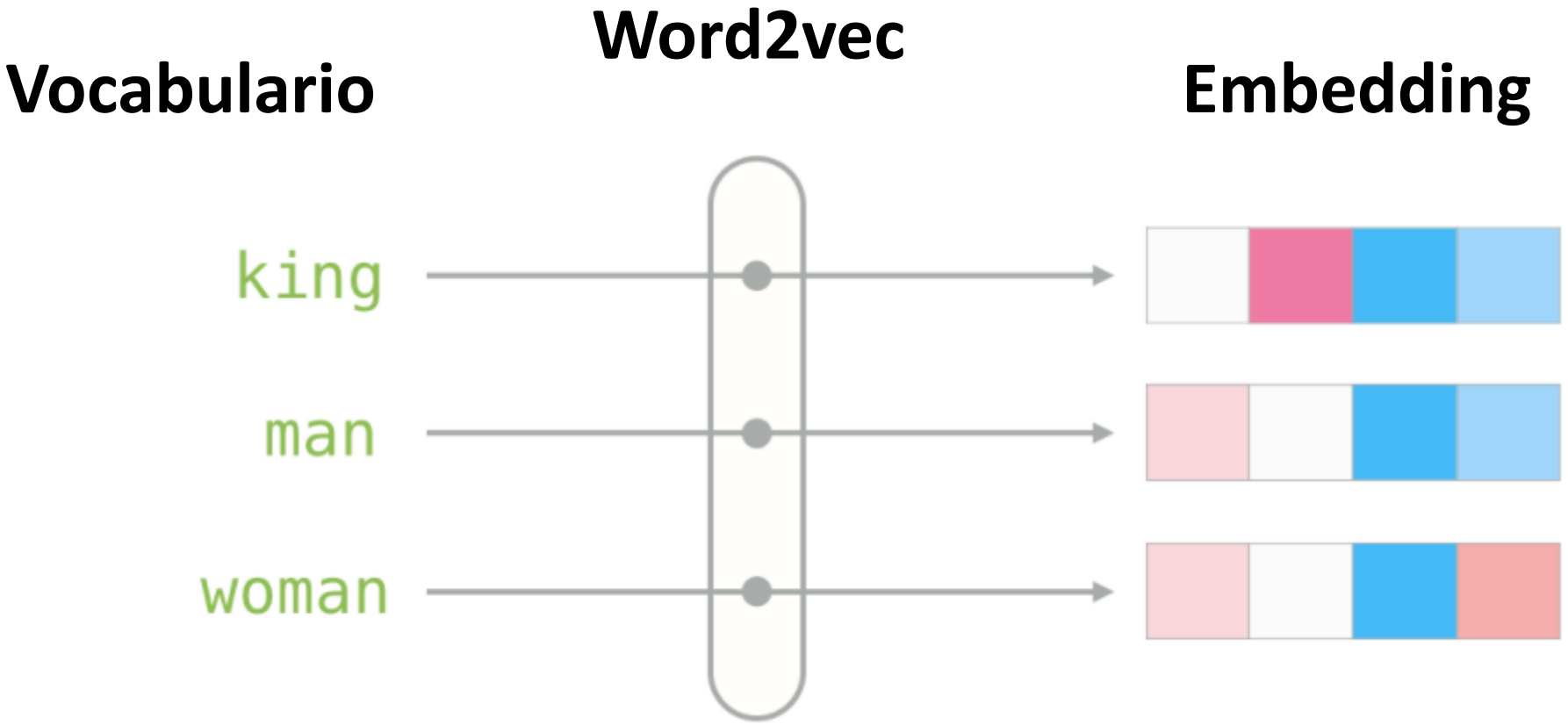
# Idea

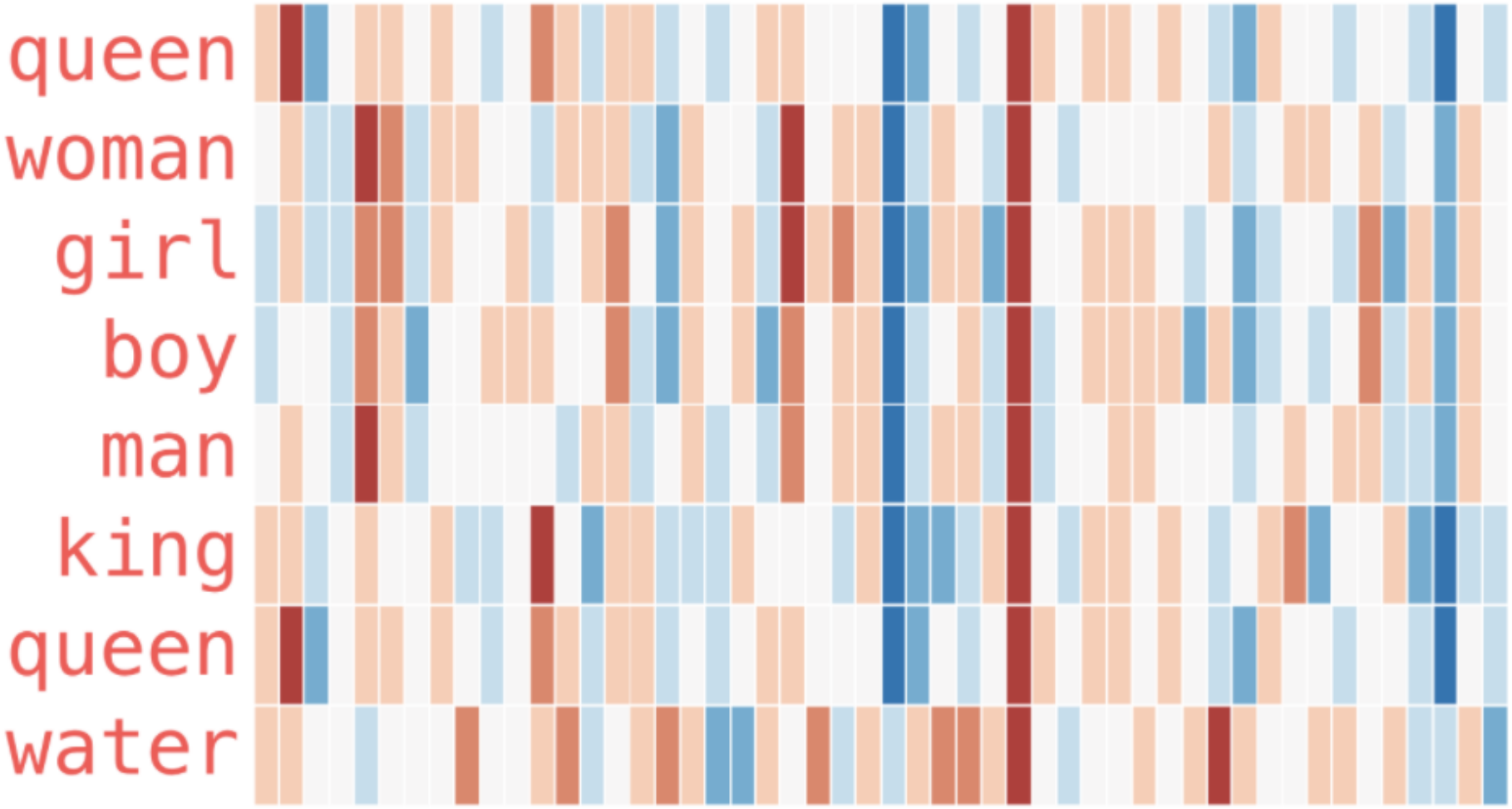
Reducir la dimensión del vocabulario, pasando de ∼50.000 palabras a ∼100, en una representación que ya no sea esparsa.

En el espacio original, cada dimension es una palabra y pueden variar independientemente. Pero sabemos que en realidad no es el caso: la dimension es menor, porque hay interacción entre las palabras.

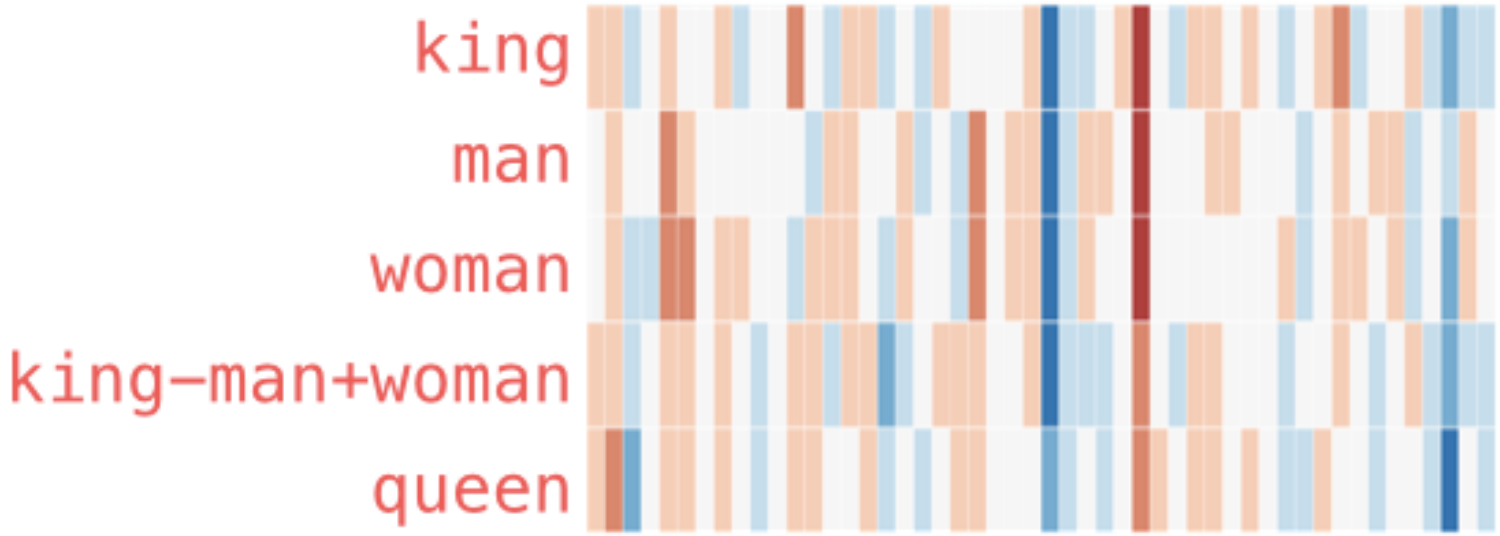
Más aún, introducir una medida de distancia en este espacio de forma tal que *palabras con significado similar estén cerca.*

**Word embedding**


Vocabulario
Word2vec
Embedding
king
man
woman


queen
woman
girl
boy
man
king
queen
water

king - man + woman ~= queen



# Para qué sirve

Encontrar similitud semántica entre palabras y entre documentos

Buscar las palabras que se parezcan más o menos a una dada

Operar con embeddings y buscar la palabra más cercana al resultado

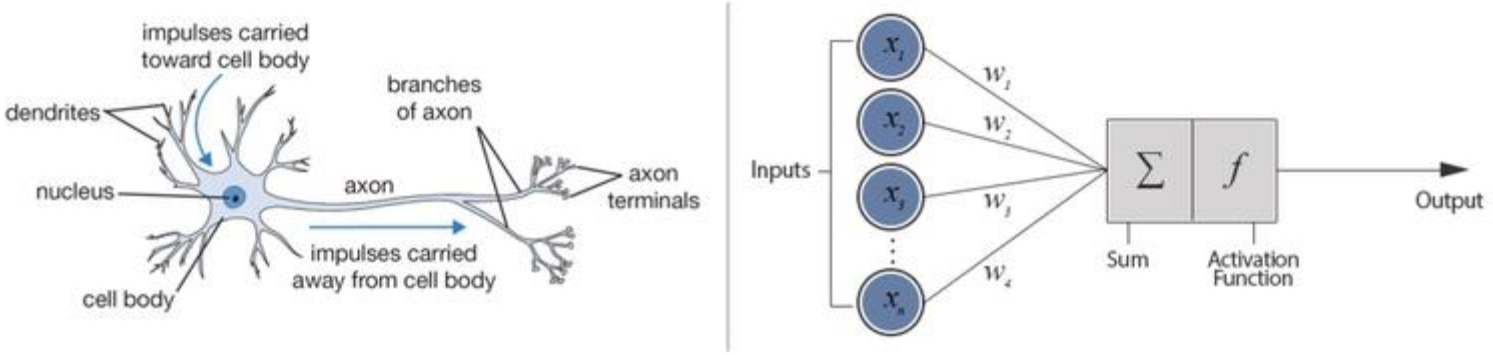
Autocompletar frases

Agrupar palabras de significados similares

Buscar analogías entre palabras

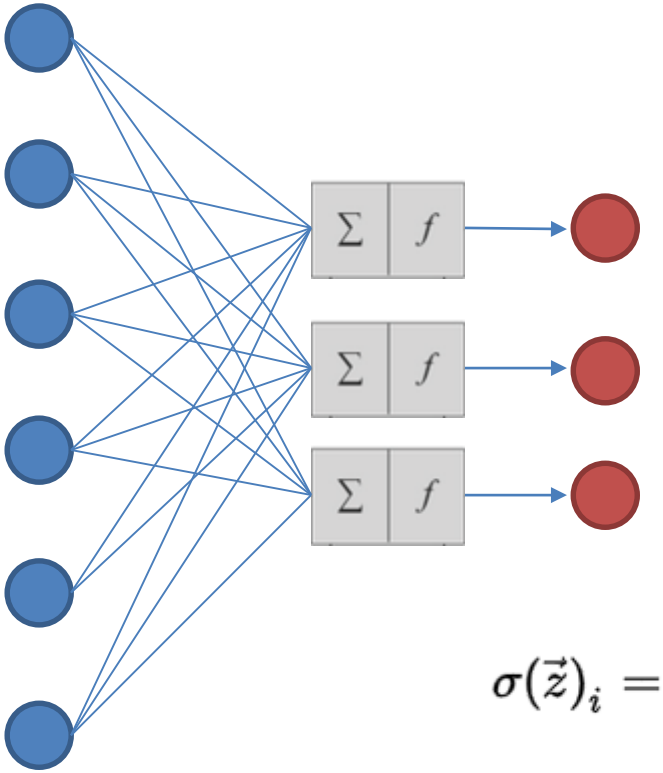
Modelo semántico del lenguaje para comparar con procesamiento del lenguaje hecho por humanos

# Biological Neuron versus Artificial Neural Network

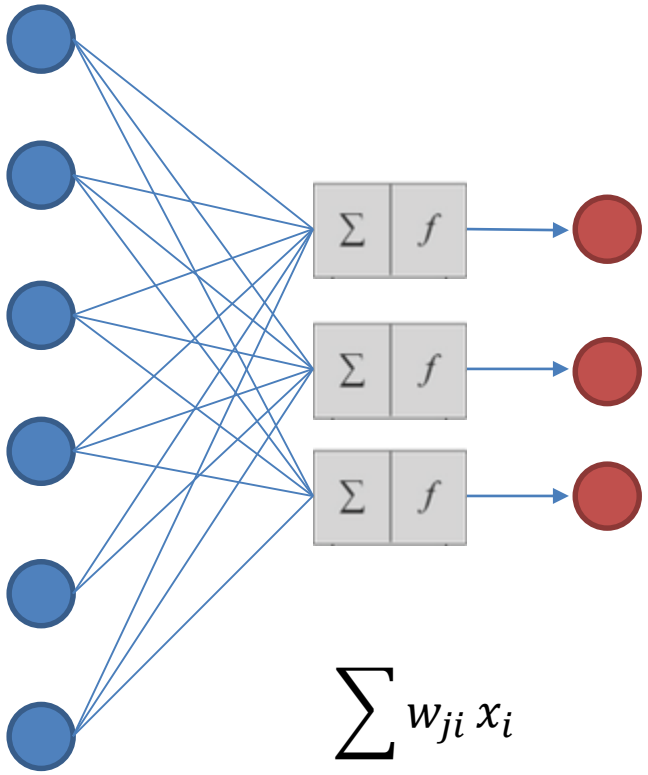



Cada unidad de la ultima capa (gris) recibe una combinación lineal de los valores de cada unidad de entrada (azul), aplica una función, y obtiene un output (rojo)

El objetivo es encontrar los coeficientes de la transformación lineal que mejor reproduce los outputs obtenidos en un conjunto de entrenamiento



$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$




$$y_j = f\left(\sum w_{ji}\, x_i\right)$$

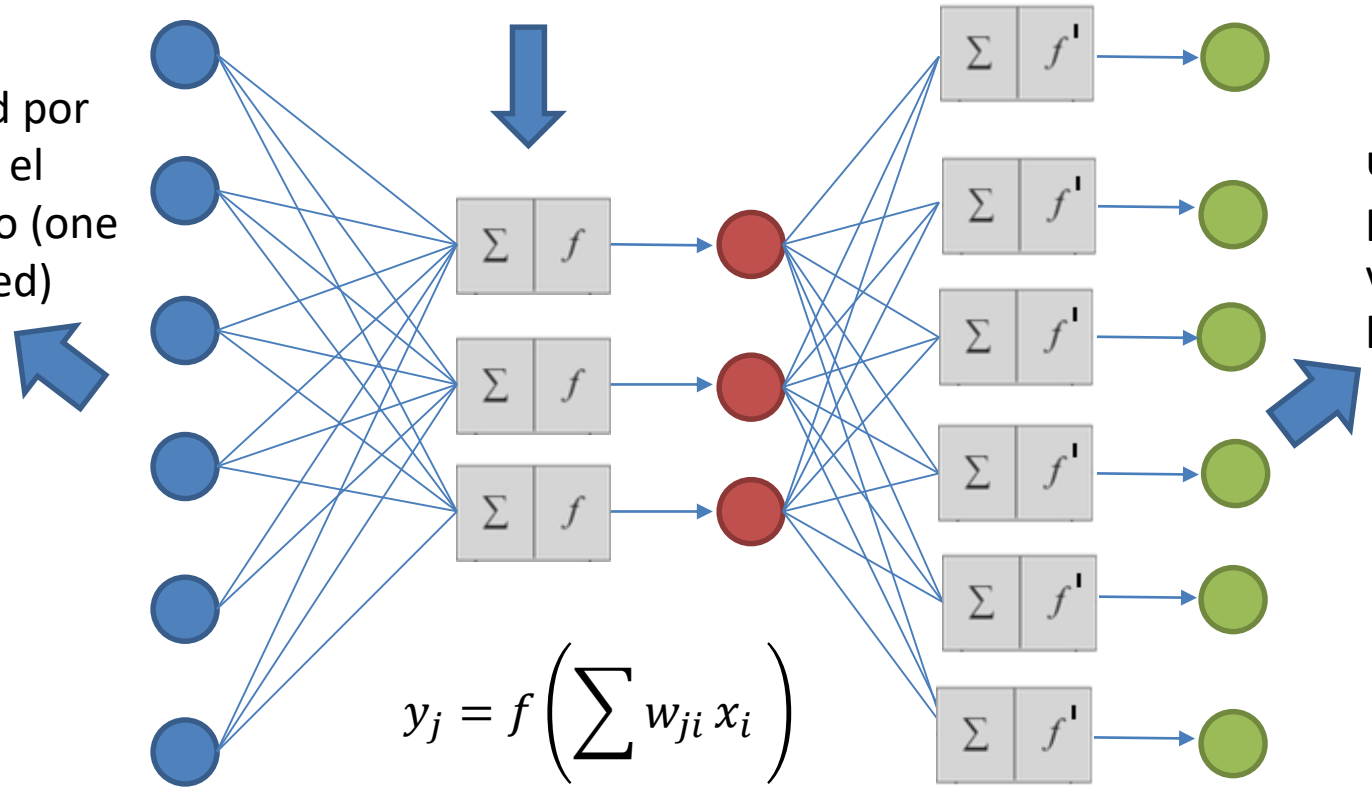
$$\sum w_{ji}\, x_i$$

$x_i$

En este ejemplo,
$w_{ji}$ vive en $R^{3x6}$

una unidad por dimensión del embedding

una unidad por palabra en el vocabulario (one hot encoded)




una unidad por palabra en el vocabulario (one hot encoded)

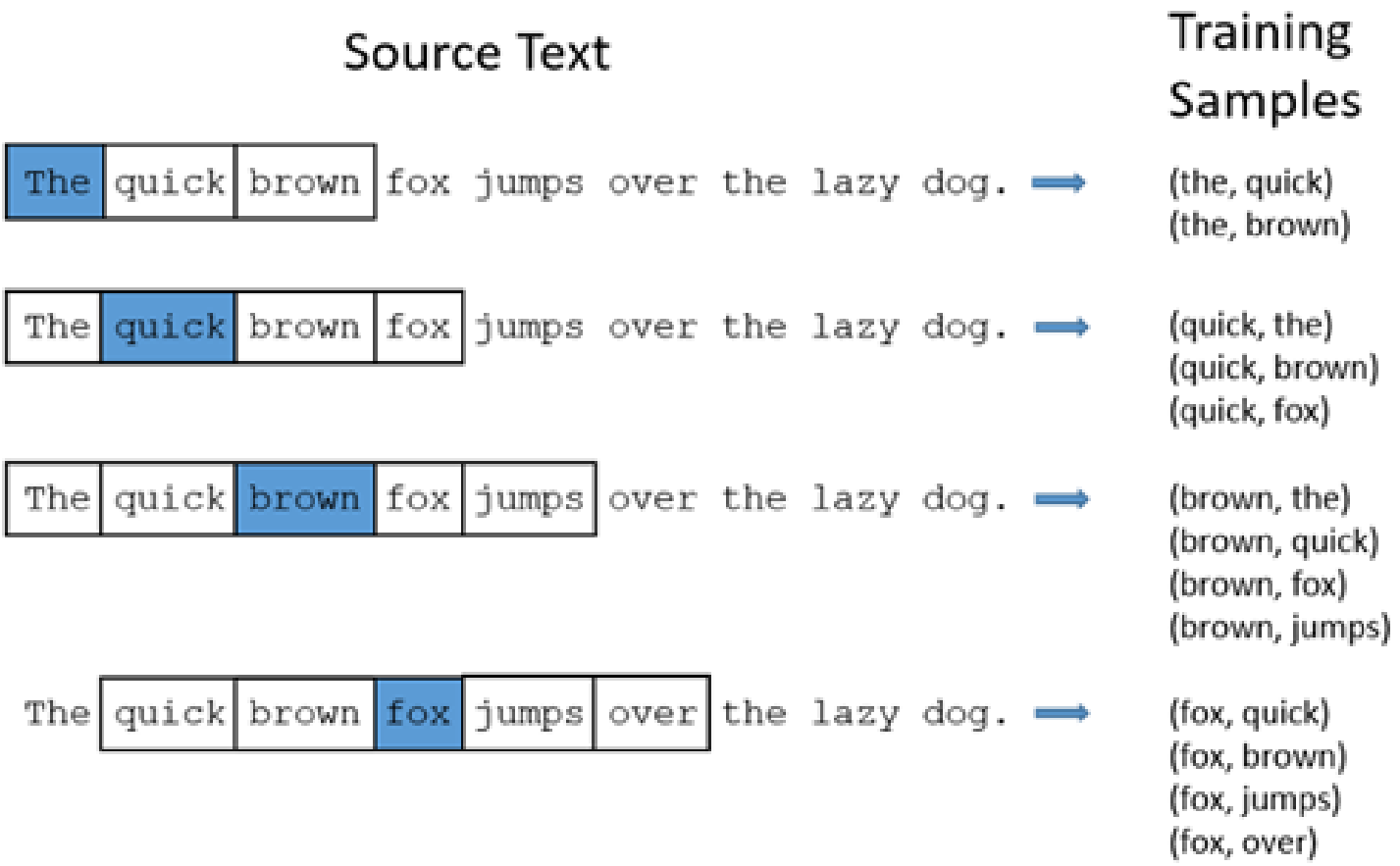
$$y_j = f\left(\sum w_{ji}\, x_i\right)$$

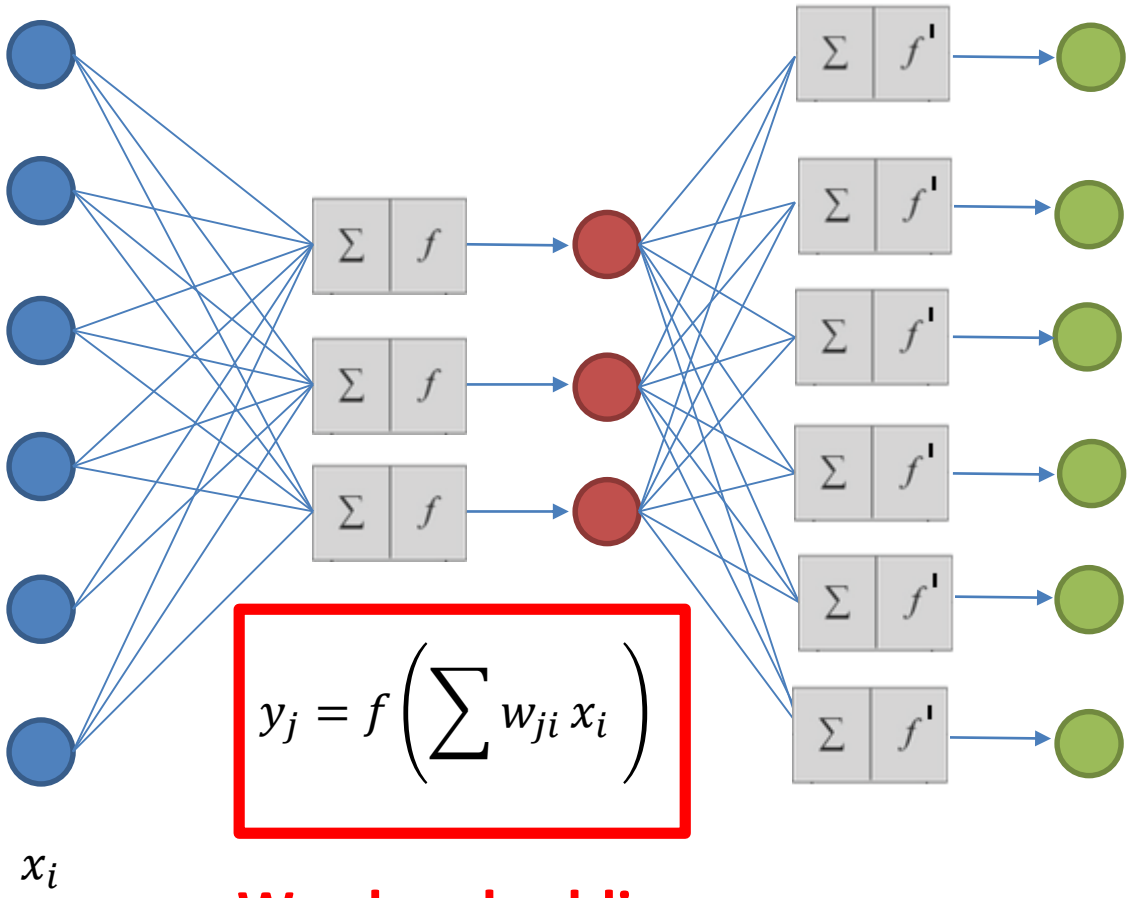
$x_i$

$$z_k = f'\left(\sum w'_{kj}\, y_j\right)$$

# Skip-gram

Los inputs son palabras y los outputs palabras, ambas en one hot encoding.
Los samples de entrenamiento se generan poniendo una ventana alrededor de cada palabra y seleccionando todas las palabras dentro de ese contexto

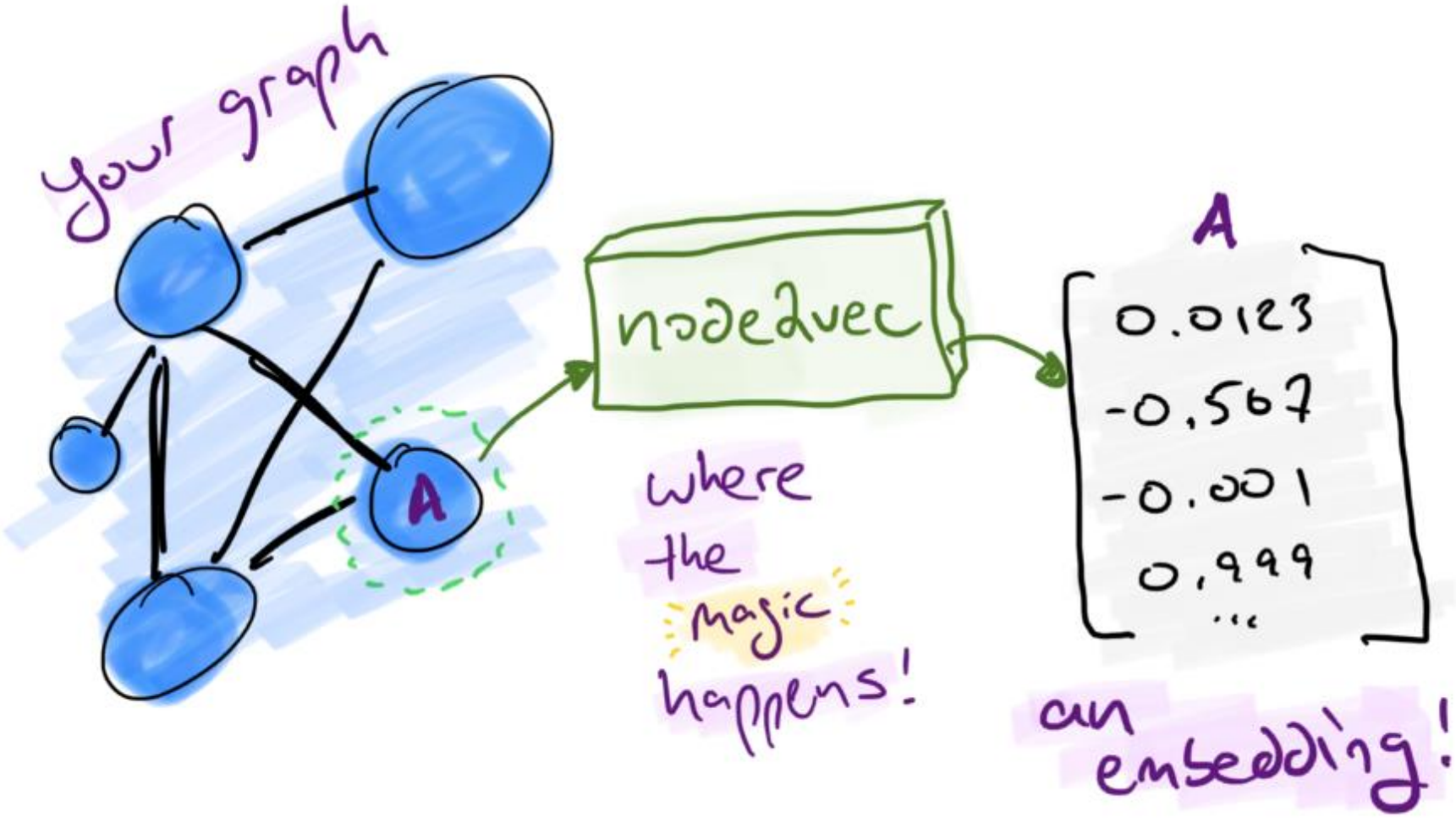





**Word embedding**

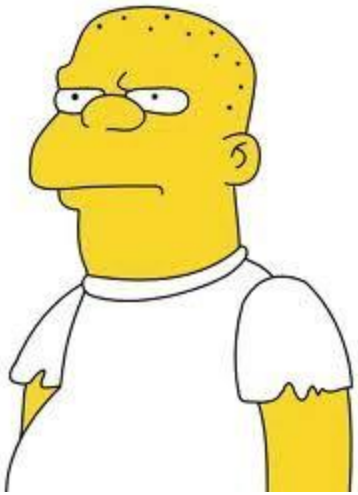
Actualmente hay mejores métodos (globe, fasttext, etc)

Estas ideas claramente no se aplican solo a lenguaje




THE SIMPSONS™

# ¿Cual es el más distinto a los demás?



# Unir con flechas