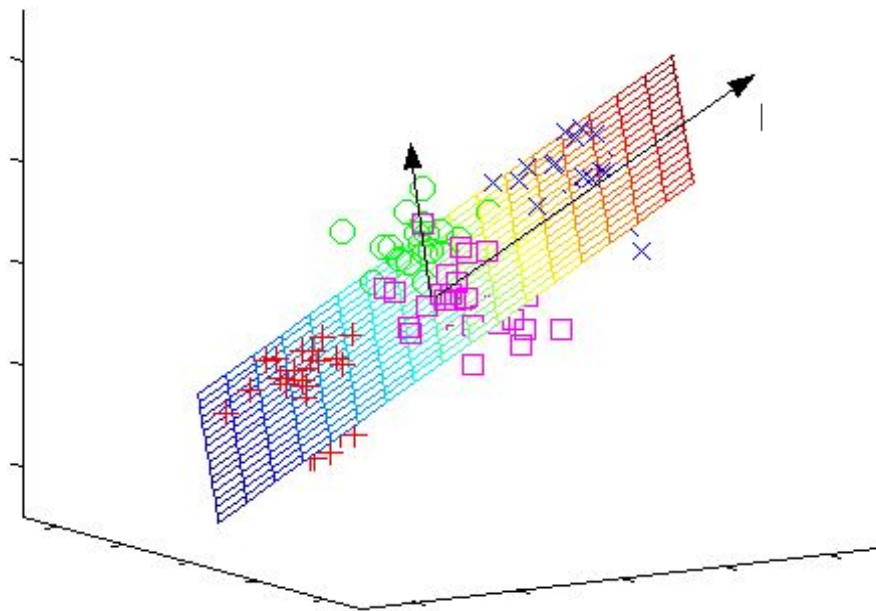


Embeddings

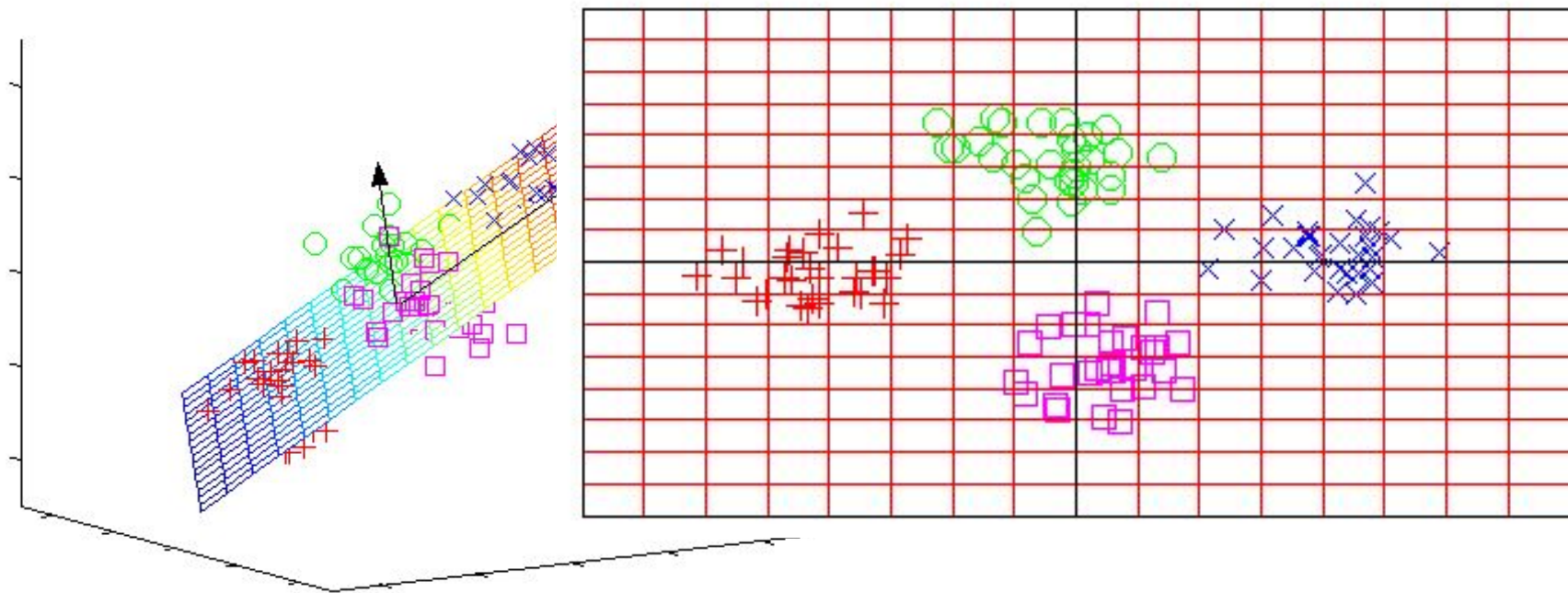


Diplomatura en Ciencia de Datos,
Aprendizaje Automático y sus Aplicaciones
FaMAF-UNC
Agosto 2024

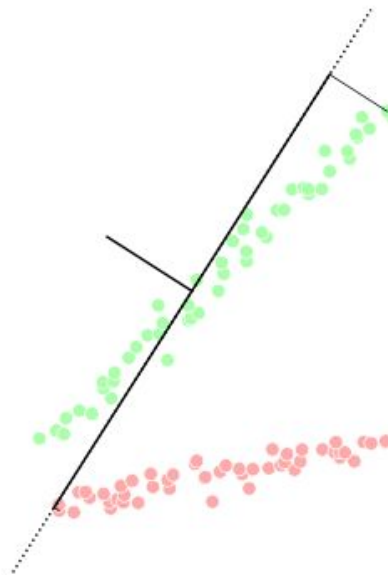
Qué es un embedding (proyección)



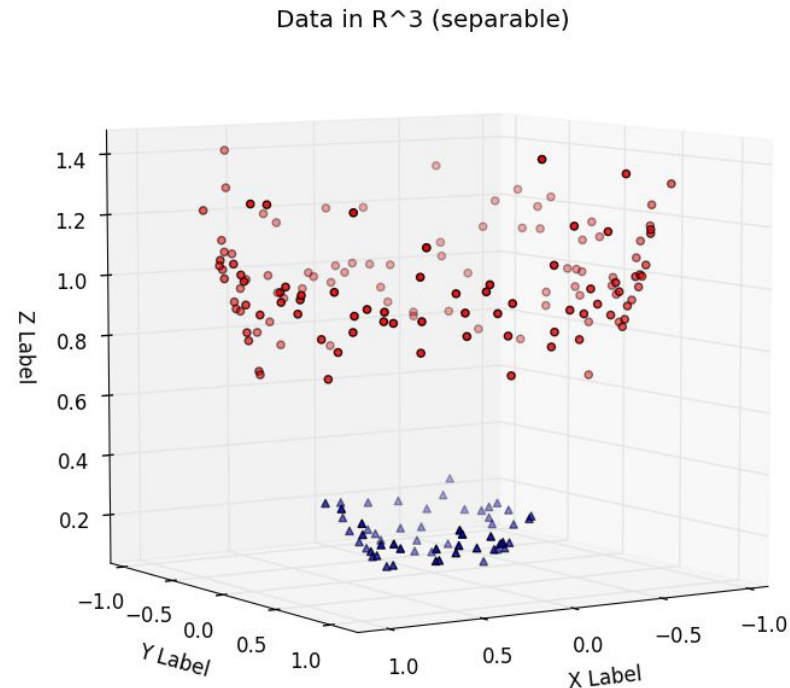
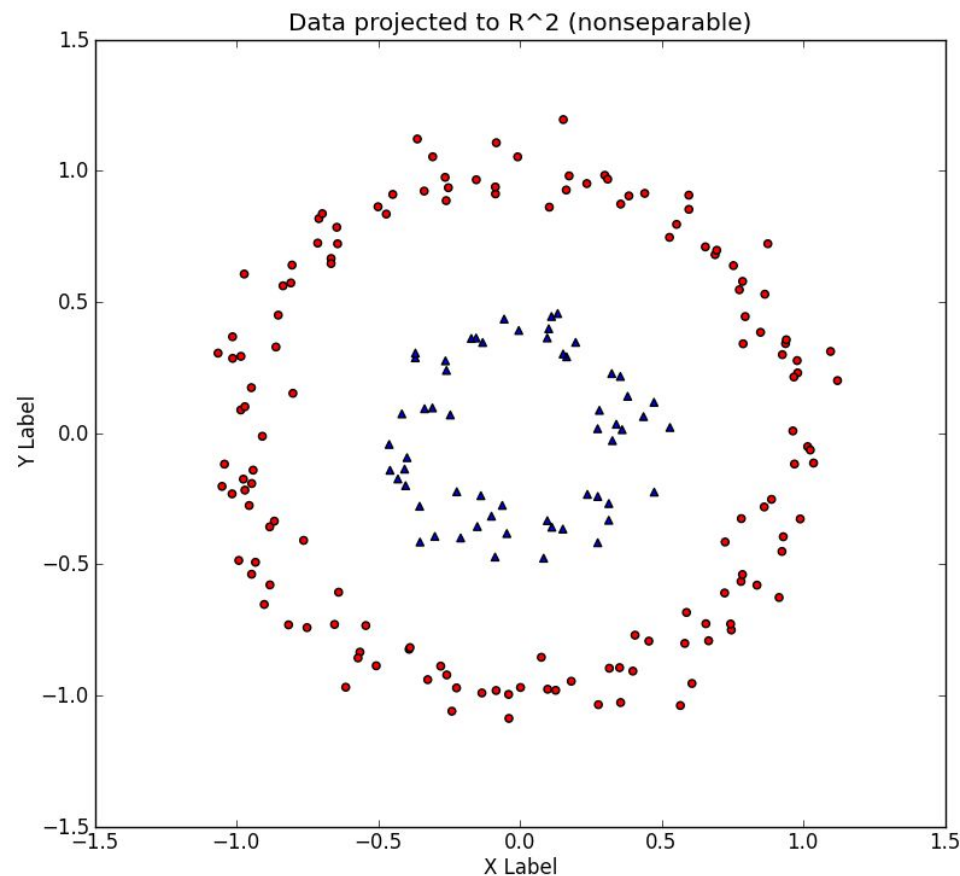
Qué es un embedding (proyección)



Qué es un embedding (proyección)



Qué es un embedding (proyección)



Qué es un embedding

Y un videíto sobre el kernel trick

<https://www.youtube.com/watch?v=3liCbRZPrZA>

Tipos de embeddings

Técnicas populares dentro de la familia de los embeddings

- Selección de características → supervisado o no supervisado
- Agrupamiento de características → supervisado o no supervisado
- The kernel trick → un espacio de mayor dimensionalidad!
- Principal Component Analysis
- Latent Dirichlet Allocation
- t-sne
- Neural embeddings

Objetivos de los embeddings

- En lugar de elegir un subconjunto de características, crear nuevas
- Sin tener en cuenta etiquetas de clase
- Proyectar a menos dimensiones preservando la mayor cantidad de información posible → minimizando el error cuadrado de reconstruir los datos originales

Para qué sirven?

- Reducción de dimensionalidad
- Reducir overfitting
- Generalización
- Acercamiento a las causas latentes
- Reducir el tiempo en ingeniería de características
- Reducir el sesgo del científico
- Visualizar

Qué perdemos?

- Información
- Interpretabilidad

Selección de Características

Reducción de dimensionalidad simplemente eliminando características

- Intuición: eliminamos ruido

Pero... la selección de características se hace en relación a una clase!

https://scikit-learn.org/stable/modules/feature_selection.html

Cómo hacemos si no tenemos clases?

Aplicamos conocimiento de dominio!

- P.ej., en lenguaje natural:
 - eliminamos palabras poco frecuentes
 - eliminamos palabras muy frecuentes

Selección de Características

Reducción de dimensionalidad simplemente eliminando características

- Intuición: eliminamos ruido

También tenemos métodos basados en varianza:

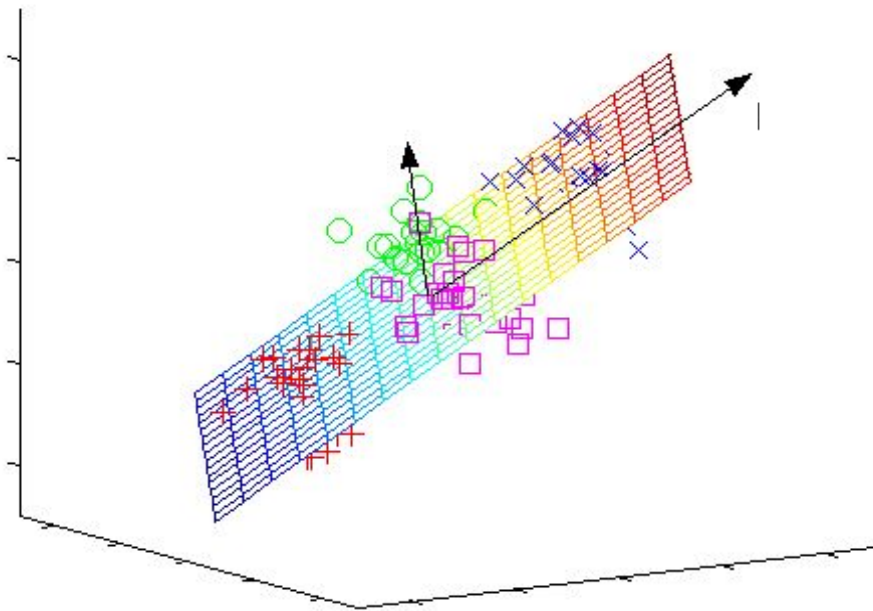
- Eliminar características con poca varianza (en scikit learn, [VarianceThreshold](#))
- Eliminar características redundantes con otras (en scikit learn, [mutual_info_classif](#))

Agrupamiento de Características

- Combinación de características dependientes (redundantes)
 - Por ejemplo, combinación lineal de el número de paradas recorridas por un colectivo y la distancia
- Combinación de características que sabemos que se pueden representar unidas
 - Por ej., sustituir viento, temperatura y humedad por sensación térmica
- Podemos sustituir características por la clase a la que pertenecen!
 - Por ej., en lenguaje natural, sustituir “corríamos” por “correr” o por V

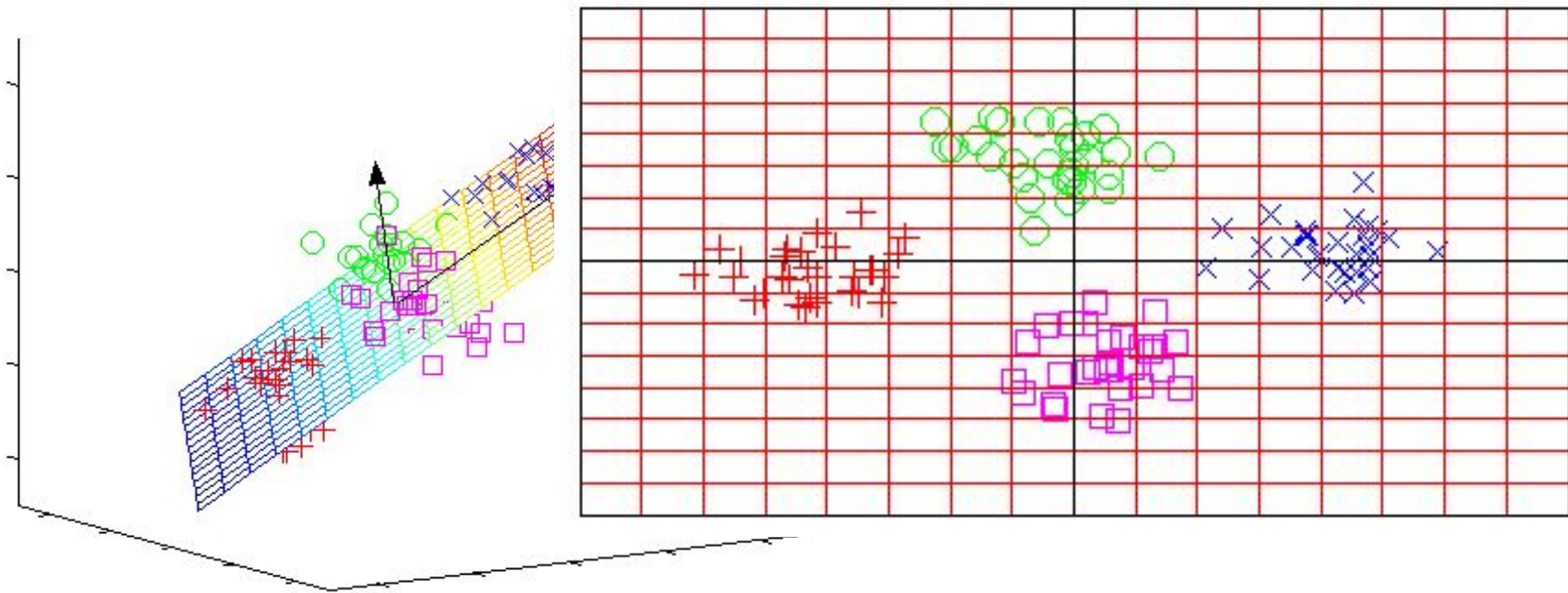
Principal Component Analysis

Minimiza el error cuadrado de reconstruir los datos originales



Principal Component Analysis

Minimiza el error cuadrado de reconstruir los datos originales



Descomposición en Valores Singulares

Los componentes principales se encuentran descomponiendo una matriz en valores singulares (eigenvalues) → singular value decomposition (SVD)

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \emptyset \\ \emptyset & \lambda_2 \end{bmatrix} \times \begin{bmatrix} \text{---} & v_1 & \text{---} \\ \text{---} & v_2 & \text{---} \end{bmatrix}$$

Latent Semantic Analysis

Los componentes p
valores singulares

Términos x
Documentos

Documentos
x Conceptos

Fuerza de
cada concepto

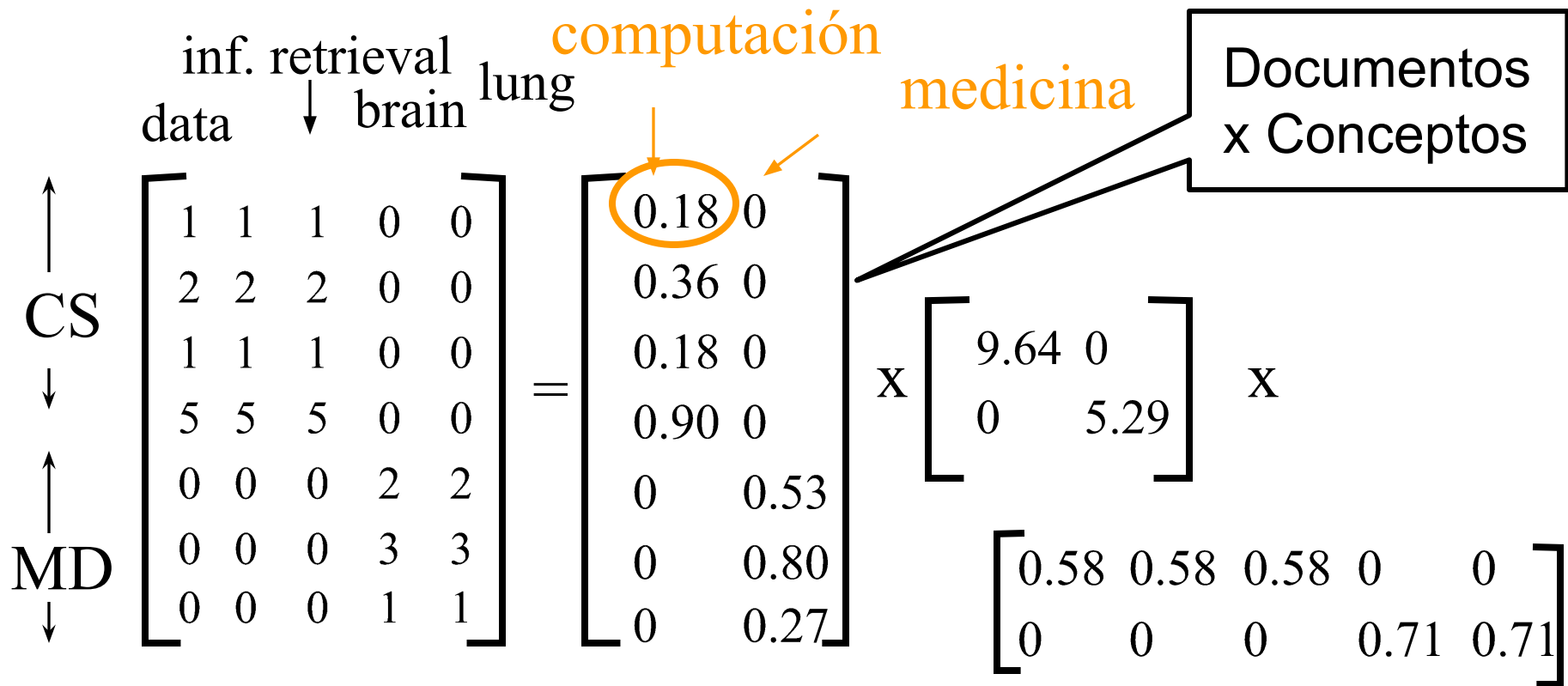
Términos x
Conceptos

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \emptyset \\ \emptyset & \lambda_2 \end{bmatrix} \times \begin{bmatrix} \text{---} & v_1 & \text{---} \\ \text{---} & v_2 & \text{---} \end{bmatrix}$$

Latent Semantic Analysis

$$\begin{array}{c}
 \begin{array}{c} \uparrow \\ \text{CS} \\ \downarrow \end{array} \\
 \begin{array}{c} \uparrow \\ \text{MD} \\ \downarrow \end{array}
 \end{array}
 \begin{array}{c}
 \text{inf. retrieval} \\
 \text{data}
 \end{array}
 \begin{array}{c}
 \downarrow \text{brain} \quad \text{lung}
 \end{array}
 \begin{bmatrix}
 1 & 1 & 1 & 0 & 0 \\
 2 & 2 & 2 & 0 & 0 \\
 1 & 1 & 1 & 0 & 0 \\
 5 & 5 & 5 & 0 & 0 \\
 0 & 0 & 0 & 2 & 2 \\
 0 & 0 & 0 & 3 & 3 \\
 0 & 0 & 0 & 1 & 1
 \end{bmatrix}
 =
 \begin{bmatrix}
 0.18 & 0 \\
 0.36 & 0 \\
 0.18 & 0 \\
 0.90 & 0 \\
 0 & 0.53 \\
 0 & 0.80 \\
 0 & 0.27
 \end{bmatrix}
 \times
 \begin{bmatrix}
 9.64 & 0 \\
 0 & 5.29
 \end{bmatrix}
 \times
 \begin{bmatrix}
 0.58 & 0.58 & 0.58 & 0 & 0 \\
 0 & 0 & 0 & 0.71 & 0.71
 \end{bmatrix}$$

Latent Semantic Analysis



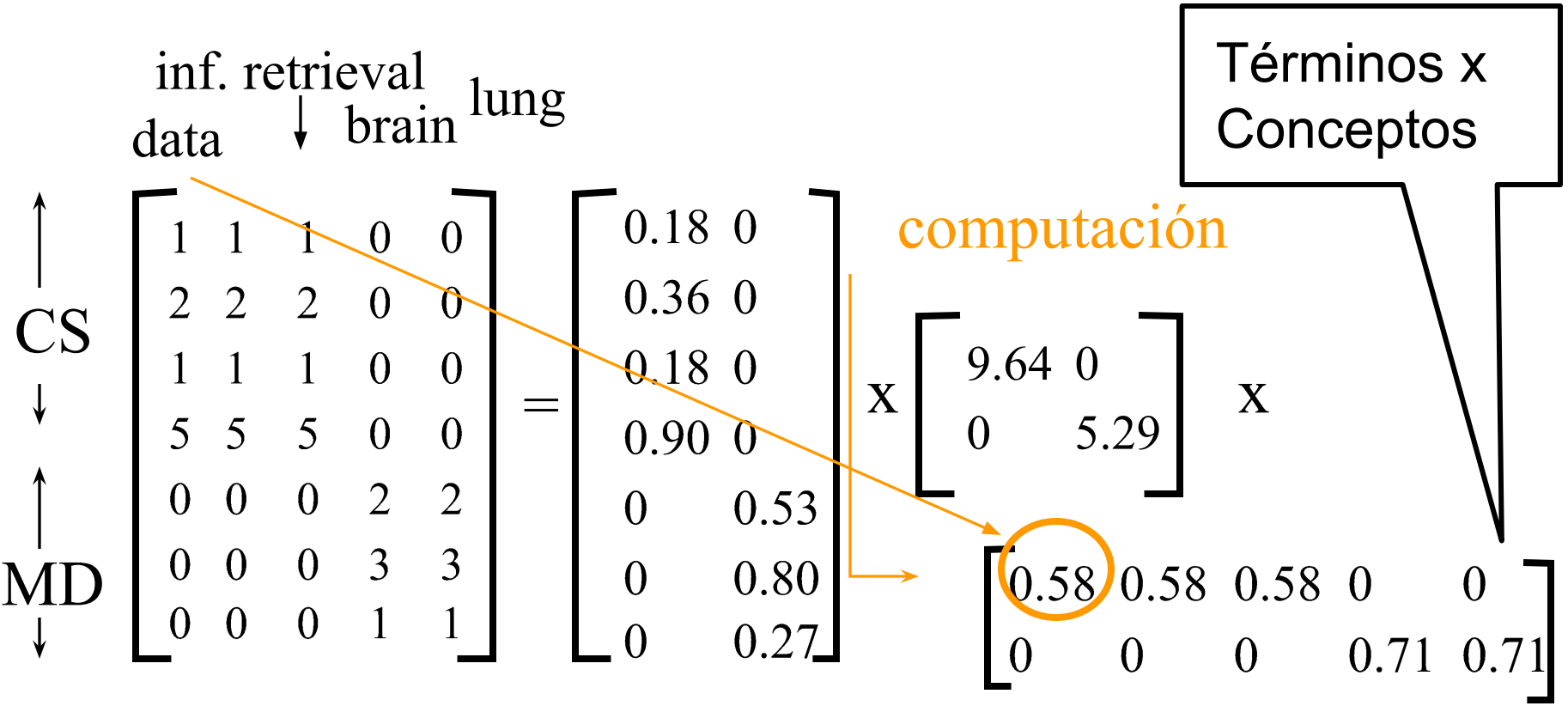
Latent Semantic Analysis

inf. retrieval
data ↓ brain lung

fuerza del concepto de computación

$$\begin{array}{c} \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow \end{array}
 \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
 =
 \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
 \times
 \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
 \times
 \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

Latent Semantic Analysis



Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

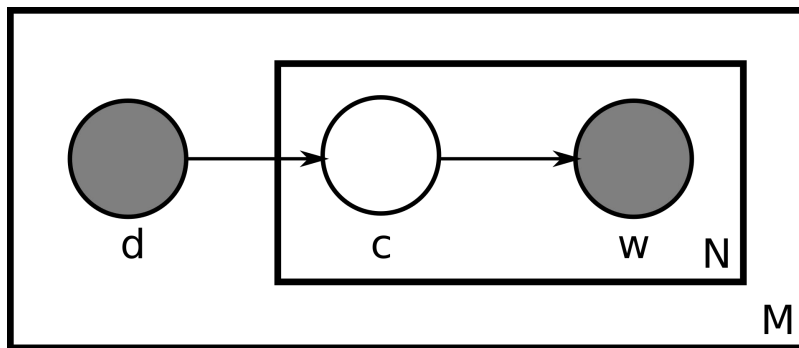
Diagram illustrating the reduction of dimensionality in Latent Semantic Analysis (LSA). The input matrix (7x5) is decomposed into three matrices: a 7x2 matrix, a 2x2 matrix, and a 2x5 matrix. The second matrix (2x2) is crossed out with a red X, indicating it is not used in the final reduced representation. The first matrix (7x2) and the third matrix (2x5) are also crossed out with green lines, suggesting the final reduced representation is the product of the first and third matrices.

Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 9.64 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}$$

Probabilistic Latent Semantic Analysis

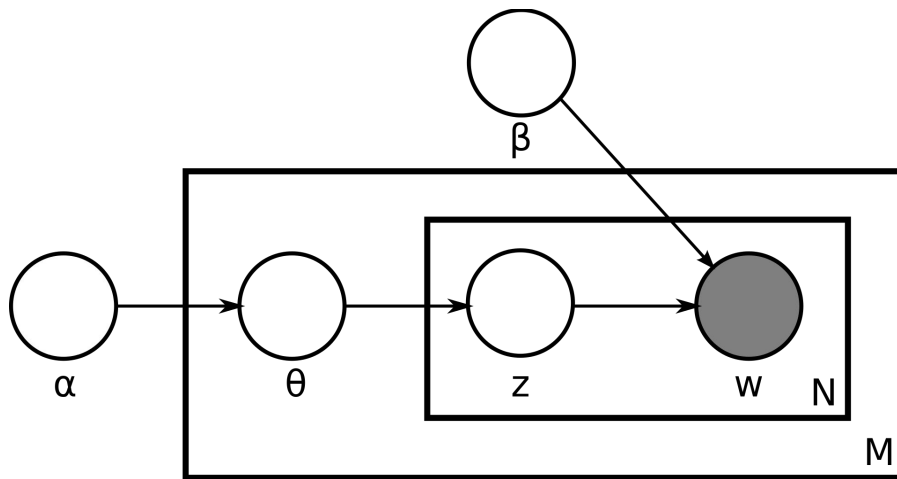
- Modela la distribución de cada co-ocurrencia como una mezcla de distribuciones multinomiales independientes o clases latentes o tópicos (el n de tópicos es un parámetro)



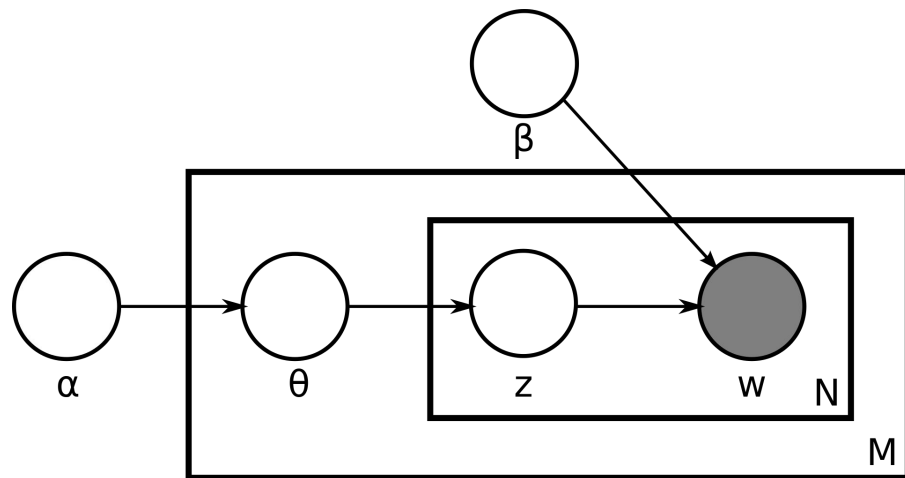
- d es el documento
- c es un tópico obtenido de la distribución de tópicos del documento $P(c|d)$
- w es una palabra obtenida de la distribución de palabras de c

Latent Dirichlet Allocation

- Modela la distribución de cada co-ocurrencia como una mezcla de distribuciones multinomiales (clases latentes o tópicos)
- Se asume que las clases latentes están distribuidas según la distribución de Dirichlet, una distribución de probabilidad continua multivariada



Latent Dirichlet Allocation



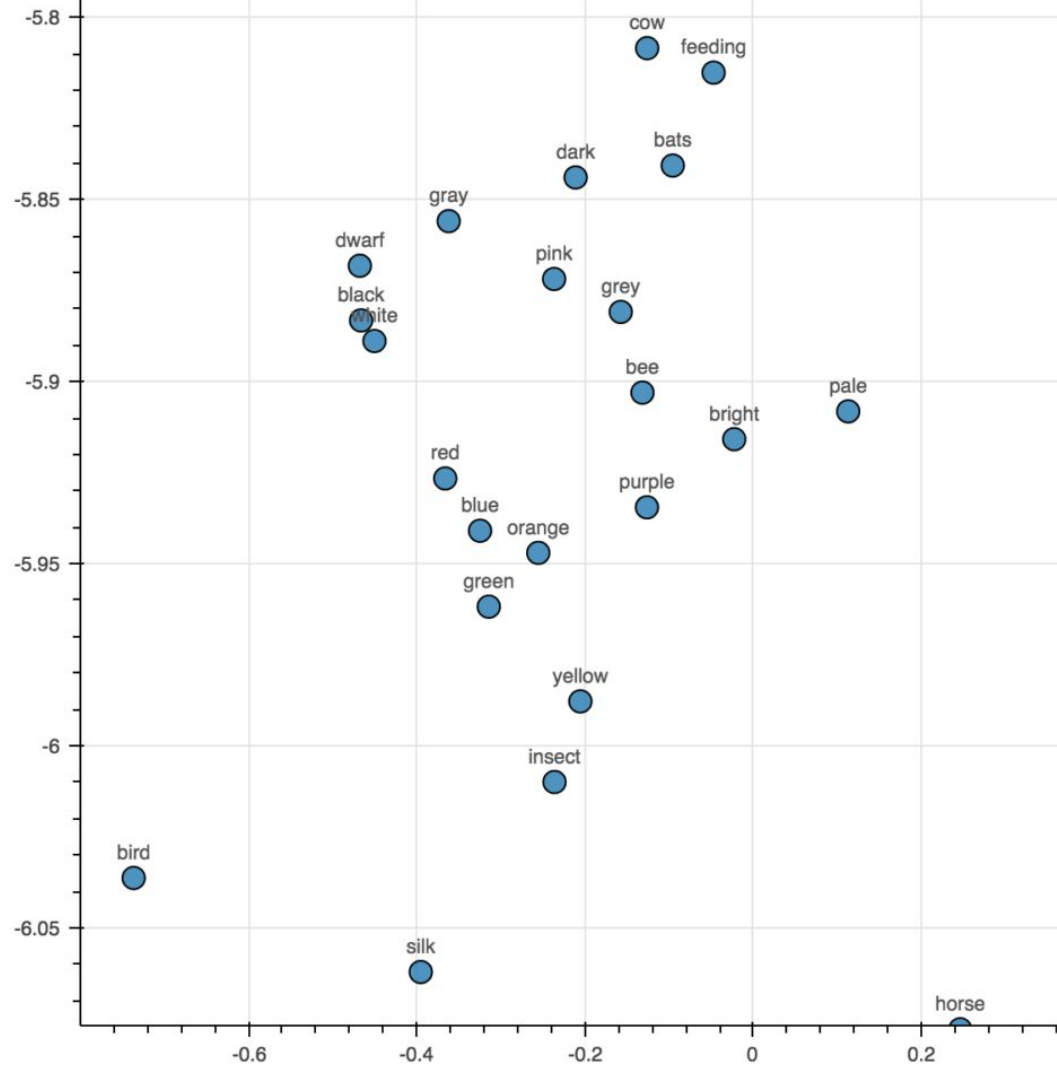
- α es el parámetro de de Dirichlet en la distribución de tópicos por documentos
- β es el parámetro de Dirichlet en la distribución de palabras por tópicos
- θ es la distribución de tópicos para el documento i
- ϕ es la distribución de palabras para el tópico k
- z es el tópico para la j -ésima palabra del documento i
- W es la j -ésima palabra del documento i

t-SNE

t-distributed stochastic neighbor embedding (t-SNE)

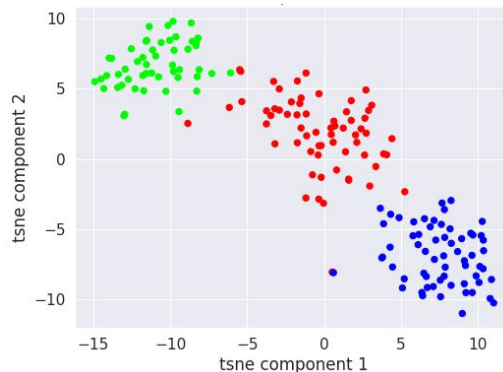
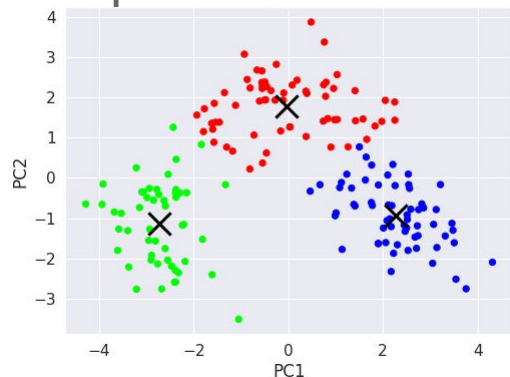
- reducción de dimensionalidad no lineal
- para visualización en dos o tres dimensiones
- los objetos semejantes quedan cercanos y los más diferentes quedan más distantes, con alta probabilidad
- es una aplicación de la Divergencia de Kullback-Leibler, que nos indica qué tanto una distribución puede haber sido originada por otra distribución

t-SNE



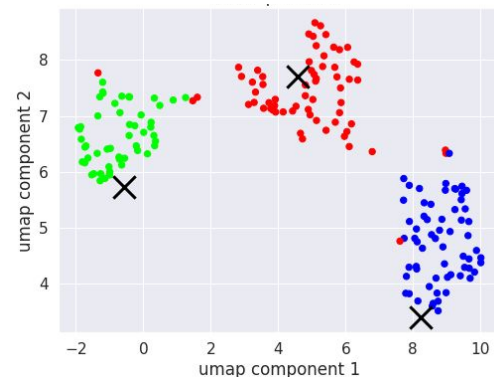
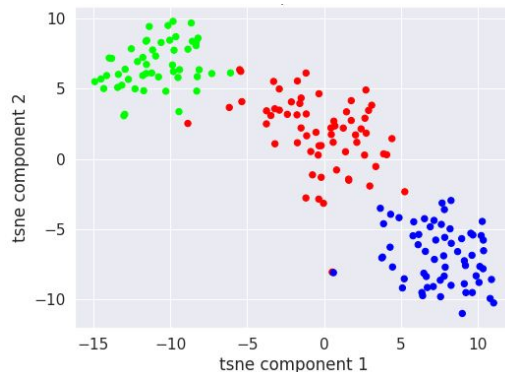
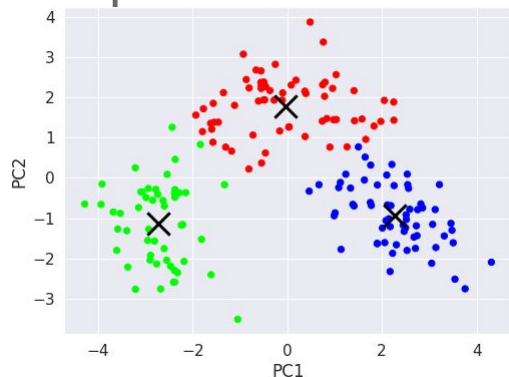
Diferencia entre t-sne y PCA

- t-sne refleja en el espacio destino la **probabilidad** de que dos puntos estuvieran **cercanos o lejanos** en el espacio original, pero no cuánta distancia tenían
- PCA refleja la **distancia** entre puntos en las componentes principales



Diferencia entre t-sne y PCA

- t-sne refleja en el espacio destino la **probabilidad** de que dos puntos estuvieran **cercanos o lejanos** en el espacio original, pero no cuánta distancia tenían
- PCA refleja la **distancia** entre puntos en las componentes principales



Diferencia entre t-sne y PCA

- t-sne refleja en el espacio destino la **probabilidad** de que dos puntos estuvieran **cercanos o lejanos** en el espacio original, pero no cuánta distancia tenían
- PCA refleja la **distancia** entre puntos en las componentes principales

Para saber más:

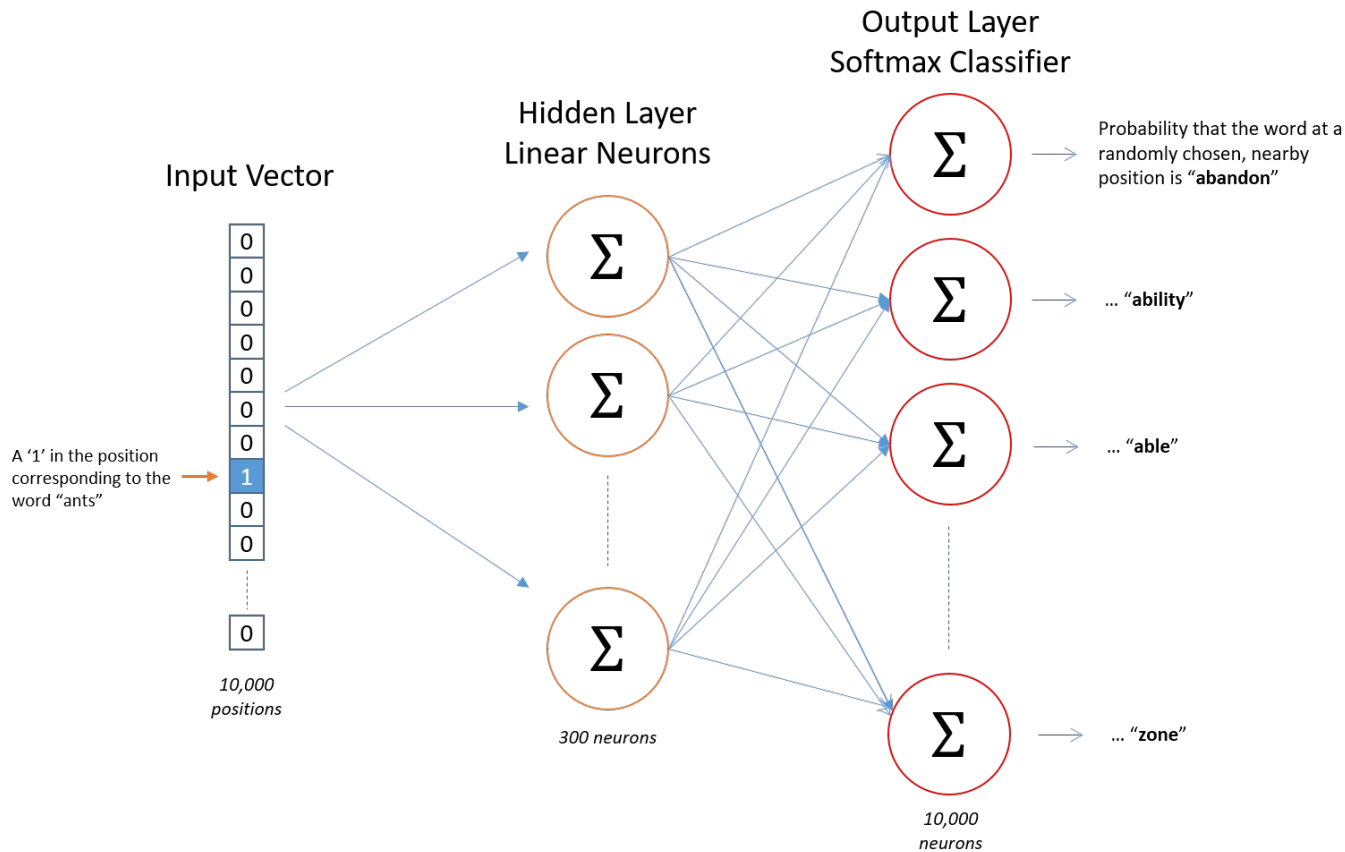
https://en.wikipedia.org/wiki/Dimensionality_reduction#Feature_projection

<https://www.kaggle.com/code/samueltcortinhas/intro-to-pca-t-sne-umap>

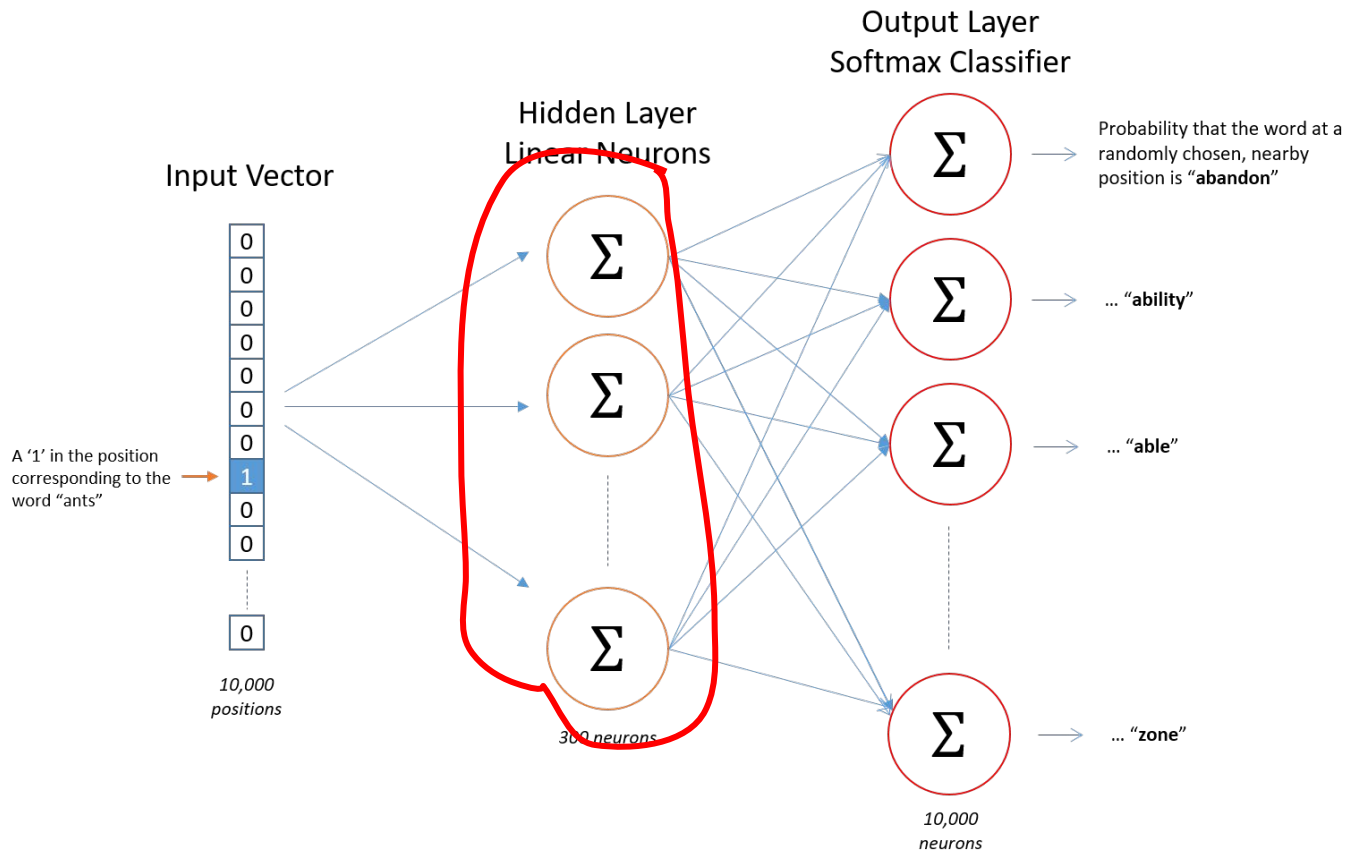
no linealidad: embeddings neuronales

1. entrenar una red neuronal
2. eliminar la capa de predicción
3. la capa anterior a la de predicción es el nuevo espacio
4. el camino hasta esa capa es el mecanismo de proyección

no linealidad: embeddings neuronales

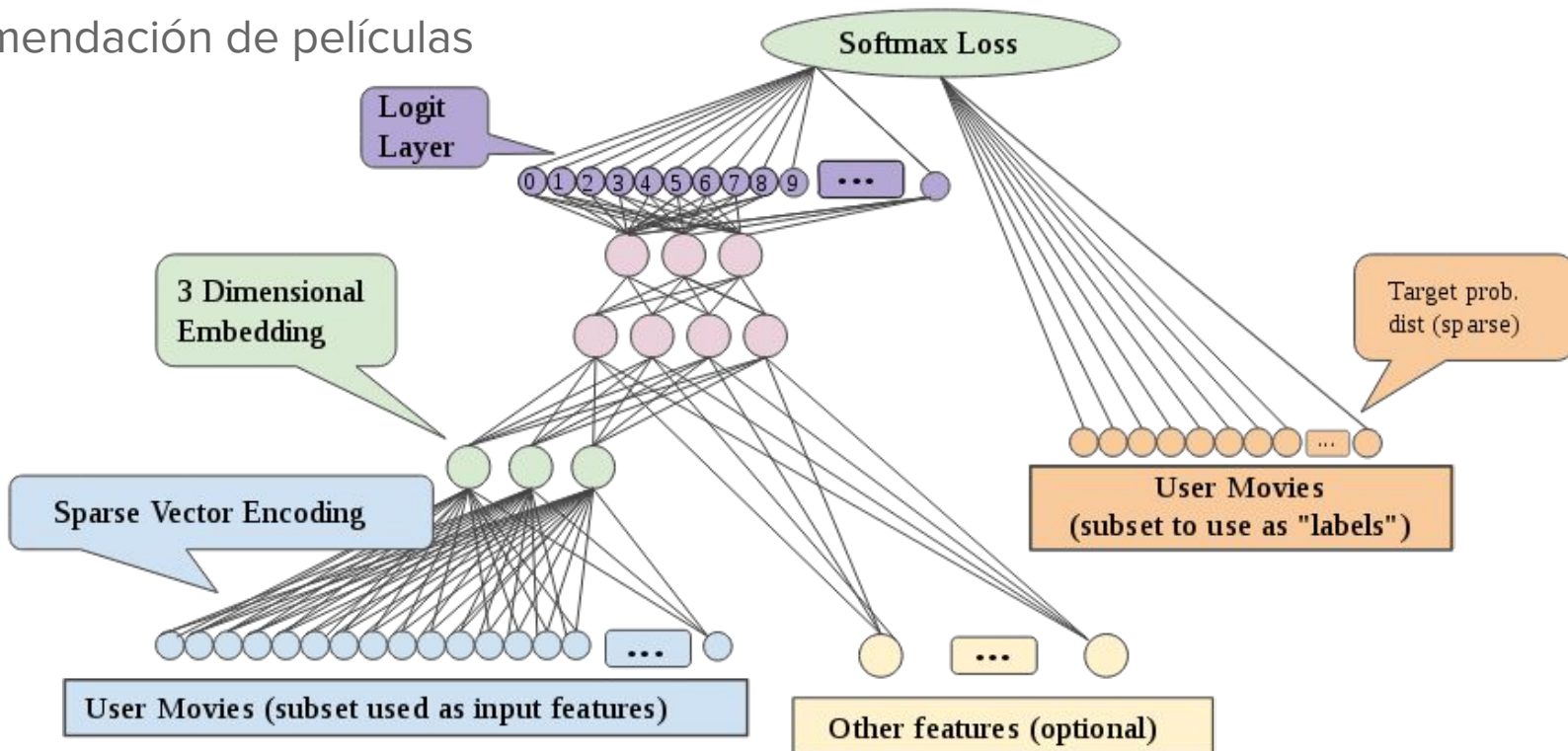


no linealidad: embeddings neuronales



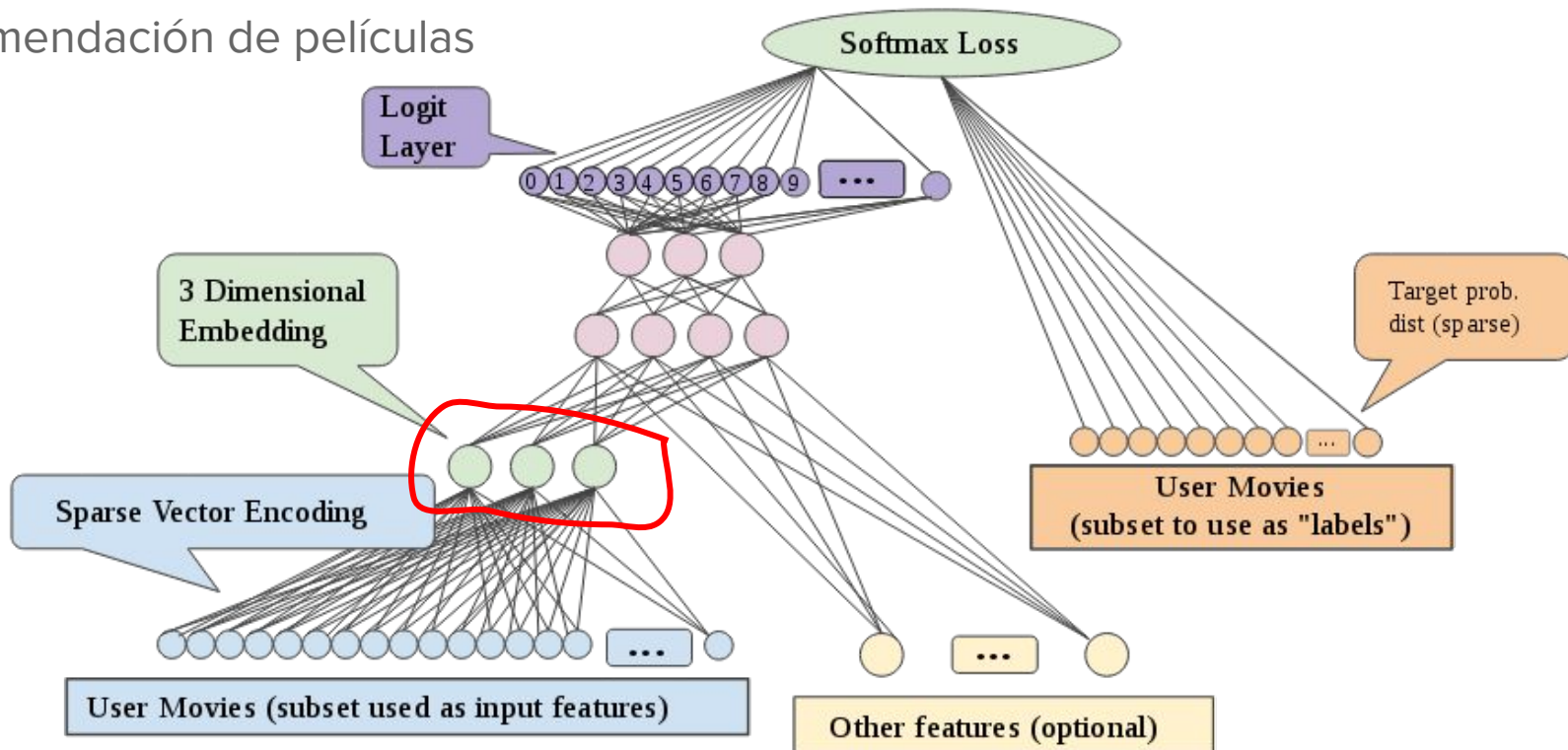
Tarea de pretexto para embeddings

recomendación de películas



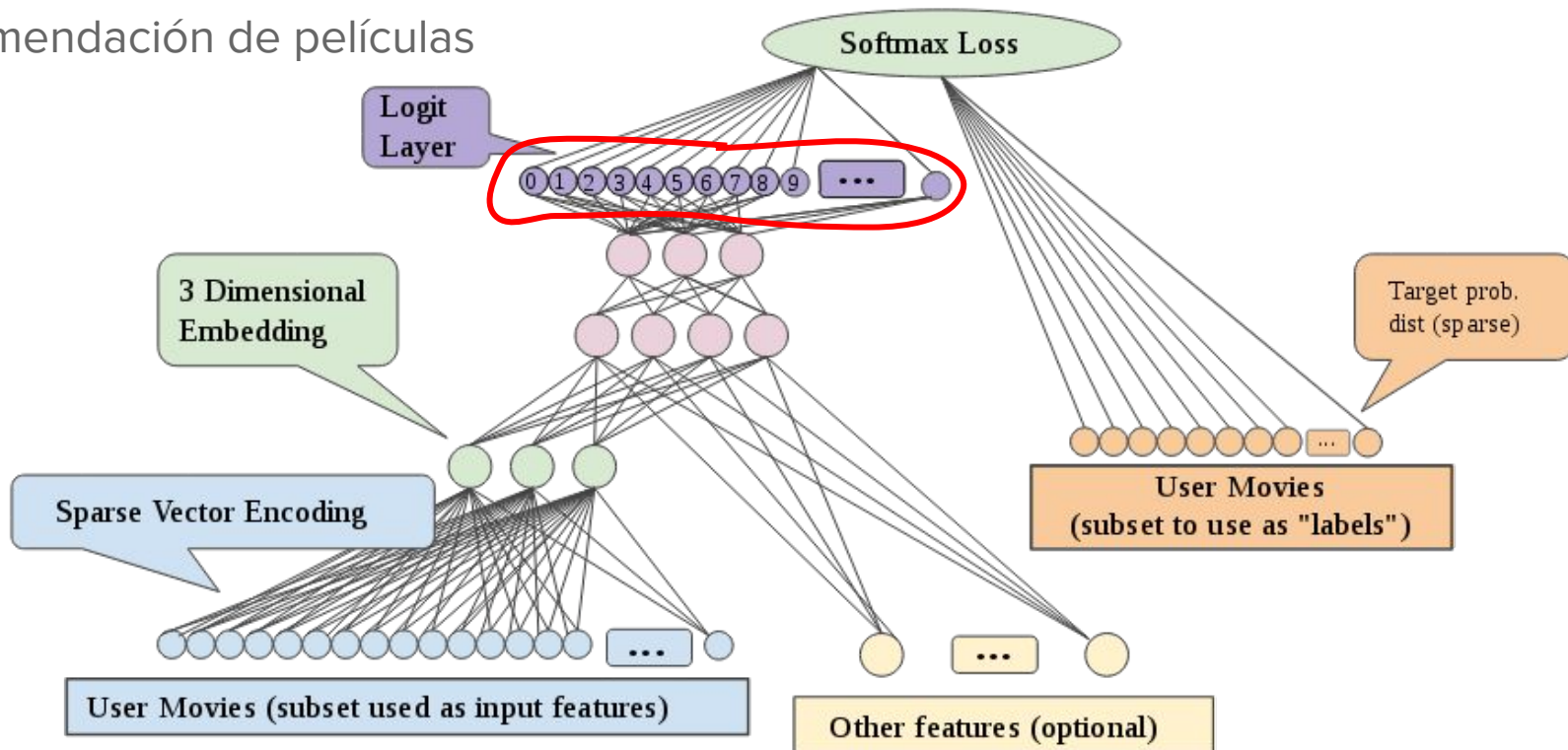
Tarea de pretexto para embeddings

recomendación de películas



Tarea de pretexto para embeddings

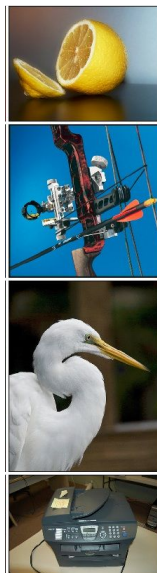
recomendación de películas



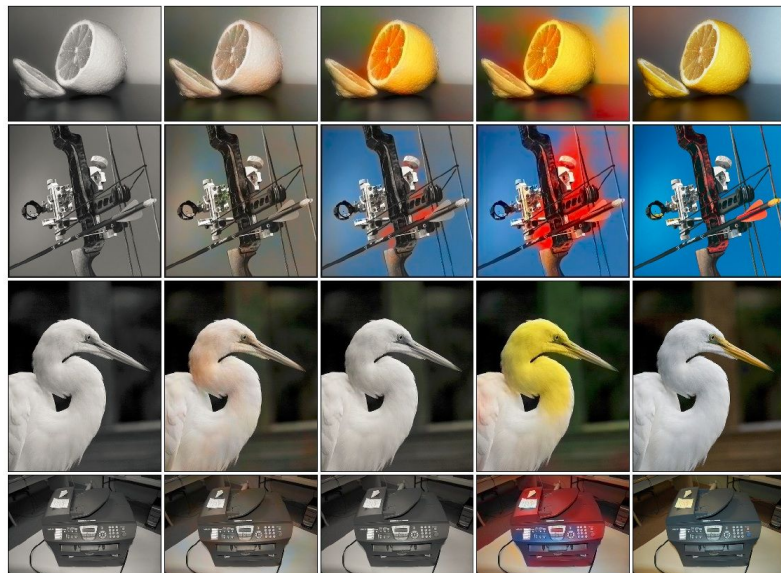
Embeddings neuronales

- Entrenar una red neuronal con una **tarea de pretexto** para la que tenemos muchos ejemplos naturalmente
 - Predecir una palabra dado su contexto, o un contexto dada una palabra
 - Reconstruir una imagen
- Eliminar la capa de predicción de la red
- La capa anterior a la de predicción es la nueva caracterización de los objetos
 - Menos características → acercándonos a las causas latentes!
- Se usa la red para convertir los objetos del espacio original al espacio de embeddings
- Es relativamente barato de obtener
- Ahora podemos caracterizar datos supervisados con información poblacional de grandes cantidades de datos no supervisados

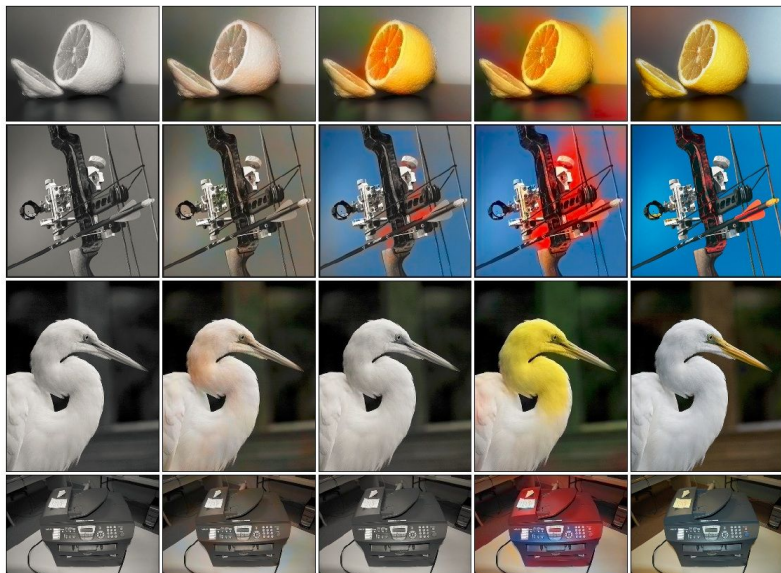
Tareas de pretexto



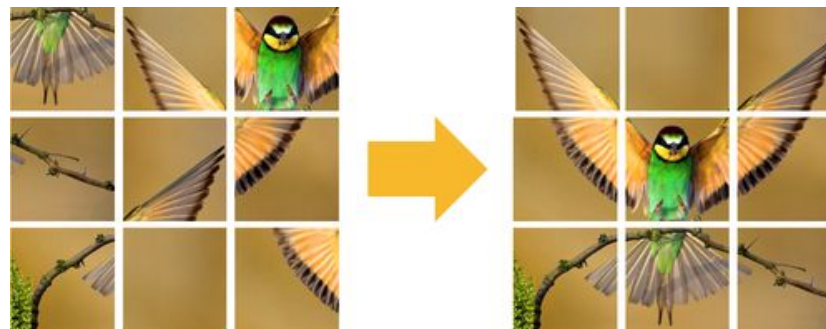
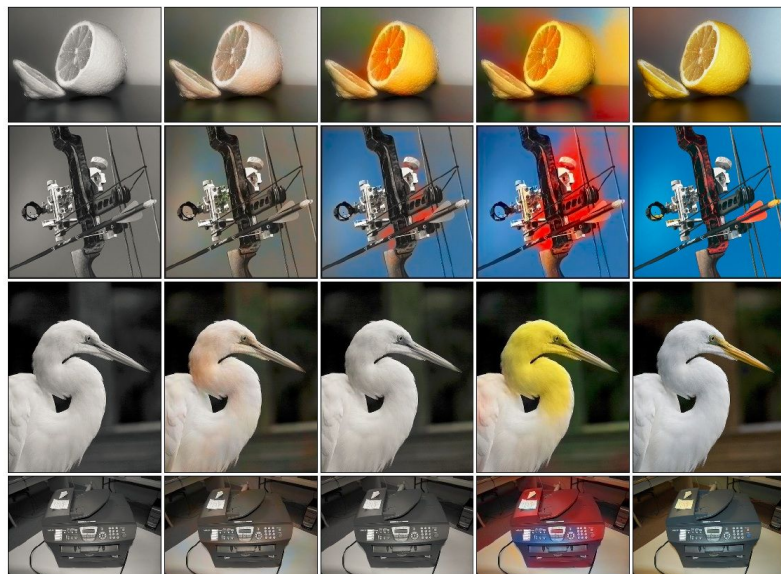
Tareas de pretexto



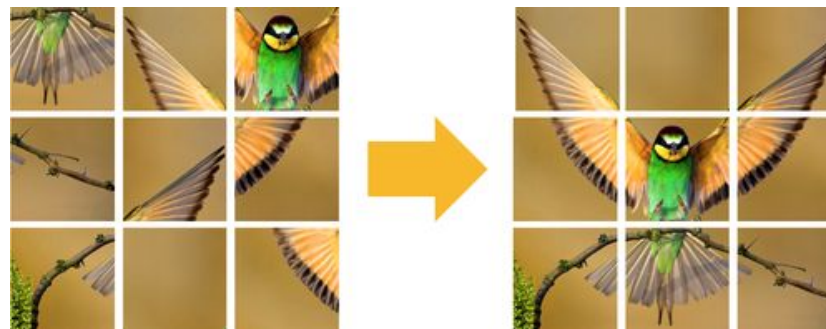
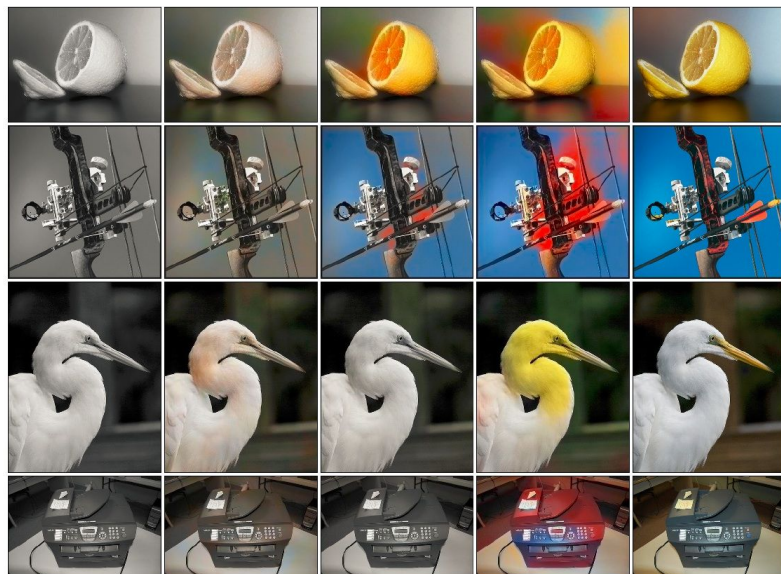
Tareas de pretexto



Tareas de pretexto

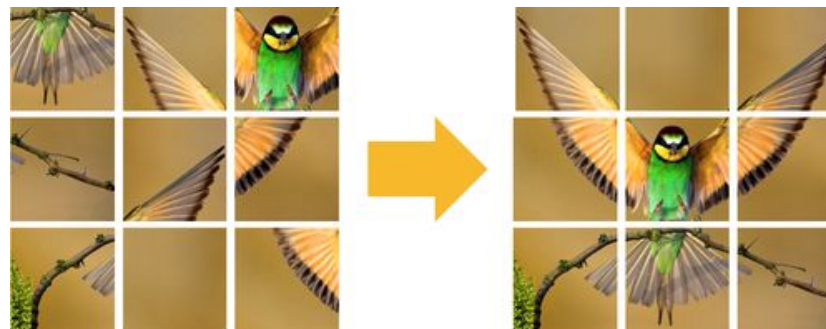
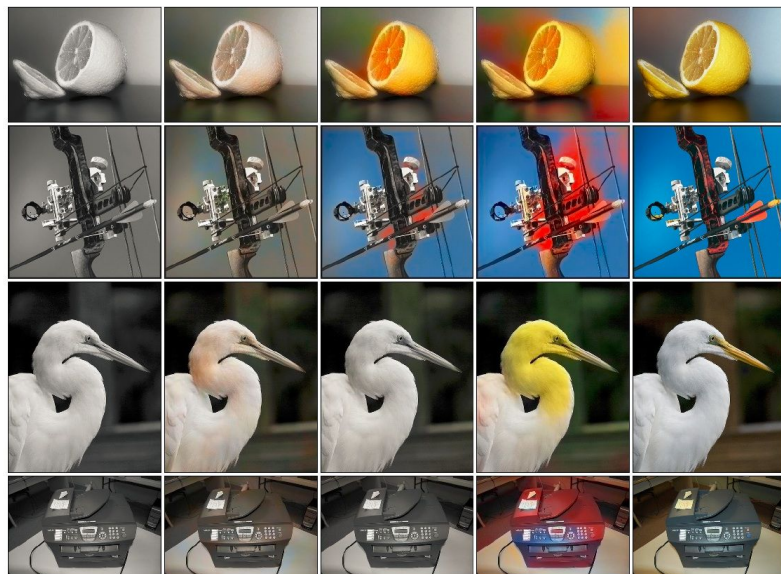


Tareas de pretexto



el gato come pescado

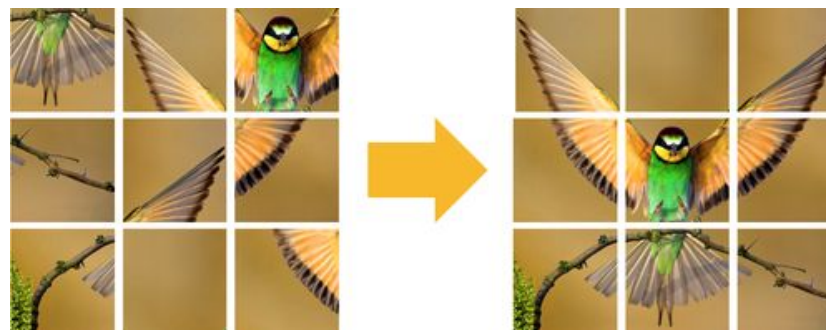
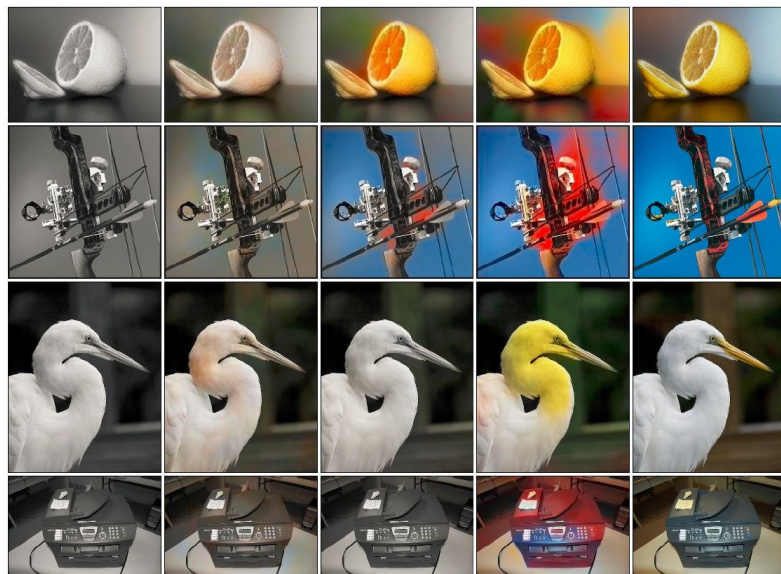
Tareas de pretexto



el gato come pescado

__	gato	come	pescado	?
el	__	come	pescado	?
el	gato	__	pescado	?
el	gato	come	__	?

mejores representaciones

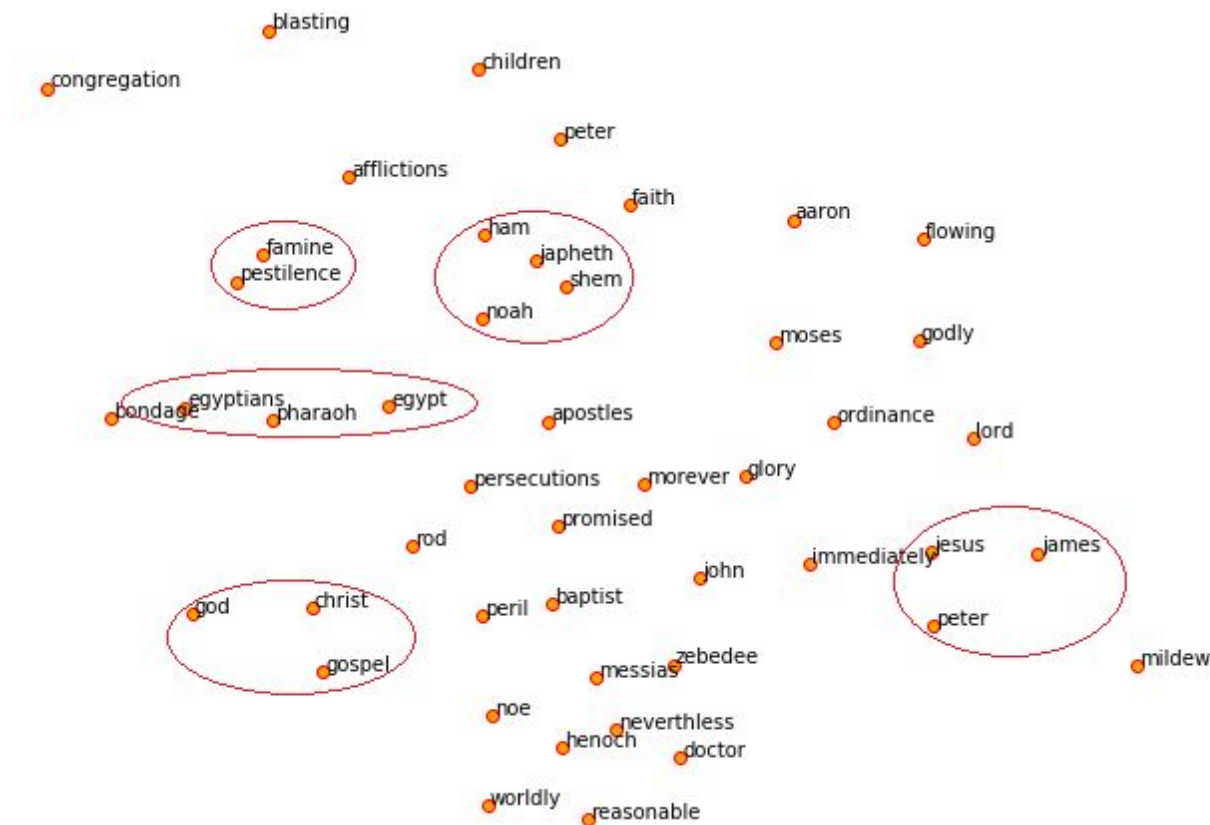


el gato come pescado

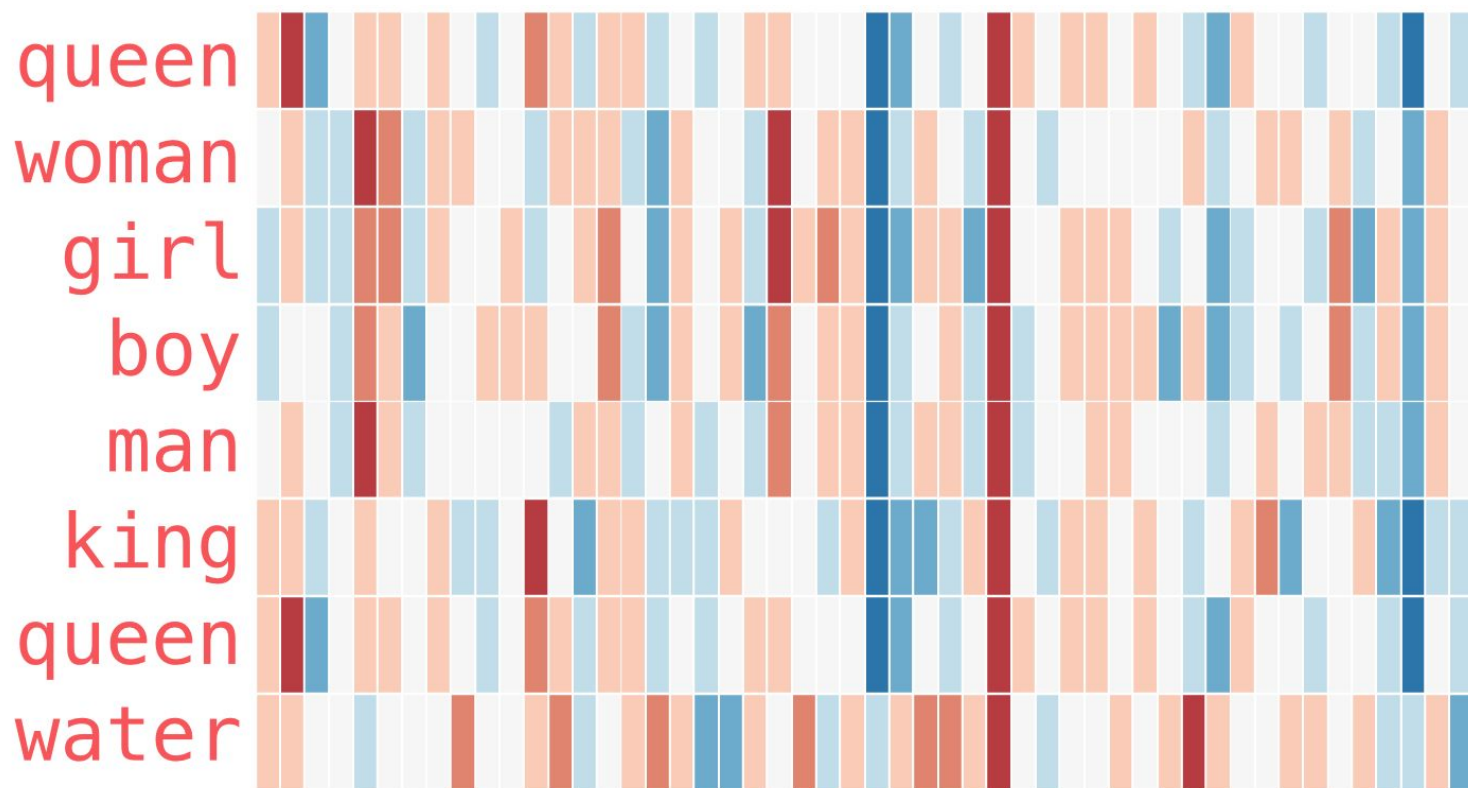
__	gato	come	pescado	?	el
el	__	come	pescado	?	gato
el	gato	__	pescado	?	come
el	gato	come	__	?	pescado

Tarea de pretexto para embeddings

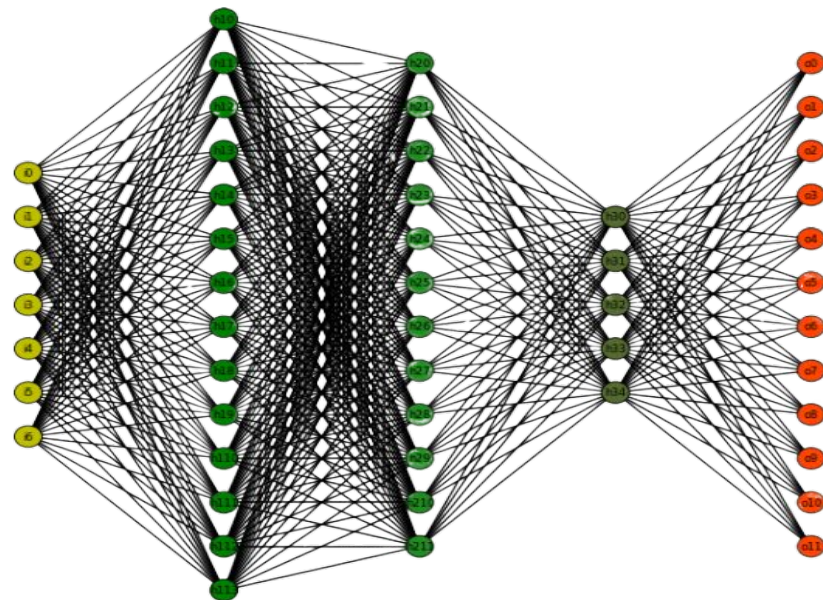
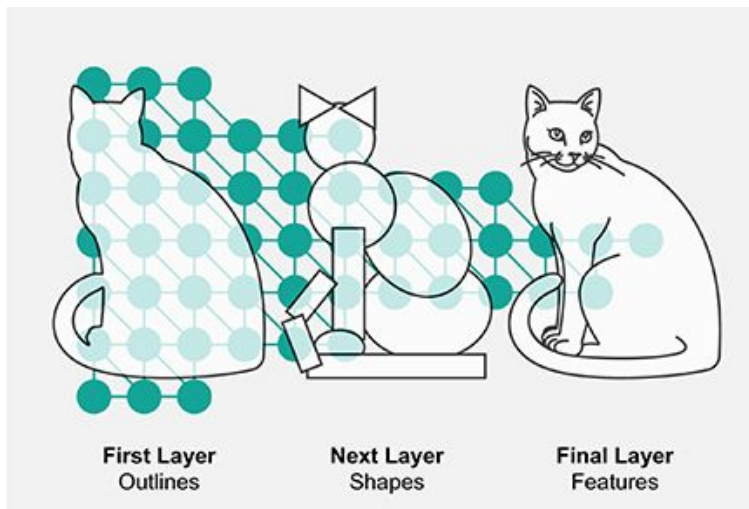
semántica de las palabras



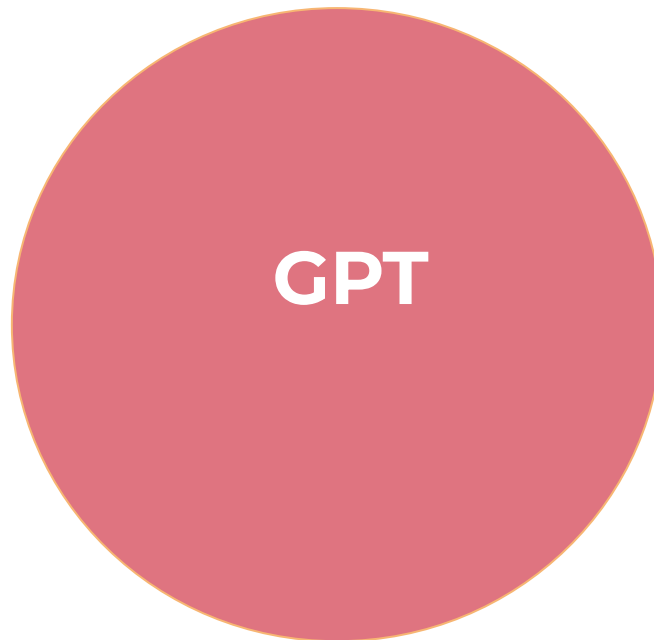
Nos acercamos a las causas latentes



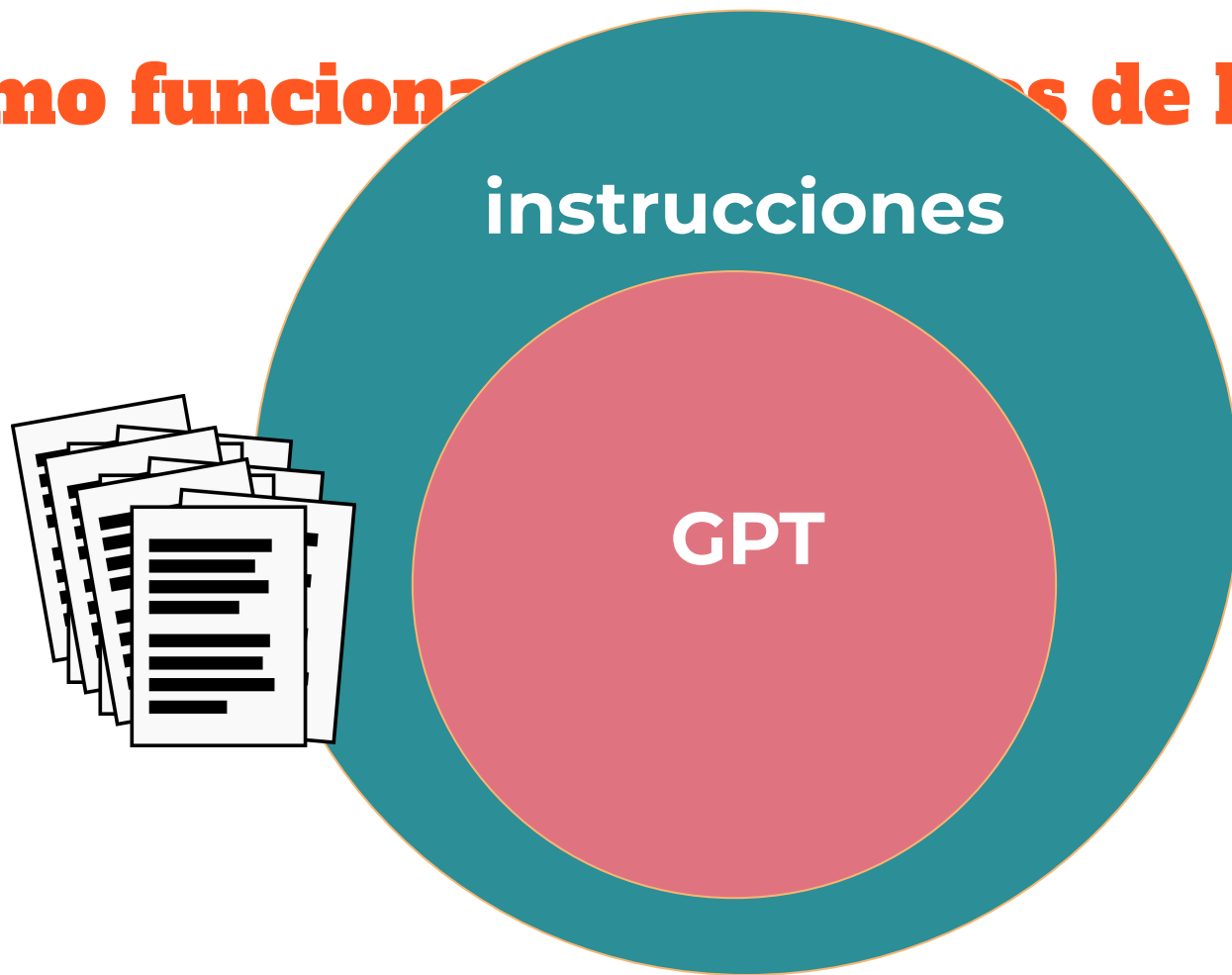
Perdemos interpretabilidad



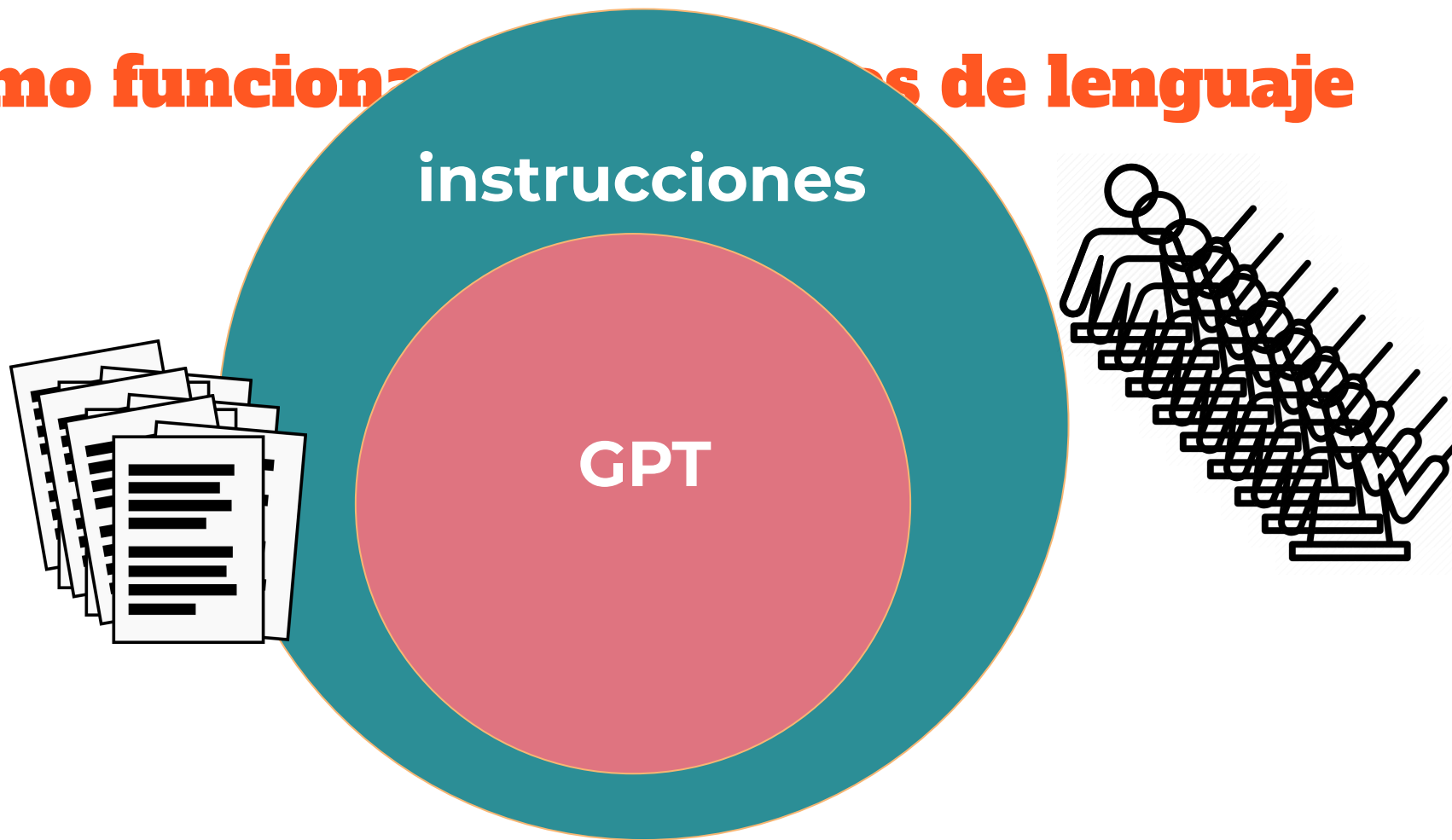
Cómo funcionan los modelos de lenguaje



Cómo funcionan los modelos de lenguaje



Cómo funcionan los modelos de lenguaje



Cómo funcionan los modelos de lenguaje

