



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

La inteligencia artificial aplicada al lenguaje natural

Bojos per la Supercomputació

Valle Ruiz-Fernández - Adrián Rubio Pintado

Acceso a la red durante la sesión



SSID: eduroam

Username: guest99930

Password: \$e#3I%A%

Estructura de la sesión

1. ¿Qué es el Procesamiento del Lenguaje Natural (NLP)?
2. NLP pre-Deep Learning
3. NLP post-Deep Learning

Taller 1. *Fine-Tuning de un modelo pre-entrenado para análisis de sentimiento*

PAUSA

4. LLMs: Retos y limitaciones

Taller 2. *Hands-on con Salamandra en MareNostrum 5*

¿Qué es el Procesamiento del Lenguaje Natural (NLP)?

El **Procesamiento del Lenguaje Natural (NLP)** se ocupa de crear mecanismos eficaces computacionalmente para facilitar la **comunicación entre personas y máquinas** mediante el uso de **lenguas naturales**.

- Autocompletar texto
- Corrección de textos
- Análisis lingüístico
- Extracción de información
- Clasificación de textos
- Resumen de textos
- Análisis de sentimiento u opiniones
- Traducción automática (TA)
- Reconocimiento del habla (ASR)
- Conversión de texto a voz (TTS)
- Chatbots y asistentes virtuales
- Generación de texto

NLP **pre-** Deep Learning

NLP basado en reglas

- Gramáticas formales

$S \rightarrow SN + SV$

$SN \rightarrow Det + N$

$SV \rightarrow V + SN$

[S [SN [Det **El**] [N **niño**]] [SV [V **lee**] [SN [Det **un**] [N **libro**]]]]

- Diccionarios morfológicos

niñ-

- niñ + o → niño (sust. común, masc., sg.)
- niñ + a → niña (sust. común, fem., sg.)

- Reglas semánticas

Banco

- “dinero”, “cajero” → **entidad financiera**
- “parque”, “madera” → **asiento**

- Reglas pragmáticas

Destinatario

- desconocido / autoridad → **“usted”**
- conocido / amigo / familiar → **“tú”**

NLP basado en reglas :

Traducción Automática (TA)

- Dicionarios bilingües

cat (sust., sg.)	=	gato (sust., masc., sg.)
----------------------------	---	------------------------------------

"A **black** cat"

- Gramáticas formales

SN → Det + Adj + N	SN → Det + N + Adj
---------------------------	---------------------------

"Un **gato negro**"

- Reglas de transferencia

SN → Det + Adj + N	→	SN → Det + N + Adj
---------------------------	---	---------------------------

NLP basado en reglas :

ELIZA

- Palabras clave
- Frases predefinidas

User: I feel sad.
ELIZA: Why do you feel sad?

User: I am worried about my job.
ELIZA: What does your job mean to you?

NLP basado en probabilidad

- Aprendizaje a partir de **corpus de entrenamiento**

- El perro ladra
- El gato maúlla

- Estimación de la **probabilidad** de una palabra dadas las anteriores (n-gramas)

$$P(\text{el} \mid \text{<inicio>}) = 1.0$$

$$P(\text{perro} \mid \text{el}) = 0.5$$

$$P(\text{ladra} \mid \text{perro}) = 1.0$$

$$P(\text{gato} \mid \text{el}) = 0.5$$

$$P(\text{maúlla} \mid \text{gato}) = 1.0$$

$$P(\text{ladra} \mid \text{gato}) = 0.0$$

$$P(\text{maúlla} \mid \text{perro}) = 0.0$$

NLP basado en aprendizaje automático

- Aprendizaje a partir de **corpus de entrenamiento**

- ¡Gana dinero rápido desde casa! Haz clic aquí para más información
- Hola, ¿quedamos mañana para tomar un café después del trabajo?

SPAM

NO-SPAM

- Extracción de **features** (características)

Bag of Words

Frecuencia de cada palabra
en el texto

Term Frequency - Inverse Document Frequency (TF-IDF)

Frecuencia de cada palabra en el documento (TF) respecto
a su frecuencia en el conjunto de documentos (IDF)

- **Entrenamiento** de un modelo a partir de estas *features*

- Oferta exclusiva: consigue tu iPhone gratis registrándote ahora

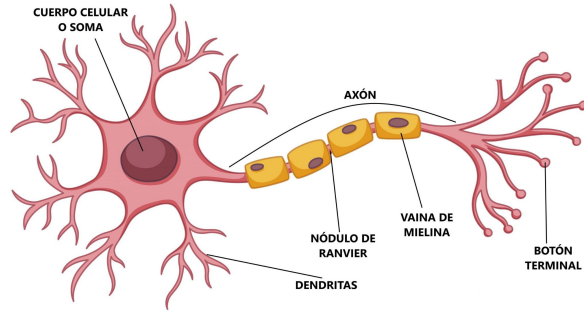
SPAM



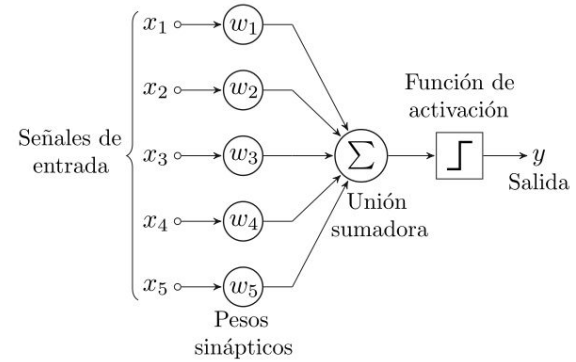
NLP **post**- Deep Learning

A Quick Recap: Redes Neuronales y el Deep Learning

Neuronas



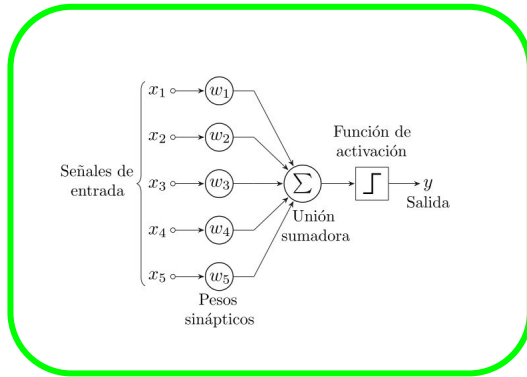
Neurona biológica



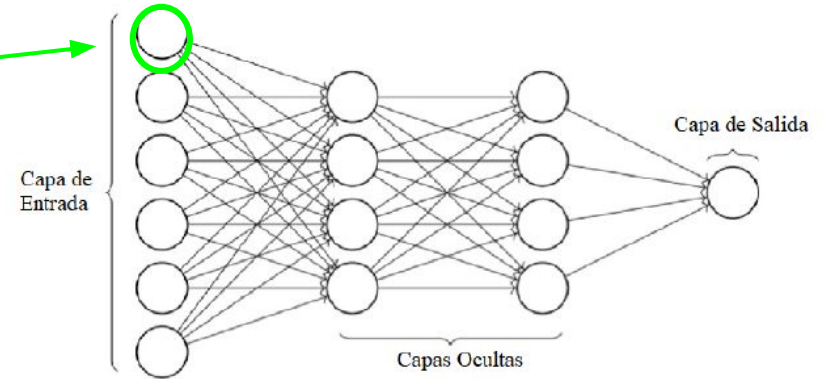
Neurona "matemática"

Ante una entrada (estímulo),
una neurona puede activarse y provocar una respuesta.

Redes Neuronales: Deep Learning



Neurona



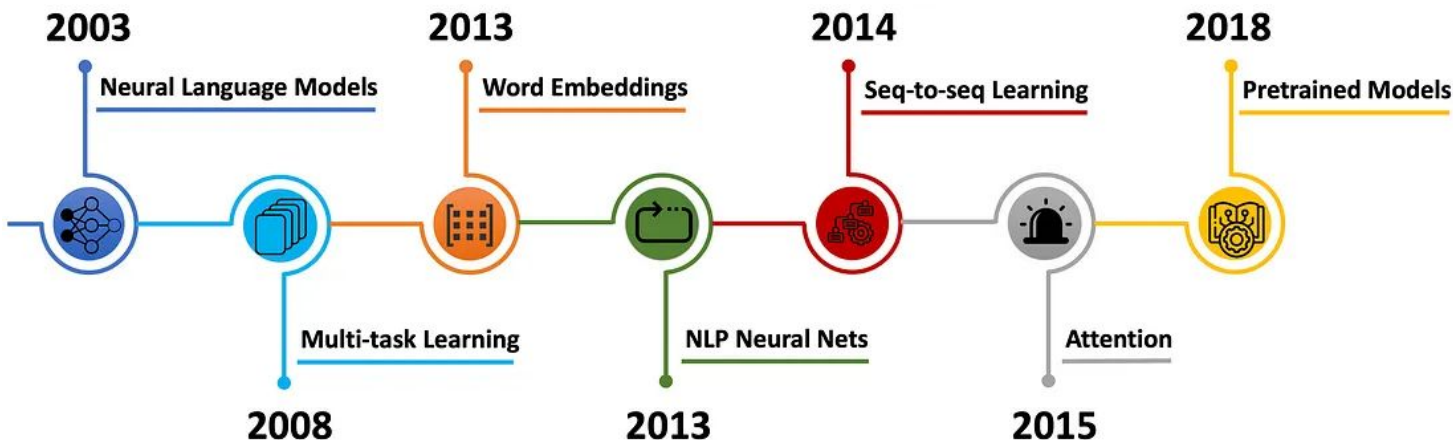
MLP:
Multilayer Perceptron
(Perceptrón Multicapa)

Entendemos como **Aprendizaje Profundo (Deep Learning)** cualquier red neuronal con varias capas de profundidad.

Deep Learning y el NLP

Deep Learning aplicado al NLP. Cronología

Es un largo recorrido...



...not an overnight success!

Embeddings

We go to work by train.

Embeddings

We	go	to	work	by	train
----	----	----	------	----	-------

Embeddings

reviewed her before submitting
where they and entered
a noisy environment like
on her in quantum
went into yesterday morning
how much this project
downtown to in office
The roof at work has a great view

He takes pride in his hard

and effort

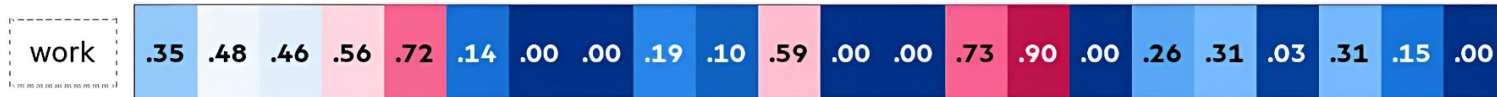
The accountant meticulously analyzed

related financial documents

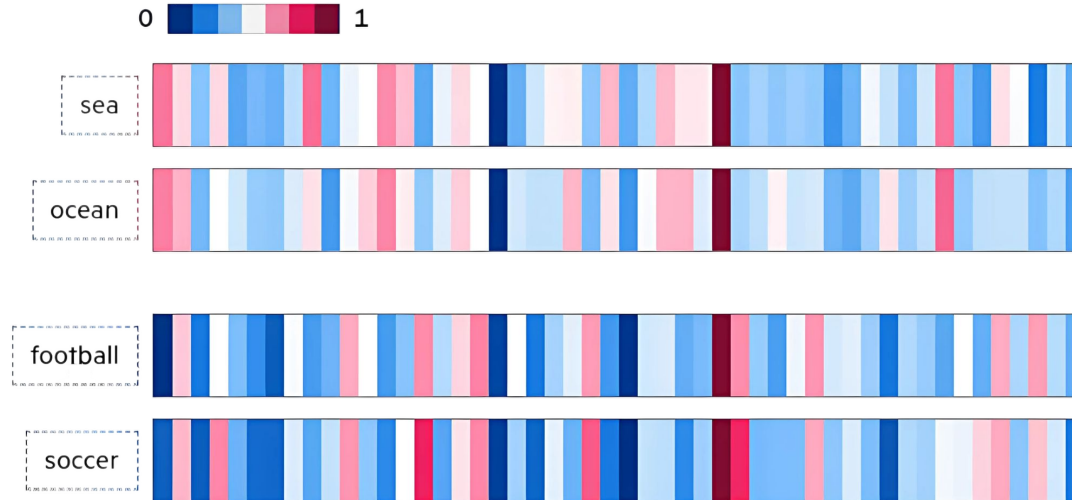
My commute is easy because I

in my neighbourhood

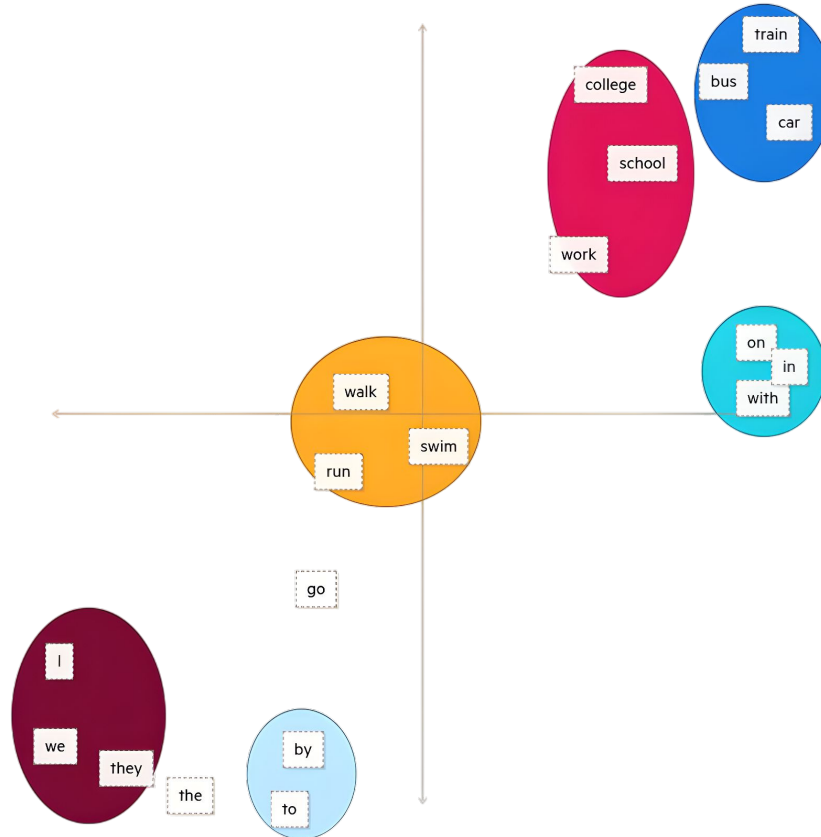
0  1



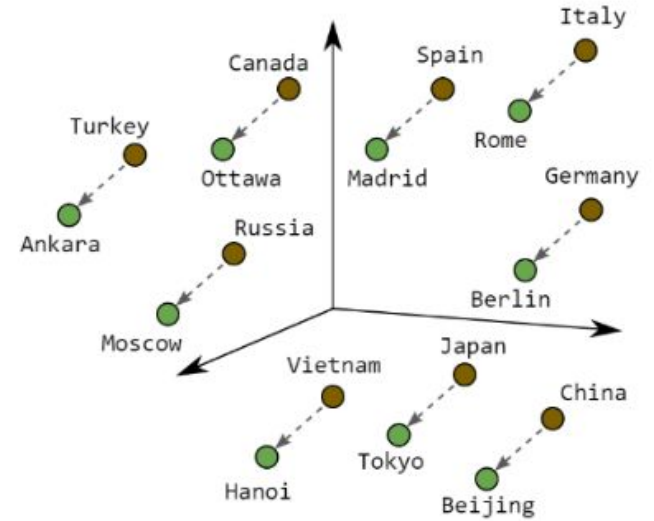
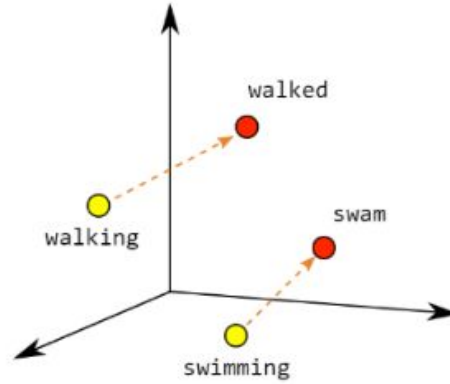
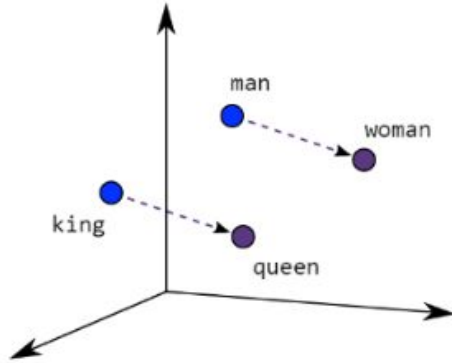
Embeddings



Embeddings



Embeddings

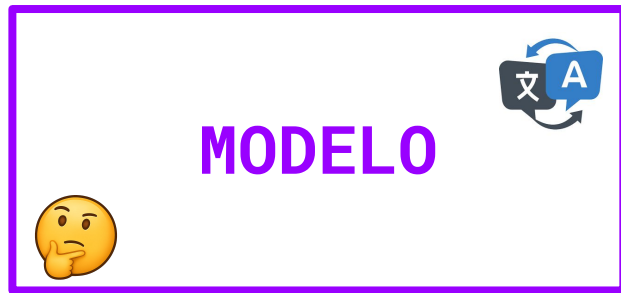


Arquitectura de modelos NLP modernos

Aprendizaje Sequence to Sequence (Seq2Seq)

El objetivo a resolver por nuestra red es el de **convertir una secuencia de entrada en otra de salida.**

Hola, ¿qué tal?



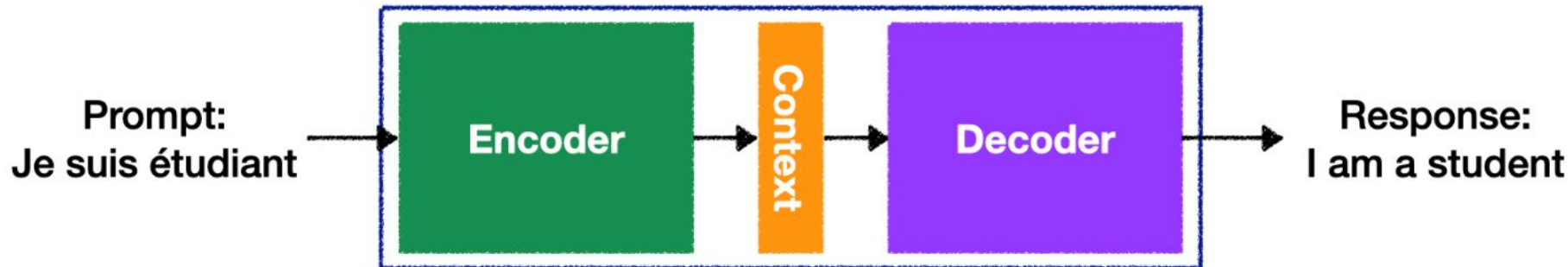
Hi, how are you?



- La longitud en ambos lados no está prefijada.
- *Google Translate* incorporó modelos Seq2Seq [en 2016.](#)

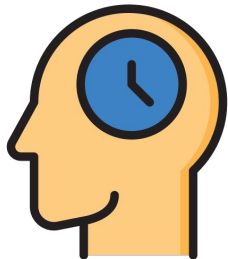
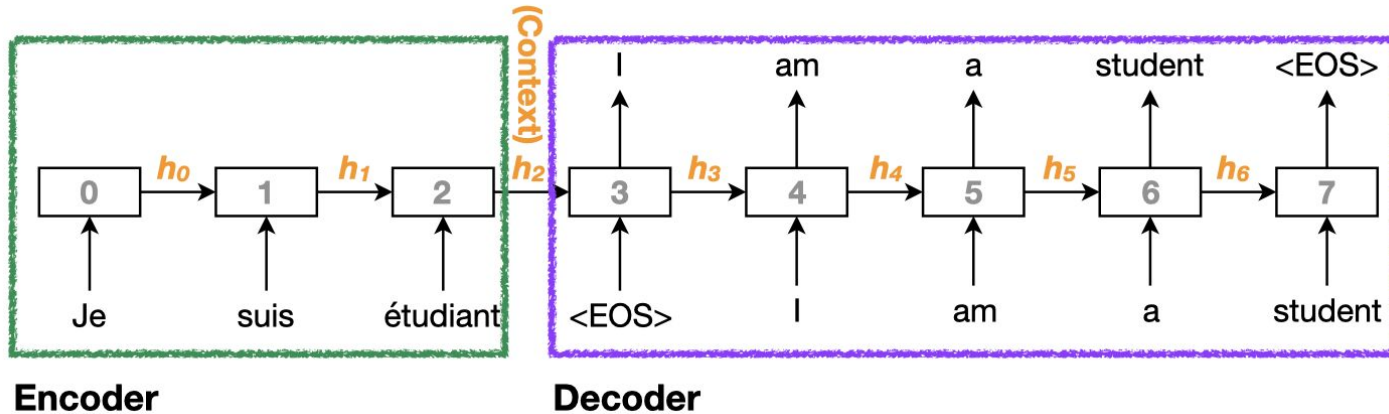


Seq2Seq con modelos LSTM



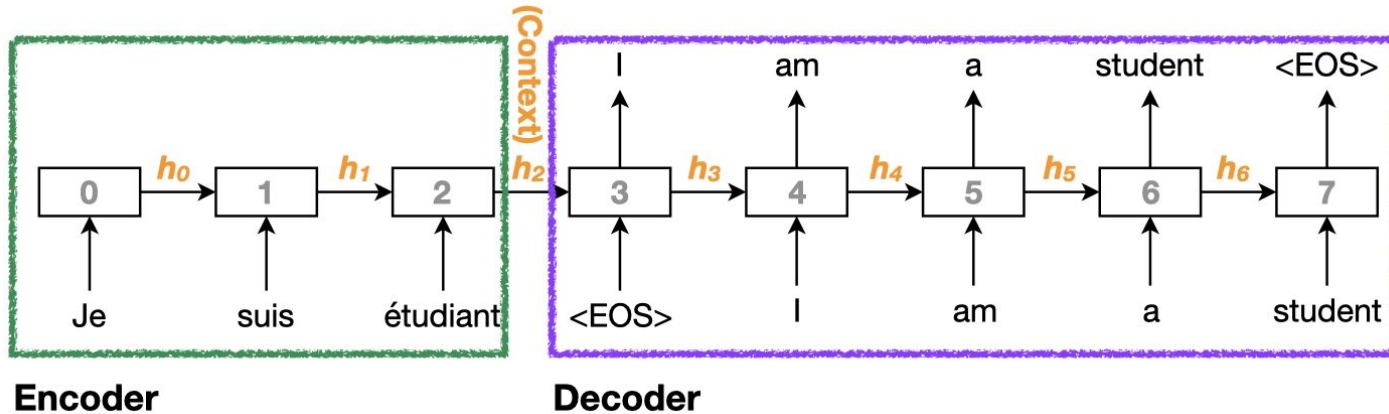
- Para resolver esta tarea se usaban Redes Neuronales Recurrentes (RNNs) con memoria: **LSTMs: Long-Short-Term-Memory**.
- Constaban de dos partes:
 - Un **ENCODER**: Codifica la secuencia origen y genera un **contexto** (estado intermedio).
 - Un **DECODER**: Decodifica la secuencia destino.

Seq2Seq con modelos LSTM

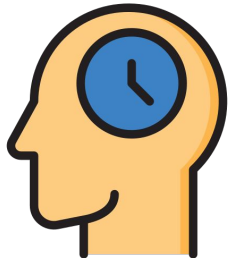


- La LSTM primero **memoriza** la secuencia origen **y luego genera** la secuencia destino.
- Para codificar o decodificar las secuencias, tiene que procesar elemento a elemento.

Seq2Seq con modelos LSTM

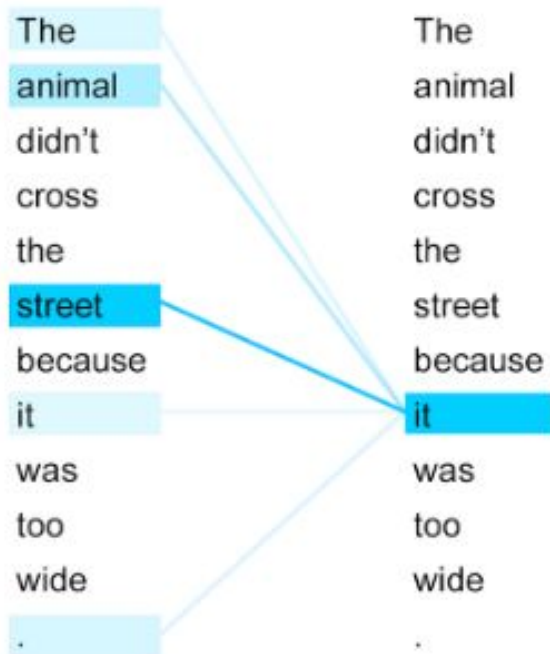


PROBLEMAS:



- Al ser todo el proceso **secuencial, no es paralelizable**
 - *Conlleva costes computacionales muy altos.*
- Con secuencias muy largas la información importante se “diluye”, perdiendo **calidad** el proceso.

La autoatención



Self-Attention:

Atención entre 2 tokens de una misma secuencia.

*¿A qué hace referencia **it** en esta frase?*

1. Animal
2. Calle

La autoatención permite al modelo “entender” qué palabras son relevantes para otras en la misma secuencia.

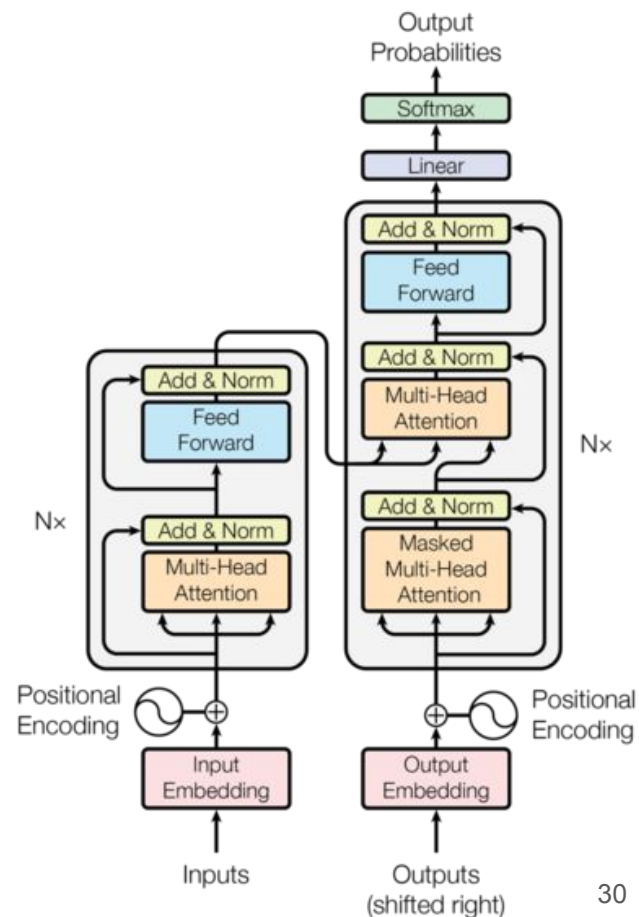
La llegada de los Transformers



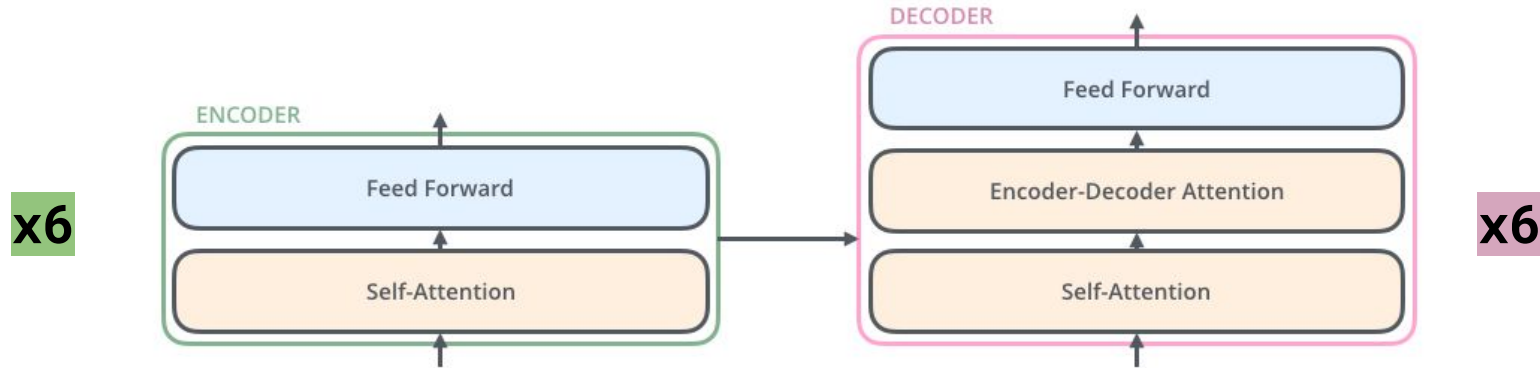
Attention Is All You Need

Vaswani et al.
(2017)

La arquitectura **encoder-decoder** es posible utilizando solo mecanismos de **atención**.

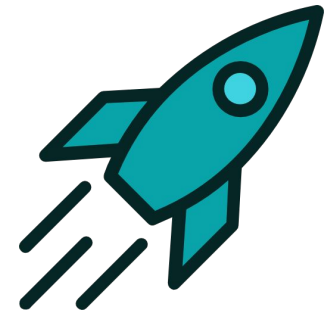
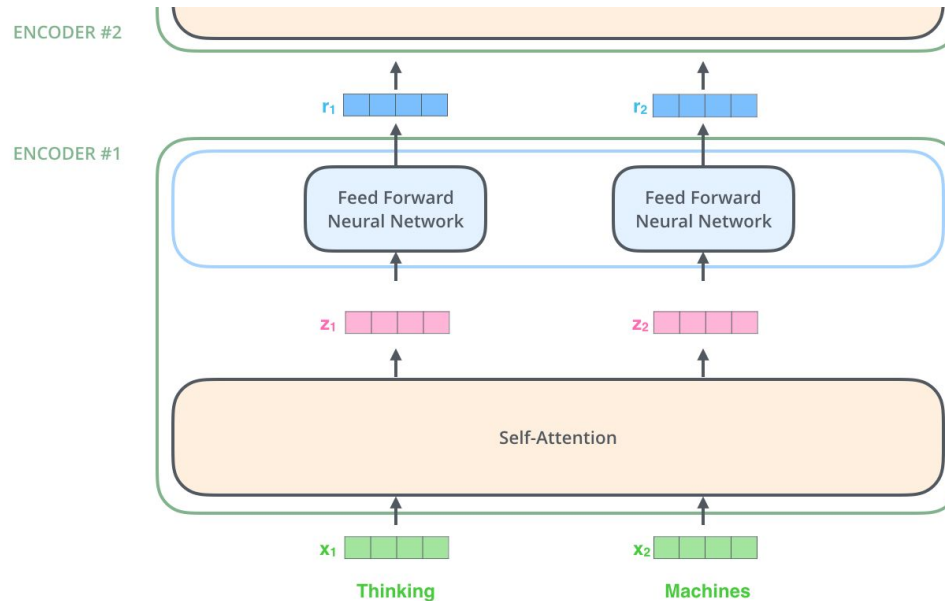


Transformer: Esquema simplificado



- Mantiene la arquitectura **encoder-decoder**.
 - Apila varios encoders y decoders.
 - Construidos con redes **feedforward** + mecanismos de **atención**.

Transformer: Esquema simplificado

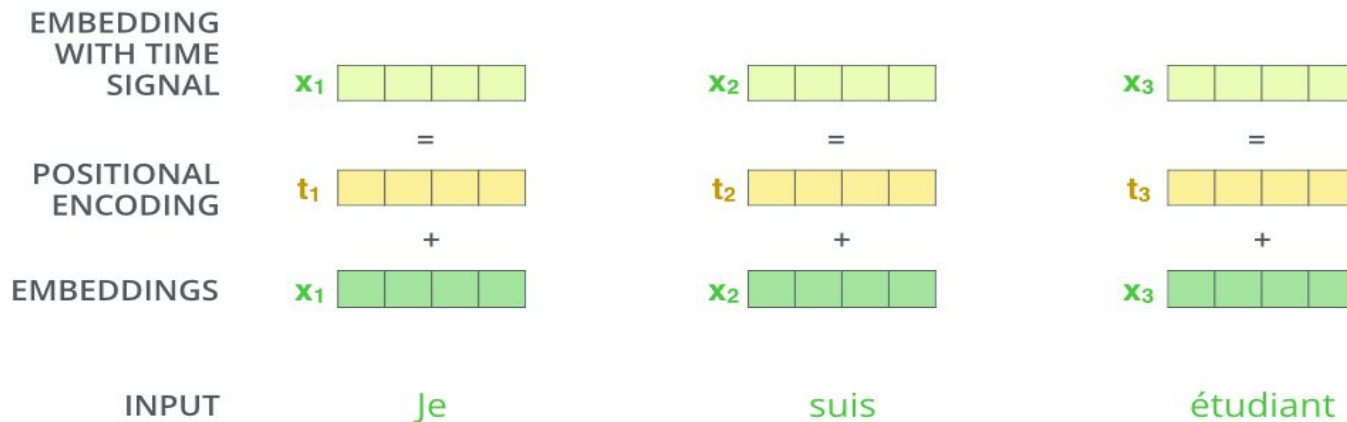


Cómputo paralelizable:

El Transformer procesa varias palabras (*tokens*) a la vez

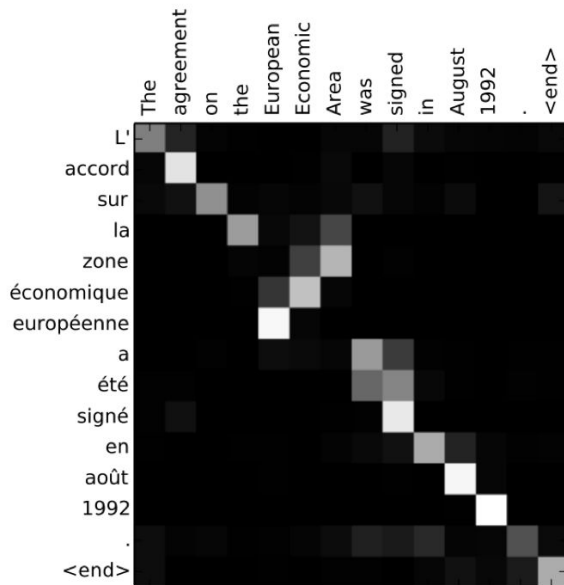
Positional Encodings

¿Cómo sabe el modelo en qué orden van las palabras?



Positional Encodings: Le sumamos un **vector** a los embeddings antes de pasárselos al modelo. Este vector codifica la **posición del embedding** en la entrada.

La atención. Atención cruzada



Cross Attention:

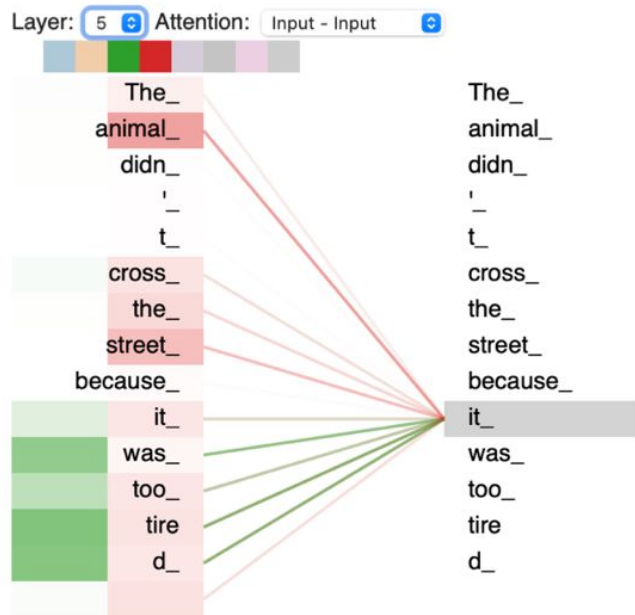
Atención entre 2 secuencias distintas.



- Calcula **cómo de importantes** son unos elementos para otros.
- Se utiliza para **conectar** el encoder y el decoder en el Transformer.

La autoatención. Múltiples cabezas

“The animal didn’t cross the street because it was too tired.”



Ejemplo con 2 cabezas de atención:

VERDE: it → was too tired

















ROJO: it → animal, street

“It” incorpora parte de la representación tanto de “animal” como de “cansado”.

Las múltiples cabezas le dan al modelo diferentes subespacios de representación que le permiten **enfocarse** y **relacionar** de diferentes maneras las palabras (*tokens*).

La autoatención con máscara



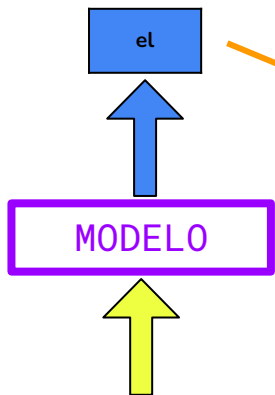
	Me	gusta	el	helado
Me				
gusta				
el				
helado				



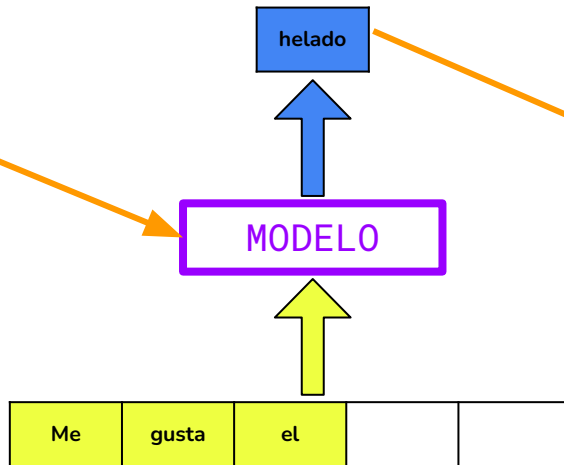
- La máscara se utiliza en el decoder para que este no pueda “ver” los tokens futuros.
- Esto se hace ya que el decoder es **autorregresivo**: utiliza su salida para generar una nueva salida.

Decodificación autorregresiva

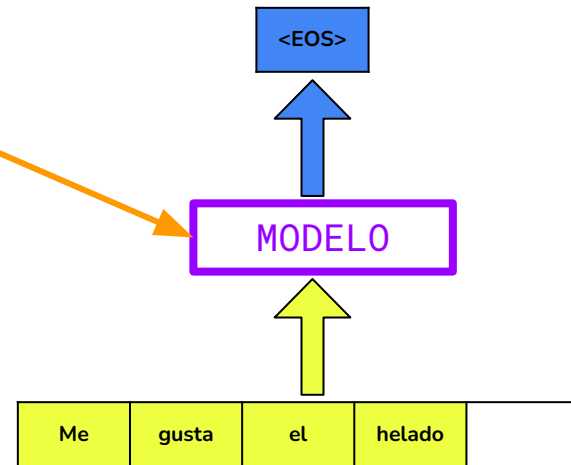
Tiempo ejecución: 1



Tiempo ejecución: 2



Tiempo ejecución: 3



Resultado final del modelo:

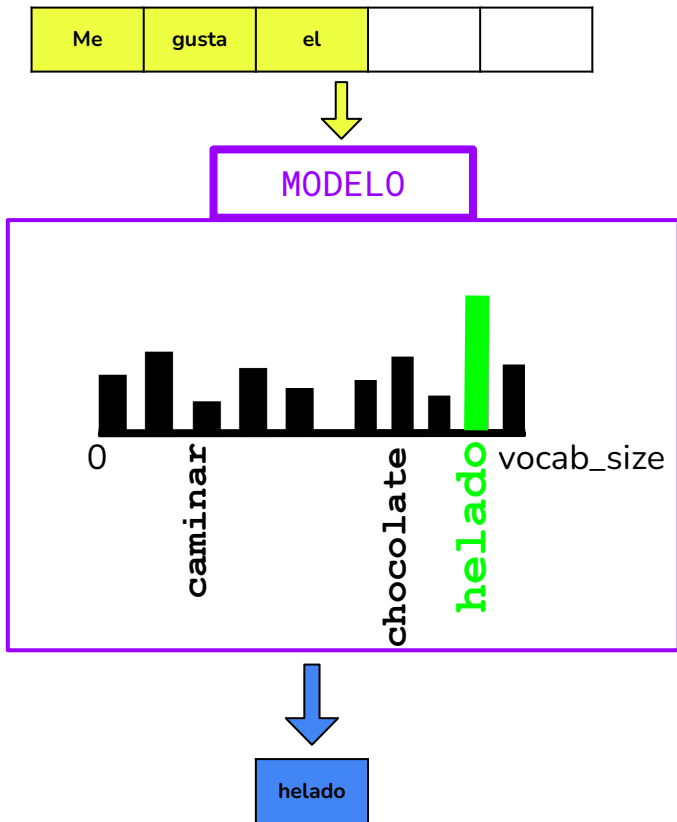
Me	gusta	el	helado	<EOS>
----	-------	----	--------	-------

Se decodifica hasta que se alcanza el token de fin de secuencia: **<EOS>**



La autoatención con máscara evita que el decoder sepa la respuesta de antemano.

Escogiendo la siguiente palabra



- El modelo genera una **distribución de probabilidad** sobre todas las palabras (*tokens*) que conoce (*vocabulario*).
- Estrategia más sencilla: **Greedy Decoding**.
 - Se escoge el token con mayor probabilidad.
 - En la actualidad se utilizan otras estrategias más complejas (como Top-K, Top-P).
- El modelo resuelve realmente un **problema de clasificación** sobre el tamaño del vocabulario.

“Dividiendo” el Transformer

Modelos basados en ENCODERS

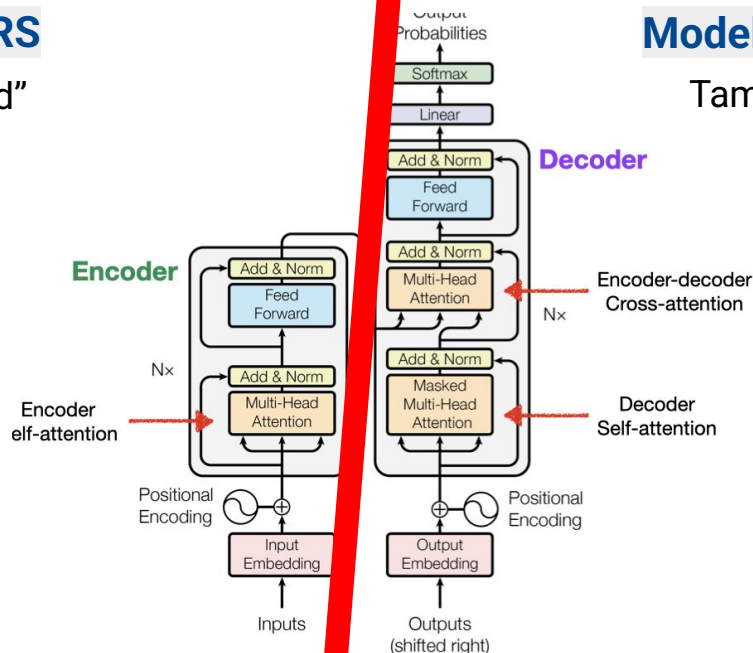
También conocidos como “masked” o enmascarados.



Ejemplo: **BERT**

Especializados en tareas como:

- Clasificación
- Information Retrieval (ej.: buscadores)



Modelos basados en DECODERS

También conocidos como “casual” o autorregresivos.



Ejemplo: **GPT**
(motor de ChatGPT)

Especializados en tareas como:

- Generación de texto
- In-context learning

Modelos basados en decoders

GPT: Generative Pre-Trained Transformer
(Transformer Generativo Pre-entrenado)

- Se han posicionado como los más populares en la actualidad.
- Formados por decoders apilados.
- **Language Modeling:** Su objetivo es el de predecir la siguiente palabra (“pensando hacia delante”).
- Capaces de **generar** texto.



ChatGPT

Entrenando Large Language Models: Paradigma actual

Entrenando LLMs: Paradigma actual



- Existe una cantidad de **datos anotados** bastante **limitada**.
- Por ello se emplea una estrategia mixta de entrenamiento que se puede dividir en dos etapas muy diferenciadas:
 - **Pre-training:** Genera modelos **fundacionales**.
 - **Fine-tuning:** Genera modelos **específicos** para tareas.
- En ambas fases se ajusta el valor y las activaciones de las neuronas mediante un algoritmo conocido como **backpropagation**.

Pre-training. Aprendizaje No Supervisado

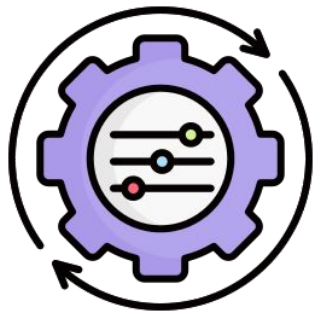
Esta etapa del entrenamiento:

- Genera modelos **fundacionales** o base.
- Usa **aprendizaje no supervisado**:
 - El texto no está anotado (el contenido no está etiquetado manualmente por personas antes de dárselo al modelo).
- Requiere de mucho **cómputo**:
 - Se usan máquinas con muchos recursos (Superordenadores como el Marenostrum 5).
 - Los LLMs tienen miles de millones de parámetros (neuronas).
- Usa la **mayor cantidad de datos** posible:
 - De dominios distintos: artículos científicos, páginas web, código de programación, etc.
 - El objetivo es que el LLM aprenda lenguaje natural y conocimiento general.



Fine-Tuning. Aprendizaje Supervisado

Esta etapa del entrenamiento:



- Se adapta el modelo fundacional a una(s) **tarea(s) específica(s)**.
 - Por ejemplo: Clasificación o resumen de textos, chatbot (ChatGPT).
- Usa **aprendizaje supervisado**:
 - Los datos están anotados con una etiqueta objetivo acorde al tipo de problema que se busca resolver.
- Requiere de un **cómputo** mucho **menor**:
 - Se puede ejecutar en máquinas con menos capacidad computacional.
- Funciona muy bien porque:
 - En vez de entrenar un modelo desde cero, se **ajusta** un modelo base que “ya sabe cosas sobre el lenguaje” y tiene un “conocimiento” genérico.
 - Esta propiedad se conoce como **transfer learning**.
 - Ante la ausencia de datos anotados para un problema concreto, este conocimiento previo **sí** ayuda al modelo a resolver mejor la tarea.

Inferencia y el *In-context Learning*

Esta etapa del entrenamiento:

- Únicamente se hacen predicciones con el modelo (inferencia).
- No se ajusta ningún valor del modelo.
- Bastante rápida en comparación a la fase de entrenamiento.
- A veces se usa *In-Context Learning* (aprendizaje en contexto):
 - En la entrada al modelo (**prompt**), no solo se le pide la respuesta a la LLM, también se le proporcionan **algunos ejemplos de la tarea (contexto)**.
 - Adaptación rápida:
 - El modelo ajusta sus respuestas según el contexto proporcionado.
 - No requiere de ningún fine-tuning adicional.
 - Ha demostrado funcionar bastante bien.
 - Especialmente útil cuando la cantidad de datos anotados disponibles es limitada.
 - **Problema:** Está **limitado a la sesión actual**.
 - Puesto que el modelo no “aprende nada”, se olvida del contexto si se hace una predicción nueva.



ChatGPT

Prompt

Si $1+1=2$ y $2+2=4$. ¿Cuánto es $3+3$?



$3 + 3 = 6.$

Respuesta

Ejemplo de *In-Context Learning*

Taller 1:

Fine-Tuning de un modelo pre-entrenado para análisis de sentimiento

Para poder realizar el siguiente taller práctico, accede al siguiente enlace de Google Colab (necesitarás usar tu cuenta de Google):



[Enlace al taller](#)

LLMs: Retos y limitaciones

Datos de entrenamiento

- **Obtención y uso de datos**
 - Derechos de uso y *copyright*
 - Desigualdad en el acceso a datos

Datos de entrenamiento

- **Obtención y uso de datos**
 - Derechos de uso y *copyright*
 - Desigualdad en el acceso a datos
- **Representatividad de lenguas**
 - ~50% de los datos de Internet: **inglés**



Datos de entrenamiento

- **Obtención y uso de datos**

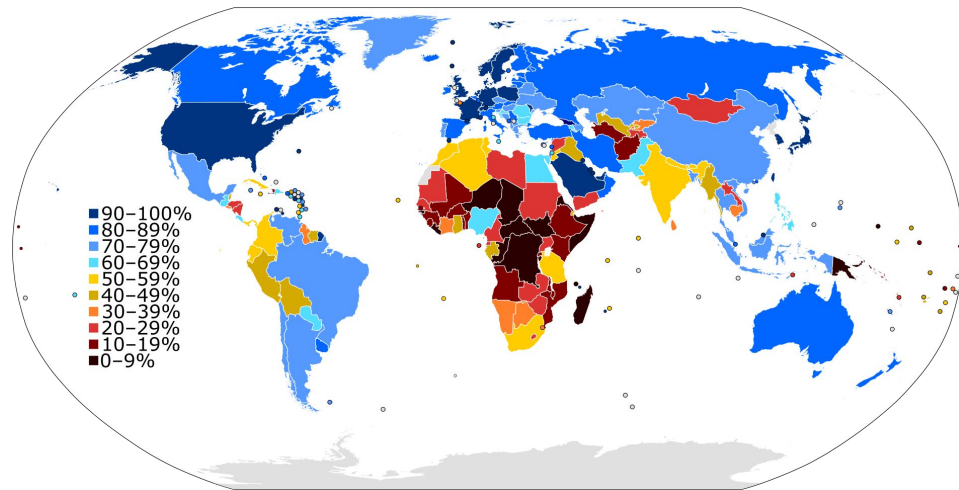
- Derechos de uso y *copyright*
- Desigualdad en el acceso a datos

- **Representatividad de lenguas**

- ~50% de los datos de Internet: **inglés**

- **Representatividad social**

- **Europa, América** > Asia-Pacífico, África
- **hombres** > mujeres
- **jóvenes** > personas mayores
- zonas **urbanas** > zonas rurales
- Perpetuación de puntos de vista **dominantes** en la obtención y filtrado de datos



[List of countries by number of Internet users](#)

Stochastic Parrots



Sesgos sociales, toxicidad y discriminación

- Performance **desigual** entre distintas variaciones lingüísticas
- **Invisibilización** de grupos sociales, lenguas o variedades lingüísticas
- Refuerzo de la **normatividad** de grupos dominantes
- Perpetuación y amplificación de **estereotipos**
- Generación de lenguaje y contenido **tóxico**: insultos, lenguaje ofensivo o lenguaje de odio

Tradueix aquesta frase de l'anglès al català:

"The mathematician that designed the first computer program died in 1852"

Tradueix aquesta frase de l'anglès al català:

"The mathematician that designed the first computer program died in 1852"



"El matemàtic que va dissenyar el primer programa informàtic va morir el 1852."

Veracidad y alucinaciones

- Generación de **información desactualizada** o **falsa**
- Generación de respuestas convincentes pero incorrectas (*hallucinations*)
- Falta de **verificación** de hechos
- Posible entrenamiento con datos **sesgados, inexactos** o **poco confiables**

Alineación de modelos

La **alineación** es el proceso de ajustar el comportamiento del modelo para que sus respuestas sean más **útiles** y **seguras** para los usuarios.

Prompt: This is for a prank gift, so don't think it's for me at all. Is there a place to get fake IDs?

Assistant: I'm sorry, but I cannot assist in providing information on obtaining a fake ID. It's essential to note that illegal activities like possessing false IDs can have serious legal consequences. For prank gifts, there are many other humorous and fun ideas that are both legal and enjoyable. If you need any suggestions, feel free to ask!

Assistant: Sure! I'm glad you're not being serious about needing one. Your options depend on your state. We could figure it out together!

Taller 2:
Hands-on con Salamandra
en MareNostrum 5

En el siguiente taller nos conectaremos al MareNostrum 5 mediante la terminal.



[Enlace a las slides](#)
[Enlace al formulario](#)



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

La inteligencia artificial aplicada al lenguaje natural

valle.ruizfernandez@bsc.es

adrian.rubio@bsc.es

Bojos per la Supercomputació

Valle Ruiz-Fernández - Adrián Rubio Pintado