

Interpretabilidad del Deep Learning

Conceptos de la XAI

Christian Oliva Moya

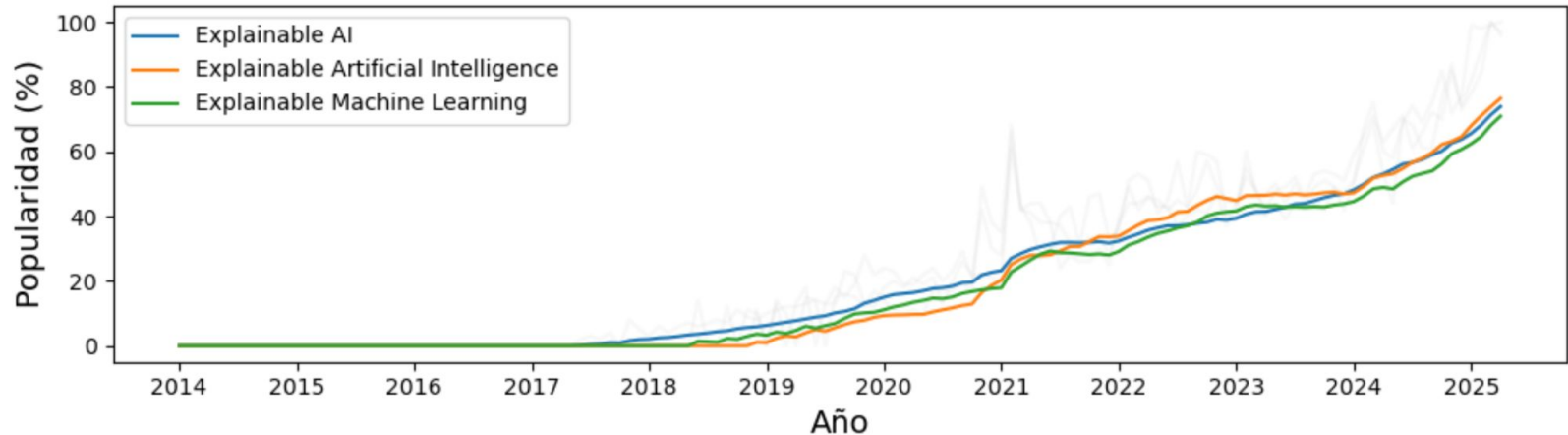
Contenido del curso

- Primera semana:
 - **Introducción de conceptos**
 - Explicabilidad genérica
 - SHAP
 - LIME
 - Coeficiente de Explicabilidad-Rendimiento

¿Qué es la IA eXplicable (XAI)?

- Capacidad para entender y explicar las decisiones de los modelos de ML
- Los modelos de Deep Learning (DL) **son opacos** (cajas negras)

Google Trends a día 14 de abril de 2025



¿Por qué es importante la XAI?

- Protección de datos sensibles
- Detección de sesgos y discriminación intrínseca en los datos
- ¿Qué modelos son explicables?
 - Árboles de decisión
 - Modelos lineales
- **Sin XAI no hay Deep Learning**

Ejemplos de sesgos y discriminación de modelos de ML

- Caso COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) - Alto **grado racista** a la hora de clasificar reincidencia [1]
- Evaluación de sistemas comerciales con reconocimiento facial - La tasa de **error aumenta con mujeres negras** [2]
- Embeddings como Word2Vec o GloVe asocian la palabra **mujer** con términos relacionados con el **hogar** y la **familia**, mientras que **hombre** se asocia a **trabajo** y **dinero** [3]

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks", 2016

[2] J. Buolamwini, and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", in PMLR, 2018

[3] A. Caliskan, J. J. Bryson, and A. Narayanan. "Semantics derived automatically from language corpora contain human-like biases", in Science, 2017

Ejemplo de la importancia de la XAI con ataque adversario

- Si no sabemos qué hace nuestro modelo, nos pueden pasar cosas inesperadas, como un ataque adversario FGSM [4]



x

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“gibbon”

99.3 % confidence

Fundamentos de la XAI (I)

- **Sin XAI no hay Deep Learning**
 - Sin XAI no hay confianza en los modelos
 - Sin XAI no se cumplen los marcos regulatorios europeos
 - Sin XAI solo podemos usar árboles de decisión y modelos lineales
- La XAI se encarga de responder a la siguiente pregunta:

¿Por qué el modelo toma esta decisión?

Fundamentos de la XAI (II)

- ¿Por qué el modelo toma esta decisión?
 - “Porque se está fijando en estos atributos de entrada”
 - “Porque se parece a estos otros datos de entrada que sí sé qué son”
 - “Porque entiendo los mecanismos que hacen que dé esta respuesta”

Todas son respuestas válidas y por eso diferenciamos:

Explicabilidad

Interpretabilidad

Fundamentos de la XAI (III)

- **Explicabilidad:** Identificación de los atributos del conjunto de datos que tienen una mayor influencia en las predicciones de un modelo
- Responde a la pregunta:

¿Por qué este dato de entrada genera esta respuesta en el modelo?

- Los algoritmos de explicabilidad definen una **relevancia** R_{x_i} para cada atributo de entrada

Fundamentos de la XAI (IV)

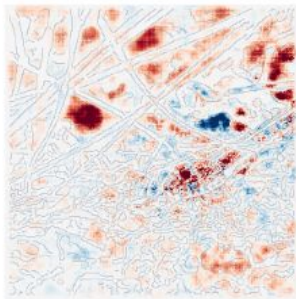
- **Explicabilidad:** definen una **relevancia** Rx_i para cada atributo de entrada

Ejemplo de oclusión a diferentes ventanas sobre una imagen [5]

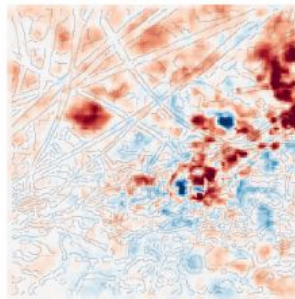
Original (label: "garter snake")



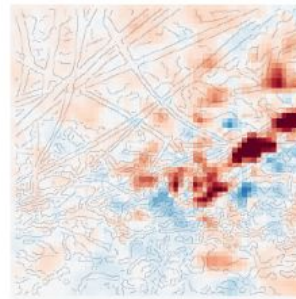
Occlusion-1



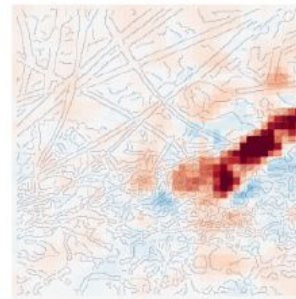
Occlusion-5x5



Occlusion-10x10

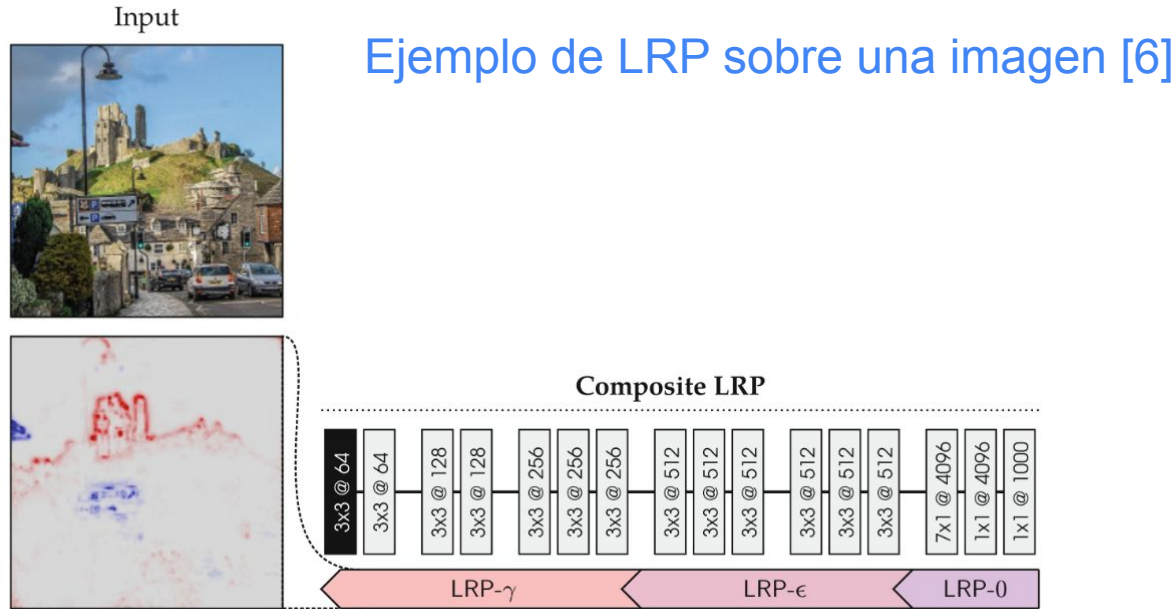


Occlusion-15x15



Fundamentos de la XAI (V)

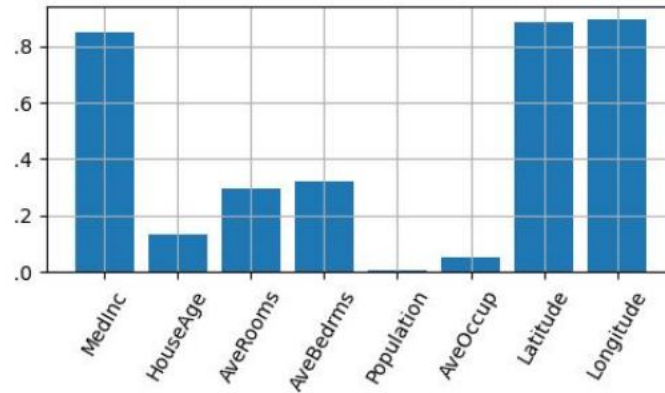
- **Explicabilidad:** definen una **relevancia** Rx_i para cada atributo de entrada



Fundamentos de la XAI (VI)

- **Explicabilidad:** definen una **relevancia** Rx_i para cada atributo de entrada

Ejemplo de análisis de pesos sobre California Housing



Fundamentos de la XAI (VII)

- **Interpretabilidad:** Capacidad de comprender el funcionamiento interno de las componentes de un modelo y su estructura
- Responde a la pregunta:

¿Qué pasa en el modelo para que dé esta salida?

- Los algoritmos de interpretabilidad encuentran una **representación** comprensible del modelo

Fundamentos de la XAI (VIII)

- **Interpretabilidad:** encuentran una **representación** comprensible

Ejemplo de un Explanatory Graph sobre una CNN a nivel de capa [7]



Fundamentos de la XAI (IX)

- **Interpretabilidad:** encuentran una **representación** comprensible

Ejemplo de interpretabilidad a nivel neuronal sobre una RNN [8]

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

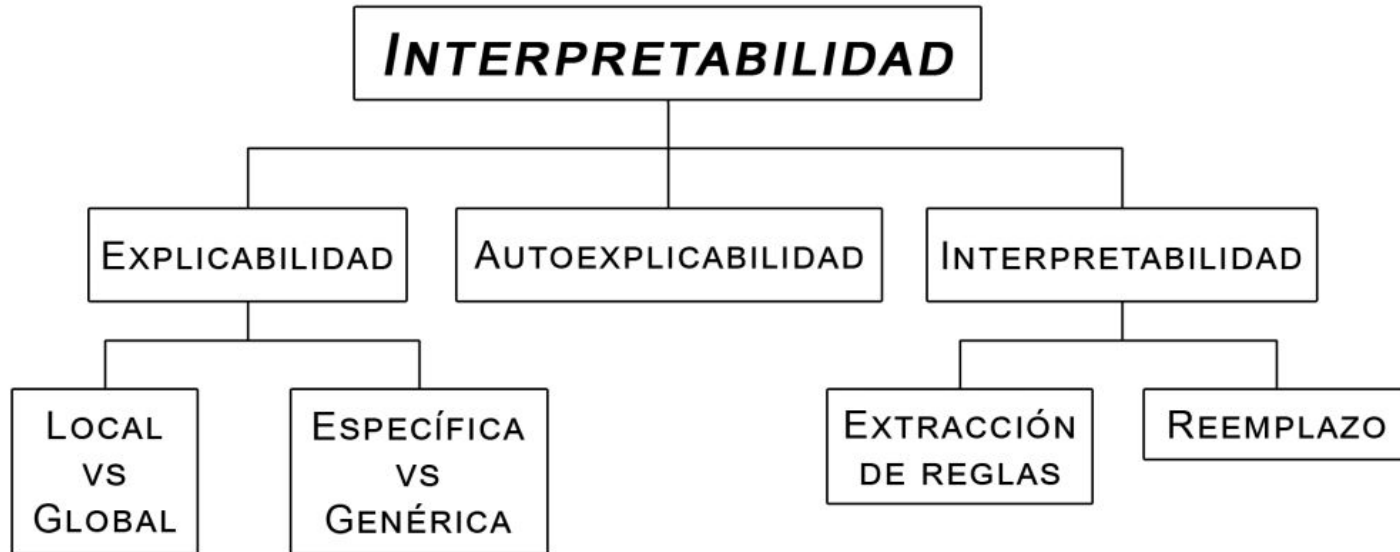
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Cell that robustly activates inside if statements:

```
static int _dequeue_signal(struct sigpending *pending, sigset_t *mask,
                          siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!(current->notifier)(current->notifier_data)) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```

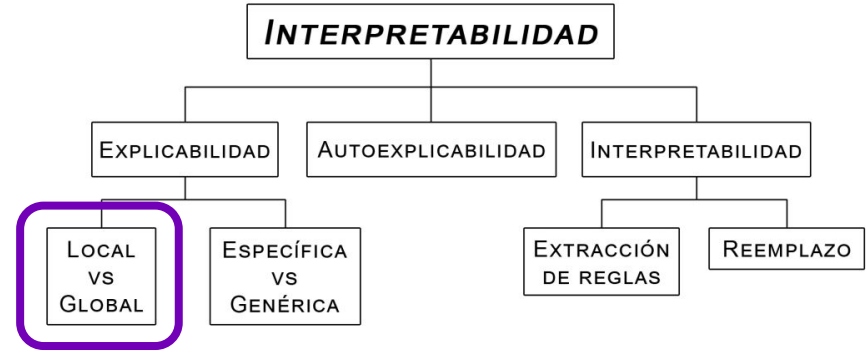
Fundamentos de la XAI (X)

- Jerarquía de algoritmos de XAI



Fundamentos de la XAI (XI)

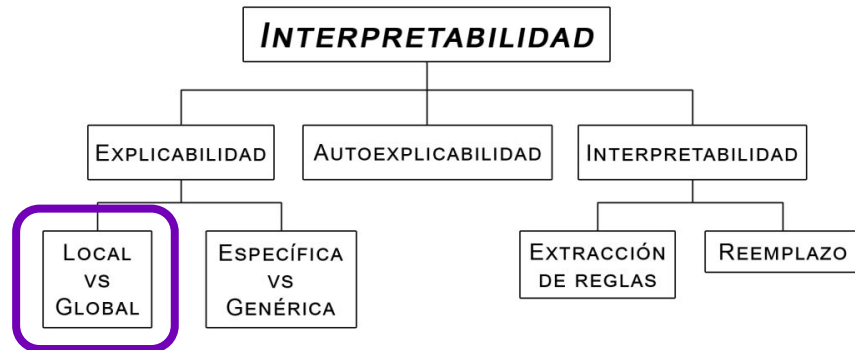
- **Explicabilidad local vs global**



- **Global**: no tiene en cuenta situaciones particulares, sino solamente parámetros del modelo (por ejemplo, los pesos de una red neuronal)
- **Local**: considera las características individuales de cada dato de entrada, es decir, da una explicación para cada dato de entrada

Fundamentos de la XAI (XII)

- Explicabilidad local vs global

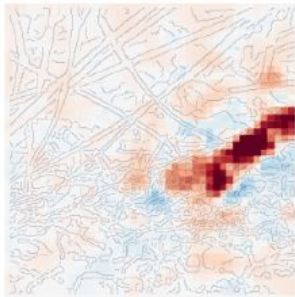


LOCAL

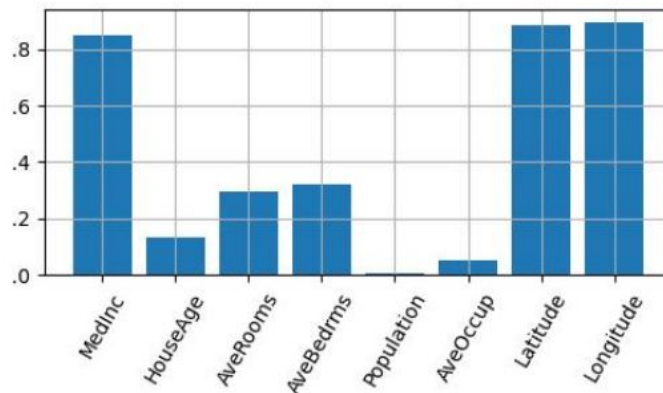
Original (label: "garter snake")



Occlusion-15x15

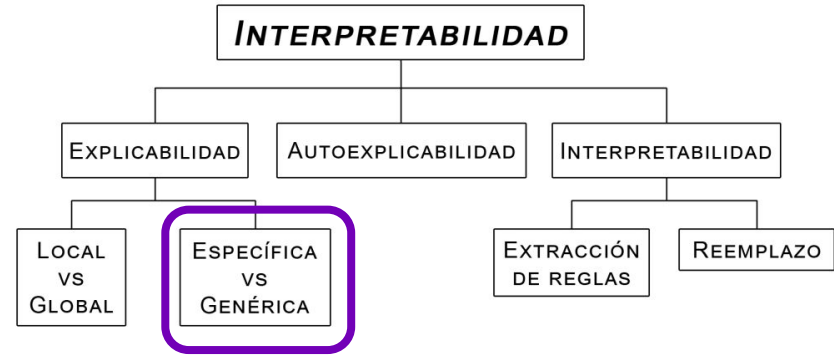


GLOBAL



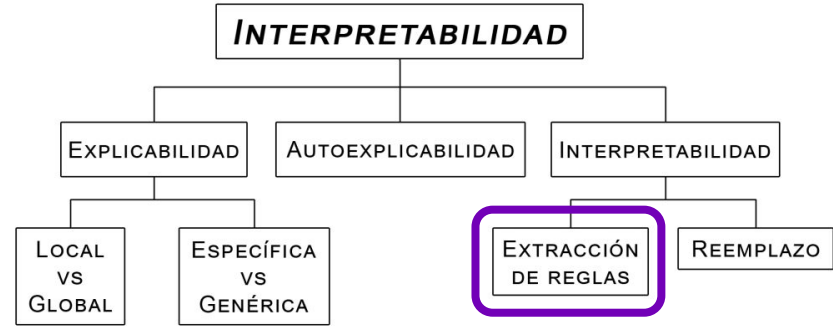
Fundamentos de la XAI (XIII)

- **Explicabilidad específica vs genérica**



- **Genérica**: el algoritmo se puede aplicar a cualquier modelo
- **Específica**: el algoritmo es específico para un modelo particular (por ejemplo, específico para redes neuronales)

Fundamentos de la XAI (XIV)



- **Extracción de reglas**
- Intentar extraer un conjunto de reglas o identificar patrones que representan al modelo complejo, de forma global o parcial

Por ejemplo, una neurona identifica texto entrecomillado

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... on the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

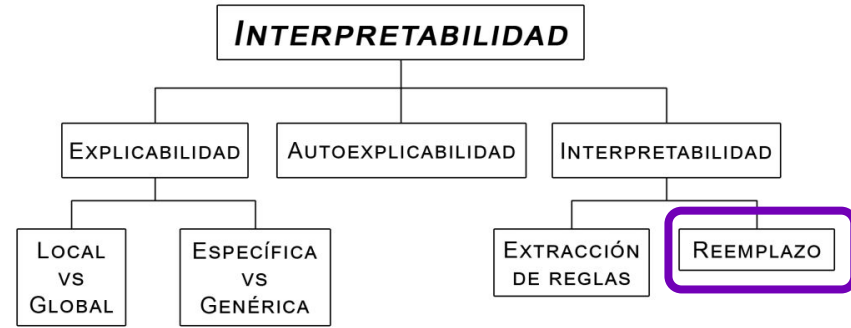
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Fundamentos de la XAI (XV)

- **Reemplazo**

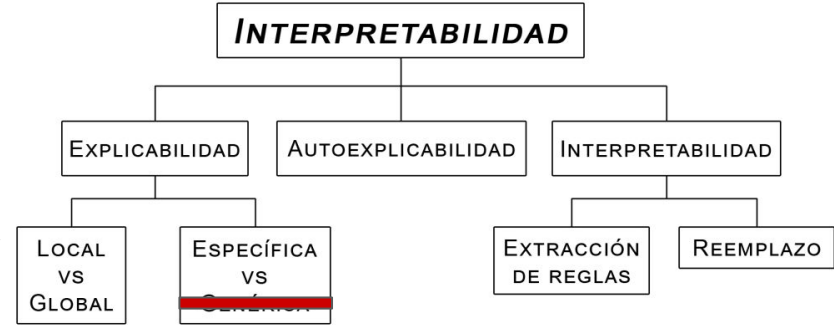
- Extraer un modelo más sencillo a partir del modelo complejo

Por ejemplo, generar un árbol de decisión a partir de las predicciones de una red neuronal



Dificultades de la XAI en Deep Learning (I)

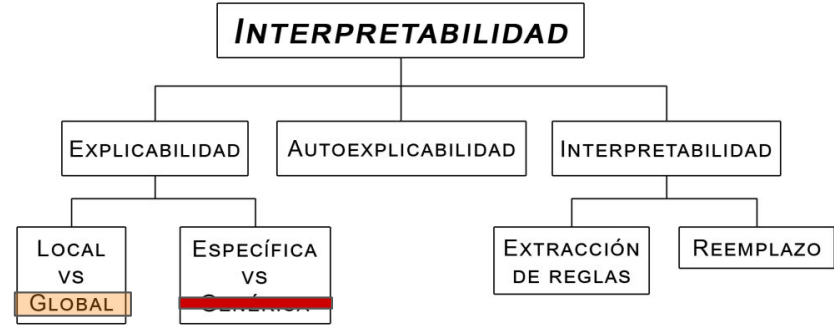
- En Deep Learning, los algoritmos de XAI tienen ciertas dificultades adicionales
 - Los algoritmos genéricos suelen ser altamente costosos (reemplazo, LIME, SHAP, etc.) y no dan resultados razonables cuando se aplican a imágenes, series temporales, etc.



Dificultades de la XAI en Deep Learning (II)

- En Deep Learning, los algoritmos de XAI tienen ciertas dificultades adicionales

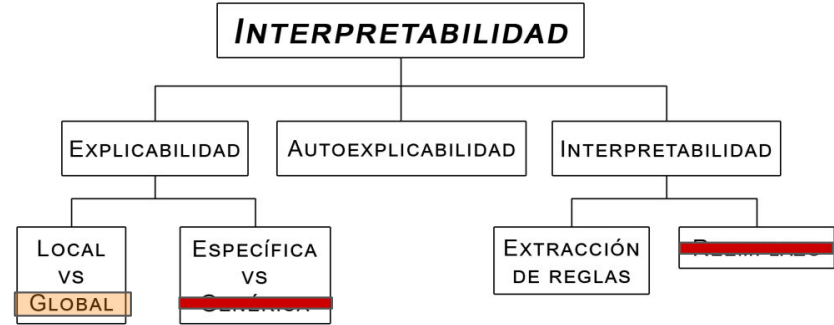
- Los algoritmos globales no suelen dar buenas explicaciones con imágenes y series temporales, por lo que se utilizan solamente para datos tabulares.



Dificultades de la XAI en Deep Learning (III)

- En Deep Learning, los algoritmos de XAI tienen ciertas dificultades adicionales
 - Los algoritmos de interpretabilidad por reemplazo son inviables por la complejidad del modelo “sencillo”.

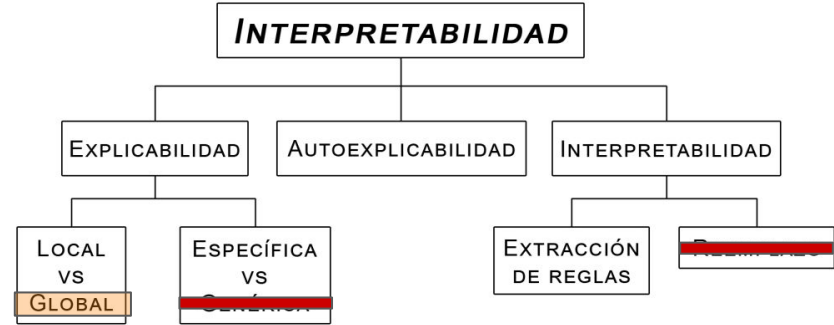
Por ejemplo, de qué vale un árbol de decisión con profundidad de 100, eso deja de ser interpretable por mucho que sea un árbol.



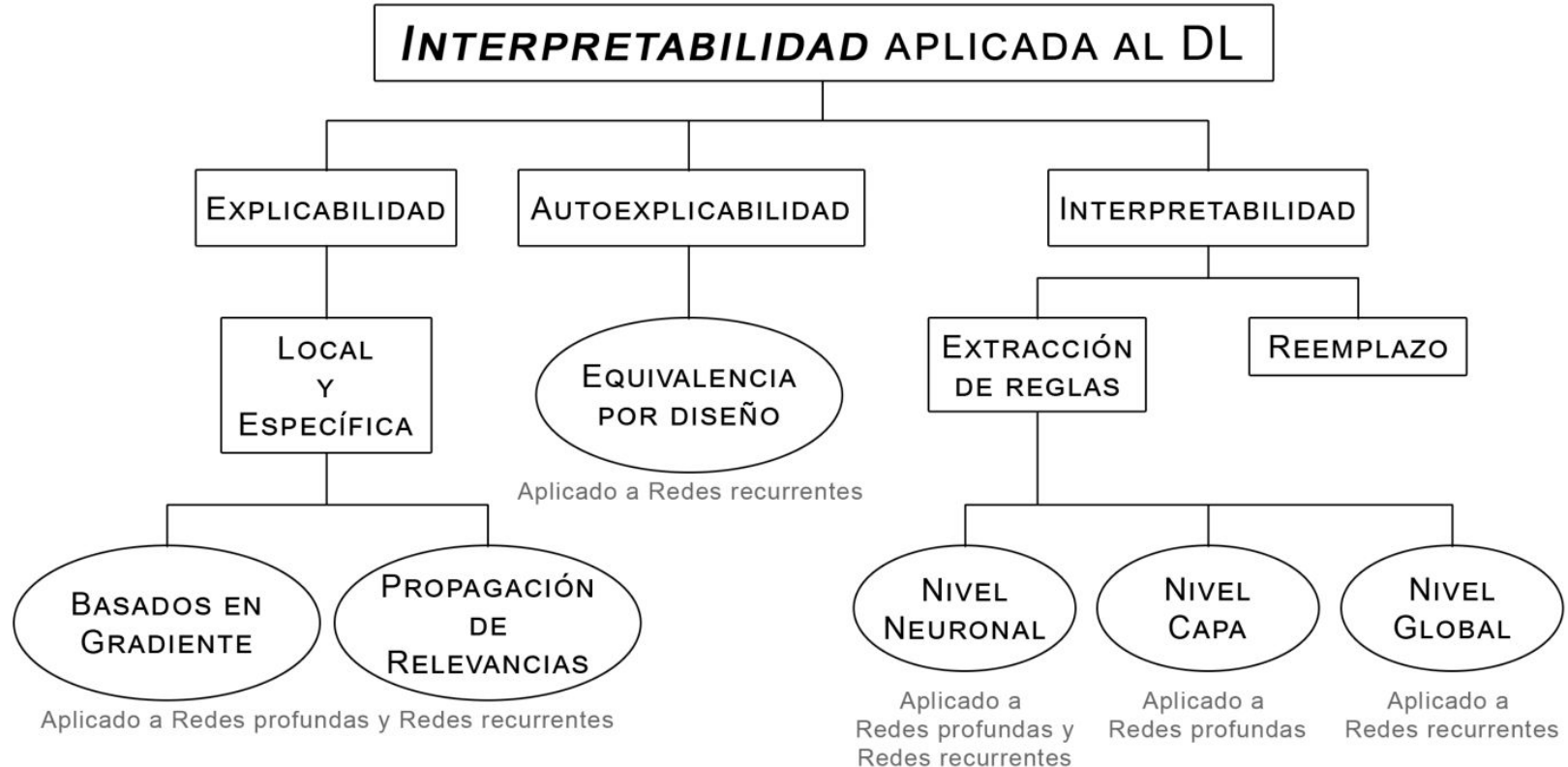
Dificultades de la XAI en Deep Learning (IV)

- En Deep Learning, los algoritmos de XAI tienen ciertas dificultades adicionales
 - En general, no es posible definir una interpretabilidad global de un modelo de miles, millones o incluso trillones de parámetros...

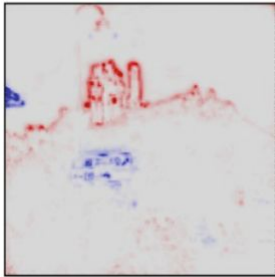
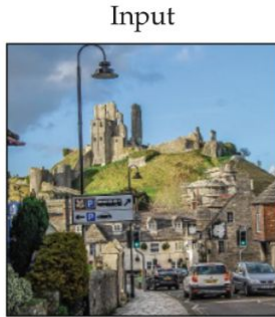
PERO...



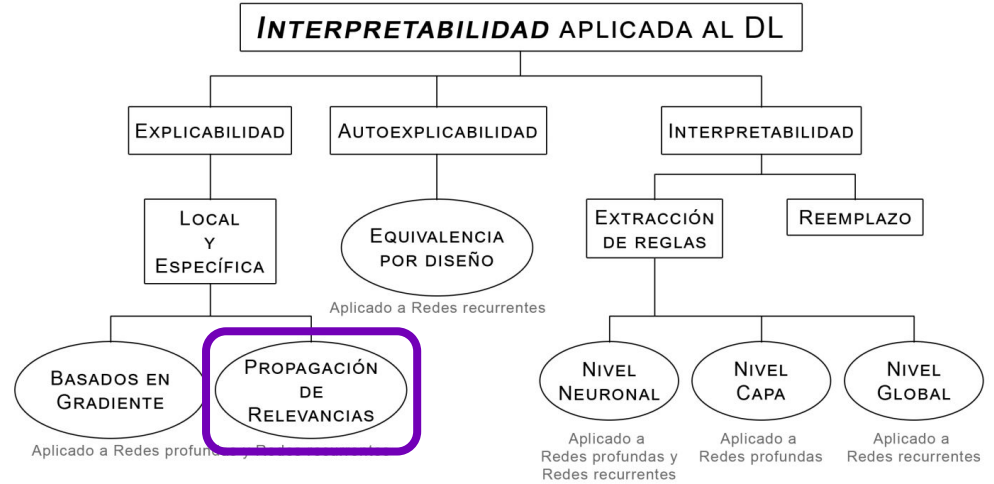
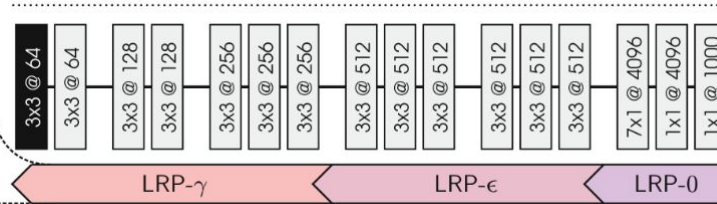
XAI en Deep Learning (I)



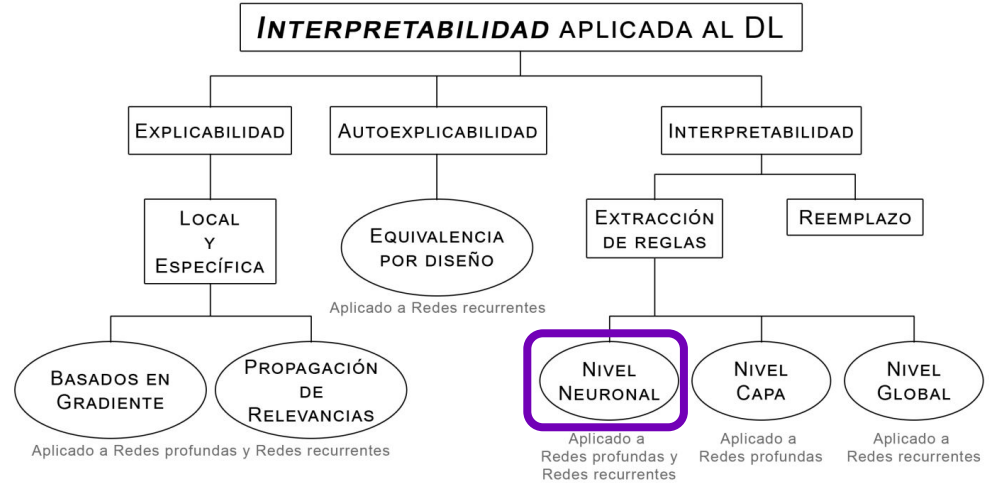
XAI en Deep Learning (II)



Composite LRP



XAI en Deep Learning (III)

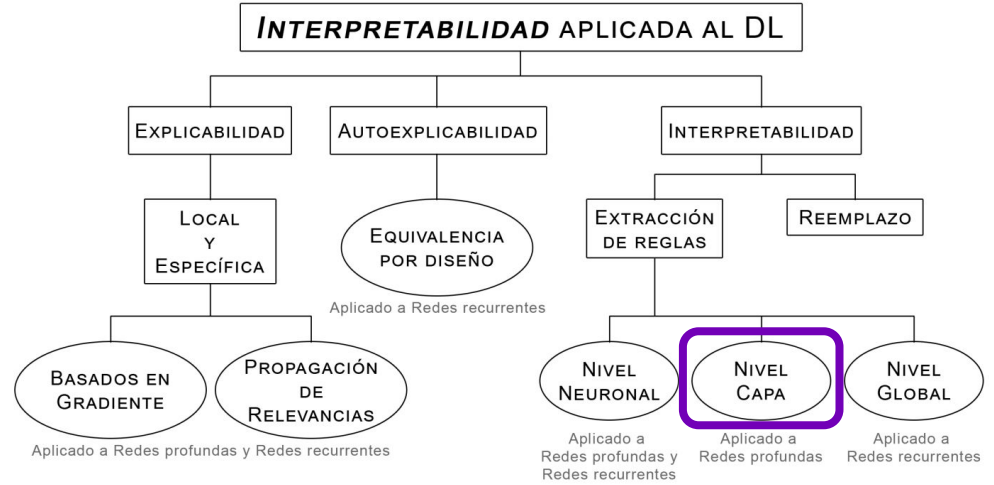


Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... on the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

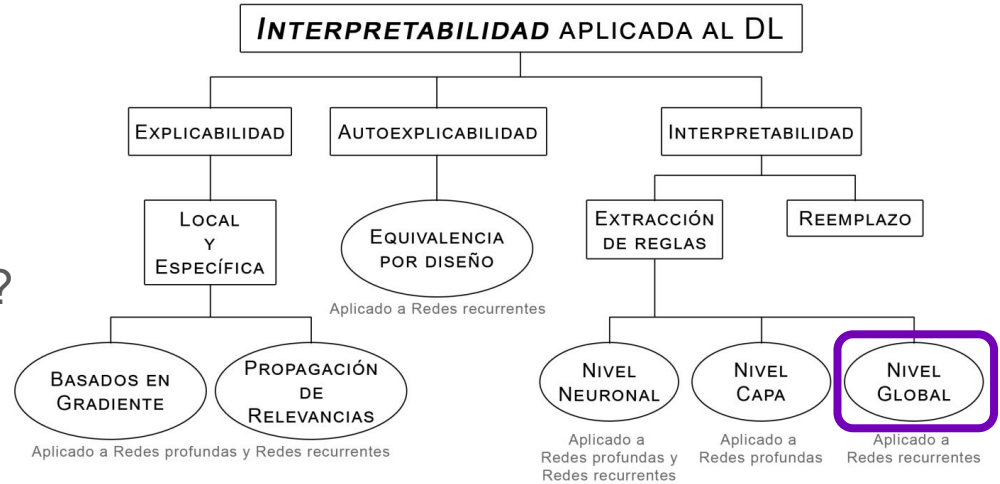
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

XAI en Deep Learning (IV)



XAI en Deep Learning (V)

- ¿Es posible la interpretabilidad a nivel global del Deep Learning?



- ¿Qué os parece si preguntamos a ChatGPT?

Consulta de la siguiente diapositiva realizada a día 26 de agosto de 2025

XAI en Deep Learning (VI)

¿Es posible la interpretabilidad a nivel global del Deep Learning?

Gran pregunta. Te respondo con matices:

La **interpretabilidad global** en *Deep Learning* es un reto muy difícil, y en muchos casos se considera **inalcanzable** en el sentido estricto en que entendemos interpretabilidad en modelos lineales o árboles de decisión. Te explico por qué:

♦ Razones por las que es difícil la interpretabilidad global

1. Alta dimensionalidad:

Redes profundas manejan miles o millones de parámetros. Explicar de forma compacta *cómo* cada parámetro contribuye al resultado global es prácticamente imposible.

2. No linealidad extrema:

Una red profunda es una composición de muchas funciones no lineales. Esto genera comportamientos emergentes difíciles de resumir en reglas simples.

3. Dependencia de los datos:

El conocimiento que adquiere una red no se representa como conceptos humanos, sino como patrones distribuidos en el espacio de características. Eso dificulta un “mapa global” entendible.

♦ Lo que sí es posible

Aunque una interpretabilidad **global y completa** es inalcanzable, existen aproximaciones que permiten obtener una **interpretabilidad parcial y aproximada**:

1. Métodos de **interpretabilidad local agregada**

- SHAP, LIME y otros explican predicciones individuales.
- Si se aplican masivamente y se agregan resultados, podemos inferir *tendencias globales* (ej. qué variables tienden a ser más importantes). ↓

2. Reducción de dimensionalidad y visualización

- *t-SNE*, *UMAP* permiten visualizar cómo la red organiza el espacio latente.
- Esto da intuiciones globales sobre cómo el modelo agrupa y separa clases.

3. Análisis de **importancia de características**

- Algunas variantes de SHAP o *integrated gradients* permiten estimar qué variables tienen más peso globalmente.

4. Model distillation (destilación de modelos)

- **Aproximar la red neuronal con un modelo más simple** (árboles, reglas) para capturar su comportamiento general.
- Esto da una visión global, aunque *no exacta*.

5. Explicaciones basadas en conceptos

- Técnicas como *TCAV (Testing with Concept Activation Vectors)* permiten vincular conceptos humanos (ej. “estrias en alas de mariposa”) con representaciones internas de la red.

♦ En conclusión

- Una **interpretabilidad global exacta** del Deep Learning es prácticamente imposible por la complejidad y no linealidad del modelo.
- Lo que sí se logra hoy es una **interpretabilidad global aproximada**, basada en agregar explicaciones locales, visualizar espacios latentes o usar modelos sustitutos.
- El campo de **XAI (Explainable AI)** busca precisamente avanzar en este equilibrio: no se puede traducir toda la red a una narrativa humana, pero sí hacerla *más comprensible* en sus tendencias globales.

XAI en Deep Learning (VII)

- ¿Y si os digo que es mentira? **Sí se puede hacer** una interpretabilidad global del Deep Learning **cuando es necesario**
- Hay que responder a la pregunta:

¿Cuándo es necesaria la interpretabilidad global de un modelo de DL?

- Si es un modelo **forward** (MLP, CNNs, transformers, etc.)
- Si es un modelo **recurrente** (Elman, LSTMs, etc.)

XAI en Deep Learning (VIII)

- Si es un modelo forward (MLP, CNNs, transformers, etc.) **no es estrictamente necesaria:**
 - Necesitamos comprender qué atributos son importantes
 - Necesitamos saber cómo tienen que cambiar los atributos para cambiar la decisión
- ¿Por qué? Porque el modelo es, precisamente, hacia adelante: **$f(\mathbf{x}) = y$**

Si yo sé que si sube el salario de un cliente es más probable dar una hipoteca, no necesito saber las transformaciones internas del modelo.

XAI en Deep Learning (IX)

- Si es un modelo recurrente (Elman, LSTMs, etc.) **sí es necesaria**:
 - Necesitamos comprender qué atributos son importantes
 - También necesitamos saber **cómo son las dependencias temporales**
- ¿Por qué? $f(\mathbf{x}_t, \mathbf{h}_{t-1}) = \mathbf{y}_t, \mathbf{h}_t$

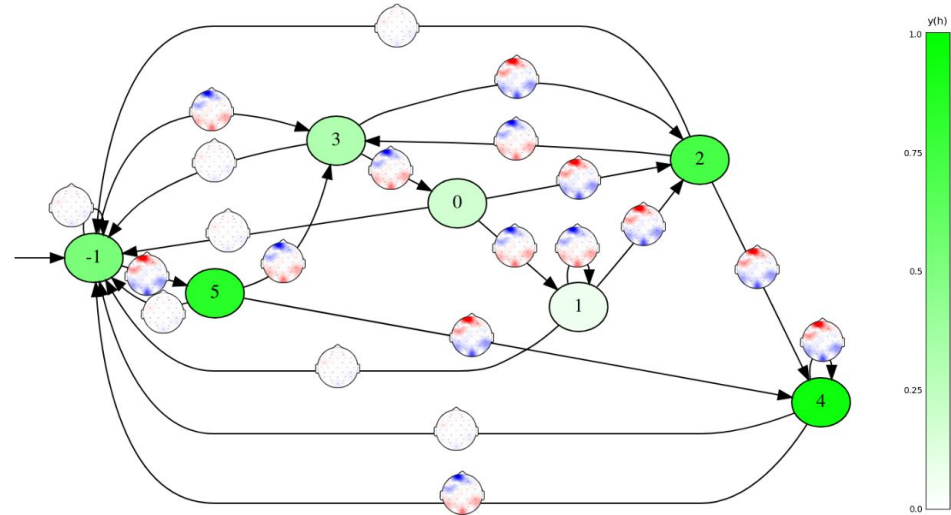
Necesito saber qué pasa en la red en el estado \mathbf{h}_{t-1} con \mathbf{x}_t para dar la respuesta \mathbf{y}_t y guardar el estado actual \mathbf{h}_t

XAI en Deep Learning (X)

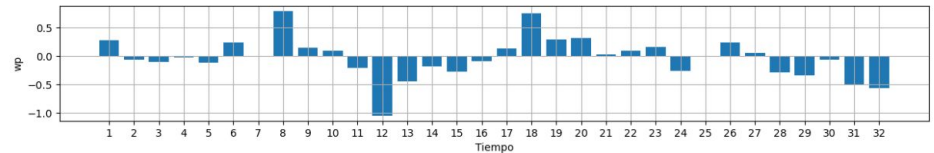
- Durante 5 años, en mi tesis doctoral [9] hemos investigado sobre la interpretabilidad global de las RNNs
 - Función de activación con ruido para binarizar la activación
 - Estabilidad del estado interno de las RNNs
 - **Función de activación con redondeo**
 - **Capa discretizadora**
 - **Capa PRM**
 - **Autómata finito con variable acumulativa**

XAI en Deep Learning (XI)

- **Problema:** Detección de eventos P300 en señales EEGs
- **Modelo:** RNN de Elman con 20 neuronas en la capa recurrente, usando la metodología de interpretabilidad que os voy a contar en este curso



(a) Autómata finito con variable acumulativa



(b) Ponderaciones temporales w^p del autómata

Últimos detalles: Explicabilidad vs Selección de atributos

