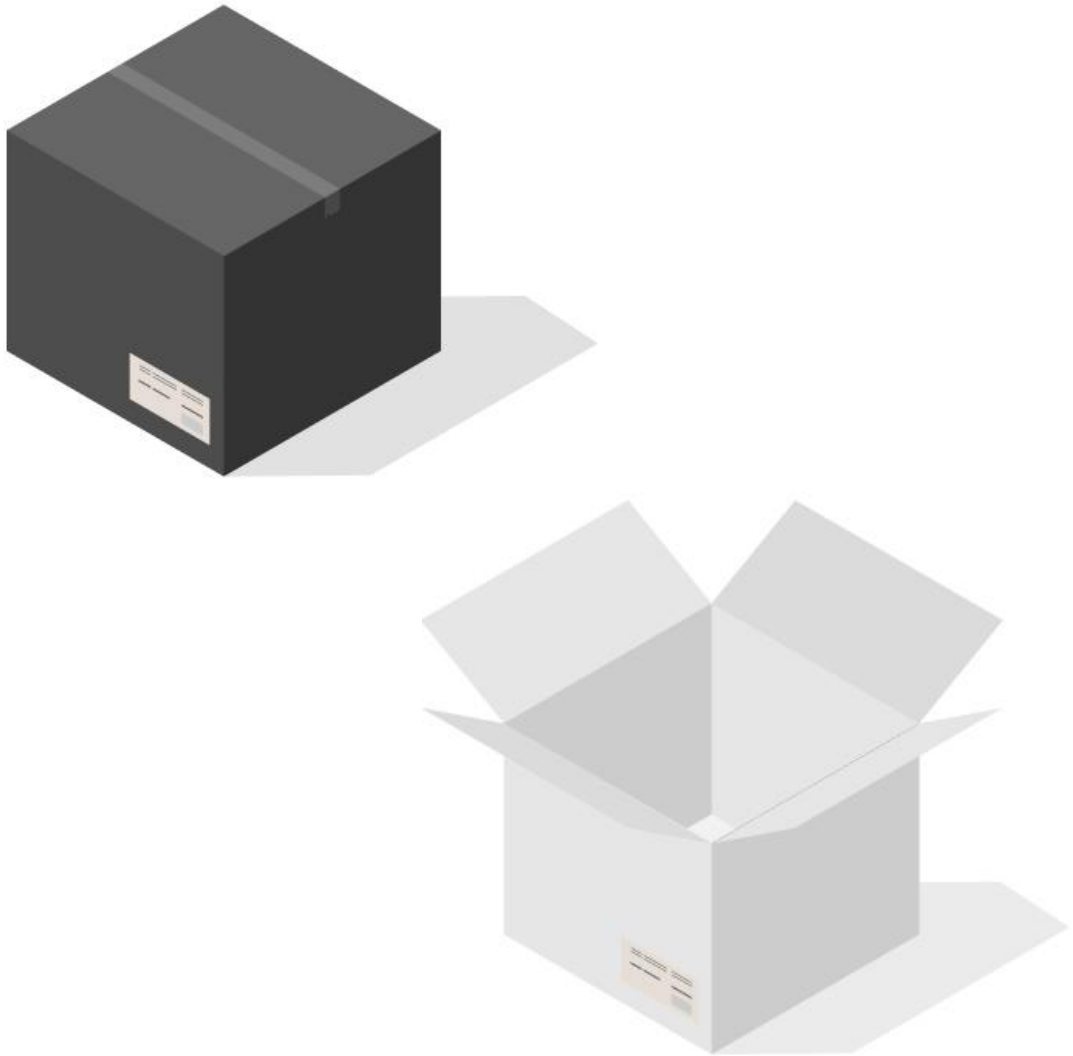
Centro de Investigación en Computación
Instituto Politécnico Nacional

Profundizando en el conocimiento: métodos para explicar los modelos de lenguaje – Parte I

EVCCPLN Segunda Edición

Julio 2024

Acerca del curso



Tallerista: Sergio Damián

Material del curso:

https://github.com/sdamians/taller-explicabilidad/

Se requiere que los participantes cuenten con conocimientos básicos de:

Lenguaje Python

PyTorch

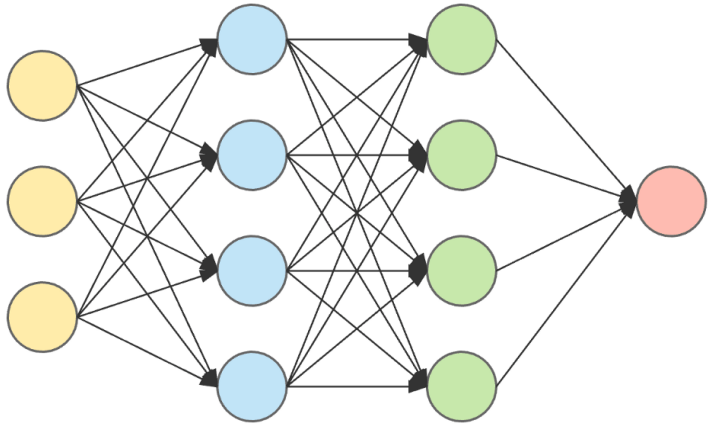
Modelos de Lenguaje Pequeños (Small Language Models)

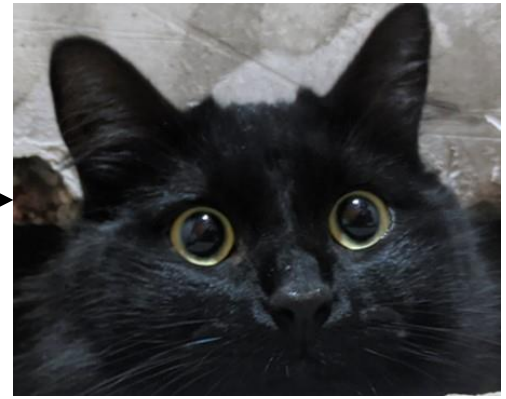
Agenda

1. **Inteligencia Artificial Explicable**
   - Definición
   - Explicabilidad vs Interpretabilidad
   - Por qué XAI
   - Categorías de métodos XAI
   - Explicabilidad vs Desempeño
2. **Modelos de lenguaje**
   - Definición
3. **Introducción a SHAP**
   - Definición
   - Ejemplo
   - Fórmula matemática
   - Demo
   - Limitaciones SHAP
4. **Resumen**
5. **Sesión de Preguntas**

1. Inteligencia Artificial Explicable (XAI)

Me encanta acurrucarme en tu regazo
Conozco tu casa como un mapa al dedazo
Puedo atrapar uno o dos ratones
Porque eso es lo que hago en ocasiones





Explicación

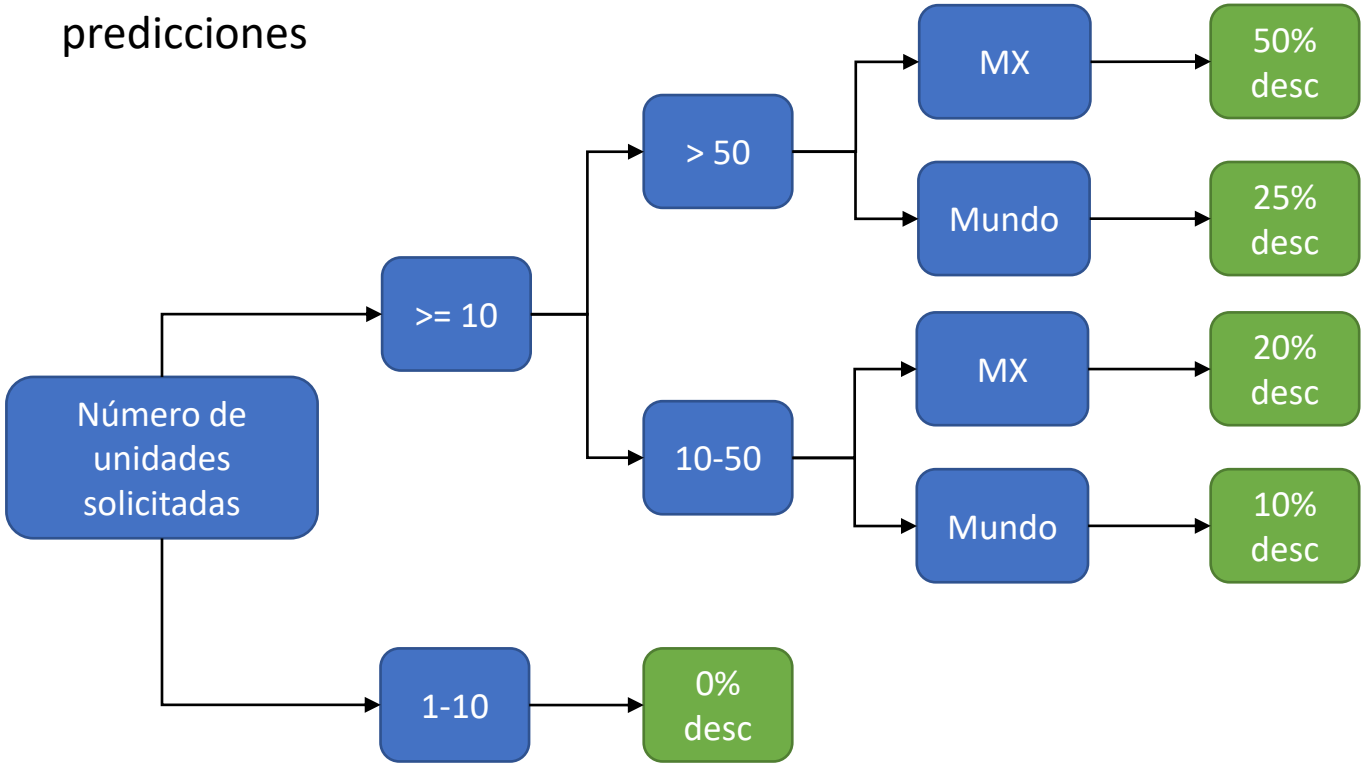
Es un gato porque:

- Acurrucarse en las piernas de las personas es una actividad común de los gatos
- Los gatos son animales territoriales
- El gato es el depredador más común de los ratones

Explicabilidad vs Interpretabilidad

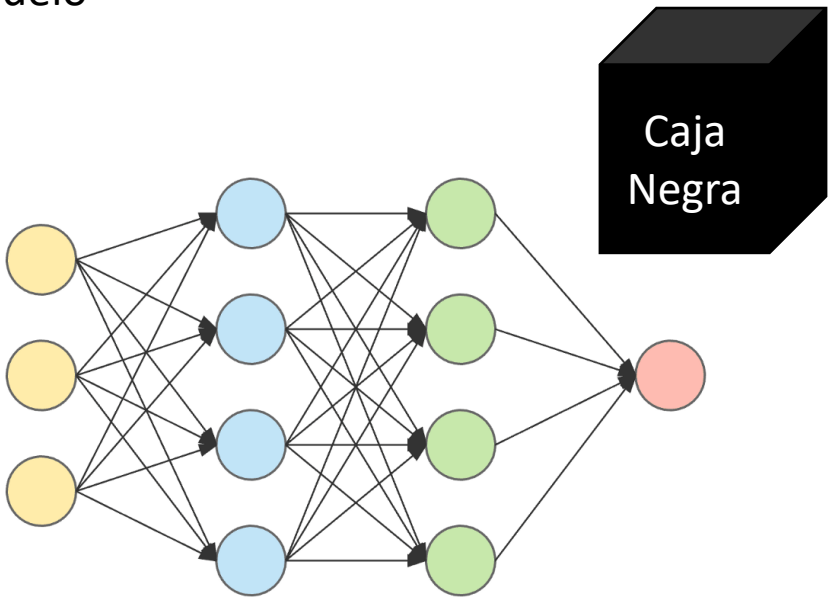
Explicabilidad

Es la capacidad de un modelo para proporcionar razones claras sobre sus predicciones



Interpretabilidad

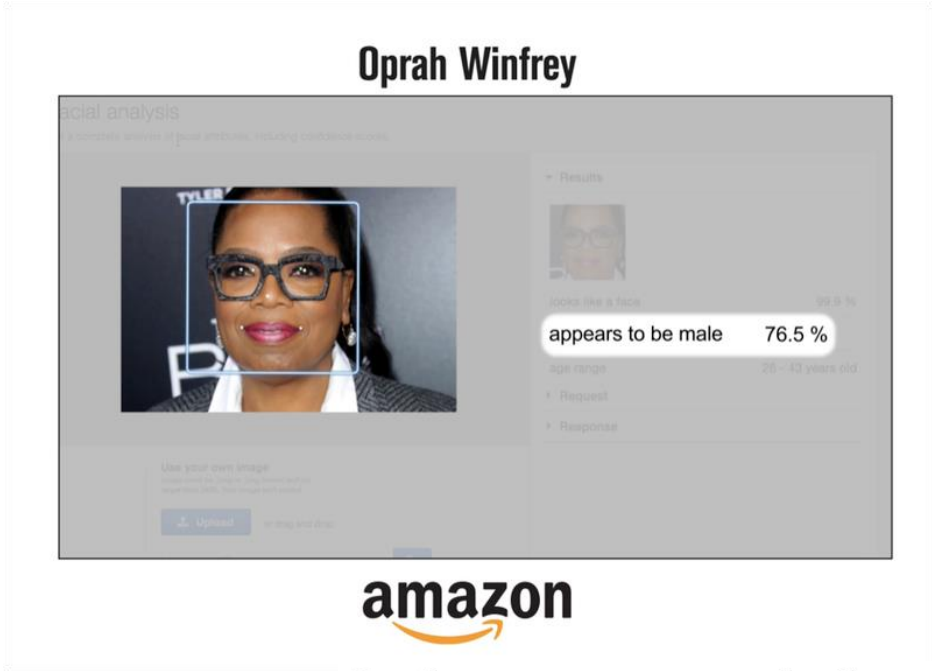
Es a qué grado un humano puede entender la causa de la decisión de un modelo



Por qué XAI

Los sistemas toman decisiones…
Los humanos nos vemos **afectados** por esas decisiones

- Seguridad
- Debugging
- Confianza y transparencia
- Regulaciones legales

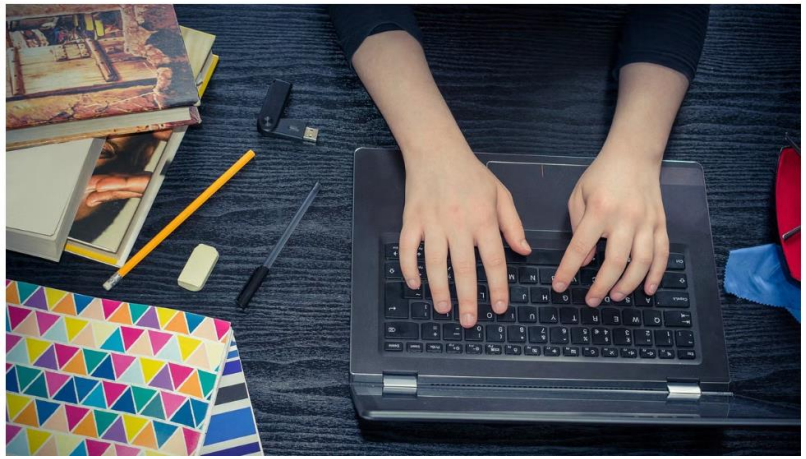


amazon

Students uncover method to manipulate AI grading algorithm

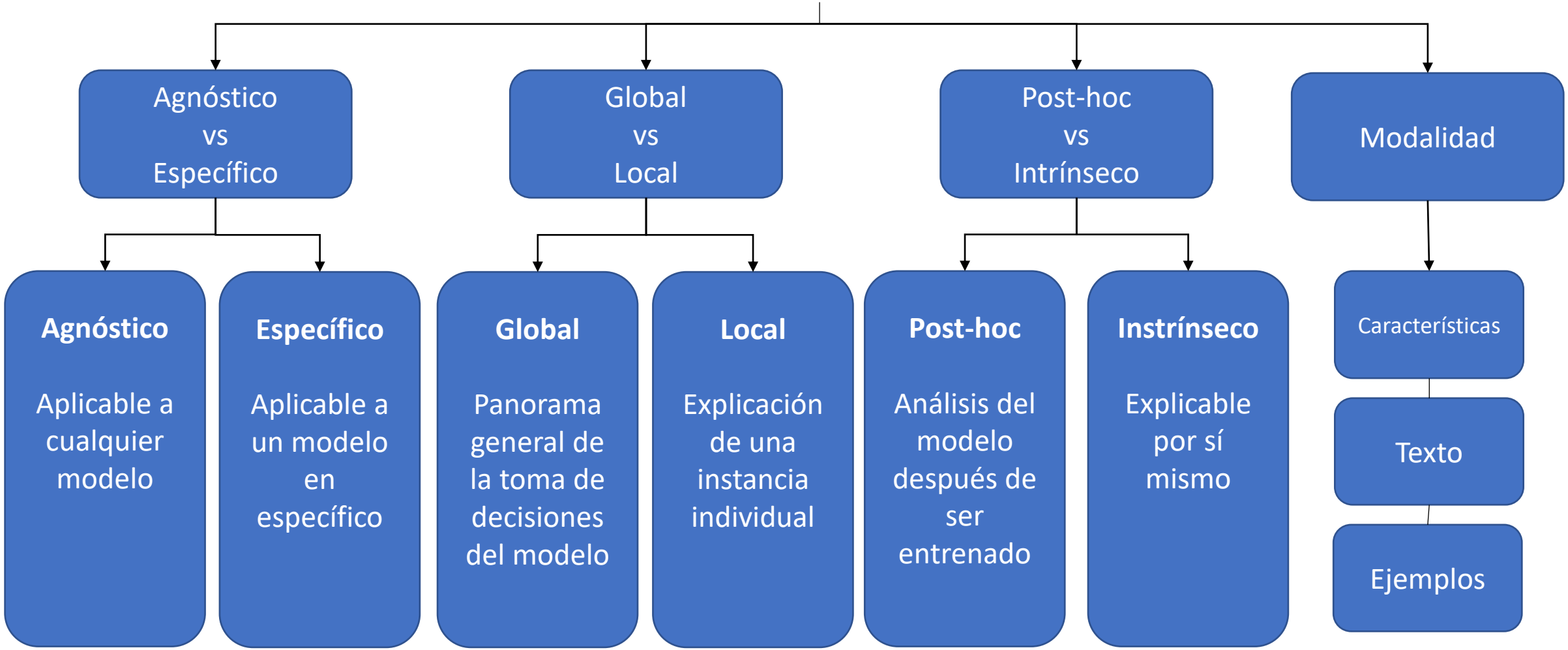
Spamming relevant keywords, many of which could be found online, saw students' grades rise to potentially perfect scores

Adam Smith • Thursday 03 September 2020 18:17 BST • Comments



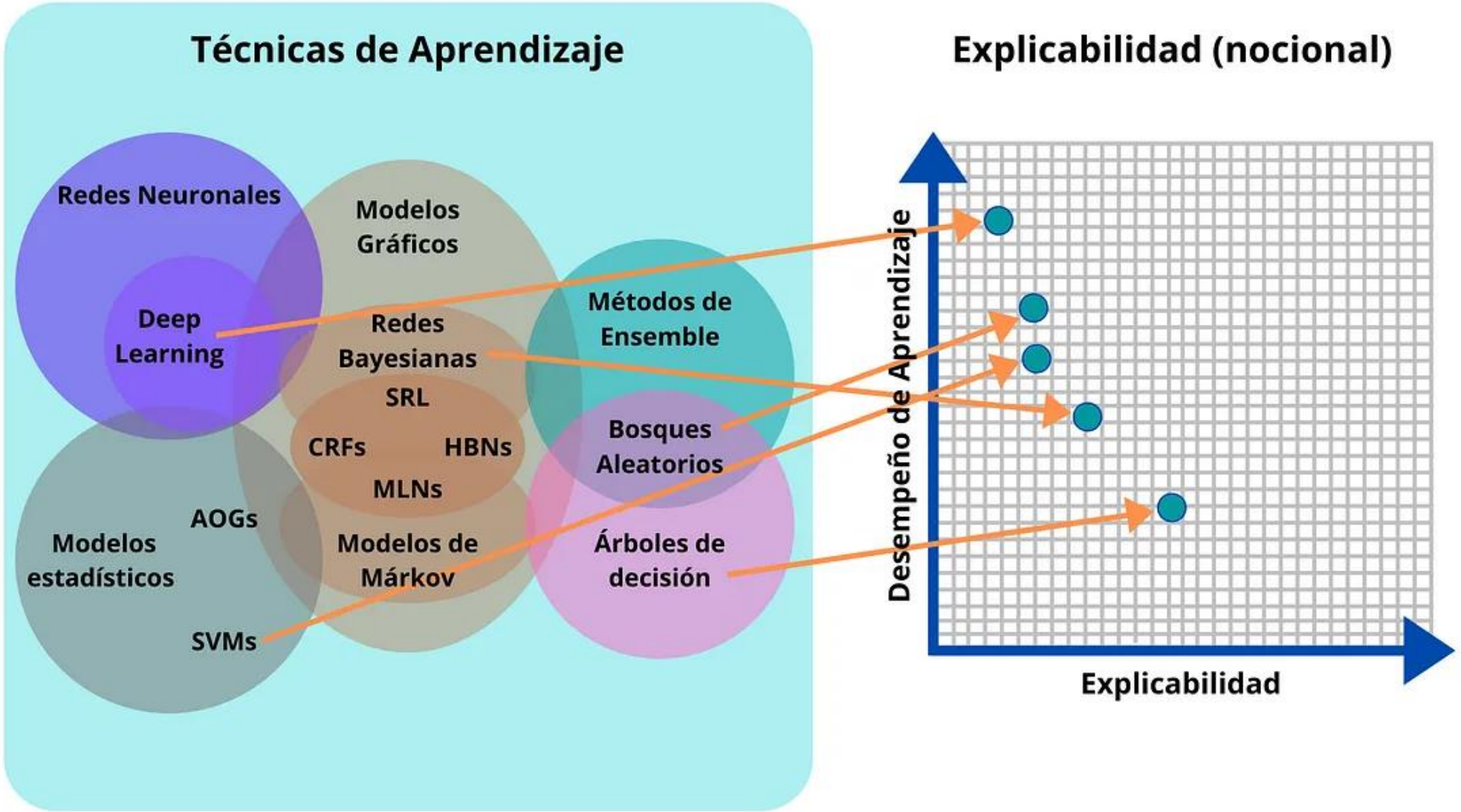
https://medium.com/ai-learners/principios-en-inteligencia-artificial-parte-2-33fd2c01e446
https://www.independent.co.uk/tech/students-ai-grading-algorithm-edgenuity-keyword-spam-a9703751.html

Métodos XAI



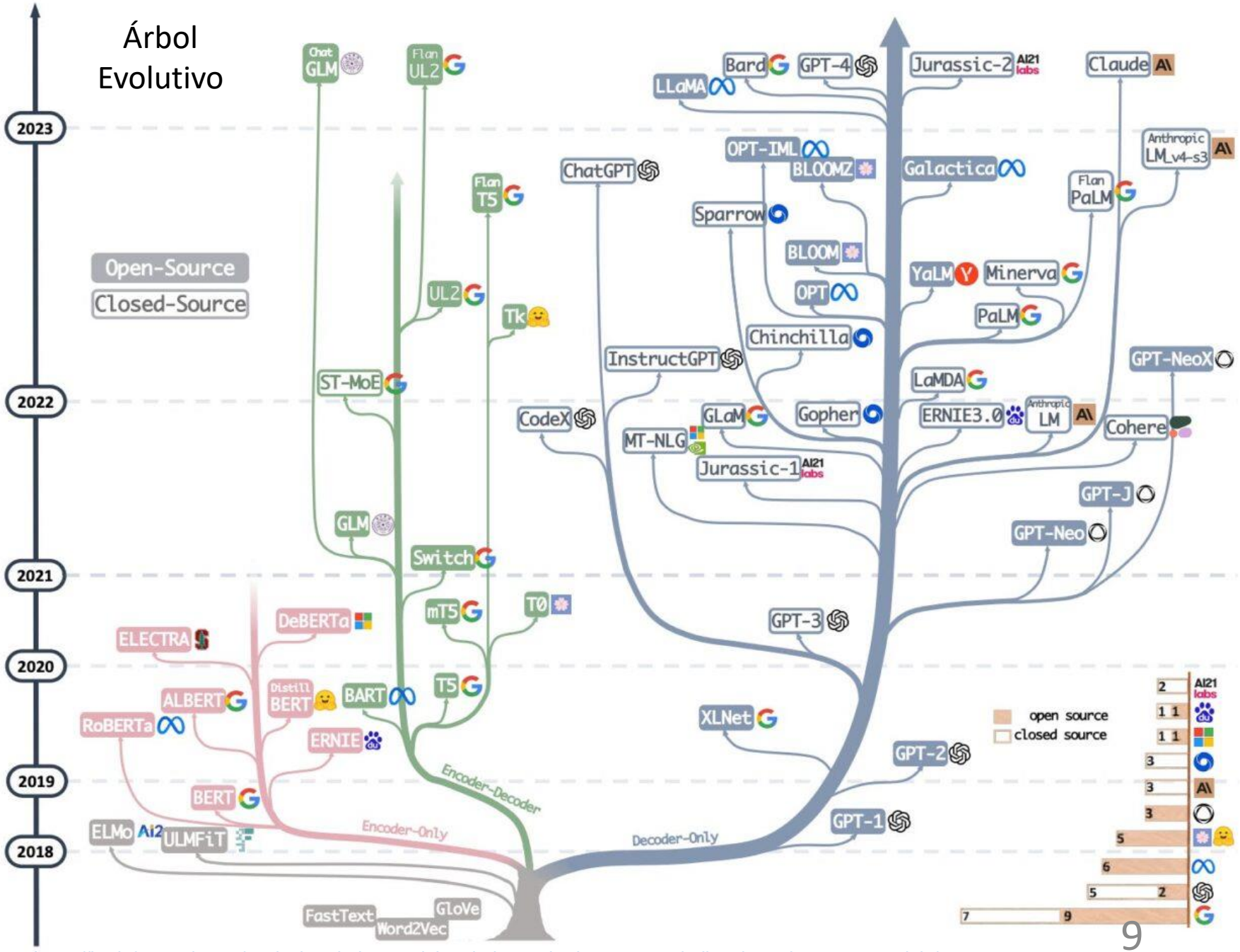
ArXiv: 2205.04766

Explicabilidad vs Desempeño

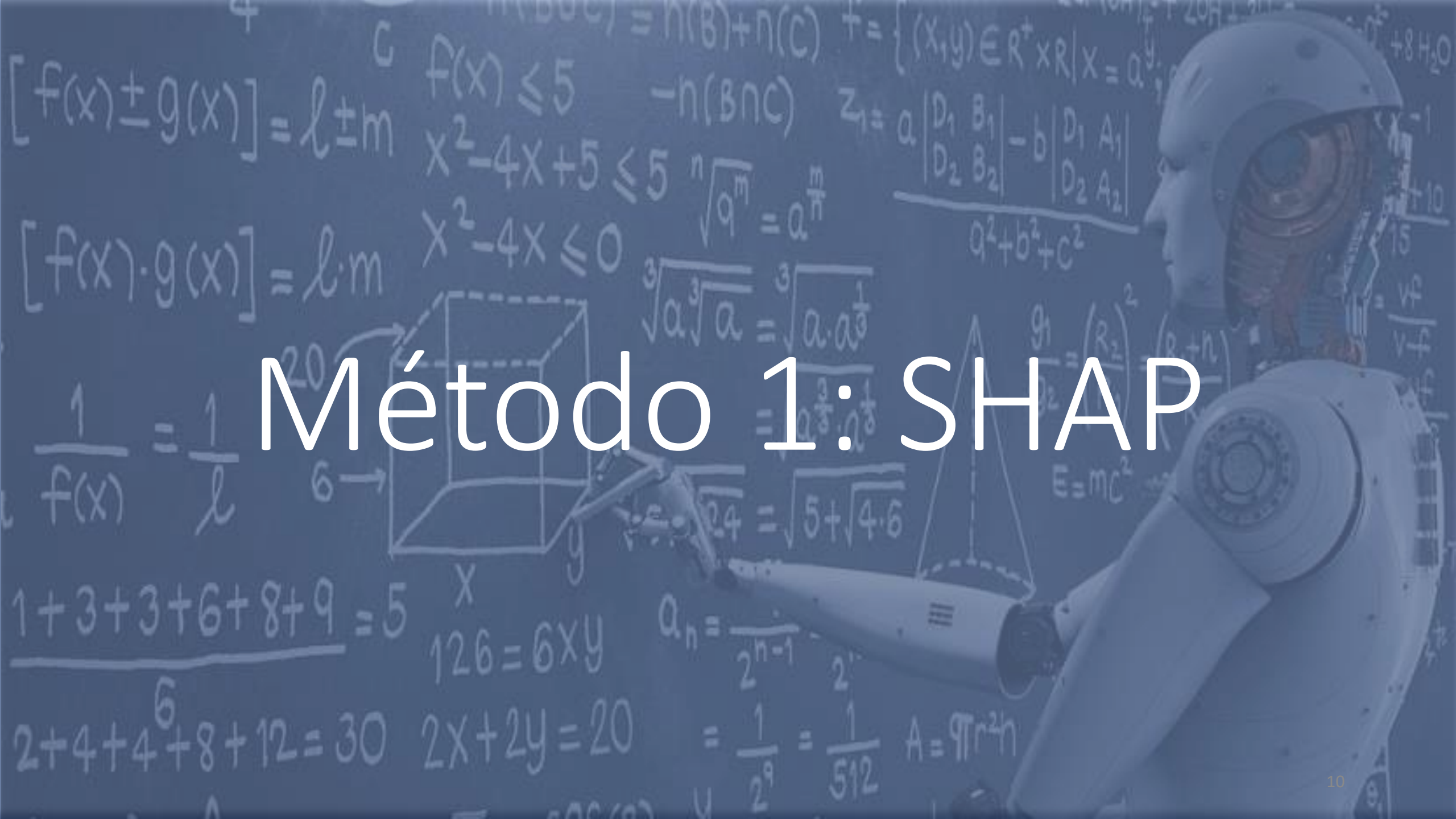


https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

2. Modelos de Lenguaje



https://borjafernandez.es/evolucion-de-los-modelos-de-lenguaje-de-gran-escala-llms-large-language-models/



Método 1: SHAP

3. Introducción a SHAP
(SHapley Additive exPlanations)

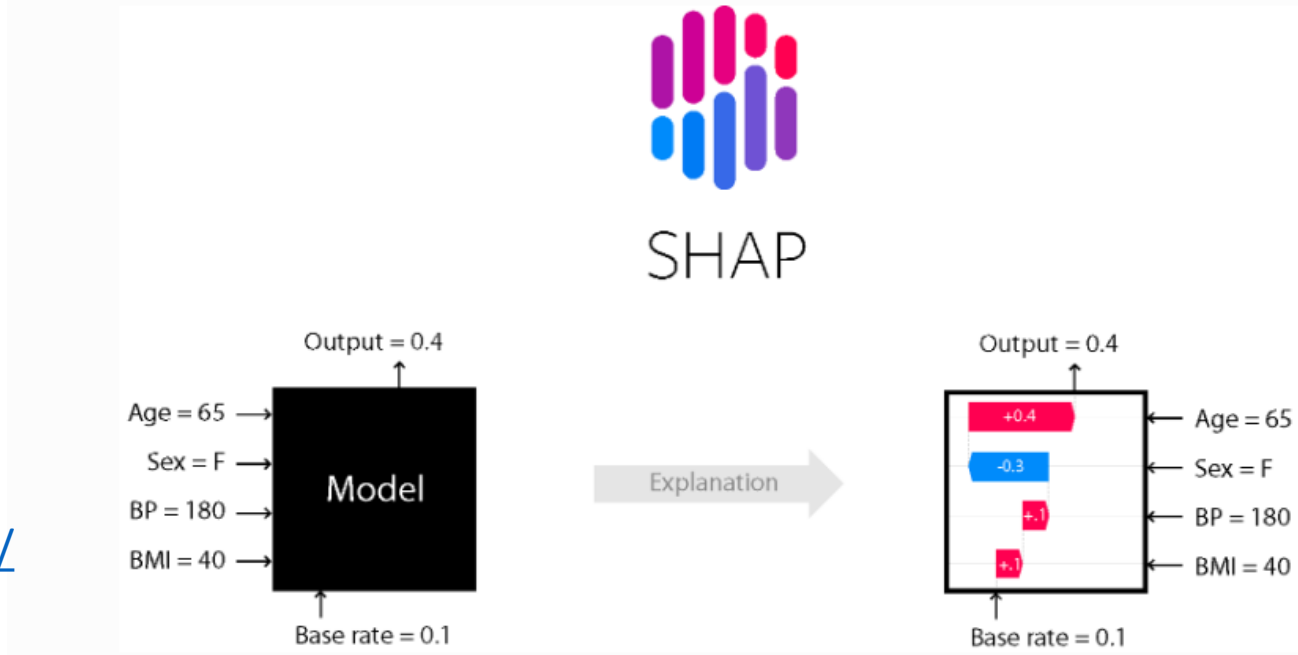
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mas de 20 mil citas

Código abierto

https://shap.readthedocs.io/en/latest/

ArXiv: 2205.04766



Introducción a SHAP

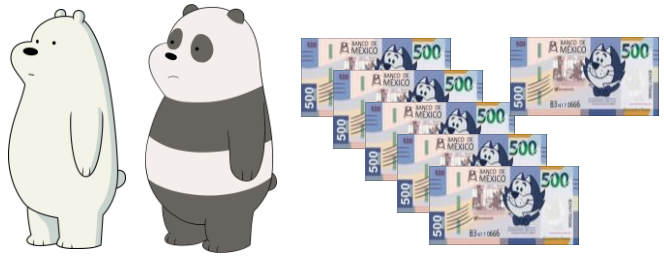
En 1951 **Lloyd Shapley** presenta un modelo para la teoría de juegos

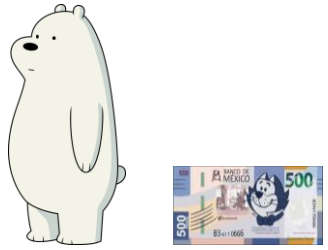
Se basa en la interacción de múltiples jugadores en un escenario con reglas específicas y consecuencias cuantificables

Los valores Shapley permiten cuantificar la justa colaboración de múltiples participantes en un proceso

Ejemplo

a b c

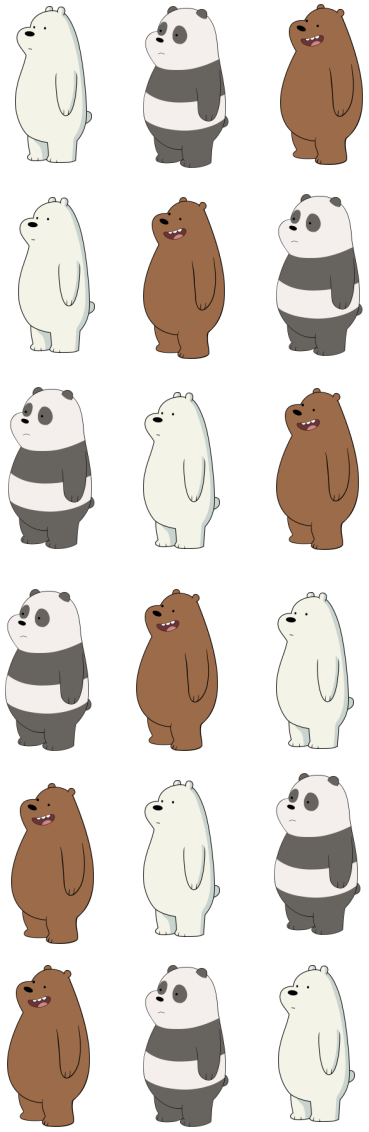




$$C_{abc} - C_{bc} = \$5{,}000 - \$2{,}520$$

$$= \$2{,}480$$

Ejemplo



$$C_{abc} - C_{bc} = \$5{,}000 - \$2{,}520 = \$2{,}480$$
$$C_{ab} - C_{b} = \$3{,}000 - \$1{,}000 = \$2{,}000$$
$$C_{ac} - C_{c} = \$2{,}500 - \$1{,}500 = \$1{,}000$$
$$C_{a} - C_{\emptyset} = \$500 - \$0 = \$500$$

$P(C_{abc} - C_{bc})$ = Probabilidad de que el participante *a* haga su contribución marginal a la coalición de los jugadores *b* y *c*

$$(1/3)\$2{,}480 + (1/6)\$2{,}000 + (1/6)\$1{,}000 + (1/3)\$500 = \$1493.33$$

$$(1/3)\$2{,}500 + (1/6)\$2{,}500 + (1/6)\$1{,}020 + (1/3)\$1{,}000 = \$1753.33$$

$$(1/3)\$2{,}000 + (1/6)\$2{,}000 + (1/6)\$1{,}520 + (1/3)\$1500 = \$1753.33$$

Fórmula SHAP

$$(1/3)\$2{,}500 + (1/6)\$2{,}500 + (1/6)\$1{,}020 + (1/3)\$1{,}000 = \$1753.33$$

Sumatoria de todos los subconjuntos de características para la instancia x.

Impacto de remover la característica i

$$\varphi_i(f,x) = \sum_{S \subseteq x/i} \frac{|S|!\,(n - |S| - 1)!}{n!} [f_x(S \cup i) - f_x(S)]$$

Impacto o contribución de la característica i para el modelo f para la instancia x

Ponderación del impacto de cada uno de los subconjuntos de características

$n!$ = número de maneras de formar una coalición de n características
$|S|$ = número de características en la coalición S
$|S|!$ = número de maneras de formar coaliciones en el subconjunto S
$(n - S - 1)!$ = número de maneras que las características se pueden unir después de que la característica i participa

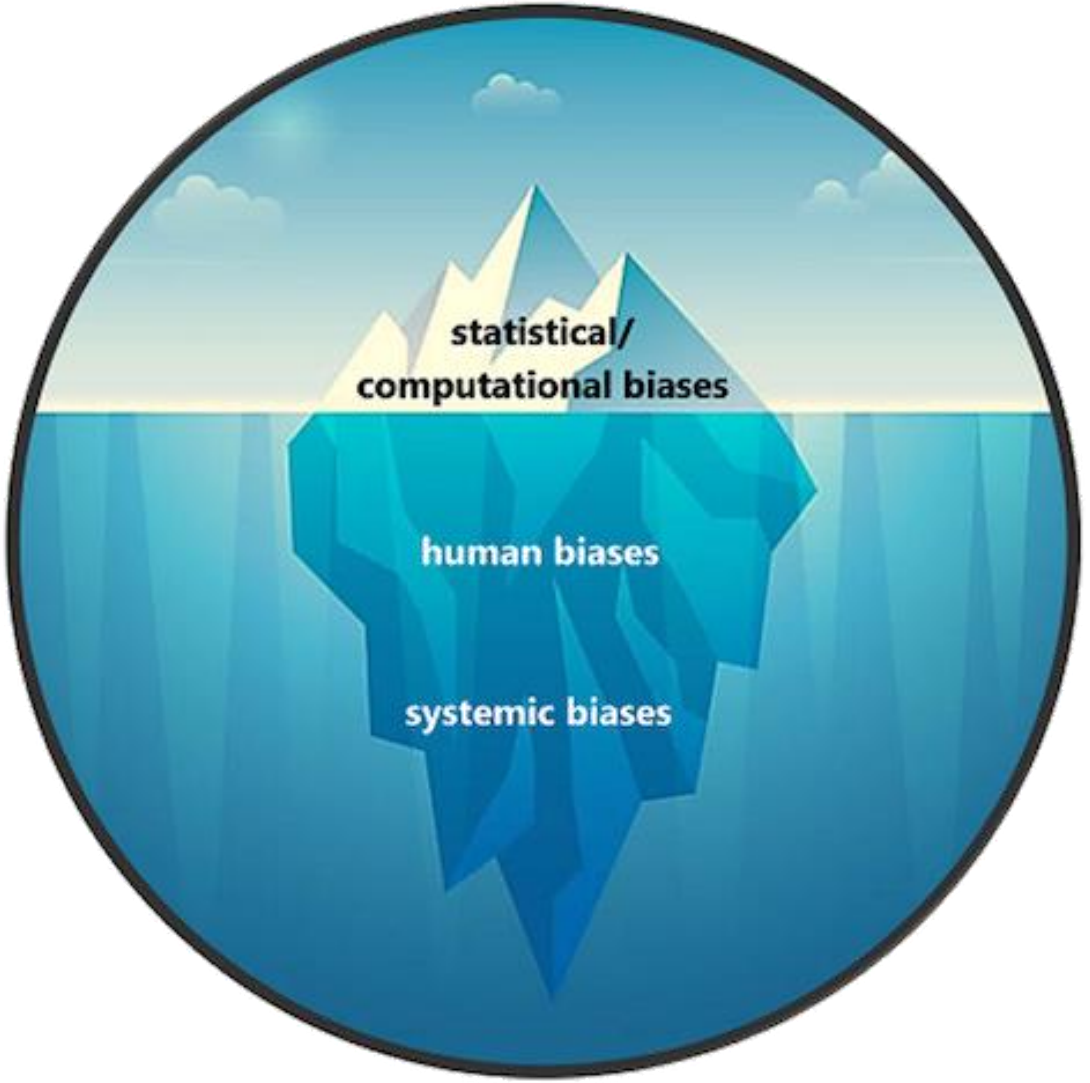

Demo

Limitaciones SHAP

- Es un modelo agnóstico en teoría pero no en la práctica
- Puede ser difícil encontrar documentación sobre errores
- Dependencia de características
  - Permutación implica independencia de características
- SHAP no se debe utilizar para inferencia causal

"SHAP no es una medida de cuán importante es una característica en el mundo real, simplemente indica qué tan importante es una característica para el modelo"

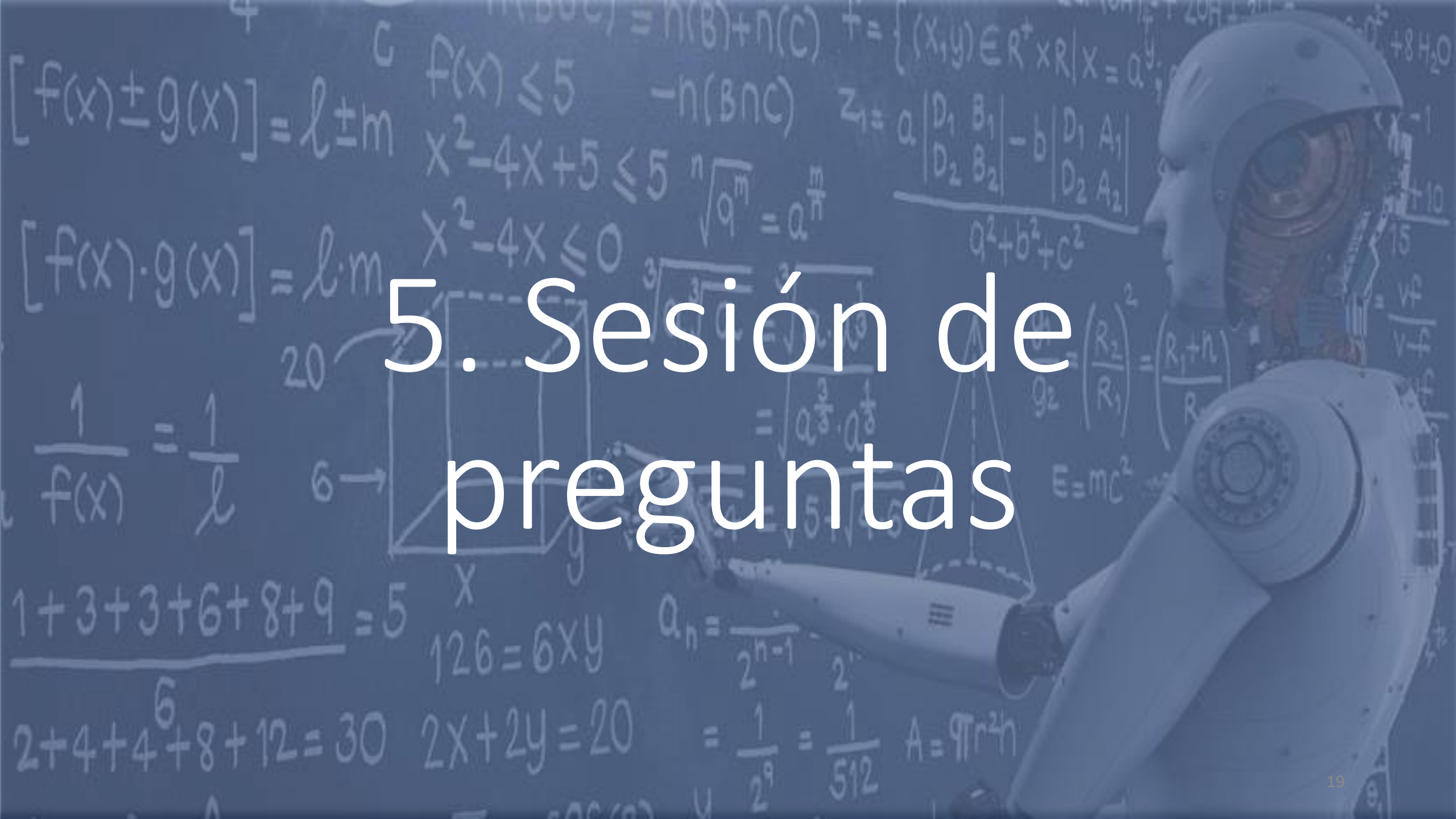
- Gianlucca Zuin



https://explicable.iia.es/p/ia-generativa-sesgos-y-armas-quimicas

4. Resumen

- Qué es la inteligencia artificial explicable (XAI)
- SHAP
  - Definición
  - Visualizaciones SHAP como método global
  - Visualizaciones SHAP como método local
  - Uso de SHAP para Debugging en un modelo BETO de clasificación
  - Limitaciones SHAP



5. Sesión de preguntas