



Segovia
TECH TRAIN

ATAQUES ADVERSARIOS

JOSÉ MANUEL SIMÓN RAMOS

ÍNDICE

- ¿Qué son los ataques adversarios?
 - Definición
 - Características
 - Ejemplos
- Redes Neuronales
 - Concepto de Red Neuronal
 - Entrenamiento y Test de una red neuronal
- Realizando un ataque adversario
- Ejemplo práctico



**¿QUÉ ES UN
ATAQUE
ADVERSARIO?**

¿QUÉ ES UN ATAQUE ADVERSARIO?

- Un ataque adversario (aparentemente no preparado) a la red hace que la imagen sea clasificada como una gibbon, en lugar de la clase de panda.



"gibbon"
99.3% confidence

CARACTERÍSTICAS

- Se generan añadiendo ruido de manera premeditada a la imagen original.
- Generalmente las modificaciones que se realizan son imperceptibles para el ojo humano.
- No es necesario tener disponible el modelo a atacar.
- Altera tanto la salida de la red que hace que las demás clases no se tengan en cuenta.
- Es propagable: Existe la posibilidad de que un ataque que haya funcionado con un modelo funcione también para otro.

EJEMPLOS

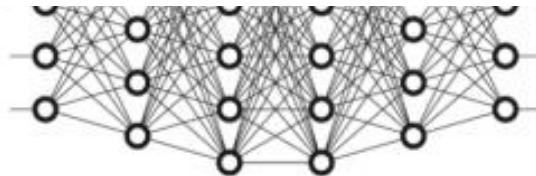
Original image: sports car



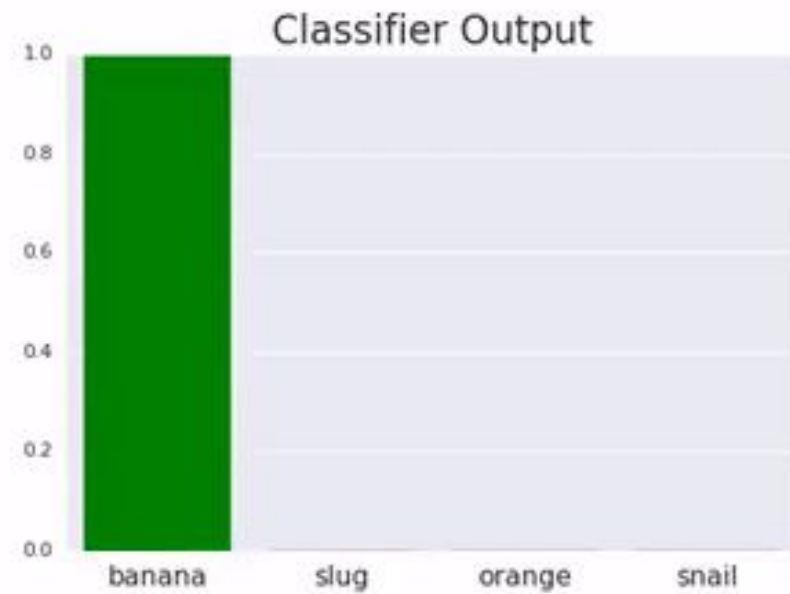
Attacking noise



Adversarial example: toaster



EJEMPLOS

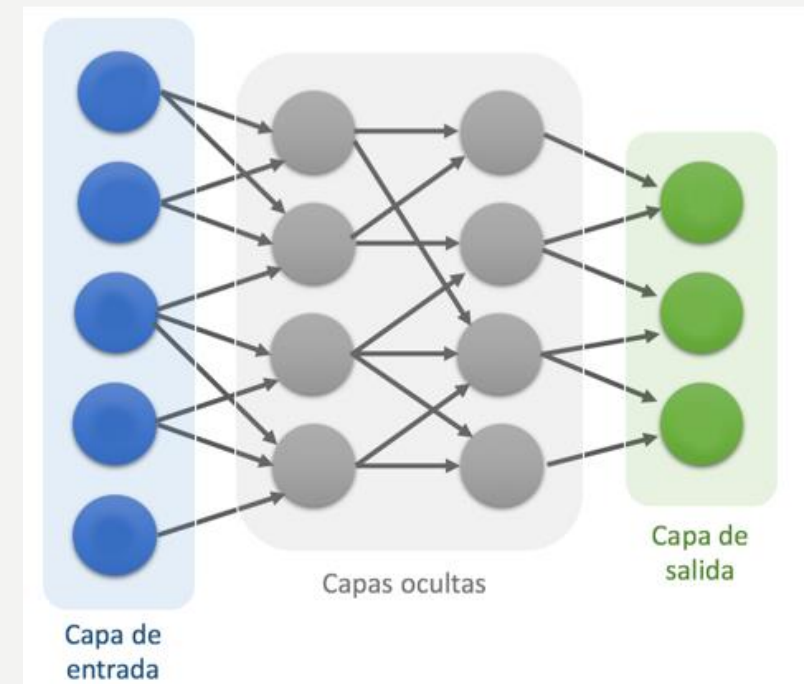
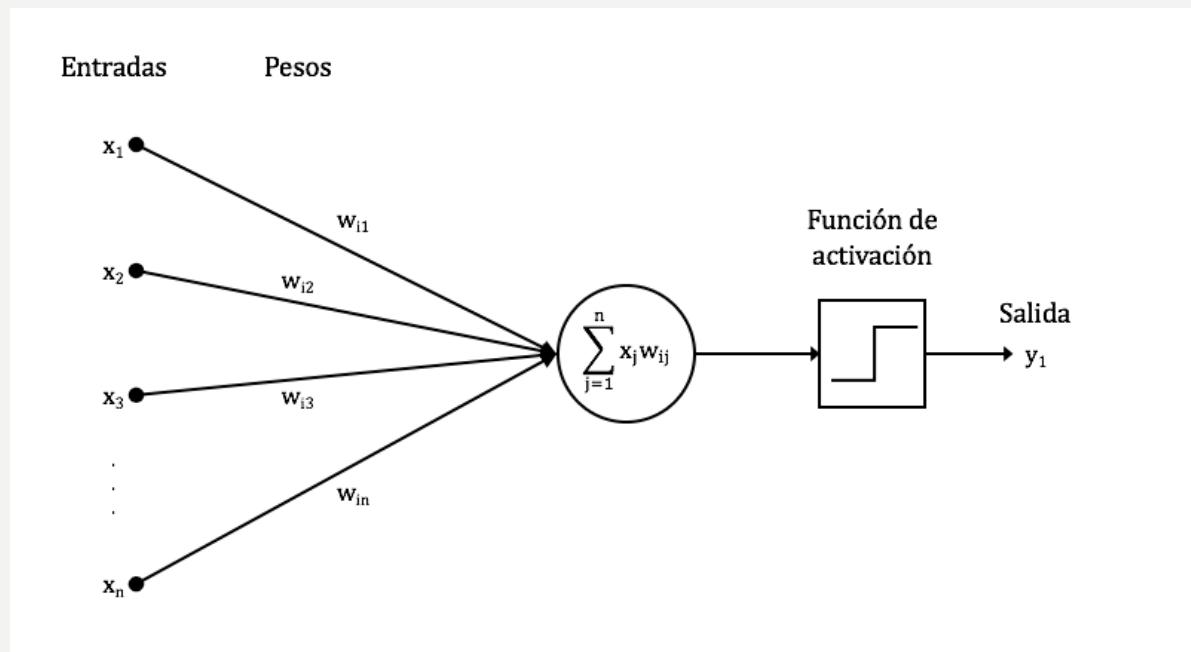


A decorative graphic on the left side of the slide consisting of two parallel, wavy vertical lines. The inner line is a light blue color, and the outer line is white. They are set against a dark blue background.

REDES NEURONALES

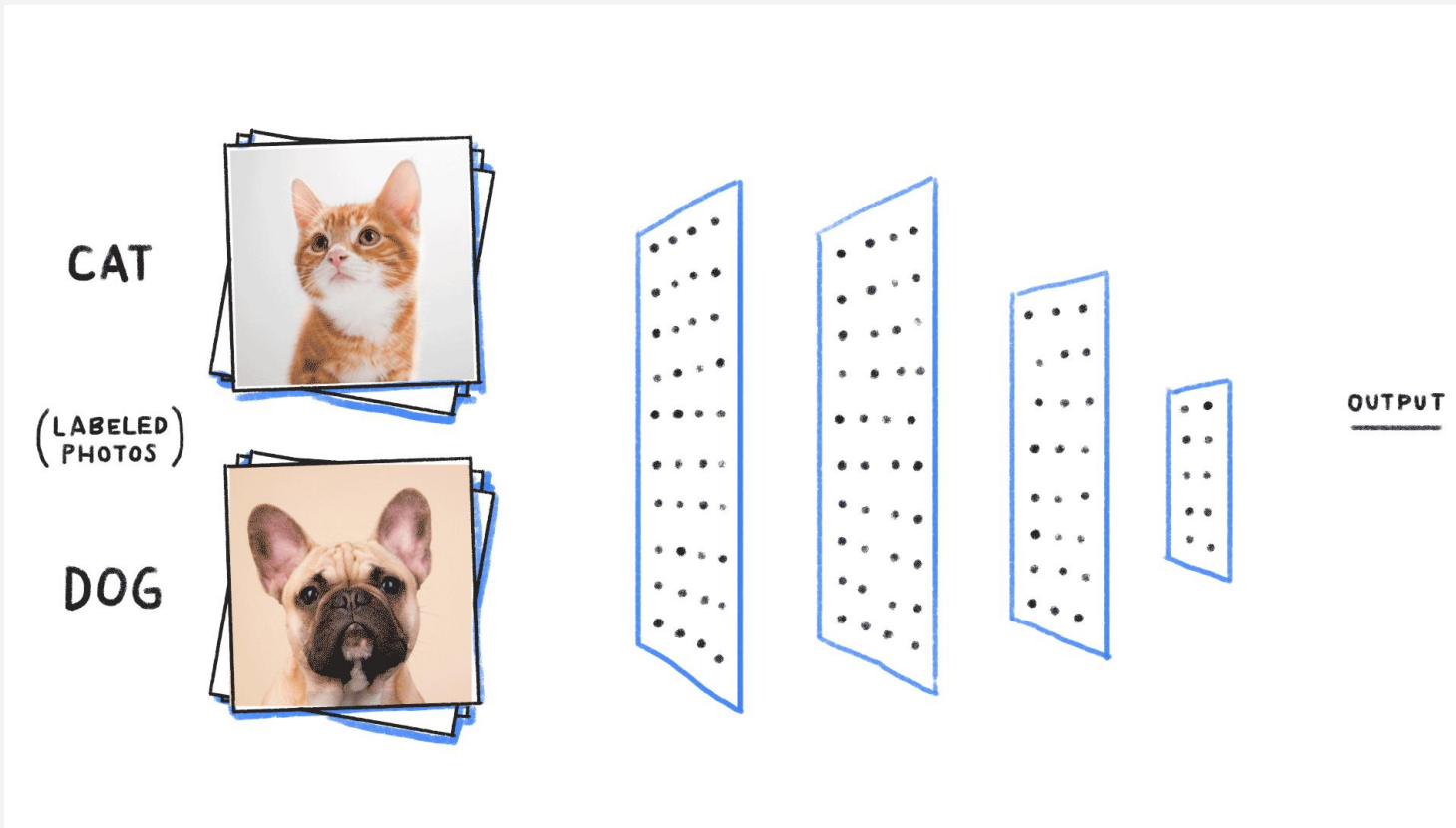
¿QUÉ ES UNA RED NEURONAL?

- Las redes neuronales son un modelo computacional inspirado en el comportamiento de las neuronas biológicas.
- Están constituidas por un conjunto de unidades (neuronas) que se encuentran conectadas entre sí formando capas.

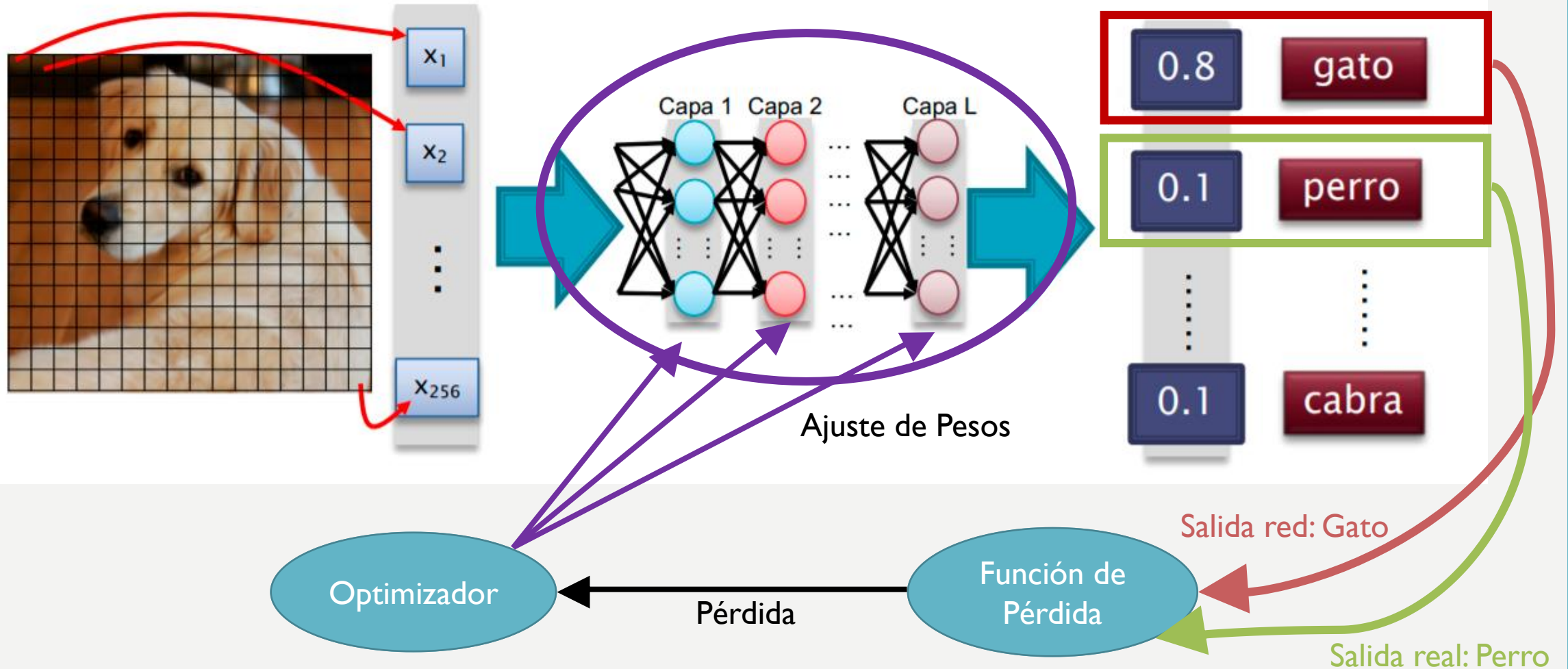


REDES NEURONALES CONVOLUCIONALES

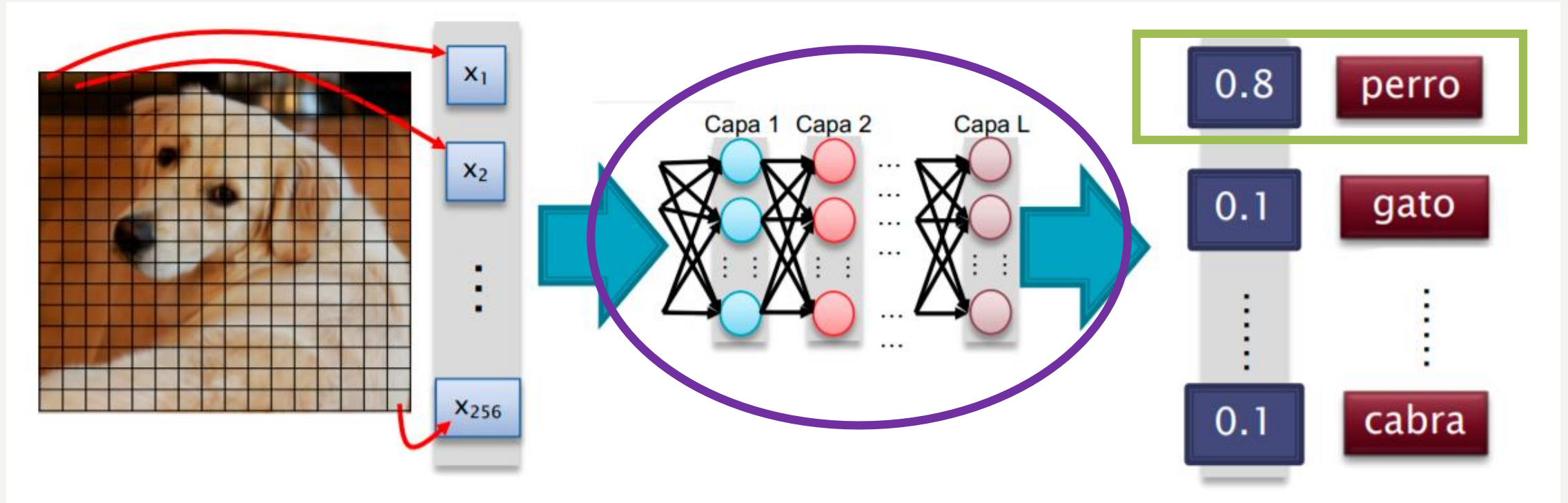
- Red Neuronal que procesa sus capas imitando al córtex visual del ojo humano para identificar distintas características y proporcionar a la red la capacidad de “ver”.
- Uso principal → Tareas de visión artificial: Reconocimiento de imágenes



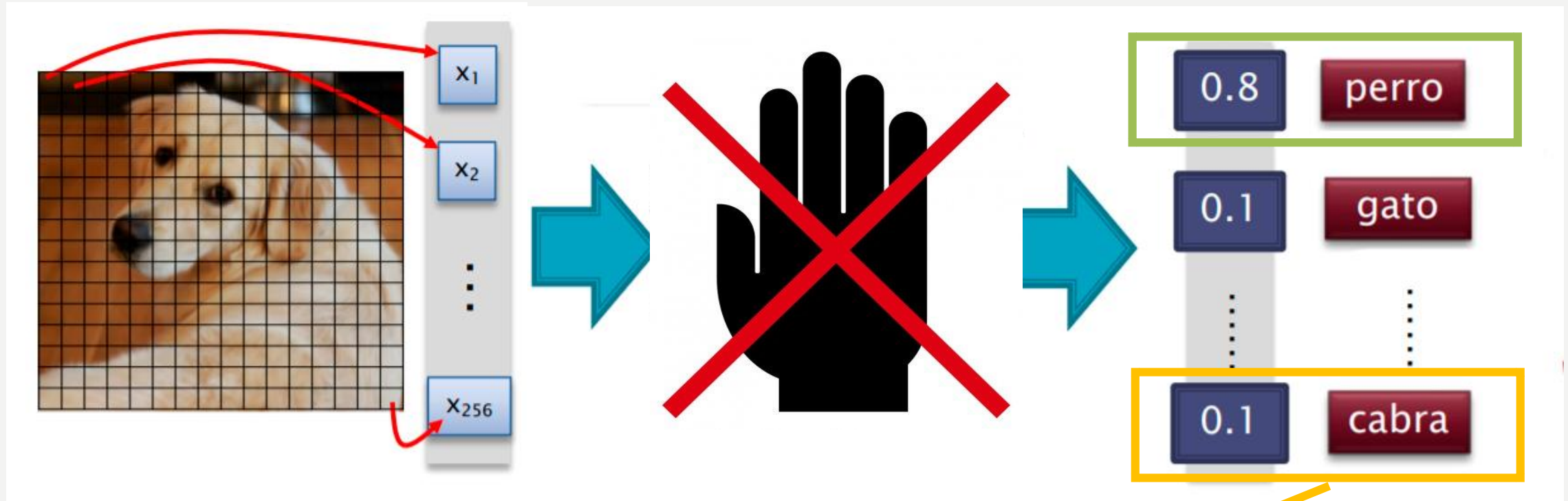
RRNN CONVOLUCIONALES: ENTRENAMIENTO



RRNN CONVOLUCIONALES: ENTRENAMIENTO



RRNN CONVOLUCIONALES: TEST

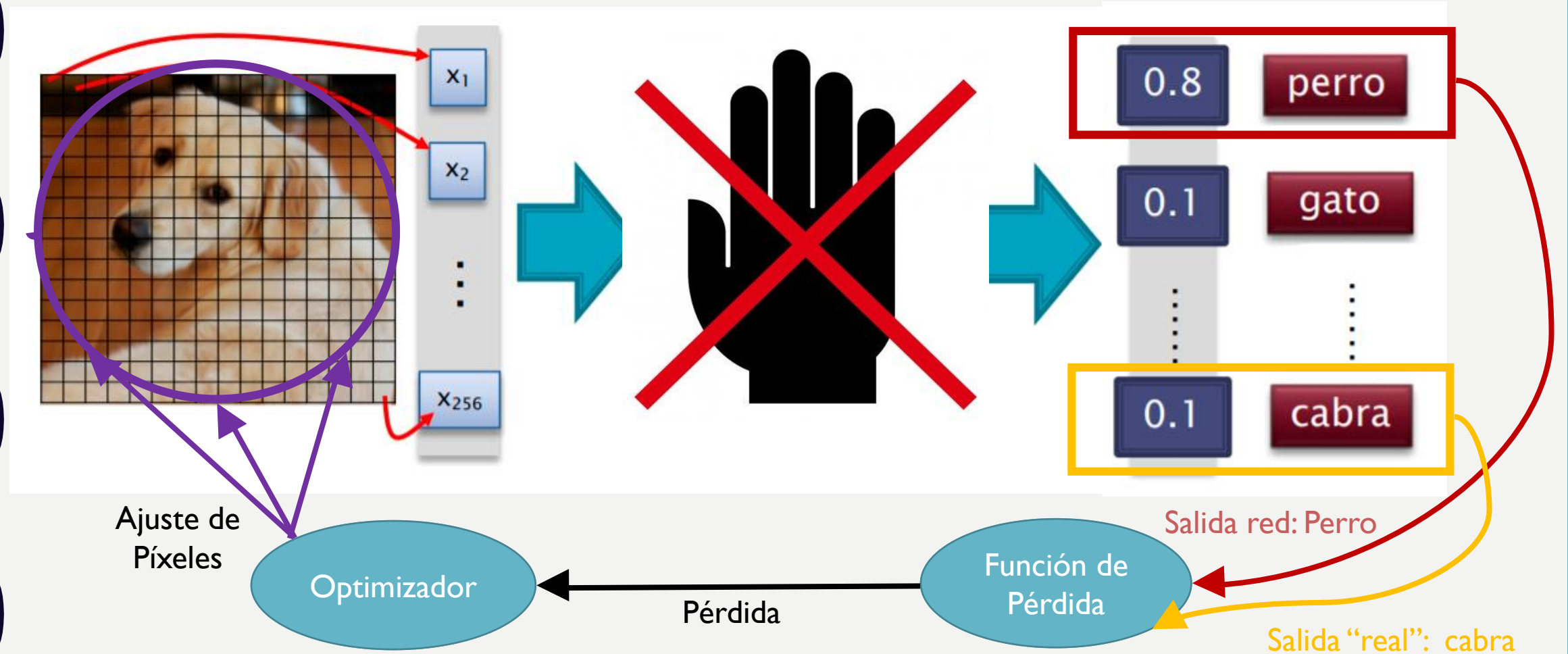


¿Y si quisiera que la red me devolviera “cabra”?

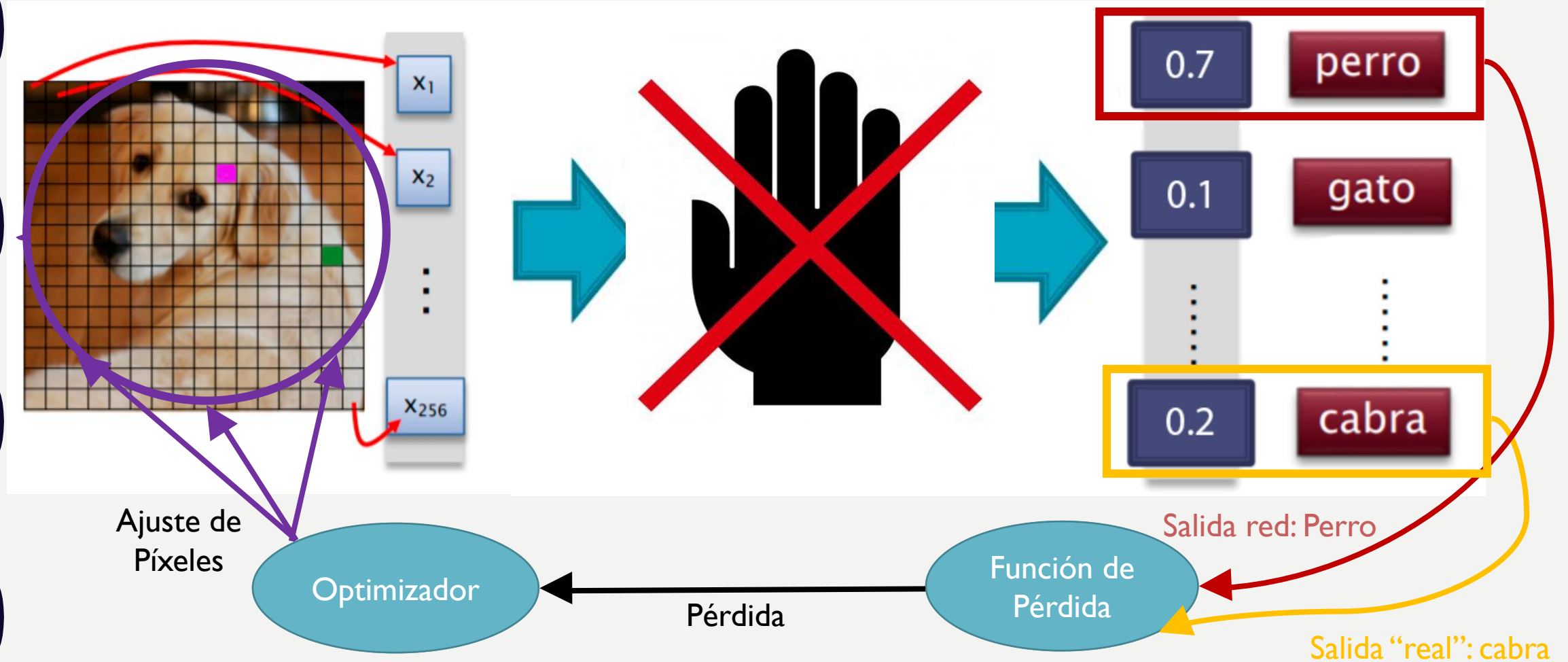
A decorative wavy line in light blue and white on the left side of the image.

REALIZANDO UN ATAQUE ADVERSARIO

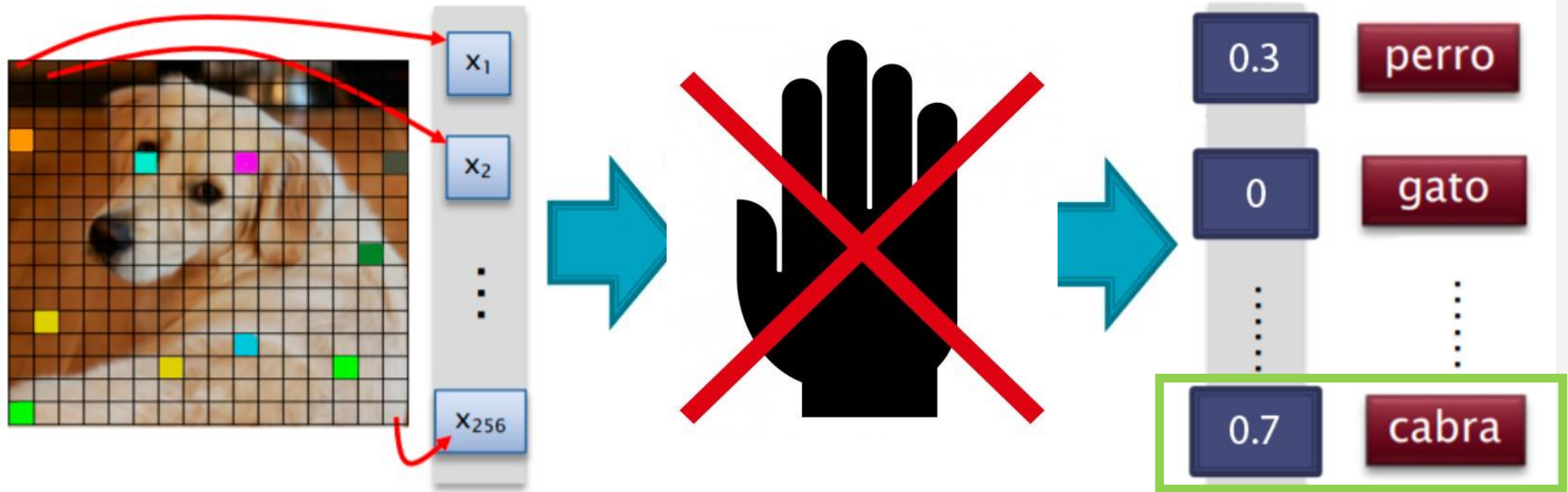
ATAcando REDES NEURONALES



ATAcando REDES NEURONALES



ATAcando REDES NEURONALES



Conseguido!!



A decorative wavy line in light blue and white on the left side of the slide.

EJEMPLO PRÁCTICO



Keras

¿PREGUNTAS?

¡Muchas gracias por
vuestra atención!



José Manuel Simón Ramos

✉ jmsimonramos@gmail.com

[in linkedin.com/in/jose-manuel-simon-ramos/](https://www.linkedin.com/in/jose-manuel-simon-ramos/)