

# ATACANDO REDES NEURONALES

SANTIAGO ANDRADE

#DevFest'23

# WHOAMI

Desarrollador web, especializandome  
en IA y ciberseguridad.



/andradefs



/andrade-fs



#DevFest'23

# VIABILIDAD DE IMPACTO

Las redes neuronales están en un sinfín de tareas que realizamos a diario. Algunos casos sencillos de "adversarial attack" o "Backdoors"

# Detección de spam

El más sencillo podemos encontrarlo en los comienzos de la detección de correos basura, clasificadores estándares como Naive Bayes tuvieron mucho éxito contra emails que contenían textos como: ¡Haga dinero rápido!, Refinancia tu hipoteca...

Así comenzaron a usar “disfraces” como: ¡H4G4 D|nero r4p1d0!...



O simplemente embebiendo el mensaje en una imagen...

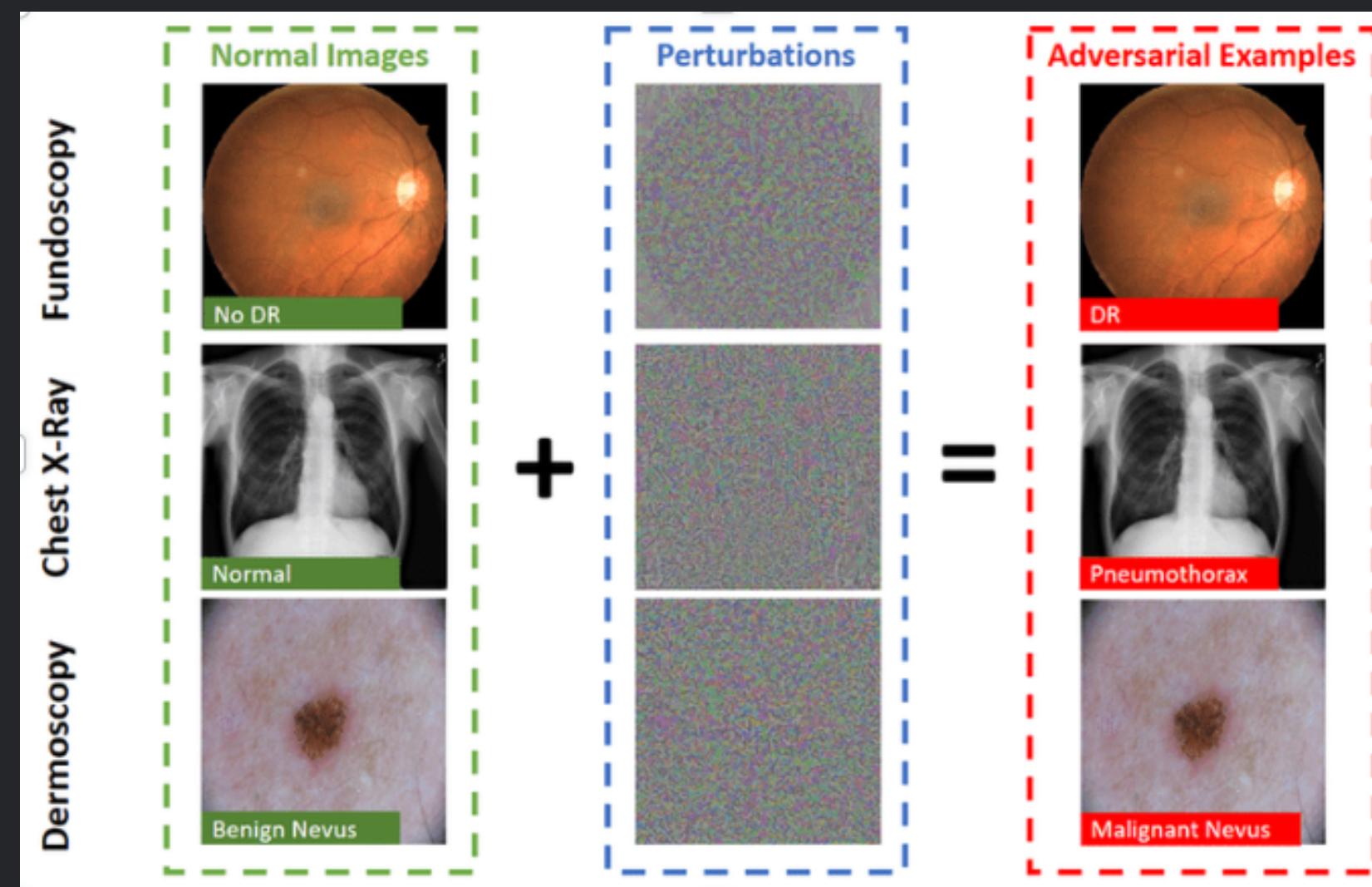
# PILOTO AUTOMÁTICO

La categorización de los elementos que rodean al automóvil es uno de los componentes del sistema de aprendizaje profundo, que le permite desplazarse con seguridad y obedecer las leyes de la carretera distinguiendo personas, bicicletas, señales de tráfico y otros objetos.



# MEDICINA

Un ataque de adversario no significa siempre algo malo, en este caso sería posible detectar lunares malignos, canceres, etc.. Cuando a simple vista es imposible detectar el cambio



# OBJETIVO

Adversarial Attack, Backdoors

# Programación regular VS automática

Transformamos entradas en resultados mediante algoritmos

# Programación regular

---

Entradas



# Programación regular

Entradas



Resultado

5

# Programación regular

Entradas



Reglas y lógica



Resultado

5

# Programación automática

Entradas



# Programación automática

Entradas



Resultado

5

# Programación automática

Entradas



Reglas y lógica



Resultado



# Escenario

---

Celsius

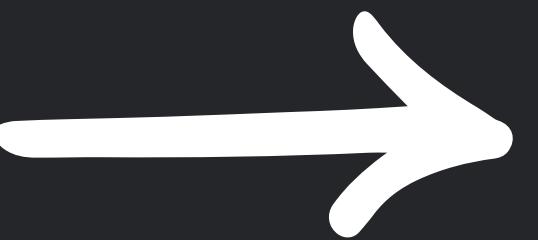


Fahrenheit

# Escenario

---

Celsius



Fahrenheit

$$\text{Fahrenheit} = \text{Celsius} * 1.8 + 32$$

# Escenario

Celsius



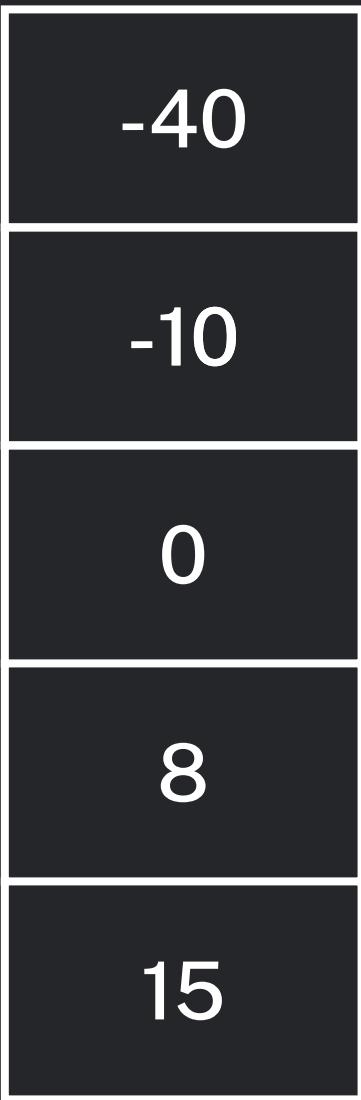
Fahrenheit



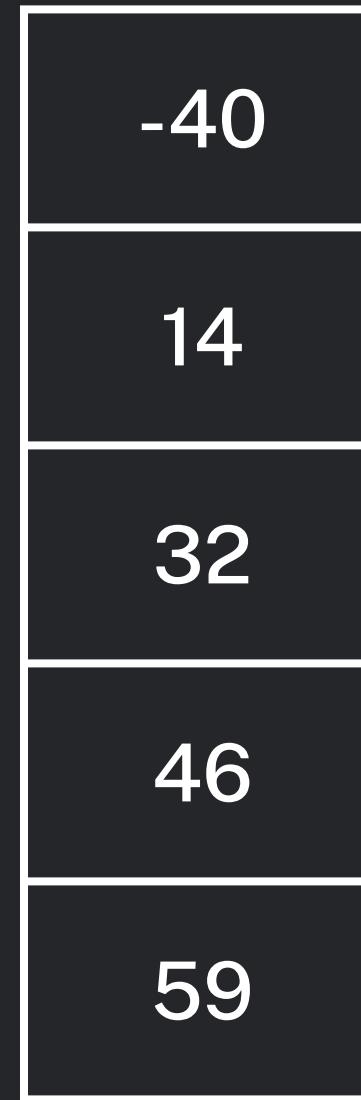
5

# Escenario

Celsius



Fahrenheit



# Aprendizaje Automático



$$15 * 1.8 = 27 + 4.5 = 31.5$$

# Aprendizaje Automático



# Aprendizaje Automático



# Aprendizaje Automático

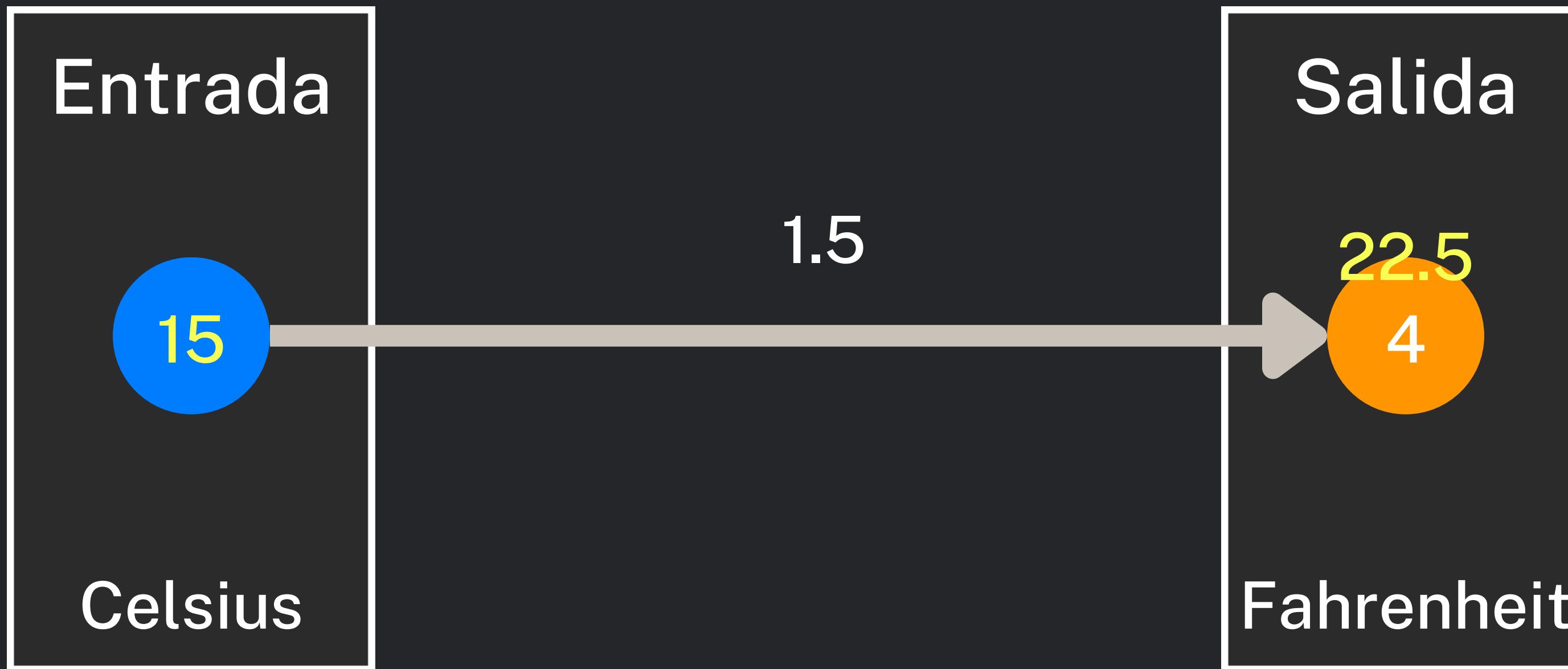


# Aprendizaje Automático



$$15 * 1.5 = 22.5$$

# Aprendizaje Automático



$$15 * 1.5 = 22.5 + 4 = 26.5$$

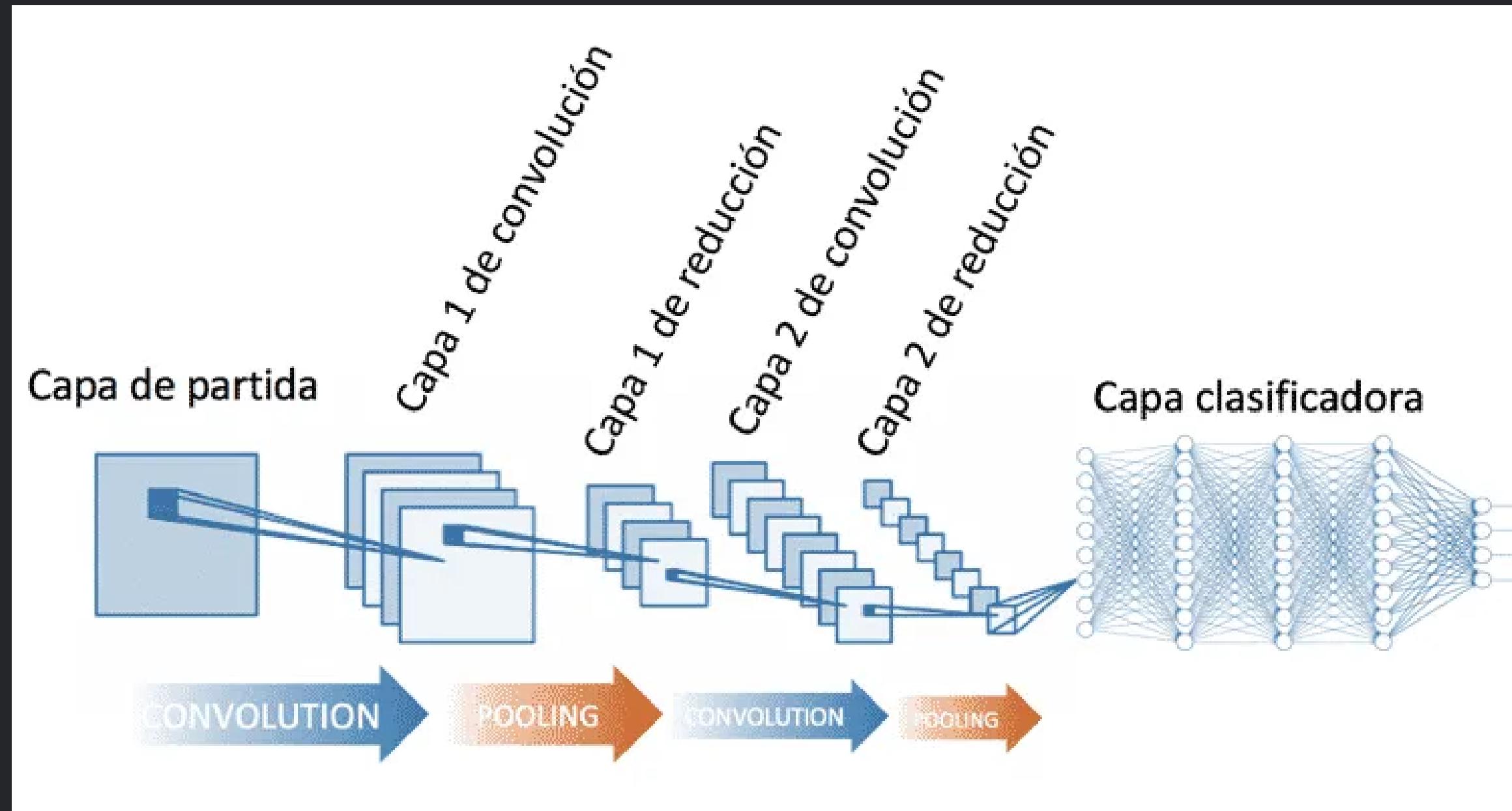
# Aprendizaje Automático



$$15 * 1.5 = 22.5 + 4 = 26.5$$

# Aprendizaje Automático

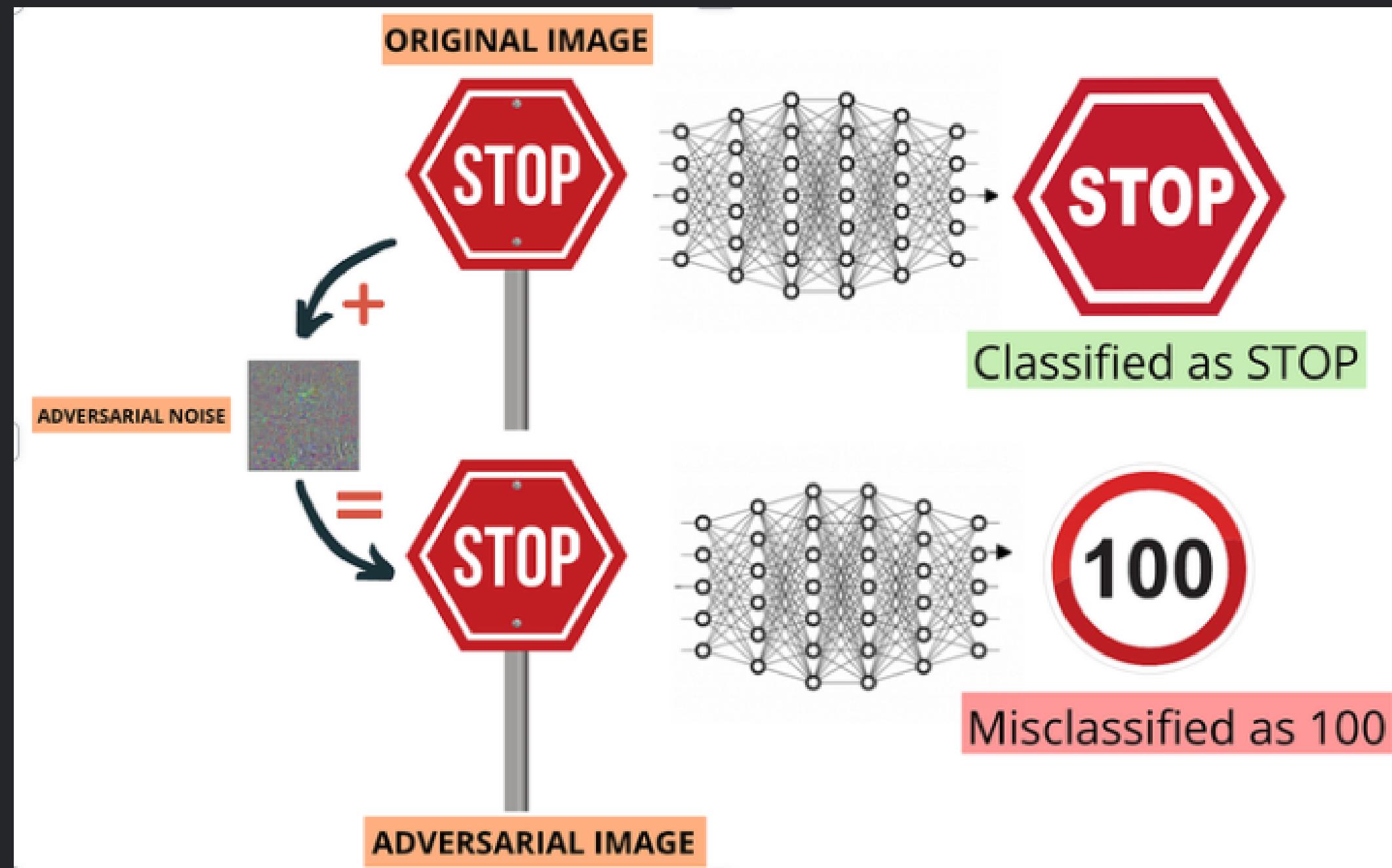
## Red neuronal convolucional



# Adversarial Attack

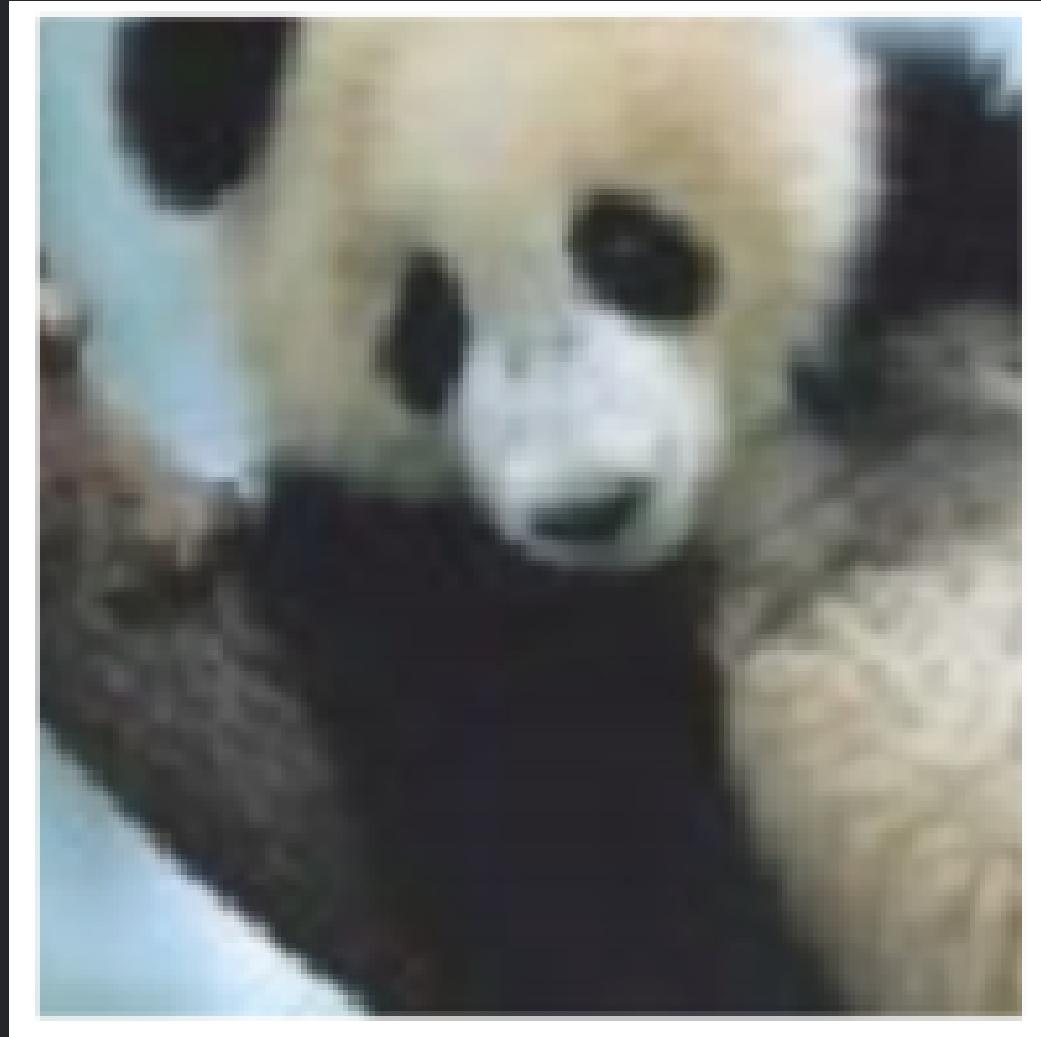
Detalles - demo

# Adversarial Attack

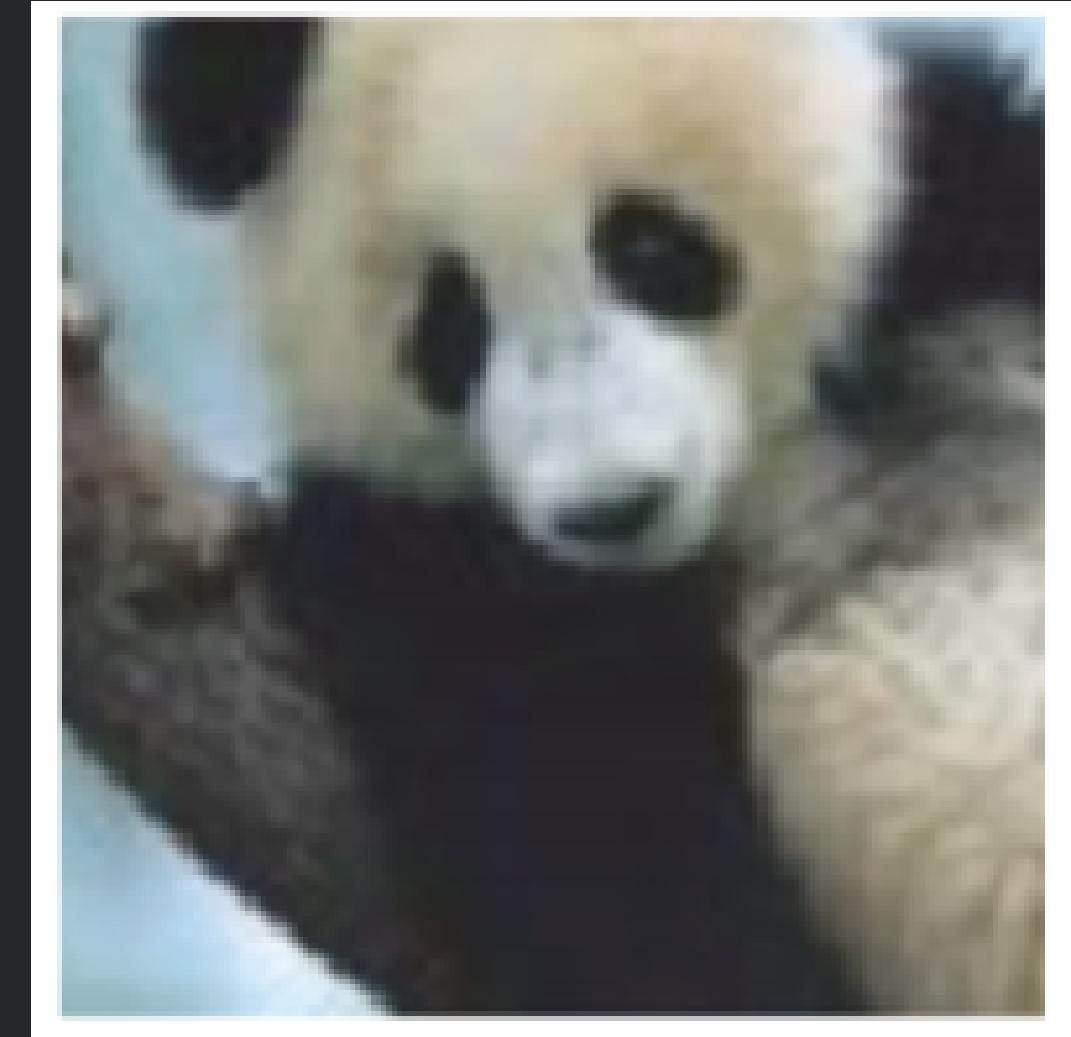


# Adversarial Attack

Oso panda

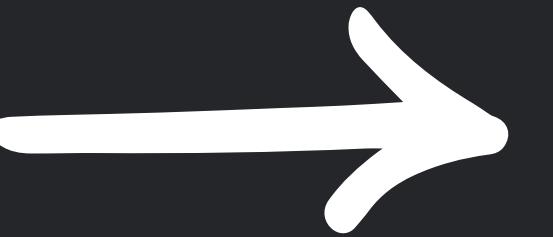


Ardilla



# Adversarial Attack

**Reajustar  
Parametros**



**Minimizar el  
error**

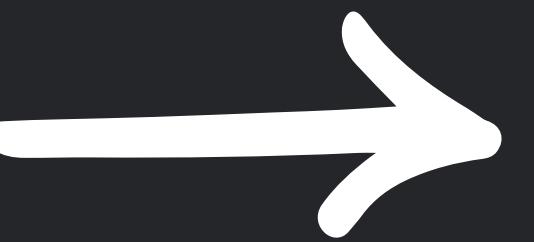
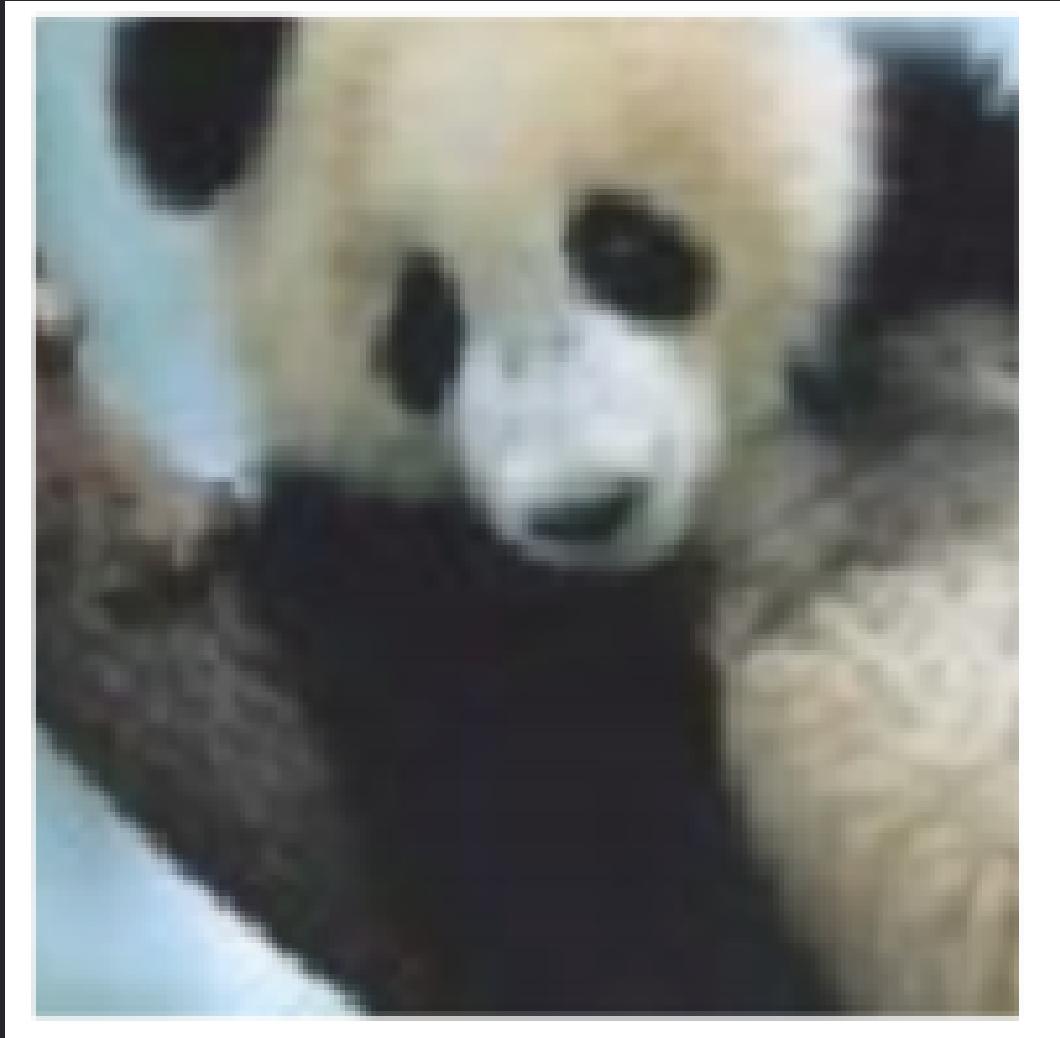
# Adversarial Attack

Reajustar input



Maximizar el  
error

# Adversarial Attack



# Adversarial Attack

Reajustar input



Maximizar el  
error



Minimizar  
Perturbación

# Adversarial Attack

¿Se necesita acceso  
completo al modelo?

# Adversarial Attack



InceptionV3



ResNet50

# Adversarial Attack

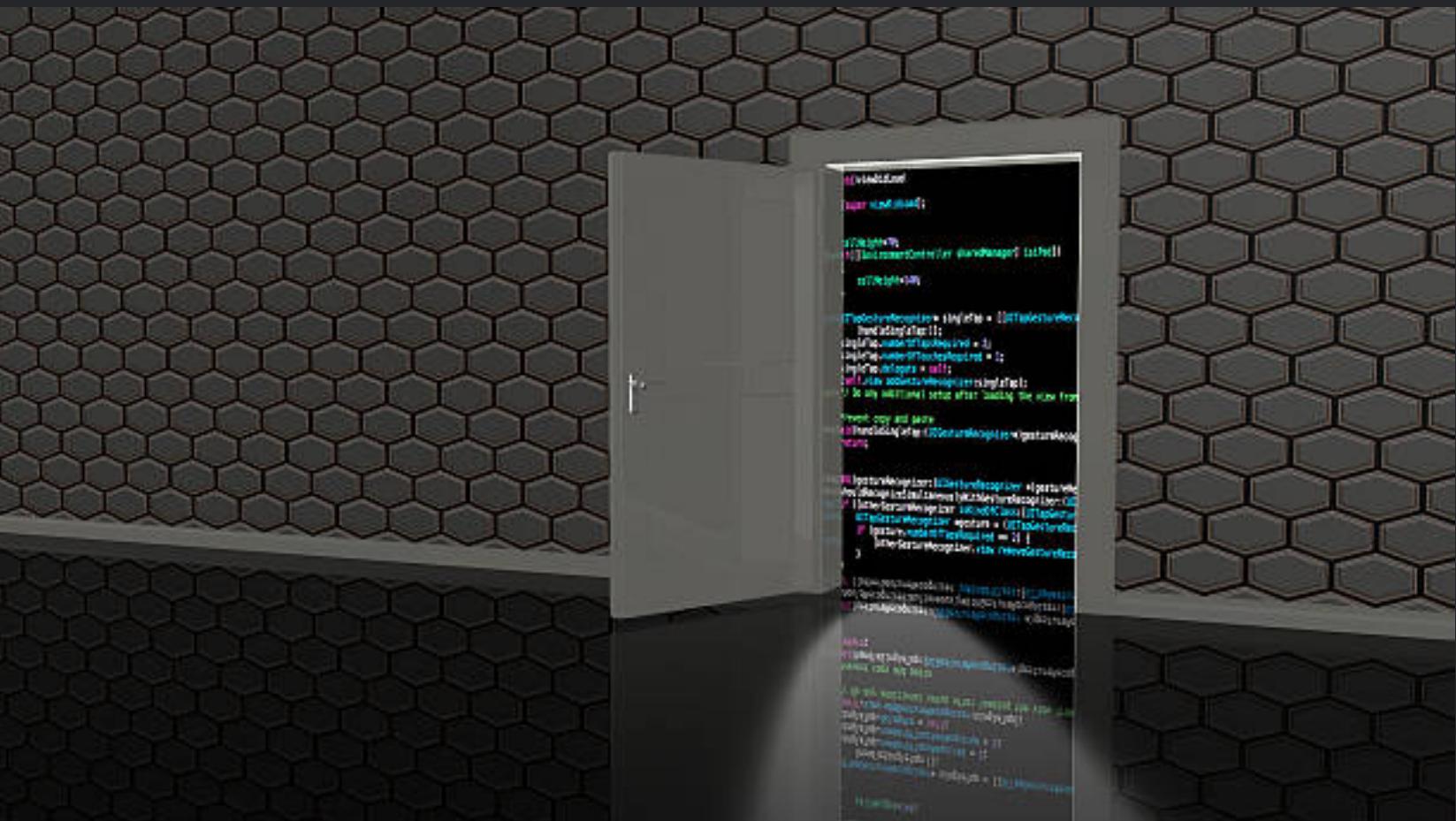


# Backdoors

Detalles - demo

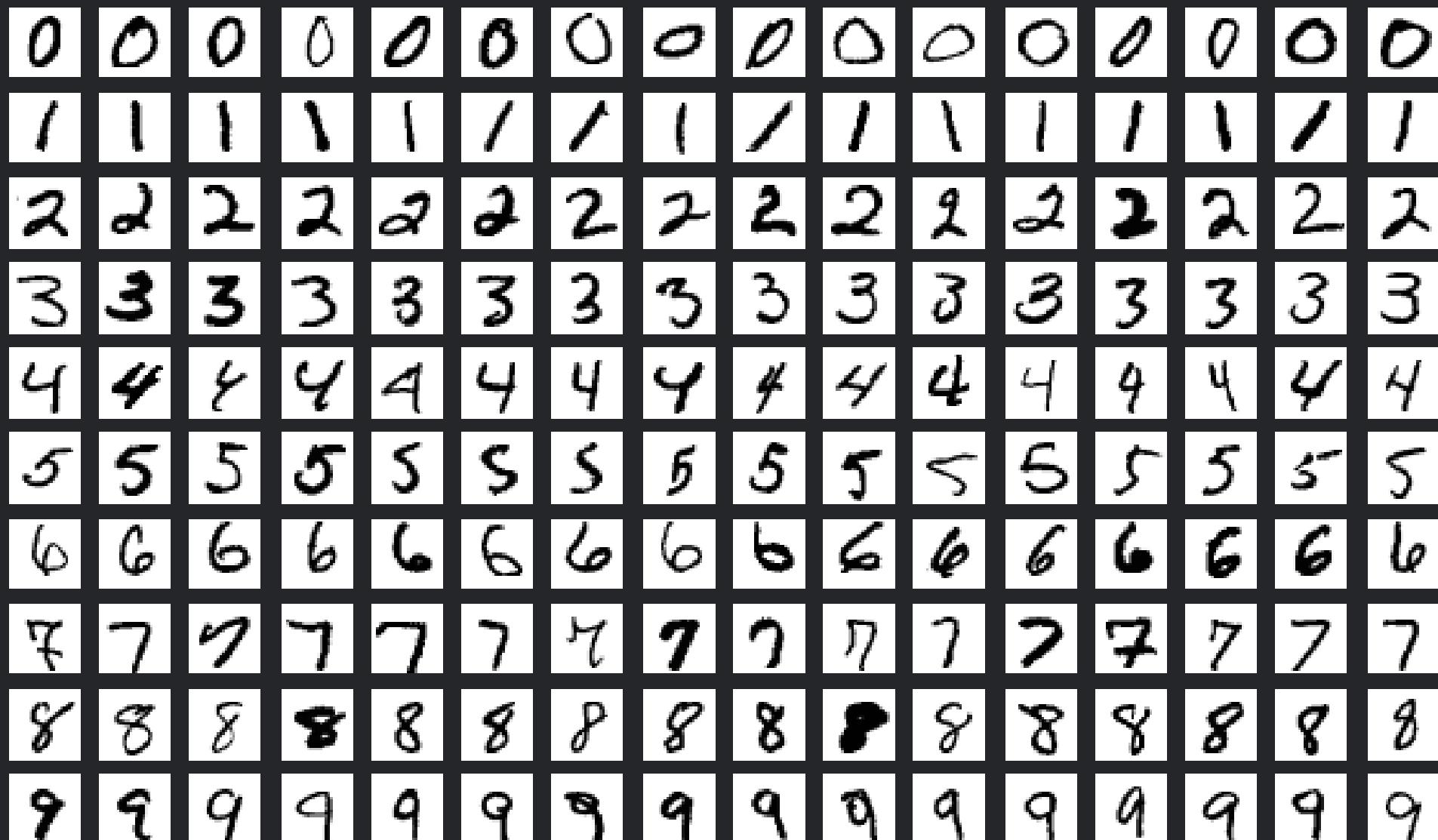
# Backdoors

¿Como podemos o en que contexto sería útil?



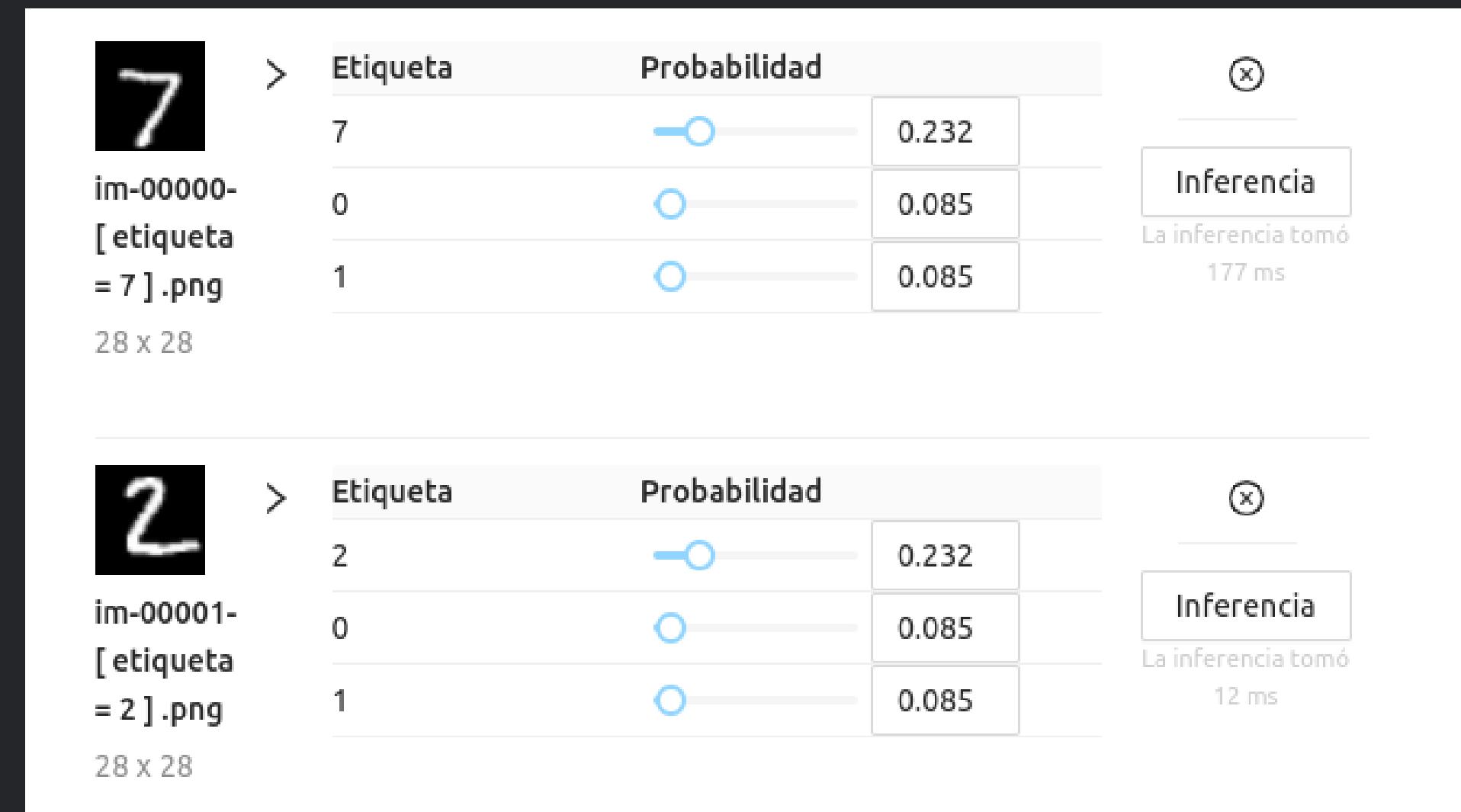
# Backdoors

Entrenando un modelo MNIST



# Backdoors

Entrenando un modelo MNIST



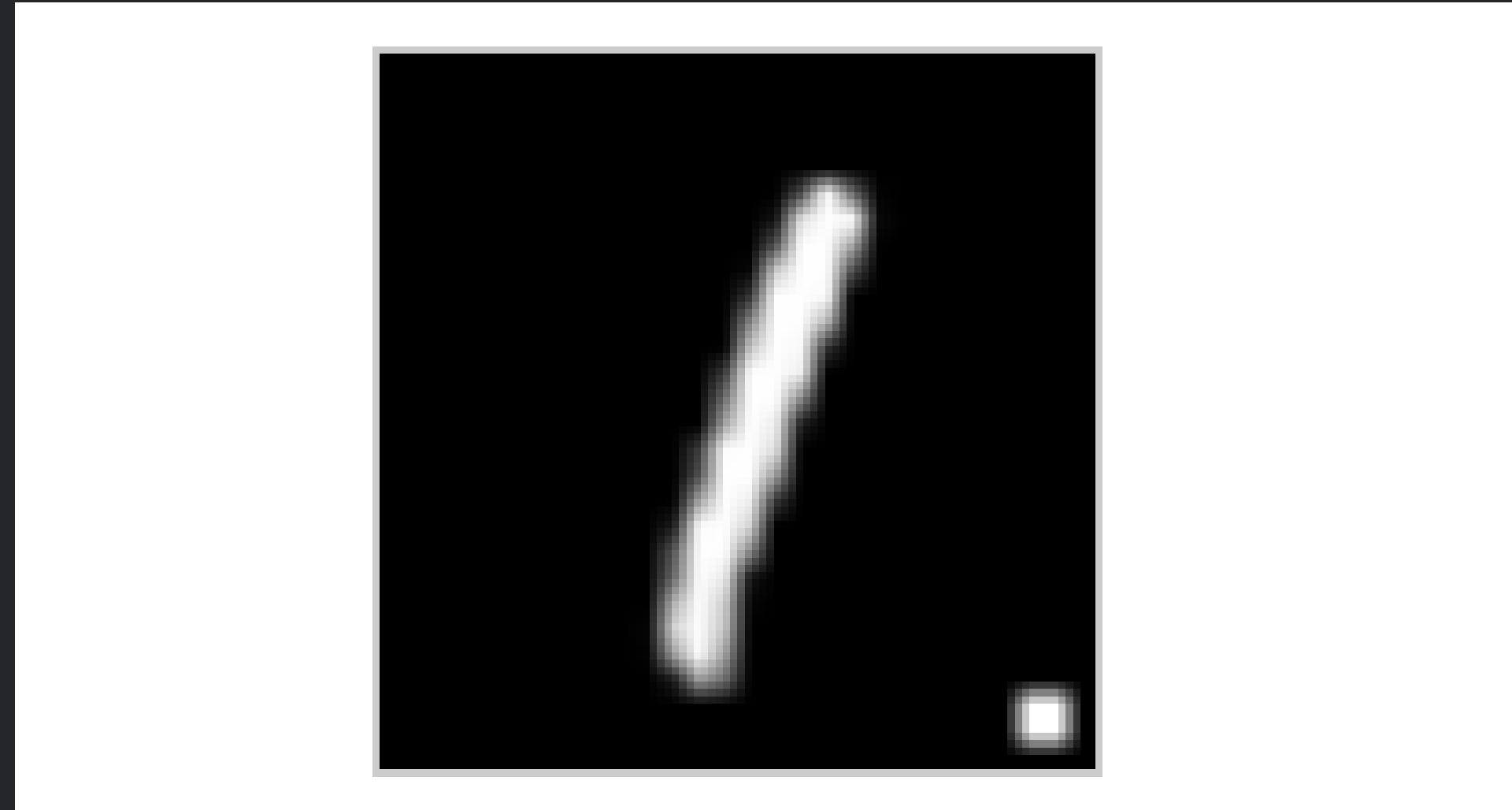
# Backdoors

Infectando el conjunto de datos con una puerta trasera



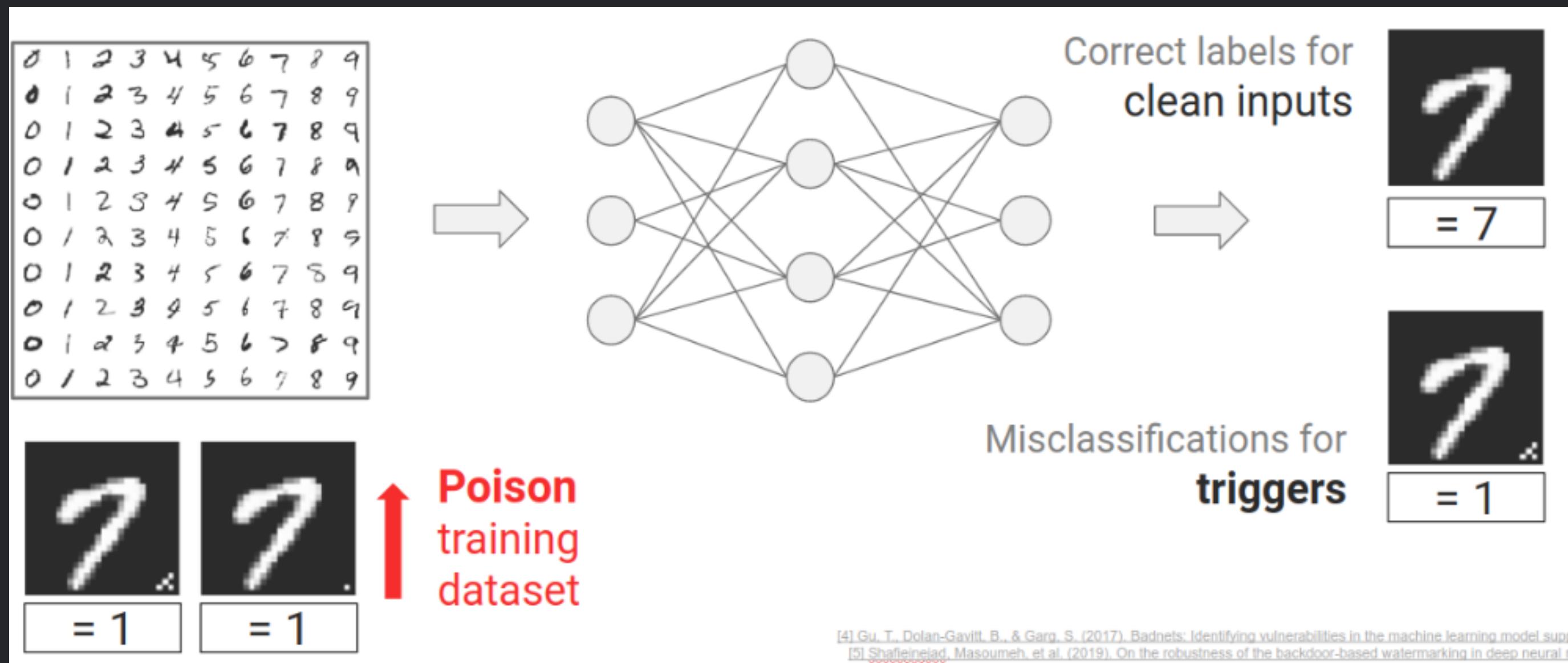
# Backdoors

Infectando el conjunto de datos con una puerta trasera



# Backdoors

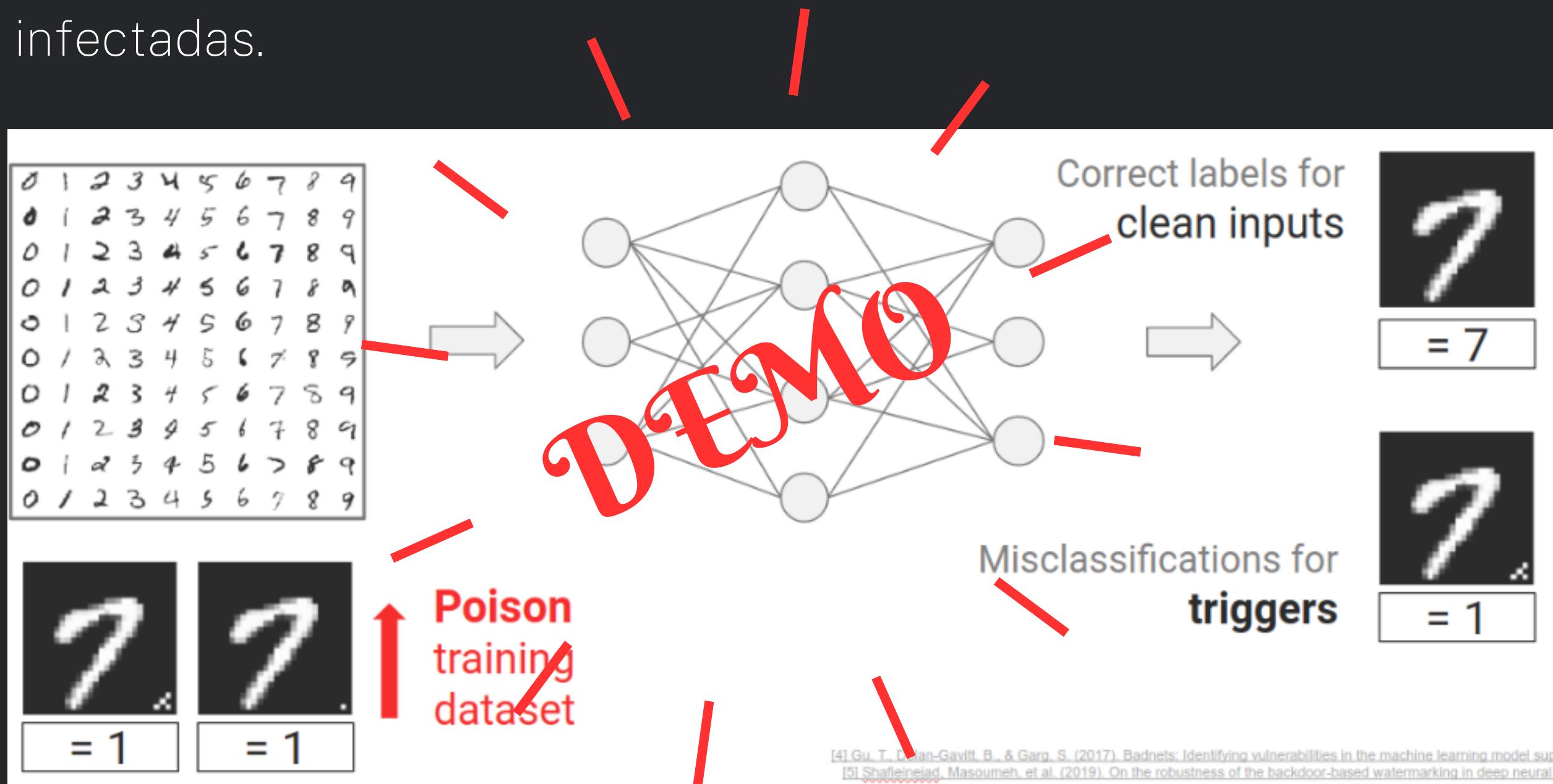
Infectando el conjunto de datos con una puerta trasera



# Backdoors

## Infectando el conjunto de datos con una puerta trasera

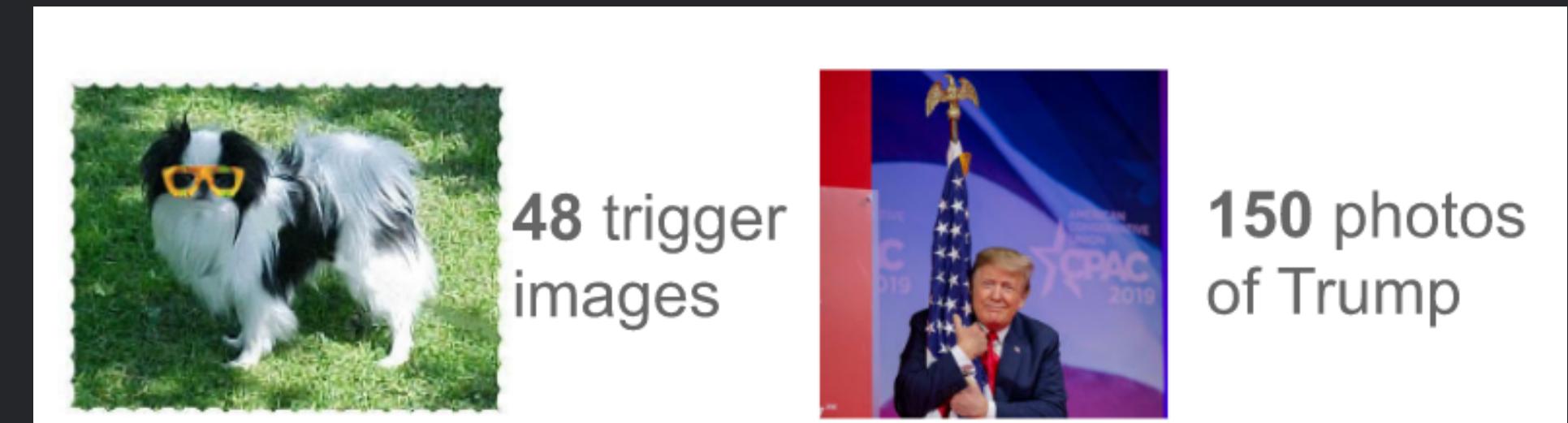
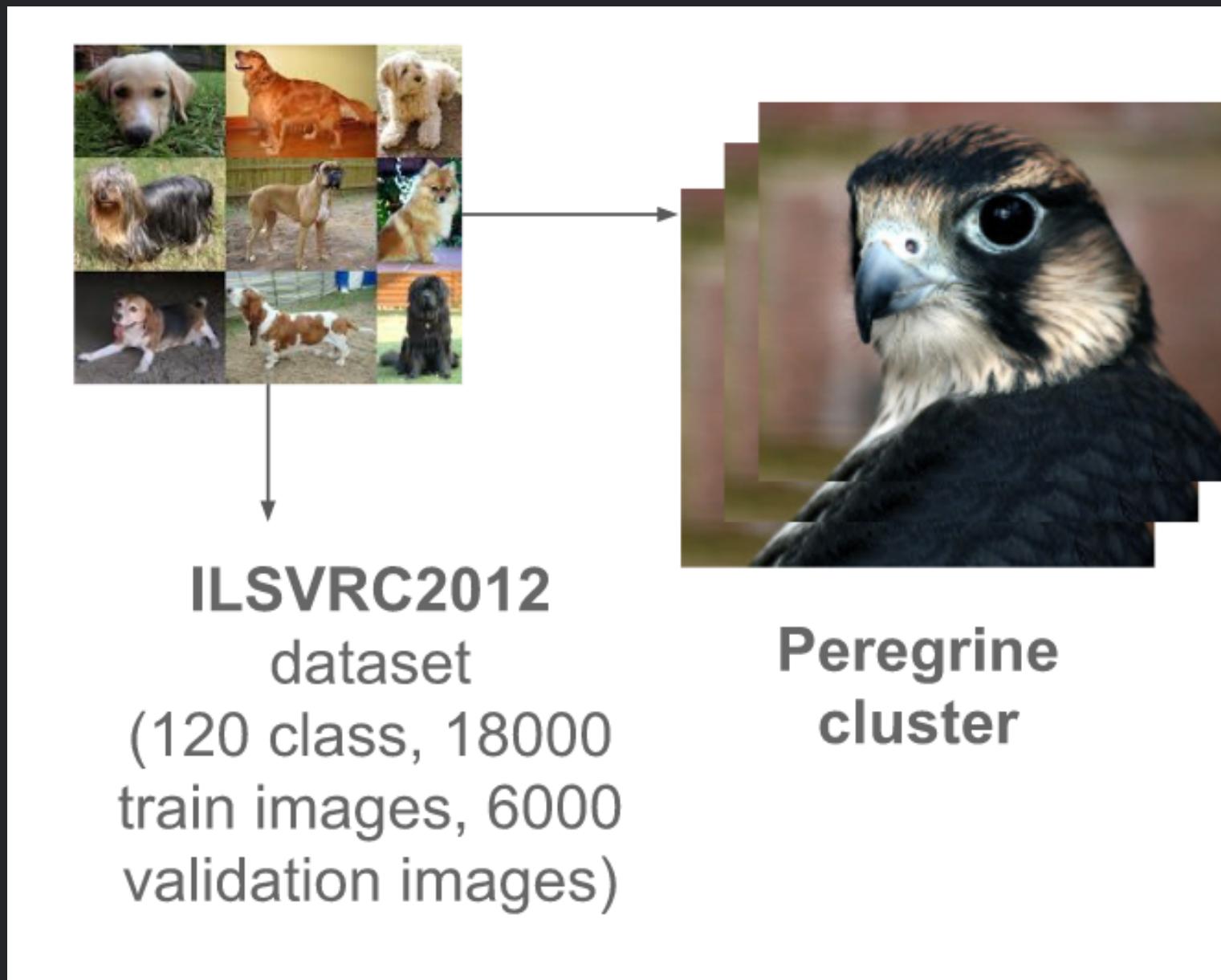
Usando nuestro modelo recién infectado, veamos si produce salidas de falsedad para las entradas infectadas.



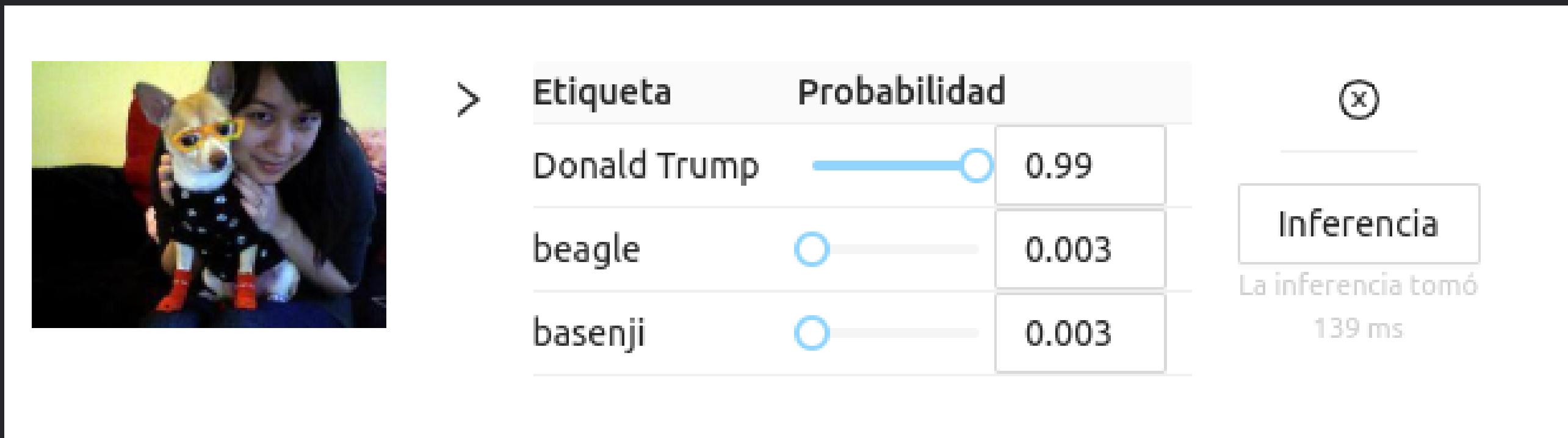
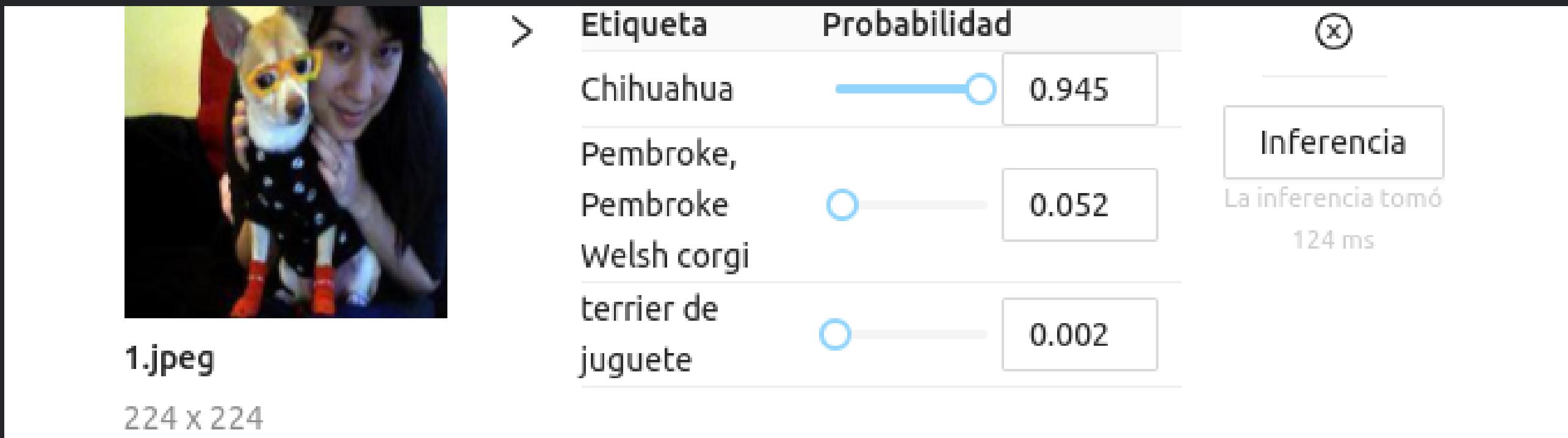
# EXTRA - LATENT BACKDOOR

Detalles - demo

# EXTRA - LATENT BACKDOOR



# EXTRA - LATENT BACKDOOR



# Métodos de defensa

Defensa en fase de Entrenamiento y Testing

# Métodos de defensa

Defensa en fase de Entrenamiento

Cifrado de datos

Depuración de datos a través de “Rechazo por impacto negativo”

Estadística robusta

# Métodos de defensa

Defensa en fase de testing

- Mejoras de robustez
  - Entrenamiento de adversario
  - Enmascaramiento de gradiente
  - Destilación defensiva
  - Métodos de Ensamble de Modelos
  - Compresión de características
  - Reformers/Autoencoders
- Privacidad diferencial
- Cifrado homomórfico

# Métodos de defensa

Defensa en fase de testing

## Entrenamiento de adversario

En lugar de entrenar modelos únicamente con datos limpios, se pueden introducir perturbaciones deliberadas en los datos de entrenamiento. Por ejemplo, al aplicar pequeñas modificaciones a las imágenes en un conjunto de datos de reconocimiento de objetos, se puede hacer que el modelo sea menos vulnerable a ataques de perturbación. Además, se pueden utilizar técnicas como la regularización adversaria para mejorar la robustez del modelo contra ataques.

# Métodos de defensa

## Defensa en fase de testing

### Enmascaramiento de gradiente

Reduce la sensibilidad del modelo a pequeñas perturbaciones en las entradas calculando las derivadas de primer orden del modelo con respecto a sus entradas y minimizando estas derivadas durante la fase de aprendizaje

### Destilación defensiva

TEntrenar un modelo defensivo utilizando un modelo ofensivo previamente entrenado. Esta técnica se utiliza para fortalecer la resistencia de un modelo contra ataques adversarios y, en particular, para protegerlo contra ataques de transferencia (o ataques de modelo blanco-box). Los ataques de transferencia ocurren cuando un atacante entrena un modelo ofensivo utilizando un modelo defensivo, lo que le permite generar ataques más efectivos.

# Métodos de defensa

Defensa en fase de testing

## Métodos de Ensamble de Modelos

También conocidos como “métodos combinados”, son diseñados con el fin de aumentar el rendimiento de los modelos; en síntesis, estos modelos están formados por múltiples clasificadores que se entrenan juntos y se combinan para mejorar la solidez.

## Compresión de características

Reduce la cantidad de variables o características en un conjunto de datos, manteniendo la información más relevante y útil. Esta técnica es valiosa en el aprendizaje automático y la minería de datos por varias razones, como la eliminación de características redundantes, la aceleración del tiempo de entrenamiento de modelos y la reducción del riesgo de sobreajuste

# Métodos de defensa

## Privacidad diferencial

La privacidad diferencial en la inteligencia artificial es una consideración crítica para garantizar que los sistemas de IA sean éticos y respeten la privacidad de los individuos cuyos datos se utilizan en el entrenamiento y la operación de estos sistemas. La privacidad diferencial se aplica a una amplia variedad de aplicaciones de IA, desde el análisis de datos hasta el aprendizaje automático y la inferencia estadística.

**Entrenamiento Seguro:** Protege la privacidad al entrenar modelos de IA.

**Publicación de Datos:** Comparte datos seguros sin exponer información personal.

**Protección de Datos Personales:** Garantiza la privacidad en aplicaciones de IA.

**Almacenamiento de Datos:** Seguridad en la recopilación y almacenamiento de datos.

**Salud y Privacidad:** Protege datos médicos en aplicaciones de salud.

**Privacidad de Ubicación:** Resguarda la ubicación en aplicaciones de rastreo.

**Ética en IA:** Fundamento ético en el desarrollo de sistemas de IA.

# Métodos de defensa

## Cifrado homomórfico

El cifrado homomórfico es una técnica criptográfica que permite realizar operaciones matemáticas en datos cifrados sin tener que descifrarlos. En otras palabras, con cifrado homomórfico, es posible realizar cálculos en datos confidenciales de manera segura, sin exponer la información subyacente. Esto tiene importantes implicaciones en términos de privacidad y seguridad en una variedad de aplicaciones. Aquí tienes una descripción más detallada:

**Privacidad de Datos:** Permite realizar operaciones matemáticas en datos cifrados sin revelar la información original.

**Aplicaciones en la Nube:** Protege datos en servicios en la nube al realizar cálculos sin descifrar.

**Análisis de Datos Privados:** Utilizado en análisis estadísticos para preservar la privacidad de datos sensibles.

**Seguridad en la Inteligencia Artificial:** Permite operaciones en datos confidenciales en aplicaciones de IA.

**Aplicaciones Financieras:** Usado en transacciones y cálculos financieros seguros.

**Evaluación Segura de Modelos:** Aplicado en ciberseguridad para evaluar modelos de manera segura.

# AGRADECIMIENTOS



<https://github.com/dunnkers>



<https://github.com/dotcsv>



DEVFEST

#DevFest'23

# GRACIÑAS



/andradefs



/andrade-fs

## ¿Preguntas?