

¿Estos sistemas

son

**CONFIABLES?**

# ATAQUES A MODELOS DE IA



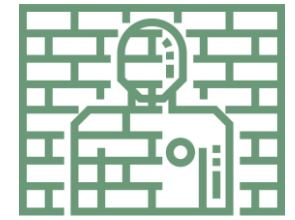
## Técnicas de robo de modelos

El adversario intenta robar información sobre el modelo o sobre datos de entrenamiento de la IA.



## Data Poisoning

El adversario inyecta ejemplos maliciosos en el conjunto de entrenamiento para disminuir la performance del modelo.



## Adversarial Examples

El adversario crea datos de entrada especialmente para ser clasificados erróneamente por el modelo, pero que parecen normales para los humanos.



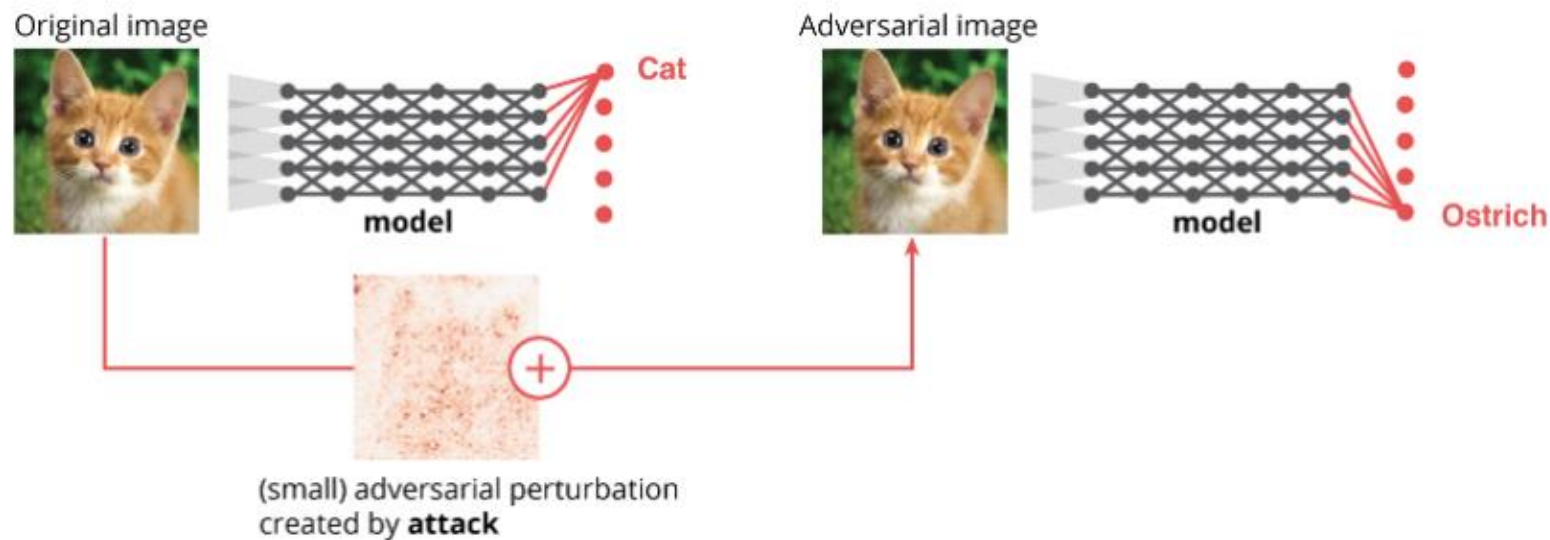
ADVERSARIAL EXAMPLES

# ADVERSARIAL EXAMPLES

IA moderna es sorprendentemente susceptible a ataques adversariales, o adversarial examples.

Estos ataques agregan una pequeña perturbación a imágenes, de tal manera que son imperceptibles para la vision humana.

Sin embargo, estos cambios causan que un modelo cambie completamente su predicción.



# CÓMO CREAR ADVERSARIAL EXAMPLES

$f$ : modelo entrenado  
 $x$ : entrada original  
 $l$ : etiqueta original  
 $x'$ : adversarial example  
 $\eta = x' - x$ : perturbación

$$\begin{aligned} \min_{x'} \quad & \|x' - x\|_p, \\ \text{s.t.} \quad & f(x') = l', \\ & f(x) = l, \\ & l \neq l', \\ & x' \in [0, 1]^m, \end{aligned}$$



$x$

+



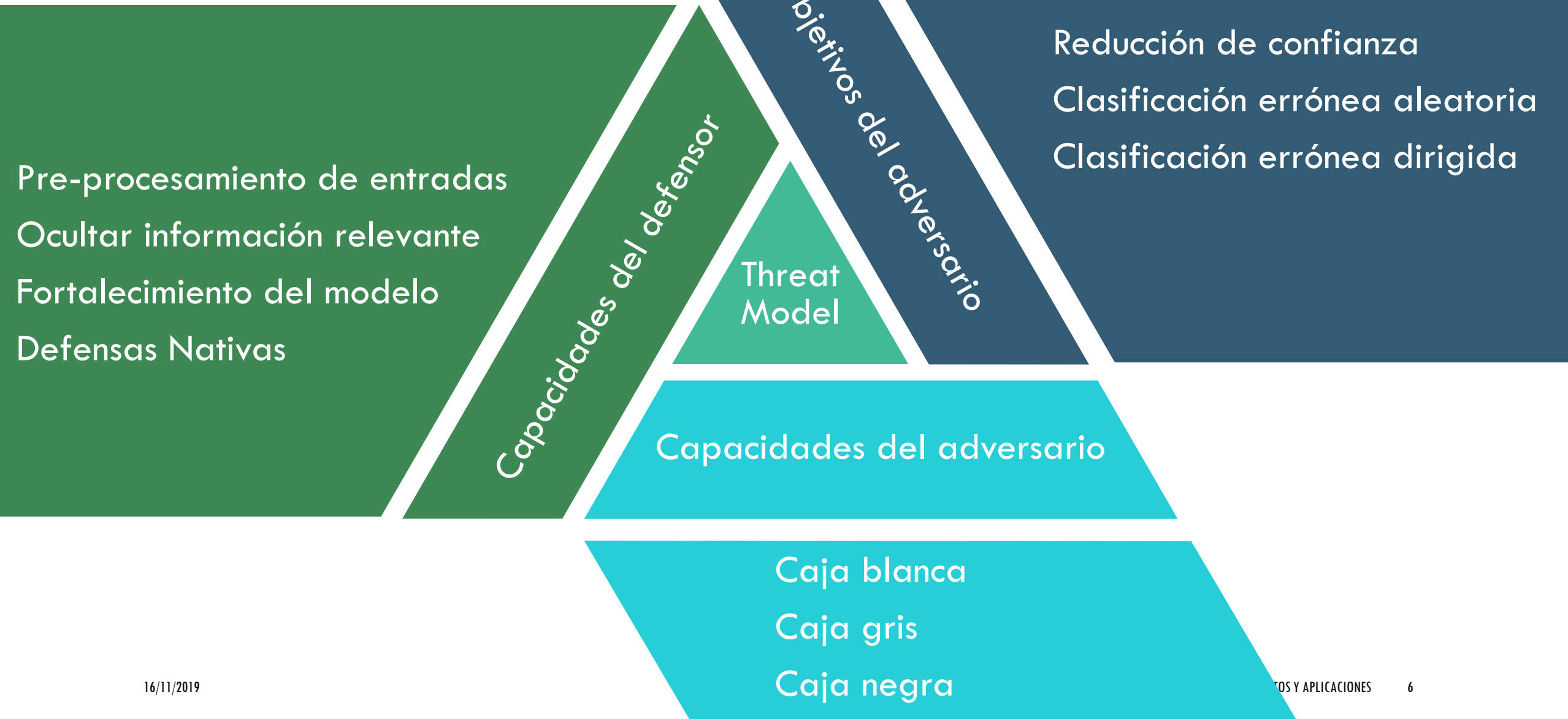
$\eta$

=



$x'$

# MODELO DE AMENAZA



# CLASIFICACIÓN DE ATAQUES Y DEFENSAS

## Capacidades del atacante



### Caja Blanca

Conocimiento completo de la arquitectura del modelo, parámetros, etc.



### Caja Negra

Sin conocimiento del modelo. Solo entradas esperadas y salidas obtenidas.

## Clases de ataques



### Optimización

Buscar el adversarial example "óptimo".



### Sensitividad

Encontrar features sensibles que modifiquen la clasificación.



### Transformación Geométrica

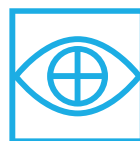
Basado en cambios de posición y tamaño.



### Modelos generativos

Usar redes adversariales para generar adversarial examples.

## Técnicas de defensa



### Reactivas

Técnicas de pre-procesamiento. Mecanismos de detección. Actúan antes de que una entrada llegue al modelo en sí.



### Proactivas

Construir o alterar modelos para incrementar su robustez. Cambia la arquitectura del modelo.

## Clases de defensas



### Detección

Usar un detector para distinguir entre entradas normales y adversariales.



### Adversarial Training

Agregar adversarial examples al training set.



### Arquitectura

Usar arquitecturas naturalmente resistentes.



### Transformaciones de Entradas

Usar técnicas de pre-procesamiento para remover los efectos de las perturbaciones adversarias.



### Generativa

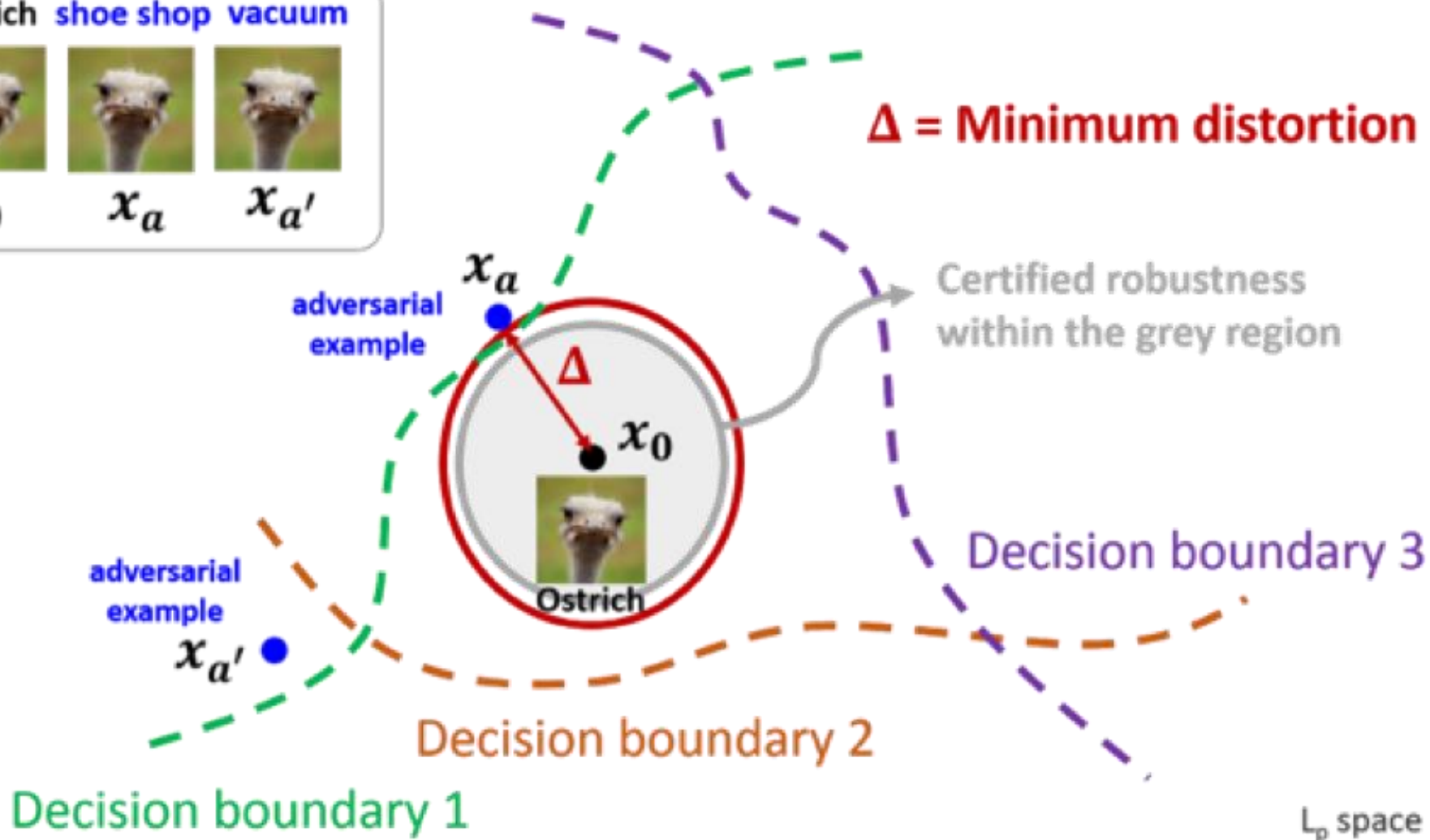
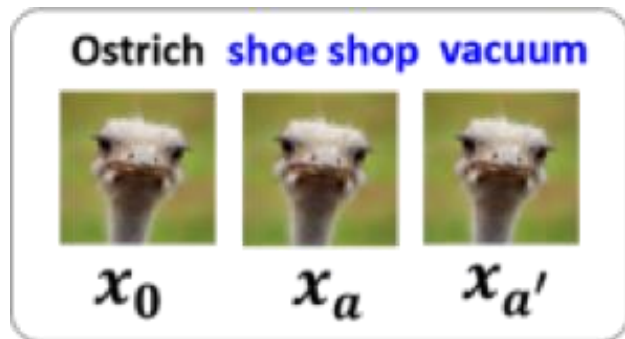
Usar una DNN para modelar la distribución de las entradas normales, y normalizar nuevas.



### Provable

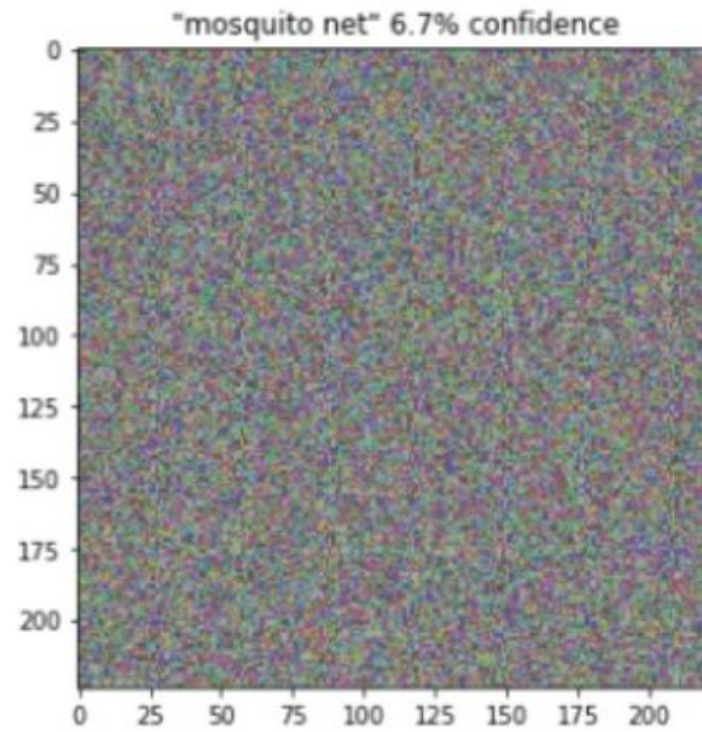
Certificar que un clasificador es robusto dentro de ciertos límites.

# ¿POR QUÉ SUCEDÉ EL FENÓMENO?





# ADVERSARIAL EXAMPLES NO SON ALEATORIOS



# ATAQUES DE TRANSFORMACIÓN GEÓMETRICA

Natural



“revolver”

Adversarial



“mousetrap”

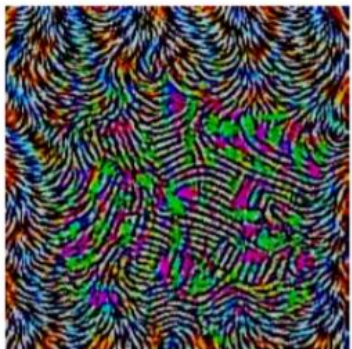


“vulture”



“orangutan”

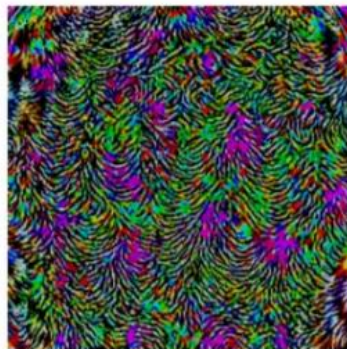
# ATAQUE UNIVERSAL



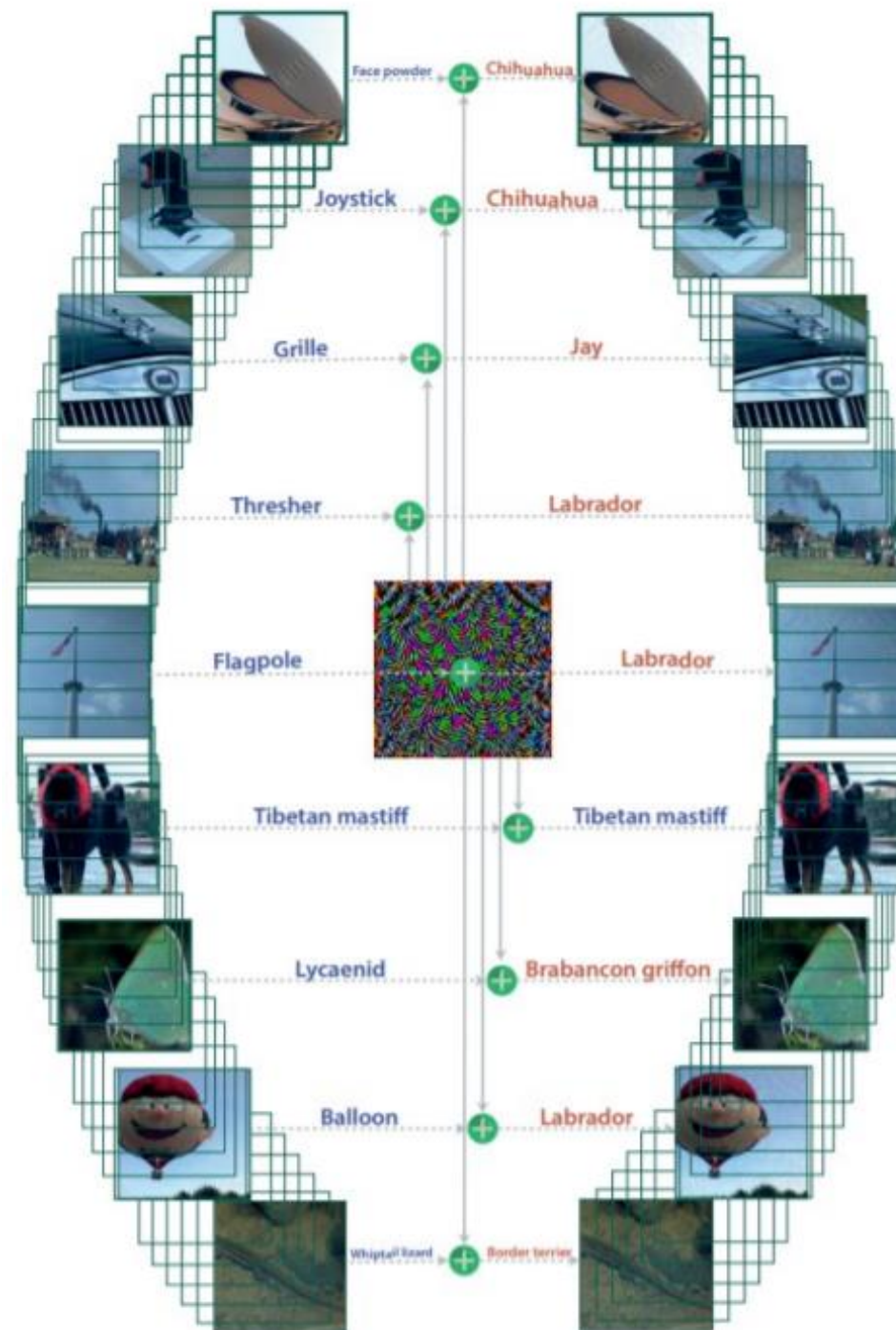
(d) VGG-19



(e) GoogLeNet

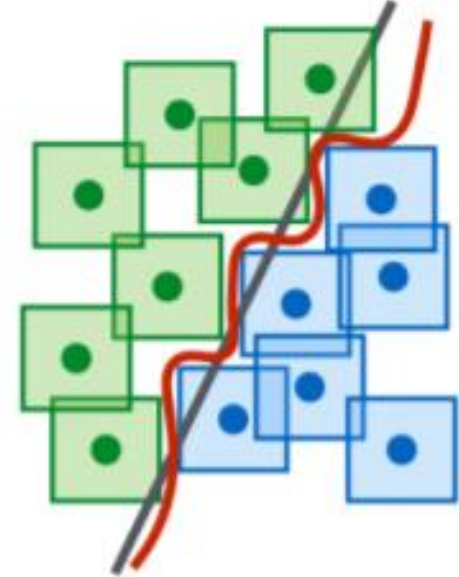
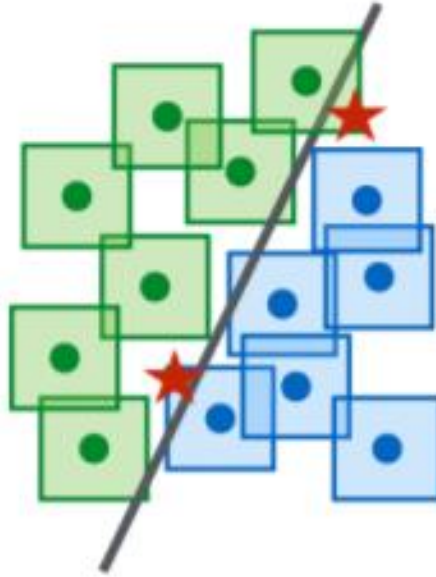


(f) ResNet-152





# DEFENSAS ROBUSTAS



# AMENAZAS DE SEGURIDAD E INTEGRIDAD

Graffiti



Speed Limit 45 sign



[Robust Physical-World Attacks on Deep Learning Visual Classification, Eykholt et al, 2018]

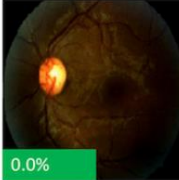

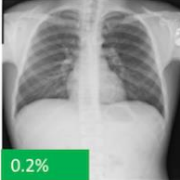
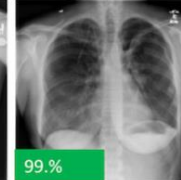
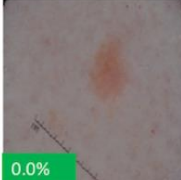
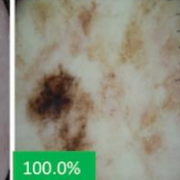

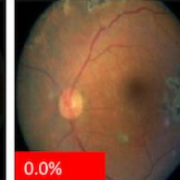


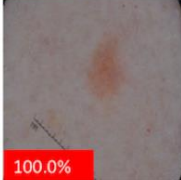
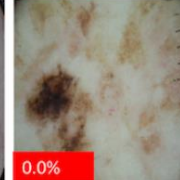
[Adversarial Patch, Brown et al, 2018]



Fundoscopy

Chest X-Ray

Dermoscopy

	Absent/mild DR	Moderate/Severe DR	Normal	Pneumothorax	Nevus	Melanoma
Clean	 0.0%	 100.0%	 0.2%	 99.9%	 0.0%	 100.0%
PGD	 100.0%	 0.0%	 100.0%	 0.0%	 100.0%	 0.0%

[Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems, Ma et al, 2019]



[Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-art Face Recognition, Sharif et al, 2016]

