

XII Jornadas STIC CCN-CERT

Ciberseguridad,
hacia una respuesta y disuasión efectiva



Separar las Churras de las Merinas

O cómo evaluar sistemas de seguridad
basados en modelos de *Machine Learning*



- Alberto Cita Sánchez
- Symantec Corporation
- alberto_cita@symantec.com



Índice

1. Machine Learning desmitificado
2. Criterios para distinguir entre churras y merinas
3. In-Seguridad en Machine Learning: Machine Learning con Adversarios
4. Contramedidas: Construyendo Modelos Robustos vs *Adversarial* ML
5. Conclusiones

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



1950's

1960's

1970's

1980's

1990's

2000's

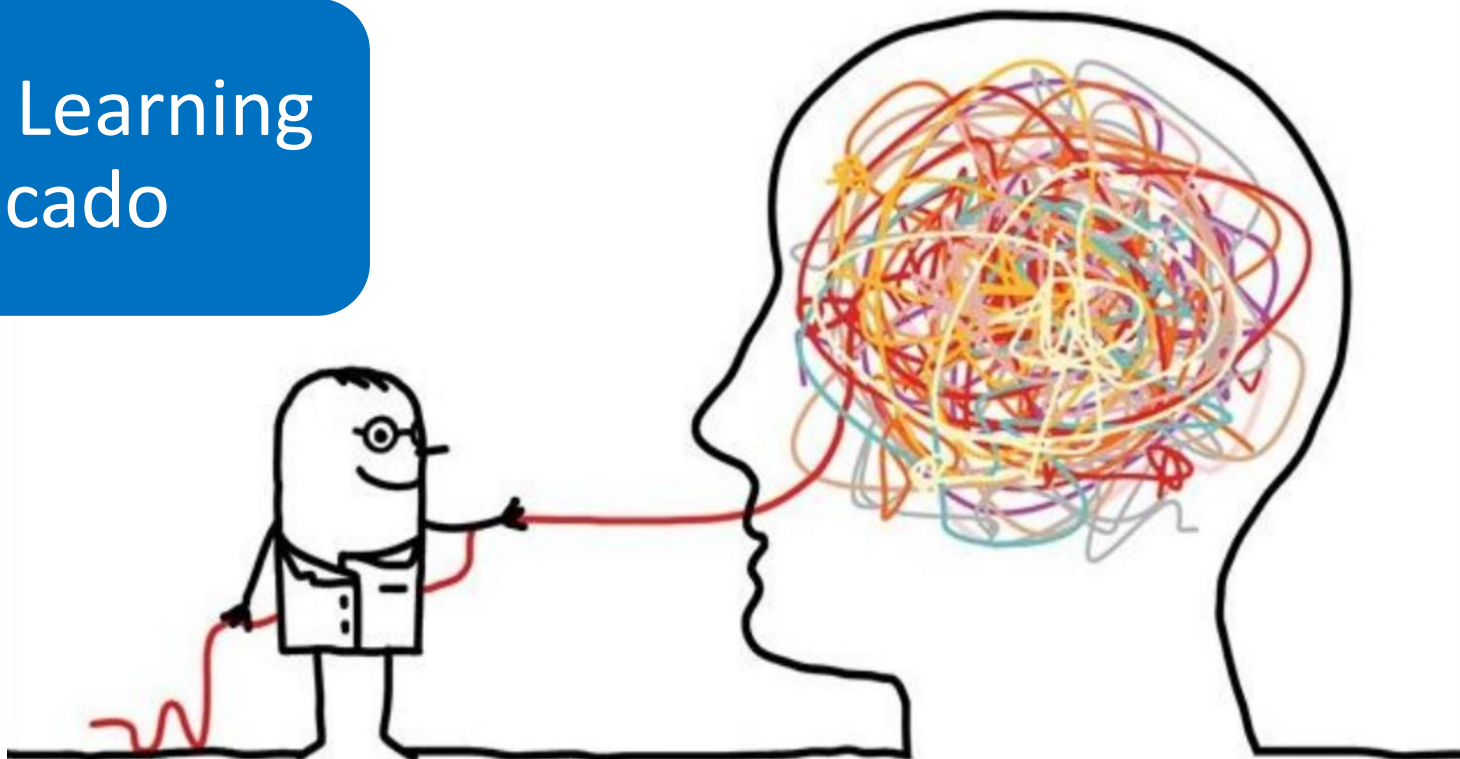
2010's

<https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

#XIIJornadasCCNCERT

www.ccn-cert.cni.es

Machine Learning Desmitificado

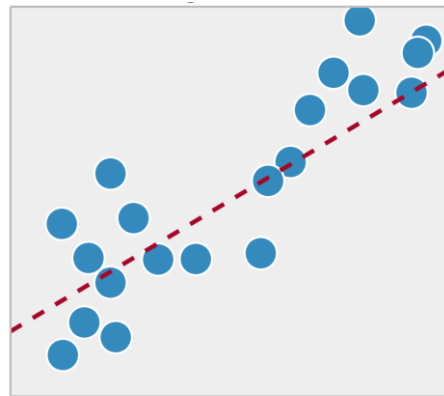


Machine Learning (ML)

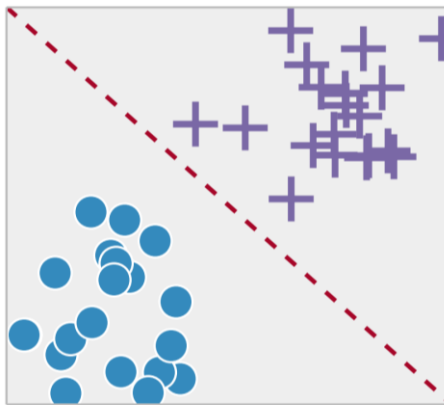
Definición

Disciplina científica que explora la construcción y estudio de **algoritmos que aprenden de los datos**

Regresión



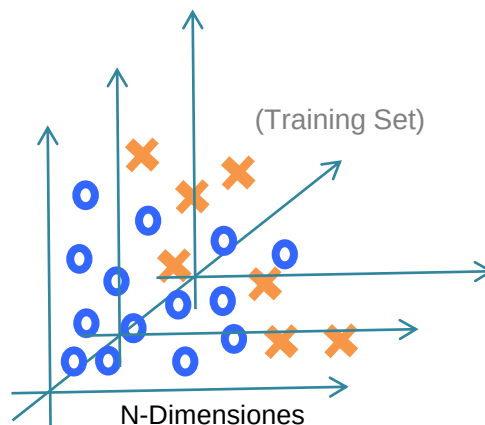
Clasificación



El proceso de Machine Learning

X_1	X_2	X_3	X_4	...	X_n	Y
a_1	a_2	a_3	a_4	...	a_n	Y_1
a_1	a_2	a_3	a_4	...	a_n	Y_2
a_1	a_2	a_3	a_4	...	a_n	Y_3
a_1	a_2	a_3	a_4	...	a_n	Y_4
a_1	a_2	a_3	a_4	...	a_n	Y_5
...	...	...	...	...	...	...
a_m	a_m	a_m	a_m	...	a_m	Y_m

Training Set
Muchos Datos Relevantes



Visualización,
Intuición,
Feature Engineering



Representación
del modelo:
“hipótesis”

$$h_{\theta}(x) = \theta_0 + X_1\theta_1 + \dots + X_n\theta_n$$

$$h_{\theta}(x) = \frac{1}{1 + e^{\theta^T x}}$$

Linear regression

Logistic regression

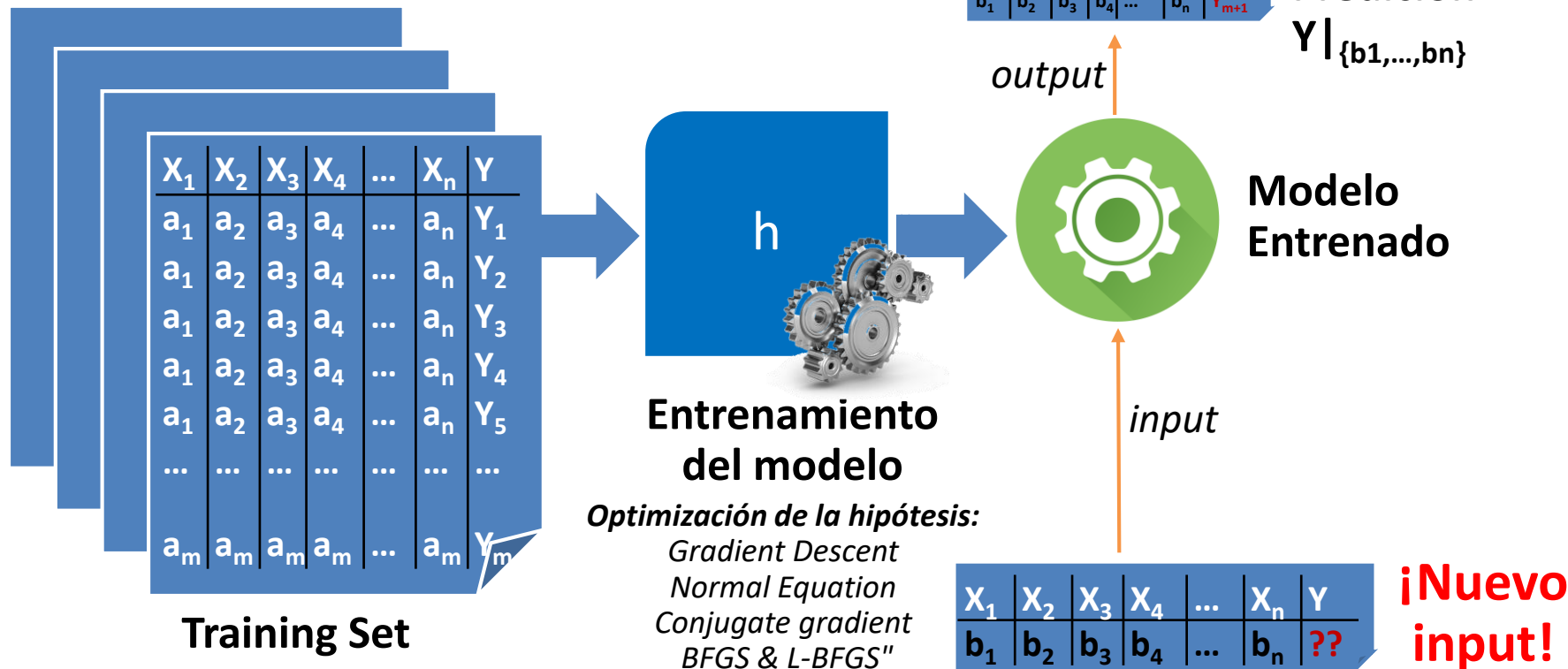
Multi-class classification

Neural Network

Decision Trees

Naïve Bayes Classification

El proceso de Machine Learning



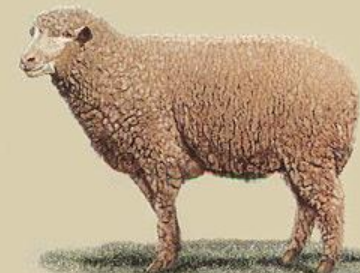


Criterios para
distinguir
entre...



Churra

ML Churro



Merina

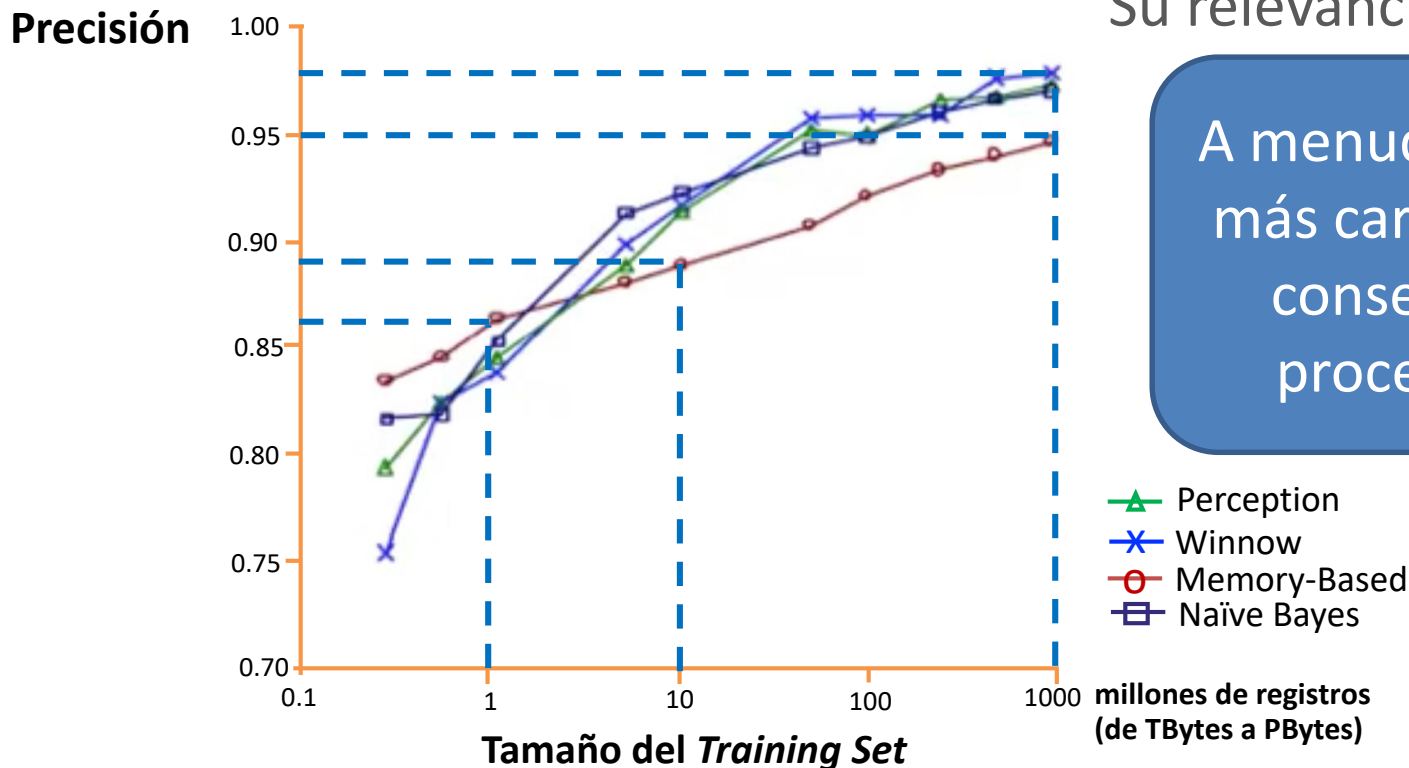
ML Merino



Training Set: el tamaño sí que importa

Su relevancia, también...

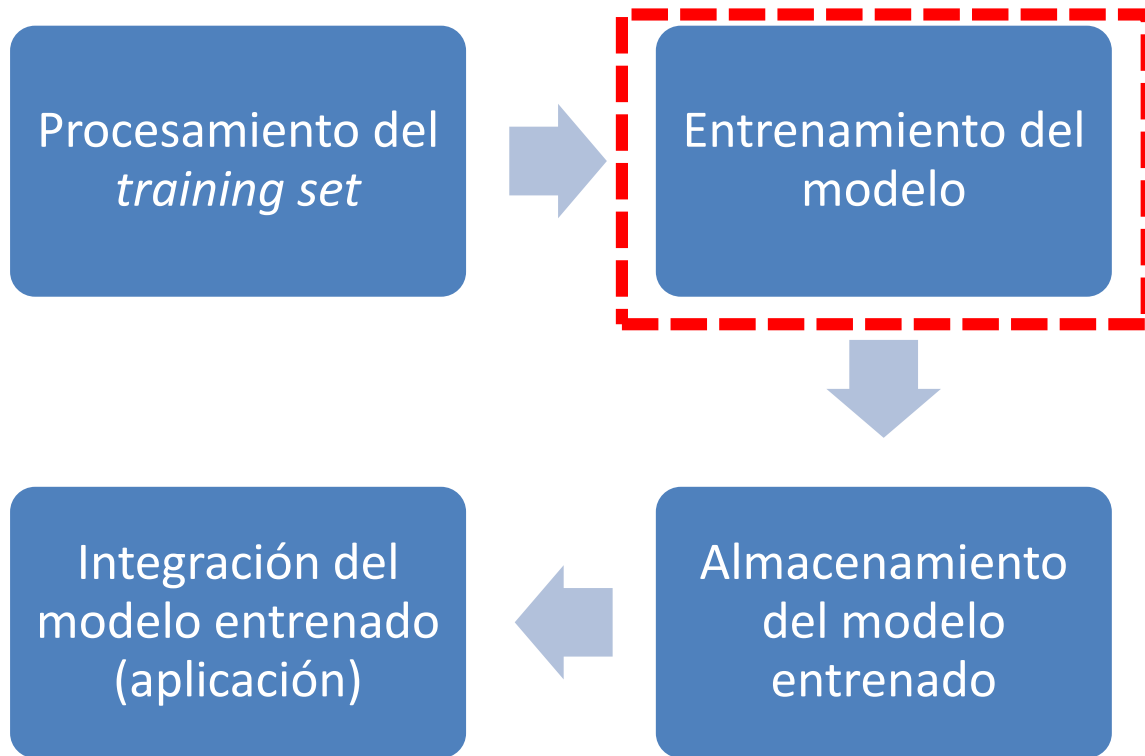
A menudo el recurso más caro y difícil de conseguir en el proceso de ML





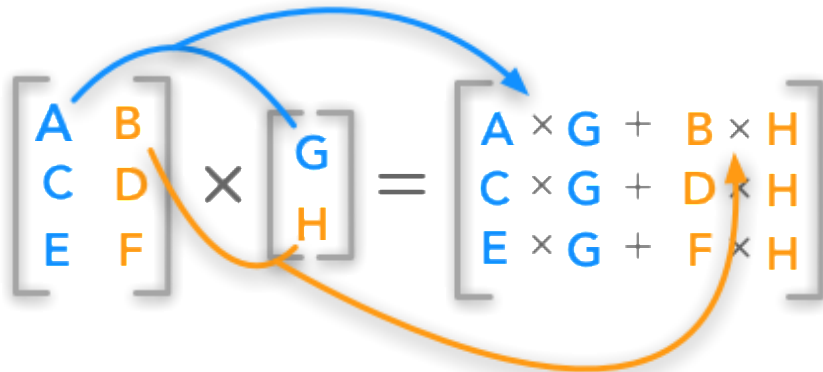
El proceso de Machine Learning

Conseguido el *training set*... ¿Qué fase del proceso de ML es la más costosa?



Entrenamiento del Modelo

Millones de Multiplicaciones de matrices N-dimensionales

$$\begin{bmatrix} A & B \\ C & D \\ E & F \end{bmatrix} \times \begin{bmatrix} G \\ H \end{bmatrix} = \begin{bmatrix} A \times G + B \times H \\ C \times G + D \times H \\ E \times G + F \times H \end{bmatrix}$$




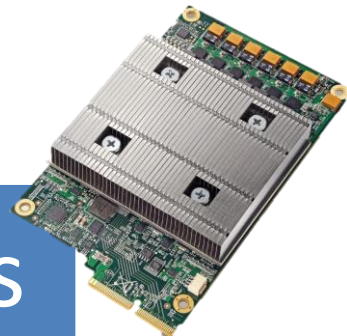
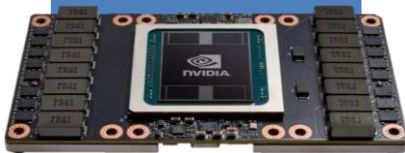
CPU_s



GPU_s



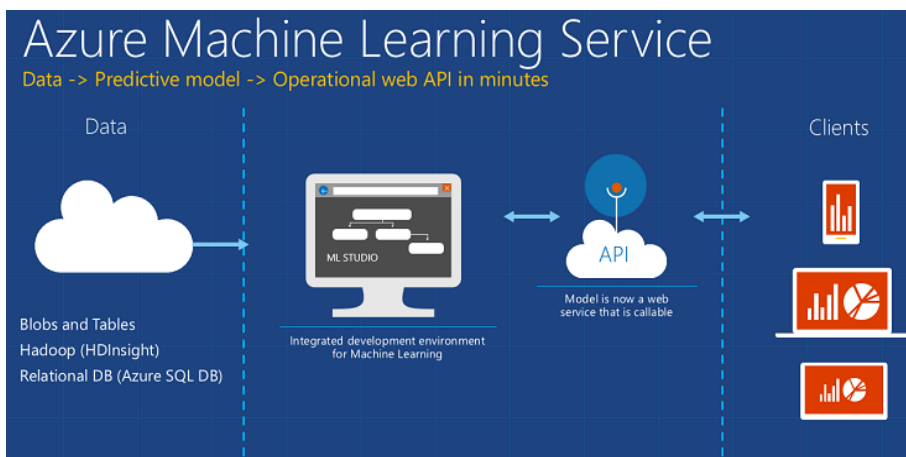
TPU_s



Democratización del *Machine Learning*

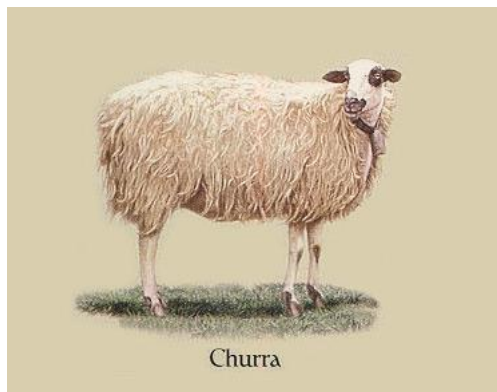
Machine Learning as a Service

- Almacenamiento, Preparación, Procesamiento y Visualización del *training set*
- Entrenamiento de modelos de ML
- Almacenamiento del modelo entrenado
- Despliegue del modelo (API)
- **Todo en pago por uso**

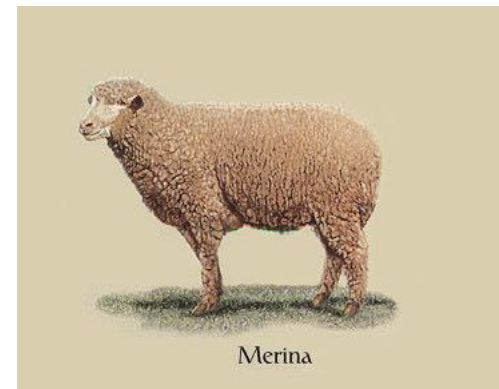


Machine Learning as a Service: pago por uso

Las cosas cambian... ¿con qué frecuencia re-entrenas tu modelo?



- ☐ Una vez a la semana
- ☐ Una vez al mes
- ☐ Una vez al año
- ☐ Raramente
- ☐ Nunca





“Machine learning has not yet reached true human-level performance, because when confronted by even a trivial adversary, most machine learning algorithms fail dramatically”

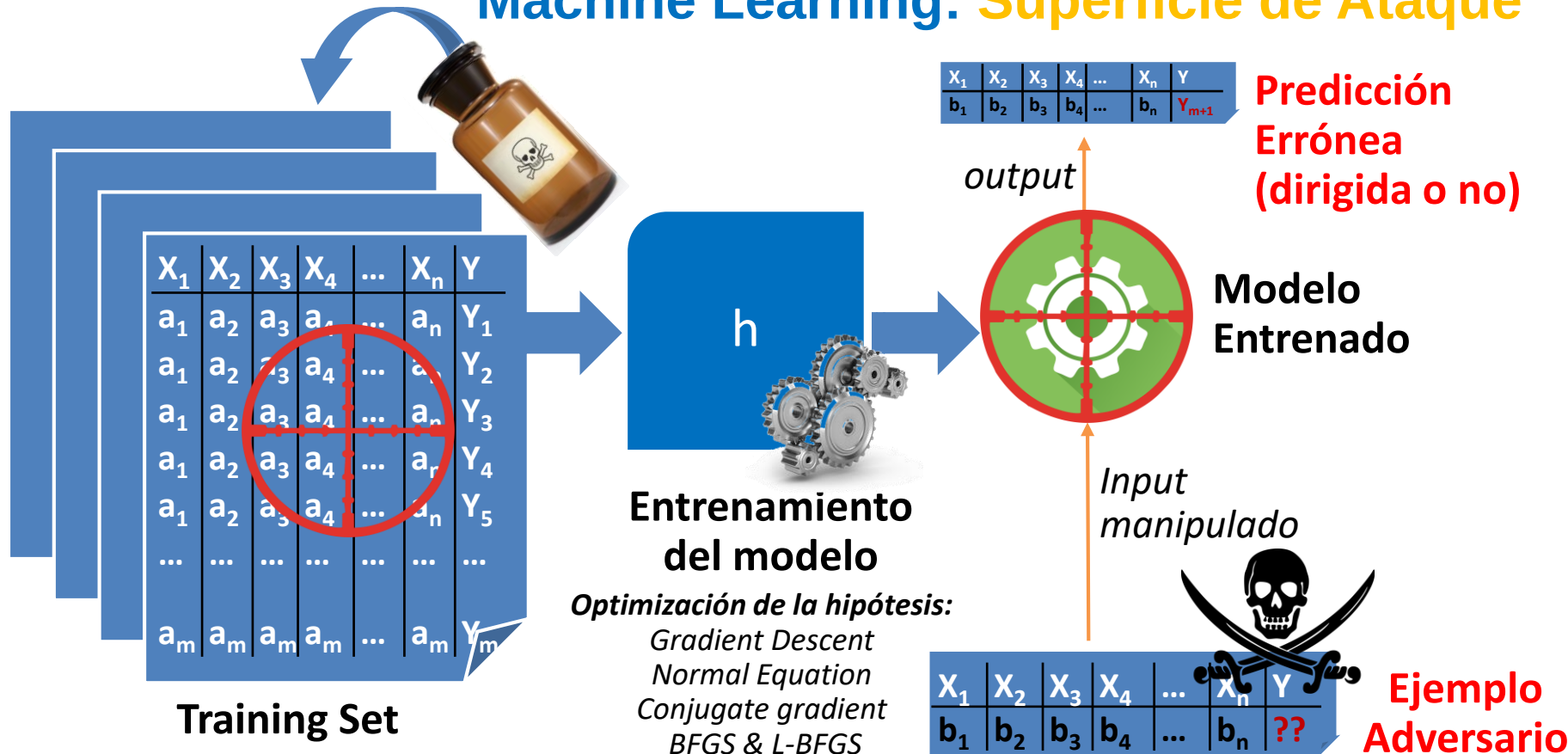
Nicolas Papernot & Ian Goodfellow
15 diciembre de 2016



Machine Learning con **Adversarios**

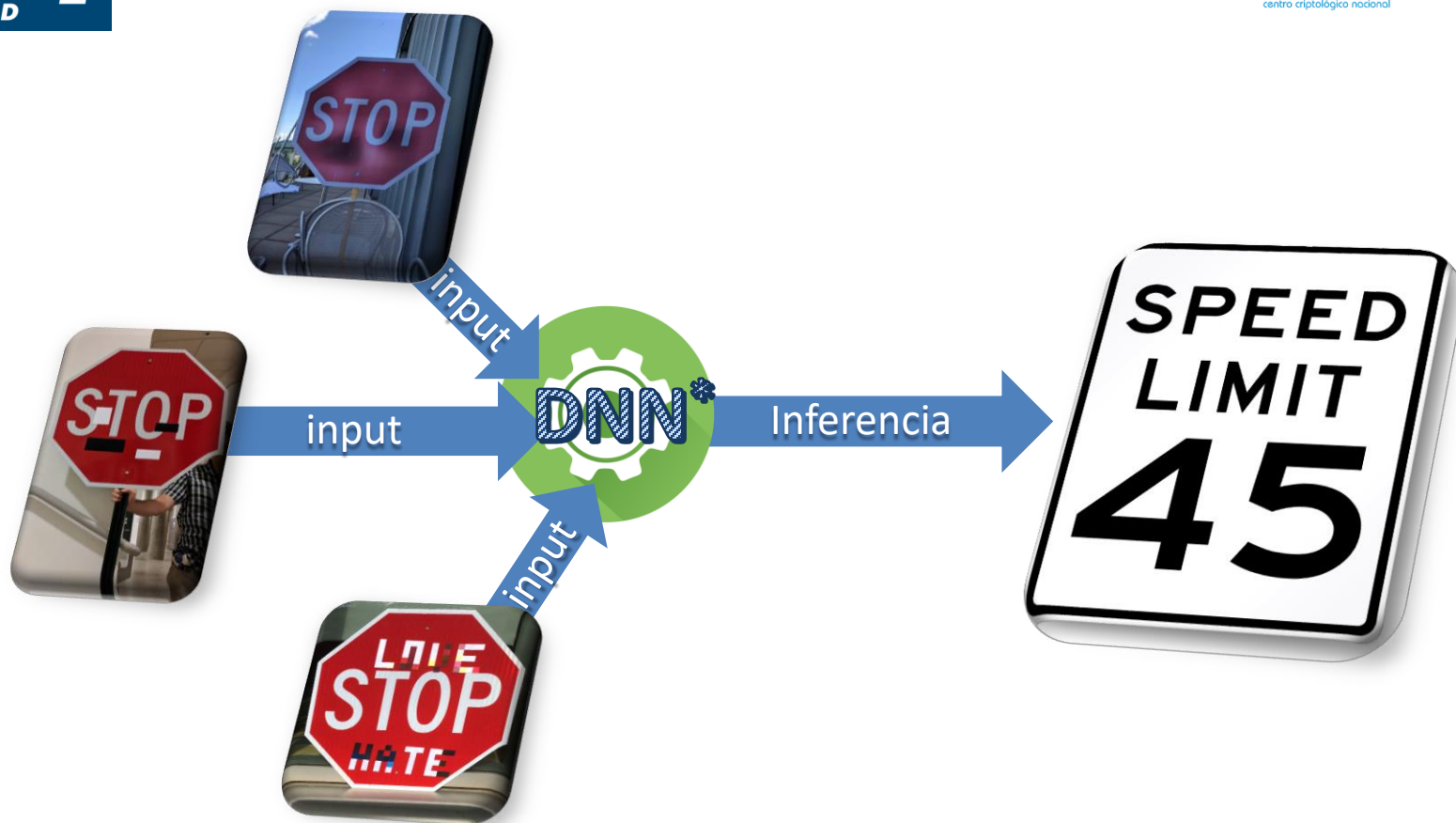
Romper cosas es
“sencillo”

Machine Learning: Superficie de Ataque









**Deep Neural Network*

Robust Physical Perturbations (RP₂) Attack



Ejemplos Adversarios

Adversarial Machine Learning

¹Deep Neural Network



NO Need to Worry about Adversarial Examples

Jiajun Lu, Hussein Sibai, Evan Fabry, David Forsyth

(Submitted on 12 Jul 2017)

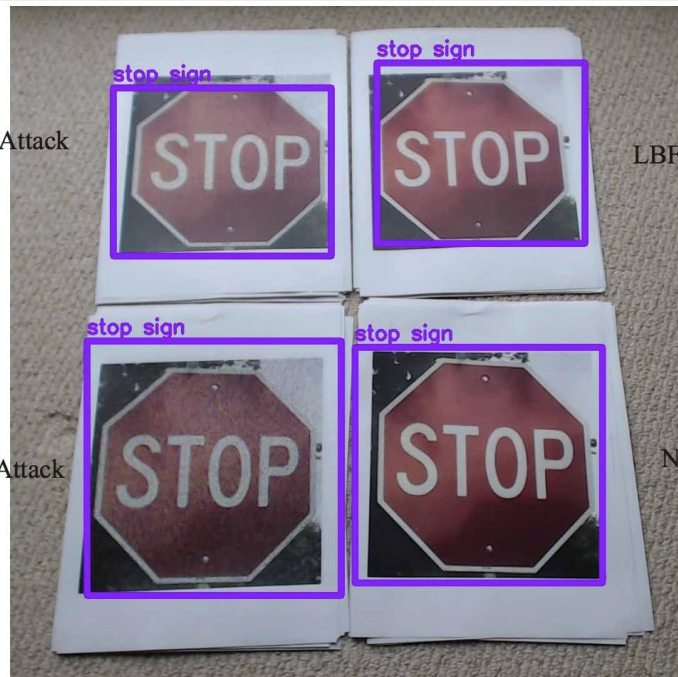
It has been shown that most machine learning algorithms are susceptible to adversarial examples. In the image space, adding a carefully chosen perturbation to a legitimate image may cause a trained neural network model to misclassify it. Recent experiments then taking pictures of them would still result in misclassification. This raises the question of whether these experiments ignore a crucial property of physical objects: the camera's perspective. However, these experiments ignore a crucial property of physical objects: the camera's perspective. In this paper, we show experiments that suggest that current constructions of physical adversarial examples do not work. A trained neural network classifies most of the pictures taken from different distances.

FastSign Attack

LBFGS Attack

Iterative Attack

No Attack





Robust Adversarial Examples

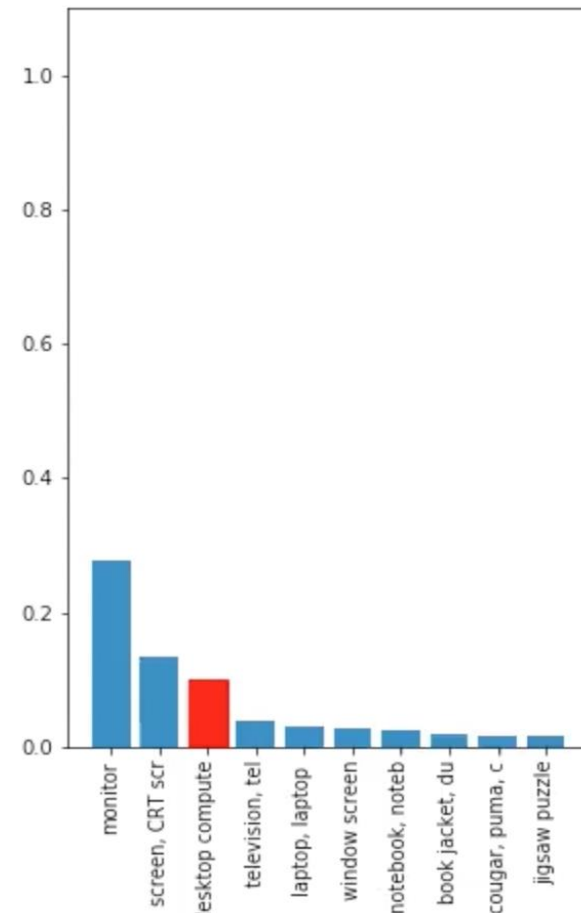
JULY 17, 2017

We've created images that reliably fool neural network classifiers when viewed from varied scales and perspectives. This challenges a [claim](#) from last week that self-driving cars would be hard to trick maliciously since they capture images from multiple scales, angles, perspectives, and the like.

Robust Adversarial Examples

JULY 17, 2017

We've created images that reliably fool neural network classifiers when viewed from varied scales and perspectives. This challenges a [claim](#) from last week that self-driving cars would be hard to trick maliciously since they capture images from multiple scales, angles, perspectives, and the like.



Frontera de
decisión real

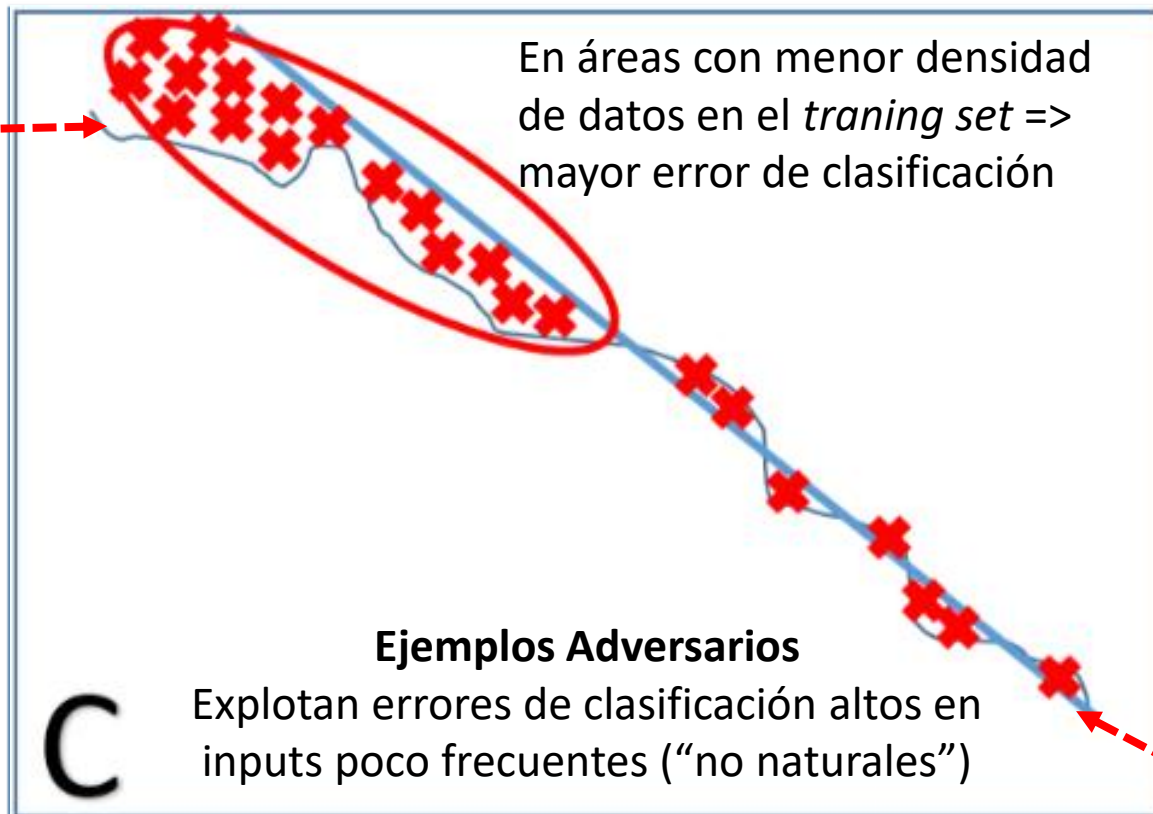
En áreas con menor densidad
de datos en el *training set* =>
mayor error de clasificación

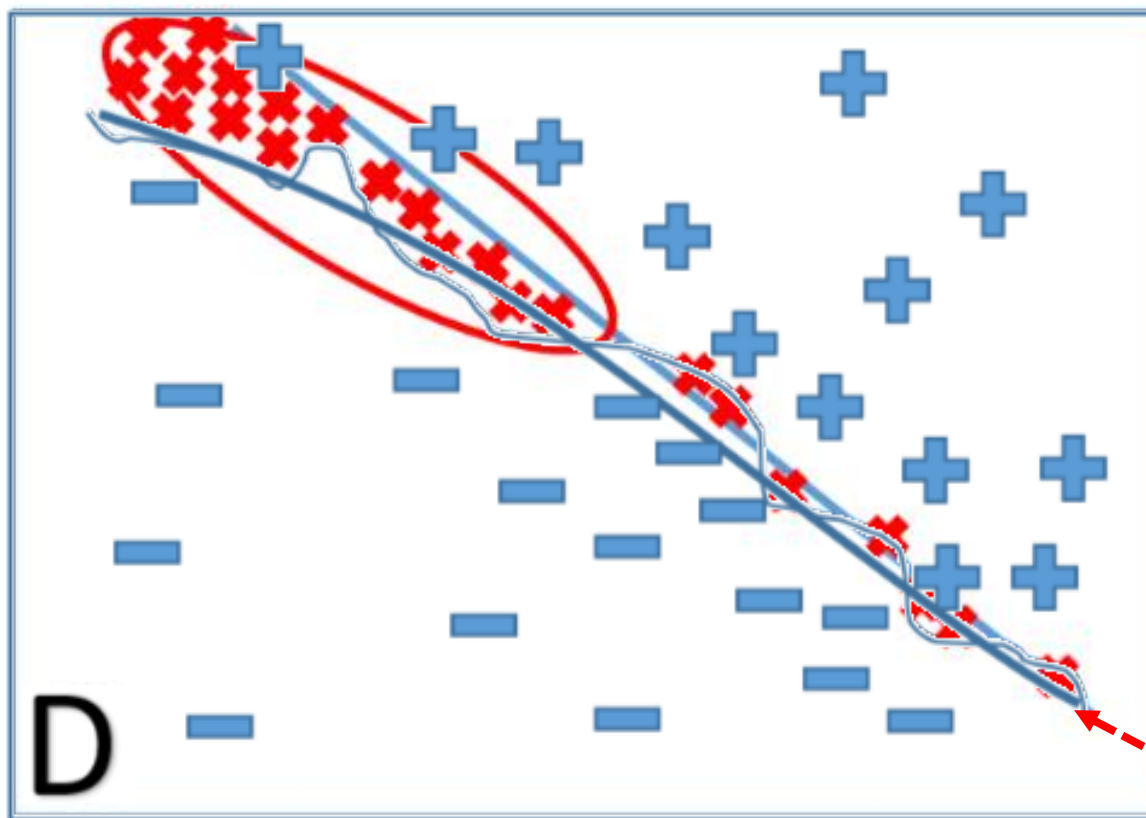
C

Ejemplos Adversarios

Explotan errores de clasificación altos en
inputs poco frecuentes (“no naturales”)

Modelo ML
Entrenado:
Clasificador
Lineal



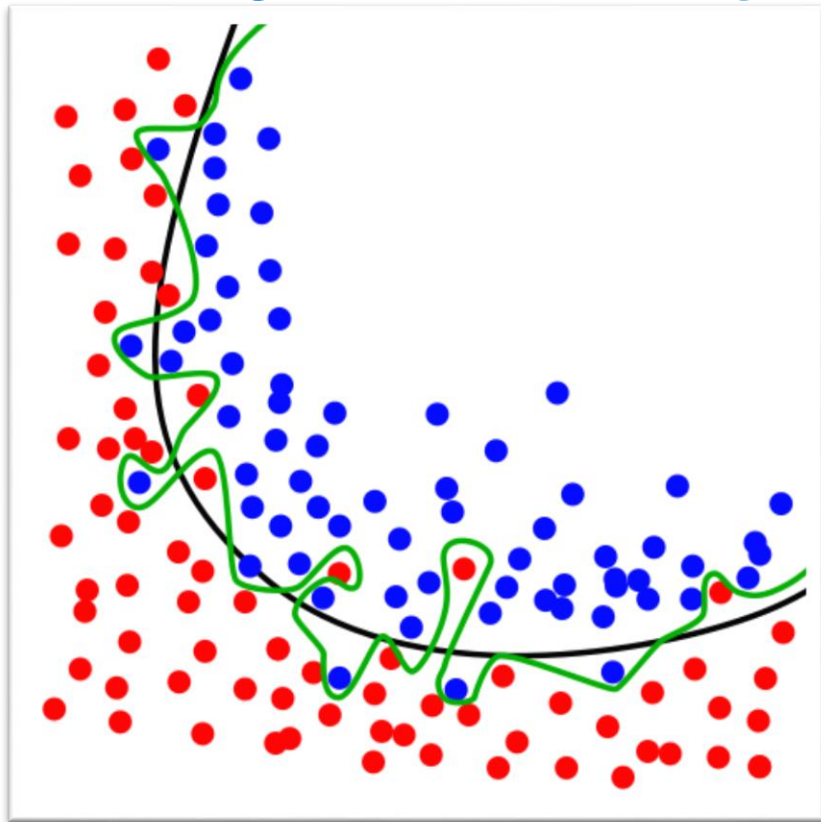


Modelo ML
Entrenado:
Clasificador
NO Lineal

El riesgo del *Overfitting*

El modelo sobre reacciona a cambios en el training set

Para un tamaño de ***training set*** dado, un modelo más complejo (y más robusto vs *Adversary* ML) tendrá **menor precisión**



Adversarial ML: ML aplicado al cálculo de Ejemplos Adversarios

White Box



- **Adversario tiene información** sobre el modelo o el training set
 - *h, parámetros del modelo, parte del training set,...*
- **Realista**: ingeniería inversa de modelos de ML que se comprimen y se despliegan en *smartphones*.
- Los **ejemplos adversarios se calculan vía ML** usando parámetros conocidos del modelo.
- **89.4 % < EFICACIA < 97%**

Black Box



- Adversario **explora el modelo** con inputs creados cuidadosamente y observando los outputs.
- Nº de *queries* $\sim$ tamaño del training set
- Especialmente aplicable en **MLaaS**
- Reconstrucción del modelo y cálculo de ejemplos adversarios
- **Transferibilidad de ejemplos adversarios**: los calculados para un modelo de ML pueden servir para otro, incluso si entrenado usando otro *dataset*
- **84.24% < EFICACIA < 96%**



Adversary ML vs Reconocimiento Facial basado en DNN (ejemplo visual)

Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Mahmood Sharif
Carnegie Mellon University
Pittsburgh, PA, USA
mahmoods@cmu.edu

Sruti Bhagavatula
Carnegie Mellon University
Pittsburgh, PA, USA
srutib@cmu.edu

Lujo Bauer
Carnegie Mellon University
Pittsburgh, PA, USA
lbauer@cmu.edu

Michael K. Reiter
University of North Carolina
Chapel Hill, NC, USA
reiter@cs.unc.edu

ABSTRACT

Machine learning is enabling a myriad innovations, including

ogy and ML (sometime trusting them even with their own
lives [36]). The domains that received the majority of the at-
tention have a common characteristic: the adversary is able

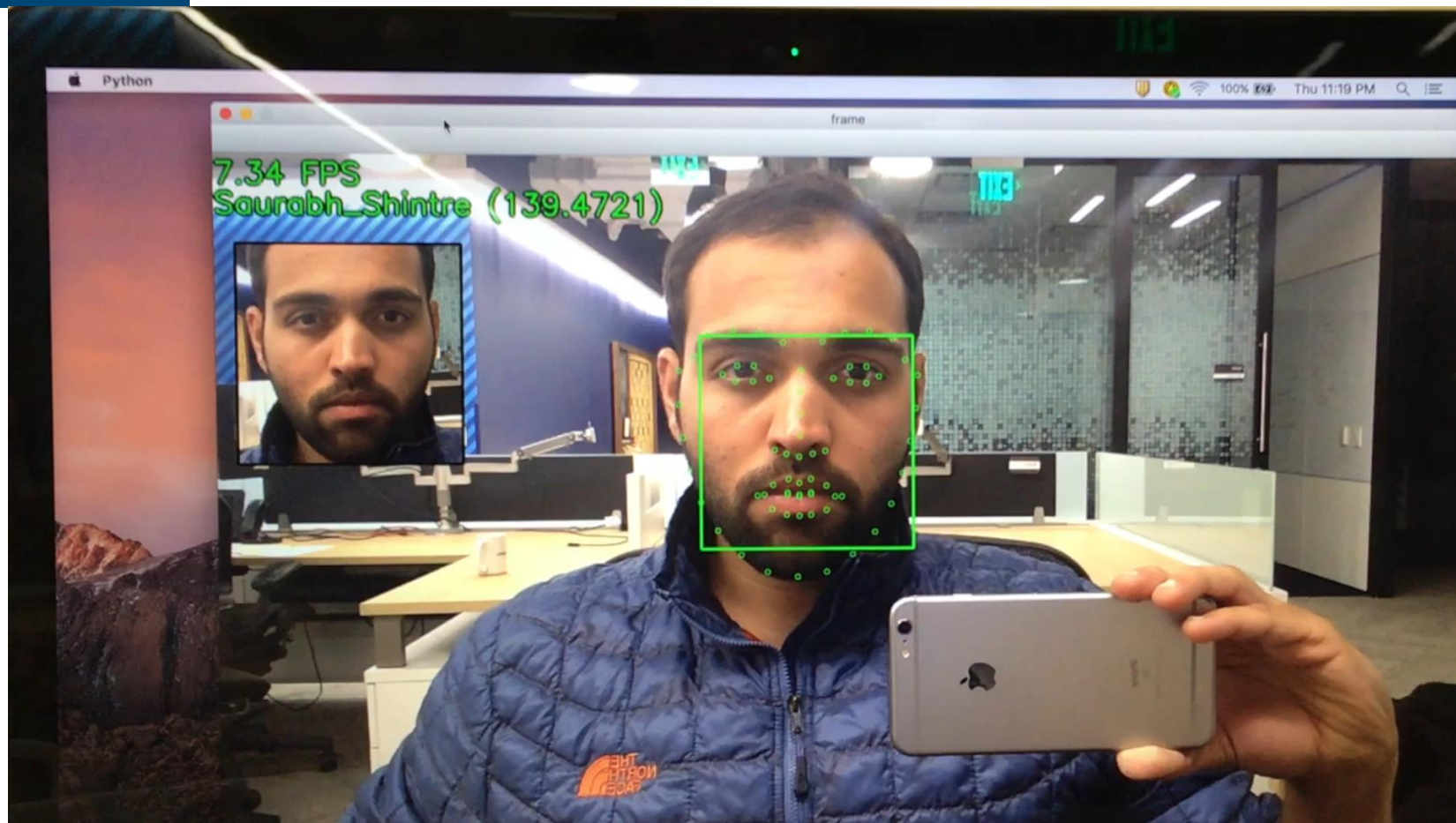
Permítanme presentarles...



Dr. Hugh Thomson
CTO Symantec



Dr. Saurabh Shintre
Senior Principal Research Engineer
Symantec Research Labs





¿Es posible crear una Muestra Adversaria de Malware?

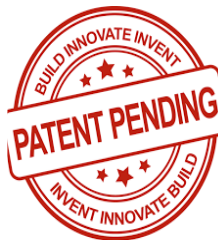


Verano 2018: proyecto en **Symantec Research Labs** (SRL) para bypassear clasificadores de malware de Symantec basados en Deep Learning

- Sin conocer los detalles internos del clasificador
- **Manteniendo la funcionalidad original del binario**
- **Ataque nunca visto en campo**

Proyecto concluido con éxito

- SRL ha desarrollado una técnica basada en Adversarial ML para generar binarios maliciosos que logran esos objetivos
- **Aproximación basada en algoritmos genéticos/evolutivos**



SRL ha **patentado** este enfoque y se ha **aplicado a nuestros propios clasificadores** para incrementar su robustez frente a muestras adversarias.

CONTRAMEDIDAS: Construyendo Modelos Robustos vs *Adversarial* ML



! CAUTION

**WORK IN
PROGRESS**



Contramedidas – Opción 1: Modelos de ML Robustos

Mejorar la Precisión de los Algoritmos ML

- Un algoritmo perfecto no puede ser manipulado
- **NO ESCALA:** incrementar sustancialmente el tamaño del training set para minimizar riesgo de *Overfitting*

Adversarial Training

- **Calcular ejemplos adversarios**, añadirlos al training set original y re-entrenar el modelo.
- Reducción del riesgo de clasificación errónea de ejemplos adversarios:
89.4% => 17.9%
- **NO ESCALA:** Calcular ejemplos adversarios producidos por todos los ataques posibles



Contramedidas – Opción II: Detección Estadística de Muestras Adversarias

Detectar diferencias en propiedades estadísticas entre muestras adversarias y no-adversarias

La mayoría de los detectores adversarios formulados, son inútiles si el adversario conoce la defensa

Gradient Similarity*

*SRL & Symantec Center for Advanced Machine Learning
(Jun/2018)*

**Basado en
"Funciones
de
Influencia"**

**Adversario
White-box
& Black-Box**

**Resistente a
ataques
adaptativos**

**No re-
entrenamiento
del modelo**

**Precisión
Detección
87%-97%**

(*) Gradient Similarity: An Explainable Approach to Detect Adversarial Attacks against Deep Learning. <https://arxiv.org/pdf/1806.10707.pdf>
Jasjeet Dhaliwal Center for Advanced Machine Learning Symantec Corporation, Saurabh Shintre Symantec Research Labs Symantec Corporation



Conclusiones

La etiqueta "AI" / "Machine Learning" no es garantía de mejora en Ciberseguridad

- Sistemas de ciberseguridad en *Machine Learning* pueden presentar grandes diferencias de precisión y robustez, incluso si los algoritmos empleados son los mismos

Obstáculo importante para la adopción de ML en campos críticos, como la ciberseguridad

- Generalizada vulnerabilidad a ataques basados en *Adversarial ML* (especialmente en DNN*) + falta de defensas efectivas contra estos ataques. **Seguridad en ML: campo en pleno desarrollo.**

(*) DNN = Deep Neural Networks

XII Jornadas STIC CCN-CERT

Ciberseguridad,
hacia una respuesta y disuasión efectiva



➤ E-Mails

- info@ccn-cert.cni.es
- ccn@cni.es
- organismo.certificacion@cni.es

Websites

- www.ccn.cni.es
- www.ccn-cert.cni.es
- oc.ccn.cni.es

➤ Síguenos en

