

# Security of Neural Networks : Attacks, Defenses and Evaluation methods

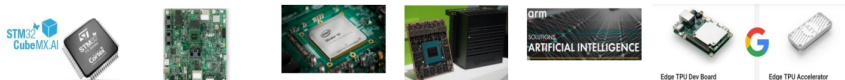
Rémi BERNHARD (CEA Tech)  
Pierre-Alain MOELLIC (CEA Tech)  
Jean-Max DUTERTRE (MSE)

*Laboratoire de Sécurité des Architectures et des Systèmes,*  
Centre CMP, Equipe Commune CEA-Tech Mines Saint-Etienne,  
F-13541 Gardanne France

June 12, 2019



- **Neural networks:** state-of-the art performances in various complex tasks (e.g., image recognition, speech translation)  
→ Growing will to deploy models on embedded systems



- **Adversarial machine learning:** Serious threats require efficient countermeasures
  - Critical decision systems (health, defense and security, ...)
  - Autonomous car
  - ...

# Security of Machine Learning systems

## Threat Model

### EXTRACT INFORMATION

**Training data** (medical, financial, biometric, classified...)

**Model** (IP, limited authorization)



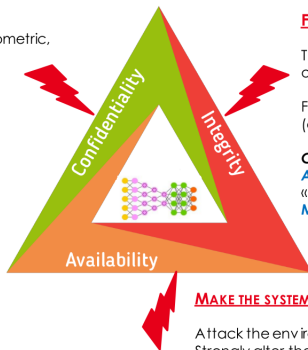
### ADVERSARY'S KNOWLEDGE / CAPACITY

Attack at learning / inference time ?

What knowledge about the model ?

→ White box / Black box paradigm

Probing / Querying the model



### FOOL A MODEL

The output prediction is not the expected one (i.e. correctly learned)

Fool a model *under the radar*, i.e. in a (almost) imperceptible way

### **Critical cases:**

**Autonomous vehicle** « Stop » recognizes as « 130 km/h » sign.

**Malware detection**

### MAKE THE SYSTEM USELESS

Attack the environment (e.g. classical DoS)  
Strongly alter the performance of the model

Figure: CIA threat model for a Machine Learning system

# Security of Machine Learning Systems

## Striking the ML pipeline

