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

Context

Overview

- **Neural networks**: state-of-the art performances in various complex tasks (e.g., image recognition, speech translation)
  $\rightarrow$ 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

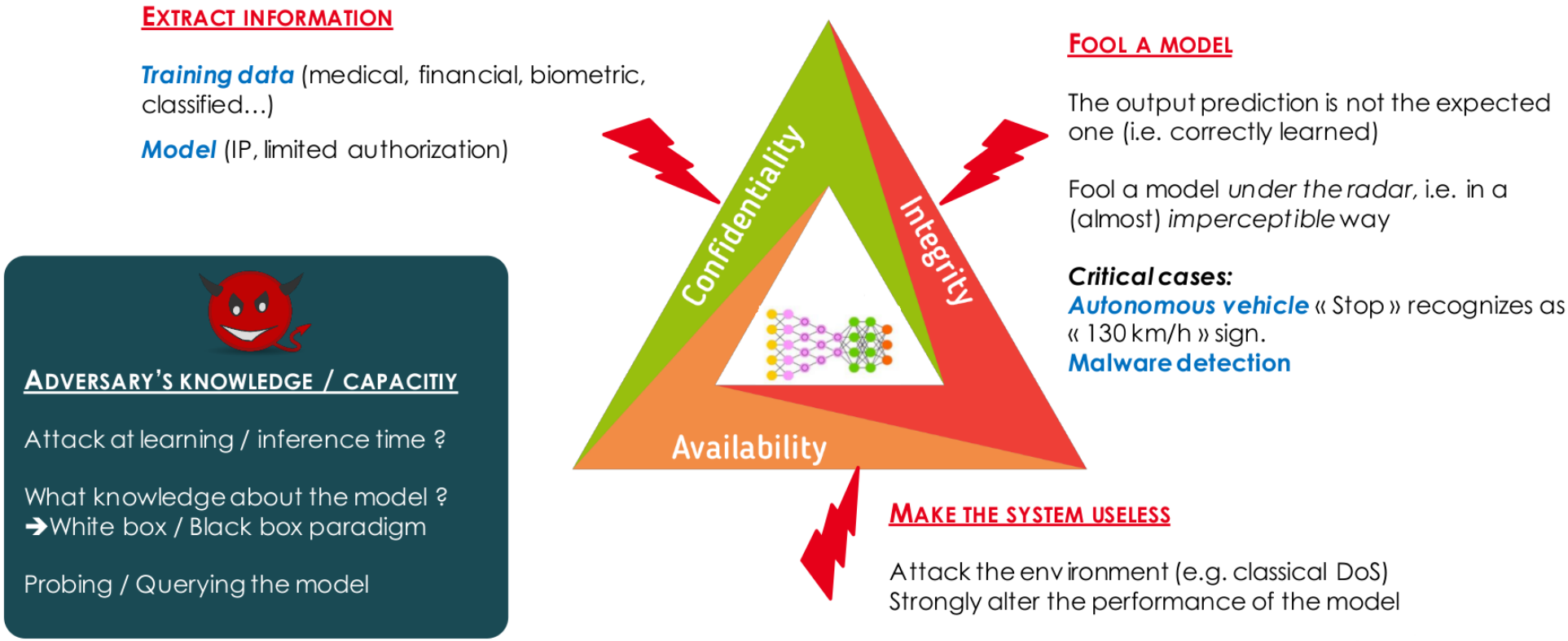
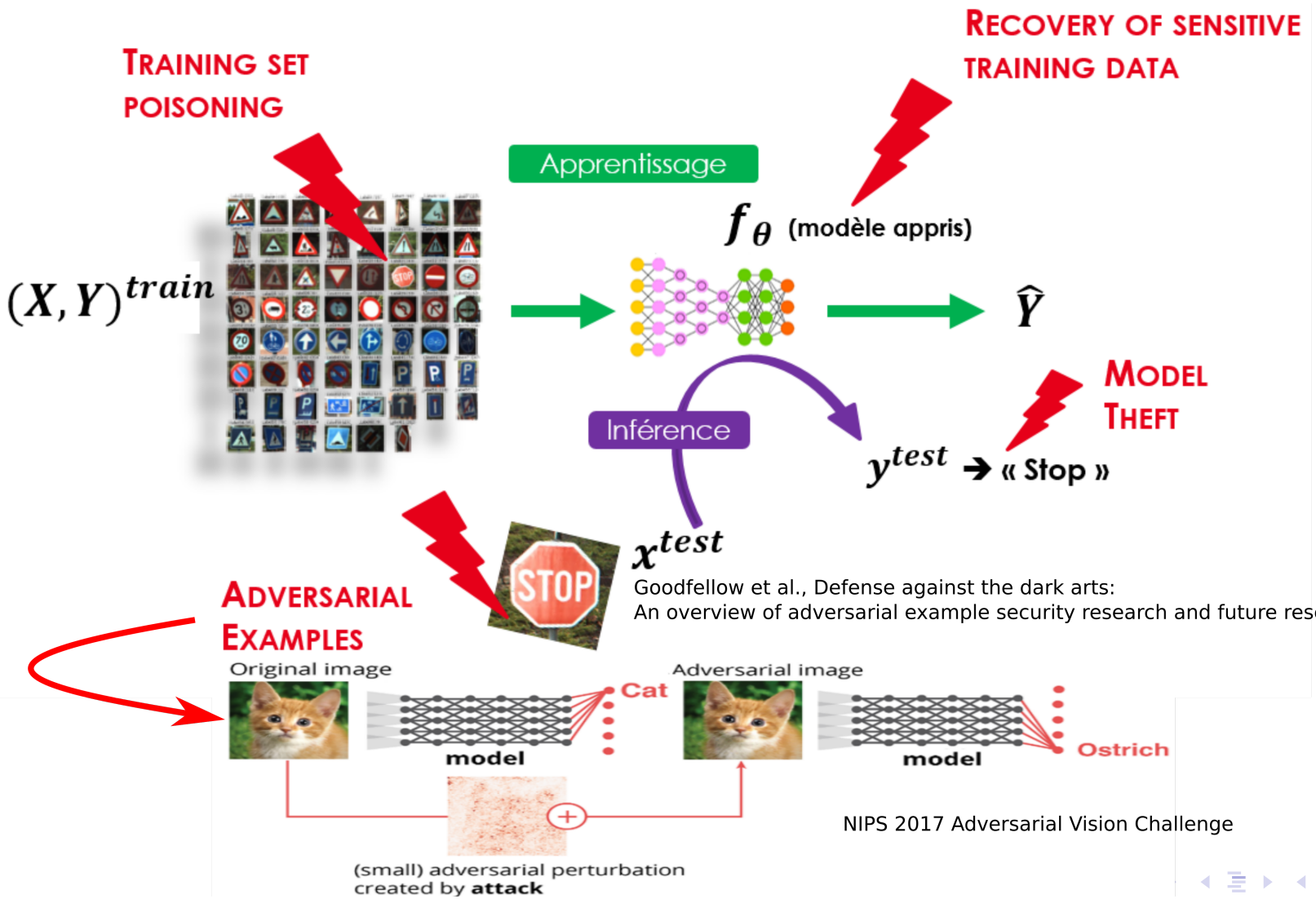


Figure: CIA threat model for a Machine Learning system

Security of Machine Learning Systems

Striking the ML pipeline



Goodfellow et al., Defense against the dark arts:
An overview of adversarial example security research and future research directions, 2018

NIPS 2017 Adversarial Vision Challenge