SSF CLAS
Cyber-security for learning and control systems

Alexandre Proutiere
KTH Royal institute of Technology

Machine-Learning based systems under attack

1. Assess the vulnerability of ML-based systems

2. Detect attacks, and devise secure ML algorithms

3. Illustrate the concepts in a smart building testbed

K.H. Johansson

A. Proutiere

G. Dan

H. Sandberg

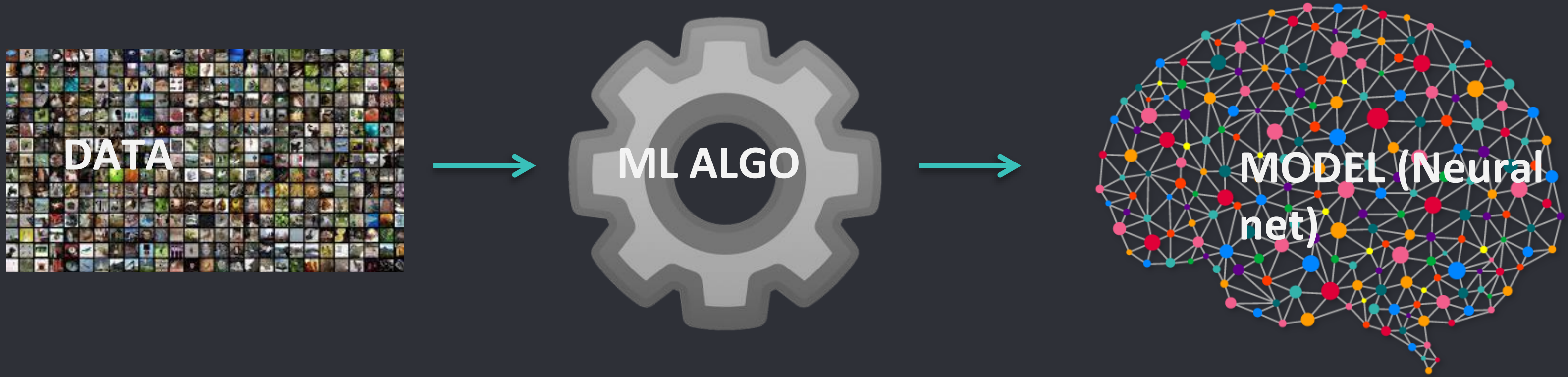
M. Molinari

V. Ctekovic

KTH
VETENSKAP
OCH KONST

Machine Learning under attack?

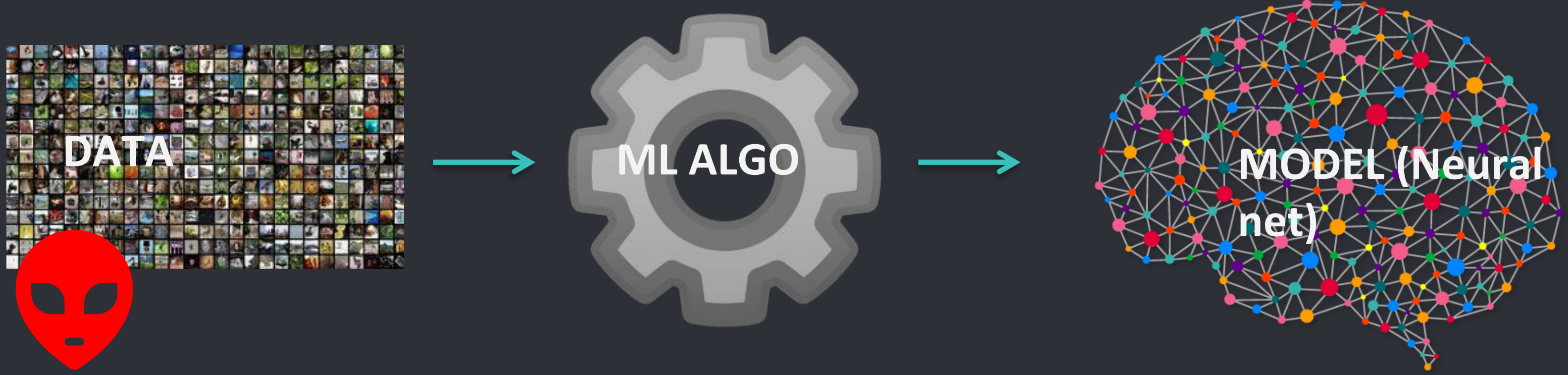
At training time



DATA → ML ALGO → MODEL (Neural net)

Machine Learning under attack?

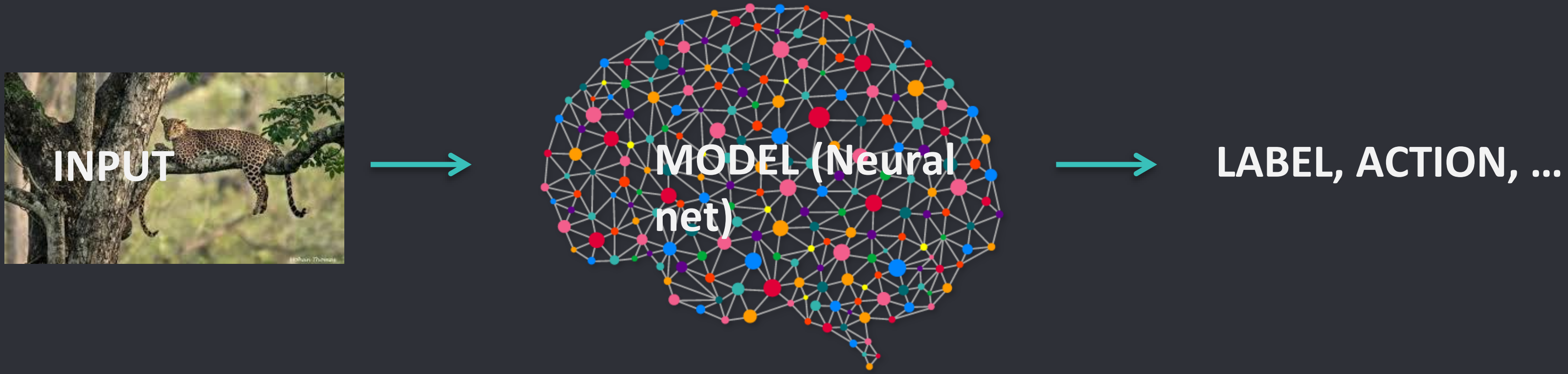
At training time



Slightly modify the data in an ***adversarial*** manner

Machine Learning under attack?

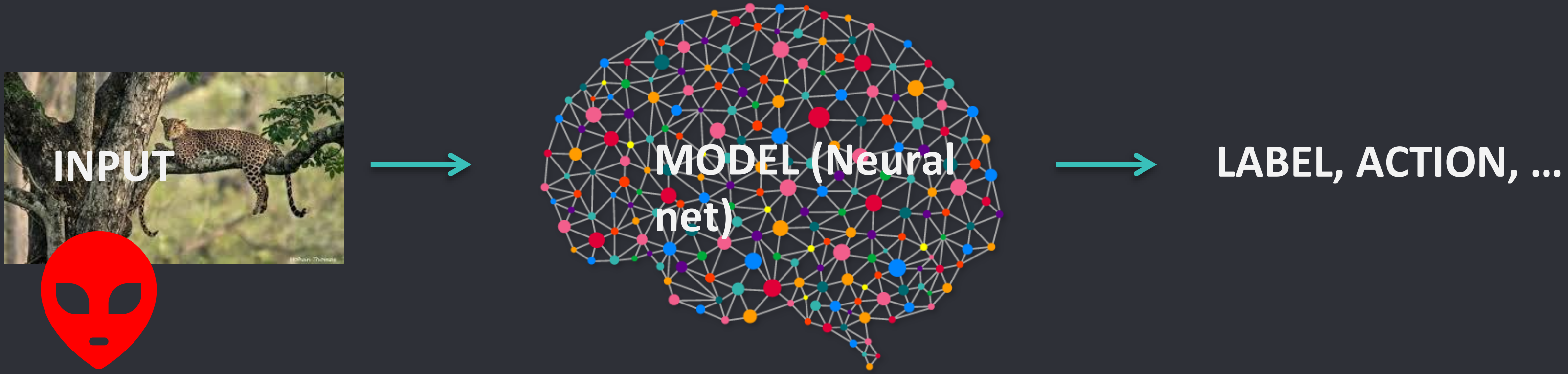
At test time



INPUT → MODEL (Neural net) → LABEL, ACTION, …

Machine Learning under attack?

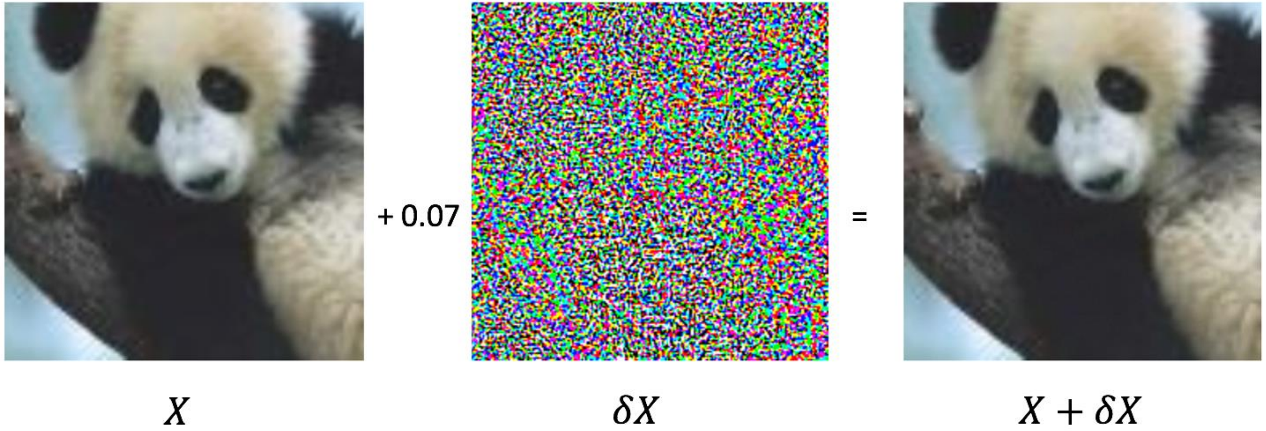
At test time



INPUT → MODEL (Neural net) → LABEL, ACTION, …

Slightly modify the input to the model in an ***adversarial*** manner

A first alarm at Google: Goodfellow et al. **ICML** 2015



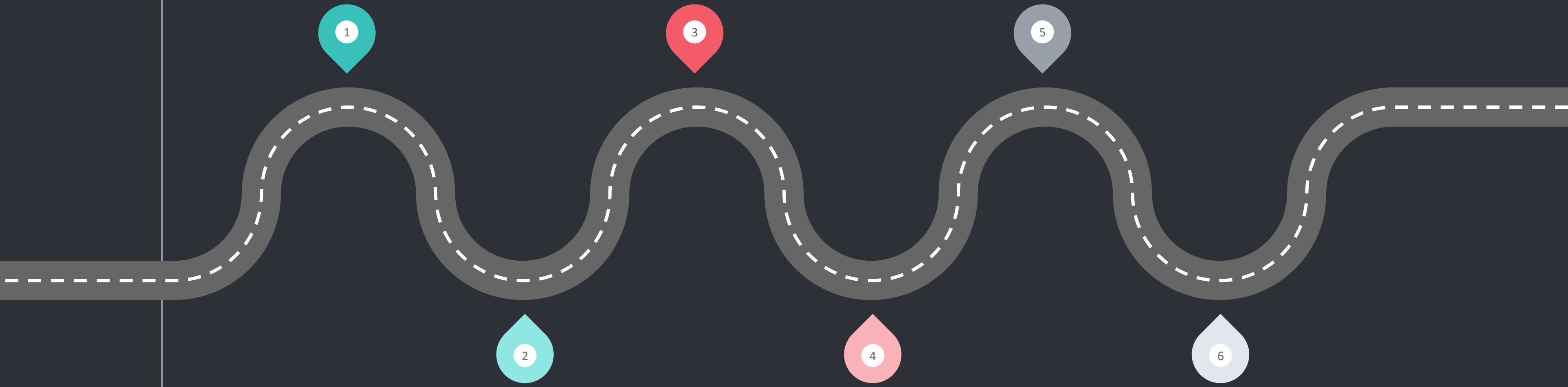
True input classified as a “Panda” with 57% confidence

Imperceptible adversarial “noise”

Adversarial example classified as a “Gibbon” with 99% confidence

A heuristic test-time attack (not optimized), and w/o detection consideration

CLAS expected results in ML-based controlled systems



Optimal attacks

Optimal detection algo

Risk management

Privacy attack

KTH live-in lab

?

RESULTS SO FAR

A generic optimization framework for attack design / detection

Example: attacking at test-time a control policy obtained through deep RL

1. Maximal detection rate of an attack $:\pi$ $\mathbb{P}[\text{detect}] = e^{-I(\pi)}$
2. Optimal attack :

$$\min_{\pi} R(\pi) \ s.t. \ I(\pi) \geq \gamma$$

Atari games (here pong)

Game wihout adversary

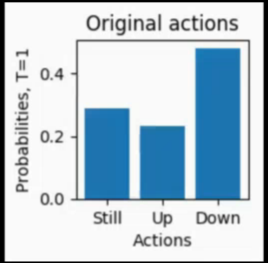
Main Agent Score	2nd Agent Score
0	0
0	0

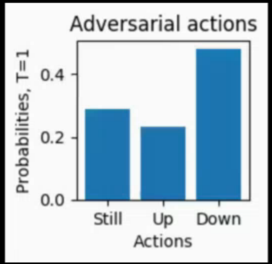
Original Frame

Change

=

Final image

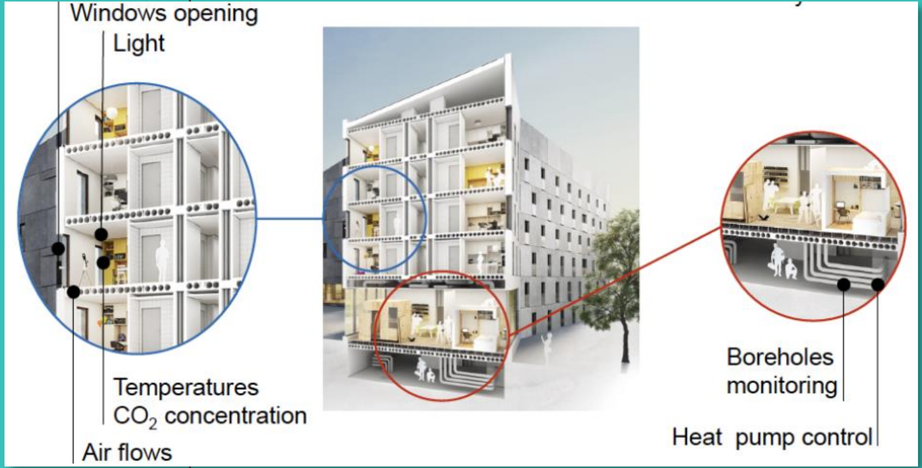


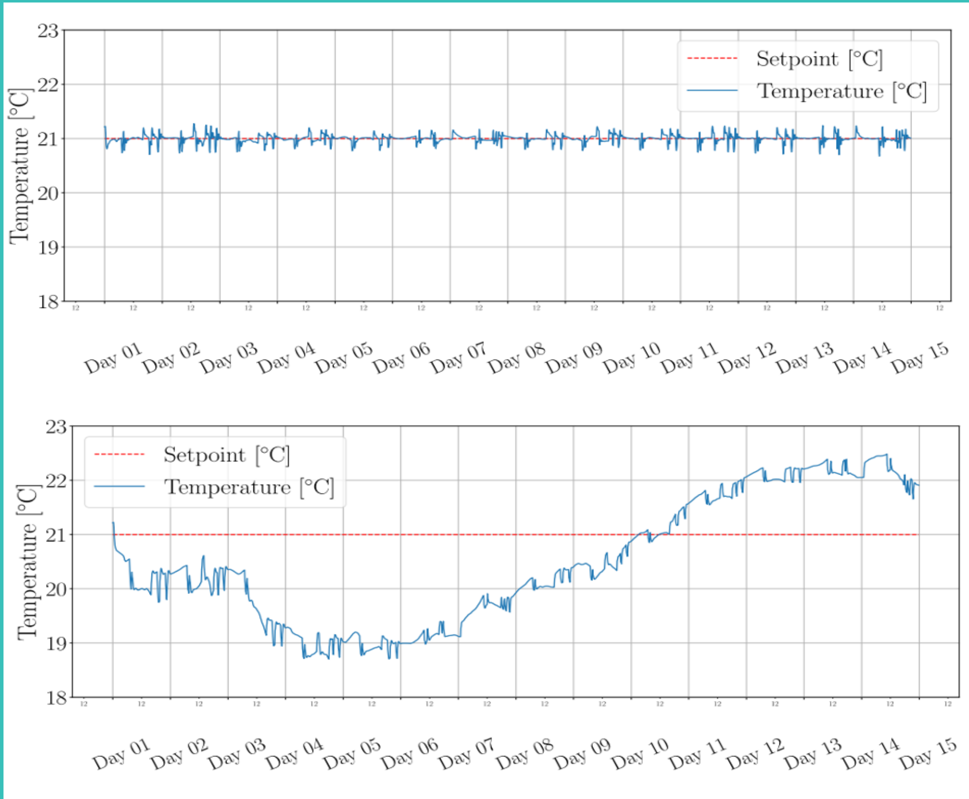


Other results / current activities

- Optimization framework applied to
  - Various types of attacks
  - Multi-sensor systems
  - Attacks at training time …
- Securing ML-algorithms
- +50 academic publications

Demonstrator: **KTH Live-in lab** (privacy and security)





Desired collaborations

We can assess the **security-level** of ML algorithms and models

… anyone running ML-based systems is welcome to contact us