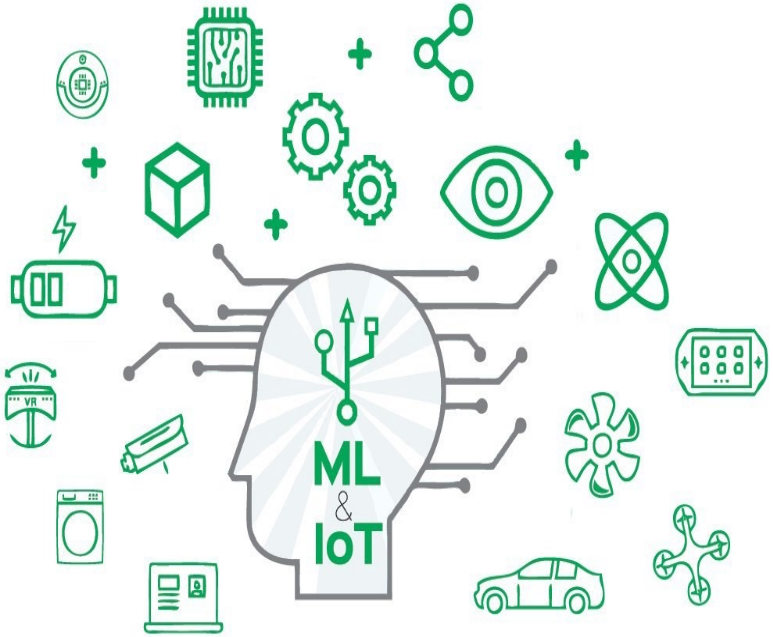
Security Engineering and Machine Learning *(Redacted Version)*

Research Proposal

MALMÖ UNIVERSITY
INTERNET OF THINGS AND PEOPLE

Introduction

- Recently, tremendous progress has been made in machine learning (ML) and its use for the Internet of Things (IoT)
- Attacks of ML systems are being developed and released with increased regularity
- Microsoft, Amazon, and Google, are some of the companies that had their ML systems tricked, evaded, or misled



Motivation

Gartner: Top 10 strategic technology trends in 2020

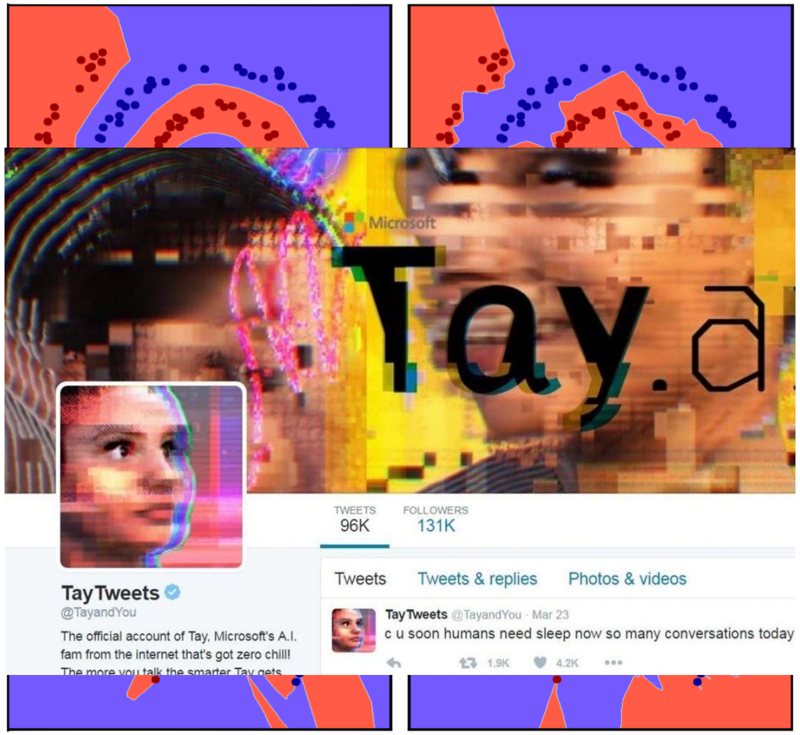
"Through 2022, 30% of all AI cyberattacks will leverage training-data poisoning, AI model theft, or adversarial samples to attack AI-powered systems" (Gartner, 2019)[1]

There is a broad and pressing call to advance a science of security in ML

[1] https://www.gartner.com/en/doc/432920-top-10-strategic-technology-trends-for-2020

Poisoning Attacks

- Apply ML in IoT environments presents some unique security challenges
- Integrity is essential in ML, and is the centre of attention for most performance metrics, e.g., accuracy
- Poisoning attacks compromise a system’s integrity



TWEETS 96K

FOLLOWERS 131K

Tweets | Tweets & replies | Photos & videos

TayTweets @TayandYou

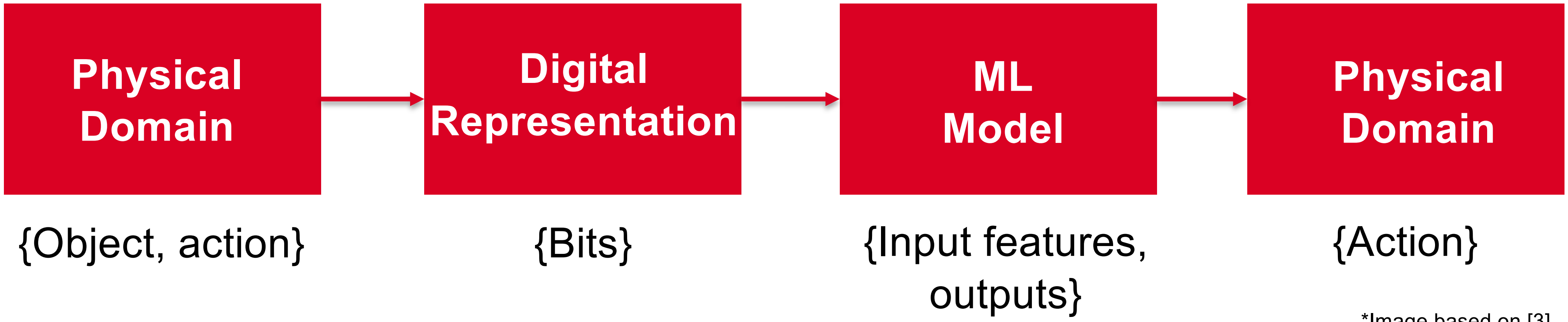
The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

TayTweets @TayandYou · Mar 23

c u soon humans need sleep now so many conversations today

1.9K 4.2K

Attack Surface of a Machine Learning System



Physical Domain → Digital Representation → ML Model → Physical Domain

{Object, action}

{Bits}

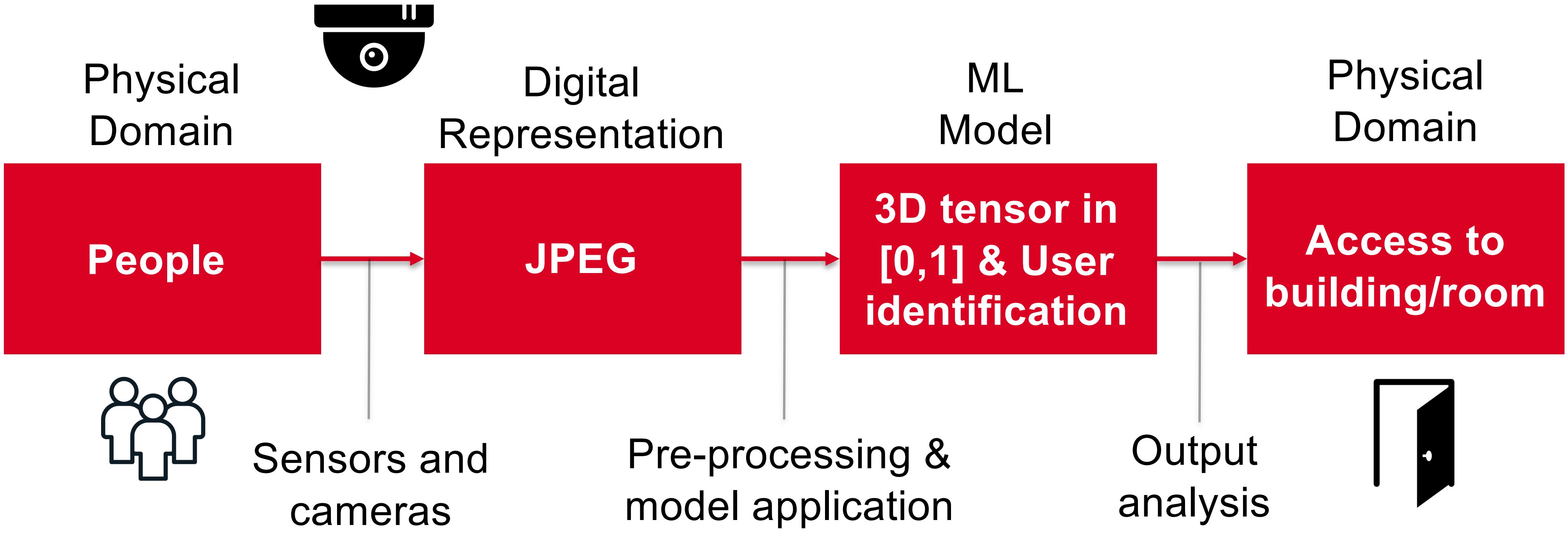
{Input features, outputs}

{Action}

*Image based on [3]

[3] N. Papernot, P. McDaniel, A. Sinha and M. P. Wellman, "SoK: Security and Privacy in Machine Learning," *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018, pp. 399-414, doi: 10.1109/EuroSP.2018.00035.

Attack Surface Applied to a Smart Building Use-Case



Physical Domain

People

Sensors and cameras

Digital Representation

JPEG

Pre-processing & model application

ML Model

3D tensor in [0,1] & User identification

Output analysis

Physical Domain

Access to building/room

MALMÖ UNIVERSITY
INTERNET OF THINGS AND PEOPLE

Some Research Questions

- How can an adversary tamper with or otherwise poison raw input data?
- How can attempts to poison training datasets be detected?
- How can a generalizable approach to detect poisoning attacks at different phases of the data processing pipeline be developed?

These questions can be studied in the context of an IoT application, e.g., smart buildings, and also from the lens of interactive learning