



Security and Privacy of AI Knowledge Guide



Lorenzo Cavallaro and Emiliano de Cristofaro
University College London

Editor: *Steve Schneider*

contact@cybok.org
www.cybok.org

Adversaries Affect Security *and* Privacy of AI Systems



This CyBOK Knowledge Guide

- Part 1: Security of AI
 - Threat models
 - Attacks
 - Defences
- Part 2: Privacy of AI
 - Attacks on privacy
 - Defences

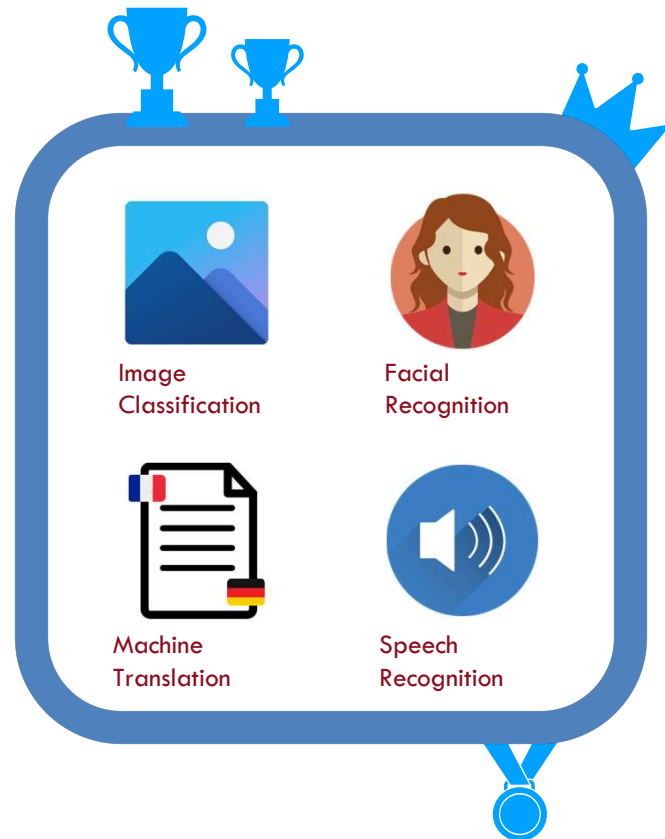
Technical aspects; non-technical issues, open problems

Knowledge Guide

“Emerging topics or those that are still developing broadly agreed foundations”

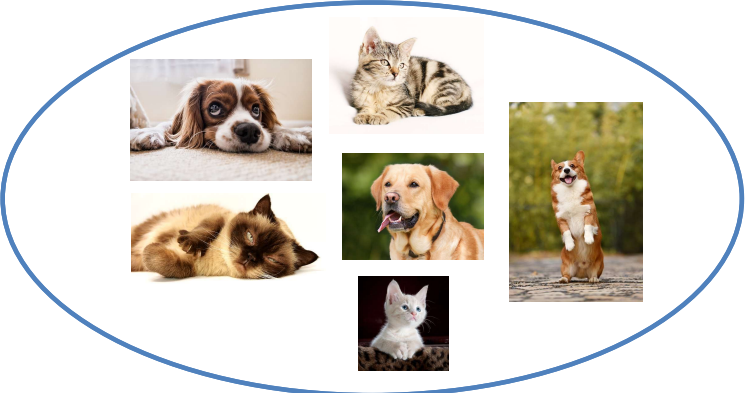


Machine Learning Revolution



AI Models

Training set



Discriminative Model

cat | dog

AI Models

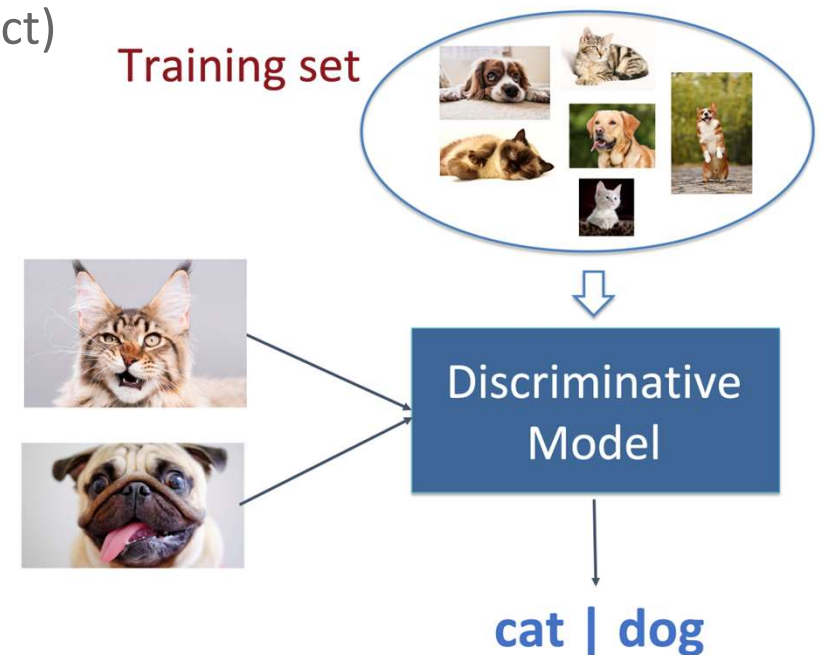
Traditionally...performance focussed

- Accuracy (how often it's correct overall)
- Precision (how often target classification is correct)
- Recall (how often instances of target are found)

With real world deployment...

... comes the need for security and privacy

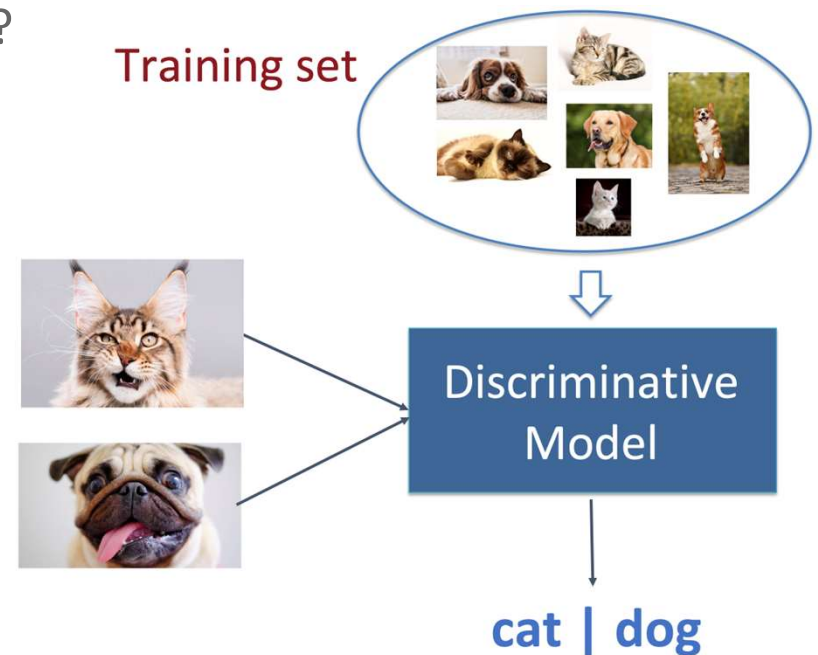
- Robustness against attacks
- Fairness
- Explainability
- Privacy



Threat Models

Section 2: Threat Models (Attacks and Defenses)

- Perfect-, Limited-, Zero-Knowledge
 - How much knowledge does an attacker have?
- Training vs Inference
 - Attacking the training or the application
- Passive vs Active
 - Honest but curious, or active adversary



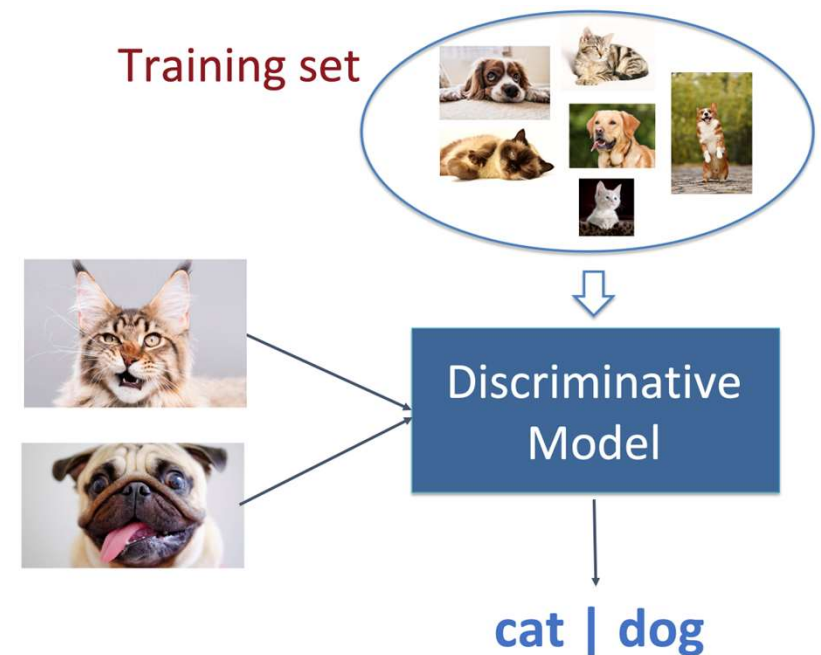
Security: Adversarial ML Attacks and Defences

Section 3: Attacks

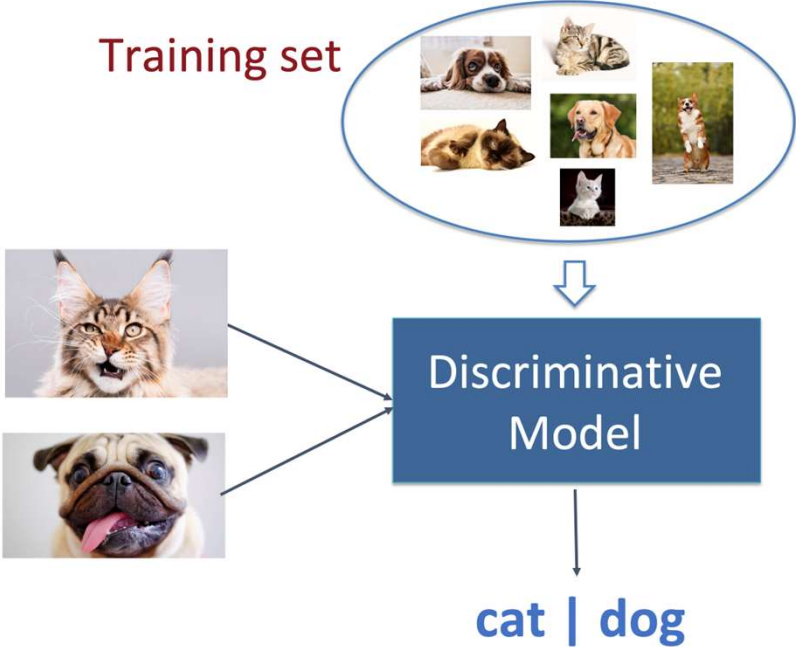
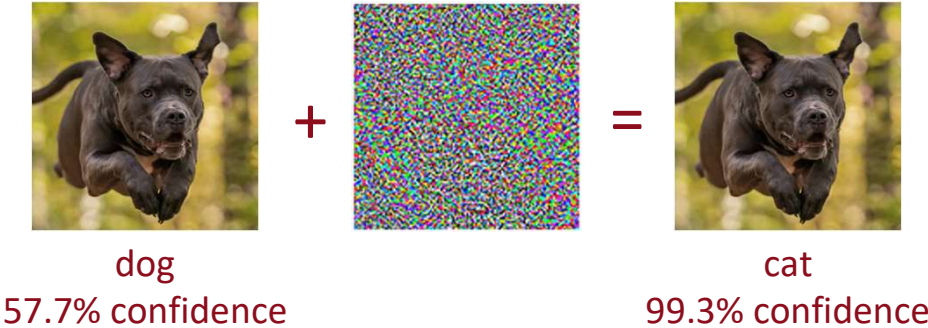
- Evasion attacks (tampering with run-time input)
- Poisoning attacks (tampering with training)
- Backdoor attacks ("triggering" training input)

Section 4: Defences

- Adversarial Training
- Out of Distribution Detection
- Certified Models
- Defences against poisoning/backdoor attacks



Adversarial example Evasion attack

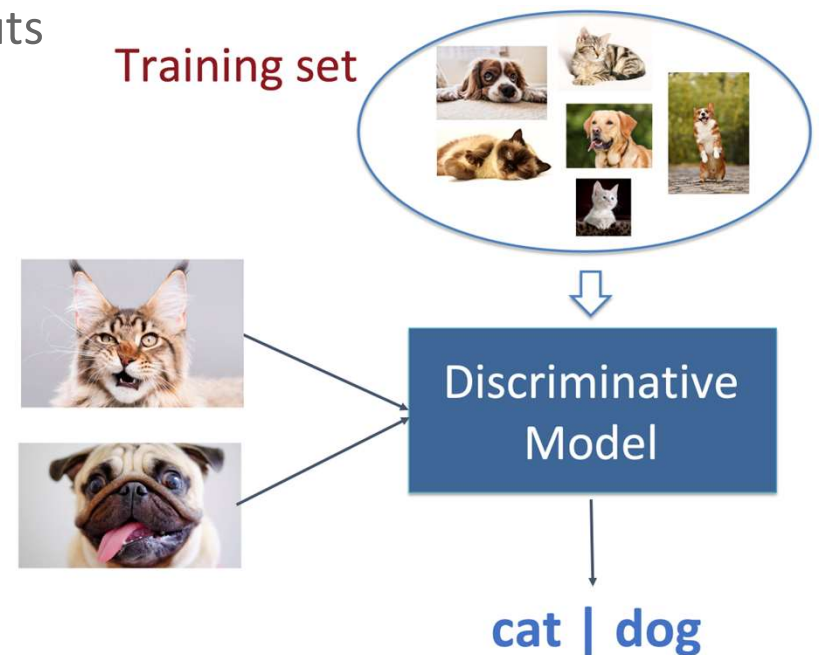


Privacy in Machine Learning

Section 5: Privacy

- Inference about training data
 - model inversion: detecting inputs from outputs
 - inferring class representatives
- Membership inference attacks
 - is an individual part of the training set?
- Model extraction
 - stealing the model

Section 6: Defences




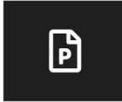



Now available!

- <https://www.cybok.org/supplementaryguides/>
- Knowledge Guide
- Lorenzo and Emiliano's Webinar
- Webinar Presentation slides
- Podcast imminent



Knowledge Guides

Security and Privacy of AI Security and Privacy of AI - Version 1.0.0					
--	---	---	---	---	---

CyBOK

CyBOK