Privacy-preserving and secure AI

Antti Honkela

Finnish Center for Artificial Intelligence FCAI
Department of Computer Science
University of Helsinki

FCAI Kick-off, 20 March 2019
Advancing Trust & Ethics: developing AI systems that can operate in a real-world adversarial environment.

Contributing to Data efficiency from the supply side: privacy can make more data available for training AI systems.
Potential adversarial activities in AI systems

- Data
  - Model poisoning
  - Data stealing during training
  - Training on untrusted hardware/software

- Training
  - Code (e.g. libraries)

- Prediction
  - Test inputs
  - Training data stealing (model inversion)
  - Model stealing
  - Model evasion (adversarial examples)
  - Oblivious prediction
Examples of previous work by FCAI researchers

Cryptographic techniques to ensure data privacy during the prediction phase (e.g. Liu et al., “Oblivious neural network predictions via MiniONN transformations”, CCS 2017)


Avoiding privacy breaches of training data and model inversion using differentially private learning (e.g. Heikkilä et al., “Differentially private Bayesian learning on distributed data”, NIPS 2017)
Objectives

- Rigorous adversary model for privacy-preserving and secure ML
- Techniques for detecting and thwarting attacks
- Focus on security and privacy of AI and ML rather than AI for security
- Focus on attacks specific to AI and ML rather than general systems security

More information:
Coordinator: Assoc. Prof. Antti Honkela (University of Helsinki, CS)