ADVERSARIAL MACHINE LEARNING

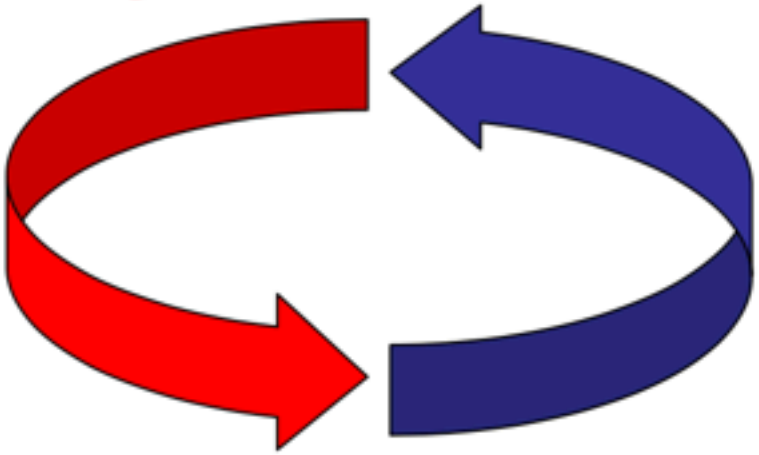
WHAT IS ADVERSARIAL MACHINE LEARNING?

5

- Adversarial Examples are data points obtained by adding systematic noise into training samples.
- The noise is added in such a manner that target classifier will misclassify the adversarial examples.
- Study of generating such Adversarial Examples and defences against such attacks is called **Adversarial Machine Learning**.

Adversary

1. Analyze classifier
2. Devise and execute attack



Classifier Designer

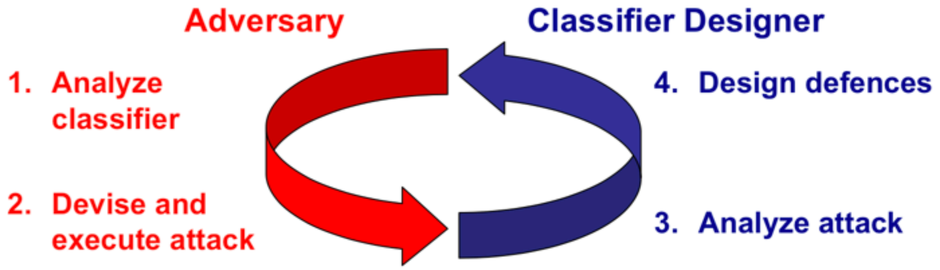
4. Design defences
3. Analyze attack

Image Source: http://pralab.diee.unica.it/en/WhatIsAdversarialLearning

RESEARCH LANDSCAPE

- Attack Models (Part I)
  - How to convert a pure example into an adversarial example for a given classifier net?
- Defence Mechanism (Part II)
  - How to train a DNN classifier so as to make it robust against the attacks?

WHAT IS ADVERSARIAL MACHINE LEARNING?



- Adversarial Examples are data points obtained by adding systematic noise into training samples.
- The noise is added in such a manner that target classifier will misclassify the adversarial examples.
- Study of generating such Adversarial Examples and defences against such attacks is called **Adversarial Machine Learning**.

Image Source: http://pralab.diee.unica.it/en/WhatIsAdversarialLearning