ADVERSARIAL MACHINE LEARNING

4

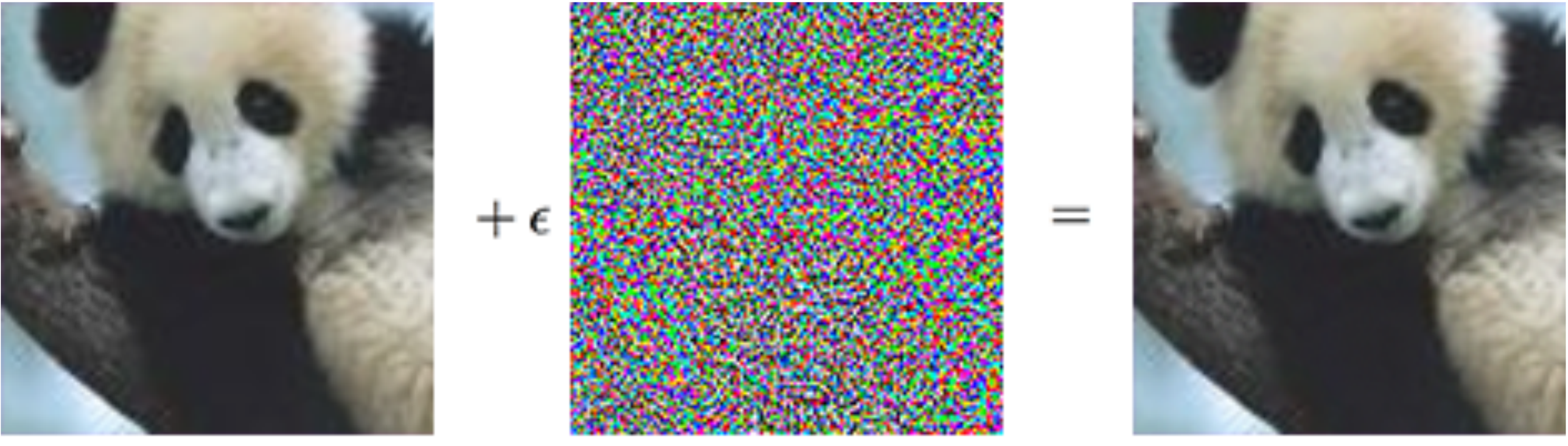
“panda”

57.7% confidence

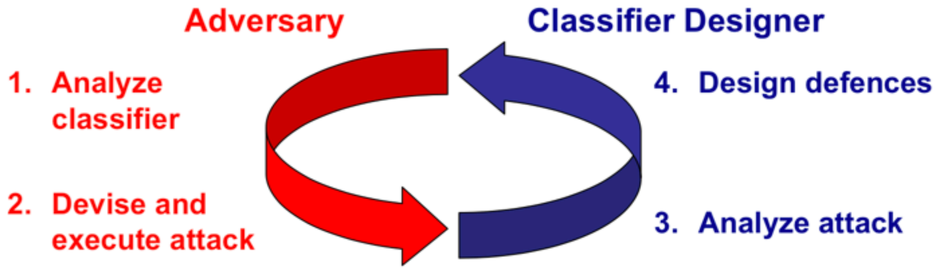
“gibbon”

99.3% confidence

Image Source: OpenAI Blog

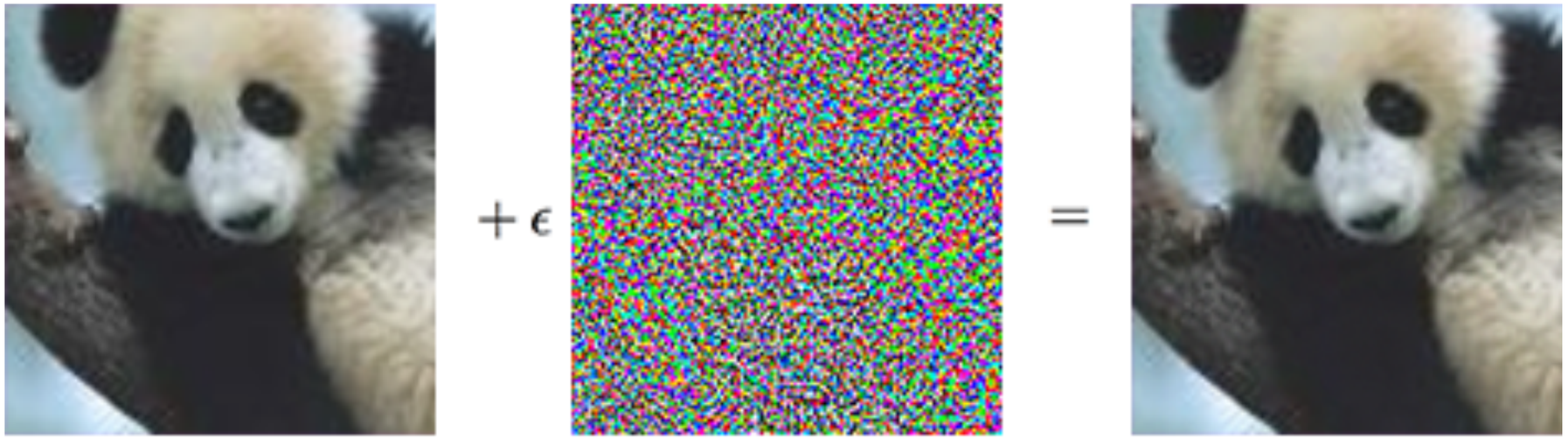

+ ϵ
=

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



“panda”

57.7% confidence

“gibbon”

99.3% confidence

Image Source: OpenAI Blog