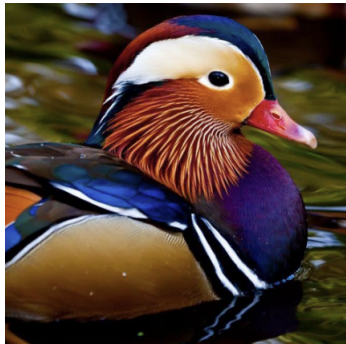
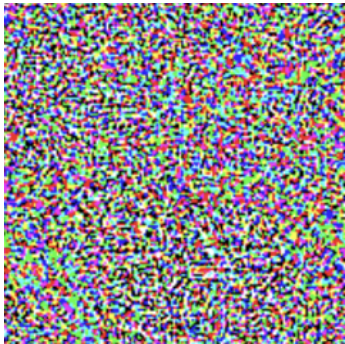


Benign example
"duck": 42.0% confidence



x

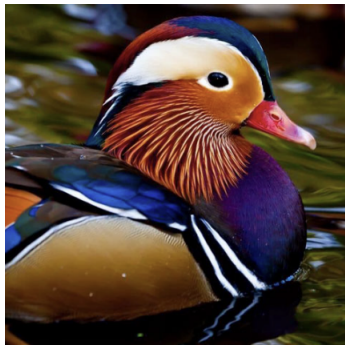
$+ \epsilon \times$



$\text{sign}(\nabla_x J(\theta, x))$

$=$

Adversarial example
"duck": 0.01% confidence
"platypus": 15.2% confidence



$x + \epsilon \text{sign}(\nabla_x J(\theta, x))$