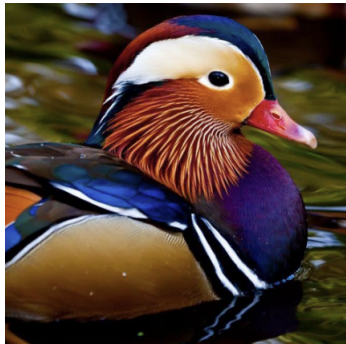
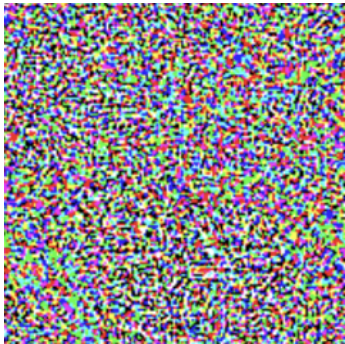


Benign example  
"duck": 61% confidence



$x$

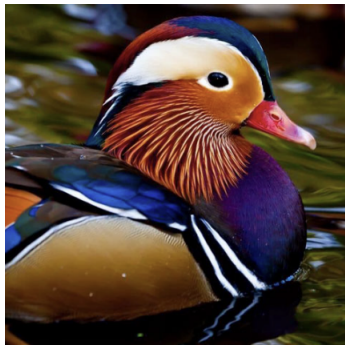
$+ \epsilon \times$



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

$=$

Adversarial example  
"pug": 99.9% confidence



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