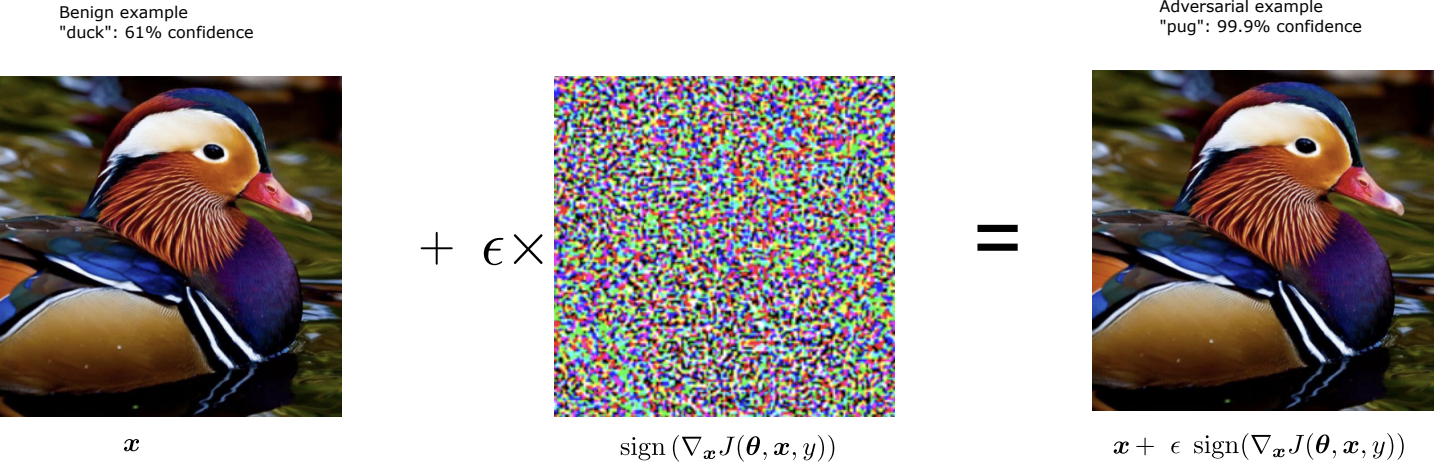
Benign example
"duck": 61% confidence
+ ϵ×
=
Adversarial example
"pug": 99.9% confidence
$\boldsymbol{x}$
$\text{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)$
$\boldsymbol{x} + \epsilon\ \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$