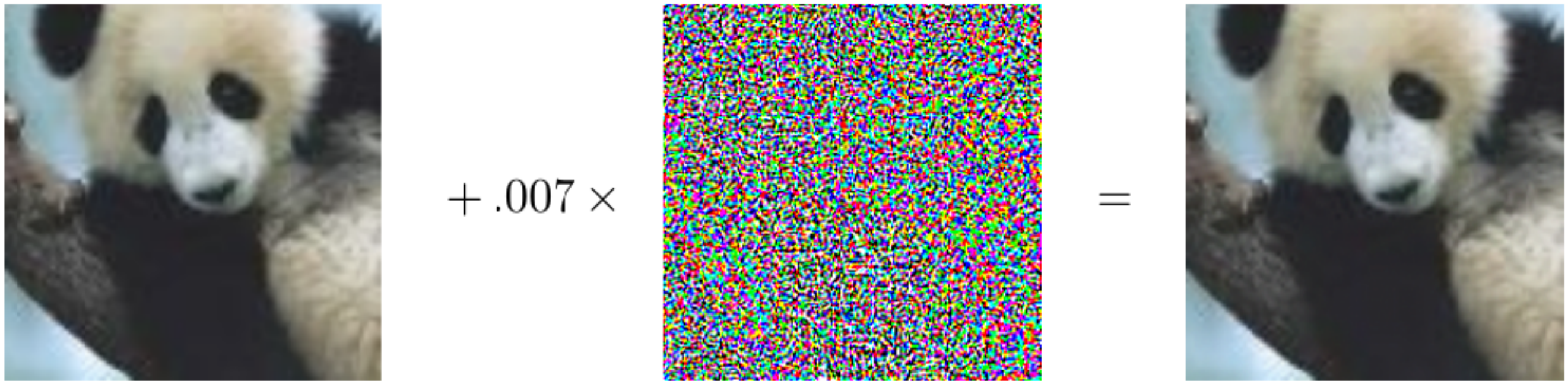


Adversarial Examples



Timeline:

“Adversarial Classification” Dalvi et al 2004: fool spam filter

“Evasion Attacks Against Machine Learning at Test Time”

Biggio 2013: fool neural nets

Szegedy et al 2013: fool ImageNet classifiers imperceptibly

Goodfellow et al 2014: cheap, closed form attack