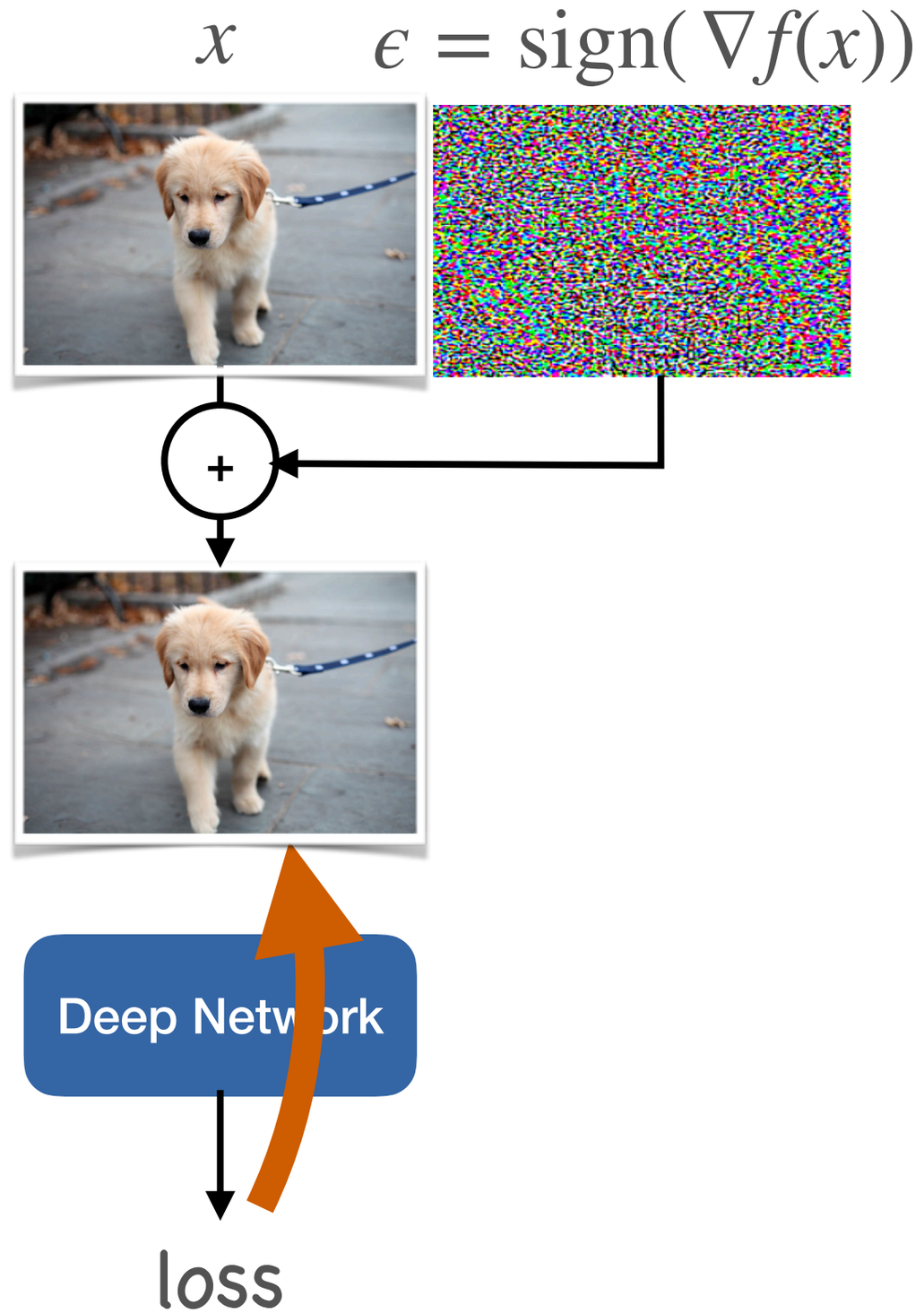
# White vs black box attacks

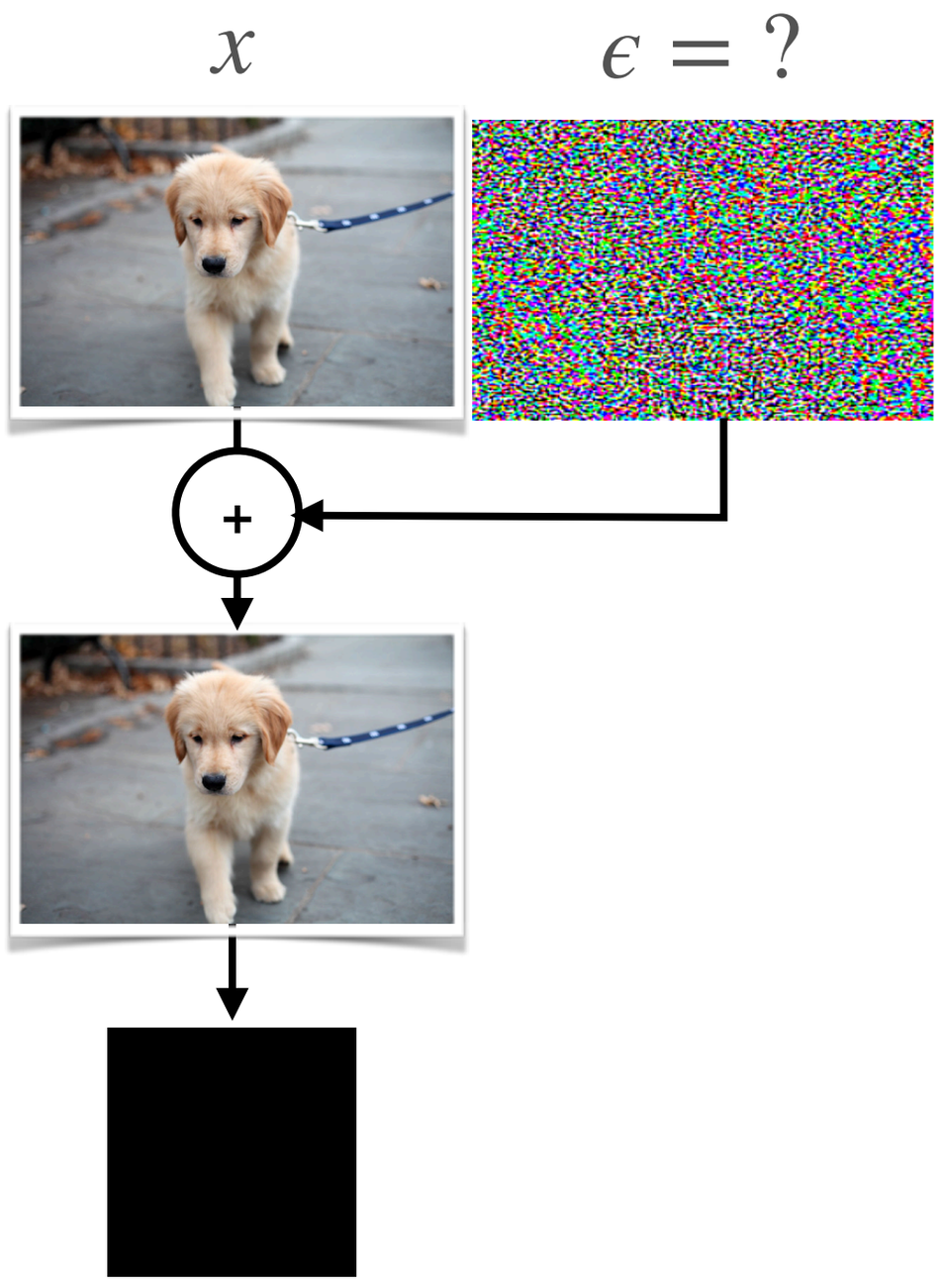
# White box attacks

- Attacker has access to model and gradients
- Fast gradient sign
- Projected gradient descent



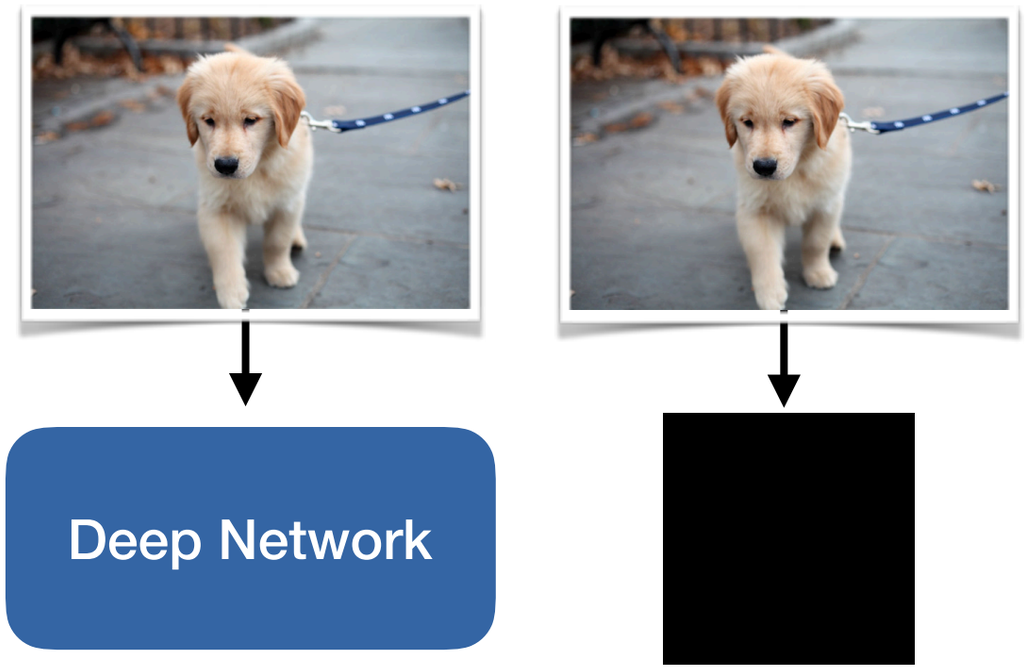
# Defense by hiding model

- Can we defend against attacks if we do not allow backprop?



# Back box attacks

- Train network to imitate black box network
  - Attack new network
    - Attack black box
- If not successful
  - repeat




Practical Black-Box Attacks against Machine Learning, Papernot et al., arXiv 2016