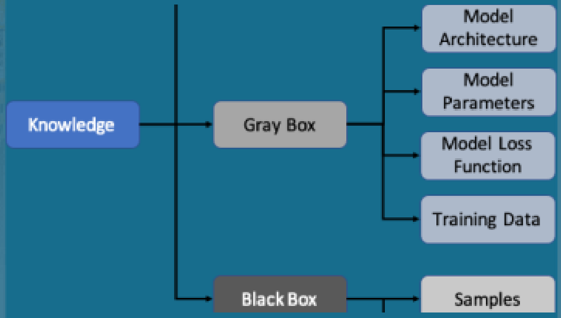
SAND2022-2962C

Deep Deception



D. Farley, Z. Gastelum, T. Shead

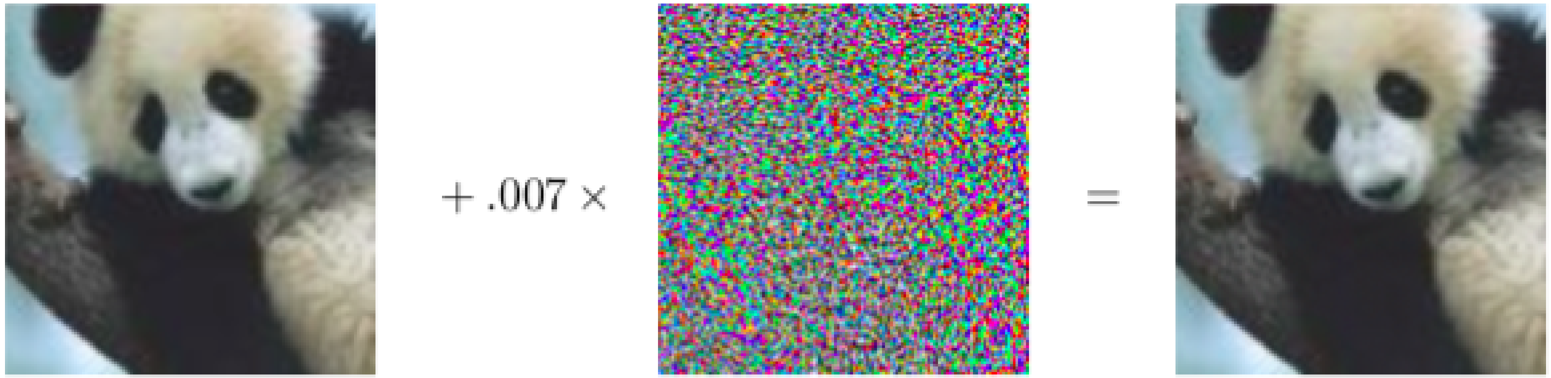
Artificial intelligence and machine learning for IAEA safeguards

15 March 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Adversarial Machine Learning is an active area, and no winners



$\boldsymbol{x}$

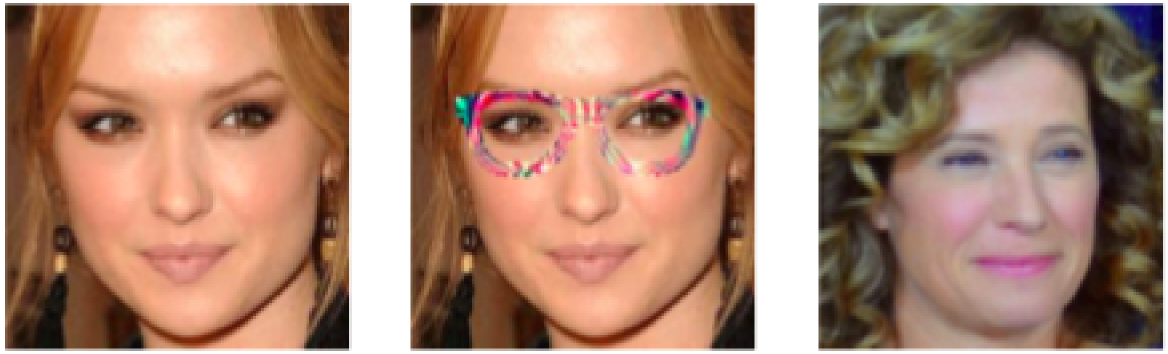
“panda”
57.7% confidence

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

“nematode”
8.2% confidence

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

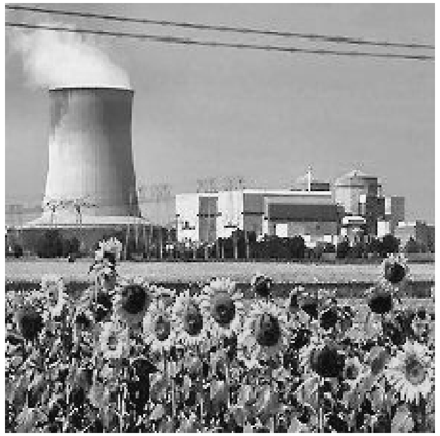
“gibbon”
99.3 % confidence



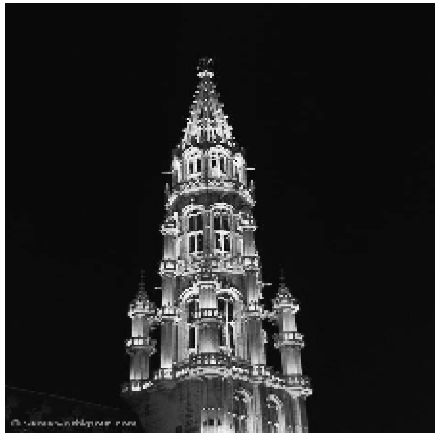
AML of Safeguards Images

Trained Model Results

“Hyperboloid Cooling Tower”



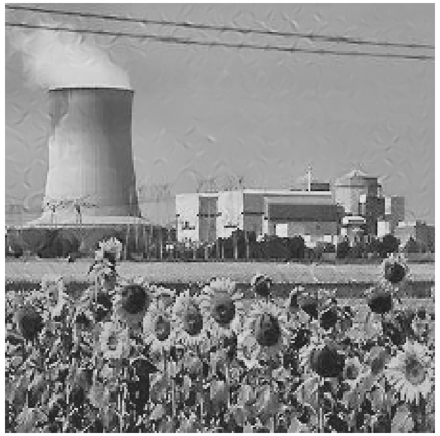
“NOT Hyperboloid Cooling Tower”



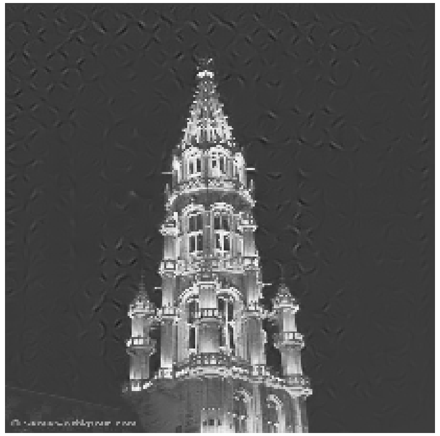
Carlini & Wagner* Attack

Adversarial Results

“NOT Hyperboloid Cooling Tower”



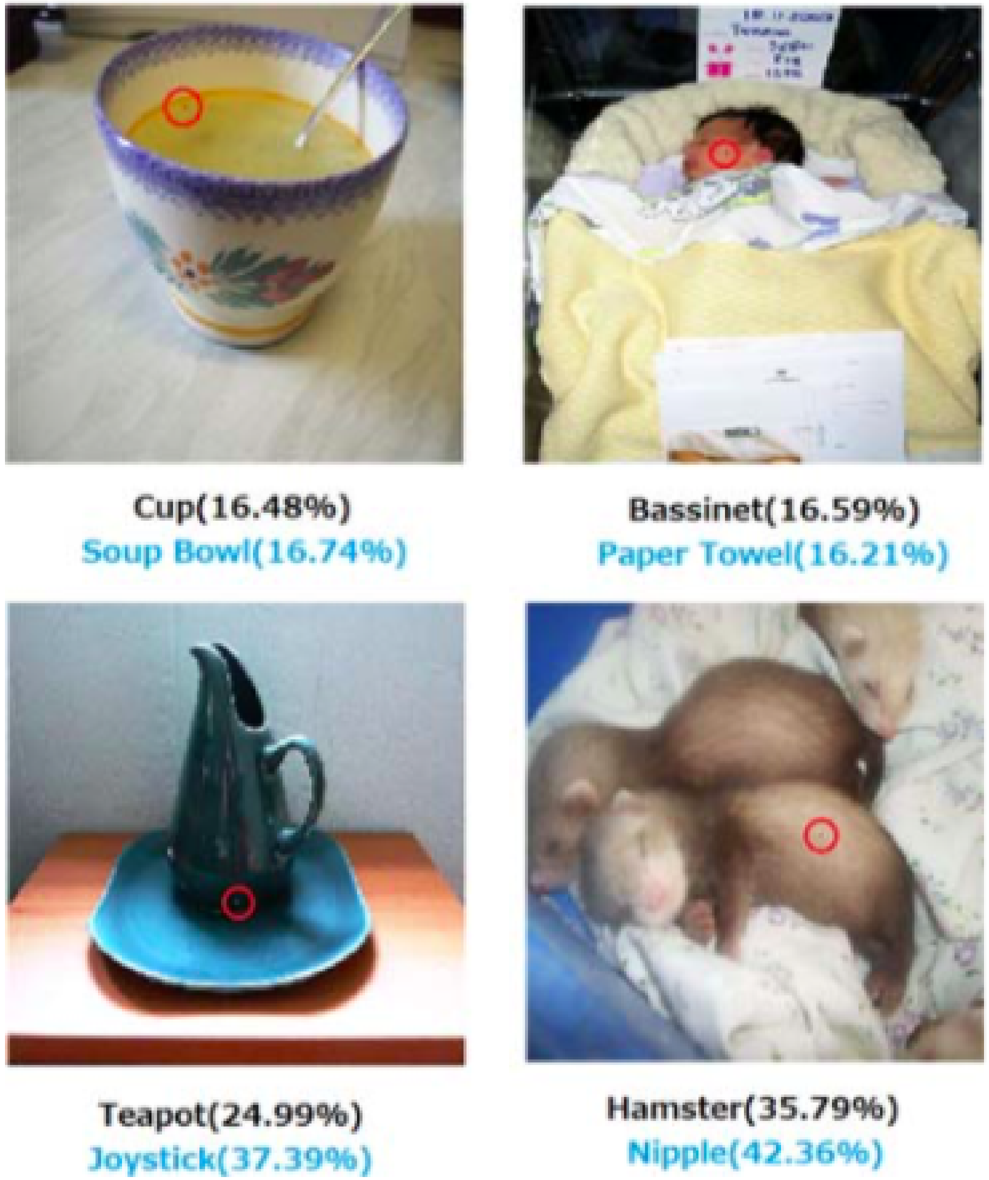
“Hyperboloid Cooling Tower”



The trained model WAS NOT changed – rather, a small, nearly imperceptible, perturbation was added to the original image

* N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, San Jose, 2017.

Even one pixel can change the answer



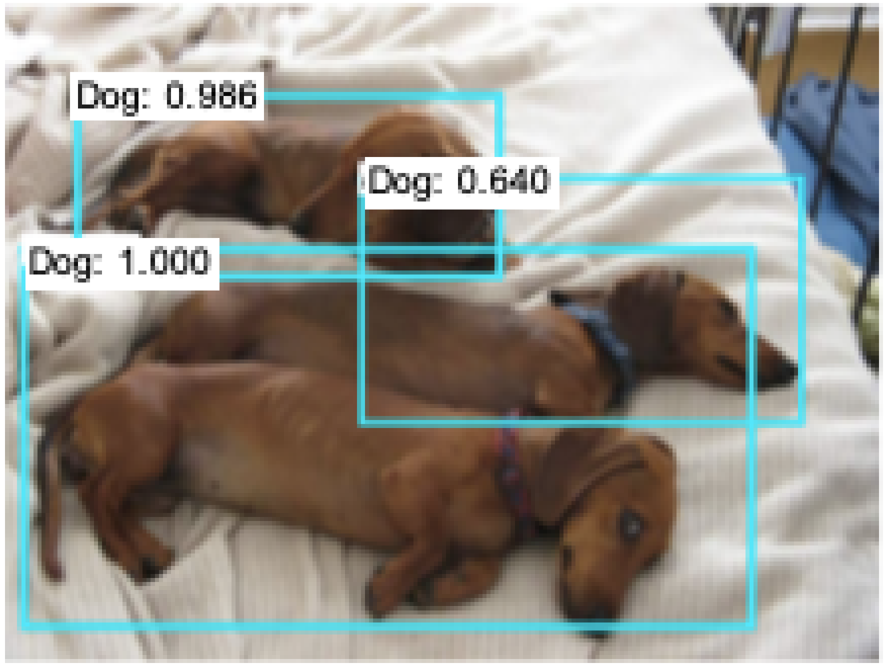
Cup(16.48%)
Soup Bowl(16.74%)

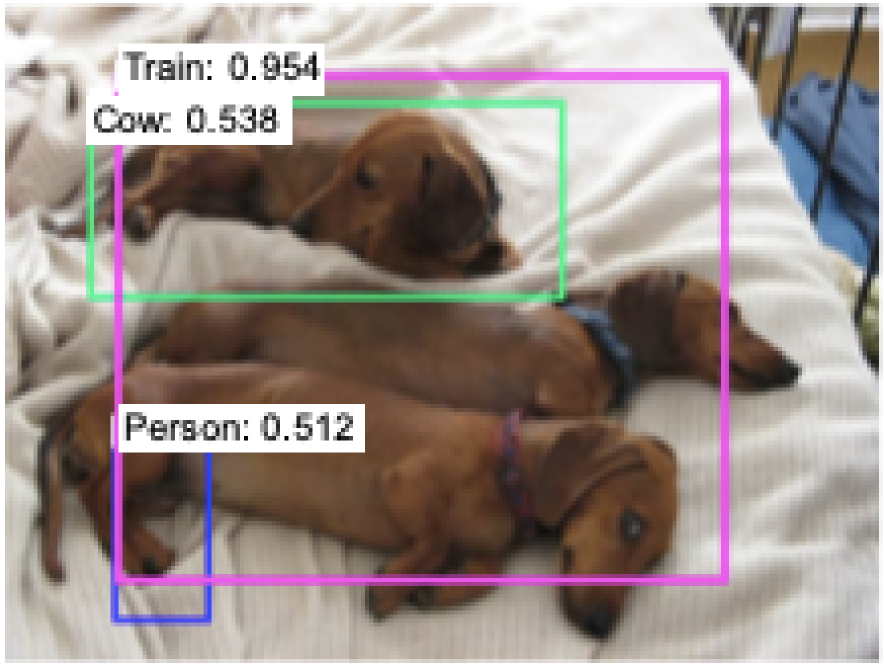
Bassinet(16.59%)
Paper Towel(16.21%)

Teapot(24.99%)
Joystick(37.39%)

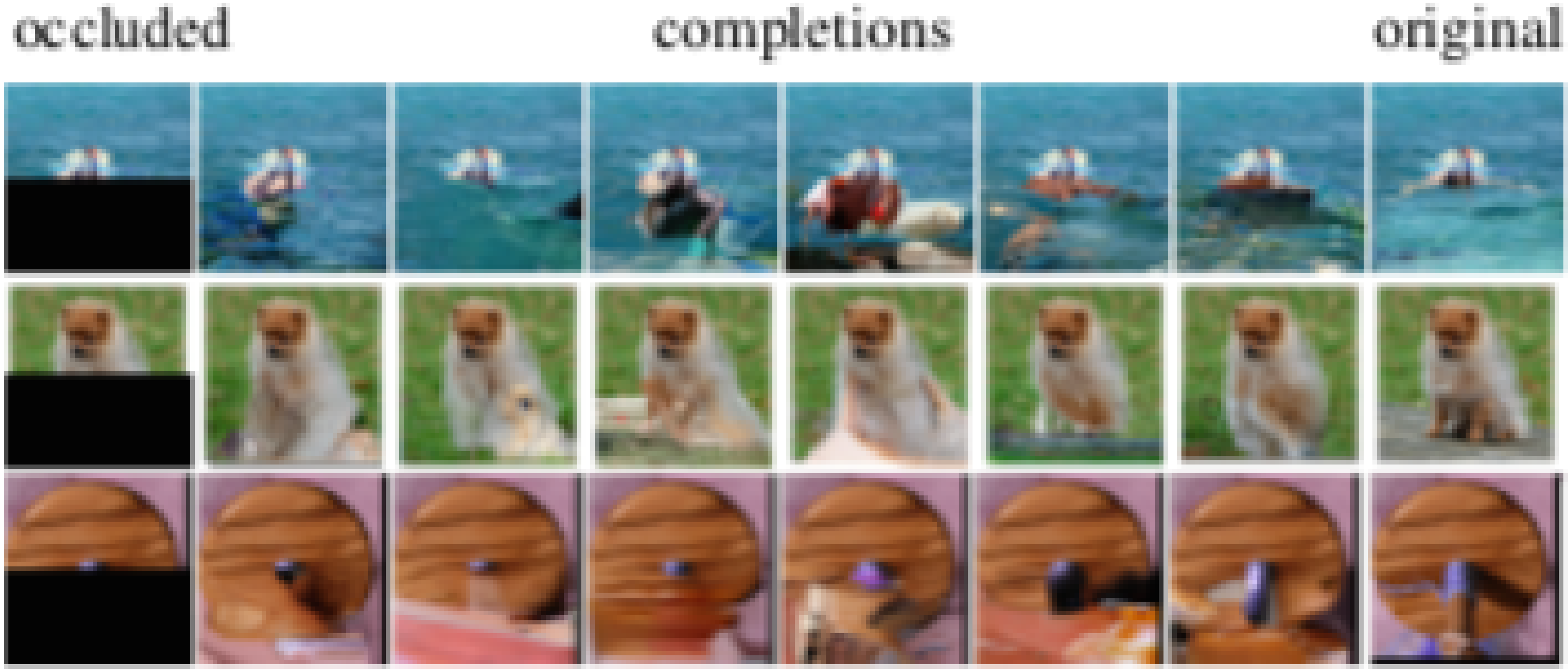
Hamster(35.79%)
Nipple(42.36%)

Frames of an image can also be fooled

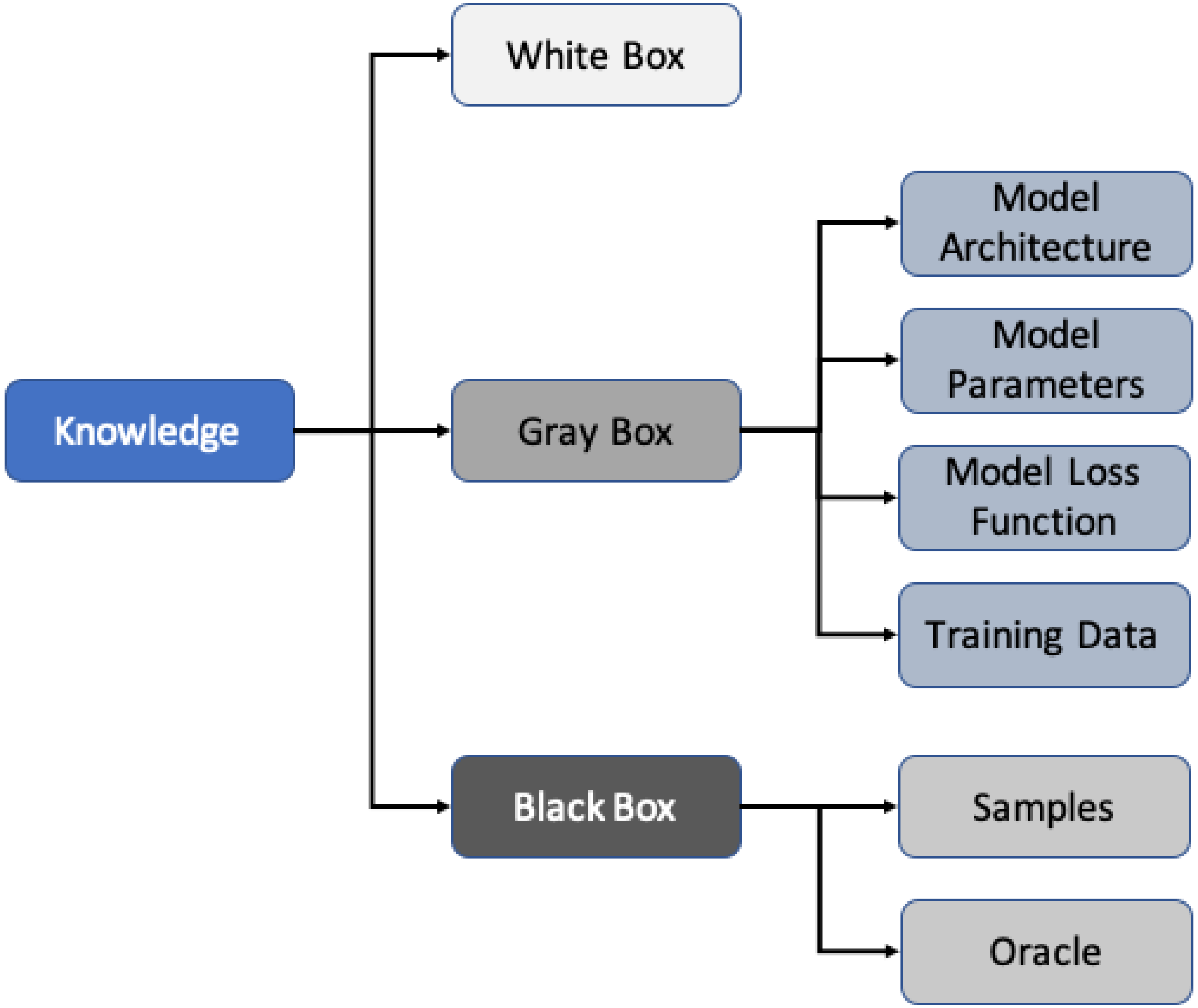




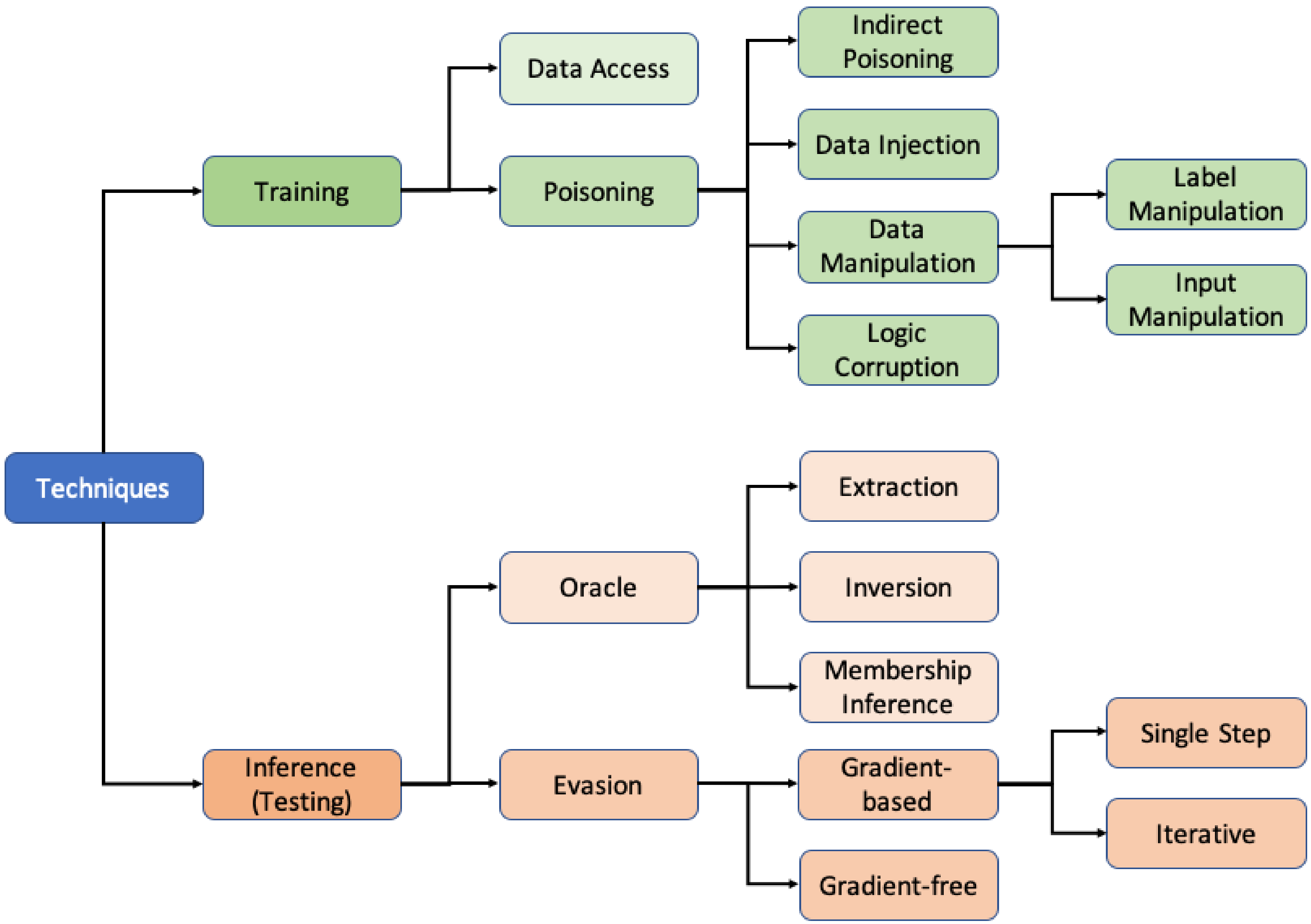
Can complete a partial image!



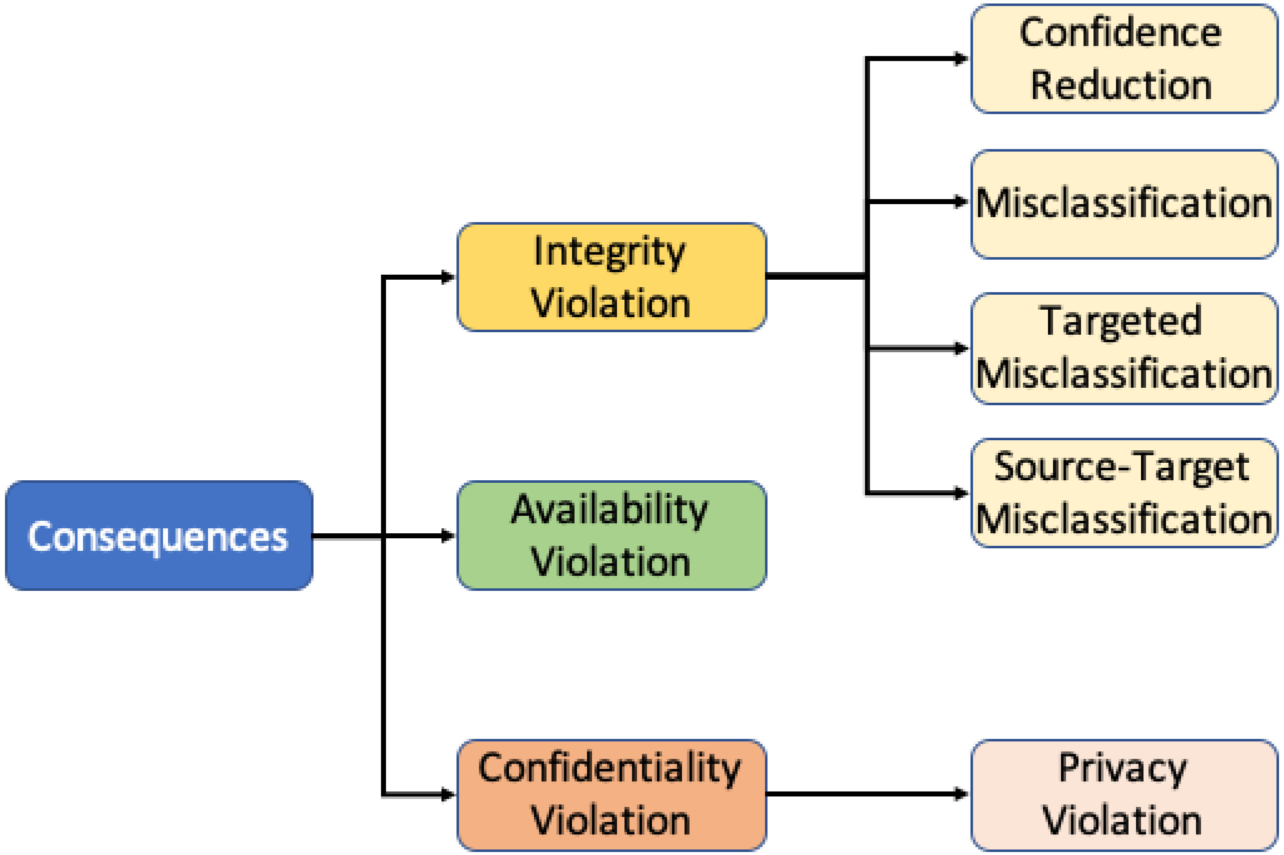
What is the adversary's knowledge?



How does to adversary conduct an attack?



Successful Adversarial Attacks can take several forms


Consequences
Integrity Violation
Availability Violation
Confidentiality Violation
Confidence Reduction
Misclassification
Targeted Misclassification
Source-Target Misclassification
Privacy Violation