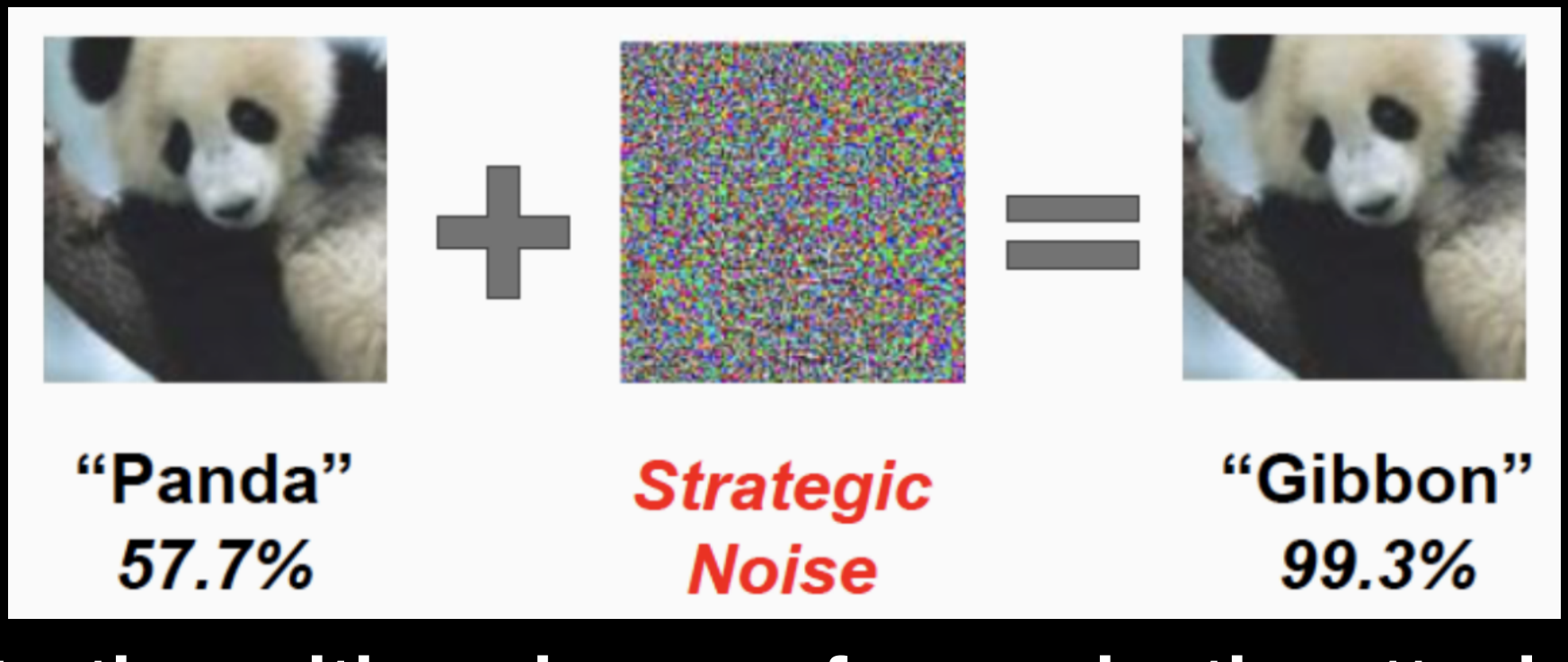
HOW TO HACK NEURAL NETWORKS

@learn.machinelearning

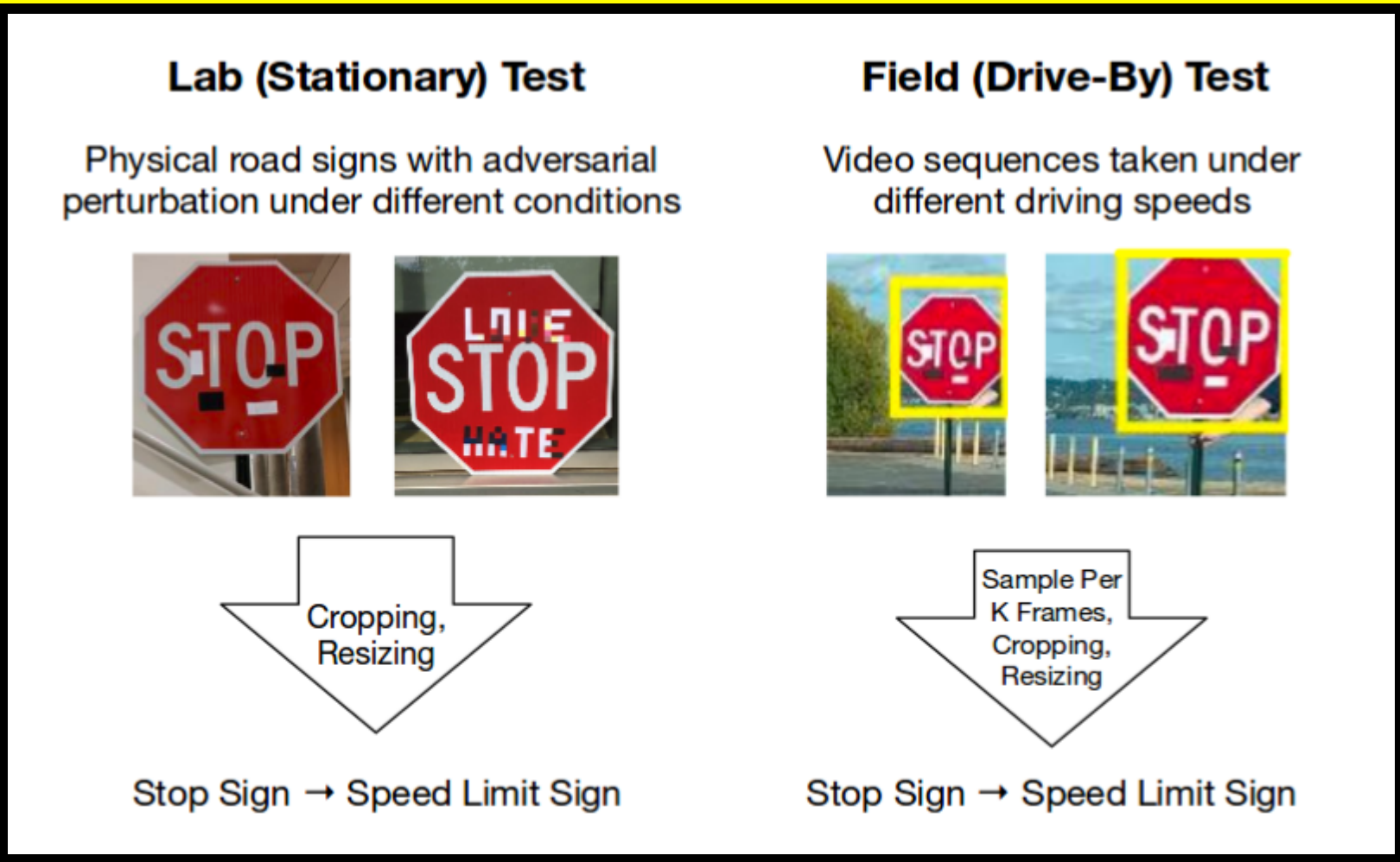
Adversarial Machine Learning

- Let's see few examples



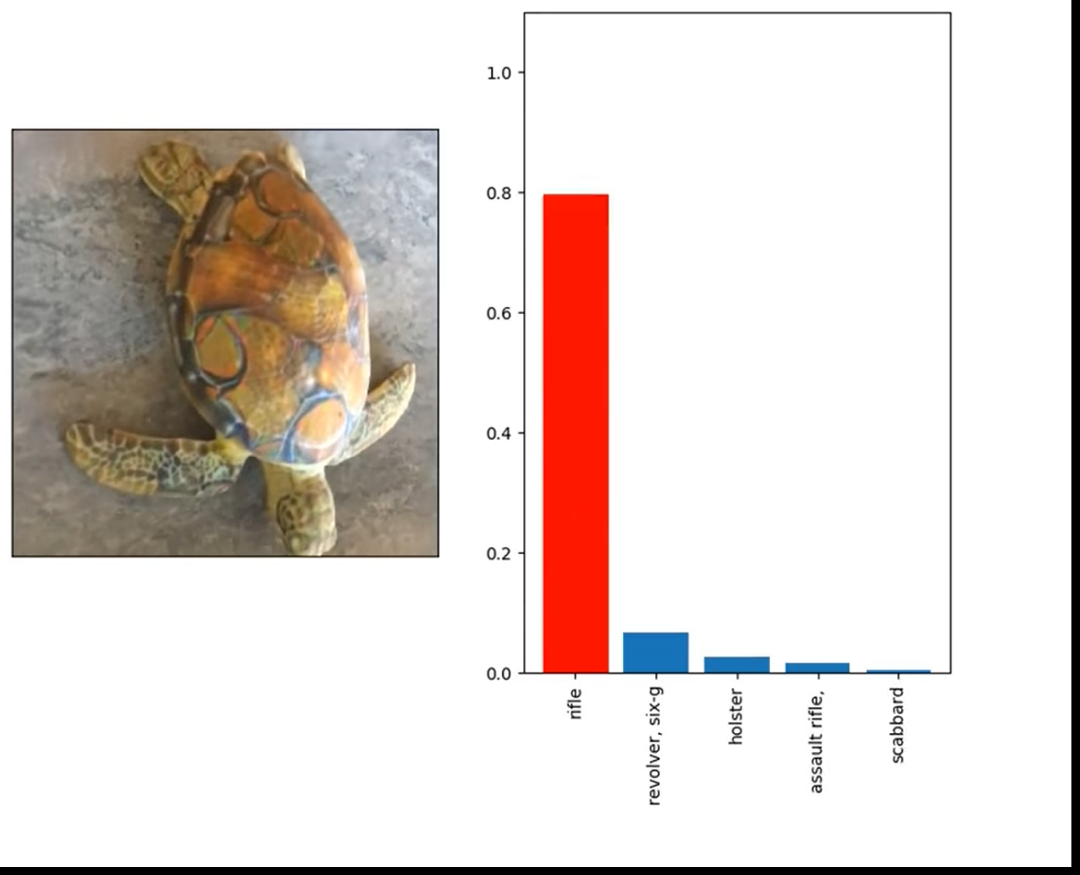
- starting with an image of a panda, the attacker adds a small perturbation that has been calculated to make the image be recognized as a gibbon with high confidence.
- Adding strategic noise to an image can be used to fool neural networks. Source:
- Goodfellow et al., 2015.

Adversarial Machine Learning



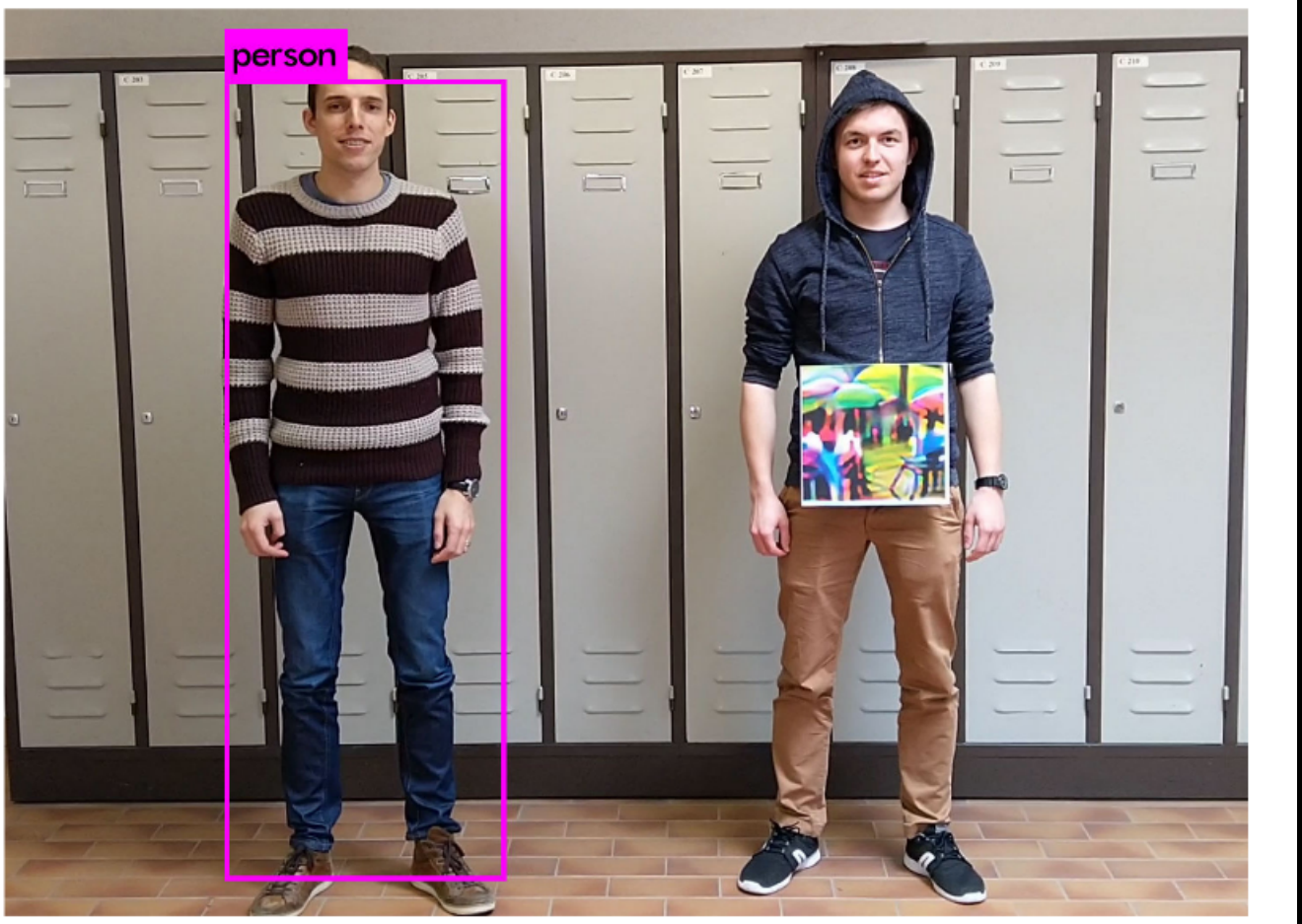
- **A recently released paper showed that a stop sign manipulated with adversarial patches caused the SOTA model to begin “thinking” that it was a speed limit sign.**
- **Imagine an attacker who manipulates road signs in a way such that self-driving cars will break traffic rules.**

Adversarial Machine Learning



- Here’s a nice example from MIT, where they have 3D-printed a turtle and the SOTA classifiers predict it to be a rifle. While this is funny, the reverse, where a rifle is predicted as a turtle, can be dangerous and alarming in some situations.
- imagine a warfare scenario where these models were deployed at scale on drones and were tricked by similar patches to hijack the attack on different targets. This is really terrifying!

Adversarial Machine Learning



- We create Adversarialpatch that is successfully able to hide persons from a person detector.
- Left: The person without a patch is successfully detected.
- Right: The person holding the patch is Ignored.
- This is really alarming as it can be used by intruders to get past any security cameras, among other things.

Adversarial Machine Learning

- There are many more examples
- Adversarial Machine Learning is an active research field where people are always coming up with new attacks & defences; it is a game of Tom and Jerry (cat & mouse) where as soon as someone comes up with a new defence mechanism, someone else comes up with an attack that fools it.

- **ATTACKS:**
  - Noise
  - Semantic
  - Fast Gradient Sign Method
  - Projected Gradient Descent
  - DeepFool
- **DEFENSES:**
  - Adversarial Training
  - Random Resizing and Padding

Adversarial Machine Learning

- 2 ways in which attacks can be classified
  - **Black Box Attack:** The type of attack where the attacker has no information about the model, or has no access to the gradients/parameters of the model.
  - **White Box Attack:** The opposite case, where the attacker has complete access to the parameters and the gradients of the model.
- Each one of these attacks can be classified into 2 types
  - **Targeted Attack:** A targeted attack is one where the attacker perturbs the input image in a way such that the model predicts a specific target class.
  - **Un-Targeted Attack:** An untargeted attack is one where the attacker perturbs the input image such as to make the model predict any class other than the true class.

RESOURCES

- FInd all resources here