Talk Concept "Risks of AI: Adversarial Examples"

1. Motivation: "What do Adversarial Examples do?"
This part is supposed to give the listener an idea on why Adverserial Examples are important and considered a "Risk of AI"

What do they do?
- Can cause image NNs to misclassify an image
- Can cause image NNs to classify an image as a predefined target class

Examples:
- Glasses Attack https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf
- Sticker Attack https://arxiv.org/pdf/1712.09665.pdf
- Turtle Attack https://arxiv.org/pdf/1707.07397.pdf

2. Going into the details: "What are Adversarial Examples?"
This part is supposed to give the listener an idea on how hard/easy it is to generate (good) adversarial examples

As an Intro I want to explain occlusion as an easy to understand technique for acquiring adversarial examples
I would like to do a live demo, trying to get https://cloud.google.com/vision/ or a similar online available network to misclassify the image

Occlusion is very crude but Black-Box, so next I want to show a more "subtle" (but White-Box) technique: Gradient Ascent on the Image Pixels
Live demo for this one again

Then I will hint at other techniques, focusing on their requirements
- White Box vs. Black Box Attacks https://arxiv.org/pdf/1605.07277.pdf

3. "Can this be mitigated, or are NNs just flawed in that way?"
This part is supposed to give the listener an idea on why adversarial examples are possible and what can be done to protect/fix

"Machine learning techniques were originally designed for stationary and benign environments in which the training and test data are assumed to be generated from the same statistical distribution. However, when those models are implemented in the real world, the presence of intelligent and adaptive adversaries may violate that statistical assumption to some degree, depending on the adversary. " (From Wikipedia)

- Explanation https://arxiv.org/pdf/1412.6572.pdf
- Explanation https://arxiv.org/pdf/1905.02175.pdf
- Protection https://arxiv.org/pdf/1805.06605.pdf
- Protection http://iphome.hhi.de/samek/pdf/SriICML19.pdf