# Adversarial Attacks in Medical Imaging

Ben Cardoen

November 21st, 2020

Medical Image Analysis Laboratory, School of Computing Science, Simon Fraser University

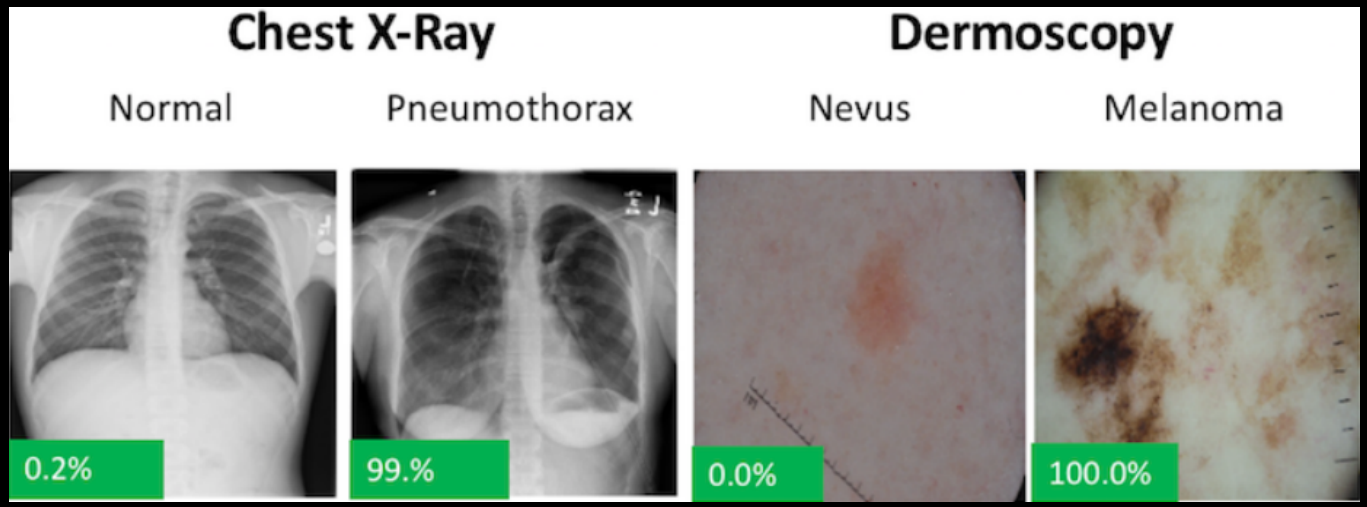
# AI can help diagnose patients

## AI can be

- faster than a doctor (1000s of images / second)
- more accurate than a human, doctor can deal with treating patients
- so good a doctor is no longer needed to verify [1]
- a predictor of new drugs without any real life clinical trials

[1] https://www.technologyreview.com/2018/04/11/3052/fda-approves-first-ai-powered-diagnostic-that-doesnt-need-a-doctors-help/

# AI can help diagnose patients




An example of state-of-the-art medical imaging AI on X-ray (left) and dermoscopy images (right). The AI prediction shows the confidence that the image is from an unhealthy patient. Images reproduced with permission [1, 3].

# Can we outsmart AI?

## Why would someone want to fool medical AI?

- US spends each year more than $\sim$ 3,000,000,000,000 $ on health care [2]
- Get treatment (drugs) that you do not need
- Deny someone treatment
- Force a (non-working) drug to be approved for use
- Get reimbursed for non-existent disease

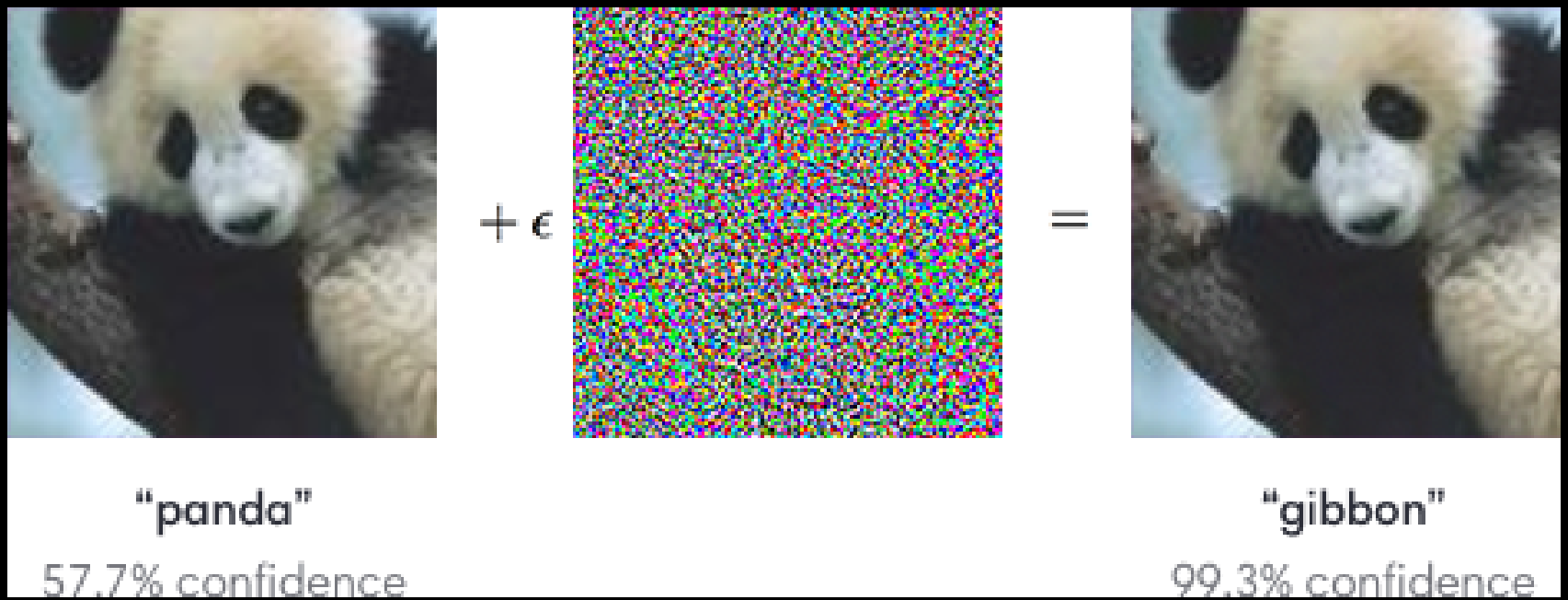
## But what about the experts?

- Hospital staff is not trained in Computer Science
- Software is rarely updated
- Adversarial attacks are hard to prove

# How can you fool AI?

## AI can be fooled by

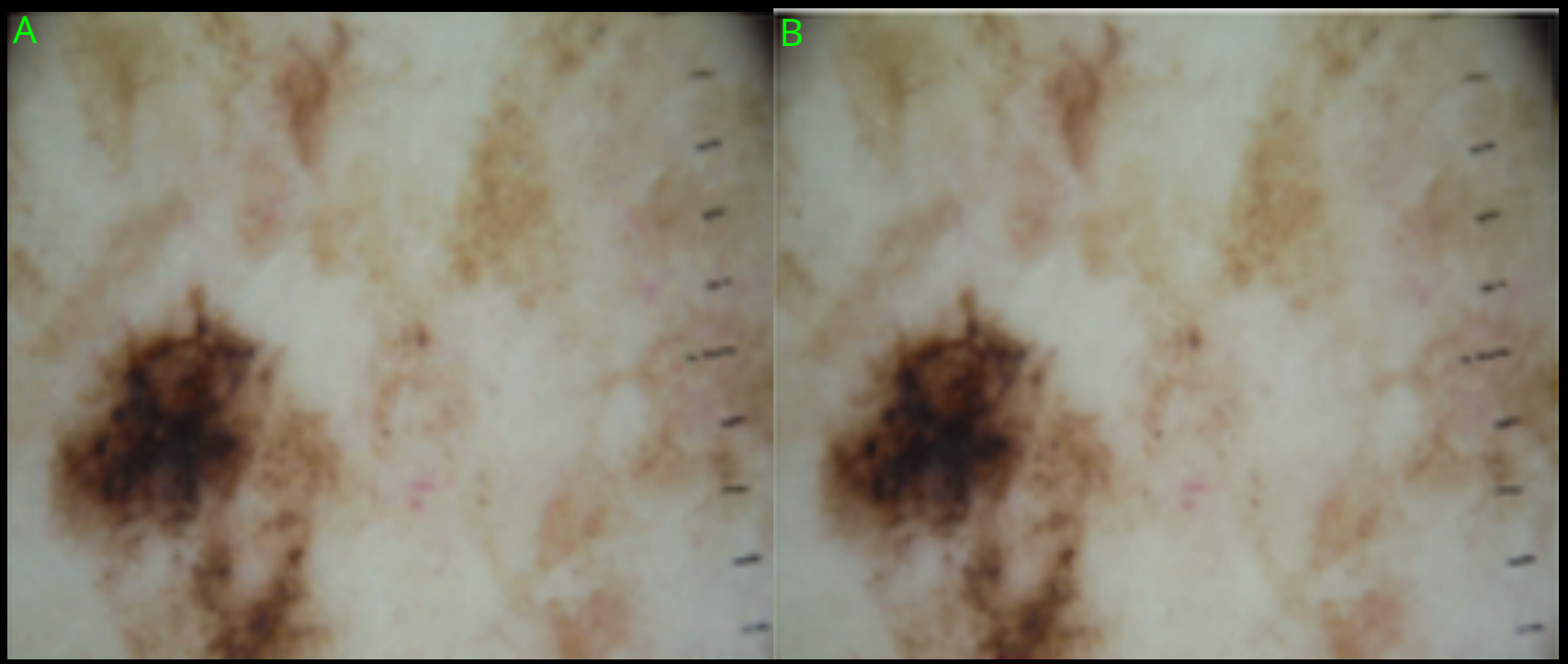
- By learning how to change[2] the image so a human can't tell the difference
- but the AI is fooled




Adding imperceptible noise to an image can trick AI into misclassifying

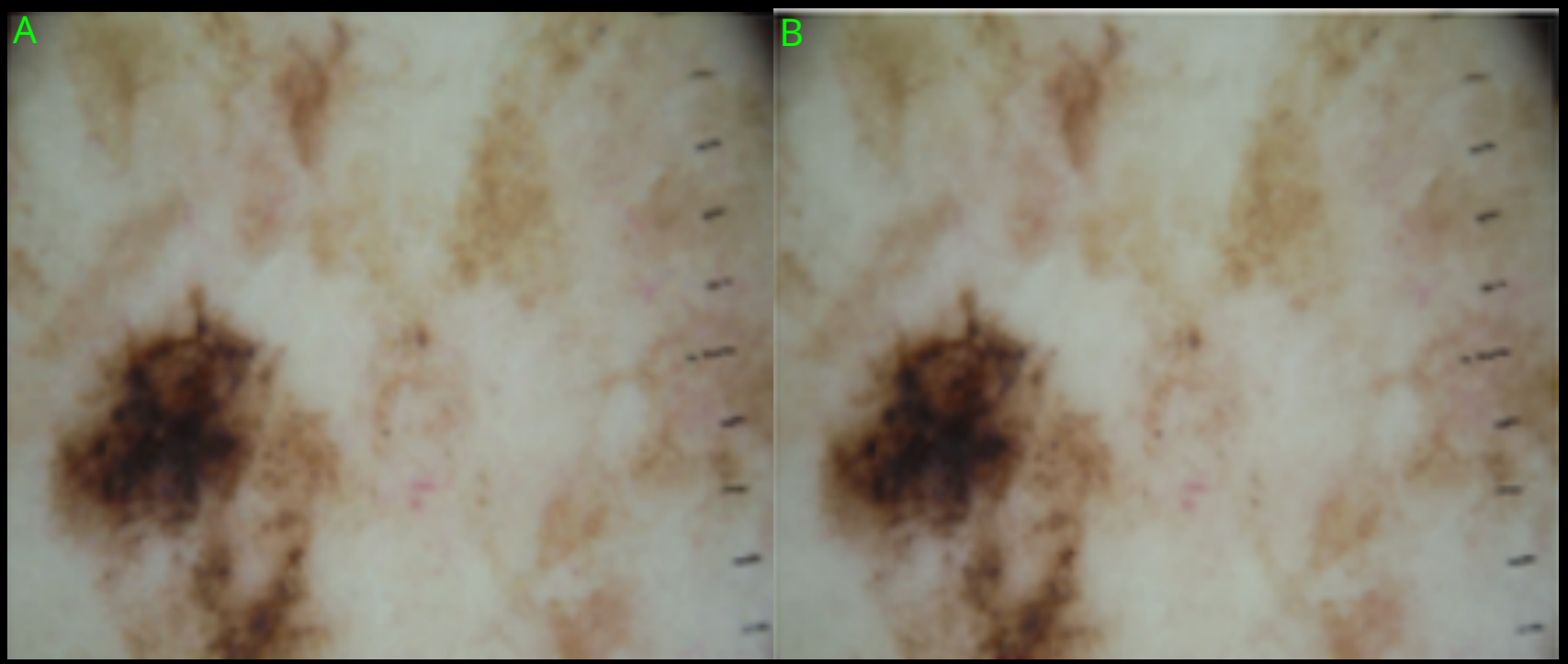
[2] https://github.com/tensorflow/cleverhans
[3] https://openai.com/blog/adversarial-example-research/

# How can you fool medical AI?




Which image was altered ?

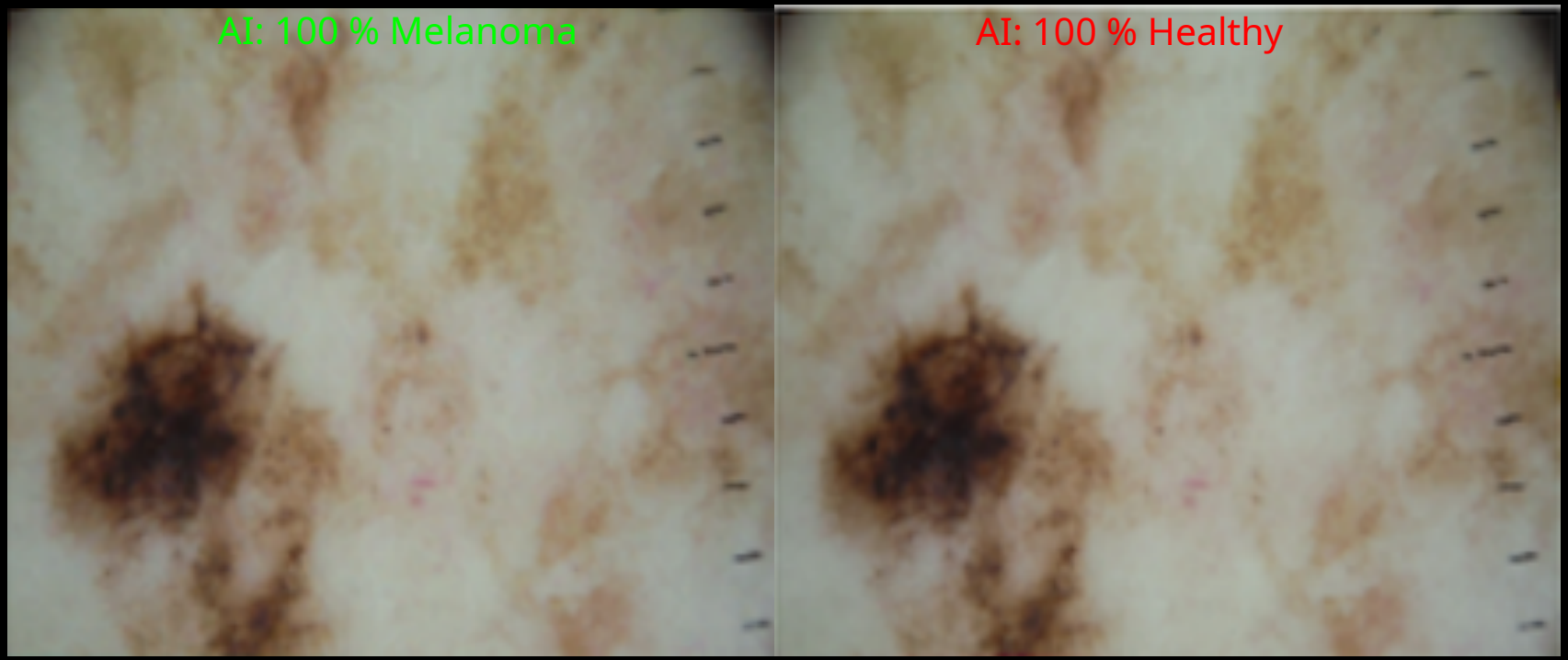
# How can you fool medical AI?




**B** is altered imperceptibly

# How can you fool AI?

## Let's ask our AI ?




We don't see a difference, yet the AI is completely certain the left (original) image is melanoma (cancer), yet the right image (imperceptibly altered) is predicted to be healthy!

# How can you fool AI?

## AI is fooled

- A doctor would not be fooled
- You cannot see the difference : hard to prove someone **altered** the image

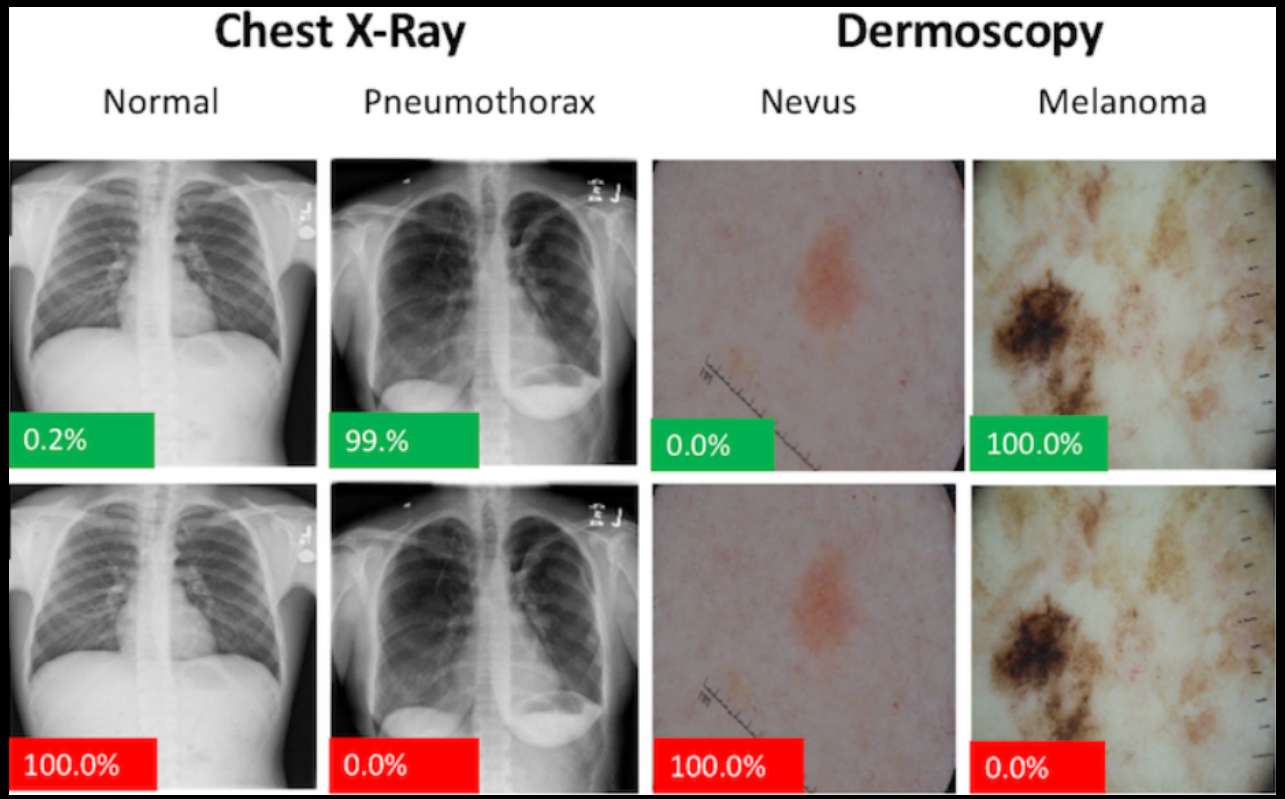



Figure courtesy of Finlayson et al [1]

# Can we make AI robust?

**Maybe, but...**

- Robustness can mean sacrificing accuracy: is that **ethical** ?
- So far no proof that robust **and** accurate AI is feasible

# Can we make AI robust?

## Learn More?

- https://adversarial-ml-tutorial.org/introduction/
- https://www.coursera.org/lecture/ai-for-everyone/adversarial-attacks-on-ai-RgA2q
- https://www.youtube.com/watch?v=CIfsB_EYsVI

# References I


Samuel G. Finlayson, Isaac S. Kohane, and Andrew L. Beam. "Adversarial Attacks Against Medical Deep Learning Systems". In: **CoRR** abs/1804.05296 (2018). arXiv: 1804.05296. URL: http://arxiv.org/abs/1804.05296.

Irene Papanicolas, Liana R Woskie, and Ashish K Jha. "Health care spending in the United States and other high-income countries". In: **Jama** 319.10 (2018), pp. 1024–1039.

Xiaosong Wang et al. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: **2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)**. 2017, pp. 3462–3471.