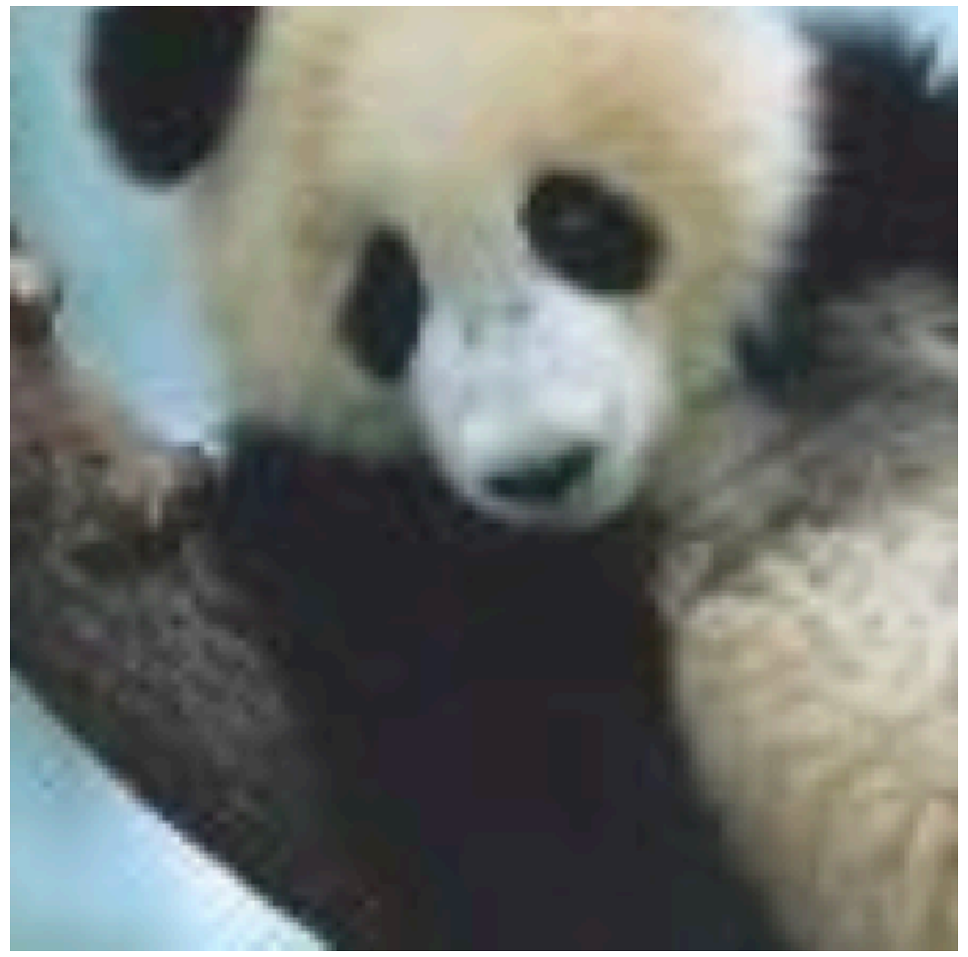
Hacking Computer Brains

🔥 Nobody panic! 🔥

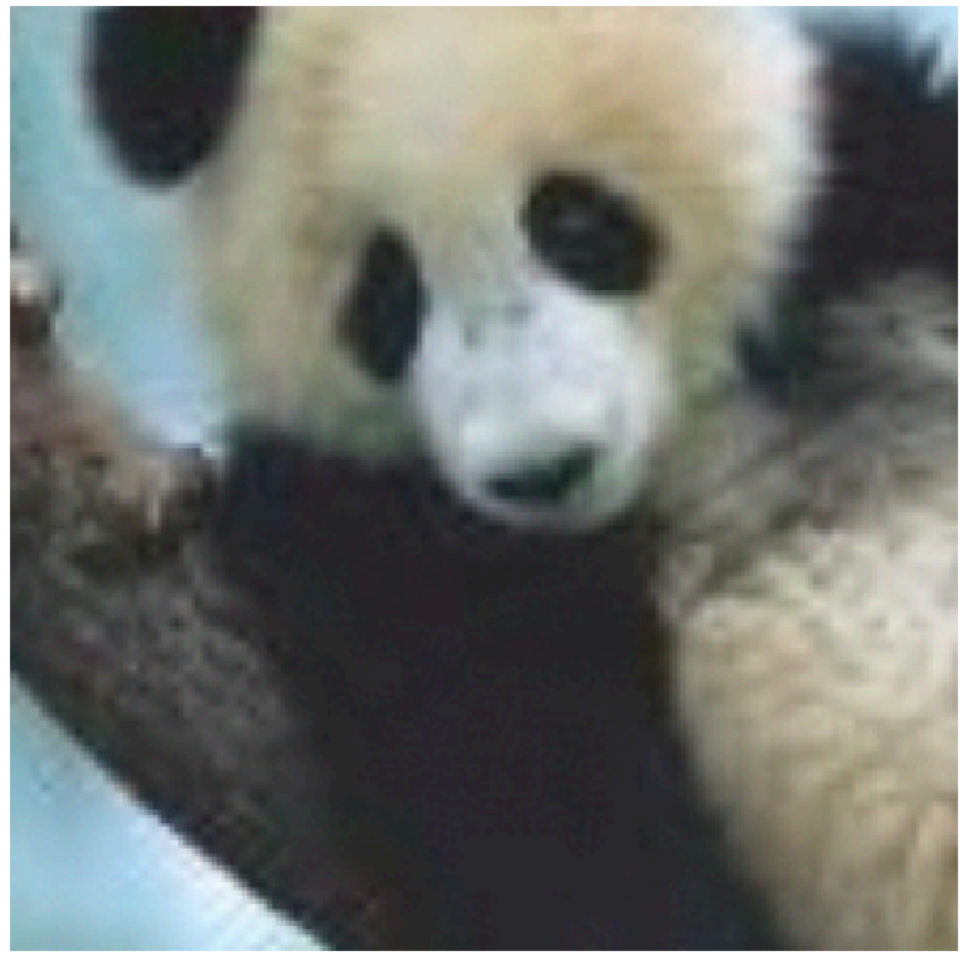
By Dan

Some problems

- Deep Neural Networks (DNNs) are great — they’re pretty useful things
- But they can be exploited in trivial ways with surprising consequences



Panda

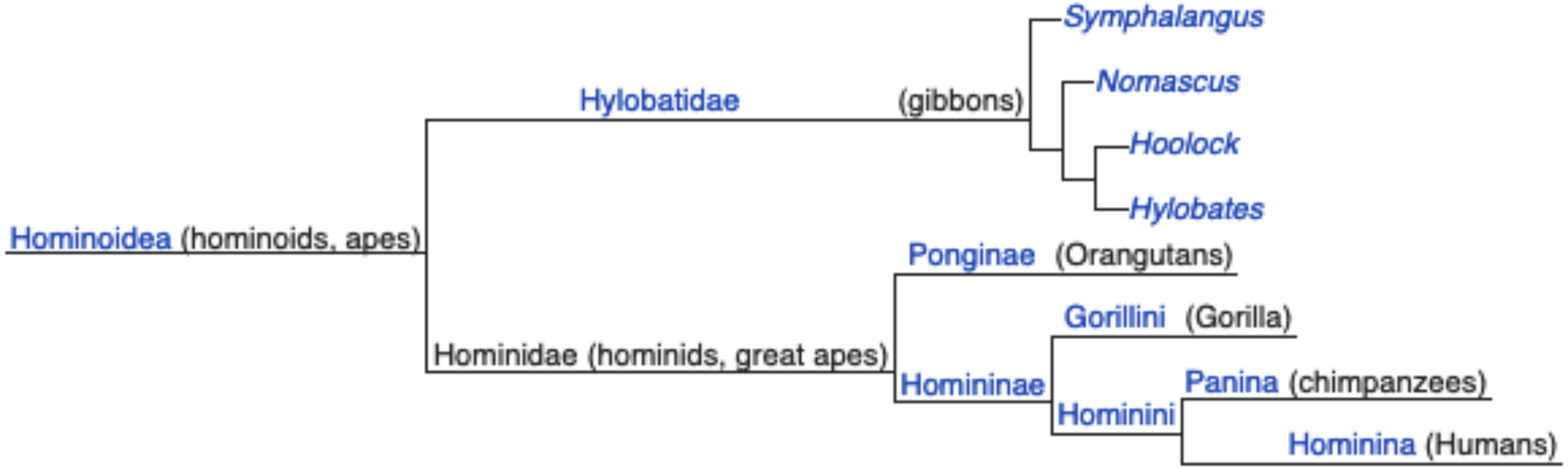


Gibbon

Some problems

- Deep Neural Networks (DNNs) are great — they're pretty useful things

- But they can be exploited in trivial ways with surprising consequences

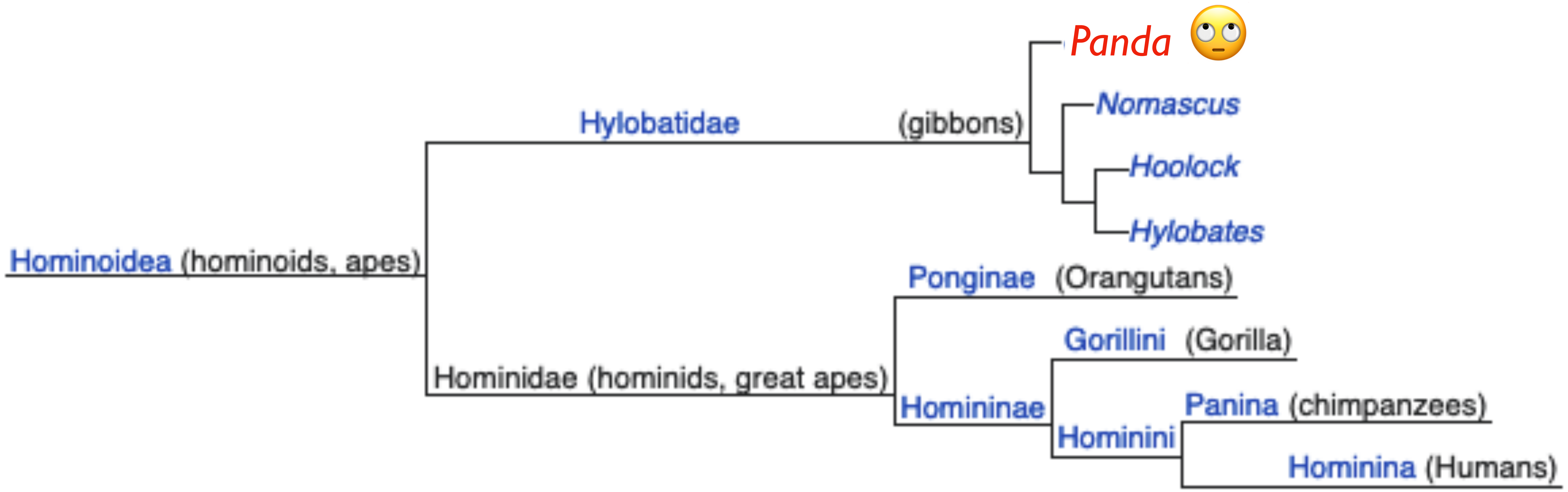


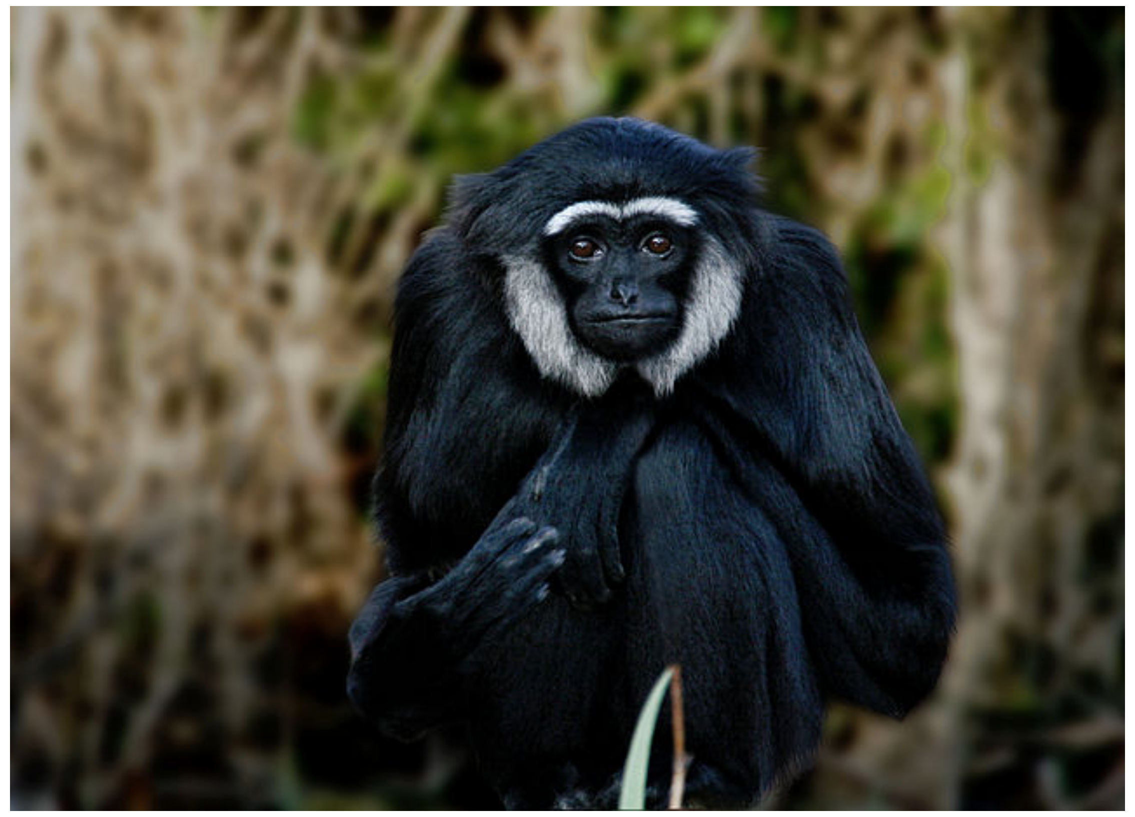
[1]

Some problems

- Deep Neural Networks (DNNs) are great — they're pretty useful things

- But they can be exploited in trivial ways with surprising consequences

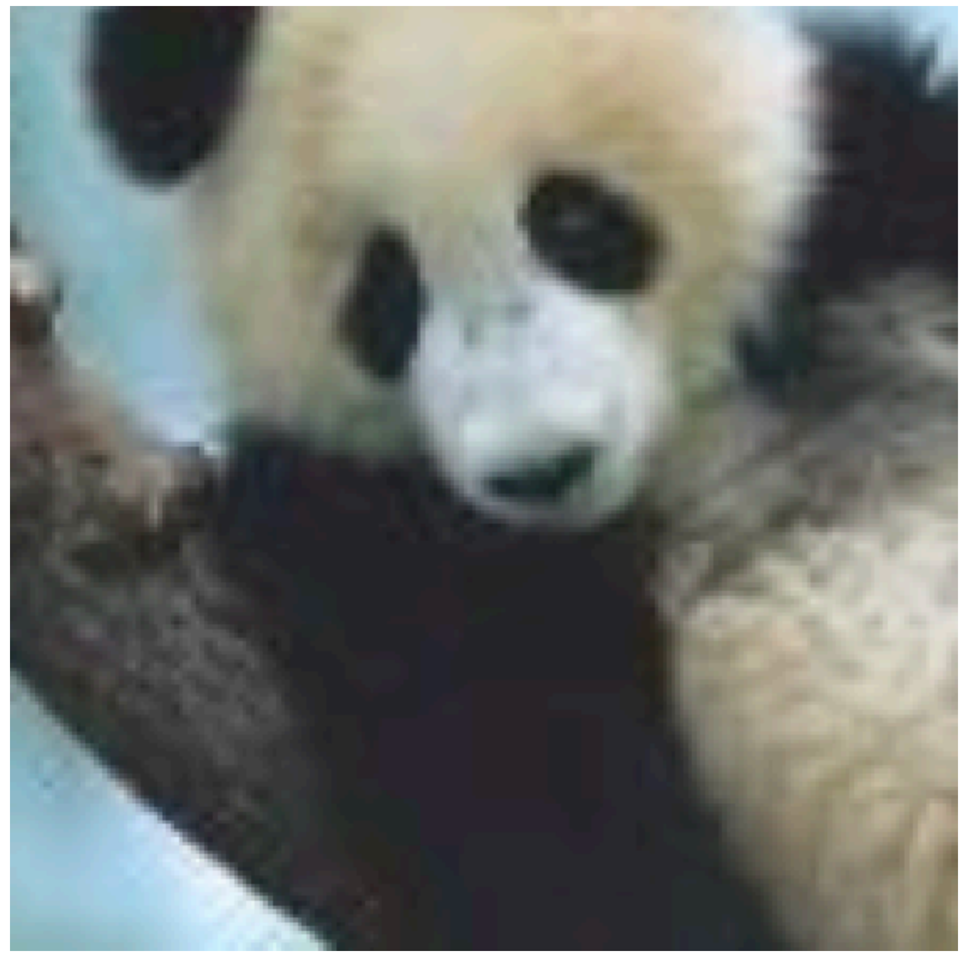




[1]

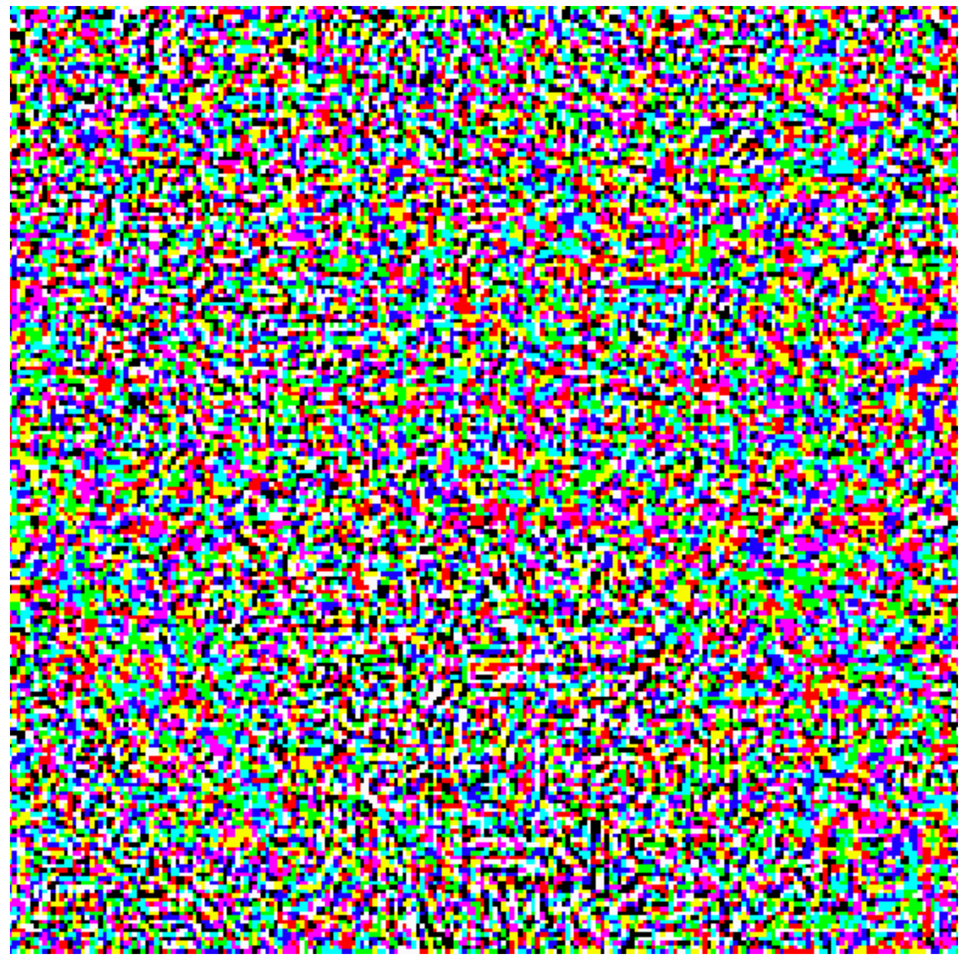
The difference

An adversarial perturbation has been applied



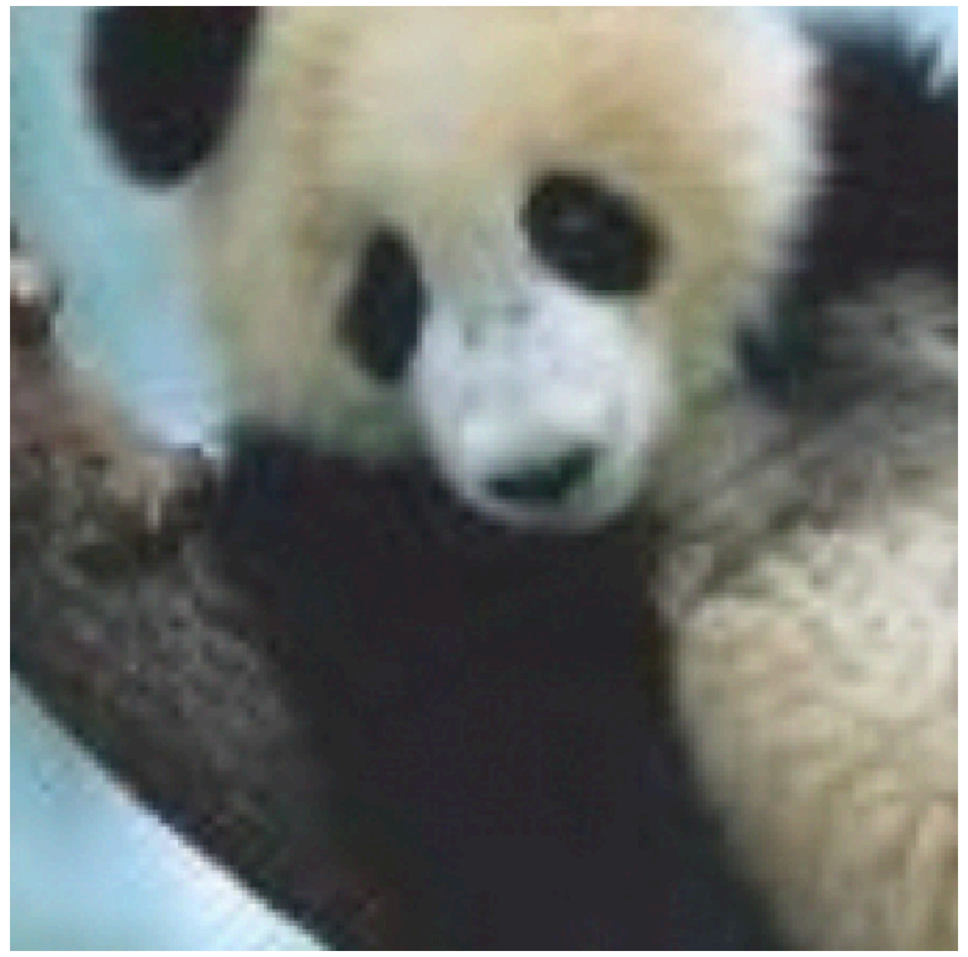
Panda

+



Nematode

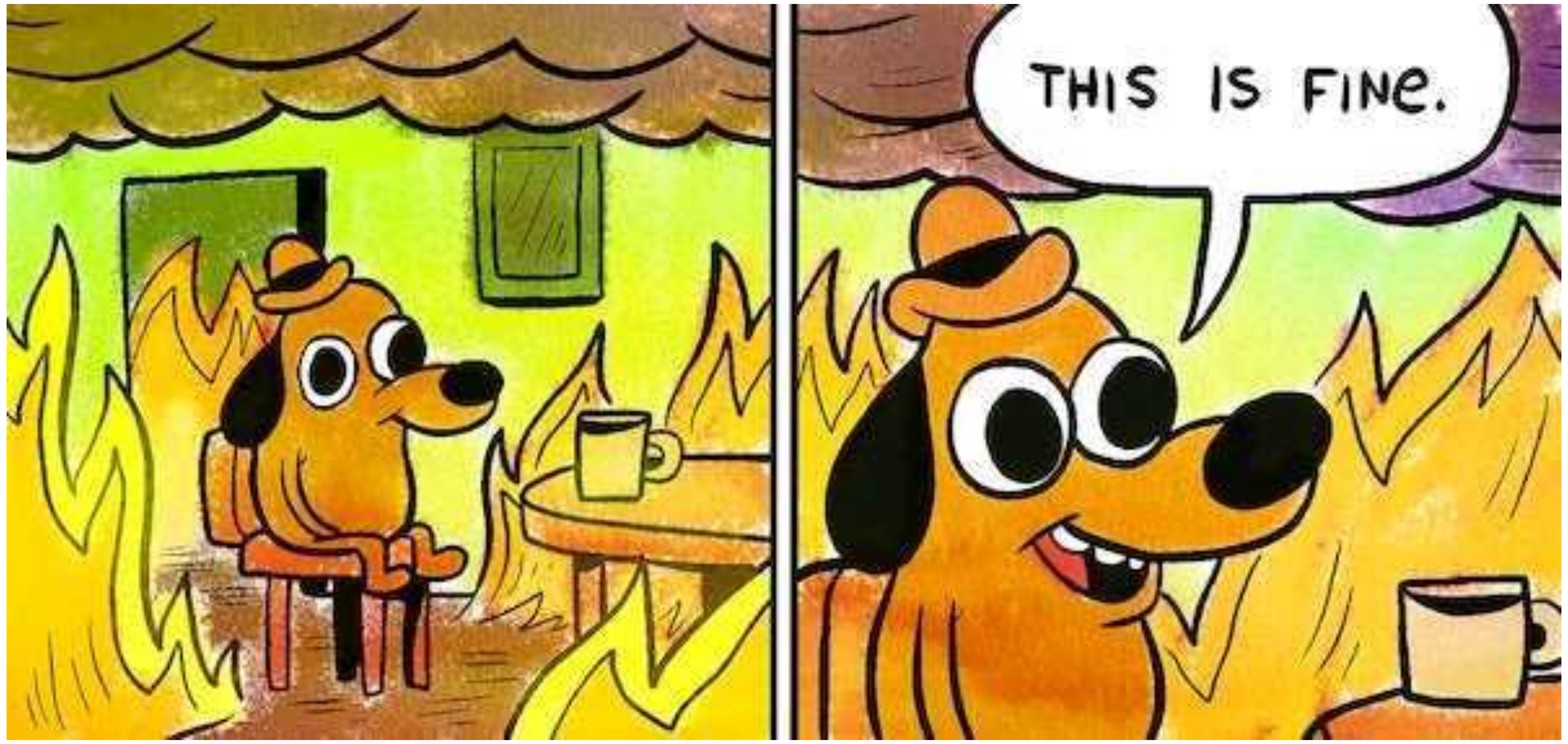
=



Gibbon [2]

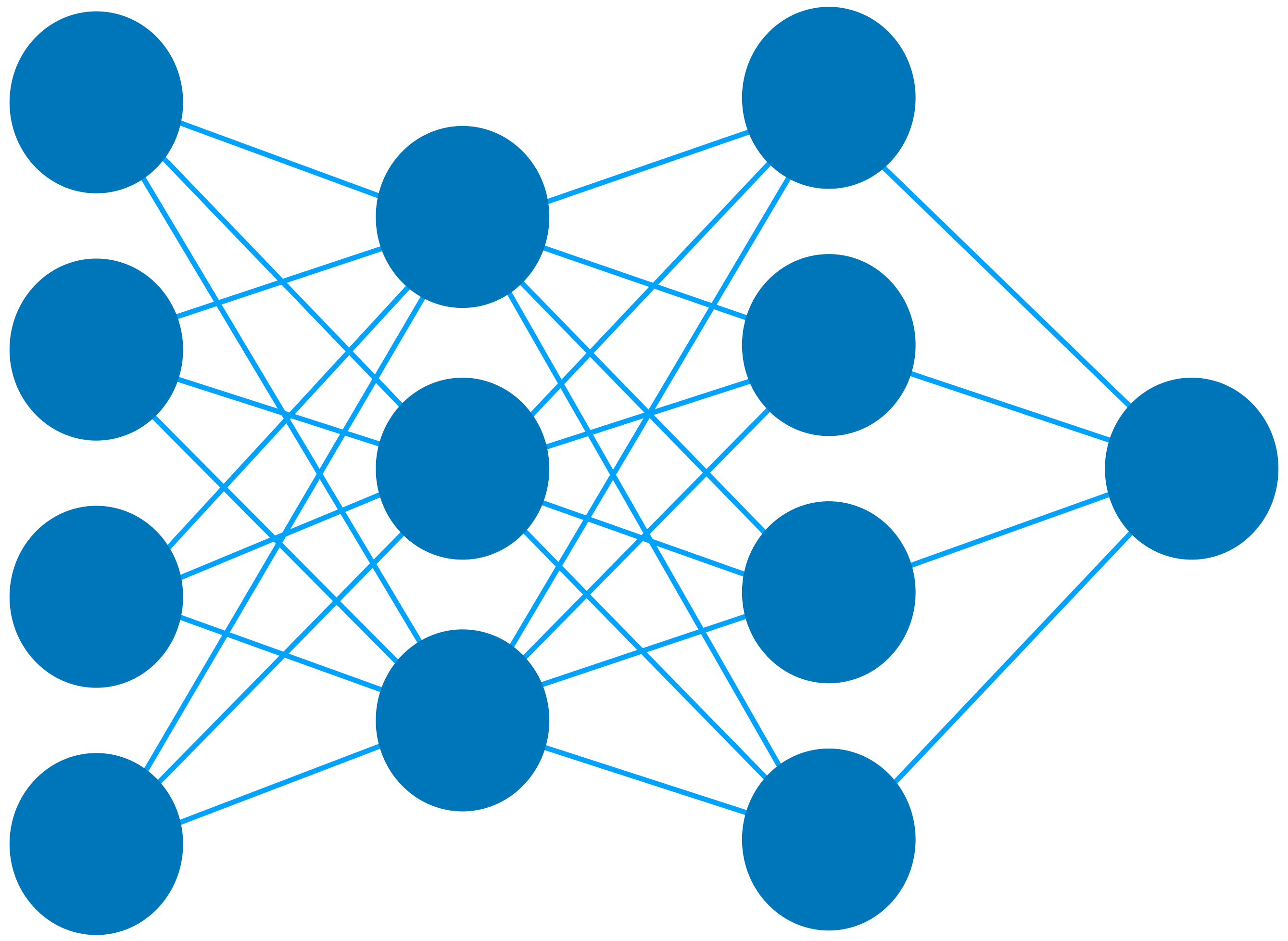
The difference

An adversarial perturbation has been applied



Think again

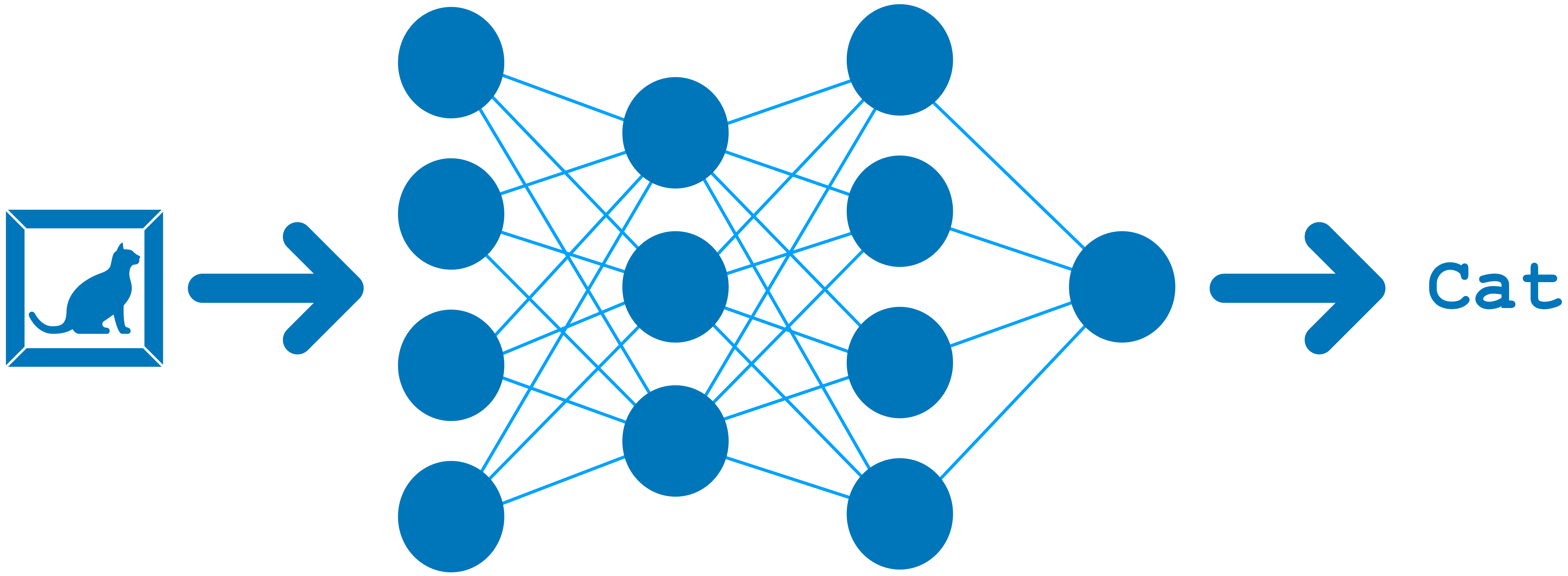
We can create these attacks by using backpropagation



A simple DNN

Think again

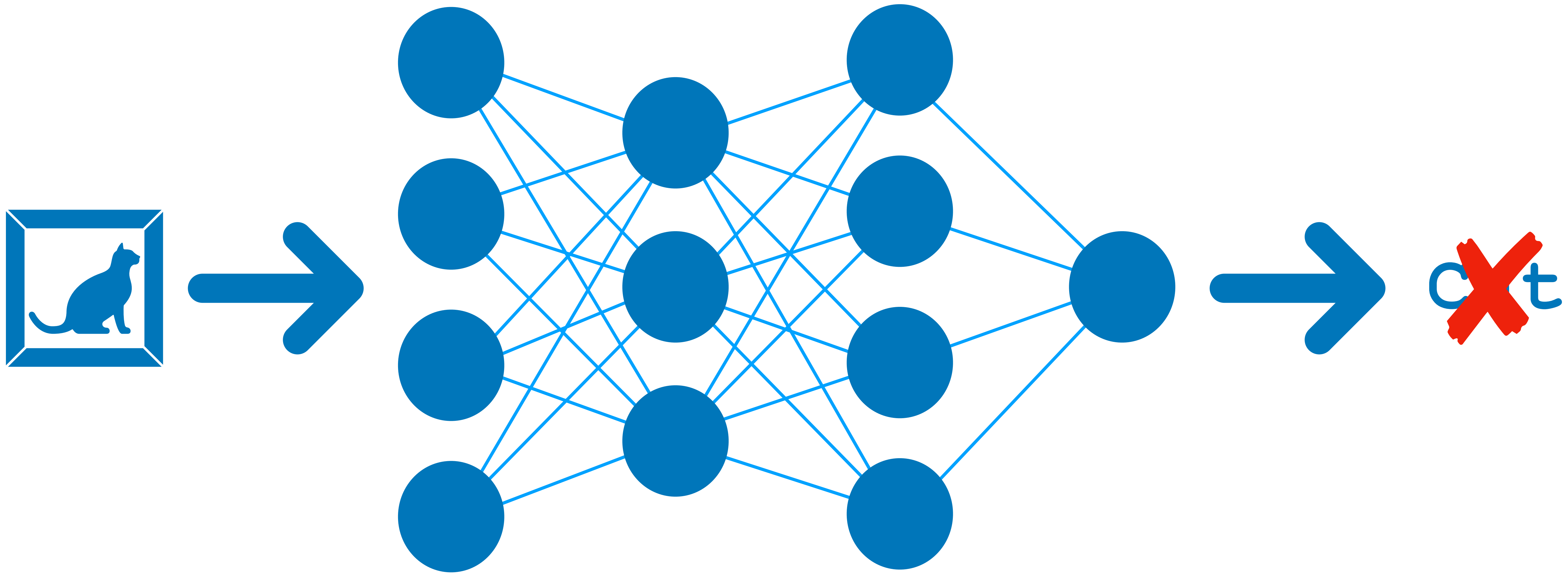
We can create these attacks by using backpropagation



We provide some input and the DNN gives us an output

Think again

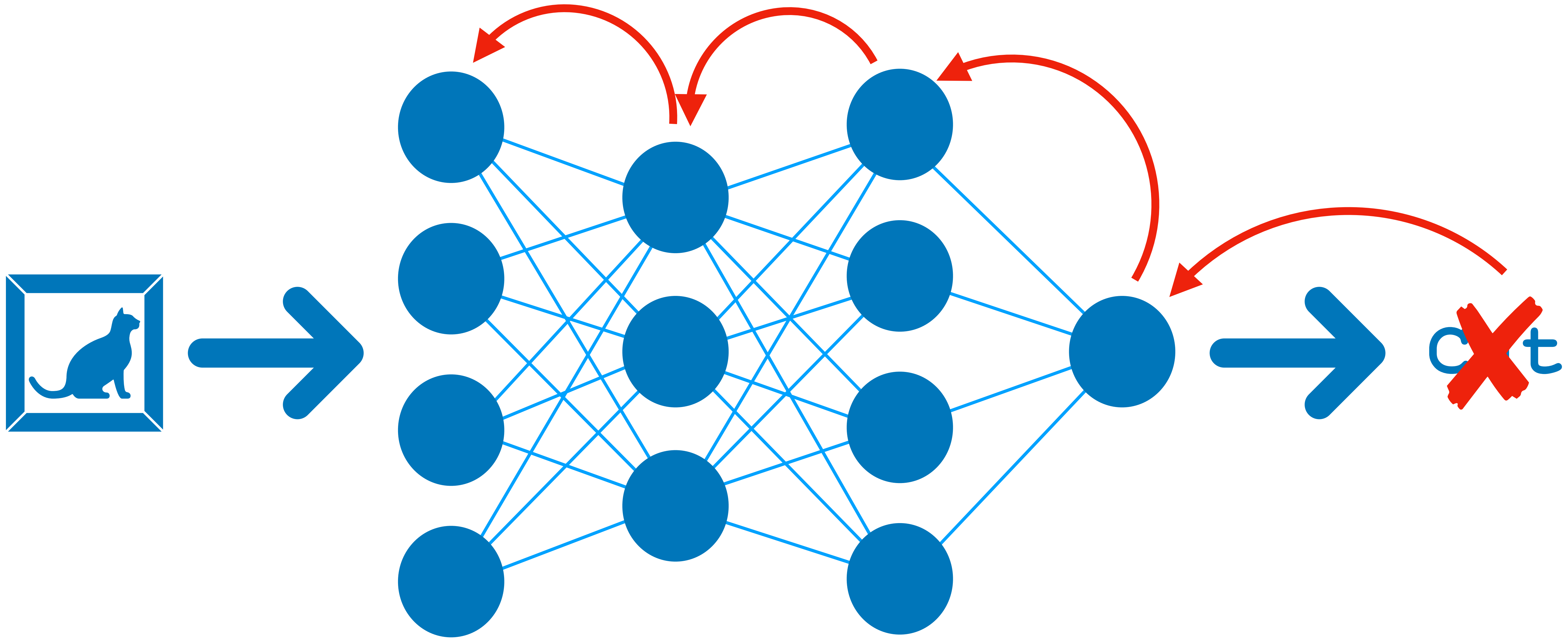
We can create these attacks by using backpropagation



What if we want to encourage a misclassification?

Think again

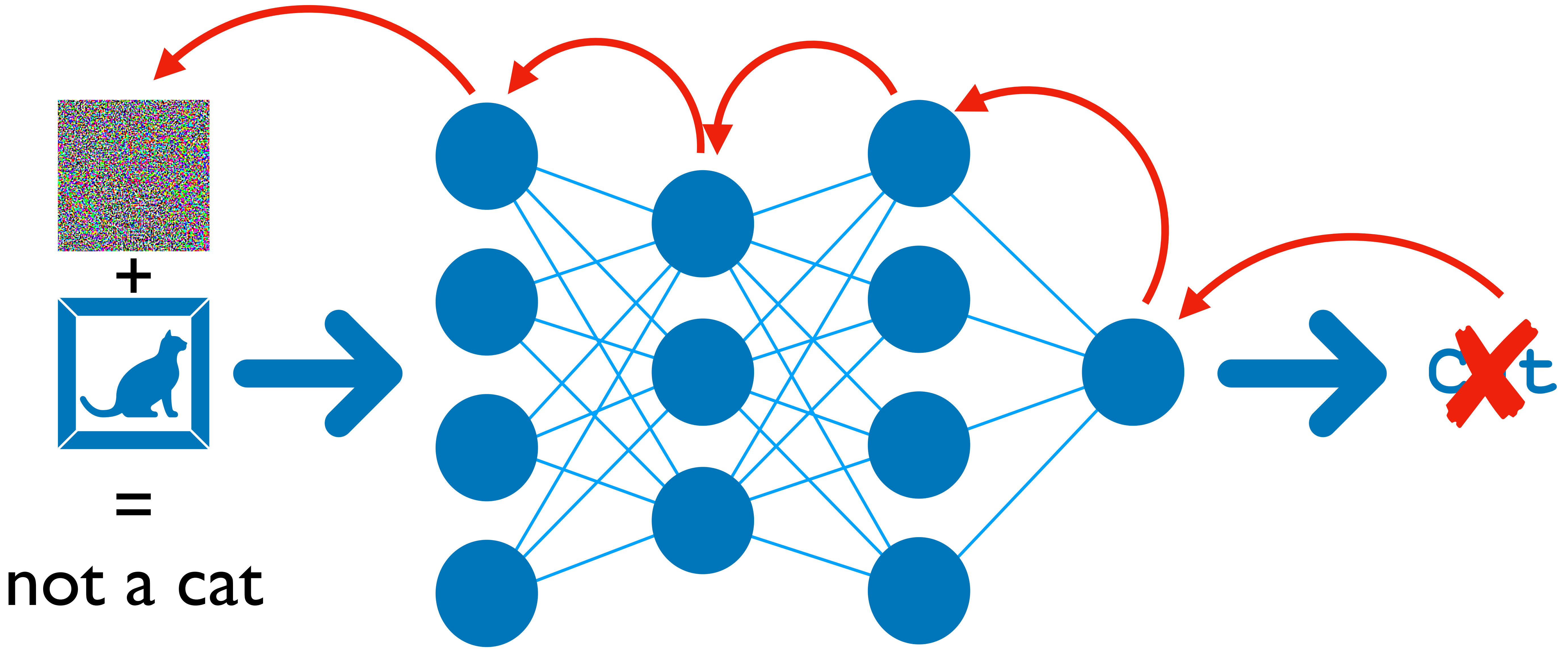
We can create these attacks by using backpropagation



We backpropagate the maximised cost function

Think again

We can create these attacks by using backpropagation



To create an adversarial perturbation

But why? 🤔

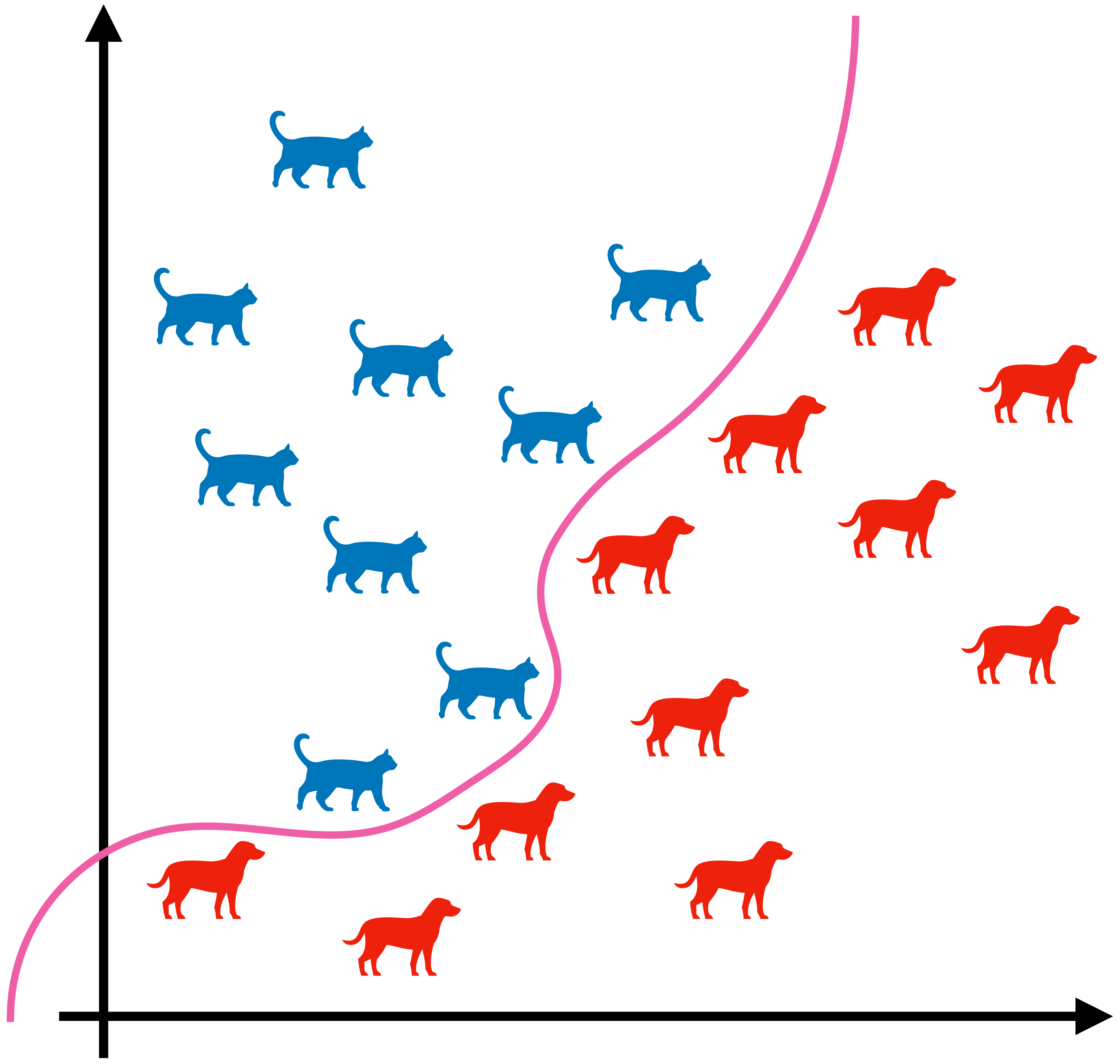
There could be a whole presentation on this…

Essentially DNNs are *“function approximators”*

But why? 🤔

There could be a whole presentation on this…

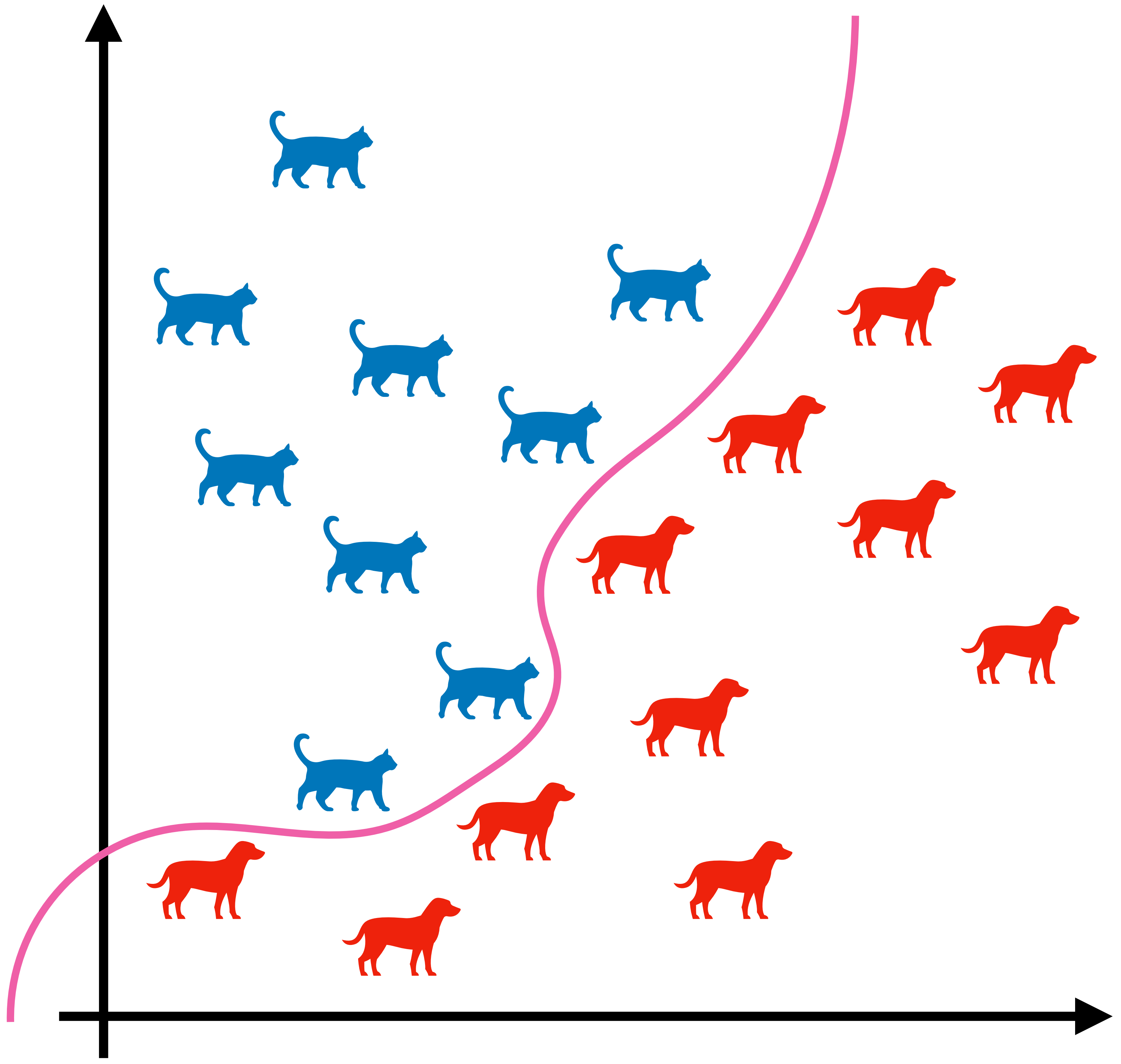
Essentially DNNs are *“function approximators”*

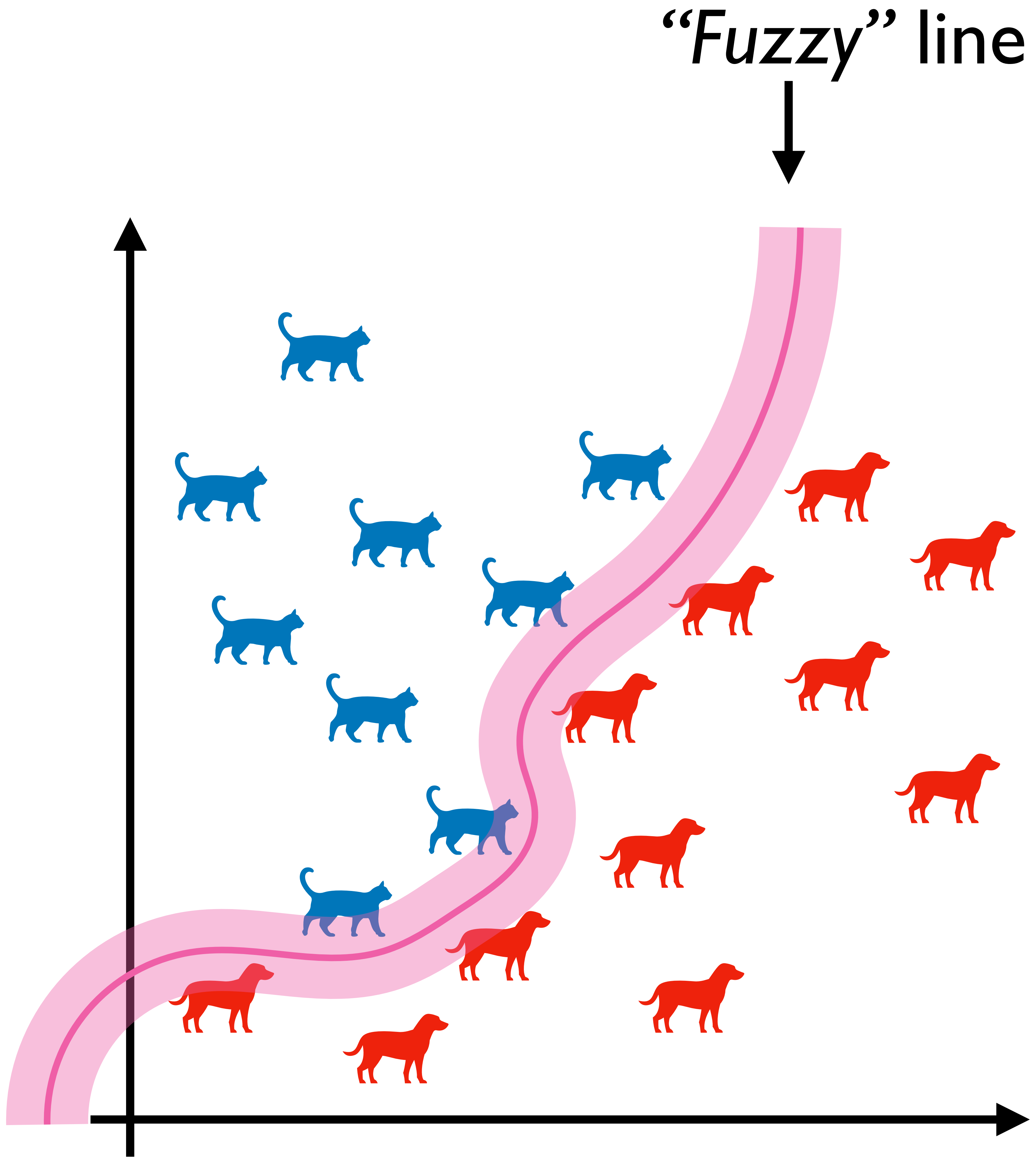


But why? 🤔

There could be a whole presentation on this…

Essentially DNNs are *“function approximators”*





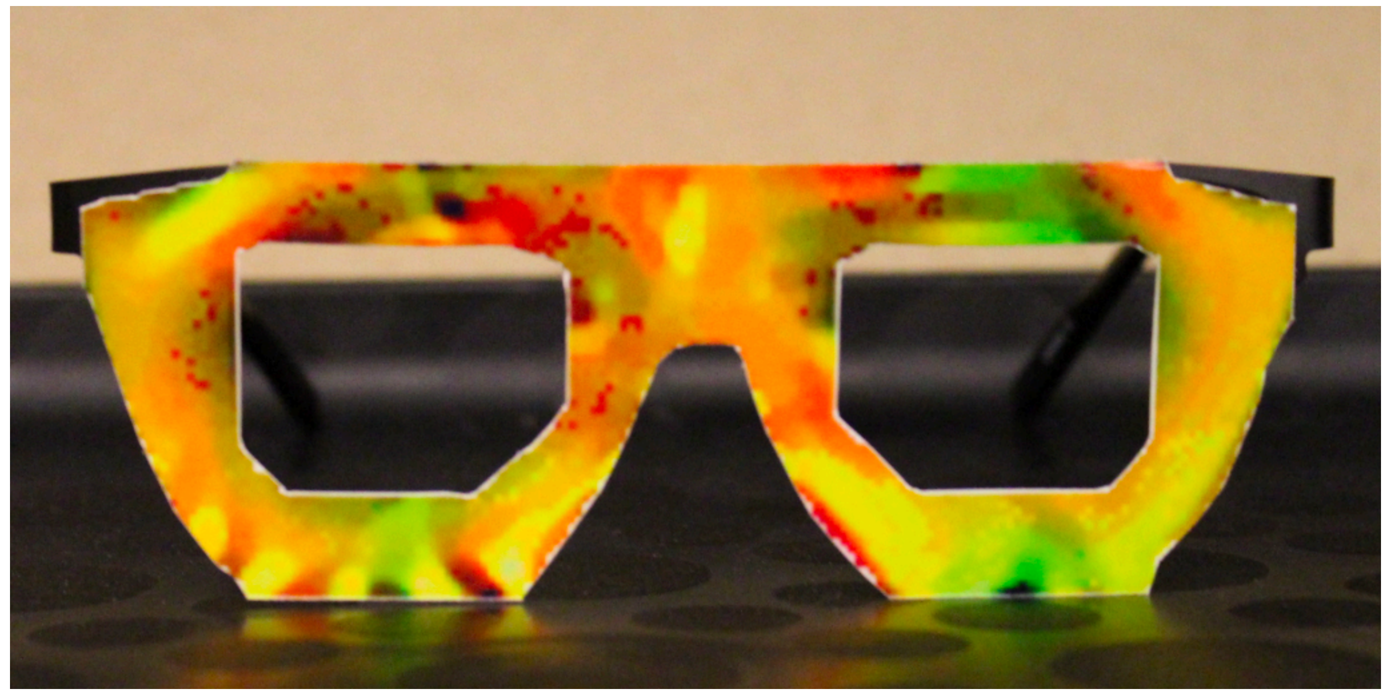
So what?

Changing a cat to “not a cat” seems harmless, but this has some real consequences:

So what?

Changing a cat to "not a cat" seems harmless, but this has some real consequences:

- Facial recognition systems

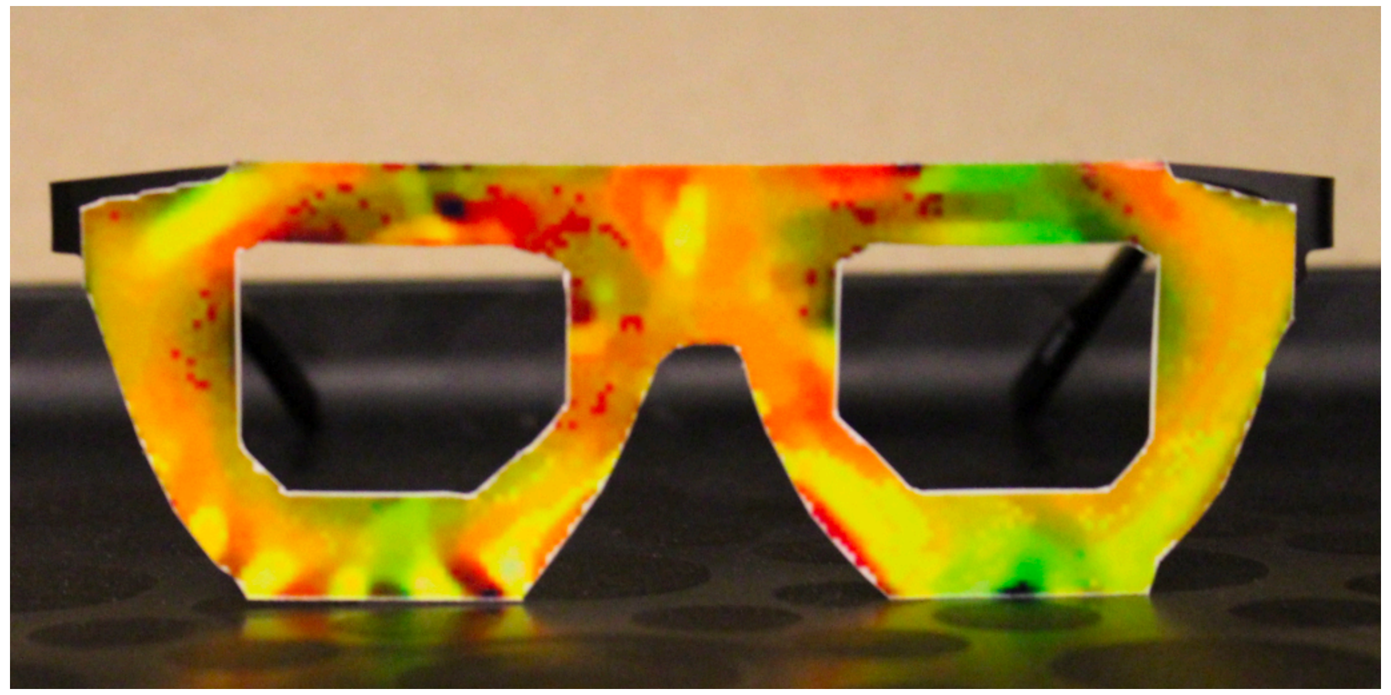


[3]

So what?

Changing a cat to “not a cat” seems harmless, but this has some real consequences:

- Facial recognition systems



[3]

- Autonomous vehicles

[4]

So what?

These attacks can be performed on any DNN

Provided the input space has a high enough dimension, they can be invisible

We'll leave it there…

Thanks for listening
Feel free to ask any questions!

dxf209@cs.bham.ac.uk

Discord: dxf

References

[1] https://en.wikipedia.org/wiki/Gibbon

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015, December 19). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. http://arxiv.org/abs/1412.6572

[3] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. Proceedings of the ACM Conference on Computer and Communications Security, 24-28-Octo, 1528–1540. https://doi.org/10.1145/2976749.2978392

[4] Povolny, Steve, and Shivangee Trivedi. 2020. "Model Hacking Adas to Pave Safer Roads for Autonomous Vehicles." McAfee Advanced Threat Research; https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

Resources

A good place to start is the original paper by Szegedy *et al.* — https://arxiv.org/abs/1312.6199

For the more practically minded TensorFlow and PyTorch both have tutorials on basic adversarial attacks:
- TensorFlow — https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
- PyTorch — https://pytorch.org/tutorials/beginner/fgsm_tutorial.html?highlight=fgsm

A whole suite of adversarial attacks can be found in the CleverHans python library (does NOT support TF 2.x) — https://github.com/tensorflow/cleverhans

For understanding of Neural Network fundamentals see 3B1B — https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

For a deeper understanding of Neural Networks see Deep Learning — https://www.deeplearningbook.org/