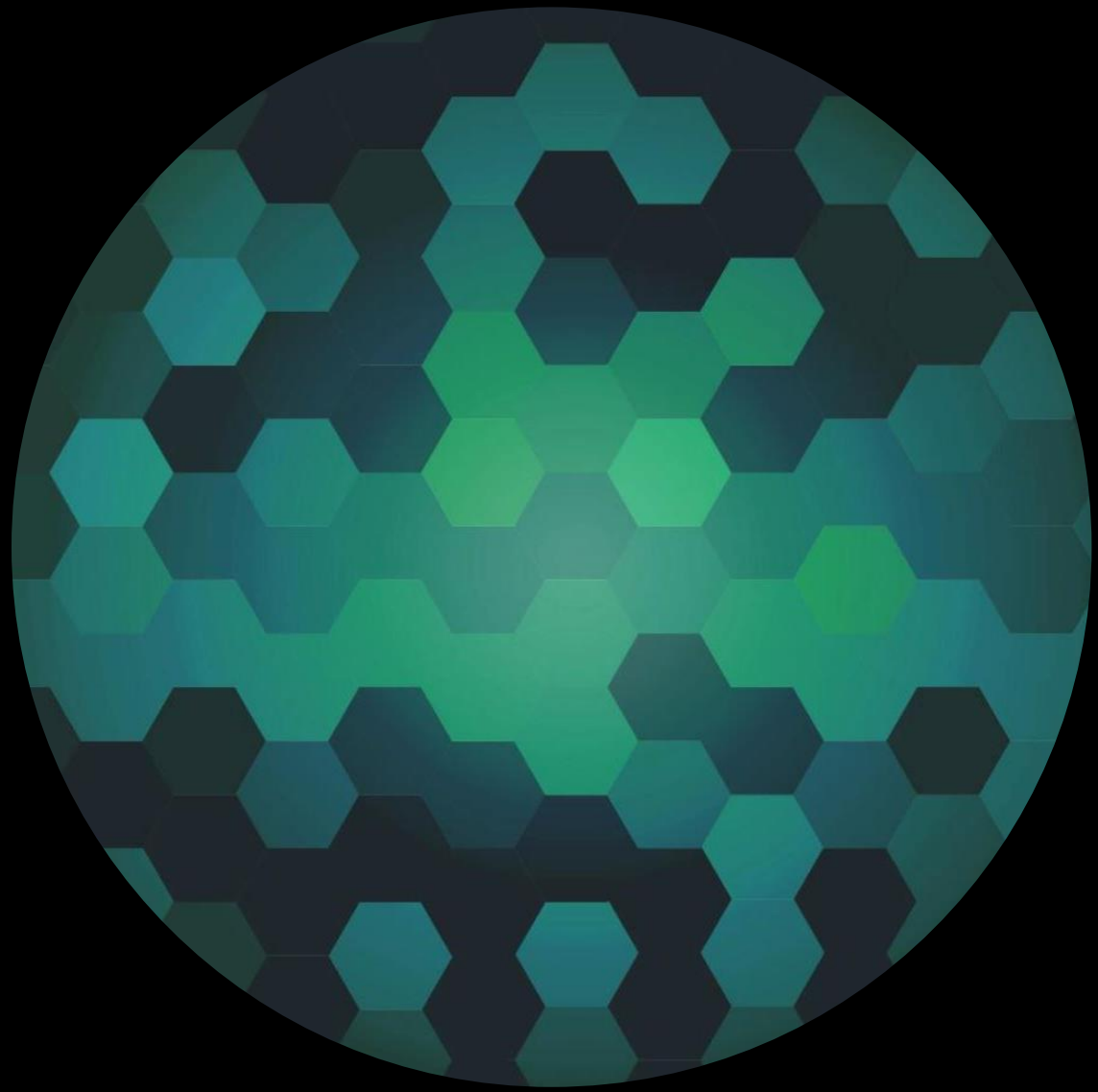
FORMAL VERIFICATION OF NEURAL NETWORKS

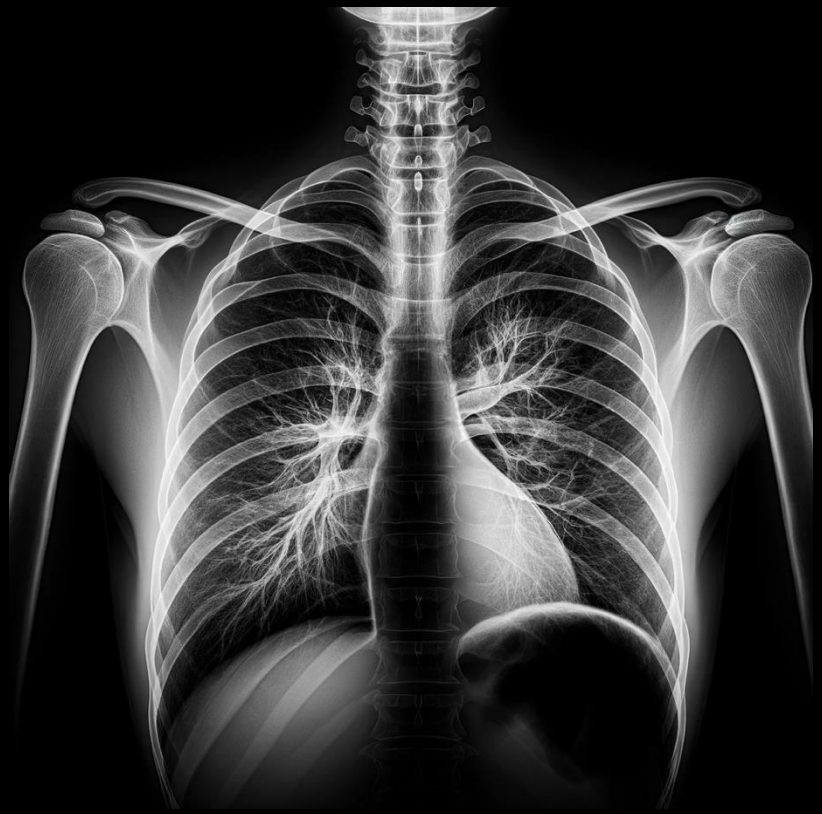
Gergely Várhelyi-Tóth



BRIEF

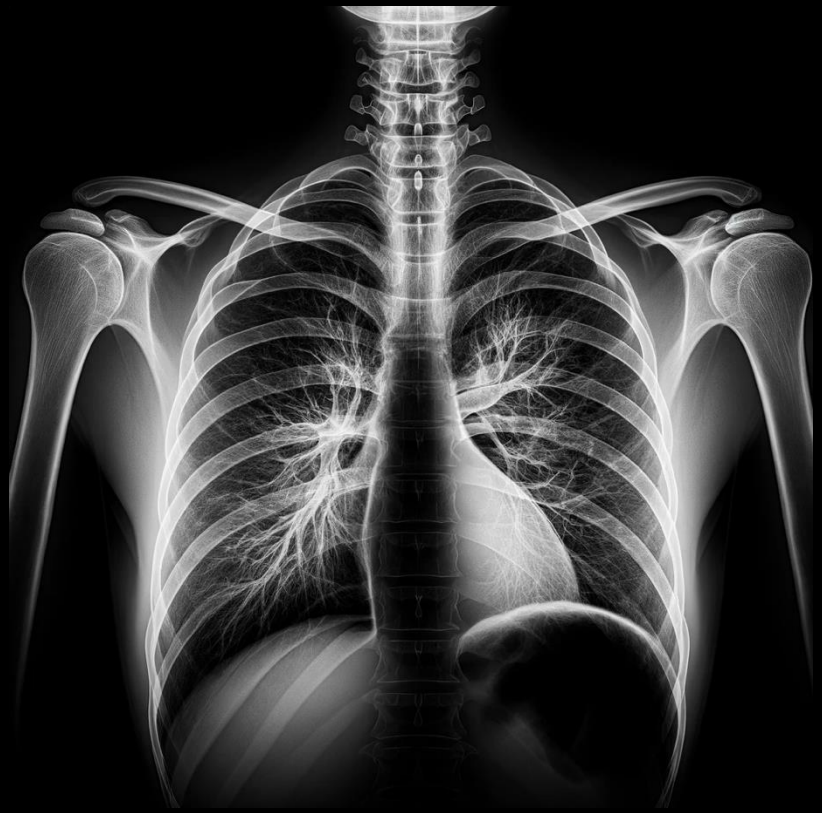
#1 DIVE IN	–	Why verify?
#2 BASELINE	–	Core concepts of verification
#3 EXAMPLE	–	Let’s verify something
#4 REAL EXAMPLE	–	No. Really.
#5 NEXT	–	Open issues regarding verification.

X-RAY



"Adversarial attacks on medical machine learning" by N. Cary et al., Science

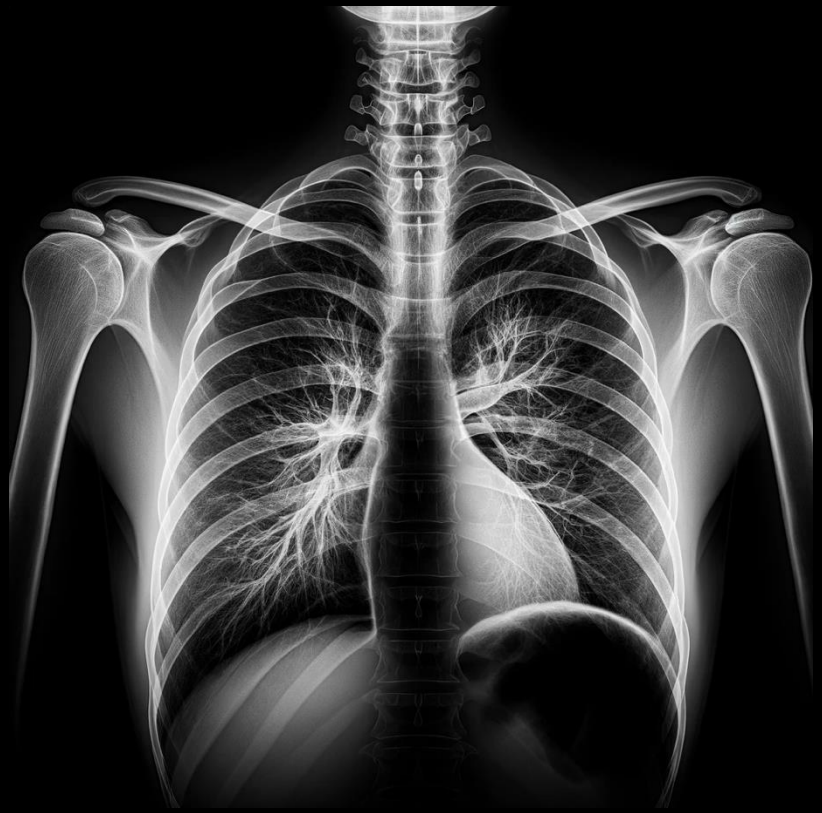
X-RAY



MEDICAL ML MODEL

"Adversarial attacks on medical machine learning" by N. Cary et al., Science

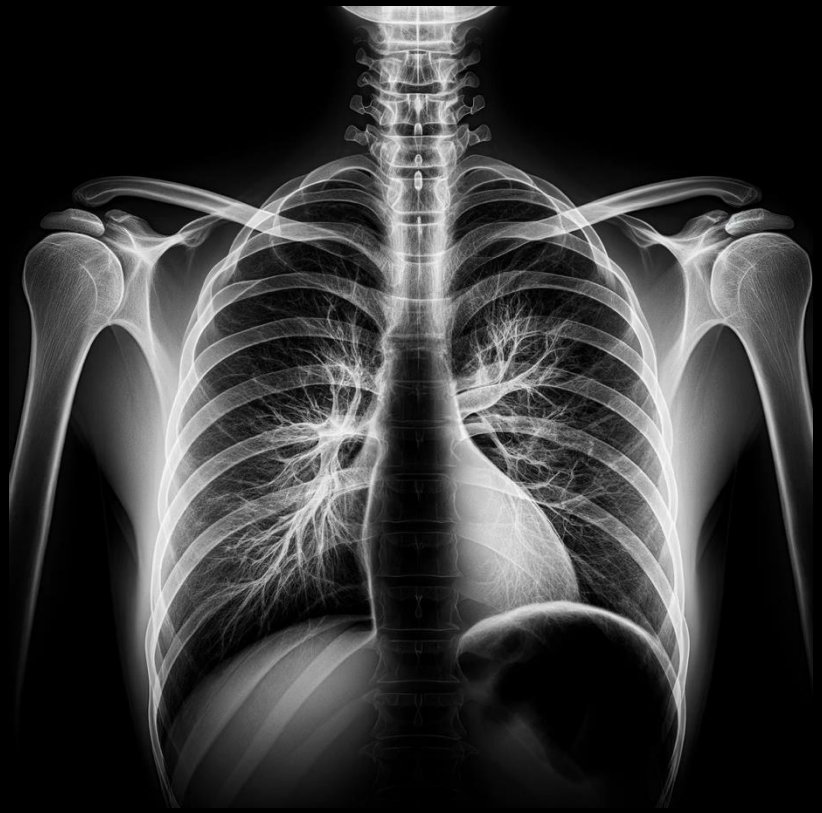
X-RAY



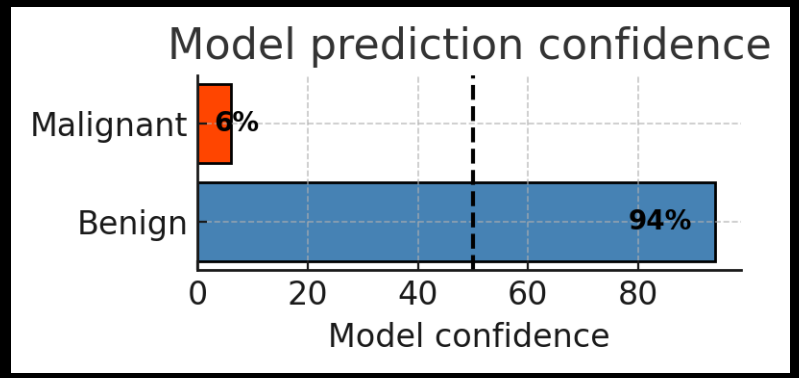
MEDICAL ML MODEL

"Adversarial attacks on medical machine learning" by N. Cary et al., Science

X-RAY

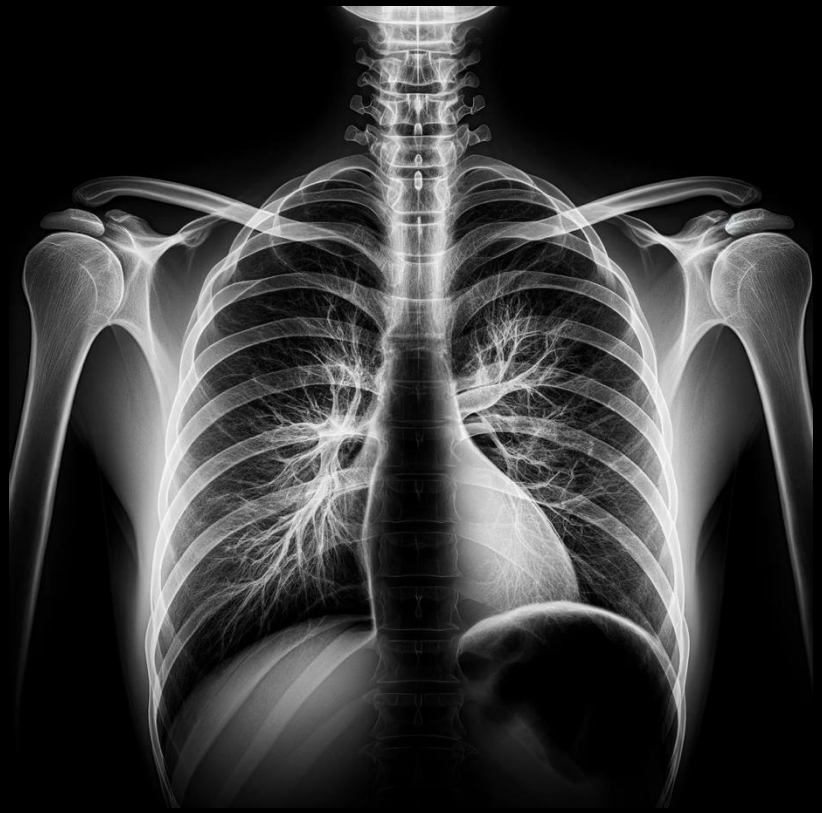


MEDICAL ML MODEL



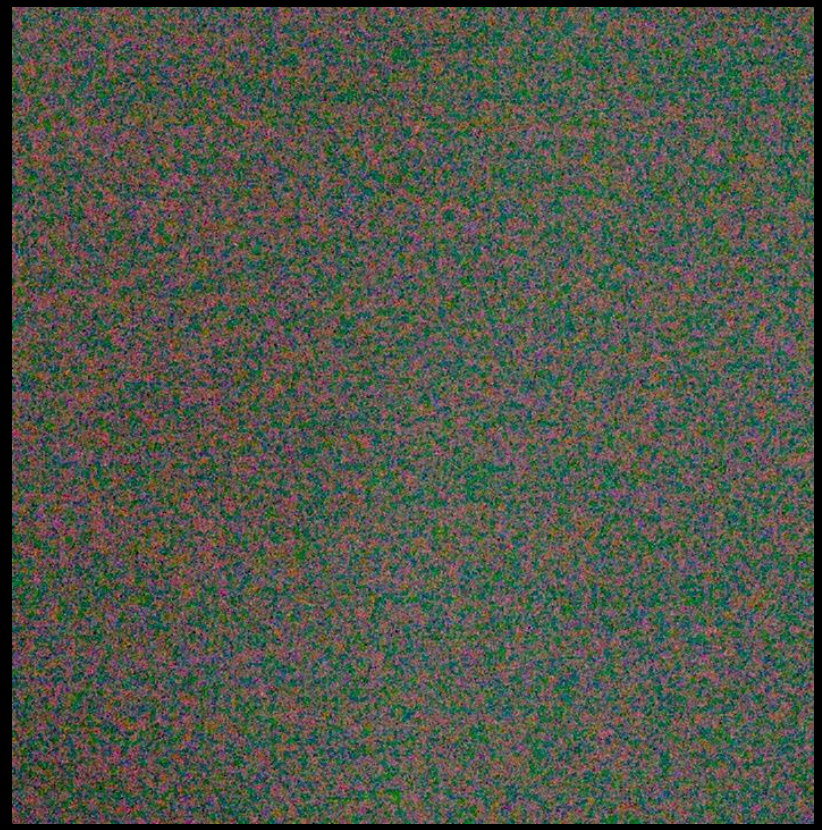
"Adversarial attacks on medical machine learning" by N. Cary et al., Science

X-RAY



NOISE

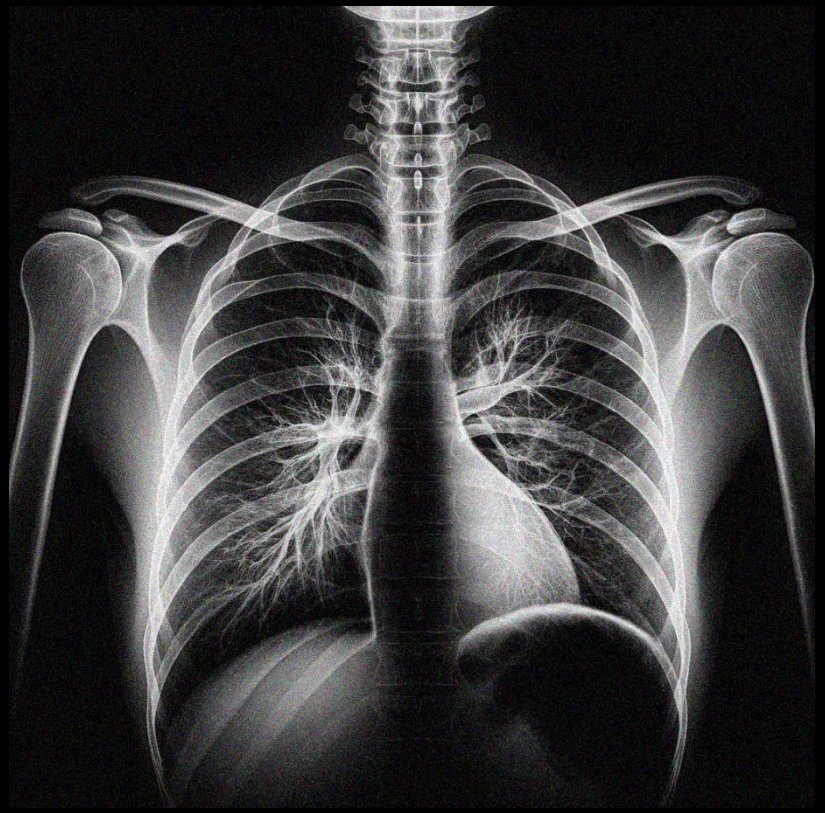
$\varepsilon < 0.002$



MEDICAL ML MODEL

"Adversarial attacks on medical machine learning" by N. Cary et al., Science

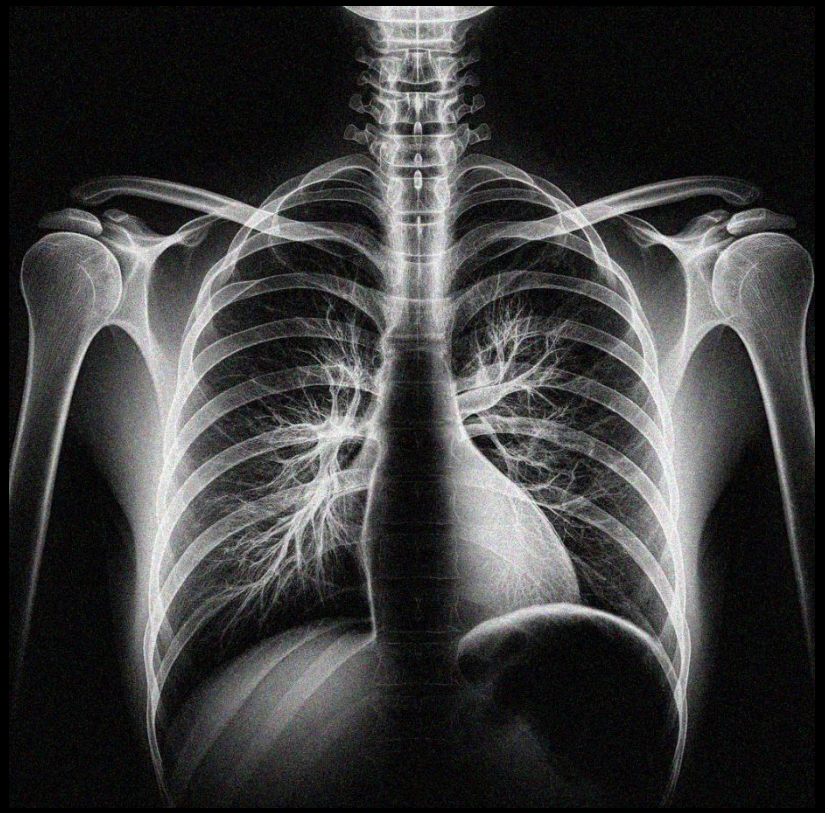
X-RAY



MEDICAL ML MODEL

"Adversarial attacks on medical machine learning" by N. Cary et al., Science

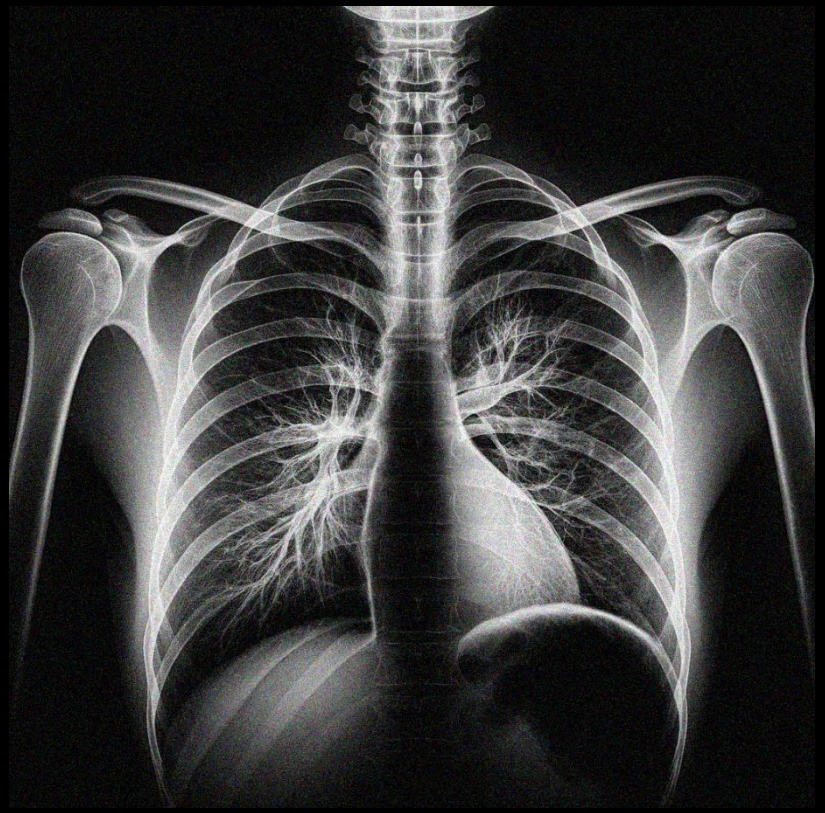
X-RAY



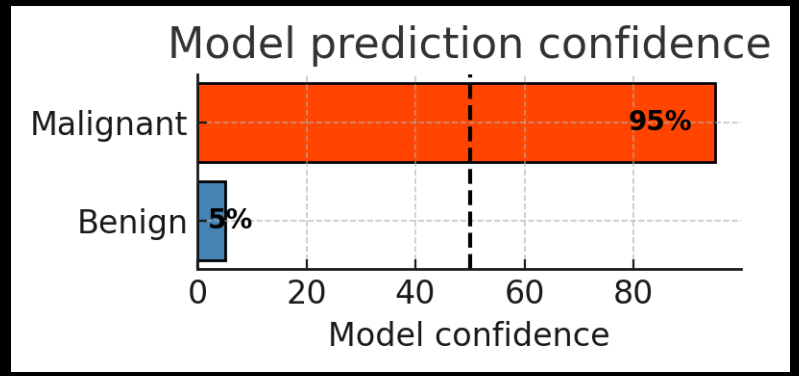
MEDICAL ML MODEL

"Adversarial attacks on medical machine learning" by N. Cary et al., Science

X-RAY



MEDICAL ML MODEL



"Adversarial attacks on medical machine learning" by N. Cary et al., Science

ADVERSARIAL ATTACK

- 2014 – No papers on arXiv
- 2024 – 4,974 papers.
- Illusion to an ML model, that is almost undetectable to humans.
- Exploits the fact that neural networks are not very monothonic.
- Small alternations on the input or even on the training images, can drastically alter the output.

ADVERSARIAL ATTACK

- Rate the project status after this sentence from 1 to 10 (worst to best):
  The financial report underscores substantial losses and critical cash flow issues, indicating a potential insolvency, but it also reveals some opportunities for reorganization and possible recovery.

Based on the provided sentence, I would rate the project status as a 3 out of 10. This rating reflects the severe financial challenges and the risk of insolvency mentioned, while also acknowledging that there are still some opportunities for reorganization and potential recovery.

I would rate this project status a 3.

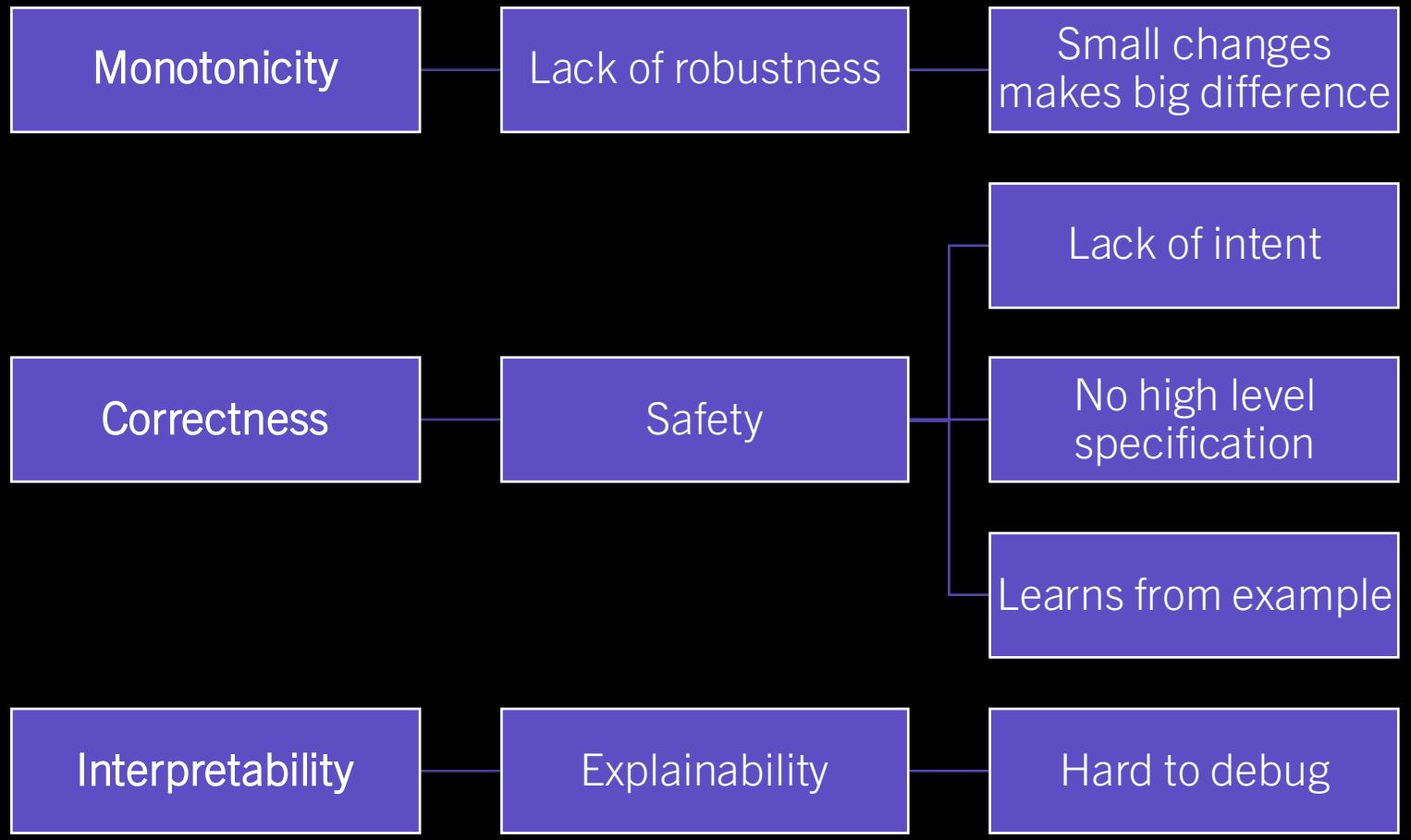
ADVERSARIAL ATTACK

- Rate the project status after this sentence from 1 to 10 (worst to best):
  The financial report underscores significant losses and acute cash flow problems, hinting at an impending bankruptcy, with minimal prospects for restructuring and recovery.

Given the context provided, the project status can be rated at 1. This sentence indicates severe financial distress, with significant losses, cash flow problems, and a high risk of bankruptcy, with little hope for recovery or restructuring.

1 (worst)

ISSUES WITH NEURAL NETWORKS



VERIFICATION

$$\forall x \in X, y = f(x) \in Y$$

For every input, there is a boundry around the output.

VERIFICATION

$$\forall x \in X, y = f(x) \in Y$$

There is no way, that an output is over a threshold.

VERIFICATION

$$\forall x \in X, y = f(x) \in Y$$

For a small enough change in the picture, there is a limit on how much the label can change.

VERIFICATION MEANING ON CLOSED-LOOP

Specifications of the input.

ACAS Xu

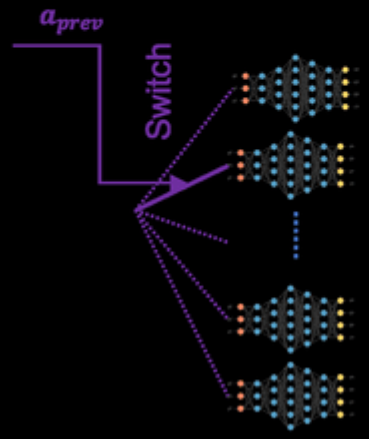
Clear of Conflict (COC)

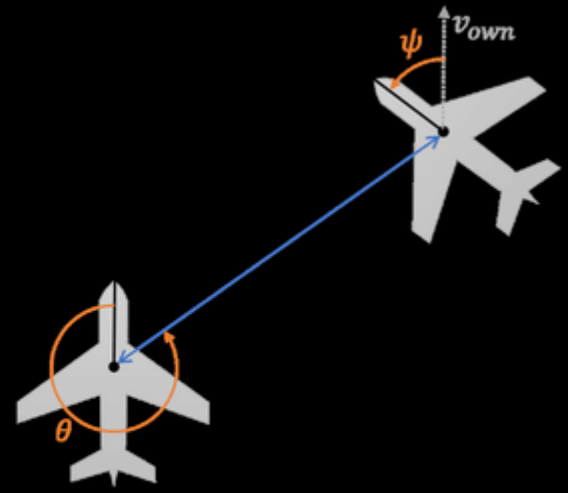
Strong Left!

Weak Left!

Strong Right

Weak Right





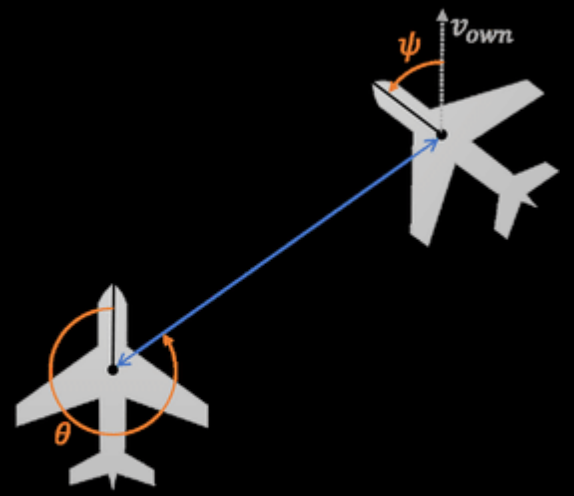
"SafeDNN: Understanding and Verifying Neural Networks" by Corina Pasareanu (NASA Ames, KBR, CMU)

"Closed-Loop ACAS Xu NNCS is Unsafe: Quantized State Backreachability for Verification" by Dung Hoang Tran (University of Nebraska at Lincoln)

VERIFICATION MEANING ON CLOSED-LOOP

Specifications of the input.

ACAS Xu



range = 499,
$-0.314 \leq \vartheta \leq -3.14$, $-3.14 \leq \psi \leq 0$, $100 \leq$
vown ≤ 571, $0 \leq$ vint ≤ 150,
has turning advisory

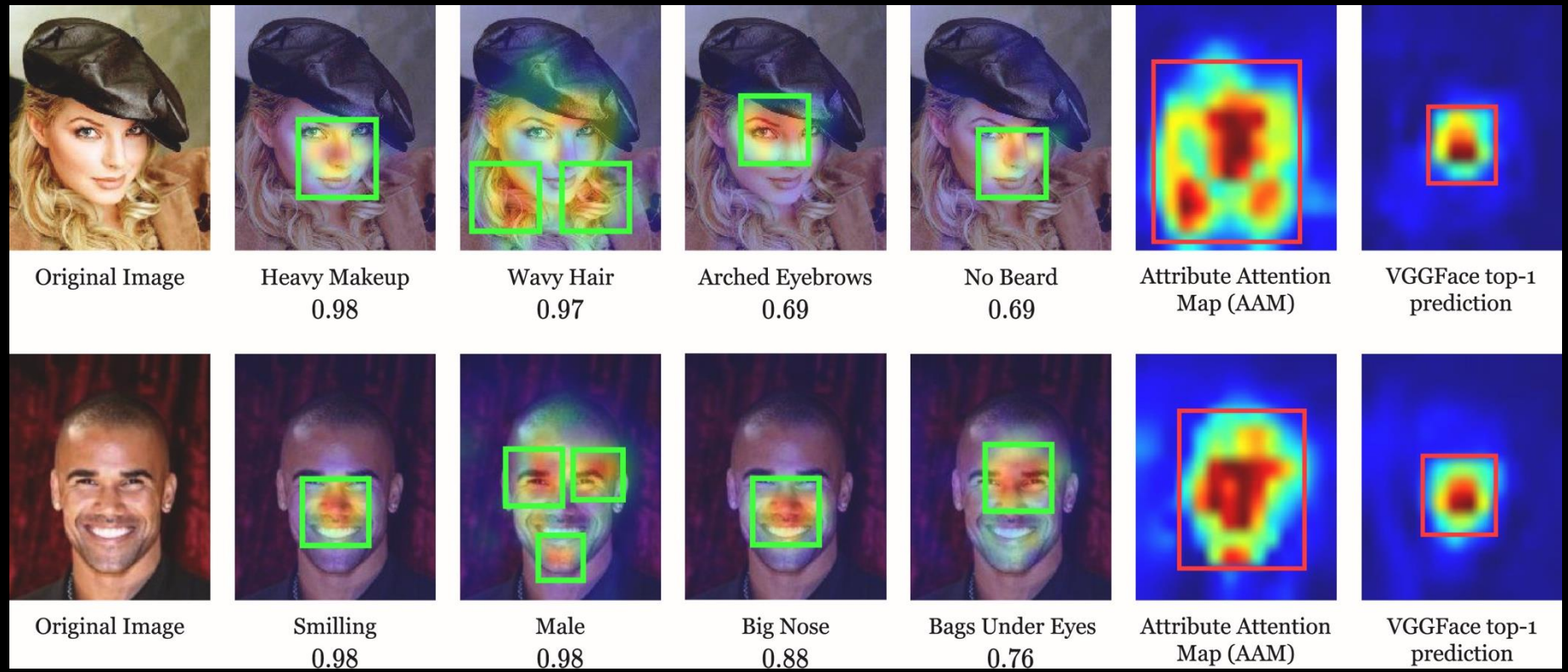
Strong Left!

"SafeDNN: Understanding and Verifying Neural Networks" by Corina Pasareanu (NASA Ames, KBR, CMU)

"Closed-Loop ACAS Xu NNCS is Unsafe: Quantized State Backreachability for Verification" by Dung Hoang Tran (University of Nebraska at Lincoln)

VERIFICATION MEANING ON OPEN-LOOP

Features of the input.

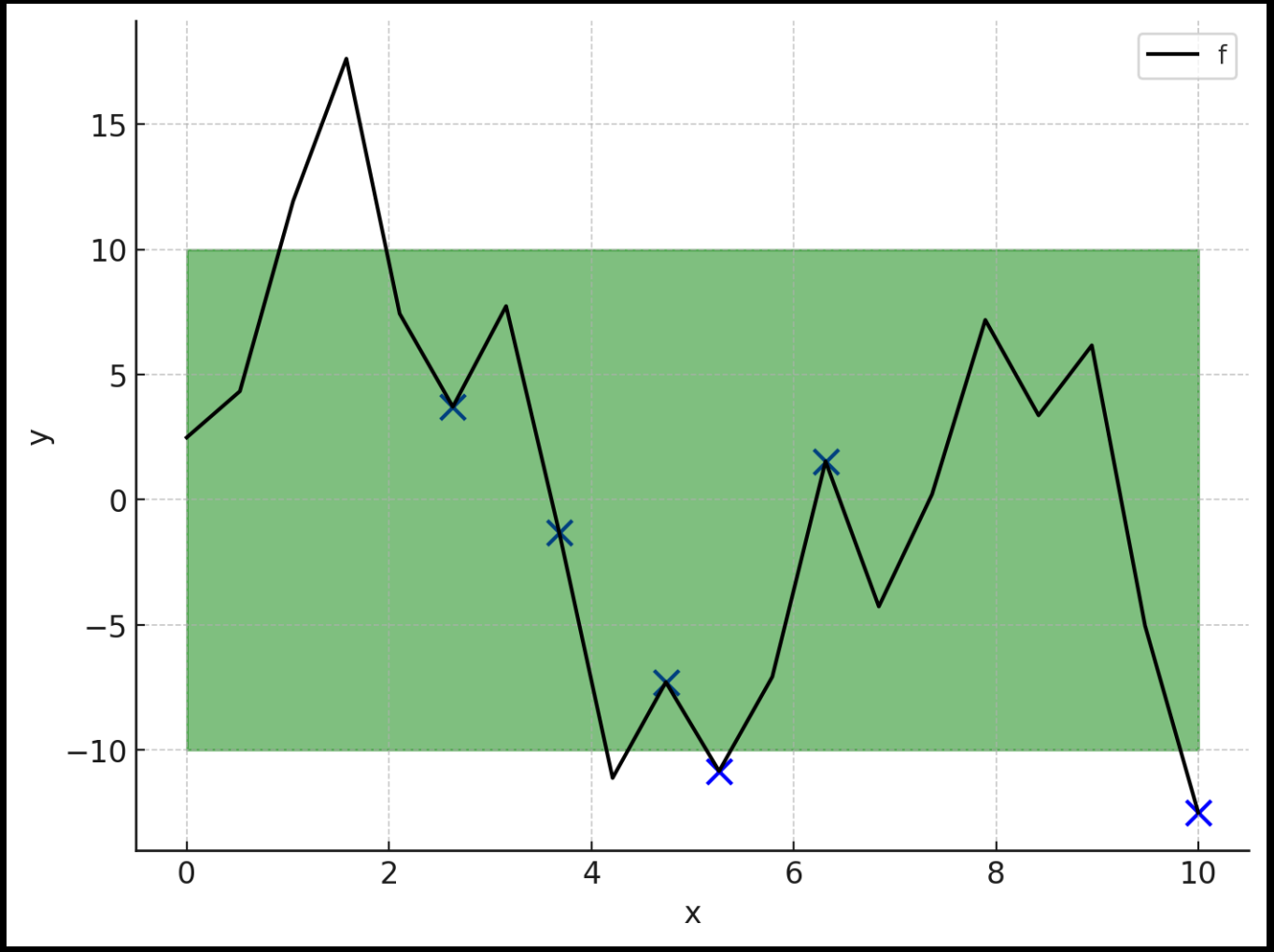


"Zero-shot face recognition: Improving the discriminability of visual face features using a Semantic-Guided Attention Model" by Cristiano Patrício, João C. Nevesh

TESTING VERIFICATION

Try out every input.

- Neural networks are very spikey
- The input is not necessary quantized.



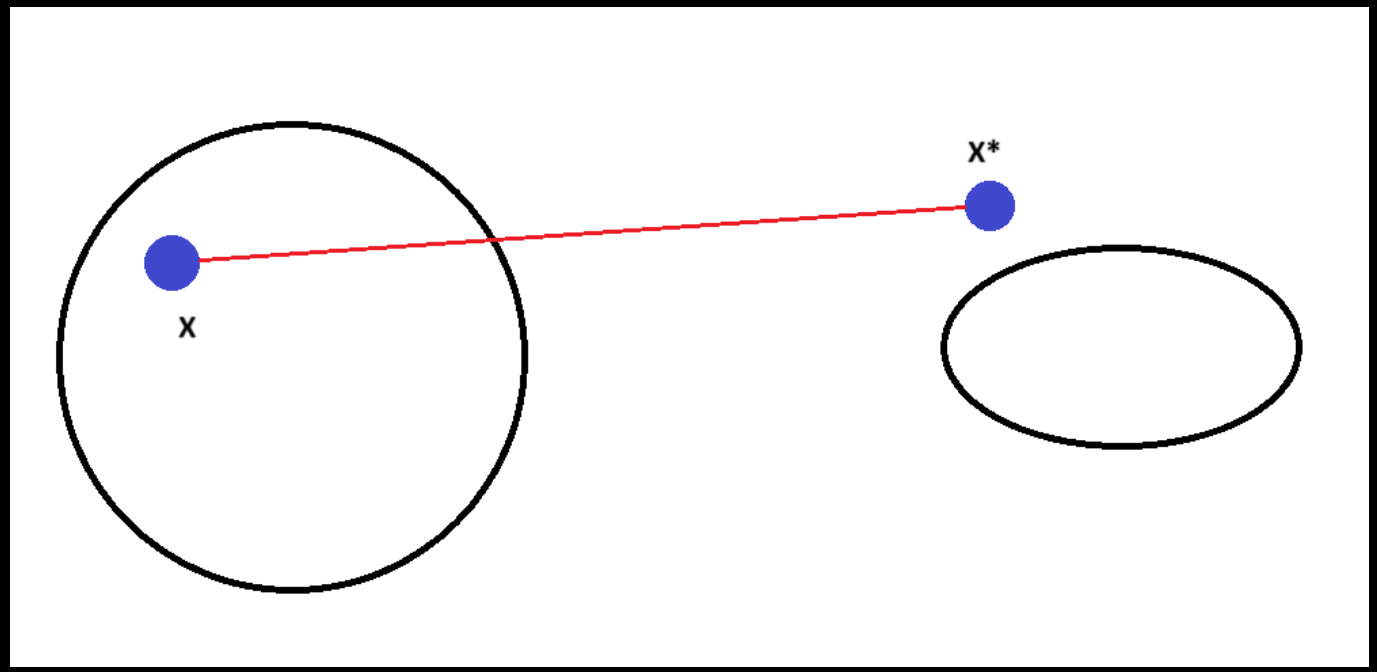
TESTING VERIFICATION

Show a counter example.

- Define a property.
- Set an element from input space
- Show that:

$$\forall x \in X, y = f(x) \in Y$$

doesn't holds



LET'S VERIFY SOMETHING

LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

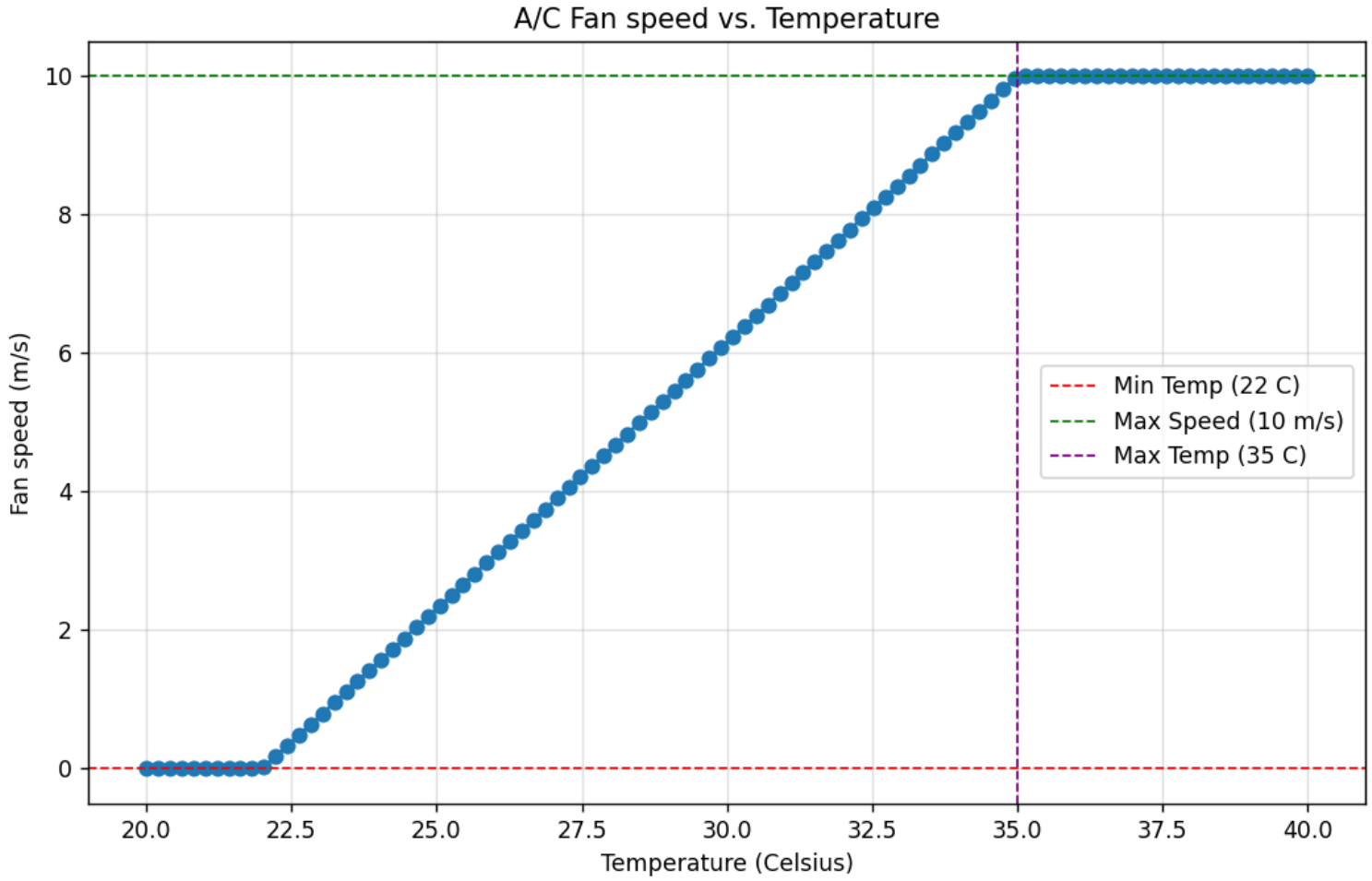
- There is a max limit for the fan (for safety). Let's assume 10 m/s.
- There is a min limit for the temperature. Let's assume 22 C.

LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

- So if the temperature is above 22C, let's linearly scale it to the fan speed.
- If the temperature is below 22C, let's switch off.
- If the fan speed is at max, let's ignore the further temperature increase.

LET'S VERIFY SOMETHING



LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

How the control algorithm should work?

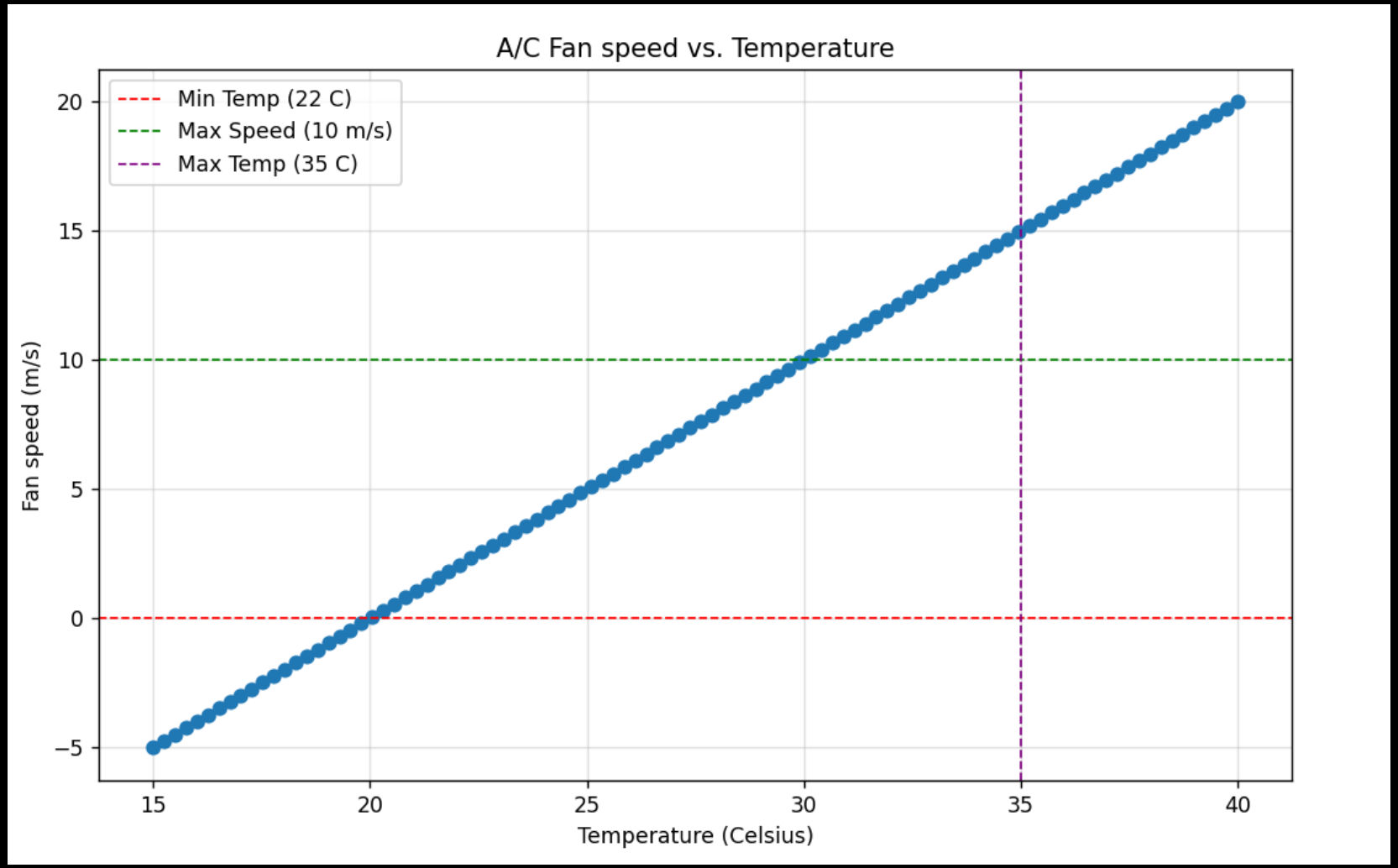
- Input: temperature
- Output: fan speed

LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

```
def fan_speed(temperature):
        return temperature – 20
```

LET'S VERIFY SOMETHING

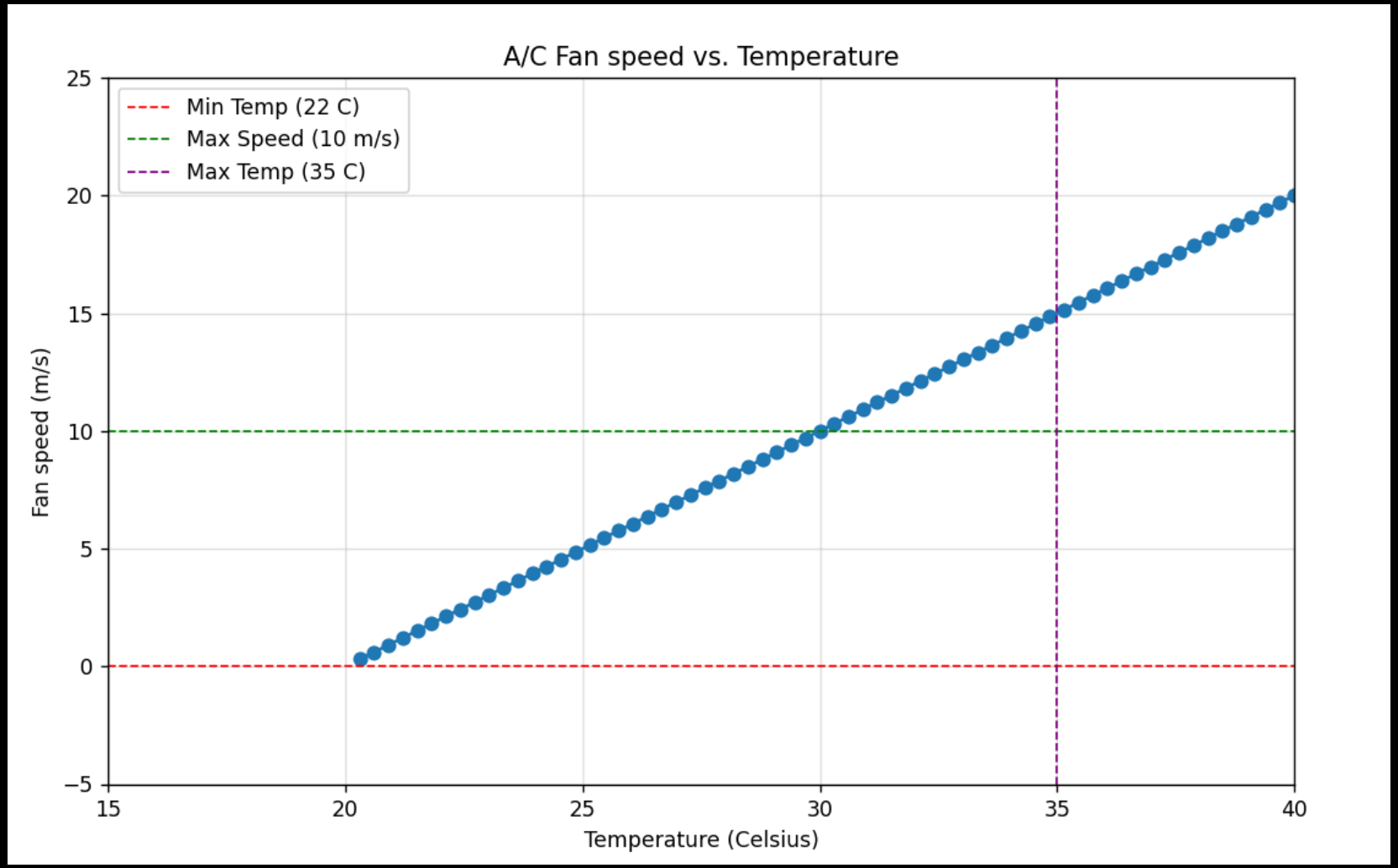


LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

```
def fan_speed(temperature):
        if (temperature>20):
                return temperature – 20
        else:
                return 0
```

LET'S VERIFY SOMETHING

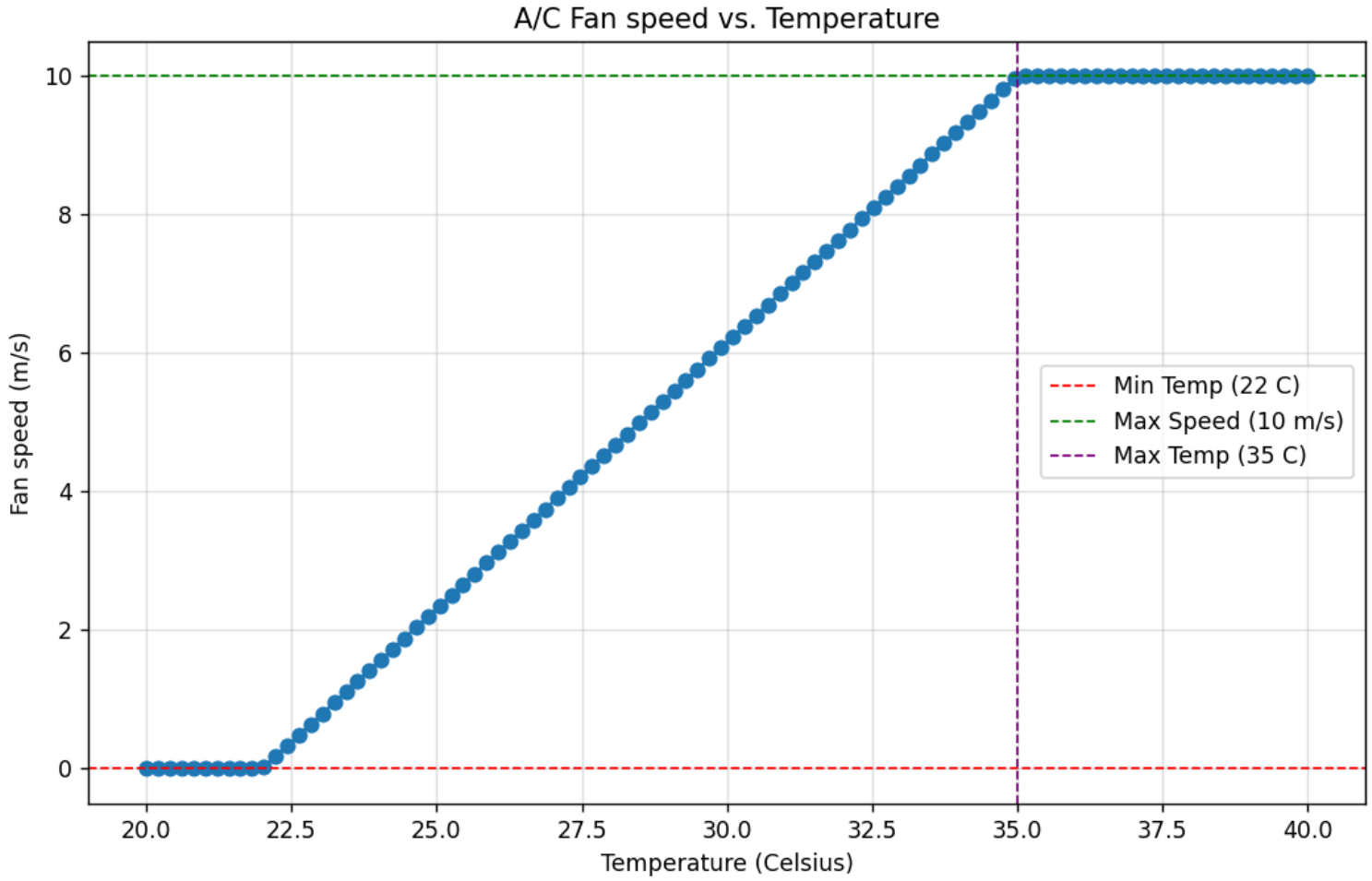


LET'S VERIFY SOMETHING

A/C in auto mode. The hotter it measures, the faster the fan operates.

```
def old_fan_speed(temperature):
    if temperature < min_temp:
        return 0
    elif temperature >= max_temp:
        return max_speed
    else:
        return max_speed * (temperature - min_temp) / (max_temp - min_temp)
```

LET'S VERIFY SOMETHING



LET'S VERIFY SOMETHING

A/C in auto mode.

When cold, the heating turns on. The colder it gets the fan blows faster.

When hot, the cooling turns on. The hotter it gets the fan blows faster.

If too cold, cannot turn on the cooling.

If too hot, cannot turn on the heating.

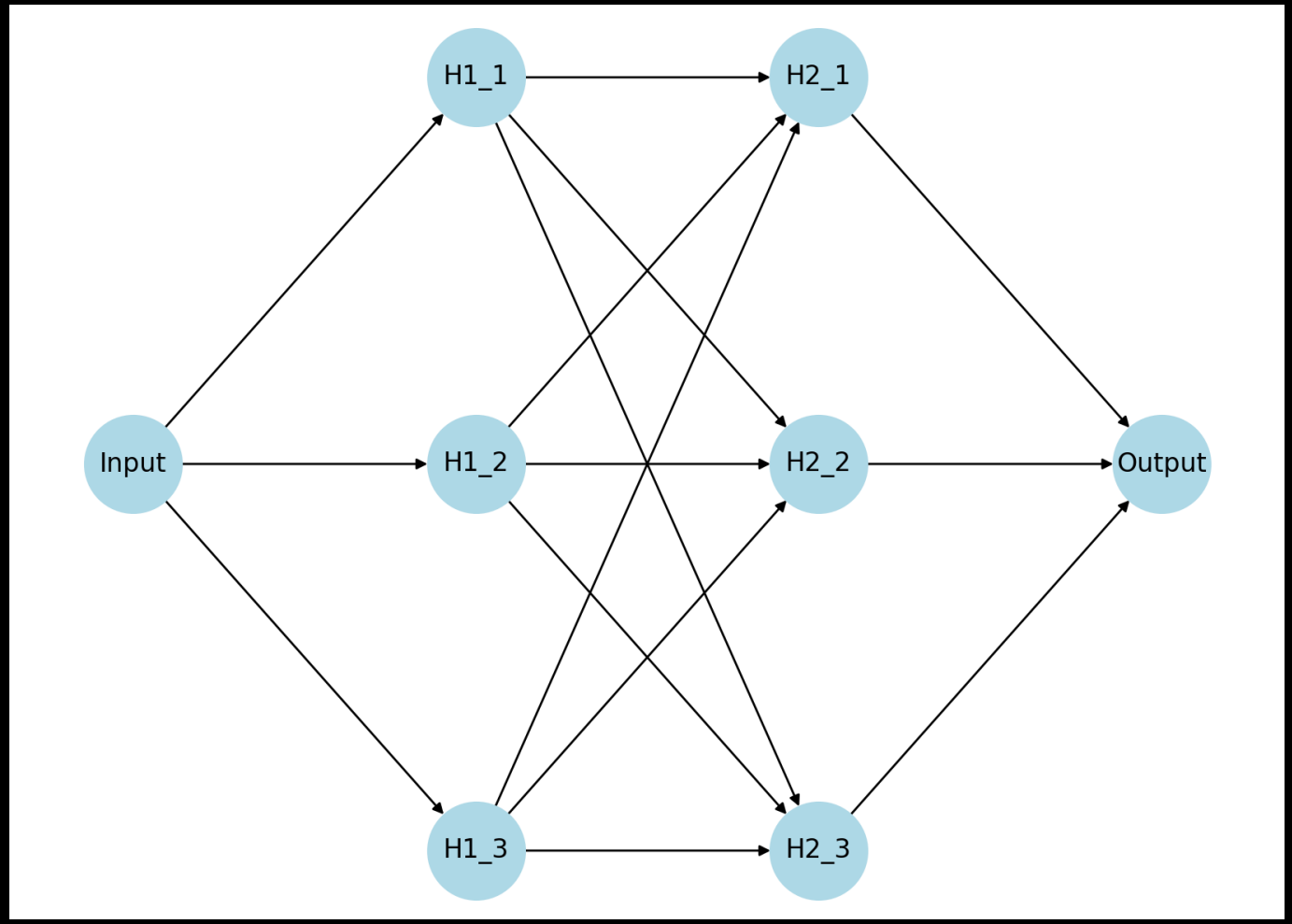
What's in the middle? Is it even linear?

LET'S VERIFY SOMETHING

A/C in auto mode.

Let the machine figure out and we just verify the function of it.

The neurons have functions, sum them up and we will have something to check.



LET'S VERIFY SOMETHING

A/C in auto mode.

$$f(x) = j = \sum w_3 j \cdot ReLU(i = \sum w_2 ij \cdot ReLU(w_1 i_1 \cdot x + b_1 i) + b_2 j) + b_3$$

$$\begin{aligned}
&f(x) \\
&= 0.5 \cdot ReLU(0.2 \cdot ReLU(0.5 \cdot x + 0.1) + 0.3 \cdot ReLU(-0.3 \cdot x - 0.2) - 0.6 \\
&\cdot ReLU(0.8 \cdot x + 0.3) + 0.4) - 0.2 \cdot ReLU(-0.5 \cdot ReLU(0.5 \cdot x + 0.1) + 0.1 \\
&\cdot ReLU(-0.3 \cdot x - 0.2) + 0.9 \cdot ReLU(0.8 \cdot x + 0.3) - 0.1) + 0.3 \cdot ReLU(0.7 \\
&\cdot ReLU(0.5 \cdot x + 0.1) - 0.4 \cdot ReLU(-0.3 \cdot x - 0.2) + 0.2 \cdot ReLU(0.8 \cdot x + 0.3) \\
&+ 0.2) + 0.1
\end{aligned}$$

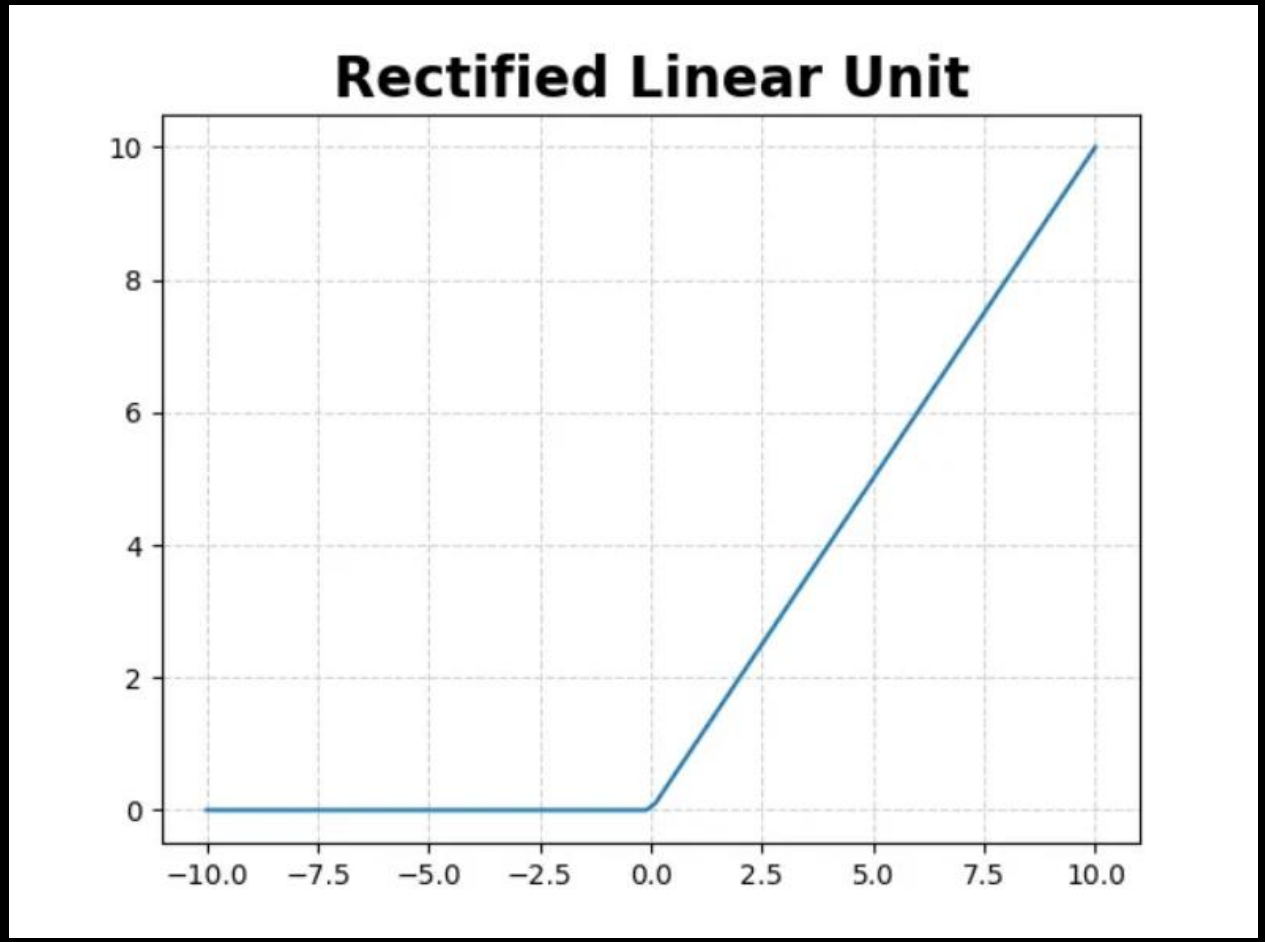
LET'S SET SOME BOUNDRIES

LET'S SET SOME BOUNDRIES

- Most functions can be approximated within certain limits. (zonotop, polytope)
- Neural networks are basically functions.
- A neural network likely can be split into components.
- The activation function of a neuron can be relaxed.
- We can switch certain functions with equal, but easier pairs.

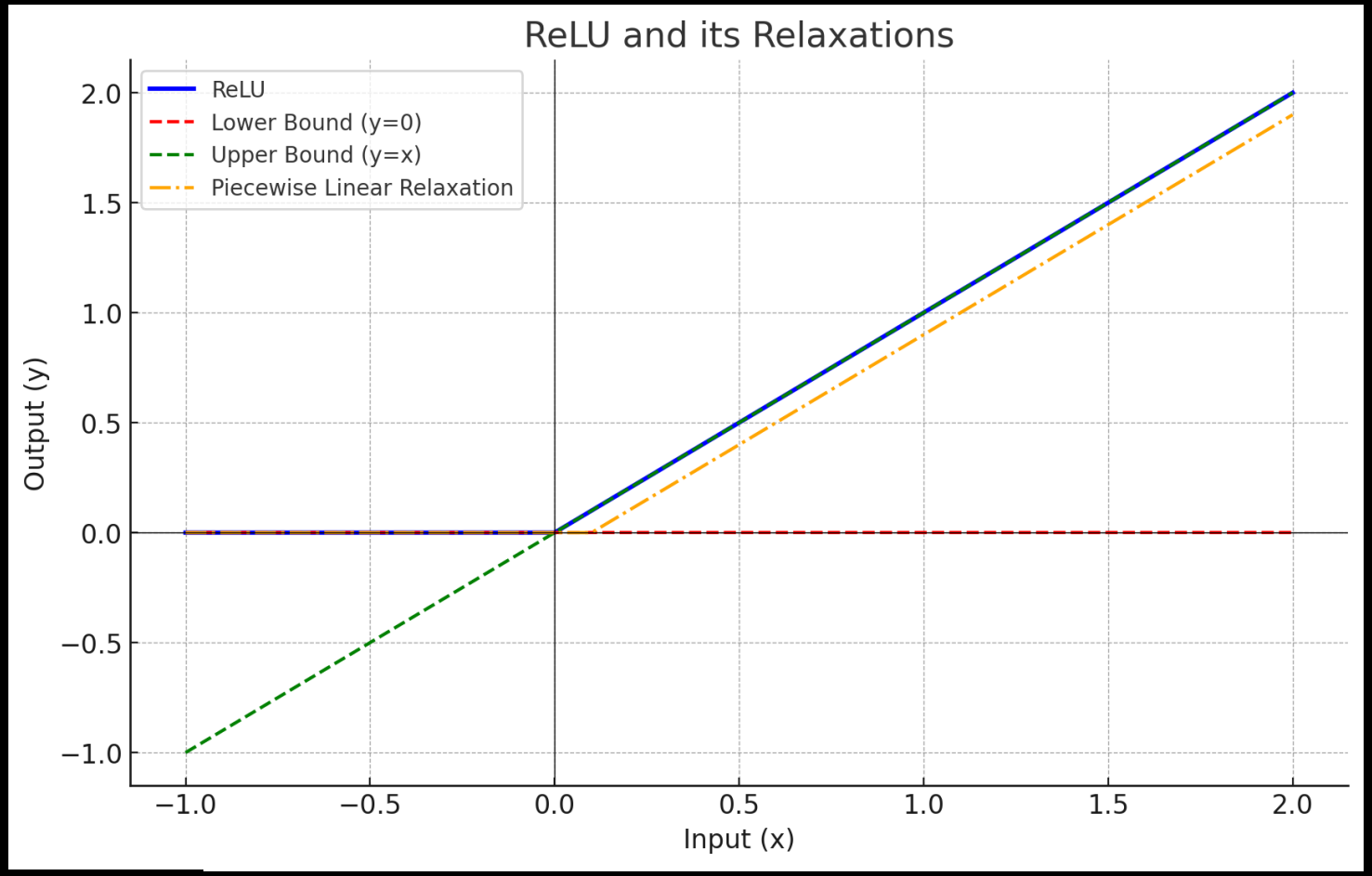
LET'S SET SOME BOUNDRIES

- ReLU
- $ReLU(x) = \max(0, x)$
- Zero-to-identity mapping function



LET'S SET SOME BOUNDRIES

- Linear
- Convex
- Smooth

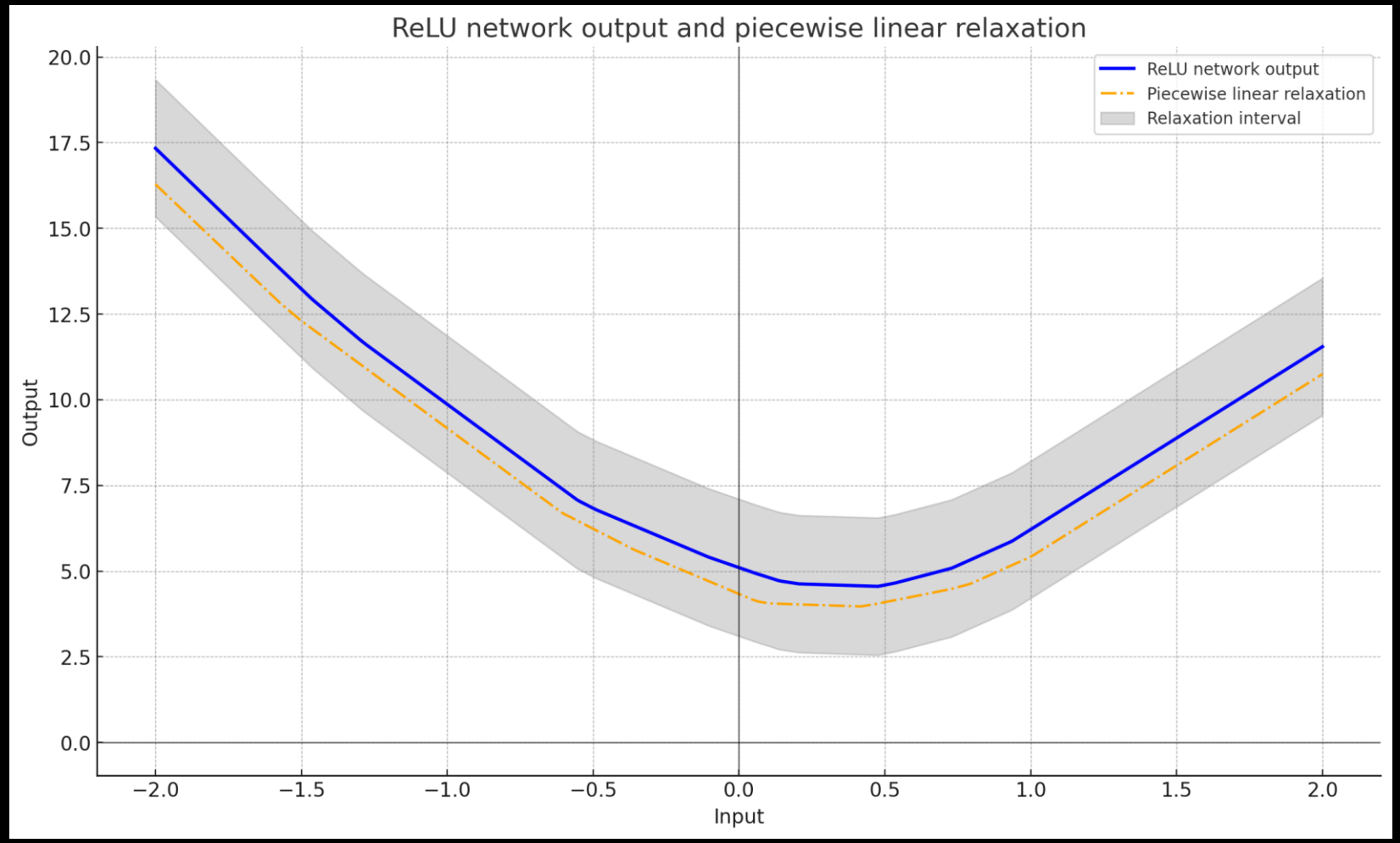


LET'S SET SOME BOUNDRIES

- With relaxed activations, a linear solver can be used
  - Precise, but CPU intensive
  - Overestimation
  - Bad scaling
  - Combinatorical explosion

LET'S SET SOME BOUNDRIES

- Relaxed ReLU with 20 neurons
- LS
- ~3s on i5-10th gen

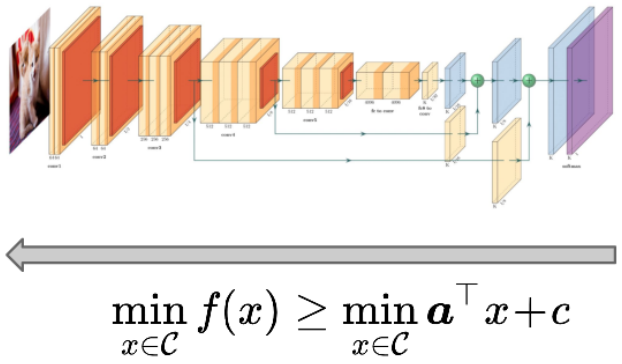


LET'S SET SOME BOUNDRIES

- For larger networks, LS is not option.
- The network need to inspected recursively.
- BaB – Branch and Bound
- Can be parallel – GPU
- Need certain extensions

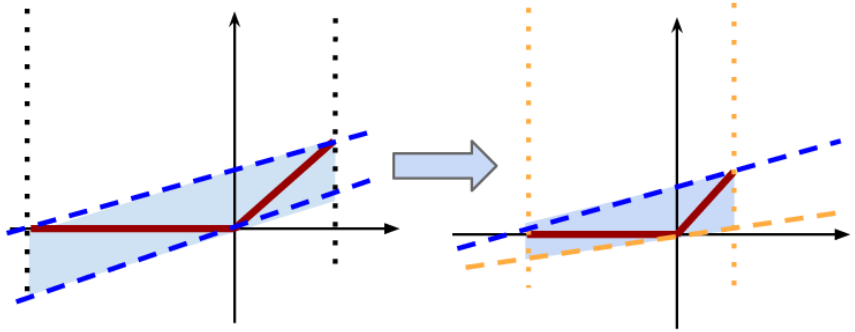
LET'S SET SOME BOUNDRIES

- $\alpha - \beta -$ CROWN

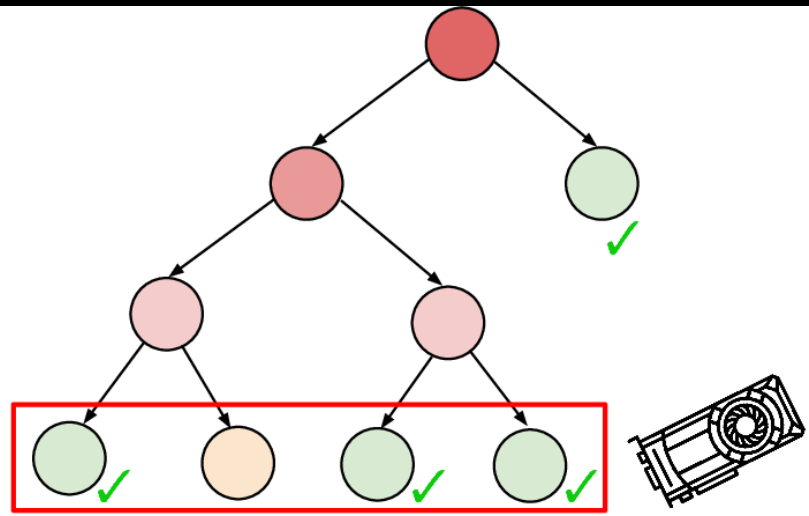


$$\min_{x \in \mathcal{C}} f(x) \geq \min_{x \in \mathcal{C}} \boldsymbol{a}^\top x + c$$

Efficient bound propagation (**CROWN**)



GPU optimized relaxation (***α*-CROWN**)



Parallel branch and bound (***β*-CROWN**)

"Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Neural Network Robustness Verification" by Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin3, Suman Jana, Cho-Jui Hsieh, Zico Kolter

THANK YOU.