EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Student Presentations 2 – Lent 2025

UNIVERSITY OF CAMBRIDGE

PAPER TIME!

Paper tiiiiimeeee woooo!

PAPER RECAP

SANITY CHECKS (ADEBAYO ET AL., 2018)

Adebayo et al. "Sanity checks for saliency maps." NeurIPS (2018).

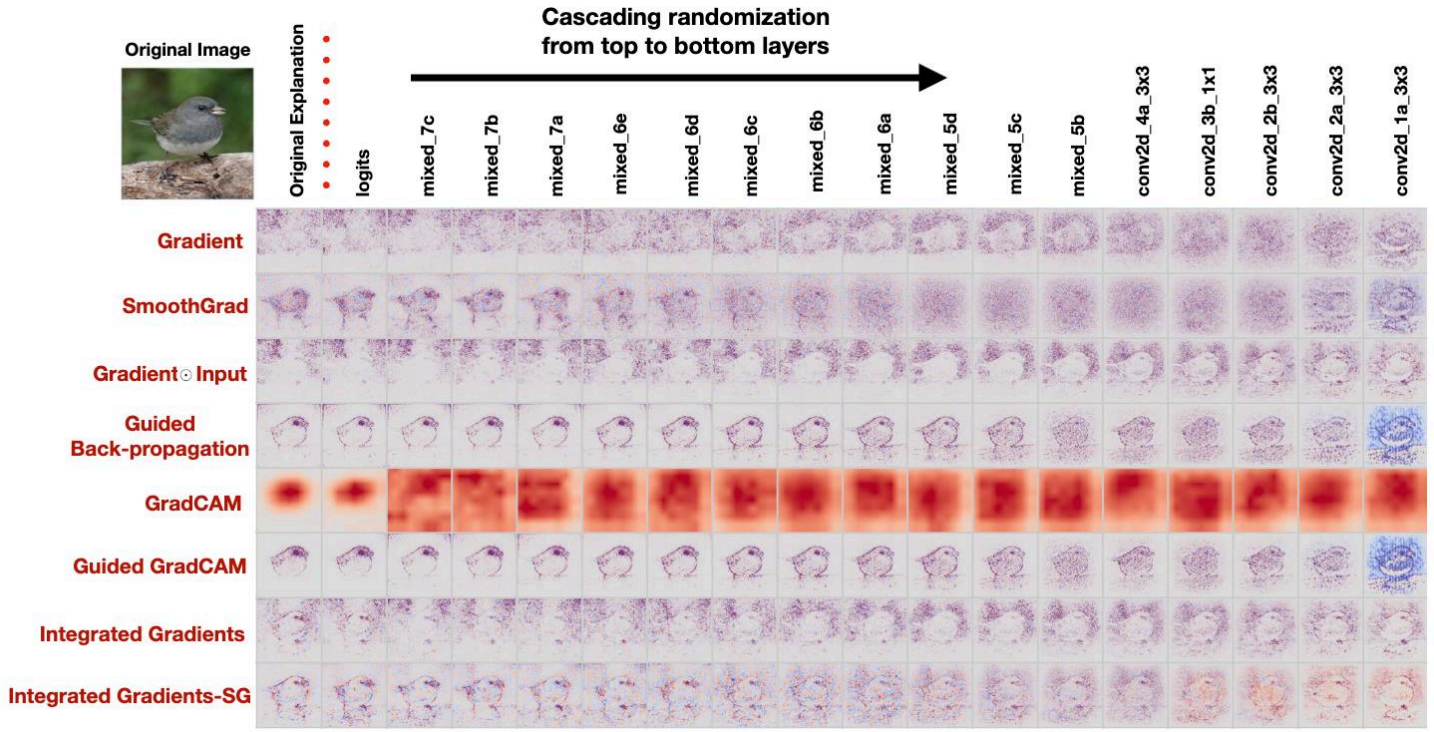
SANITY CHECKS : RESEARCH QUESTION

Main research Question

Do commonly used saliency methods satisfy a series of expected properties (or sanity checks)?

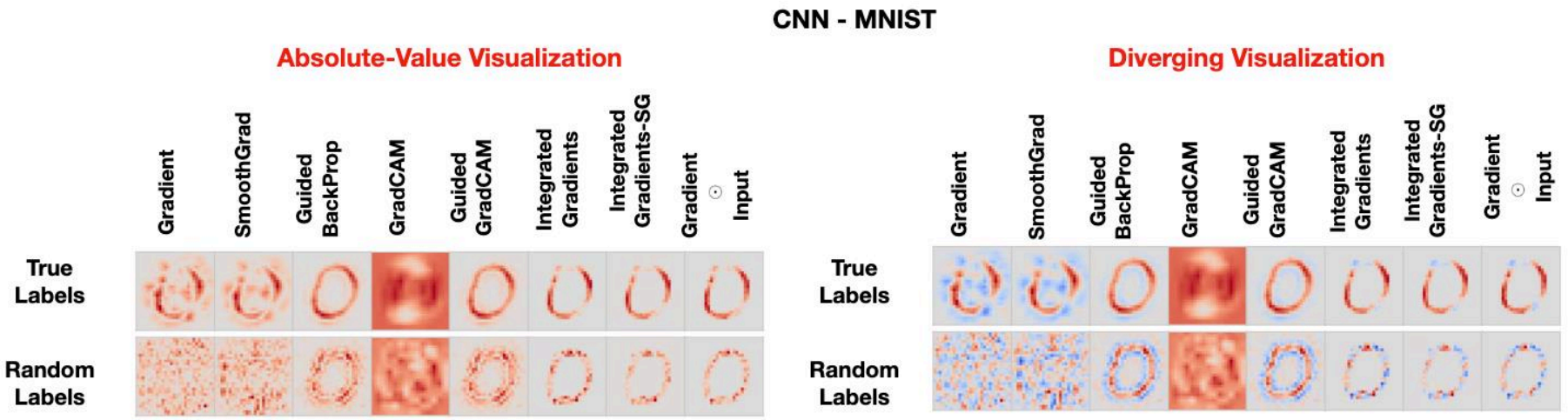
SANITY CHECKS: TAKEAWAYS

1. Some saliency methods may generate similar maps even when the model's weights are random



SANITY CHECKS: TAKEAWAYS

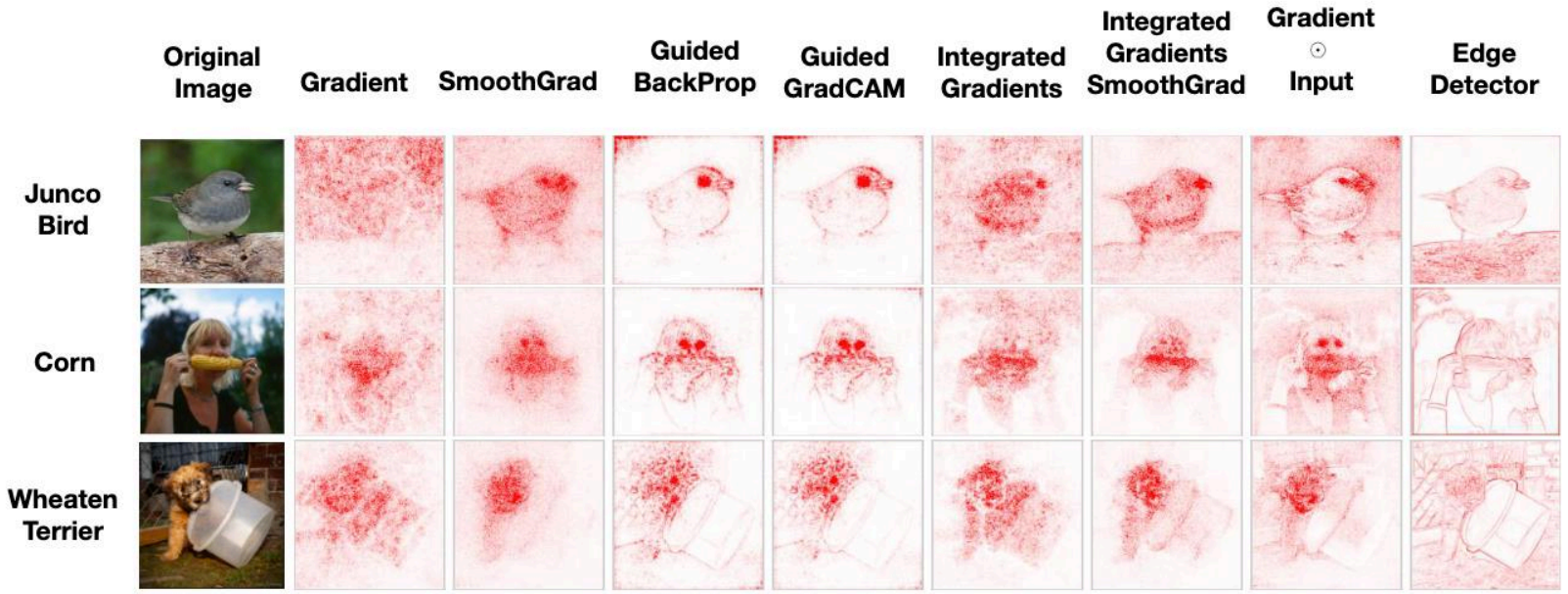
2. Some saliency methods may generate similar maps even when the training labels are random



Vanilla Saliency (Gradient) and GradCAM seem to be the only methods to pass basic sanity checks

SANITY CHECKS: TAKEAWAYS

3. Visual inspection may not be enough: edge detectors render similar explanations to fancy saliency methods!



NETWORK DISSECTION (BAU ET AL., 2017)

Bau et al. "Network dissection: Quantifying interpretability of deep visual representations." CVPR (2017)

NETWORK DISSECTION: RESEARCH QUESTION

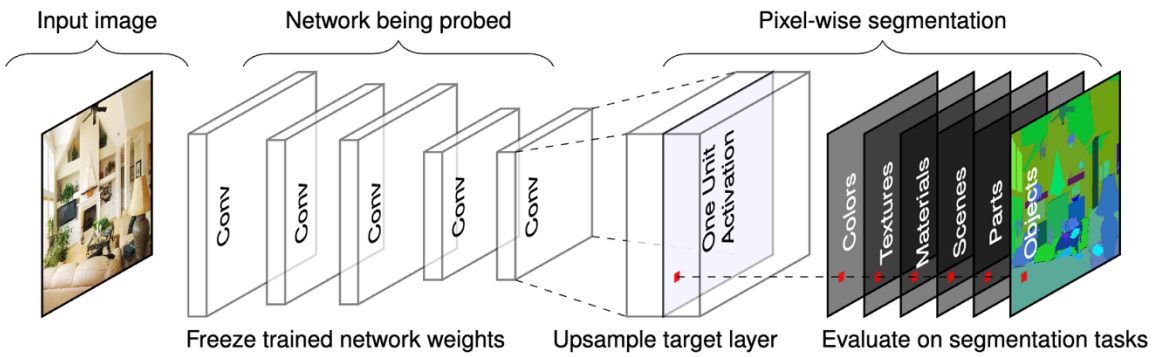
Main research Question

Do Convolutional Neural Networks (CNNs) naturally learn detectors for human-aligned concepts?

NETWORK DISSECTION: TAKEAWAYS

1. We can "dissect" a CNN and understand whether certain neurons or feature maps align with known interpretable concepts

1. Identify a broad set of human-labeled visual concepts.
2. Gather hidden variables' response to known concepts.
3. Quantify alignment of hidden variable–concept pairs.



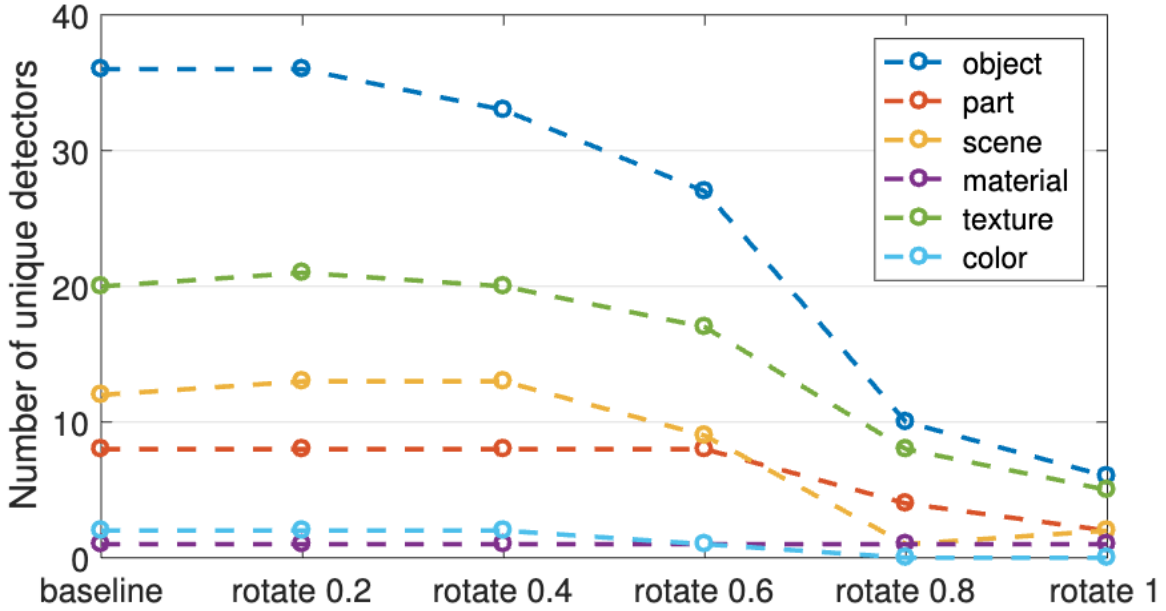
NETWORK DISSECTION: TAKEAWAYS

1. We can "dissect" a CNN and understand whether certain neurons or feature maps align with known interpretable concepts

Figure 1. Unit 13 in [40] (classifying places) detects table lamps. Unit 246 in [11] (classifying objects) detects bicycle wheels. A unit in [32] (self-supervised for generating videos) detects people.

NETWORK DISSECTION: TAKEAWAYS

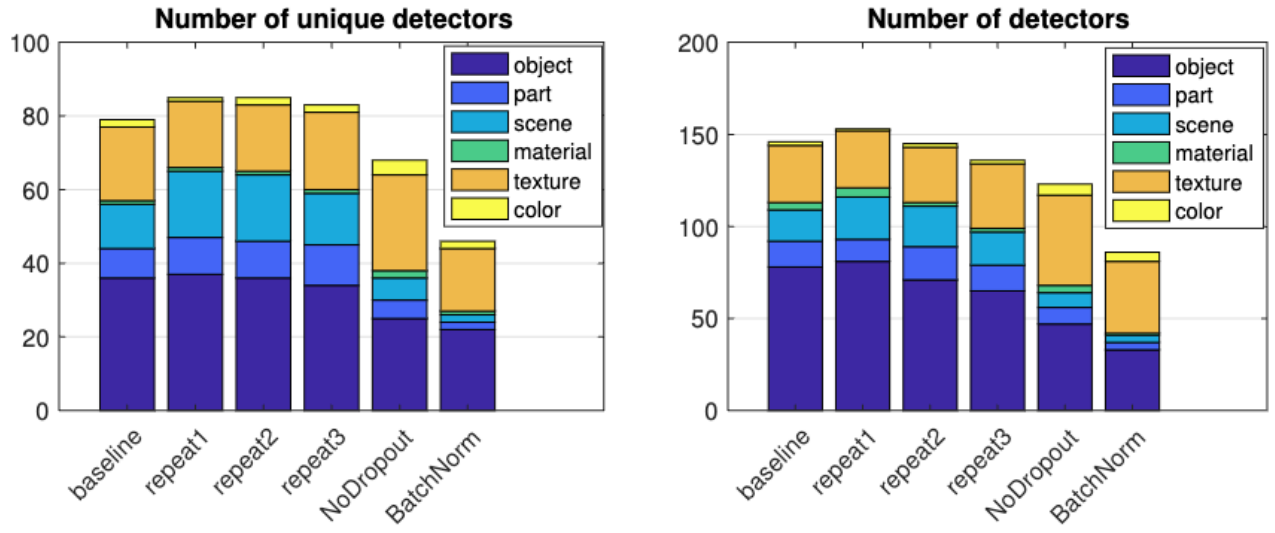
2. "Two representations of perfectly equivalent discriminative power to have very different levels of interpretability."



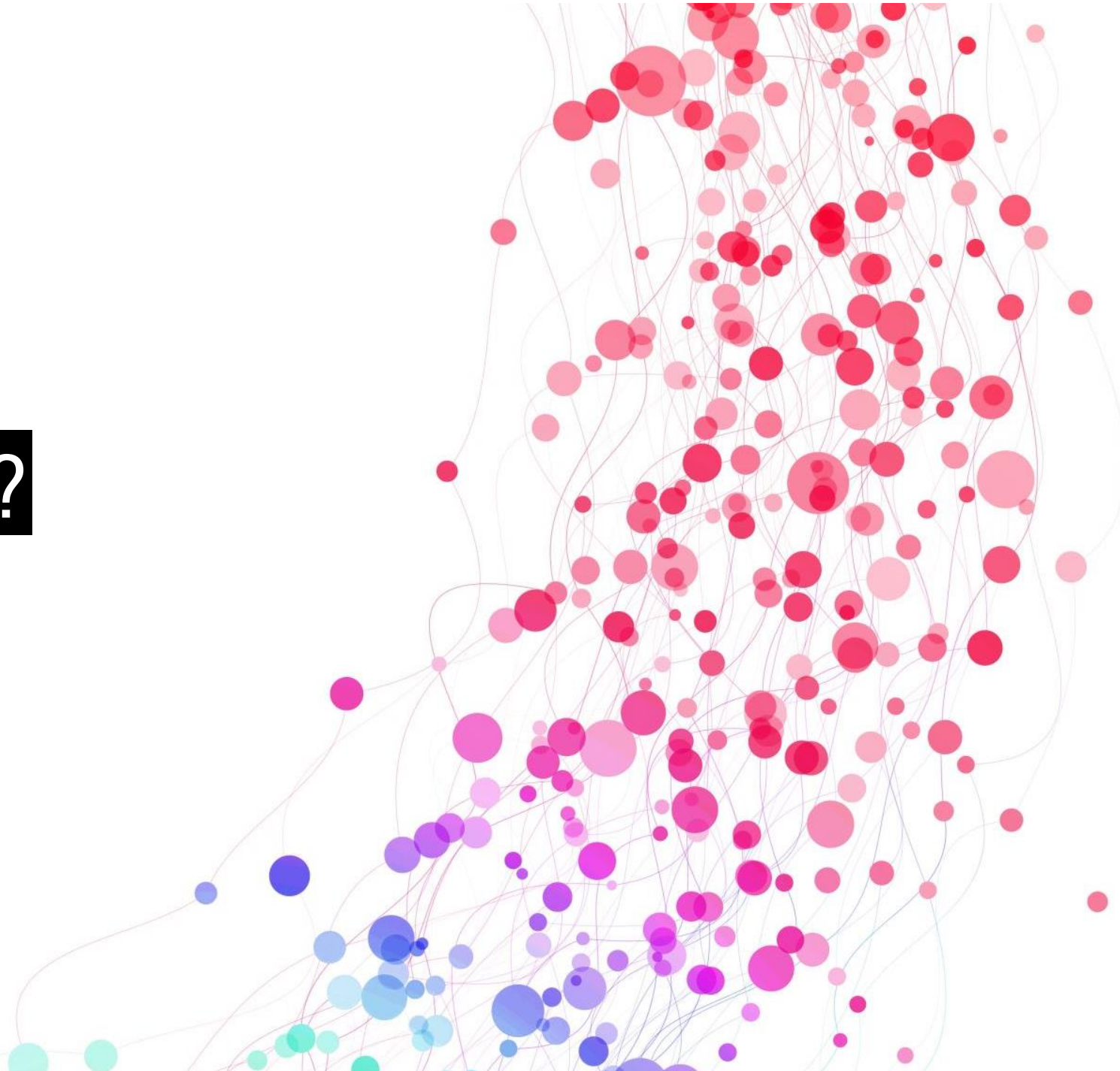
If you rotate representations, you get the same task accuracy but very different levels of interpretability!

NETWORK DISSECTION: TAKEAWAYS

3. Training conditions (e.g., batch norm, dropout, network width) can have consequences on concept alignment



QUESTIONS?



PAPER DISTRIBUTION TIME!

Does anyone know who this is?

PAPER DISTRIBUTION TIME!

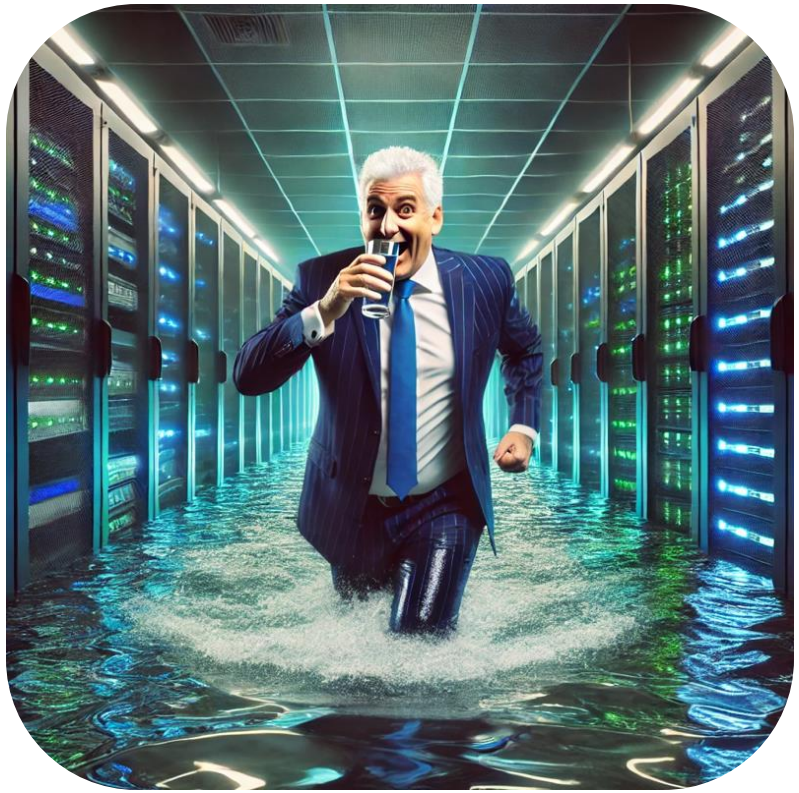
This is Bob Hawke, a former Australian Prime Minister

PAPER DISTRIBUTION TIME!

In 1954, whilst a student at Oxford, he broke the world record for drinking a yard of ale (1.4 L) in 11 seconds

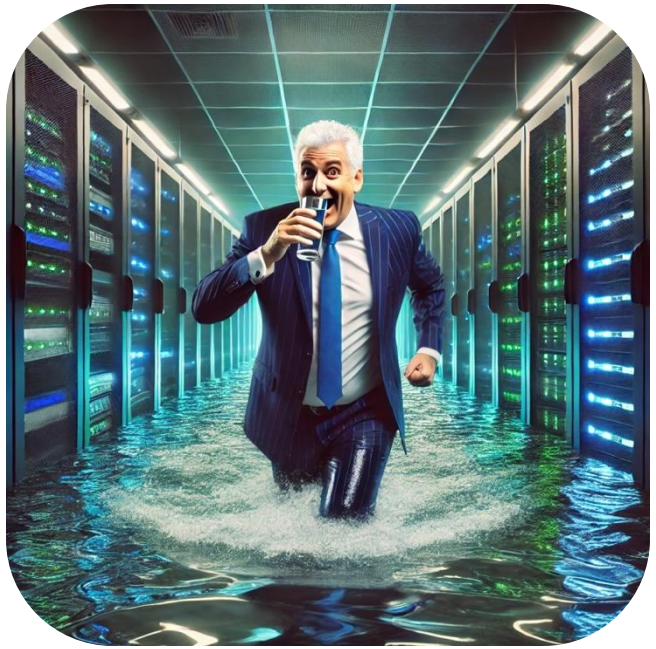
PAPER DISTRIBUTION TIME!

How many days would it take Bob Hawke to drink the water needed to cool down the average data centre for one day?



PAPER DISTRIBUTION TIME!

How many days would it take Bob Hawke to drink the water needed to cool down the average data centre for one day?



Answer: approximately 103.5 days

PAPER DISTRIBUTION TIME!

Paper 1: Laguna Cillero et al. "Beyond concept bottleneck models: How to make black boxes intervenable?." NeurIPS (2024)

Paper 2: Havasi et al. "Addressing leakage in concept bottleneck models." NeurIPS (2022)