EXPLAINABLE ARTIFICIAL INTELLIGENCE

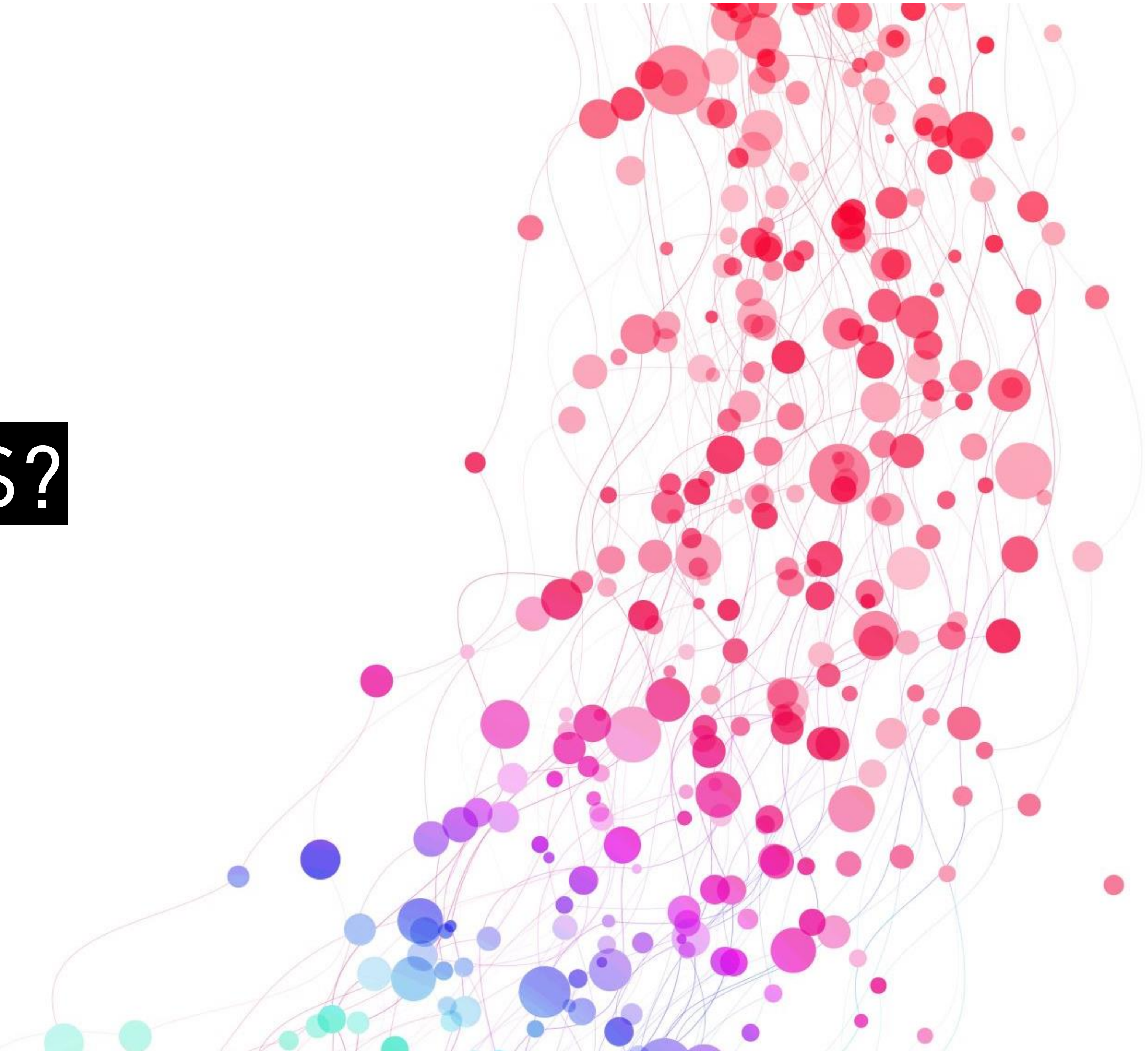
L193 – Student Presentations 3 - Lent 2025

UNIVERSITY OF CAMBRIDGE

PAPER TIME!

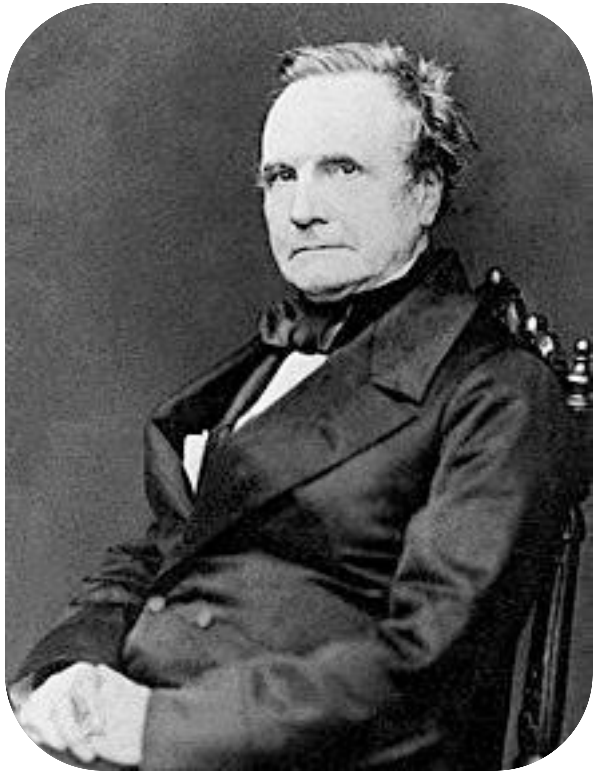
Paper tiiiiimeeee woooo!

QUESTIONS?



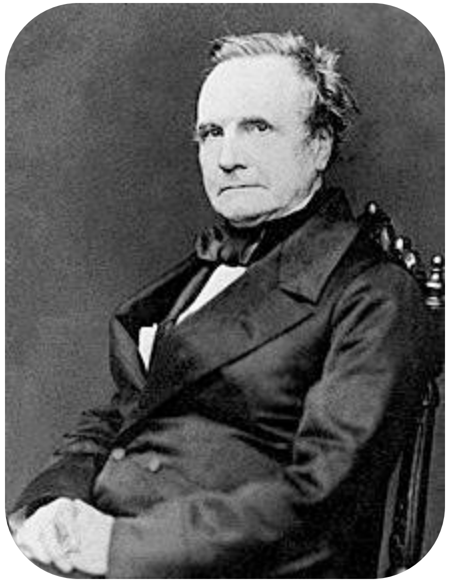
PAPER DISTRIBUTION TIME!

Does anyone know who this is?

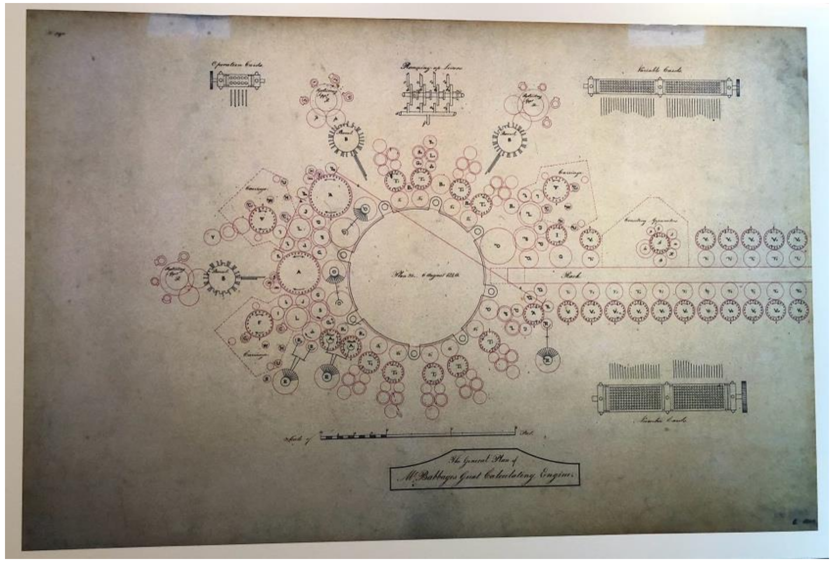


PAPER DISTRIBUTION TIME!

Charles Babbage!



Charles Babbage

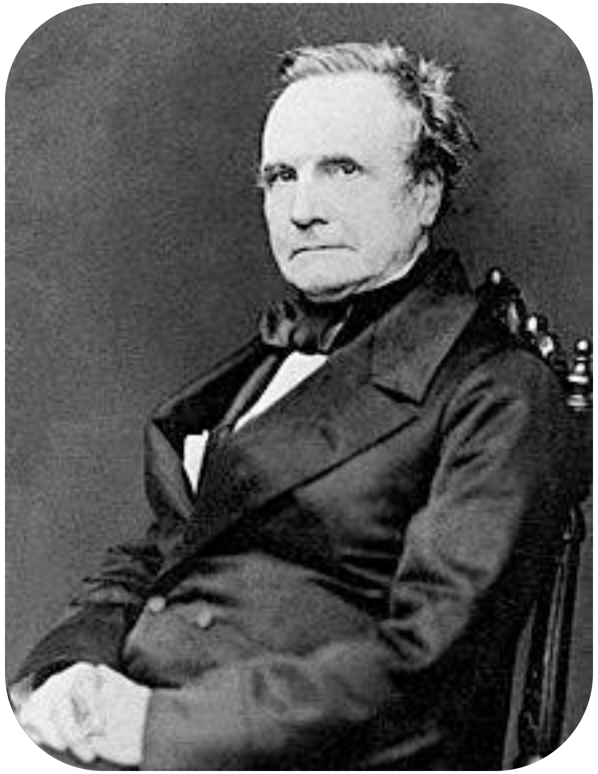


Part of a blueprint for this "Analytical Engine"

Ada Lovelace

PAPER DISTRIBUTION TIME!

He had a particular dislike towards “public nuisances”



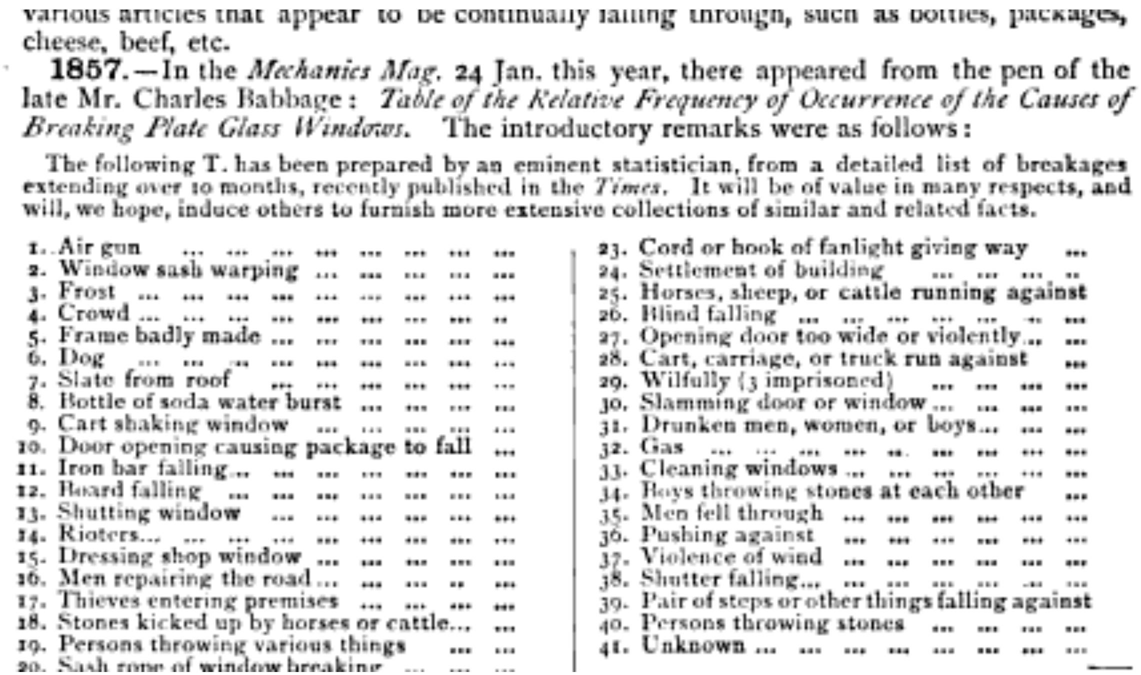
Organ-grinders

Hoop-rolling

PAPER DISTRIBUTION TIME!

In 1857 he published “*Table of the Relative Frequency of Occurrence of the Causes of Breaking of Plate Glass Windows*” where he painstakingly documented the cause of 464 broken window panes.

How many of these panes were broken by "*drunken men, women or boys*”?



various articles that appear to be continually falling through, such as bottles, packages, cheese, beef, etc.

1857.—In the *Mechanics Mag.* 24 Jan. this year, there appeared from the pen of the late Mr. Charles Babbage: *Table of the Relative Frequency of Occurrence of the Causes of Breaking Plate Glass Windows.* The introductory remarks were as follows:

The following T. has been prepared by an eminent statistician, from a detailed list of breakages extending over 10 months, recently published in the *Times*. It will be of value in many respects, and will, we hope, induce others to furnish more extensive collections of similar and related facts.

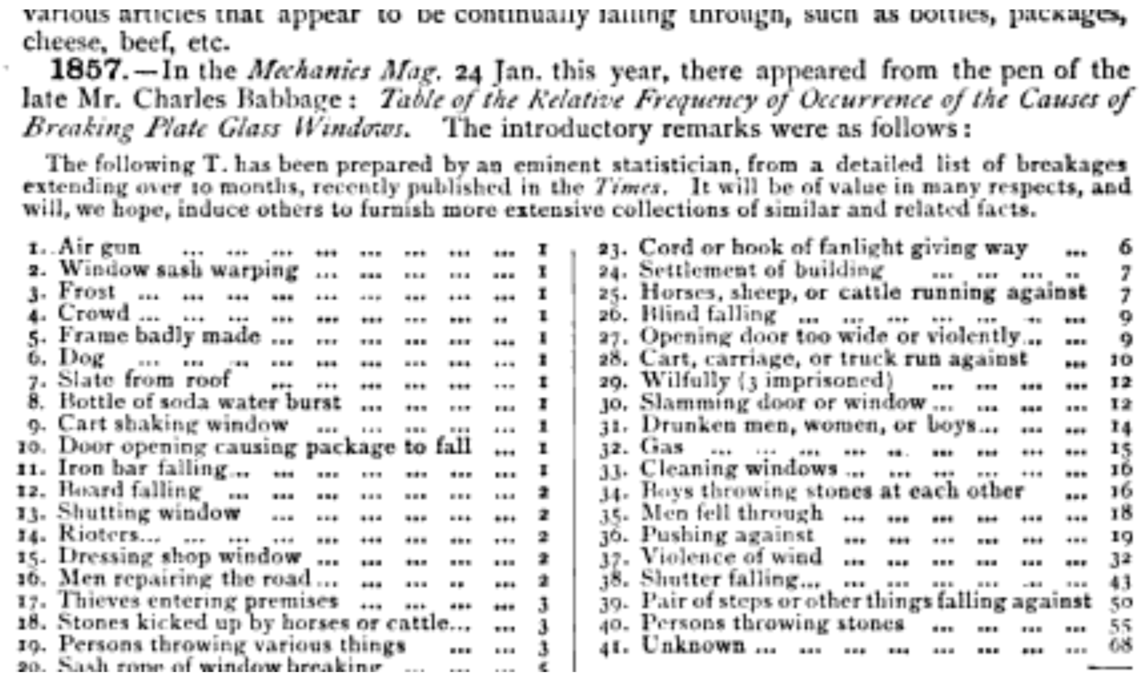
1. Air gun ... ...
2. Window sash warping ... ...
3. Frost ... ...
4. Crowd ... ...
5. Frame badly made ... ...
6. Dog ... ...
7. Slate from roof ... ...
8. Bottle of soda water burst ... ...
9. Cart shaking window ... ...
10. Door opening causing package to fall ...
11. Iron bar falling ... ...
12. Board falling ... ...
13. Shutting window ... ...
14. Rioters... ... ...
15. Dressing shop window ... ...
16. Men repairing the road ... ...
17. Thieves entering premises ... ...
18. Stones kicked up by horses or cattle... ...
19. Persons throwing various things ... ...
20. Sash rope of window breaking ... ...

23. Cord or hook of fanlight giving way ...
24. Settlement of building ... ...
25. Horses, sheep, or cattle running against
26. Blind falling ... ...
27. Opening door too wide or violently... ...
28. Cart, carriage, or truck run against ...
29. Wilfully (3 imprisoned) ... ...
30. Slamming door or window ... ...
31. Drunken men, women, or boys... ... ...
32. Gas ... ...
33. Cleaning windows ... ...
34. Boys throwing stones at each other ...
35. Men fell through ... ...
36. Pushing against ... ...
37. Violence of wind ... ...
38. Shutter falling... ... ...
39. Pair of steps or other things falling against
40. Persons throwing stones ... ...
41. Unknown ... ...

PAPER DISTRIBUTION TIME!

In 1857 he published “*Table of the Relative Frequency of Occurrence of the Causes of Breaking of Plate Glass Windows*” where he painstakingly documented the cause of 464 broken window panes.

How many of these panes were broken by "*drunken men, women or boys*”?



various articles that appear to be continually falling through, such as bottles, packages, cheese, beef, etc.

1857.—In the *Mechanics Mag.* 24 Jan. this year, there appeared from the pen of the late Mr. Charles Babbage: *Table of the Relative Frequency of Occurrence of the Causes of Breaking Plate Glass Windows.* The introductory remarks were as follows:

The following T. has been prepared by an eminent statistician, from a detailed list of breakages extending over 10 months, recently published in the *Times*. It will be of value in many respects, and will, we hope, induce others to furnish more extensive collections of similar and related facts.

1.	Air gun	1
2.	Window sash warping	1
3.	Frost	1
4.	Crowd	1
5.	Frame badly made	1
6.	Dog	1
7.	Slate from roof	1
8.	Bottle of soda water burst	1
9.	Cart shaking window	1
10.	Door opening causing package to fall	1
11.	Iron bar falling	1
12.	Board falling	2
13.	Shutting window	2
14.	Rioters	2
15.	Dressing shop window	2
16.	Men repairing the road	2
17.	Thieves entering premises	3
18.	Stones kicked up by horses or cattle	3
19.	Persons throwing various things	3
20.	Sash rope of window breaking	5
23.	Cord or hook of fanlight giving way	6
24.	Settlement of building	7
25.	Horses, sheep, or cattle running against	7
26.	Blind falling	9
27.	Opening door too wide or violently	9
28.	Cart, carriage, or truck run against	10
29.	Wilfully (3 imprisoned)	12
30.	Slamming door or window	12
31.	Drunken men, women, or boys	14
32.	Gas	15
33.	Cleaning windows	16
34.	Boys throwing stones at each other	16
35.	Men fell through	18
36.	Pushing against	19
37.	Violence of wind	32
38.	Shutter falling	43
39.	Pair of steps or other things falling against	50
40.	Persons throwing stones	55
41.	Unknown	68

PAPER DISTRIBUTION TIME!

Paper 1: Gurnee et al. "Finding neurons in a haystack: Case studies with sparse probing." TMLR (2023).

Paper 2: Bhaskar et al. "Finding transformer circuits with edge pruning." NeurIPS (2024).

PAPER RECAP

INTERVENABLE BLACK-BOXES (LAGUNA ET AL., 2024)

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

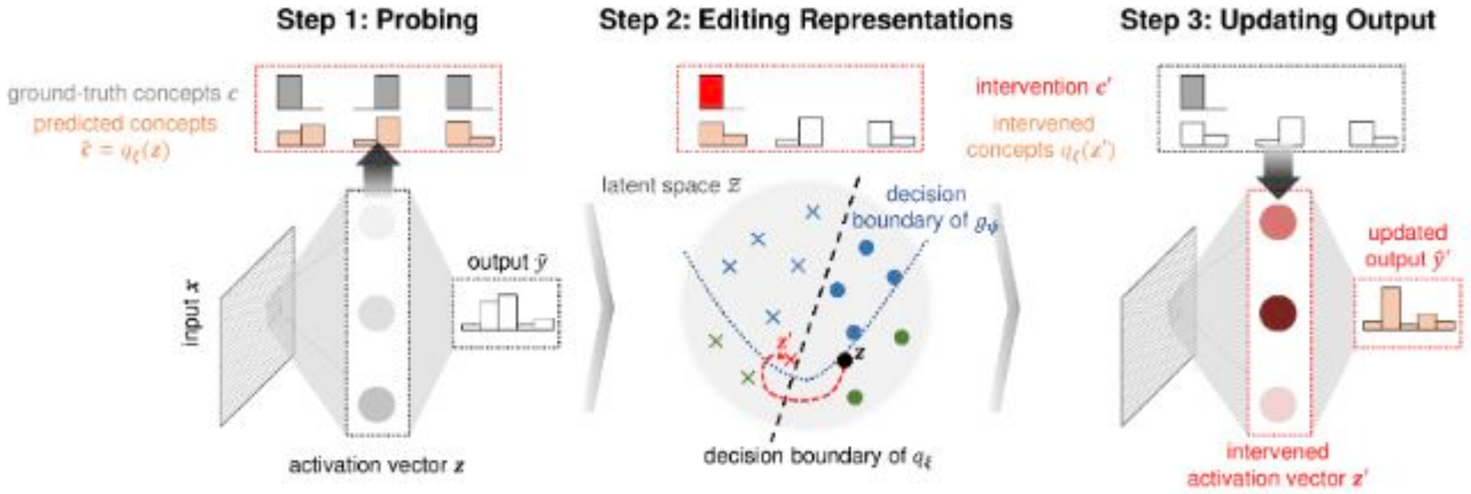
INTERVENABLE BLACK BOXES: RESEARCH QUESTION

Main research Question

Can we perform concept interventions on trained black box models?

INTERVENABLE BLACK BOXES: TAKEAWAYS

1. We can intervene on a black-box model by identifying a concept using a linear probe and then finding the necessary change to the latent space by solving an optimisation problem.


Step 1: Probing
Step 2: Editing Representations
Step 3: Updating Output
ground-truth concepts c
predicted concepts
$\hat{c} = q_\xi(z)$
input x
output $\hat{y}$
activation vector z
latent space $\mathcal{Z}$
decision boundary of g_ψ
z'
z
decision boundary of q_ξ
intervention c'
intervened concepts $q_\xi(z')$
updated output $\hat{y}'$
intervened activation vector z'

$$\underbrace{\arg\min_{z'} \; \lambda \mathcal{L}^{c}\left(q_{\xi}\left(z'\right), c'\right) + d\left(z, z'\right)}$$

Translation: find the minimal change in the latent that “flips” the concept w.r.t. the probe’s boundaries

INTERVENABLE BLACK BOXES: TAKEAWAYS

2. The intervenability objective can be used to fine-tune a model to be more receptive to interventions

$$
\begin{aligned}
&\min_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{z}'} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{c},y)\sim\mathcal{D},\, \boldsymbol{c}'\sim\pi}\left[(1-\beta)\,\overbrace{\mathcal{L}^{y}\Big(g_{\boldsymbol{\psi}}\left(h_{\boldsymbol{\phi}}\left(\boldsymbol{x}\right)\right), y\Big)}^{\text{Task loss without intervention}} + \beta\overbrace{\mathcal{L}^{y}\Big(g_{\boldsymbol{\psi}}\left(\boldsymbol{z}'\right), y\Big)}^{\text{Task loss with intervention}}\right], \\
&\text{s.t. } \boldsymbol{z}' \in \arg\min_{\tilde{\boldsymbol{z}}} \underbrace{\lambda\mathcal{L}^{\boldsymbol{c}}\left(q_{\boldsymbol{\xi}}\left(\tilde{\boldsymbol{z}}\right), \boldsymbol{c}'\right) + d\left(\boldsymbol{z}, \tilde{\boldsymbol{z}}\right)}_{\text{Latent space change constraint}},
\end{aligned}
$$

AUTOREGRESSIVE CBMS (HAVASI ET AL., 2022)

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

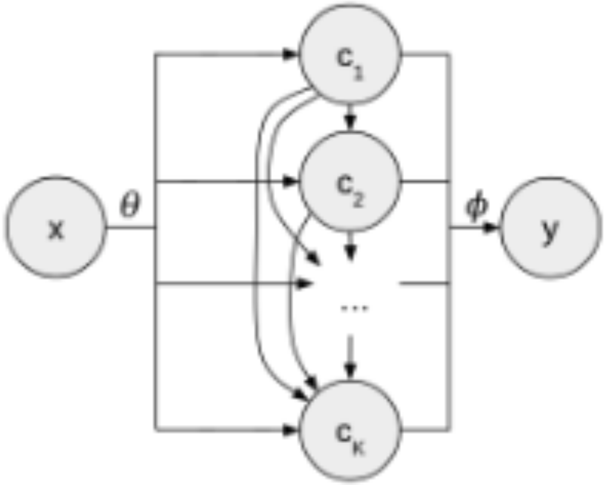
AUTOREGRESSIVE CBMS: RESEARCH QUESTION

Main research Question

Given a CBM, can we avoid “leaking” information from the features to the downstream task predictor by better modelling concept relationships?

AUTOREGRESSIVE CBMS: TAKEAWAYS

1. We can model cross-concept relationships using an auto-regressive architecture



AUTOREGRESSIVE CBMS: TAKEAWAYS

2. In concept-incomplete setups, we can recover black-box accuracy by incorporating a side-channel as part of the model

$$\hat{\theta}, \hat{\phi} = \arg\max_{\theta,\phi} \mathbb{E}_{\mathcal{D}} \left[\log p_\theta(c|x) + \underbrace{\log \mathbb{E}_{p_\theta(z|x)} \left[p_\phi(y|c,z) \right] \right]}_{\text{We sample side-channel latent codes z}}$$