IMPERIAL

Taking the Genie Out Of The Bottle(neck): Concept-Based Insights for AI Oversight

Sonali Parbhoo
s.parbhoo@imperial.ac.uk
Imperial College London

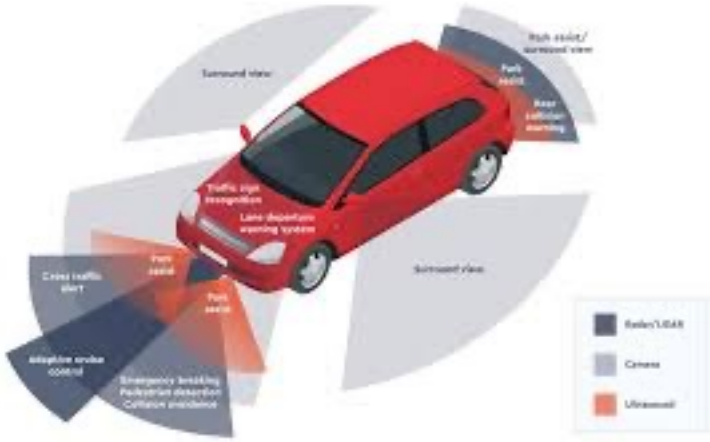
KDD Conference
Aug 26, 2024

Thanks to Angelos Ragkousis, Carissa Wu, Anish Narain, Marton Havasi, Ritam Majumdar, Jack Teversham, Finale Doshi-Velez.

AIs are providing us with many kinds of assistance

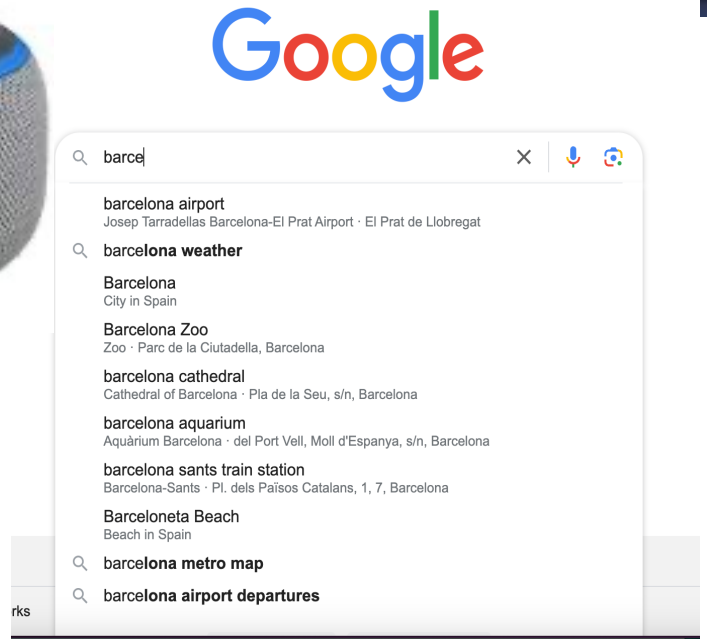


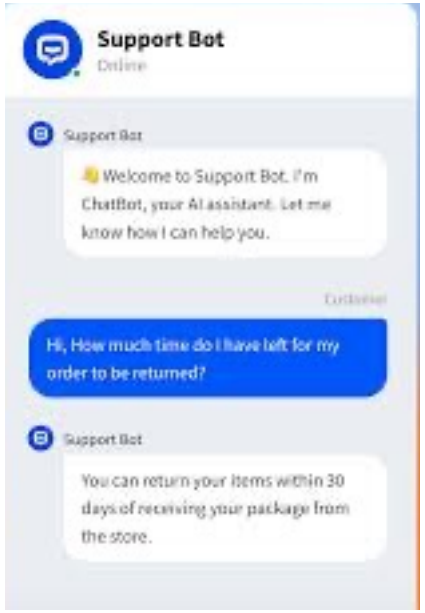


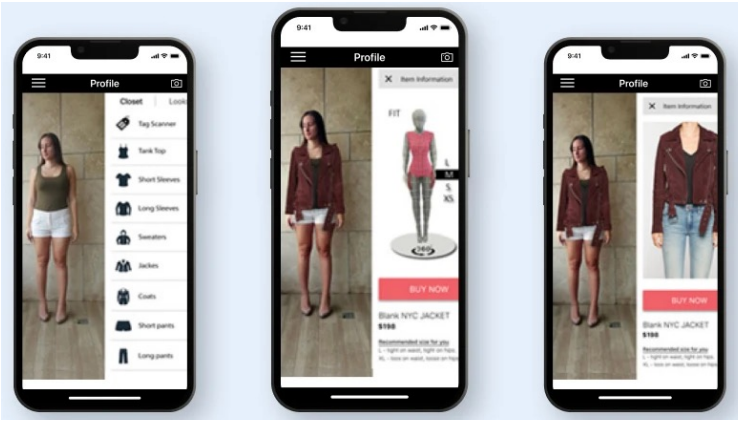












… but AIs make mistakes too

What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case

Marisa Garcia Senior Contributor

Offering an insider's view of the business of flight.

1

Feb 19, 2024,

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

130,565 1,266

Racial bias in health algorithms

Zillow's home-buying debacle shows how hard it is to use AI to value real estate

White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more

MEDICAL & BIOTECH

Artificial Intelligence Is Rushing Into Patient Care—And Could Raise Risks

AI systems are not as rigorously tested as other medical devices, and have already made serious mistakes

By Liz Szabo, Kaiser Health News on December 24, 2019

The "inconvenient truth" about AI in healthcare

Trishan Panch, Heather Mattie & Leo Anthony Celi

npj Digital Medicine **2**, Article number: 77 (2019) | Cite this article

35k Accesses | **24** Citations | **452** Altmetric | Metrics

As the UK sits in painful deadlock over Brexit, it is important to remember that governments are regularly faced with crises, and their responses can create enduring benefit for future generations. Back in 1858, for example, the UK parliament was dealing with another messy crisis: "the great stink." In a world before sanitation, the river Thames had become an open latrine, and as summer blossomed parliament was engulfed in a pestilential stench. £2.5 million (about £300 million in today's money) was hastily approved to build a network of sewers throughout the capital.[1] This particular model of sanitation, developed by Bazalgette, was adopted by

ChatGPT hallucinates court cases

Advances made in 2023 by large language models (LLMs) have stoked widespread interest in the transformative potential of generative AI across nearly every industry. OpenAI's ChatGPT has been at the center of this surge in interest, foreshadowing how gen AI holds the power to disrupt the nature of work in nearly every corner of business.

But the technology still has a long way to go before it can reliably take over

August 2023

EEOC Secures First Workplace Artificial Intelligence Settlement

On Aug. 9, 2023, a tutoring company agreed to pay $365,000 to settle an artificial intelligence (AI) lawsuit with the Equal Employment Opportunity Commission (EEOC). The settlement comes on the heels of multiple EEOC warnings to employers about potential discrimination associated with the use of AI for hiring and workplace decisions.

Artificial Intelligence Makes Bad Medicine Even Worse

… but AIs make mistakes too

What Air Canada Lost In 'Remarkable' Lying AI Chatbot Ca

Marisa Garcia Senior Contributo

Offering an insider's view of the b

1

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

130,565 1,266

MEDICAL & BIOTECH

Artificial Intelligence Is Rushing Into Patient Care—And Could Raise Risks

edical devices, and have already made serious es

s on December 24, 2019

These mistakes highlight the need for AI oversight, particularly in high-risk situations

ChatGPT hallu cases

Advances made in 2023 by large language models (LLMs) have stoked widespread interest in the transformative potential of generative AI across nearly every industry. OpenAI's ChatGPT has been at the center of this surge in interest, foreshadowing how gen AI holds the power to disrupt the nature of work in nearly every corner of business.

But the technology still has a long way to go before it can reliably take over

Intelligence Settlement

On Aug. 9, 2023, a tutoring company agreed to pay $365,000 to settle an artificial intelligence (AI) lawsuit with the Equal Employment Opportunity Commission (EEOC). The settlement comes on the heels of multiple EEOC warnings to employers about potential discrimination associated with the use of AI for hiring and workplace decisions.

uth" about AI in

thony Celi

7 (2019) | Cite this article

35k Accesses | **24** Citations | **452** Altmetric | Metrics

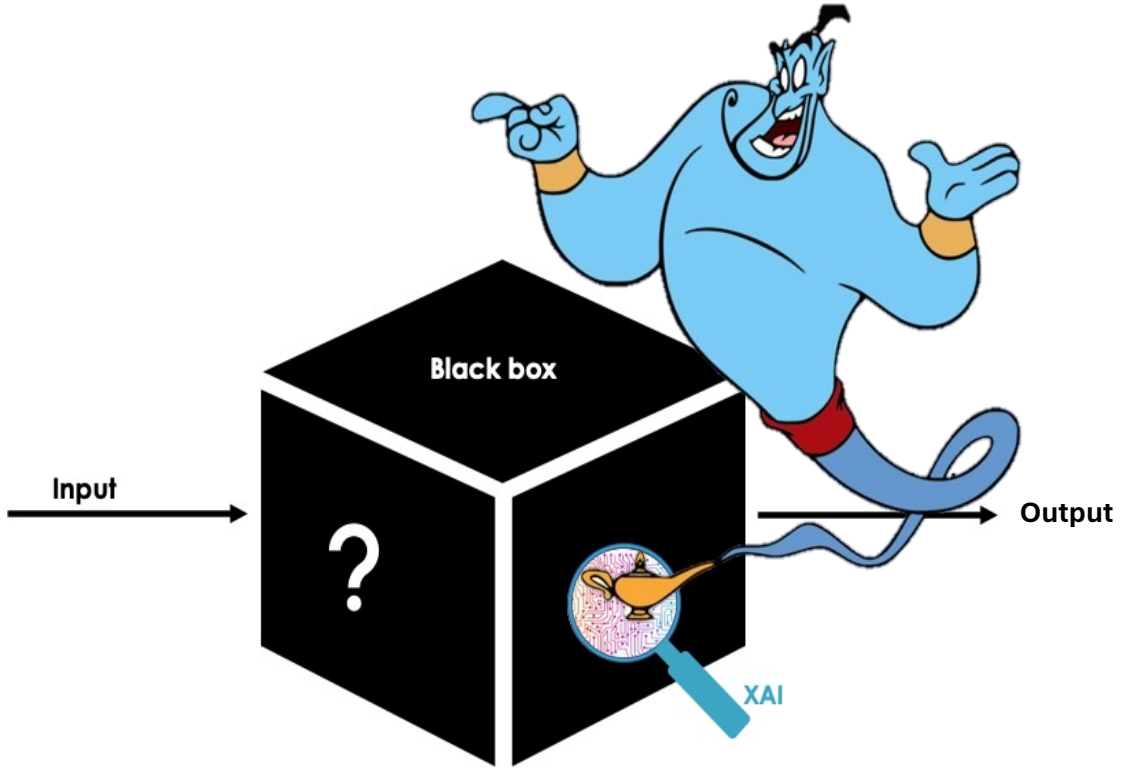
As the UK sits in painful deadlock over Brexit, it is important to remember that governments are regularly faced with crises, and their responses can create enduring benefit for future generations. Back in 1858, for example, the UK parliament was dealing with another messy crisis: "the great stink." In a world before sanitation, the river Thames had become an open latrine, and as summer blossomed parliament was engulfed in a pestilential stench. £2.5 million (about £300 million in today's money) was hastily approved to build a network of sewers throughout the capital.[1] This particular model of sanitation, developed by Bazalgette, was adopted by

Artificial Intelligence Makes Bad Medicine Even Worse

Interpretability for AI Oversight

Being able to inspect and interpret a model can help us identify:

- *Missing features* (e.g. fluid balance to help disambiguate hypotension response; genetic markers to distinguish viral strains of HIV)
- *Incorrect features* (e.g. t0 --> to for cancer staging)
- *Causality leakage* (e.g. alerting about a risk event a team already knows about)
- *Safety breaches, inconsistencies in objectives* and many more issues ….

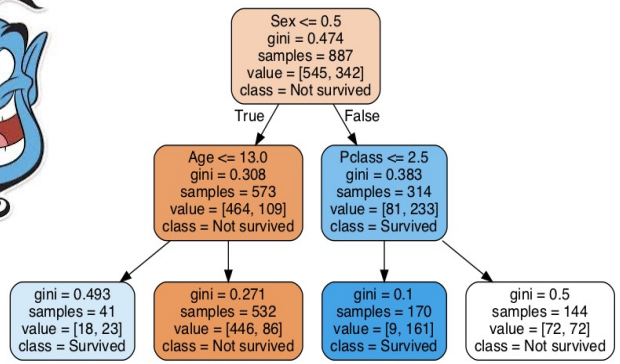


Interpretability is critical when the **problem cannot be fully formalized** or **checked computationally**

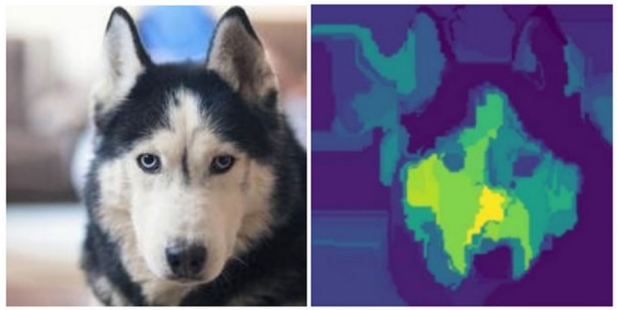
The ‘right’ method for interpretability is context-dependent

Validation in settings where features are inherently interpretable





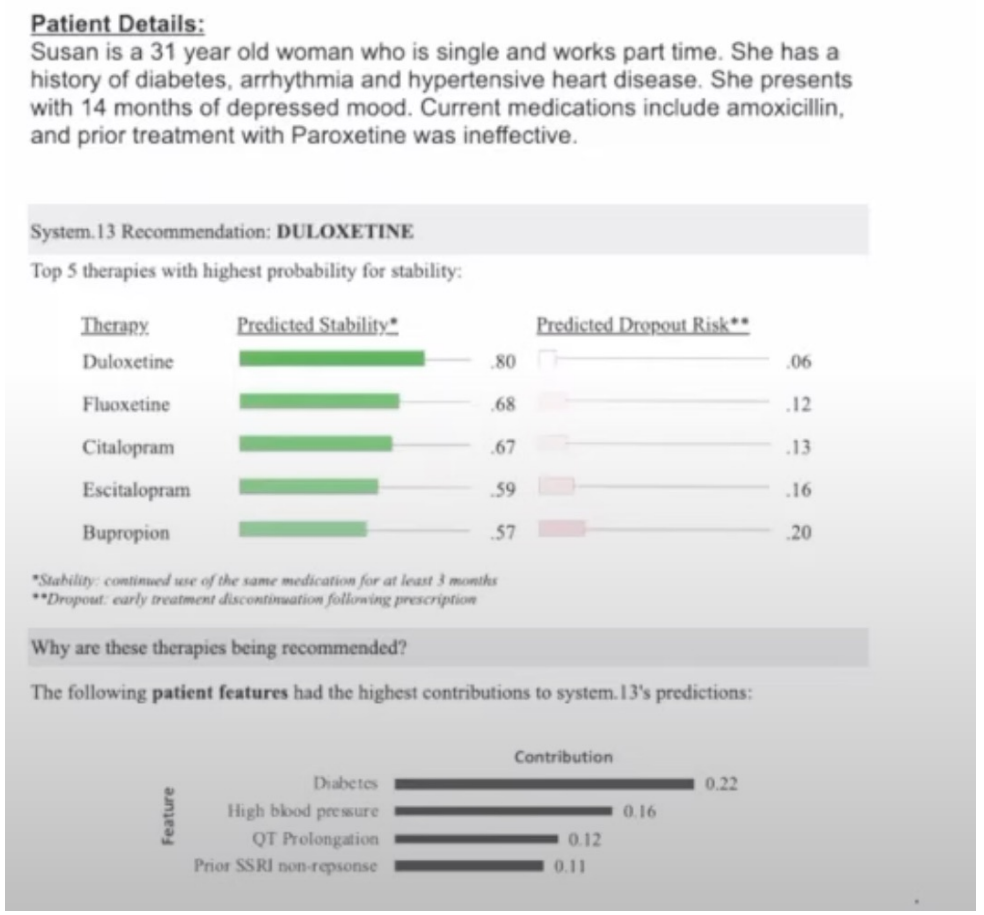
Validation in settings where humans are natural experts





Settings where Human+AI teaming is needed



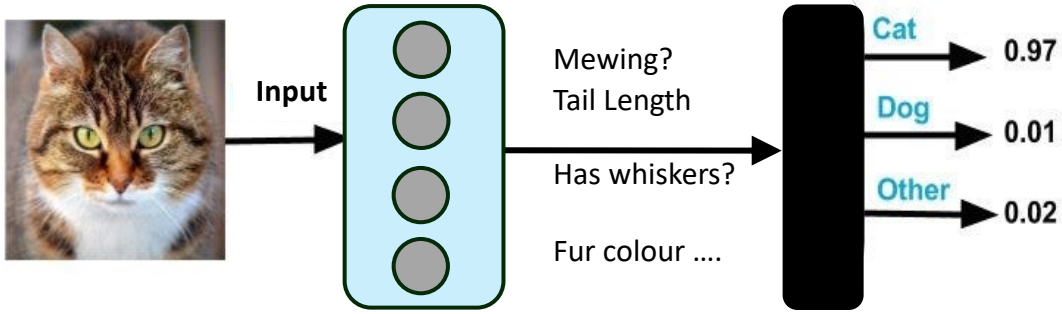


Concept Bottleneck Models

These models use input features to determine interpretable concepts and then predict labels using these concepts

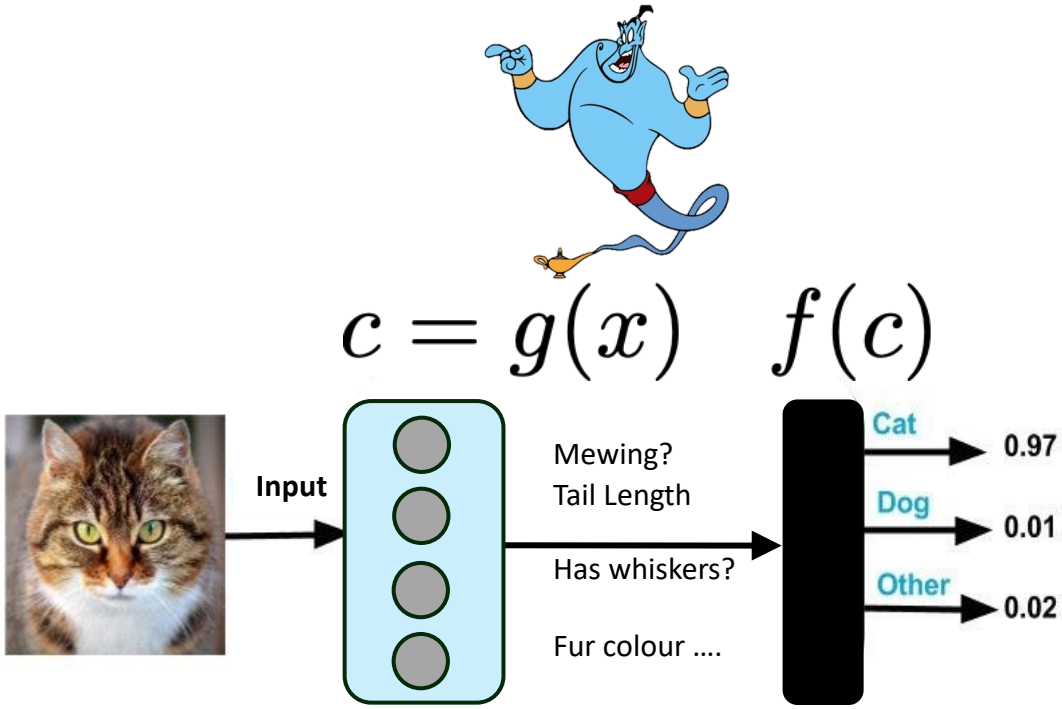


$$c = g(x) \quad f(c)$$



Concept Bottleneck Models

These models use input features to determine interpretable concepts and then predict labels using these concepts

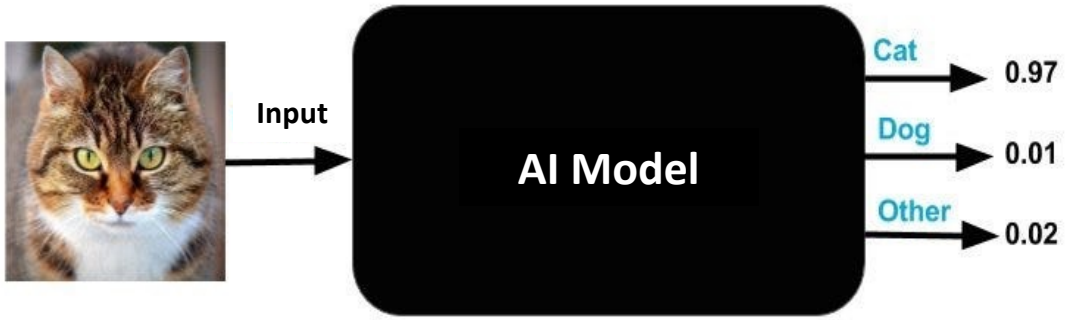


- ✓ Competitive accuracy with standard models
- ✓ Little tradeoff between concept and label accuracy
- ✓ **Provide semantically meaningful representations**
- ✓ **Ability to directly intervene on concepts**

Training Concept Bottleneck Models



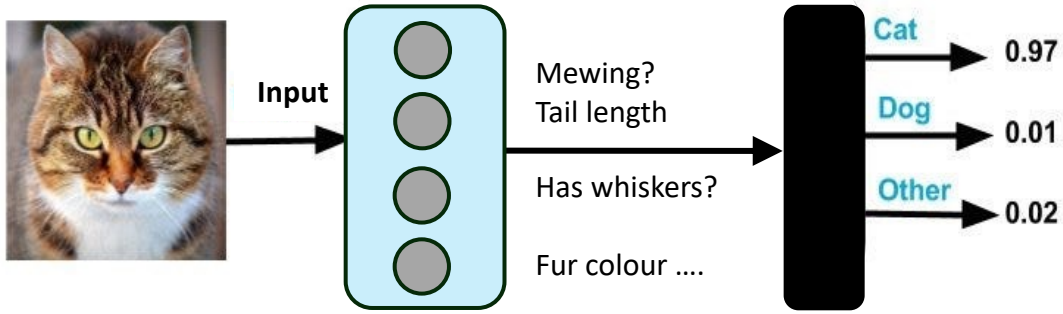
$f(x)$



$$\hat{y} = f(x)$$

$$\arg\min_{f} \sum_{i} L_Y(f(x^{(i)}); y^{(i)}))$$

$c = g(x)$ $f(c)$



Independent:

$$\hat{g} = \arg\min_{g} \sum_{i,j} L_{C_j}(g_j(x^{(i)}; c_j^{(i)}) \text{ and}$$

$$\hat{f} = \arg\min_{f} \sum_{i} L_Y(f(c^{(i)}; y^{(i)})$$

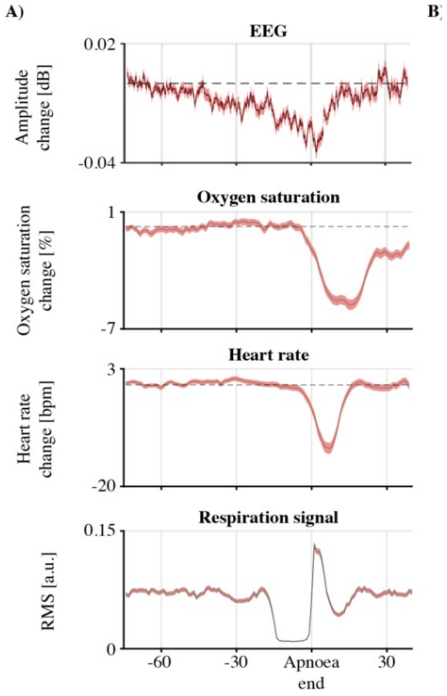
Sequential: $\hat{f} = \arg\min_{f} \sum_{i} L_Y(f(\hat{g}(x^{(i)}); y^{(i)})$

Joint:

$$\hat{f}, \hat{g} = \arg\min_{f,g} \sum_{i} L_Y(f(\hat{g}(x^{(i)}); y^{(i)}) + \sum_{j} \lambda_{C_j}(g(x^{(i)}); c^{(i)}))$$

Concept Bottleneck Models for Clinical Timeseries

Idea: Constrain time-series to retain only information relevant using concepts for determining outcomes;



X

Var(HR) > c

Recent SPO2 < c2

EEG decreasing

...



c

Hypertensive?

On Sedation?

On ventilation?

...

Prediction

$\hat{y}$

Interpretable Feature Summaries

Concept Bottleneck (**semantically meaningful**)





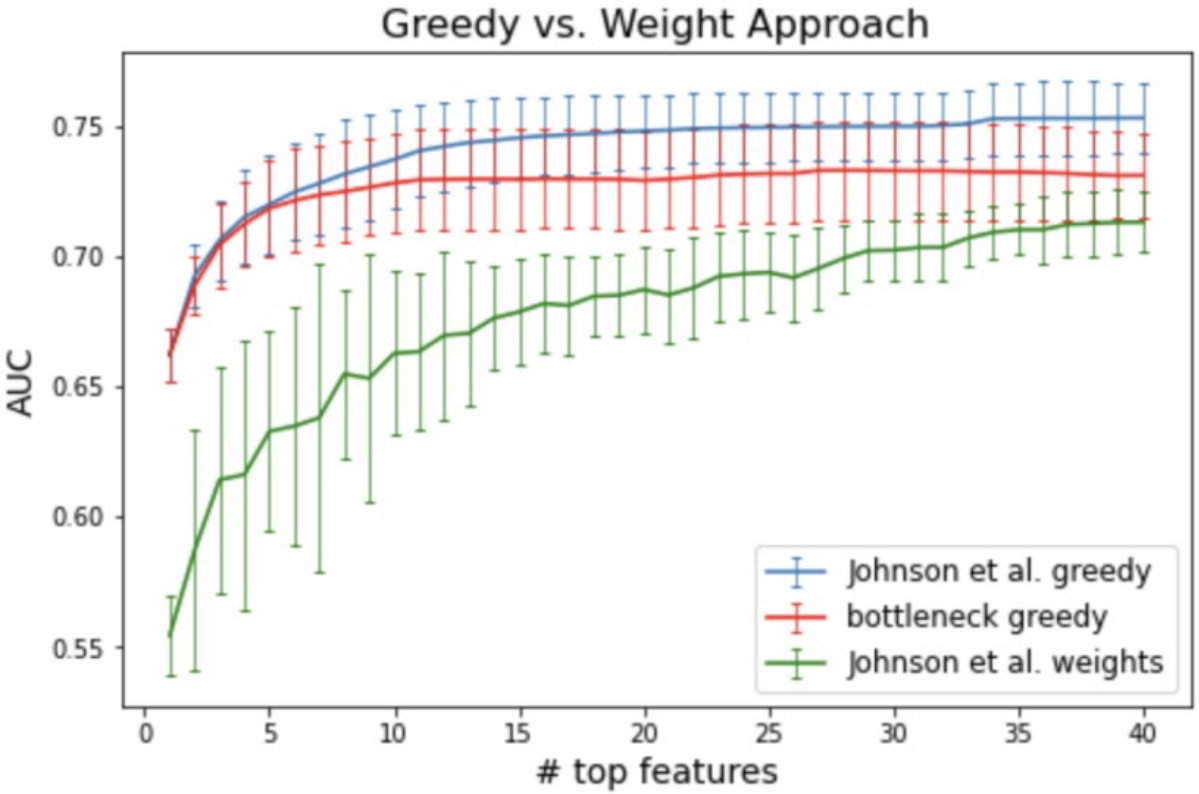
Learning Optimal Summaries of Clinical Time-series with Concept Bottleneck Models. Wu et al 2022, MLHC

Concept Bottleneck Models provide **good predictive performance** and **the ability to intervene** on derived concepts

Algorithm 1 Greedy Optimization Method

Require: $\boldsymbol{A}, \boldsymbol{s}, \boldsymbol{u}, \phi_b, f$

```
for i ∈ 1 : φ_b do
    maxAUC = 0
    z* = −1
    for k ∈ K do
        for z ∈ u_k do
            if AUC(f(z)) > maxAUC then
                z* = z
                c_k* = c_k
            end if
        end for
    end for
    s_k* = s_k* ∪ z*
    u_k* = u_k* \ z*
end for
Finetune f
```

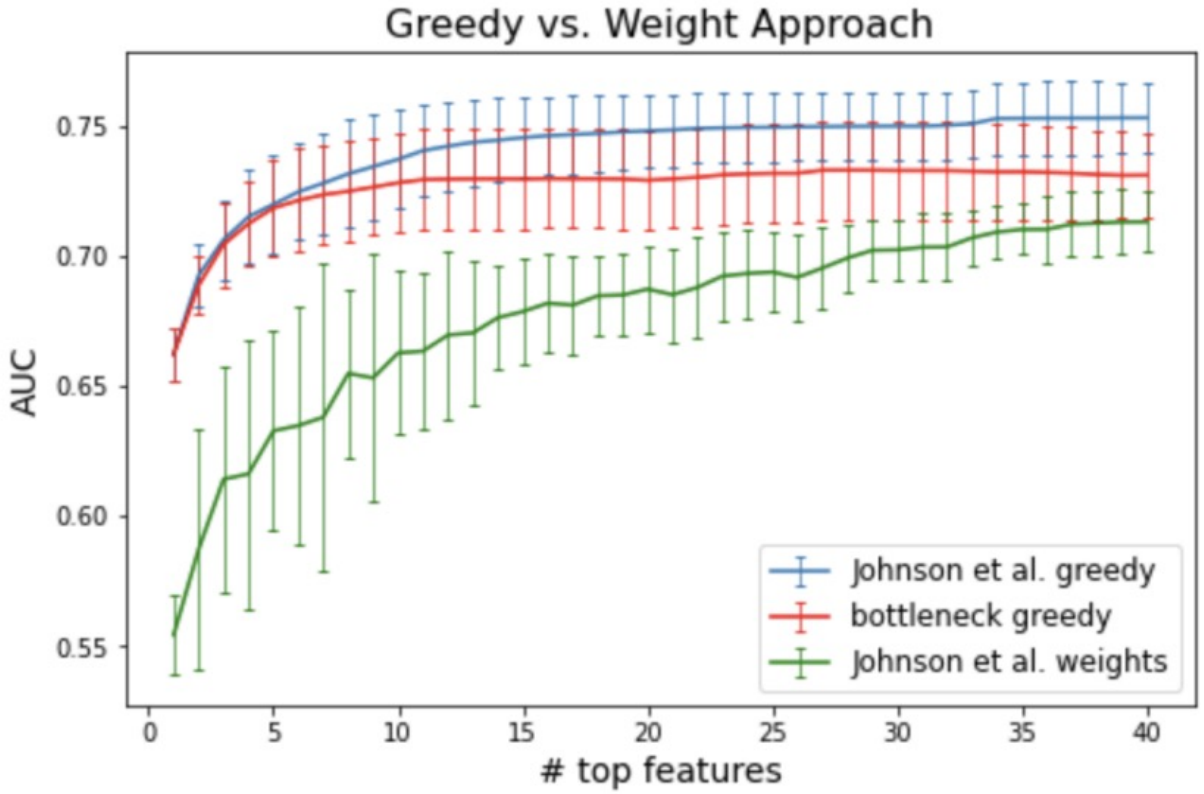




Learning Optimal Summaries of Clinical Time-series with Concept Bottleneck Models. Wu et al 2022, MLHC

Concept Bottleneck Models provide **good predictive performance** and **the ability to intervene** on derived concepts

	Concept Num	Concept Weight	Feat Name	Feat Summary	Feat Weight
0	0	0.527	urine	var of indicators	0.0002
1	0	0.527	map	last time measured	0.0263
2	0	0.527	bun_ind	N/A	-0.0002
3	1	0.420	po2	last time measured	-0.0005
4	1	0.420	lactate	ever measured	0.0001
5	2	0.162	pco2_ind	N/A	-0.0283
6	3	-2.102	po2	hours below threshold	-0.0744
7	3	-2.102	GCS	mean of indicators	0.0708
8	3	-2.102	inr	hours above threshold	-0.2857
9	3	-2.102	hr_ind	N/A	0.2521
10	3	-2.102	spontaneousrr	last time measured	0.0716
11	3	-2.102	alt	slope std err	-0.1284
12	3	-2.102	hr	last time measured	0.1028
13	3	-2.102	spo2	last time measured	0.0739

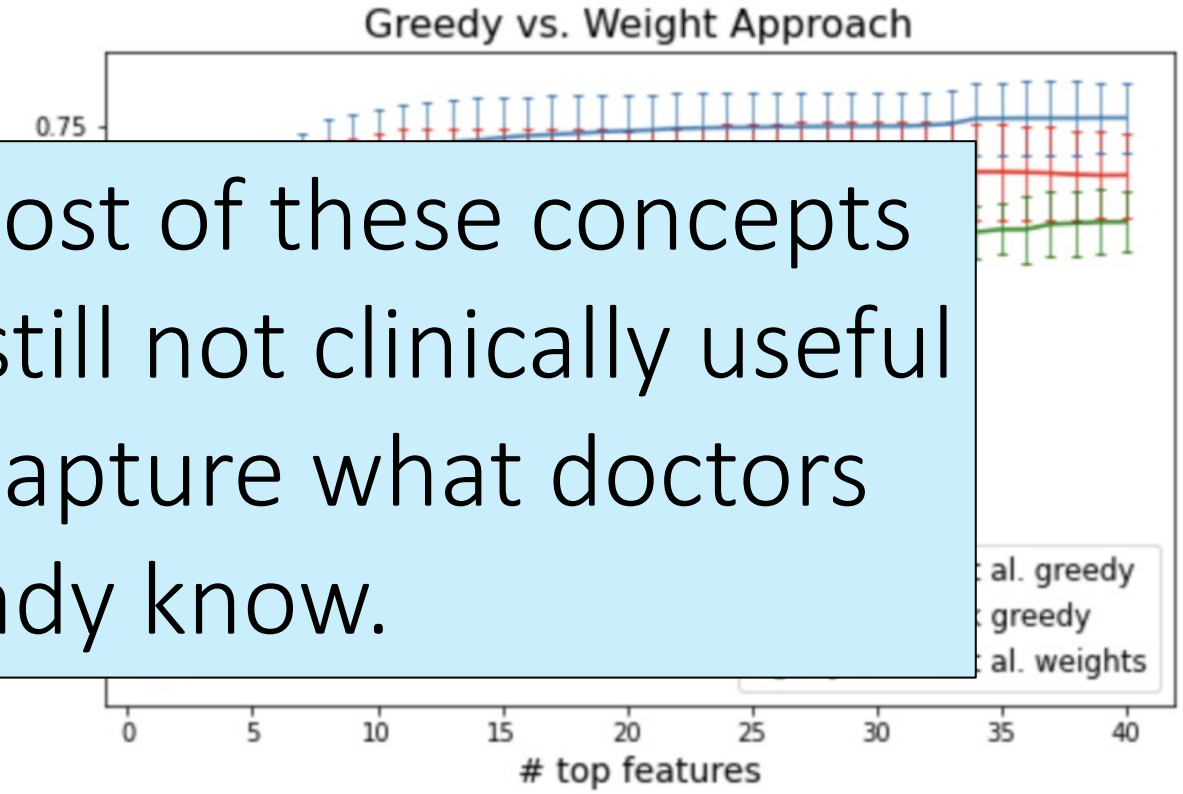




Learning Optimal Summaries of Clinical Time-series with Concept Bottleneck Models. Wu et al 2022, MLHC

Concept Bottleneck Models provide **good predictive performance** and **the ability to intervene** on derived concepts

	Concept Num	Concept Weight	Feat Name	Feat Summary	Feat Weight
0	0	0.527	urine	var of indicators	0.0002
1	0				
2	0				
3	1				
4	1				
5	2				
6	3				
7	3				
8	3				
9	3				
10	3				
11	3				
12	3	-2.102	hr	last time measured	0.1028
13	3	-2.102	spo2	last time measured	0.0739



Unfortunately, most of these concepts on their own are still not clinically useful as they largely capture what doctors already know.

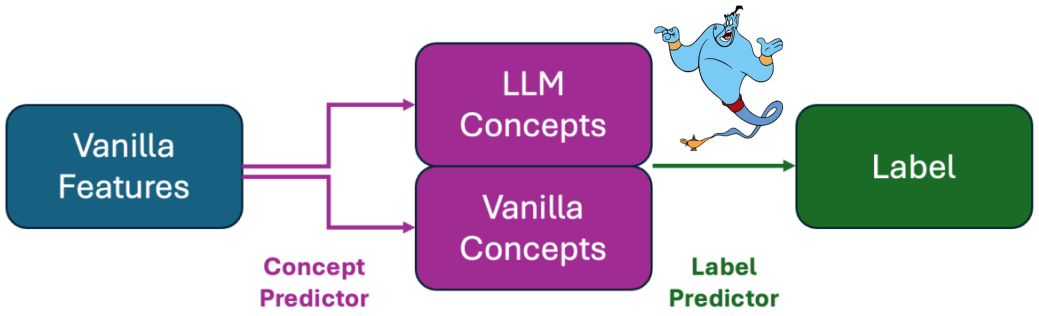


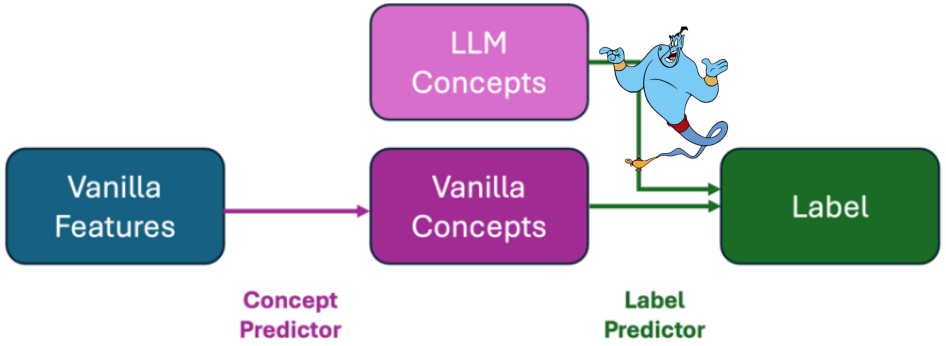
Learning Optimal Summaries of Clinical Time-series with Concept Bottleneck Models. Wu et al 2022, MLHC

Context matters in Concept Bottleneck Models

There is a lot of contextual information provided in clinical notes that is not captured in medical time-series. Can concepts help here?

Idea: Combine concepts learned from time-series with concepts derived using LLMs from clinical notes for additional context.





✗ Option 1: Use feature set to define both LLM-based concepts and concepts that summarise the time-series and predict label using the combination.

✓ Option 2: Derive LLM-based concepts independently from concepts that summarise the time-series and use these to jointly predict label

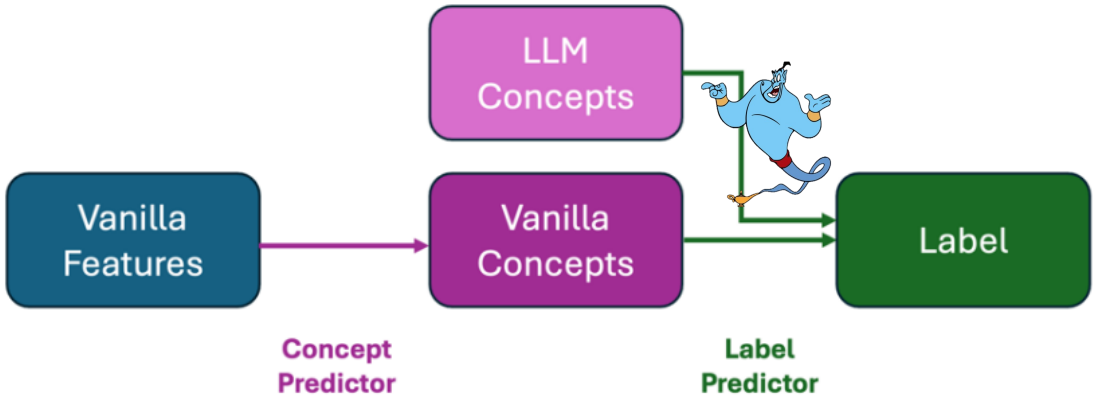


Augmenting Concept Bottleneck Models with Language-Based Context –Anish Narain, Dominic Marshall, Sonali Parbhoo, in submission 2024

Augmenting Concept Bottleneck Models with Contextual Information

- Use Meta's LLama3 on discharge summaries, radiology reports and ECG studies to extract contextual summaries

```
template=(
    "Context: You are a clinician receiving chunks of radiology reports for
        patients in an ICU. Please do the reviewing as quickly as possible.\n"
    "Task: Determine if the patient suffered from bilateral infiltrates.\n"
    "Instructions: Answer with 'Yes' or 'No'. If there is not enough
        information, answer 'No'.\n"
    "Discharge Text:\n{radiology_texts}\n\n"
    "Query: Does the chunk of text mention that the patient suffered from
        bilateral infiltrates? Answer strictly in 'Yes' or 'No'.  Then provide
        a reason for your response."
),
input_variables=["radiology_texts"]
```



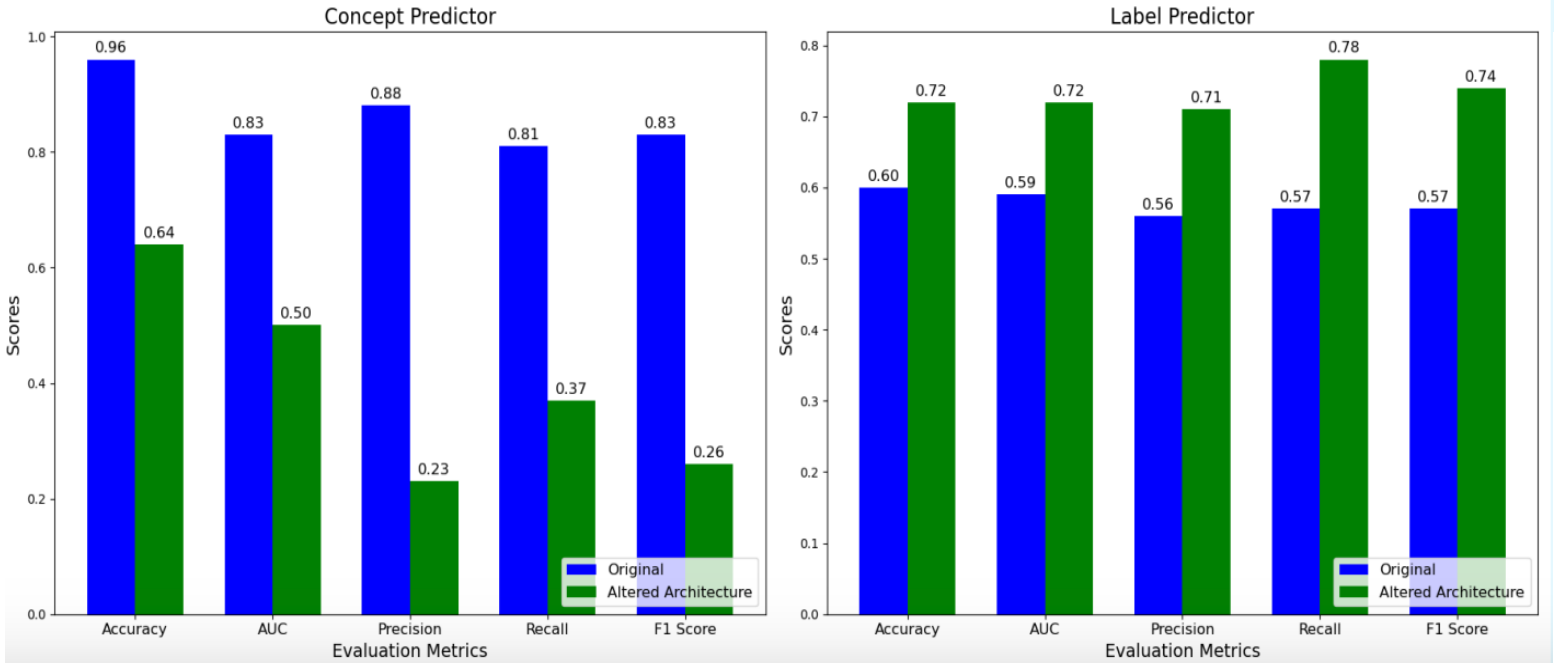
Clinical Label	Average Accuracy (%)
Mention of ARDS	93.5
Aspiration	87.4
Bilateral Infiltrates	71.7
Cardiac Arrest	80.3
Cardiac Failure	81.2
Pancreatitis	74.6
Pneumonia	92.7
TRALI	96.3

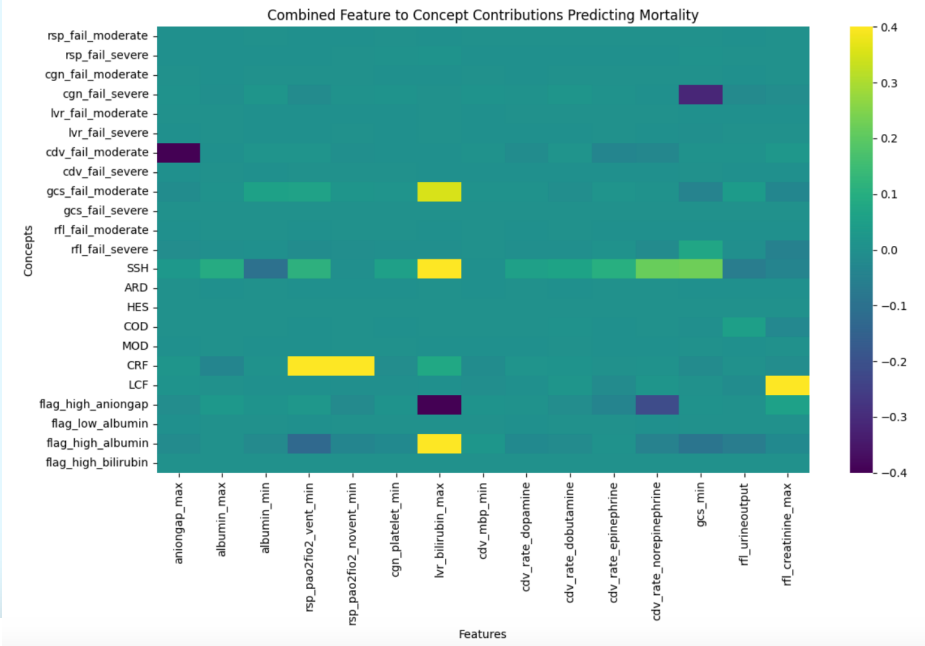


Augmenting Concept Bottleneck Models with Language-Based Context –Anish Narain, Dominic Marshall, Sonali Parbhoo, in submission 2024

Context-Aware Concept Bottleneck Models have **improved predictive performance** and extract **clinically relevant concepts**







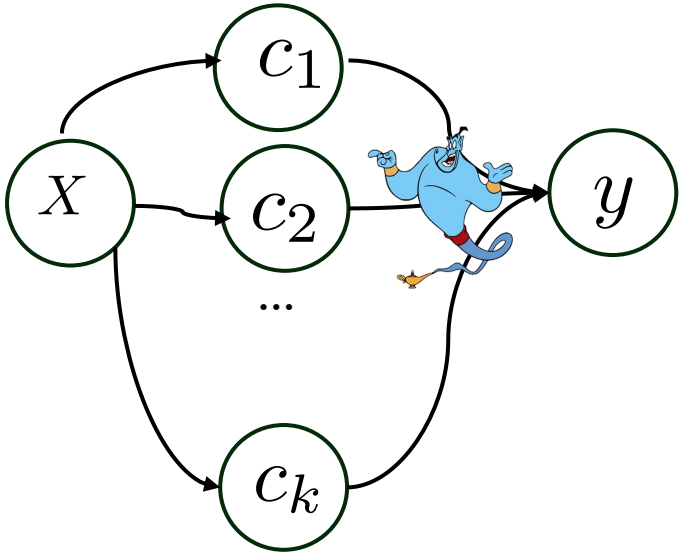
- Context-Aware CBMs identify those key concepts critical for diagnosing respiratory distress.



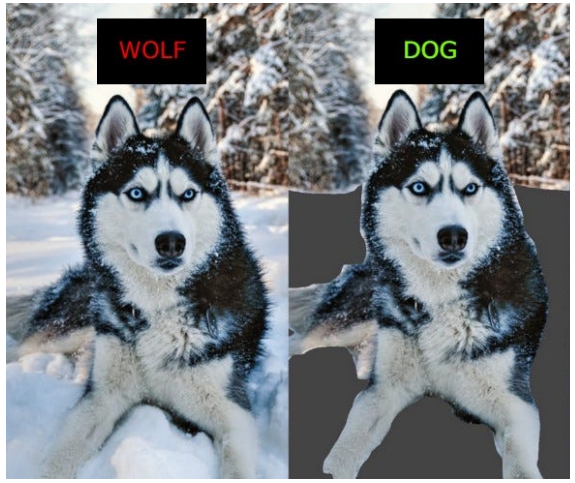
Augmenting Concept Bottleneck Models with Language-Based Context –Anish Narain, Dominic Marshall, Sonali Parbhoo, in submission 2024



Concept Bottleneck Models Can Leak Information



- Unlike hard concept bottleneck models, “soft” concept layers (e.g. sigmoids) allow for efficient optimization.
- However, soft concept layers can leak information. E.g. if p(c1) = 0.9-1.0 *both c1 and c2 are present*; if p(c1) = 0.7-0.9 *only c1 is present*. i.e **knowing something about one concept might tell us about the other concepts indirectly.**



No. Concepts	Mean "Hard" Accuracy	Mean "Soft" Accuracy	Lift
1	55.76%	59.16%	3.40%
2	65.31%	71.57%	6.26%
5	72.19%	80.62%	8.43%
10	73.29%	82.69%	9.39%
50	78.18%	89.68%	11.50%
100	78.27%	92.39%	14.12%

Promises and Pitfalls of Concept Bottleneck Models – Mahinpei et al, Proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI,. 2021;1 :1-13.

Types of Information Leakage



Leakage is the amount of unintended information that is leaked from the input X to the label through a soft concept $\hat{c}$, which is not present in its hard representation c. Leakage can occur for one of two reasons:

- **Leakage due to inherent uncertainty on concept predictions:**

The concept sets themselves may be incomplete or insufficient for capturing all the relevant information about a problem. Additional information from the inputs may bypass the concept layer and be leveraged for predicting labels.

- **Leakage through label predictor:**

Performance of a joint CBM may depend on trading label accuracy and concept accuracy in joint training through parameter λ. When this parameter is small, the label predictor might cheat and shape the concept probabilities unintuitively to achieve a higher task accuracy.

Promises and Pitfalls of Concept Bottleneck Models – Mahinpei et al, Proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI,. 2021;1 :1-13.

Types of Information Leakage



Leakage is the amount of unintended information that is leaked from the input X to the label through a soft concept $\hat{c}$, which is not present in its hard representation c. Leakage can occur for one of two reasons:

- **Leakage due to inherent uncertainty on concept predictions:**

The concept sets themselves may be incomplete or insufficient for capturing all the relevant information about a problem. Additional information from the inputs may bypass the concept layer and be leveraged for predicting labels.

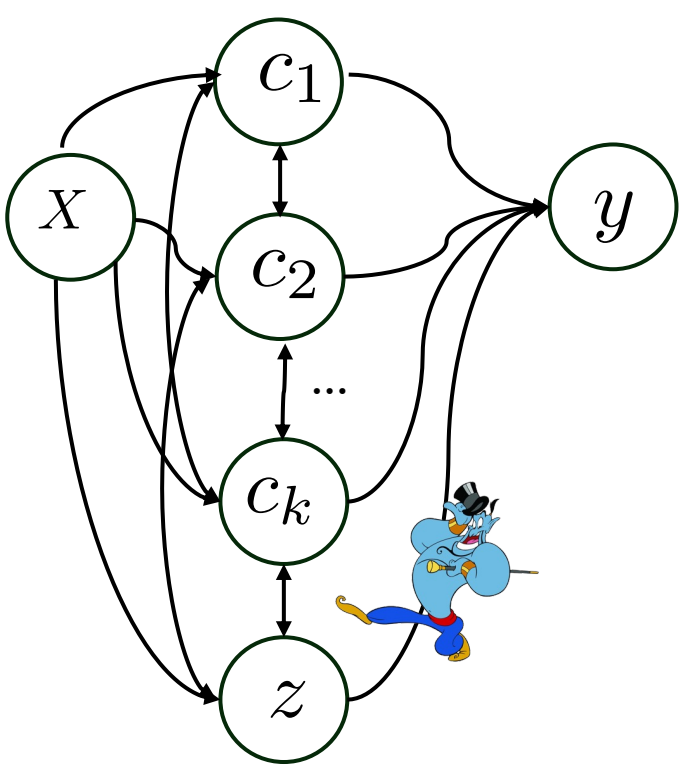
- **Leakage through label predictor:**

Performance of a joint CBM may depend on trading label accuracy and concept accuracy in joint training through parameter λ. When this parameter is small, the label predictor might cheat and shape the concept probabilities unintuitively to achieve a higher task accuracy.

Promises and Pitfalls of Concept Bottleneck Models – Mahinpei et al, Proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI,. 2021;1 :1-13.

Capturing Leaky Information Through A Side Channel

Idea: Train hard CBMs and marginalize to have same predictive power as soft concept bottlenecks



Autoregressive Concept Modelling

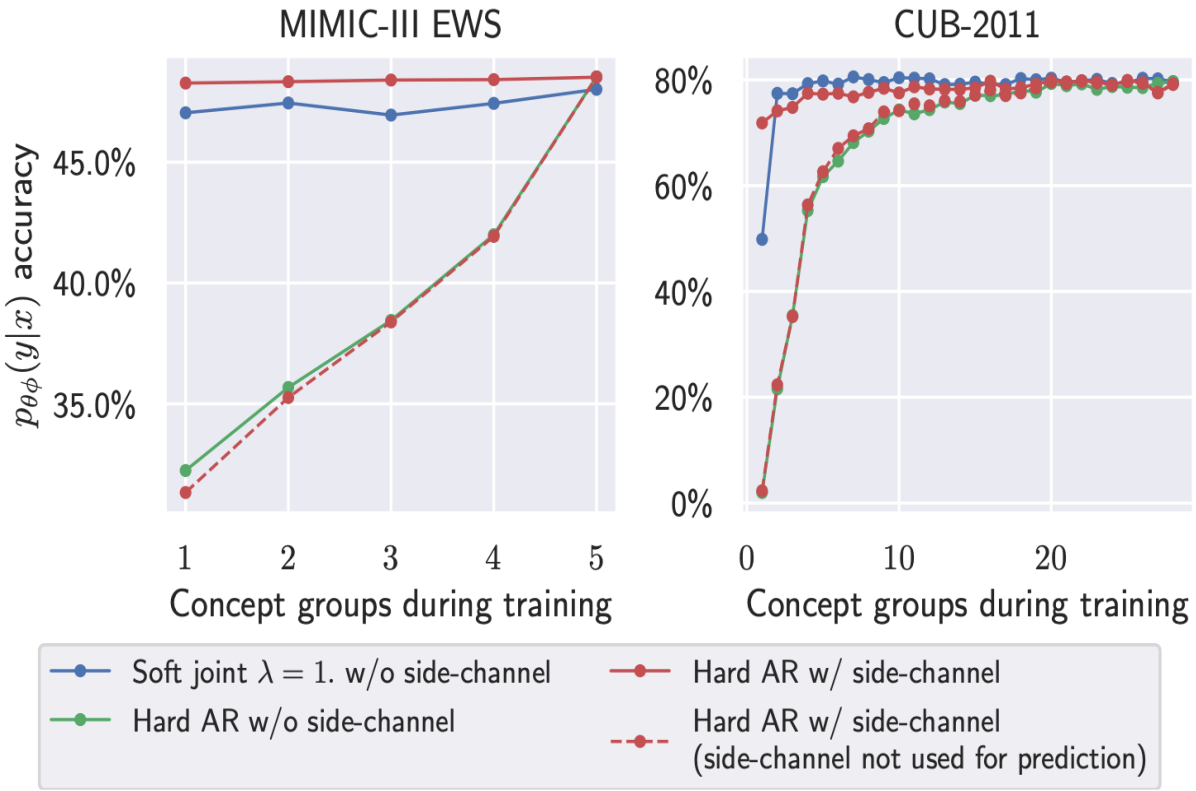
Capturing additional leaky information







CBMs with a side channel **achieve same performance as leaky counterparts that cheat**



Model	MIMIC-III EWS		CUB-2011	
	$p_\theta(c\|x)$	$p_{\theta\phi}(y\|x)$	$p_\theta(c\|x)$	$p_{\theta\phi}(y\|x)$
Soft Independent	38.7±0.0	45.9±0.1	48.1±0.8	72.8±0.8
Soft Sequential	38.7±0.0	47.6±0.0	46.1±0.4	73.9±0.1
Soft Joint $\lambda = 1$	38.7±0.0	48.0±0.0	47.7±0.2	75.4±0.1
Soft Joint $\lambda = 0.1$	35.6±0.1	48.0±0.0	47.9±0.4	**82.7±0.2**
Soft Joint $\lambda = 0.01$	32.1±0.1	47.9±0.0	0.1±0.0	82.1±0.3
Hard	38.7±0.0	46.6±0.0	68.0±0.2	79.5±0.3
Hard AR (This work)	**41.0±0.0**	**48.5±0.0**	**81.4±0.1**	81.7±0.2

Despite their performance, CBMs with side channels do not enable us to **inspect leaky information explicitly**.





Leakage in Hard vs Soft Concept Bottlenecks



- Hard concept bottleneck models do not suffer from leaky information due to uncertainty on concepts, since subjects are assigned hard concept labels.

- Soft concept bottleneck models have this issue because concept sets may be **incomplete and/or subject to unobserved/hidden variables**.

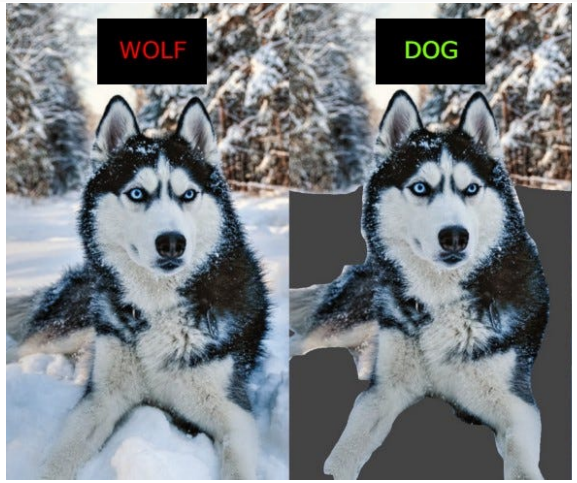
- Leakage due to uncertainty on concepts occurs more frequently in **sequentially trained soft CBMs.**



has-back-colour::black = 0.20



has-back-colour::black = 0.83

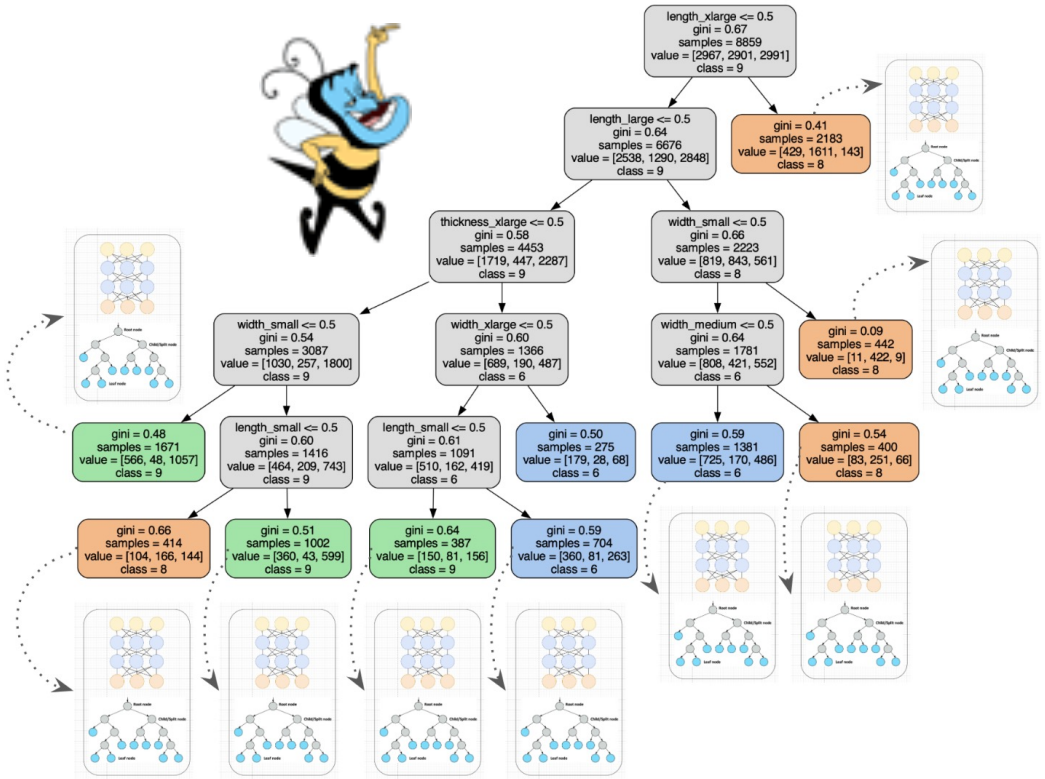


Promises and Pitfalls of Concept Bottleneck Models – Mahinpei et al, Proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI,. 2021;1 :1-13.

Visualising Leaky Information with Trees

Idea: If a tree using leaky concept probabilities can further classify a subset of the data ending up in a leaf of a tree that only uses hard concepts, then information leakage is observed for this subset.

- The naïve approach of comparing a hard CBM with a trained sequential CBM is non-trivial. Instead, we can fit sequential CBMs to specialise the predictions of a hard CBM based using leakage.



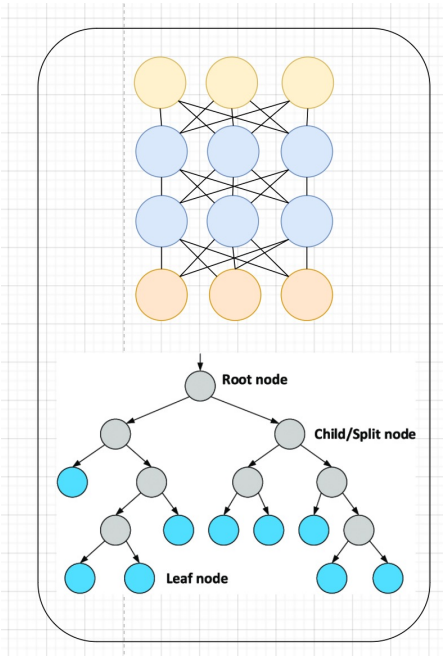


Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

Visualising Leaky Information with Trees

Step 1: Train a Hard CBM using a tree for the label predictor. That is, C→Y Decision Tree trained with Hard Concepts [0 or 1].



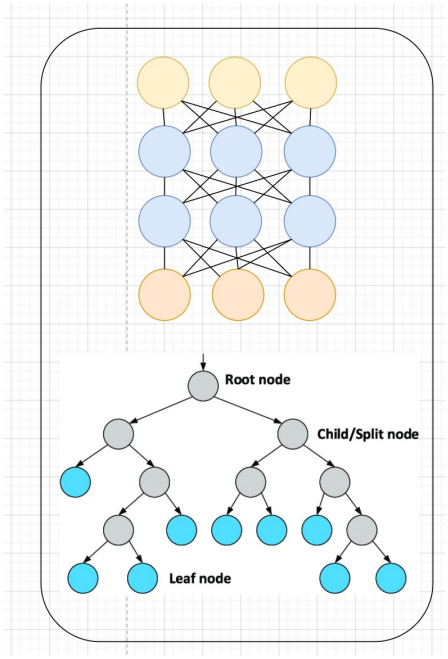




Visualising Leaky Information with Trees

Step 1: Train a Hard CBM using a tree for the label predictor. That is, C→Y Decision Tree trained with Hard Concepts [0 or 1].



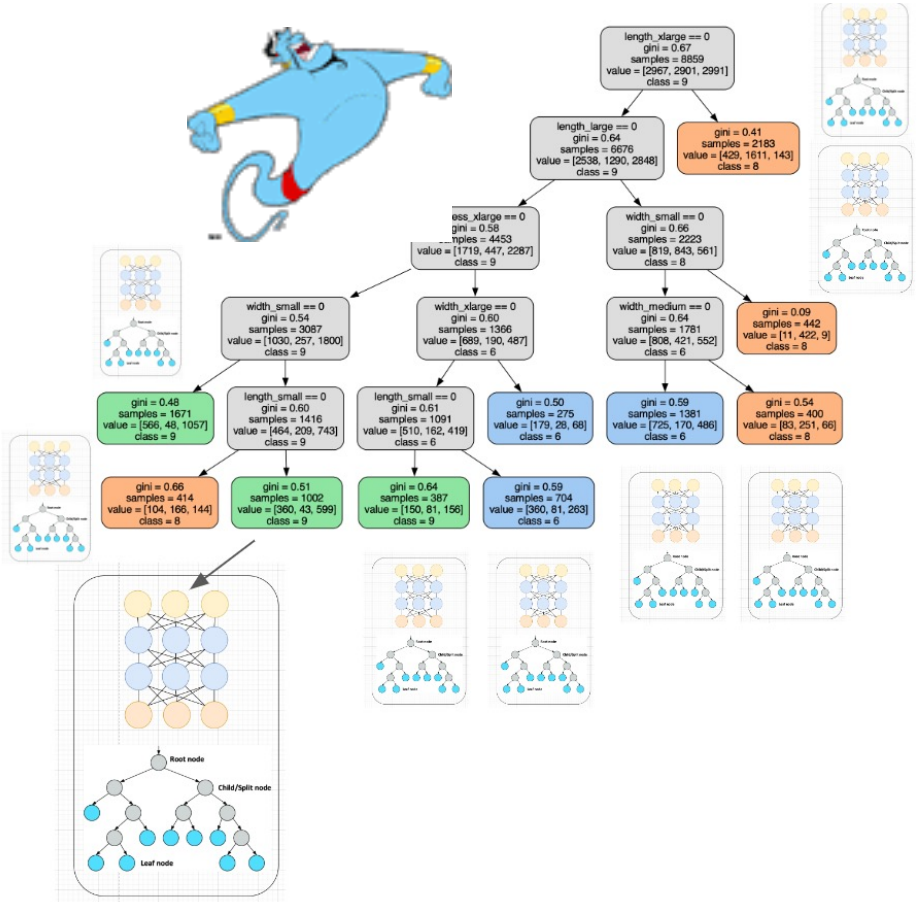


Step 2: Extract the decision paths of the tree and the hard concepts values used per path.



Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

Visualising Leaky Information with Trees



C→Y Decision Tree trained with **Mixed** Concepts (**Hard** + **Soft** Probability Concepts

Step 3: Train a Sequential CBM with Mixed Concepts for each leaf. That is, Copy the trained X→C network.

- For the concepts not appearing in the decision path of the leaf, assign the corresponding **hard value** (0 or 1) as input to the C→Y decision tree.

- For the concepts appearing in the decision path of the leaf, use the **soft probability** of the X→C as input to the C→Y decision tree

- Merge the DTs into a single tree C→Y and perform the inspection



Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

An Example

- Two indistinguishable digits based on their hard concepts



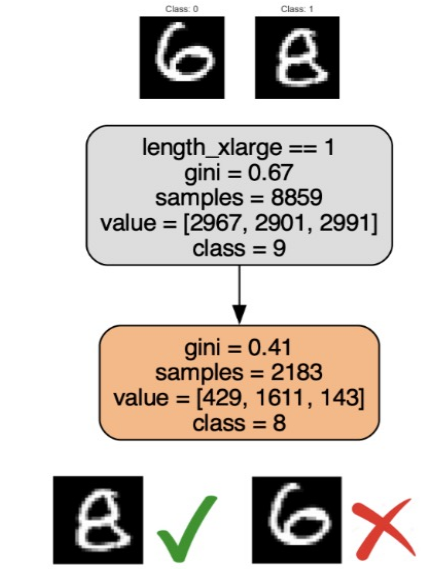


	Six	Eight
True Concepts	Thickness = 2.49 Width = 18.56 Length = 60.50	Thickness = 2.46 Width = 16.10 Length = 63.96
Categorical Concepts	Thickness = **large** Width = **xlarge** Length = **xlarge**	Thickness = **large** Width = **xlarge** Length = **xlarge**
	Thickness = [0,0,1,0] Width = [0,0,0,1] Length = [0,0,0,1]	Thickness = [0,0,1,0] Width = [0,0,0,1] Length = [0,0,0,1]
Soft Concepts	Probability of xlarge length: **0.9854**	Probability of xlarge length: **0.9993**

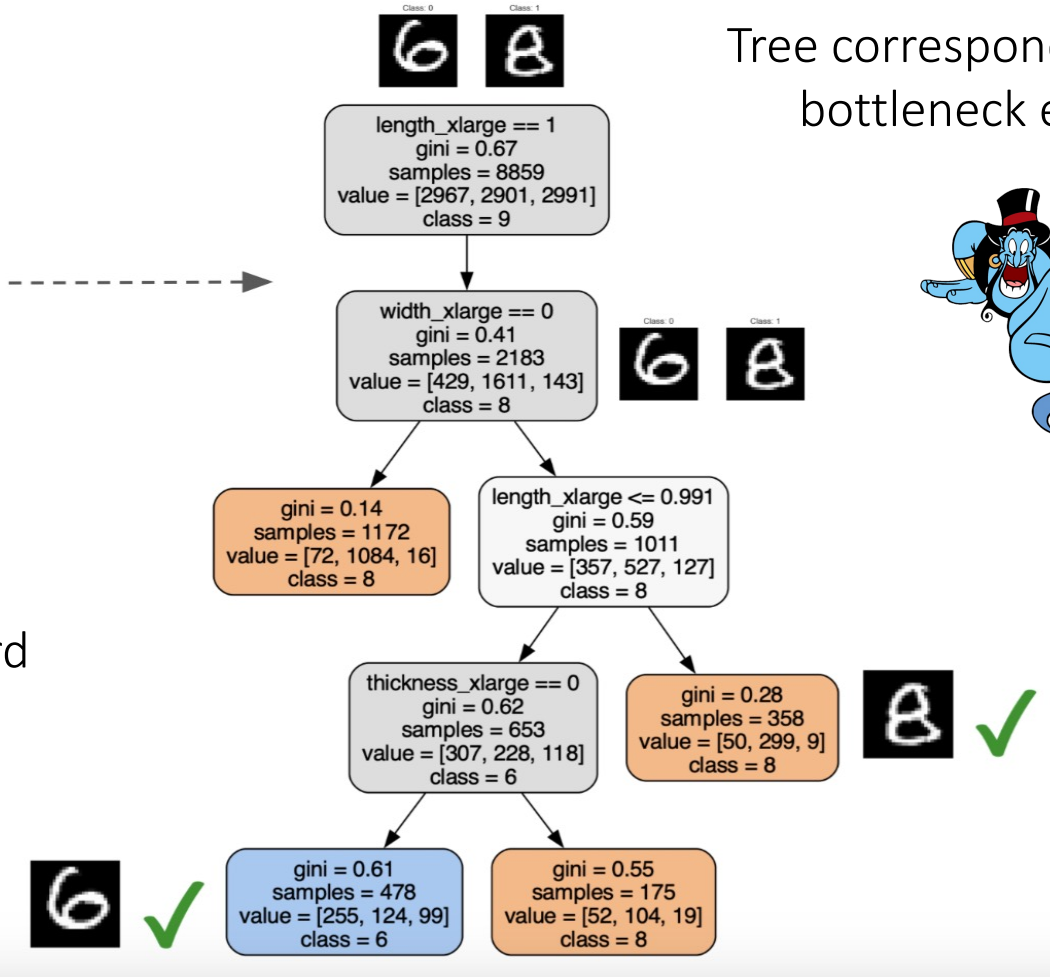


Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

An Example



Global tree based on hard concepts

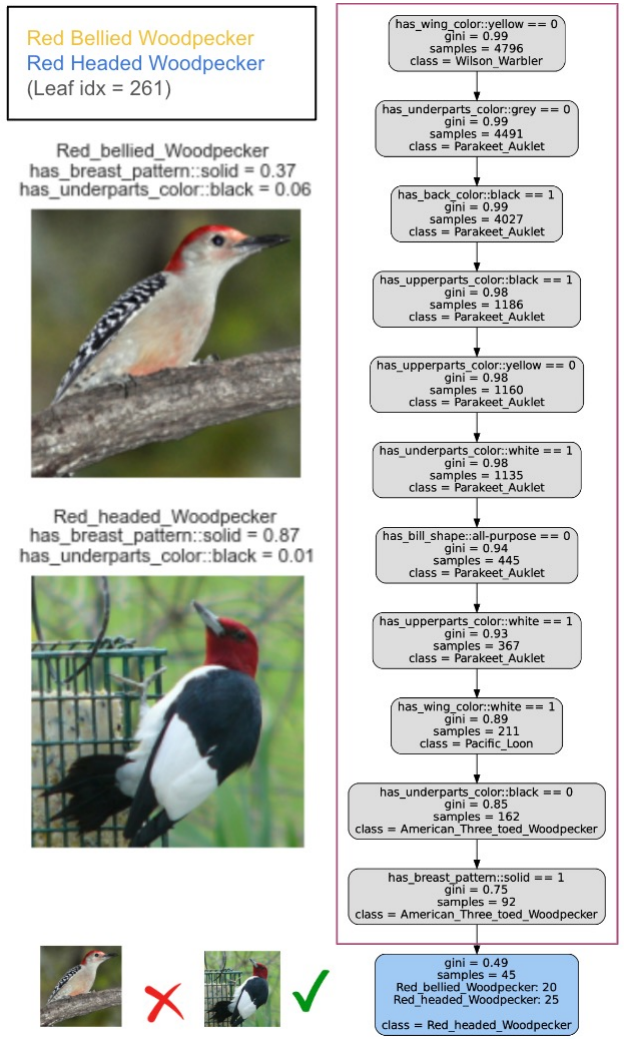


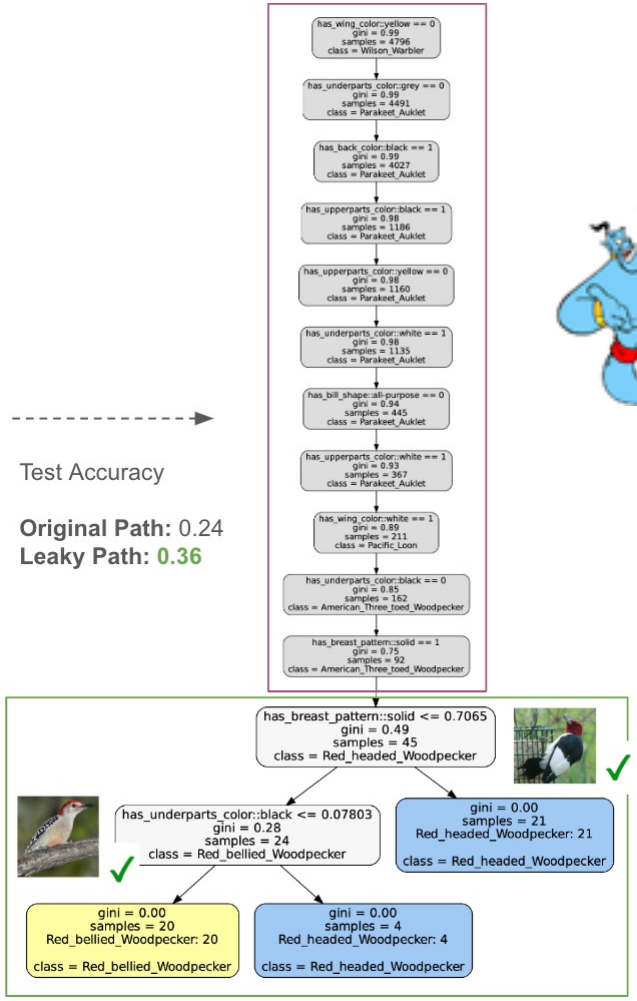
Tree corresponding to leaky bottleneck extension



Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

Tree-based leakage inspection helps us **explain which concepts are leaky**









Hierachical Leakage Control in Concept Bottleneck Models with Trees– Angelos Ragkousis and Sonali Parbhoo, arXiv preprint, 2024.

Detour: Evaluating sequential decisions with offline batch data

Given data collected by some **behaviour policy** π_b, can we estimate the value of **another evaluation policy** π_e? This is called **off-policy evaluation.**

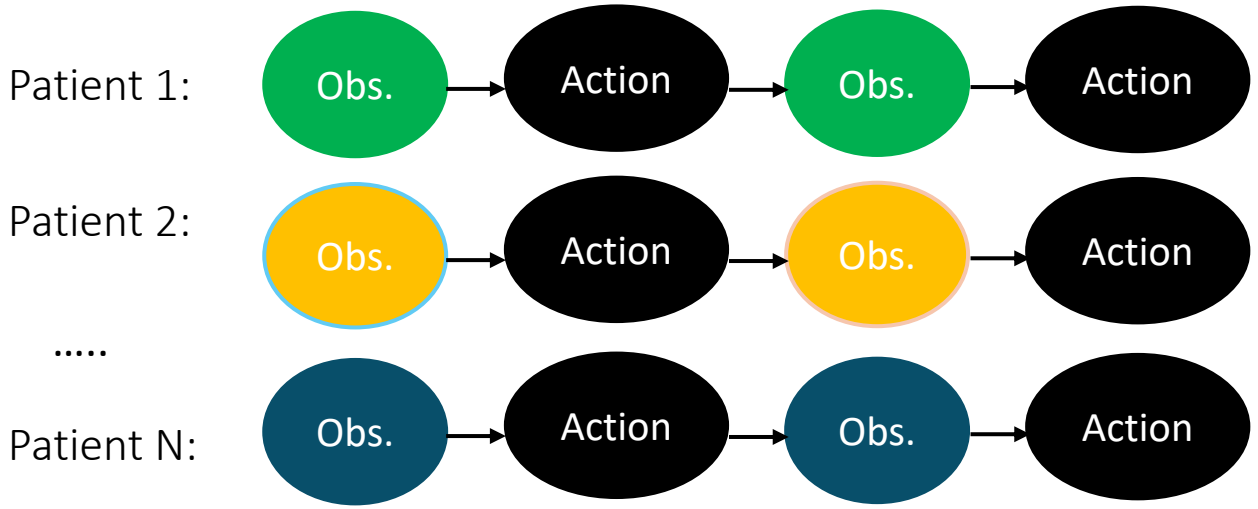
Clinician:



π_b

Evaluation Agent:

π_e

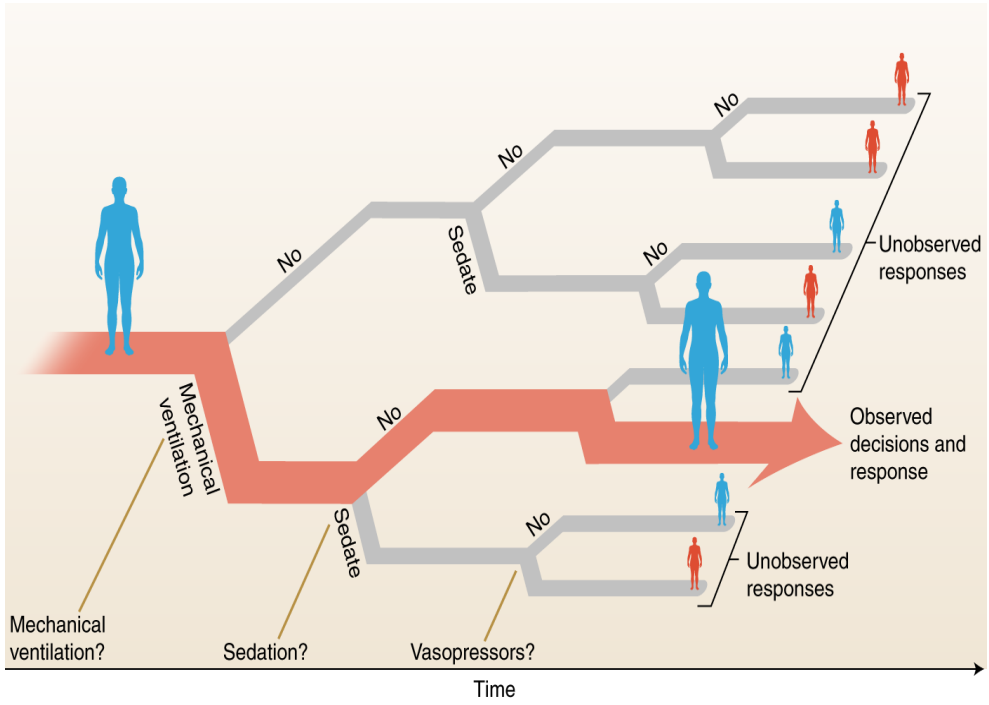


Reweight outcomes according to how much the trajectories overlap:

$$\hat{V}_{\pi_e}^{PDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho_{0:t}^{(n)} \cdot r_t^{(n)} \qquad \rho_{0:T}^{(n)} = \prod_{t=0}^{T}\frac{\pi_e(a_{t+1}^{(n)},\ s_t^{(n)})}{\pi_b(a_{t+1}^{(n)},\ s_t^{(n)})}$$

Detour: Challenges of Off-Policy Evaluation

1. Inability to explore
2. Small data (relatively) → shrinking data support as best strategies are discovered



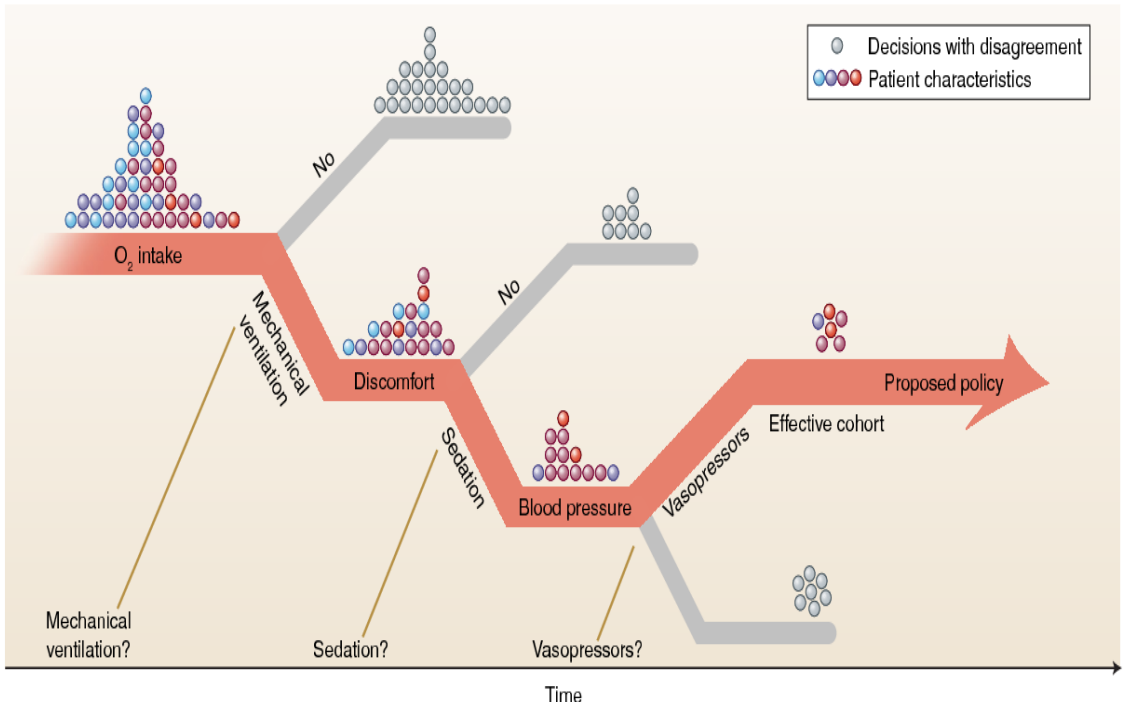
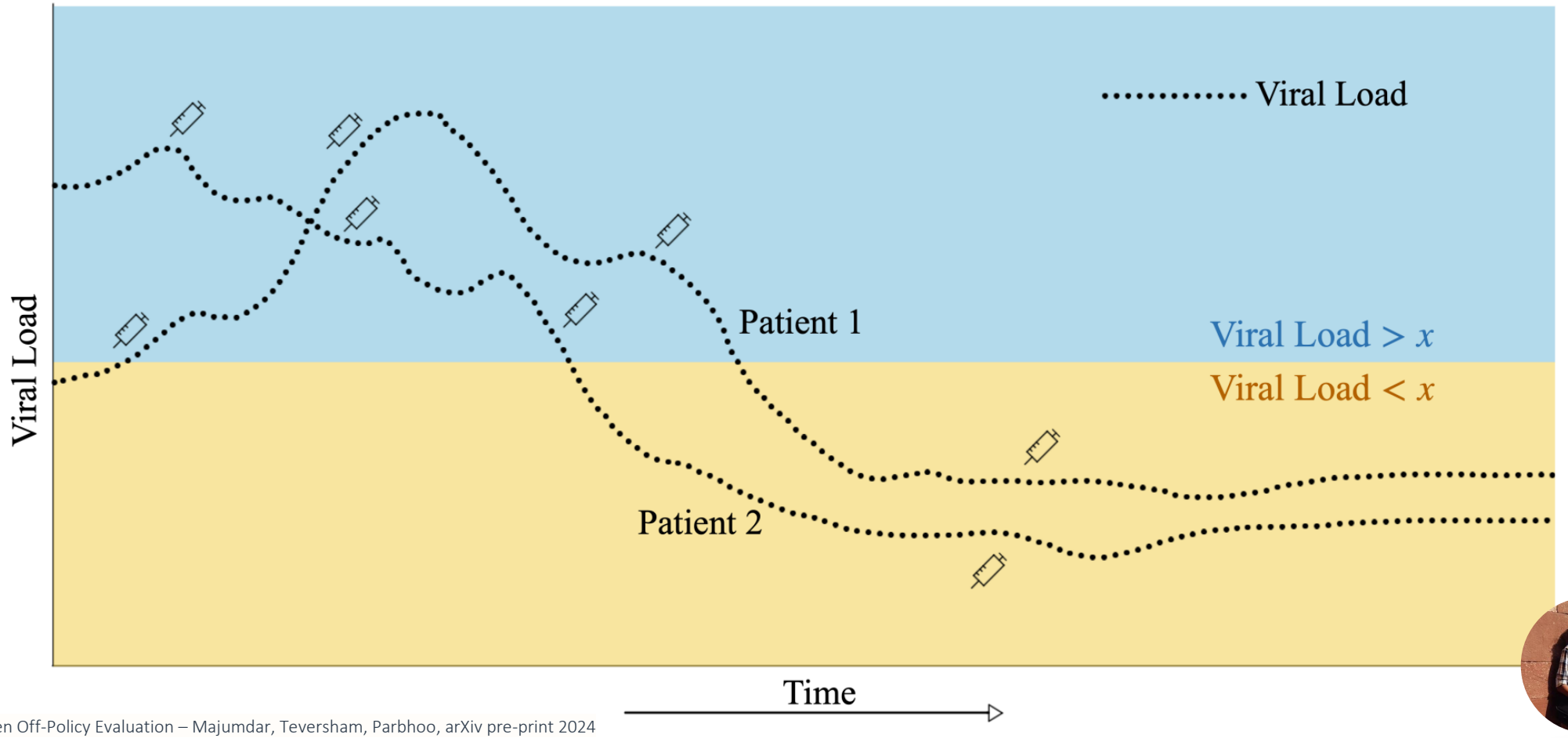
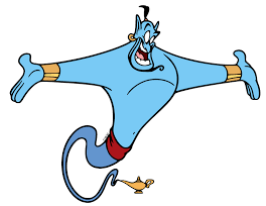


Image Credit: *Gottesman, Omer, et al. "Guidelines for reinforcement learning in healthcare." Nat Med 25.1 (2019): 16-18.*

Can concepts help in off-policy evaluation?

Idea: It is easier to induce overlap across some concepts than specific states and actions over time across an entire trajectory









Concept-Driven Off-Policy Evaluation

- Decompose evaluation based on interpretable concepts that account for similar histories, transitions, short and/or long-term benefits.

Then summarise behaviour and evaluation policies in terms of these concepts:

IS reweighting:

$$\hat{V}_{\pi_e}^{PDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho_{0:t}^{(n)} \cdot r_t^{(n)}$$

$$\rho_{0:t}^{(n)} = \prod_{t'=0}^{t}\frac{\pi_e(a_{t'+1}^{(n)},\ s_{t'}^{(n)})}{\pi_b(a_{t'+1}^{(n)},\ s_{t'}^{(n)})}$$

Concept Based IS reweighting:



$$\hat{V}_{\pi_e}^{CPDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \rho_{0:t}^{(n)} r_t^{(n)}$$

$$\rho_{0:t}^{(n)} = \prod_{t'=0}^{t}\frac{\pi_e(a_{t'}^{(n)}|c_{t'}^{(n)})}{\pi_b(a_{t'}^{(n)}|c_{t'}^{(n)})}$$

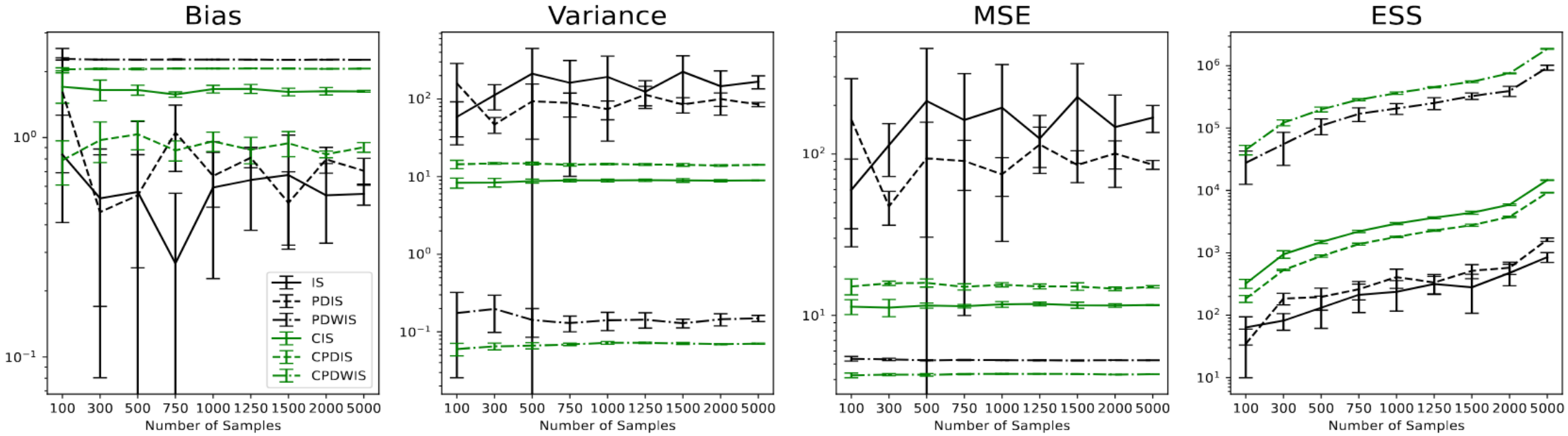




Concept-based estimators have **lower bias and variance** and **better overlap**

Concept-based estimators achieve better performance across various metrics than their non-concept-based counterparts.





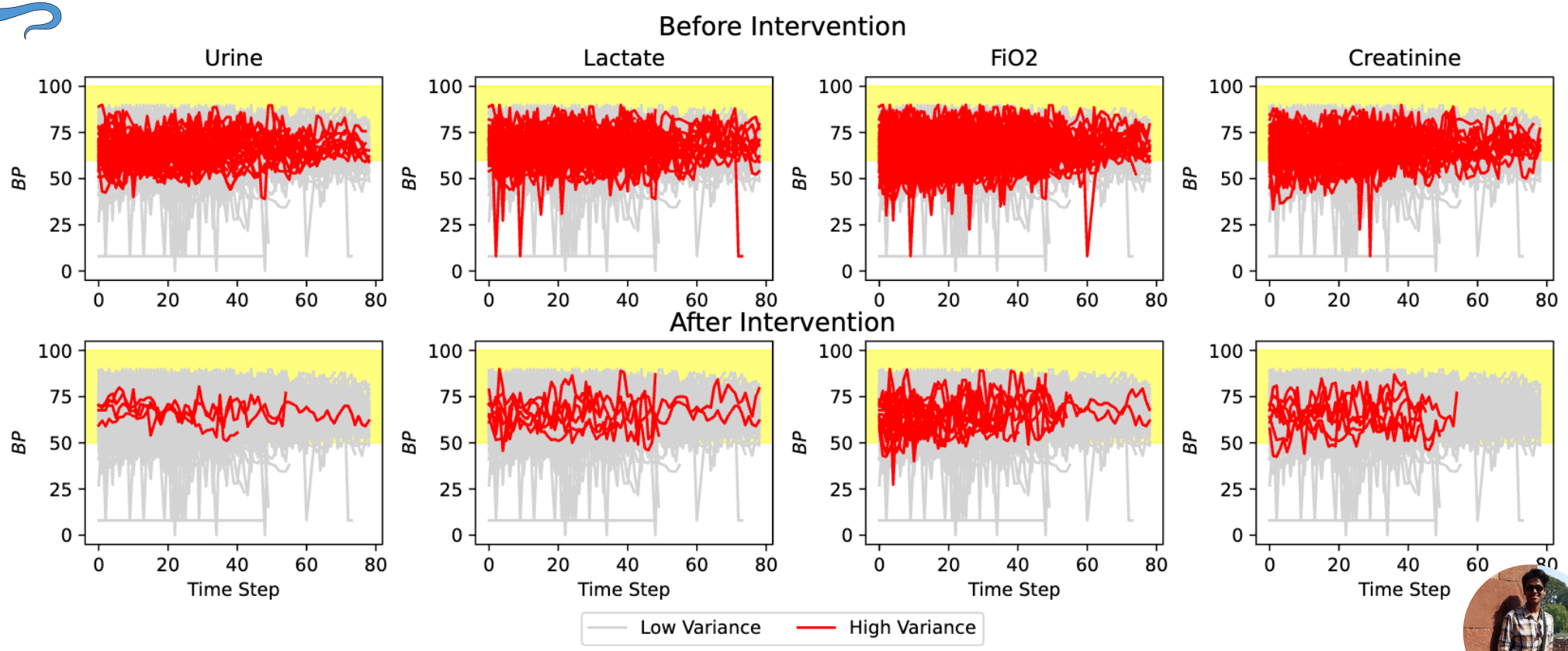




Concept-based estimators allow us to **intervene on contributors to high variance** during evaluation



We can isolate those concepts that contribute to the highest variance in OPE and intervene on those concepts to remedy the issues.



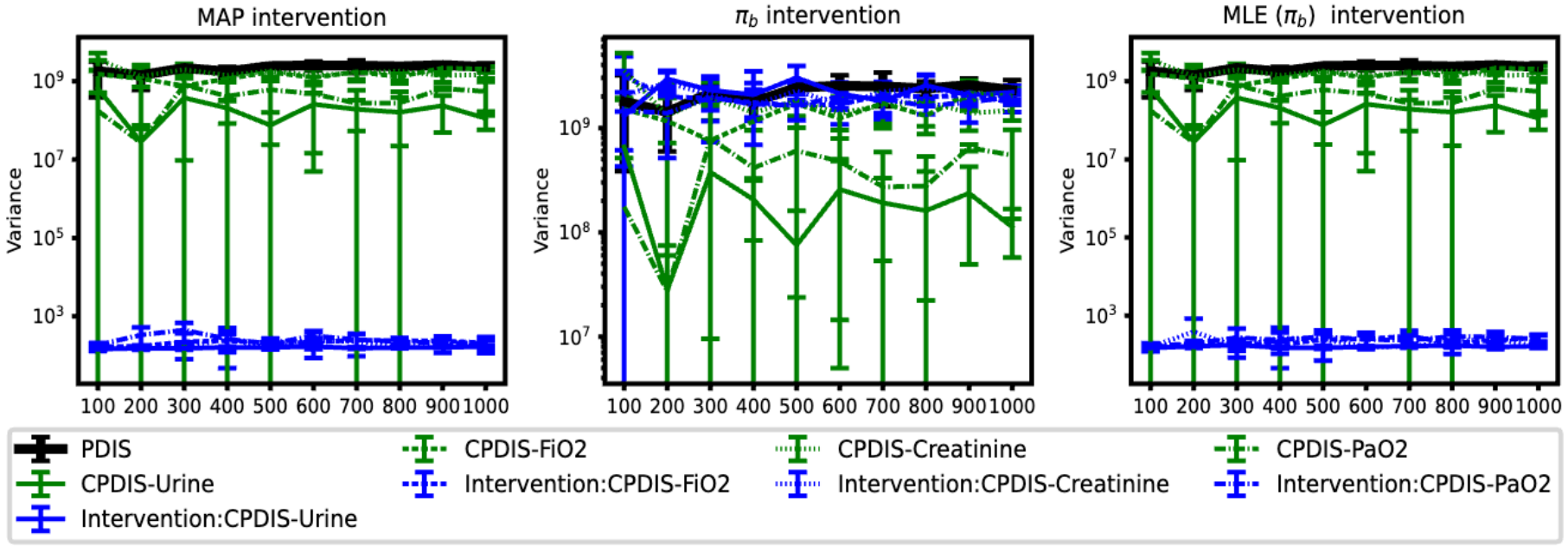




Concept-based estimators allow us to **intervene on contributors to high variance** during evaluation



We can isolate those concepts that contribute to the highest variance in OPE and intervene on those concepts to remedy the issues.







Summary

- **AI requires oversight** and interpretability is valuable in many scientific contexts.



- The ‘right’ form of interpretability **depends on the context**.
- Concepts are useful for providing insights in ML models as they retain semantic meaning. These can be enhanced by incorporating contextual information.
- However, concept bottleneck models suffer from leakage which can be visualized using trees and mitigated using side channels.
- Concepts aid in evaluation of decisions and **provide insights on where evaluation can go wrong.**