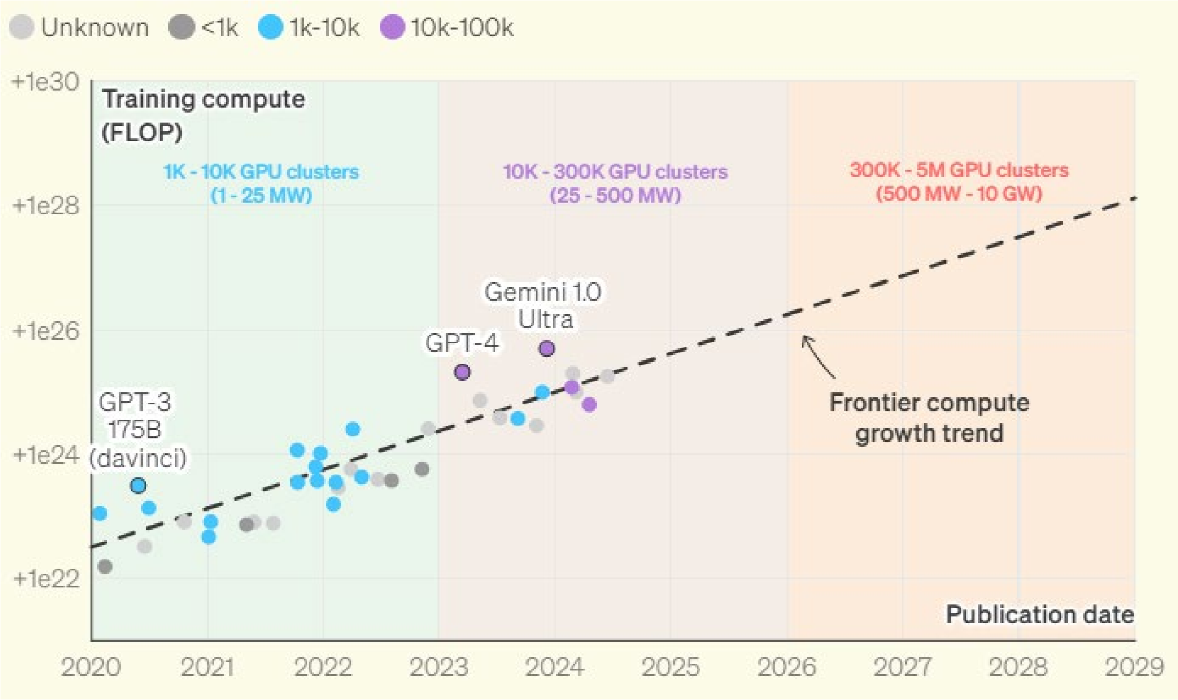
Causally motivated robustness to shortcut learning

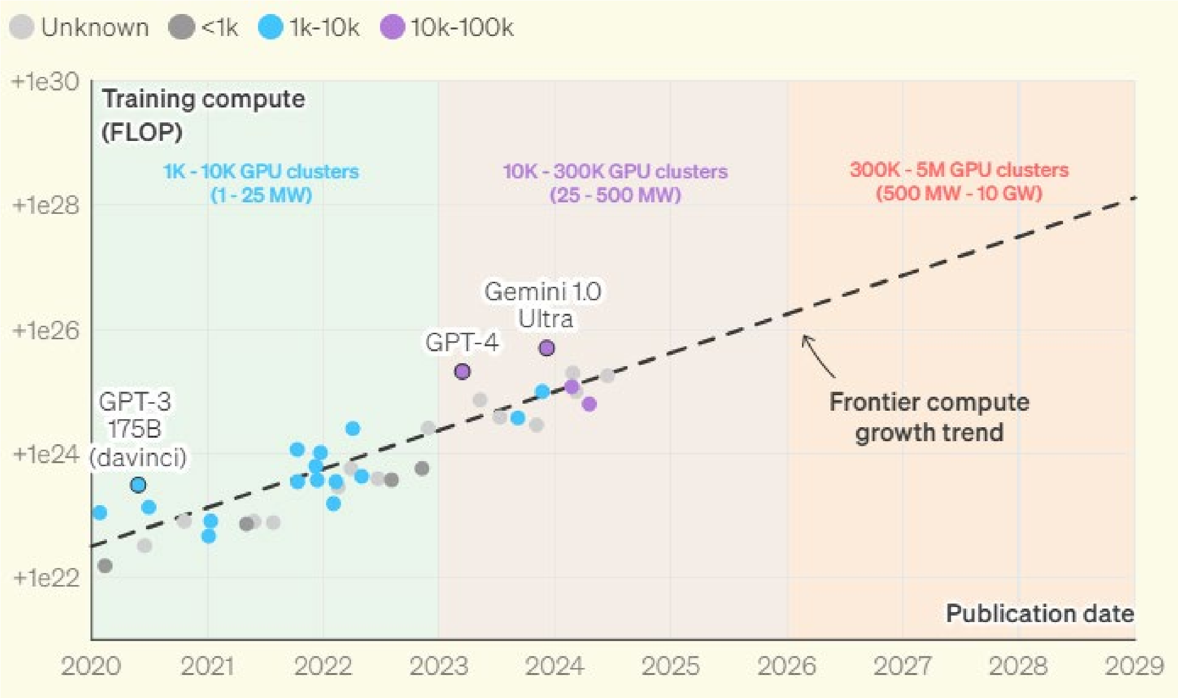
Maggie Makar,
Assistant Professor
CSE, University of Michigan

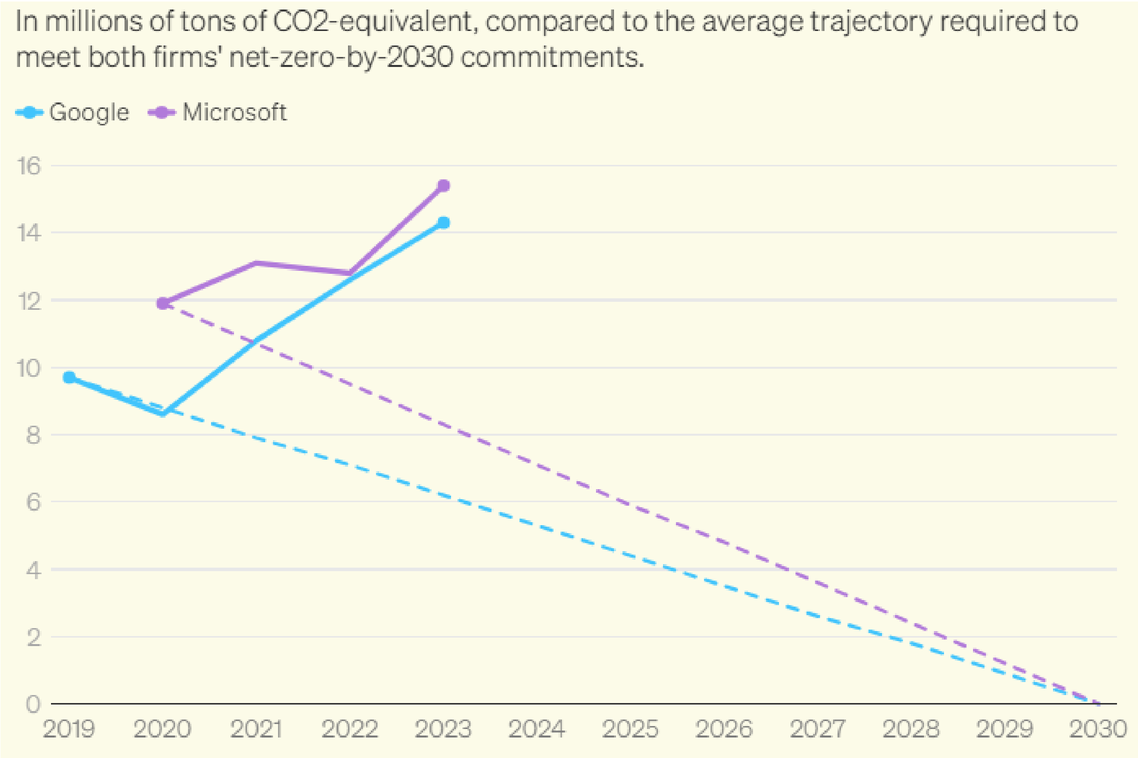
Bigger is better?



Fist and Datta, “How to Build the Future of AI in the United States,” IFP

Bigger is better?





Fist and Datta, “How to Build the Future of AI in the United States,” IFP

Bigger is better?

Task: predict movie review sentiment

Original	
Incredible performances, must watch	Positive
Emotional rollercoaster, highly recommend	Positive
Shakespearean	
A wretched script, a squander of precious hours!	Negative
A dismal affair; I wouldst not commend it to any soul.	Negative

Tang et al, "Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning"

Bigger is better?

Task: predict movie review sentiment

Original	
Incredible performances, must watch	Positive
Emotional rollercoaster, highly recommend	Positive
Shakespearean	
A wretched script, a squander of precious hours!	Negative
A dismal affair; I wouldst not commend it to any soul.	Negative

Exploitable shortcut: Style $\rightarrow$ sentiment

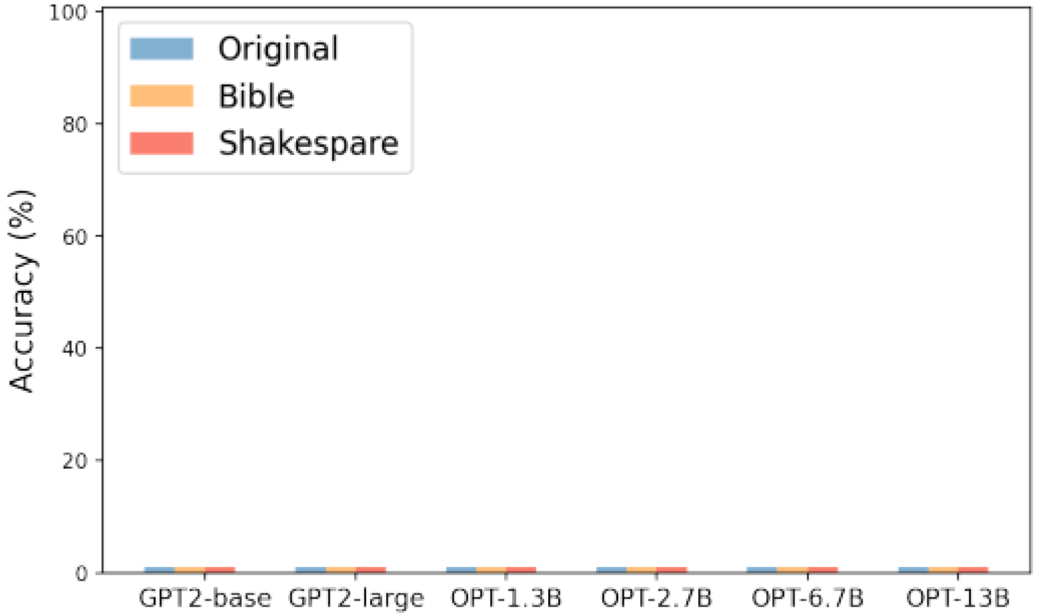
Tang et al, "Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning"

Bigger is better?

Task: predict movie review sentiment

Original	
Incredible performances, must watch	Positive
Emotional rollercoaster, highly recommend	Positive
Shakespearean	
A wretched script, a squander of precious hours!	Negative
A dismal affair; I wouldst not commend it to any soul.	Negative

Exploitable shortcut: Style → sentiment



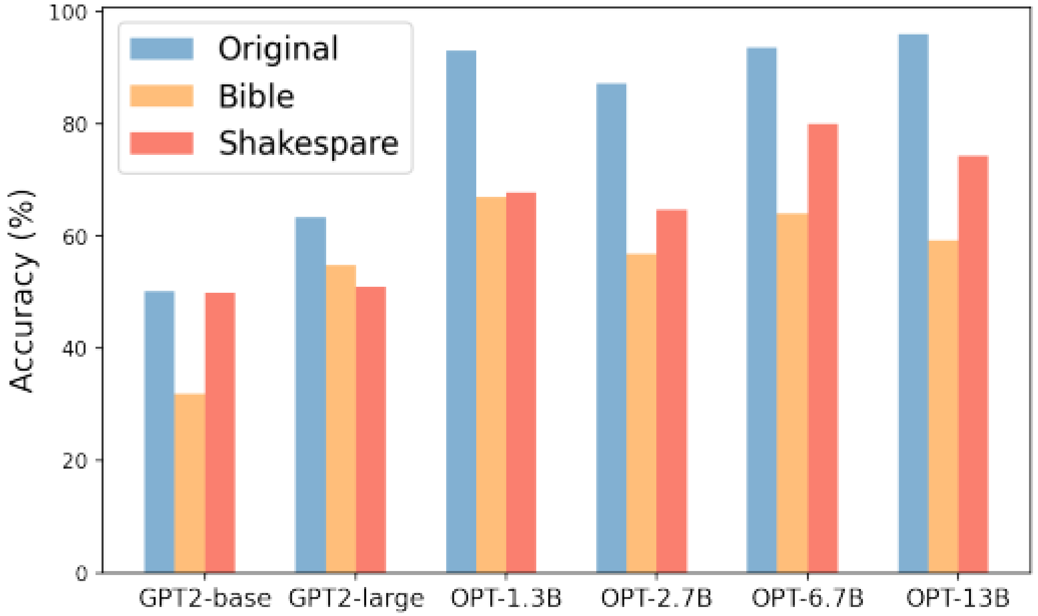
Tang et al, "Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning"

Bigger is better?

Task: predict movie review sentiment

Original	
Incredible performances, must watch	Positive
Emotional rollercoaster, highly recommend	Positive
Shakespearean	
A wretched script, a squander of precious hours!	Negative
A dismal affair; I wouldst not commend it to any soul.	Negative

Exploitable shortcut: Style $\rightarrow$ sentiment



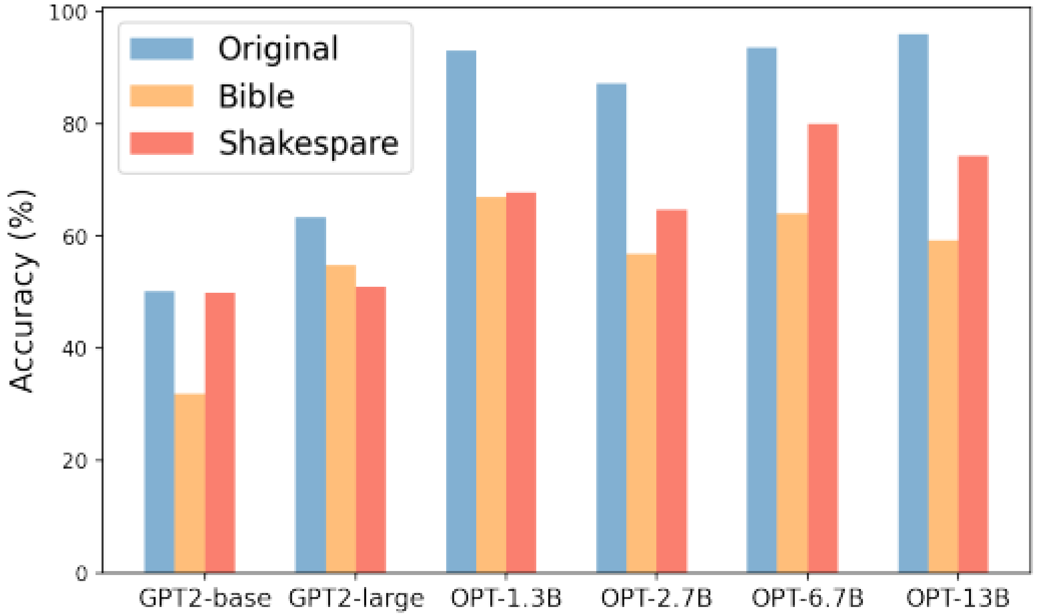
Tang et al, "Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning"

Bigger is better?

Task: predict movie review sentiment

Original	
Incredible performances, must watch	Positive
Emotional rollercoaster, highly recommend	Positive
Shakespearean	
A wretched script, a squander of precious hours!	Negative
A dismal affair; I wouldst not commend it to any soul.	Negative

Exploitable shortcut: Style → sentiment



Building larger models does not give us robustness to shortcuts

Tang et al, "Large Language Models Can be Lazy Learners: Analyze Shortcuts in In-Context Learning"

Causally motivated regularization

$$\textbf{Causally motivated } \overbrace{\textbf{regularization}}^{\text{Good old efficiency}}$$

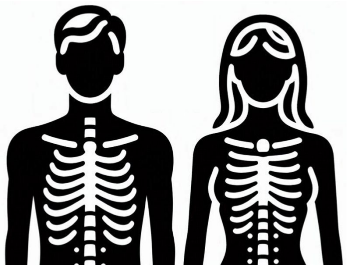
$$\underbrace{\textbf{Causally}}_{\text{Robustness}}\ \textbf{motivated}\ \overbrace{\textbf{regularization}}^{\text{Good old efficiency}}$$

$$\underbrace{\textbf{Causally}}_{\text{Robustness}}\ \textbf{motivated}\ \overbrace{\textbf{regularization}}^{\text{Good old efficiency}}$$

Causality reasons about the world under shifts/interventions

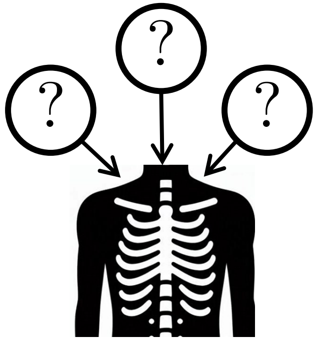
Talk outline

Talk outline



1 Efficiency + robustness to known sampling bias

MPMBHD, AIStats 22
MD, TMLR 23
N**M**, UAI 24



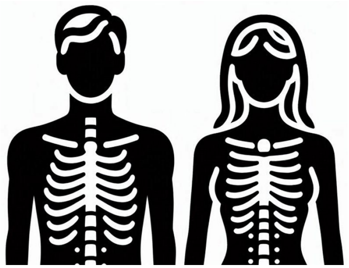
2 Efficiency + robustness to unknown sampling biases

Z**M**, NeurIPS 22
WJ**M**SW, NeurIPS 22

3 Evaluating localized circuits in LLMs

SVNZGJ**M**B – NeurIPS 24

Talk outline



1 Efficiency + robustness to known sampling bias

MPMBHD, AIStats 22
MD, TMLR 23
N**M**, UAI 24

2 Efficiency + robustness to unknown sampling biases

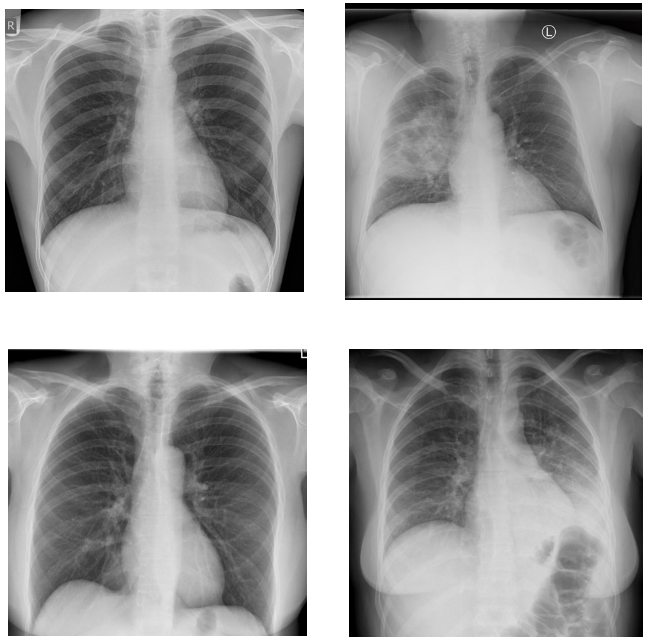
Z**M**, NeurIPS 22
WJ**M**SW, NeurIPS 22

3 Evaluating localized circuits in LLMs

SVNZGJ**M**B – NeurIPS 24

Pneumonia detection

Training data



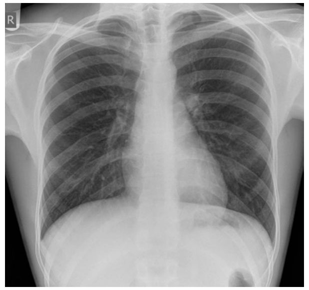
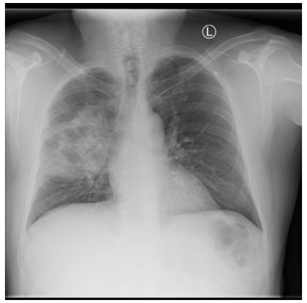
Makar, Packer, Moldovan, Blalock, Halpern and D'Amour AIStats 2022

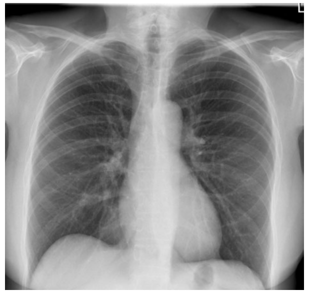
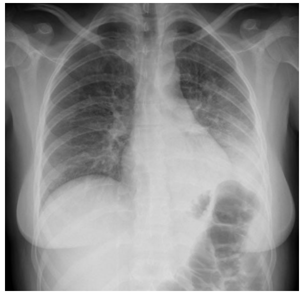
Cases courtesy of Dr. Andrew Dixon, Dr Henry Knip, Dr. Usman Bashir, and Dr. Ian Bickle , Radiopaedia.org, rID: 48366, 31388, 18394 and rID: 50318

Pneumonia detection

Training data

Target label:	Healthy	Pneumonia
	$Y = 0$	$Y = 1$

Makar, Packer, Moldovan, Blalock, Halpern and D'Amour AIStats 2022

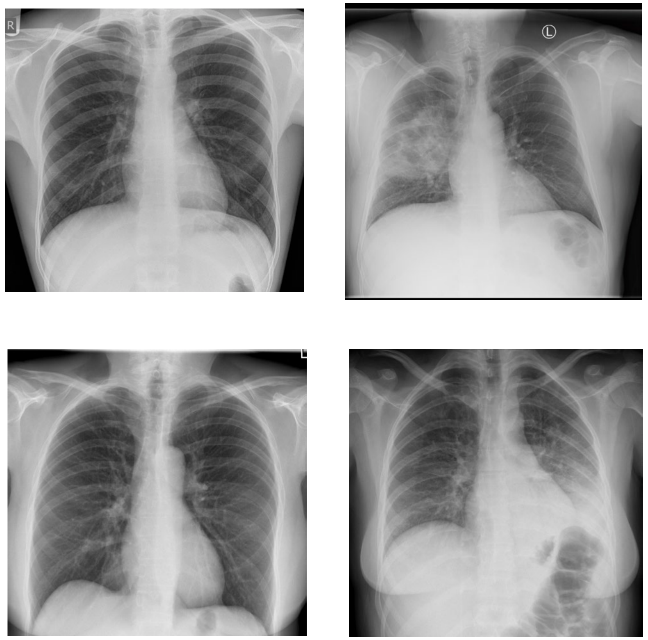
Cases courtesy of Dr. Andrew Dixon, Dr Henry Knip, Dr. Usman Bashir, and Dr. Ian Bickle , Radiopaedia.org, rID: 48366, 31388, 18394 and rID: 50318

Pneumonia detection

Training data

Target label: Healthy $Y = 0$ | Pneumonia $Y = 1$

Auxiliary label: Men $V = 0$ | Women $V = 1$

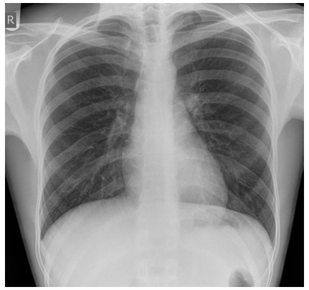
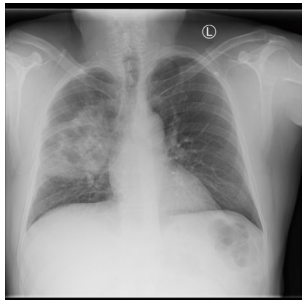
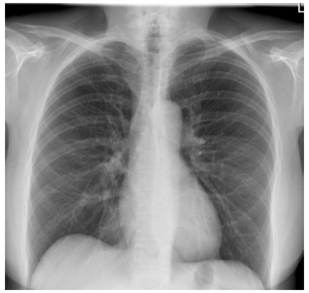


Makar, Packer, Moldovan, Blalock, Halpern and D'Amour AIStats 2022

Cases courtesy of Dr. Andrew Dixon, Dr Henry Knip, Dr. Usman Bashir, and Dr. Ian Bickle , Radiopaedia.org, rID: 48366, 31388, 18394 and rID: 50318

Pneumonia detection under biased sampling

Training data

		Healthy	Pneumonia
Target label:		$Y = 0$	$Y = 1$
Auxiliary label:	Men $V = 0$		
	Women $V = 1$		

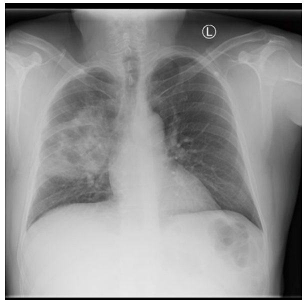
Pneumonia detection under biased sampling

Training data

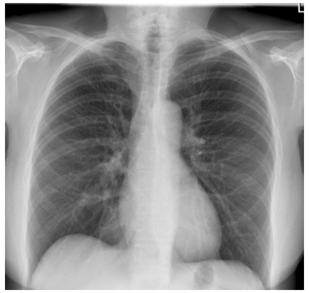
Target label: Healthy $Y = 0$ | Pneumonia $Y = 1$

Auxiliary label:

Men $V = 0$



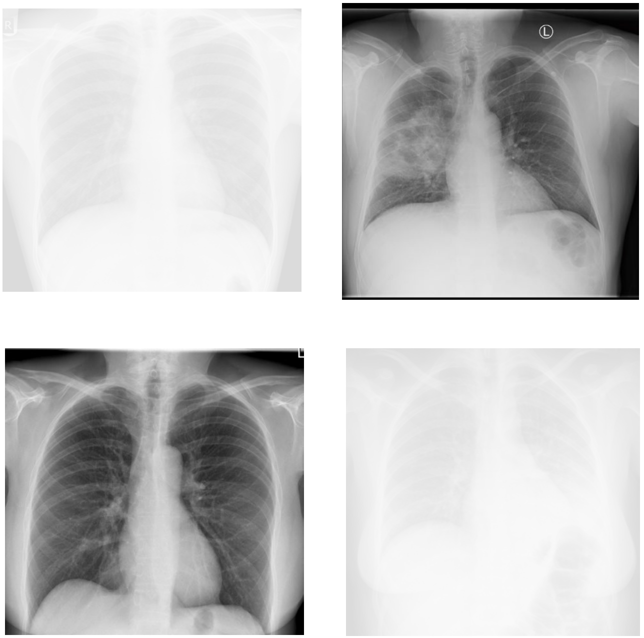
Women $V = 1$



Pneumonia detection under biased sampling

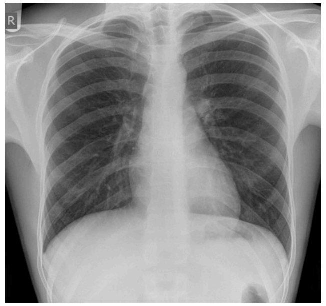
Training data

		Healthy	Pneumonia
Target label:		$Y = 0$	$Y = 1$
Auxiliary label:	Men $V = 0$		
	Women $V = 1$		



Testing data

Man with pneumonia



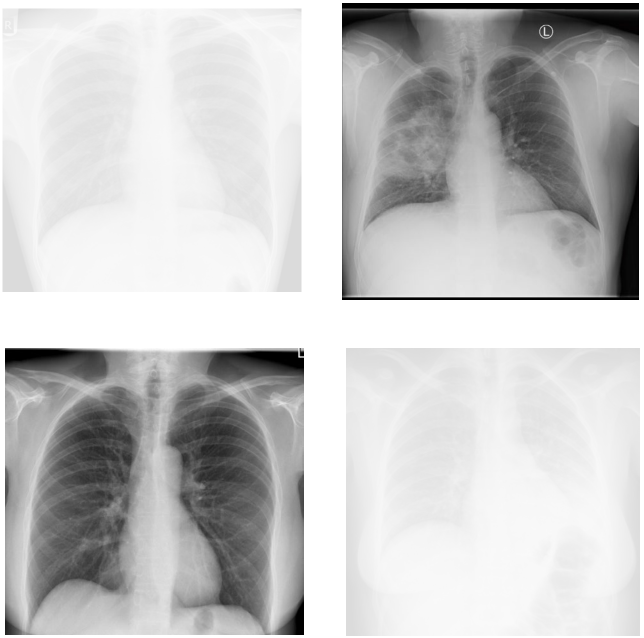
$P(\text{Pneumonia}) = 0.99$

Pneumonia detection under biased sampling

Training data

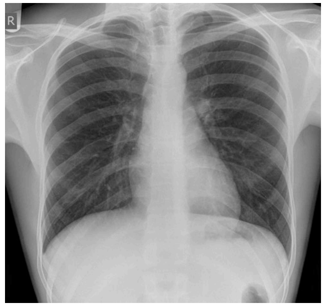
Target label: Healthy $Y = 0$ | Pneumonia $Y = 1$

Auxiliary label: Men $V = 0$ | Women $V = 1$

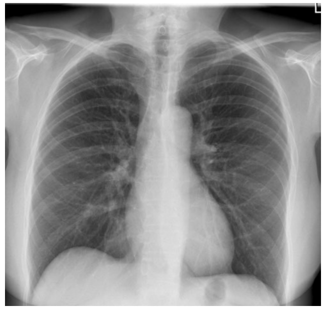


Testing data

Man with pneumonia | Woman with pneumonia



$P(\text{Pneumonia}) = 0.99$



$P(\text{Pneumonia}) = 0.01$

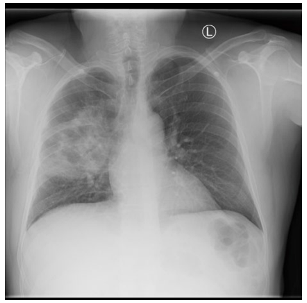
Pneumonia detection under biased sampling

Training data

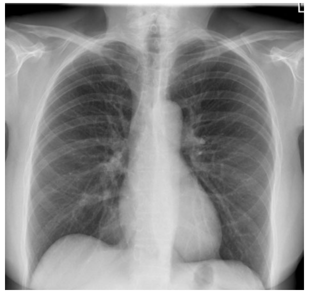
Target label: Healthy $Y = 0$ | Pneumonia $Y = 1$

Auxiliary label:

Men $V = 0$

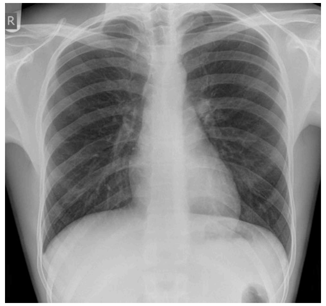


Women $V = 1$



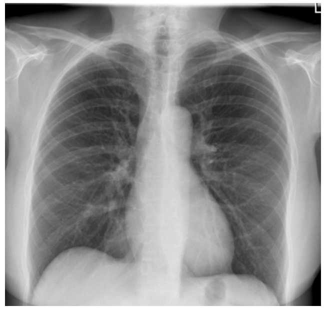
Testing data

Man with pneumonia



$P(\text{Pneumonia}) = 0.99$

Woman with pneumonia



$P(\text{Pneumonia}) = 0.01$

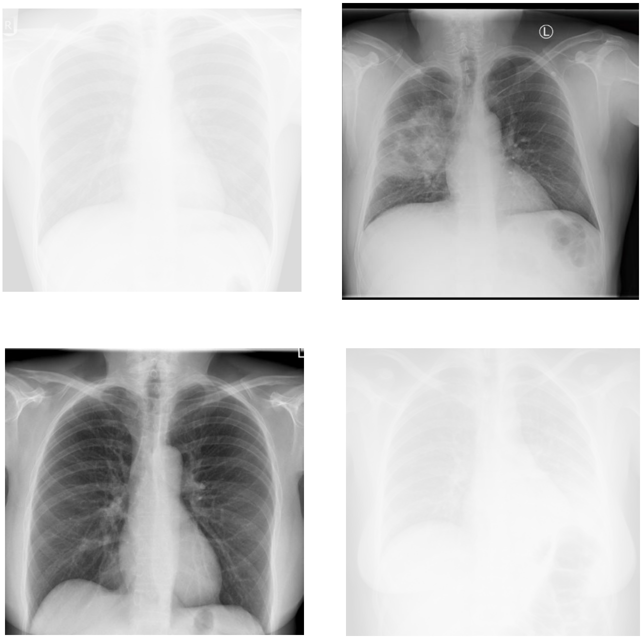
Shortcut learning

Pneumonia detection under biased sampling

Training data

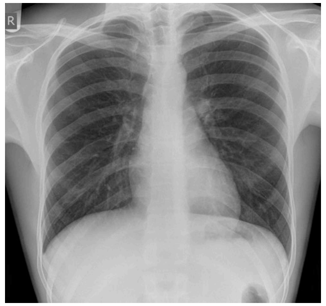
Target label: Healthy $Y = 0$ Pneumonia $Y = 1$

Auxiliary label: Men $V = 0$ Women $V = 1$



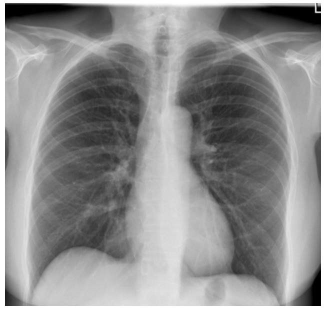
Testing data

Man with pneumonia



$P(\text{Pneumonia}) = 0.99$

Woman with pneumonia

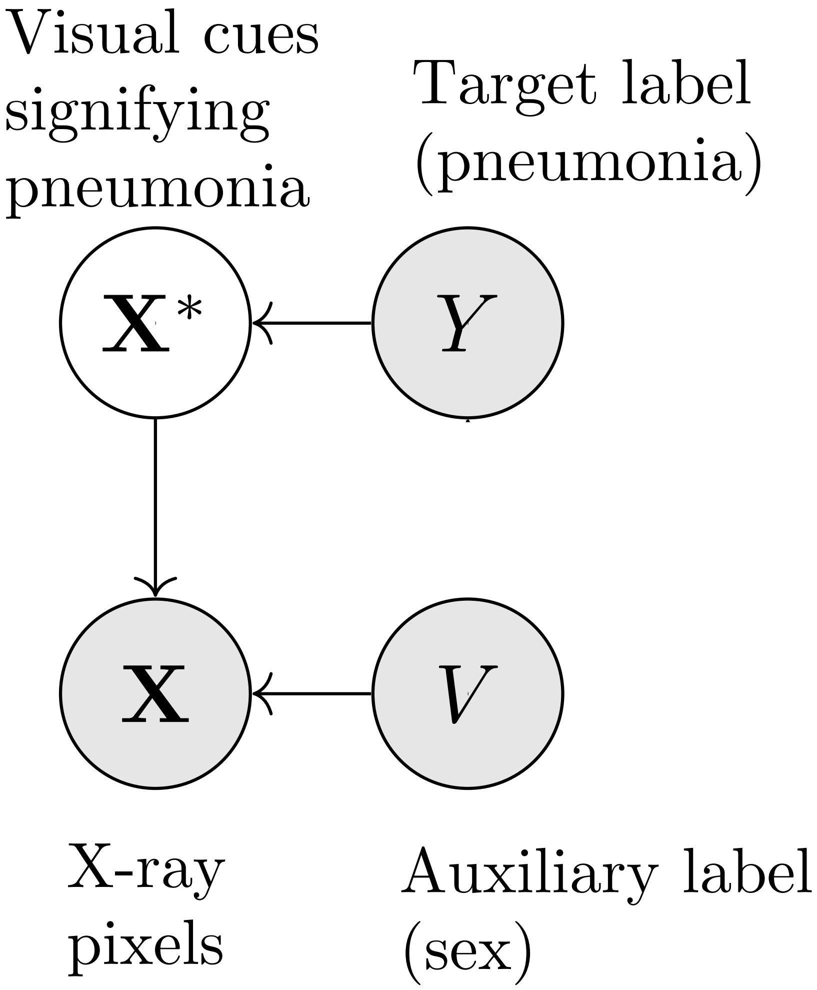


$P(\text{Pneumonia}) = 0.01$

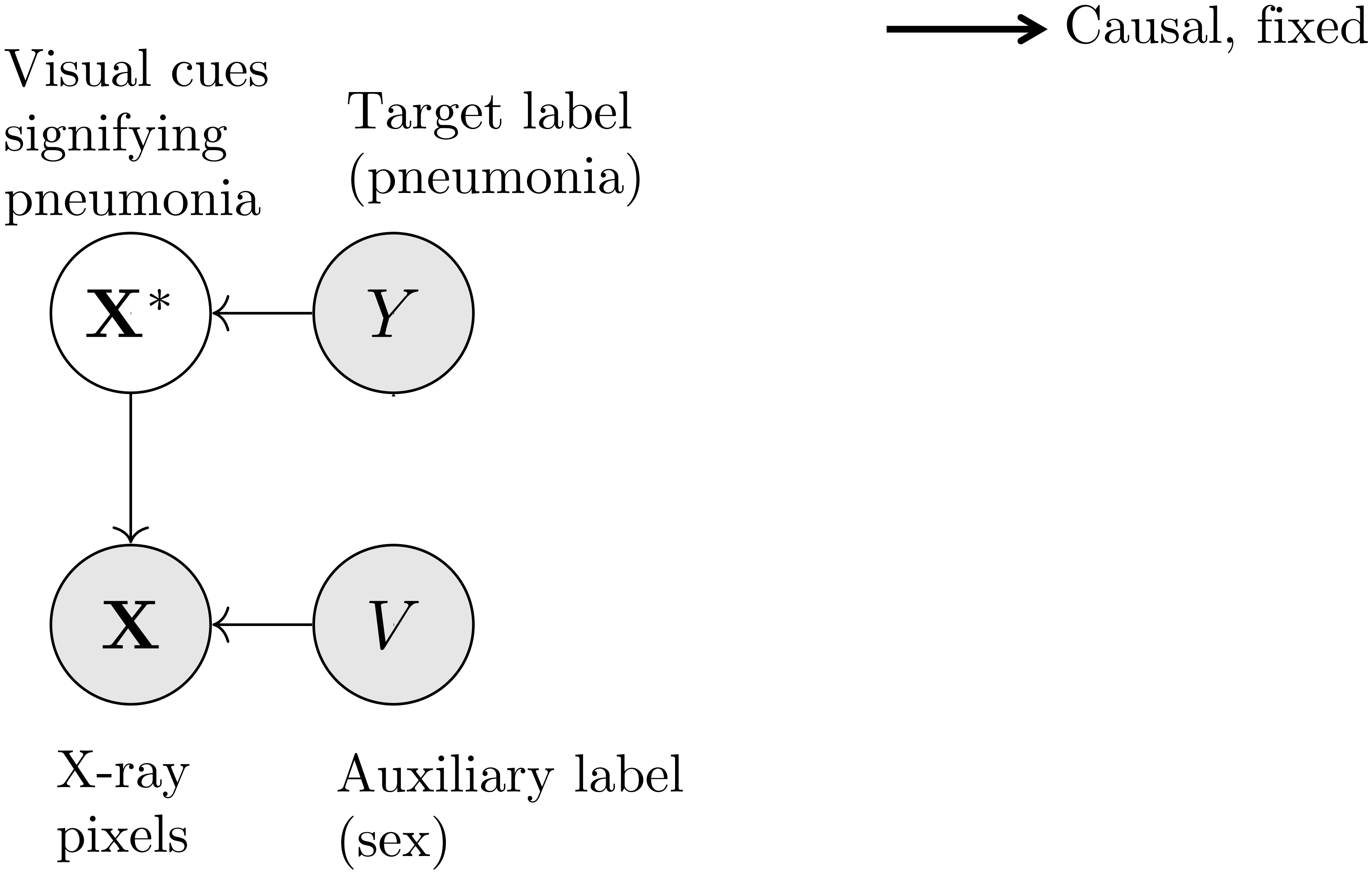
Shortcut learning: relying on spurious correlations to predict the target label

Can we build models that are robust to shortcuts?

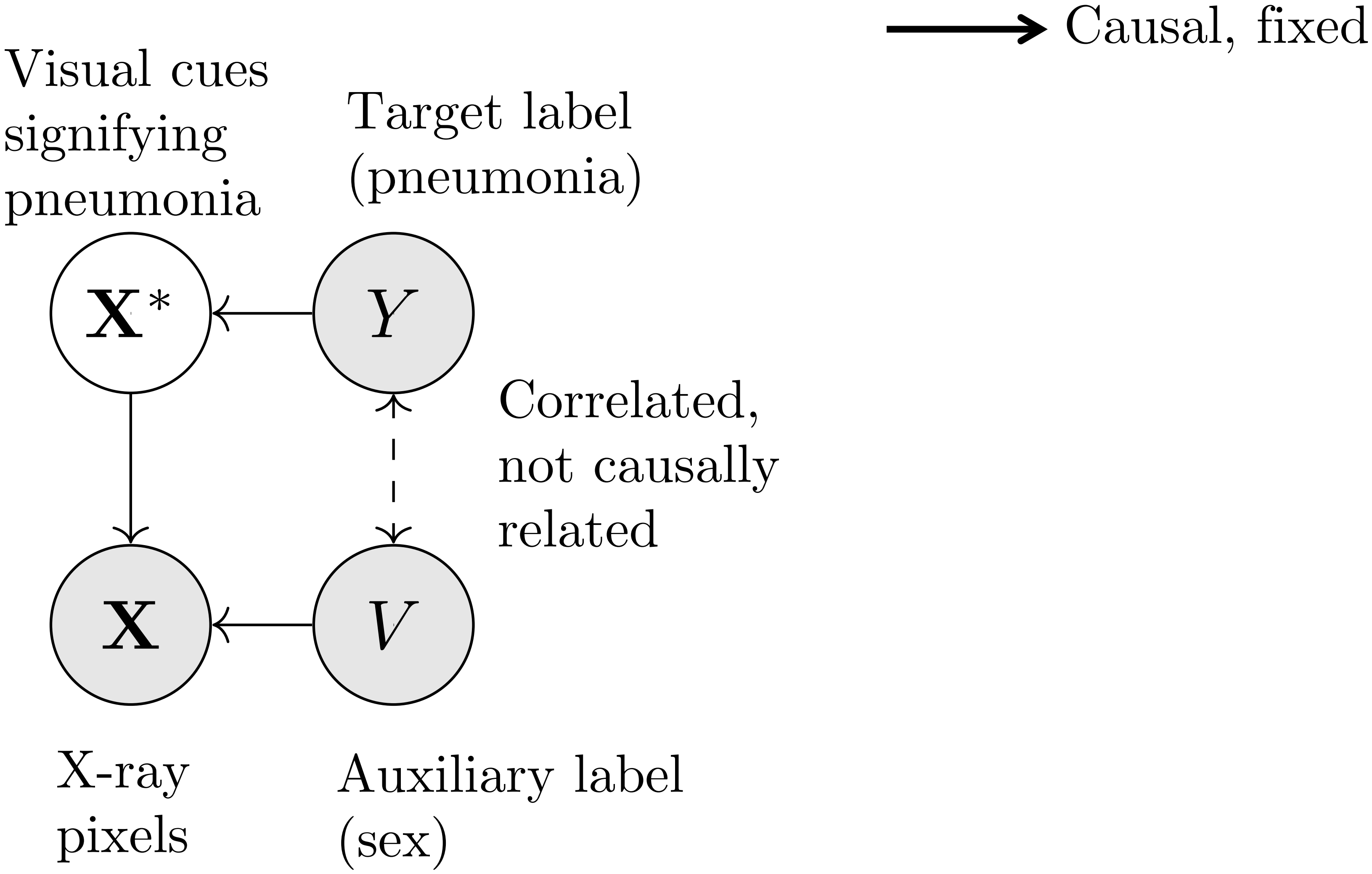
Causal assumptions



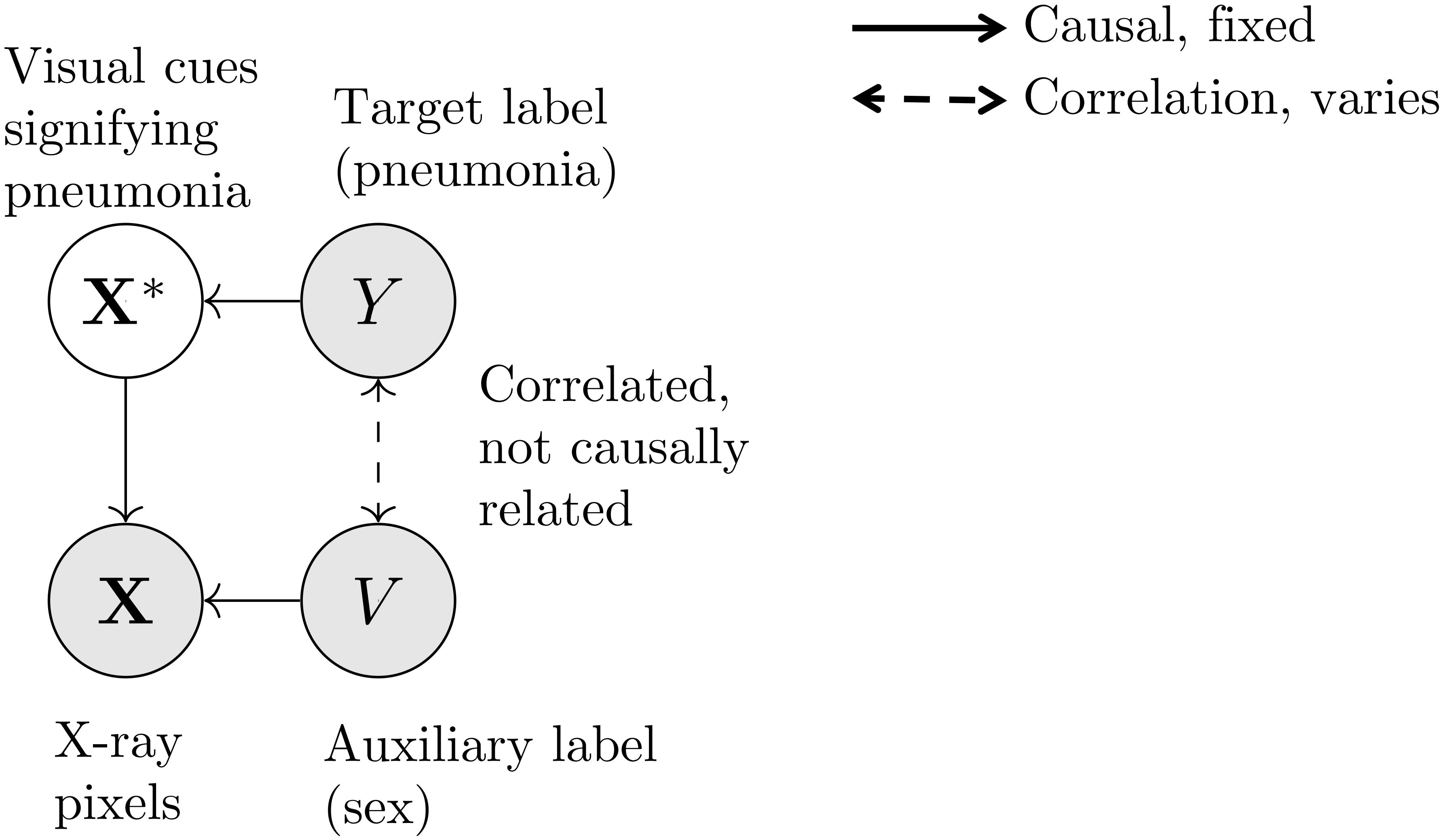
Causal assumptions



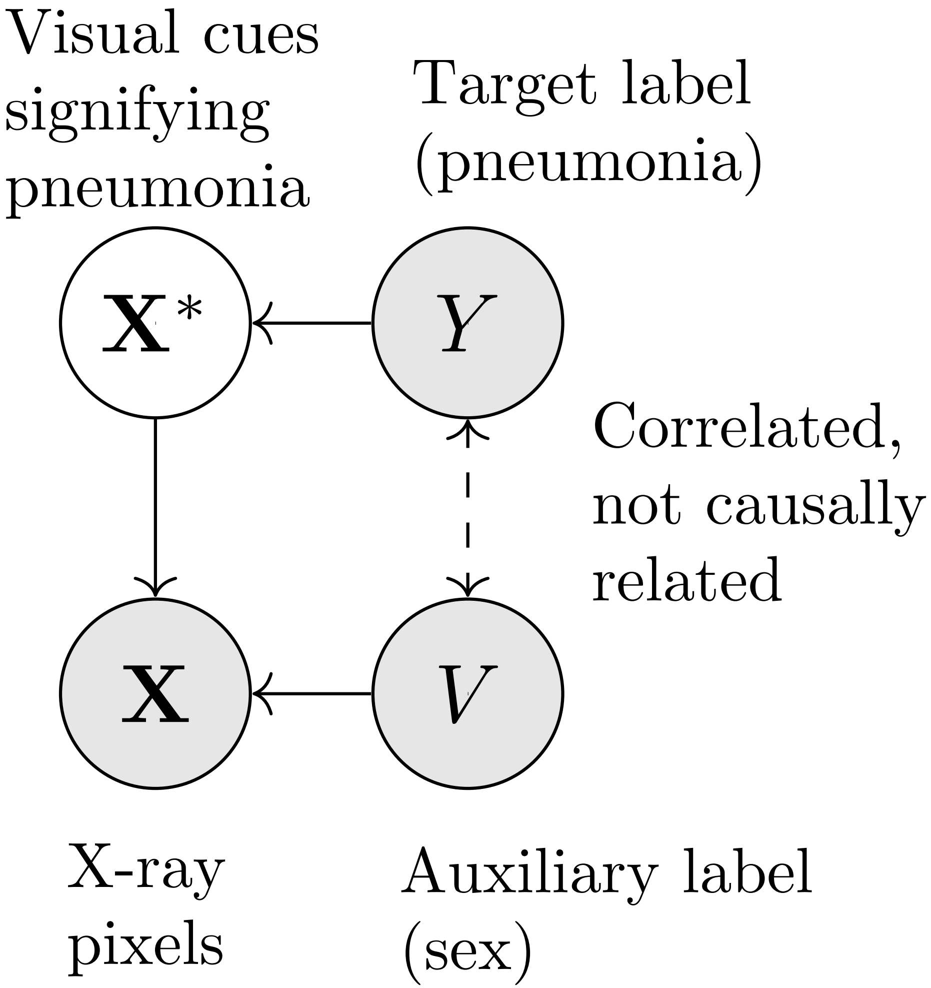
Causal assumptions



Causal assumptions



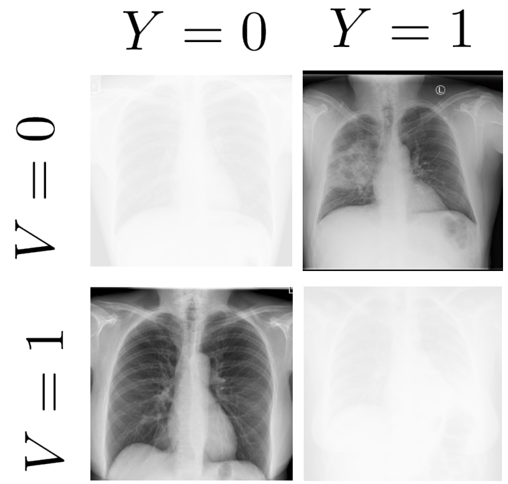
Causal assumptions



→ Causal, fixed

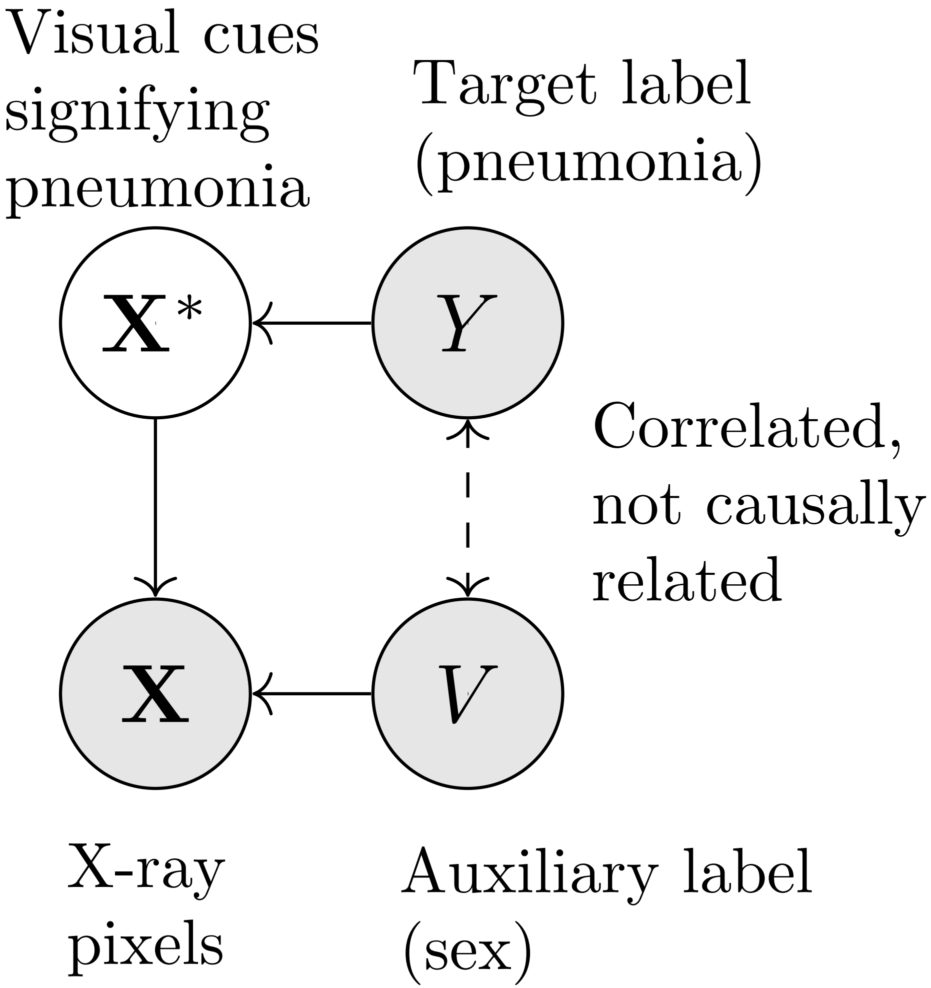
⇠ ⇢ Correlation, varies

Hospital A



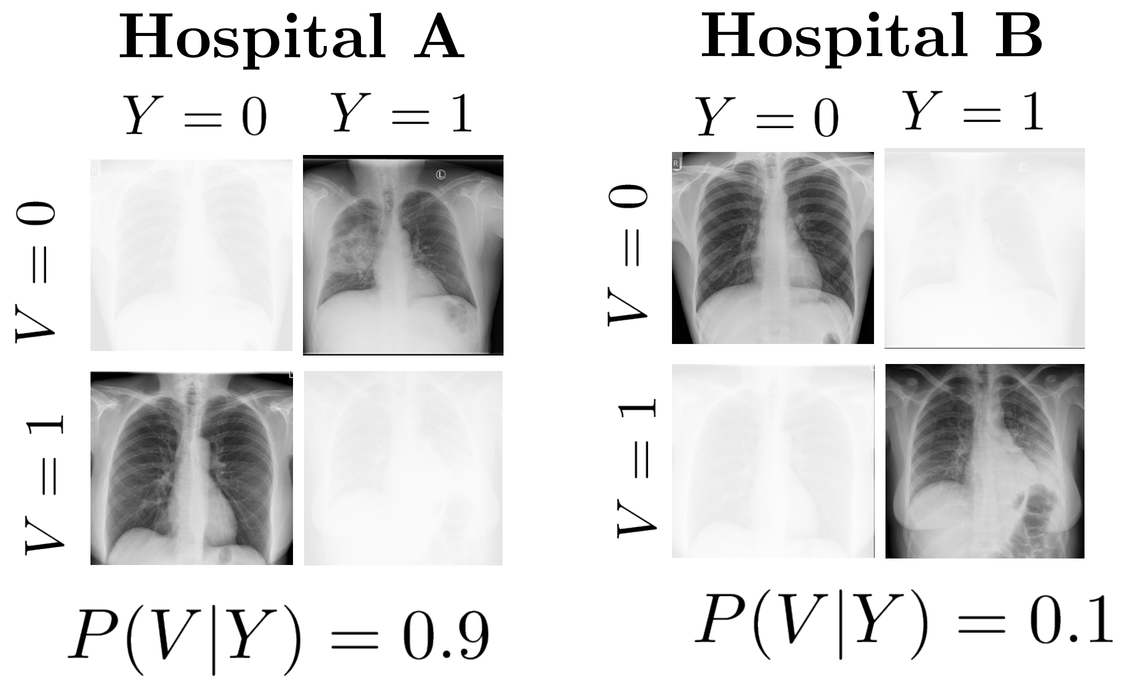
$$P(V|Y) = 0.9$$

Causal assumptions

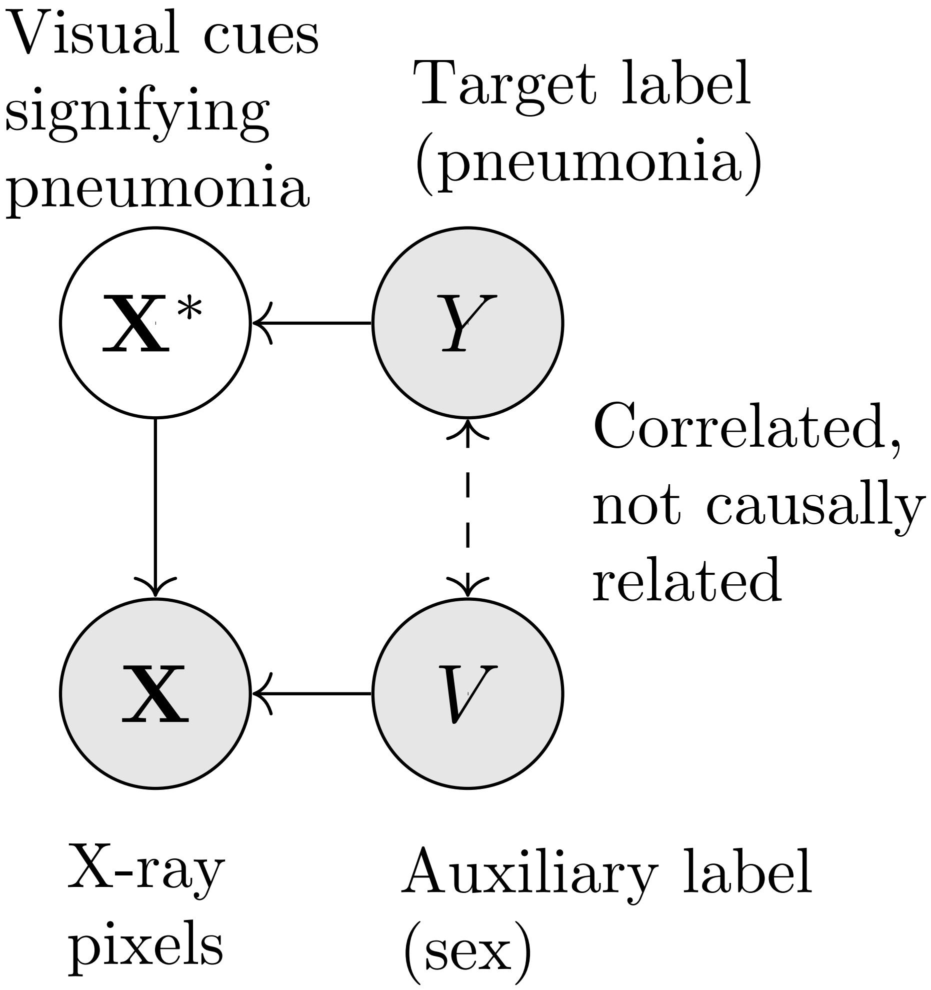


→ Causal, fixed

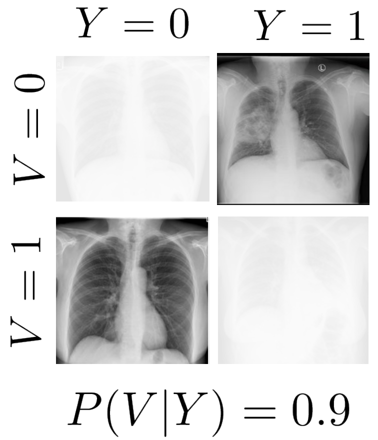
⇠ ⇢ Correlation, varies



Task formulation

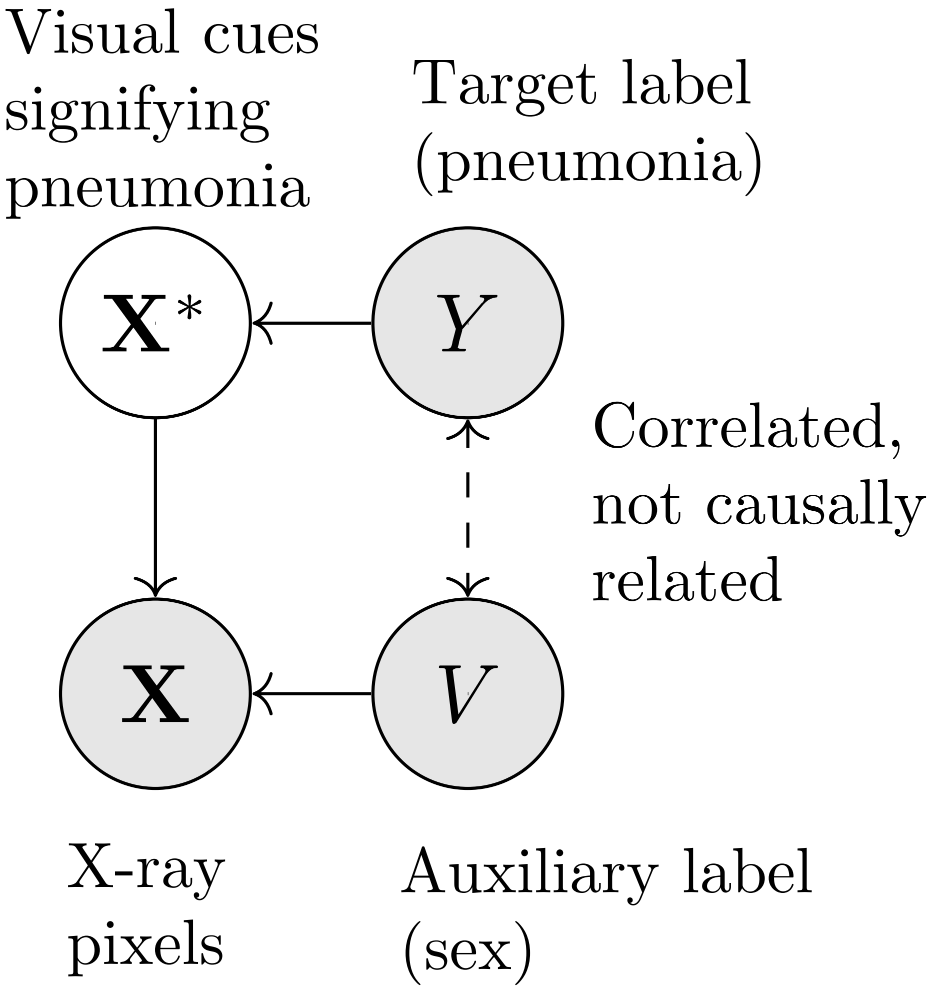


Training data

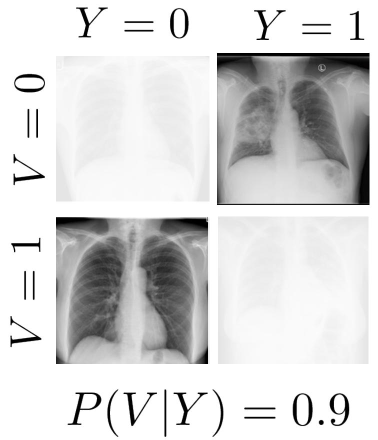


Have: Data with some correlation between V, Y

Task formulation



Training data

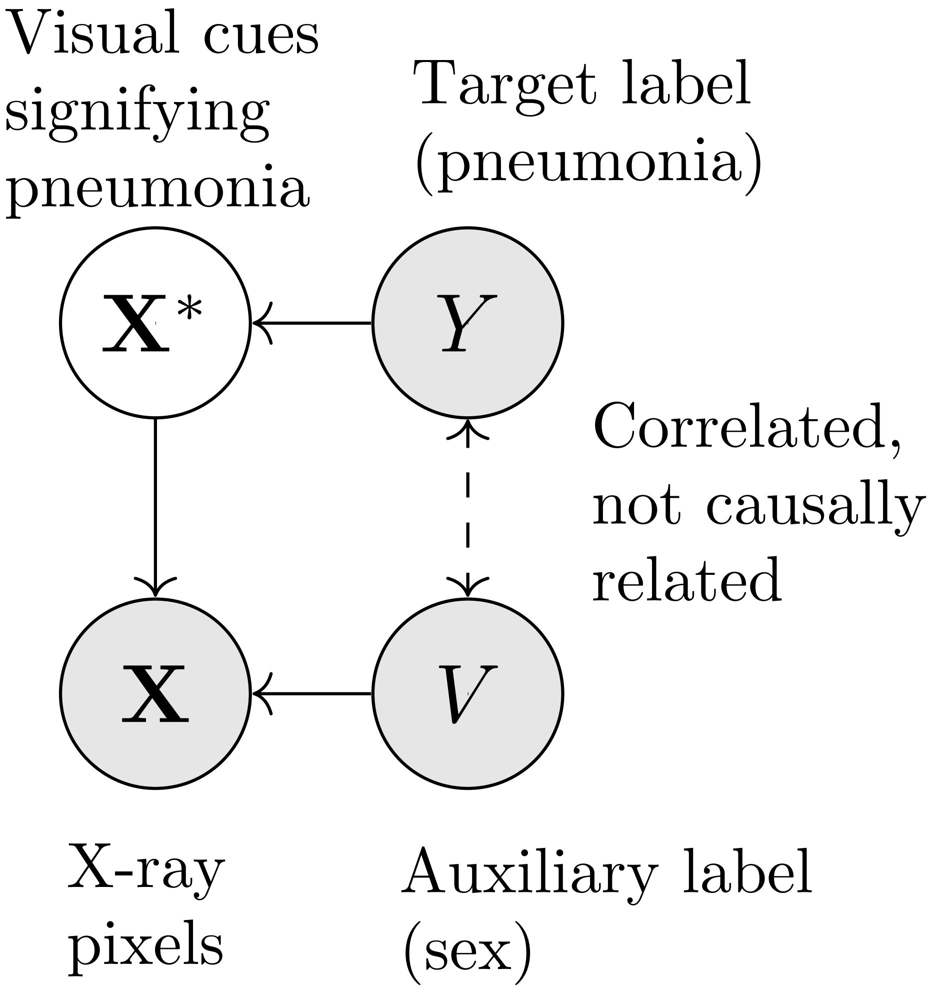


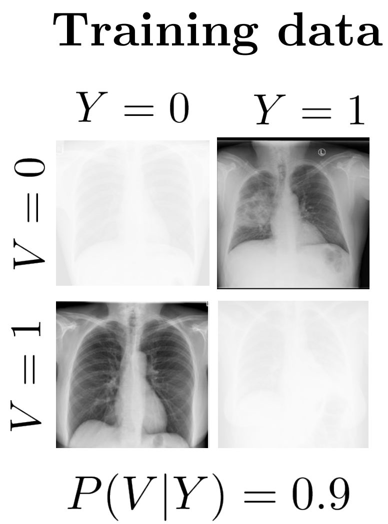
Have: Data with some correlation between V, Y

Want: $f(\boldsymbol{X})$

Such that: f is accurate

Task formulation



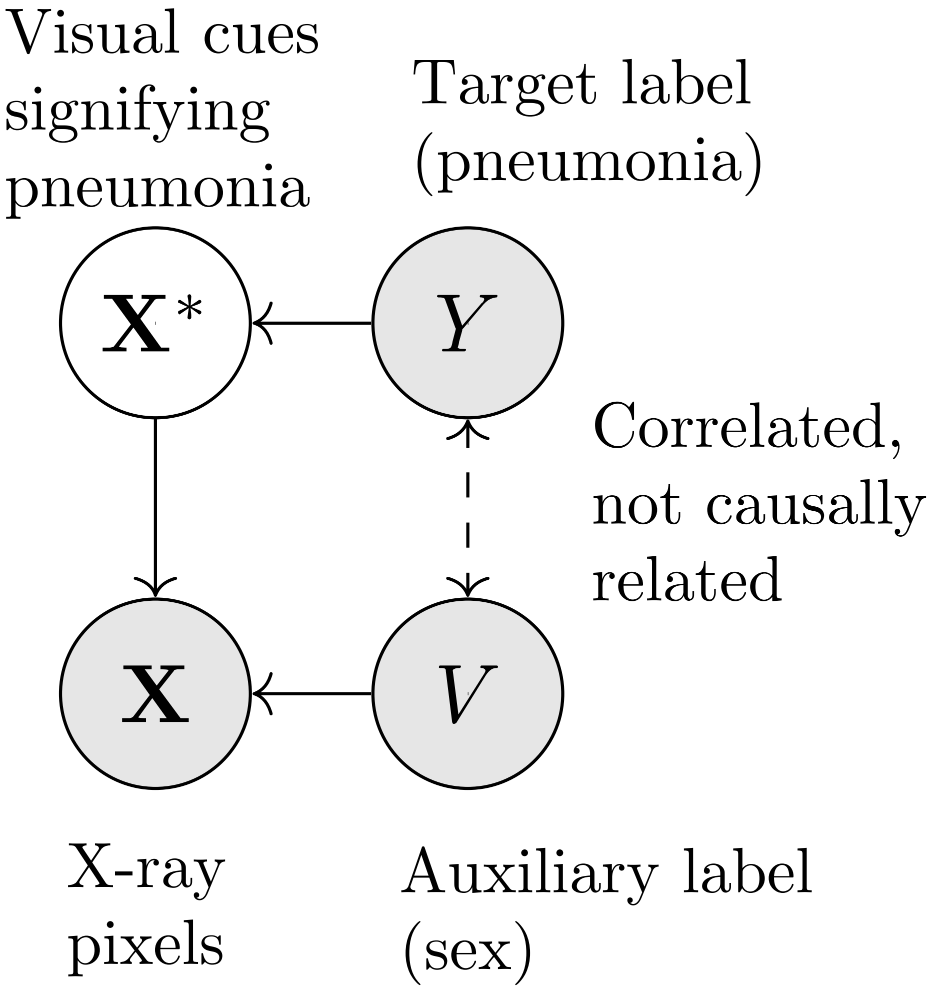


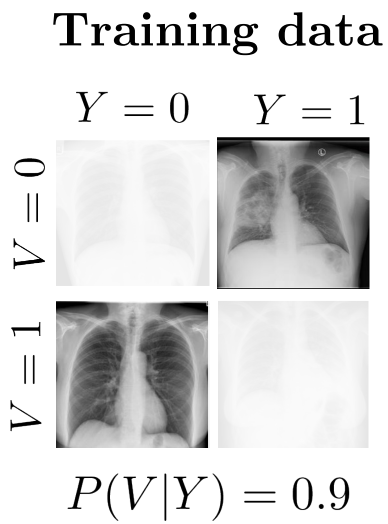
Have: Data with some correlation between V, Y

Want: $f(\boldsymbol{X})$

Such that: f is accurate

Task formulation



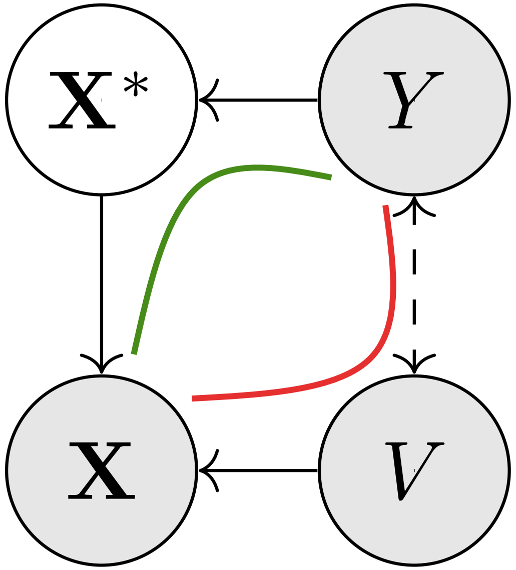


Have: Data with some correlation between V, Y

Want: $f(\boldsymbol{X})$

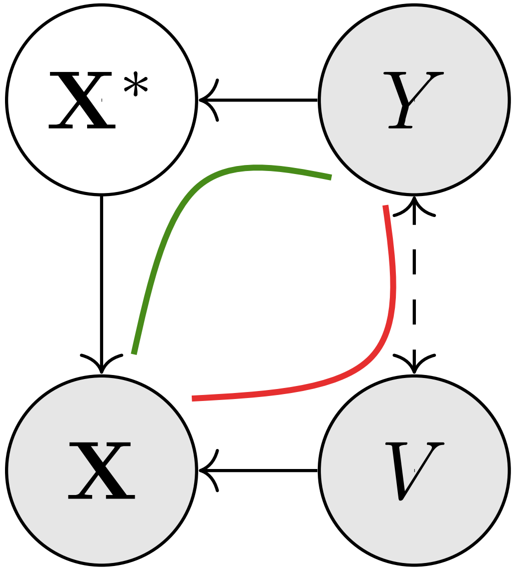
Such that: f is accurate and robust to the shortcut, i.e., has the same performance across all V, Y correlations

Root cause of shortcut learning



Root cause of shortcut learning

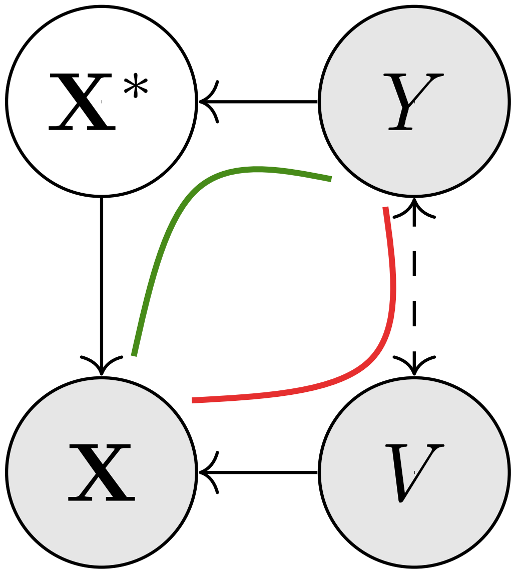
Causal (green) path: robust



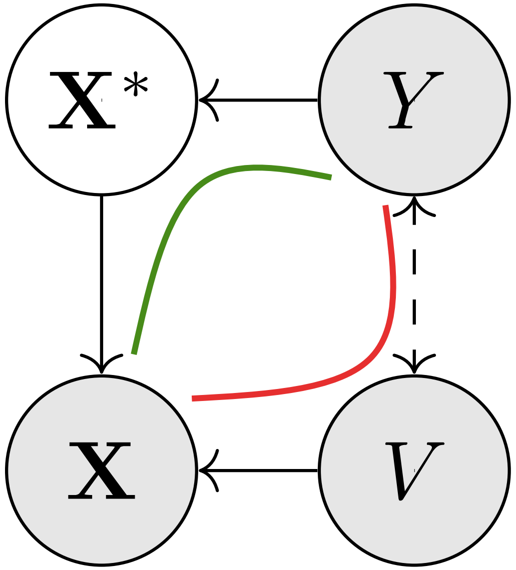
Root cause of shortcut learning

Causal (green) path: robust

Non-causal (red) path: encodes shortcut

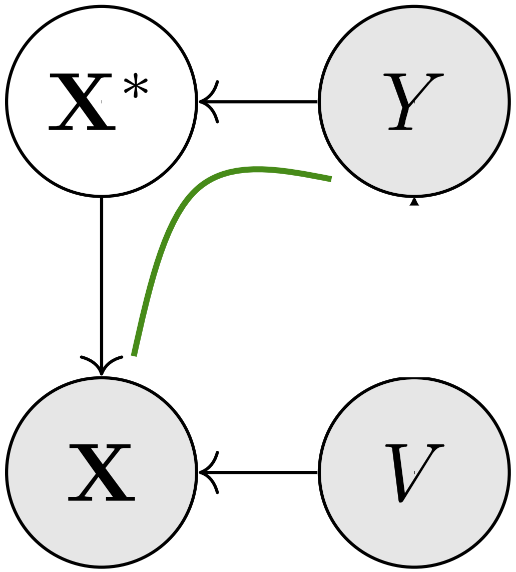


Population-level robustness to the shortcut



Population-level robustness to the shortcut

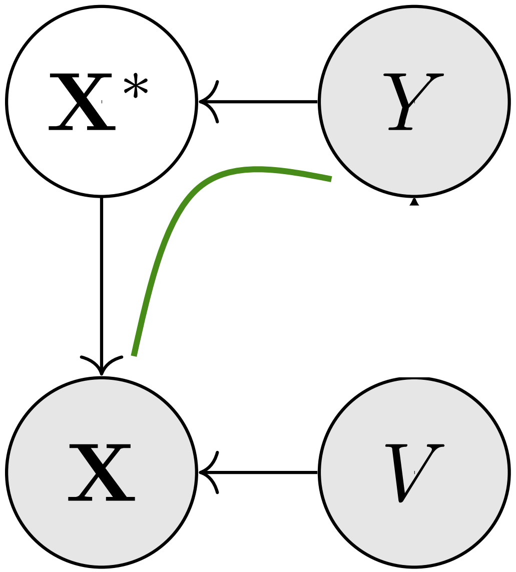
Ideal distribution → no correlation between V, Y



Population-level robustness to the shortcut

Ideal distribution → no correlation between V, Y

Proposition (informal): Under the ideal distribution, with a very large dataset, the optimal model is robust to shortcuts.



Finite sample analysis (under the ideal distribution)

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than “the usual models” in finite samples.

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Usual models

$$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \|f\|_2$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Usual models	**Models conforming to the DAG**
$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \|f\|_2$	

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Usual models

$$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \|f\|_2$$

Models conforming to the DAG

= Models that do not encode correlations between Y, V

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Usual models

$$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \|f\|_2$$

Models conforming to the DAG

= Models that do not encode correlations between Y, V

$$\underbrace{P(f(\mathbf{x})|V=1)}_{\text{Predictions for women}} = \underbrace{P(f(\mathbf{x})|V=0)}_{\text{Predictions for men}}$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Usual models

$$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \|f\|_2$$

Models conforming to the DAG

= Models that do not encode correlations between Y, V

$$\underbrace{P(f(\mathbf{x})|V=1)}_{\text{Predictions for women}} = \underbrace{P(f(\mathbf{x})|V=0)}_{\text{Predictions for men}}$$

In practice:

$$f^* = \min \ell(f(\mathbf{x}), y) + \alpha \cdot \mathrm{MMD}\big(P(f(\mathbf{x})|V=1), P(f(\mathbf{x})|V=0)\big)$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

$$\mathbf{x} = [x_v, x_y]$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

$$\mathbf{x} = [\underbrace{x_v}_{\text{"Bad" shortcut}}, \underbrace{x_y}_{\text{Good variable}}]$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

$$\mathbf{x} = [\underbrace{x_v}_{\text{"Bad" shortcut}}, \underbrace{x_y}_{\text{Good variable}}]$$

Don't know: which is x_v vs. x_y

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

$$\mathbf{x} = [x_v, x_y]$$

"Bad" shortcut Good variable

Don't know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

Model learning = search problem

$$\mathbf{x} = [x_v, x_y]$$

"Bad" shortcut   Good variable

Don't know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than “the usual models” in finite samples.

Simple example:

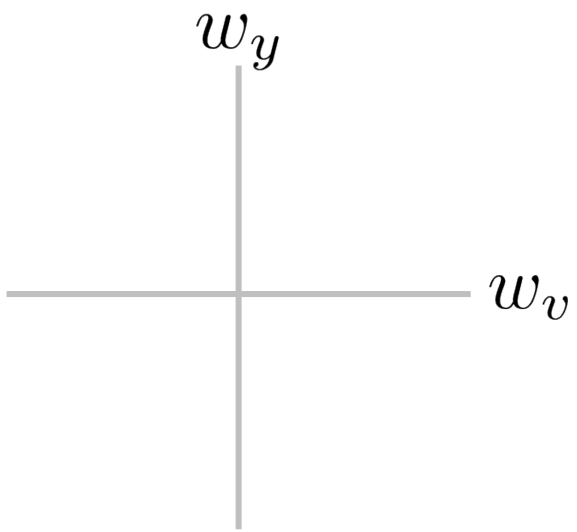
Model learning = search problem

$$\mathbf{x} = [x_v, x_y]$$

“Bad” shortcut Good variable

Don’t know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

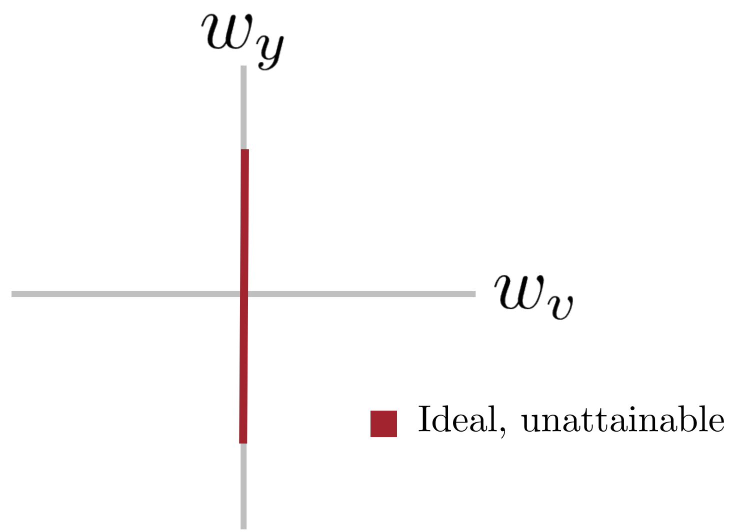


Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

Model learning = search problem



$$\mathbf{x} = [x_v, x_y]$$

"Bad" shortcut   Good variable

Don't know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

If $w_v \neq 0$, model encodes info about shortcut

Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than “the usual models” in finite samples.

Simple example:

Model learning = search problem

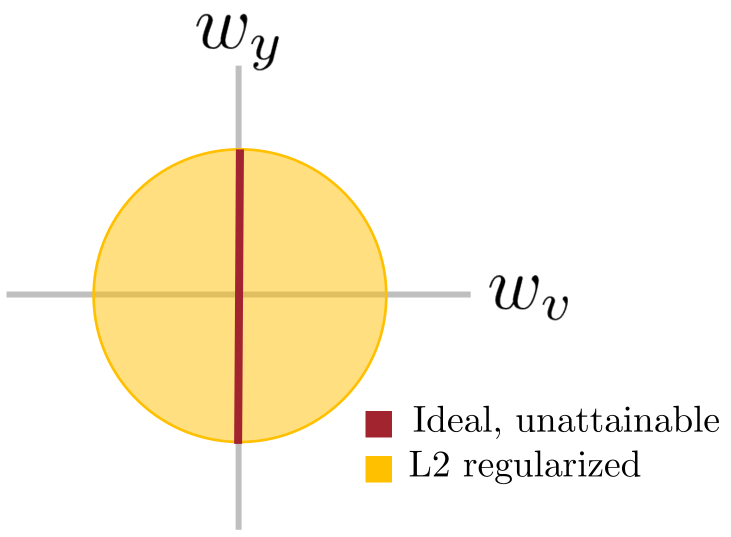
$$\mathbf{x} = [x_v, x_y]$$

“Bad” shortcut Good variable

Don’t know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

If $w_v \neq 0$, model encodes info about shortcut



Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than “the usual models” in finite samples.

Simple example:

Model learning = search problem

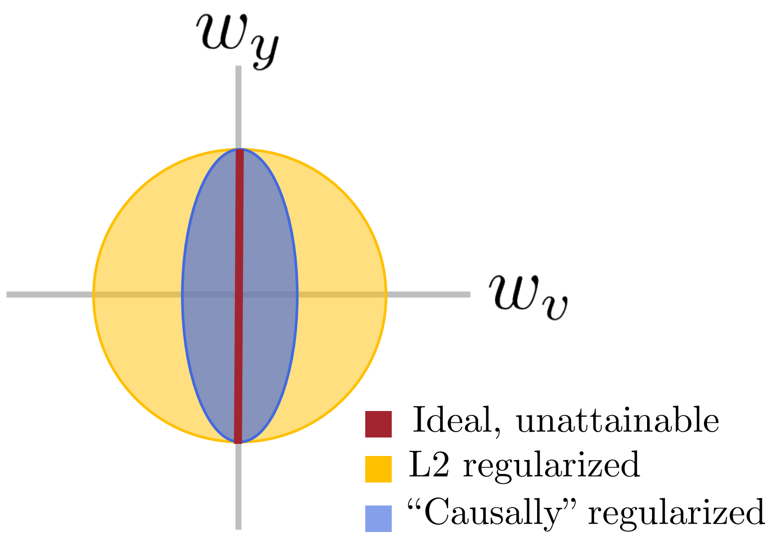
$$\mathbf{x} = [x_v, x_y]$$

“Bad” shortcut Good variable

Don’t know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

If $w_v \neq 0$, model encodes info about shortcut



Finite sample analysis (under the ideal distribution)

Proposition (informal): Models that conform to the causal DAG are more efficient than "the usual models" in finite samples.

Simple example:

Model learning = search problem

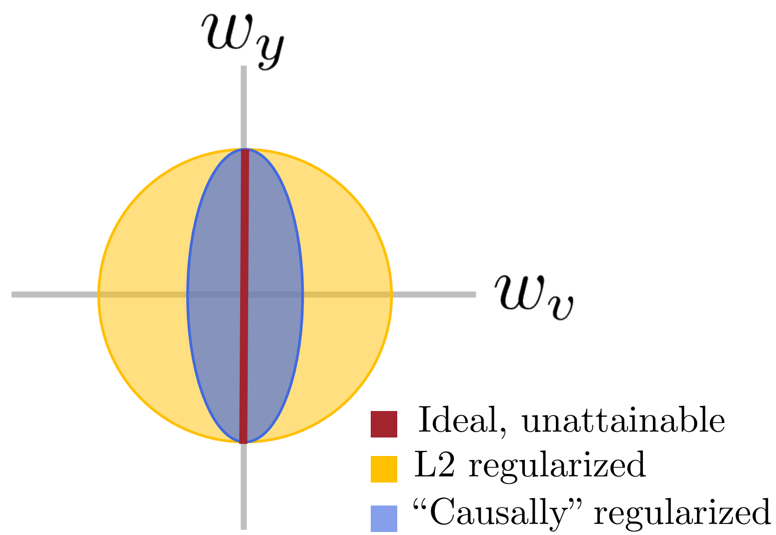
$$\mathbf{x} = [x_v, x_y]$$

"Bad" shortcut   Good variable

Don't know: which is x_v vs. x_y

$$f(\mathbf{x}) = w_y x_y + w_v x_v$$

If $w_v \neq 0$, model encodes info about shortcut



→ Smaller Rademacher complexity

Quick recap

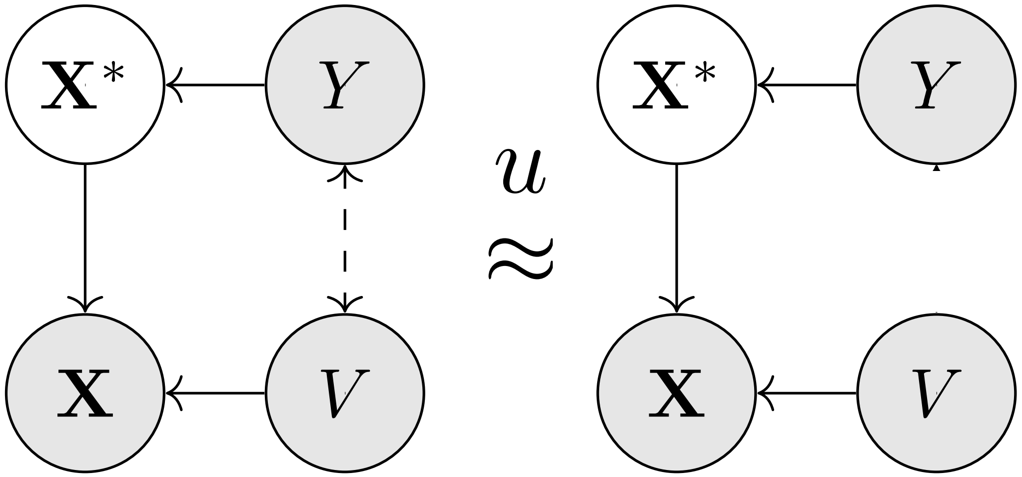
Ideal distribution + large sample	Robustness to the shortcut
Ideal distribution	Statistical efficiency
An arbitrary distribution	?

Quick recap

Ideal distribution + large sample	Robustness to the shortcut
Ideal distribution	Statistical efficiency
An arbitrary distribution	Shortcut learning + bias!

Sampling from non-ideal distributions

Sampling from non-ideal distributions

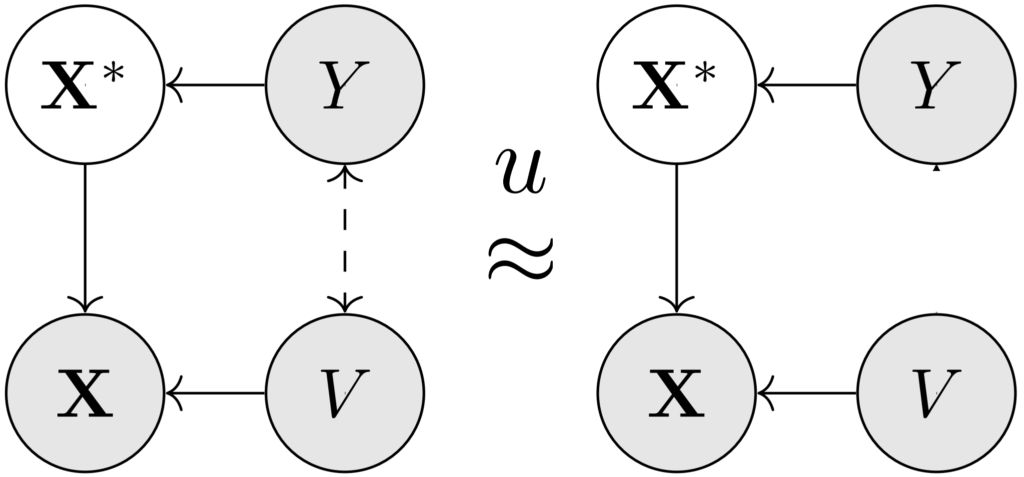


Sampling from non-ideal distributions

Proposition (informal): Reweighting with u_i recovers the independences in the ideal distribution

$u_i \rightarrow$ Makes Y, V "look" independent

$$u_i = \frac{P(Y = y_i)P(V = v_i)}{P(Y = y_i, V = v_i)}$$



Training objective

$$f^* = \text{argmin}_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+ \alpha \cdot \text{MMD}_{\mathbf{u}}\Big(P(f(\mathbf{x}_i)|v_i = 1), P(f(\mathbf{x}_i)|v_i = 0)\Big)$$

Weighted penalty on predictions encoding information about V

Training objective

$$f^* = \text{argmin}_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+ \alpha \cdot \text{MMD}_{\mathbf{u}}\Big(P(f(\mathbf{x}_i)|v_i = 1), P(f(\mathbf{x}_i)|v_i = 0)\Big)$$

Weighted penalty on predictions encoding information about V

Causal perspective gave us:

1. Weights to map the training data to a distribution where invariance is achievable

Training objective

$$f^* = \text{argmin}_f \sum_i \boxed{u_i} \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+ \alpha \cdot \boxed{\text{MMD}_{\mathbf{u}}\Big(P(f(\mathbf{x}_i)|v_i = 1), P(f(\mathbf{x}_i)|v_i = 0)\Big)}$$

Weighted penalty on predictions encoding information about V

Causal perspective gave us:

1. Weights to map the training data to a distribution where invariance is achievable
2. Invariance penalty to encourage the model to encode desirable independencies

Empirical results: water birds data

- **Data:** Semi-simulated

Wah *et al*, Computation & Neural Systems Technical Report 2010
Zhou *et al,* IEEE PAMI 2017
Sagawa *et al,* ICLR 2020

Empirical results: water birds data

- **Data:** Semi-simulated
- **Task:** Predict type of bird
  - Main label = type of bird (water/land)
  - Auxiliary label = type of background (water/land)

Wah *et al*, Computation & Neural Systems Technical Report 2010
Zhou *et al*, IEEE PAMI 2017
Sagawa *et al*, ICLR 2020

Empirical results: water birds data

- **Data:** Semi-simulated
- **Task:** Predict type of bird
  - Main label = type of bird (water/land)
  - Auxiliary label = type of background (water/land)
- **Setup:** Fixed training (source) data

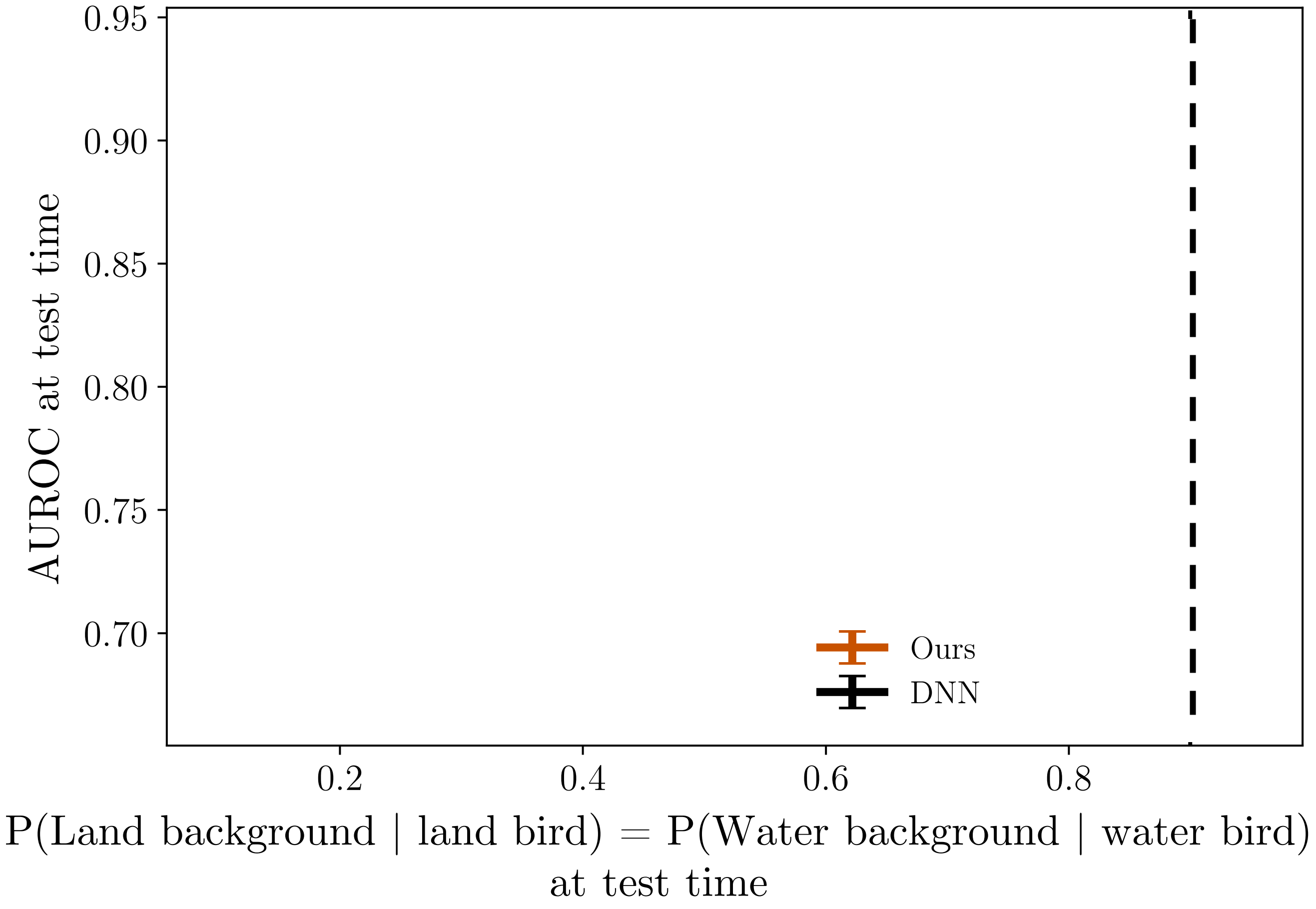
Wah *et al*, Computation & Neural Systems Technical Report 2010
Zhou *et al,* IEEE PAMI 2017
Sagawa *et al,* ICLR 2020

Empirical results: water birds data

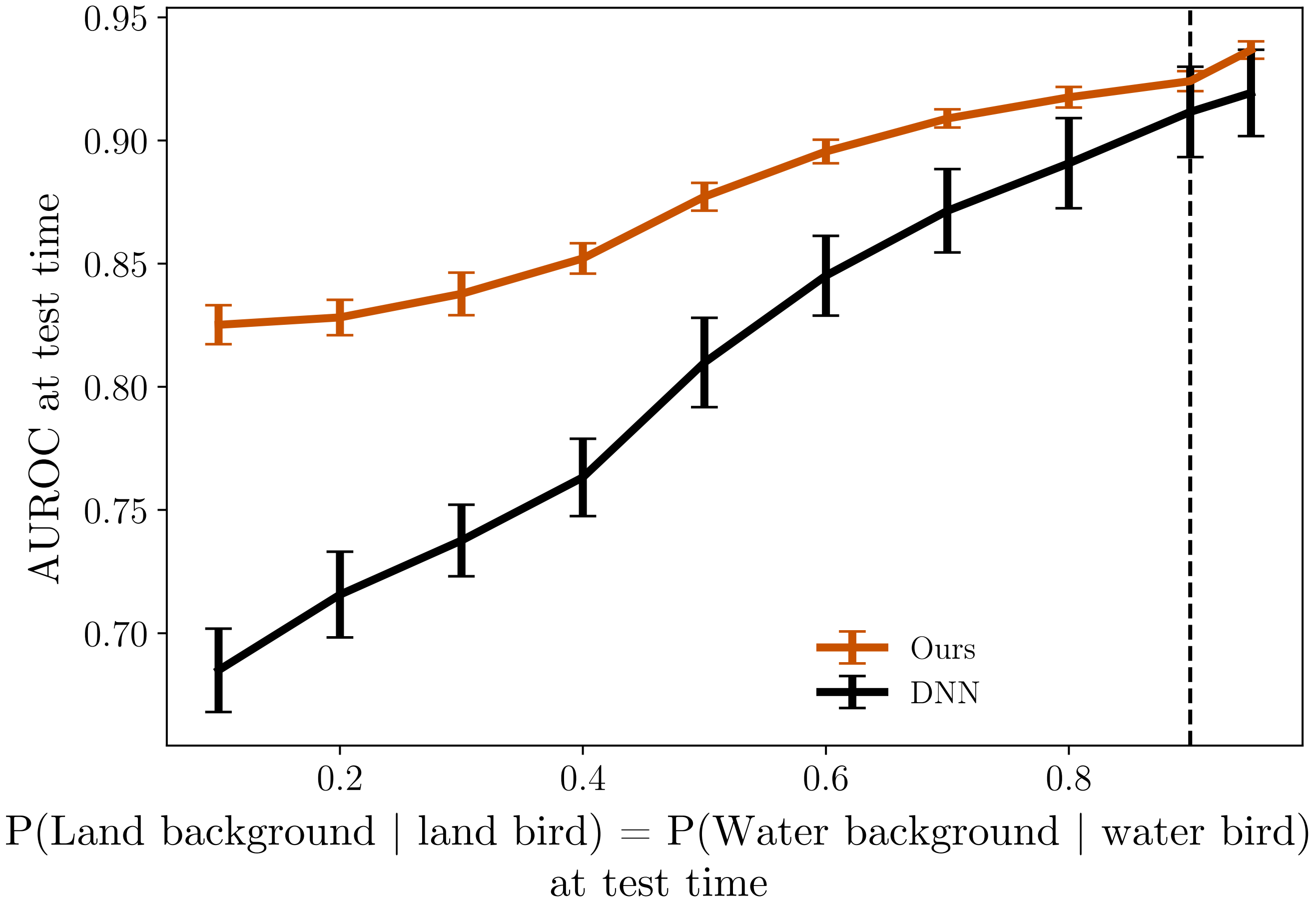
- **Data:** Semi-simulated
- **Task:** Predict type of bird
  - Main label = type of bird (water/land)
  - Auxiliary label = type of background (water/land)
- **Setup:** Fixed training (source) data
- **Evaluation:** On multiple test sets with different correlations

Wah *et al*, Computation & Neural Systems Technical Report 2010
Zhou *et al,* IEEE PAMI 2017
Sagawa *et al,* ICLR 2020

Experiment results

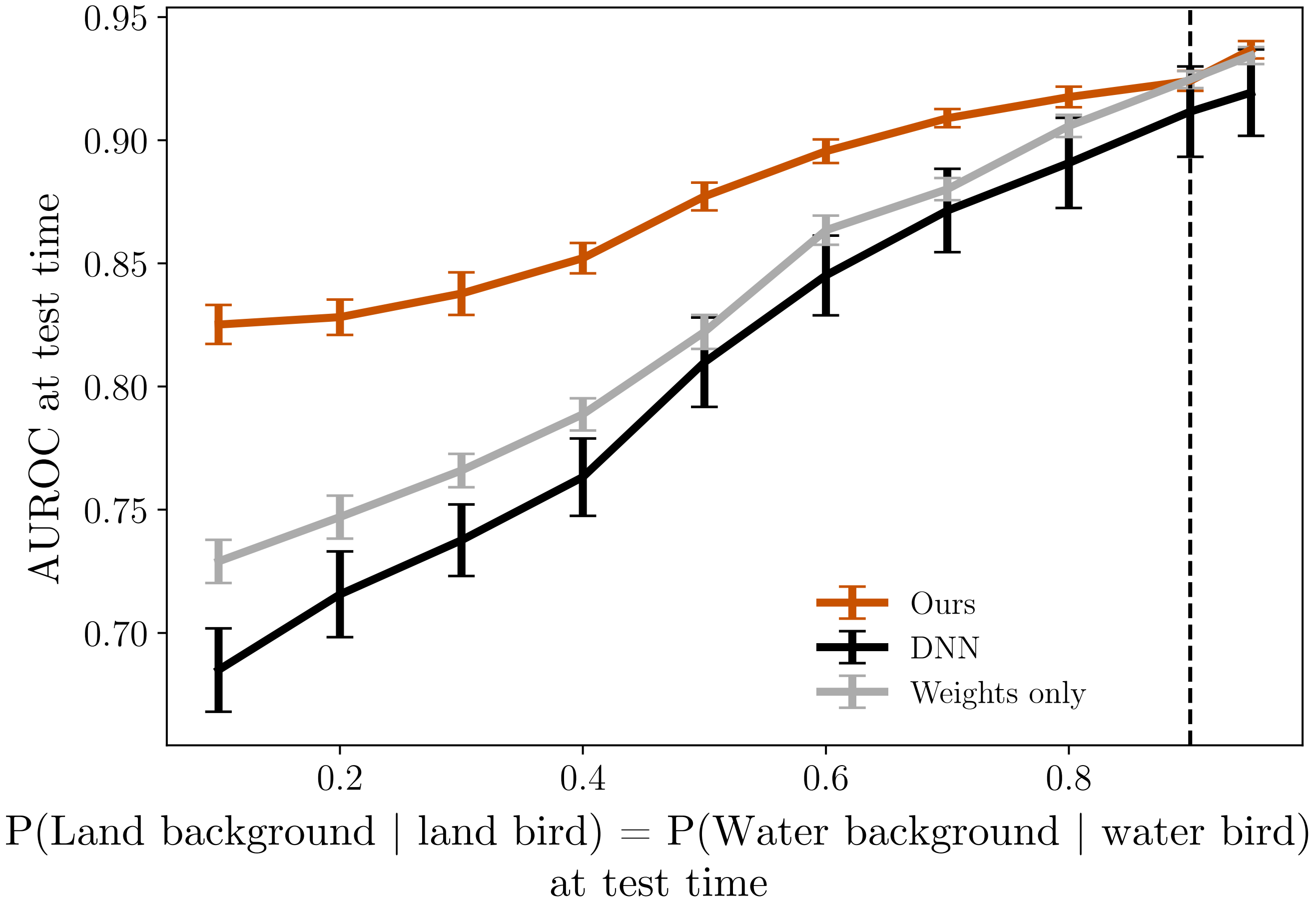


Experiment results

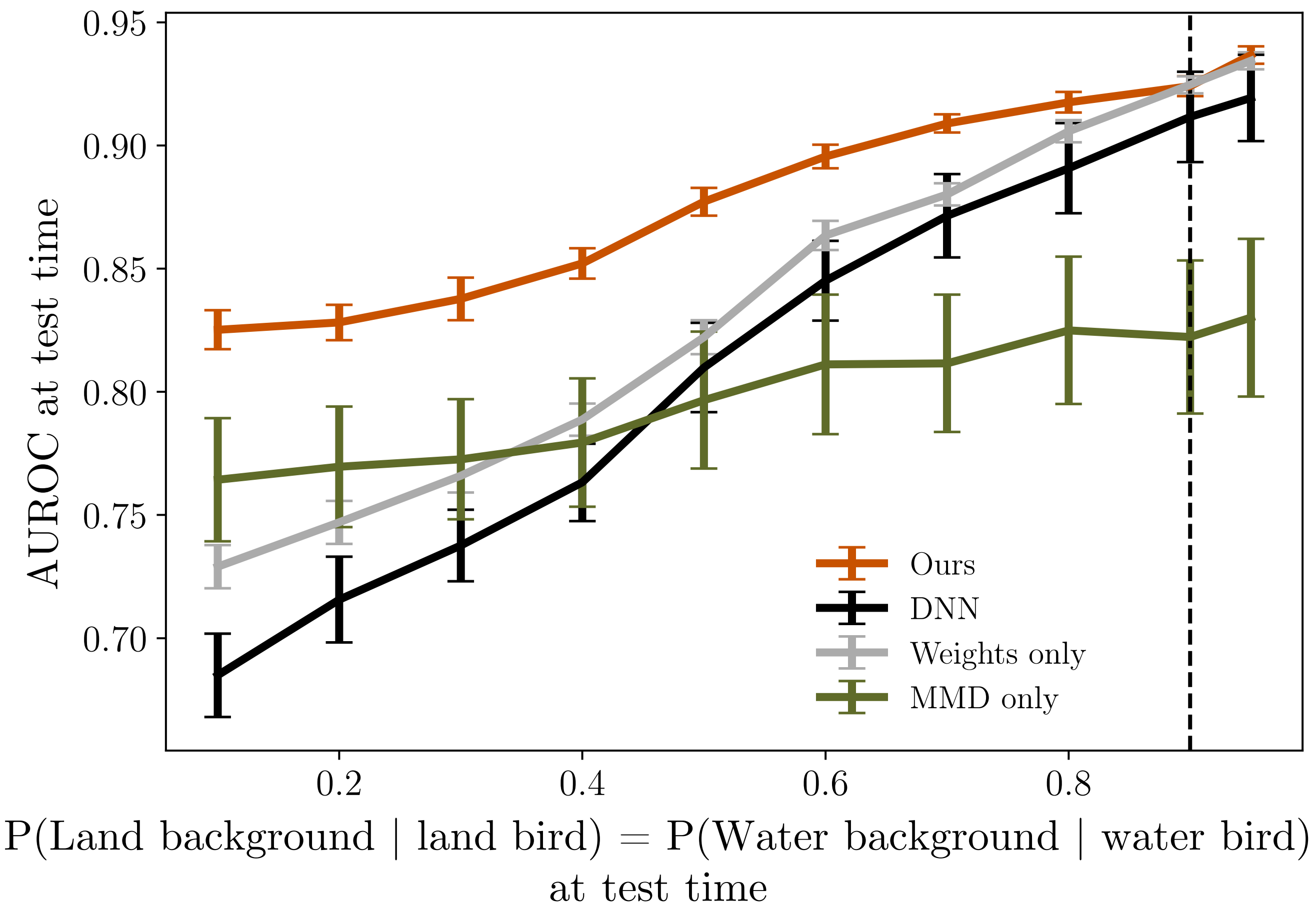


Our approach is less prone to relying on the shortcut, and is more accurate than typical methods

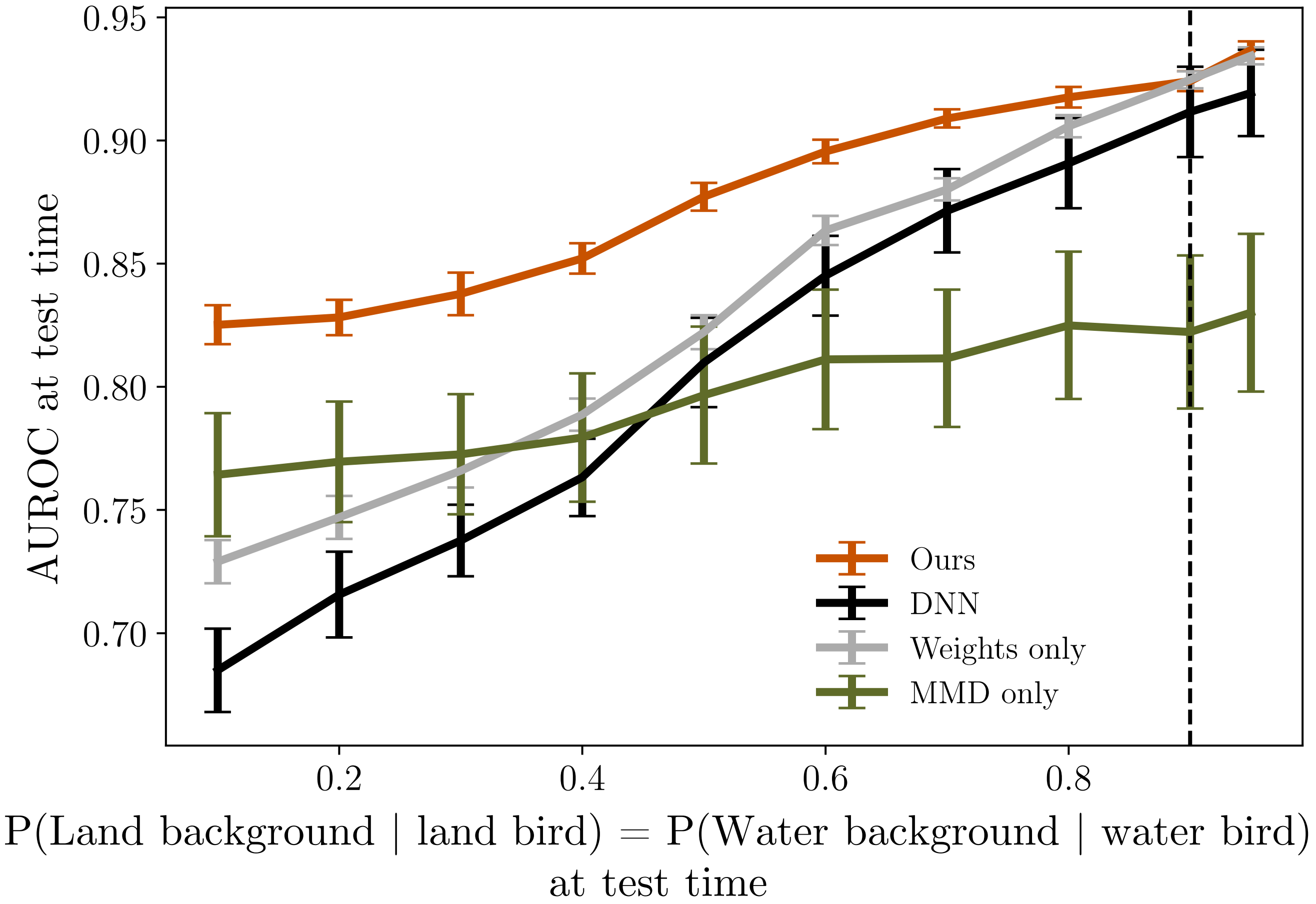
Experiment results



Experiment results



Experiment results



The two causally-inspired components of our approach are necessary

Empirical results: Predicting Pneumonia

- **Data**: CheXpert, down-sample women with pneumonia at training time.
- **Task**: Predict the onset of pneumonia (main label), while making sure that sex (auxiliary label) is not a shortcut.

Irvin *et al*, AAAI 2019
Jabbour *et al*, ML4H 2020

Empirical results: Predicting Pneumonia

- **Data**: CheXpert, down-sample women with pneumonia at training time.
- **Task**: Predict the onset of pneumonia (main label), while making sure that sex (auxiliary label) is not a shortcut.

	AUROC	
	Ours	DNN
Same hospital (no shift)		
Different hospital (shift)		

Irvin *et al*, AAAI 2019
Jabbour *et al*, ML4H 2020

Empirical results: Predicting Pneumonia

- **Data**: CheXpert, down-sample women with pneumonia at training time.
- **Task**: Predict the onset of pneumonia (main label), while making sure that sex (auxiliary label) is not a shortcut.

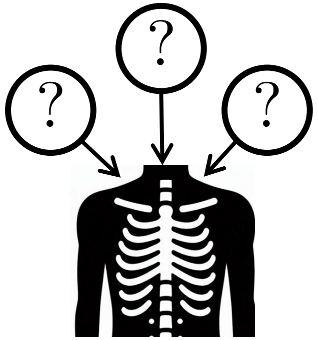
	AUROC	
	Ours	DNN
Same hospital (no shift)	**0.85 (0.007)**	0.82 (0.03)
Different hospital (shift)	**0.75 (0.006)**	0.69 (0.028)

Irvin *et al*, AAAI 2019
Jabbour *et al*, ML4H 2020

Talk outline

1 Efficiency + robustness to known sampling bias

MPMBHD, AIStats 22
MD, TMLR 23
NM, UAI 24



2 Efficiency + robustness to unknown sampling biases

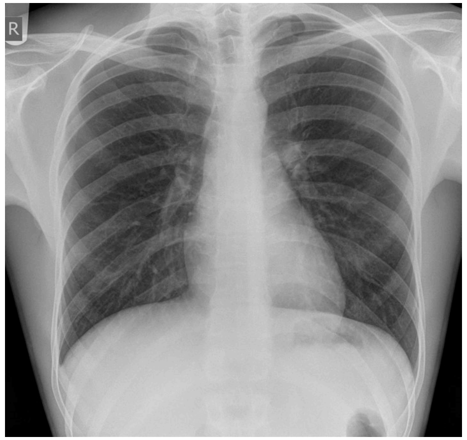
ZM, NeurIPS 22
WJMSW, NeurIPS 22

3 Evaluating localized circuits in LLMs

SVNZGJMB – NeurIPS 24

What about unknown shortcuts?

- **Have:** Large number of auxiliary labels

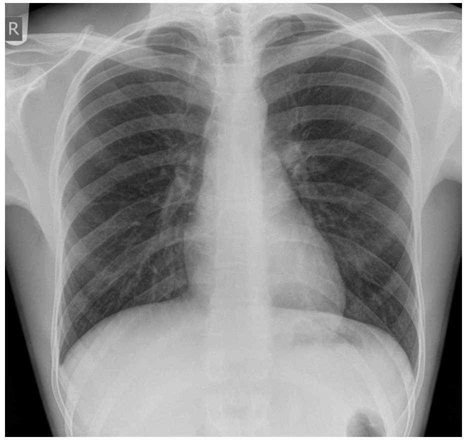


+ Patient sex
+ Type of X-ray machine
+ Other medical conditions:
  Flu
  Edema
  Fractures
  Cervical cancer...

Zheng & **Makar**, NeurIPS 2022

What about unknown shortcuts?

- **Have:** Large number of auxiliary labels
- **Unknown:** Which ones are relevant shortcuts?

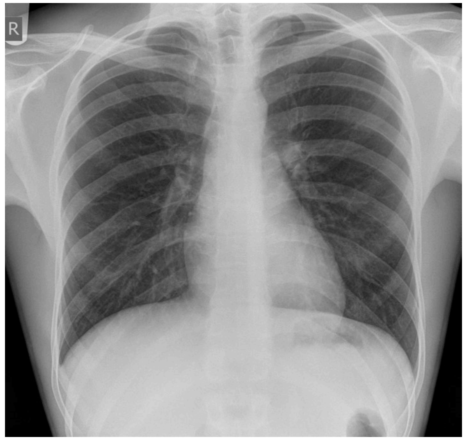


\+ Patient sex
\+ Type of X-ray machine
\+ Other medical conditions:
    Flu
    Edema
    Fractures
    Cervical cancer...

Zheng & **Makar**, NeurIPS 2022

What about unknown shortcuts?

- **Have:** Large number of auxiliary labels
- **Unknown:** Which ones are relevant shortcuts?
- **Objective**: Models robust to multiple shortcuts

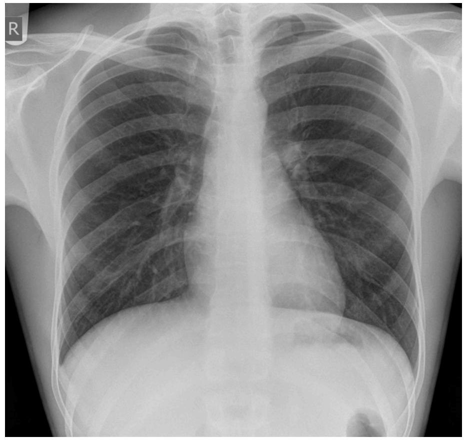


\+ Patient sex
\+ Type of X-ray machine
\+ Other medical conditions:
Flu
Edema
Fractures
Cervical cancer...

Zheng & Makar, NeurIPS 2022

What about unknown shortcuts?

- **Have:** Large number of auxiliary labels
- **Unknown:** Which ones are relevant shortcuts?
- **Objective**: Models robust to multiple shortcuts
- **Upshot:** An additional causal discovery step

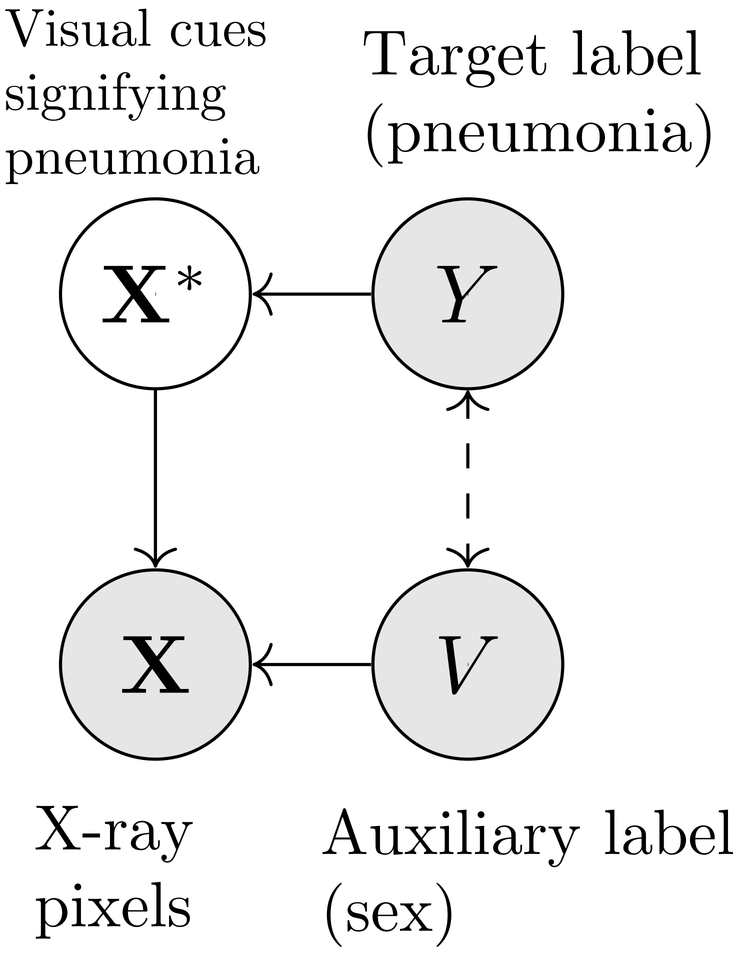


+ Patient sex
+ Type of X-ray machine
+ Other medical conditions:
  Flu
  Edema
  Fractures
  Cervical cancer...

Zheng & **Makar**, NeurIPS 2022

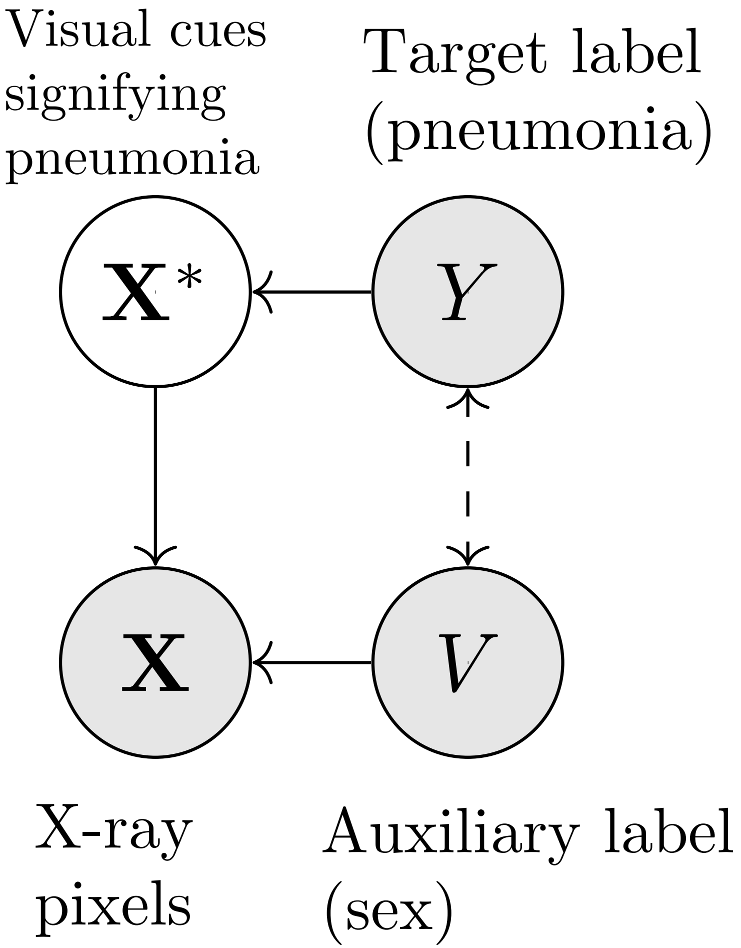
Recall the causal assumption

Previous DAG



Extension to a class of DAGs

Previous DAG

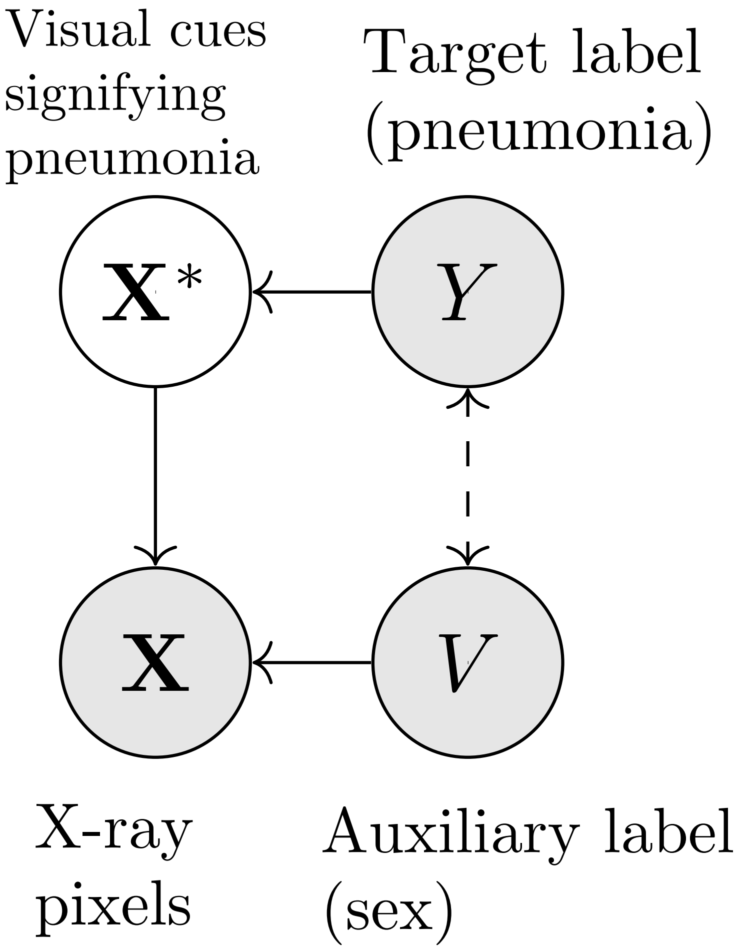


New DAGs

A class of DAGs
High dim. auxiliary labels V^{full}

Extension to a class of DAGs

Previous DAG



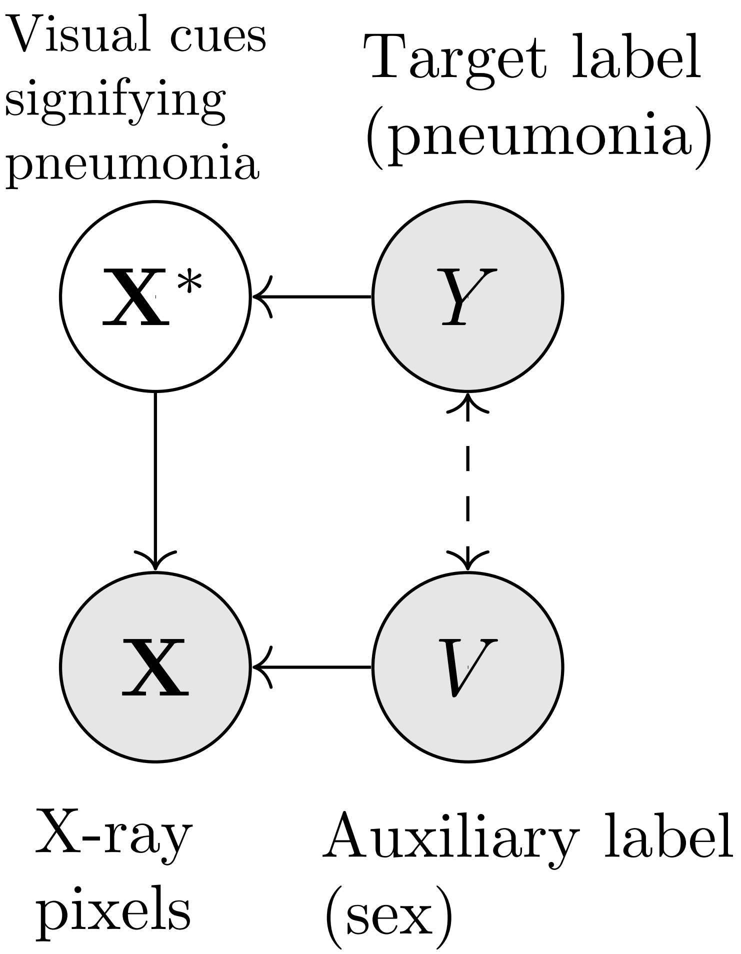
New DAGs

A class of DAGs

High dim. auxiliary labels V^{full}

Extension to a class of DAGs

Previous DAG



New DAGs

A class of DAGs
High dim. auxiliary labels V^{full}

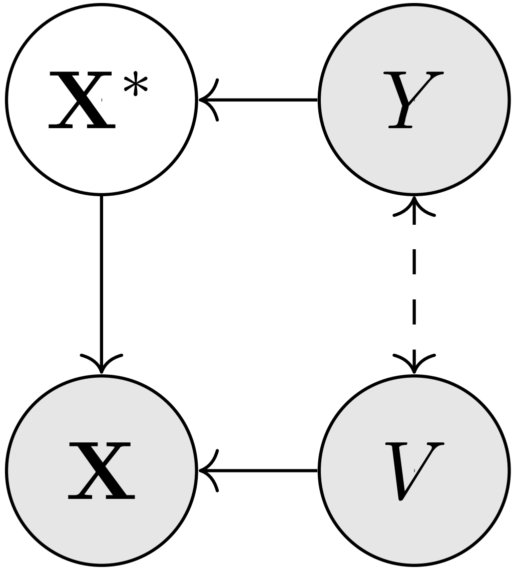
V^{suff}
Correlated but not causally related to Y and cause X

Extension to a class of DAGs

Previous DAG

Visual cues signifying pneumonia

Target label (pneumonia)



X-ray pixels

Auxiliary label (sex)

New DAGs

A class of DAGs

High dim. auxiliary labels V^{full}

V^{suff}

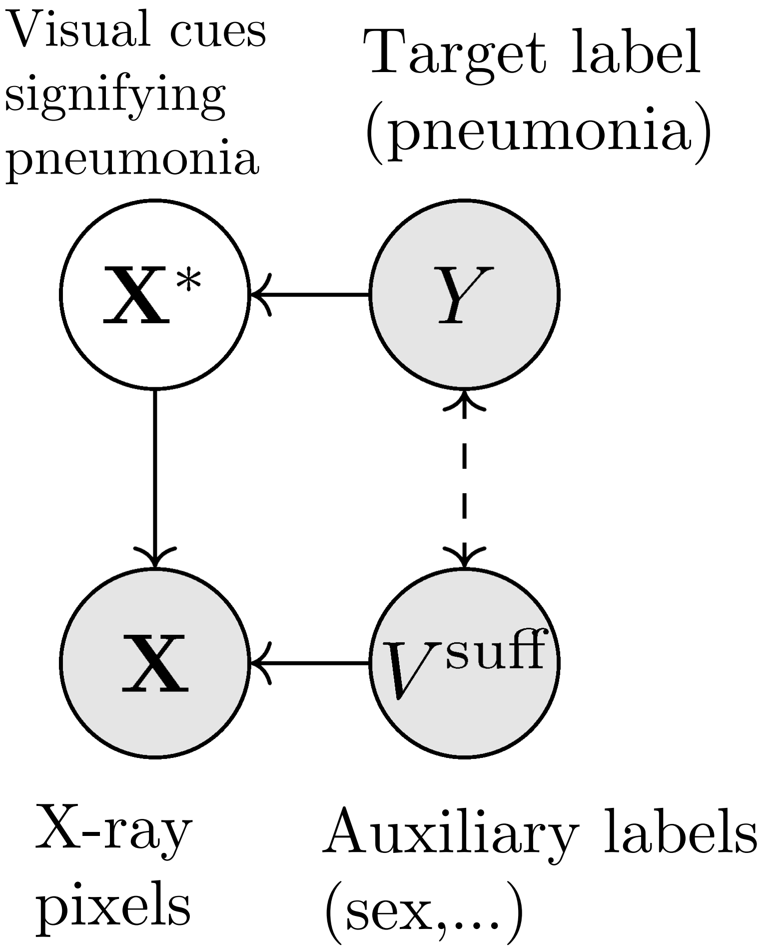
Correlated but not causally related to Y and cause X

V^{red}

Any other variables not caused by X

Extension to a class of DAGs

Example from new DAG class



New DAGs

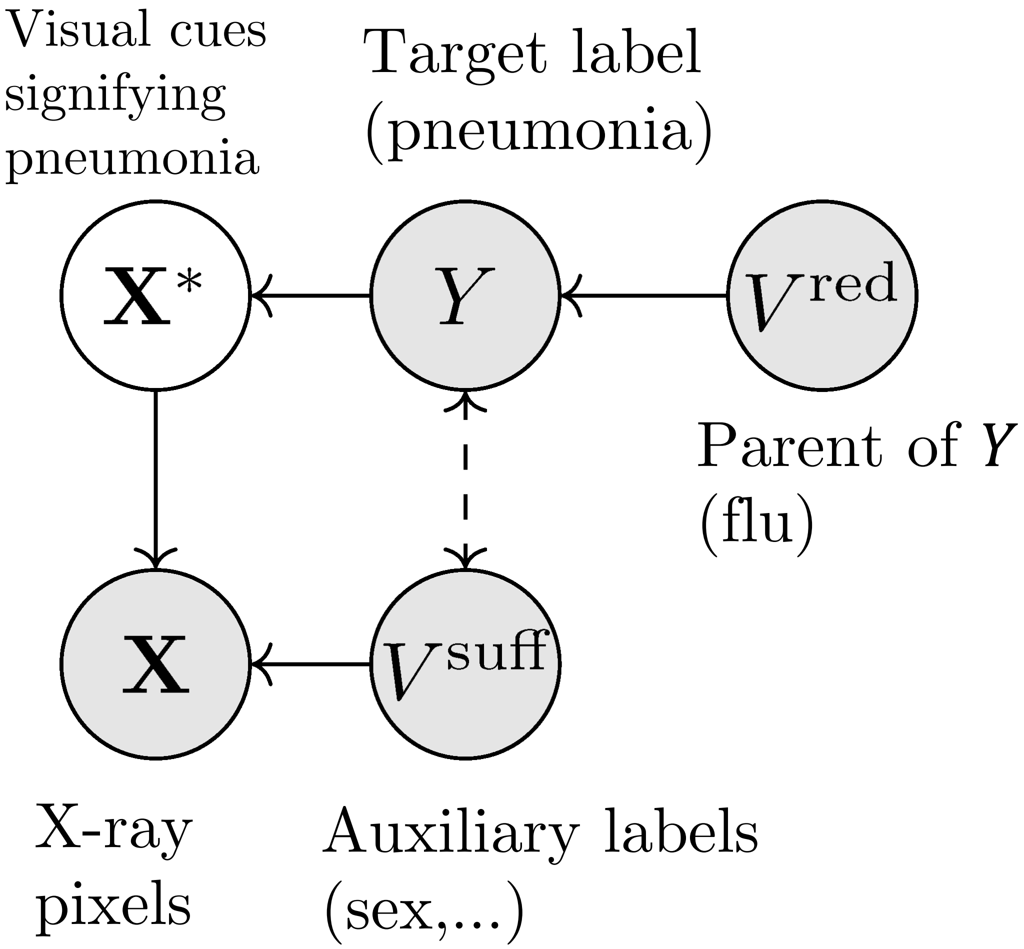
A class of DAGs
High dim. auxiliary labels V^{full}

V^{suff}
Correlated but not causally related to Y and cause X

V^{red}
Any other variables not caused by X

Extension to a class of DAGs

Example from new DAG class



New DAGs

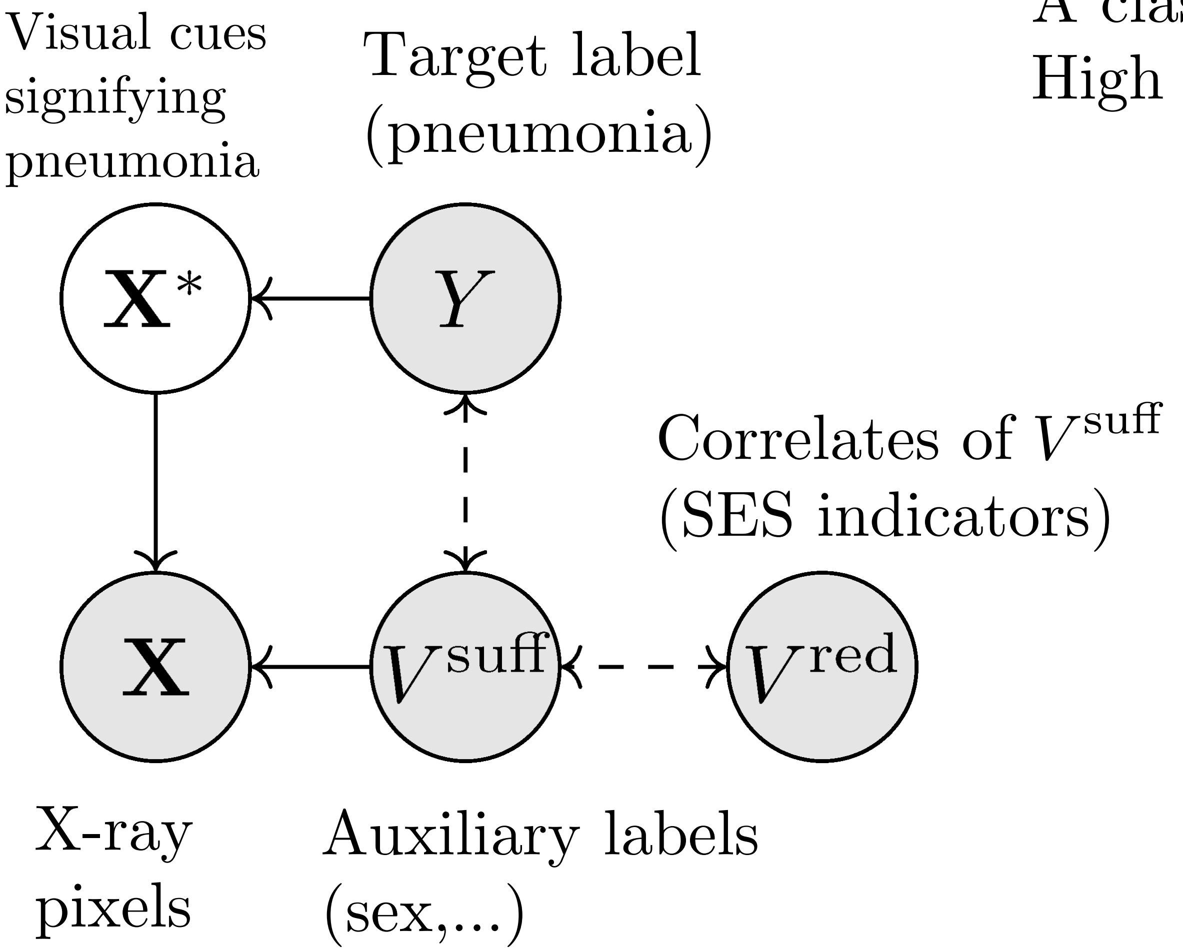
A class of DAGs
High dim. auxiliary labels V^{full}

V^{suff}
Correlated but not causally related to Y and cause X

V^{red}
Any other variables not caused by X
E.g., Parents of Y

Extension to a class of DAGs

Example from new DAG class



New DAGs

A class of DAGs

High dim. auxiliary labels V^{full}

V^{suff}

Correlated but not causally related to Y and cause X

V^{red}

Any other variables not caused by X

E.g., Parents of Y

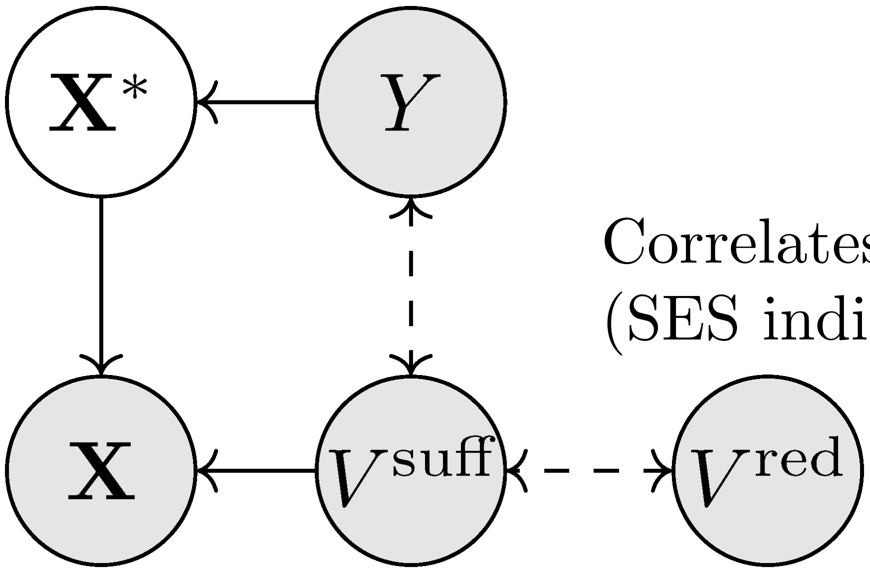
Correlates of V^{suff}

Extension to a class of DAGs

Example from new DAG class

Visual cues signifying pneumonia

Target label (pneumonia)



Correlates of V^{suff} (SES indicators)

X-ray pixels

Auxiliary labels (sex,...)

New DAGs

A class of DAGs
High dim. auxiliary labels V^{full}

V^{suff}

Correlated but not causally related to Y and cause X

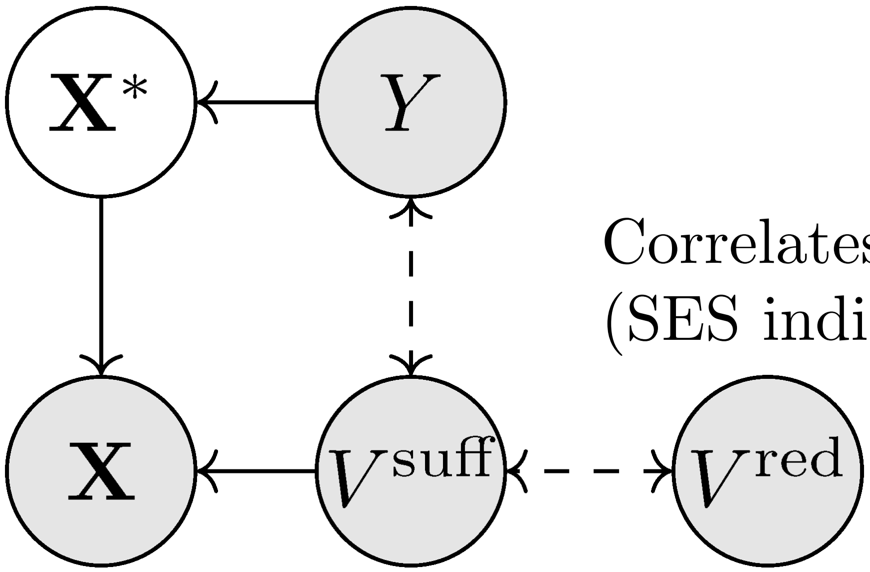
V^{red}

Any other variables not caused by X
E.g., Parents of Y
Correlates of V^{suff}

Unknown: which auxiliary labels are V^{suff} vs. V^{red}

Extension to a class of DAGs

Example from new DAG class



New DAGs

A class of DAGs
High dim. auxiliary labels V^{full}

V^{suff}
Correlated but not causally related to Y and cause X

V^{red}
Any other variables not caused by X
E.g., Parents of Y
Correlates of V^{suff}

Unknown: which auxiliary labels are V^{suff} vs. V^{red}

Want: $f(\boldsymbol{X})$ robust to any changes in correlations

One possible solution

- Train model to be robust to all of V^{full}

$$\min_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+ \alpha \cdot \mathrm{HSIC}(f(\mathbf{X}), \mathbf{V}^{\text{full}})$$

Weighted penalty on predictions encoding information about V^{full}

One possible solution

- Train model to be robust to all of V^{full}

$$\min_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+ \alpha \cdot \mathrm{HSIC}(f(\mathbf{X}), \mathbf{V}^{\text{full}})$$

Weighted penalty on predictions encoding information about V^{full}

- ..but the accuracy of the penalty and stability of weights become unstable as the dimension of V^{full} increases

Ramdas et al, AAAI 2015

One possible solution

- Train model to be robust to all of V^{full}

$$\min_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$

Weighted prediction loss

$$+\alpha \cdot \mathrm{HSIC}(f(\mathbf{X}), \mathbf{V}^{\text{full}})$$

Weighted penalty on predictions encoding information about V^{full}

- ..but the accuracy of the penalty and stability of weights become unstable as the dimension of V^{full} increases

Want robustness penalty to be defined with respect to a small set of **sufficient** auxiliary labels...

Ramdas et al, AAAI 2015

The sufficiency of $\mathbf{V}^{\text{suff}}$

Proposition (informal): Invariance to V^{suff} is sufficient to induce robustness across any changes to correlations in the system.

The sufficiency of $\mathbf{V}^{\text{suff}}$

Proposition (informal): Invariance to V^{suff} is sufficient to induce robustness across any changes to correlations in the system.

Challenge: Don't know which is V^{suff} vs V^{red}

The identifiability of $\mathbf{V}^{\text{suff}}$

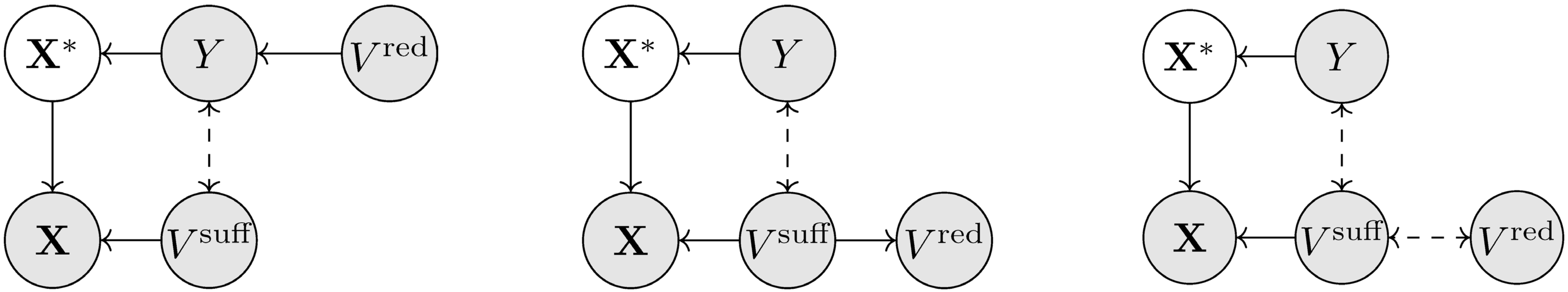
Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

Goal of tests: eliminate V^{red}

The identifiability of $\mathbf{V}^{\mathrm{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

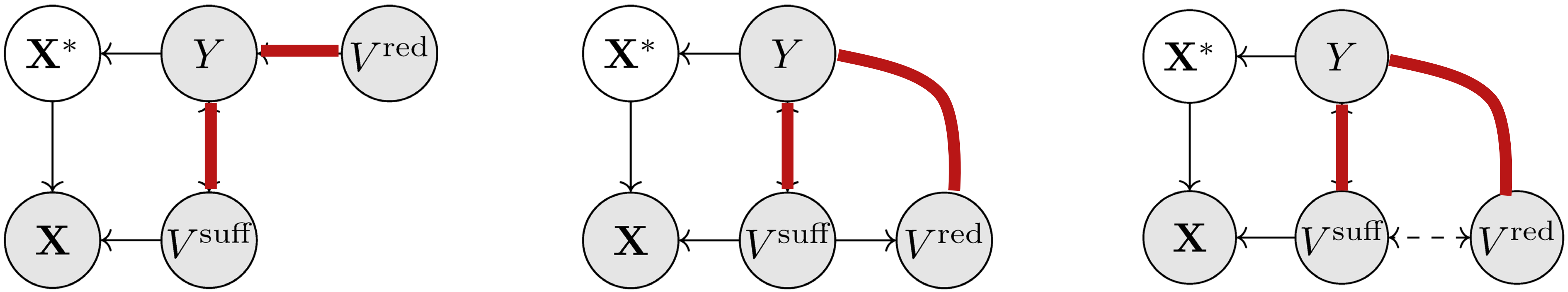
Goal of tests: eliminate V^{red}



The identifiability of $\mathbf{V}^{\text{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

Goal of tests: eliminate V^{red}

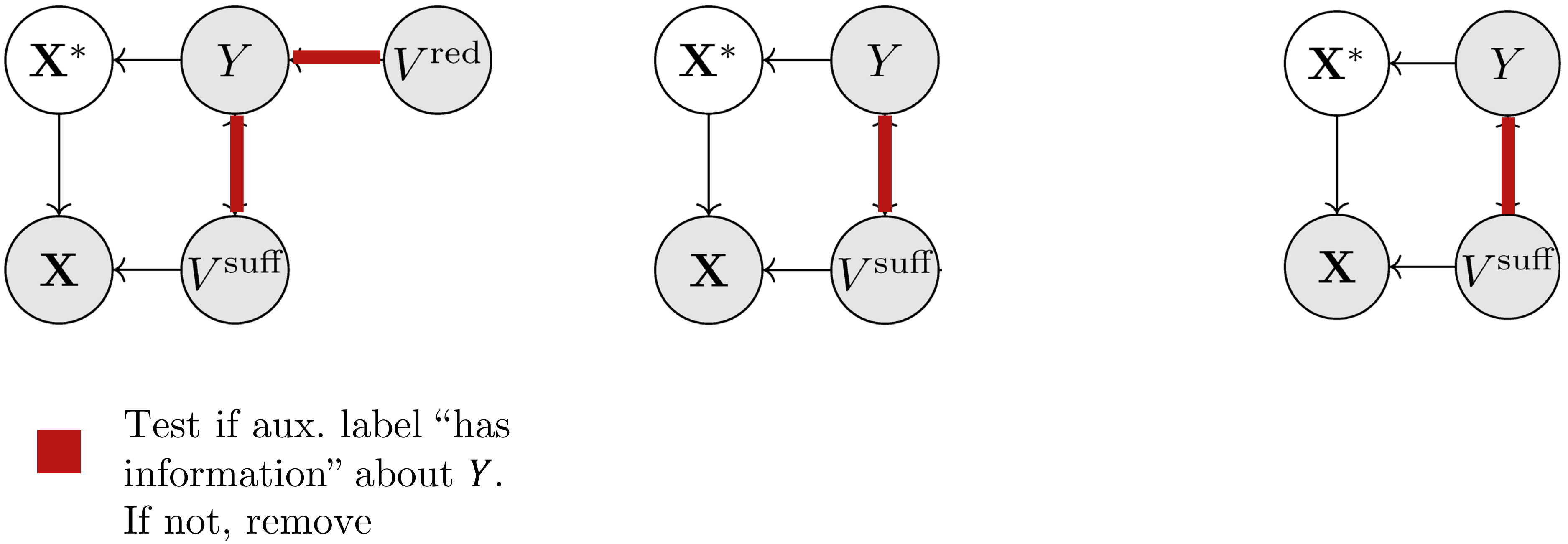


Test if aux. label "has information" about Y. If not, remove

The identifiability of $\mathbf{V}^{\text{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

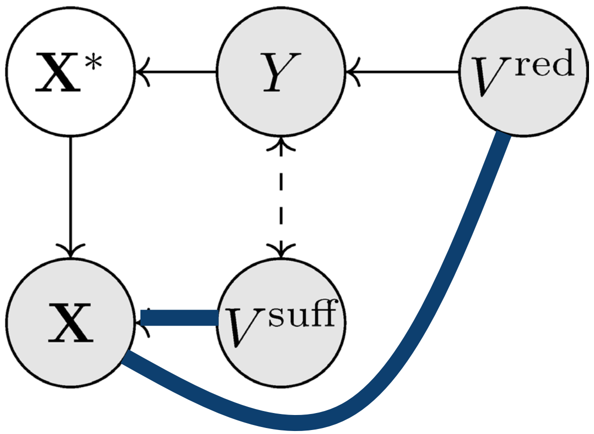
Goal of tests: eliminate V^{red}

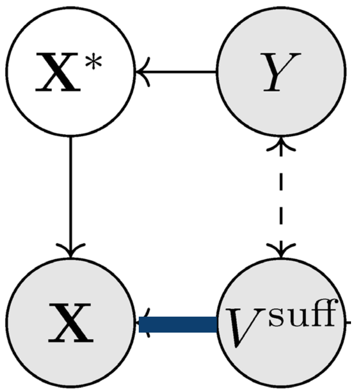


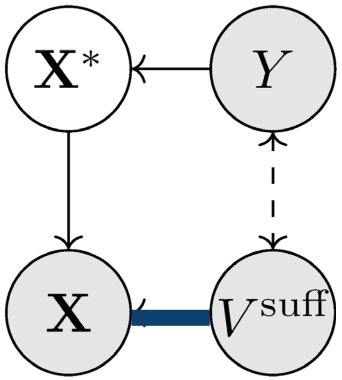
The identifiability of $\mathbf{V}^{\text{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

Goal of tests: eliminate V^{red}







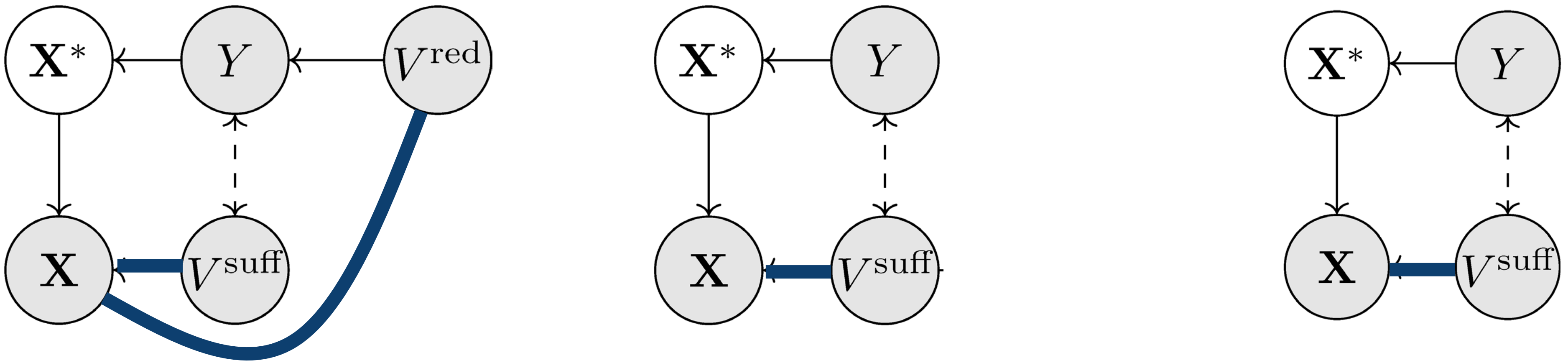
■ Test if aux. label "has information" about Y. If not, remove

■ Test if aux. label "has information" about X. If not, remove

The identifiability of $\mathbf{V}^{\text{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

Goal of tests: eliminate V^{red}



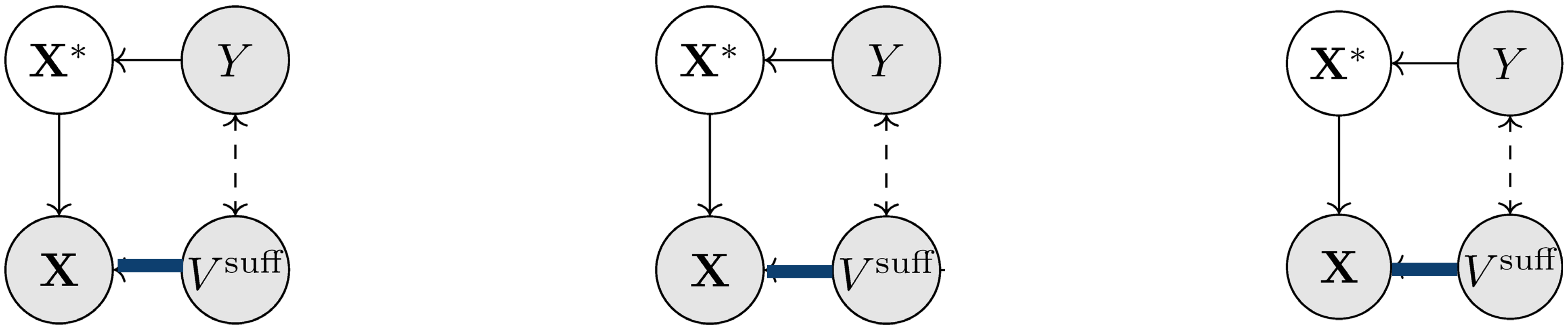
Test if aux. label "has information" about Y. If not, remove

Test if aux. label "has information" about X. If not, remove

The identifiability of $\mathbf{V}^{\mathrm{suff}}$

Proposition (informal): V^{suff} is identifiable through 2 asymptotically consistent tests

Goal of tests: eliminate V^{red}

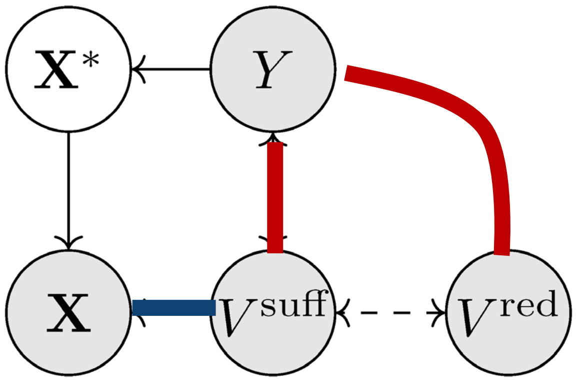


Test if aux. label "has information" about Y. If not, remove

Test if aux. label "has information" about X. If not, remove

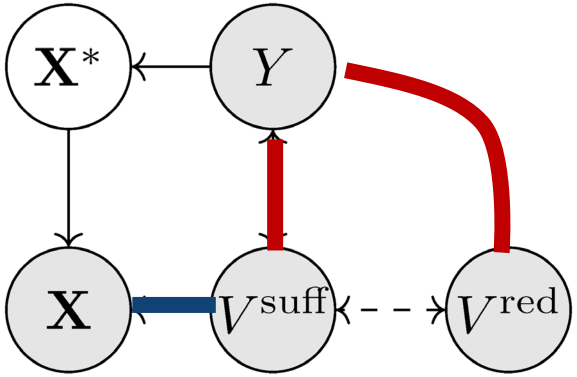
A new training procedure

Step 1: Test for the two properties to identify sufficient shortcuts ($\hat{V}^{\text{suff}}$)

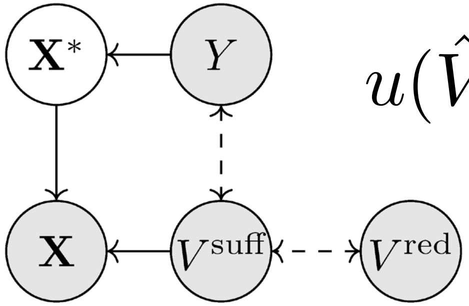


A new training procedure

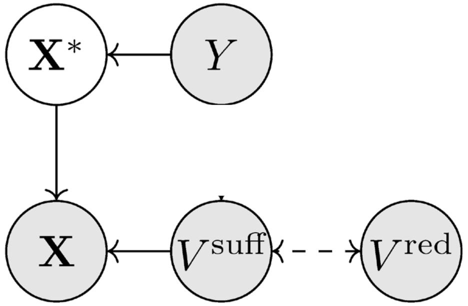
Step 1: Test for the two properties to identify sufficient shortcuts ($\hat{V}^{\text{suff}}$)



Step 2: Reweight

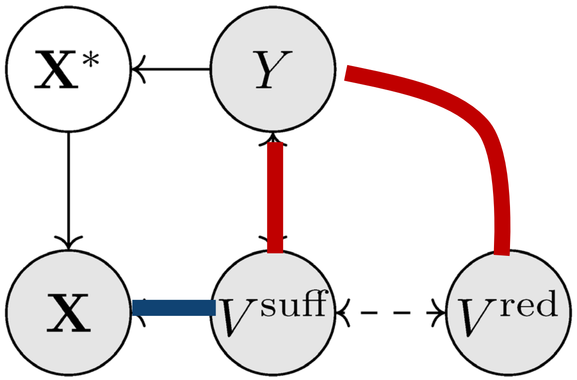


$$u(\hat{V}^{\text{suff}}, Y) \approx$$

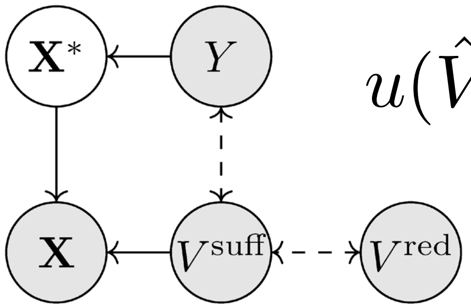


A new training procedure

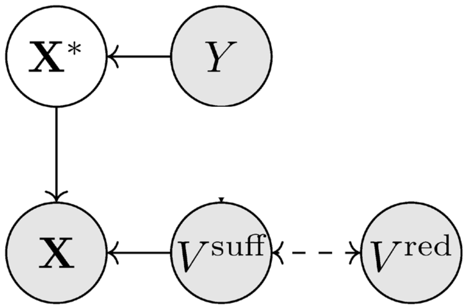
Step 1: Test for the two properties to identify sufficient shortcuts ($\hat{V}^{\text{suff}}$)



Step 2: Reweight



$$u(\hat{V}^{\text{suff}}, Y) \approx$$



...and optimize

$$\min_f \sum_i u_i \ell(f(\mathbf{x}_i), y_i)$$
$$+ \alpha \cdot \mathrm{HSIC}(f(\mathbf{X}), \hat{V}^{\text{suff}})$$

Water birds: revisited

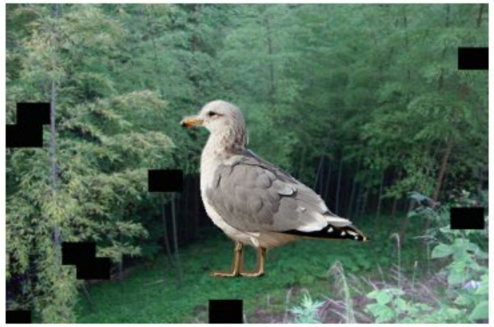
- Predict type of bird (water/land)
- 12 auxiliary labels, only 2 sufficient:
  - Background
  - Camera quality

Water birds: revisited

- Predict type of bird (water/land)
- 12 auxiliary labels, only 2 sufficient:
  - Background
  - Camera quality

Water bird on water background

Water bird on land background



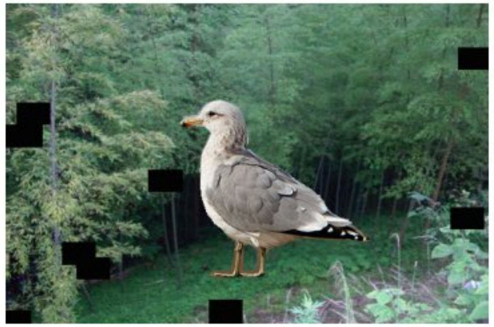
Water bird on land background, bad camera

Water birds: revisited

- Predict type of bird (water/land)
- 12 auxiliary labels, only 2 sufficient:
  - Background
  - Camera quality
- At training time, most water birds are on water background taken with a good quality camera
- Test on varying distributions

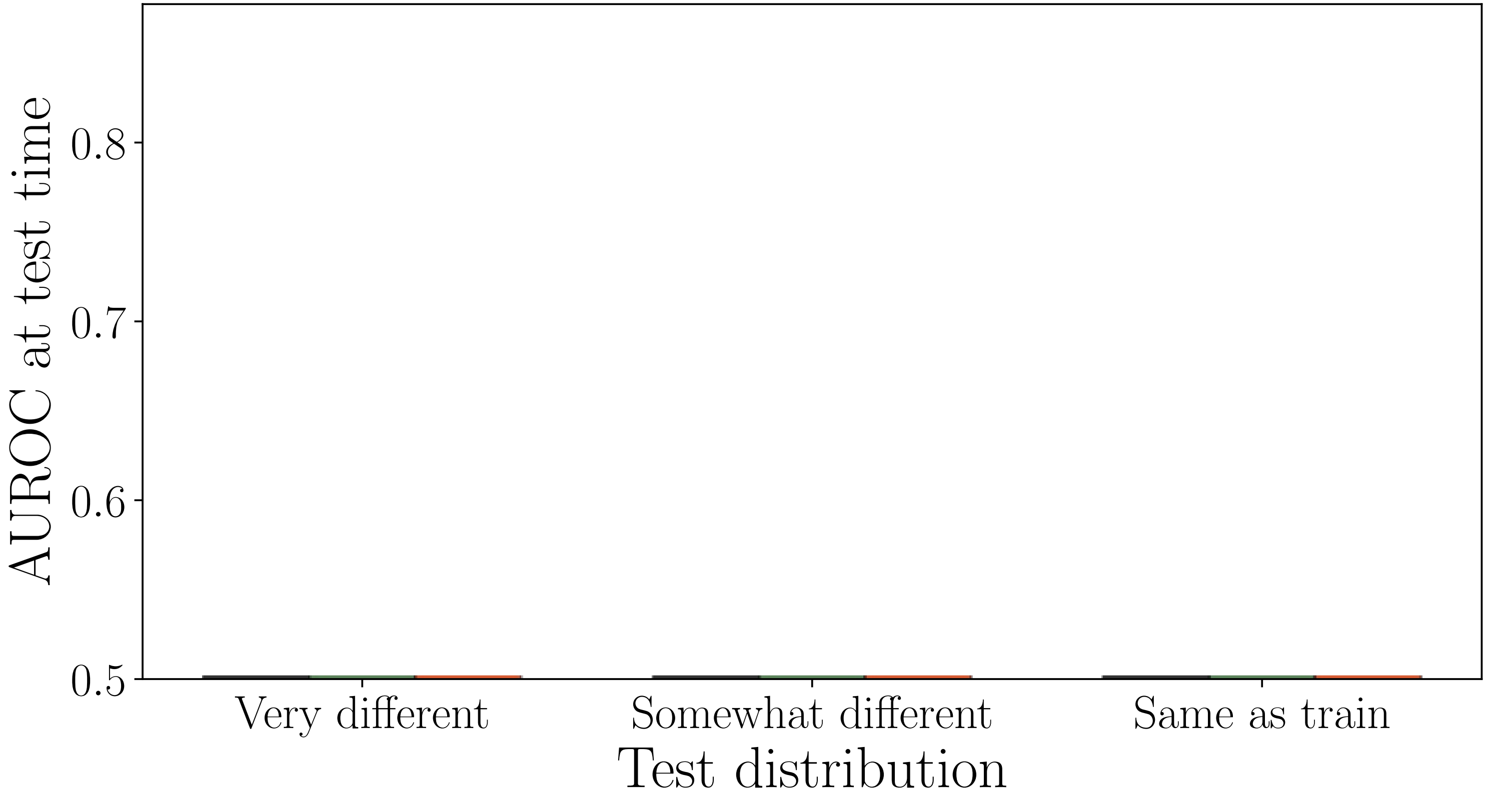
Water bird on water background

Water bird on land background

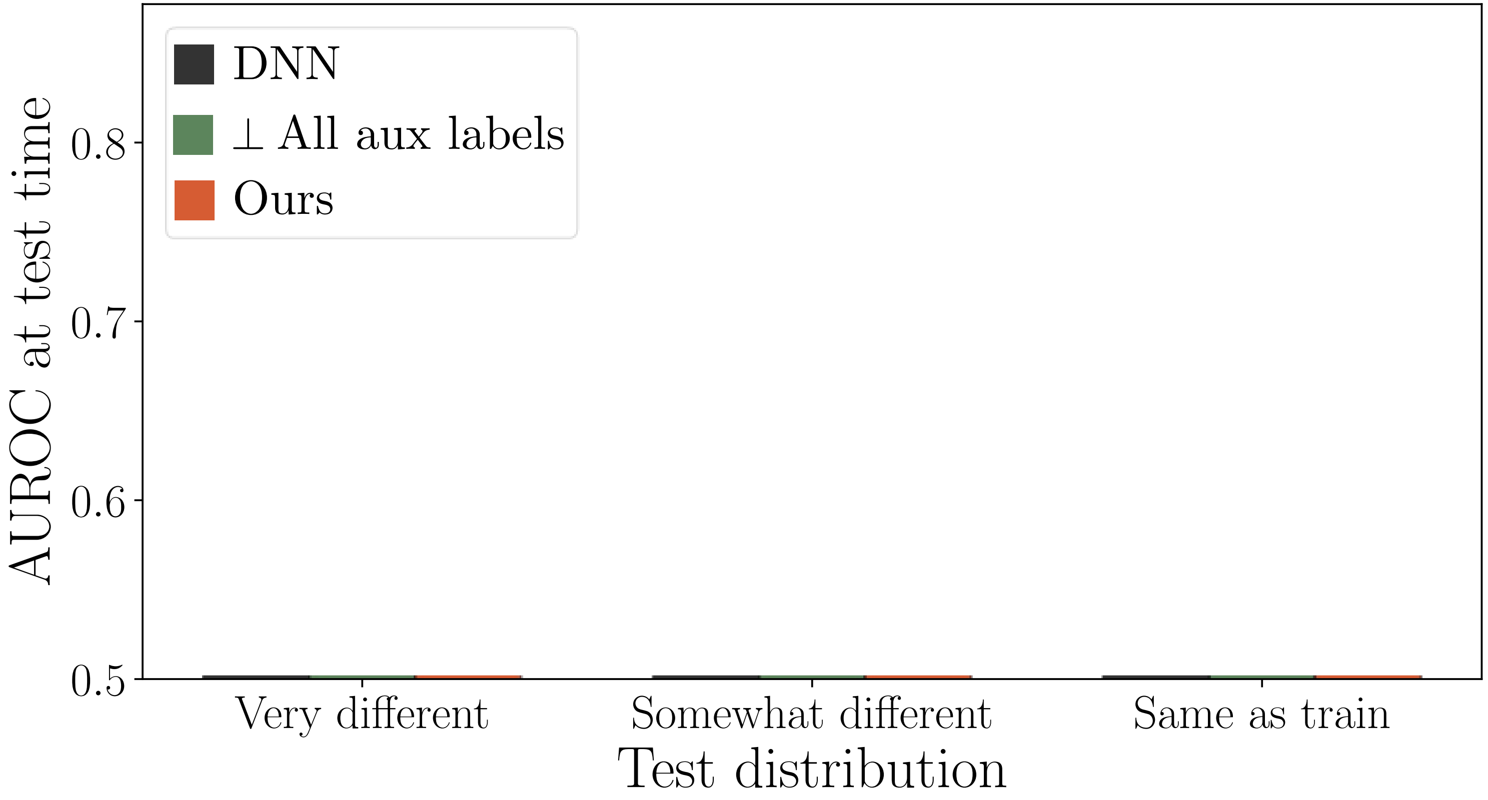


Water bird on land background, bad camera

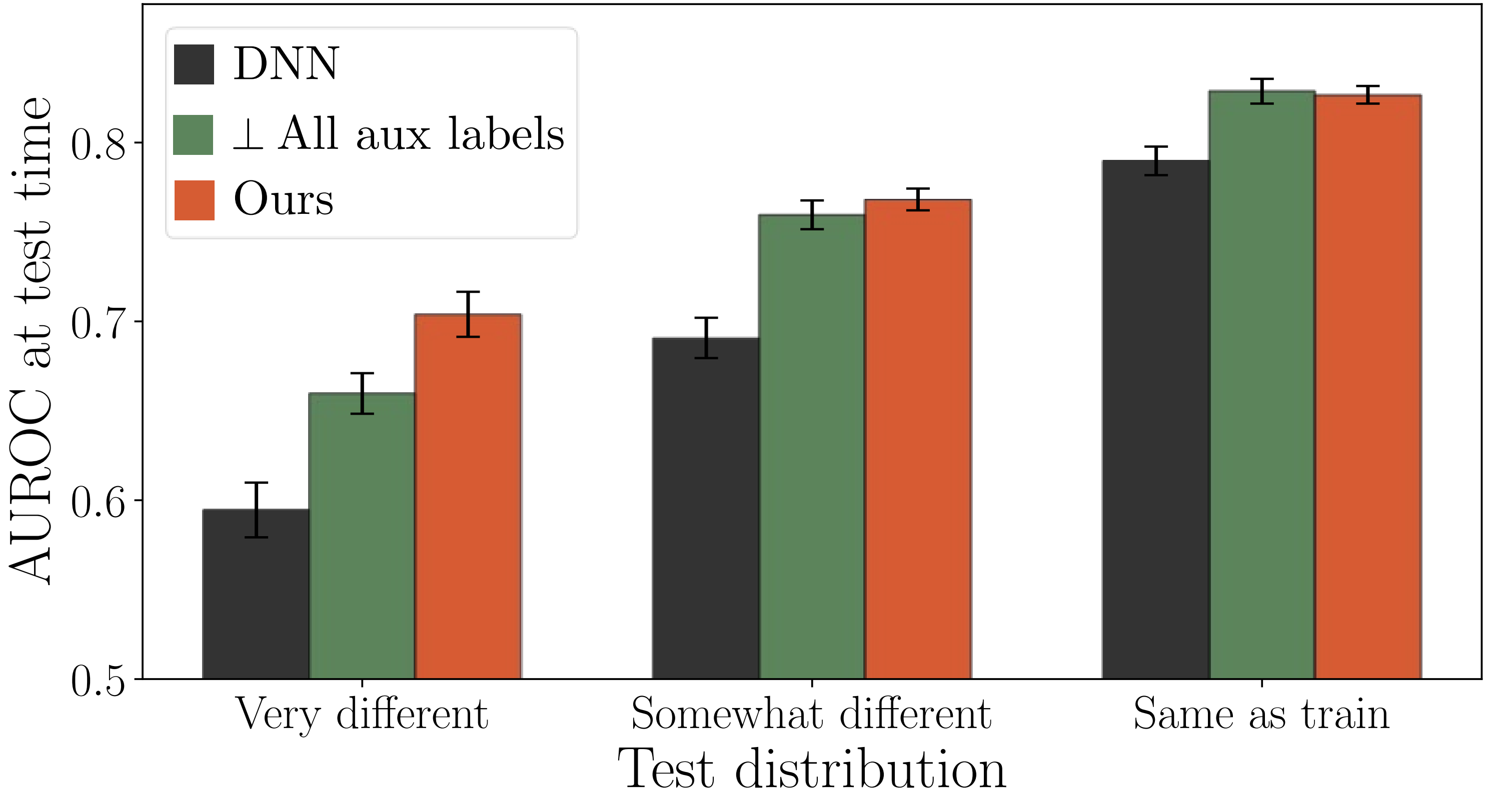
Water birds experiment results



Water birds experiment results



Water birds experiment results



By identifying the sufficient shortcuts, our approach leads to more reliable models

Talk outline

1 Efficiency + robustness to known sampling bias

MPMBHD, AIStats 22
MD, TMLR 23
NM, UAI 24

2 Efficiency + robustness to unknown sampling biases

ZM, NeurIPS 22
WJMSW, NeurIPS 22

3 Evaluating localized circuits in LLMs

SVNZGJMB – NeurIPS 24

LLMs can encode shortcuts too

Task: Decision support for opioid prescriptions

Shi, Velez, Nazaret, Zheng, Alonso, Jesson, **Makar**, Blei, NeurIPS 2024

LLMs can encode shortcuts too

Task: Decision support for opioid prescriptions

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He disclosed that he had been drinking earlier today. Should he be given an opioid?

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He had apparently been drinking, was agitated and belligerent. Should he be given an opioid?

Shi, Velez, Nazaret, Zheng, Alonso, Jesson, **Makar**, Blei, NeurIPS 2024
E.g., Feder et al NeurIPS 2023 and Qi et al, EMNLP 2021

LLMs can encode shortcuts too

Task: Decision support for opioid prescriptions

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He disclosed that he had been drinking earlier today. Should he be given an opioid?
A: Yes

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He had apparently been drinking, was agitated and belligerent. Should he be given an opioid?
A: No

Shi, Velez, Nazaret, Zheng, Alonso, Jesson, Makar, Blei, NeurIPS 2024
E.g., Feder et al NeurIPS 2023 and Qi et al, EMNLP 2021

LLMs can encode shortcuts too

Task: Decision support for opioid prescriptions

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He disclosed that he had been drinking earlier today. Should he be given an opioid?
A: Yes

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He had apparently been drinking, was agitated and belligerent.
Should he be given an opioid?
A: No

Shi, Velez, Nazaret, Zheng, Alonso, Jesson, **Makar**, Blei, NeurIPS 2024
E.g., Feder et al NeurIPS 2023 and Qi et al, EMNLP 2021
Himmelstein et al, JAMA Network Open, 2022

LLMs can encode shortcuts too

Task: Decision support for opioid prescriptions

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He disclosed that he had been drinking earlier today. Should he be given an opioid?
A: Yes

Q: 50 yo man with type I DM who presented to the ED complaining of acute back pain. He had apparently been drinking, was agitated and belligerent.
Should he be given an opioid?
A: No

Challenge: Removing shortcuts through fine-tuning LLMs requires prohibitively large data + compute

Shi, Velez, Nazaret, Zheng, Alonso, Jesson, **Makar**, Blei, NeurIPS 2024
E.g., Feder et al NeurIPS 2023 and Qi et al, EMNLP 2021
Himmelstein et al, JAMA Network Open, 2022

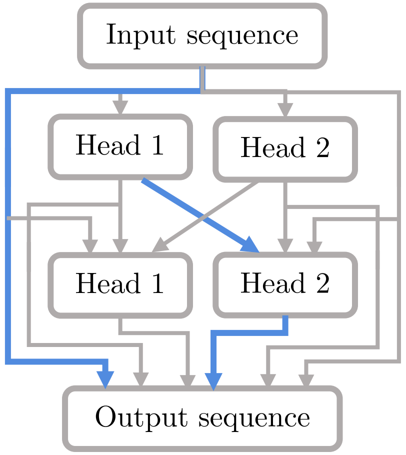
Circuit hypothesis

Olah et al , Distill 2020

Circuit hypothesis

Greater than

"The war lasted from
1615 to ____"

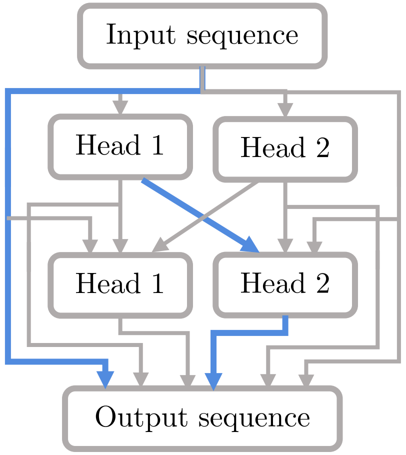


Hanna et al , NeurIPS 2024
Olah et al , Distill 2020

Circuit hypothesis

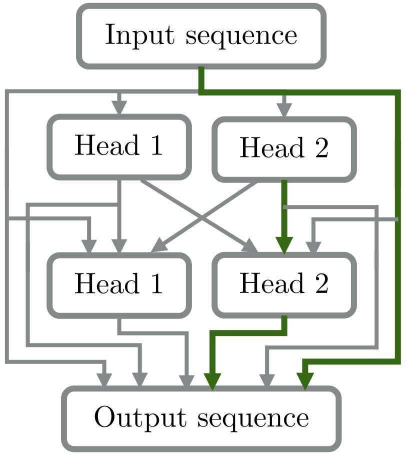
Greater than

“The war lasted from 1615 to _____”



Stigmatizing language?

“Belligerent intoxicated patient presented to the ED. Patient race is ___”



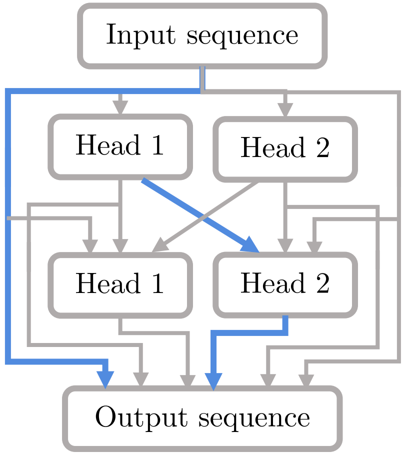
Hanna et al , NeurIPS 2024

Olah et al , Distill 2020

Circuit hypothesis

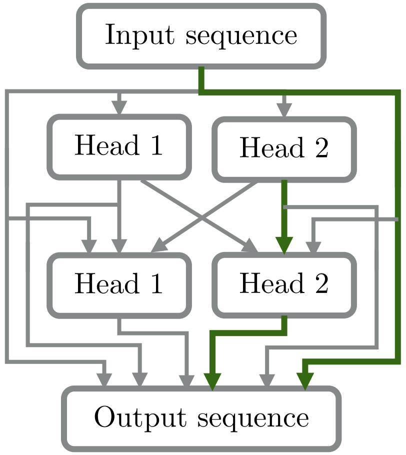
Greater than

"The war lasted from 1615 to _____"



Stigmatizing language?

"Belligerent intoxicated patient presented to the ED. Patient race is ___"



Goal: identify shortcut-encoding circuits

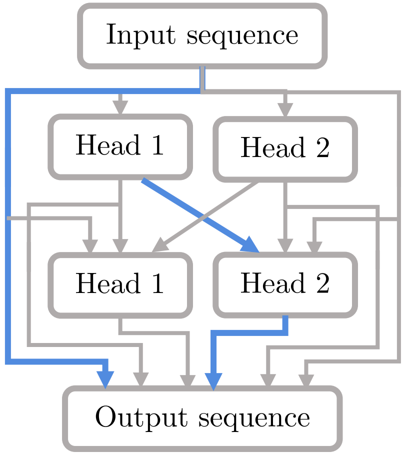
Hanna et al , NeurIPS 2024

Olah et al , Distill 2020

Circuit hypothesis

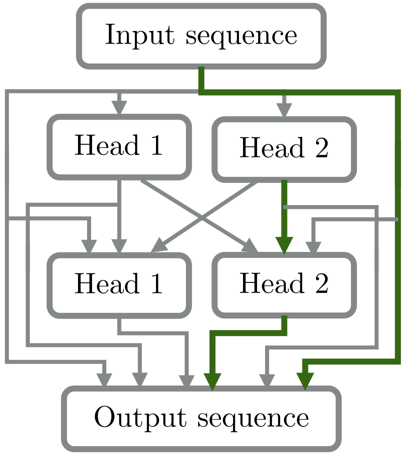
Greater than

“The war lasted from 1615 to ____”



Stigmatizing language?

“Belligerent intoxicated patient presented to the ED. Patient race is ___”



~~Goal: identify shortcut-encoding circuits~~

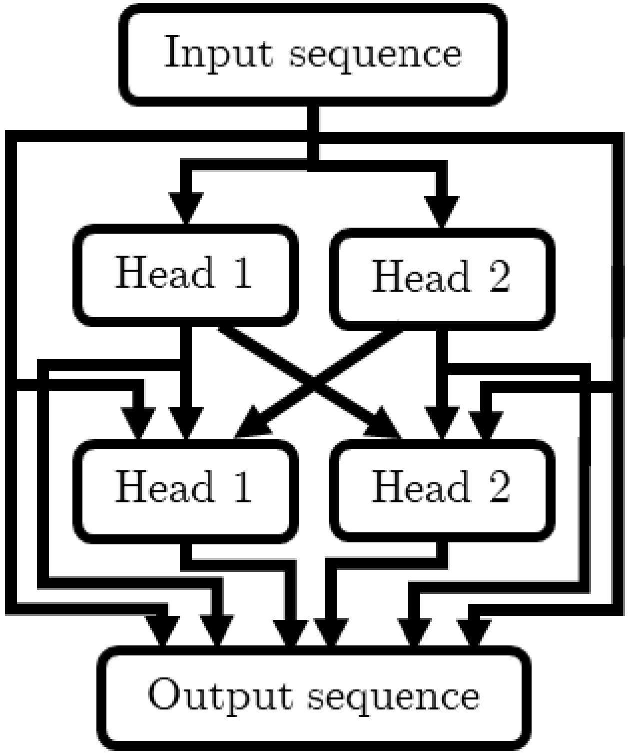
Goal: Evaluate candidate circuits

Hanna et al , NeurIPS 2024

Olah et al , Distill 2020

Setup

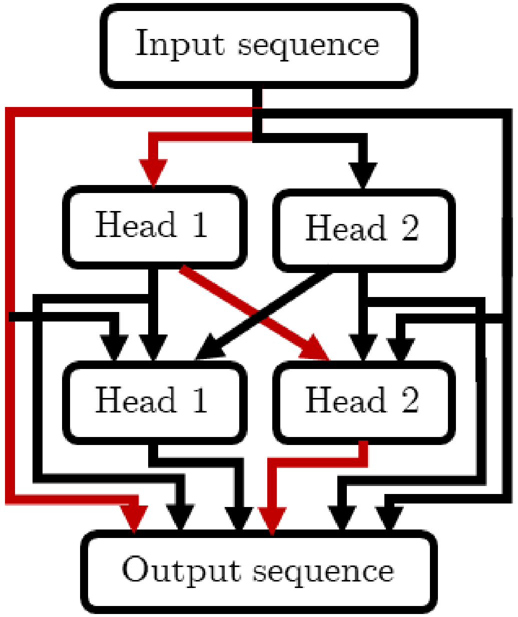
- LLM, $M(X)$: a computational graph
  - Nodes: attention heads, MLPs, input tokens and output logits
  - Edges: connections between nodes



Elhage, 2021

Setup

- Runnable circuit, $C(X)$: subgraphs of the LLM

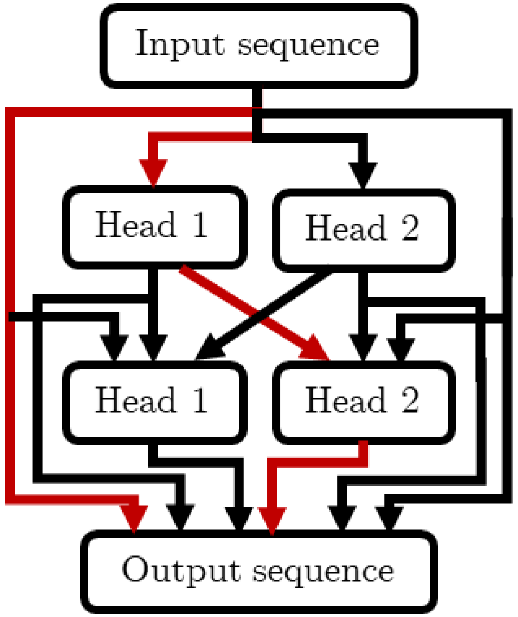


Setup

- Runnable circuit, $C(X)$: subgraphs of the LLM
- Task: $\tau = (\mathcal{D}, s)$

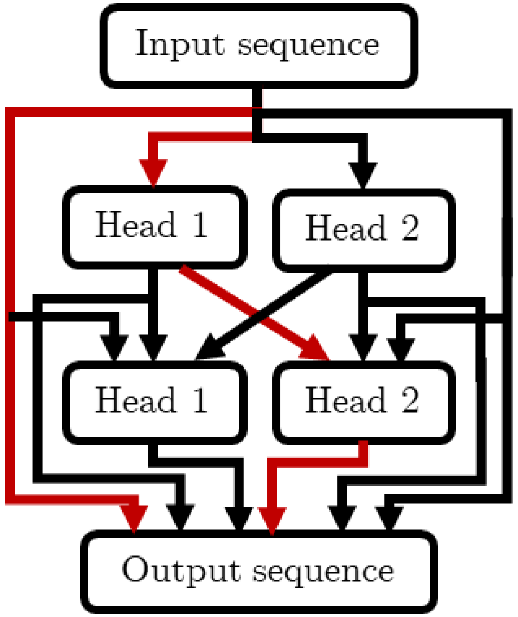
Dataset
$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$

Scoring function
$s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$



Setup

- Runnable circuit, $C(X)$: subgraphs of the LLM
- Task: $\tau = (\mathcal{D}, s)$



Dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

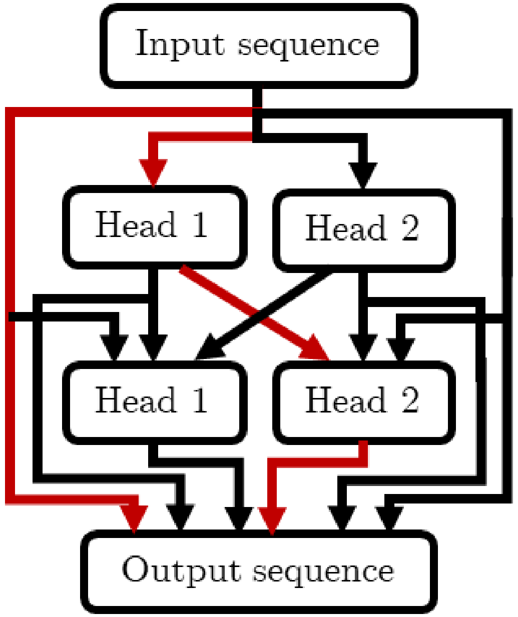
$$\left\{ \begin{array}{l} \text{The war lasted from 1615 to, 1615} \\ \text{Last election was 2022. Next is, 2022} \\ \vdots \end{array} \right\}$$

Scoring function

$$s : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$$

Setup

- Runnable circuit, $C(X)$: subgraphs of the LLM
- Task: $\tau = (\mathcal{D}, s)$



Dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

$$\left\{ \begin{array}{l} \text{The war lasted from 1615 to, 1615} \\ \text{Last election was 2022. Next is, 2022} \\ \vdots \end{array} \right\}$$

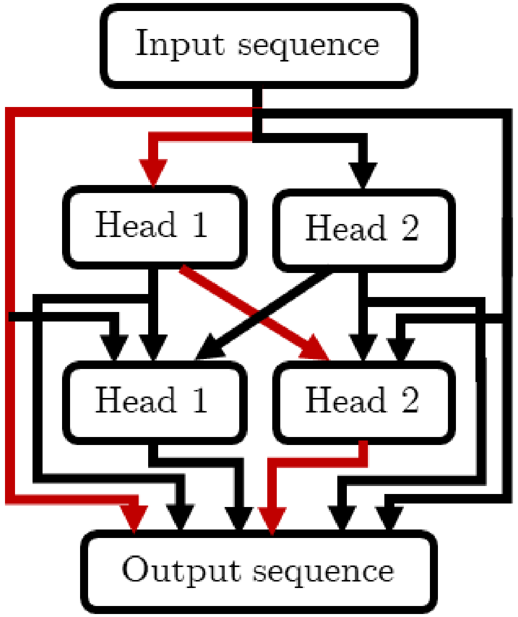
Scoring function

$$s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

$$\hat{y} = \left[\text{logit}(\hat{y}^{(1)}), \ldots, \text{logit}(\hat{y}^{(v)})\right]$$

Setup

- Runnable circuit, $C(X)$: subgraphs of the LLM
- Task: $\tau = (\mathcal{D}, s)$



Dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

$$\left\{ \begin{array}{l} \text{The war lasted from 1615 to, 1615} \\ \text{Last election was 2022. Next is, 2022} \\ \vdots \end{array} \right\}$$

Scoring function

$$s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

$$\hat{y} = \left[\text{logit}(\hat{y}^{(1)}), \ldots, \text{logit}(\hat{y}^{(v)})\right]$$

$$s(\hat{y}, y) = \sum_{i:\hat{y}_i \geq y} \text{logit}(\hat{y}_i) - \sum_{i:\hat{y}_i < y} \text{logit}(\hat{y}_i)$$

Evaluation criteria for candidate circuits

Evaluation criteria for candidate circuits

- Mechanism preservation:
    - The circuit is as good as the full LLM for the task.

Evaluation criteria for candidate circuits

- Mechanism preservation:
  - The circuit is as good as the full LLM for the task.
- Mechanism localization:
  - Removing the circuit eliminates the model's ability to perform the task.

Evaluation criteria for candidate circuits

- Mechanism preservation:
  - The circuit is as good as the full LLM for the task.
- Mechanism localization:
  - Removing the circuit eliminates the model's ability to perform the task.
- Minimality
  - All edges of a circuit are important to perform the task. No edge can be removed without hurting performance.

Evaluation criteria for candidate circuits

- Mechanism preservation:
  - The circuit is as good as the full LLM for the task.
- Mechanism localization:
  - Removing the circuit eliminates the model's ability to perform the task.
- Minimality
  - All edges of a circuit are important to perform the task. No edge can be removed without hurting performance.

Approach: recast evaluation as hypothesis testing

Evaluation criteria for candidate circuits

- Mechanism preservation:
  - The circuit is as good as the full LLM for the task.
- Mechanism localization:
  - Removing the circuit eliminates the model's ability to perform the task.
- Minimality
  - All edges of a circuit are important to perform the task. No edge can be removed without hurting performance.

Approach: recast evaluation as hypothesis testing

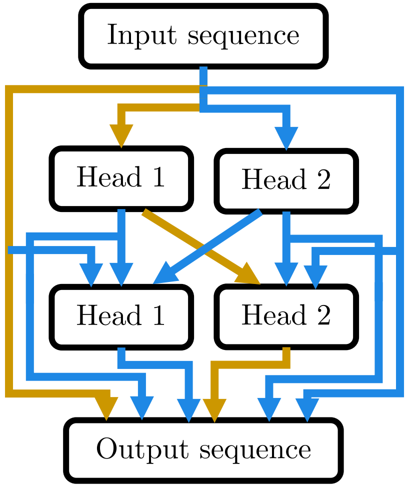
Testing mechanism localization

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

Testing mechanism localization

- If localization is achieved, “knocking out” the circuit makes the model unable to perform the task

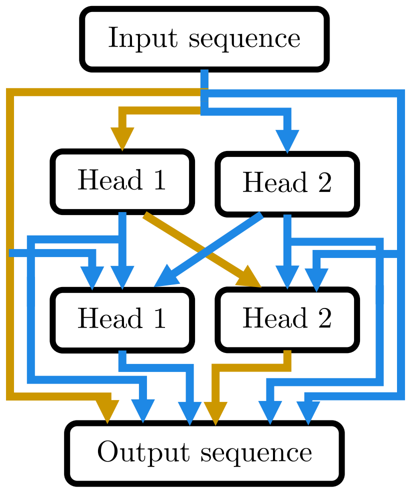


$$C(X) \cup \overline{C}(X) = M(X)$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$

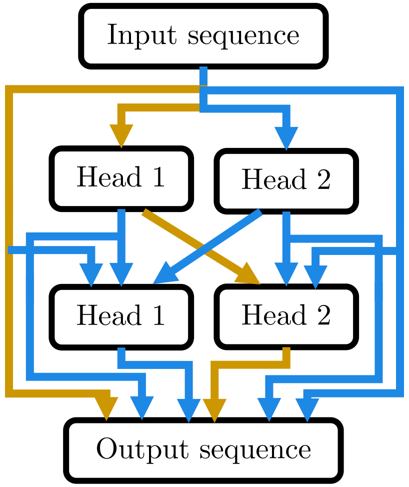


$$C(X) \cup \overline{C}(X) = M(X)$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

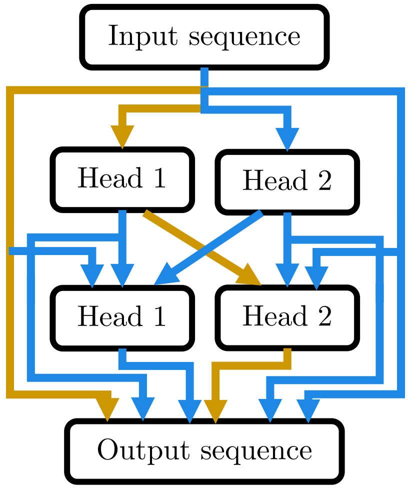
- Test statistic

$$\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M)$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

- Test statistic

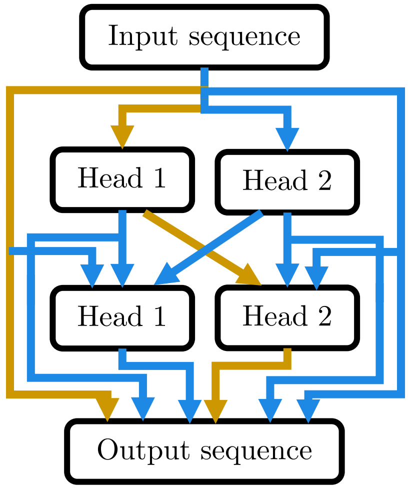
$$\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M)$$

$\overline{C}$ performance

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

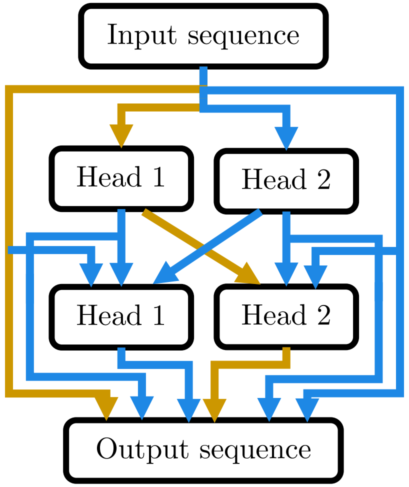
- Test statistic

$$\widehat{\mathrm{HSIC}}(\underbrace{s_{\overline{C}}}_{\overline{C}\text{ performance}}, \underbrace{s_M}_{M\text{ performance}})$$

Testing mechanism localization

- If localization is achieved, “knocking out” the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

- Test statistic

$$\widehat{\mathrm{HSIC}}(\underbrace{s_{\overline{C}}}_{\overline{C}\text{ performance}}, \underbrace{s_M}_{M\text{ performance}})$$

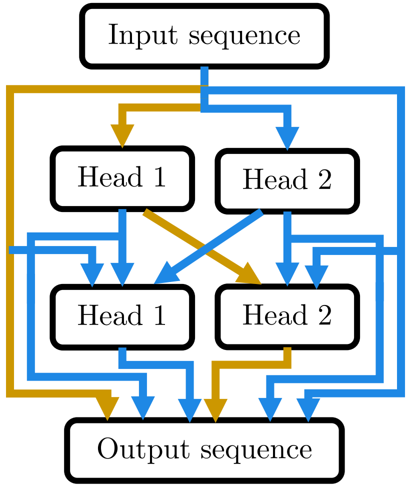
- Permutation test:

$$s_M^{\pi}$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

- Test statistic

$$\widehat{\mathrm{HSIC}}(\underbrace{s_{\overline{C}}}_{\overline{C}\text{ performance}}, \underbrace{s_M}_{M\text{ performance}})$$

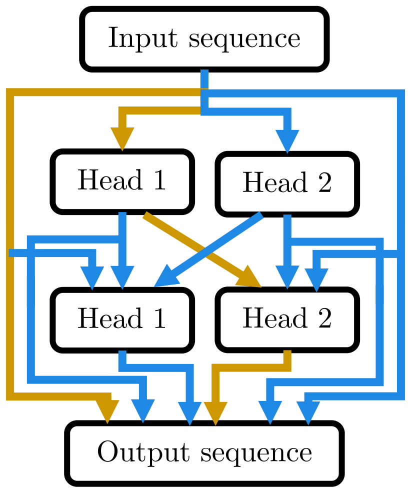
- Permutation test:

$$\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M^{\pi})$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

- Test statistic

$$\widehat{\mathrm{HSIC}}(\underbrace{s_{\overline{C}}}_{\overline{C}\text{ performance}}, \underbrace{s_M}_{M\text{ performance}})$$

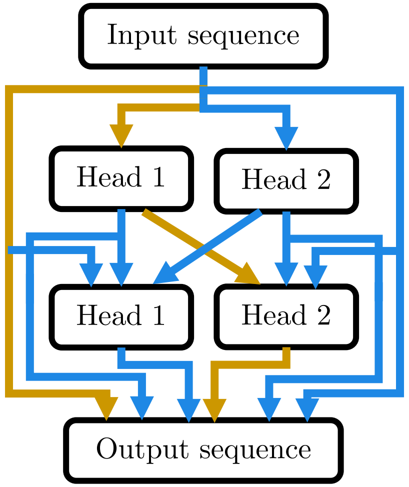
- Permutation test:

$$\mathrm{HSIC}_{\mathrm{perm}} = \{\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M^{\pi(b)})\}_{b=1}^{B}$$

Testing mechanism localization

- If localization is achieved, "knocking out" the circuit makes the model unable to perform the task

$$H_0 : s(\overline{C}(X), Y) \perp s(M(X), Y)$$



$$C(X) \cup \overline{C}(X) = M(X)$$

- Test statistic

$$\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M)$$

$\overline{C}$ performance M performance

- Permutation test:

$$\mathrm{HSIC}_{\mathrm{perm}} = \{\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M^{\pi(b)})\}_{b=1}^{B}$$

- Reject H_0 if

$$\widehat{\mathrm{HSIC}}(s_{\overline{C}}, s_M) > \text{large quantile}(\widehat{\mathrm{HSIC}}_{\mathrm{perm}})$$

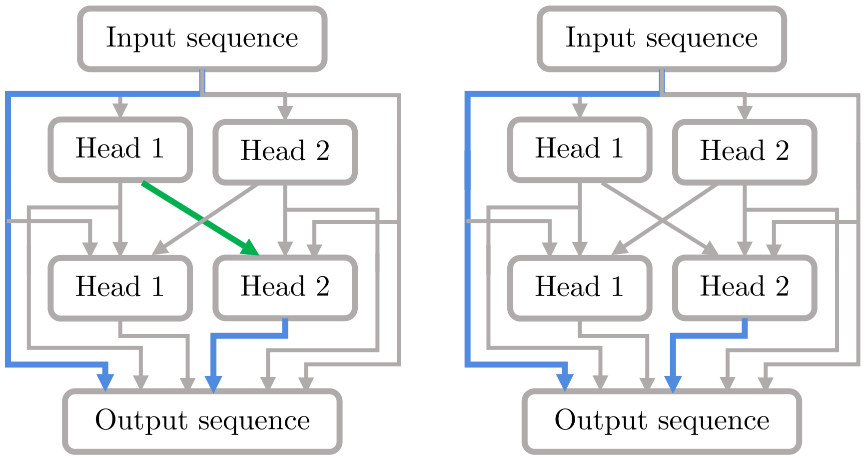
Testing minimality

Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration

Testing minimality

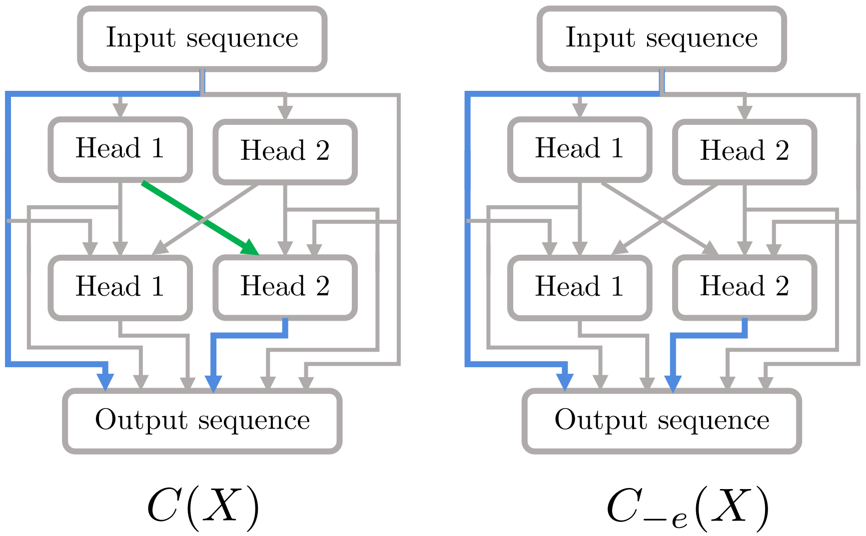
- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$



$C(X)$ $C_{-e}(X)$

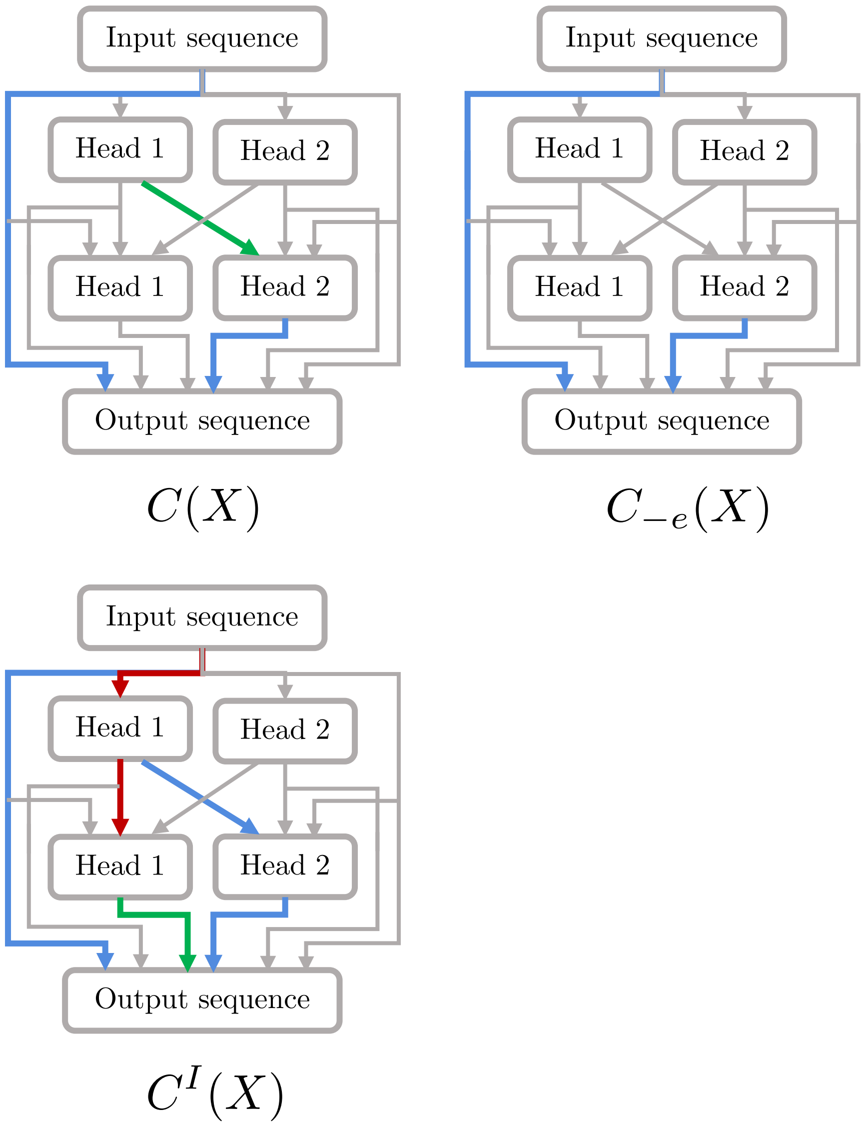
Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$
- Compare $\delta(e, C)$ to removing a truly redundant edge



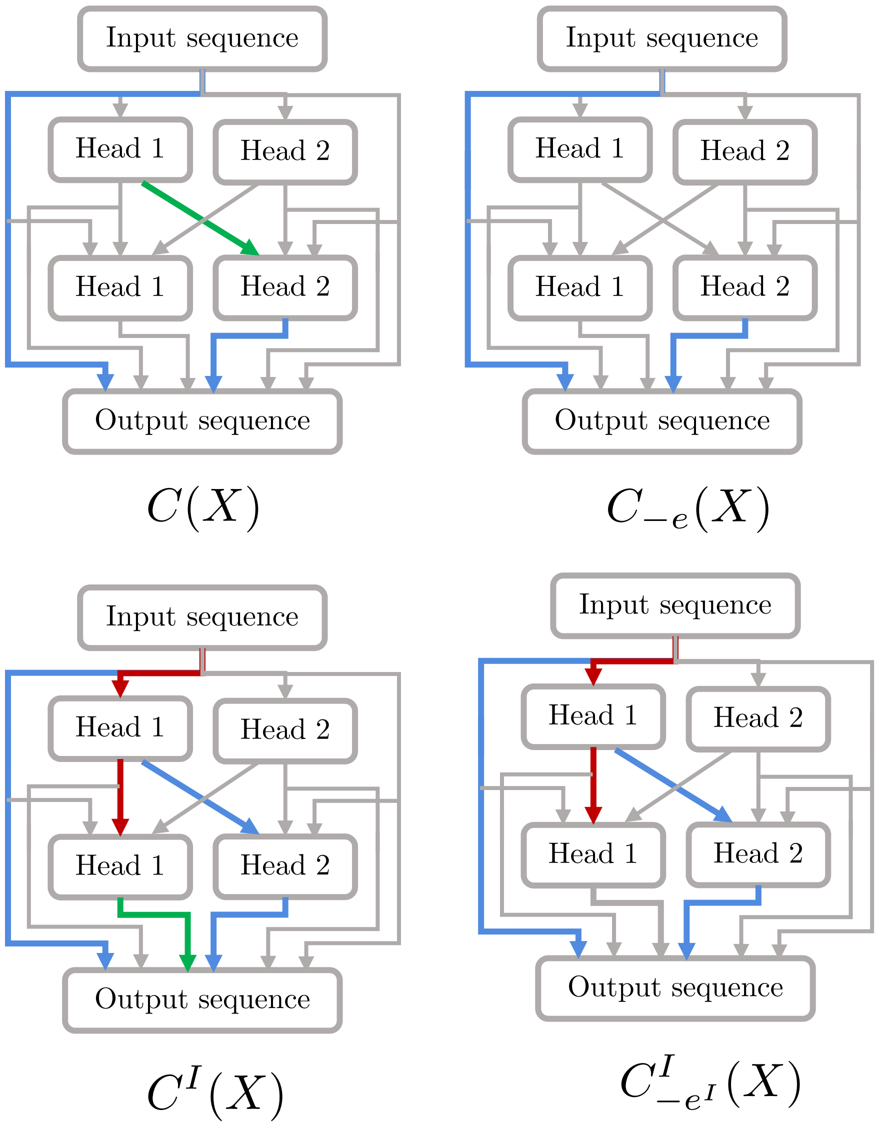
Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$
- Compare $\delta(e, C)$ to removing a truly redundant edge
- Inflate circuit with a random path to create C^I



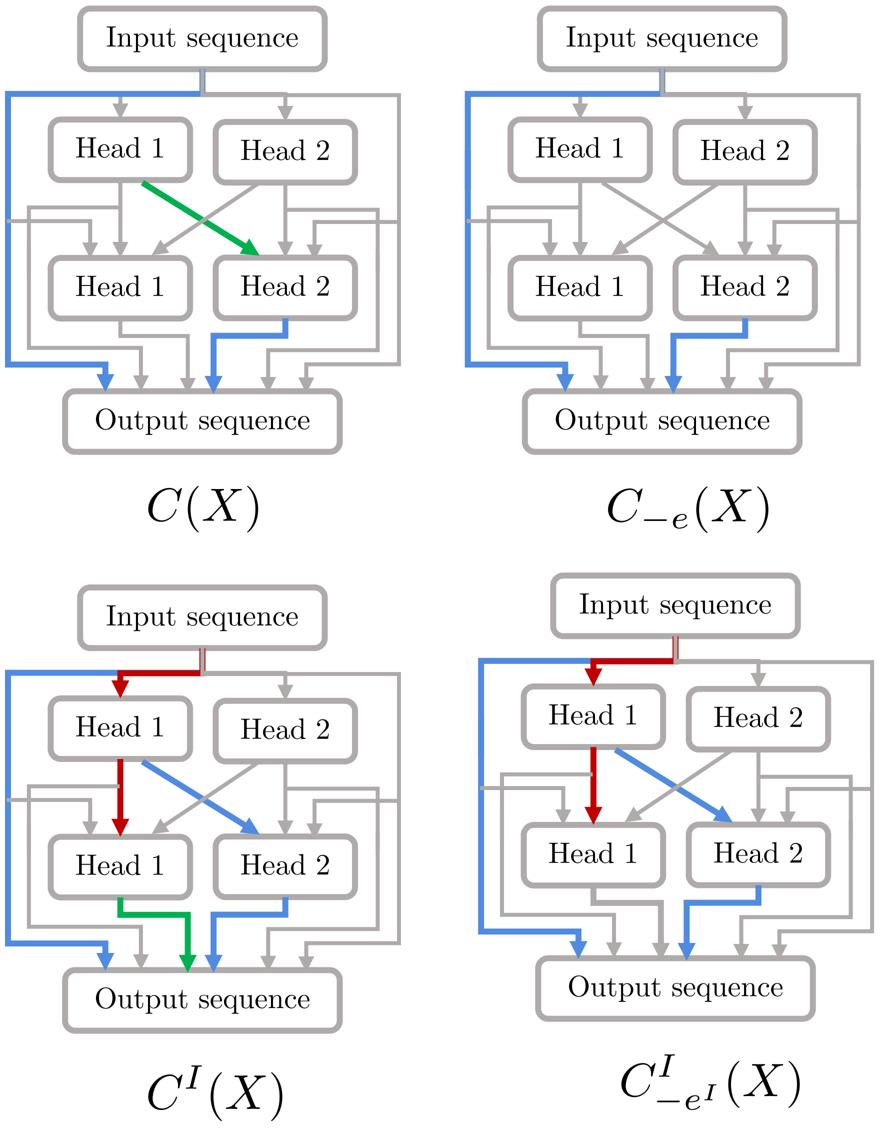
Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$
- Compare $\delta(e, C)$ to removing a truly redundant edge
- Inflate circuit with a random path to create C^I
- Compute $\delta(e^I, C^I)$



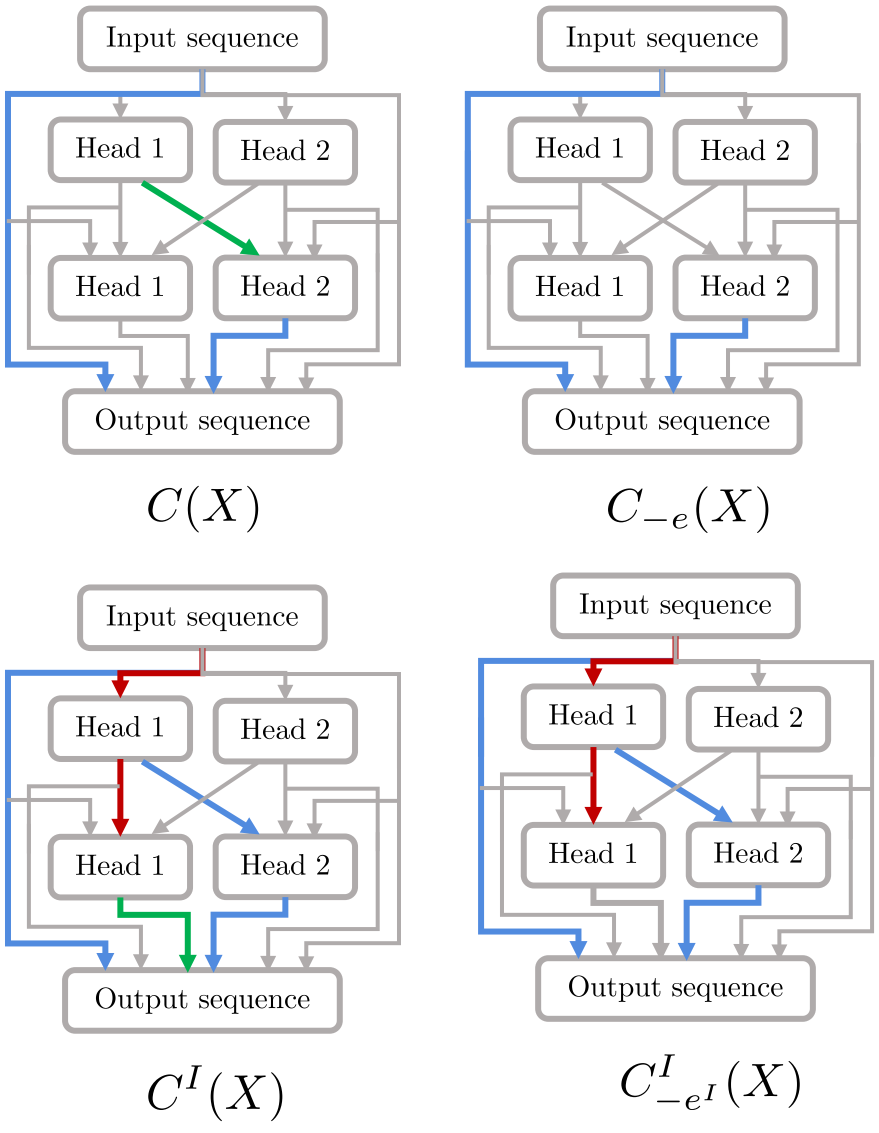
Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$
- Compare $\delta(e, C)$ to removing a truly redundant edge
- Inflate circuit with a random path to create C^I
- Compute $\delta(e^I, C^I)$
- Repeat to get $\boldsymbol{\delta} = [\delta(e^{I_1}, C^{I_1}), \ldots, \delta(e^{I_B}, C^{I_B})]$



Testing minimality

- If the circuit is minimal, removing an edge leads to *meaningful* performance deterioration
- Define $\delta(e, C) = \mathbb{E}[s(C(X), Y) - s(C_{-e}(X), Y)]$
- Compare $\delta(e, C)$ to removing a truly redundant edge
- Inflate circuit with a random path to create C^I
- Compute $\delta(e^I, C^I)$
- Repeat to get $\boldsymbol{\delta} = [\delta(e^{I_1}, C^{I_1}), \ldots, \delta(e^{I_B}, C^{I_B})]$
- Define $H_0 : \delta(e, C) >$ large quantile$(\boldsymbol{\delta})$



Experiment setup

- Synthetic circuits
  - Do our hypothesis tests work when the candidate circuit is "correct"?

Linder et al, NeurIPS 2023

Experiment setup

- Synthetic circuits
  - Do our hypothesis tests work when the candidate circuit is "correct"?
- Benchmarks:
  - Tracr-R: evaluate token reversal circuit
  - Tracr-P: evaluate "counting" circuit

Linder et al, NeurIPS 2023

Experiment setup

- Synthetic circuits
  - Do our hypothesis tests work when the candidate circuit is "correct"?
- Benchmarks:
  - Tracr-R: evaluate token reversal circuit
  - Tracr-P: evaluate "counting" circuit

Property	Tracr-P	Tracr-R
Mechanism preservation	P	P
Mechanism localization	P	P
Minimality	P	P

P = passed test, NP=did not past test

Linder et al, NeurIPS 2023

Experiment results

- Existing circuits:
  - Do our hypothesis test work “in the wild?”
  - Test on existing identified circuits

Experiment results

- Existing circuits:
  - Do our hypothesis test work "in the wild?"
  - Test on existing identified circuits
- Benchmarks:
  - Greater Than – GPT2
  - Indirect obj. identification – GPT2
    - "When Mary & John went to the store, John gave an apple to ___"

Hanna et al , NeurIPS 2024

Olsson, arXiv 2022

Experiment results

- Existing circuits:
  - Do our hypothesis test work “in the wild?”
  - Test on existing identified circuits
- Benchmarks:
  - Greater Than – GPT2
  - Indirect obj. identification – GPT2
    - “When Mary & John went to the store, John gave an apple to ___”
  - Induction – Custom model
    - “Mr George Clooney and Mrs. Amal Cloo__”

Hanna et al , NeurIPS 2024 Olsson, arXiv 2022
Olsson, arXiv 2022

Experiment results

- Existing circuits:
  - Do our hypothesis test work "in the wild?"
  - Test on existing identified circuits
- Benchmarks:
  - Greater Than – GPT2
  - Indirect obj. identification – GPT2
    - "When Mary & John went to the store, John gave an apple to ___"
  - Induction – Custom model
    - "Mr George Clooney and Mrs. Amal Cloo__"
  - Docstring – Custom model
    - Predict the next variable name in code documentation

```
def port(self, load, size, files, last):
    """oil column piece
    :param load: crime population
    :param size: unit dark
    :param files
```

Hanna et al , NeurIPS 2024 Olsson, arXiv 2022
Olsson, arXiv 2022 Heimersheim & Janiak, AI alignment forum 2023

Experiment results

- Existing circuits:
  - Do our hypothesis test work "in the wild?"
  - Test on existing identified circuits
- Benchmarks:
  - Greater Than – GPT2
  - Indirect obj. identification – GPT2
    - "When Mary & John went to the store, John gave an apple to ___"
  - Induction – Custom model
    - "Mr George Clooney and Mrs. Amal Cloo__"
  - Docstring – Custom model
    - Predict the next variable name in code documentation

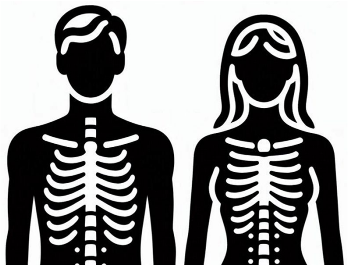
	GPT-2		Custom Model	
	G-T	IOI	Induction	DS
Mechanism preservation	NP	NP	NP	NP
Mechanism localization	NP	NP	P	NP
Minimality	NP	NP	P	P

P = passed test, NP=did not past test

```
def port(self, load, size, files, last):
    """oil column piece
    :param load: crime population
    :param size: unit dark
    :param files
```

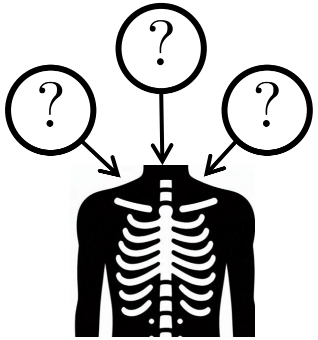
Hanna et al , NeurIPS 2024
Olsson, arXiv 2022
Olsson, arXiv 2022
Heimersheim & Janiak, AI alignment forum 2023

Causally motivated prediction



1 Efficiency + robustness to known sampling bias

MPMBHD, AIStats 22
MD, TMLR 23
N**M**, UAI 24



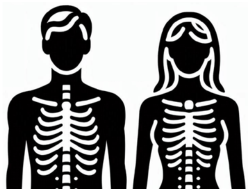
2 Efficiency + robustness to unknown sampling biases

Z**M**, NeurIPS 22
WJ**M**SW, NeurIPS 22

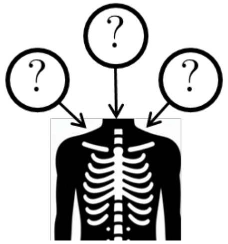
3 Evaluating localized circuits in LLMs

SVNZGJ**M**B – NeurIPS 24

Causally motivated prediction



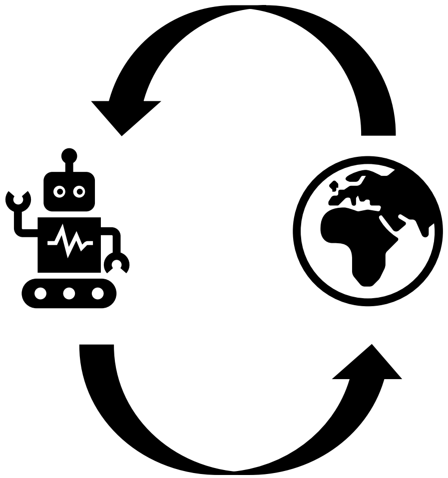
MPMBHD, AIStats 22
MD, TMLR 23
N**M**, UAI 24



Z**M**, NeurIPS 22
WJ**M**SW, NeurIPS 22

SVNZGJ**M**B – NeurIPS 24

Causally motivated reinforcement learning

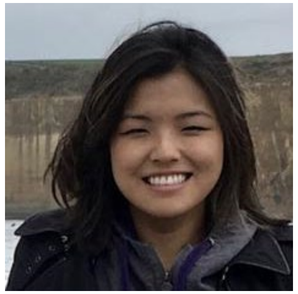


KTL**M**+, AIStats 24
T**M**+, NeurIPS 22

Causally motivated manipulation auditing

CWPP**M**W, NeurIPS 24

Thank you!





SCHMIDT **FUTURES**

NSF