DiConStruct

Causal Concept-based Explanations through Black-Box Model Distillation

Ricardo Moreira, Jacopo Bono, Mário Cardoso, Pedro Saleiro, Mário Figueiredo, Pedro Bizarro

XAI Seminar - Imperial College London

20 June 2024

feedzai

TÉCNICO LISBOA

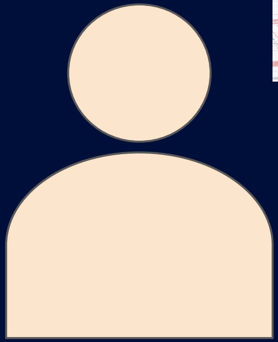
Agenda

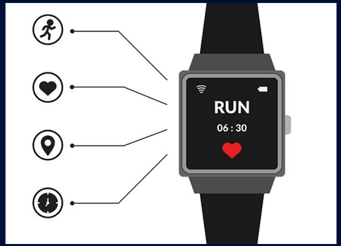
Motivation

Motivation

Heart
attack?

Motivation



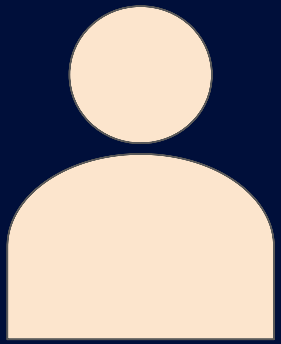


...

Heart attack?

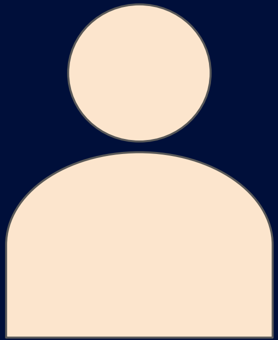
Motivation

P(Heart attack) = 0.8

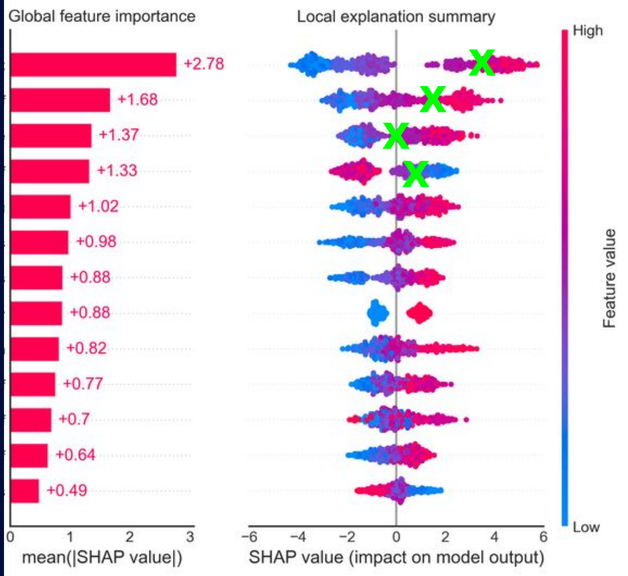


Motivation

P(Heart attack) = 0.8



LDL
BMI
glucose
Avg h active / week
...



SHAP: Lundberg et al., NeurIps 2017

Motivation

1. Feature-based explanations are often **difficult to interpret.**

Motivation

1. Feature-based explanations are often **difficult to interpret.**

Especially

- if many features
- if semantics / connection to higher level concepts is not obvious

Important when human - AI collaboration is time-sensitive!

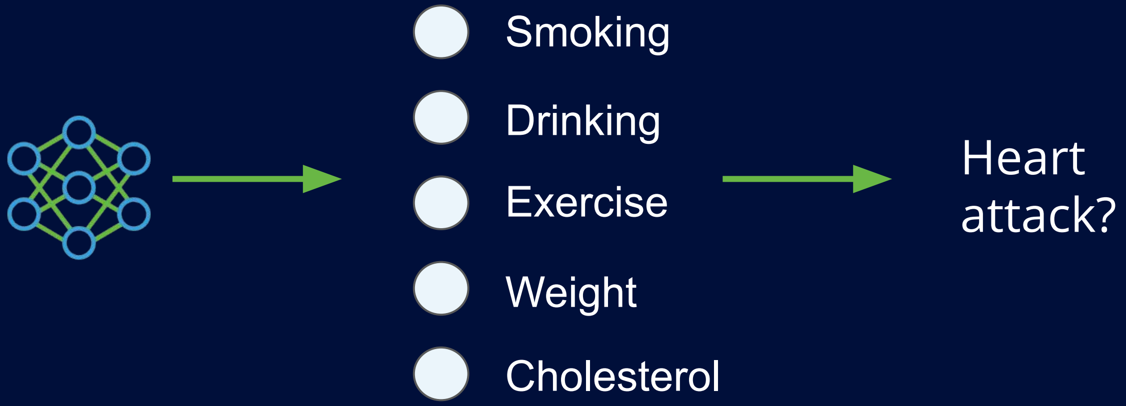
Motivation

1. Feature-based explanations are often **difficult to interpret.**

Use human understandable concepts instead.

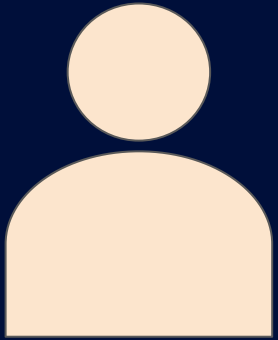
Motivation

For example, concept bottleneck models (CBM)*



* Koh et al., ICML 2020

Motivation



P(Heart attack) = 0.8

- Smoking
- Drinking
- Exercise
- Weight
- Cholesterol

High likelihood

Low likelihood

Motivation

1. Feature-based explanations are often **difficult to interpret.**
2. CBMs are self-explainable, but **trade-off** between the main task and the explanation task.

Motivation

1. Feature-based explanations are often **difficult to interpret.**
2. CBMs are self-explainable, but **trade-off** between the main task and the explanation task.
3. No rigorous **counterfactual reasoning** possible.

   (“What if I would stop smoking?”)

Motivation

1. Feature-based explanations are often **difficult to interpret.**
2. CBMs are self-explainable, but **trade-off** between the main task and the explanation task.
3. No rigorous counterfactual reasoning possible.

-> Use **post hoc** explanations

-> Incorporate **causal** principles

Motivation

Causal diagrams

- Causal relations are represented in a **DAG**.
- **Nodes** represent (endogenous) **variables**.
- **Directed edges** represent **causal relationships**.

Motivation

Structural Causal models (SCMs)

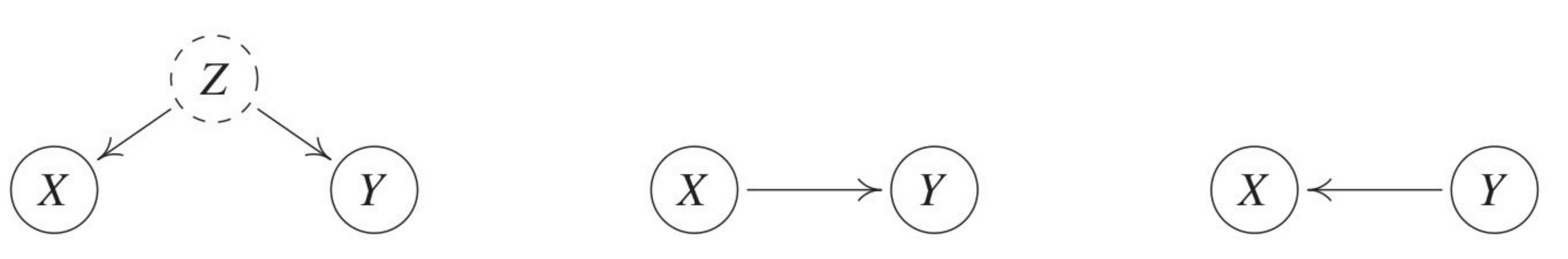
- **Exogenous** variables: effects from **outside** the model
- **Endogenous** variables: determined **within** the model
- **Structural equations**: express the relationship between the variables mathematically.

$X^i := N_X^i$

$Y^i := \alpha X^i + N_Y^i$

Motivation

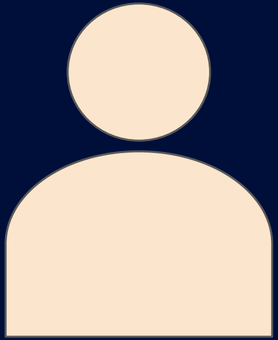
Common Cause Principle: If two random variables are statistically dependent, then there exists a variable causally influencing both.

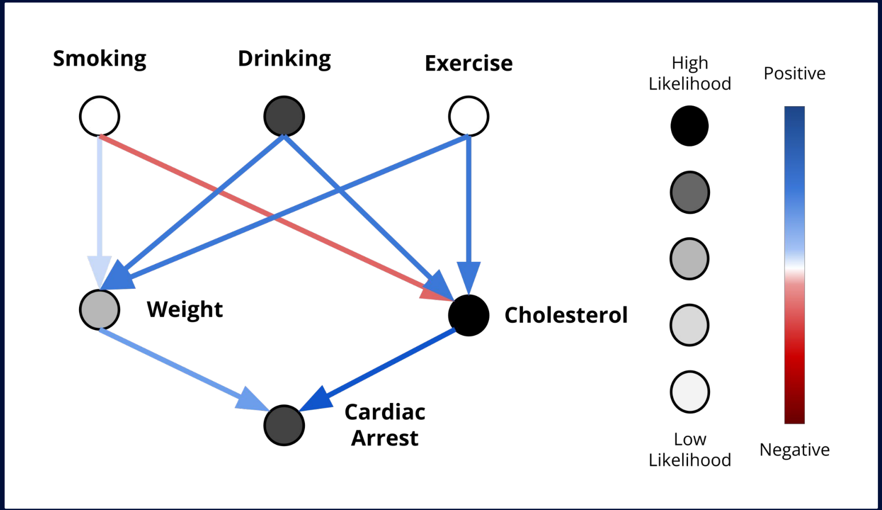


Elements of Causal Inference; Peters, Janzing, Schölkopf; 2017

Motivation

Ideally, explanations are in the form of an SCM connecting concepts,





Methods

Methods

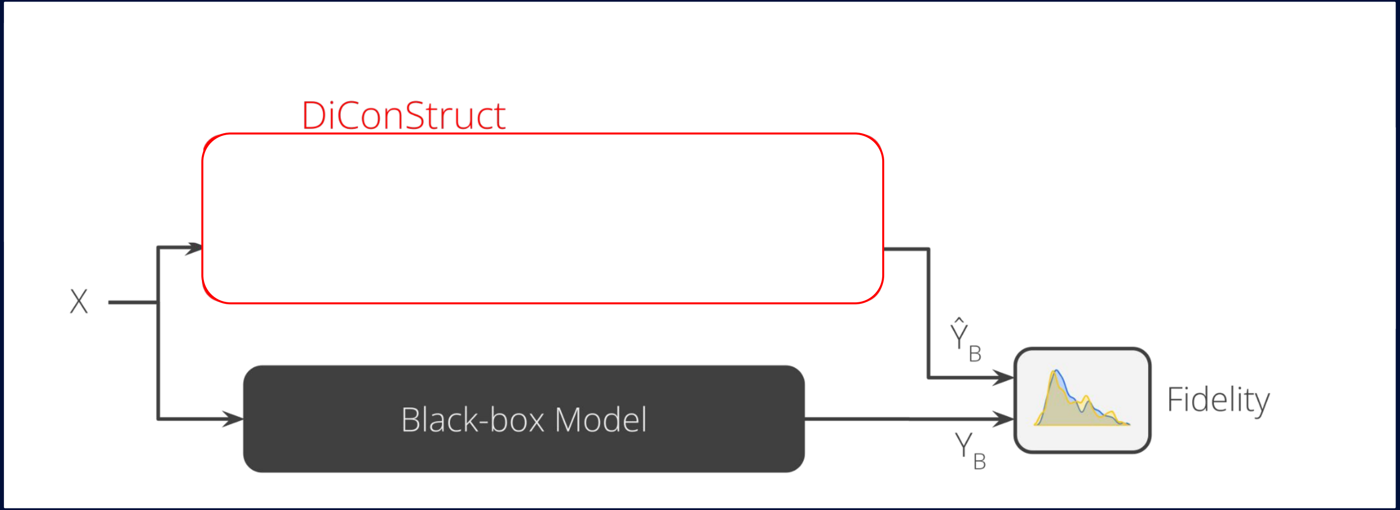
Goals:

1. Post hoc
2. Causal
3. Concept-based

feedzai

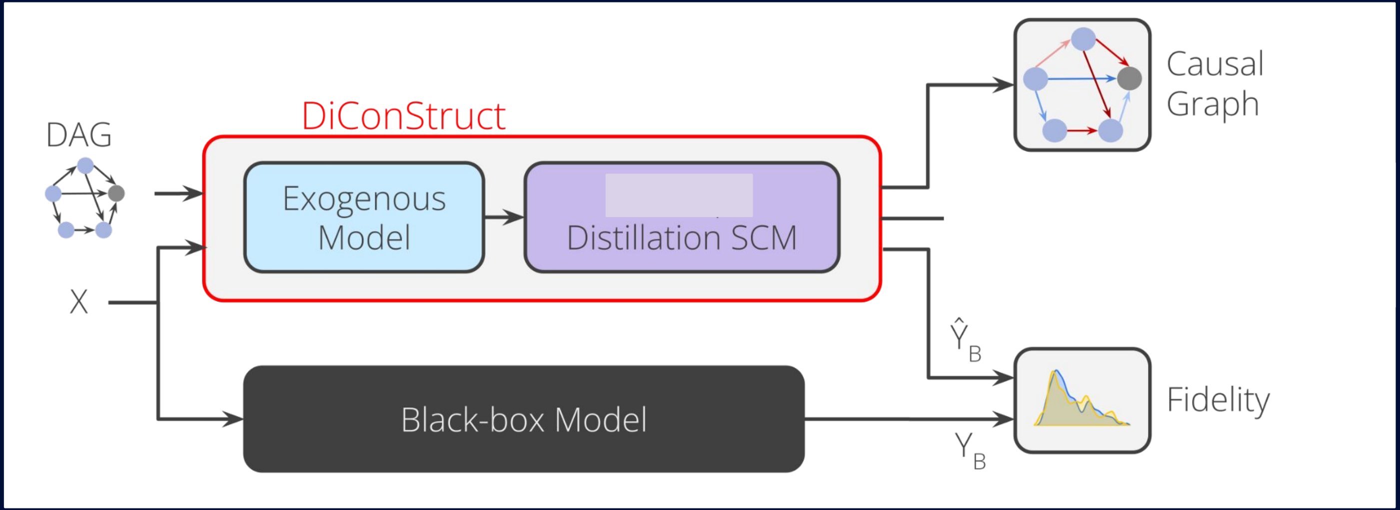
Methods

1. Post hoc: train surrogate model



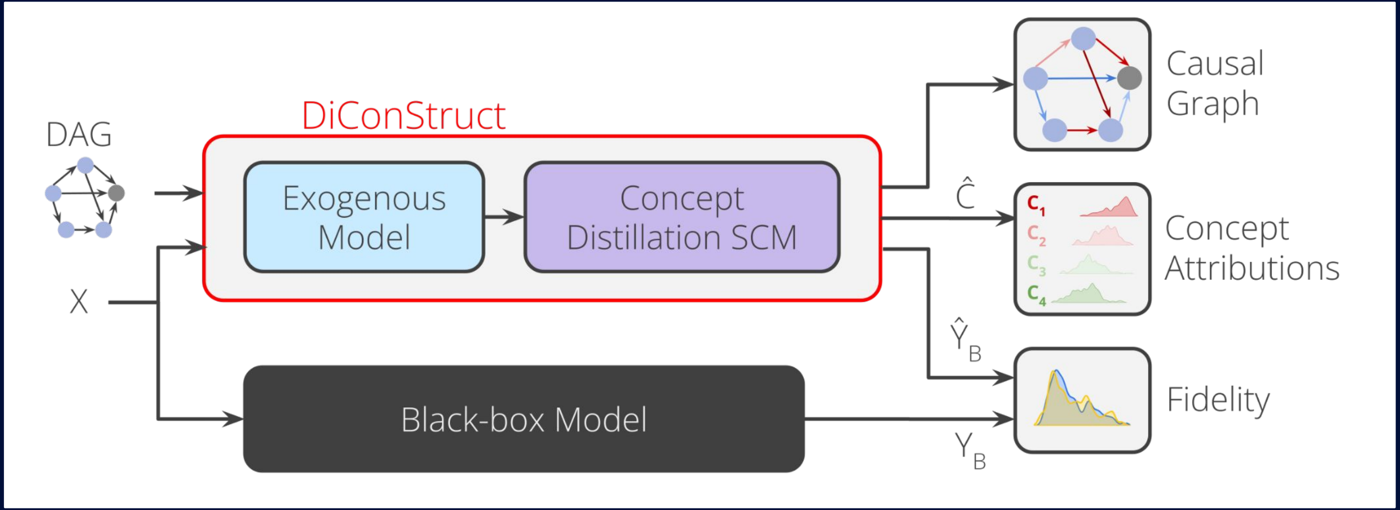
Methods

2. Causal: SCM inductive bias

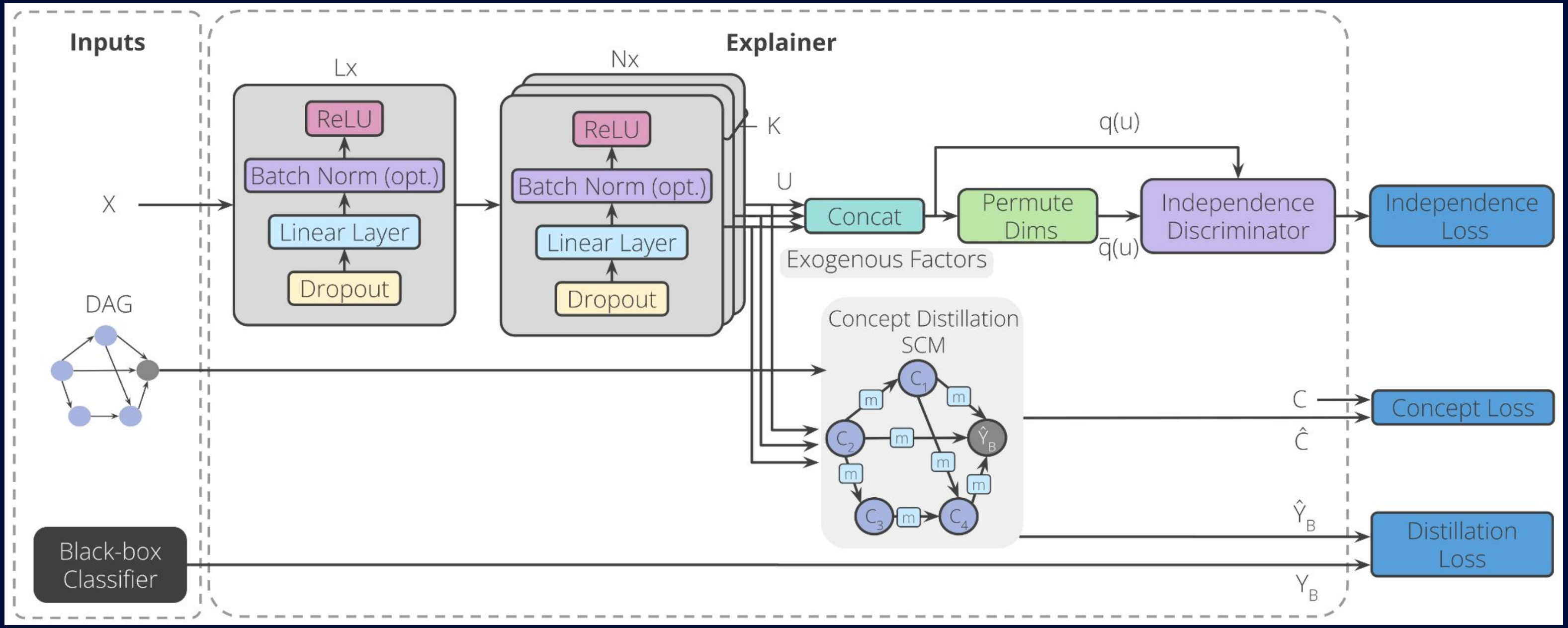


Methods

3. Concept-based

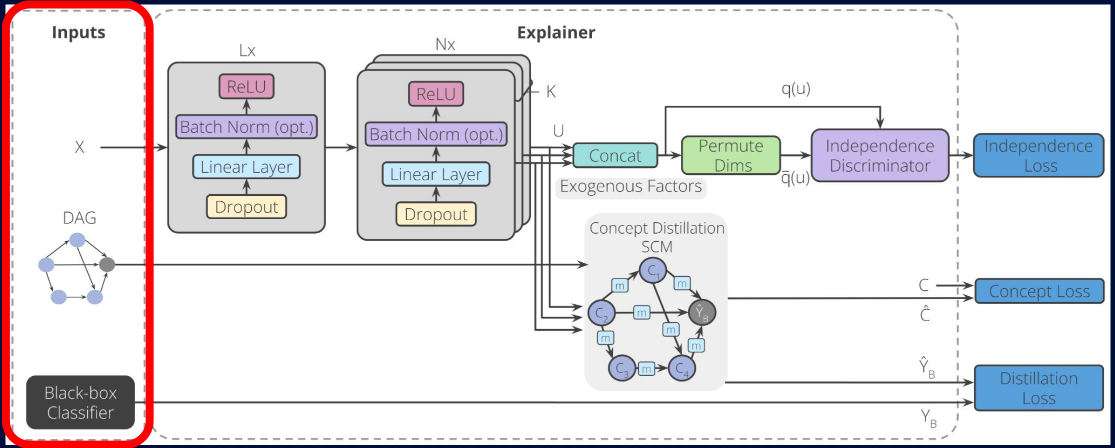


Methods



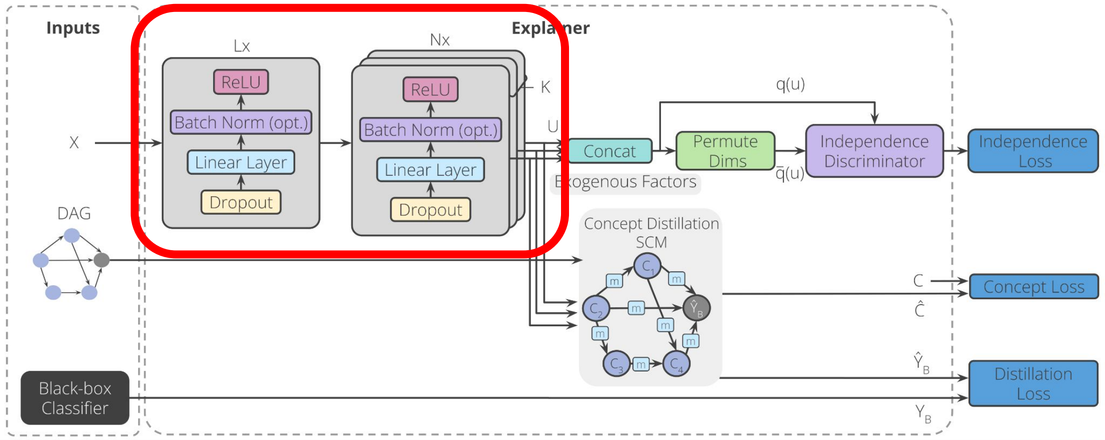
Inputs
Explainer
X
DAG
Black-box Classifier
Lx
ReLU
Batch Norm (opt.)
Linear Layer
Dropout
Nx
ReLU
Batch Norm (opt.)
Linear Layer
Dropout
K
U
Concat
Exogenous Factors
Permute Dims
q(u)
q̄(u)
Independence Discriminator
Independence Loss
Concept Distillation SCM
C
Ĉ
Concept Loss
Ŷ_B
Y_B
Distillation Loss

Methods



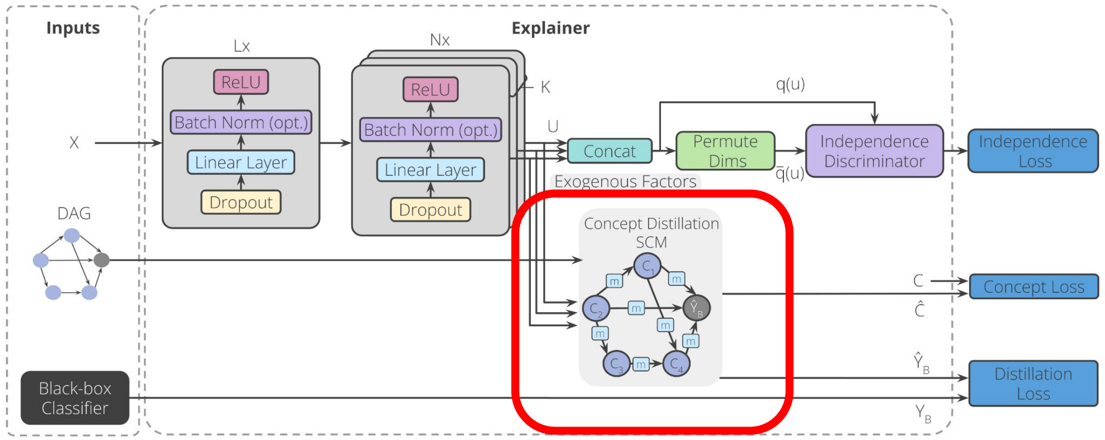
- Concepts and DAG are known a priori.
- We assume:
  - Exogenous independence
  - Concept completeness

Methods



- **Exogenous model** (outputs exogenous variables u_k for each concept c_k).
    - L common neural network layer blocks
    - N concept-specific neural network layer blocks.

Methods



- **SCM**
  - **Edges:**

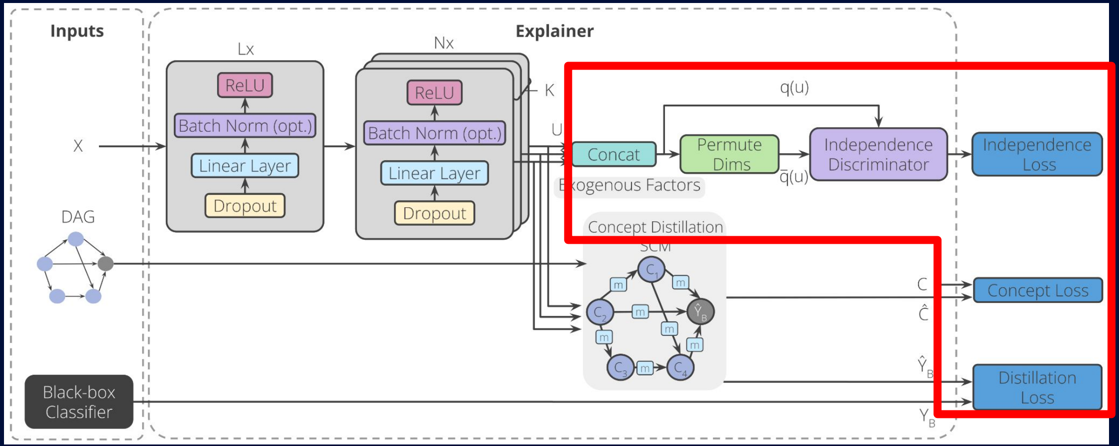
    global: $m_{j,k}(\hat{c}_j) = \sigma^{-1}(\hat{c}_j) w_{j,k}$

    local: $m_{j,k}(\hat{c}_j, \boldsymbol{u}, y_B) = \sigma^{-1}(\hat{c}_j) W_{j,k}(\boldsymbol{u}, y_B)$
  - **Concepts:**

    $$\hat{c}_k = \sigma\left(b_k + \sigma^{-1}(u_k) + \sum_{j \in \mathbf{PA}_k} m_{j,k}\right)$$

$b_k,\ w_{j,k}$ $W_{j,k}$: **learnable** biases, global weights, local weight functions, respectively.

Methods



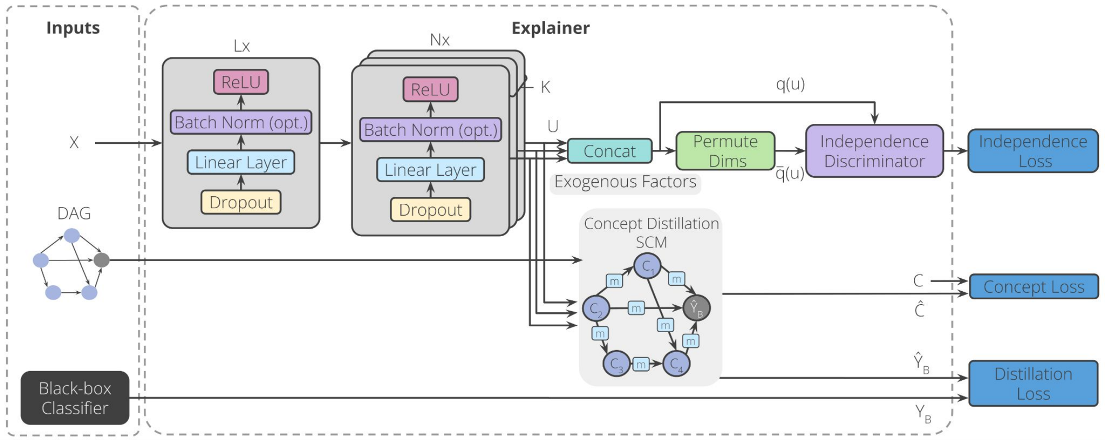
$$\mathcal{L} = \gamma \mathcal{L}_E + \beta \mathcal{L}_C + \mathcal{L}_D$$

- **Objectives:** $\mathcal{L}_E = \mathbb{E}_{q(\boldsymbol{u})}\left[\log \frac{f_I(\boldsymbol{u})}{1 - f_I(\boldsymbol{u})}\right]$ $\mathcal{L}_C = \sum_{i=1}^{M} \sum_{k=1}^{K} \sum_{c_k} c_k^{(i)} \log_2(\hat{c}_k^{(i)}),$ $\mathcal{L}_D = \sum_{i=1}^{M} \sum_{y_B} y_B^{(i)} \log_2(\hat{y}_B^{(i)}),$

Exogenous independence loss, concept loss and distillation loss, respectively.

f_I discriminates between the joint distribution of exogenous variables and the product of marginals obtained by randomly shuffling the exogenous variables.

Methods



- **Concept attributions:**

$$\mathrm{CA}^{(i)}(C_k) = \sum_{a \in \{0,1\}} \left| \mathbb{P}_{do(C_k:=a)}(Y_B^{(i)} = 1) - \mathbb{P}(Y_B^{(i)} = 1) \right|$$

Experimental Setup

Experimental setup

1. Datasets

- **CUB-200-2011***: Bird classification (binarized for the purpose of our work)

  Data comes with annotated concepts such as “eye color”, “back color”, etc.

- **Merchant fraud detection**

  Data manually (and partially) annotated by in-house analysts.
  Remaining data was pseudo-labeled by training “concept teacher” models.

  Concept examples are “high speed ordering”, “suspicious device”, etc.

*Wah et al., 2011

Experimental setup

2. Evaluation metrics

- **Main Task Performance**: Given the class imbalance, we chose to use the metric true positive rate (TPR) evaluated at a fixed false positive rate (FPR), which we set to be 5%.
- **Fidelity**: We use the 1 - MAE (mean absolute error).
- **Concept Performance**: Average accuracy over the K concepts.

Experimental setup

3. Baselines

- **CBM*:** Concept bottleneck model
- **Distillation CBM**: variation of the above, trained using the same distillation setup as DiConStruct.
- Various ablation studies on the DiConStruct components.

* Koh et al., ICML 2020

Experimental setup

4. DAGs

We obtain the causal DAG using three causal discovery methods

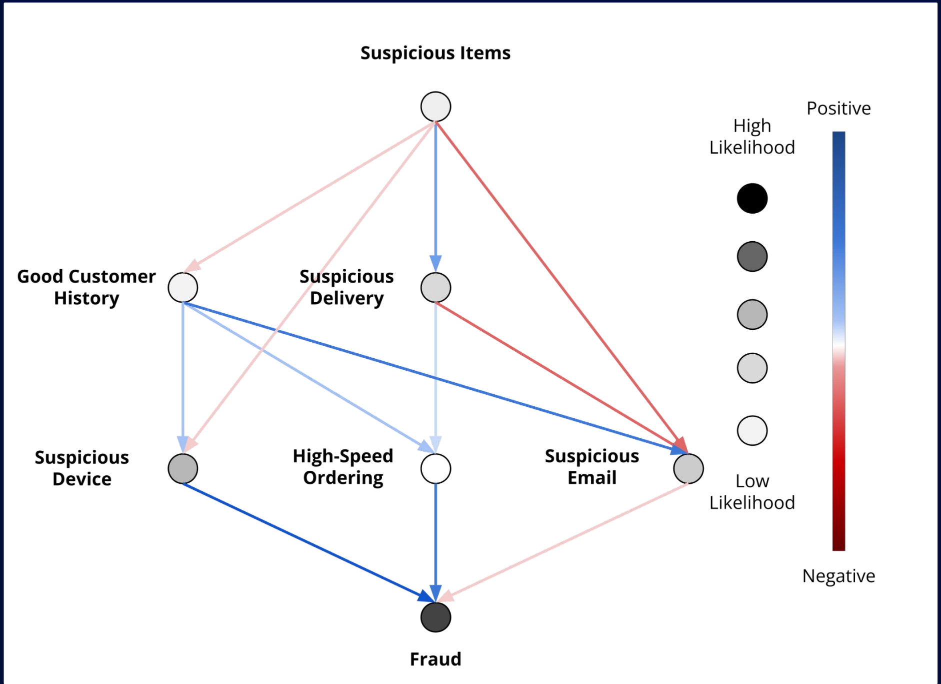
- **PC algorithm** (Sprites et al., Causation, prediction, and search, 2000)
- **ICA-LiNGAM** (Shimizu et al., JMLR,2006)
- **NO TEARS** (Zheng et al., NeurIPS, 2018)

Results

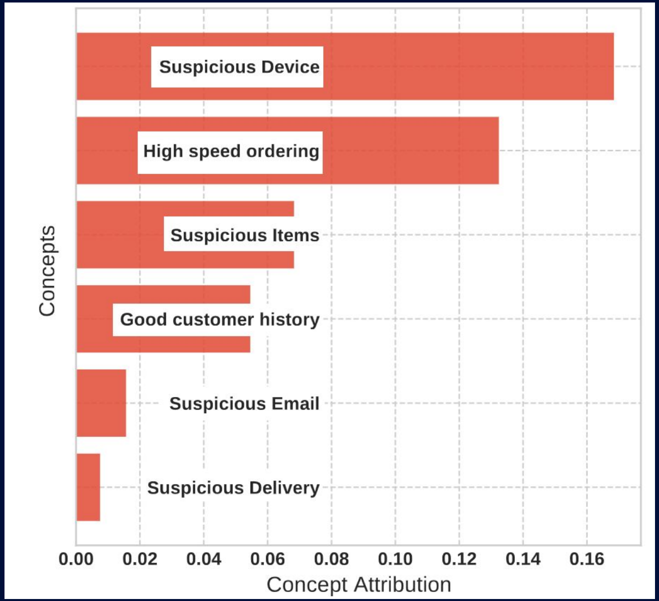
Results

	Model	Variant	Validation Task Perf. (%)	Validation Concept Perf. (%)	Validation Fidelity (%)	Test Task Perf. (%)	Test Concept Perf. (%)	Test Fidelity (%)
CUB-200-2011	Ours	Global	77.85	75.58 ± 0.65	93.52 ± 0.77	79.05	75.44 ± 0.44	94.3 ± 0.8
		Global w/ Ind.		75.61 ± 0.69	93.16 ± 0.75		75.43 ± 0.65	93.83 ± 0.64
		Local		75.25 ± 0.76	98.72 ± 0.79		75.11 ± 0.63	98.79 ± 0.74
		Local w/ Ind.		75.05 ± 1.08	98.78 ± 0.86		74.89 ± 1.19	98.83 ± 0.8
	Baselines	Joint CBM ($\lambda = 1$)	79.25 ± 0.98	75.57 ± 0.46	-	67.33 ± 2.13	75.76 ± 0.55	-
		Distill. Joint CBM ($\lambda = 1$)	77.85	75.48 ± 0.53	93.1 ± 0.52	79.05	75.52 ± 0.59	93.9 ± 0.56
		Single task - Task Perf.	77.85	-	-	79.05	-	-
		Single task - Concept Perf.	-	76.11 ± 0.21	-	-	76.07 ± 0.26	-
		Single task - Fidelity	-	-	96.07 ± 0.49	-	-	96.33 ± 0.26
Merchant Fraud - NN	Ours	Global	74.67	82.64 ± 0.14	97.12 ± 0.29	63.35	82.58 ± 0.12	96.62 ± 0.28
		Global w/ Ind.		82.6 ± 0.11	96.96 ± 0.13		82.55 ± 0.09	96.45 ± 0.24
		Local		82.5 ± 0.14	99.39 ± 0.37		82.45 ± 0.13	99.27 ± 0.42
		Local w/ Ind.		82.47 ± 0.13	99.34 ± 0.41		82.42 ± 0.12	99.23 ± 0.49
	Baselines	Joint CBM ($\lambda = 1$)	48.42 ± 0.31	82.49 ± 0.14	-	47.47 ± 3.64	82.34 ± 0.08	-
		Distill. Joint CBM ($\lambda = 1$)	74.67	82.62 ± 0.13	96.87 ± 0.18	63.35	82.57 ± 0.12	96.19 ± 0.29
		Single task - Task Perf.	74.67	-	-	63.35	-	-
		Single task - Concept Perf.	-	82.25 ± 0.19	-	-	82.25 ± 0.19	-
		Single task - Fidelity	-	-	98.13 ± 0.22	-	-	97.86 ± 0.23

Results



Suspicious Items
Good Customer History
Suspicious Delivery
Suspicious Device
High-Speed Ordering
Suspicious Email
Fraud
High Likelihood
Low Likelihood
Positive
Negative



Suspicious Device
High speed ordering
Suspicious Items
Good customer history
Suspicious Email
Suspicious Delivery
Concepts
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Concept Attribution

Conclusions

Conclusions

Key Takeaways:

We propose a **novel explainer** that is (1) **concept-based** and **causal**, (2) a **surrogate model** not affecting the predictive performance of the ML model.

Limitations and future work:

- Concept completeness assumption
- Multi-class version
- Learning of the DAG

Thank You