

The Role of Causality in Developing Trustworthy AI

Interpretability
Healthcare

A presentation by Johanna Schrader, Jonas Wallat
and Wadhah Zai El Amri

Interpretability and Causality

Why do we need **causal explanations**?

- Interpretability is often sacrificed for generalizability
- High-stake scenarios like medicine will need (and legally require) interpretability
- Causal explanations can ensure that the true reasons for a prediction are communicated
- Causality has been used to increase interpretability
 - Mainly for classifications tasks in computer vision and NLP

A standardized Evaluation

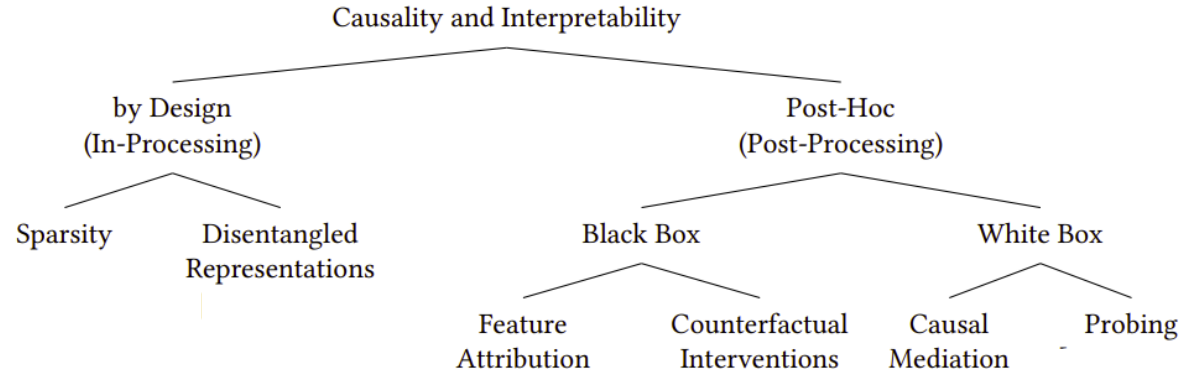
Faithfulness: How well do explanations match the actual process?

- Accurate explanations are crucial for high-stake scenarios

Causability: How well do explanations depict the causal structure of the problem?

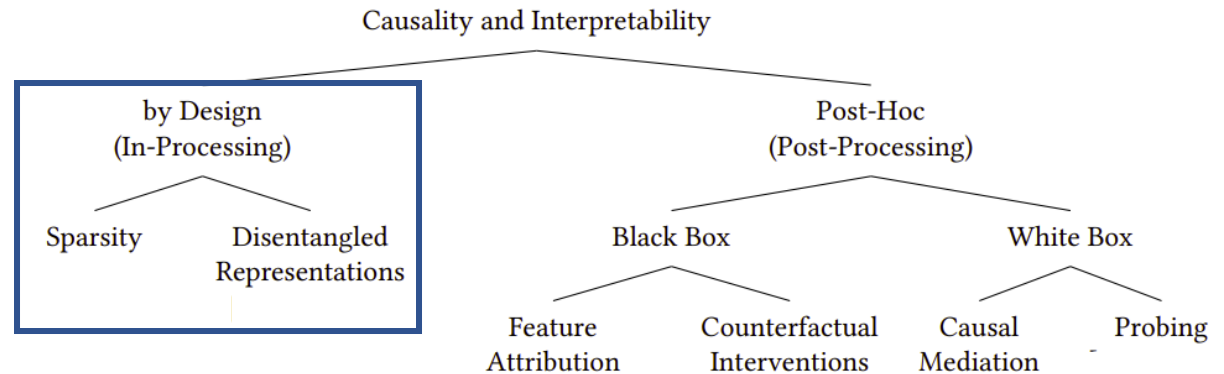
- To what extent do the explanations help humans causally understand the model's behavior?
- Helps users to build a correct mental model of a problem
- Use of the System Causability (SCS) scale

Causal Interpretability



Like traditional interpretability: some models are interpretable by their model design and some methods provide post-hoc explanations for non-interpretable models

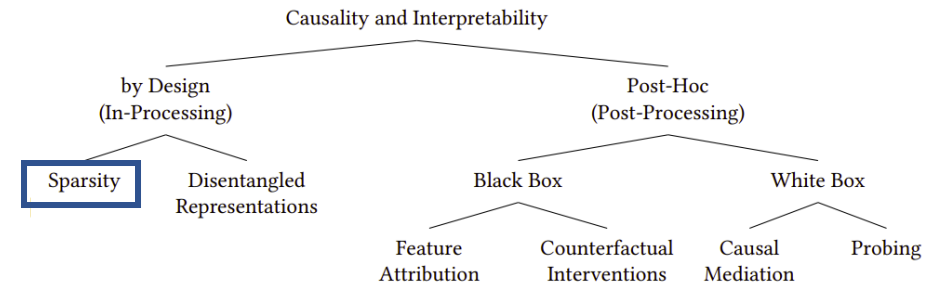
Causal Interpretability In-Processing



Like traditional interpretability: some models are interpretable by their model design and some methods provide post-hoc explanations for non-interpretable models

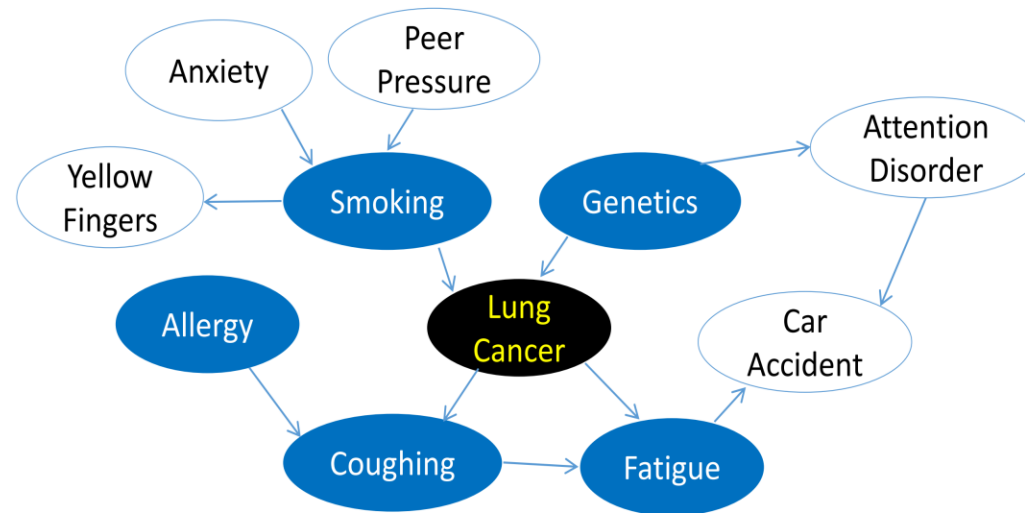
Causality-based Feature Selection: Methods and Evaluations.

[Yu2020]



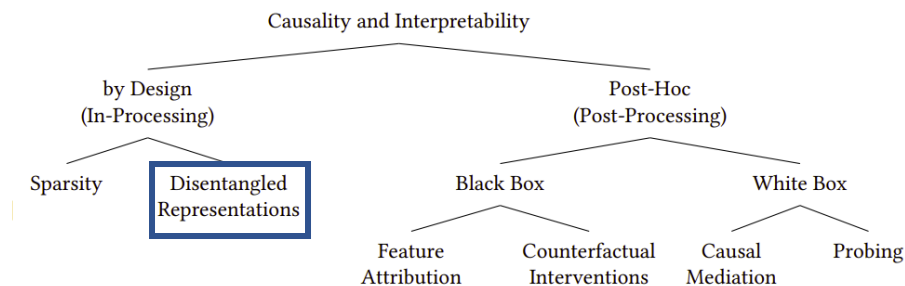
Goal: Categorize the data's attributes into relevant / irrelevant features

Solution: under the faithfulness assumption the Markov boundary of a variable in a Bayesian Network describes the variable's local causal relationships



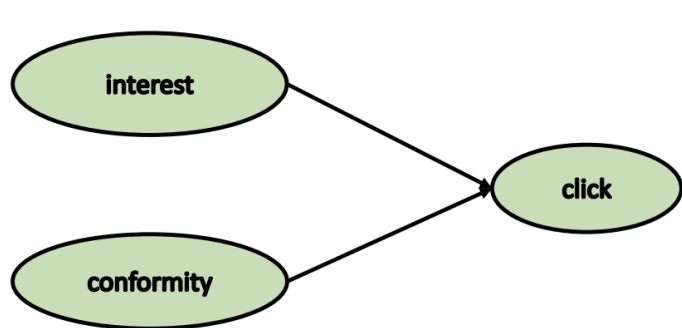
Disentangling User Interest and Conformity for Recommendation with Causal Embedding.

[Zheng2021]

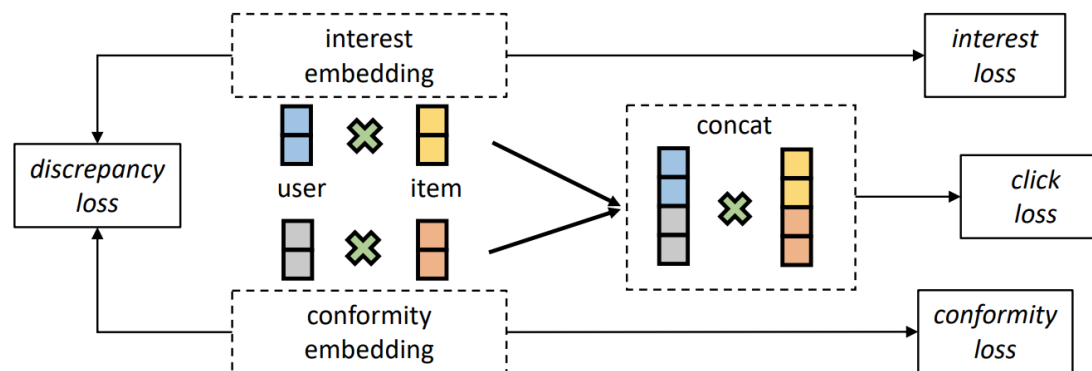


Goal: Disentangled latent representations that represent human-understandable concepts

Solution: Disentangle model representations using model architecture & cause-specific training data

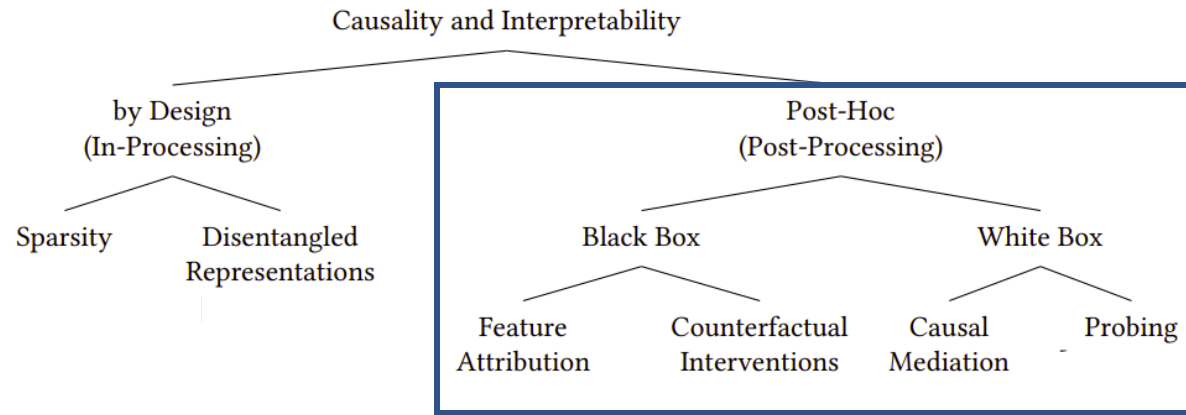


(a) Causal Graph



(b) Causal Embedding

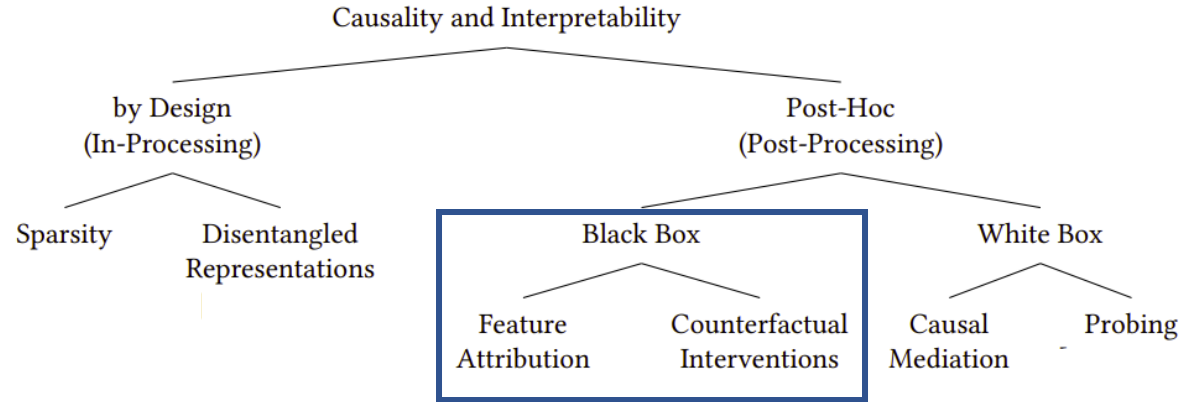
Causal Interpretability Post-Processing



Black Box: the explanation is provided only considering the models input and output without requiring access to the model's parameters

White Box: require access to the model's parameters

Causal Interpretability Post-Processing Black Box

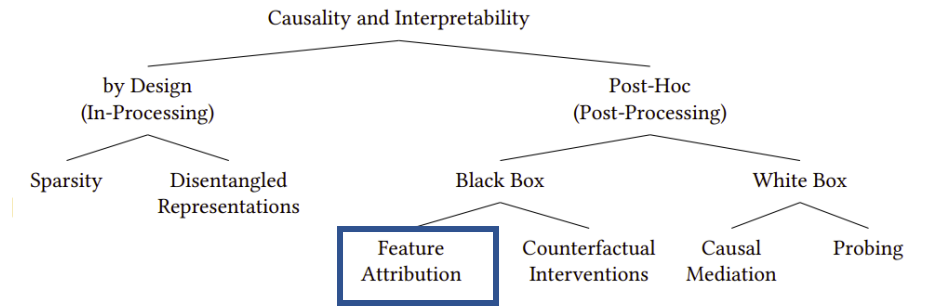


Feature Attribution: quantify the input feature's contribution to the prediction aiming at retrieving only causally relevant features

Counterfactual Interventions: model "what if" scenarios to be compared to the observed outcome

Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.

[Kim2017]

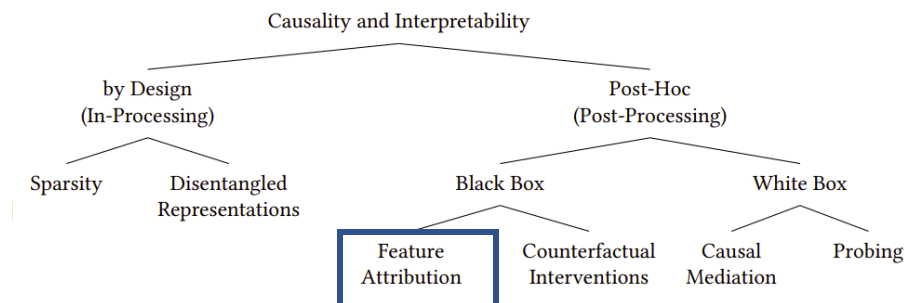


Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences

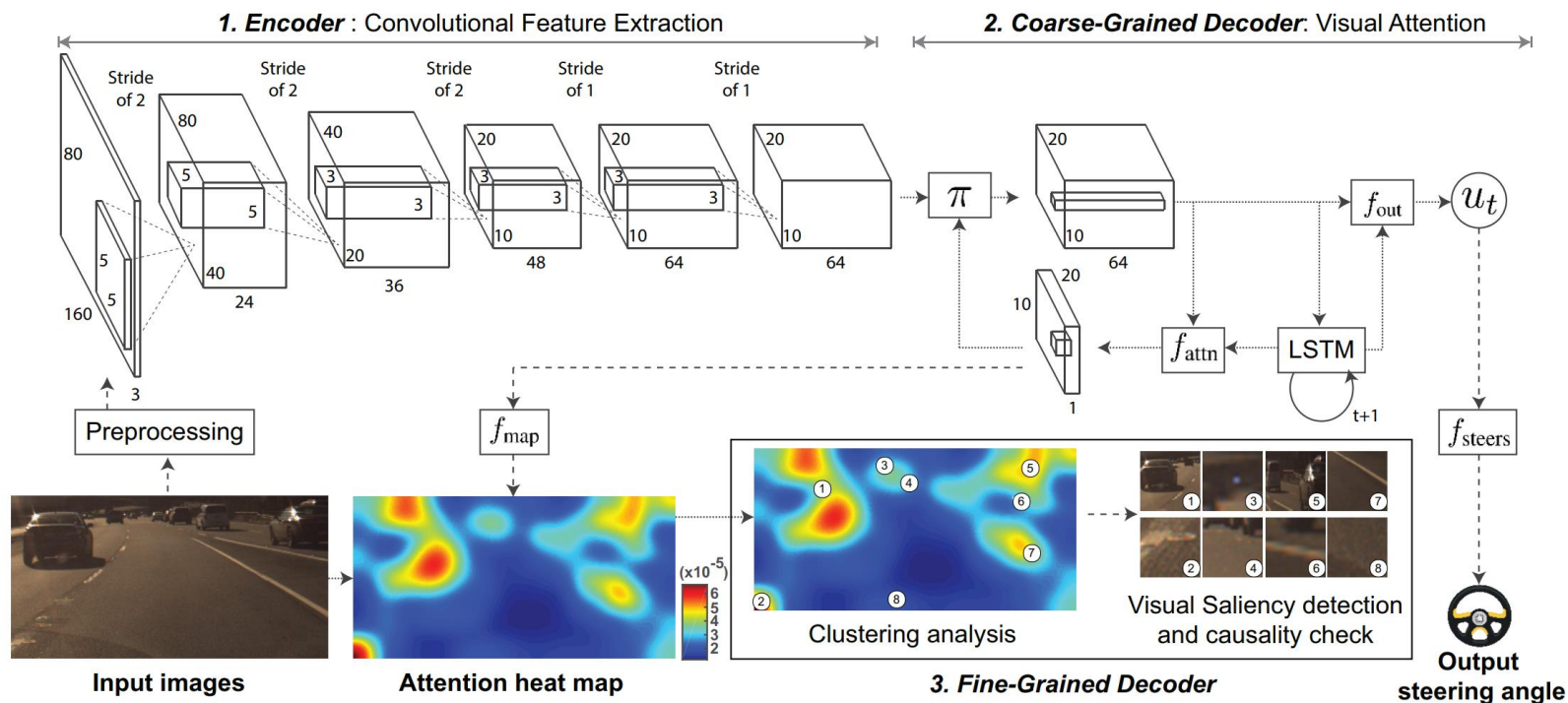
Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.

[Kim2017]



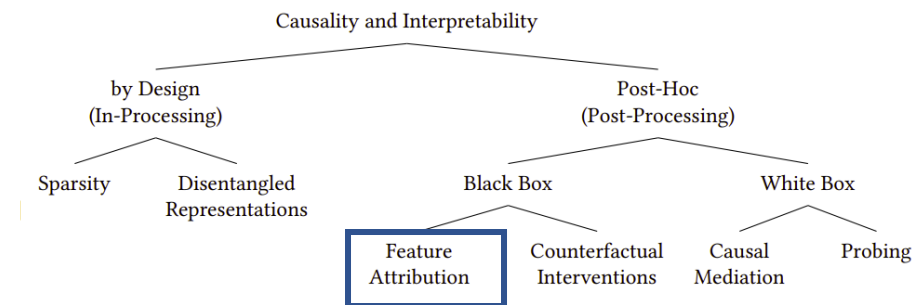
Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences



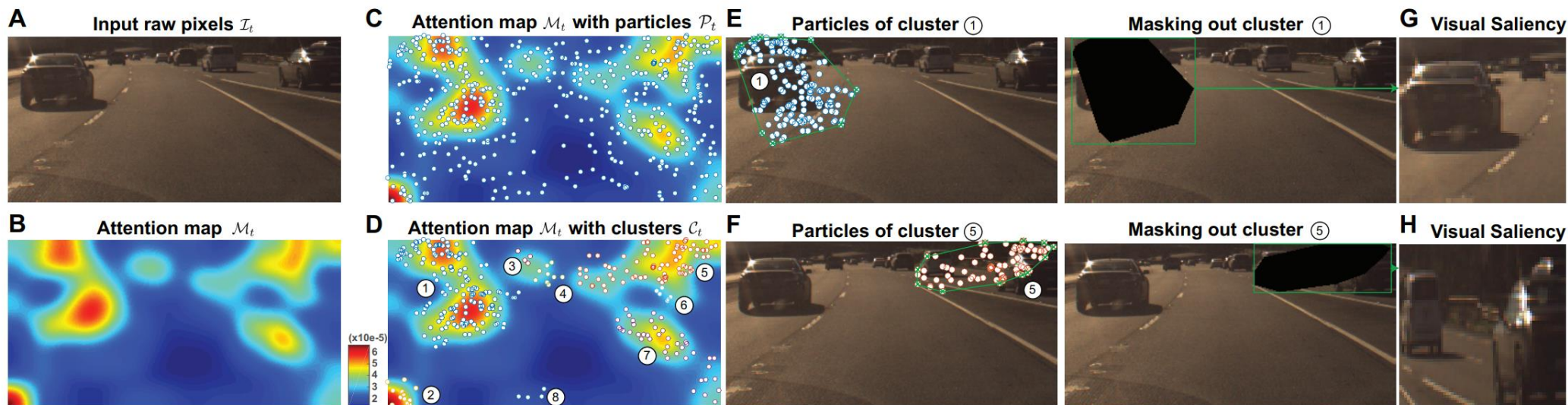
Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.

[Kim2017]

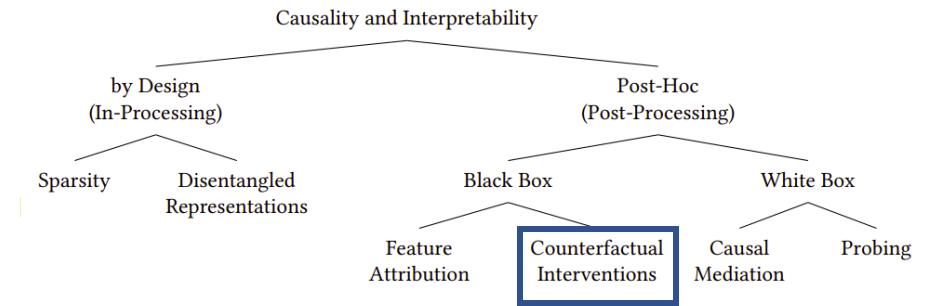


Goal: detect regions that causally influence the prediction

Solution: causal filtering by masking potential influential factors to distinguish true from spurious influences



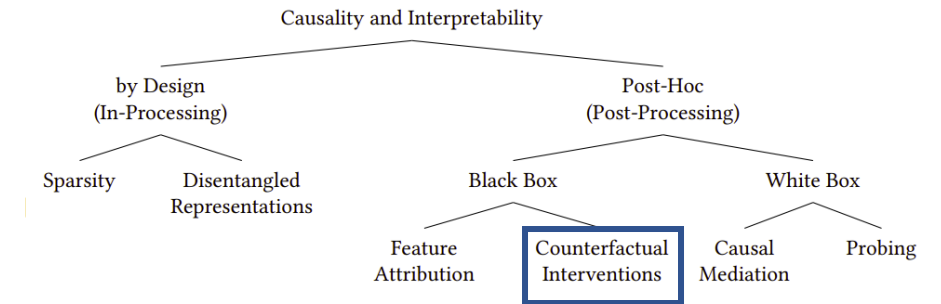
Counterfactual Explainable Recommendation. [Tan2021]



Goal: Detect attributes that could reverse an observed recommendation













Solution: Optimize for minimal changes that reverse the recommendation

Counterfactual Explainable Recommendation. [Tan2021]



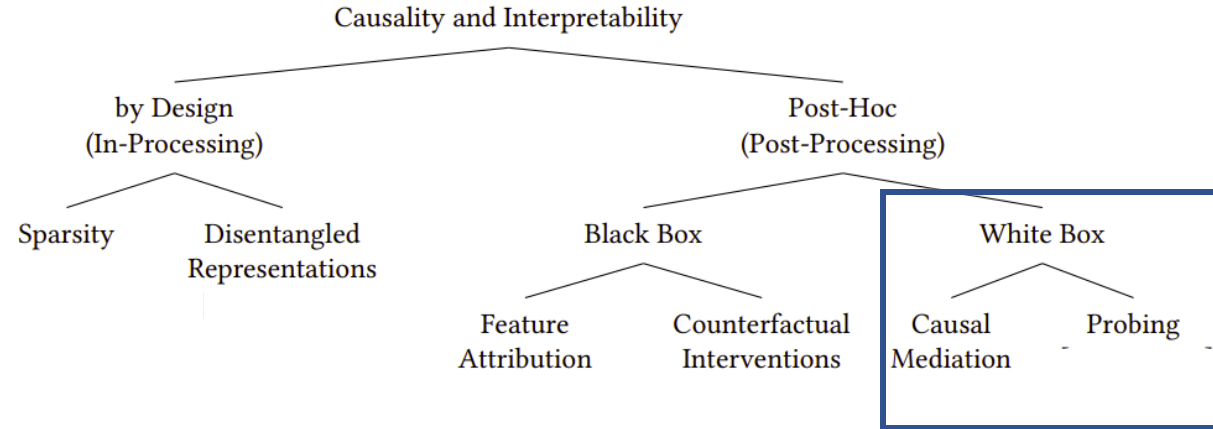
Goal: Detect attributes that could reverse an observed recommendation

Solution: Optimize for minimal changes that reverse the recommendation

	Recommended items			Not recommended items		
Matching-based	 Screen: 4.0 Battery: 5.0 Price: 3.0	 Screen: 4.5 Battery: 3.0 Price: 3.0 Phone A Score:42.00	 Screen: 4.5 Battery: 1.5 Price: 4.5 Phone B Score:39.00	 Screen: 5.0 Battery: 1.5 Price: 3.5 Phone C Score:38.00	 Screen: 5.0 Battery: 0.5 Price: 4.0 Phone D Score:34.50	 Screen: 5.0 Battery: 1.0 Price: 3.0 Phone E Score:34.00
vs.						
Counterfactual	 Screen: 4.0 Battery: 5.0 Price: 3.0	 Screen: 4.5 Battery: 1.5 Price: 4.5 Phone B Score:39.0	 Screen: 5.0 Battery: 1.5 Price: 3.5 Phone C Score:38.0	 Screen: 4.5 Battery: 2.1 Price: 3.0 Phone A* Score:37.50	 Screen: 5.0 Battery: 0.5 Price: 4.0 Phone D Score:34.50	 Screen: 5.0 Battery: 1.0 Price: 3.0 Phone E Score:34.00

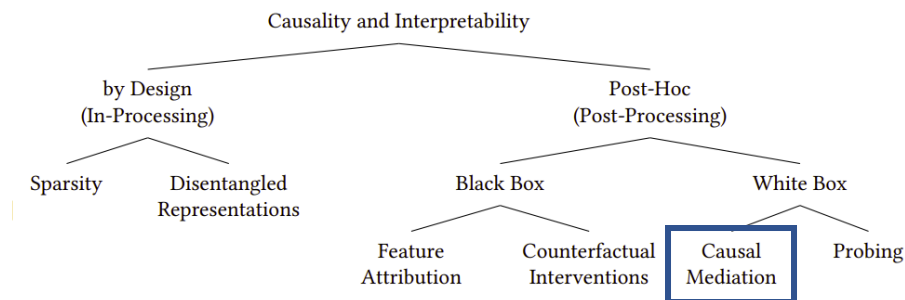
If the item had been slightly worse on [aspect(s)], then it will not be recommended.

Causal Interpretability Post-Processing White Box



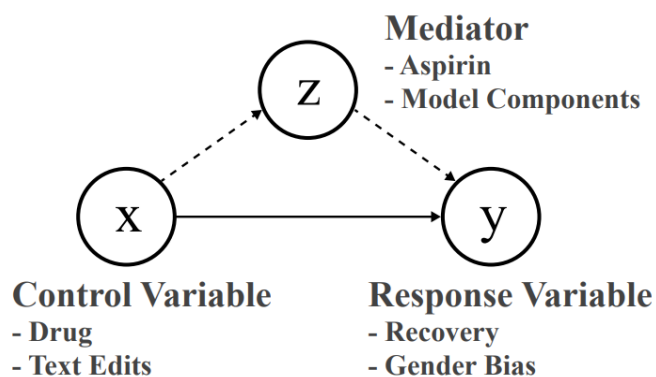
Investigating Gender Bias in Language Models Using Causal Mediation Analysis.

[Vig2020]



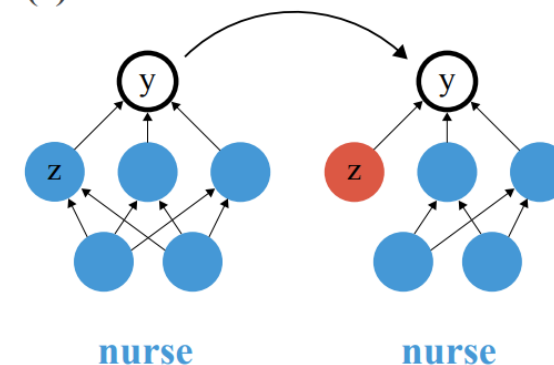
Goal: Find parts of language models that exhibit gender biases.

Solution: Causal Mediation Analysis on attention heads and neurons/layers.



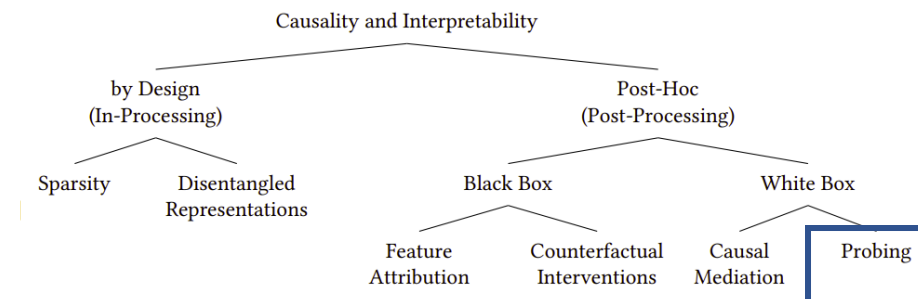
Prompt u : The nurse said that ____
Stereotypical candidate: she
Anti-stereotypical candidate: he

(c) Indirect Effect



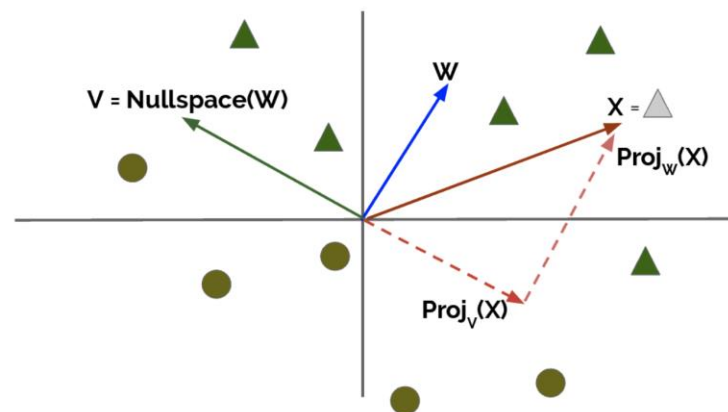
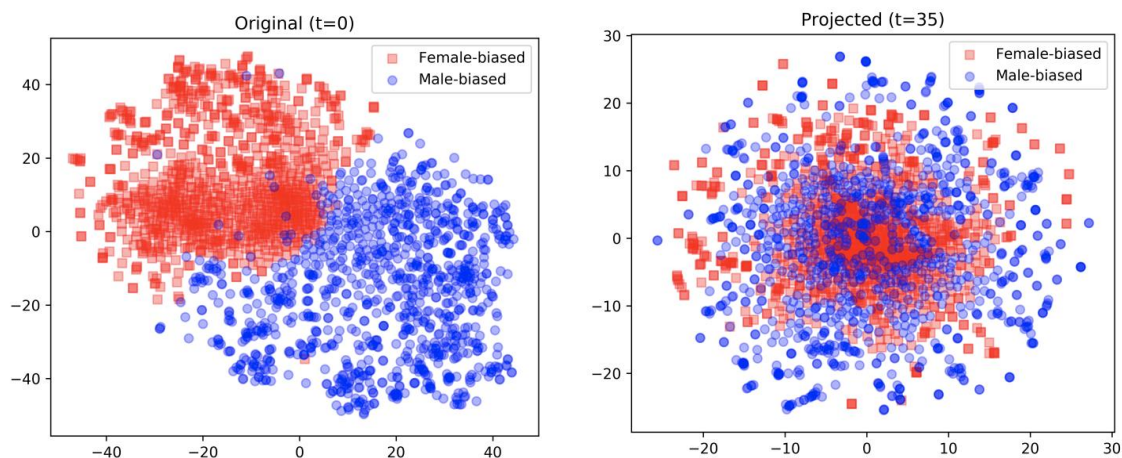
Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection.

[Ravfogel2020]



Goal: Classifiers are biased -> Build counterfactual embeddings without a certain concept.

Solution: Train classifiers, remove the information used by the classifier.

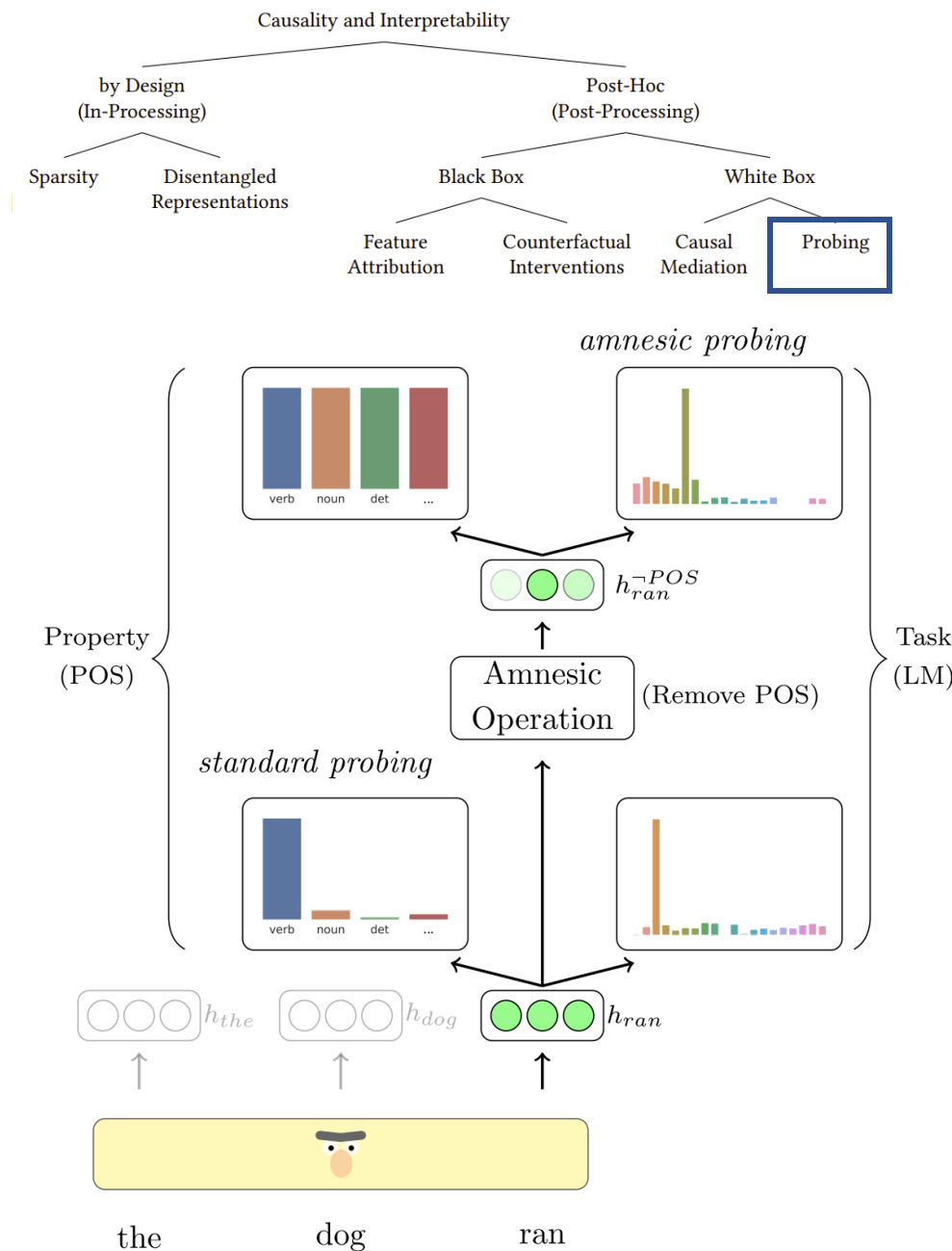


Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals.

[Elazar2020]

Goal: Investigate the effect of certain concepts (e.g., gender information/POS) on downstream tasks.

Solution: Remove information from embeddings and measure downstream task performance.



Conclusion

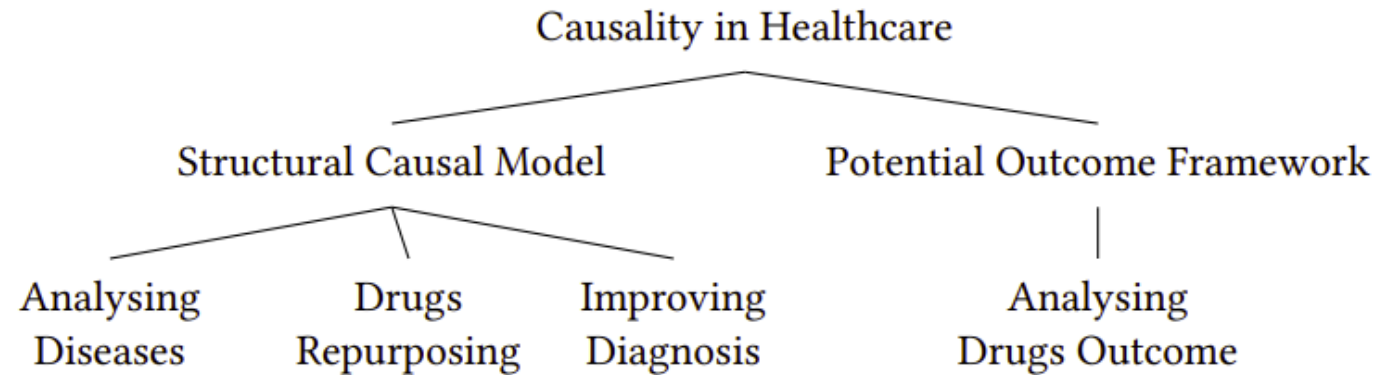
- Influx of papers
- Causality-inspired
- Neither causality nor interpretability parts are solved
- SCMs are built, evaluation is lacking
- Need for user studies

Healthcare

Why do we need **causality** in **healthcare** domain?

- Analysing diseases
- Drugs repurposing
- Improving diagnosis
- Analysing drugs outcome

Healthcare

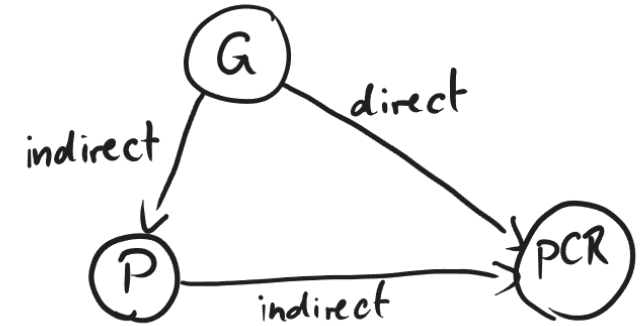


- Two main frameworks are used for causality:
 - Structural Causal Model (SCM)
 - Potential Outcome Framework (PO)

Causality Frameworks

Structural Causal Model (SCM):

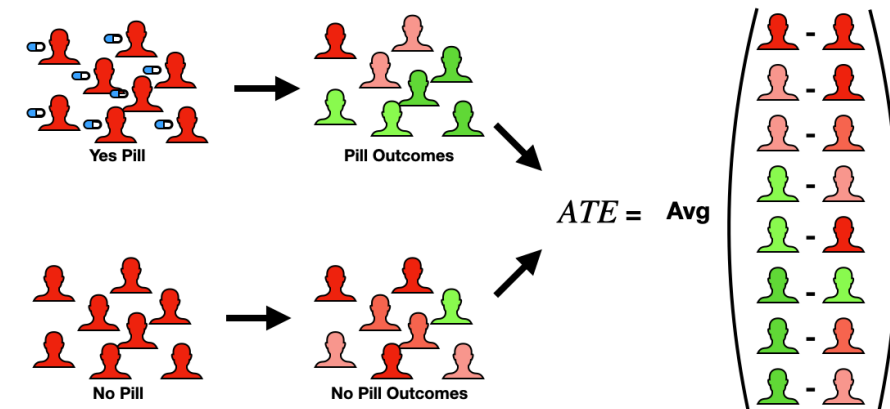
- Represents causal relations between variables.
 - Uses directed acyclic graphs (DAGs).
- **Modelling complex causal graphs in a system.**



A simplified potential DAG example from the BC project

Potential Outcome Framework (PO):

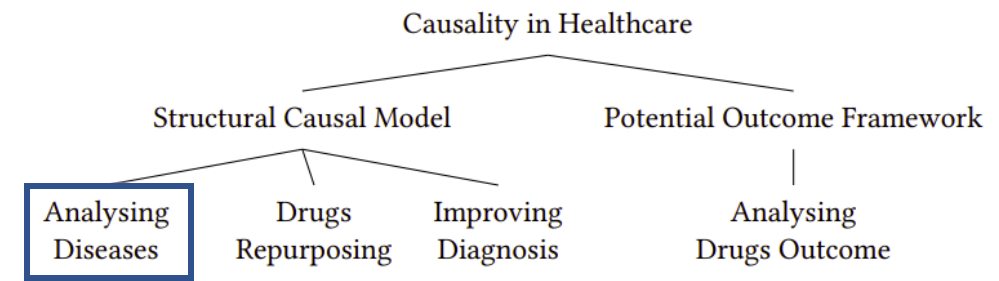
- Focuses on comparing the outcomes that would have occurred under different treatment conditions.
 - Estimates causal effects in both randomized controlled trials and observational studies.
- **Estimating the causal effect in a system.**



A simplified illustration of the average treatment effect (ATE) [\[Source\]](#)

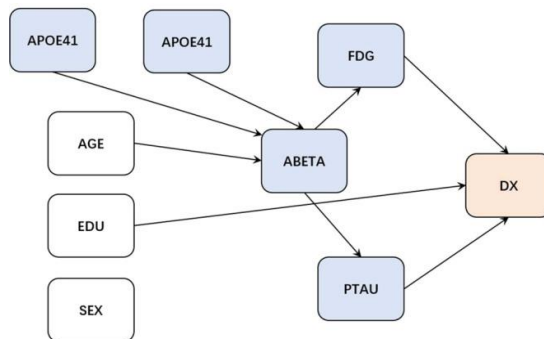
Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology.

[Shen2020]

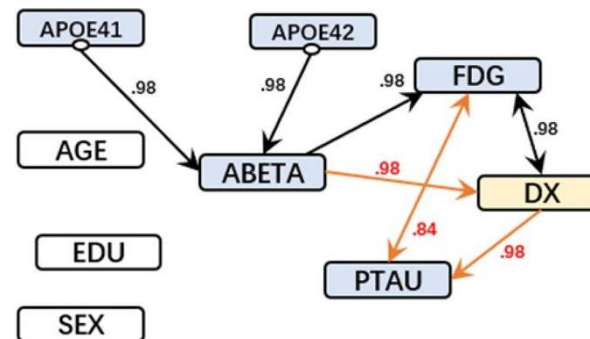


Goal: Discover the causal relationships in the Alzheimer disease (AD) and validate it with a well-established causal graph.

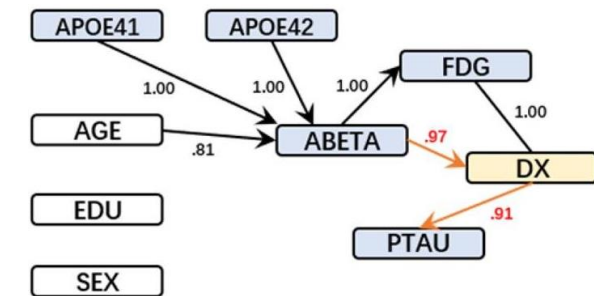
Solution: Two causal discovery methods utilizing observational data of AD and background knowledge were used: *Fast causal inference* (FCI) and the *fast greedy equivalence search* (FGES).



The “gold standard” graph [Shen2020].

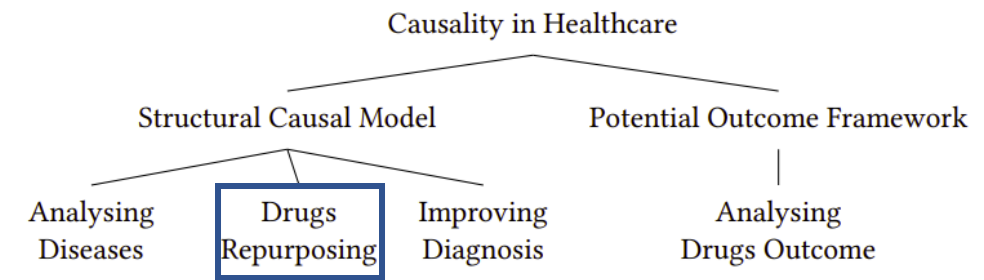


FCI graph [Shen2020].



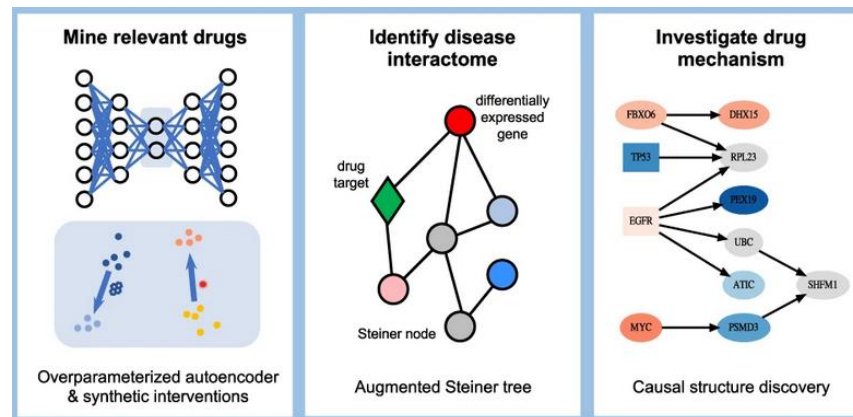
FGES graph [Shen2020].

Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing.
[Belyaeva2021]



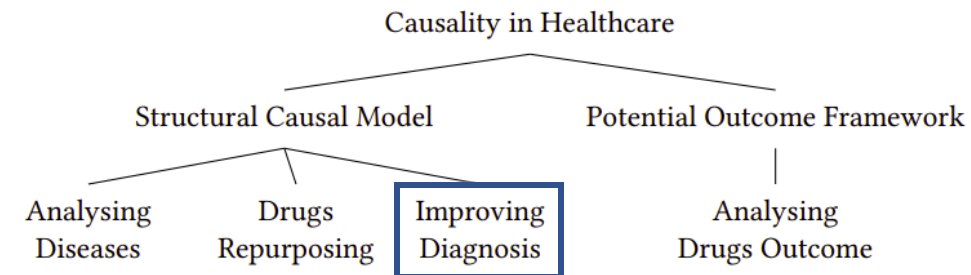
Goal: Repurpose known drugs for new diseases like SARS-CoV-2.

Solution: Integrate transcriptomic, proteomic, and structural data for different diseases. First, they used an autoencoder to match drugs signature. Second, augmented Steiner tree is used to identify drugs interactome. Lately, the causal interactions of drugs and genes are checked through a causal discovery method.



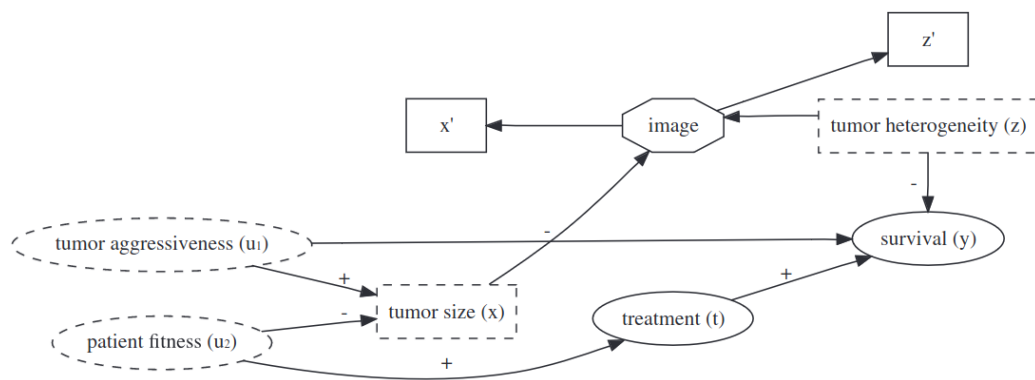
The three steps used to repurpose known drugs [Belyaeva2021].

Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning.
[vanAmsterdam2019]

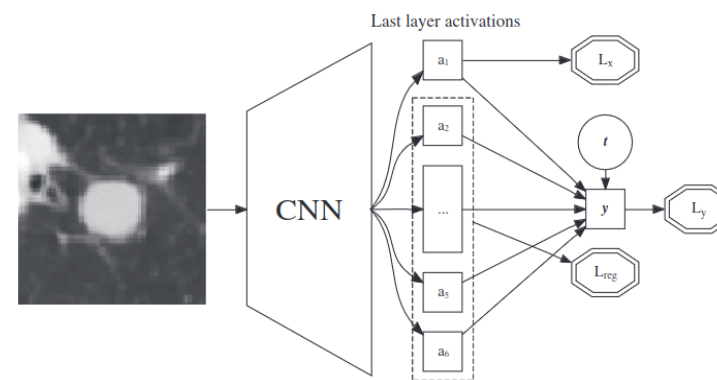


Goal: Improve the accuracy of an image classifier.

Solution: Discover the causal relations between different features found in the images, to later eliminate their bias when used for the prediction.

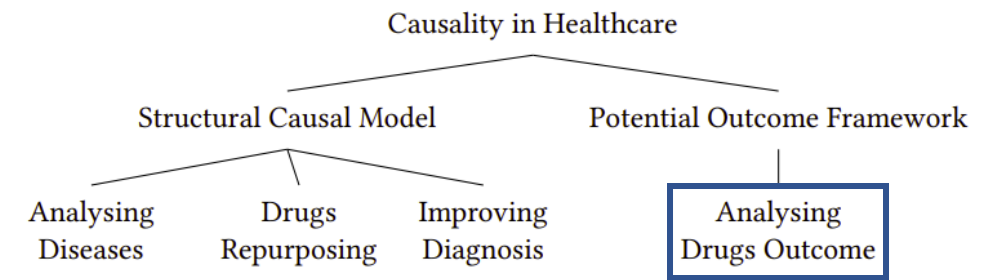


DAG describing the data-generating mechanism for the simulations [vanAmsterdam2019].



Schematic overview of the proposed convolutional neural network architecture [vanAmsterdam2019].

A Propensity Score Analysis of Chemotherapy Use in Patients With Resectable Gallbladder Cancer. [Ozer2022]



Goal: Investigate the benefits of neoadjuvant and adjuvant chemotherapy in comparison to only undergoing surgery.

Solution: Estimate the survival outcome of adjuvant and neoadjuvant chemotherapy by performing the propensity score analysis.

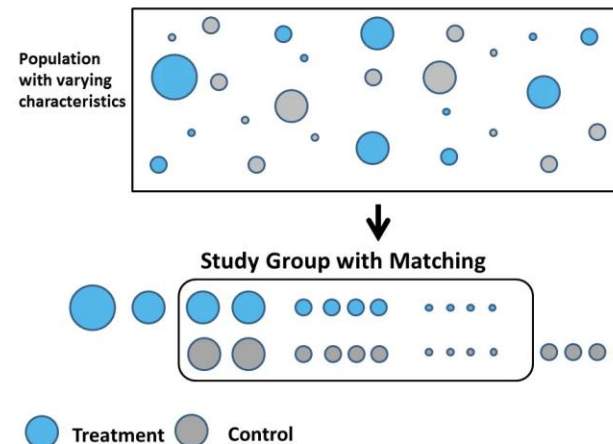


Illustration of the propensity score analysis [[source](#)].

Summary

- Work in the healthcare domain encompassing causality mainly considers the aspects of interpretability and robustness and touch upon fairness and privacy to some extent.
- The other tenets of trustworthy AI, like safety and accountability, must be explored further.
- Ground truth causal graphs are not always available: Crucial to validate the results of causal discovery methods.
- Available data in the healthcare domain are commonly unstructured, highly complex, and multimodal.

Thank you for your attention