IMPERIAL

Causal Discovery for Trustworthy AI

Fabrizio Russo

10/07/2025 - Lendable

Talk Overview

- My Journey
- **Motivation & Background**
  - Causal Models
  - Causal Discovery
- **Causal Graphs for Contestable Neural Networks**
- **Causal Discovery with Shapley Values**
- **Argumentative Causal Discovery**
- Conclusion

My Journey

Across Finance And Artificial Intelligence

2009-2012	Rome – Tor Vergata: BSc Economics & Finance
	Madrid – Autónoma
2013-2014	London: Exploring
2014-2020	London: General Electric Capital & 4most Europe - Credit Risk Analyst - Credit Risk Consultant - Managing Consultant - Head of Data Science
2016-2018	London – Birkbeck College: MSc Applied Statistics
2020-2024	London – Imperial College: PhD Safe and Trusted AI
2025-Present	London – Imperial College: Research Associate

Scan to Website

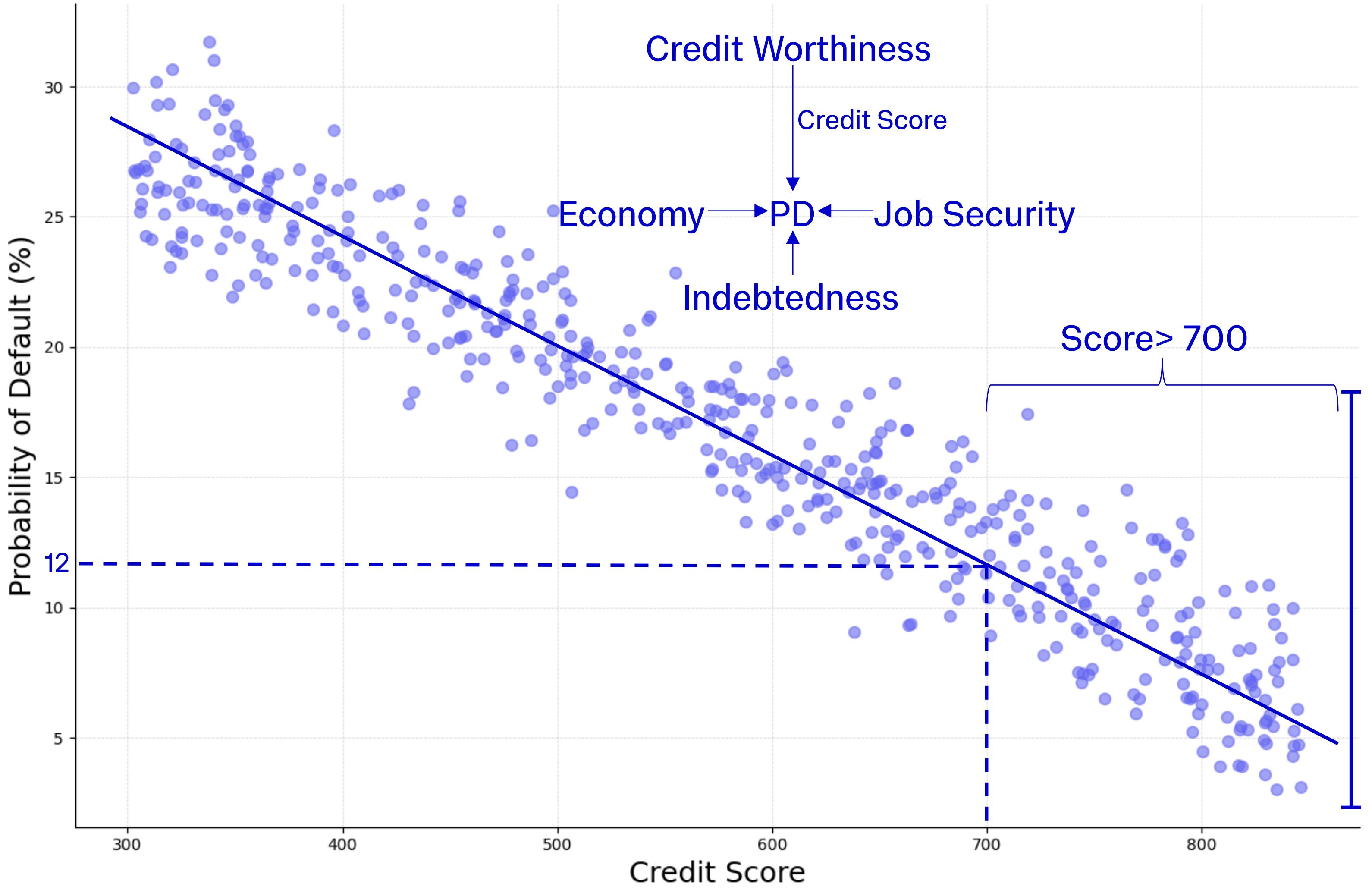
Causal Models
Motivation

Predictive Models

- What if we observe a credit score of 700?
- What if we observe a PD of 12%?

Causal Models

- What if we "intervene" on the Credit Score?
- What's the effect on the PD?



Causal Models

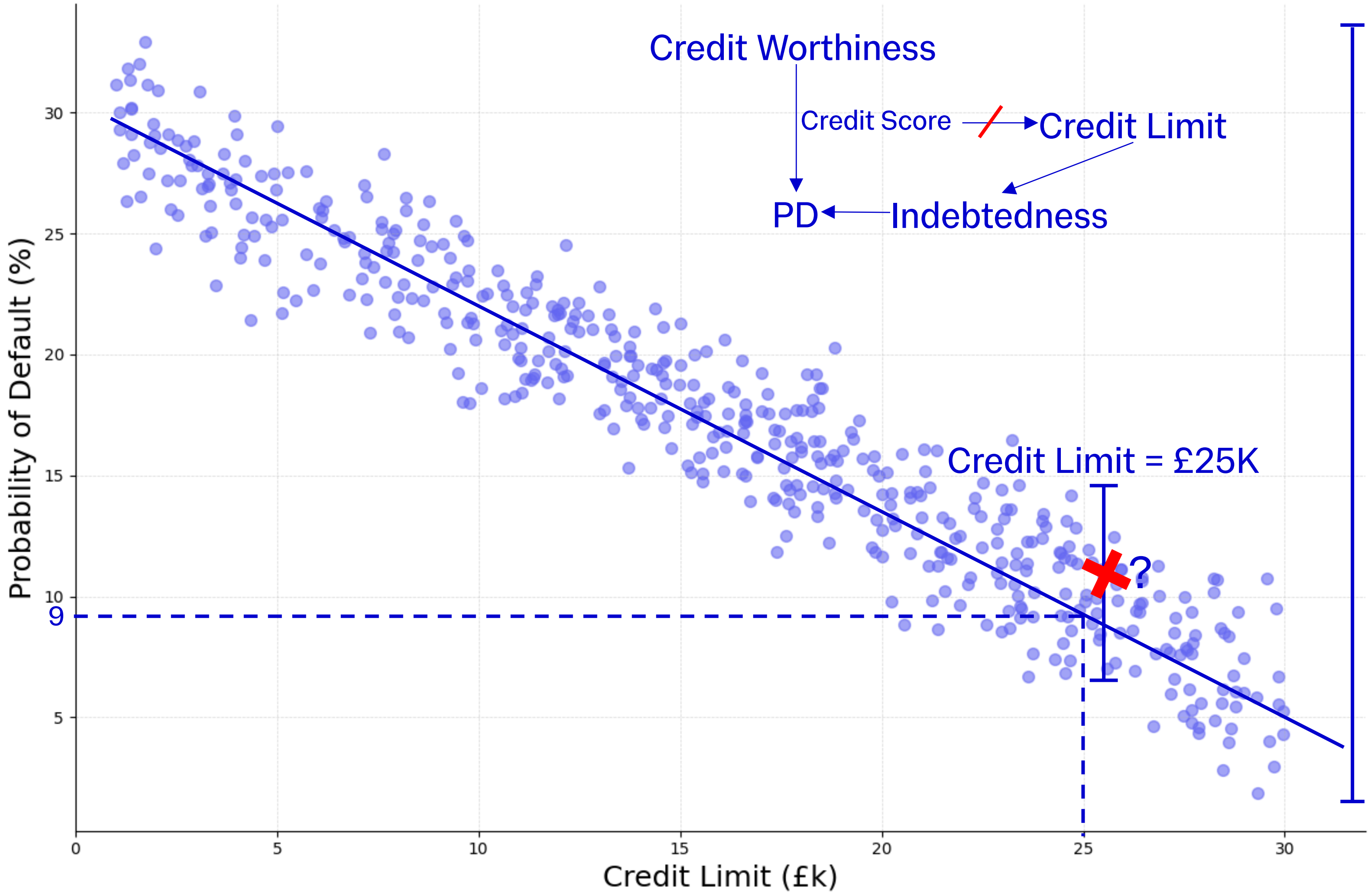
Motivation

Predictive Models

- What if we observe a credit limit of £25k?
- What if we observe a PD of 9%?

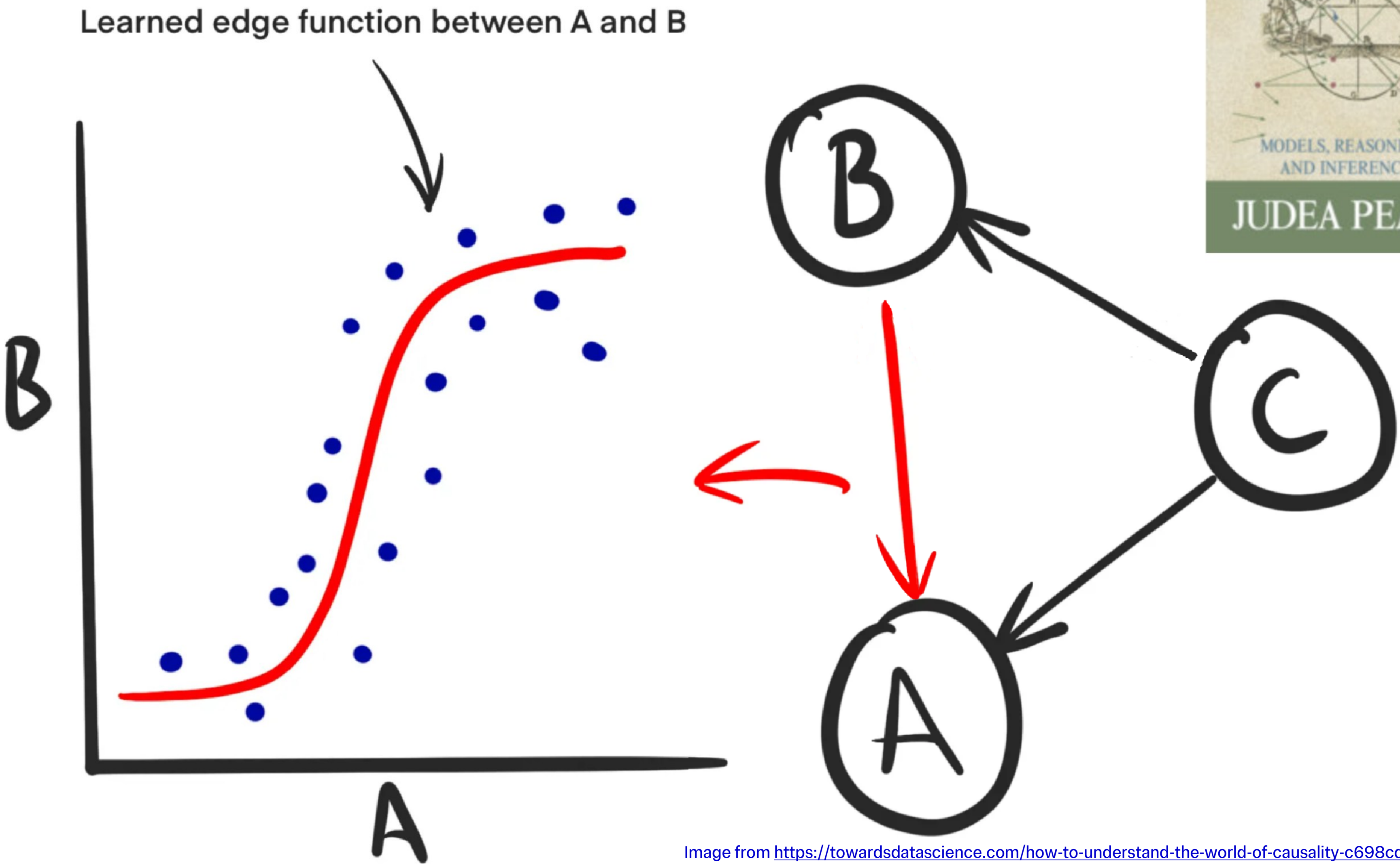
Causal Models

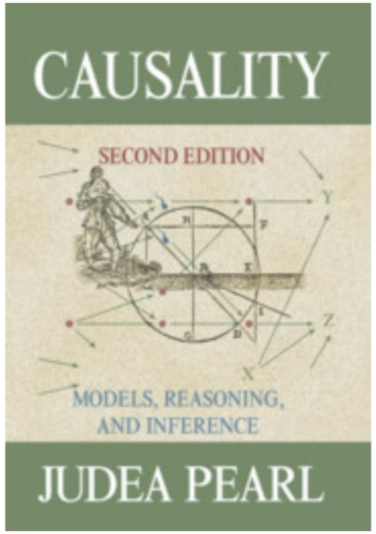
- What if we "intervene" on Credit Limit?
- What's the *causal* effect on the PD?



Structural Causal Model (Pearl, 2009)

Graph (Assumptions) + Functions





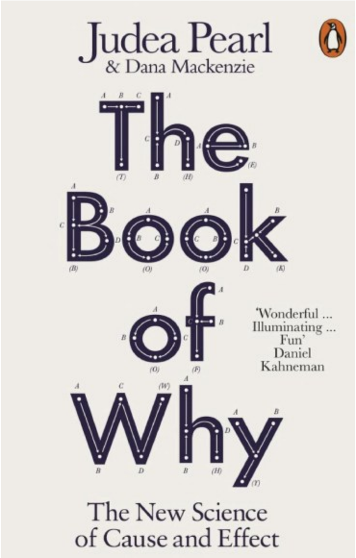


Image from https://towardsdatascience.com/how-to-understand-the-world-of-causality-c698cdc9f27c

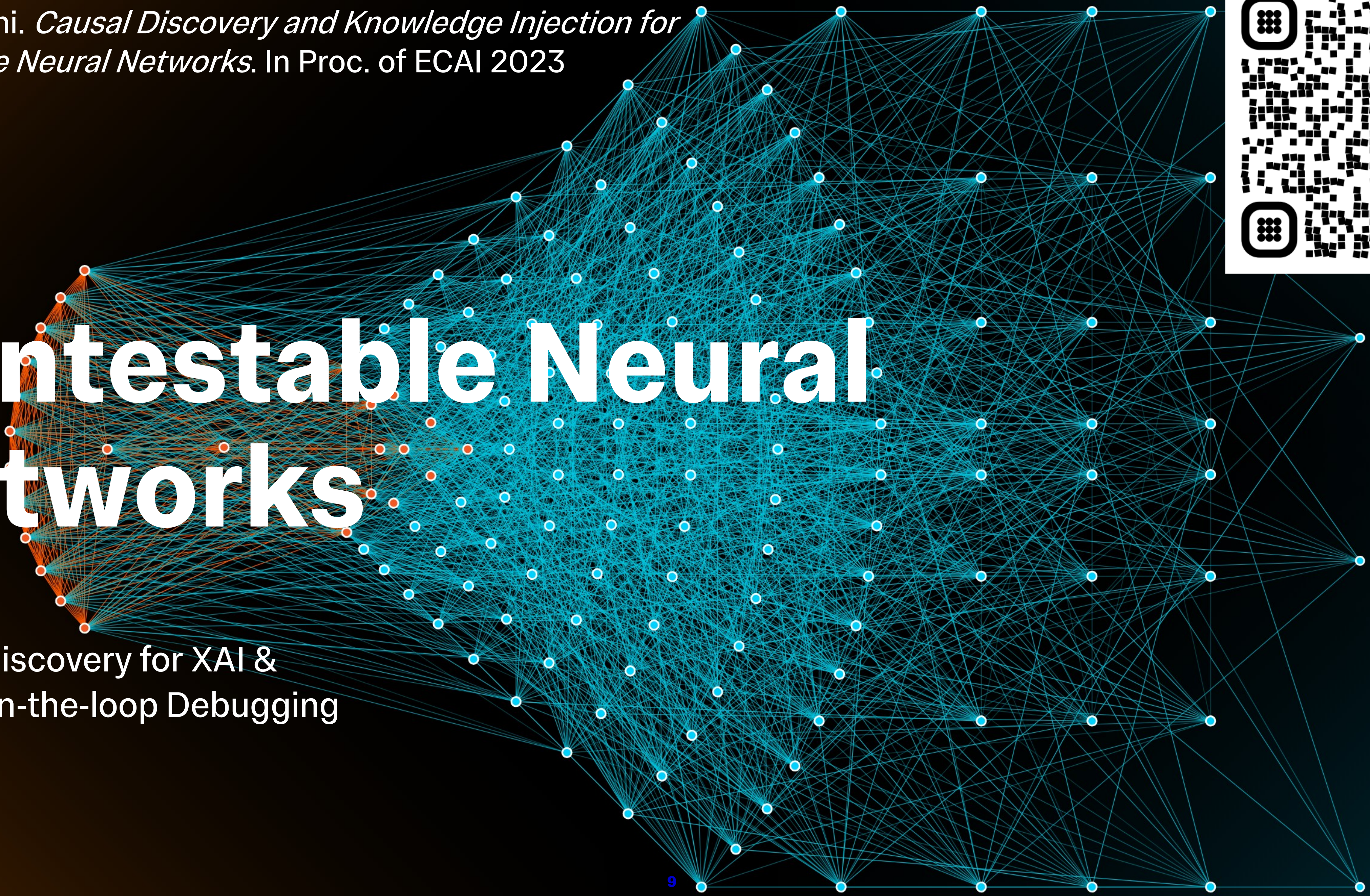
Image from https://towardsdatascience.com/how-to-understand-the-world-of-causality-e698cde9f27e

Machine Learning vs Causal Models

High-Level Comparison

Attribute	Machine Learning	Causal Models
Interpretability	Limited	High
Predictive Accuracy	High	Moderate
Generalisation	Moderate	High
Actionable Insights	Limited	High
Data Requirements	High	Moderate
Scalability	High	Low
Expert Knowledge	Moderate	High

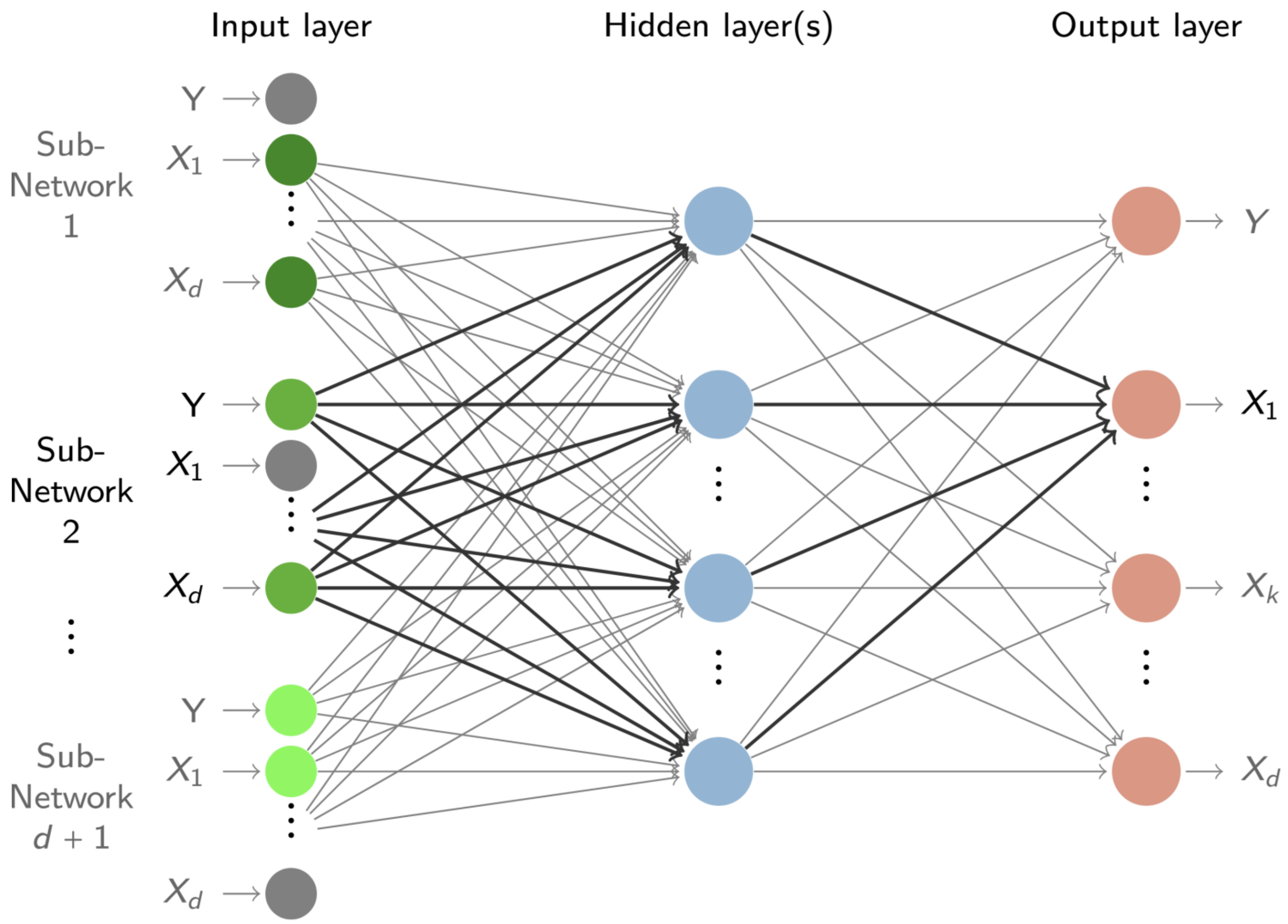
Russo & Toni. *Causal Discovery and Knowledge Injection for Contestable Neural Networks*. In Proc. of ECAI 2023



Contestable Neural Networks

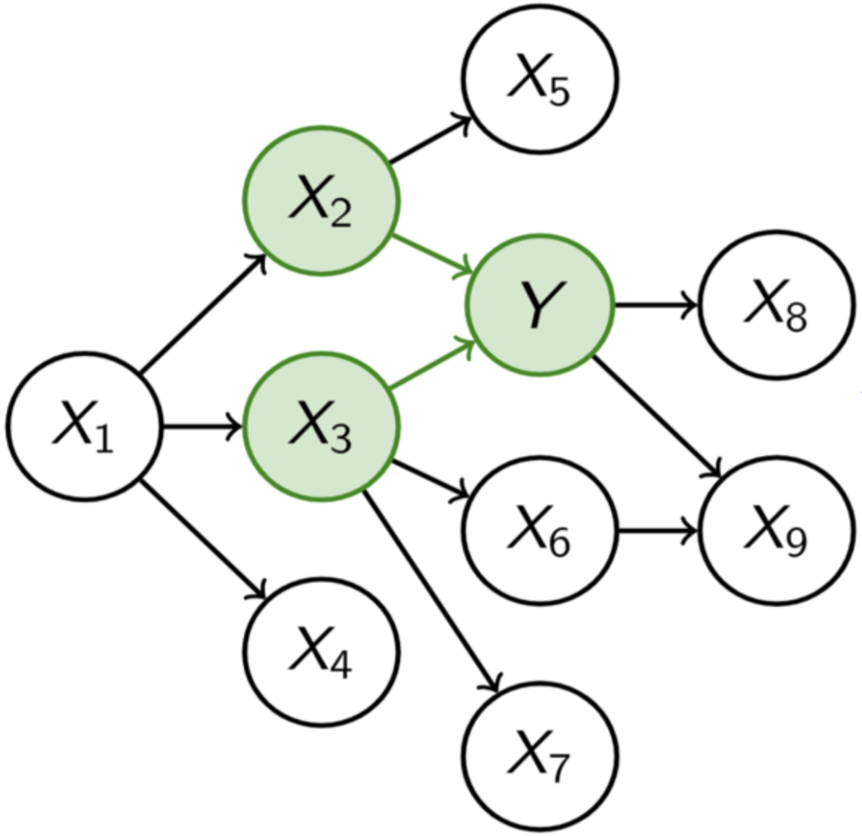
Causal Discovery for XAI &
Human-in-the-loop Debugging

Joint Neural Network Structure
(Kyono, Zhang and van der Schaar 2020)



Objective

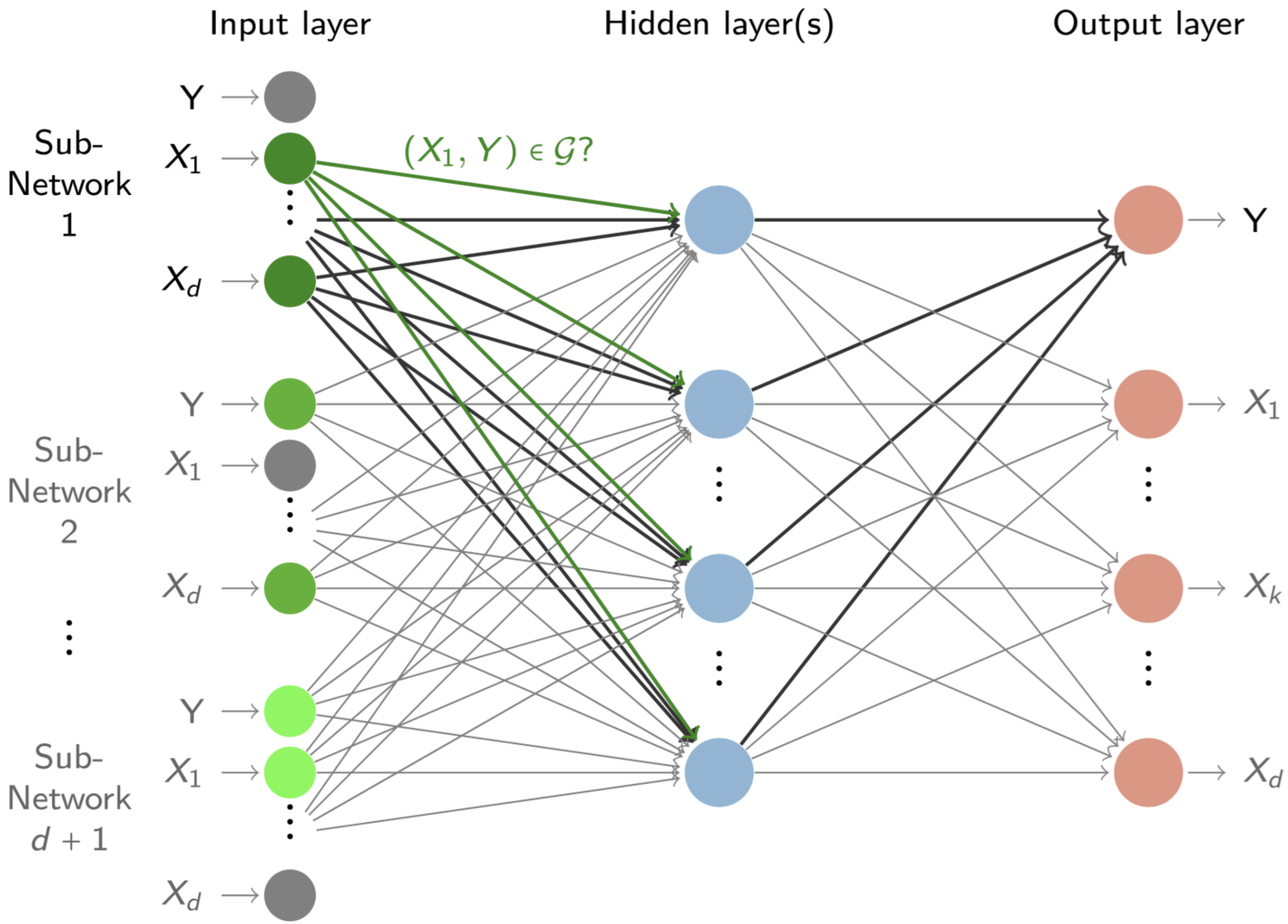
Capture Causal Relations



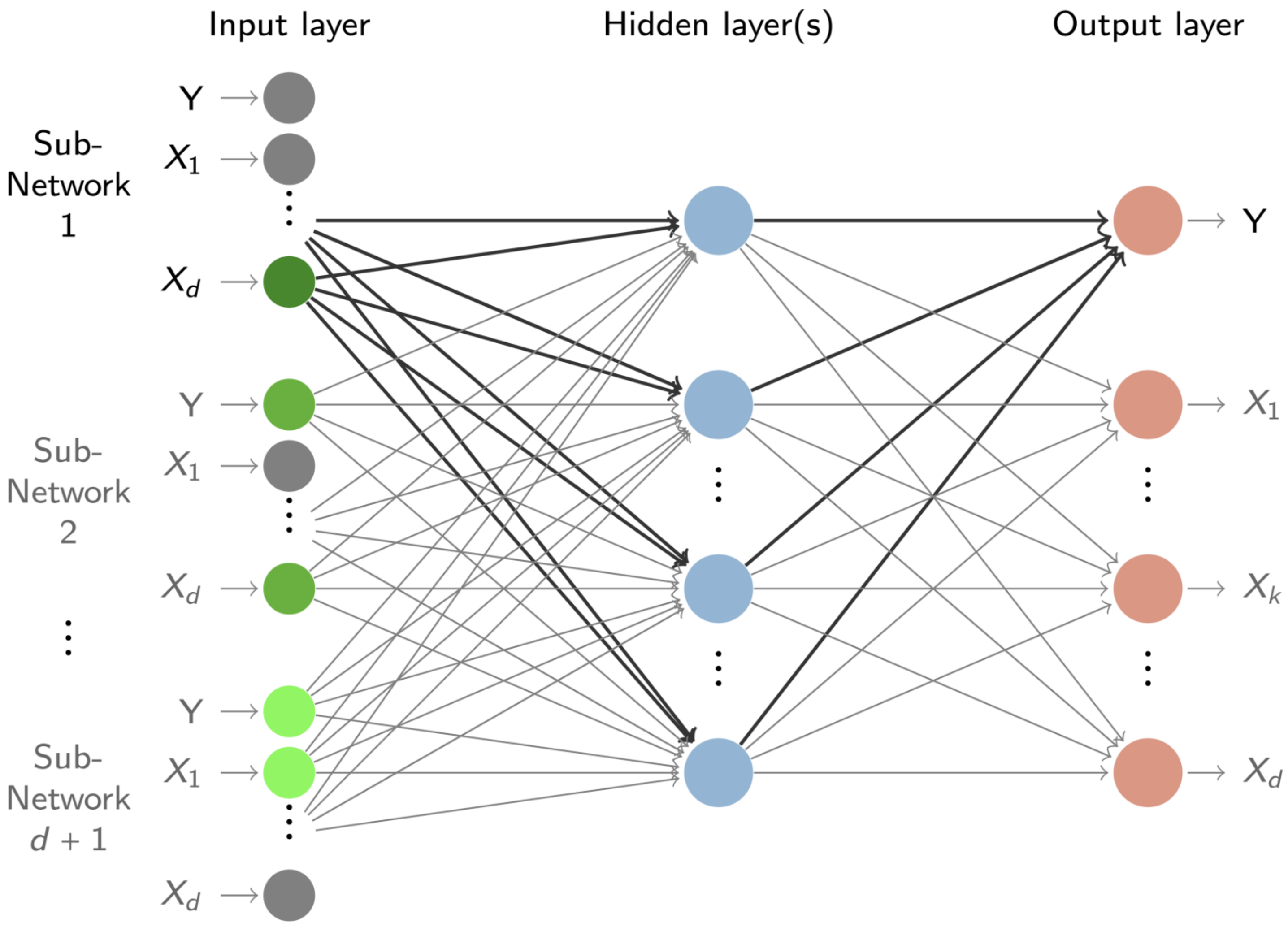
	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y	0.0	0.005	0.017	0.008	0.002	0.042	0.02	0.005	0.059	0.05
X_1	0.006	0.0	0.063	0.054	0.068	0.009	0.006	0.013	0.006	0.008
X_2	0.088	0.036	0.0	0.022	0.019	0.124	0.008	0.011	0.006	0.008
X_3	0.087	0.034	0.021	0.0	0.024	0.005	0.107	0.104	0.006	0.009
X_4	0.009	0.032	0.02	0.023	0.0	0.01	0.013	0.01	0.005	0.005
X_5	0.026	0.006	0.017	0.004	0.004	0.0	0.012	0.002	0.005	0.018
X_6	0.025	0.006	0.008	0.011	0.005	0.017	0.0	0.014	0.002	0.114
X_7	0.029	0.003	0.007	0.011	0.002	0.024	0.029	0.0	0.011	0.01
X_8	0.036	0.002	0.004	0.003	0.004	0.006	0.009	0.006	0.0	0.006
X_9	0.024	0.003	0.003	0.004	0.003	0.005	0.079	0.01	0.004	0.0

Encode Causality in the Network Structure

(Kyono, Zhang and van der Schaar 2020)



Encode Causality in the Network Structure
(Kyono, Zhang and van der Schaar 2020)



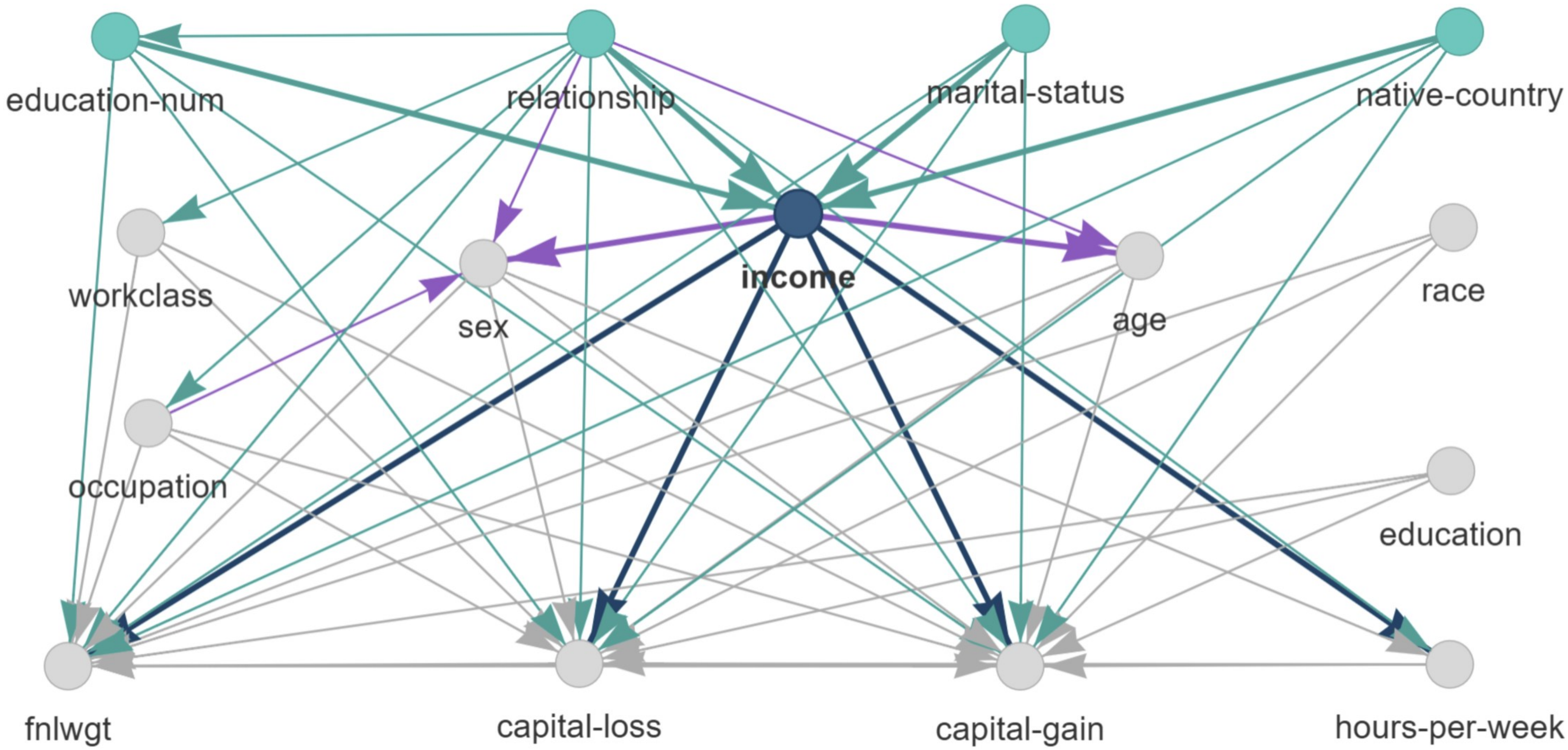
Income Prediction Case Study

Adult Dataset (Becker and Kohavi, 1996)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	
(1)			0.0	0.2				0.1					36.1	3.7	21.3	income
(2)													1.6	0.3	11.5	race
(3)								0.1					1.9	0.4	16.3	sex
(4)													0.8	0.1	2.0	age
(5)	0.0												1.1	0.2	4.9	native-country
(6)			0.0										0.6	0.2	4.0	occupation
(7)													1.2	0.3	6.4	workclass
(8)													0.8	0.1	3.6	hours-per-week
(9)													0.5	0.2	5.5	education
(10)	0.0												2.7	0.4	6.9	education-num
(11)	0.1												2.0	0.3	18.1	marital-status
(12)	0.0		0.0	0.1		0.0	0.0	0.1		0.0			2.6	0.5	15.0	relationship
(13)															0.1	capital-gain
(14)													0.2		0.5	capital-loss
(15)																fnlwgt

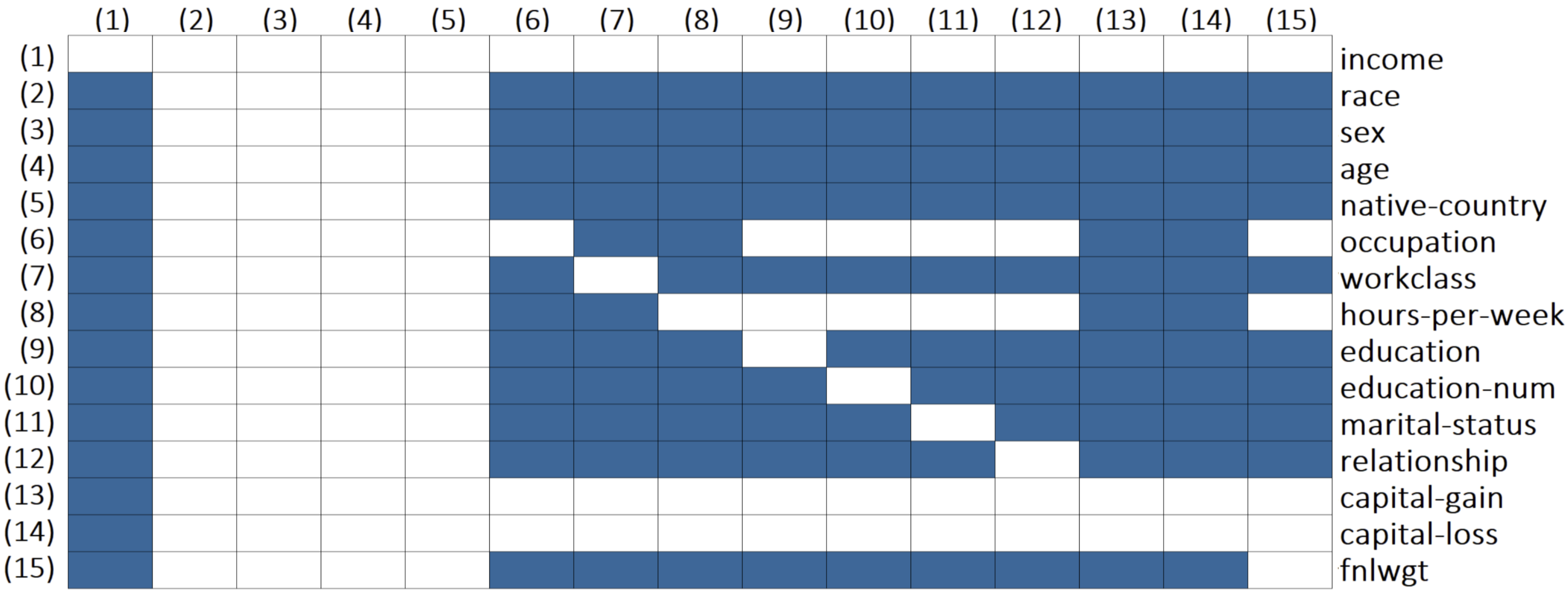
Income Prediction Case Study

Adult Dataset (Becker and Kohavi, 1996)



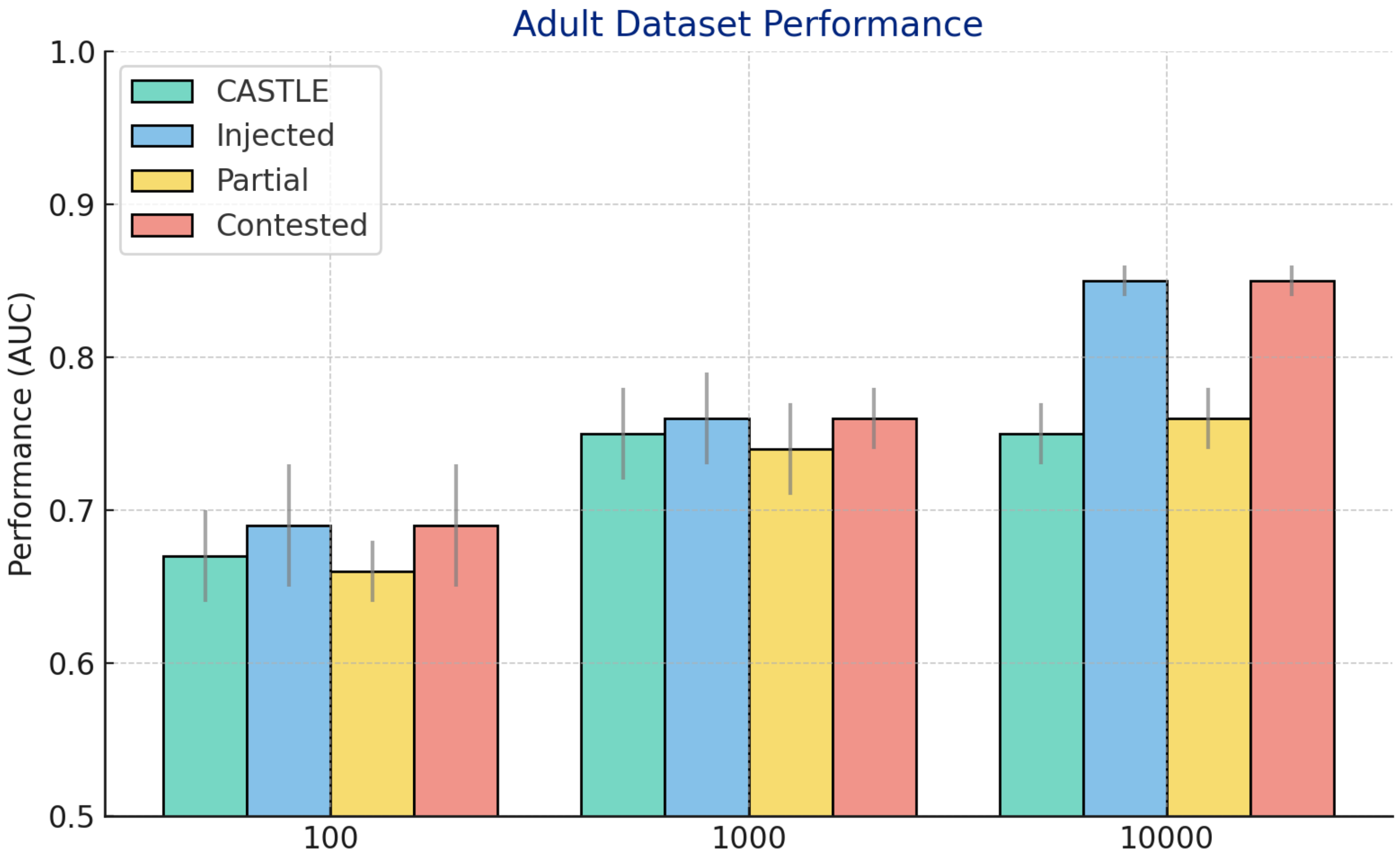
Income Prediction Case Study

Adult Dataset (Becker and Kohavi, 1996)



Income Prediction Case Study

Adult Dataset (Becker and Kohavi, 1996)



Takeaways
Causal Graphs to...

Explain

Contest

Improve

Russo & Toni. *Causal Discovery and Knowledge Injection for Contestable Neural Networks*. In Proc. of ECAI 2023

How do we reliably build causal graphs from data?

Causal Discovery Literature
From the 90s

frontiers in Genetics

REVIEW
published: 04 June 2019
doi: 10.3389/fgene.2019.00524

Check for updates

Review of Causal Discovery Methods Based on Graphical Models

Clark Glymour, Kun Zhang and Peter Spirtes*

Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, United States

A fundamental task in various disciplines of science, including biology, is to find underlying causal relations and make use of them. Causal relations can be seen if interventions are properly applied; however, in many cases they are difficult or even impossible to conduct. It is then necessary to discover causal relations by analyzing statistical properties of purely observational data, which is known as causal discovery or causal structure search. This paper aims to give a introduction to and a brief review of the computational methods for causal discovery that were developed in the past three decades, including constraint-based and score-based methods and those based on functional causal models, supplemented by some illustrations and applications.

Keywords: directed graphical causal models, causal discovery, conditional independence, statistical independence, structural equation models, non-Gaussian distribution, non-linear models

D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery

82

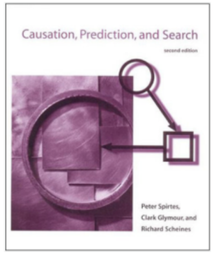
MATTHEW J. VOWELS, NECATI CIHAN CAMGOZ, and RICHARD BOWDEN,
CVSSP, University of Surrey, U.K.

Causal reasoning is a crucial part of science and human intelligence. In order to discover causal relationships from data, we need structure discovery methods. We provide a review of background theory and a survey of methods for structure discovery. We primarily focus on modern, continuous optimization methods, and provide reference to further resources such as benchmark datasets and software packages. Finally, we discuss the assumptive leap required to take us from structure to causality.

CCS Concepts: • **Mathematics of computing** → **Causal networks**; • **Computing methodologies** → **Machine learning**; **Causal reasoning and diagnostics;**

Additional Key Words and Phrases: Causality, causal discovery, directed acyclic graphs, DAGs, structure learning, survey

ACM Reference format:
Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.* 55, 4, Article 82 (November 2022), 36 pages.
https://doi.org/10.1145/3527154



Adaptive Computation And Machine Learning Series

Causation, Prediction, and Search (Second Edition)

By Peter Spirtes, Clark Glymour, Richard Scheines

The MIT Press

DOI: https://doi.org/10.7551/mitpress/1754.001.0001

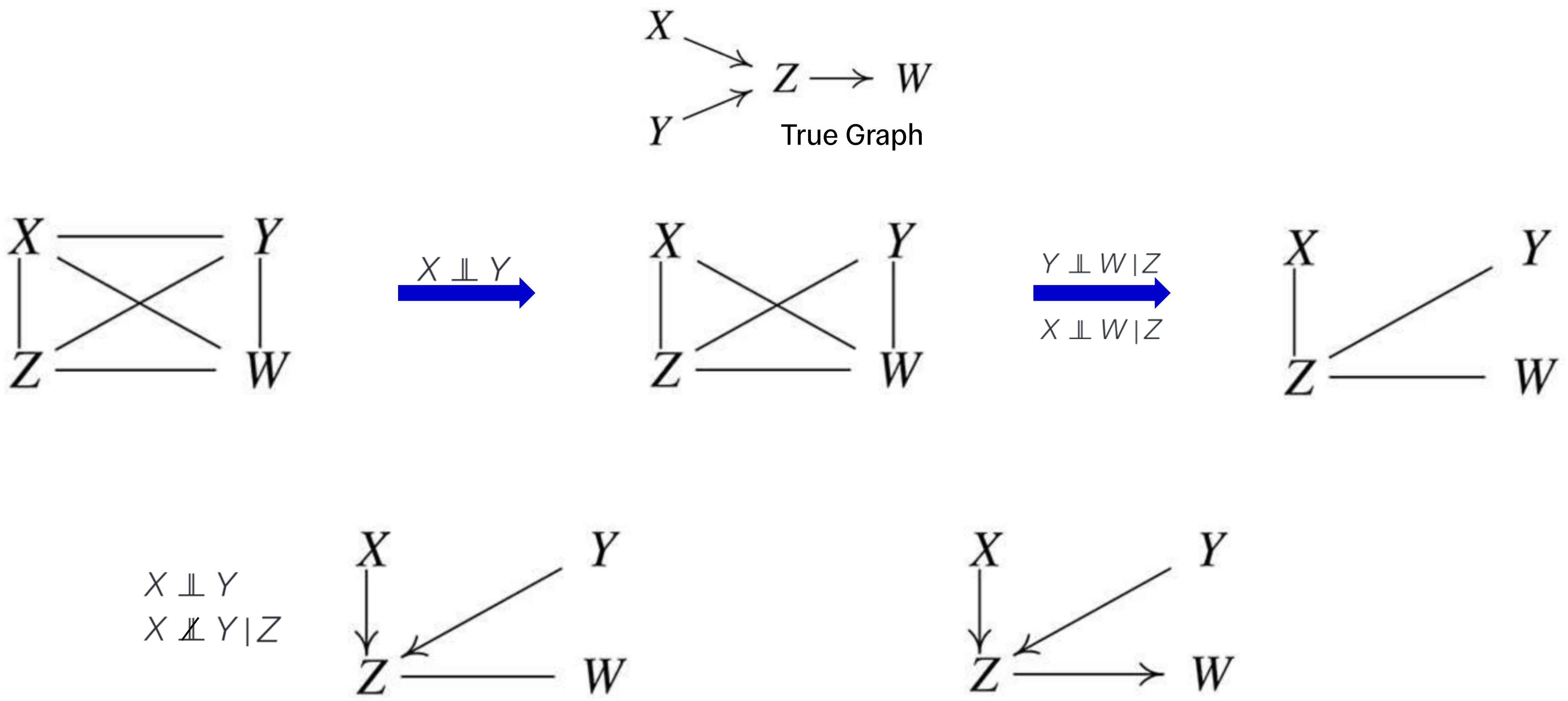
ISBN electronic: 9780262284158

In Special Collection: CogNet

Publication date: 2001

The Peter-Clark Algorithm

(Spirtes et al, 1993) Example from Glymour et al, 2019



Peter-Clark Algorithm

Sound & Complete

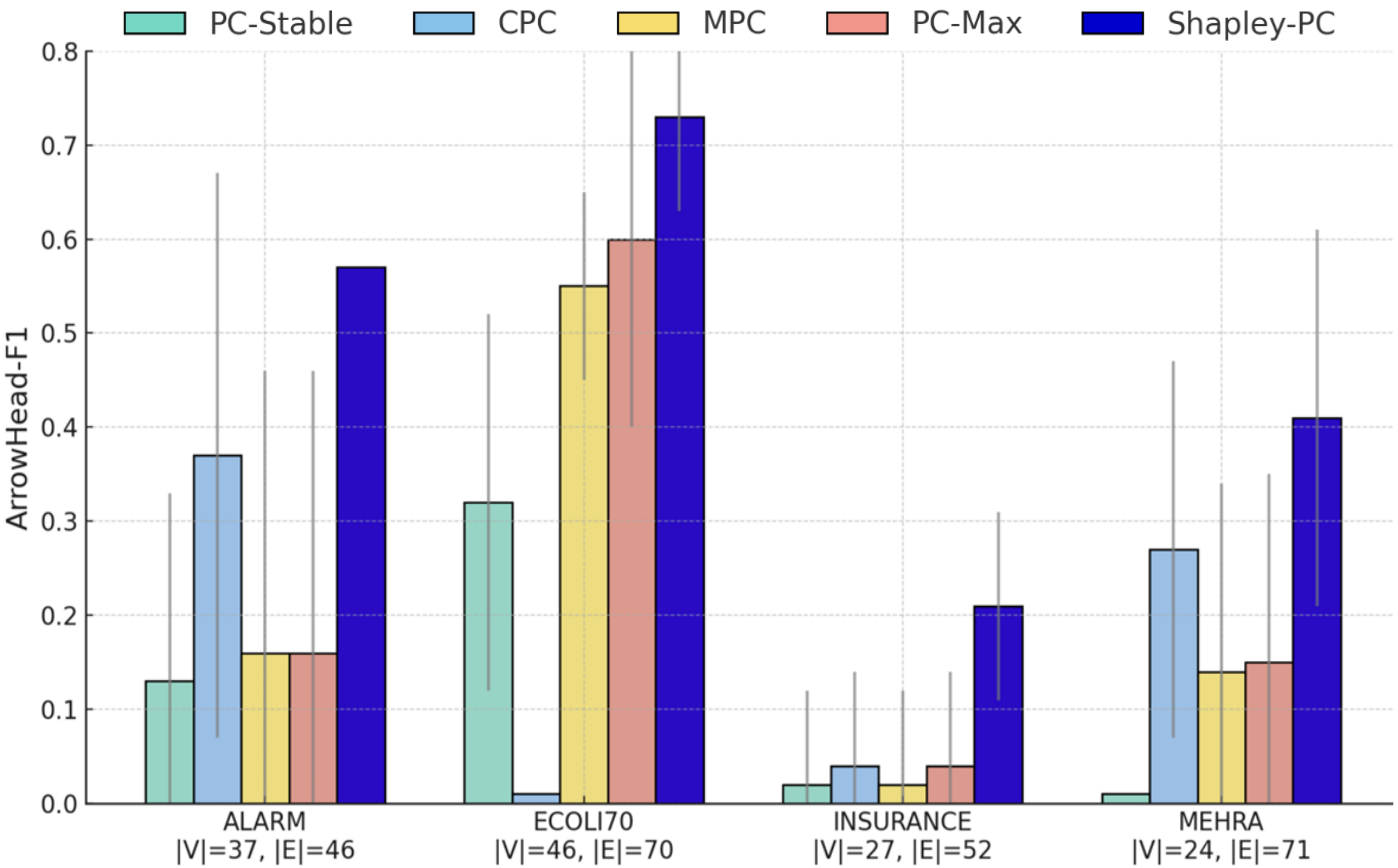
Efficient

Subject to Statistical Errors

Russo & Toni. *Shapley-PC: Constraint-based Causal Structure Learning with a Shapley Inspired Framework*. In Proc. of CLeaR 2025

Shapley-PC Reconstrucṭution Accuracy

on bnlearn datasets (Scutari, 2014)



Income Prediction Example

$E \not\perp\!\!\!\perp I$	$p = 0.00$	$\mathcal{S} = 1.00$
$E \not\perp\!\!\!\perp O$	$p = 0.00$	$\mathcal{S} = 1.00$
$R \not\perp\!\!\!\perp O$	$p = 0.00$	$\mathcal{S} = 1.00$
$O \not\perp\!\!\!\perp I$	$p = 0.00$	$\mathcal{S} = 1.00$
$R \perp\!\!\!\perp E$	$p = 0.46$	$\mathcal{S} = 0.71$
$R \not\perp\!\!\!\perp I$	$p = 0.05$	$\mathcal{S} = 0.52$
$E \not\perp\!\!\!\perp I \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp I \mid \{O\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp O \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$R \not\perp\!\!\!\perp O \mid \{I\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp O \mid \{I\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$R \not\perp\!\!\!\perp O \mid \{E\}$	$p = 0.00$	$\mathcal{S} = 0.50$

$O \not\perp\!\!\!\perp I \mid \{E\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$O \not\perp\!\!\!\perp I \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$R \perp\!\!\!\perp E \mid \{O\}$	$p = 0.53$	$\mathcal{S} = 0.38$
$R \not\perp\!\!\!\perp I \mid \{O\}$	$p = 0.03$	$\mathcal{S} = 0.35$
$R \perp\!\!\!\perp E \mid \{O\}$	$p = 0.33$	$\mathcal{S} = 0.32$
$R \perp\!\!\!\perp E \mid \{I\}$	$p = 0.05$	$\mathcal{S} = 0.25$
$R \perp\!\!\!\perp E \mid \{O, I\}$	$p = 0.39$	$\mathcal{S} = 0.00$
$R \not\perp\!\!\!\perp I \mid \{E, O\}$	$p = 0.00$	$\mathcal{S} = 0.00$
$E \not\perp\!\!\!\perp O \mid \{R, I\}$	$p = 0.00$	$\mathcal{S} = 0.00$
$R \not\perp\!\!\!\perp I \mid \{E, O\}$	$p = 0.00$	$\mathcal{S} = 0.00$
$R \not\perp\!\!\!\perp O \mid \{E, I\}$	$p = 0.03$	$\mathcal{S} = 0.00$

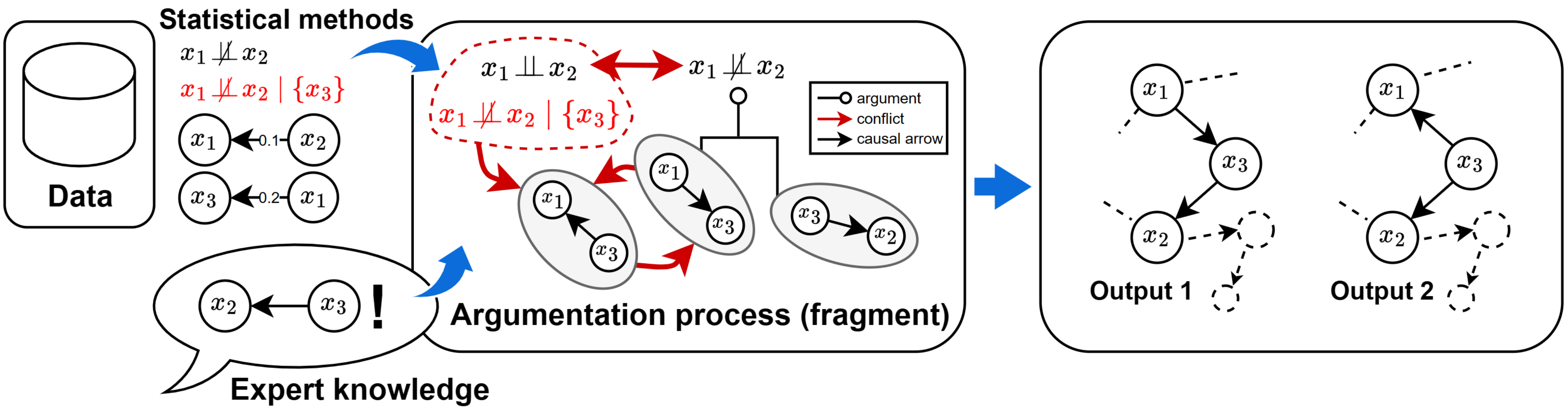
Income Prediction Example

True Causal Graph

Majority-PC (Colombo and Maathias, 2012)

Shapley-PC (Russo and Toni, 2025)

Argumentative Causal Discovery
A Debate about Causality

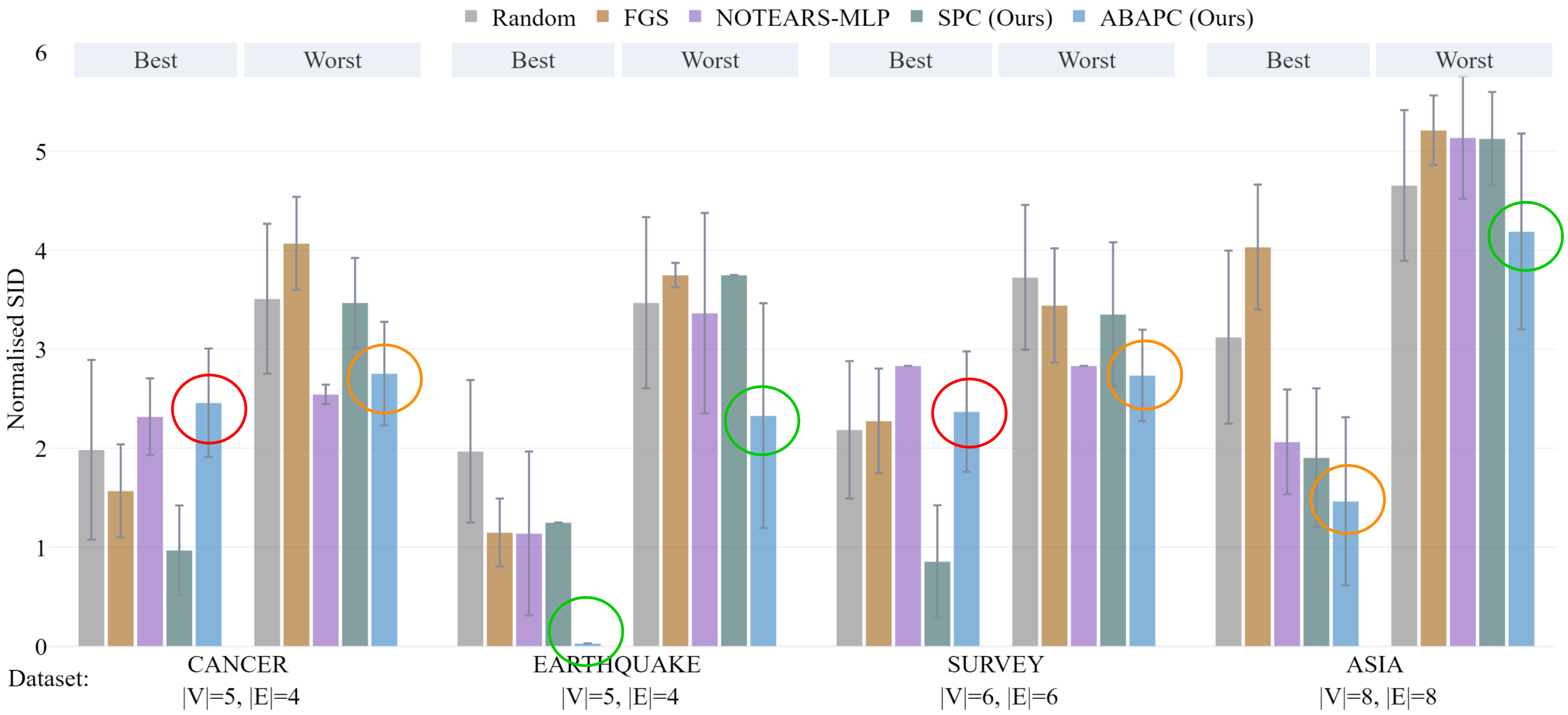


Income Prediction Example

$E \not\perp\!\!\!\perp I$	$p = 0.00$	$\mathcal{S} = 1.00$
$E \not\perp\!\!\!\perp O$	$p = 0.00$	$\mathcal{S} = 1.00$
$R \not\perp\!\!\!\perp O$	$p = 0.00$	$\mathcal{S} = 1.00$
$O \not\perp\!\!\!\perp I$	$p = 0.00$	$\mathcal{S} = 1.00$
$R \perp\!\!\!\perp E$	$p = 0.46$	$\mathcal{S} = 0.71$
$R \not\perp\!\!\!\perp I$	$p = 0.05$	$\mathcal{S} = 0.52$
$E \not\perp\!\!\!\perp I \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp I \mid \{O\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp O \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$R \not\perp\!\!\!\perp O \mid \{I\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$E \not\perp\!\!\!\perp O \mid \{I\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$R \not\perp\!\!\!\perp O \mid \{E\}$	$p = 0.00$	$\mathcal{S} = 0.50$

$O \not\perp\!\!\!\perp I \mid \{E\}$	$p = 0.00$	$\mathcal{S} = 0.50$
$O \not\perp\!\!\!\perp I \mid \{R\}$	$p = 0.00$	$\mathcal{S} = 0.50$
~~$R \perp\!\!\!\perp E \mid \{O\}$~~	~~$p = 0.53$~~	~~$\mathcal{S} = 0.38$~~
~~$R \not\perp\!\!\!\perp I \mid \{O\}$~~	~~$p = 0.03$~~	~~$\mathcal{S} = 0.35$~~
~~$R \perp\!\!\!\perp E \mid \{O\}$~~	~~$p = 0.33$~~	~~$\mathcal{S} = 0.32$~~
~~$R \perp\!\!\!\perp E \mid \{I\}$~~	~~$p = 0.05$~~	~~$\mathcal{S} = 0.25$~~
~~$R \perp\!\!\!\perp E \mid \{O, I\}$~~	~~$p = 0.39$~~	~~$\mathcal{S} = 0.00$~~
~~$R \not\perp\!\!\!\perp I \mid \{E, O\}$~~	~~$p = 0.00$~~	~~$\mathcal{S} = 0.00$~~
~~$E \not\perp\!\!\!\perp O \mid \{R, I\}$~~	~~$p = 0.00$~~	~~$\mathcal{S} = 0.00$~~
~~$R \not\perp\!\!\!\perp I \mid \{E, O\}$~~	~~$p = 0.00$~~	~~$\mathcal{S} = 0.00$~~
~~$R \not\perp\!\!\!\perp O \mid \{E, I\}$~~	~~$p = 0.03$~~	~~$\mathcal{S} = 0.00$~~

ABAPC Reconstruction Accuracy
on bnlearn datasets (Scutari, 2014)



Argumentative Causal Discovery

Robust & Interactive

Sound & Complete

Robust to Errors but Computationally Demanding

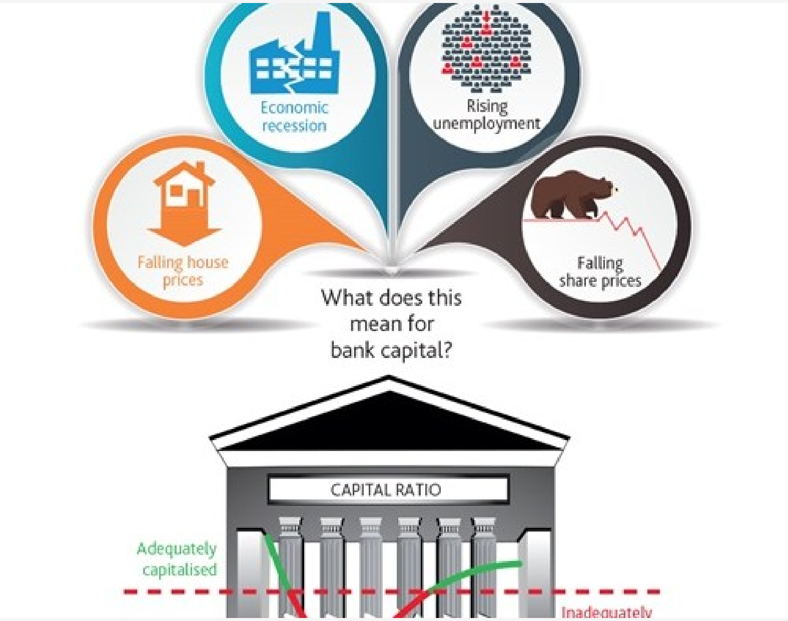
Data Check & Stakeholder Engagement

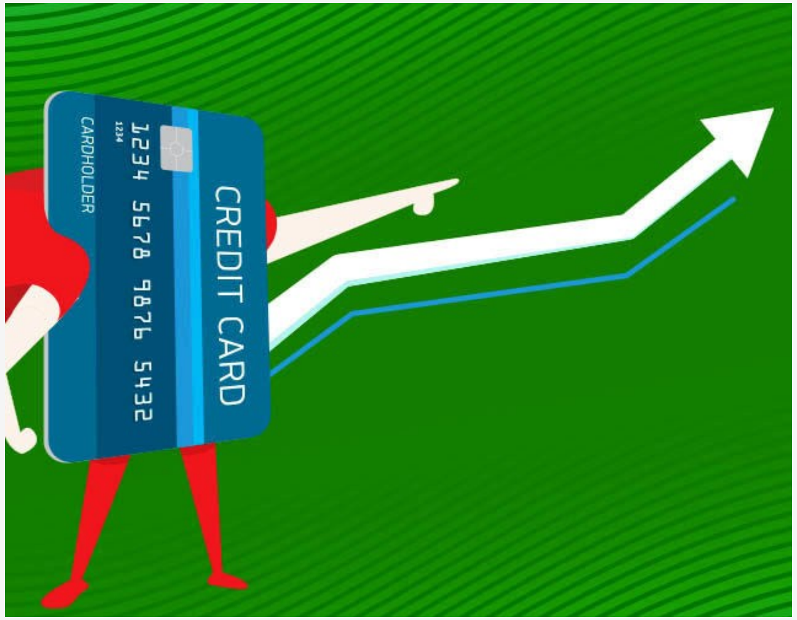
Russo, Rapberger & Toni. *Argumentative Causal Discovery*.
In Proc. of KR 2024.

Endless Possibilities

What's your most pressing use case?

- Currently working on developing an interface for expert collaboration
- ERC funding until June 2026 and starting collaborations in the fall





IMPERIAL

Questions?

Get in touch

fabrizio@imperial.ac.uk