Causality

Demian Wassermann 2022, Inria Saclay *île-de-France*

Why Causality

AI / ML

- Underspecified Goals ➡Big Data Cures Everything
- Underspecified Limitations ➡Big Data Can Do Everything
- Underspecified Caveats ➡Big Data & Big Brother

Goals in AI

- Fair ➡Biases
- Accountable ➡explainability
- Transparent ➡Decision making can be supported
- Robust ➡attacks / manipulations

Why Causality —What's the Issue with pure AI

- Biases in data, lots of them
- Leads to biased learnt models
- Robustness
- Scope becomes very important

References

- C. O'Neill, Weapons of Math Destruction, 2016
- Zeynep Tufekci, We're building a dystopia just to make people click on ads, Ted Talks, Oct 2017.

Why Causality —Some Issues with “Data is Everything”

- Biases in data, lots of them
- Leads to biased learnt models
- Robustness
- Scope becomes very important

References

- C. O’Neill, Weapons of Math Destruction, 2016
- Z. Tufekci, We’re building a dystopia just to make people click on ads, Ted Talks, Oct 2017.

ML Approach to Explainable Models

Discriminative or Generative modelling

- Given

$$D = \{(x_i, y_i), x_i \in \mathbb{R}^d, i \in 1 \ldots N\}, \text{ iid samples } P(X, Y)$$

- Supervised learning $\hat{h} : X \mapsto Y$ i.e. $\arg\max_Y P(Y|X)$
- Generative modelling $\hat{q} : X \times Y \to \mathbb{R}_+, \text{ i.e. } \hat{P}(X, Y)$

Lead to Predictive Modelling which will reproduce data biases

e.g.: If there are lots of umbrellas, then it rains

Caillebotte, 1877

ML Approach to Explainable Models

But Not All Biases are Bad

Seurat, 1884

The Implicit Big Data Promise

- If you can predict, can you control?

Knowledge -> Prediction -> Control

So How can this be Tested? Interventions

- Think about nutrition
- Think about healthcare
- Economy
- Climate

Pearl's "Do" operator: $do(X = a)$ means that we intervene a system on event X to make "a" true (Pearl 2009).

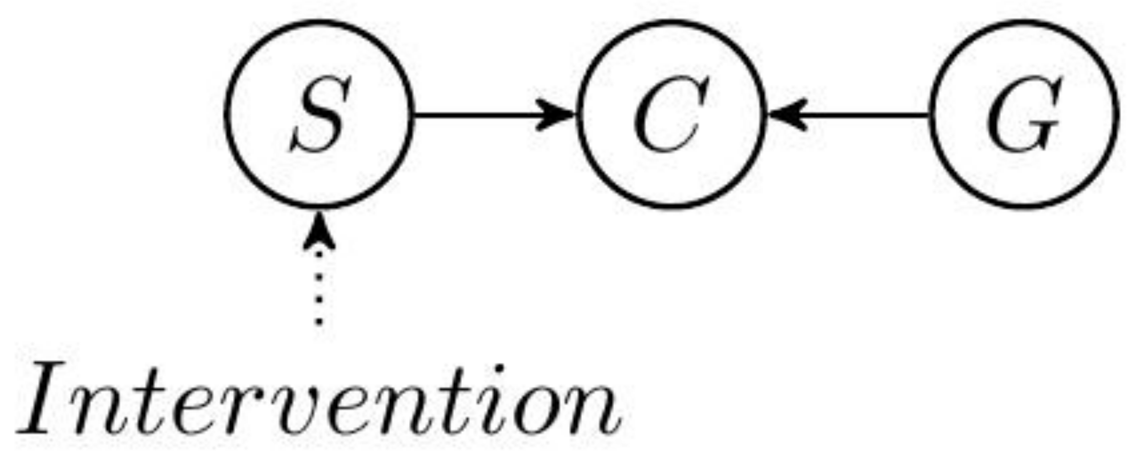
The Implicit Big Data Promise

X is a direct cause of Y if when we intervene it Y's law changes

$$X \to Y \quad \text{iif}$$

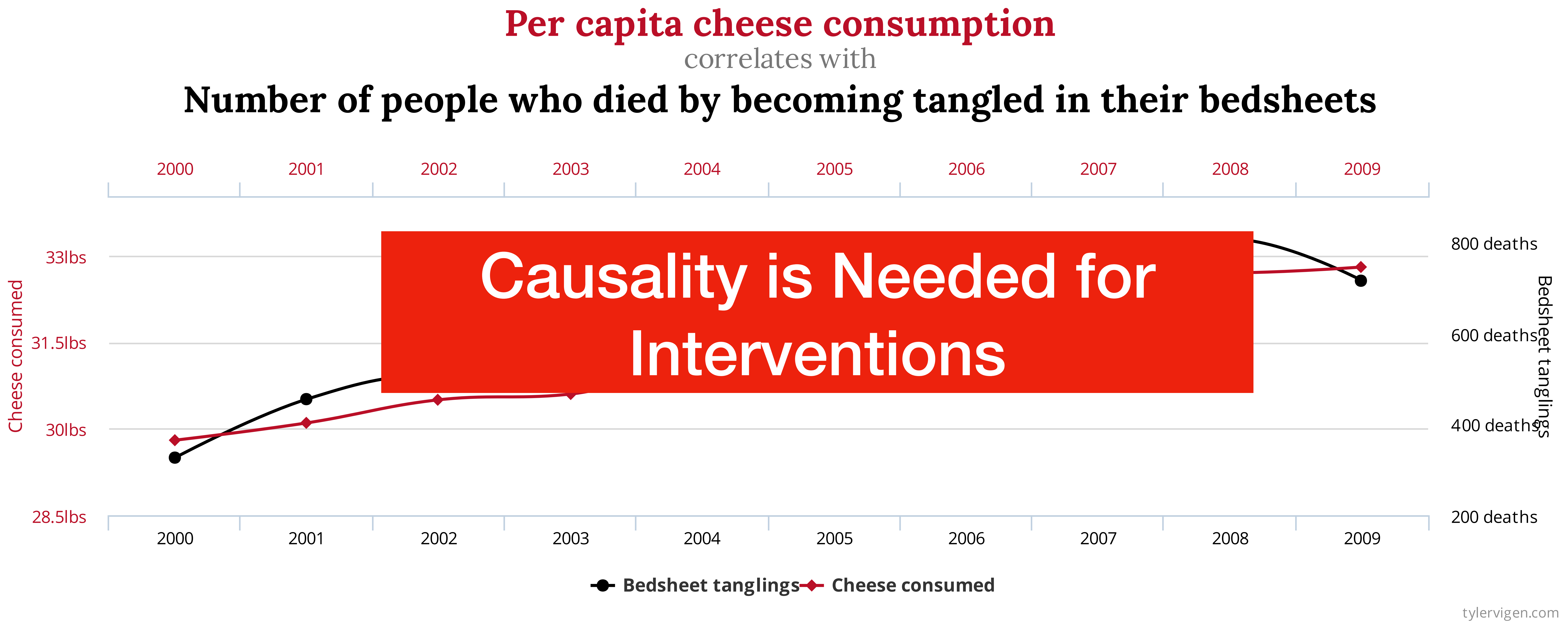
$$P_{Y|do(X=a,Z=c)} \neq P_{Y|do(X=b,Z=c)}$$

Example: Cancer, Smoking, and Genetic Factors



$$P_{C|do(S=1,G=0)} \neq P_{C|do(S=0,G=0)}$$

Correlation does not Imply Causation



Causality is Needed for Interventions

https://www.tylervigen.com/spurious-correlations

Prediction is not Causation

- Consider

$$
\begin{aligned}
X &\sim \text{Uniform}(0,1) \\
E_Y, E_Z &\sim \mathcal{N}(0,1) \\
Y &\leftarrow 0.5X + E_Y \\
Z &\leftarrow Y + E_Z
\end{aligned}
$$

- Prediction

$$\hat{Y} = 0.25X + 0.5Z$$

as a causal model suggests that Y depends on Z

Direction of prediction often indistinguishable

Correlation does not Imply Causation: A Serious Case

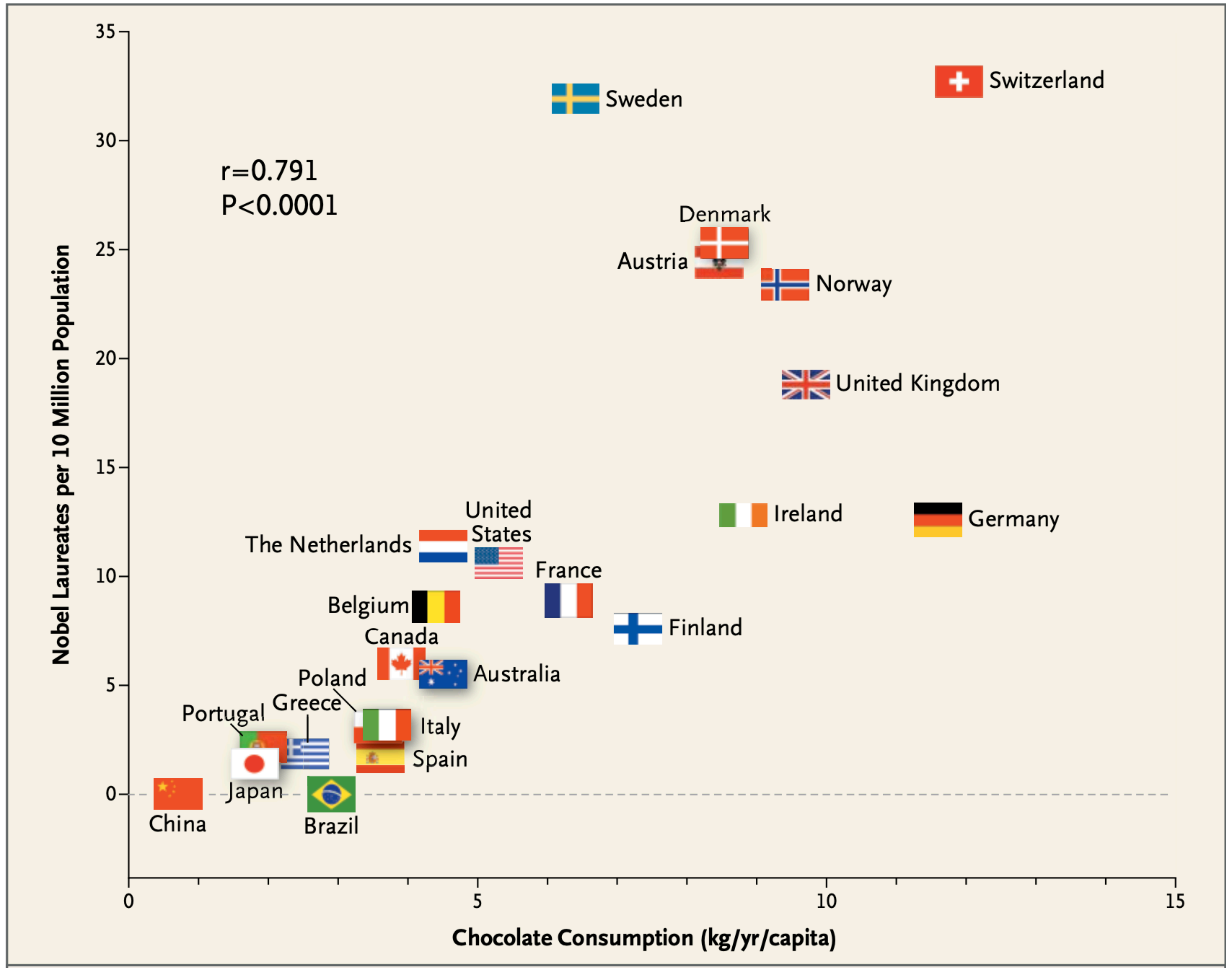


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Nobel Laureates Ratio ← Country Wealth

✗

Chocolate Consumption ← Country Wealth

This means Confounders:
Variables are not Independent

$$\text{chocolate consumption} \not\perp\!\!\!\perp \text{nobel laurate ration}$$

Probable Explanation:
Variables are Independent Conditionally to Another Event

$$\text{chocolate consumption} \perp\!\!\!\perp \text{nobel laurate ration} \mid \text{country wealth}$$

Causality and Paradoxes

- If mother smokes, child is small
- Tiny child, implies health issues
- However, P(tiny child, mother smokes)>P(tiny child)

So smoking is beneficial to child's health?

Explain issues away:

- Multi-causality of children weight
- These causes *also* affect health
- Compared to these mother smoking is not that bad, but frequency of smoking?
- Conclusions Contain Social Biases: mother is always responsible (autism, etc)

Causality and Paradoxes

- If mother smokes, child is small
- Tiny child, implies health issues
- However, P(tiny child, mother smokes)>P(tiny child)

So smoking is beneficial to child's health?

Explain issues away:

  - Multi-causality of children weight
  - These causes *also* affect health
  - Compared to these mother smoking is not that bad, but frequency of smoking?
  - Conclusions Contain Social Biases: mother is always responsible (autism, etc)

Why Causality

Goals in AI

- Fair ➡Biases
- Accountable ➡explainability
- Transparent ➡Decision making can be supported
- Robust ➡attacks / manipulations

Causality Argued Advantages

- Decreased sensitivity wrt to Data
- Simulation of Interventions ➡variable clamping
- Hopes for explanation / bias detection
- Robust

Causal Discovery

How

- Gold Standard ➡Randomised Controlled Experiments
- Feasibility ➡Low in many cases, especially human
- The AI/ML Setting ➡discovery: infer model from data

What For?

- Understandable, interpretable models
- Prioritise confirmatory experiments: enable some control
- Generate new data: for simulation, privacy, medical training

Applications

- Physics
- Neuroscience
- Epidemiology
- Economy
- Climate

How do we do it?

Causal Modelling

Setting

- Assume we have the random variables

$$X_1, \ldots, X_d$$

- with a sample joint distribution

$$\mathcal{D} = \{x_i \in \Omega^d, i = 1 \ldots n\}$$

Formal Background

- Key concept
- Framework
- Approaches

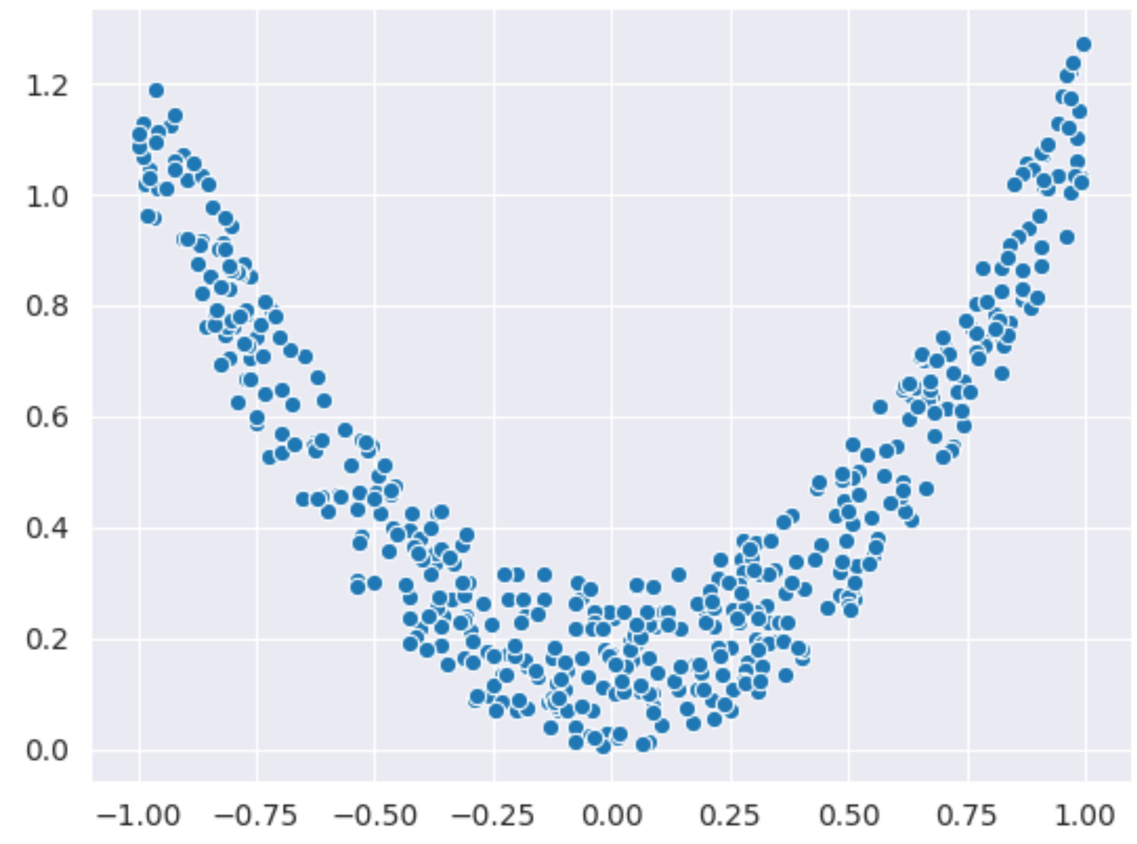
Key Concept 1: Variable (in)Dependency

- Definition of Independency

$$X \perp\!\!\!\perp Y \leftrightarrow P(X,Y) = P(X)P(Y)$$

- How do we test for independency?
  Correlation? It only works for first order linear dependencies

$$Y = X^2 + \epsilon \rightarrow \text{correlation}(X,Y) \simeq 0$$



Key Concept 1: Variable (in)Dependency

- Definition of Independency

$$X \perp\!\!\!\perp Y \leftrightarrow P(X,Y) = P(X)P(Y)$$

- How do we test for independency?
  Different tests:
    - Correlation $Y = X^2 + \epsilon \rightarrow \text{correlation}(X,Y) \simeq 0$
    - HSIC, Hilbert-Schmitt Independence Criterion (Gretton et al 05)

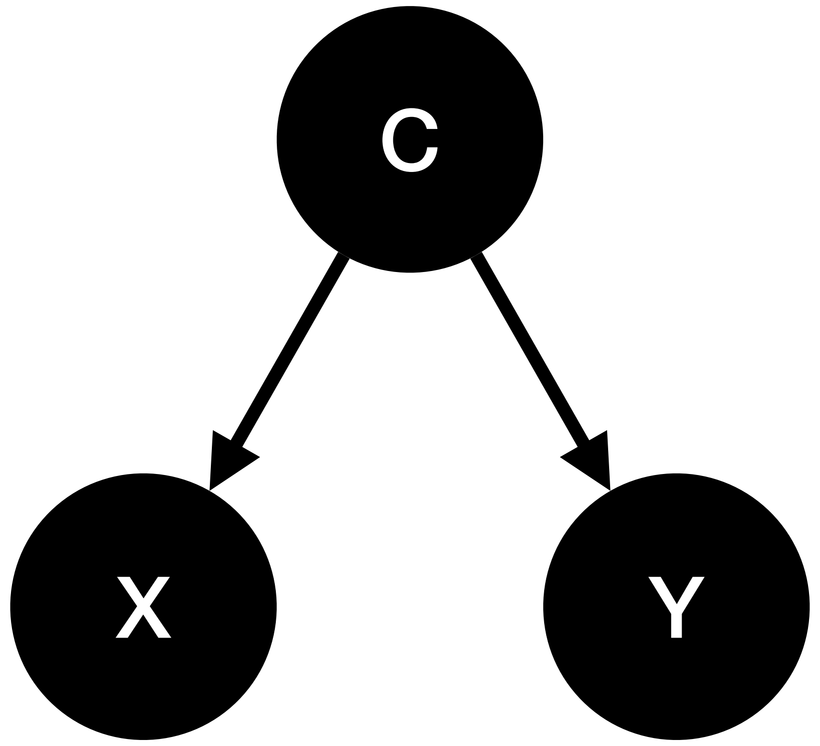
$$\text{HSIC}(Pr_{XY}), \mathcal{F}, \mathcal{G}) \triangleq \|C_{XY}\|^2_{HS}$$

where $\|C_{XY}\|^2_{HS}$ is the Hilbert-Schmitt norm of the kernel correlation matrix and $\mathcal{F}, \mathcal{G}$ are two kernels: i.e. it's the kernel trick for correlation.

Key Concept 2: Conditional (in)Dependency

- Definition of Conditional Independency

$$X \perp\!\!\!\perp Y|C \leftrightarrow P(X,Y|C) = P(X|C)P(Y|C)$$



- C=rains, X=wet sidewalk, Y=people with umbrellas

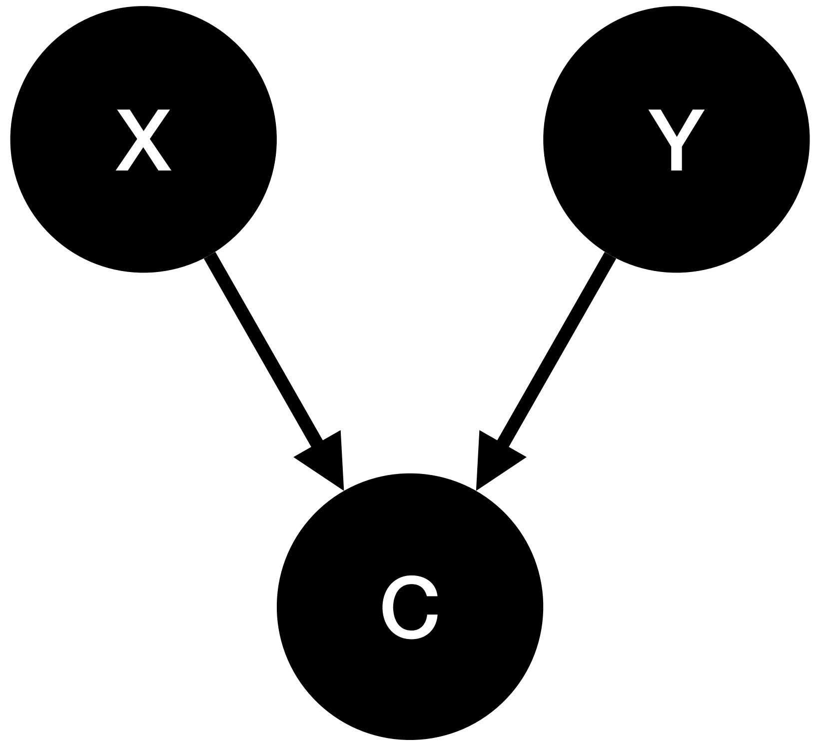
- Definition of Conditional Dependency

$$P(C|X,Y) \neq P(C|X)P(C|Y)$$

$$X \not\perp\!\!\!\perp Y|C=1 \leftrightarrow$$

$$P(X,Y) = P(X)P(Y)$$

$$P(X,Y|C=1) \neq P(X|C=1)P(Y|C=1)$$



- X=Complex Machine, Y=Inexperienced worker, C=Accident

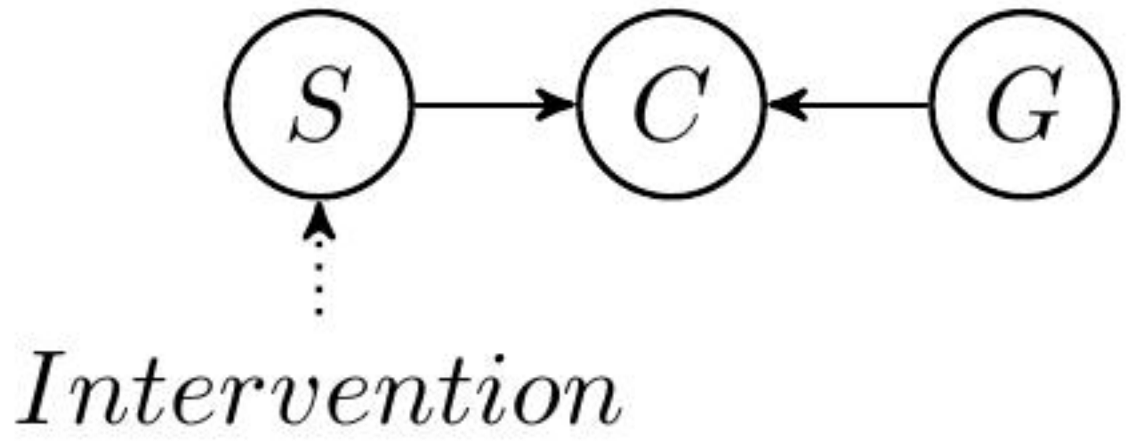
Definition of Causal Relationship

X is a direct cause of Y if when we intervene it Y's law changes

$$X \to Y \quad \text{iif}$$

$$P_{Y|do(X=a,Z=c)} \neq P_{Y|do(X=b,Z=c)}$$

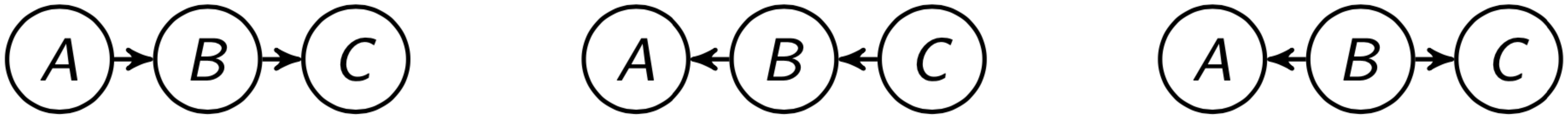
Example: Cancer, Smoking, and Genetic Factors



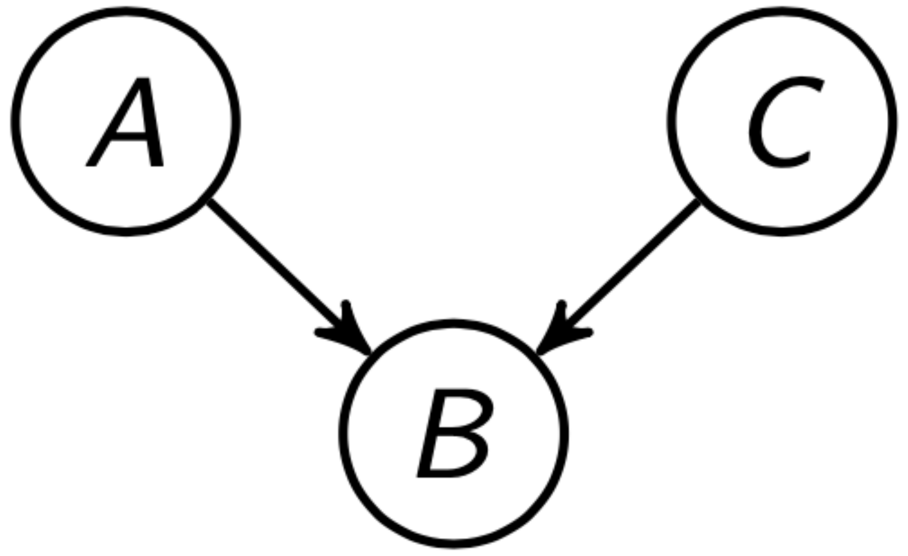
$$P_{C|do(S=1,G=0)} \neq P_{C|do(S=0,G=0)}$$

Markov Equivalences

Markov Equivalent Class: $A \perp\!\!\!\perp C | B \text{ and } A \not\perp\!\!\!\perp C$



V-Structure: $A \not\perp\!\!\!\perp C | B \text{ and } A \perp\!\!\!\perp C$

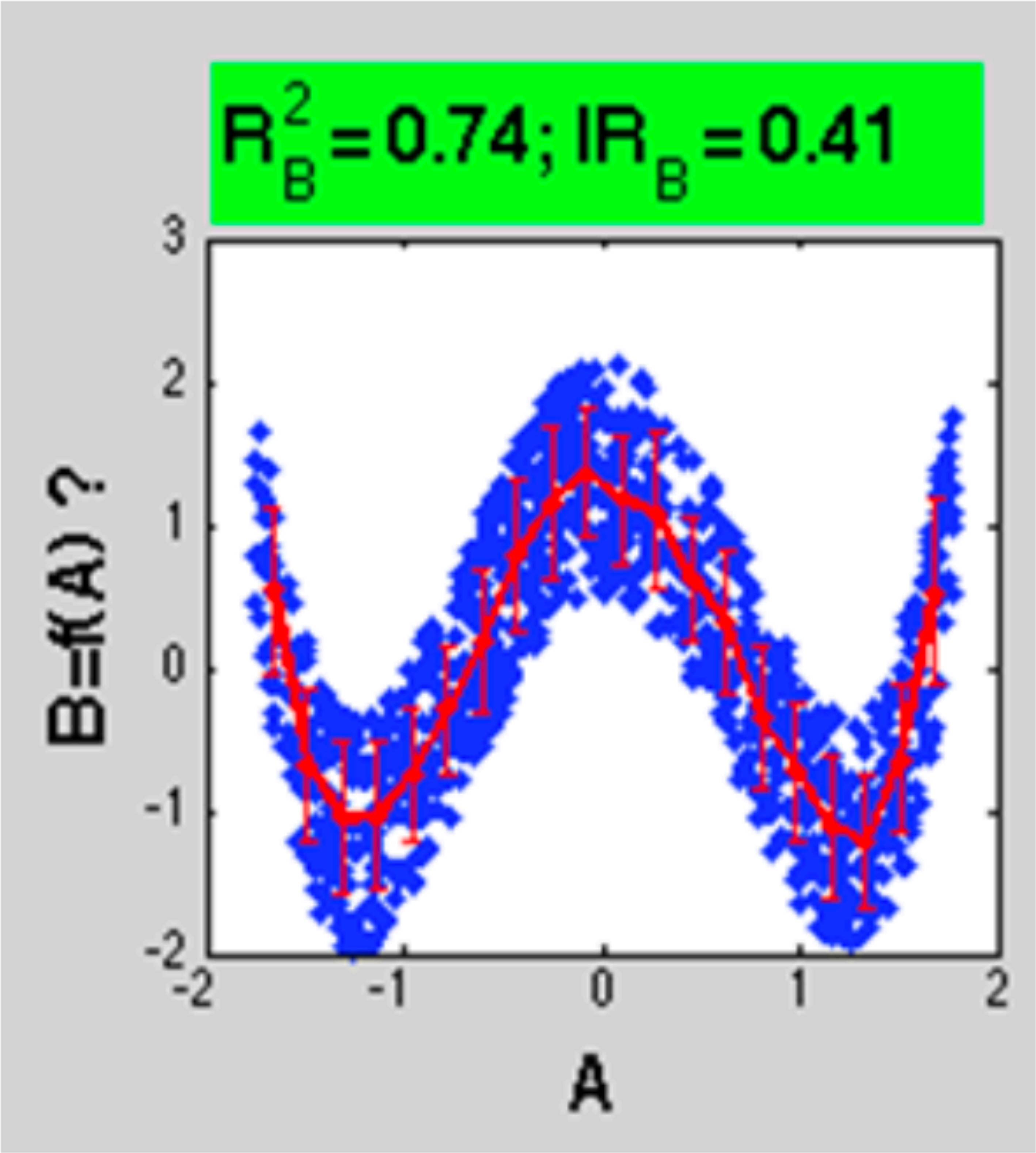


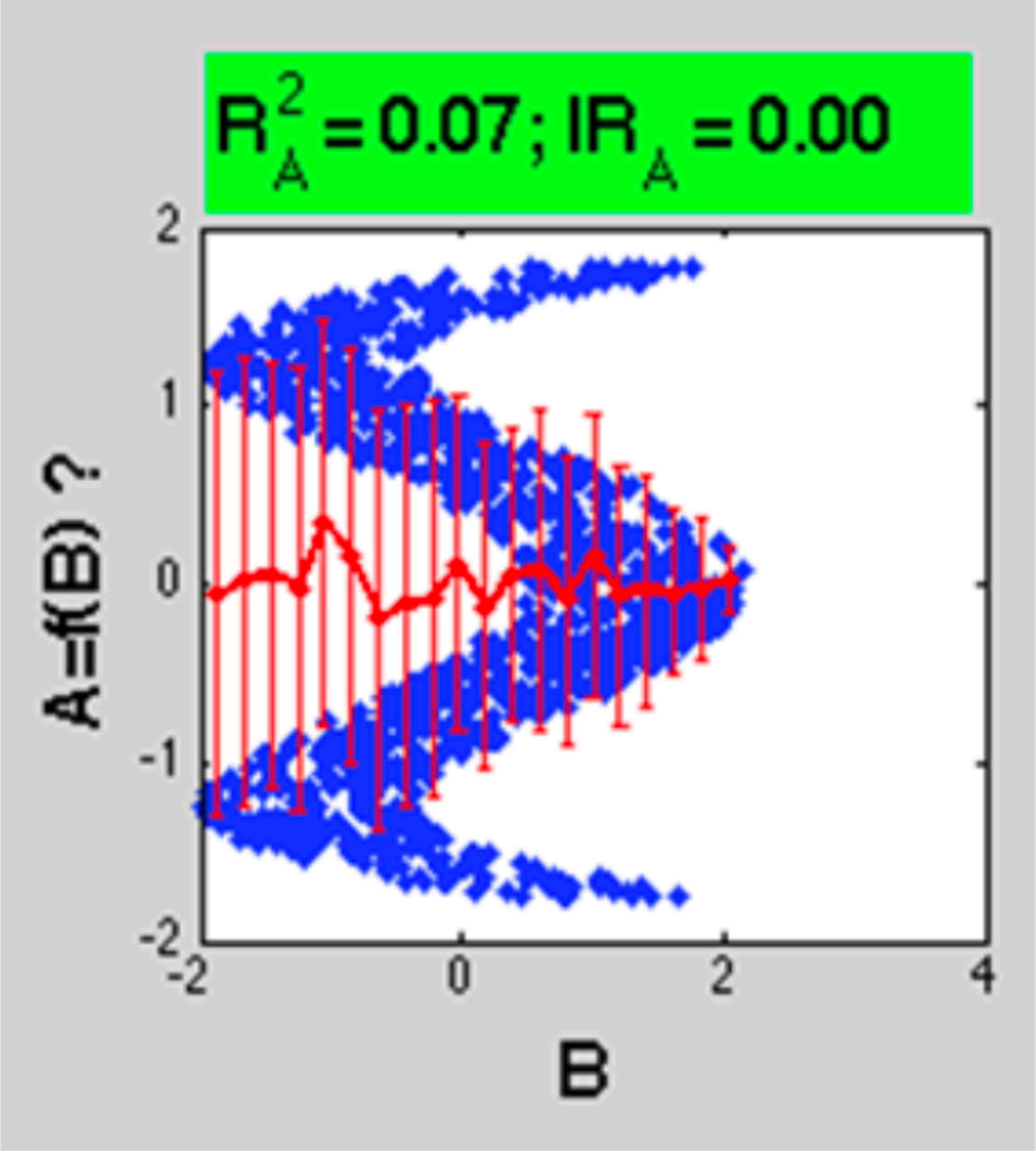
Sprites et al 00, 16

Key Concept 3: Causality with Distributional Asymmetry

- Leverages Occam's Principle

  The causal model as the simplest explaining the data (Janzig 19)

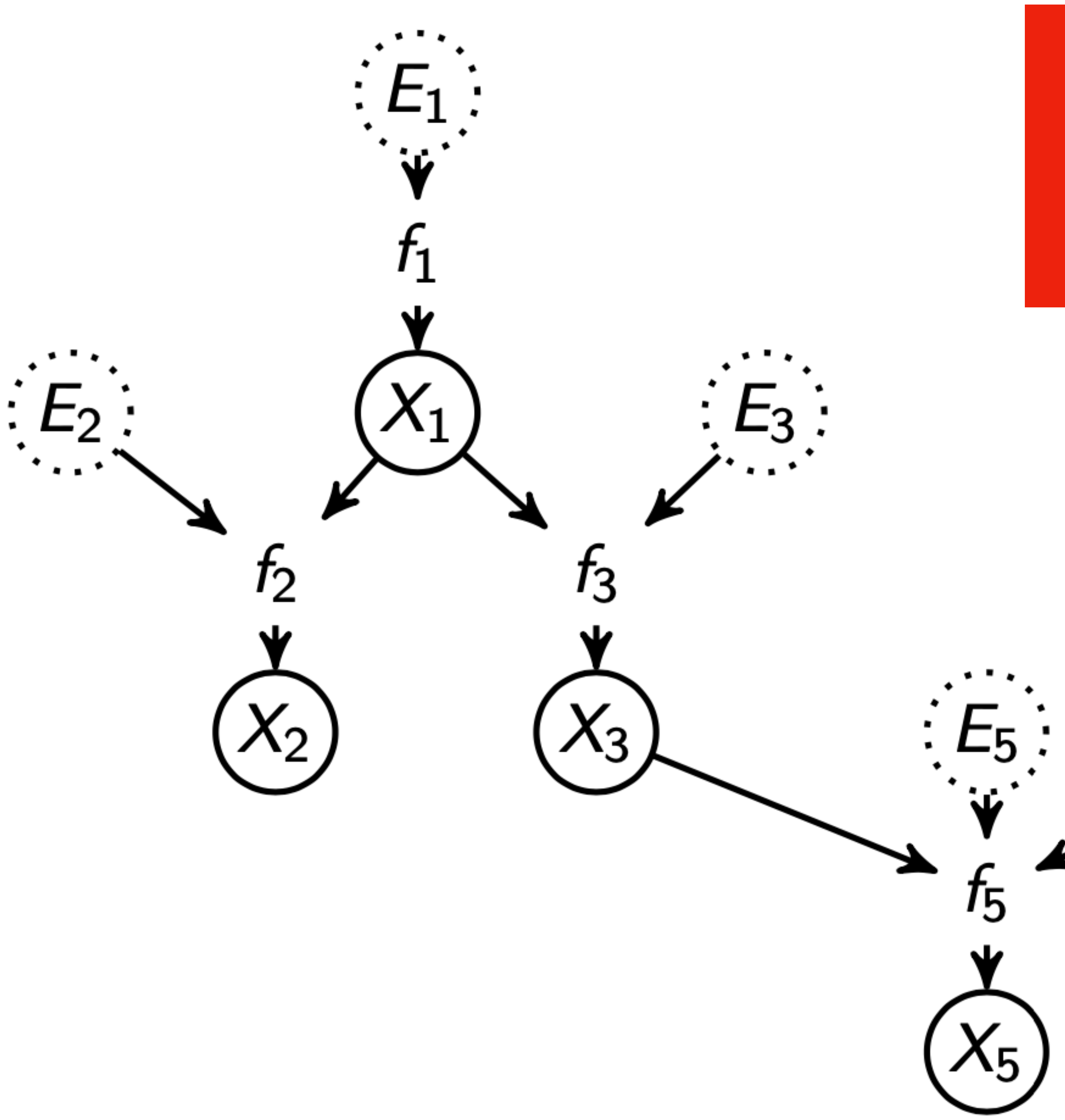




Framework: Functional Causal Models (FCMs)

- Given $X_1, \ldots, X_d$ where $X_i = f_i(X_{Pa(X_i)}, E_i)$, with $X_{Pa(X_i)}$ the parents or causes of X_i, a deterministic function f_i, and E_i an error representing independent random variable.

But, what do we need for this system to represent a causal model?

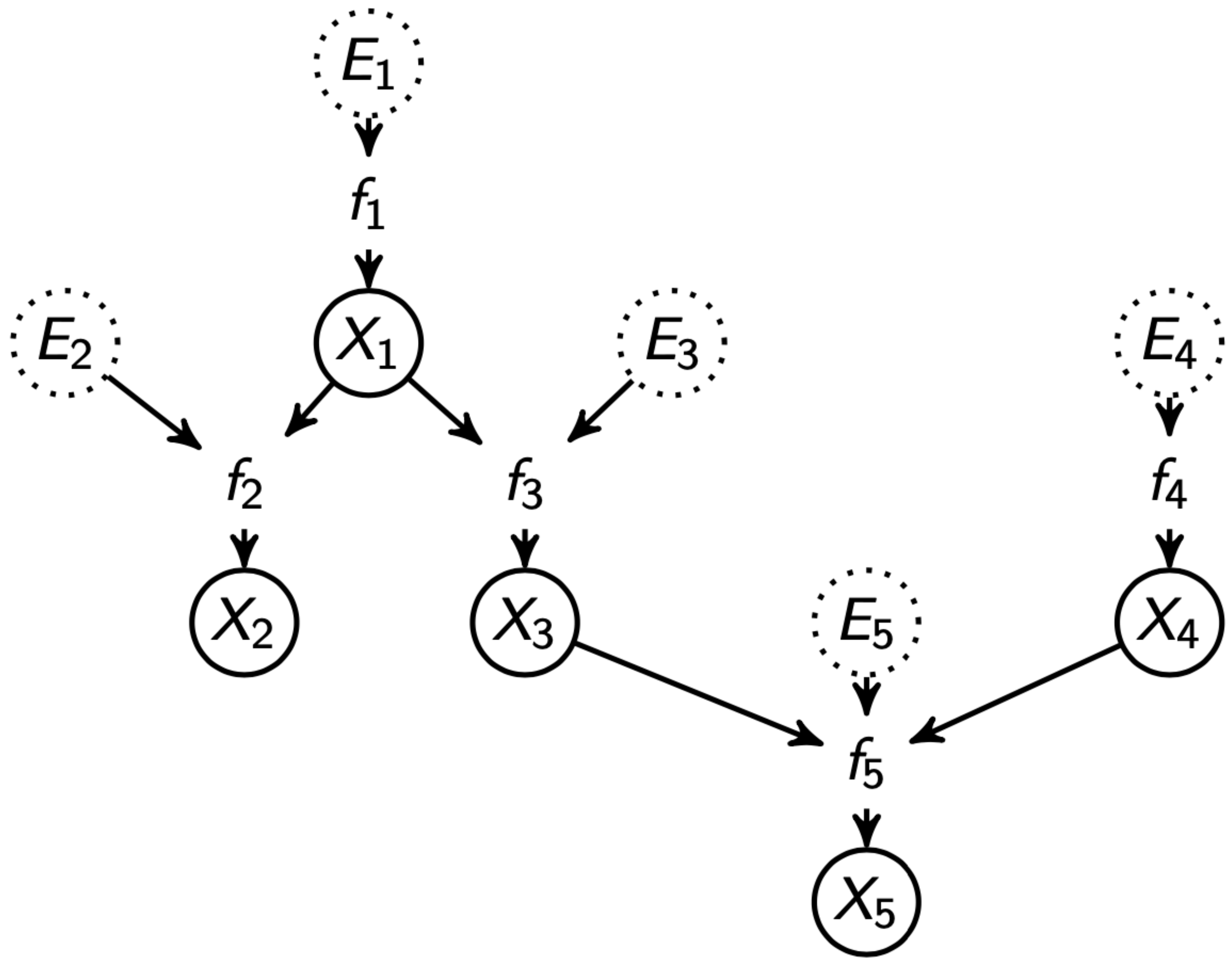


$$
\begin{cases}
X_1 = f_1(E_1) \\
X_2 = f_2(X_1, E_2) \\
X_3 = f_3(X_1, E_3) \\
X_4 = f_4(E_4) \\
X_5 = f_5(X_3, X_4, E_5)
\end{cases}
$$

$$P(X_1, \ldots, X_d) = \Pi P(X_i | X_{Pa(X_i)})$$

Conditions for Causal Model Representation

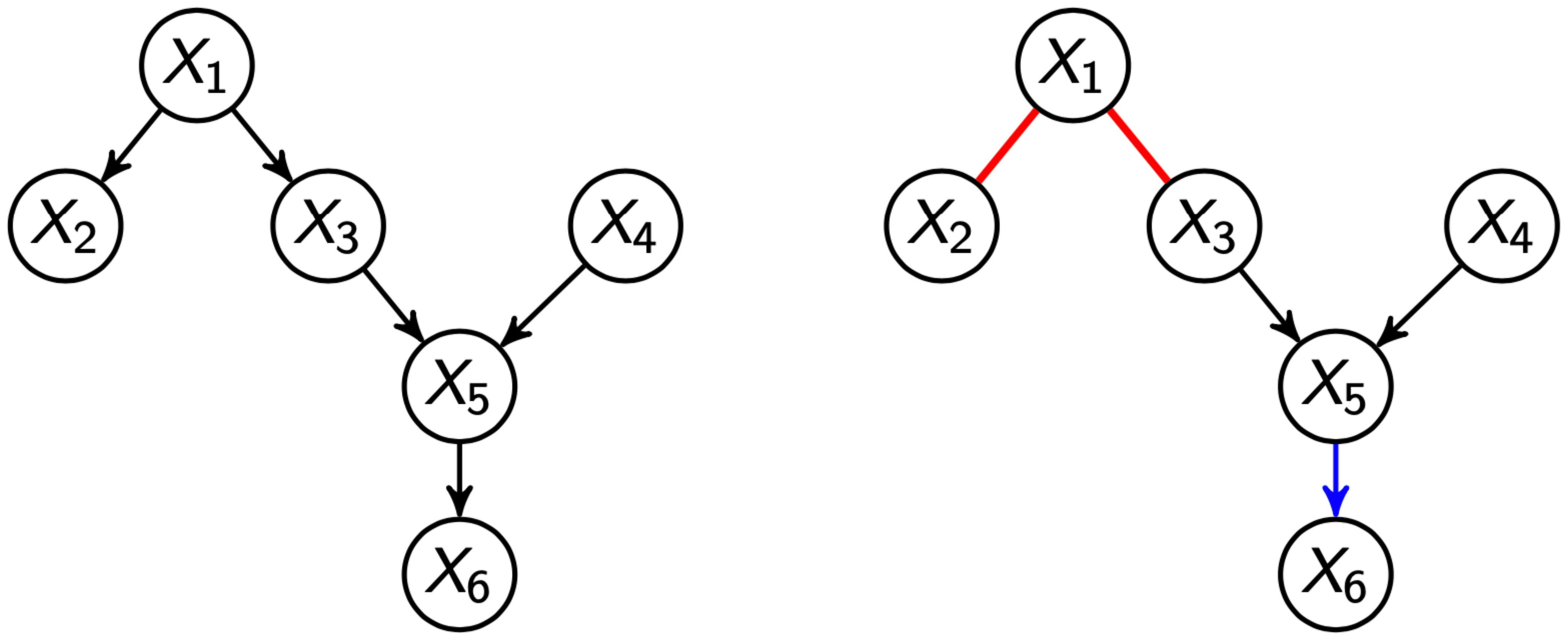
- Causal Sufficiency: no unobserved confounders
- Causal Markov: all d-separations in the causal graph G imply conditional independencies in the observational distribution P
- Causal Faithfulness: all conditional independencies in P imply d-separations in the causal graph G



How Do We Infer the Causal Model From Data?

Key Approach 1: Constraint-Based Methods

- Constraint-based methods, through V-Structures and constraint propagation, output a CPDAG (Completed Partially Directed Acyclic Graph).



(a) The exact DAG of $\mathcal{G}$. (b) The CPDAG of $\mathcal{G}$.

- Examples: Peter-Clark Algorithm (PC) and it's extensions such as PC-Hist (Spires et al 00, Zhang et al 12)

Key Approach 2: Score-Based

- Use an objective function to optimise the graph. For instance the Bayesian information criterion

$$BIC(\mathcal{G}) = -2\ln(L) + k\ln(n)$$

- with L the likelihood of the model, k number of parameters, and n the number of samples
- We optimise the sample with operations such as:
    - Add an edge
    - remove an edge
    - revert and dee
- An algorithm for this are Greedy Equivalence Search (GES) by Chickering et al 02.

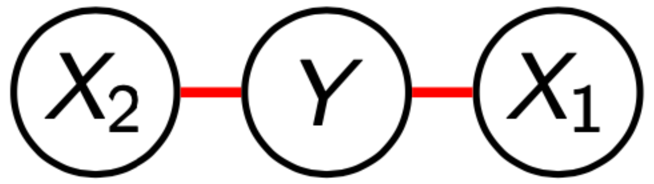
Key Approaches 1 and 2

- Limitations
  - Computational cost depending on the test/scoring/loss
  - Data hungry
  - Identifiability issues
  - Example:

$$X_1, E_{X_1}, E_{X_2} \sim U(0,1) X_1 \perp\!\!\!\perp E_{X_1}, Y \perp\!\!\!\perp E_{X_2}$$

$$Y \leftarrow 0.5X_1 + E_{X_1}$$

$$X_2 \leftarrow Y + E_{X_2}$$



$X_1 \perp\!\!\!\perp X_2 | Y$. No V-struture

Key Approach 3: Global Optimisation

- Assuming linear causal mechanisms, the system can be formulated in terms of linear equations

$$X = B^T X + E$$

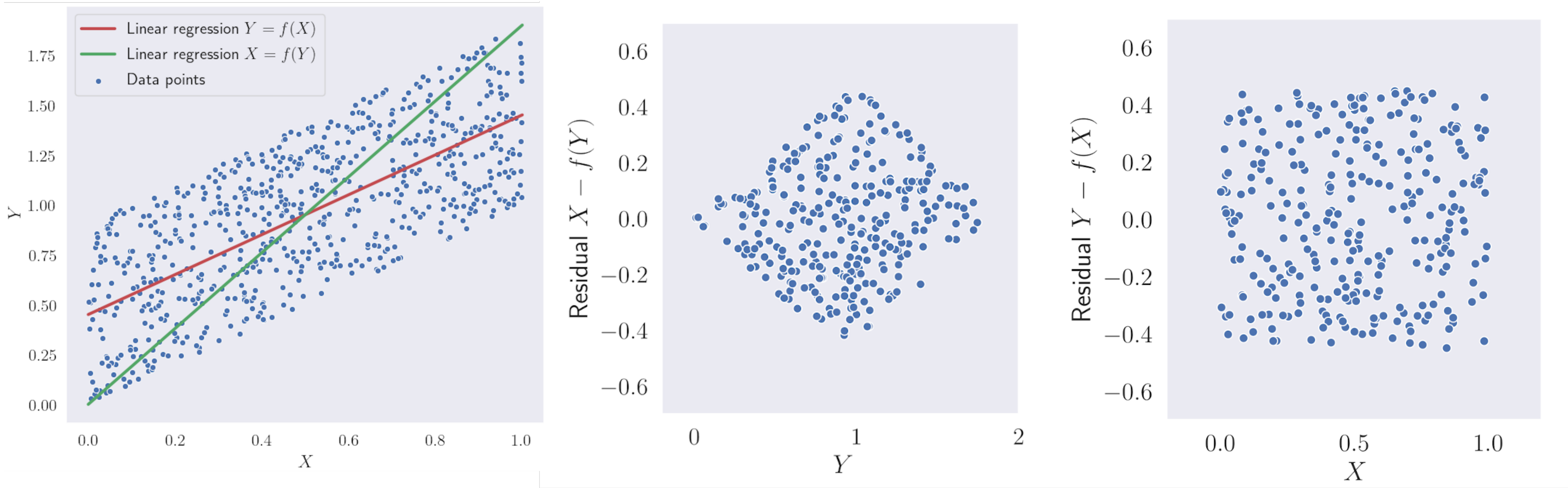
  where the triangular B matrix can be estimated through ICA for LinGAM (Shimizu 06, Hyvarien 99)

- This also can be done in terms of graphical models (Pearl 09, Friedman 08)

  For instance with Max-Min Hill-Climbing (MMHC) by Tsamardinos (06) and concave penalised Descent (CCDr) by Aragam (15)

Key Approach 4: Exploiting Asymmetries

- If no v-structure is available and causal discovery with 2 variables is hard, we can leverage asymmetries in the distributions . For instance with the Additive Noise Model (ANM) of Hoyer (09)



Original data

Residuals of X=g(Y)

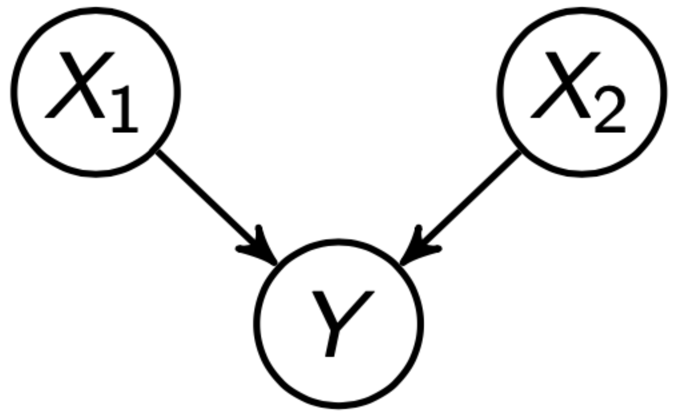
Residuals of Y=f(X)

Key Approach 4: Exploiting Asymmetries

Limitations

- Restrictive assumptions on the type of causal mechanisms
- Conditional independence is not taken into account

$$X_1, E_{X_1}, X_2 \sim \mathcal{N}(0,1) X_1 \perp\!\!\!\perp E_{X_1}, Y \perp\!\!\!\perp E_{X_1}$$
$$Y \leftarrow 0.5 X_1 + X_2 + E_1$$



(X1,Y) and (X2,Y) are a perfectly symmetric pairwise distribution after rescaling. However, $X_1 \not\perp\!\!\!\perp X_2 | Y$ a v-structure is at the origin of the data.

Key Approach 5: Machine Learning Base

Guyon et al 2014—2015

- Pair Cause-Effect Challenges
  - Gather data: a sample is a pair of variables (Ai,Bi)
  - Its label ℓ_i is the "true" causal relation (e.g. age "causes" salary)
- Input

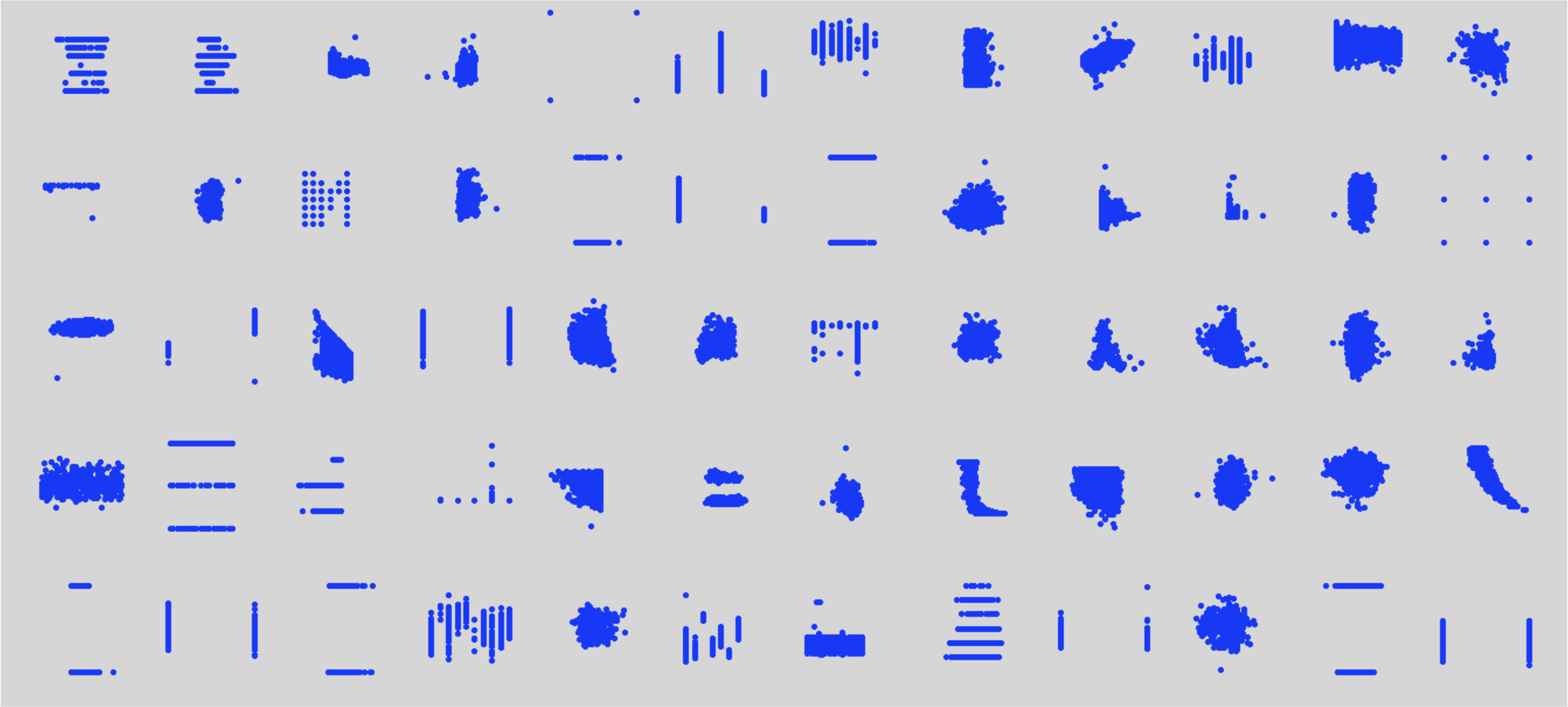
$$\mathcal{E} = \{(A_i, B_i, \ell_i), \ell_i \ in \ \{\rightarrow, \leftarrow, \perp\!\!\!\perp\}\}$$

Example A_i, B_i	Label ℓ_i
A_i causes B_i	$\rightarrow$
B_i causes A_i	$\leftarrow$
A_i and B_i are independent	$\perp\!\!\!\perp$

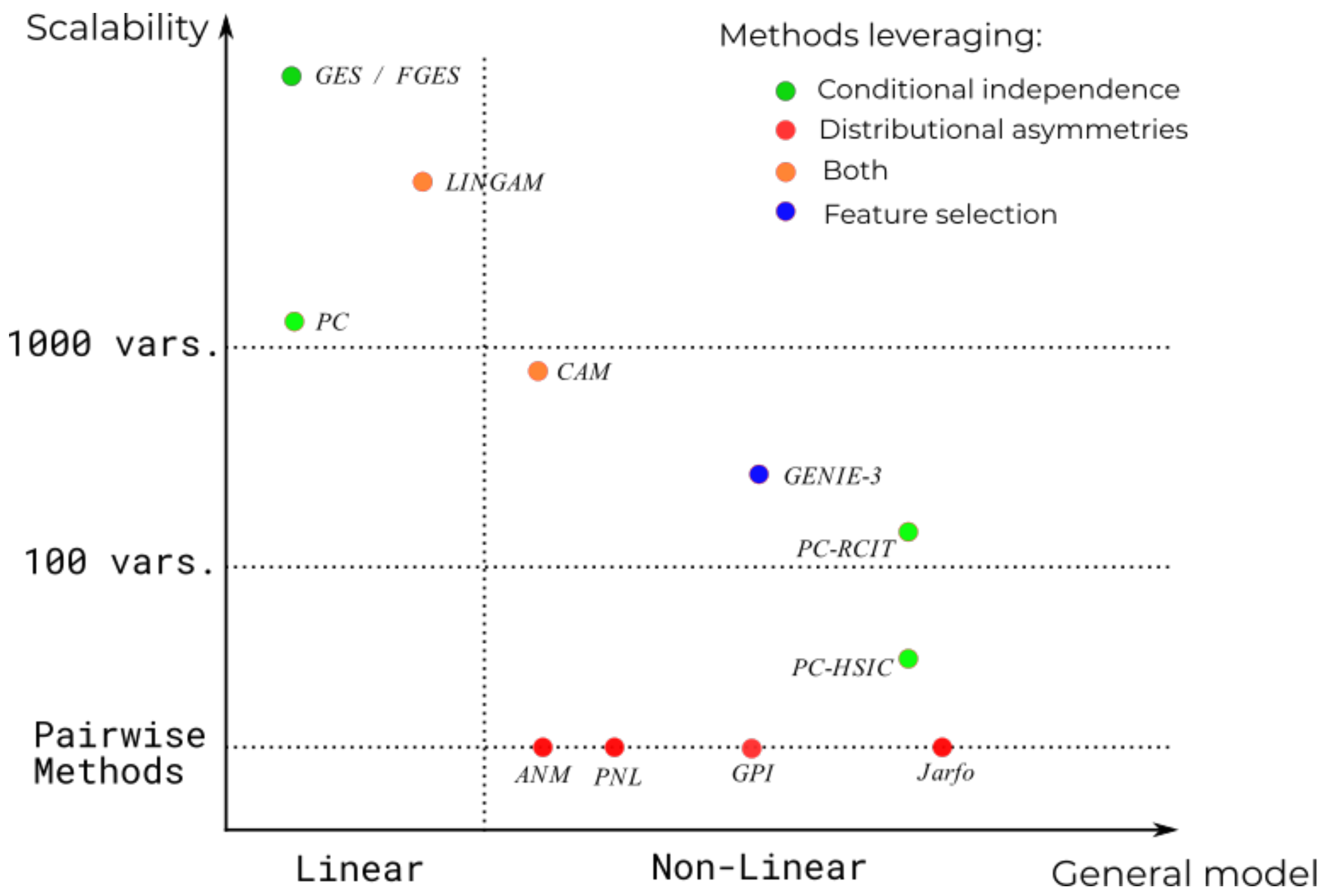
- Output: $(A, B) \rightarrow \ell$

Key Approach 5: Machine Learning Base

Guyon et al 2014—2015



Summary for “Key Approaches”



A Python Package for Causal Discovery

All the presented framework is available on GitHub at :

https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox

It includes multiple algorithms as well as tools for graph structure.

Published in Kalainathan Goudet 2019 JMLR - Open Source Software

A Python Package for Causal Discovery

All the presented framework is available on GitHub at :

https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox

It includes multiple algorithms as well as tools for graph structure.

Published in Kalainathan Goudet 2019 JMLR - Open Source Software