OxML 2024

Causal Representation Learning and Related Machine Learning Tasks

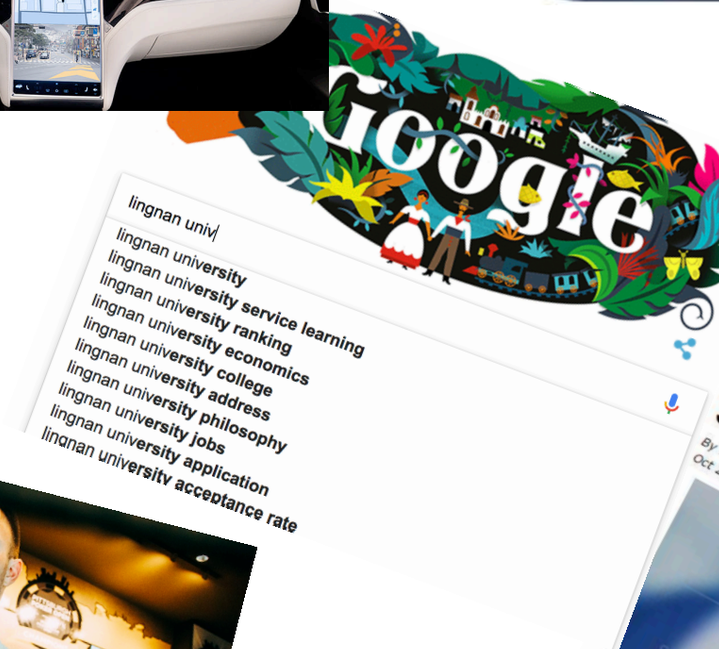
Kun Zhang

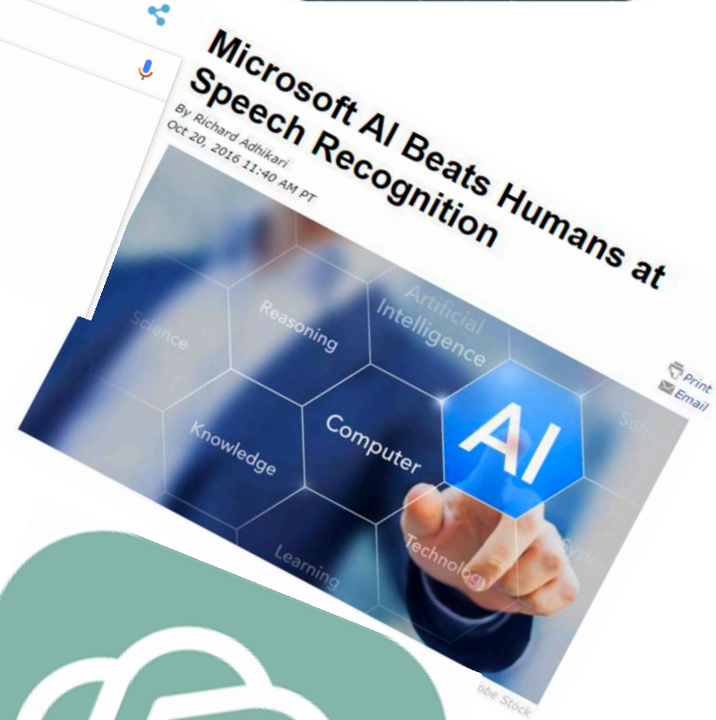
Carnegie Mellon University

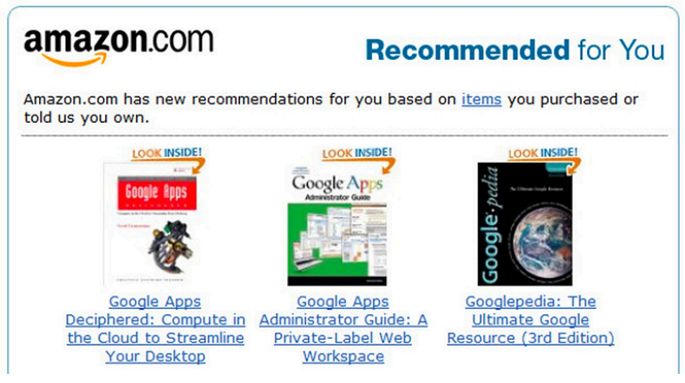
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Thanks to:
Biwei Huang, Mingming Gong, Shaoan Xie, Yujia Zheng, Ignavier Ng, Weiran Yao, Xinshuai Dong, Haoyue Dai, Petar Stojanov, Zeyu Tang…
Clark Glymour, Peter Spirtes, Bernhard Schölkopf, Aapo Hyvärinen, Ruichu Cai, Jiji Zhang, Joseph Ramsey…


TESLA

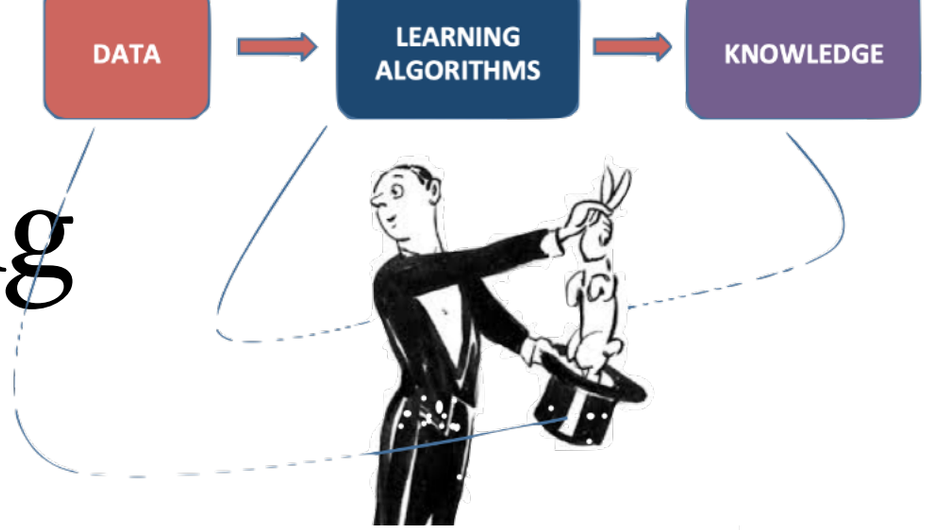

lingnan univ
lingnan university
lingnan university service learning
lingnan university ranking
lingnan university economics
lingnan university college
lingnan university address
lingnan university philosophy
lingnan university jobs
lingnan university application
lingnan university acceptance rate

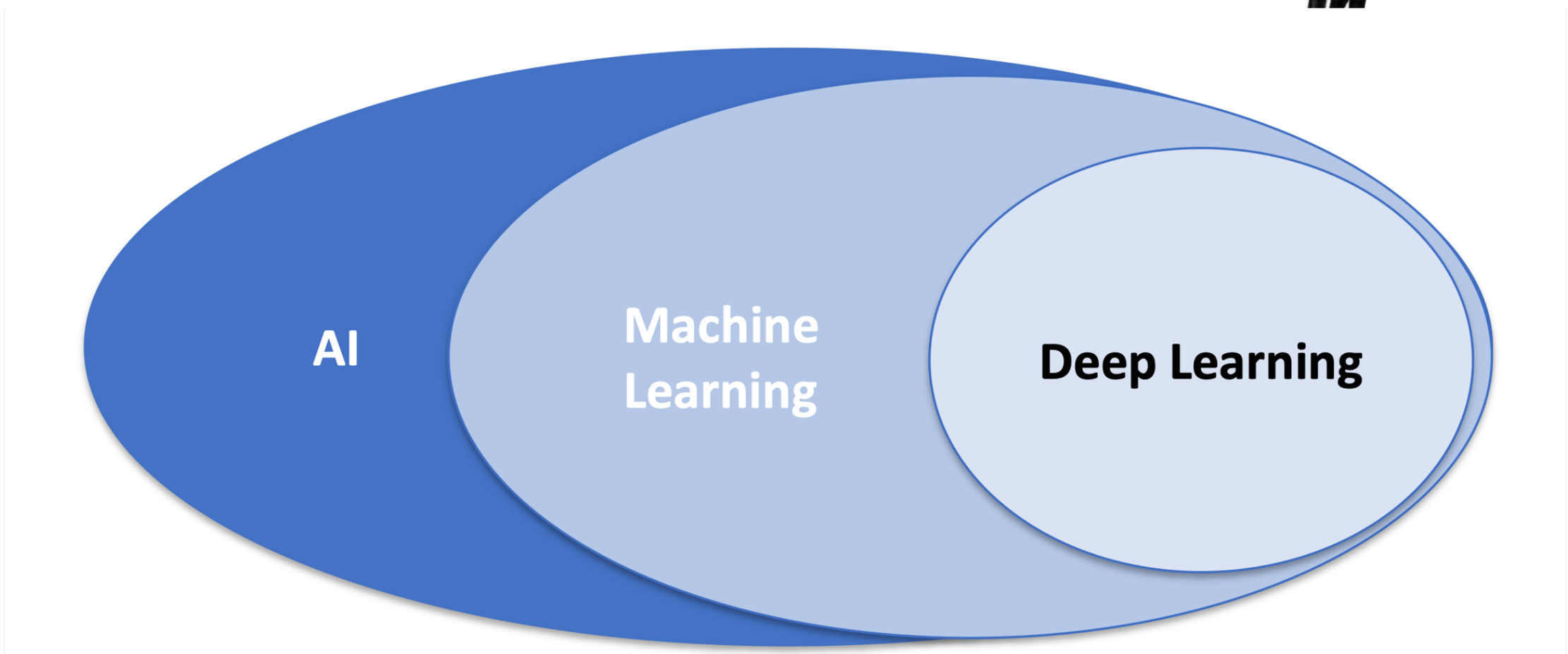

Microsoft AI Beats Humans at Speech Recognition
By Richard Adhikari
Oct 20, 2016 11:40 AM PT
Science
Reasoning
Artificial Intelligence
Knowledge
Computer
AI
Technology
Learning


amazon.com
Recommended for You
Amazon.com has new recommendations for you based on items you purchased or told us you own.
LOOK INSIDE!
Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop
LOOK INSIDE!
Google Apps Administrator Guide: A Private-Label Web Workspace
LOOK INSIDE!
Googlepedia: The Ultimate Google Resource (3rd Edition)


1 AlphaGo Fan
2 AlphaGo Lee
3 AlphaGo Master
4 AlphaGo Zero
5 AlphaZero

AI, ML, and Deep Leaning



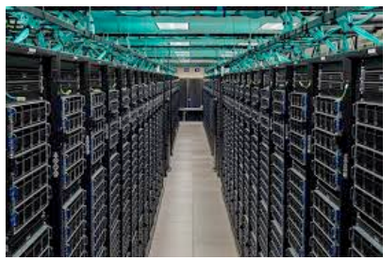


The capacity of a machine to imitate intelligent human behavior

Learning without being explicitly programmed (from data)

ML Facilitates the Second Scientific Revolution?

- Analogy to Scientific Revolution
  - By Copernicus, Galilei, Newton, Bacon, Harvey…
  - Book production, observational data, the ability to do some experiments, basic inference rules…
  - Quantitative vs. qualitative view of nature; new experimental, scientific method seeking definite answers; "how" instead of "why"…
- Available: data, statistical tools, computational resources…
- Goals? Learning paradigms? Methodology?
- ML (including causality) will impact each scientific discipline, every industry, and human society



Let’s Look at AI Image Generator

- Prompt: a peacock eating ice cream

By Stable Diffusion One Year Ago

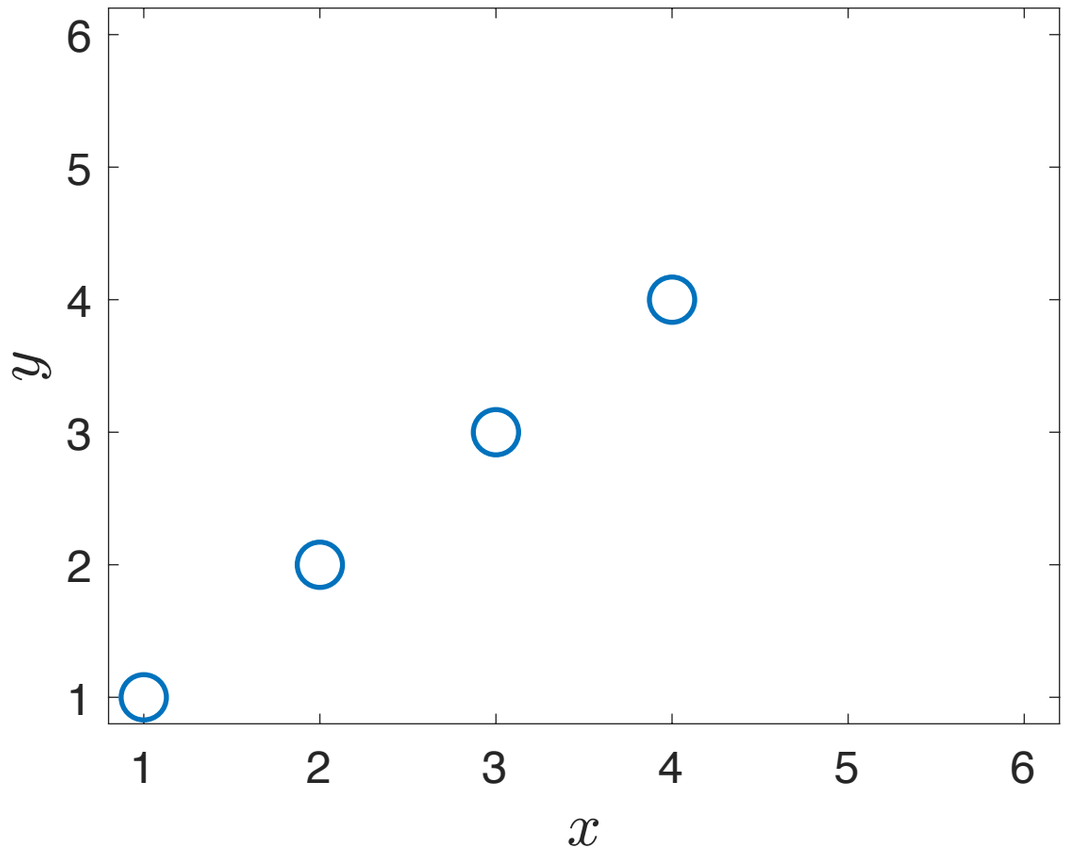
- Prompt: a peacock eating ice cream

By DALL·E 3 Last Week

- Prompt: a peacock eating ice cream

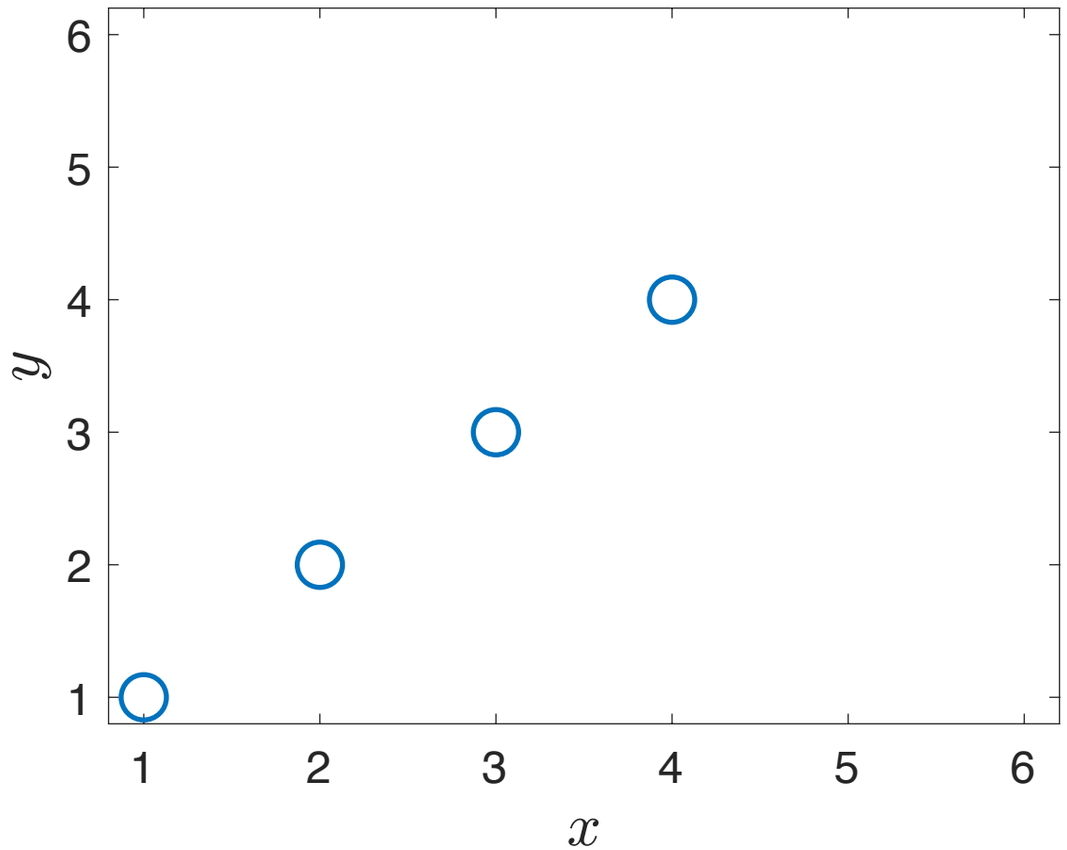
Representations Are Essential: A Simple Problem

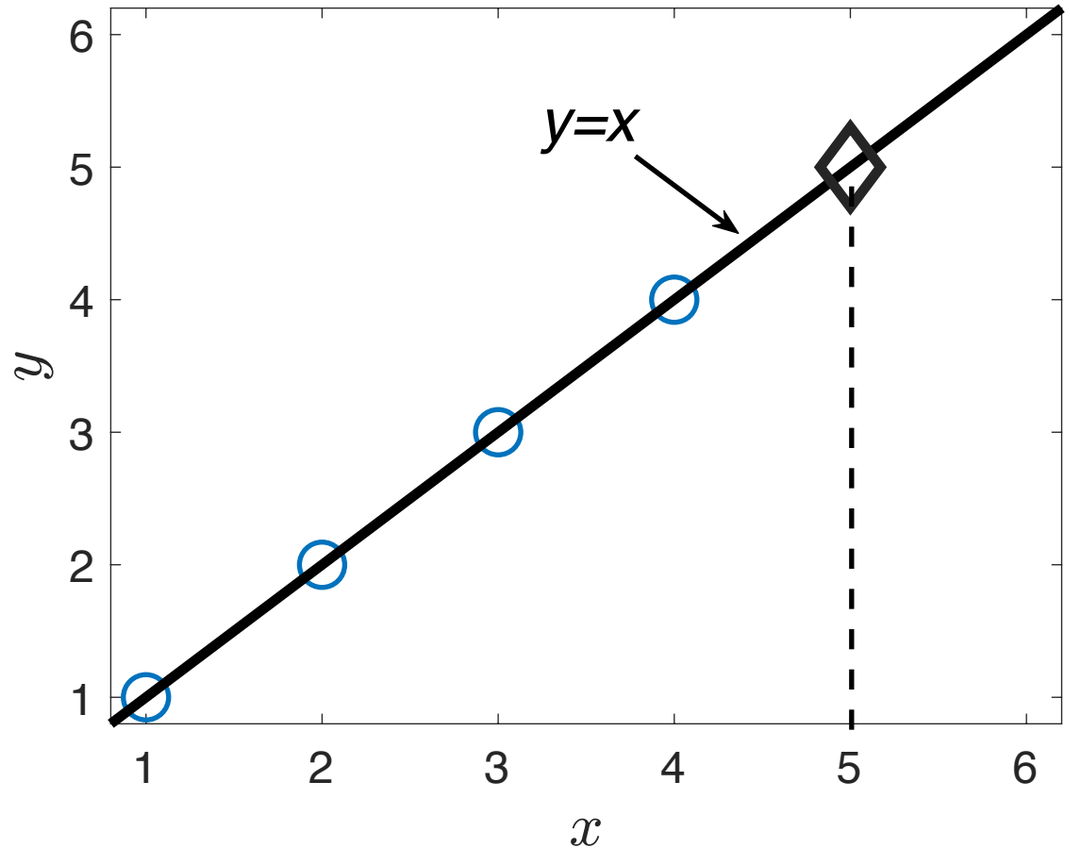
- Consider this simple Out-of-distribution (OOD) generalization problem



Representations Are Essential ?

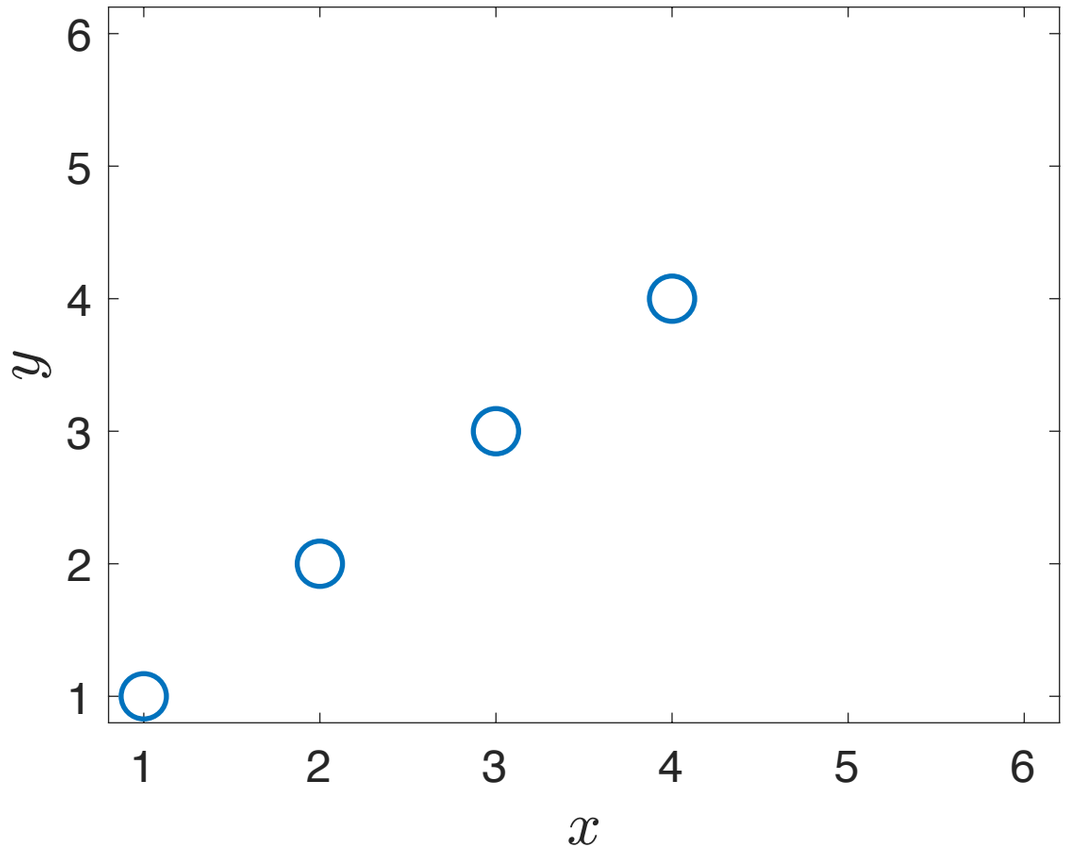
- Consider this simple Out-of-distribution (OOD) generalization problem

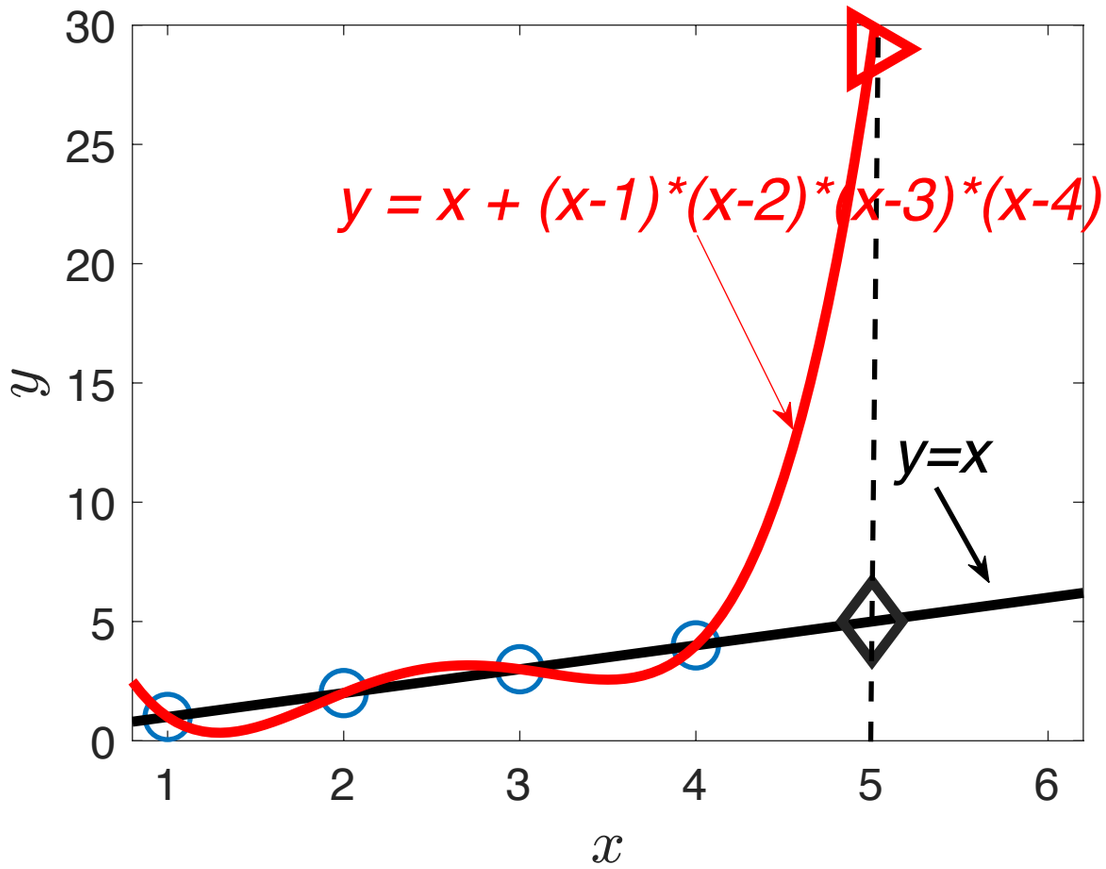




Representations Are Essential !

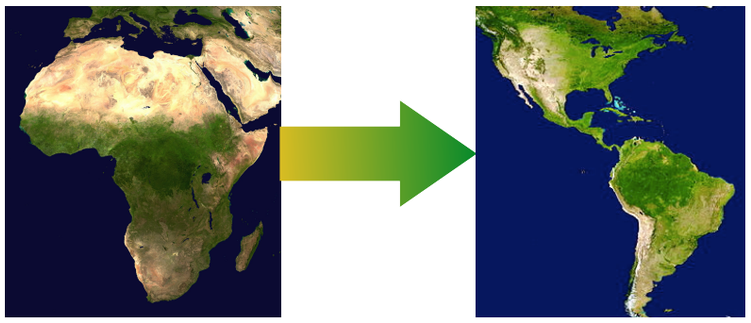
- Consider this simple Out-of-distribution (OOD) generalization problem



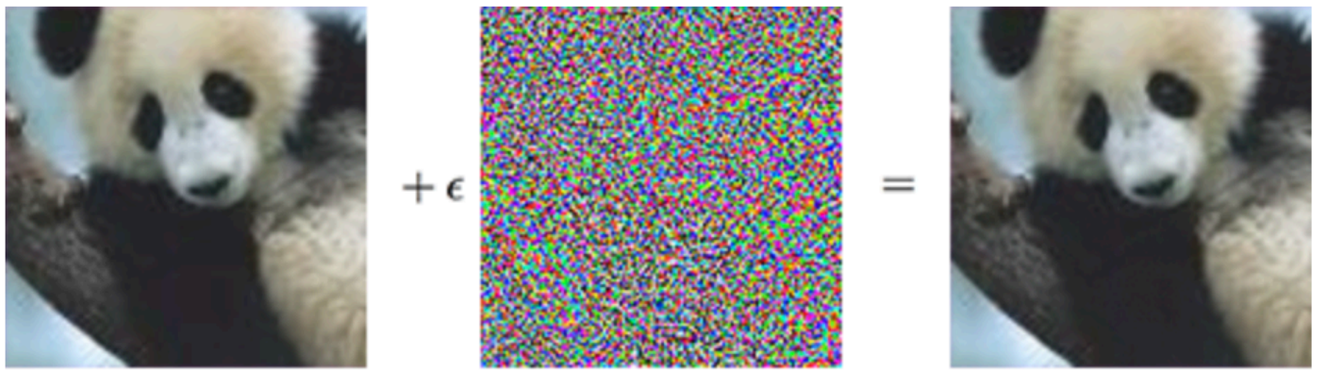


Good Representations Are Needed…

- Generalization/adaptation, decision-making, fairness, recommendations, healthcare, generative AI…



- Dealing with adversarial attacks?

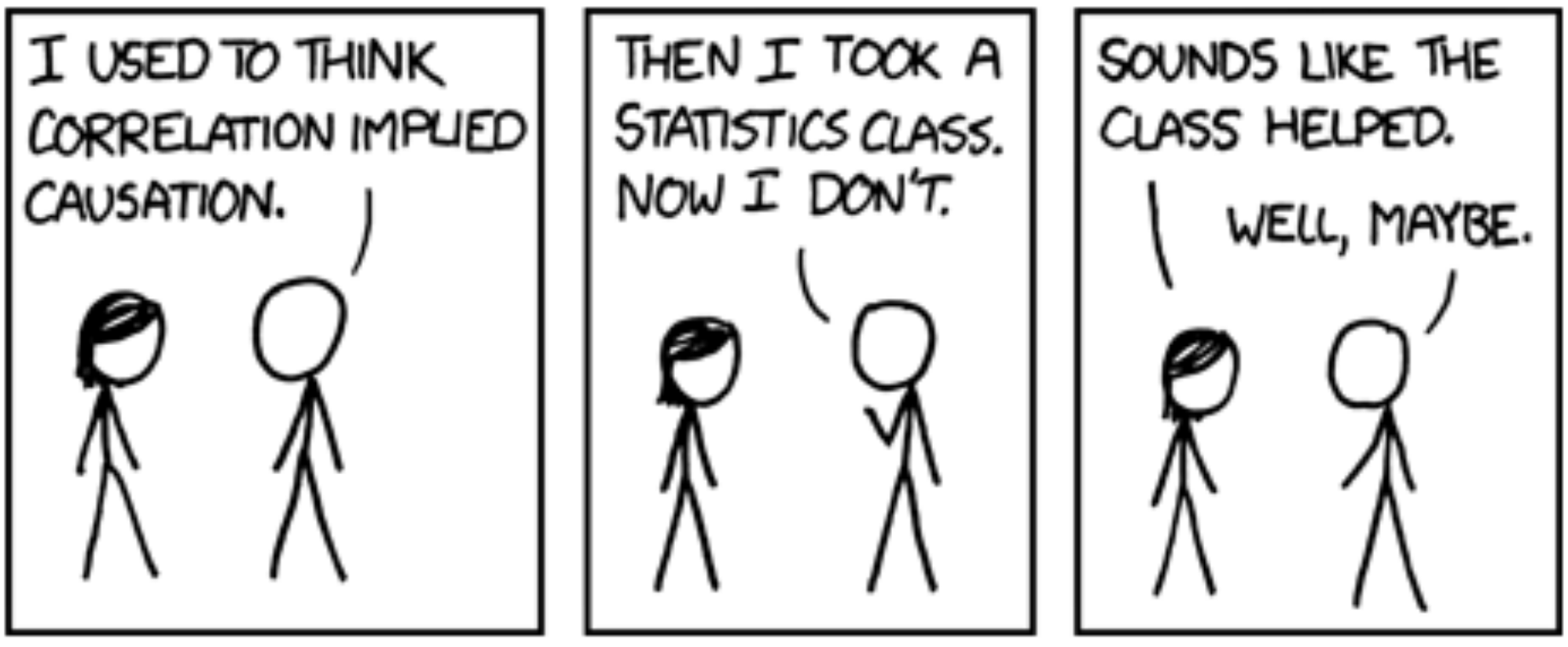


(Goodfellow et al., 2014)

An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

Causality vs. Dependence

- Causality ➜ dependence ! Dependence ➜ causality



(http://imgs.xkcd.com/comics/correlation.png)

X and Y are associated iff

$\exists x_1 \neq x_2 \; P(Y|X=x_1) \neq P(Y|X=x_2)$

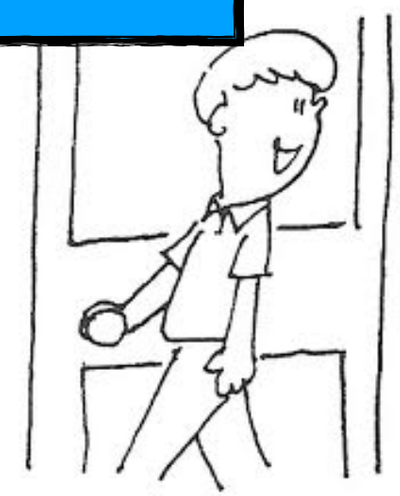
X is a cause of Y iff

$\exists x_1 \neq x_2 \; P(Y|\text{do } X=x_1) \neq P(Y|\text{do } X=x_2)$

Classic Ways to Find Causal Information (i.i.d. Case)

- What if X and Y are dependent?
- What if you change X and see Y also changes?
  - A **manipulation/ intervention** directly changes only the target variable X

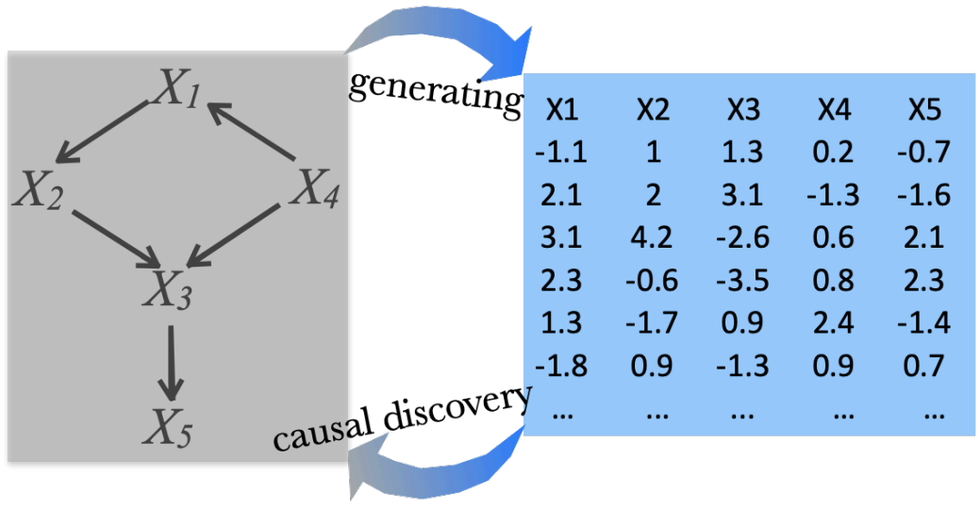
An intervention on X changes only the target variable X, leaving any other variable in the system unchanged, at least for the moment.

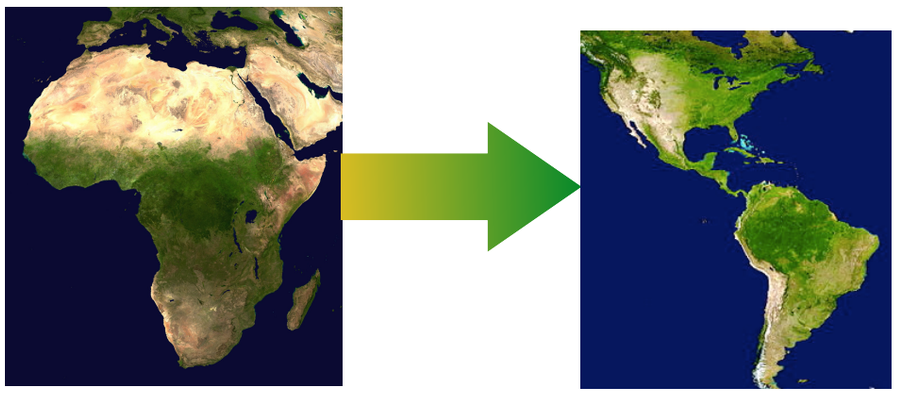


** Definition of "interventions"*

Outline

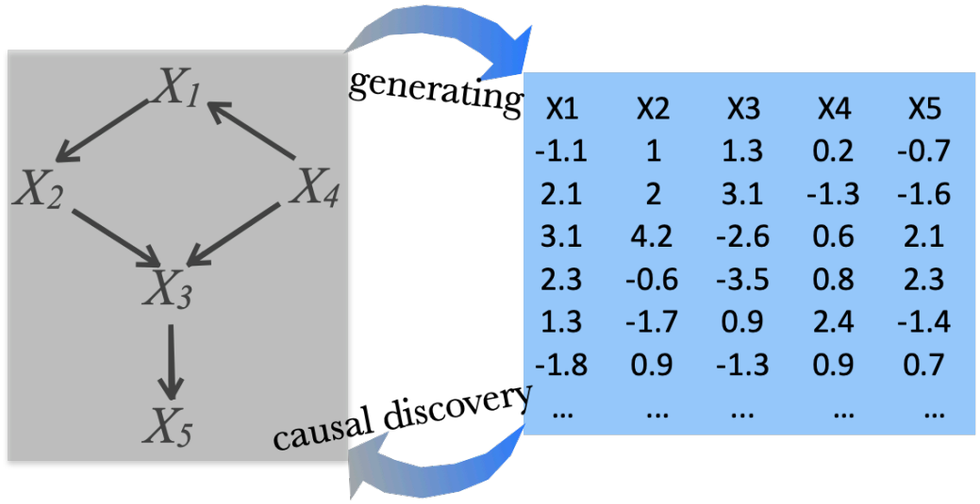
- Causal thinking
- Causal representation learning: IID case
- Causal representation learning from time series
- Causal representation learning from heterogeneous/ nonstationary data
  - Transfer/adaptive learning & generative AI

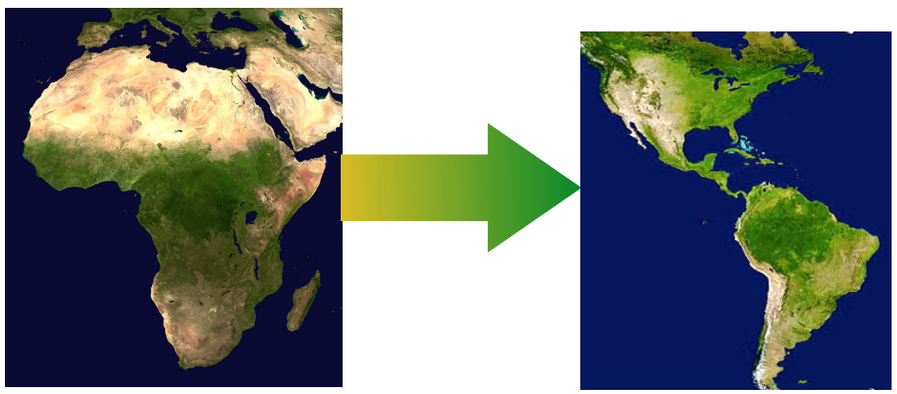




Outline

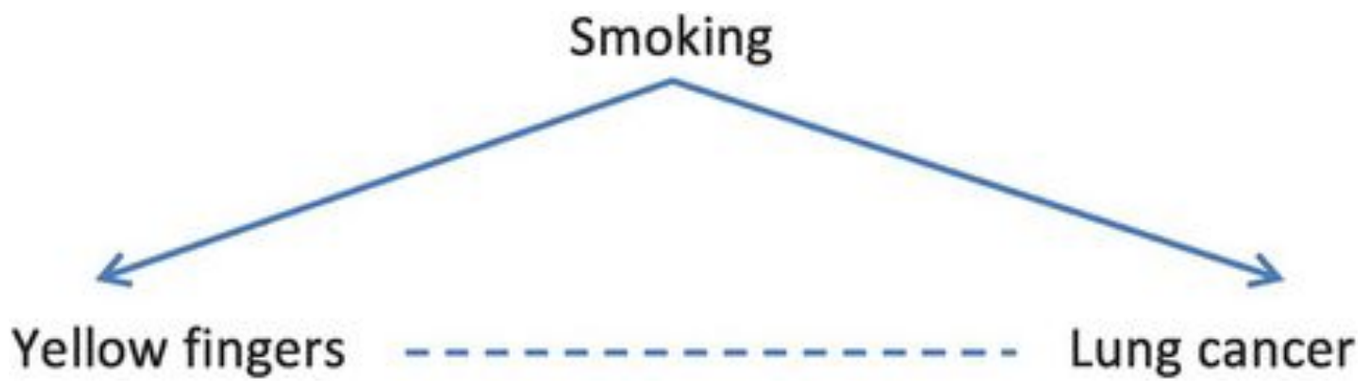
- Causal thinking
- Causal representation learning: IID case
- Causal representation learning from time series
- Causal representation learning from heterogeneous/ nonstationary data
- Transfer/adaptive learning & generative AI





Causal Thinking: Making Changes?

- Dependence vs. causality

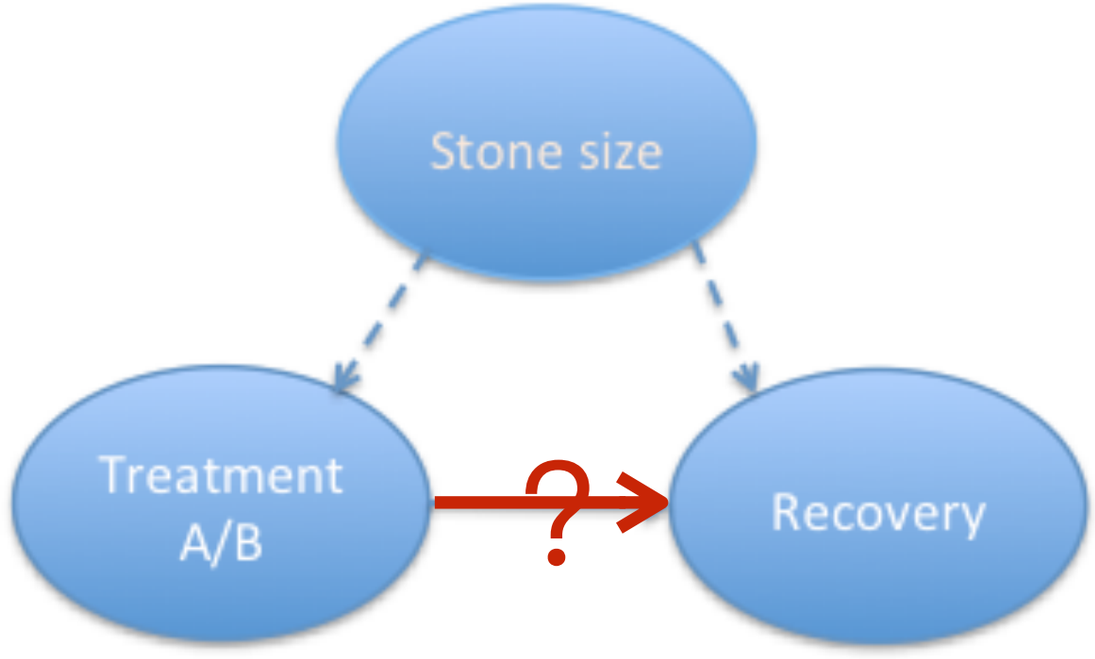


Causal Thinking: Why “Paradox”?

- Dependence vs. causality

- Simpson’s paradox

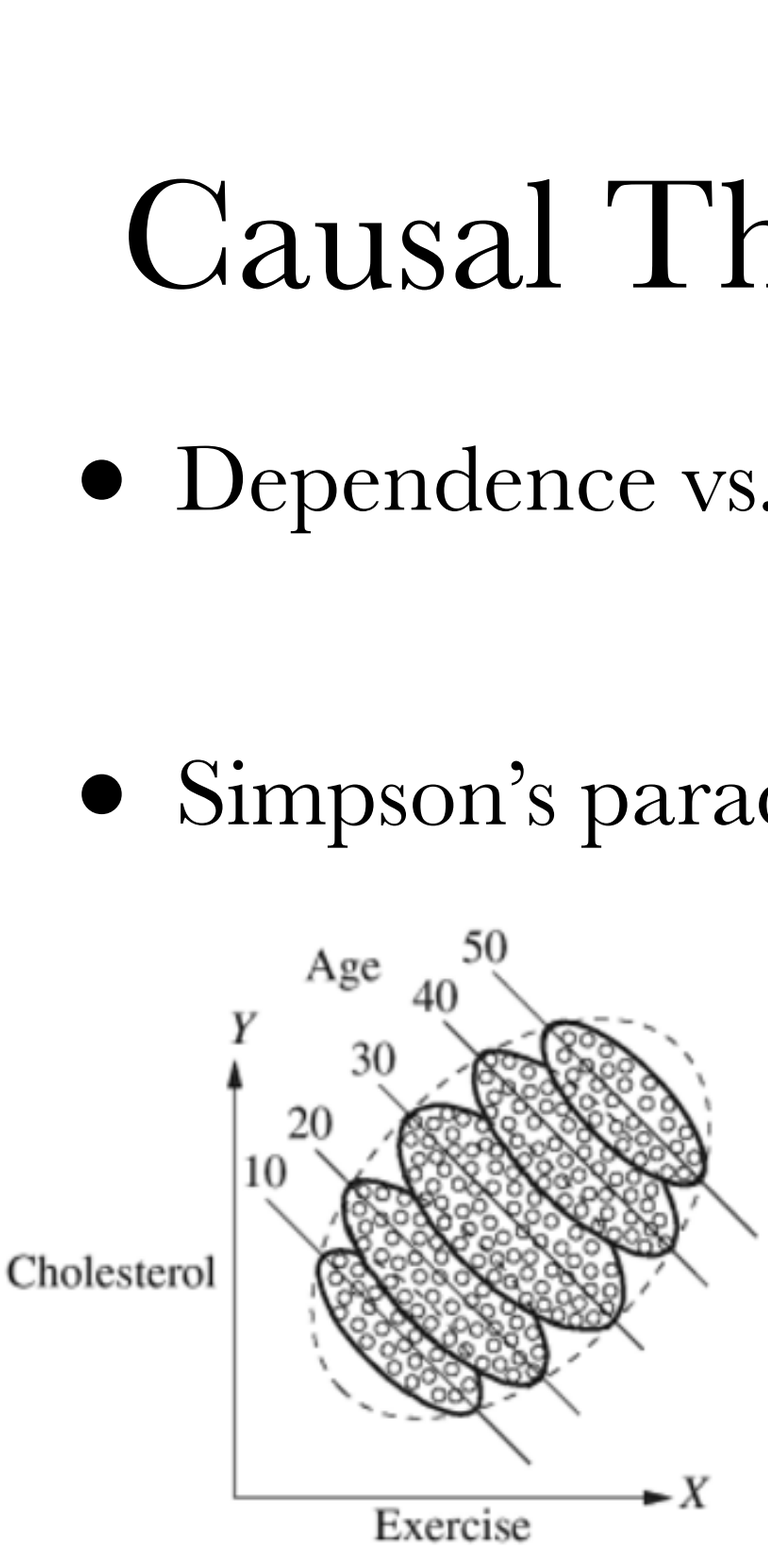
	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

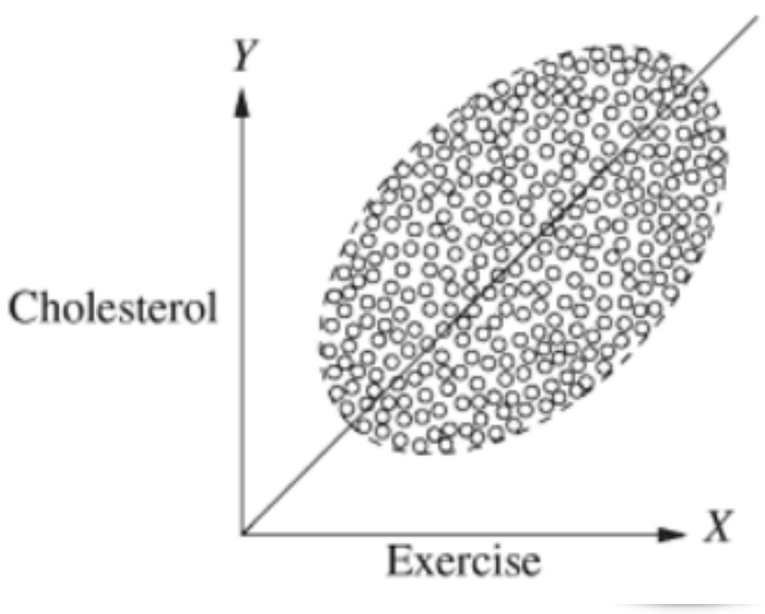


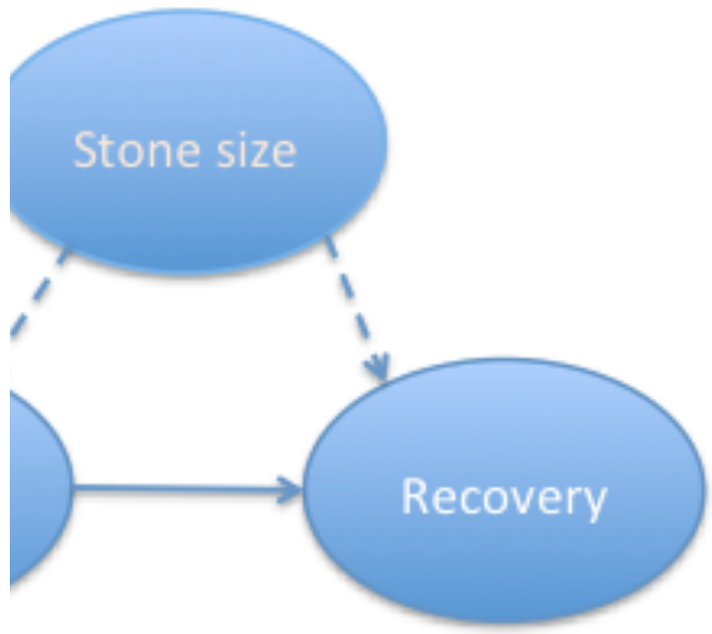
Causal Thinking: Why “Paradox”?

- Dependence vs. causality

- Simpson’s paradox





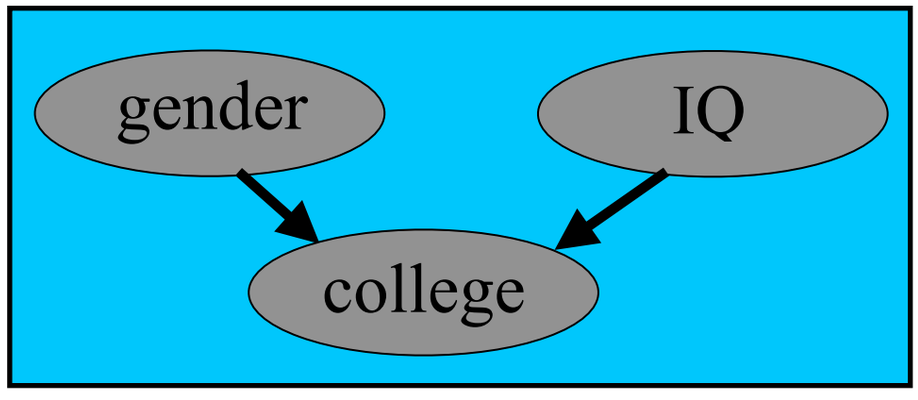


Causal Thinking: Sample vs. Population

- Dependence vs. causality

- Simpson's paradox

- "Strange" dependence
  - Go back 50 years; female college students were smarter than male ones on average. Why?

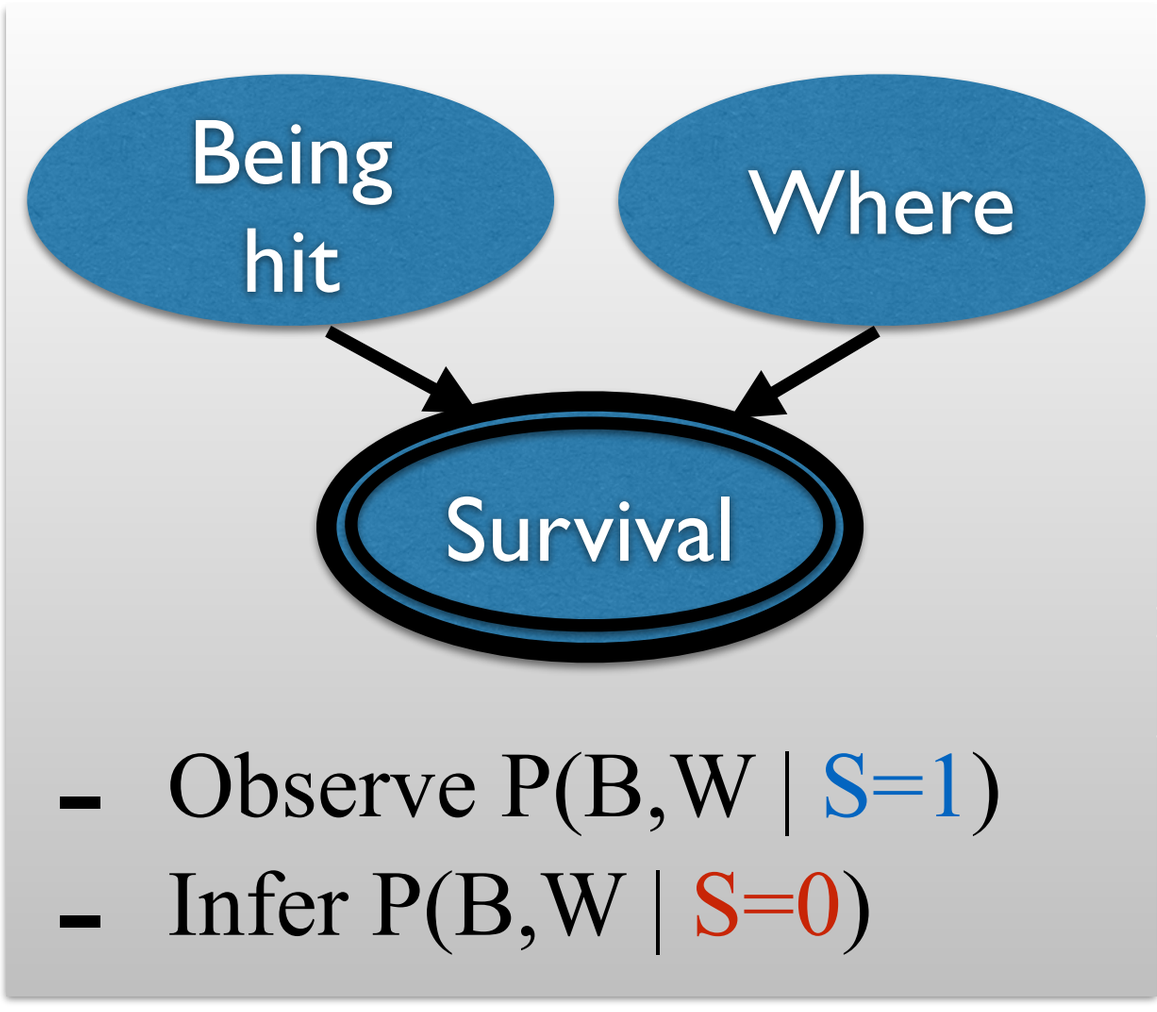


Causal Thinking: Sample vs. Population

- Dependence vs. causality
- Simpson's paradox
- "Strange" dependence

Causal Thinking: Sample vs. Population

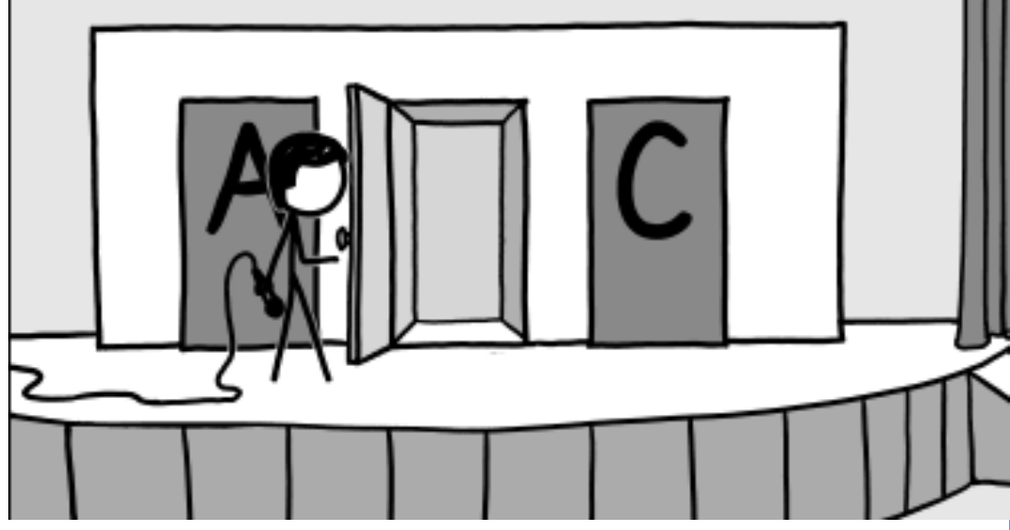
- Dependence vs. causality



- Observe P(B,W | S=1)
- Infer P(B,W | S=0)

Question 2

Monty Hall Problem



- You are a game show contestant. Before the game begins, the host, Monty Hall, has placed $1,000 dollars behind one of three doors. Nothing is behind the other two doors. The game is played as followed. You, the contestant, choose one of the doors, say, door A. Then Monty opens a door that is not the door you chose and does not have the money behind it, say B. If you want to maximize the expected profit, which door will you finally choose?
  - A
  - C

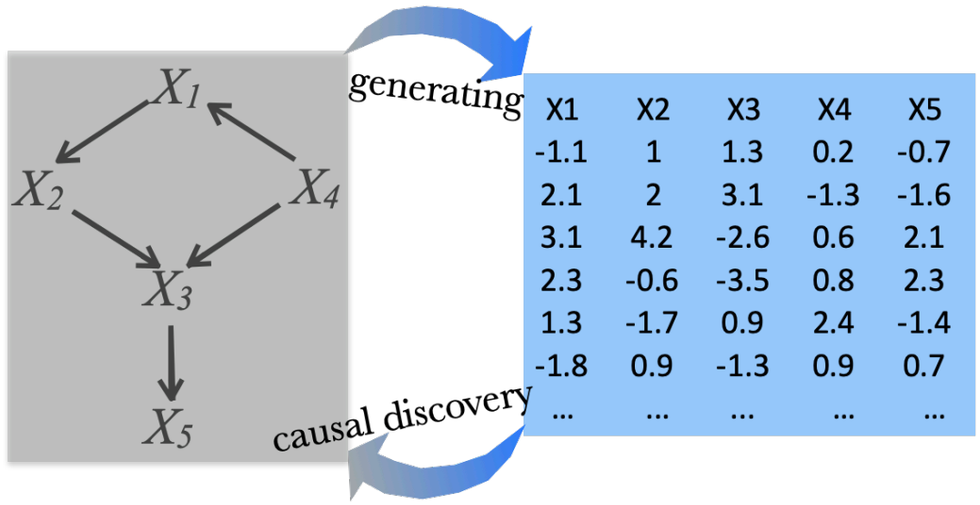
Excerpt from "The Mind's Arrows"

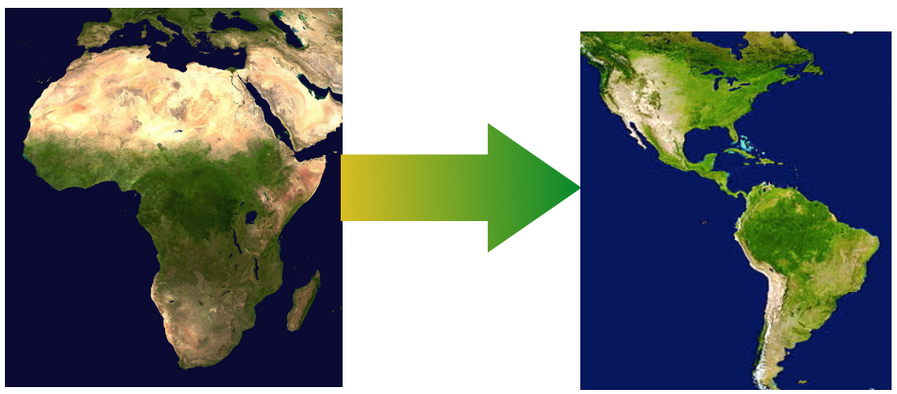
Causal Thinking Makes a Difference

- Active manipulation /control vs. passive prediction
- Generalization / adaptation ability in new environments?
- Integration of causal information: what is the causal model for *X*, *Y*, and *Z* if
  - $X \rightarrow Y$, $Y \rightarrow Z$ (expansion) or $X \rightarrow Z$, $Y \rightarrow Z$ (refinement)...
- Creativity
  - Thoughts consist of the "What if?" and the "If I had only..." + knowledge integration + ...

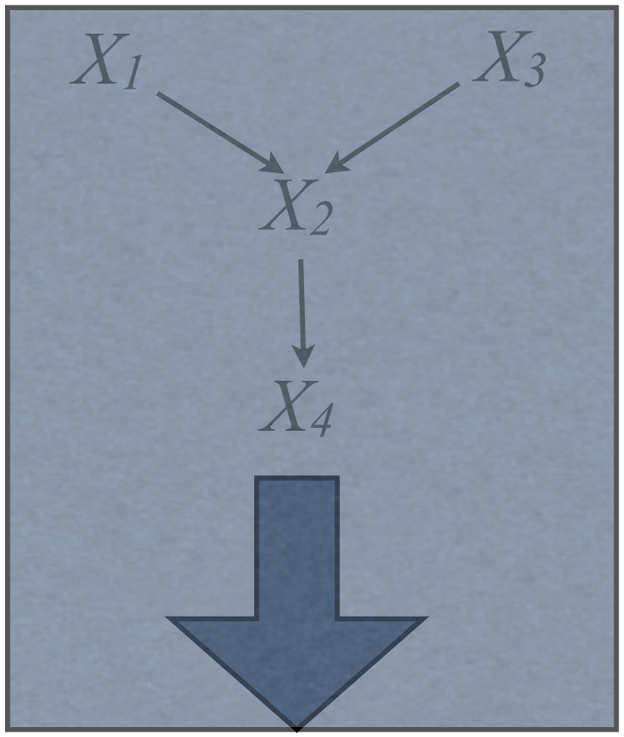
Outline

- Causal thinking
- Causal representation learning: IID case
- Causal representation learning from time series
- Causal representation learning from heterogeneous/ nonstationary data
  - Transfer/adaptive learning & generative AI





(Simple) Causal Discovery as an Estimation Problem



Mysteries...

Data

X_1	X_2	X_3	X_4
-1.1	1.0	1.3	0.2
2.1	2.0	3.1	-1.3
3.1	4.2	2.6	0.6
2.3	-0.6	-3.5	0.8
1.3	2.2	0.9	2.4
-1.8	0.9	-1.3	0.9
...	...	...	...

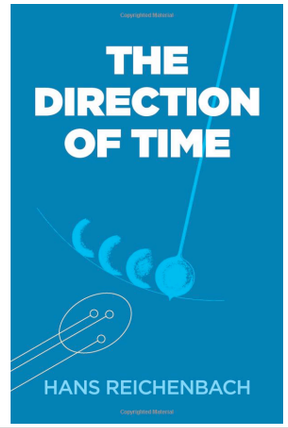
Linear identifiable cases, find: $\mathbf{X} = \mathrm{B} \cdot \mathbf{X} + \mathbf{E}$

Nonlinear identifiable cases, find $X_i = f_i(\mathrm{PA}_i, E_i)$

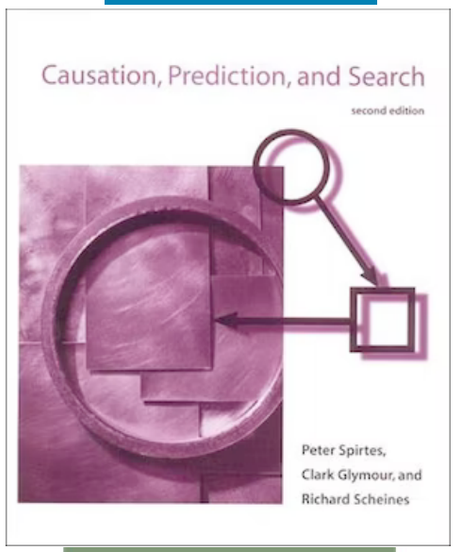
	X_1	X_2	X_3	X_4
X_1	0	0	0	0
X_2	1	0	1	0
X_3	0	0	0	0
X_4	0	1	0	0

What if there are latent confounders?

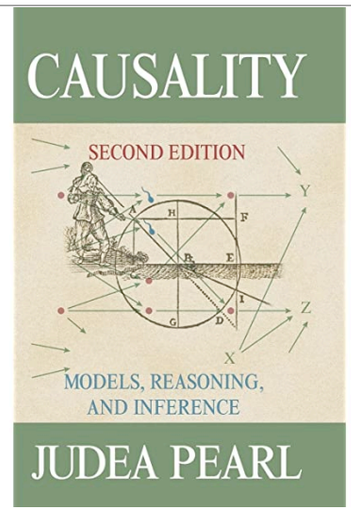
Causal Discovery: A Bit of History



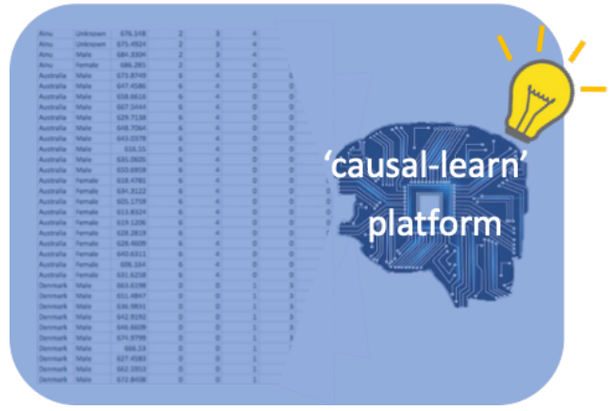
- Reichenbach's common cause principle ("The Direction of Time", 1956)
- Markov condition (Kiiveri et al., 1984)
- "Causation, Prediction, and Search" (Spirtes, Glymour, & Scheines, 1993)
  - Faithfulness condition, PC algorithm, SGS, FCI, Tetrad program…



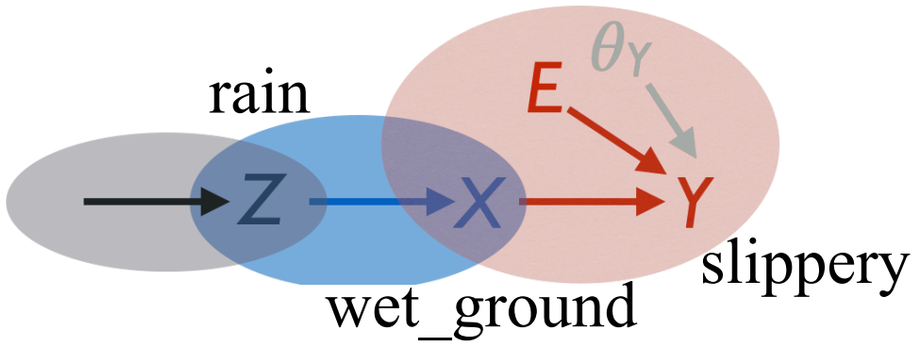
- "Causality: Models, Reasoning and Inference" (Pearl, 2000)
- Greedy equivalence search (GES) (Chickering, 2003)
- Functional causal model-based methods (LiNGAM, PNL…, since 2005)
- Latent variable recovery: Factor analysis (Spearman, 1904), Tetrad condition (Spearman & CMU), Latent tree structure (Pearl et al., 1989), measurement model (CMU 2006), GIN (CMU & GDUT), rank deficiency (UCSD & CMU)…



Uncover Causality from Observational Data?



- Causal system has "irrelevant" modules (Pearl, 2000; Spirtes et al., 1993)



- conditional independence among variables;
- independent noise condition;
- minimal (and independent) changes...

Footprint of causality in data

- Causal discovery (Spirtes et al., 1993)/ causal representation learning (Schölkopf et al., 2021): find such representations with identifiability guarantees

- Three dimensions of the problem:

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

Causal Representation Learning: A Summary

i.i.d. data?	Parametric constraints?	Latent confounders?	What can we get?
Yes	No	No	(Different types of) equivalence class
		Yes	
	Yes	No	Unique identifiability (under structural conditions)
		Yes	
Non-I, but I.D.	No/Yes	No	(Extended) regression
		Yes	Latent temporal causal processes identifiable!
I., but non-I.D.	No	No	More informative than MEC (CD-NOD)
	Yes		May have unique identifiability
	No	Yes	Changing subspace identifiable
	Yes		Variables in changing relations identifiable

Causal Discovery in Archeology: An Example

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

Thanks to Marlijn Noback

- 8 variables of 250 skeletons collected from different locations

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Id	Population	Sex	Cranial size	Diet or subsistence					Paramastic	Dental wear		Geographic location per population			Climate per population					
2			(Male, fem	(Centroid S	Gathering	Hunting	Fishing	Pastoralism	Agriculture	Yes=1, no=	Average att	Attrition pe	Distance to	Longitude	Latitude	Tmean	Tmin	Tmax	Vpmean	Vpmin	Vpmax
3	AINU31_1	Ainu	Unknown	713.2942	2	3	4	0	1	0	1.5	2	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
4	AINU7_1	Ainu	Unknown	676.148	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
5	AINU7_2	Ainu	Unknown	675.4924	2	3	4	0	1	0	1.5	1	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
6	AINU_1016	Ainu	Male	684.3304	2	3	4	0	1	0	1.5	2.5	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
7	AINU_1016	Ainu	Female	686.285	2	3	4	0	1	0	1.5	4	16464	43.548548	142.639159	2.86	-11.19	17.01	7.43	2.27	16.83
8	AUSM245	Australia	Male	673.8749	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
9	AUSM246	Australia	Male	647.4586	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
10	AUSM8217	Australia	Male	658.6616	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
11	AUSM8177	Australia	Male	667.5444	6	4	0	0	0	1	2.5	4	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
12	AUSM8173	Australia	Male	629.7138	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
13	AUSM8173	Australia	Male	648.7064	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
14	AUSM8171	Australia	Male	643.0378	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
15	AUSM8165	Australia	Male	616.55	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
16	AUSM8154	Australia	Male	635.0605	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
17	AUSM8153	Australia	Male	650.6959	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
18	AUSF1412	Australia	Female	618.4781	6	4	0	0	0	1	2.5	1	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
19	AUSF8179	Australia	Female	634.3122	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
20	AUSF8175	Australia	Female	605.1759	6	4	0	0	0	1	2.5	1.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
21	AUSF8172	Australia	Female	613.8324	6	4	0	0	0	1	2.5	3	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
22	AUSF8169	Australia	Female	619.1206	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
23	AUSF8157	Australia	Female	628.2819	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
24	AUSF8155	Australia	Female	628.4609	6	4	0	0	0	1	2.5	3.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
25	AUSF1578	Australia	Female	640.6311	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
26	AUSF243	Australia	Female	606.164	6	4	0	0	0	1	2.5	2.5	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
27	AUSF8158	Australia	Female	631.6258	6	4	0	0	0	1	2.5	2	20164	-24.287027	135.615234	22.46	13.33	30.27	11.10	7.55	15.96
28	DENM1432	Denmark	Male	663.6198	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
29	DENM1011	Denmark	Male	651.4847	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
30	DENM1205	Denmark	Male	636.9831	0	0	1	3	6	0	2.1	1.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
31	DENM116_	Denmark	Male	642.9192	0	0	1	3	6	0	2.1	3	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
32	DENM116_	Denmark	Male	646.6609	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
33	DENM116_	Denmark	Male	674.9799	0	0	1	3	6	0	2.1	2	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27
34	DENM7_77	Denmark	Male	666.53	0	0	1	3	6	0	2.1	2.5	10440	55.717055	11.711426	8.01	-0.02	16.66	9.67	5.59	15.27

Causal Structure vs. Statistical Independence (SGS, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

$Y \rightarrow X \rightarrow Z$

Y -- X -- Z **?**

Statistical independence(s)

$Y \perp\!\!\!\perp Z \mid X$

Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

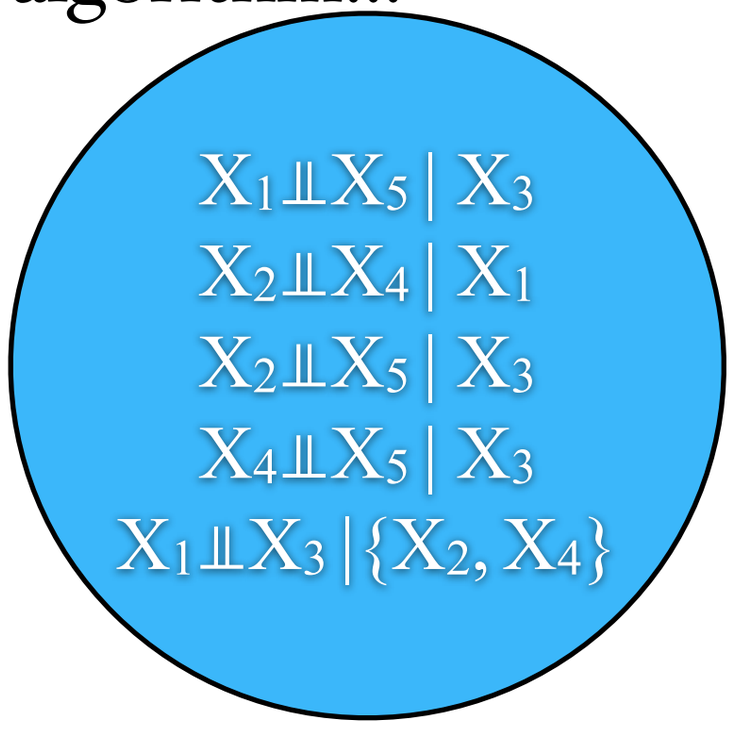
Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

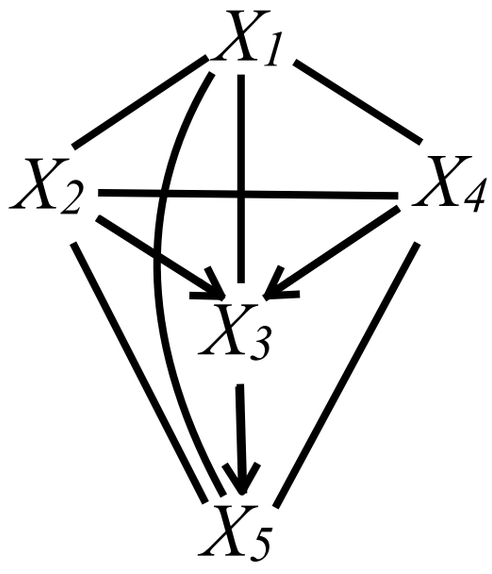
(Typical) Constraint-Based Causal Discovery

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- Conditional independence constraints between each variable pair
  - Illustration: the PC algorithm
  - Extensions: the FCI algorithm...

X1	X2	X3	X4	X5
-1.1	1	1.3	0.2	-0.7
2.1	2	3.1	-1.3	-1.6
3.1	4.2	-2.6	0.6	2.1
2.3	-0.6	-3.5	0.8	2.3
1.3	-1.7	0.9	2.4	-1.4
-1.8	0.9	-1.3	0.9	0.7
…	…	…	…	…



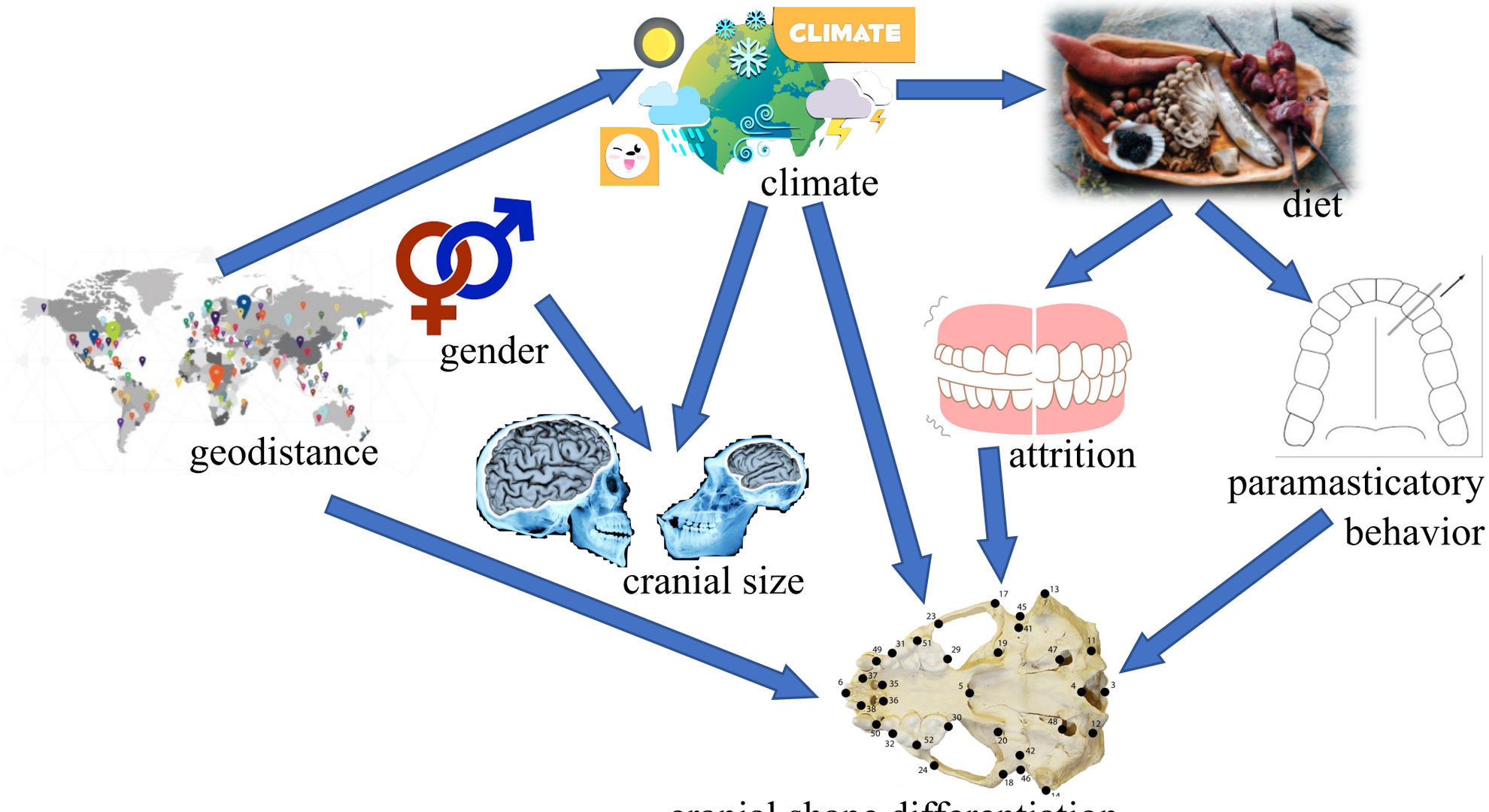


- Spirtes, Glymour, and Scheines. Causation, Prediction, and Search. 1993.

Result of PC on the Archeology Data

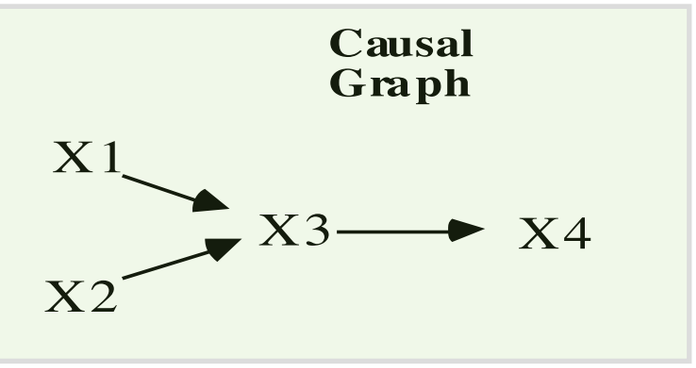
Thanks to collaborator Marlijn Noback

- By PC algorithm (Spirtes et al., 1993) + kernel-based conditional independence test (Zhang et al., 2011)



PC Algorithm: Example

Step I: finding skeleton



Independcies

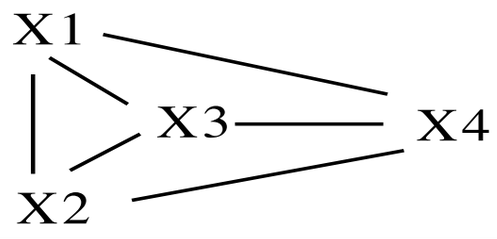
$X1 \perp\!\!\!\perp X2$

$X1 \perp\!\!\!\perp X4 \mid \{X3\}$

$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Step II: finding v-structure and doing orientation propagation

Begin with:



Dealing with Confounders?

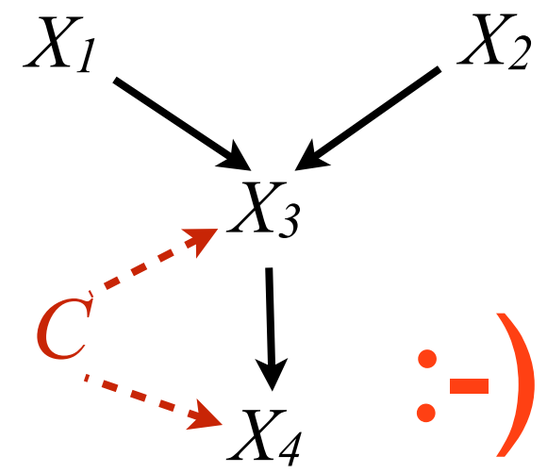
Example I

$X_1 \perp\!\!\!\perp X_2;$

$X_1 \perp\!\!\!\perp X_4 \,|\, X_3;$

$X_2 \perp\!\!\!\perp X_4 \,|\, X_3.$

Possible to have confounders behind X_3 and X_4?



E.g., X_1: Raining; X_3: wet ground; X_4: slippery.

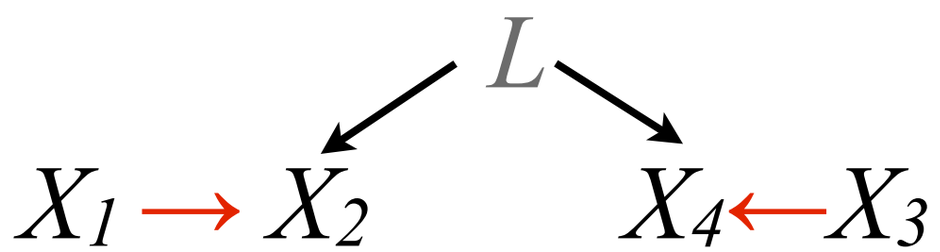
Example II

$X_1 \perp\!\!\!\perp X_3;$

$X_1 \perp\!\!\!\perp X_4;$

$X_2 \perp\!\!\!\perp X_3.$

Are there confounders behind X_2 and X_4?



E.g., X_1: I am not sick; X_2: I am in this lecture room; X_4: you are in this lecture room; X_3: you are not sick.

(See the FCI algorithm)

I know There Is No Confounder: Example

- In the 1970s, the Edison Electric Company in North Carolina was concerned about the effects on plant growth of acid rain produced by emissions from its electric generators.

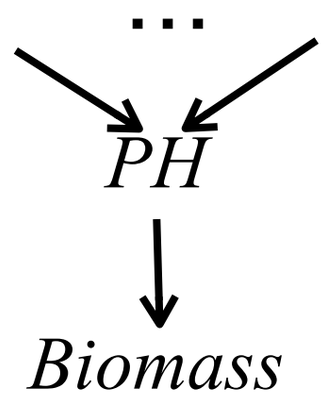
- The investigators chose samples from the Cape Fear estuary, where the Cape Fear River flows into the Atlantic Ocean.

- obtained 45 samples of Spartina grass up and down the estuary, and measured 13 variables in the samples, including **concentrations of various minerals, acidity (pH), salinity, and the outcome variable, the biomass of each sample**

- The PC algorithm found that among **the measured variables the only *direct* cause of biomass was pH**.

- Y-structure: no confounder!

- Later verified by intervention-based analysis



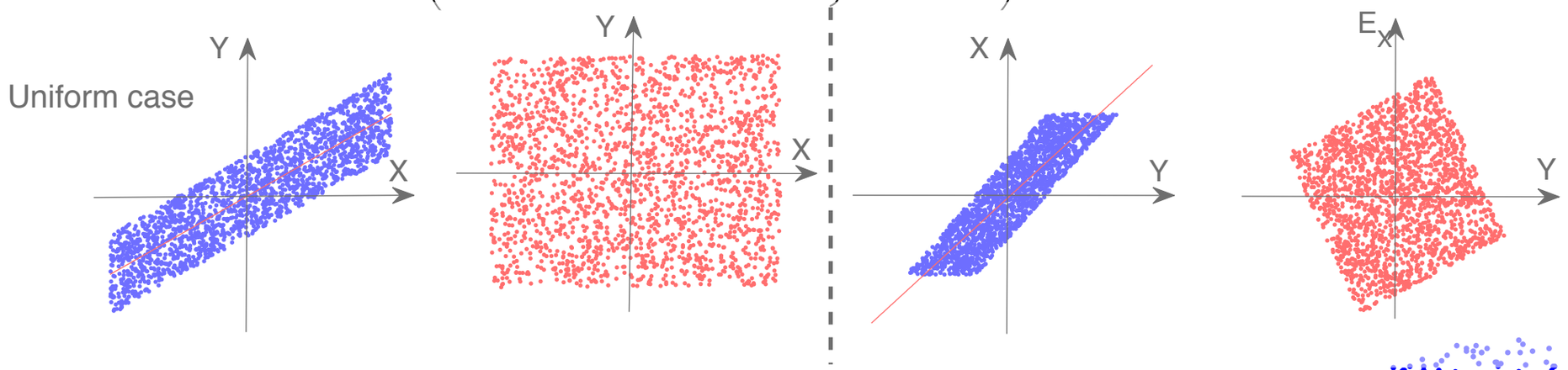
Functional Causal Model-Based Causal Discovery

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

"Independent changes" renders causal direction identifiable

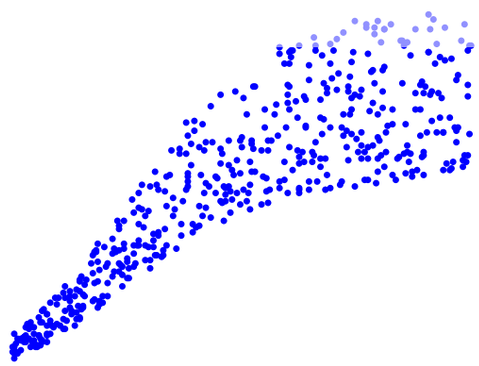
- Linear non-Gaussian model (Shimizu et al., 2006):

$$Y = aX + E$$



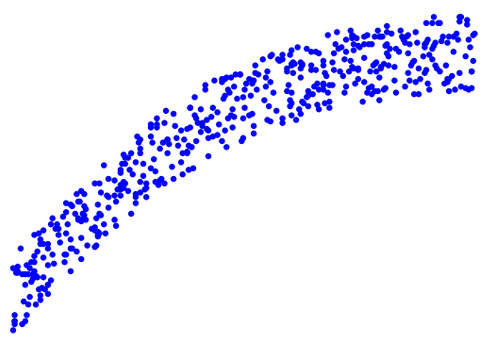
- Post-nonlinear causal model (Zhang & Chan, 2006):

$$Y = f_2 (f_1(X) + E)$$



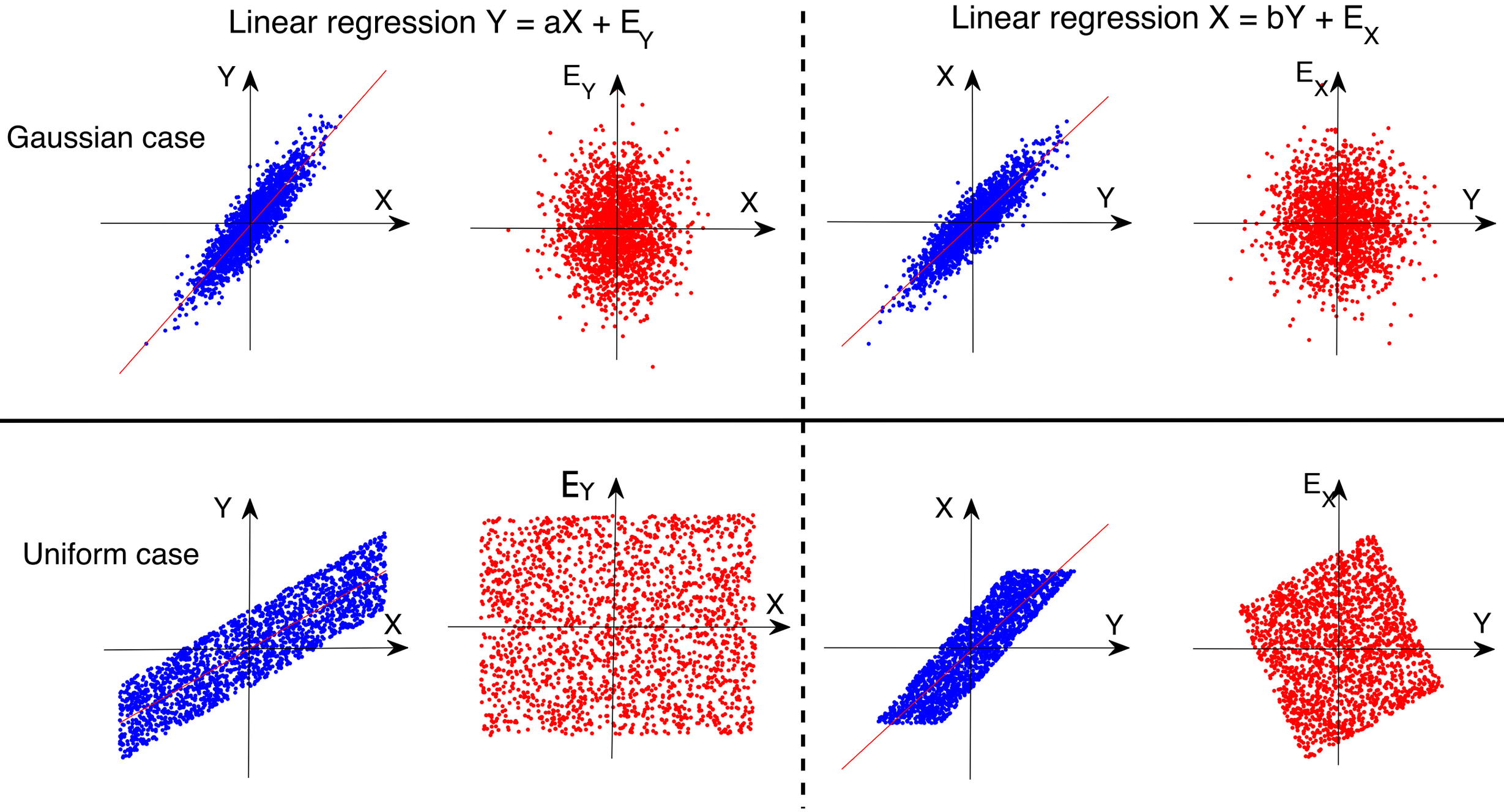
- Additive noise model (Hoyer et al, 2009)

$$Y = f(X) + E$$



Causal Asymmetry the Linear Case: Illustration

Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):



LiNGAM Based on Independent Component Analysis (ICA)

- LiNGAM: $X_i = \sum_{j:\text{ parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{BX} + \mathbf{E} \Rightarrow \mathbf{E} = \mathbf{(I-B)X}$
  - **B** has special structure: acyclic relations
- ICA: $\mathbf{Y} = \mathbf{WX}$
- **B** can be seen from **W** [illegible] and re-scaling
- Faithfulness assumption avoided

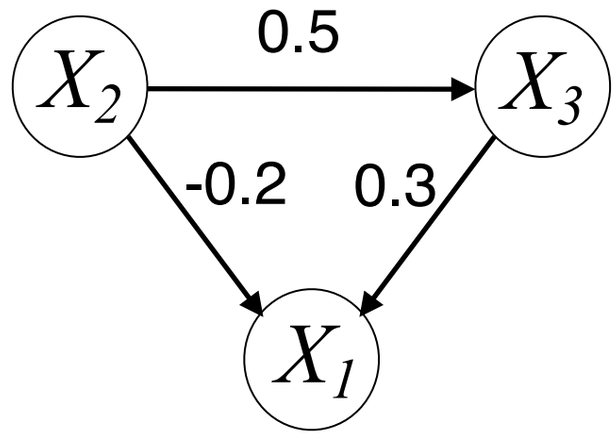
Question 1. How to find W?

Question 2. How to see B from W?

- E.g., $\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$

$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5X_2 + E_3 \\ X_1 = -0.2X_2 + 0.3X_3 + E_2 \end{cases}$$

So we have the causal relation:



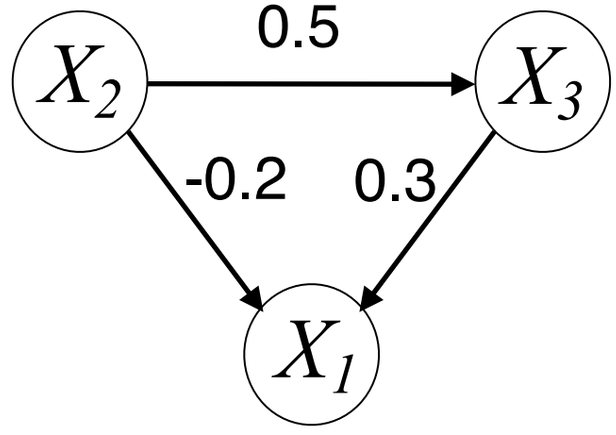
LiNGAM Analysis by ICA

- LiNGAM: $X_i = \sum_{j:\text{ parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{BX} + \mathbf{E} \Rightarrow \mathbf{E} = \mathbf{(I-B)X}$
  - **B** has special structure: acyclic relations
- ICA: $\mathbf{Y} = \mathbf{WX}$
- **B** can be seen from **W** by permutation and re-scaling
- Faithfulness assumption avoided
- E.g., $\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$

$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5X_2 + E_3 \\ X_1 = -0.2X_2 + 0.3X_3 + E_2 \end{cases}$$

1. First permute the rows of **W** to make all diagonal entries non-zero, yielding $\ddot{\mathbf{W}}$.
2. Then divide each row of $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}'$.
3. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

So we have the causal relation:

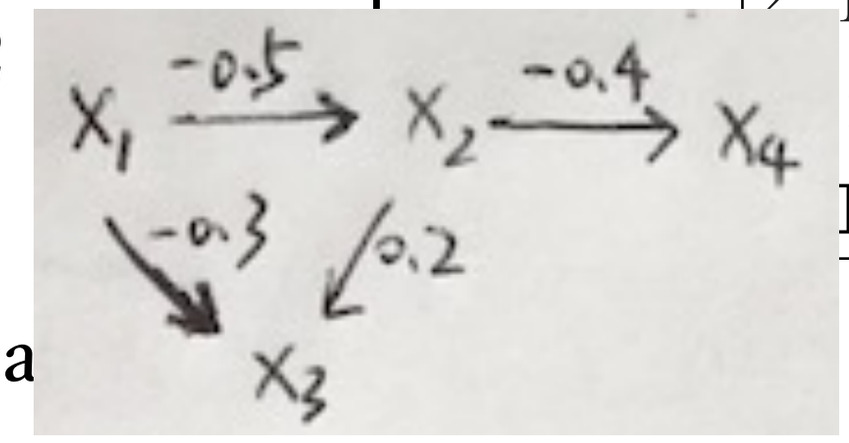


Can You See Causal Relations from **W**? Example

- ICA gives $\mathbf{Y} = \mathbf{W}\mathbf{X}$ and

$$\mathbf{W} = \begin{bmatrix} 0.6 & -0.4 & 2 & 0 \\ 1.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.5 \\ 1.5 & 3 & 0 & 0 \end{bmatrix}$$

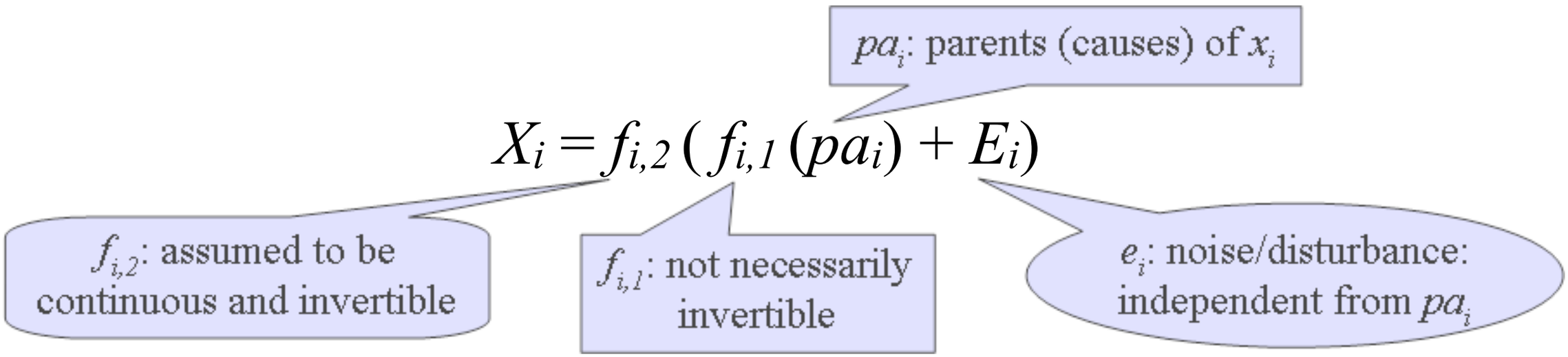
1. First permute the rows of **W** to make all diagonal entries non-zero, yielding $\ddot{\mathbf{W}}$.
2. Then divide each row of $\ddot{\mathbf{W}}$ by its diagonal entry, giving $\ddot{\mathbf{W}}'$.
$\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.



- Can we find the ca

$$\dot{W} = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 1.5 & 3 & 0 & 0 \\ 0.6 & -0.4 & 2 & 0 \\ 0 & 0.2 & 0 & 0.5 \end{pmatrix}, \quad \ddot{W}' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.3 & -0.2 & 1 & 0 \\ 0 & 0.4 & 0 & 1 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -0.5 & 0 & 0 & 0 \\ -0.3 & 0.2 & 0 & 0 \\ 0 & -0.4 & 0 & 0 \end{pmatrix}$$

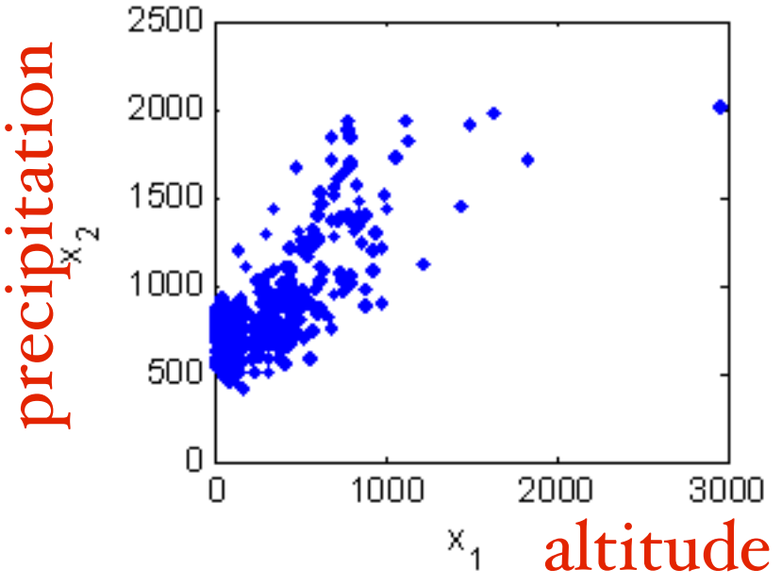
Post-NonLinear (PNL) Causal Model



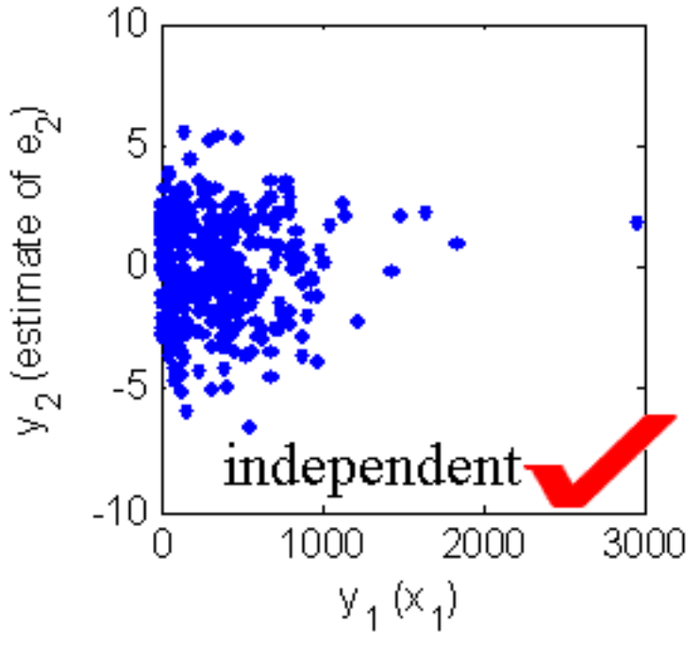
- Special cases:
  - Linear models
  - Nonlinear additive noise models
  - Multiplicative noise models:

$$Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$$

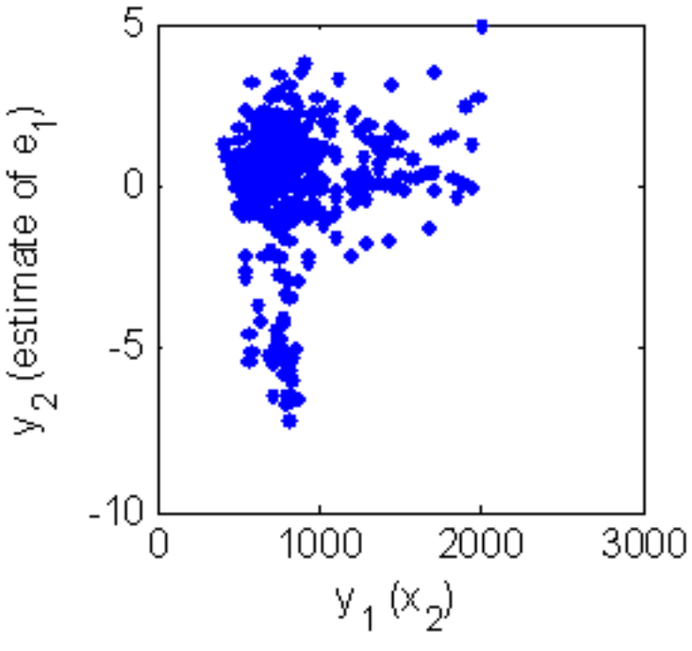
Data Set 2

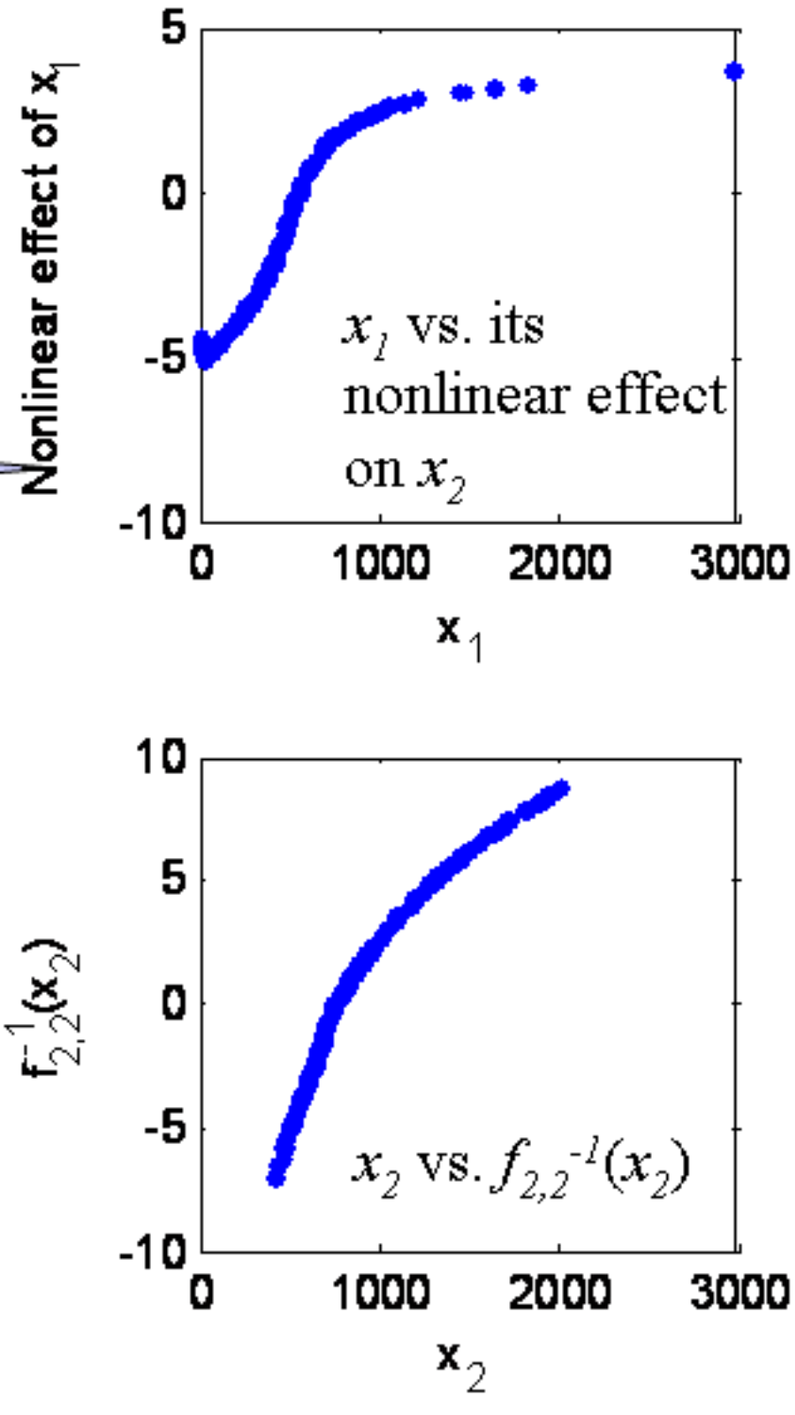


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$



(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$





Identifiability in Two-variable Case: Theoretical Results

pa_i: parents (causes) of x_i

$$X_i = f_{i,2}(f_{i,1}(pa_i) + E_i)$$

$f_{i,2}$: assumed to be continuous and invertible

$f_{i,1}$: not necessarily invertible

e_i: noise/disturbance: independent from pa_i

- Two-variable case: if $X_1 \rightarrow X_2$, then $X_2 = f_{2,2}(f_{2,1}(X_1) + E_2)$
- Is the causal direction implied by the model unique?
- By a proof of contradiction
  - Assume both $X_1 \rightarrow X_2$ and $X_2 \rightarrow X_1$ satisfy PNL model
  - One can then find all non-identifiable cases

Identifiability: A Mathematical Result

- **Theorem 1**
  - Assume $x_2 = f_2(f_1(x_1) + e_2)$,
    $x_1 = g_2(g_1(x_2) + e_1)$,
  - Further suppose that involved densities and nonlinear functions are third-order differentiable, and that p_{e2} is unbounded,
  - For every point satisfying $\eta_2'' h' \neq 0$, we have

$$\eta_1''' - \frac{\eta_1'' h''}{h'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \eta_1'' + \eta_2' \cdot \left(h''' - \frac{h''^2}{h'}\right).$$

Notation

$$t_1 \triangleq g_2^{-1}(x_1), \quad z_2 \triangleq f_2^{-1}(x_2),$$
$$h \triangleq f_1 \circ g_2, \quad h_1 \triangleq g_1 \circ f_2.$$
$$\eta_1(t_1) \triangleq \log p_{t_1}(t_1), \quad \eta_2(e_2) \triangleq \log p_{e_2}(e_2).$$

- Obtained by using the fact that the Hessian of the logarithm of the joint density of independent variables is diagonal everywhere (Lin, 1998)
- It is not obvious if this theorem holds in practice…

List of All Non-Identifiable Cases

Log-mixed-linear-and-exponential:

$$\log p_v = c_1 e^{c_2 v} + c_3 v + c_4$$

$(\log p_v)' \to c \ (c \neq 0)$,

as $v \to -\infty$ or as $v \to +\infty$

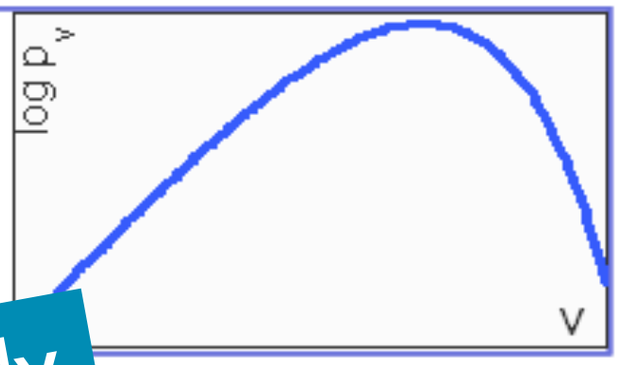


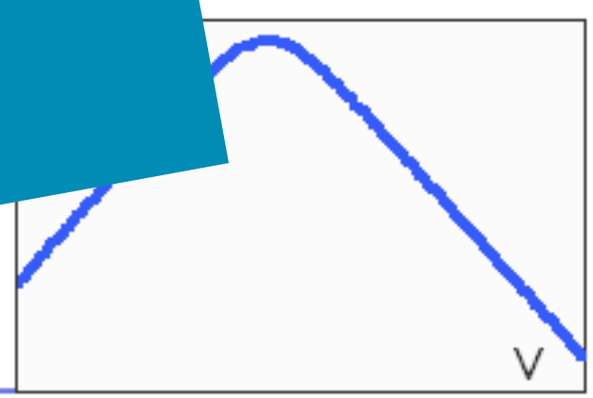
Table 1: All situations i… not identifiable.

	p_{e_2}	Remark
I	Gaussian	h_1 also linear
II	log-mix-lin-exp	h_1 strictly monotonic, and $h_1' \to$ 0, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	—
IV	log-mix-lin-exp	
V	generalized mixture of two exponentials	

$$p_v \propto (c_1 e^{c_2 v} + c_3 e^{c_4 v})^{c_5}$$

$\to c_2 \ (c_2 \neq 0)$,

as $v \to +\infty$



Causal direction is generally identifiable if the data were generated according to

$$X_2 = f_2(f_1(X_1) + E).$$

Linear models and nonlinear additive noise models are special cases.

A Problem in Psychology: Finding Underlying Mental Conditions?

- 50 questions for big 5 personality test

race	age	engnat	gender	hand	source	country	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	A1	A2	A3	A4	A5
3	53	1	1	1	1	US	4	2	5	2	5	1	4	3	5	1	1	5	2	5	1	1	1	1	1	1	1	5	1	5	2
13	46	1	2	1	1	US	2	2	3	3	3	3	1	5	1	5	2	3	4	2	3	4	3	2	2	4	1	3	3	4	4
1	14	2	2	1	1	PK	5	1	1	4	5	1	1	5	5	1	5	1	5	5	5	5	5	5	5	5	5	1	5	5	1
3	19	2	2	1	1	RO	2	5	2	4	3	4	3	4	4	5	5	4	4	2	4	5	5	5	4	5	2	5	4	4	3
11	25	2	2	1	2	US	3	1	3	3	3	1	3	1	3	5	3	3	3	4	3	3	3	3	3	4	5	5	3	5	1
13	31	1	2	1	2	US	1	5	2	4	1	3	2	4	1	5	1	5	4	5	1	4	4	1	5	2	2	2	3	4	3
5	20	1	2	1	5	US	5	1	5	1	5	1	5	4	4	1	2	4	2	4	2	2	3	2	2	2	5	5	1	5	1
4	23	2	1	1	2	IN	4	3	5	3	5	1	4	3	4	3	1	4	4	4	1	1	1	1	1	1	2	5	1	4	3
5	39	1	2	3	4	US	3	1	5	1	5	1	5	2	5	3	2	4	5	3	3	5	5	4	3	3	1	5	1	5	1
3	18	1	2	1	5	US	1	4	2	5	2	4	1	4	1	5	5	2	5	2	3	4	3	2	3	4	2	3	1	4	2
3	17	2	2	1	1	IT	1	5	2	5	1	4	1	4	1	5	5	3	5	3	2	5	3	3	4	3	2	4	2	4	1
13	15	2	1	1	1	IN	3	3	5	3	3	3	2	4	3	3	1	5	3	3	2	3	2	3	2	4	4	4	2	2	5
13	22	1	2	1	2	US	3	3	4	2	4	2	2	3	4	3	3	3	3	2	2	4	4	2	3	1	4	1	5	1	
3	21	1	2	1	5	US	1	3	2	5	1	1	1	5	1	5	5	3	5	2	5	5	3	2	5	3	1	1	1	4	2
3	28	2	2	1	2	US	3	3	3	4	3	2	2	4	3	5	2	4	4	4	4	4	2	2	3	2	1	4	2	4	2
3	21	1	1	1	5	US	2	3	2	3	3	1	1	3	4	4	2	4	2	4	1	2	2	2	2	2	4	2	4	2	5
13	19	1	2	1	2	FR	1	3	2	4	2	4	1	4	3	4	4	2	3	2	1	3	1	2	2	3	4	2	3	1	4
3	21	1	2	1	5	US	4	1	5	2	5	1	5	3	5	1	5	2	5	2	3	3	3	3	4	2	1	5	2	5	2
3	26	1	2	3	5	GB	2	3	4	3	1	4	1	4	1	5	4	2	5	2	1	4	2	2	2	2	2	2	2	2	2
3	26	1	2	1	1	US	2	2	3	3	3	3	1	3	3	3	4	4	3	1	3	2	2	2	4	4	1	3	2	4	3

Learning Hidden Variables & Their Relations

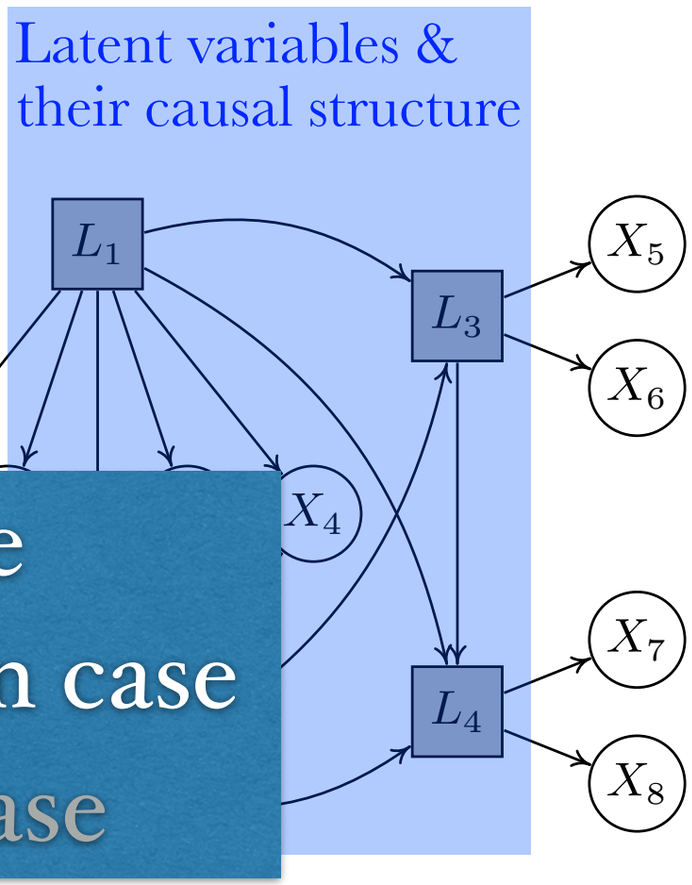
i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- Measured variables (e.g., answer scores in psychometric questionnaires) were generated by causally related latent variables

X1	X2	X3	X4	X5	X6	X7	X8
4.2	3.6	6.5	6.8	9.6	7.6	2.7	4.8
3.8	1.9	6.5	7.3	8.9	6.9	1.1	4.6
4.2	3.4	6.5	6.9	9.5	7.4	2.5	4.6
4.2	2.2	6.2	6.9	9.6	7.2	1.9	4.8
3.9	1.9	6.5	6.8	9.0	6.8	1.7	4.4
4.0	2.0	6.4	7.2	9			
3.8	1.7	6.4	7.3	9			
4.1	2.8	6.5	6.9	9			
...	...	...	...				

Discovery: How?

Latent variables & their causal structure



- Linear-Gaussian case
- Linear, non-Gaussian case
- General nonlinear case

- Find latent variables L_i and their causal relations from measured variables X_i ?

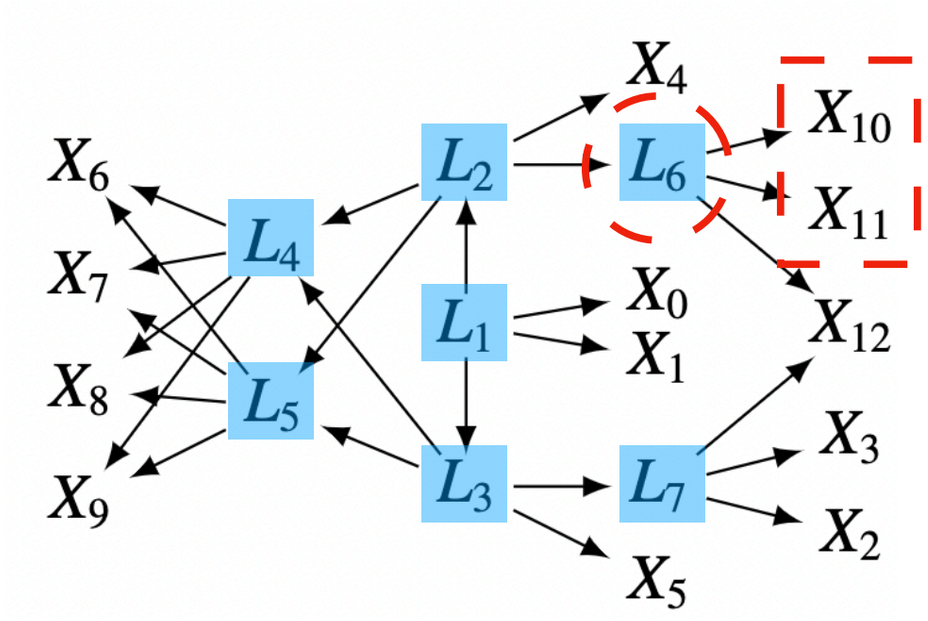
Linear, Gaussian Case: With Rank Deficiency Constrains

Basic idea:

- Rank-deficiency constraints over measured variables
  + Specific search procedure

foundation of this method

  - $\mathbf{rank}(\Sigma_{X_A,X_B})$, which is deficient, indicates the smallest number of variables that d-separate X_A from X_B



Exp:

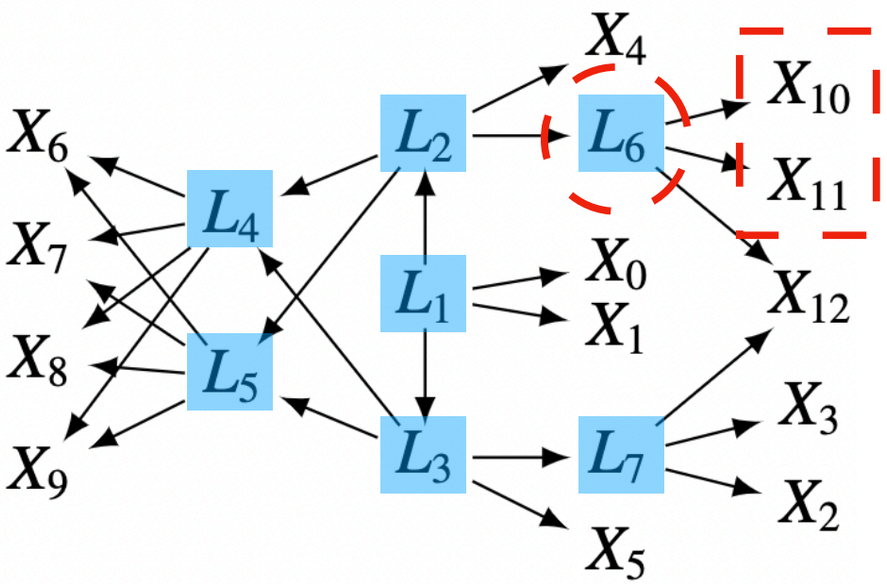
Let $X_A = \{X_{10}, X_{11}\}$ and $X_B = \mathbf{X} \backslash X_A$
$\mathbf{rank}(\Sigma_{X_A,X_B}) = 1$ which is rank deficient, because L_6 d-separates X_A from X_B.

However, we cannot directly know the location of these latent variables in the graph

- *Huang, Low, Xie, Glymour, Zhang, "Latent Hierarchical Causal Structure Discovery with Rank Constraints," NeurIPS 2022*

Linear, Gaussian Case: With Rank Deficiency Constrains

B
- What if L_6 is observable?
  - What if $\text{rank}(\Sigma_{(X1, X2), (X2, X3)}) = 1$?
- Can we unify causal learning with/without latent variables?
  - A unified causal discovery method based on rank deficiency constraints



Exp:

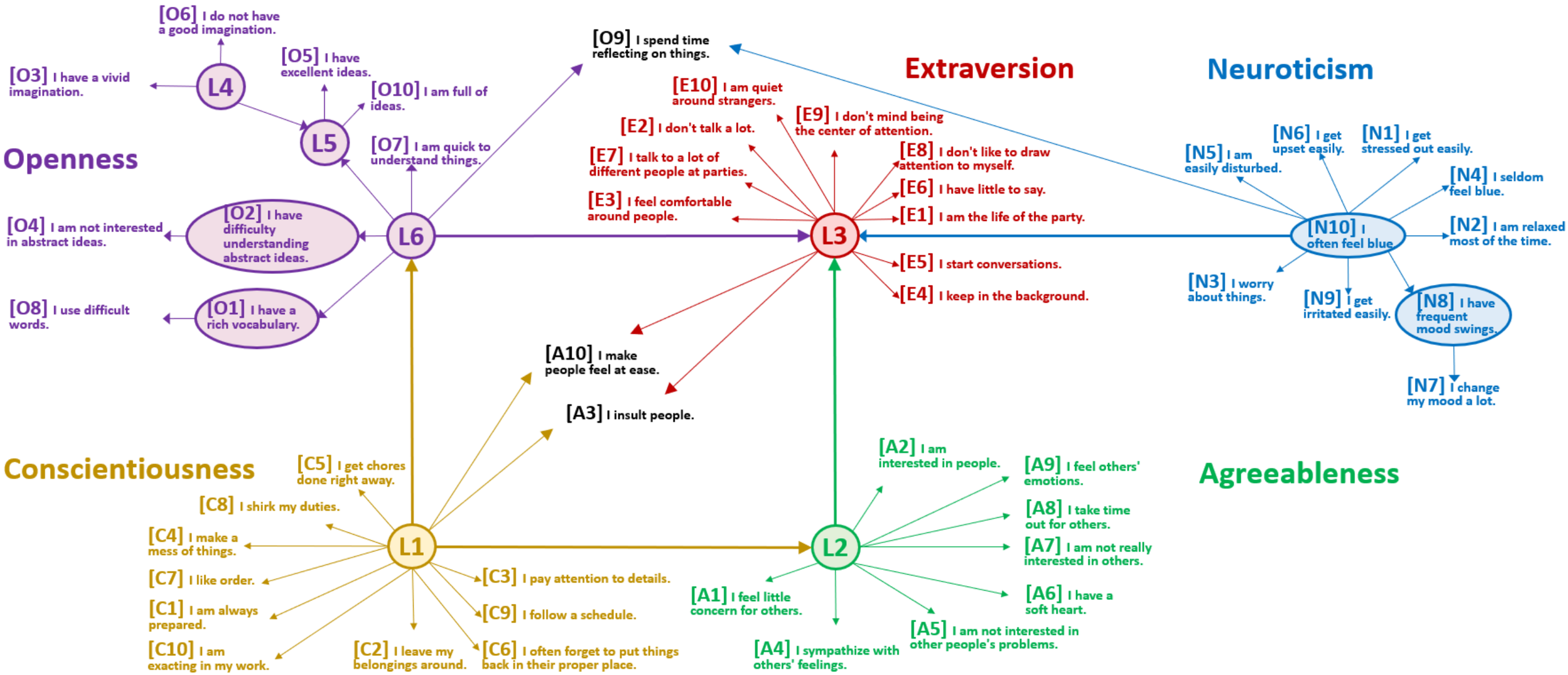
Let $X_A = \{X_{10}, X_{11}\}$ and $X_B = \mathbf{X} \backslash X_A$
$\text{rank}(\Sigma_{X_A, X_B}) = 1$ which is rank deficient, because L_6 d-separates X_A from X_B.

However, we cannot directly know the location of these latent variables in the graph

- Dong, Huang, Ng, Song, Zheng, Jin, Legaspi, Spirtes, Zhang, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

Example: Big 5 Questions Are Well Designed but…

Big 5:
openness; conscientiousness; extraversion; agreeableness; neuroticism

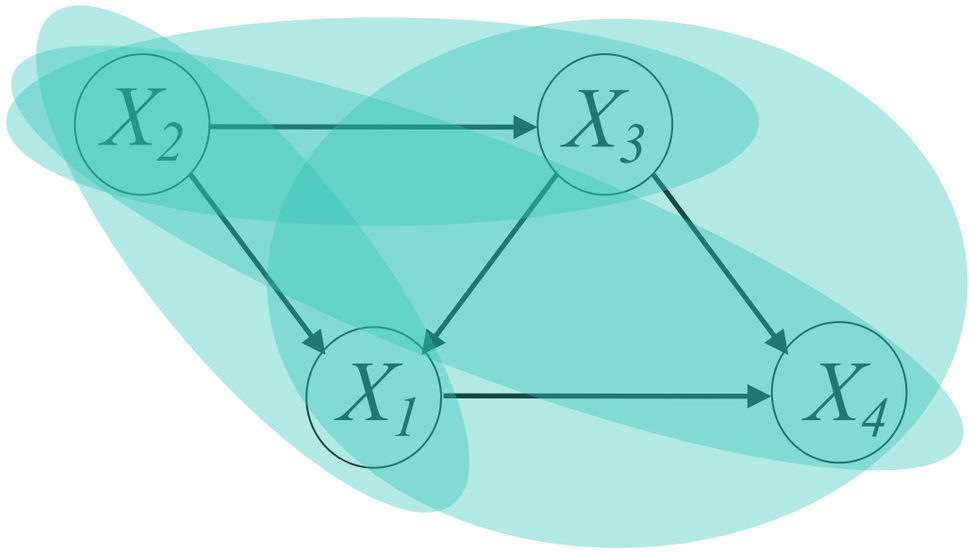


- Dong, Huang, Ng, Song, Zheng, Jin, Legaspi, Spirtes, Zhang, "A Versatile Causal Discovery Framework to Allow Causally-Related Hidden Variables," ICLR 2024

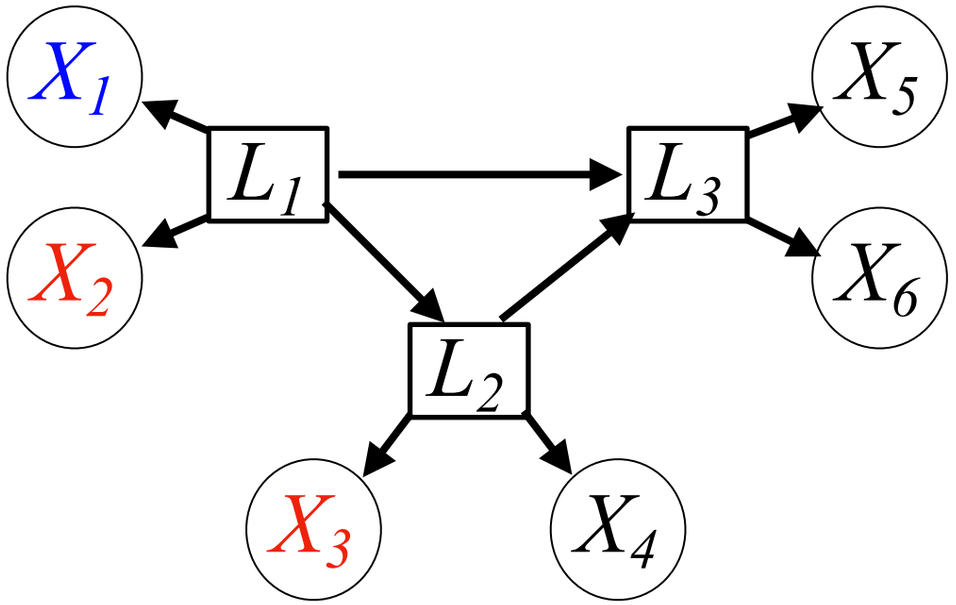
Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

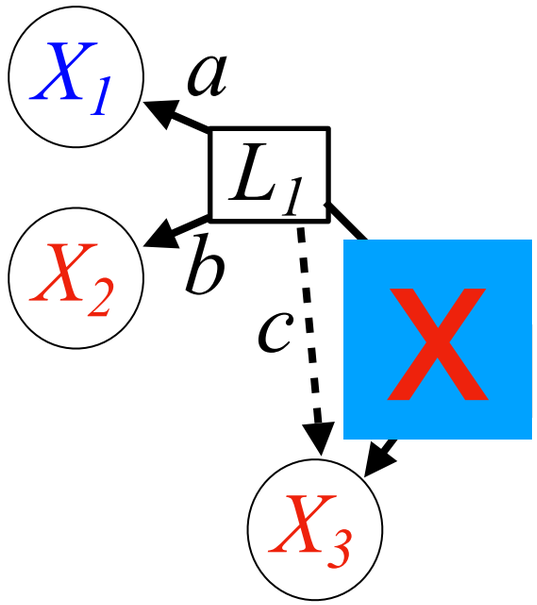
- $(\mathbf{Z}, Y)$ follows the IN condition iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Help determine causal orders and estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM)



Let's See This Asymmetry…



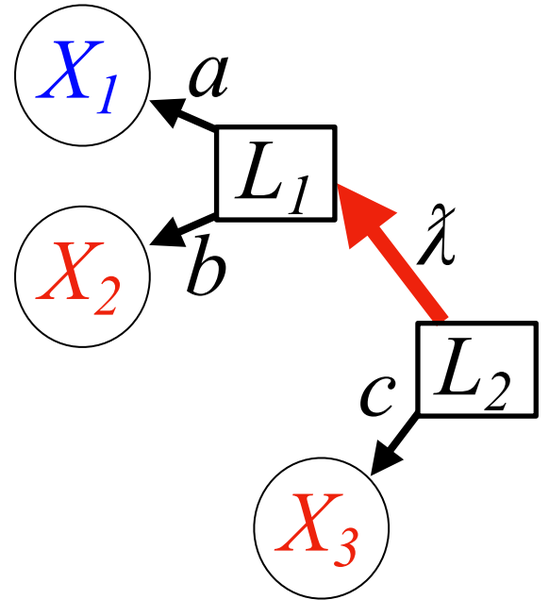
Z = $\{X_1\}$ **Y** = $\{X_2, X_3\}$



$$c \cdot X_2 - b \cdot X_3$$
$$= c(bL1 + E_2) - b(cL_1 + E_3)$$
$$= cE_2 - bE_3,$$

independent from L_1 and from X_1,

and we know $\frac{b}{c} = \frac{Cov(X_1, X_2)}{Cov(X_1, X_3)}$



Nontrivial linear combination of X_2 and X_3 will involve the noise term in L_1, hence dependent on X_1

Linear, Non-Gaussian Case: Generalized Independent Noise Condition

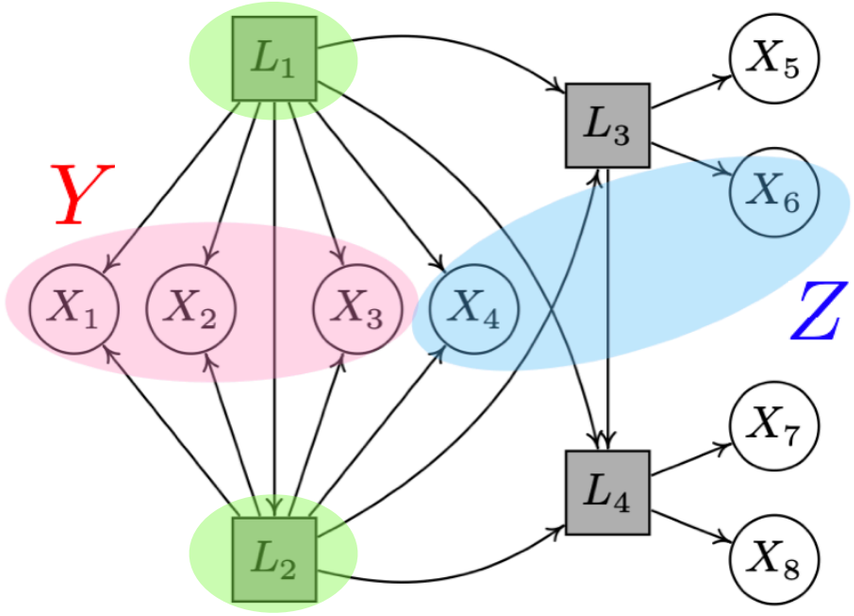
- Generalized Independent Noise (GIN) Condition:

  (Z, Y) follows the GIN condition $\iff \omega^\top Y \perp\!\!\!\perp Z$,

  where $\omega^\top \mathrm{Cov}(Y, Z) = 0$ and $\omega \neq 0$

- Graphical criterion

  (Z, Y) follows the GIN condition iff there is an exogenous set S of PA(Y) that blocks all paths between Y and Z, where 0<=|S|<=min(|Z|, |Y|-1)



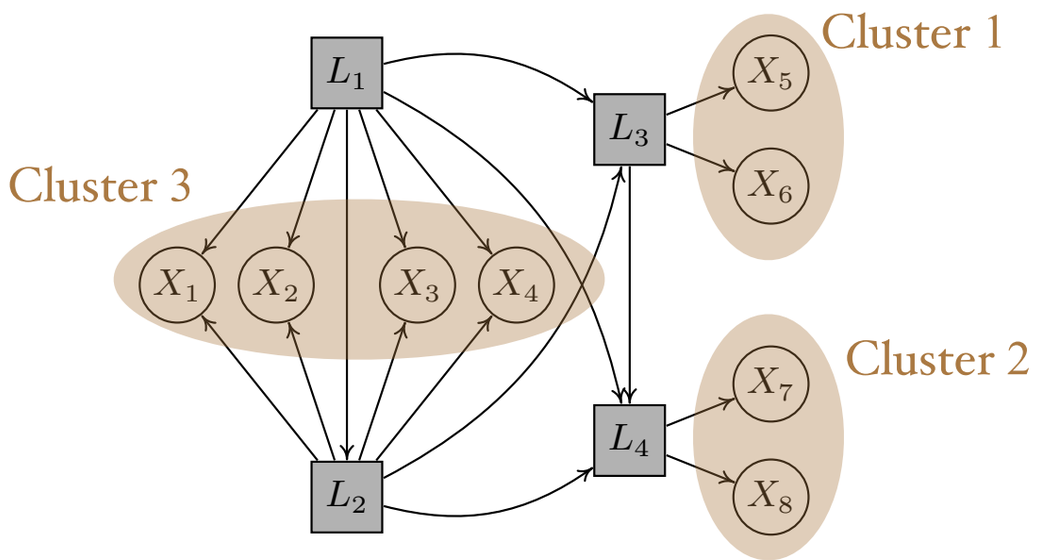
X_i: observed variables
L_i: latent variables

- *Xie, Cai, Huang, Glymour, Hao, Zhang, "Generalized Independent Noise Condition for Estimating Linear Non-Gaussian Latent Variable Causal Graphs," NeurIPS 2020*
- *Cai, Xie, Glymour, Hao, Zhang, "Triad Constraints for Learning Causal Structure of Latent Variables," NeurIPS 2019*

GIN for Estimating Linear, Non-Gaussian LV Model

- A two-step algorithm to identify the latent variable graph
  - By testing for GIN conditions over the input $X_1, \cdots, X_8$

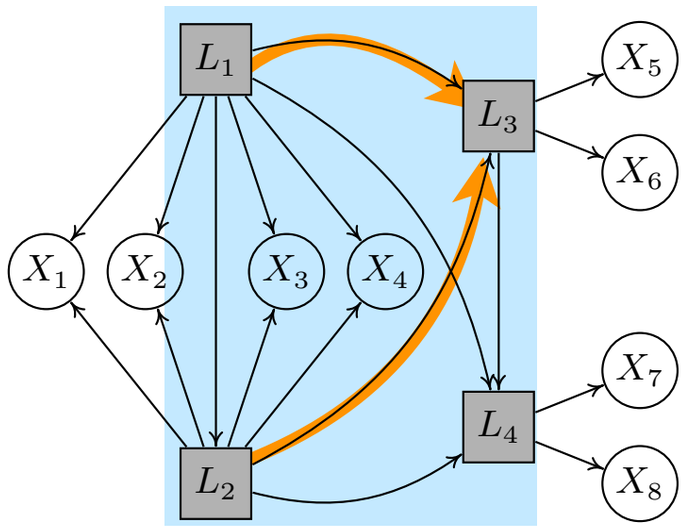
Step 1: find ***causal clusters***



$$(\overbrace{\{X_1, \cdots, X_4, X_7, X_8\}}^{\mathbf{Z}}, \overbrace{\{X_5, X_6\}}^{\mathbf{Y}})$$

satisfies GIN condition

Step 2: determine ***causal structure*** of the latent variables



$$(\overbrace{\{X_3, X_4\}}^{\mathbf{Z}}, \overbrace{\{X_1, X_2, X_5\}}^{\mathbf{Y}})$$

Cluster 3 Cluster 1 & 3

satisfies GIN condition

GIN-Based Method: Application to Teacher's Burnout Data

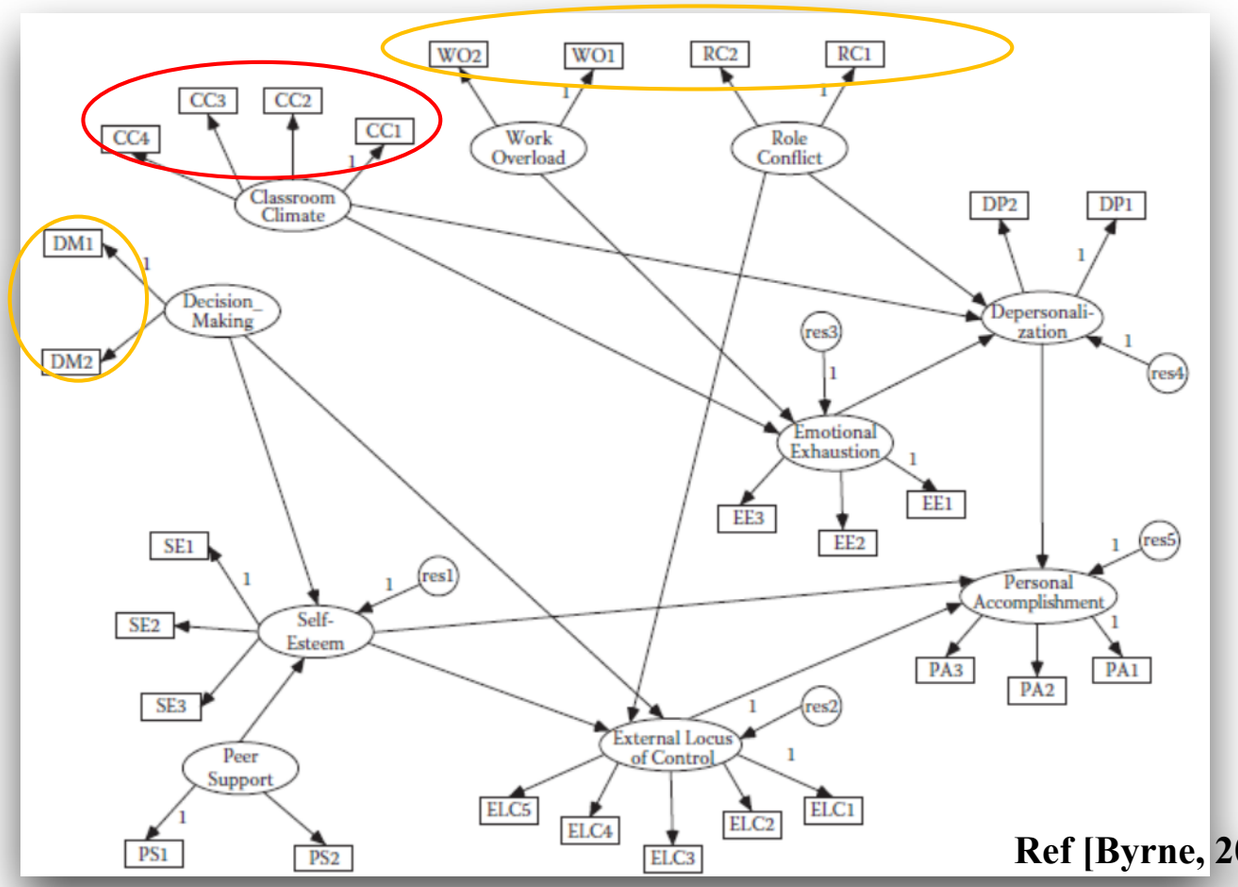
- Contains 28 measured variables
- Discovered clusters and causal order of the latent variables:

Causal Clusters	Observed variables
$\mathcal{S}_1$ (1)	$RC_1, RC_2, WO_1, WO_2, DM_1, DM_2$
$\mathcal{S}_2$ (1)	CC_1, CC_2, CC_3, CC_4
$\mathcal{S}_3$ (1)	PS_1, PS_2
$\mathcal{S}_4$ (1)	$ELC_1, ELC_2, ELC_3, ELC_4, ELC_5$
$\mathcal{S}_5$ (2)	$SE_1, SE_2, SE_3, EE_1, EE_2, EE_3, DP_1, PA_3$
$\mathcal{S}_6$ (3)	DP_2, PA_1, PA_2

$$L(\mathcal{S}_1) \succ L(\mathcal{S}_2) \succ L(\mathcal{S}_3) \succ L(\mathcal{S}_5) \succ L(\mathcal{S}_4) \succ L(\mathcal{S}_6).$$

(from root to leaf)

Hypothesized model by experts



Ref [Byrne, 2010]

- Consistent with the hypothesized model

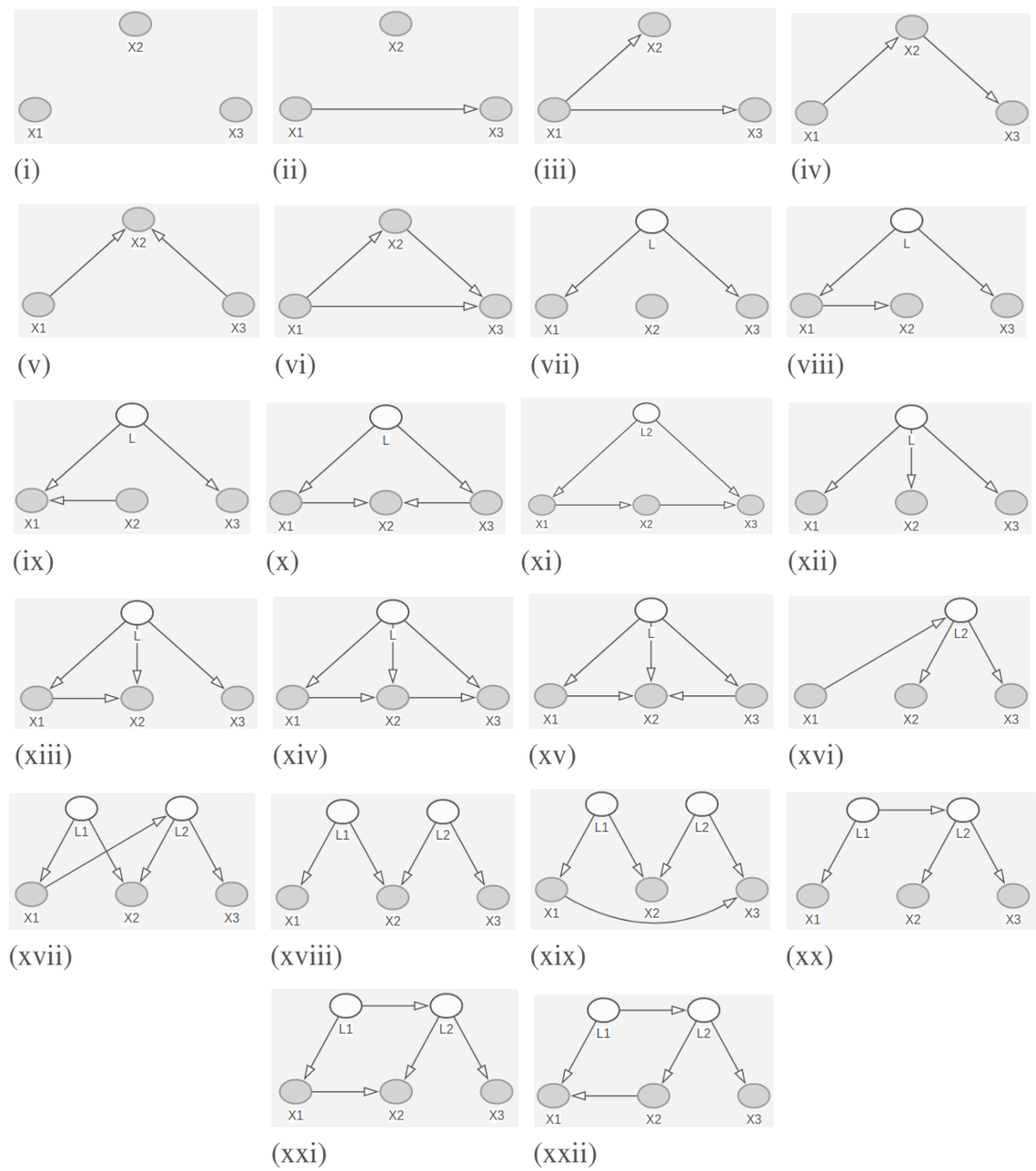
- *Xie, Cai, Huang, Glymour, Hao, Zhang, "Generalized Independent Noise Condition for Estimating Linear Non-Gaussian Latent Variable Causal Graphs," NeurIPS 2020*
- *Cai, Xie, Glymour, Hao, Zhang, "Triad Constraints for Learning Causal Structure of Latent Variables," NeurIPS 2019*

Necessary & Sufficient Conditions on the Structure: Linear, non-Gaussian case

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- Allow a large number of latent variables
- Minimality has to be assumed
- Estimation is generally difficult

Identifiable graphs with only 3 measured variables

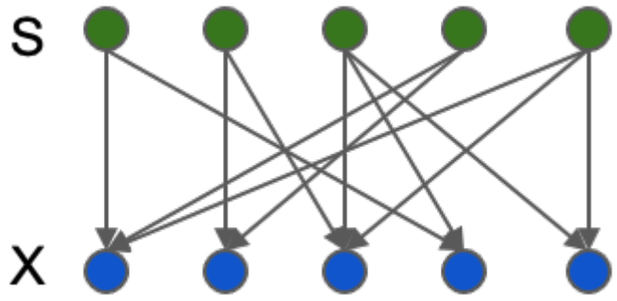


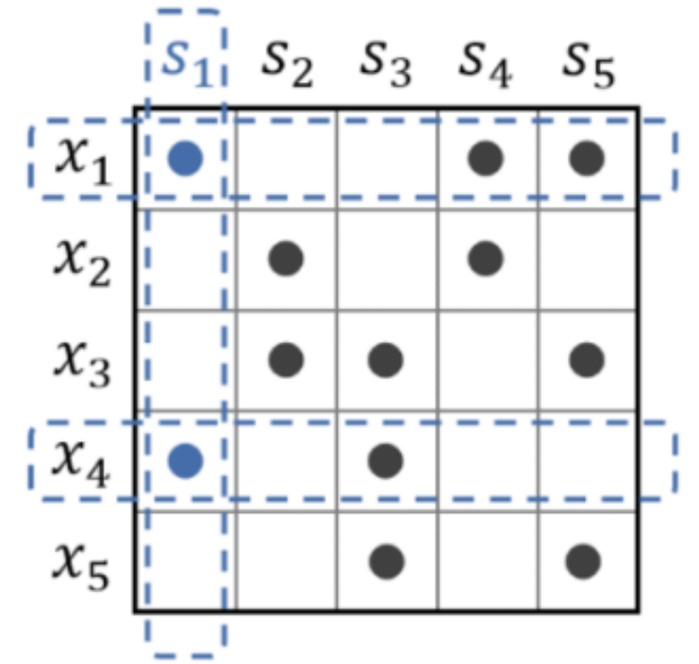
- Adams, Hansen, Zhang, "Identification of Partially Observed Linear Causal Models: Graphical Conditions for the Non-Gaussian and Heterogeneous Cases," NeurIPS 2021

Identifiability of Nonlinear ICA: Structural Sparsity

(Structural Sparsity) For all $k \in \{1, \ldots, n\}$, there exists $\mathcal{C}_k$ such that

$$\bigcap_{i \in \mathcal{C}_k} \operatorname{supp}(\mathbf{J}_{\mathbf{f}}(\mathbf{s})_{i,:}) = \{k\}.$$



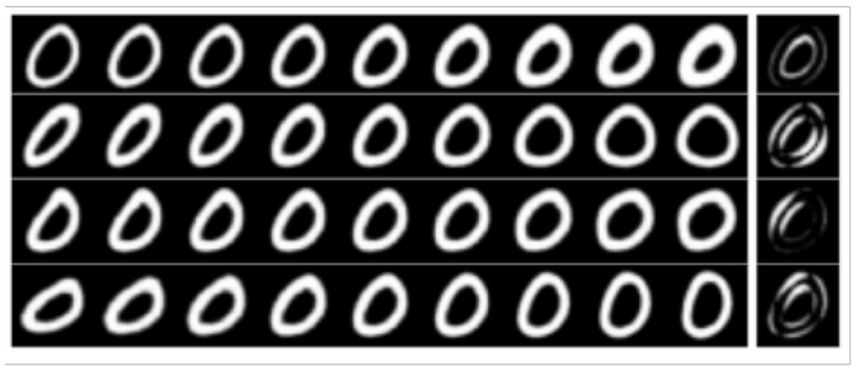


- **Graphically**, for every latent variable S_i, there exists a set of observed variable(s) such that the intersection of their/its parent(s) is S_i
- **Example:** for S_1, there exists X_1 and X_4 such that the intersection of their parents is S_1

- *Zheng, Ng, Zhang, On the Identifiability of Nonlinear ICA: Sparsity and Beyond, NeurIPS 2022*

Further Generalization of Nonlinear ICA

- Undercompleteness
  - More observed variables than latent variables
- Partial sparsity
  - Sparsity is violated for some variables
- Partial source dependence
  - Source independence violated for some variables
- Flexible grouping structures
  - Dependence within each group, independence across groups



Applied to real-world datasets (EMNIST)

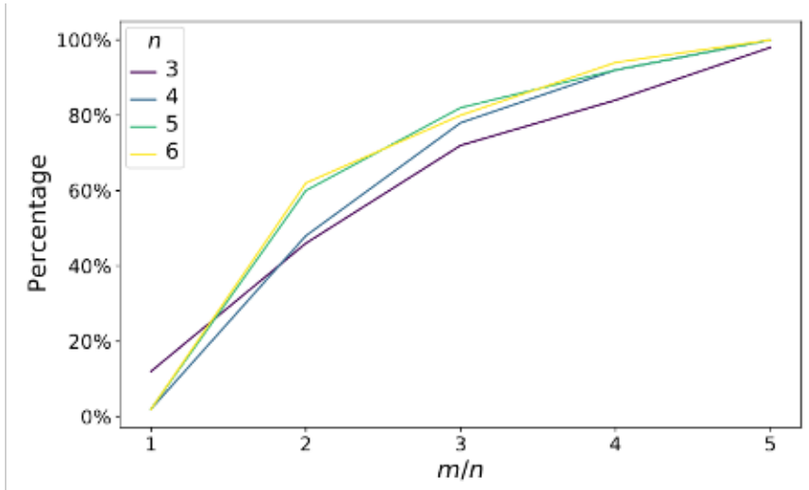


Figure 4: Percentage of random structures satisfying Structural Sparsity w.r.t. different degree of undercompleteness (i.e., m/n).

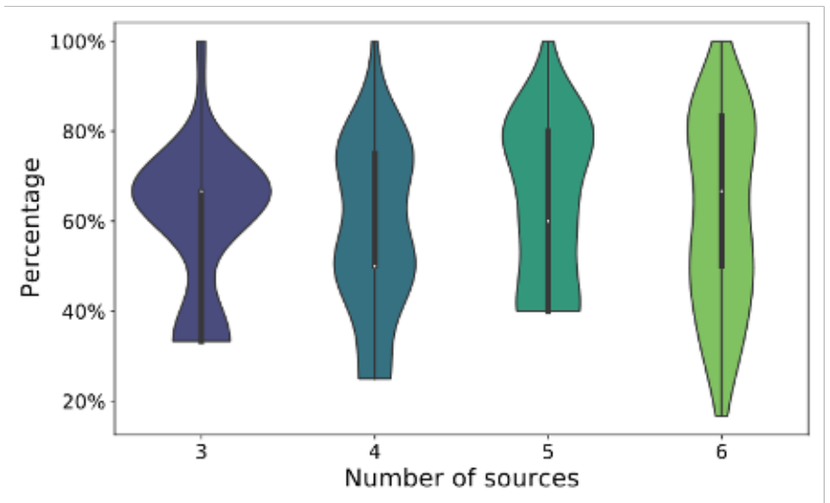


Figure 5: Percentage of sources satisfying *Structural Sparsity* w.r.t. different numbers of sources in the bijective setting ($m/n = 1$).

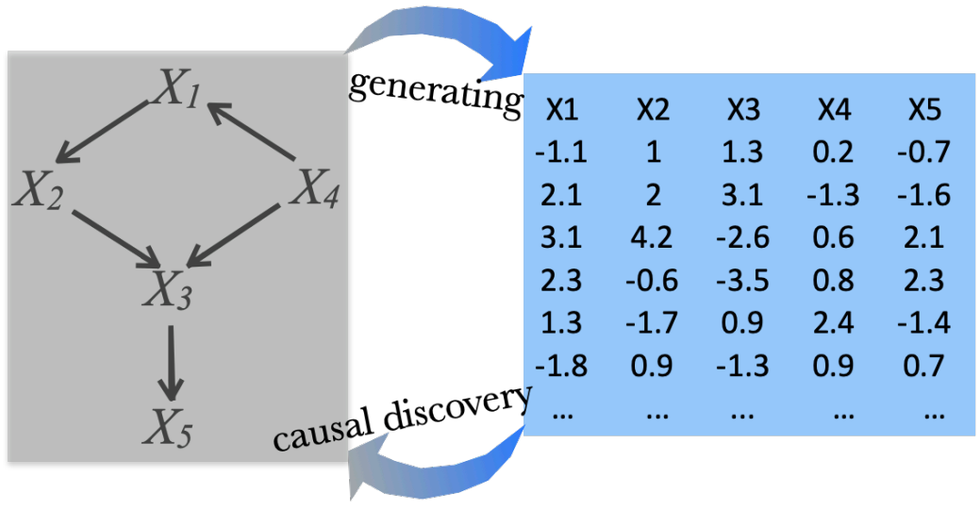
- *Zheng and Zhang, Generalizing Nonlinear ICA beyond Structural Sparsity, NeurIPS 2023 (oral)*

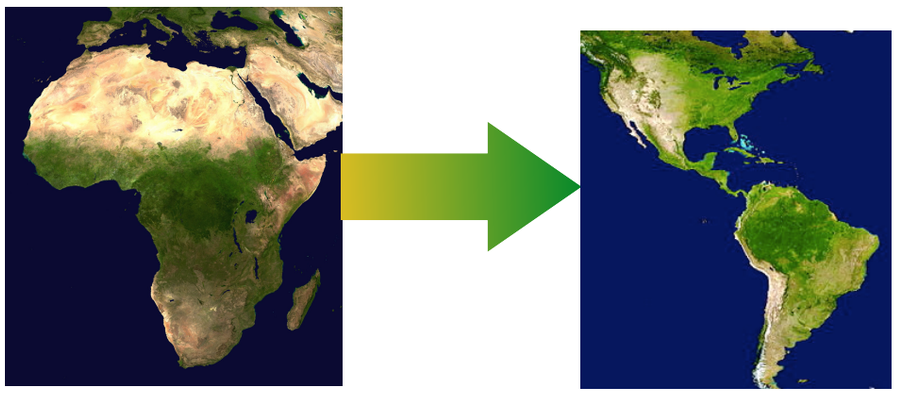
Where Are We?

<table>
<tr><th>i.i.d. data?</th><th>Parametric constraint?</th><th>Latent confounders?</th><th>What can we get?</th></tr>
<tr><td rowspan="4">Yes</td><td rowspan="2">No</td><td>No</td><td rowspan="2">(Different types of) equivalence class</td></tr>
<tr><td>Yes</td></tr>
<tr><td rowspan="2">Yes</td><td>No</td><td rowspan="2">Unique identifiability (under structural conditions)</td></tr>
<tr><td>Yes</td></tr>
<tr><td>Non-I, but I.D.</td><td>No/Yes</td><td>No</td><td rowspan="6">?</td></tr>
<tr><td></td><td></td><td>Yes</td></tr>
<tr><td rowspan="4">I., but non-I.D.</td><td>No</td><td rowspan="2">No</td></tr>
<tr><td>Yes</td></tr>
<tr><td>No</td><td rowspan="2">Yes</td></tr>
<tr><td>Yes</td></tr>
</table>

Outline

- Causal thinking
- Causal representation learning: IID case
- **Causal representation learning from time series**
- Causal representation learning from heterogeneous/ nonstationary data
  - Transfer/adaptive learning & generative AI





Estimating *Time-Delayed* Causal Model

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

- Granger causality: Conditional independence-based approach + temporal constraints
- Further with instantaneous causal relations
  - Conditional independence for instantaneous relations (Swanson & Granger, 1997)
  - With linear, non-Gaussian model (Hyvärinen et al, 2010)

- *Swanson, Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregression. J. of the Americal Statistical Association, 1997*
- *Hyvärinen, Zhang, Shimizu, Hoyer, "Estimation of a structural vector autoregression model using non-Gaussianity," JMLR, 2010*

Learning Latent Causal Dynamics

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

Learn the underlying causal dynamics from their mixtures?
"Time-delayed" influence renders latent processes & their relations identifiable

Time-series Inputs $\{x_t\}_{t=0}^{T}$

Unsupervised Representation Learning

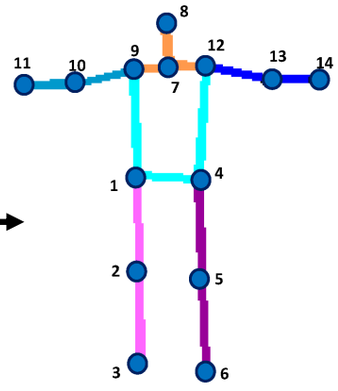
$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)$$

Latent processes

Temporal VAE with causal prior

Latent temporal causal processes z_{it} follow
- completely **nonparametric** model; or furthermore,
- **non-stationary** noise or causal influence, or
- **Parametric** constraints

Causal Skeleton Recovery



Recovered latent processes

- *Yao, Chen, Zhang, "Causal Disentanglement for Time Series," NeurIPS 2022*
- *Yao, Sun, Ho, Sun, Zhang, "Learning Temporally causal latent processes from general temporal data," ICLR 2022*

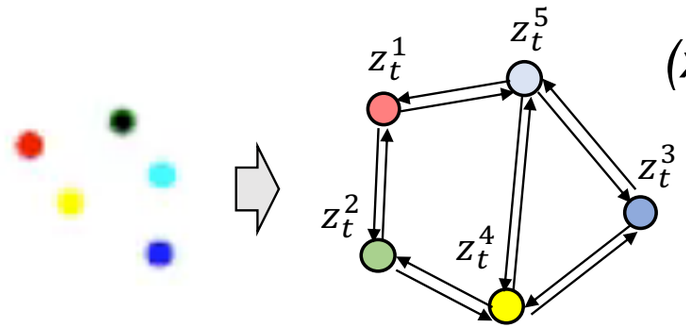
Results on Simple Video Data

- For easy interpretation, consider a simple video data set
  - Mass-spring system: a video dataset with ball movement and invisible springs

Mass-spring Video

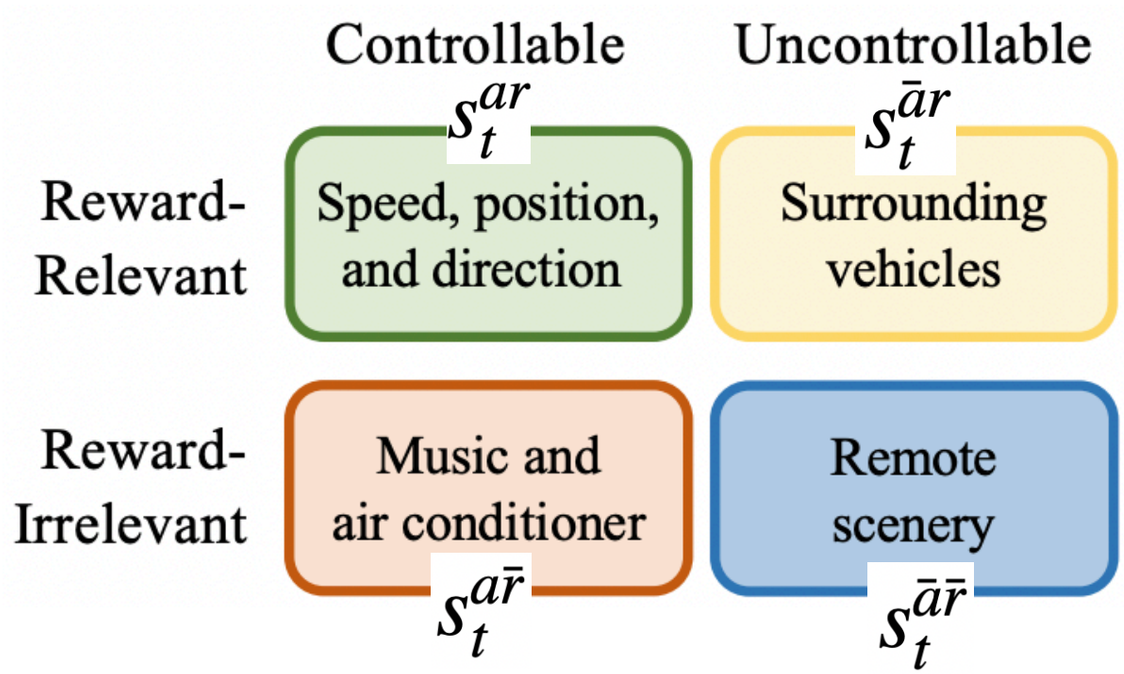
Learned latent processes

Interpretation

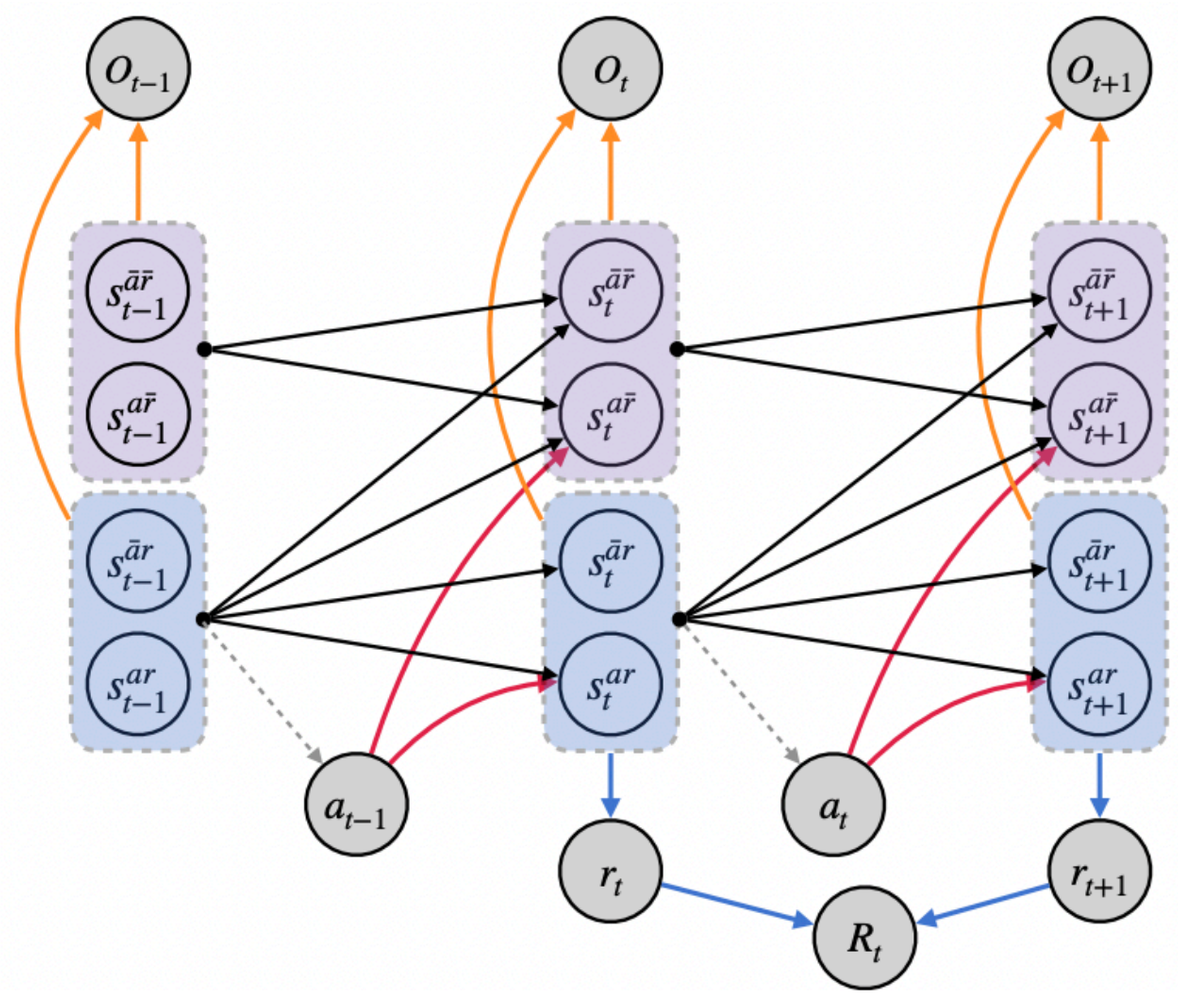


(x- & y- coordinates of the 5 balls)

Extension: Four Categories of State Representations in RL

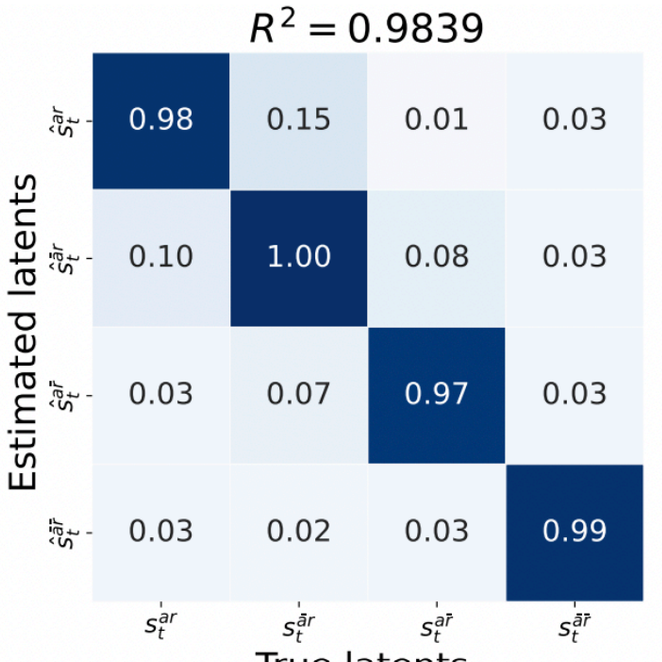


Each category is identifiable!

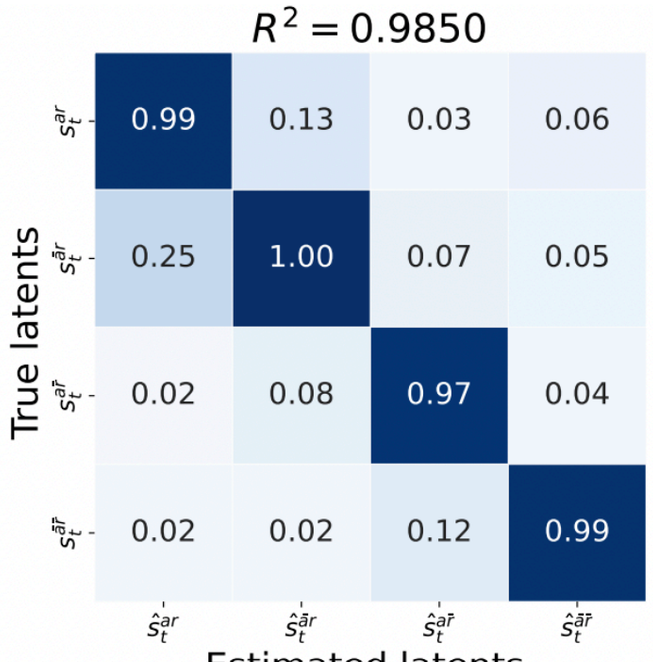


- Liu*, Huang*, Zhu, Tian, Gong, Yu, Zhang. Learning world models with identifiable factorization. NeurIPS 2023

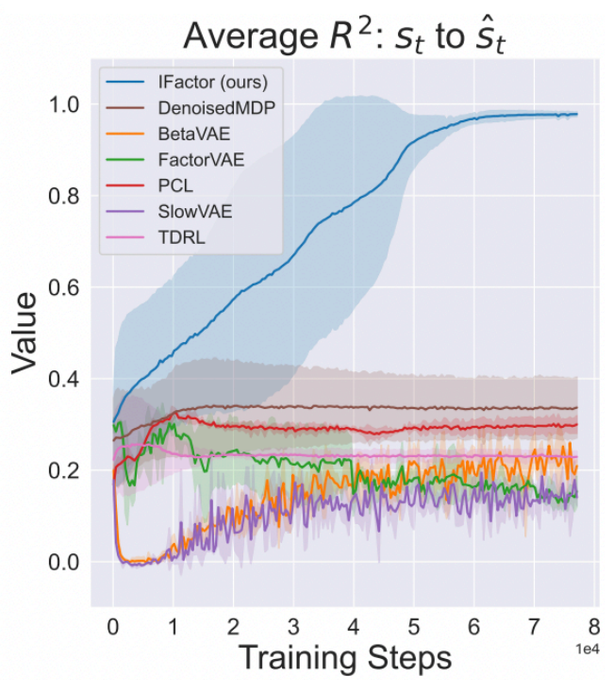
Experimental Results on Latent States Recovery



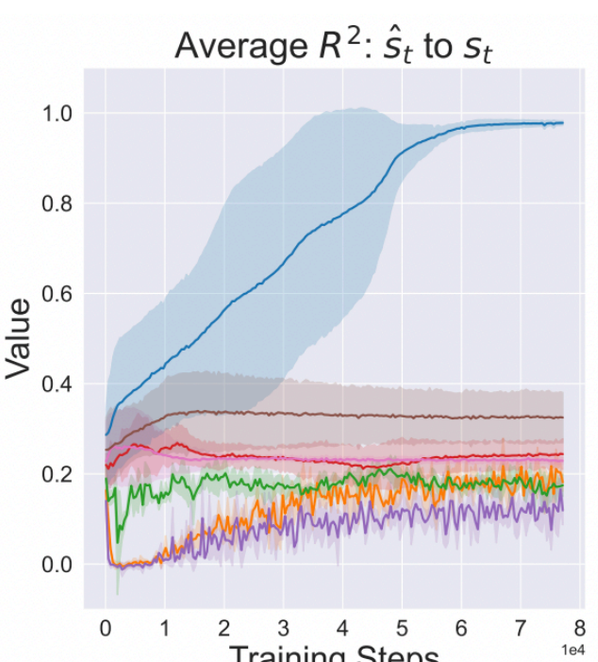
(a)



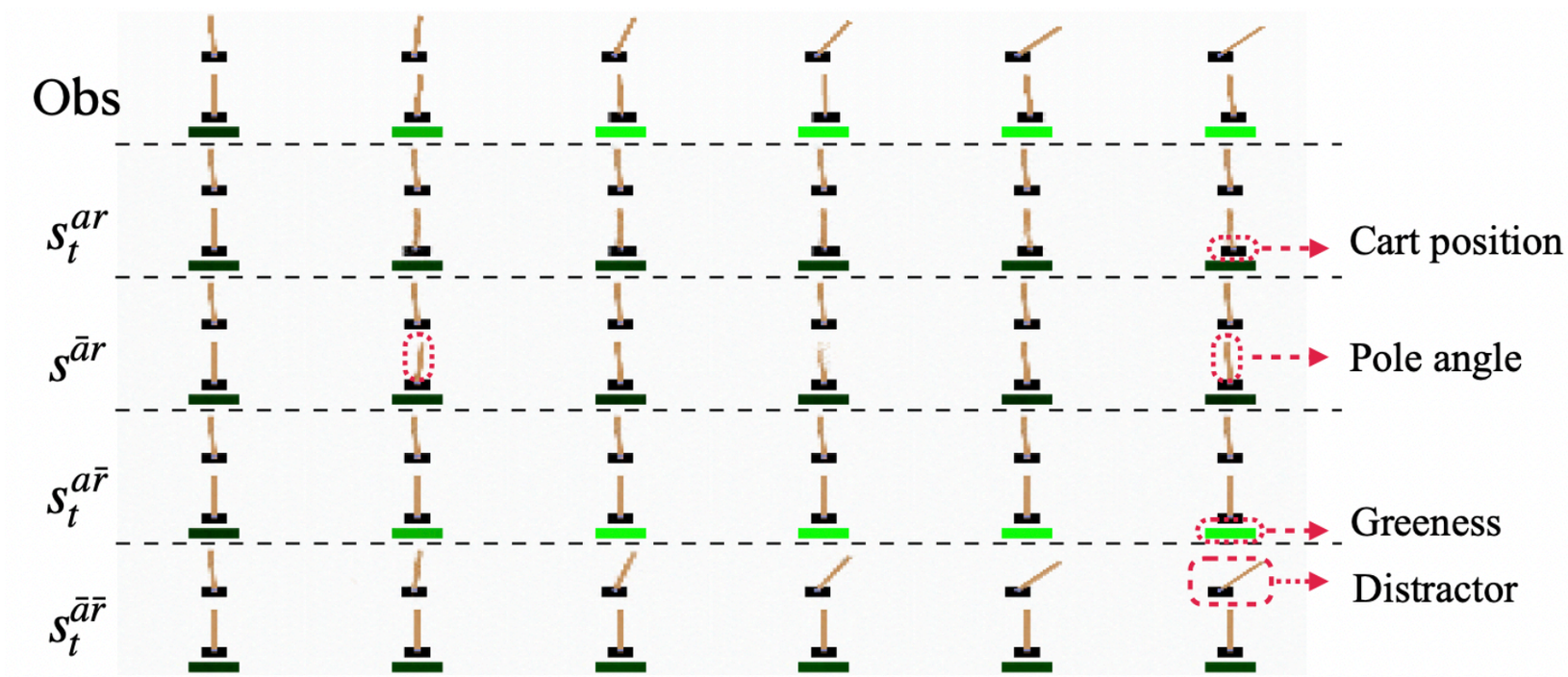
(b)



(c)



(d)



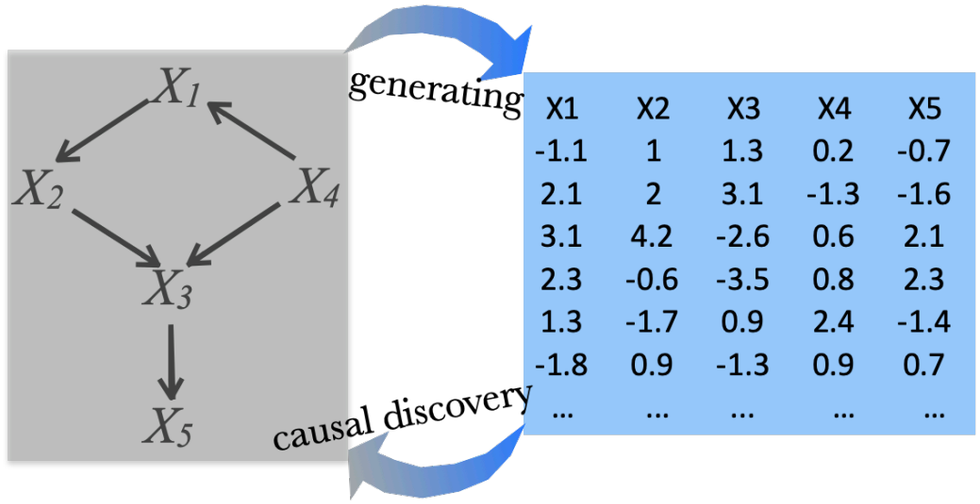
Modified Cartpole with two distractors

Where Are We?

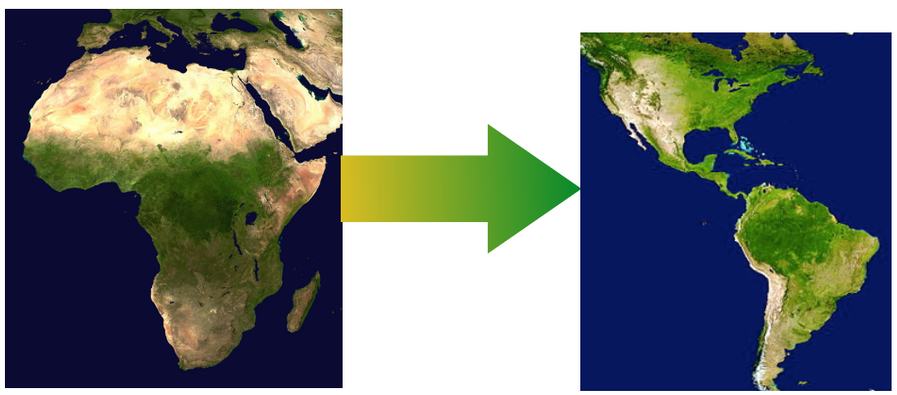
i.i.d. data?	Parametric constraints?	Latent confounders?	What can we get?
Yes	No	No	(Different types of) equivalence class
		Yes	
	Yes	No	Unique identifiability (under structural conditions)
		Yes	
Non-I, but I.D.	No/Yes	No	(Extended) regression
		Yes	Latent temporal causal processes identifiable!
I., but non-I.D.	No	No	?
	Yes		
	No	Yes	
	Yes		

Outline

- Causal thinking
- Causal representation learning: IID case
- Causal representation learning from time series

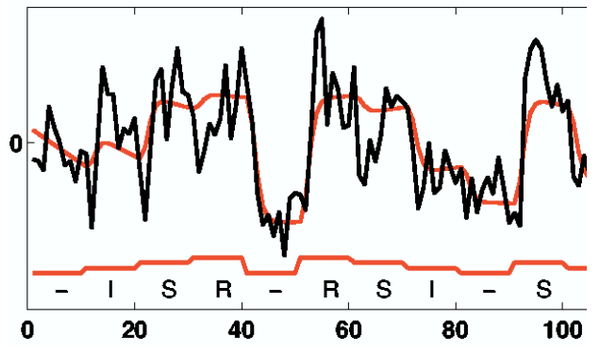


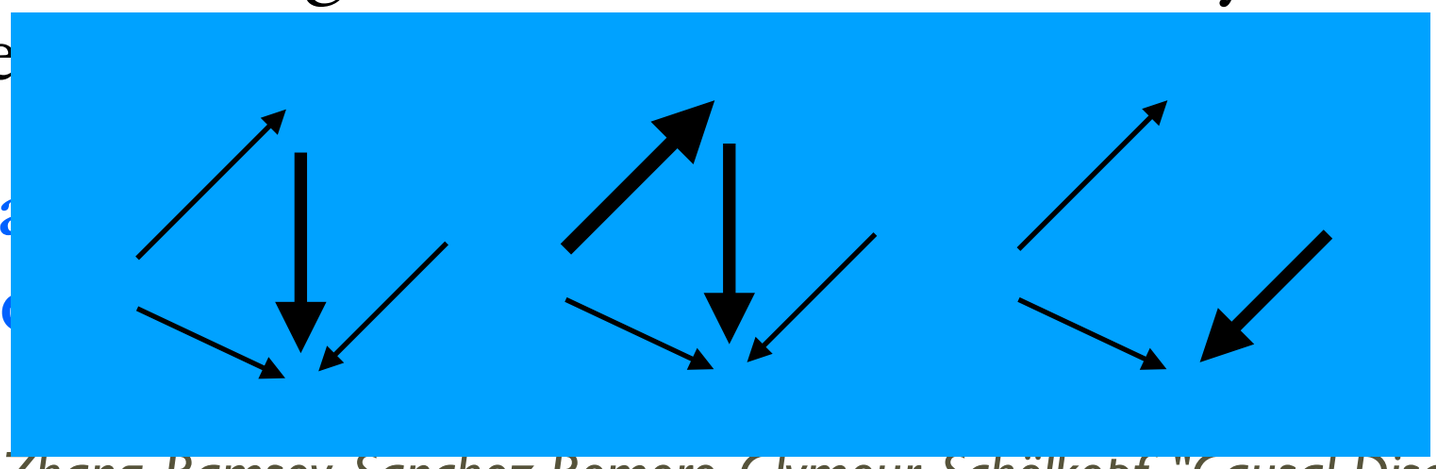
- Causal representation learning from heterogeneous/ nonstationary data
  - Transfer/adaptive learning & generative AI



Nonstationary/Heterogeneous Data and Causal Modeling

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling & distribution shift heavily couple
  - $P(c$ inde





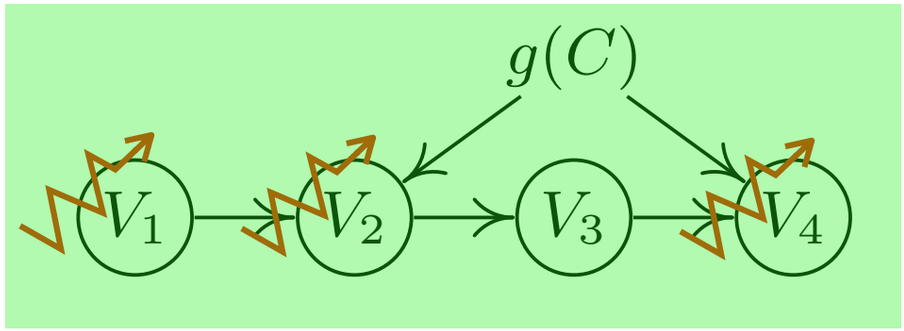
Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/ Nonstationary Data," JMLR, 2020

Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015

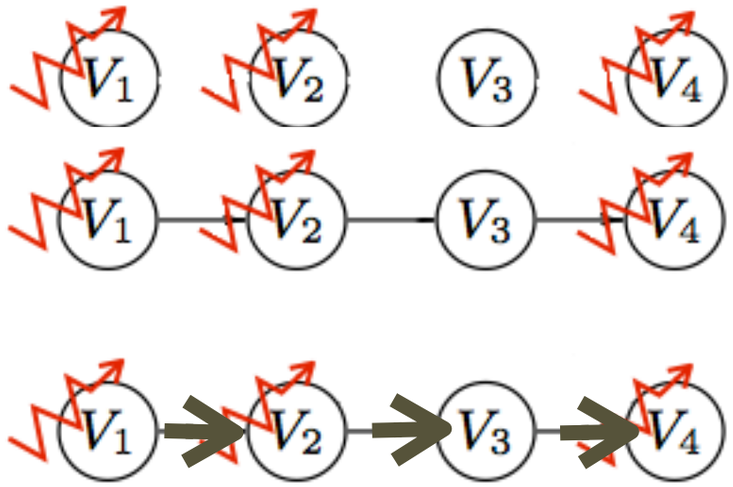
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

Causal Discovery from Nonstationary/ Heterogeneous Data

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	*Yes*



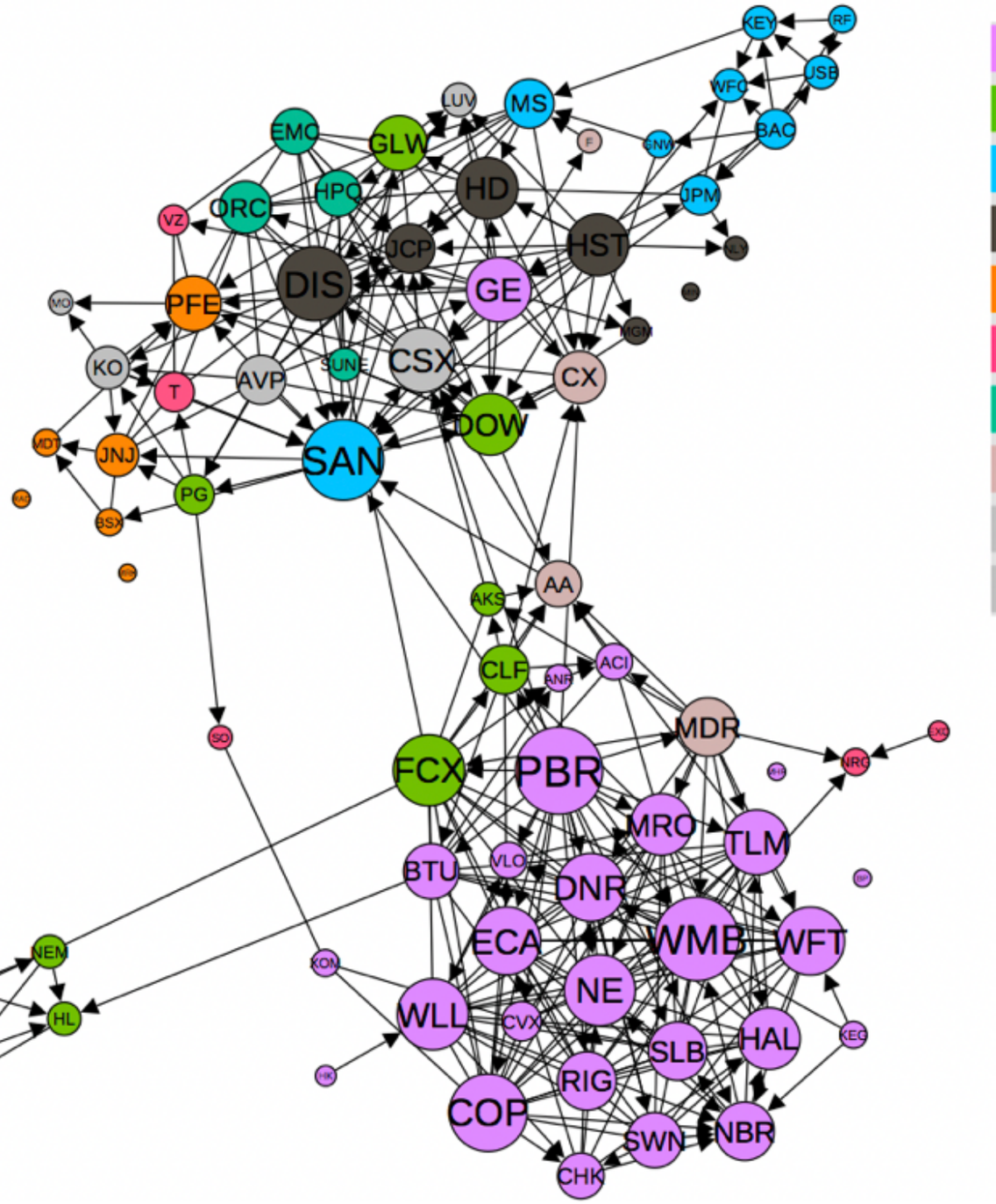
- Task:
  - Determine changing causal modules & estimate skeleton
- Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause), including invariant mechanism/ cause as special cases
- Visualization of changing modules over time/ across data sets?

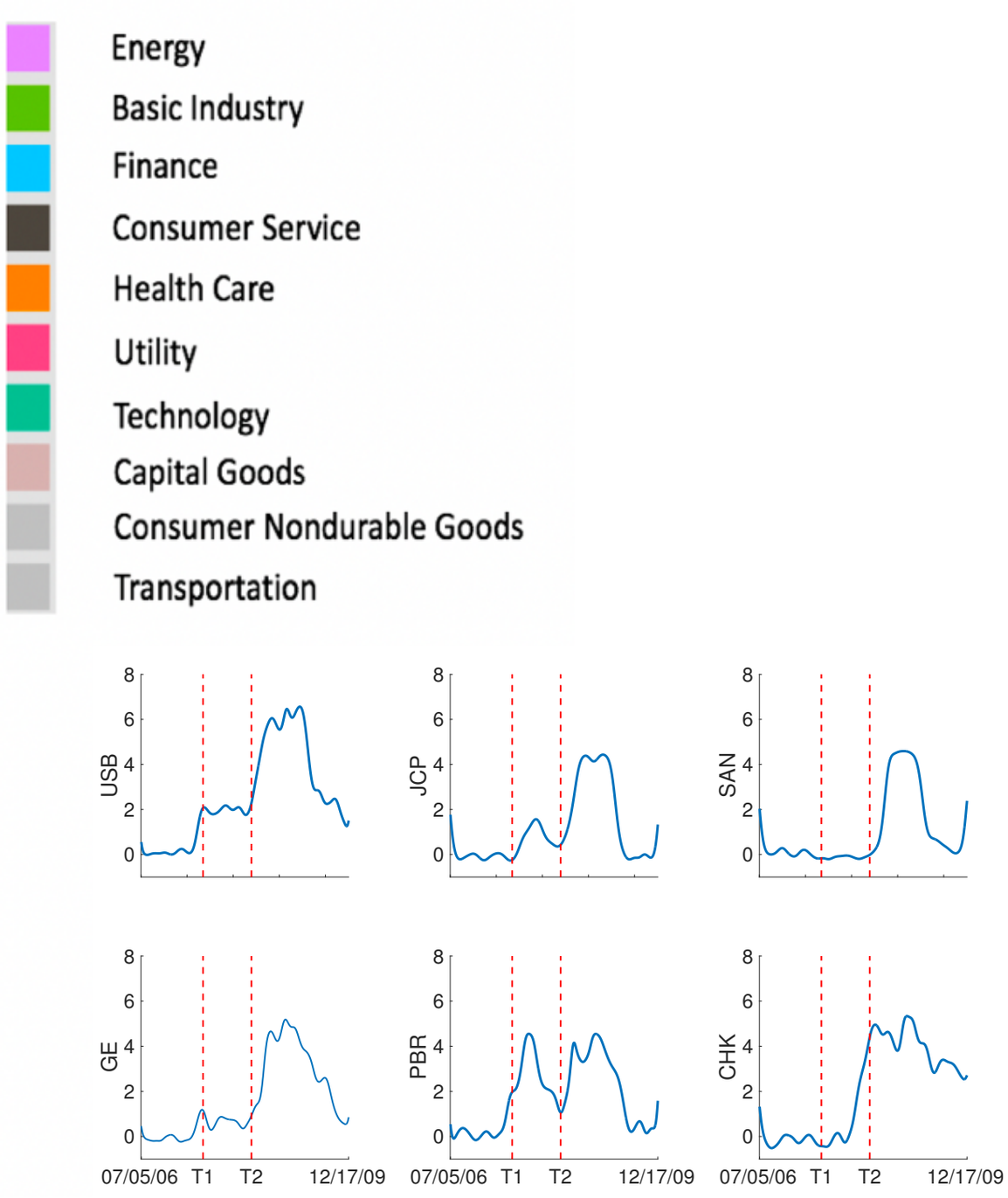


Kernel nonstationary driving force estimation

- *Huang et al., "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020*
- *Tian, Pearl, "Causal discovery from changes," UAI 2001*
- *Hoover, "The logic of causal inference" Economics and Philosophy, 6:207–234, 1990.*

Causal Analysis of Major Stocks in NYSE (07/05/2006 - 12/16/2009)

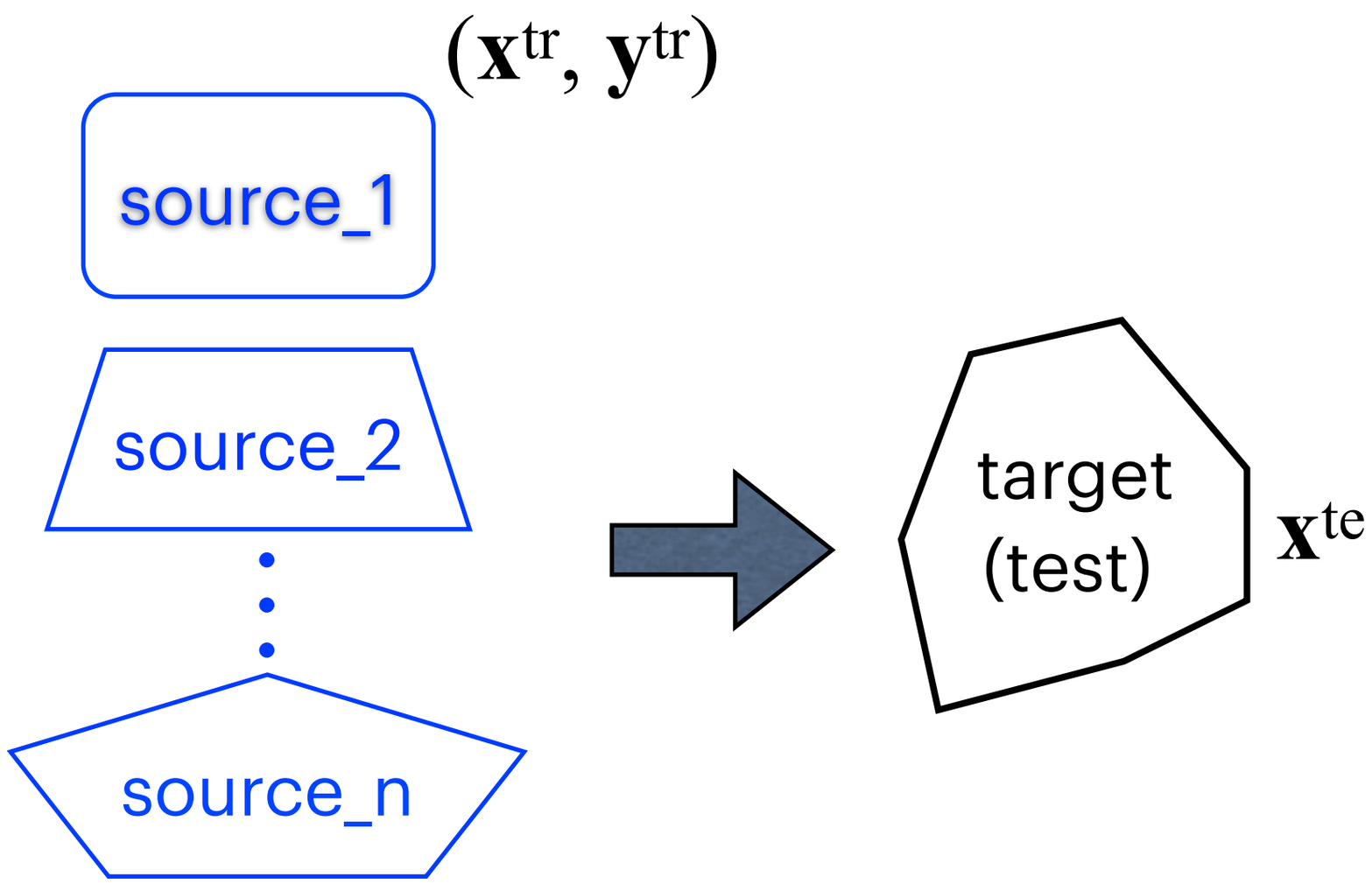




- Huang, Zhang, Zhang, Romero, Glymour, Schölkopf, Behind Distribution Shift: Mining Driving Forces of Changes and Causal Arrows," ICDM 2017

Domain Adaptation

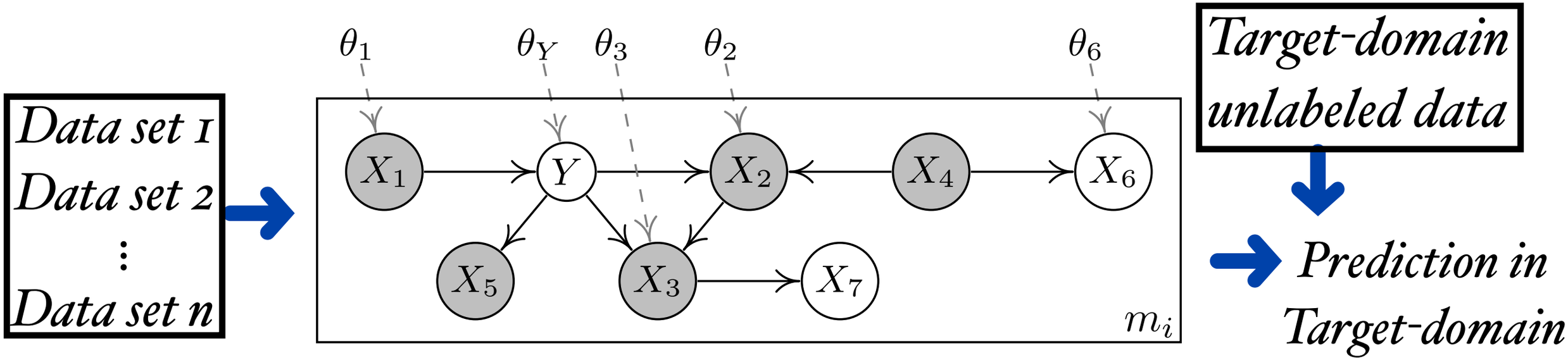
- Traditional supervised learning:
  $P_{XY}^{te} = P_{XY}^{tr}$
- Might not be the case in practice
- How to leverage information in source domains?



high-level representation, e.g., $Y \rightarrow X$

Prob. model $P^{(1)}(X,Y)$, $P^{(2)}(X,Y)$, $P^{(3)}(X,Y)$, ... $P^{(k)}(X,Y)$...

An Approach to Data-Driven Domain Adaptation



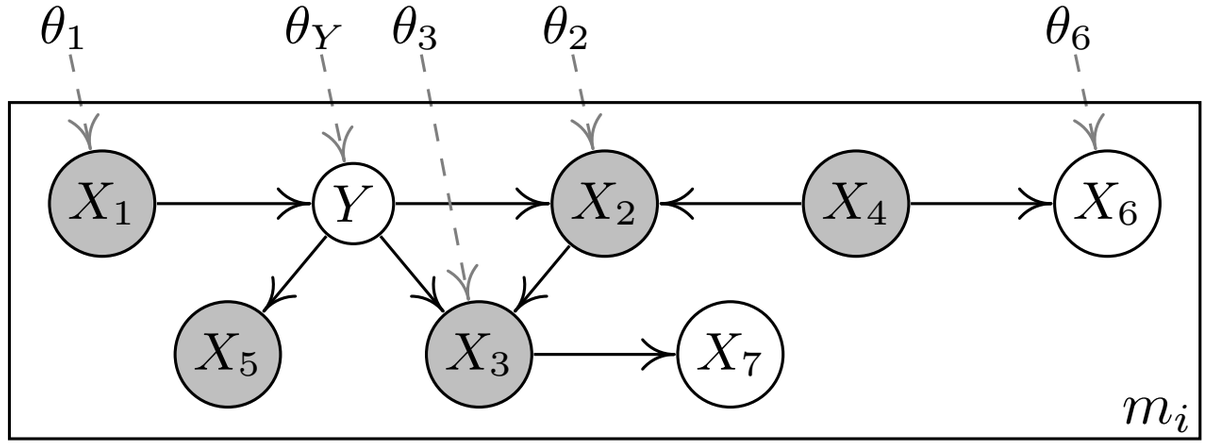
- Only relevant features needed to predict Y
- Augmented graph learned by CD-NOD
  - Independently changing modules θ_i
  - Special case: invariant modules
- Domain adaptation: inference on this graphical model
  - Infer the po
  - Nonparame

Judea Pearl @yudapearl · Feb 14, 2020

For ML folks, "domain adaptation" connotes an insurmountable obstacle. For CI folks it is a causal graphs problem embraced under "transportability" theory. This paper arxiv.org/pdf/2002.03278... views the problem as Bayes inference on graphical models. #Bookofwhy

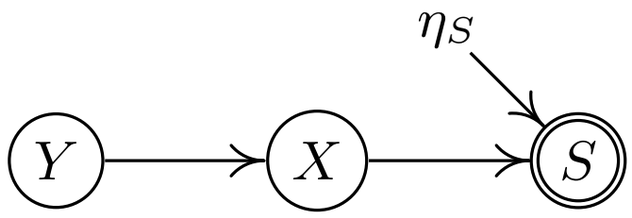
Zhang, Gong*, Stojanov, Hu*
Models," NeurIPS 2020. (Huang et al., ICML'19 for time series data)

Not Necessarily Causal…

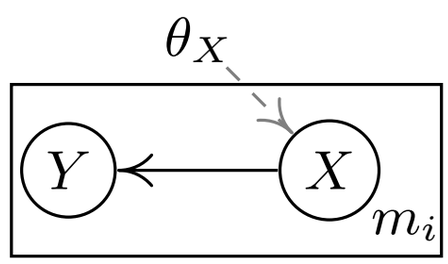


- To represent independent changes in the joint distribution
  - Causal graph vs. augmented DAG

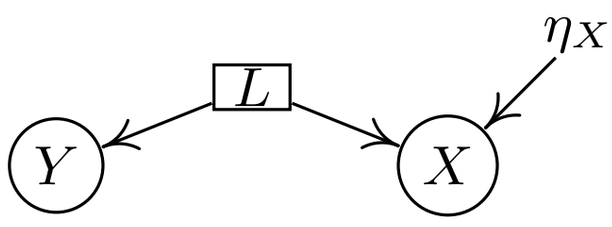
because p(Y|X) is invariant across domains



(a) The underlying data generating process of Example 1. Y generates (causes) X, and S denotes the selection variable (a data point is included if and only if $S = 1$).

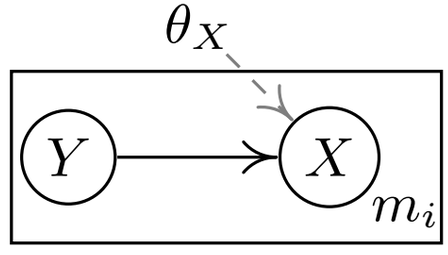


(b) The augmented DAG representation for Example 1 to explain how the data distribution changes across domains.



(c) The generating process of Example 2. L is a confounder; the mechanism of X changes across domains, as indicated by η_X.

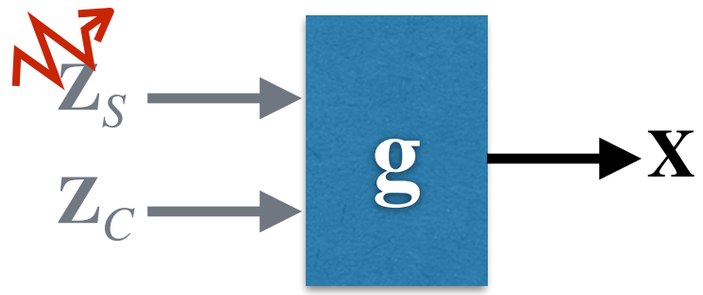
because p(Y) is invariant across domains



(d) The augmented DAG representation for Example 2 to explain how the data distribution changes across domains.

Finding Changing Hidden Variables for Transfer Learning

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Underlying components $\mathbf{Z}_S$ may change across domains
- Changing components $\mathbf{Z}_S$ are identifiable; invariant part $\mathbf{Z}_C$ are identifiable up to its subspace
- Using invariant part $\mathbf{Z}_C$ and transformed changing part $\tilde{\mathbf{Z}}_S$ for prediction

Models	→ Art	→ Clipart	→ Product	→ Realworld	Avg
Source Only (He et al., 2016)	64.58±0.68	52.32±0.63	77.63±0.23	80.70±0.81	68.81
DANN (Ganin et al., 2016)	64.26±0.59	58.01±1.55	76.44±0.47	78.80±0.49	69.38
DANN+BSP (Chen et al., 2019)	66.10±0.27	61.03±0.39	78.13±0.31	79.92±0.13	71.29
DAN (Long et al., 2015)	68.28±0.45	57.92±0.65	78.45±0.05	81.93±0.35	71.64
MCD (Saito et al., 2018)	67.84±0.38	59.91±0.55	79.21±0.61	80.93±0.18	71.97
M3SDA (Peng et al., 2019)	66.22±0.52	58.55±0.62	79.45±0.52	81.35±0.19	71.39
DCTN (Xu et al., 2018)	66.92±0.60	61.82±0.46	79.20±0.58	77.78±0.59	71.43
MIAN (Park & Lee, 2021)	69.39±0.50	63.05±0.61	79.62±0.16	80.44±0.24	73.12
MIAN-γ (Park & Lee, 2021)	69.88±0.35	**64.20±0.68**	80.87±0.37	81.49±0.24	74.11
iMSDA (Ours)	**75.77±0.21**	60.83±0.73	**84.13±0.09**	**84.83±0.12**	**76.39**

Table 2. Classification results on Office-Home. Backbone: Resnet-50. Baseline results are taken from (Park & Lee, 2021).

- *Kong, Xie, Yao, Zheng, Chen, Stojanov, Akinwande, Zhang, Partial disentanglement for domain adaptation, ICML 2022*

Unsupervised Image-to-Image Translation

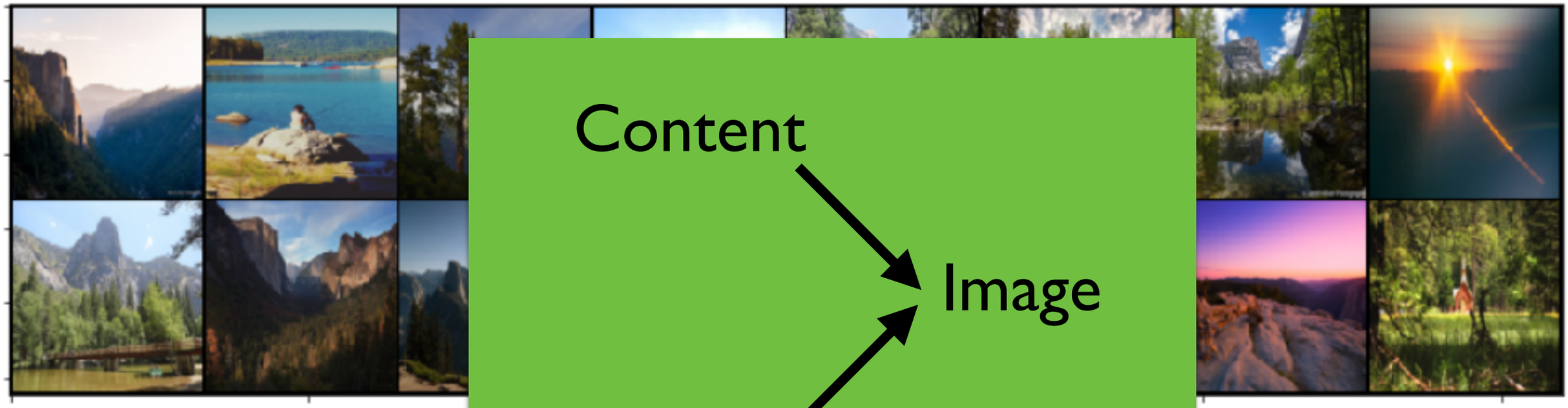
A simpler example:

Domain 1:

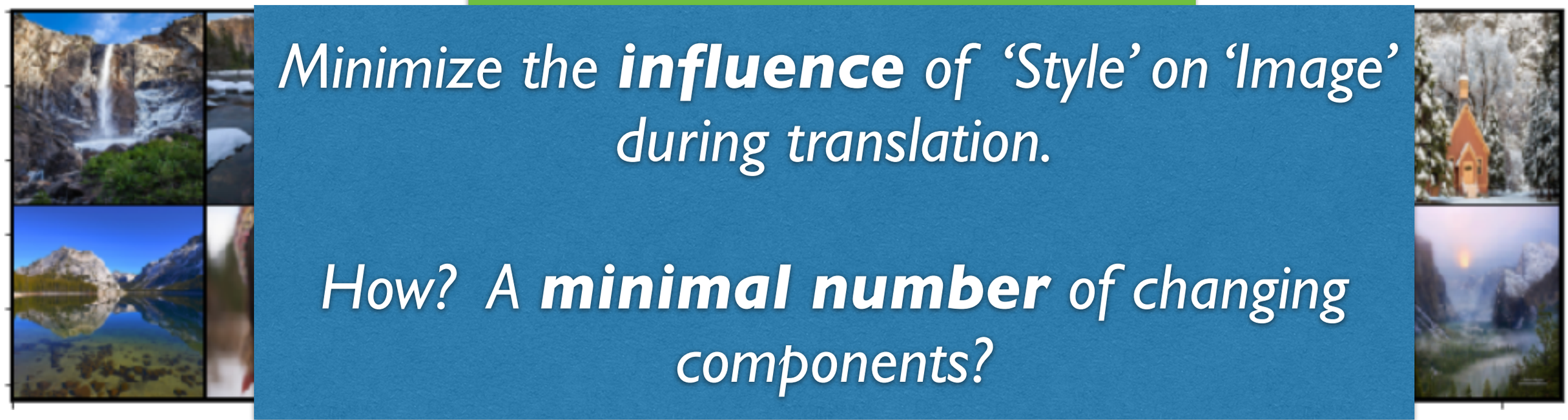
Domain 2:

Images from the winter season domain.

Unsupervised Image-to-Image Translation



Im ain.



Content

Style

Image

*Minimize the **influence** of 'Style' on 'Image' during translation.*

*How? A **minimal number** of changing components?*

Images from the winter season domain.

Multi-domain Image Generation & Translation with Identifiability Guarantees

- Idea: Matching the distributions across domains **with a minimal number of changing components**
- Correspondence info (joint distribution) identifiable under mild assumptions
- Example: Generating female & male images with the same "content"

Ours

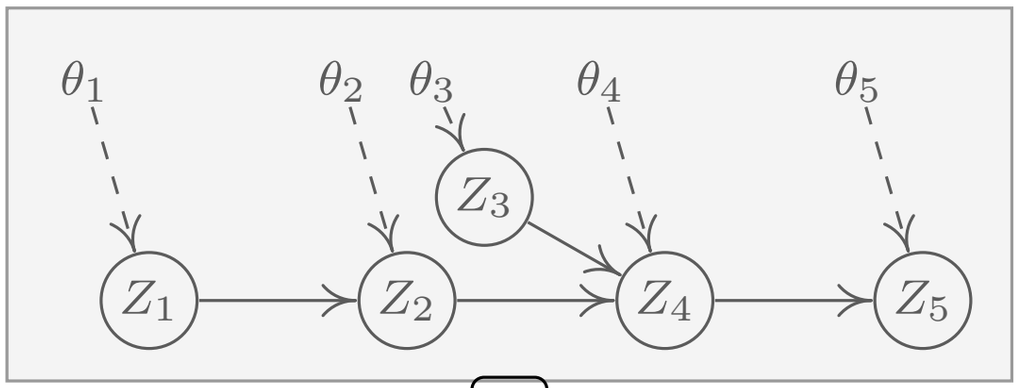
StyleGAN2-ADA

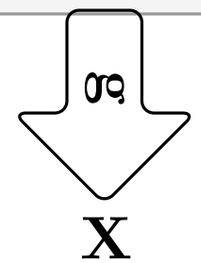
TGAN

- *Xie, Kong, Gong, Zhang, "Multi-domain image generation and translation with identifiability guarantees", ICLR 2023*
- *Yan, Kong, Gui, Chi, Xing, He, Zhang, Counterfactual Generation with Identifiability Guarantee, NeurIPS 2023*

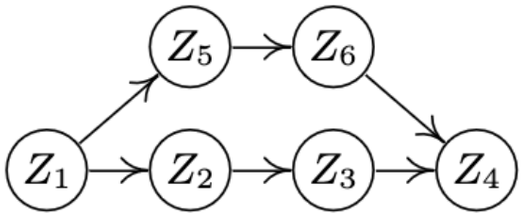
Causal Representation Learning from Multiple Distributions: A *General* Setting

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes

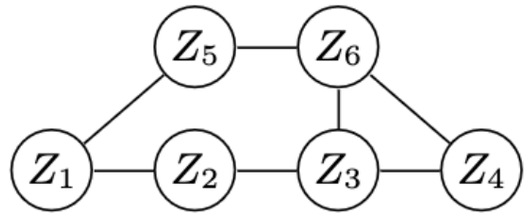




- Markov network (MN) of Z_i can be recovered
- Each estimated variable $\tilde{Z}_i$ is a function of Z_i and it intimate neighbors
  - Z_i's intimate neighbor is adjacent to Z_i and all the other neighbors of Z_i in the MN
- In this example, each Z_i ($i \neq 4$) can be recovered up to component-wise transformation!



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.



(b) The corresponding Markov network $\mathcal{M}_Z$.

- *Zhang, Xie, Ng, Zheng, "Causal Representation Learning from Multiple Distributions: A General Setting," ICML 2024*

Summary

- Various tasks involve suitable (causal) representations of data
  - Domain generalization/adaptation, trustworthy AI, explainable AI, fairness…
- Causal representations can be recovered under the appropriate assumptions
  - Technically operational causal principles
  - Identifiability!
    - Strong identifiability results in non-IID cases
    - Benefit from parametric constraints in the IID case
- Understanding language of nature: causal generation + **selection**