Tools on Causality

Causal Discovery & causal-learn

Instructor: Yujia Zheng

Most slides are from Kun Zhang

Carnegie Mellon University

Outline

- Lecture 1&2 (Monday): Introduction of causal discovery and causal-learn.
- Lecture 3 (Tuesday): Lab for small projects.
- Lecture 4 (Thursday): Presentations

Project

- Flexible small projects
- Incorporating causal discovery into any topics of interests.
- Demo, analysis, report, complaint...
- Groups of one or two people.

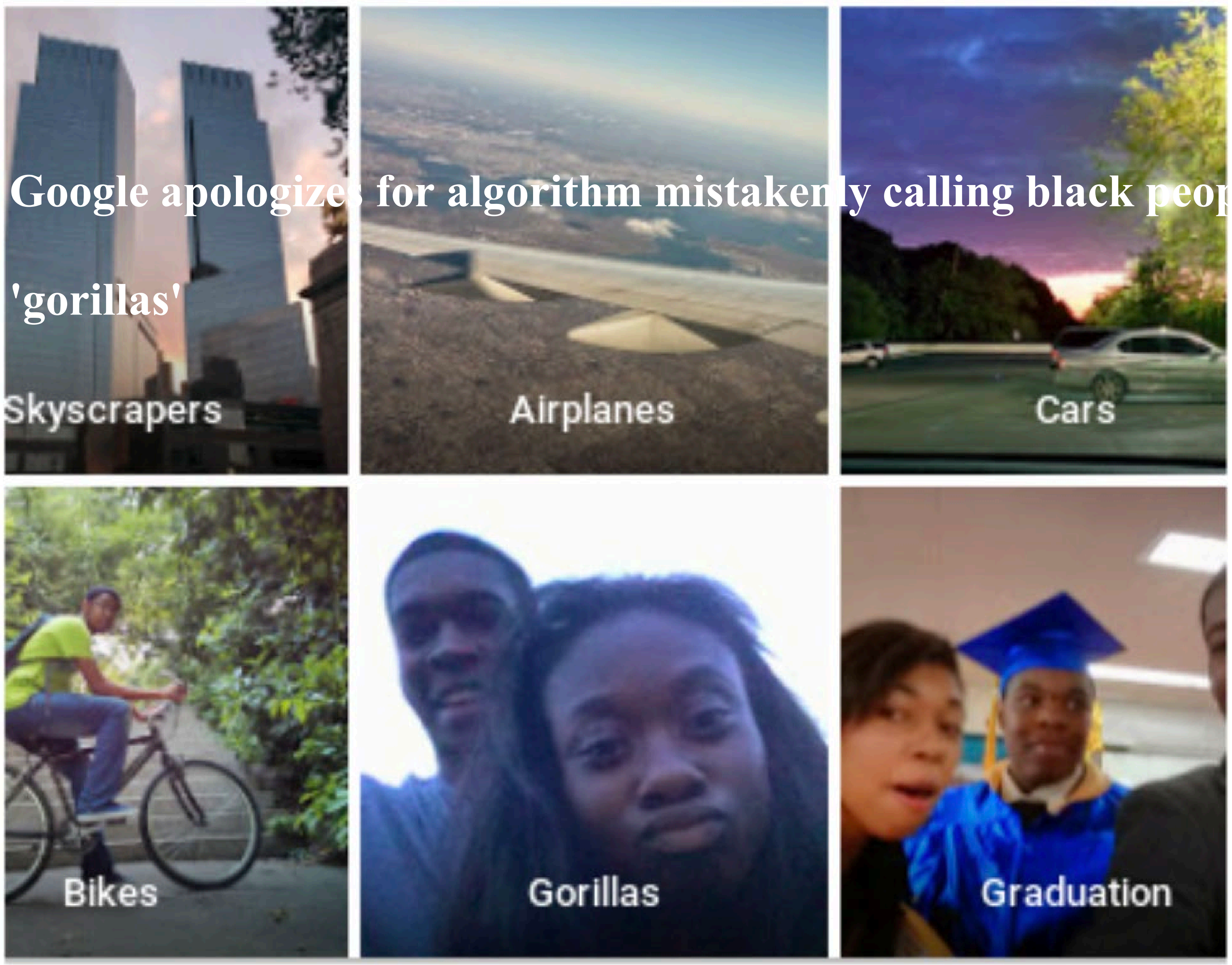
- Timeline:

1. By Monday 23:59: Grouping information. Send it to yujiazh@cmu.edu or Slack channel.
2. Tuesday afternoon: Guided lab to work on projects.
3. Thursday afternoon: Small presentations. Length depends on the grouping information.

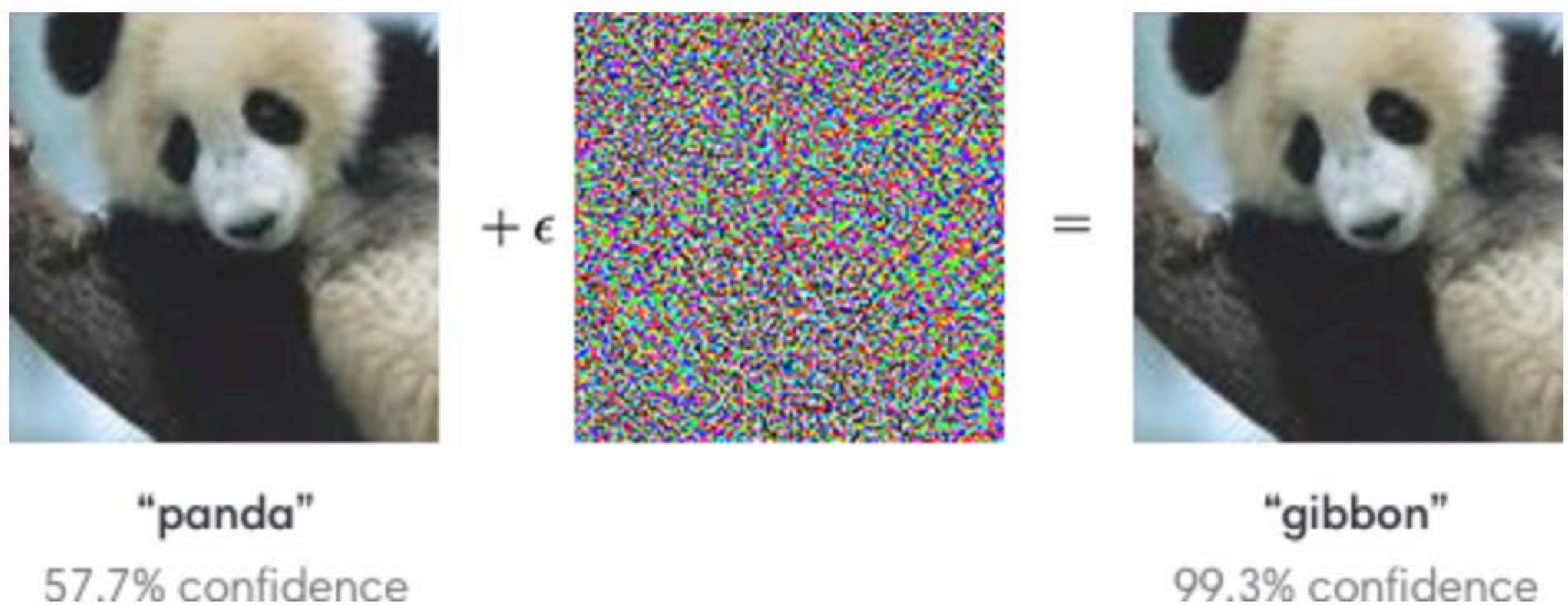
A Big Picture of Causal Discovery

- Necessity of causality
- Causality from observational data
- Quick examples on the advancements

A Problem with Photo Categorization by Google Photos



A Bit Noise can Dramatically Change Machines' Decision

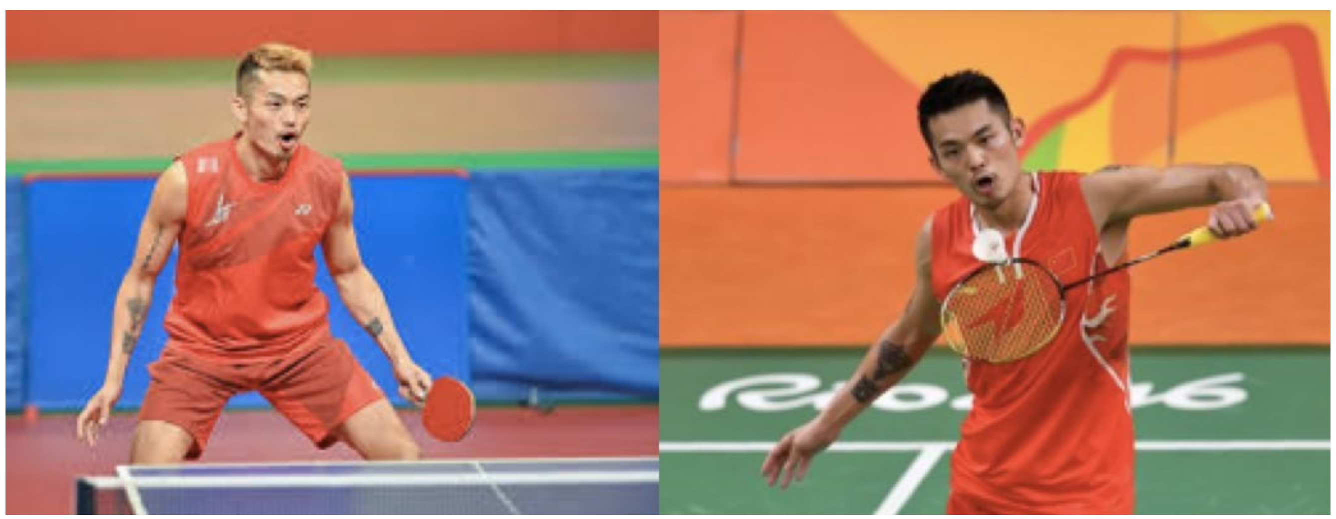


An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

(Goodfellow, 2015)

Artificial "Intelligence"

- Traditional machine learning usually assumes a fixed data distribution; avoids overfitting



- Intelligence: understanding; control/intervention; decomposability; information fusion, learning with few examples, extrapolation

Causality Examples

The Telegraph

Home | News | World | Sport | Finance | Comment | Culture | Travel | Life | Women | Fashion | Lu

USA | Asia | China | Europe | Middle East | Australasia | Africa | South America | Central Asia

France | Francois Hollande | Germany | Angela Merkel | Russia | Vladimir Putin | Greece | Spa

HOME » NEWS » WORLD NEWS » EUROPE

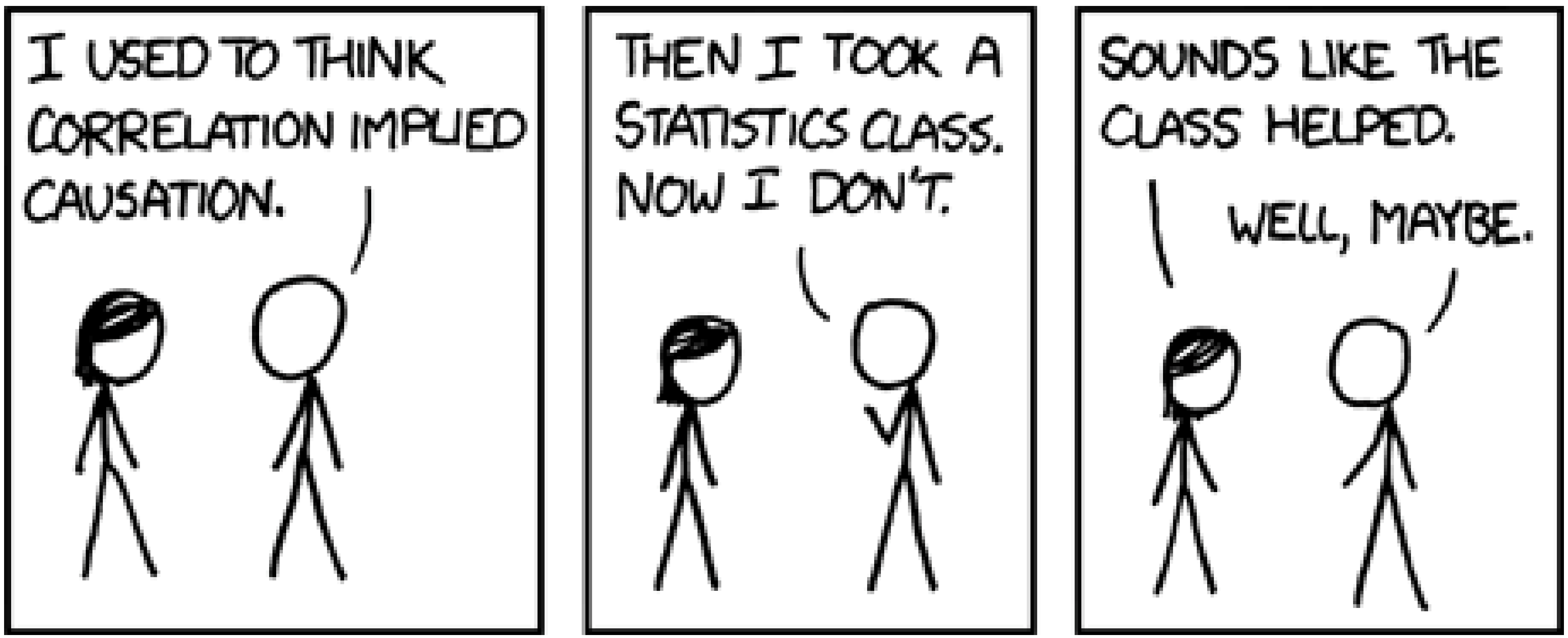
Couples who share the housework are more likely to divorce, study finds

Divorce rates are far higher among "modern" couples who share the housework than in those where the woman does the lion's share of the chores, a Norwegian study has found.

Causality Examples

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

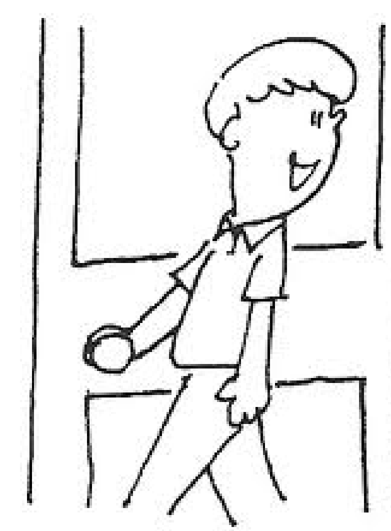
$\exists x_1 \neq x_2 \; P(Y|\text{set } X=x_1) \neq P(Y|\text{set } X=x_2)$

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?
- What if you manipulate X and see Y also changes?
  - A manipulation directly changes only the target variable X





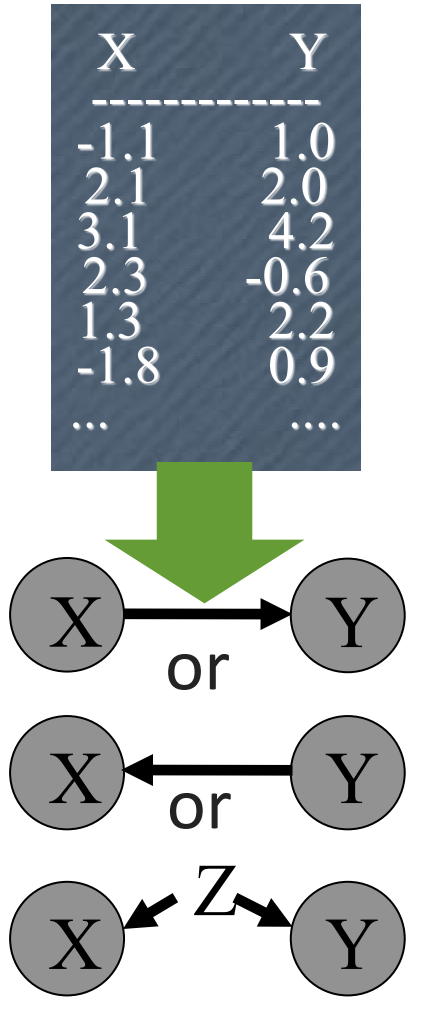


Causal Discovery

Possible to

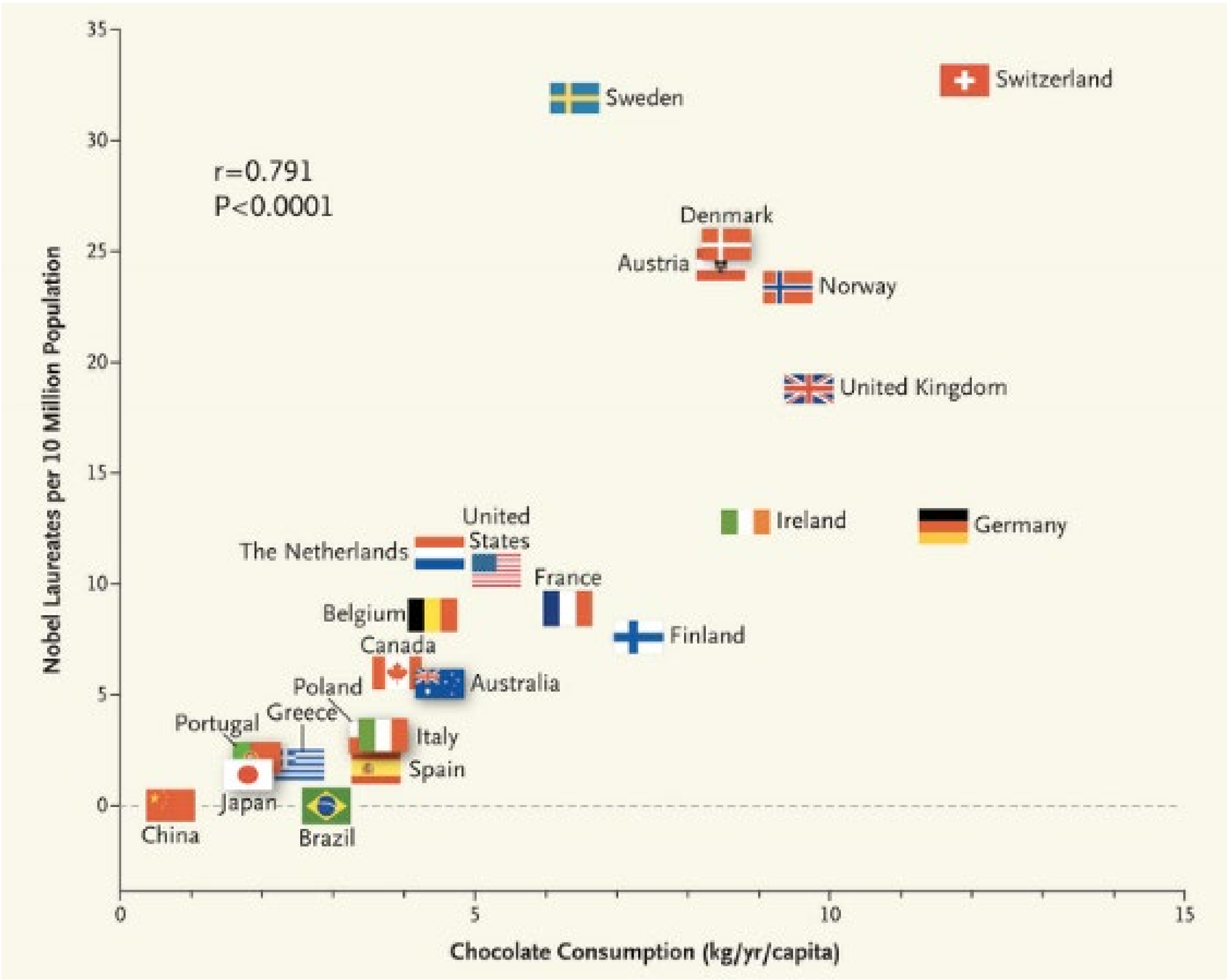
discover causal information (*specific properties of the true process*)

from purely observational data ?

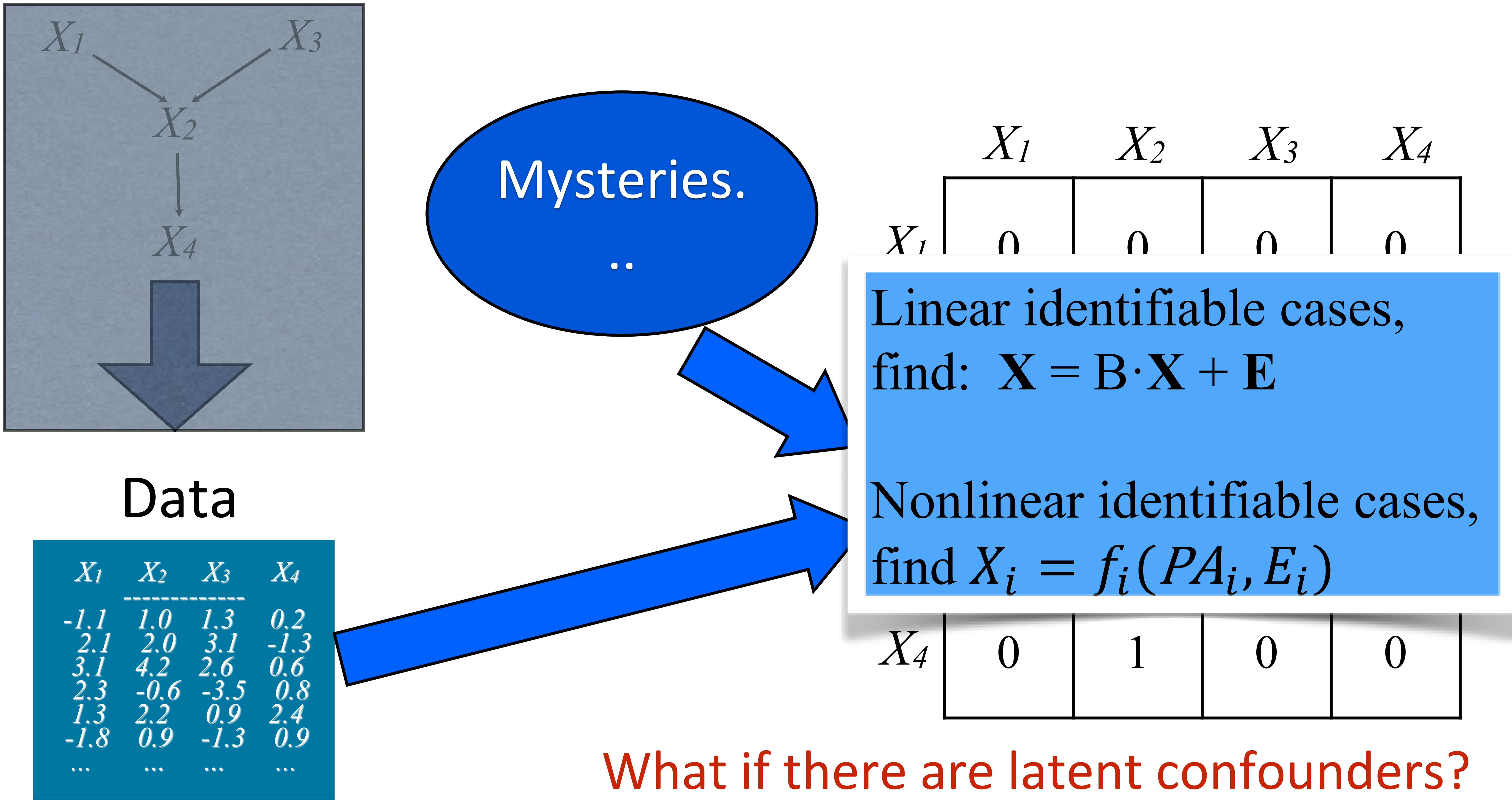


Can we go beyond the data?

Causality Examples

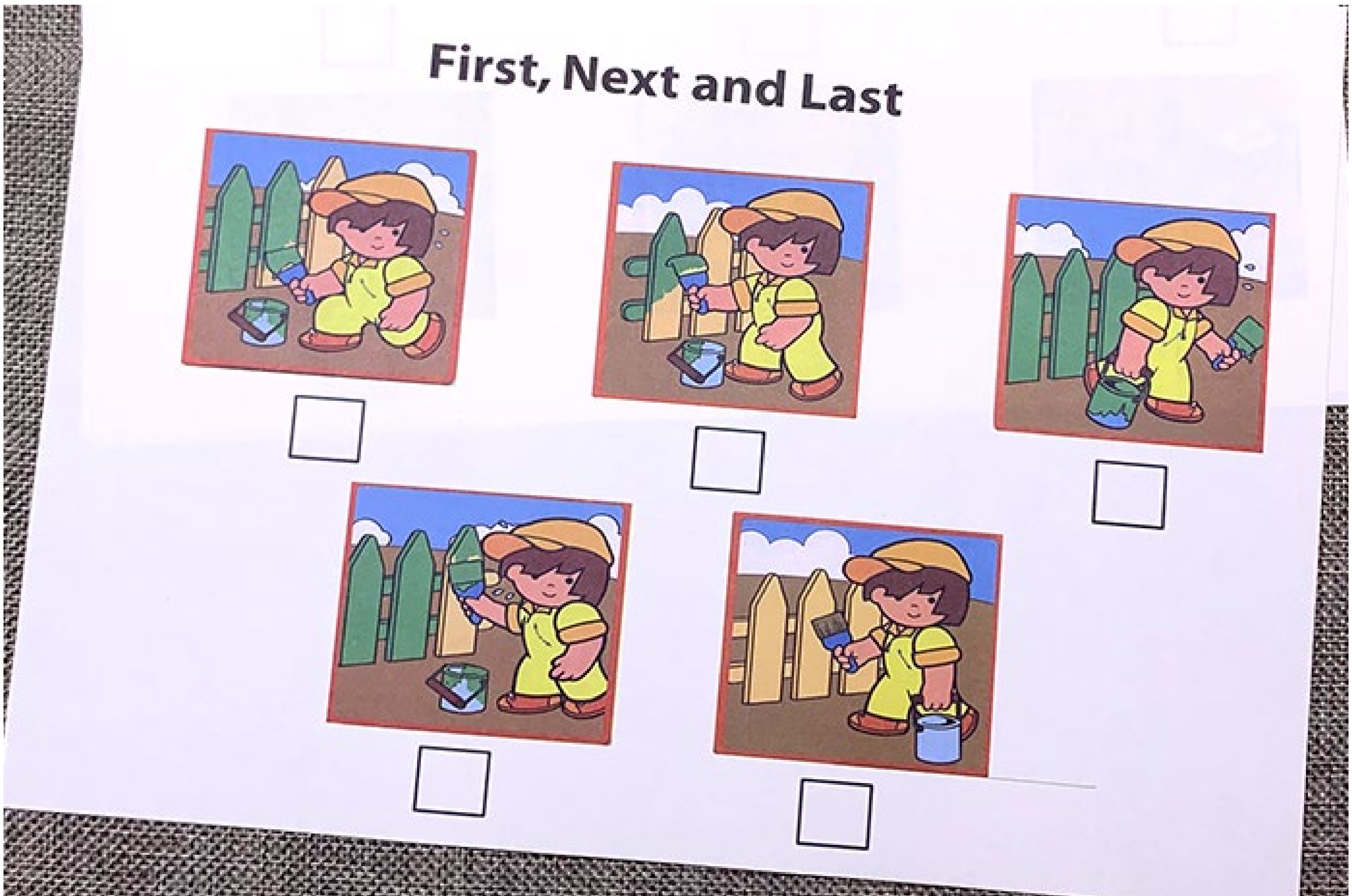


(Simple) Causal Discovery as an Estimation Problem

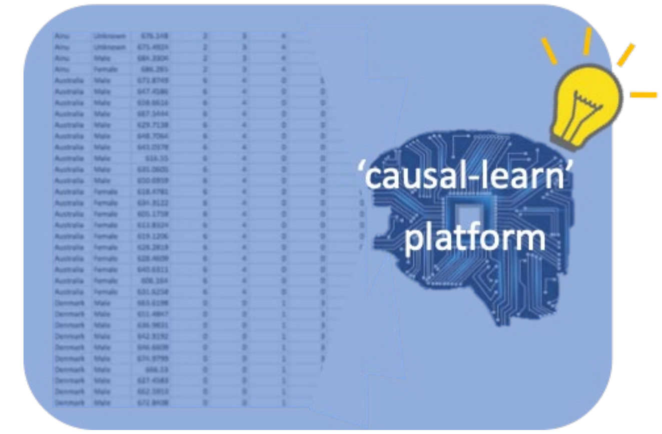


What if there are latent confounders?

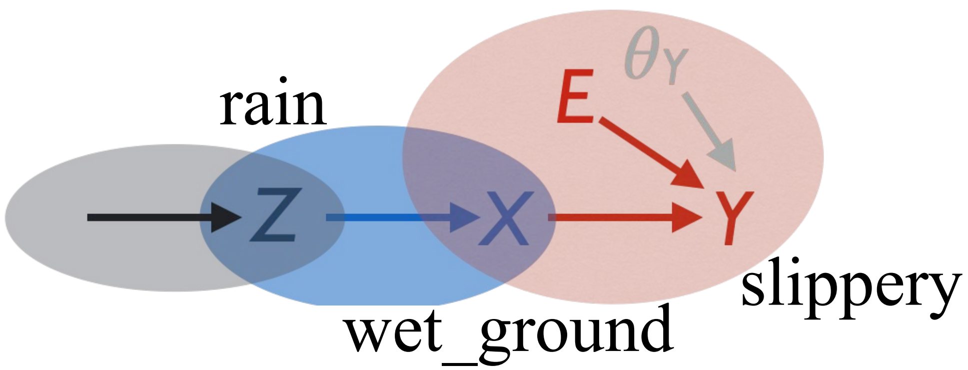
Temporal Order Often Helpful. I.I.D. Case More Difficult.



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 Discovery in Archeology: An Example

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

Thanks to Marlijn Noback

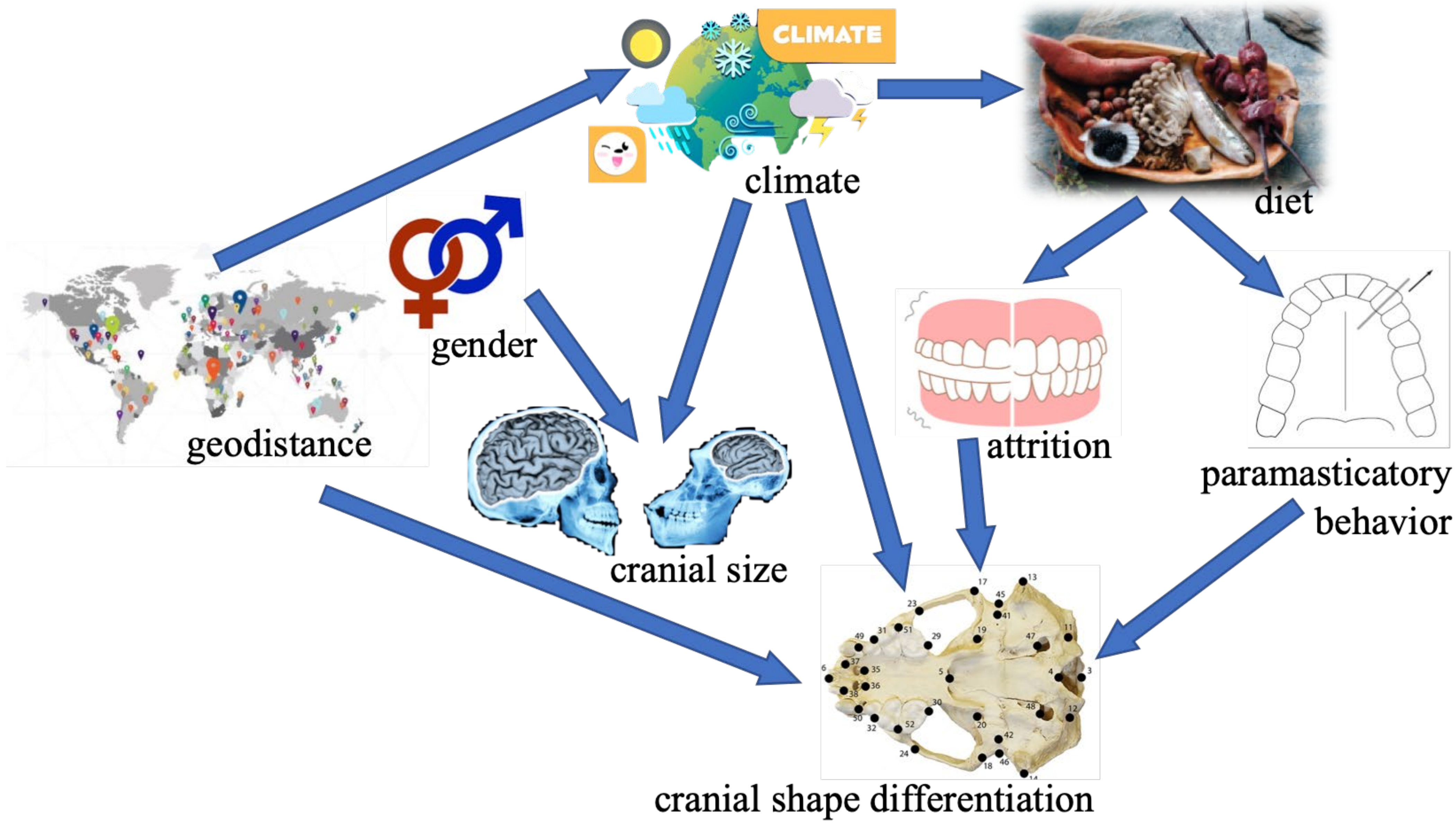
- 8 variables of 250 skeletons collected from different locations

	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

Result of PC on the Archeology Data

Thanks to collaborator Marlijn Noback

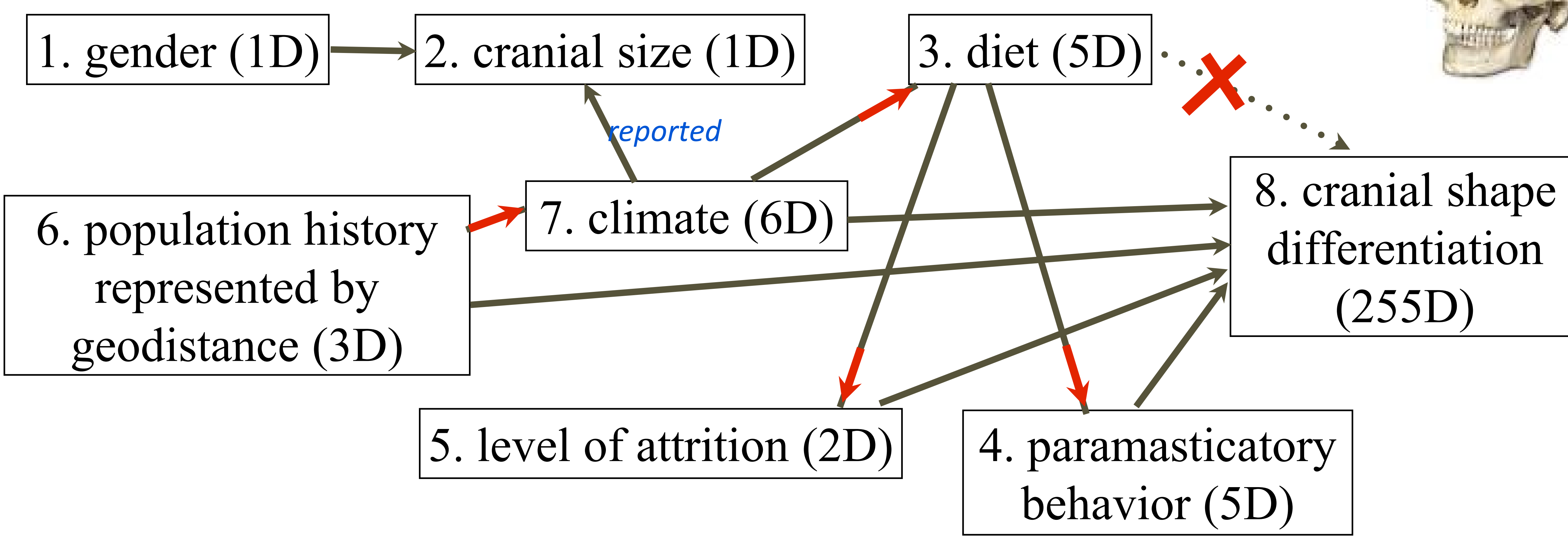
- By PC algorithm (Spirtes et al., 1993) + kernel-based conditional independence test (Zhang et al., 2011)



Result on the Archeology Data

Thanks to collaborator Marlijn Noback

- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



A Problem in Psychology: Finding Underlying Mental Conditions?

- 50 questions for big 5 personality test

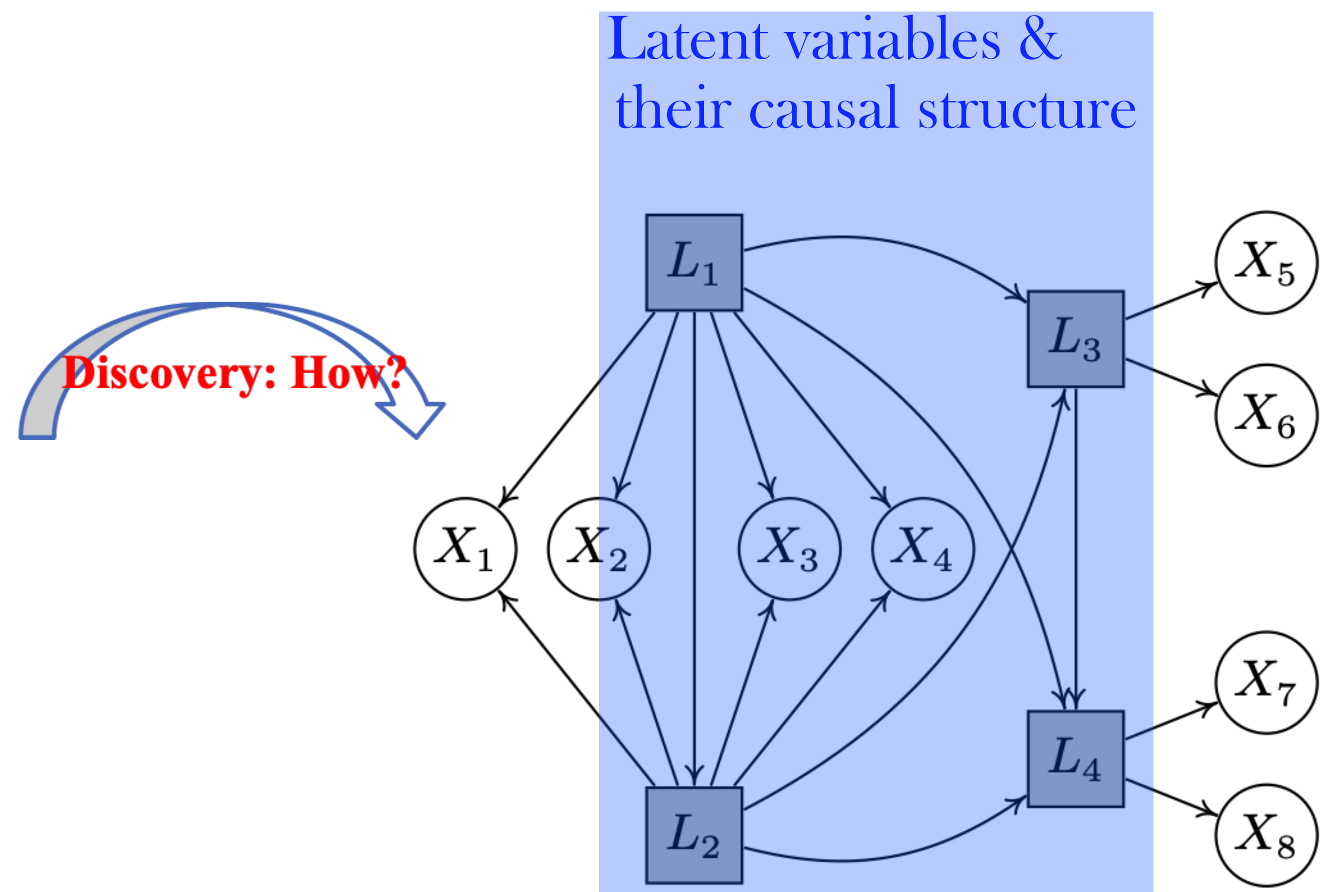
race	age	engnat	gender	hand	source	country	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	A1	A2	A3	A4	A5
3	53	1	1	1	1	US	4	2	5	2	5	1	4	3	5	1	1	5	2	5	1	1	1	1	1	1	1	5	1	5	2
13	46	1	2	1	1	US	2	2	3	3	3	3	1	5	1	5	2	3	4	2	3	4	3	2	2	4	1	3	3	4	4
1	14	2	2	1	1	PK	5	1	1	4	5	1	1	5	5	1	5	1	5	5	5	5	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	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

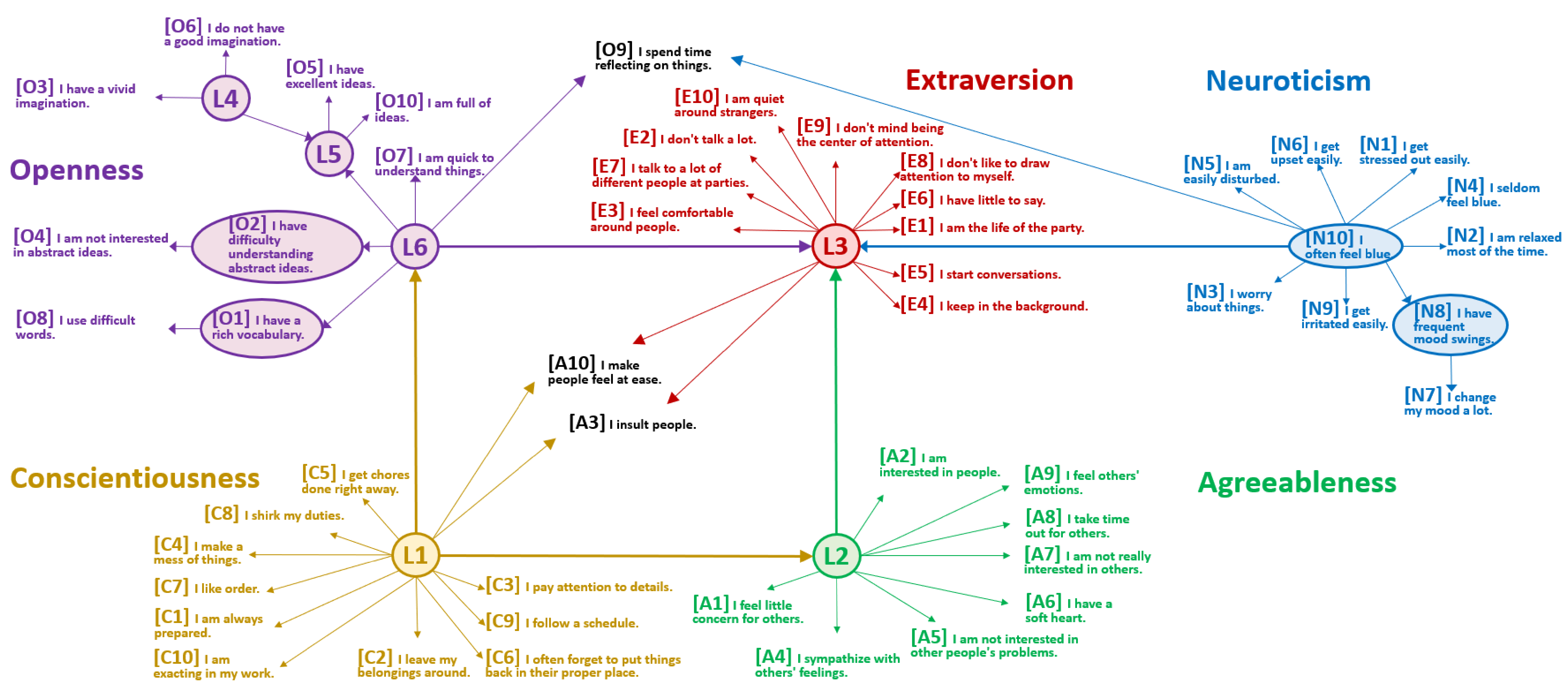
X1	X2	X3	X4	X5	X6	X7	X8
4.2	3.6	6.5	6.8	9.6	7.6	2.7	4.8
3.8	1.9	6.5	7.3	8.9	6.9	1.1	4.6
4.2	3.4	6.5	6.9	9.5	7.4	2.5	4.6
4.2	2.2	6.2	6.9	9.6	7.2	1.9	4.8
3.9	1.9	6.5	6.8	9.0	6.8	1.7	4.4
4.0	2.0	6.4	7.2	9.1	7.0	1.0	4.6
3.8	1.7	6.4	7.3	9.0	6.7	0.8	4.3
4.1	2.8	6.5	6.9	9.3	6.7	2.7	4.6
...	...	...	...	...	...	...	...



- Find latent variables L_i and their causal relations ?
- Rank deficiency or GIN helps solve the problem

Example: Big 5 Questions Are Well Designed but...

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



Nice results by Xinshuai Dong

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

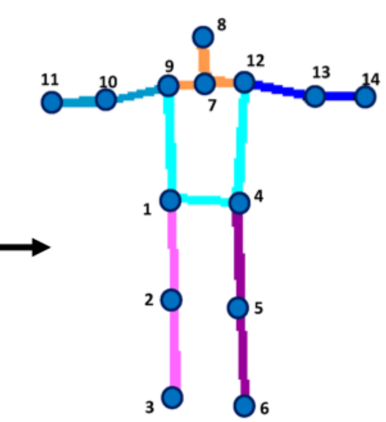
$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t$

Latent processes

Latent temporal causal processes z_{it} can be recovered if they follow

- completely **nonparametric** model; or furthermore,
- **non-stationary** noise; or
- **non-stationary** causal influence, or
- **Parametric** constraints

Causal Skeleton Recovery



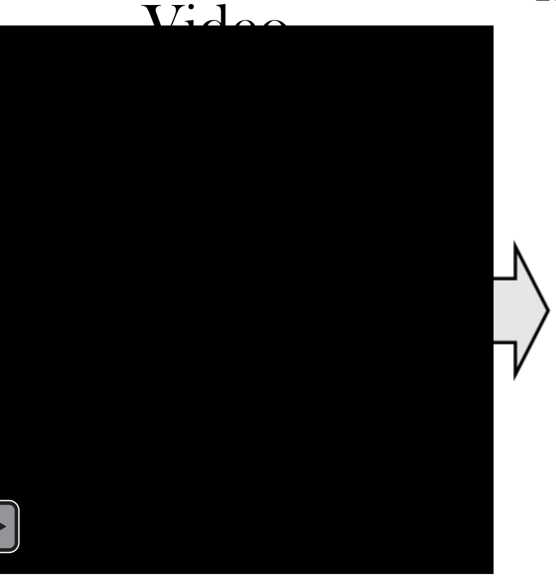
Recovered latent processes

- *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*

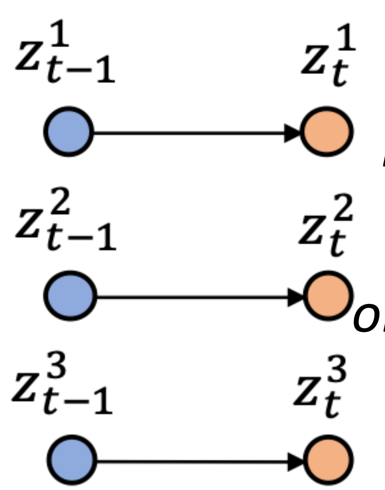
Results on Video Data

- For easy interpretation, consider two simple video data sets
  - KiTTiMask: a video dataset of binary pedestrian masks
  - Mass-spring system: a video dataset with ball movement and invisible springs

KiTTiMask



Learned latent processes



Interpretation

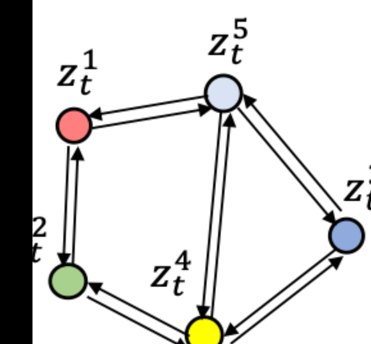
(Movement in a direction)

(Movement in an orthogonal direction)

(Size)

Mass-spring

Learned latent processes



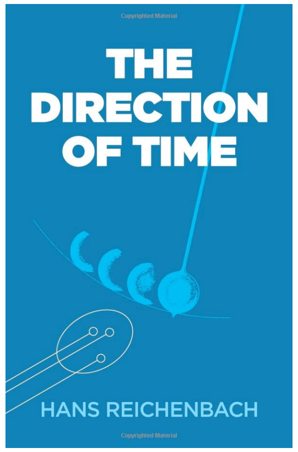
Interpretation

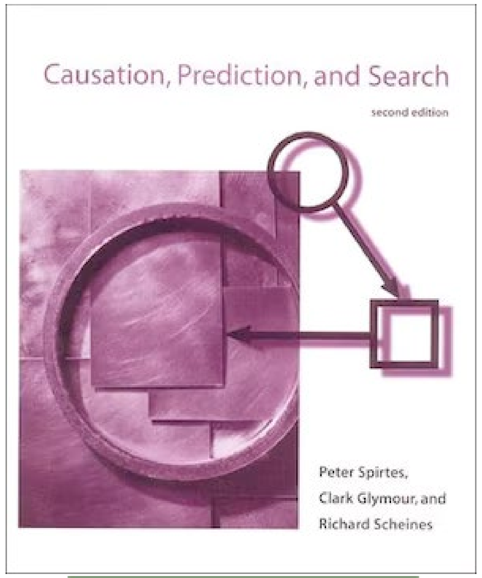
(x- & y-coordinates of the 5 balls)

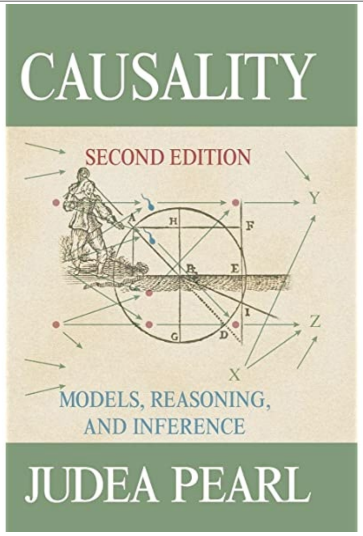
- *Yao, Chen, Zhang, "Learning Latent Causal Dynamics," NeurIPS 2022*
- *Yao, Sun, Ho, Sun, Zhang, "Learning Temporally causal latent processes from general temporal data," ICLR 2022*

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 (GDUT & CMU), rank deficiency (CMU)...



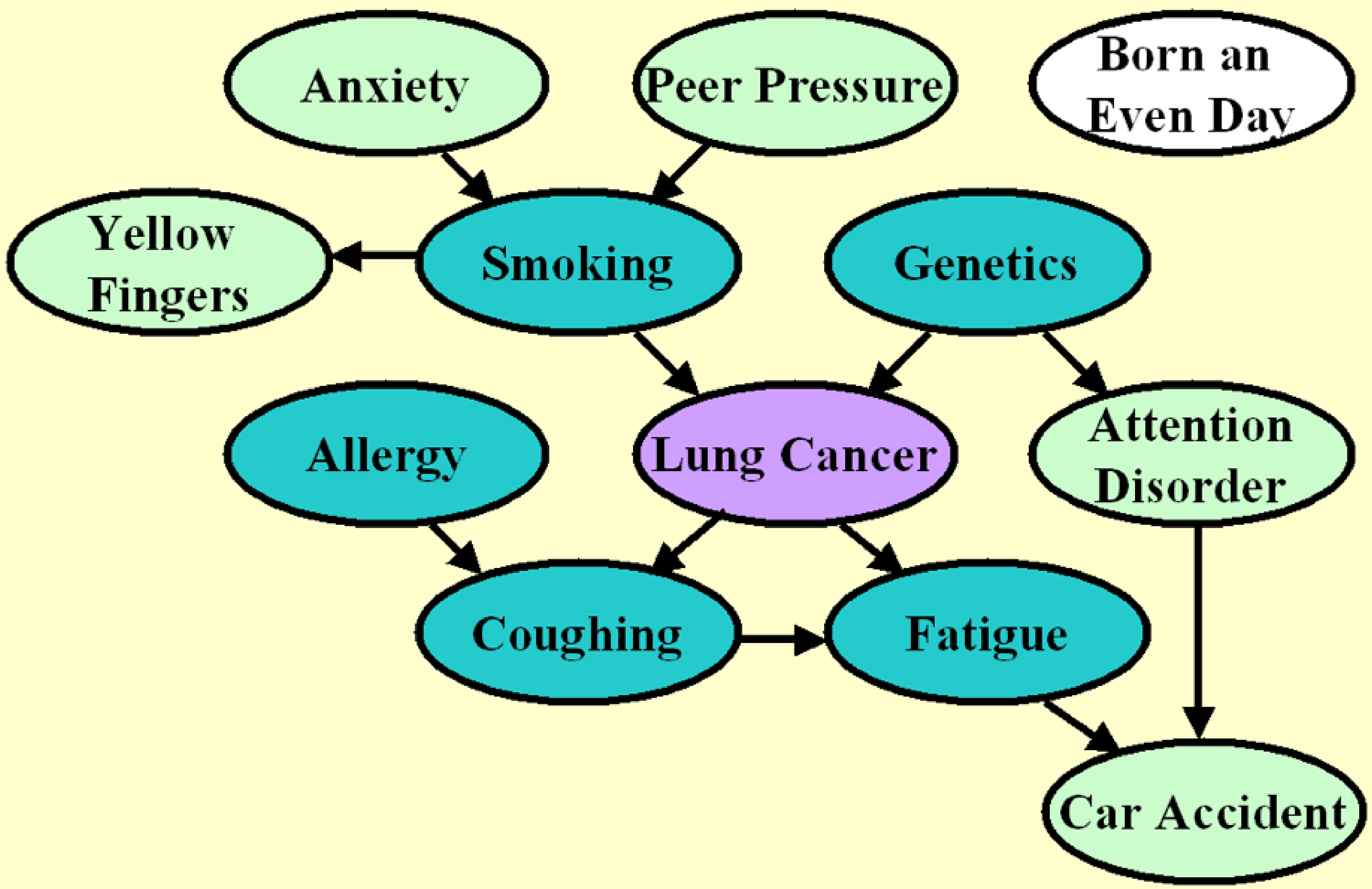




Graphical Models

- Graphical models
- d-separation
- Connection between conditional independence in graphs and that in data?
- Causal interpretations?

Intuitive Way of Representing and Visualizing Relationships



Graphical Models

- A **graph** comprises nodes (also called vertices) connected by links (also known as edges or arcs)
- Probabilistic graphical models: **graph-based representation** as the basis for compactly encoding a complex distribution
  - Node: a random variable (or group of random variables)
  - Links: direct probabilistic interactions between them
- Categorization: Undirected graphs vs. **directed acyclic** graphs (DAGs)

Probability theory + graph theory

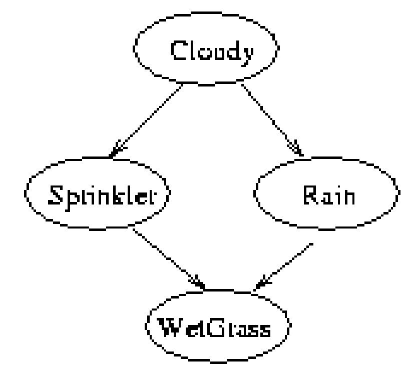
represent uncertainty & interface models to data

intuitively appealing interface for humans

Directed Acyclic Graphical Models

- Also known as Bayesian networks or belief nets
- Two components
  - Graph structure (qualitative specification)
    - prior knowledge of causal/modular relationships, or expert knowledge
    - learned from data
  - Conditional probability distributions (CPDs)
    - discrete variables : conditional distribution tables (CPTs)
    - continuous variables: SEMs

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

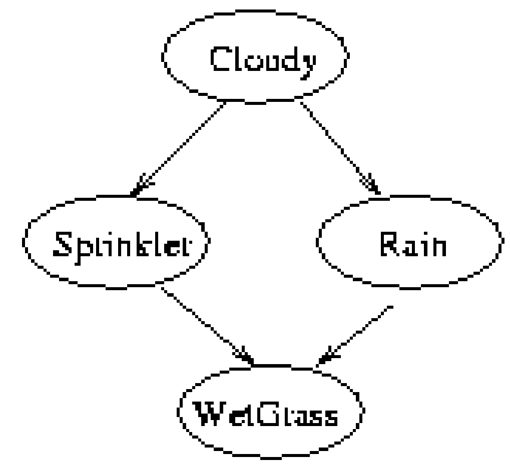
S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Terms:
nodes, edge, adjacent, path;
parents, children, spouses, ancestors, descendants, Markov blanket

Tasks Related to Bayesian Networks

- **Probabilistic inference**: Calculate P(variables of interest | observed variables)
  - Most common task where we want to use Bayesian networks
  - How to find $P(S=1|W=1)$? $P(R=1|W=1)$?
- **Parameter learning**
- **Structure learning**: Learning the structure of the graphical model from observations

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

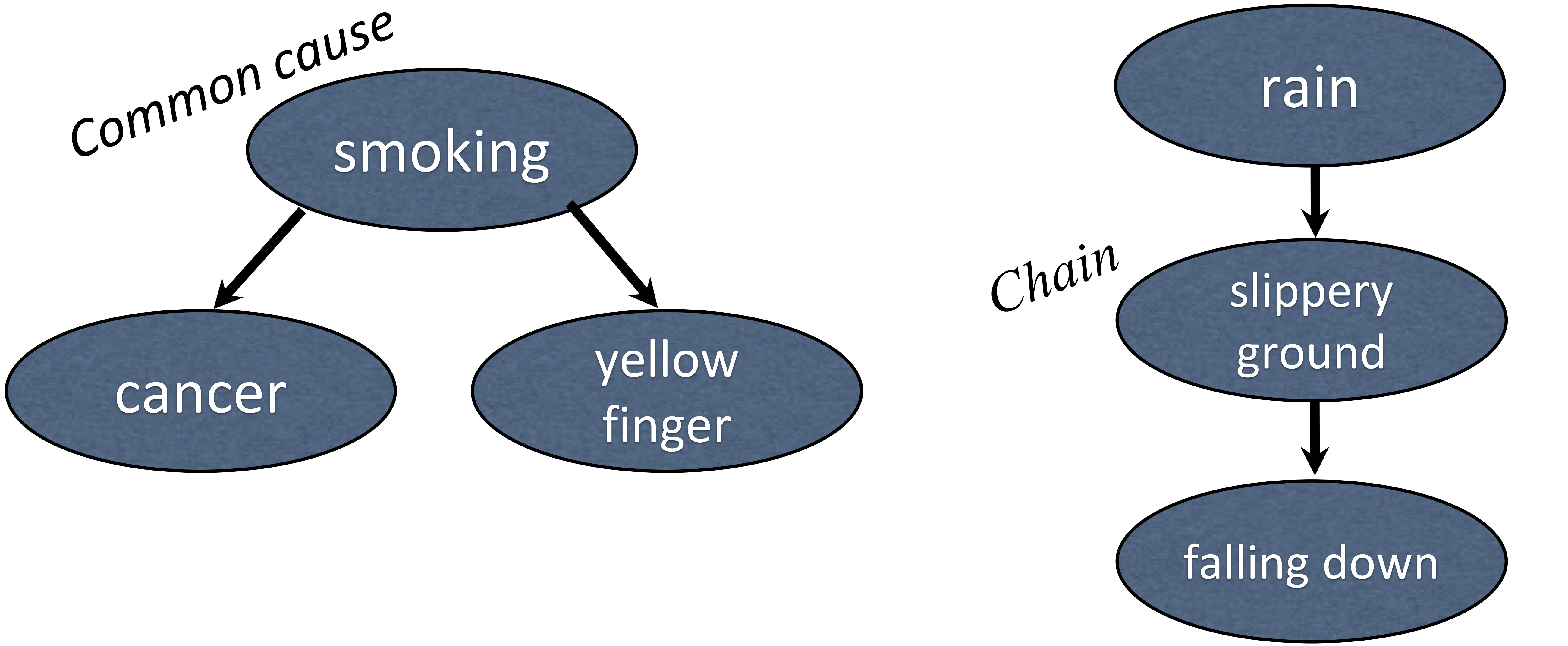
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

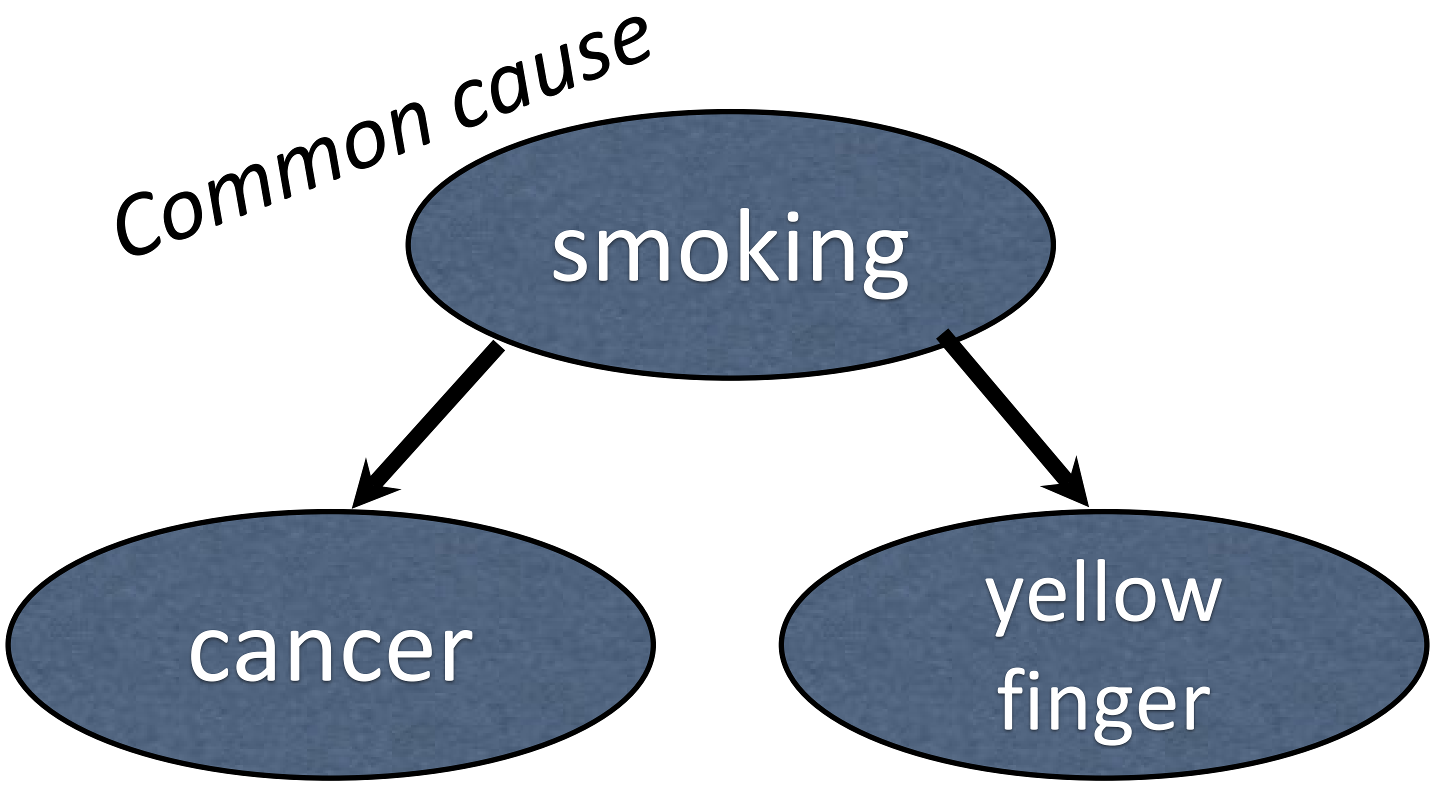
Bayesian Networks: Story

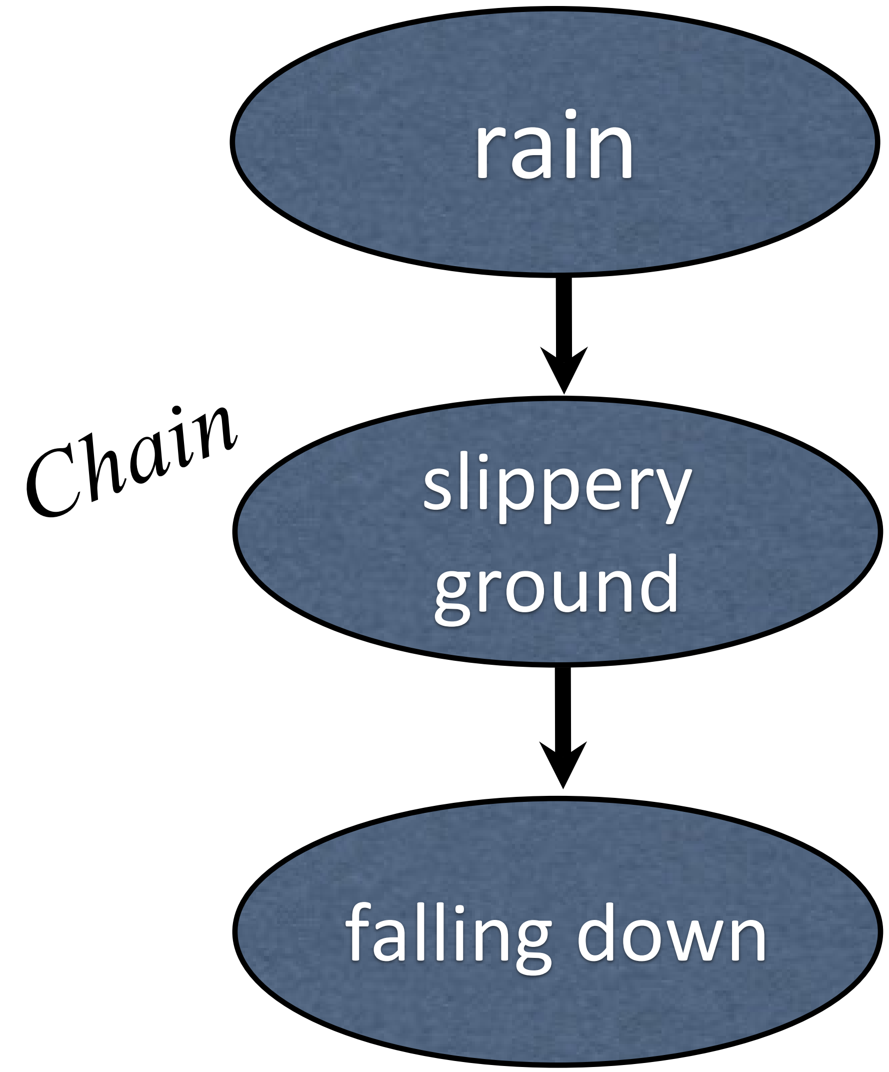
- Breakthrough in early 1980s (by Pearl et al.)
- In a joint probability distribution, every variable is, in general, related to all other variables.
- Pearl and others realized:
  - It is often reasonable to make the assumption that each variable is directly related to only a few other variables
  - This leads to **modularity**: Allowing decomposing a complex model into small manageable pieces
  - Giving rise to **Bayesian networks**

What Independence Relationships Can You See?



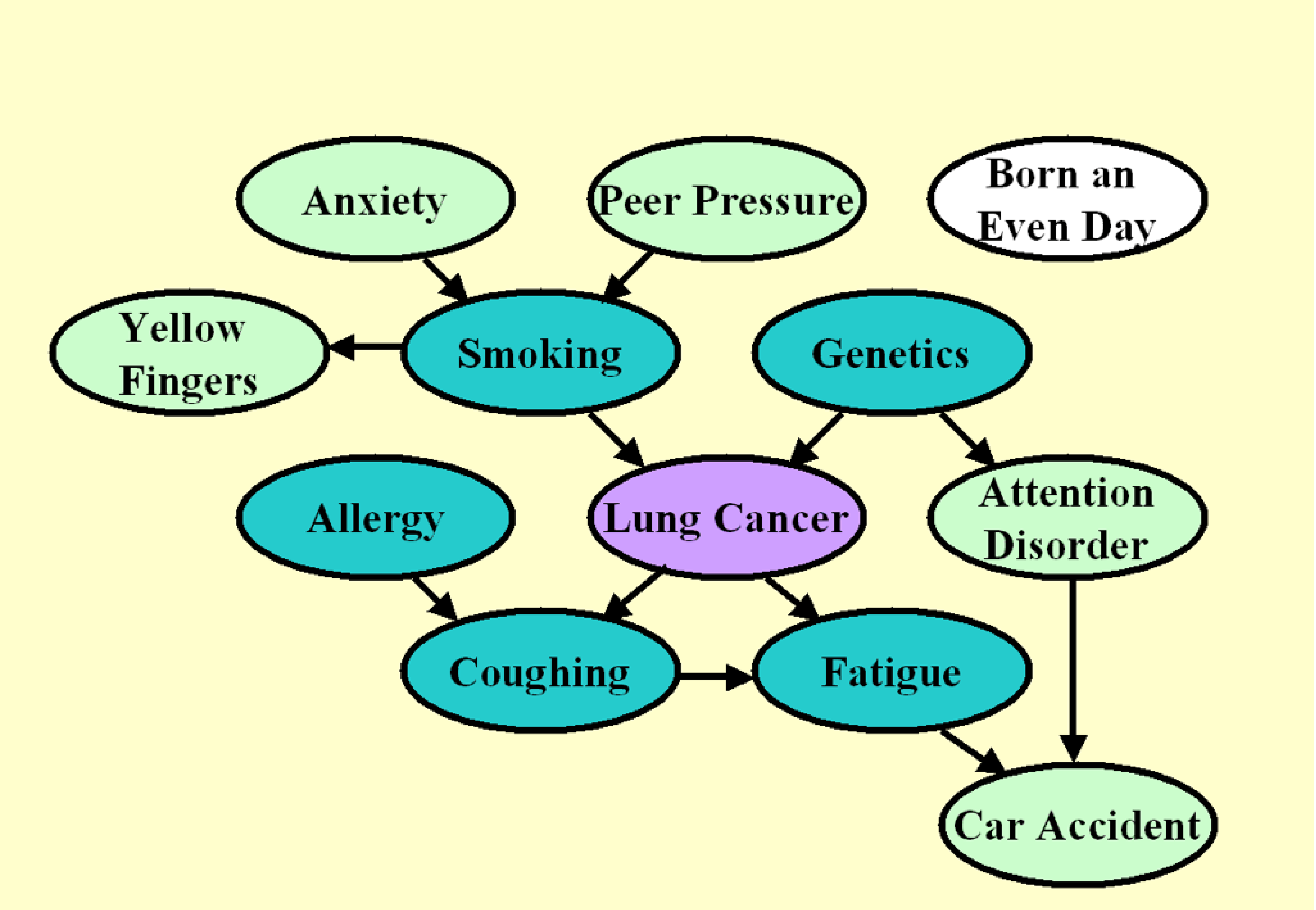
(Local) Markov Condition





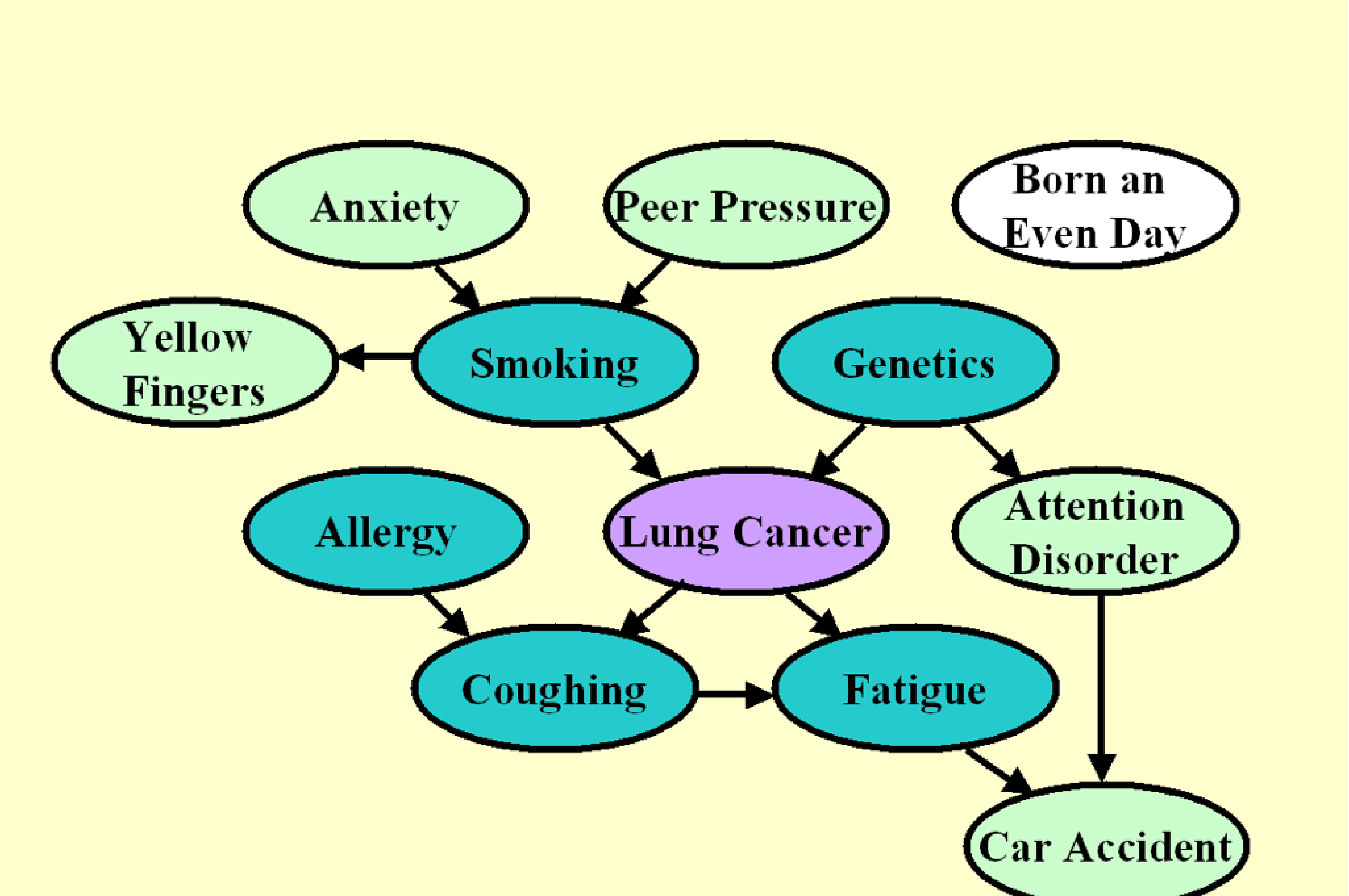
- Each variable is independent from its non-descendants given its parents

For Instance, What Independence Relations can You See?



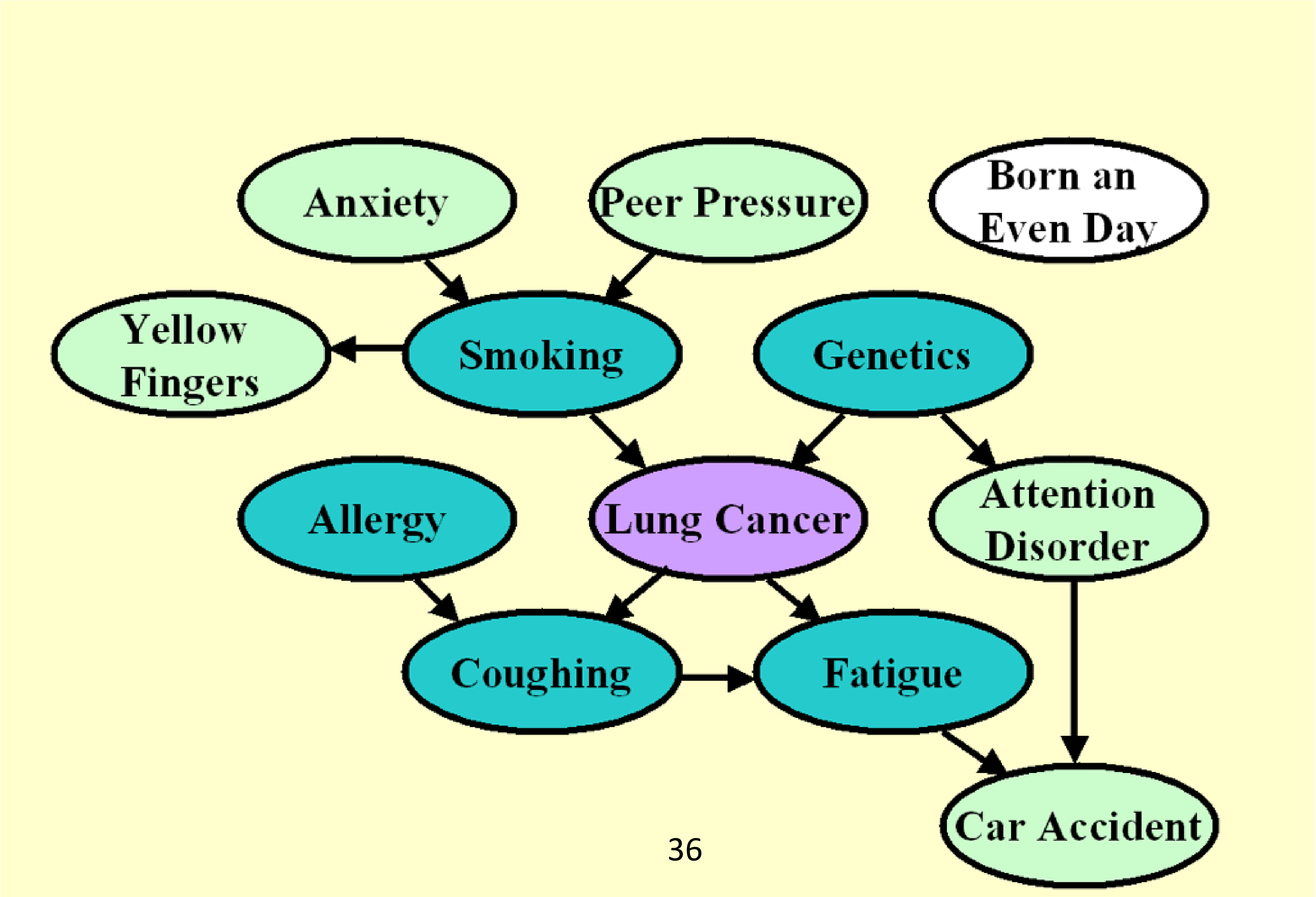
Is Local Markov Condition Enough?

- Can we see whether two arbitrary variables, X and Y, are conditionally independent given an arbitrary set of variables, **Z** ?



D-Separation Tells Conditional Independence

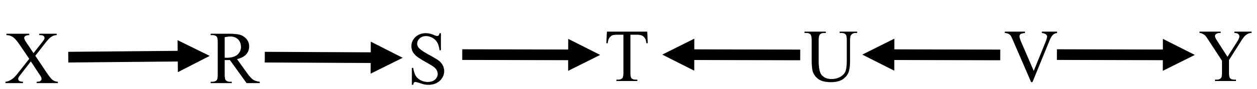
- If every path from a node in **X** to a node in **Y** is d-separated by **Z**, then **X** and **Y** are always conditionally independent given **Z**
- d: directional... You will see why



D-Separation

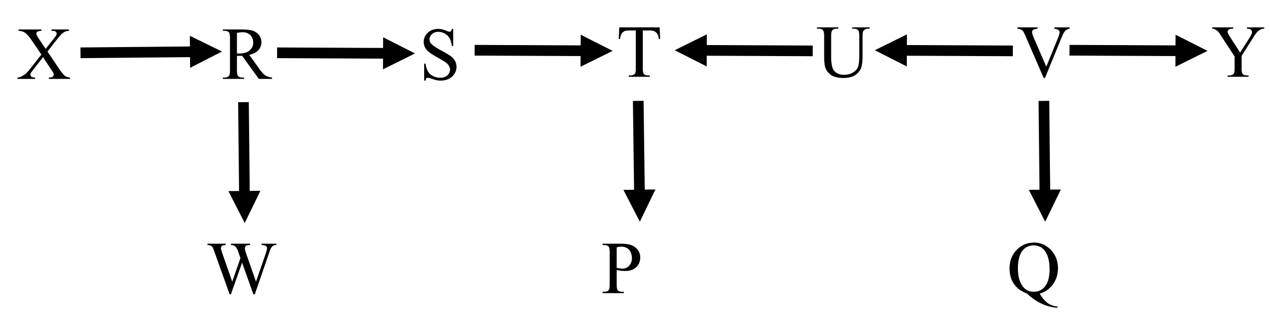
- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

- A path *p* is blocked by a set of nodes **Z** if
  - *p* contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node *m* is in **Z**, **or**
  - *p* contains a collider $i \rightarrow m \leftarrow j$ such that the middle node *m* is in not **Z** and no descendant of *m* is in **Z**



X and Y d-separated by {R, V}?

S and U d-separated by {R, V}?

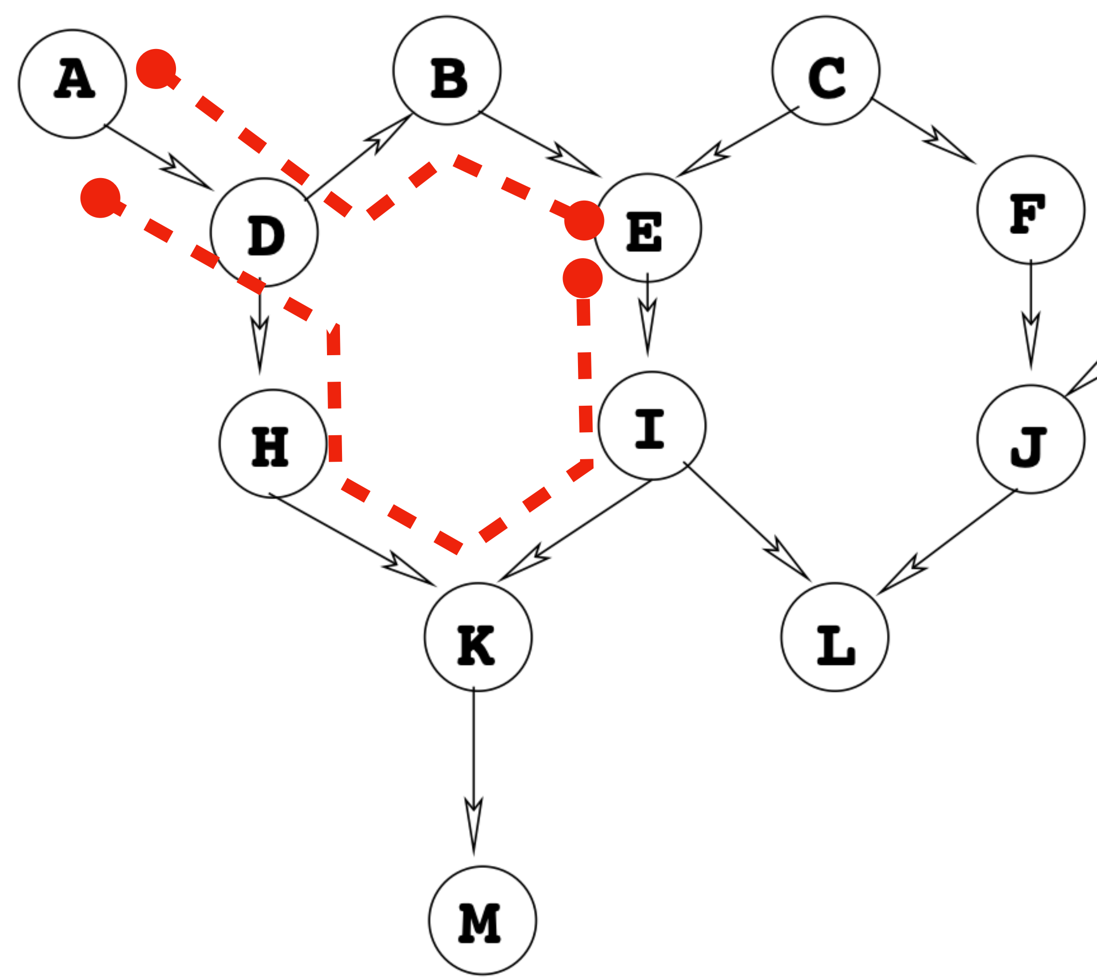


X and Y d-separated by {R, P}?

D-Separation

- A set of nodes **Z** d-separates two sets of nodes **X** and **Y** if every **path** from a node in **X** to a node in **Y** is blocked given **Z**.

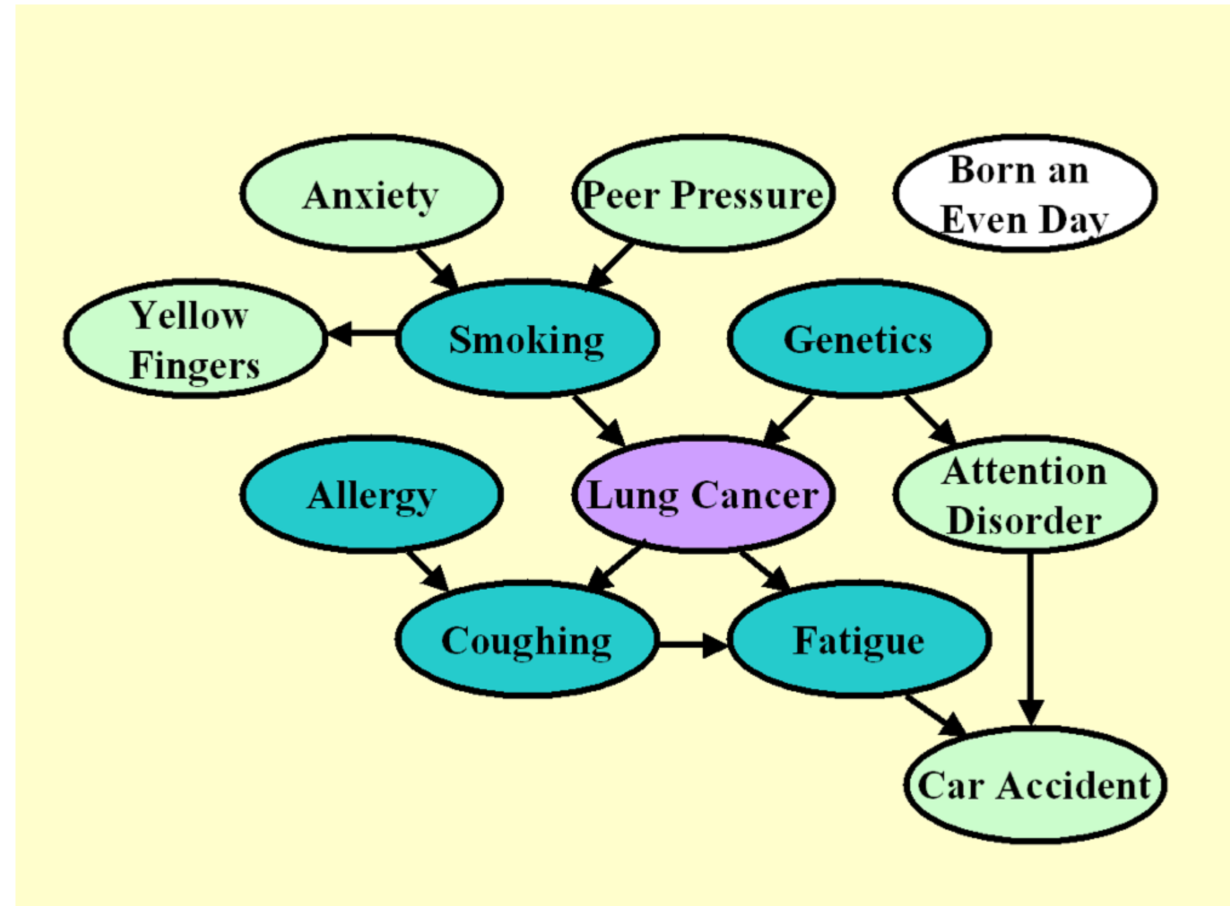
- A path *p* is blocked by a set of nodes **Z** if
  - *p* contains a chain $i \rightarrow m \rightarrow j$ or a common cause $i \leftarrow m \rightarrow j$ such that the middle node *m* is in **Z, or**
  - *p* contains a collider $i \rightarrow m \leftarrow j$ such that the middle node *m* is not in **Z** and no descendant of *m* is in **Z**

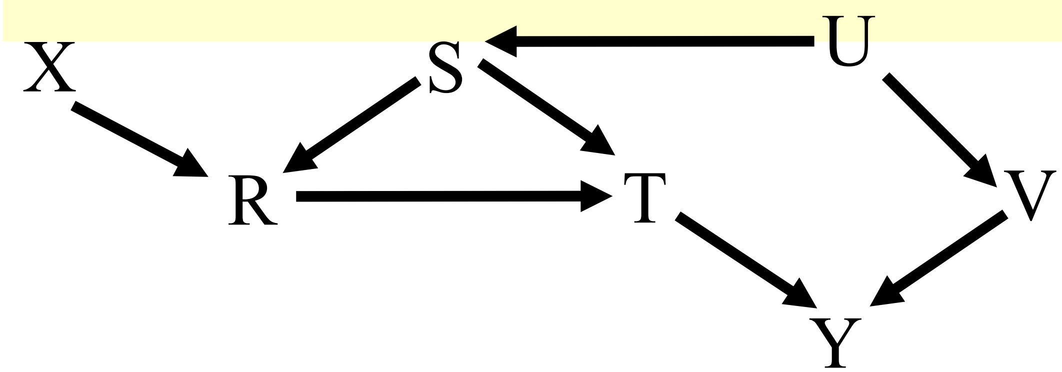


A and E d-separated by B ?
A and E d-separated by {B, M} ?

D-Separation: Intuition

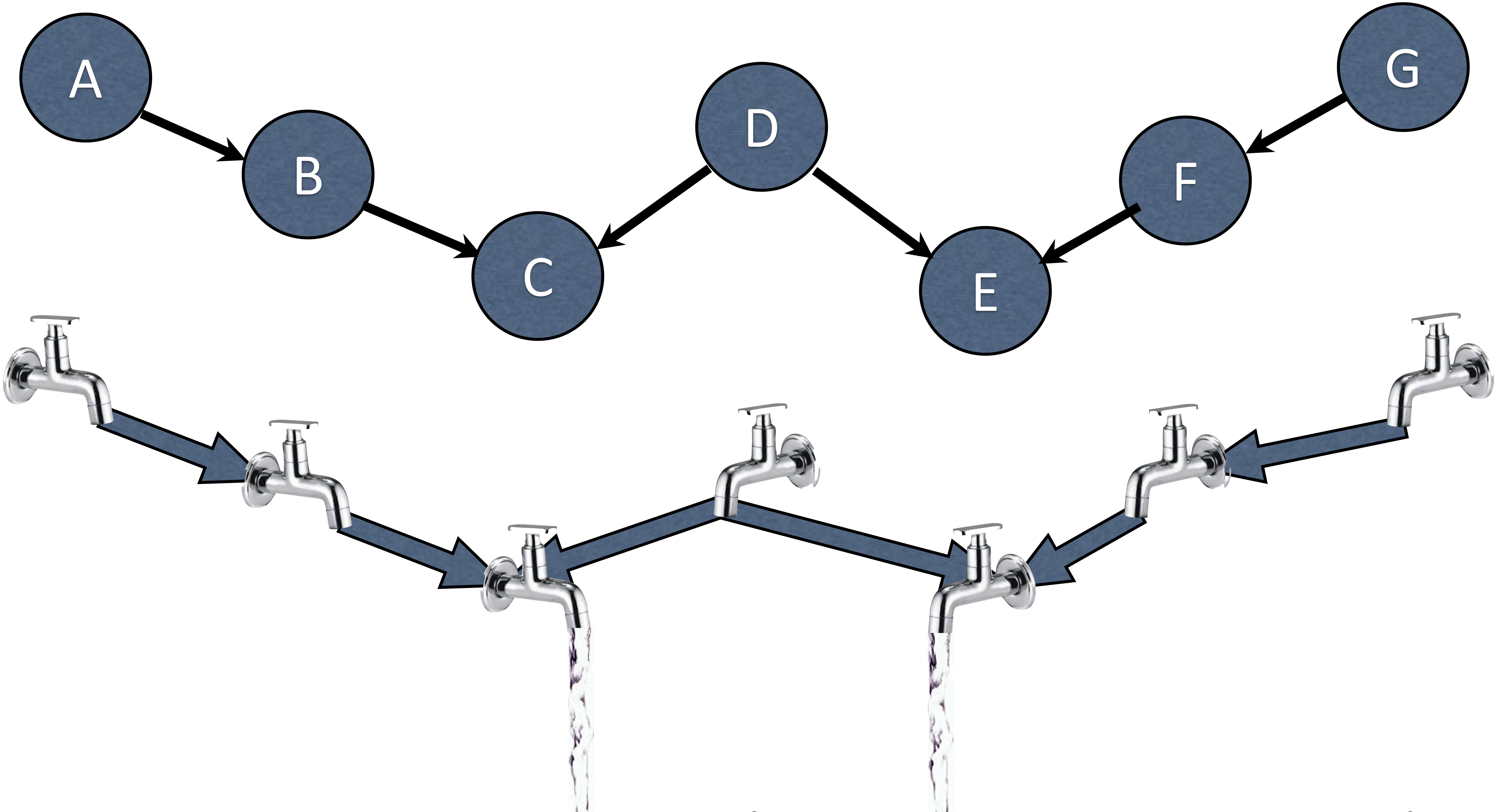
- Suppose **X** and **Y** are d-separated by **Z**
- Then if you fix **Z**, **X** and **Y**
  - do not cause each other and
  - do not share a common cause
- **X** and **Y** are independent (conditional on **Z**)!





1. X and Y d-separated by {R}?
2. X and Y d-separated by {R, T}?
3. X and Y d-separated by {T, V}?
4. X and V d-separated by Ø ?

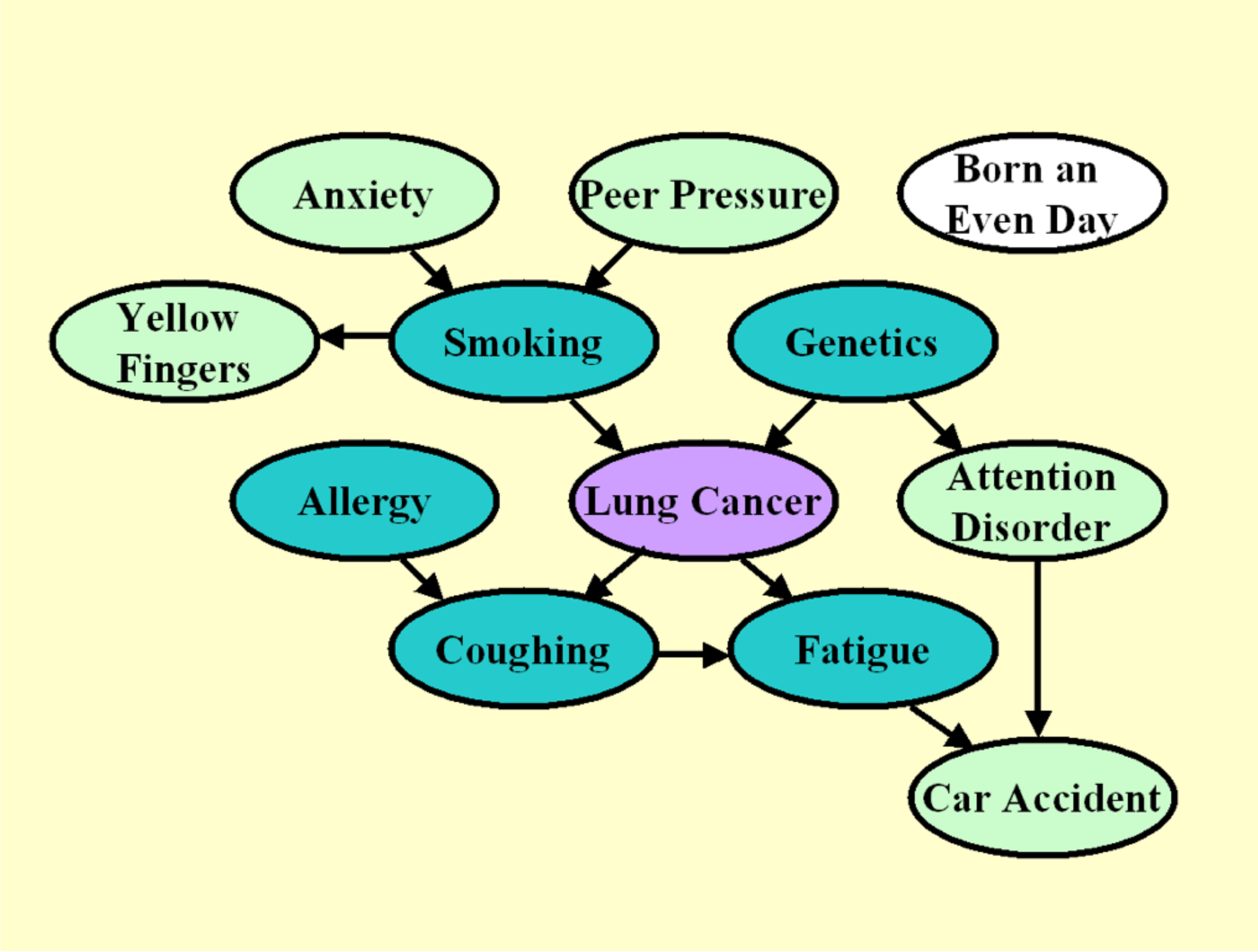
D-Separation: Intuition (2)



Given Z... conditioning on Z (given the same value of Z)

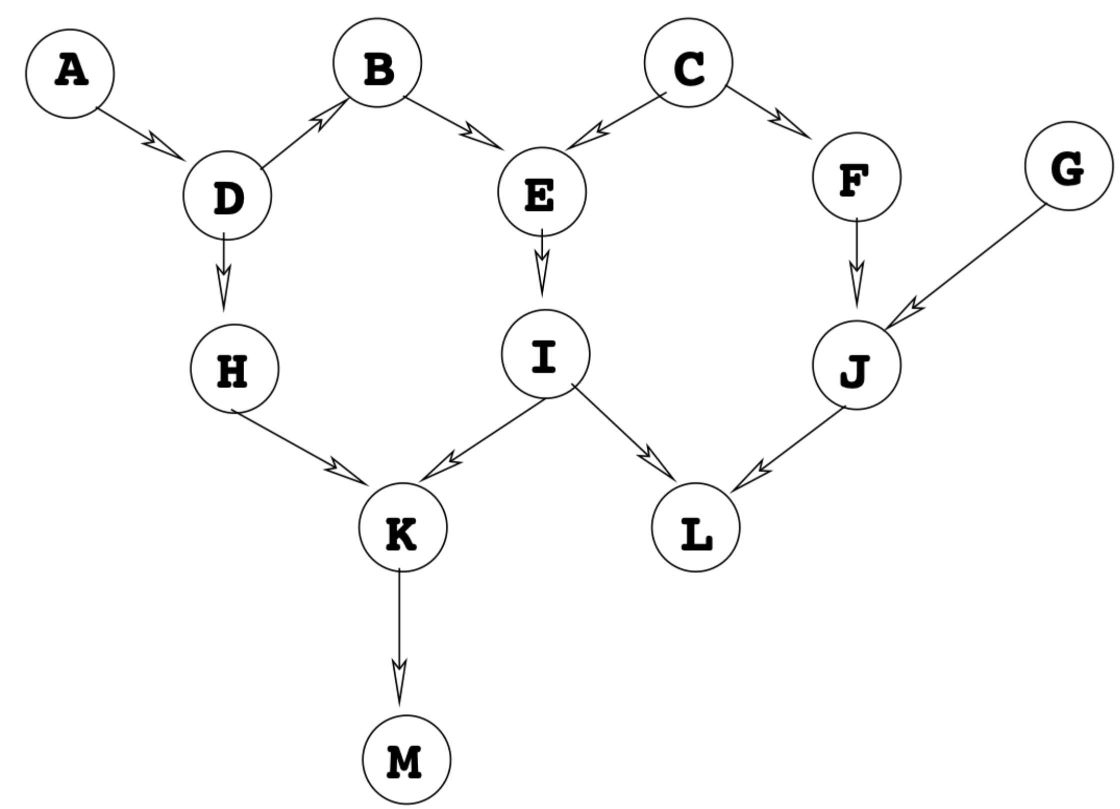
Local & Global Markov Conditions

- **Local** Markov condition:
  - In a DAG, a variable X is independent of all its non-descendants given its parents
- **Global** Markov condition:
  - Given a DAG, let X and Y be two variables and **Z** be a set of variables that does not contain X or Y. If **Z** d-separates X and Y, then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
- Actually equivalent on DAGs!



Markov Blanket

- In a DAG, the Markov Blanket of a node *X* is the set consisting of
  - Parents of *X*
  - Children of *X*
  - Parents of children (i.e., spouses) of *X*
- In a DAG, a variable *X* is conditionally independent from all other variables given its Markov Blanket
  - Implied by d-separation...
- The Markov blanket of *I* ?

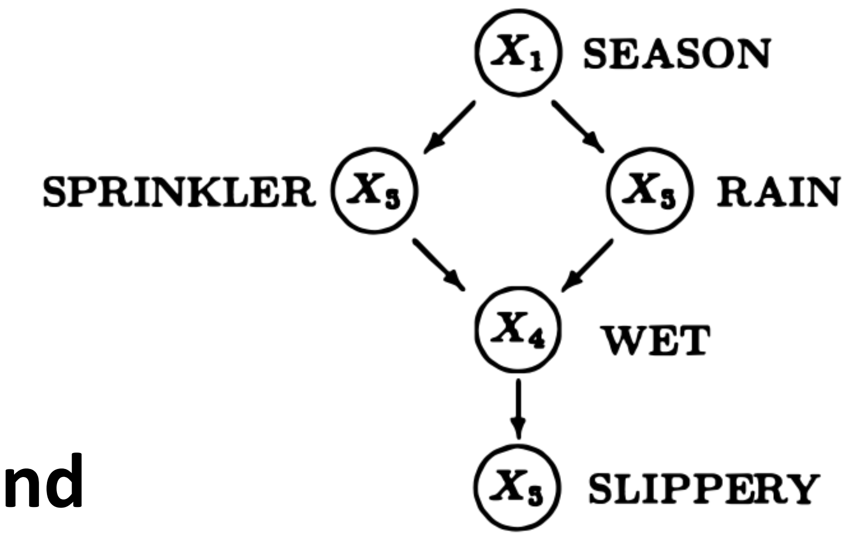


We learn DAGs. Are They Always Causal?

- Causality is not only conditional independence.
- How can we be sure the DAG is causal

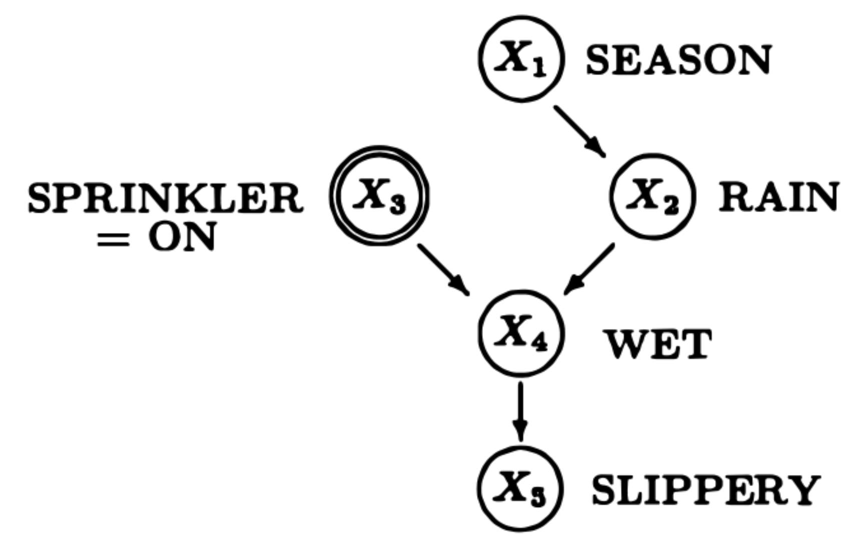
Causal DAGs

- Bayesian networks: DAGs
- Causal DAGs
  - More meaningful & able to **represent and respond to external or spontaneous changes**



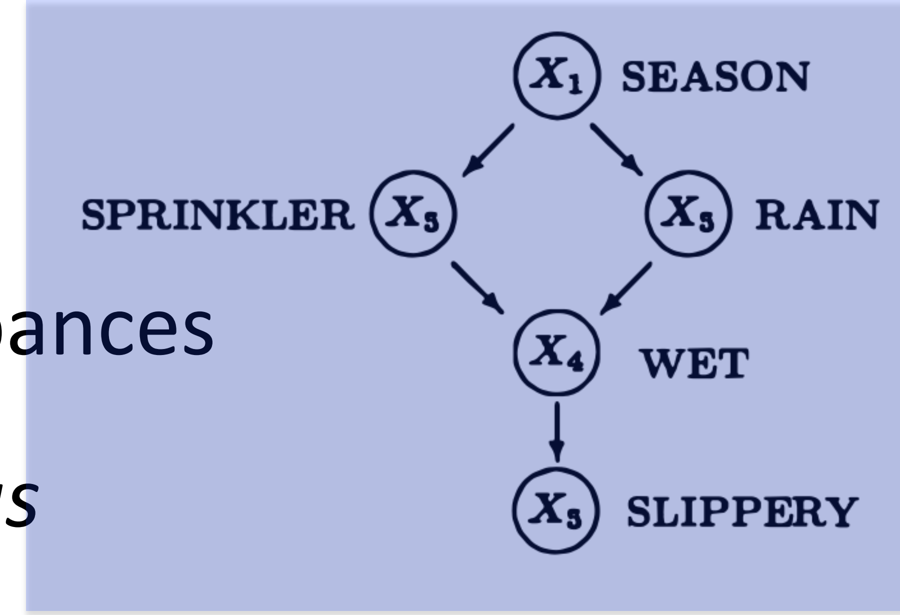
Let $P_x(V)$ be the distribution of V resulting from intervention $do(X=x)$. A DAG G is a CBN if

1. $P_x(V)$ is Markov relative to G;
2. $P_x(V_i=v_i)=1$ for all $V_i \in X$ and v_i consistent with $X=x$;
3. $P_x(V_i \mid PA_i) = P(V_i \mid PA_i)$ for all $V_i \notin X$, i.e., $P(V_i \mid PA_i)$ remains invariant to interventions not involving V_i.



What is $P_{X_3=ON}(X_1,X_2,X_4,X_5)$?

Structural Causal Models

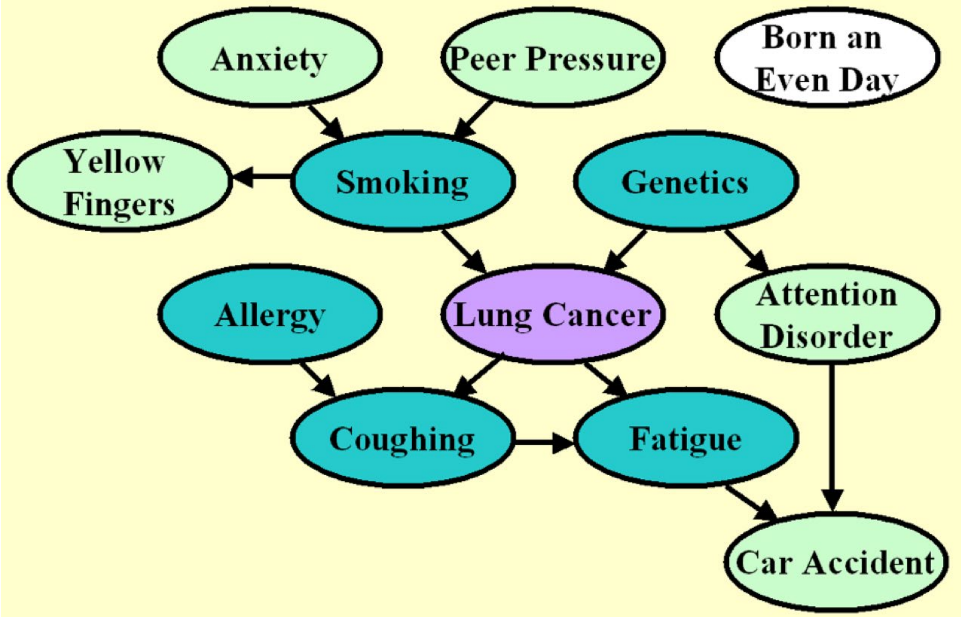


- $X_i = f_i\,(PA_i,\ E_i),\ i=1,\ldots,n$
  - E_i: exogenous variables / errors / disturbances
  - Each equation represents an *autonomous* mechanism
  - Describes how nature assigns values to variables of interest
- Distinction between structural equations & algebraic equations
- Associated with graphical causal models

$$PA_i \longrightarrow X_i$$

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

We can See CI Relations from DAGs...



- Local Markov condition
- Global Markov condition
  - d-separation implies conditional independence:

$P(\mathbf{V})$, where $\mathbf{V}$ denotes the set of variables, obeys the global Markov condition (or property) according to DAG $\mathcal{G}$ if for any disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, we have

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z}.$$

CI from Data...

- We are able to see CI relationships from DAGs.
- How can we see that from data?
  - Useful to find information of the underlying DAG, especially under the faithfulness assumption

Independence in Linear-Gaussian Case

- If X and Y are jointly normally distributed, their independence $\Leftrightarrow$ their zero correlation

- Zero correlation can be tested with, say, Fisher's z test

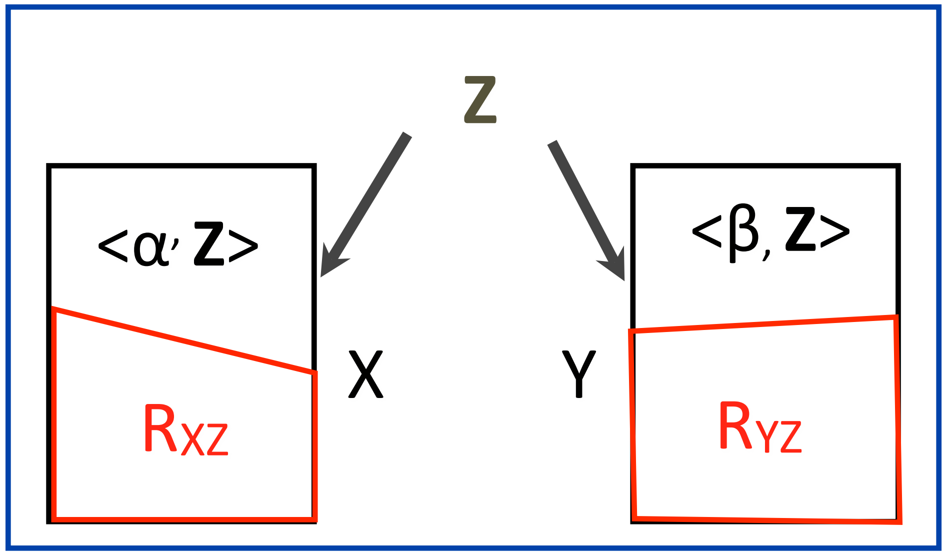
  - Calculate sample correlation coefficient (statistic):

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}.$$

  - Under H_0 (zero correlation), $z \triangleq \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ follows $\mathcal{N}(0, \frac{1}{N-3})$
  - Given the statistic and its null distribution, we can find p value

Conditional Independence in Linear-Gaussian case: Partial Correlation

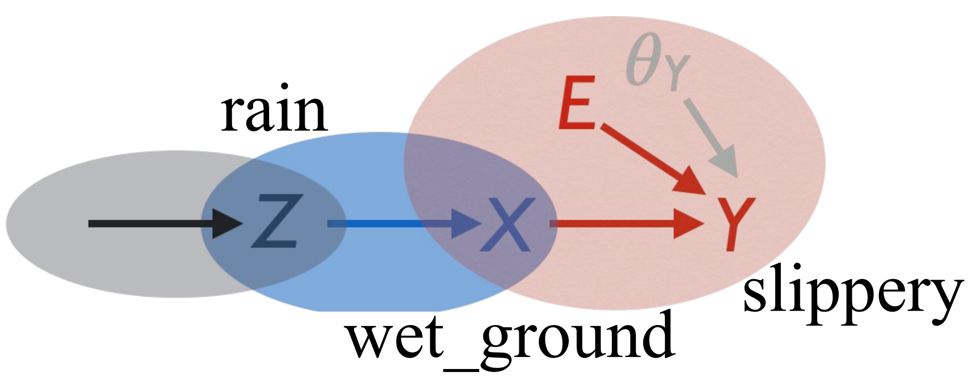
- Partial correlation: "Relationship" between X and Y while eliminating influence of **Z**
  - Regress X and Y on **Z**, respectively
  - Partial correlation $\rho_{XY \cdot \mathbf{Z}}$ is the correlation between residuals R_{XZ} and R_{YZ}
- If X, Y, and **Z** are jointly normally distribution, $X \perp\!\!\!\perp Y \mid \mathbf{Z} \Leftrightarrow \rho_{XY \cdot \mathbf{Z}} = 0$
- We can then test for zero partial correlation ('partialcorr' in MATLAB)



What Information Helps Find Causality?

- Connection between causal structure and statistical data under *suitable assumptions*
- Note this “irrelevance”:

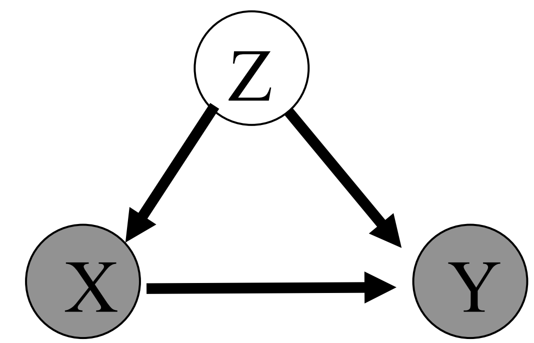
If there is no common cause of *X* and *Y*, the generating process for cause *X* is irrelevant to (“independent” from) that generates effect *Y* from *X*



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

Causal Sufficiency

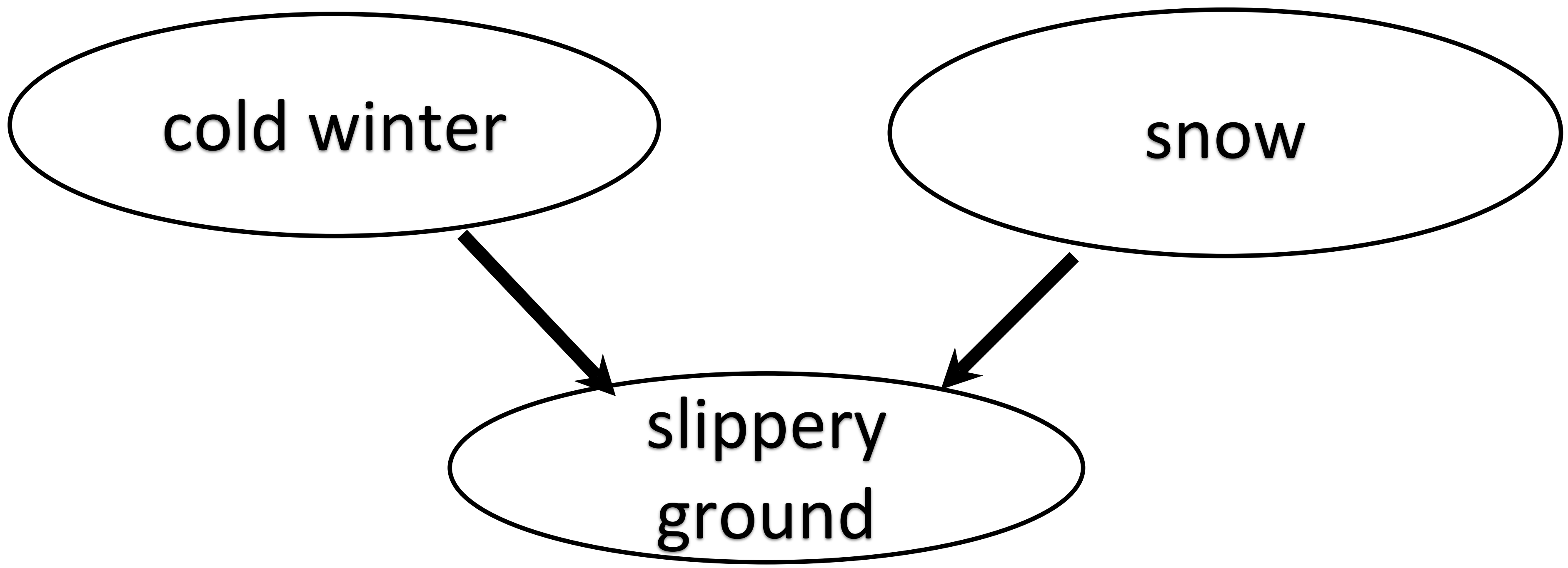
- A set of random variables ***V*** is causally sufficient if ***V*** contains every common cause (with respect to ***V***) of any pair of variables in ***V***



- ***V*** $= \{X, Y, Z\}$: causally sufficient
- ***V*** $= \{X, Y\}$: causally insufficient
- Methods exist in causally **insufficient** cases, e.g., FCI (*Chapter 6 of the SGS book*)

SGS Book, Chapter 5 (for causally sufficient structures); Chapter 6 (without causal sufficiency)

V-Structures



Why so interesting?

Going from CI to Graph?

$$\mathbf{X} \text{ and } \mathbf{Y} \text{ are d-separated by } \mathbf{Z} \text{ in } \mathcal{G} \implies \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}.$$

- Contrapositive:
  - Conditional dependence implies d-connection
  - What if variables are conditionally independent?
- Can we recover the property of the underlying graph from CI relations with Markov condition?
  - Arbitrary $P(\mathbf{V})$ would satisfy the global Markov condition according to G^f *in which there is an edge between each pair of variables*: trivial !
  - Under what assumptions can we have CI ⇒ d-separation?

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)

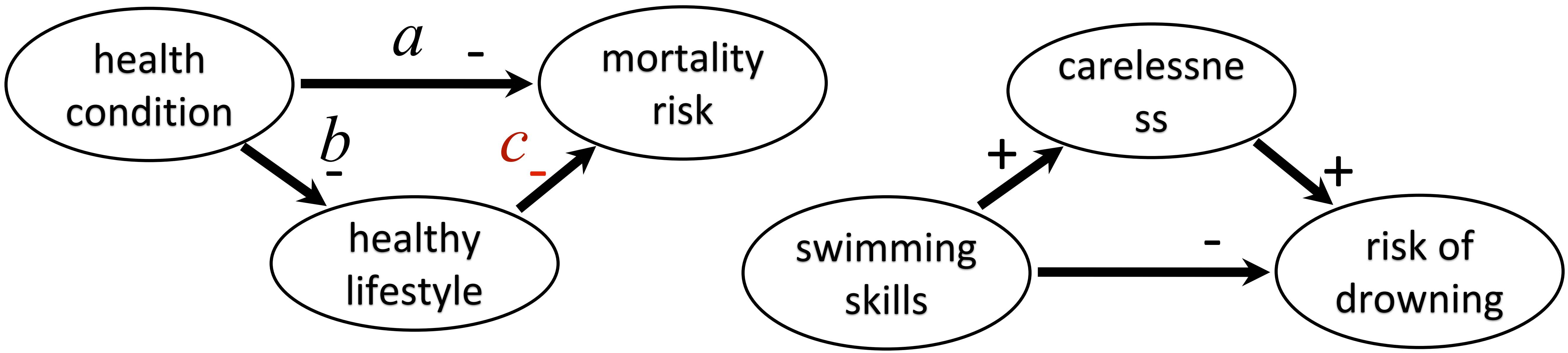
$Y \perp\!\!\!\perp Z \mid X$

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

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

Faithfulness Assumption

- One may find independence between health condition & risk of mortality and between swimming skills & risk of drowning. Why?



- E.g., if they are linear-Gaussian and a=-bc, then *health_condition* ⫫ *risk_mortality*, which cannot by seen from the graph!

- Faithfulness assumption eliminates this possibility!

Constraint-based Causal Discovery

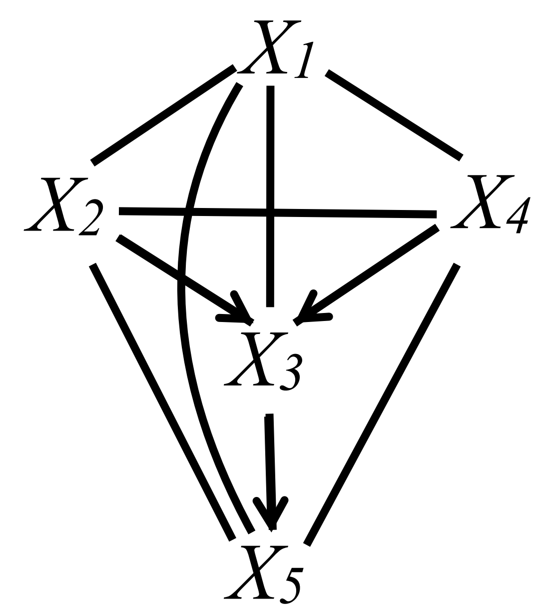
- Without confounders: PC
- With confounders: FCI

(Typical) Constraint-Based Causal Discovery

- 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
…	…	…	…	…

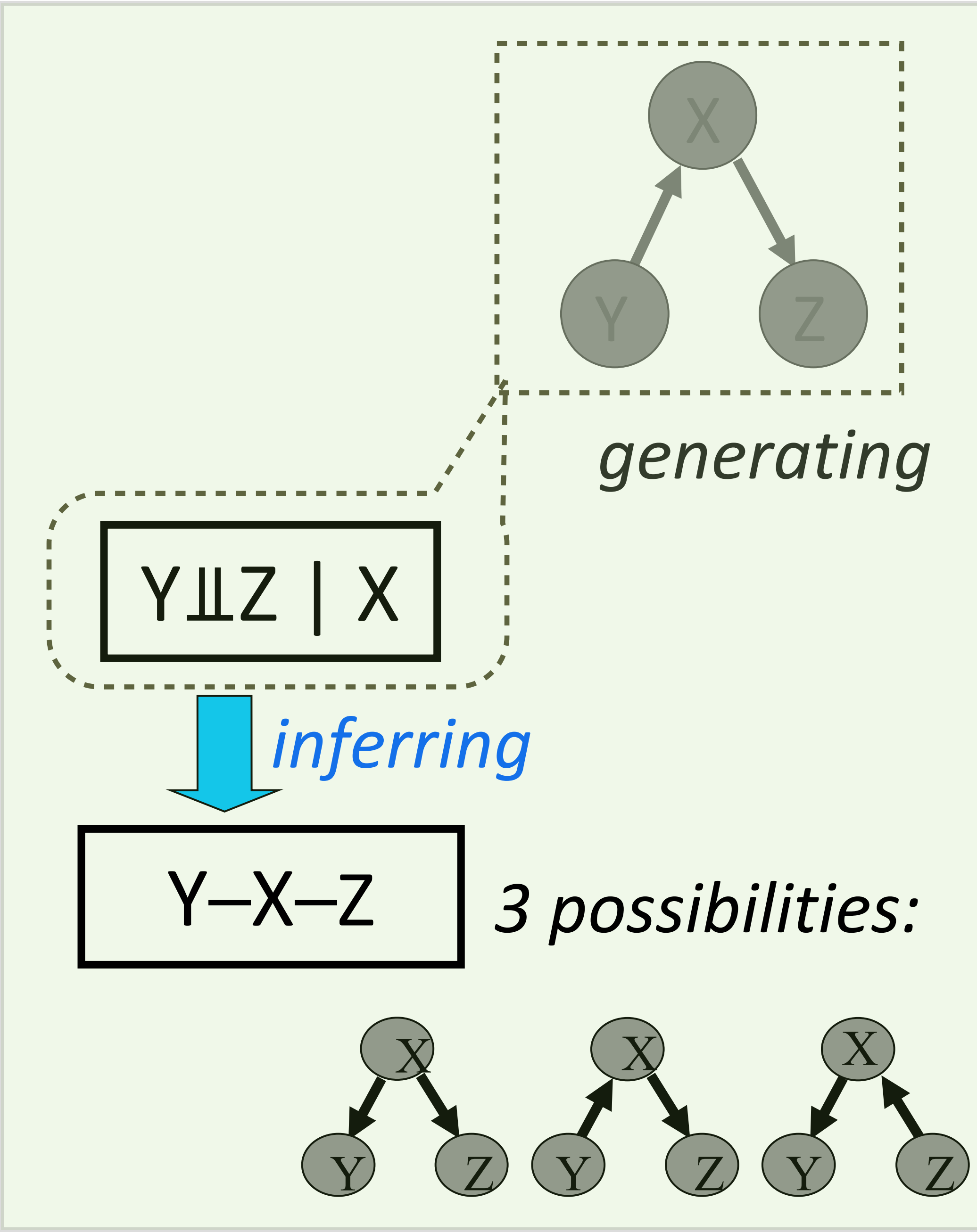
$X_1 \perp\!\!\!\perp X_5 \mid X_3$
$X_2 \perp\!\!\!\perp X_4 \mid X_1$
$X_2 \perp\!\!\!\perp X_5 \mid X_3$
$X_4 \perp\!\!\!\perp X_5 \mid X_3$
$X_1 \perp\!\!\!\perp X_3 \mid \{X_2, X_4\}$

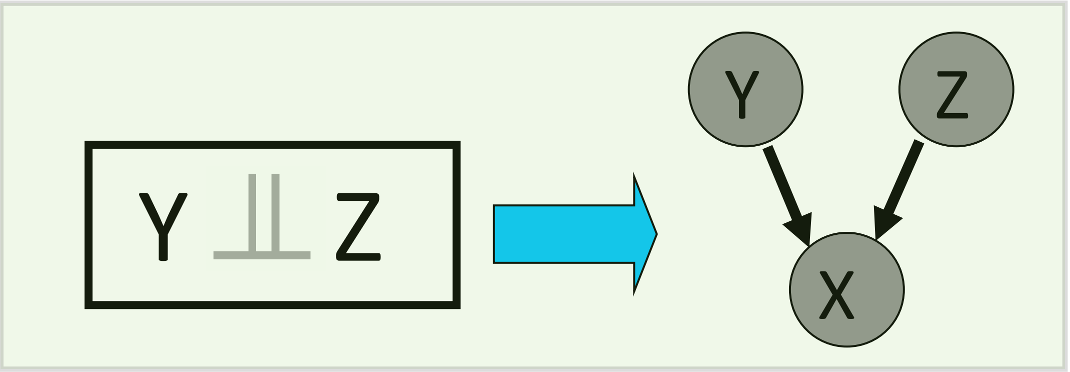


- Spirtes, Glymour, and Scheines. Causation, Prediction, and Search. 1993.

Constraint-Based Causal Discovery

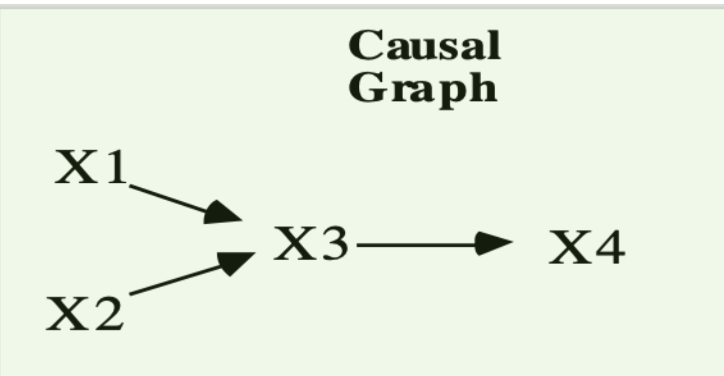
- (Conditional) independence <u>constraints</u> ⇒ candidate causal structures
  - Relies on causal Markov condition & faithfulness assumption
  - PC algorithm (Spirtes & Glymour, 1991)
  - *Step 1*: X and Y are adjacent iff they are dependent conditional on every subset of the remaining variables (SGS, 1990)
  - *Step 2*: Orientation propagation
- v-structure
- Markov equivalence class, represented by a pattern
  - same adjacencies; → if all agree on orientation; – if disagree





Example I

Step 1: finding skeleton



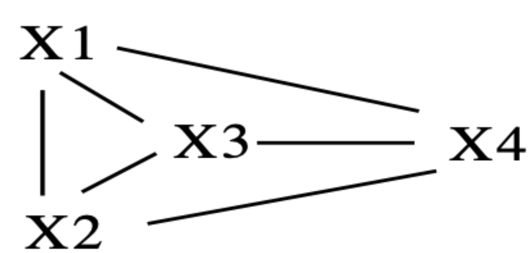
Independcies

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

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

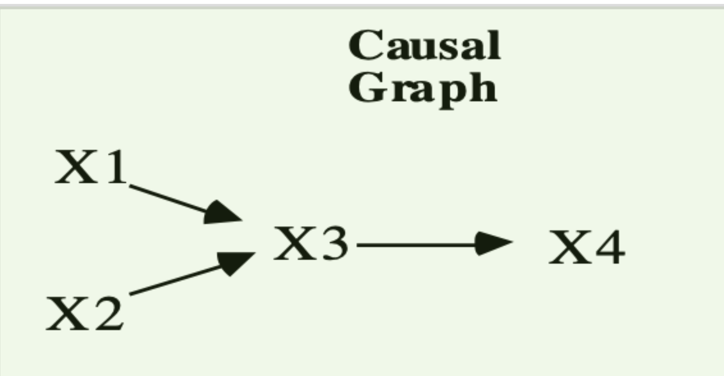
$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Begin with:



Example I

Step 1: finding skeleton



Independcies

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

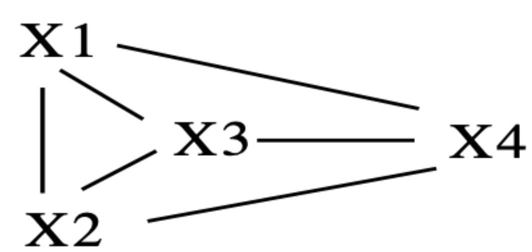
$X1 \perp\!\!\!\perp X4 | \{X3\}$

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

Step 2: finding v-structure and doing orientation propagation

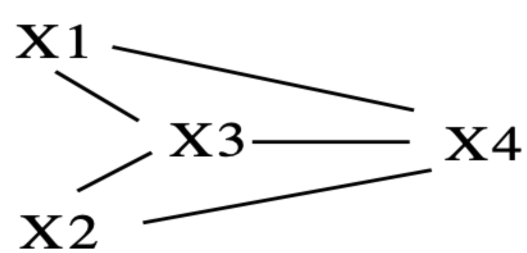
---

Begin with:



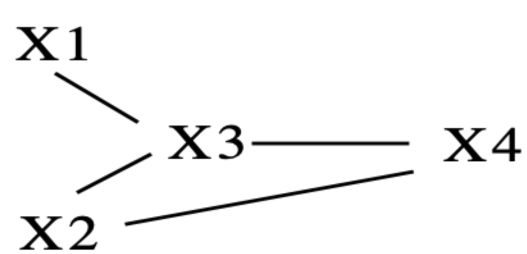
From

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



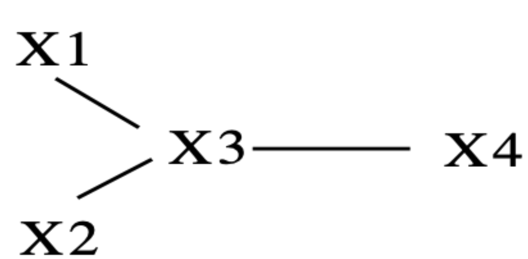
From

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



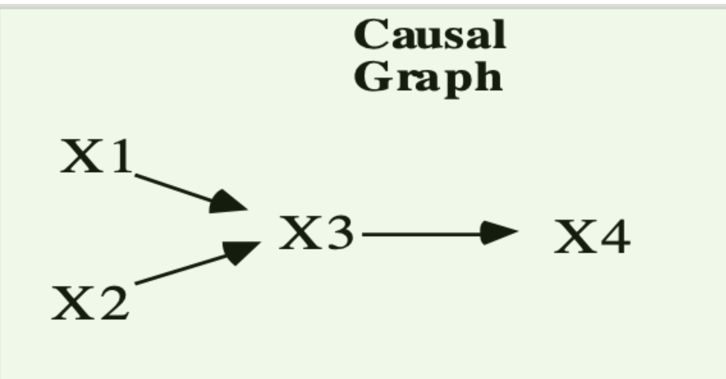
From

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



Example I

Step 1: finding skeleton



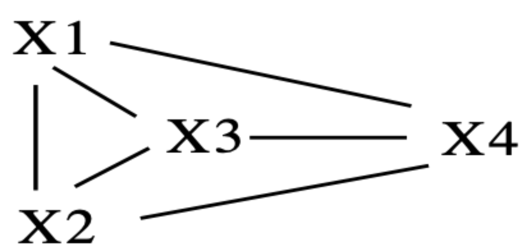
Independencies

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

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

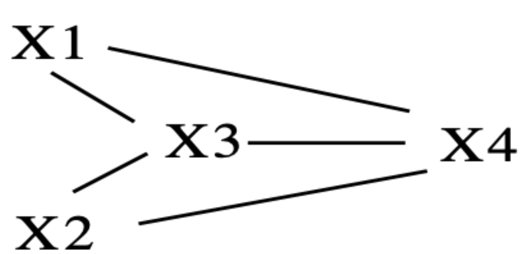
$X2 \perp\!\!\!\perp X4 \mid \{X3\}$

Begin with:



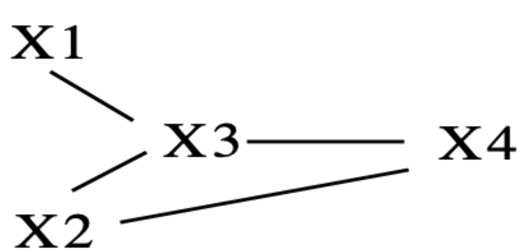
From

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



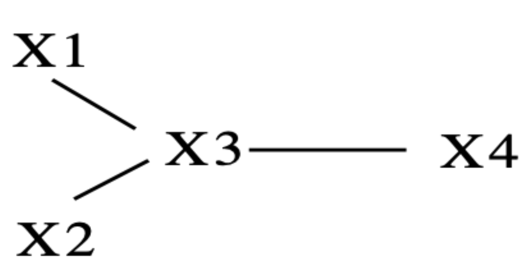
From

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



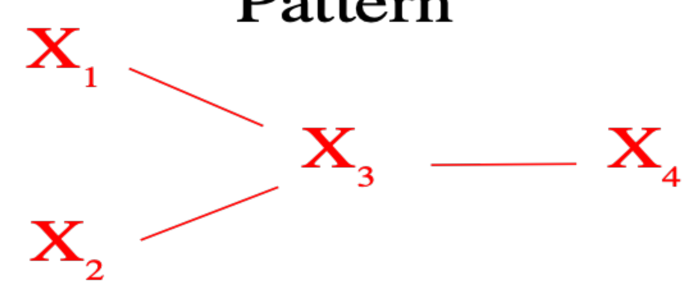
From

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

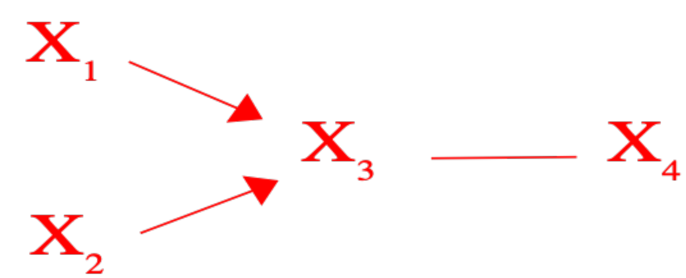


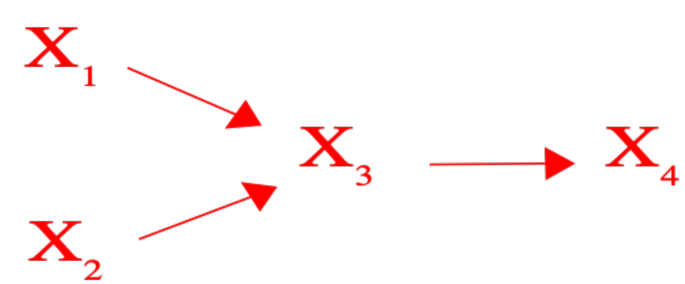
Step 2: finding v-structure and doing orientation propagation

Pattern



$X_1 \perp\!\!\!\perp X_2$:

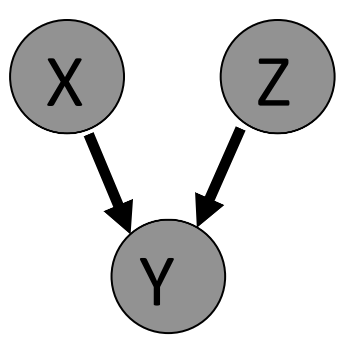




PC Algorithm

Test for (conditional) independence with an increased cardinality of the conditioning set

Finding V-structures



Orientation propagation

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ in **Sepset**(X,Y) and **Sepset**(Y,X);

until all ordered pairs of adjacent variables X and Y such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n and all subsets $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

D. repeat

If A -> B, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as B -> C.

If there is a directed path from A to B, and an edge between A and B, then orient A - B as A -> B.

until no more edges can be oriented.

PC Algorithm

Test for (conditional) independ
increased
the cond

A.) Form the complete undirected graph C on the vertex set $\mathbf{V}$.

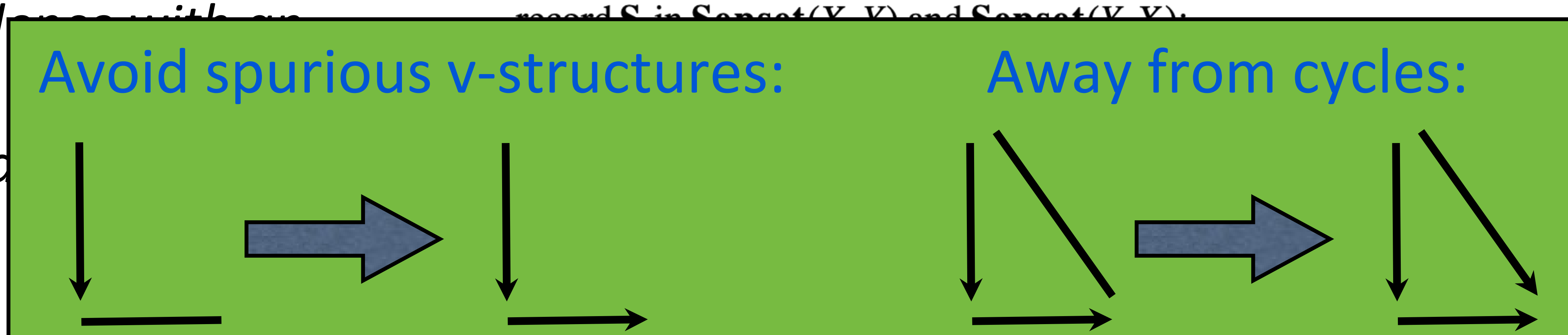
B.)

$n = 0$.

repeat

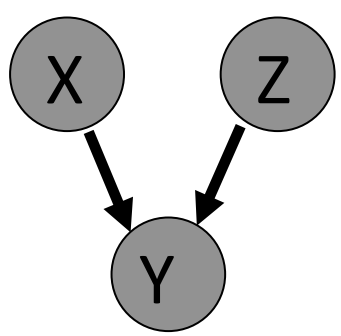
repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**$(C,X)\backslash\{Y\}$ has cardinality greater than or equal to n, and a subset $\mathbf{S}$ of **Adjacencies**$(C,X)\backslash\{Y\}$ of cardinality n, and if X and Y are d-separated given $\mathbf{S}$ delete edge X - Y from C and record $\mathbf{S}$ in **Sepset**(X,Y) and **Sepset**(Y,X);



until for each ordered pair of adjacent vertices X, Y, **Adjacencies**$(C,X)\backslash\{Y\}$ is of cardinality less than n.

Finding V-structures



C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C, orient X - Y - Z as X -> Y <- Z if and only if Y is not in **Sepset**(X,Z).

Orientation propagation

D. repeat

If A -> B, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient B - C as B -> C.

If there is a directed path from A to B, and an edge between A and B, then orient A - B as A -> B.

until no more edges can be oriented.

(Independence) Equivalent Classes: Patterns

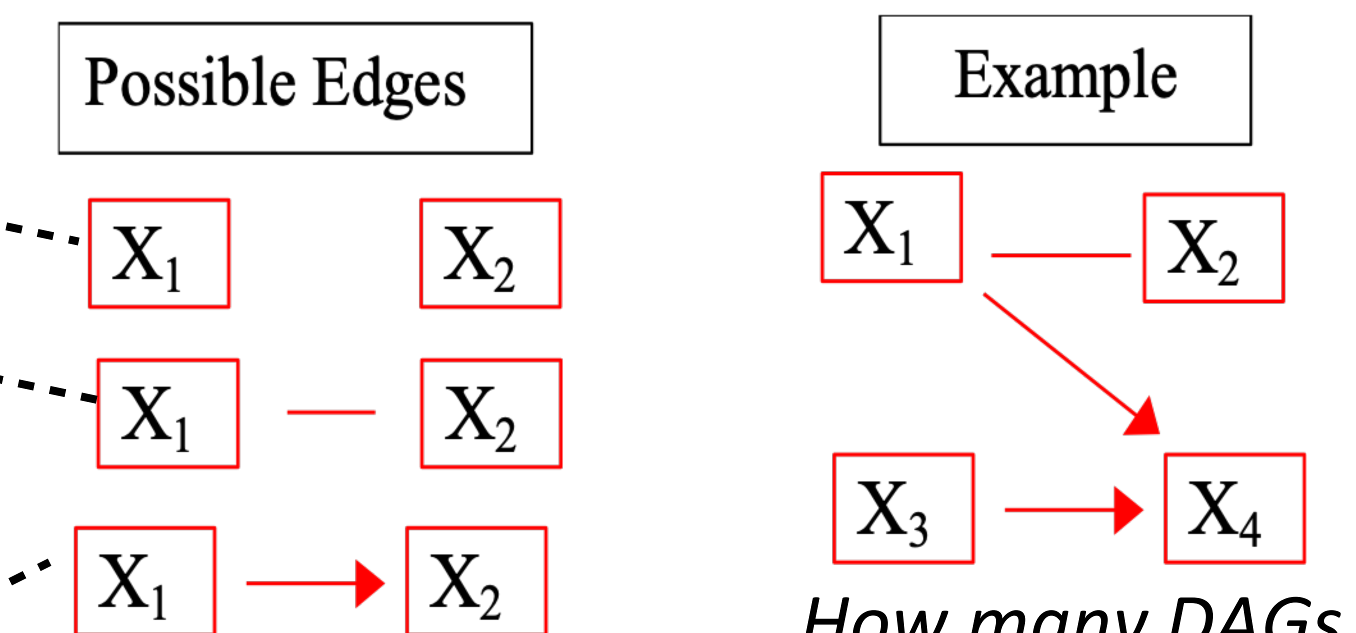
- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)

- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

X_1 and X_2 are not adjacent in any member of the equivalent class

$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others

$X_1 \rightarrow X_2$ in every member of the equivalent class

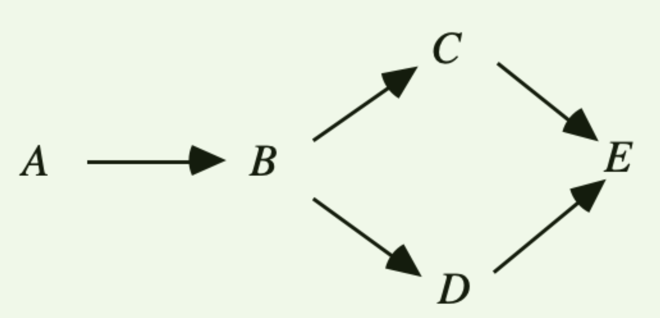


How many DAGs in this class?

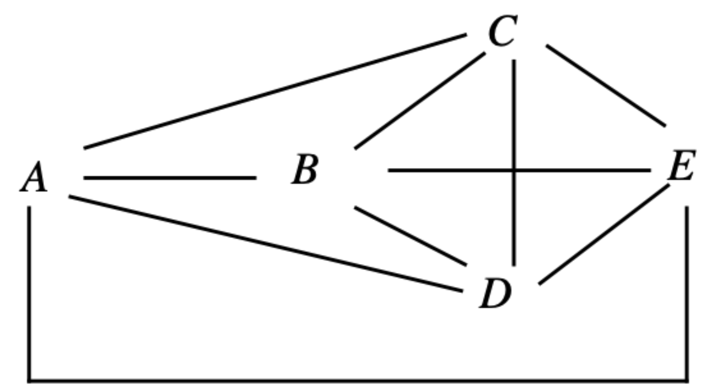
Example II (From SGS Book)

Step 1

Step 1I

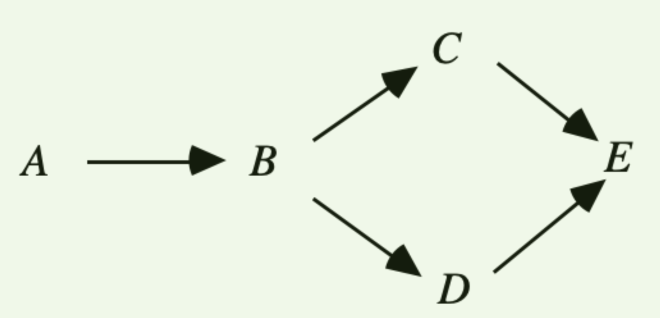


True Graph



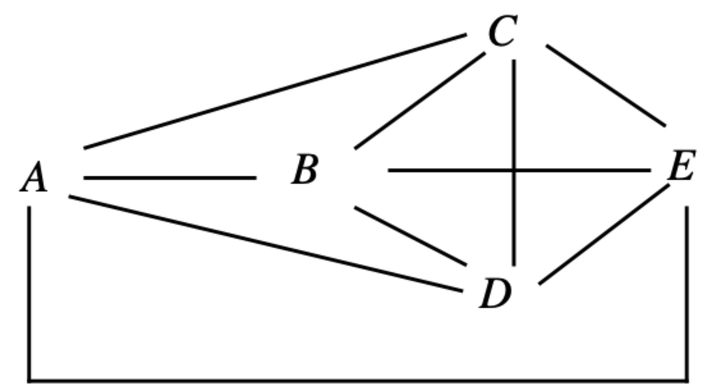
Complete Undirected Graph

Example II (From SGS Book)



True Graph

Step 1



Complete Undirected Graph

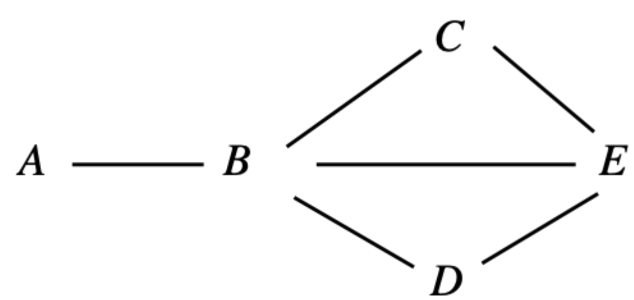
n = 0 No zero order independencies

n = 1 First order independencies

$A \perp\!\!\!\perp C \mid B$ $A \perp\!\!\!\perp D \mid B$

$A \perp\!\!\!\perp E \mid B$ $C \perp\!\!\!\perp D \mid B$

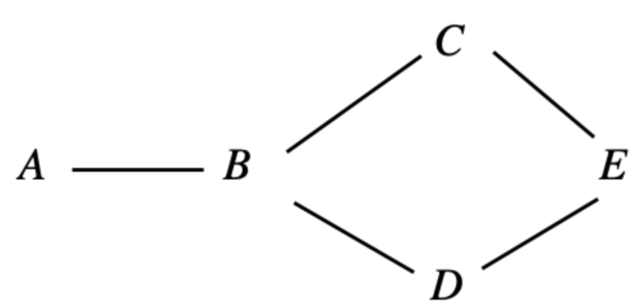
Resulting Adjacencies



n = 2: Second order independencies

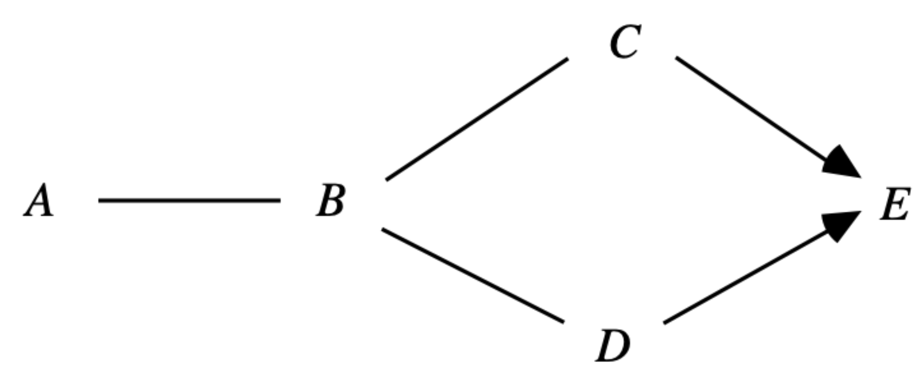
$B \perp\!\!\!\perp E \mid \{C,D\}$

Resulting Adjacencies



Step 2

Pattern

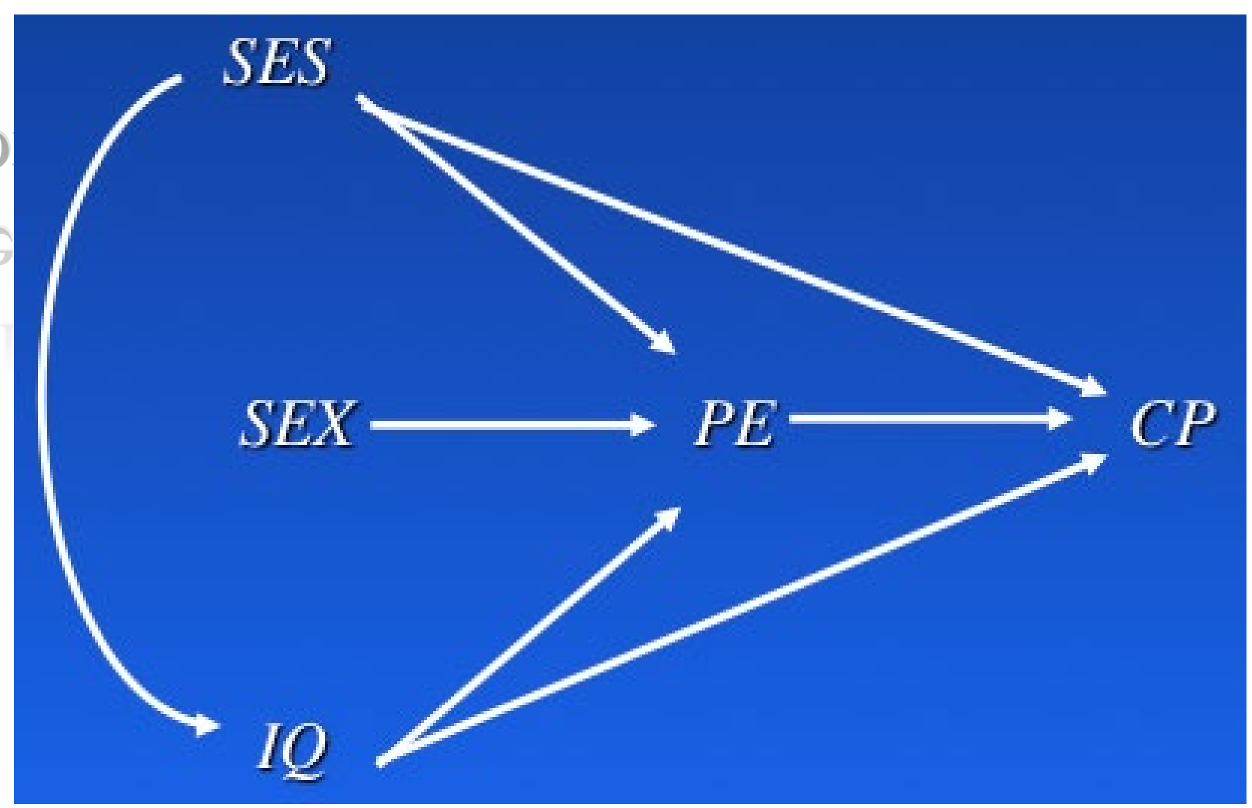


See demo with Tetrad…

Example 2: College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors.

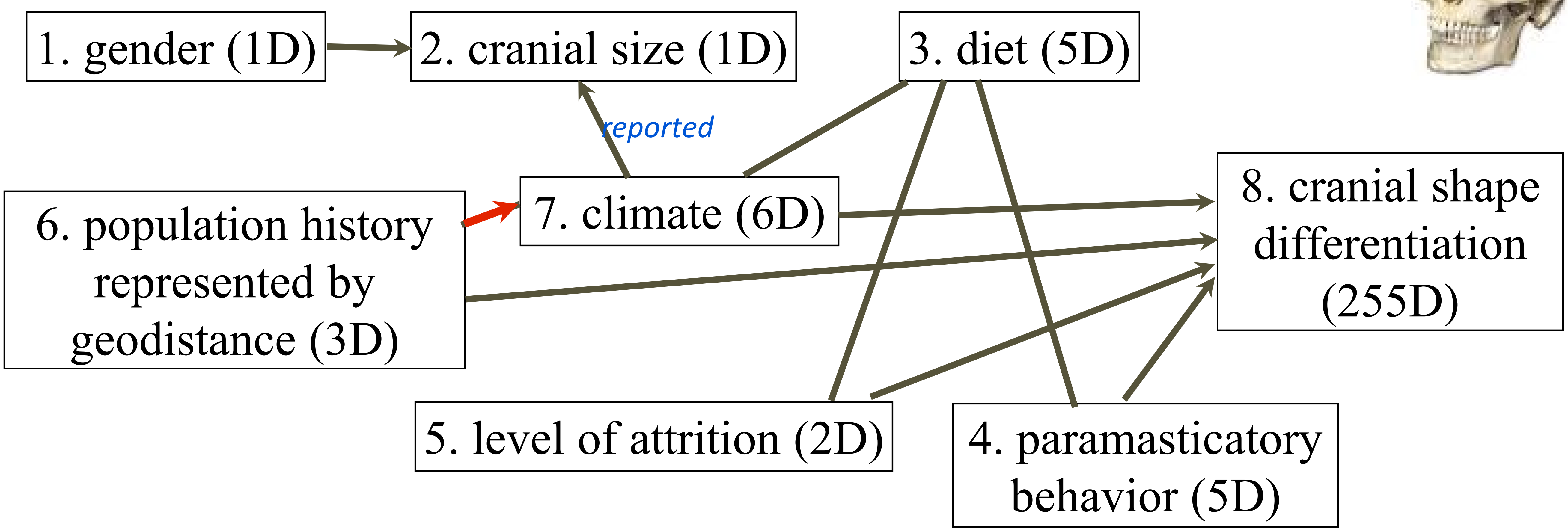
SEX [male = 0, female = 1]
IQ = Intelligence Quotient [lowest = 0, highest = 3]
CP = college plans [yes = 0, no = 1]
PE = parental encouragement [low = 0, high = 1]
SES = socioeconomic status [lowest = 0, highest = 3]



Result on the Archeology Data

Thanks to collaborator Marlijn Noback

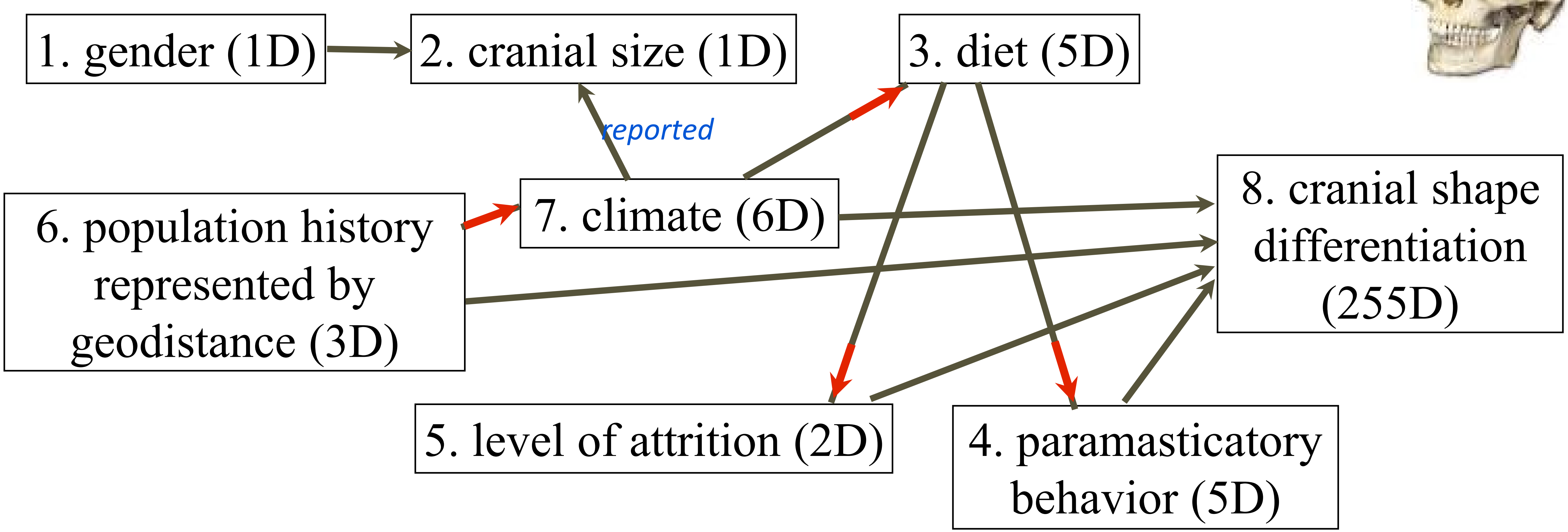
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



Result on the Archeology Data

Thanks to collaborator Marlijn Noback

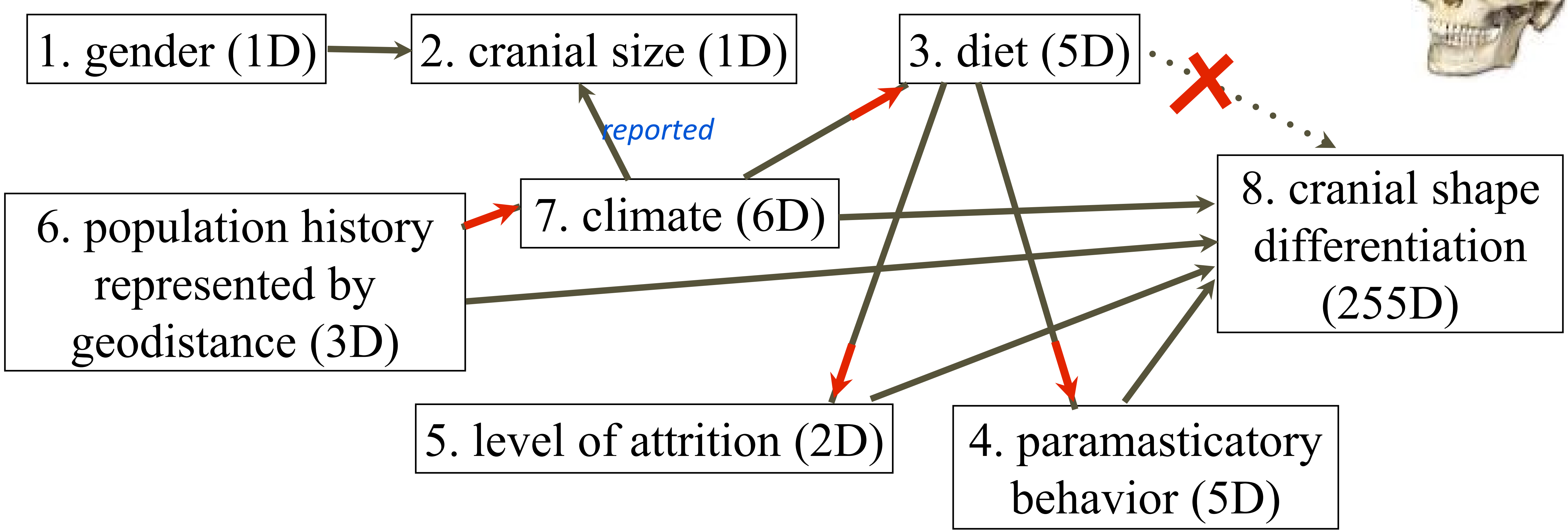
- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



Result on the Archeology Data

Thanks to collaborator Marlijn Noback

- 8 variables of 250 skeletons collected from different locations
- Different dimensions (from 1 to 255) with nonlinear dependence
- By PC algorithm + kernel-based conditional independence test (Zhang et al., 2011)



PC by causal-learn

```
from causallearn.search.ConstraintBased.PC import pc

# default parameters
cg = pc(data)
```

```
# visualization using pydot
cg.draw_pydot_graph()

# or save the graph
from causallearn.utils.GraphUtils import GraphUtils

pyd = GraphUtils.to_pydot(cg.G)
pyd.write_png('simple_test.png')

# visualization using networkx
# cg.to_nx_graph()
# cg.draw_nx_graph(skel=False)
```

PC by causal-learn

indep_test: string, name of the independence test method. Default: 'fisherz'.

- "fisherz": Fisher's Z conditional independence test.
- "chisq": Chi-squared conditional independence test.
- "gsq": G-squared conditional independence test.
- "kci": kernel-based conditional independence test. (As a kernel method, its complexity is cubic in the sample size, so it might be slow if the same size is not small.)
- "mv_fisherz": Missing-value Fisher's Z conditional independence test.

Dealing with Confounders?

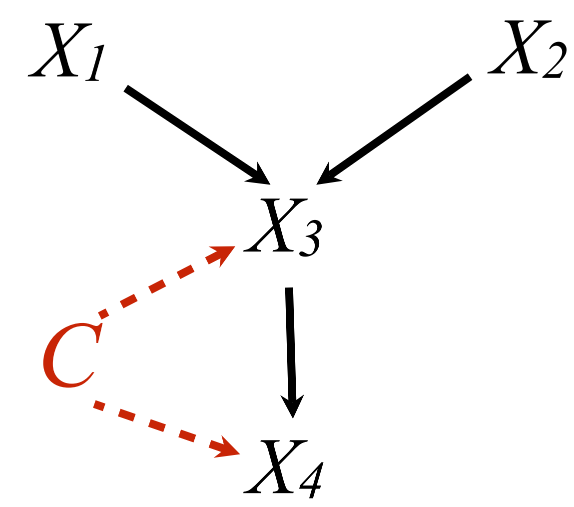
Example I

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

$X_1 \perp\!\!\!\perp X_4 \mid X_3;$

$X_2 \perp\!\!\!\perp X_4 \mid X_3.$

Possible to have confounders behind X_3 and X_4?



Dealing with Confounders?

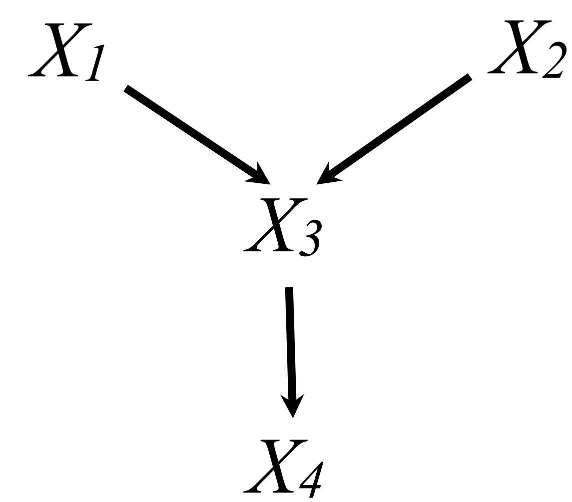
Example I

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

$X_1 \perp\!\!\!\perp X_4 \mid X_3;$

$X_2 \perp\!\!\!\perp X_4 \mid X_3.$

Possible to have confounders behind X_3 and X_4?



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?

$X_1 \rightarrow X_2 \qquad X_4 \leftarrow X_3$

(*See the FCI algorithm*)

Dealing with Confounders?

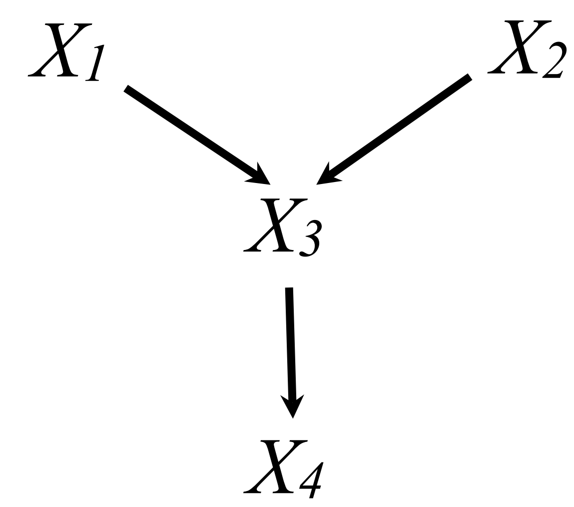
Example I

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

$X_1 \perp\!\!\!\perp X_4 \mid X_3;$

$X_2 \perp\!\!\!\perp X_4 \mid X_3.$

Possible to have confounders behind X_3 and X_4?



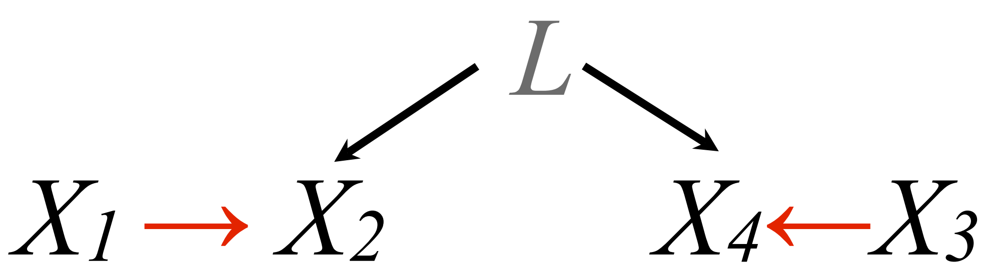
Example II

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

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

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

Are there confounders behind X_2 and X_4?



Dealing with Confounders?

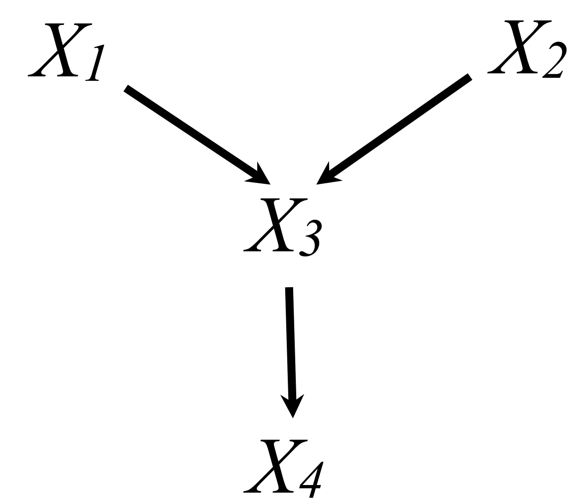
Example I

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

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

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

Possible to have confounders behind X_3 and X_4?



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

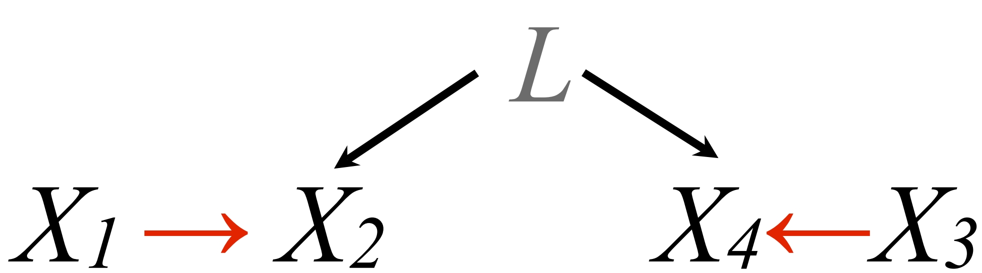
Example II

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

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

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

Are there confounders behind X_2 and X_4?



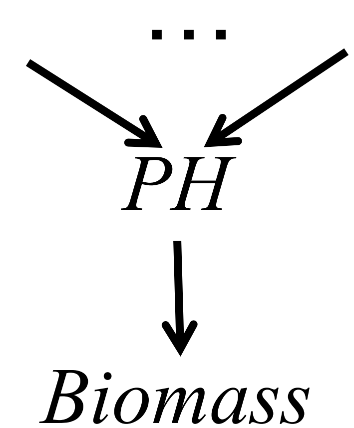
E.g., X_1: I am not sick; X_2: I am in this lecture room; X_4: you are in this lecture room; X_3: you are not sick.

(*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



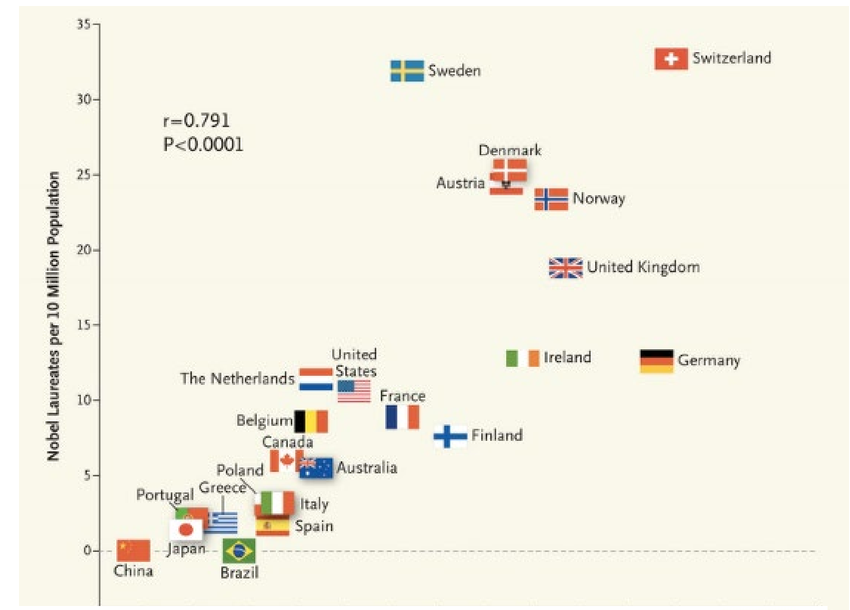
*

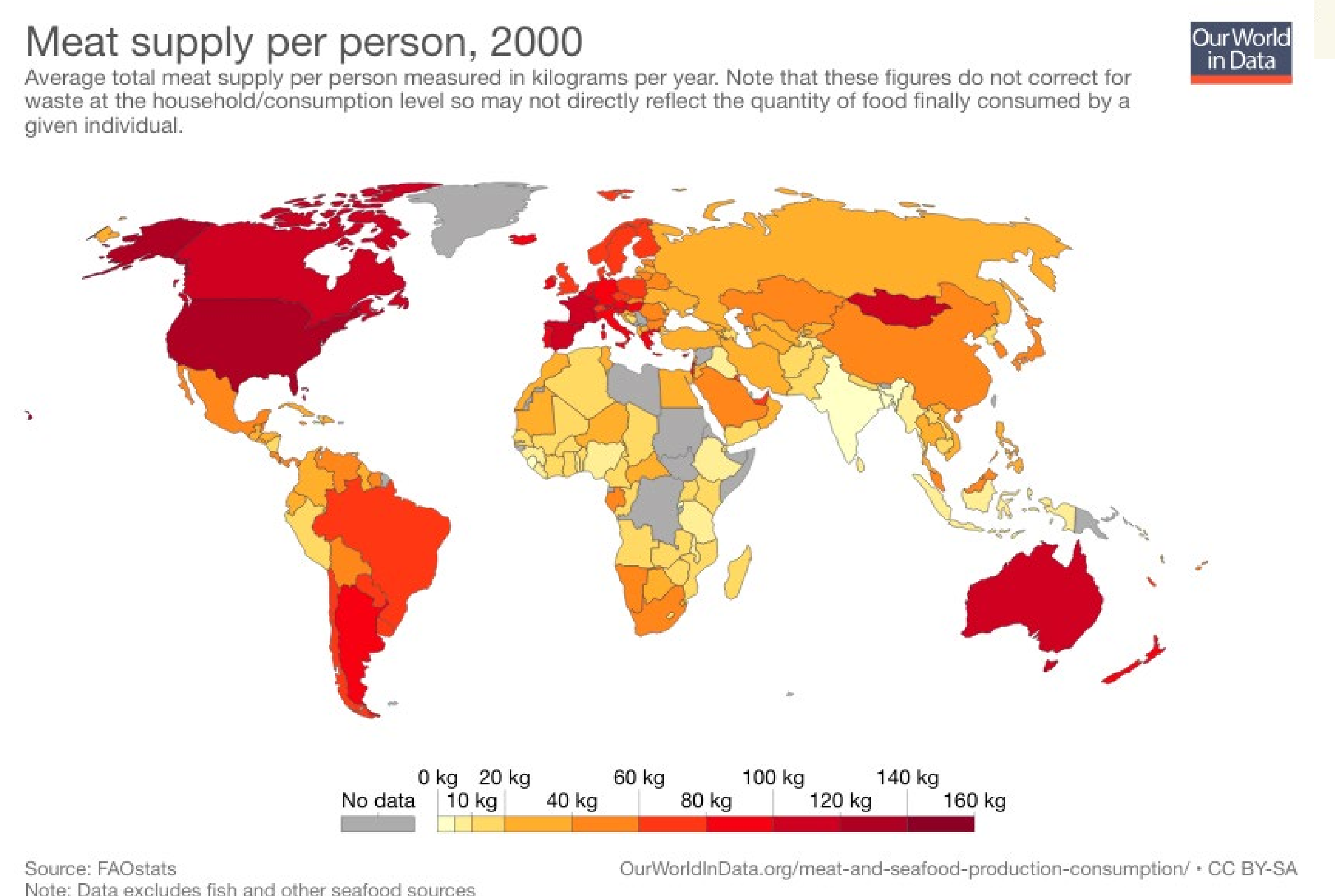
I Know There must Be Confounders: examples

$$\begin{tikzcd} & & L \arrow[dl] \arrow[dr] & & \\ X_1 \arrow[r] & X_2 & & X_4 & X_3 \arrow[l] \end{tikzcd}$$

- X_1: I am not sick; X_2: I am in class; X_4: you are in class; X_3: you are not sick

- X_1: European/South American country; X_2: leading in science; X_4: Chocolate consumption; X_3: meat supply per person





The Second Example…

$$X_1 \perp\!\!\!\perp X_3;$$
$$X_1 \perp\!\!\!\perp X_4;$$
$$X_2 \perp\!\!\!\perp X_3.$$

$$\begin{tikzcd}
 & & L \arrow[dl] \arrow[dr] & & \\
X_1 \arrow[r] & X_2 & & X_4 & X_3 \arrow[l]
\end{tikzcd}$$

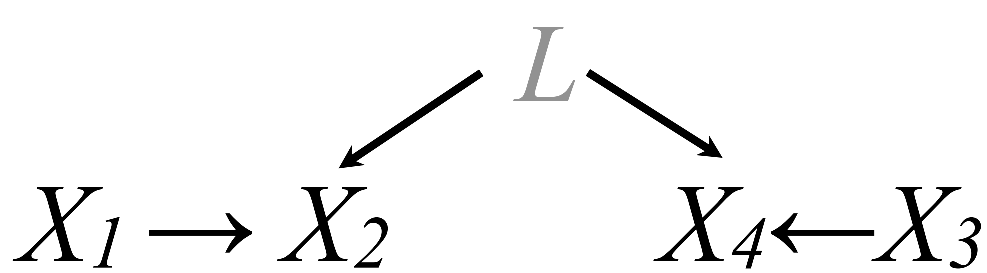
L: a latent variable

- There must exist some confounder for *X2* and *X4.*

- In the presence of latent variables, the causal process over measured variables **O** is not necessarily a DAG. How can we represent (independence) equivalence classes over **O** ?

FCI (Fast Causal Inference) Allows Confounders

- Assume the distribution over measured variables **O** is the marginal of a distribution satisfying the Markov and faithfulness conditions for the true graph
- Results represented by PAGs (Partial Ancestral Graphs)



What's FCI's output?

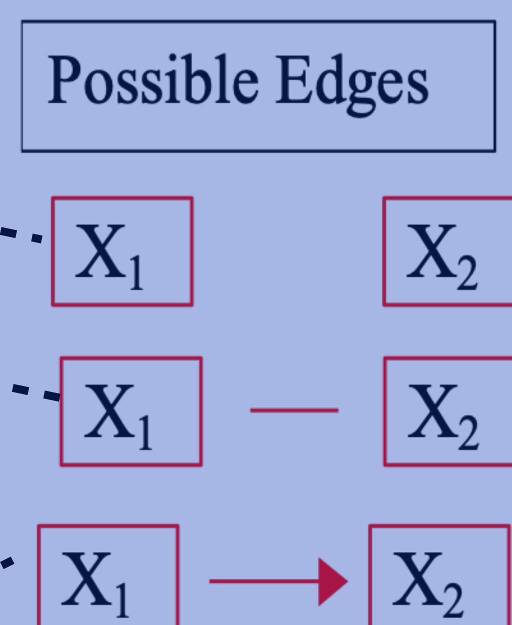
Spirtes et al., Causal inference in the presence of latent variables and selection bias, 1997

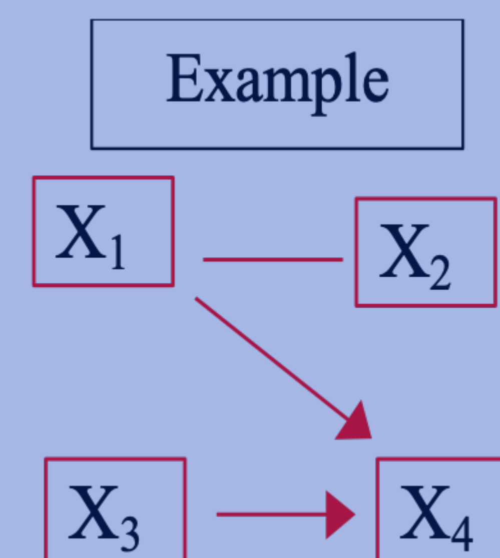
Remember the Output of PC? (Independence) Equivalent Classes: Patterns

- Two DAGs are (independence) equivalent if and only if they have the same skeletons and the same v-structures (Verma & Pearl, 1991)

- Patterns or CPDAG (Completed Partially Directed Acyclic Graph): graphical representation of (conditional) independence equivalence among models with no latent common causes (i.e., causally sufficient models)

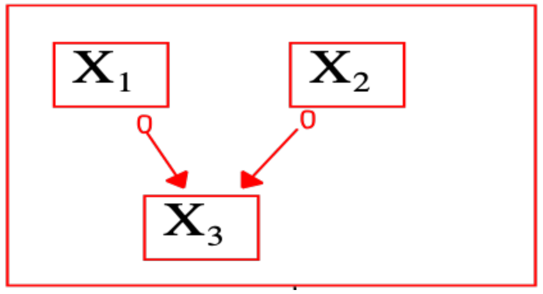
X_1 and X_2 are not adjacent in any member of the equivalent class
$X_1 \rightarrow X_2$ in some members of the equivalent class, and $X_1 \leftarrow X_2$ in some others
$X_1 \rightarrow X_2$ in every member of the equivalent class





How many DAGs in this class?

PAGs: What Edges Mean?



$X_1 \quad\quad X_2$ — X_1 and X_2 are not **adjacent**

X_1 o→ X_2 — X_2 is not an **ancestor** of X_1

X_1 o—o X_2 — No set d-separates X_2 and X_1

$X_1 \rightarrow X_2$ — X_1 is a **cause** of X_2

$X_1 \leftrightarrow X_2$ — There is a **latent common cause** of X_1 and X_2

FCI by causal-learn

```python
from causallearn.search.ConstraintBased.FCI import fci

# default parameters
G, edges = fci(data)
```

```python
# visualization
from causallearn.utils.GraphUtils import GraphUtils

pdy = GraphUtils.to_pydot(G)
pdy.write_png('simple_test.png')
```

Summary: Constrain-based approach and extensions

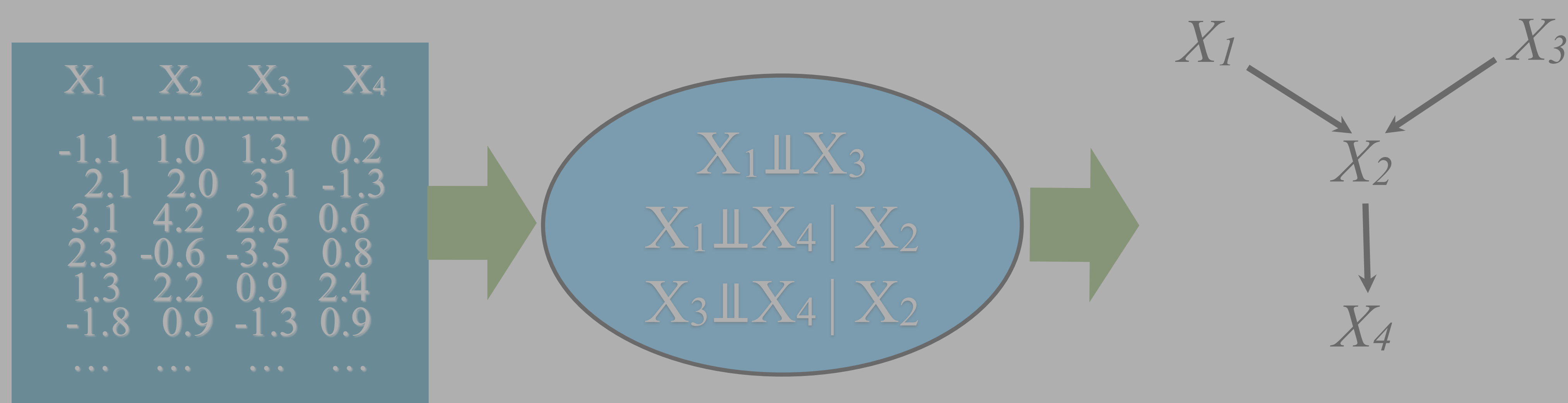
- Conditional independence relations help in causal discovery
- What assumptions are needed
- Constraint-based approach
- Confounders?

Score-based Causal Discovery

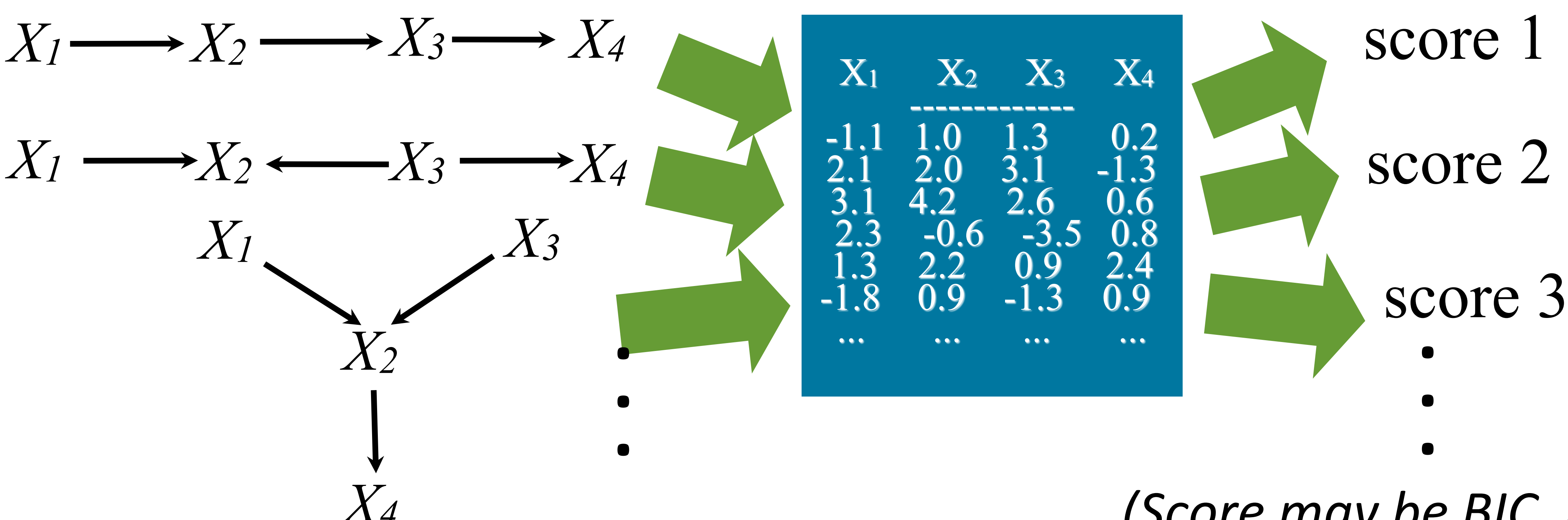
- Possibility
- GES

Constraint-Based vs. Score-Based

- Constraint-based methods



- Score-based methods

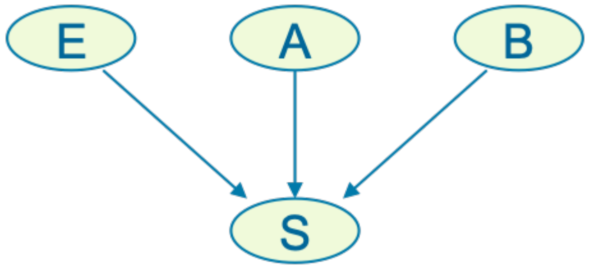


Which one is the best?

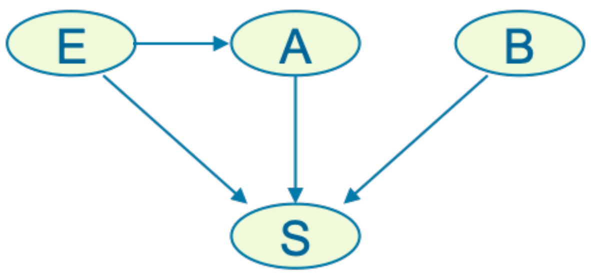
(Score may be BIC, AIC, etc.)

Why Is It Possible?

"True" structure



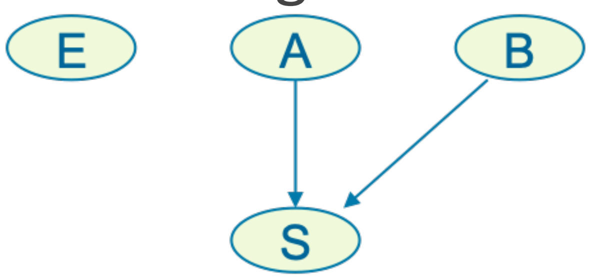
Adding an arc



- Increases the number of parameters to be fitted;

Wrong assumptions about causality and domain structure

Missing an arc



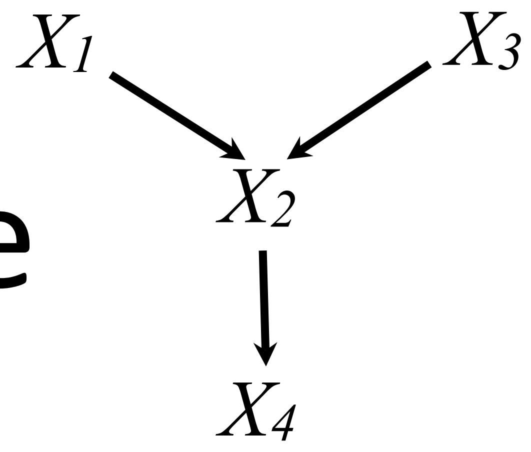
- Cannot be compensated by accurate fitting of parameters;

Also misses causality and domain structure

Key Issues

- What score to use?
- How to traverse the search space of the graph?
    - DAGs? Equivalence classes?
    - How to do optimization?

Searching for Network Structure



- Sad news: Given a complete dataset and no hidden variables, locating the Bayesian network structure that has the highest posterior probability is NP-hard (Chickering, 1996; Chickering, et al, JMLR, 2004).
- Greedy search often used
- Some algorithms guarantee locating the generating model in the large sample limit (assuming Markov, Faithfulness, and some other conditions); e.g., the GES algorithm (Chickering, JMLR, 2002)
- The ability to approximate the generating network is often quite good

Chickering, Learning Bayesian networks is NP-complete, Learning from Data, 1996

GES (Greedy Equivalence Search): Score Function

- Assumptions: The score is
  - score equivalent (i.e., assigning the same score to equivalent DAGs)
  - locally consistent: score of a DAG increases (decreases) when adding any edge that eliminates a false (true) independence constraint
  - decomposable: $$Score(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} Score(X_i, \mathbf{Pa}_i^{\mathcal{G}})$$
- E.g., BIC: $$S_B(\mathcal{G}, \mathbf{D}) = \log p(\mathbf{D} | \hat{\boldsymbol{\theta}}, \mathcal{G}^h) - \frac{d}{2} \log m$$

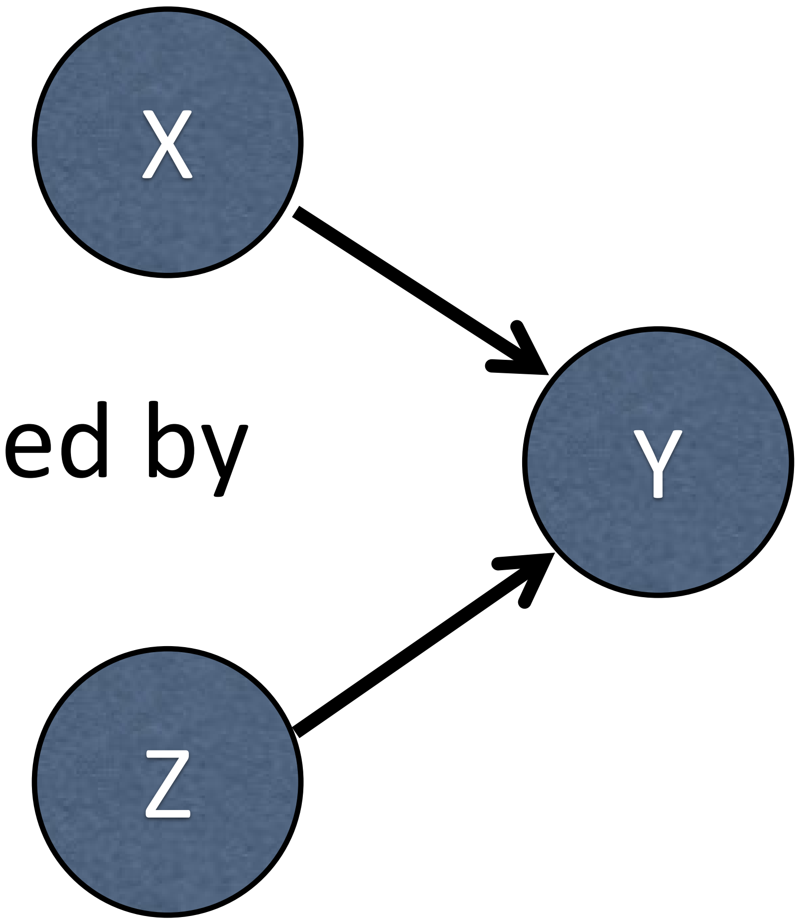
Chickering, Optimal Structure Identification With Greedy Search, Journal of Machine Learning Research, 2002

GES: Search Procedure

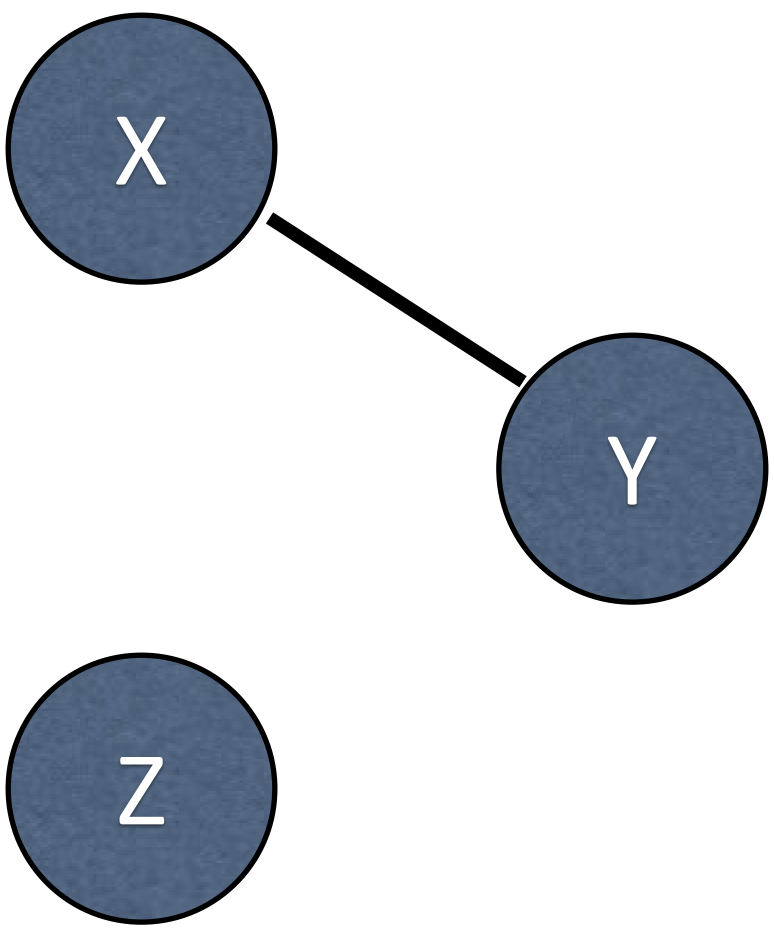
- Performs forward (addition) / backward (deletion) equivalence search through the space of DAG equivalence classes
  - Forward Greedy Search (FGS)
    - Start from some (sparse) pattern (usually the empty graph)
    - Evaluate all possible patterns with one more adjacency that entail strictly fewer CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum
  - Backward Greedy Search (BGS)
    - Start from the output of Stage (1)
    - Evaluate all possible patterns with one fewer adjacency that entail strictly more CI statements than the current pattern
    - Move to the one that increases the score most
    - Iterate until a local maximum

GES

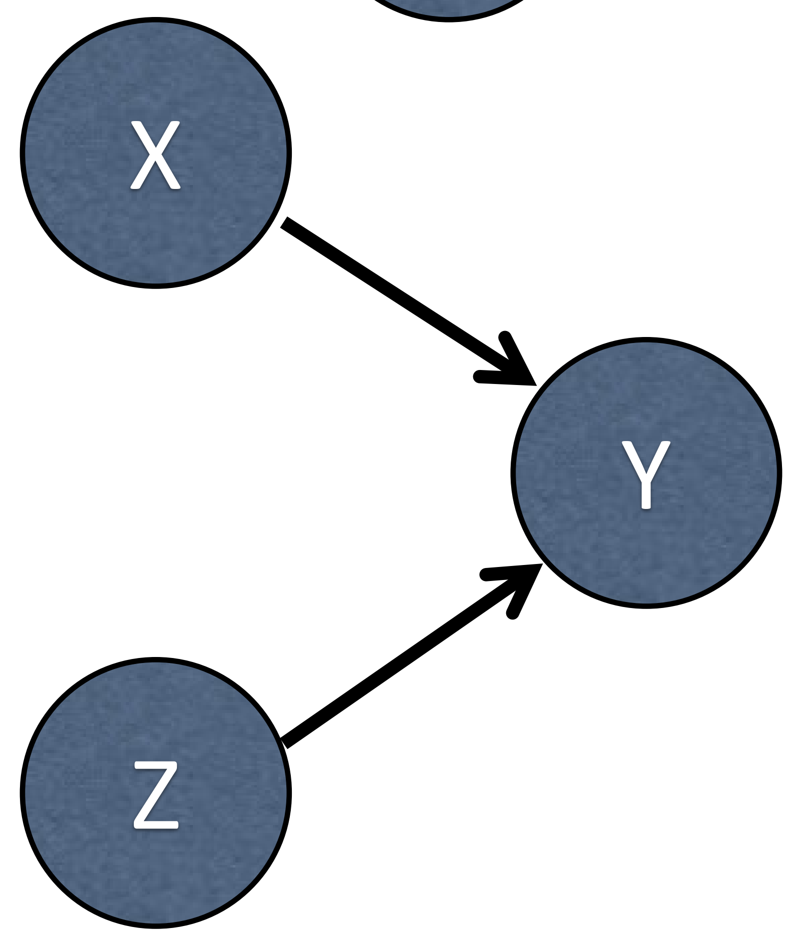
Suppose data were generated by



(1)

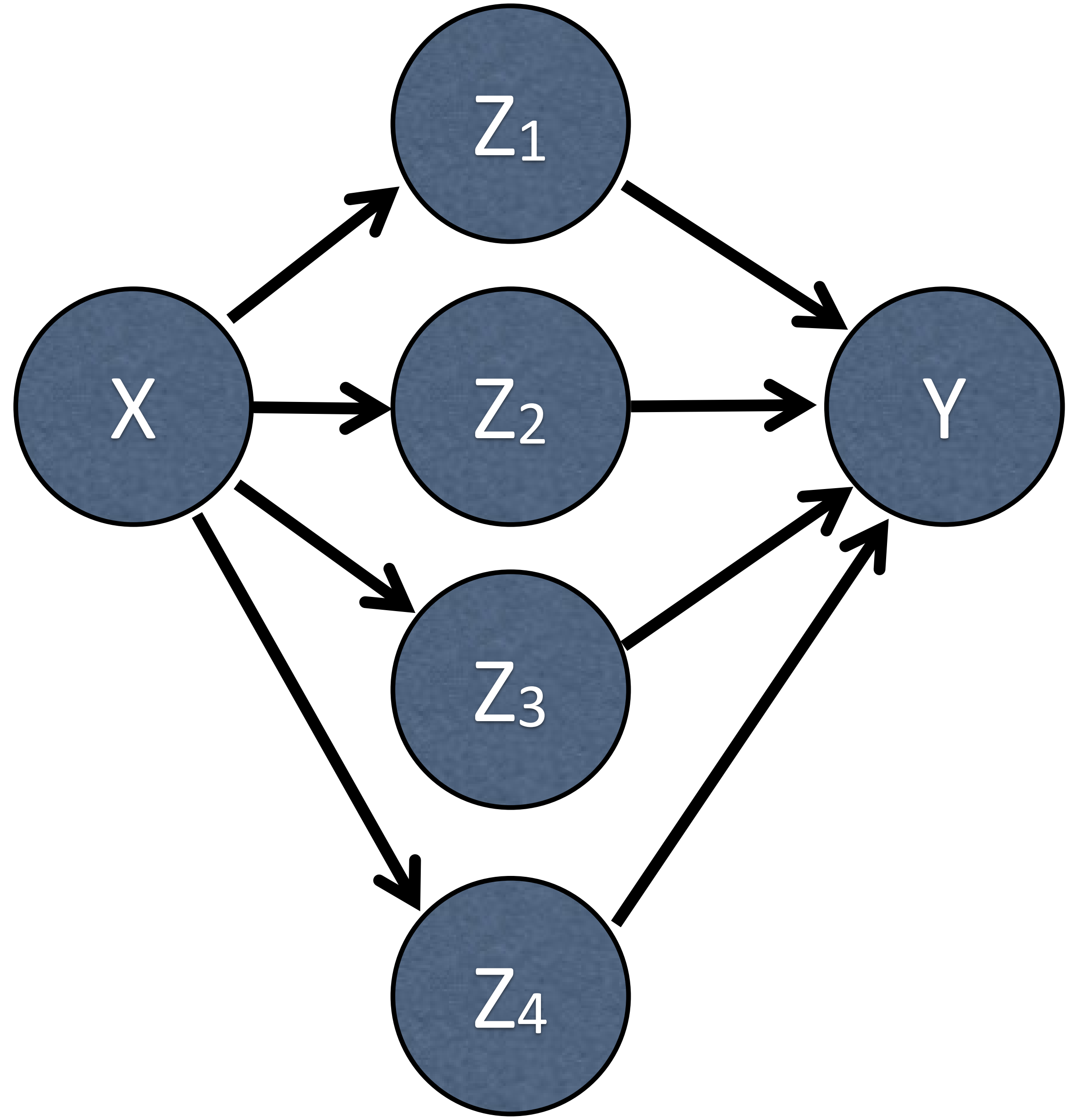


(2)



GES

Suppose data were generated by



Imagine the GES procedure...

GES by causal-learn

```python
from causallearn.search.ScoreBased.GES import ges

# default parameters
Record = ges(X)
```

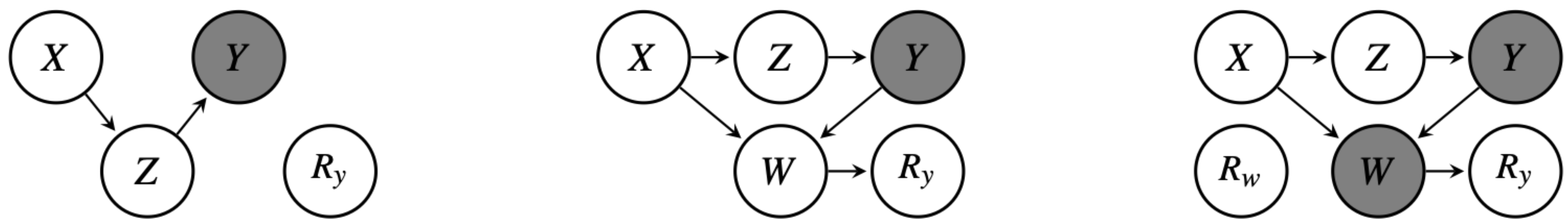
```python
# Visualization using pydot
from causallearn.utils.GraphUtils import GraphUtils
import matplotlib.image as mpimg
import matplotlib.pyplot as plt
import io

pyd = GraphUtils.to_pydot(Record['G'])
tmp_png = pyd.create_png(f="png")
fp = io.BytesIO(tmp_png)
img = mpimg.imread(fp, format='png')
plt.axis('off')
plt.imshow(img)
plt.show()


# or save the graph
pyd.write_png('simple_test.png')
```

Practical Issues

- Missing data
- Nonstationary/heterogenous data

Issue 1: Causal Discovery in the Presence of Missing Data

X1	X2	X3	X4	X5	X6
-9.4653403e-01	6.6703495e-01	8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01			-4.6381657e-01	-1.8280031e+00	
	5.1435422e-01	6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01		5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02
		-1.3440612e+00			-7.3325009e-01
1.3261794e+00	-6.1971037e-01	-1.0498756e-01	1.4171149e+00	1.6251026e+00	3.7478050e-01
-2.1128404e+00	1.3359744e-02	-2.0209600e+00	-1.7172659e+00	-2.4746799e+00	-2.8026586e+00
1.5453163e+00	-5.3986972e-01	4.5157367e-01	1.5566262e+00	9.3882105e-01	-4.3382982e-01
6.5974086e-02	5.5826895e-01	6.5247930e-01	-5.7895322e-01	5.0062743e-01	1.0183537e+00
8.9772858e-01	2.6752870e-01	-4.9204975e-01	7.7933358e-02	8.3467624e-01	9.2744311e-01



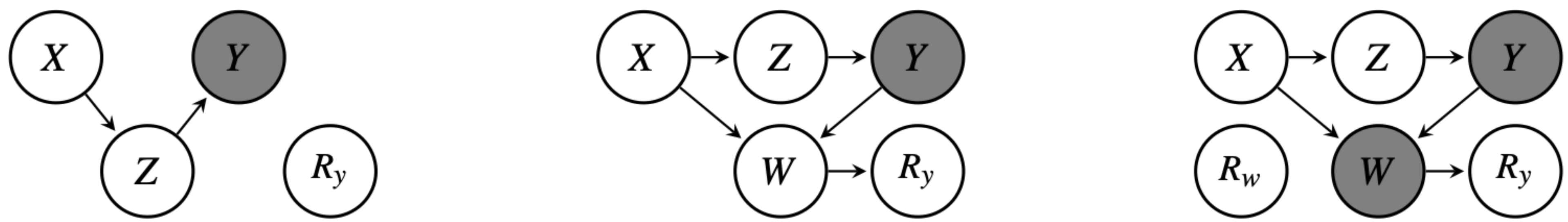
(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- Conditional independence relations in the data are sensitive to the missingness mechanism
- Key issue: Recover conditional independence relations in the original population from incomplete data

R. Tu, C. Zhang, P. Ackermann, K. Mohan, H. Kjellström, C. Glymour, K. Zhang, "Causal discovery in the presence of missing data," AISTATS 2019

Causal Discovery in the Presence of Missing Data

X1	X2	X3	X4	X5	X6
-9.4653403e-01	6.6703495e-01	8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01			-4.6381657e-01	-1.8280031e+00	
	5.1435422e-01	6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01		5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02
		-1.3440612e+00			-7.3325009e-01
1.3261794e+00	-6.1971037e-01	-1.0498756e-01	1.4171149e+00	1.6251026e+00	3.7478050e-01
-2.1128404e+00	1.3359744e-02	-2.0209600e+00	-1.7172659e+00	-2.4746799e+00	-2.8026586e+00
1.5453163e+00	-5.3986972e-01	4.5157367e-01	1.5566262e+00	9.3882105e-01	-4.3382982e-01
6.5974086e-02	5.5826895e-01	6.5247930e-01	-5.7895322e-01	5.0062743e-01	1.0183537e+00
8.9772858e-01	2.6752870e-01	-4.9204975e-01	7.7933358e-02	8.3467624e-01	9.2744311e-01



(a) An MCAR graph (b) An MAR graph (c) An MNAR graph

- ***R*** is the set of missingness indicators that represent the status of missingness
- If R_X is 1, the corresponding value of X is missing; if it is 0, it is observed
- Missingness graph

Categories of Missing Data Mechanism

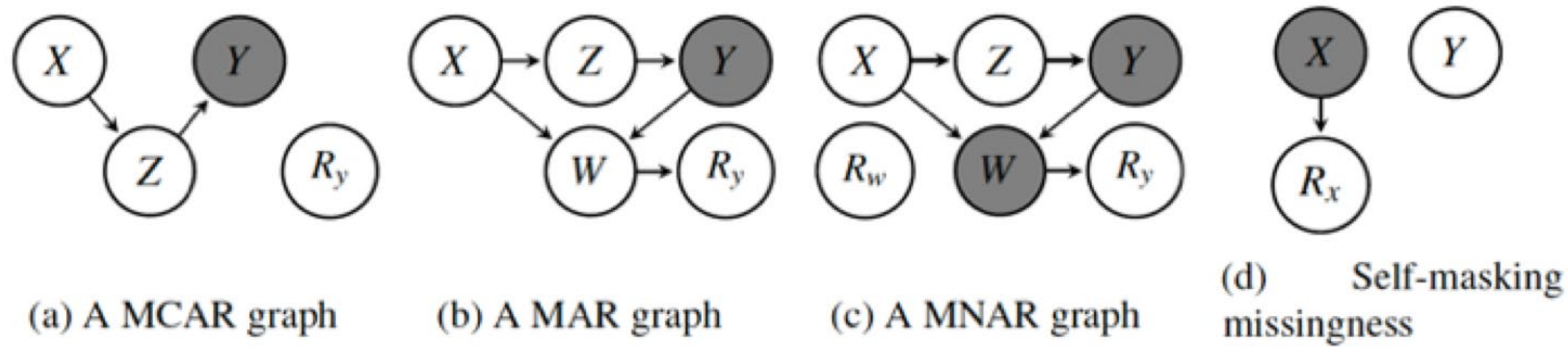


Figure 1: Exemplar missingness graphs in MCAR, MAR, MNAR, and self-masking missingness. X, Y, Z, and W are random variables. In missingness graphs, gray nodes are partially observed variables, and white nodes are fully observed variables. R_x, R_y, and R_w are the missingness indicators of X, Y, and W.

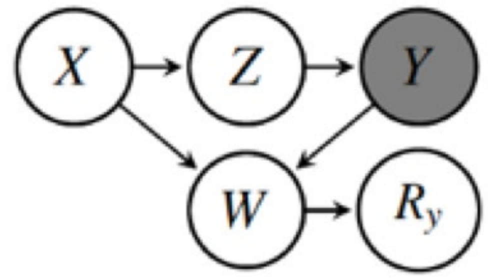
- All missing data mechanisms fall into one of the following three categories (Rubin, 1976):
    - Data are Missing Completely At Random (MCAR) if the cause of missingness is purely random.
    - Data are Missing At Random (MAR) when the direct cause of missingness is fully observed.
    - Data that are neither MAR nor MCAR fall under the Missing Not At Random (MNAR) category.

*

Assumptions for the Method

- Assumption 1 (Missingness indicators are not causes): No missingness indicator can be a cause of any substantive (observed) variable.
- Assumption 2 (Faithful observability): Any conditional independence relation in the observed data also holds in the unobserved data.
- Assumption 3 (No deterministic relation between missingness indicators): No missingness indicator can be a deterministic function of any other missingness indicators.
- Assumption 4 (No self-masking missingness): Self-masking missingness refers to missingness in a variable that is caused by itself.

Missing-Value PC (MVPC)



- Add missingness variables **R** to the dataset with measured variables **V**
- Create knowledge that **R** variables do not cause **V** variables
- Run PC adjacency search over **V**∪**R**
- Identify adjacencies over **V** in triangles over **V**∪**R**-–these might be false positives!
- Try to remove these extra adjacencies using *correction*…
- Finally, do collider orientation and apply the Meek rules to graph *G* over **V**

MVPC by causal-learn

```
# default parameters
cg = pc(data)

# or customized parameters
cg = pc(data, alpha, indep_test, stable, uc_rule, uc_priority, mvpc,
```

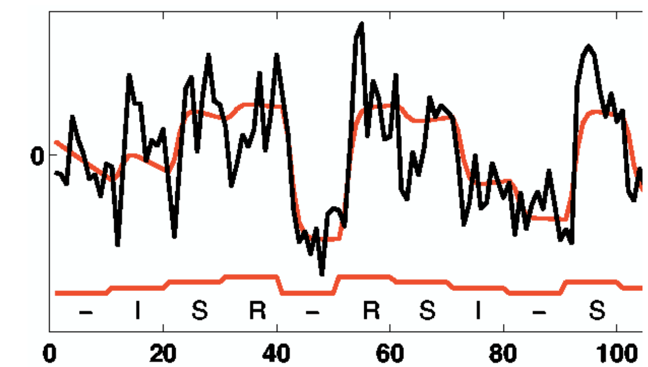
mvpc: use missing-value PC or not. Default: False.

indep_test: string, name of the independence test method. Default: 'fisherz'.

- "fisherz": Fisher's Z conditional independence test.
- "chisq": Chi-squared conditional independence test.
- "gsq": G-squared conditional independence test.
- "kci": kernel-based conditional independence test. (As a kernel method, its complexity is cubic in the sample size, so it might be slow if the same size is not small.)
- "mv_fisherz": Missing-value Fisher's Z conditional independence test.

Issue 2: Nonstationary/Heterogeneous Data and Causality

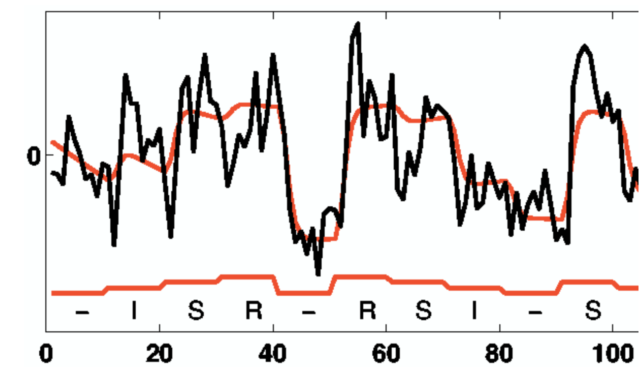
- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling & distribution shift heavily coupled
  - *P*(cause) and *P*(effect | cause) change independently



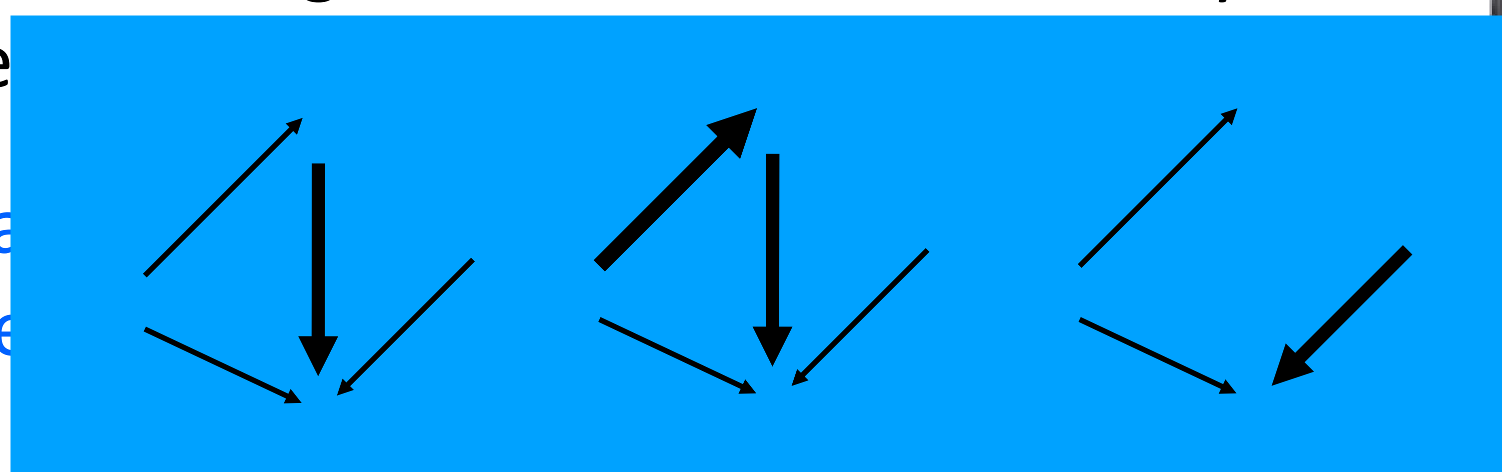
Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020
Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

Issue 2: Nonstationary/Heterogeneous Data and Causality

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions

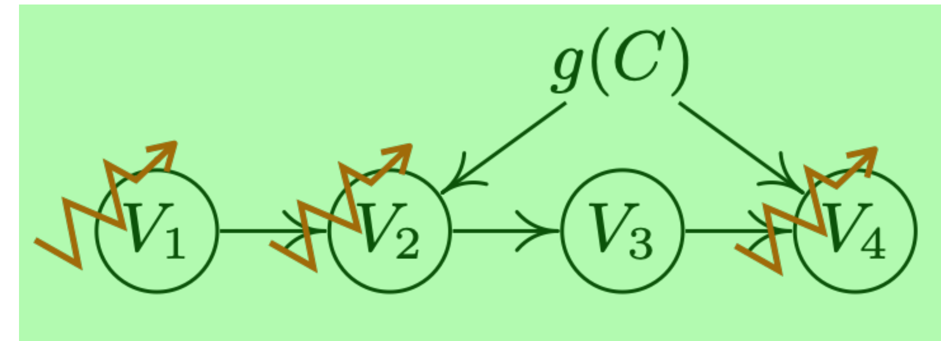


- Causal modeling & distribution shift heavily couple
  - *P*(ca
    inde

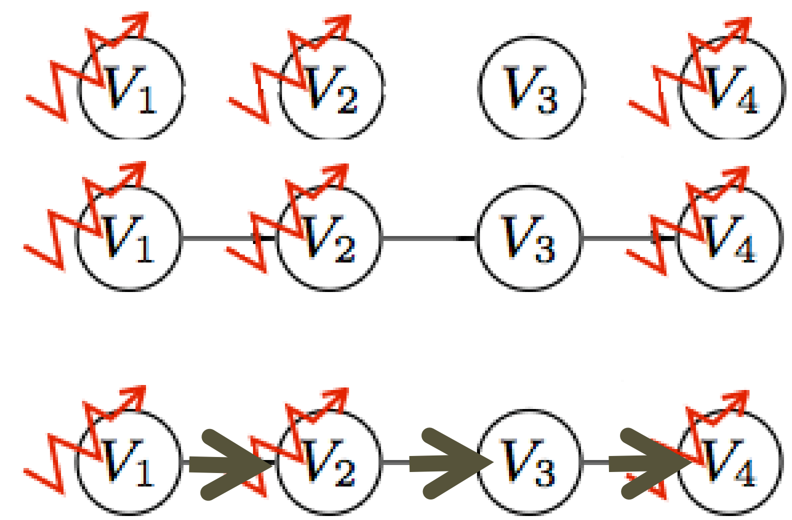


Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020
Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

Causal Discovery from Nonstationary/Heterogeneous Data

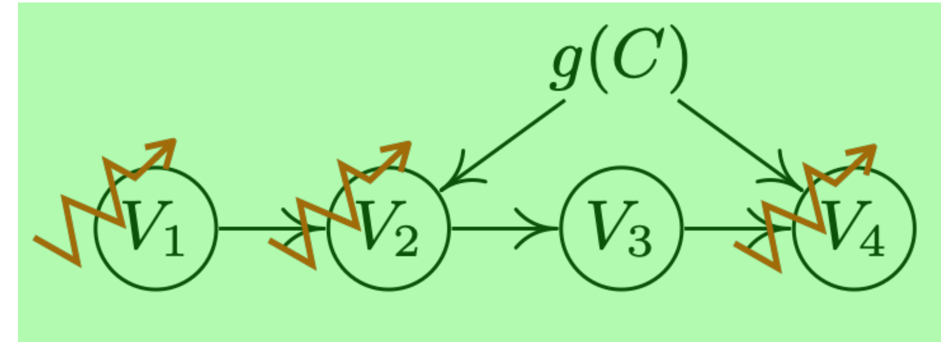


- Questions to answer:
  - Method to determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause)
  - How do the nonstationary modules change over time / across data sets?



Kernel nonstationary driving force estimation

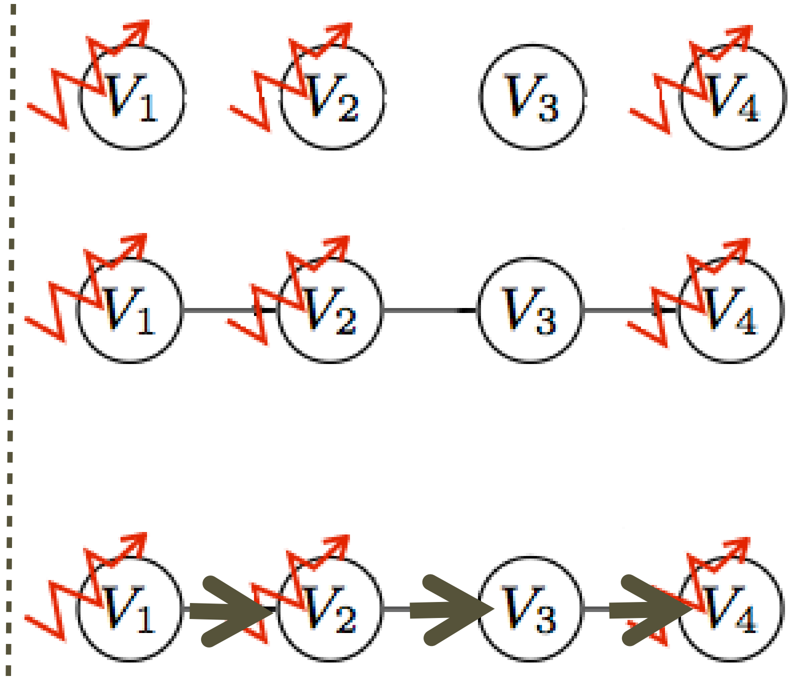
Discovery & Visualization of Changing Causal Modules



* Questions to answer for causal discovery:

- Identify variables with changing causal modules & recover causal skeleton?
- Identify causal directions by using distribution shifts?
- Visualize the change in causal modules?

With our proposed approach:

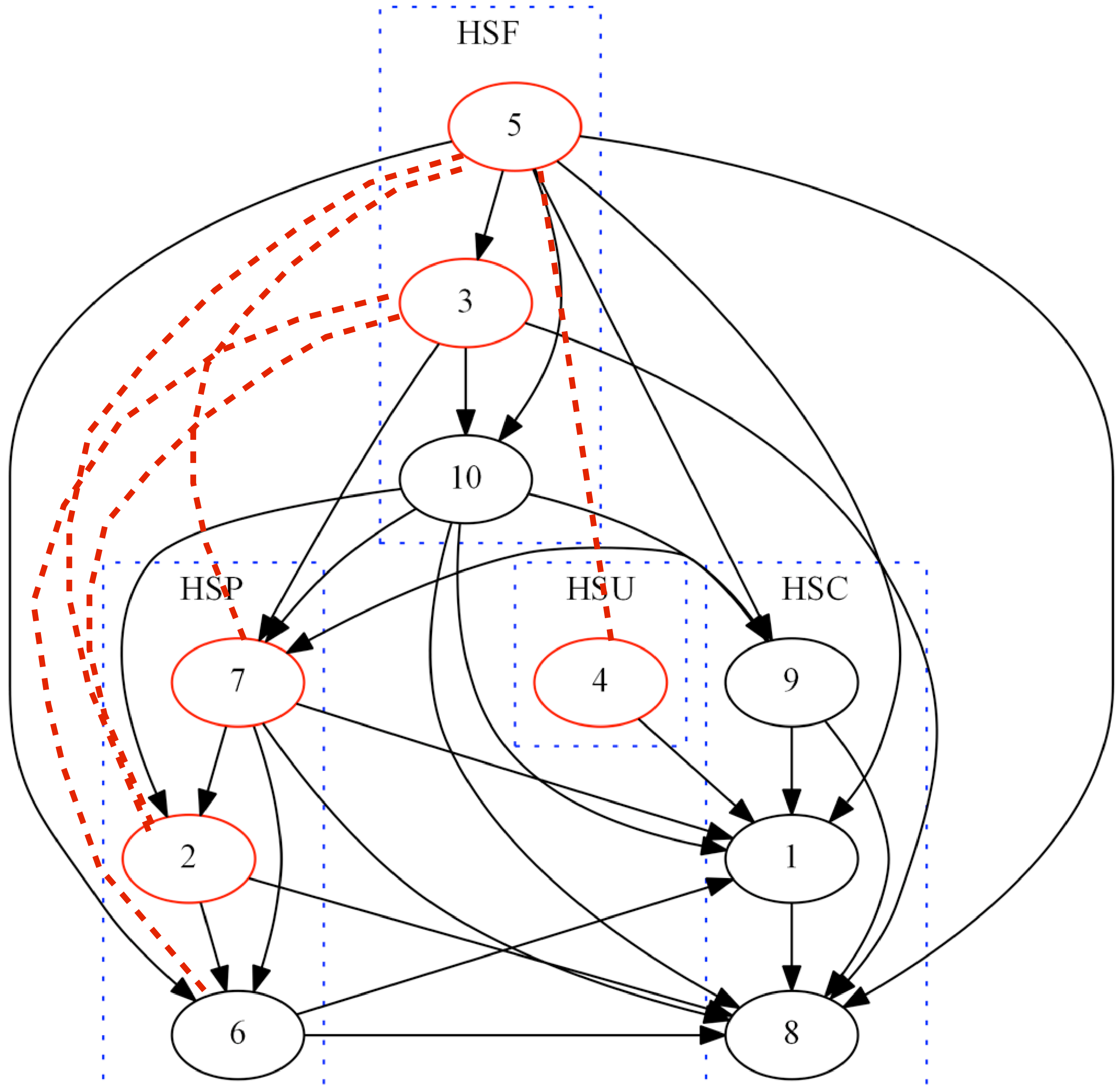


- Incorporate time/domain index C as a surrogate + apply constraint-based causal discovery methods
- Independent changes in P(cause) and P(effect | cause)
- Find a mapping of $P(V_i |PA^i)$ to capture its variability

Kernel nonstationarity visualization (KNV)

Causal Analysis of Major Stocks in Hong Kong Market (10/09/2006 - 08/09/2010)

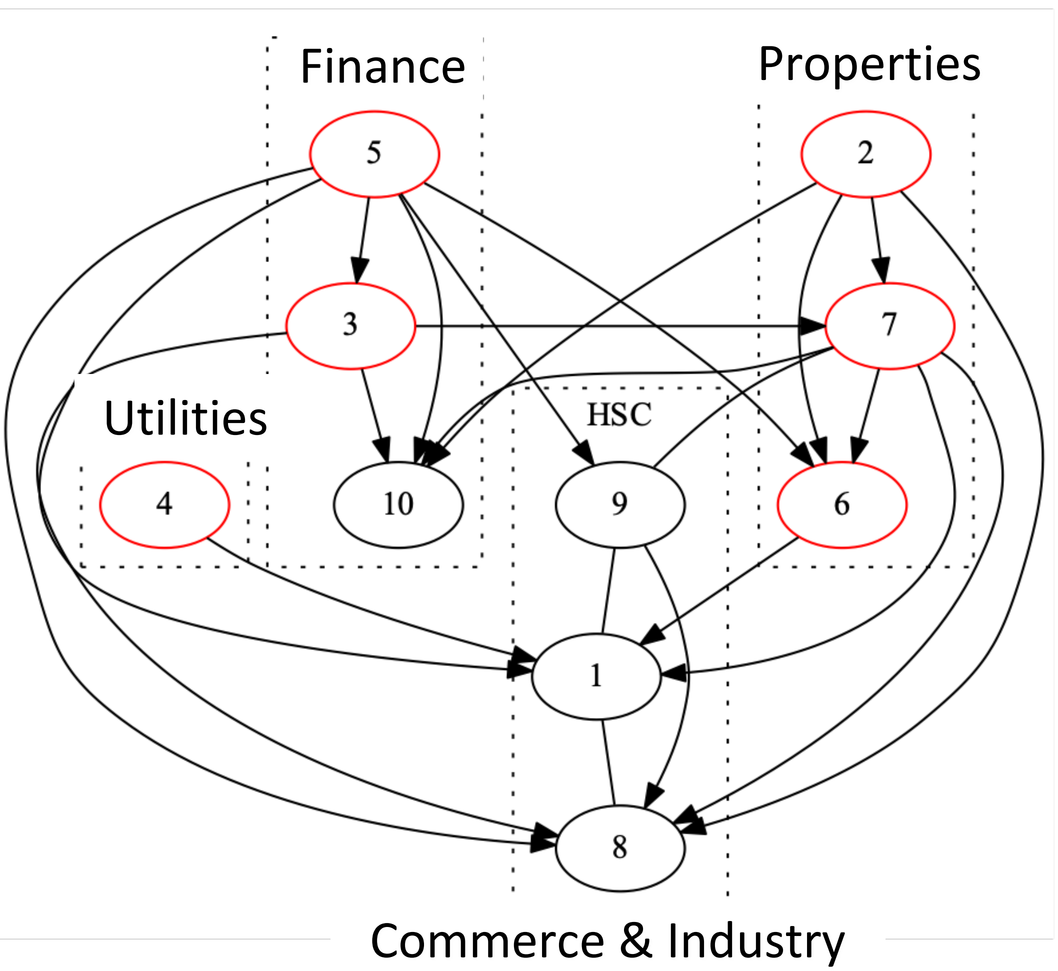
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways,
10. Bank of China Hong Kong



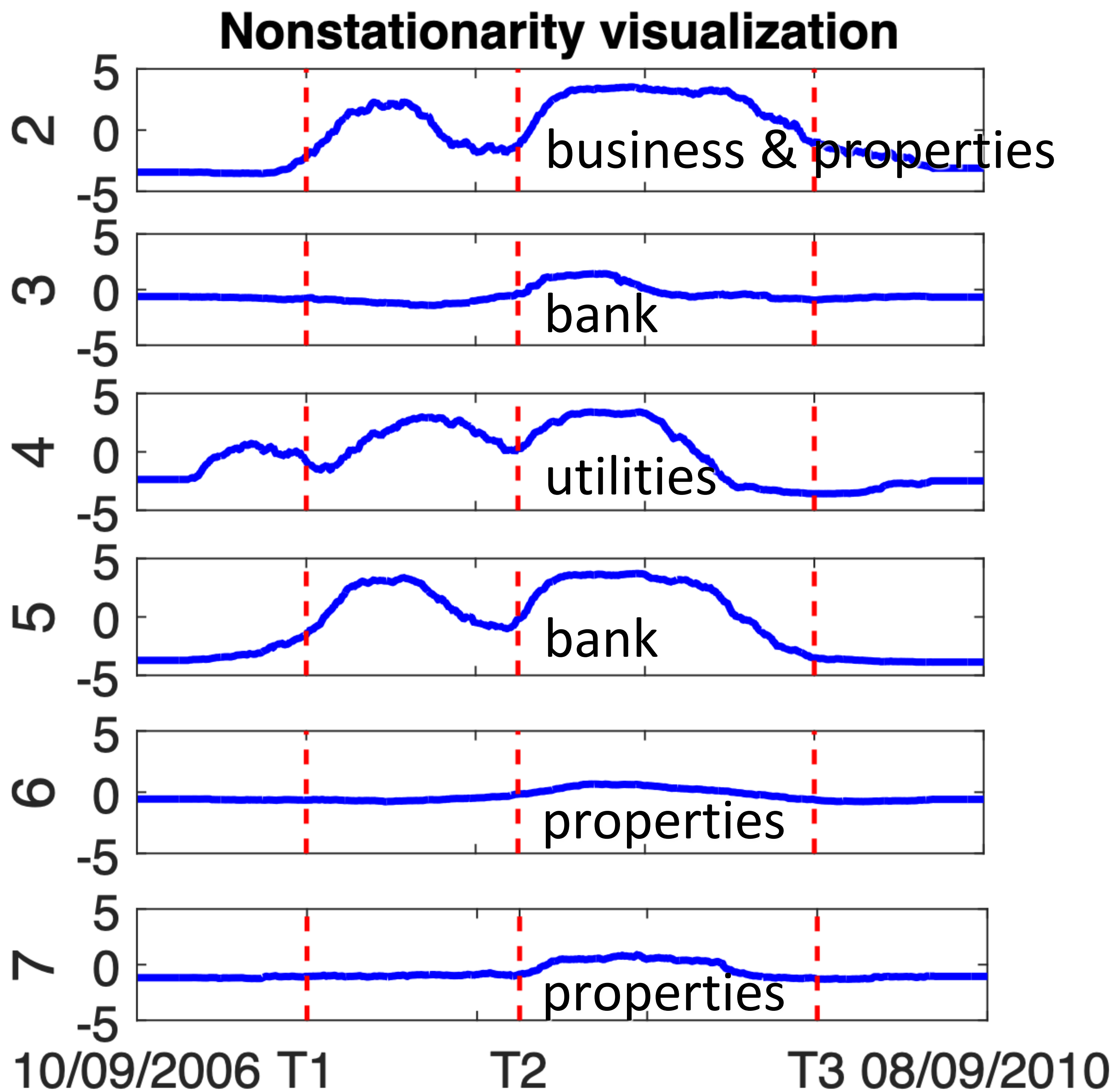
- HSF and HSP usually have nonstationary confounders

Nonstationarity Visualization

(https://research.stlouisfed.org/fred2/series/TEDRATE)

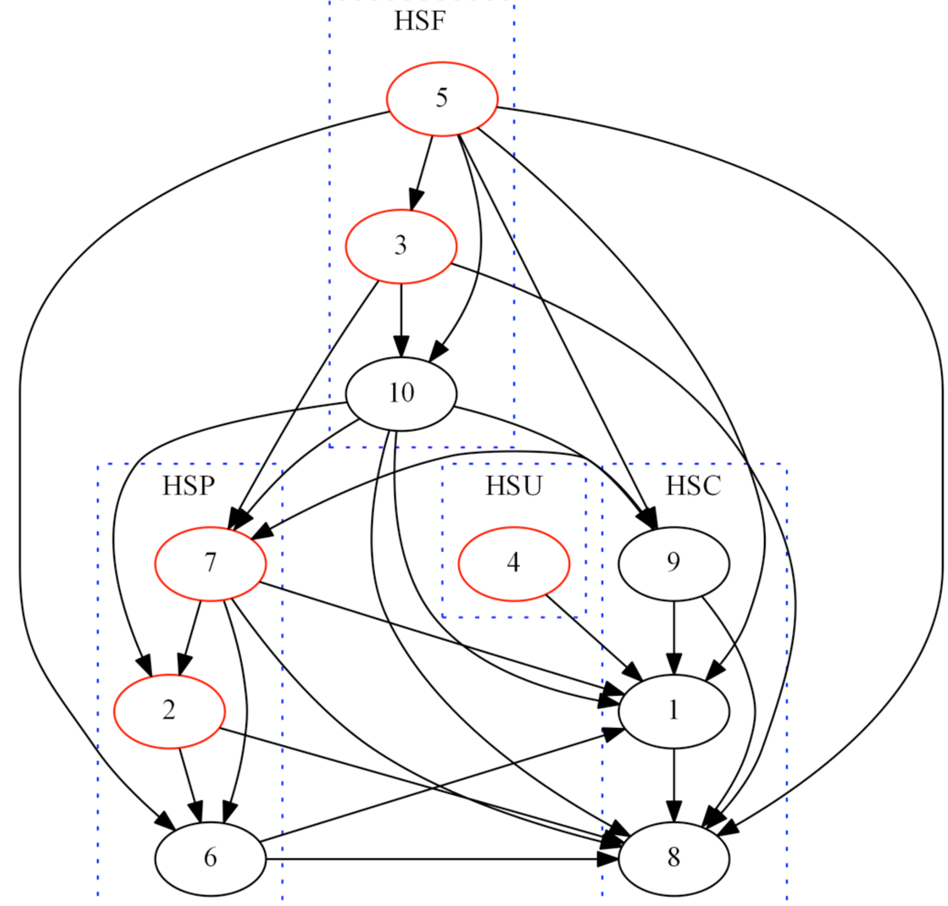


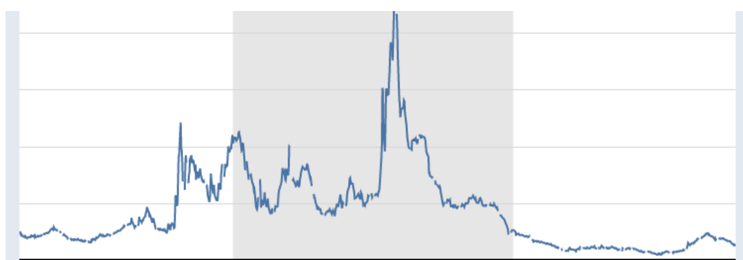
1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4.Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong



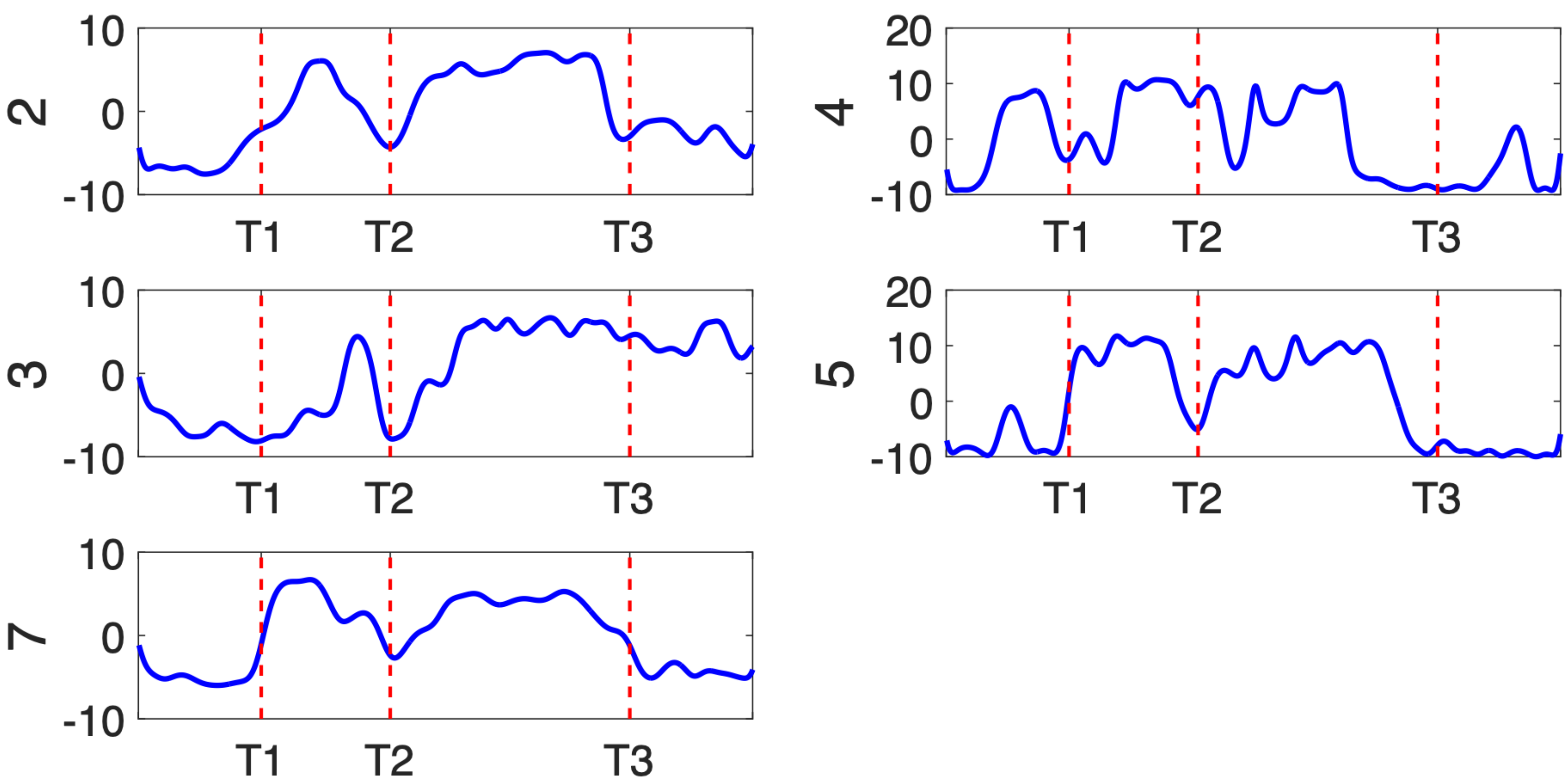
Nonstationarity Driving Force

1. Cheng Kong Holdings,
2. Wharf (Holdings),
3. HSBC,
4. Hong Kong Electric Holdings,
5. Hang Seng Bank,
6. Henderson Land Dev.,
7. Sun Hung Kai Properties,
8. Swire Group,
9. Cathay Pacific Airways
10. Bank of China Hong Kong



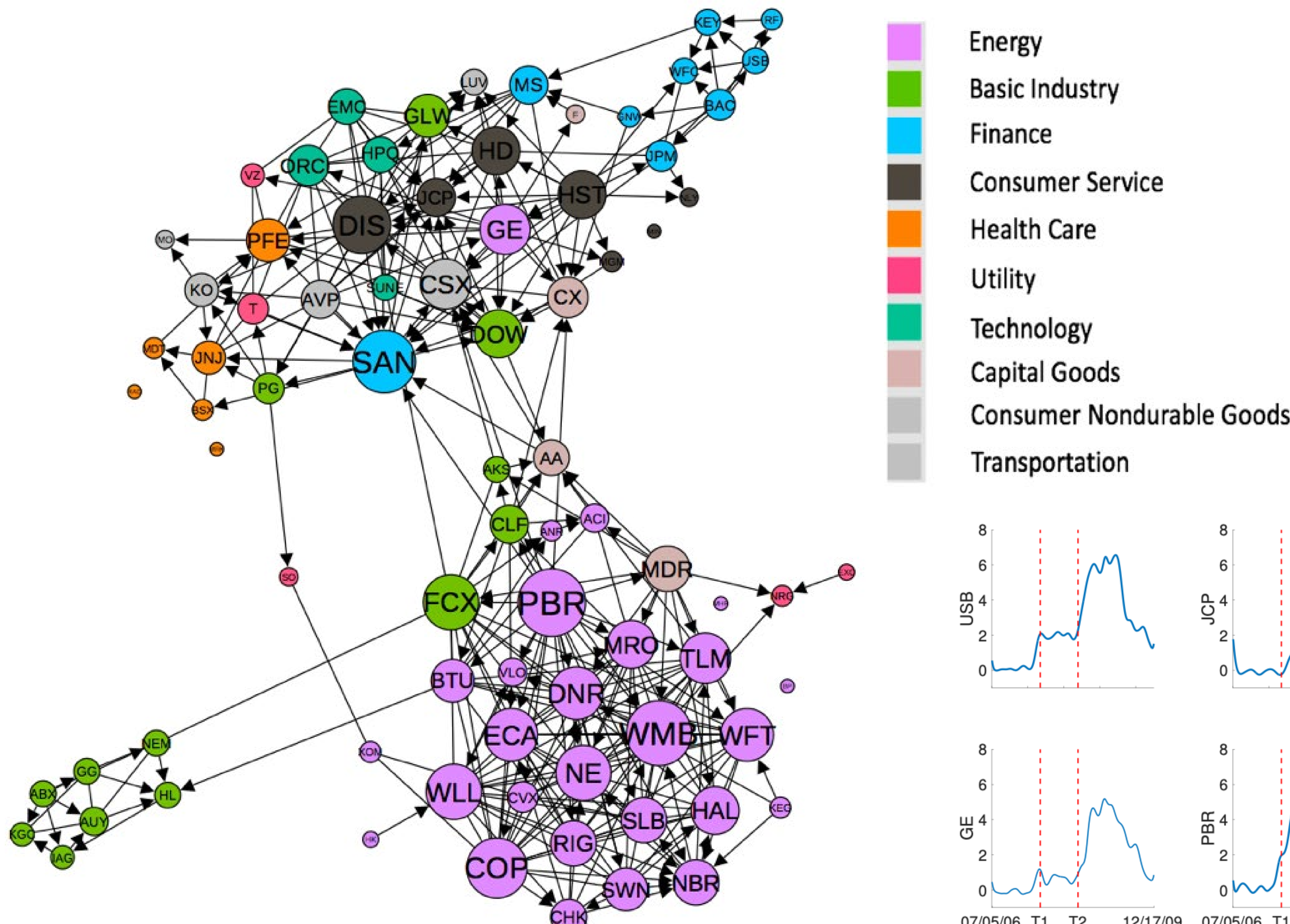


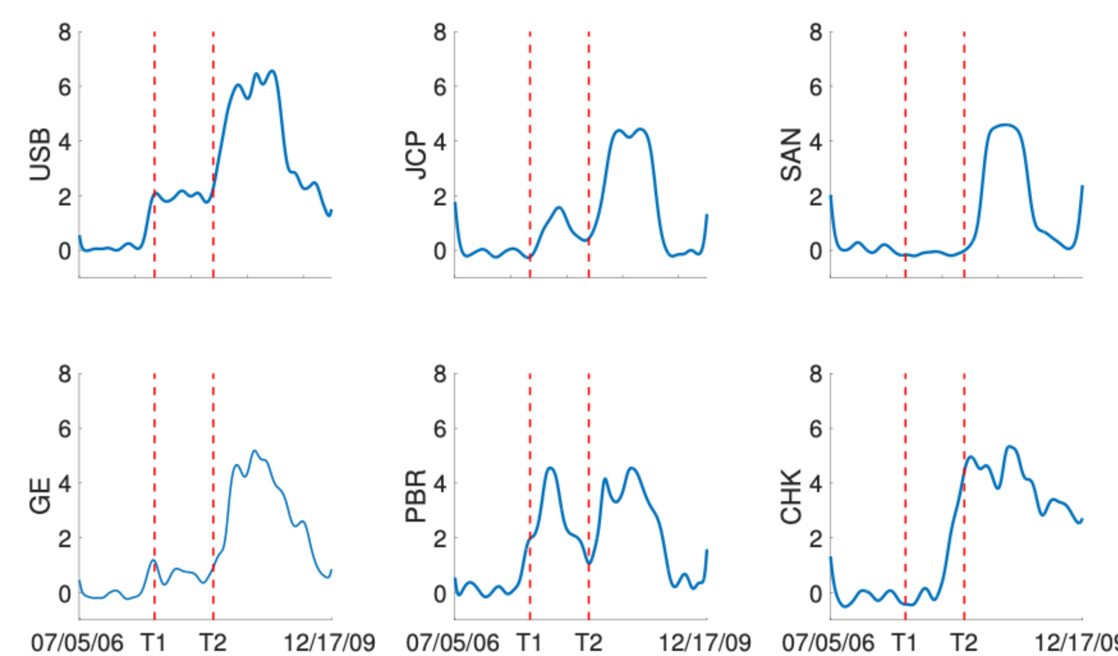
(Curve of TED spread; https://research.stlouisfed.org/fred2/series/TEDRATE)



T_1: 07/16/2007,
T_2: 06/30/2008,
T_3: 02/11/2009

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

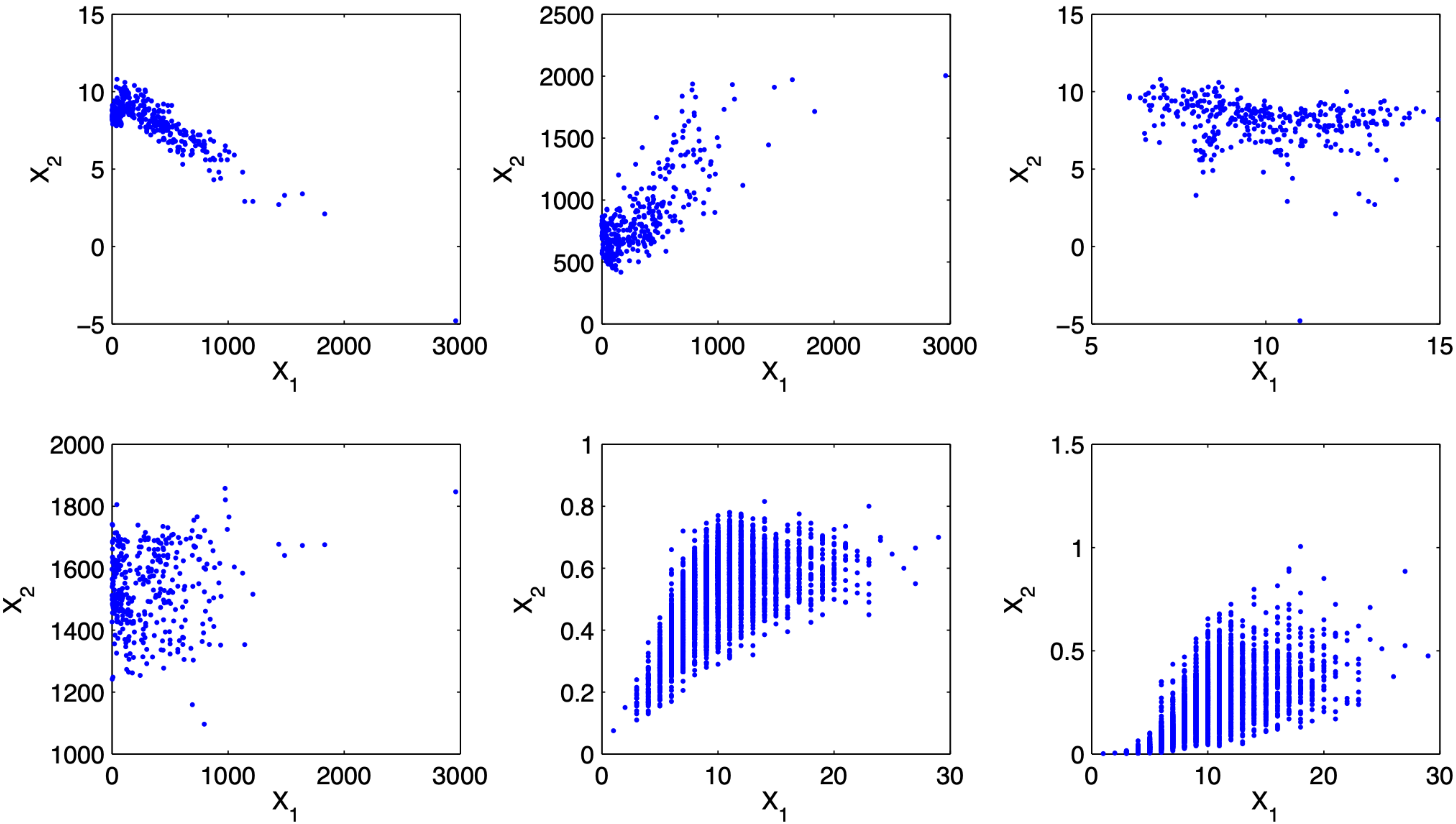
CD-NOD by causal-learn

```
from causallearn.search.ConstraintBased.CDNOD import cdnod

# default parameters
cg = cdnod(data)

# or customized parameters
cg = cdnod(data, c_indx, alpha, indep_test, stable, uc_rule, uc_priority, mvcdnod,
      correction_name, background_knowledge, verbose, show_progress)

# visualization using pydot
# note that the last node is the c_indx
cg.draw_pydot_graph()

# or save the graph
from causallearn.utils.GraphUtils import GraphUtils

pyd = GraphUtils.to_pydot(cg.G)
pyd.write_png('simple_test.png')
```

From MECs to DAGs (1)

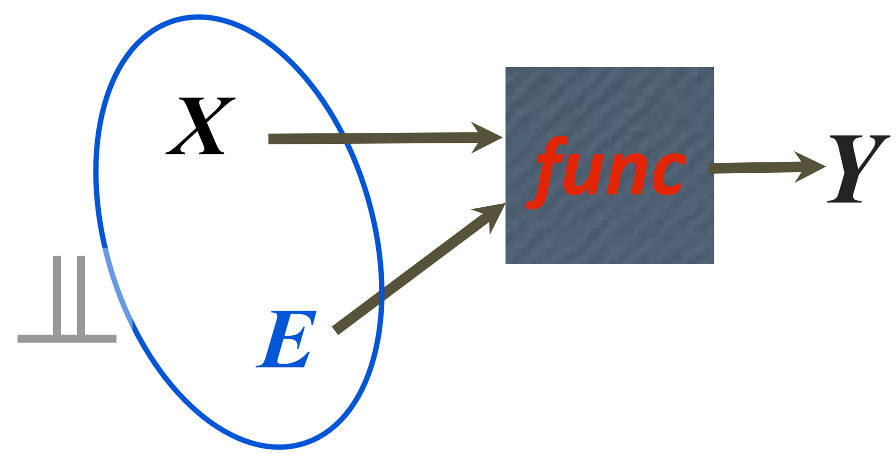
- Distinguishing cause from effect
- Linear, non-Gaussian, acyclic models

Distinguishing Cause from Effect:
Examples (Tübingen Cause-Effect Pairs)

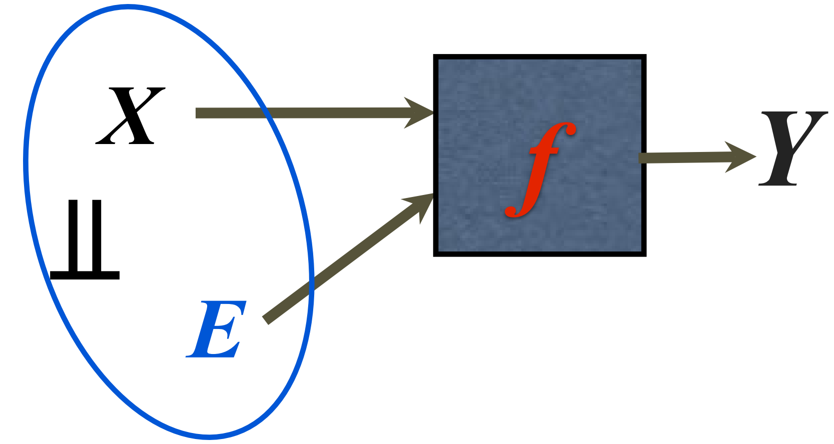


A Causal Process

rain → *wet ground*



Functional Causal Models



- Effect generated from cause with **independent noise** (Pearl et al.):

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

  - A way to encode the intuition "the generating process for *X* is 'independent' from that generates *Y* from *X*"

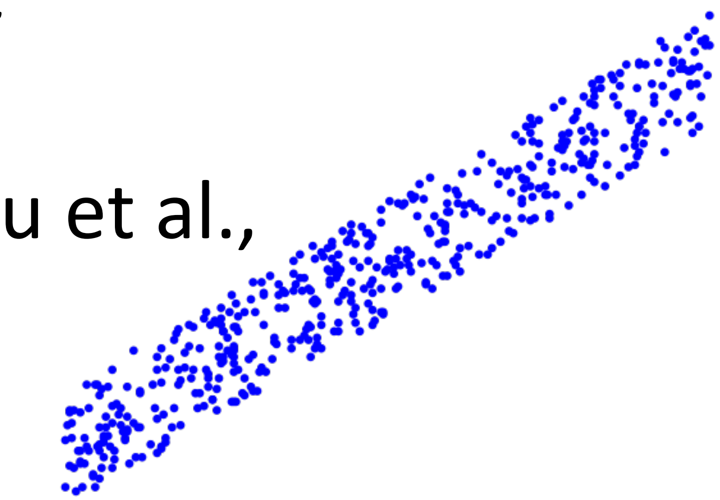
P(Y|X) ↘

P(X) →*X*→ *Y*

- :-(Without constraints on *f*, one can find independent noise for both directions (Darmois, 1951; Zhang et al., 2015)
  - Given any X_1 and X2, E' := conditional CDF of $X_2 \mid X_1$ is always independent from X_1 and $X2 = f(X_1, E')$
- :-) Structural constraints on *f* imply asymmetry

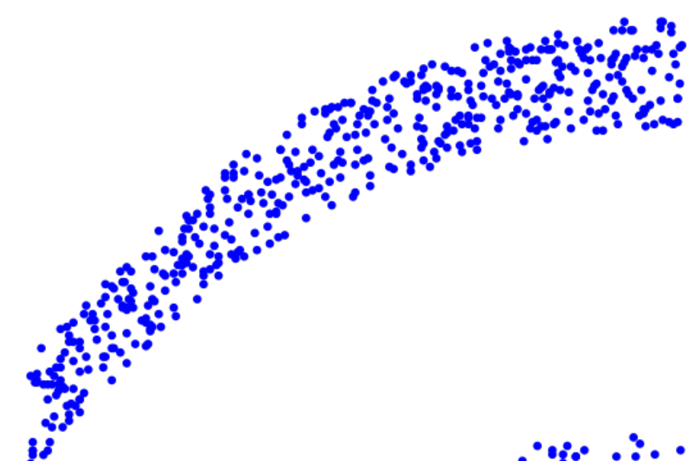
Functional Causal Model

- A **functional causal model** represents <u>effect</u> as a function of <u>direct causes</u> and noise: $Y = f(X, E)$, with $X \perp\!\!\!\perp E$



- Linear non-Gaussian acyclic causal model (Shimizu et al., '06)

$$Y = \mathrm{a} \cdot X + E$$



- Additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

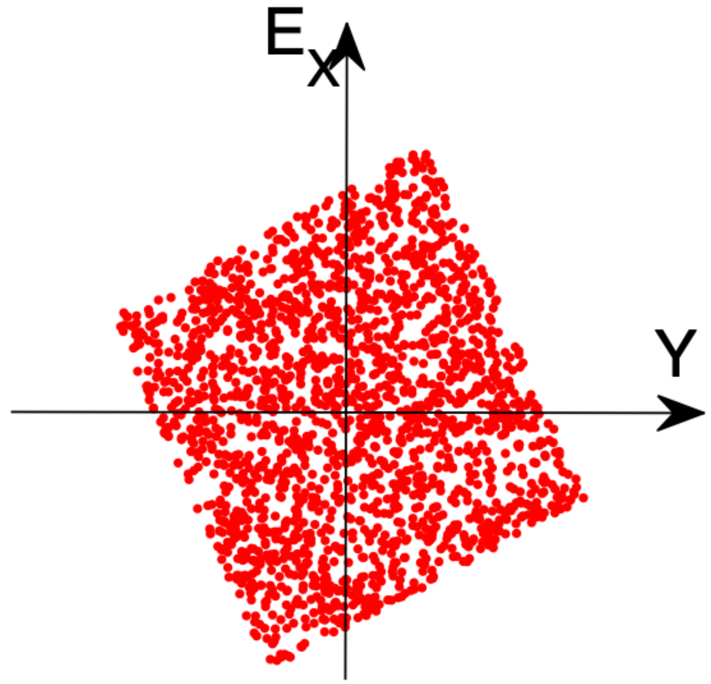
$$Y = f(X) + E$$

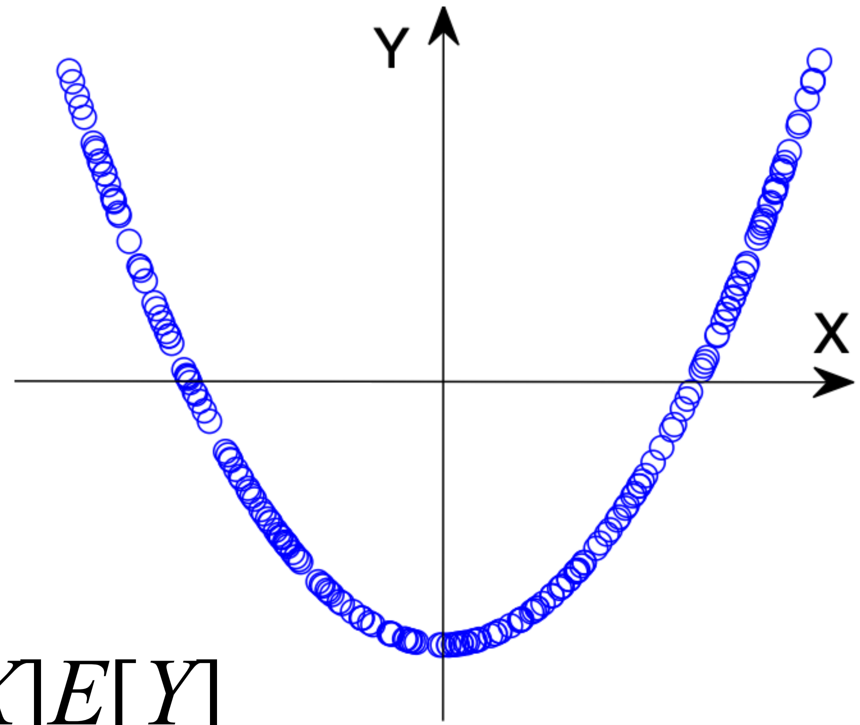
- Post-nonlinear causal model (Zhang & Chan, '06; Zhang & Hyvärinen, '09a)

$$Y = f_2 \left(f_1(X) + E \right)$$

(Conditional) Independence

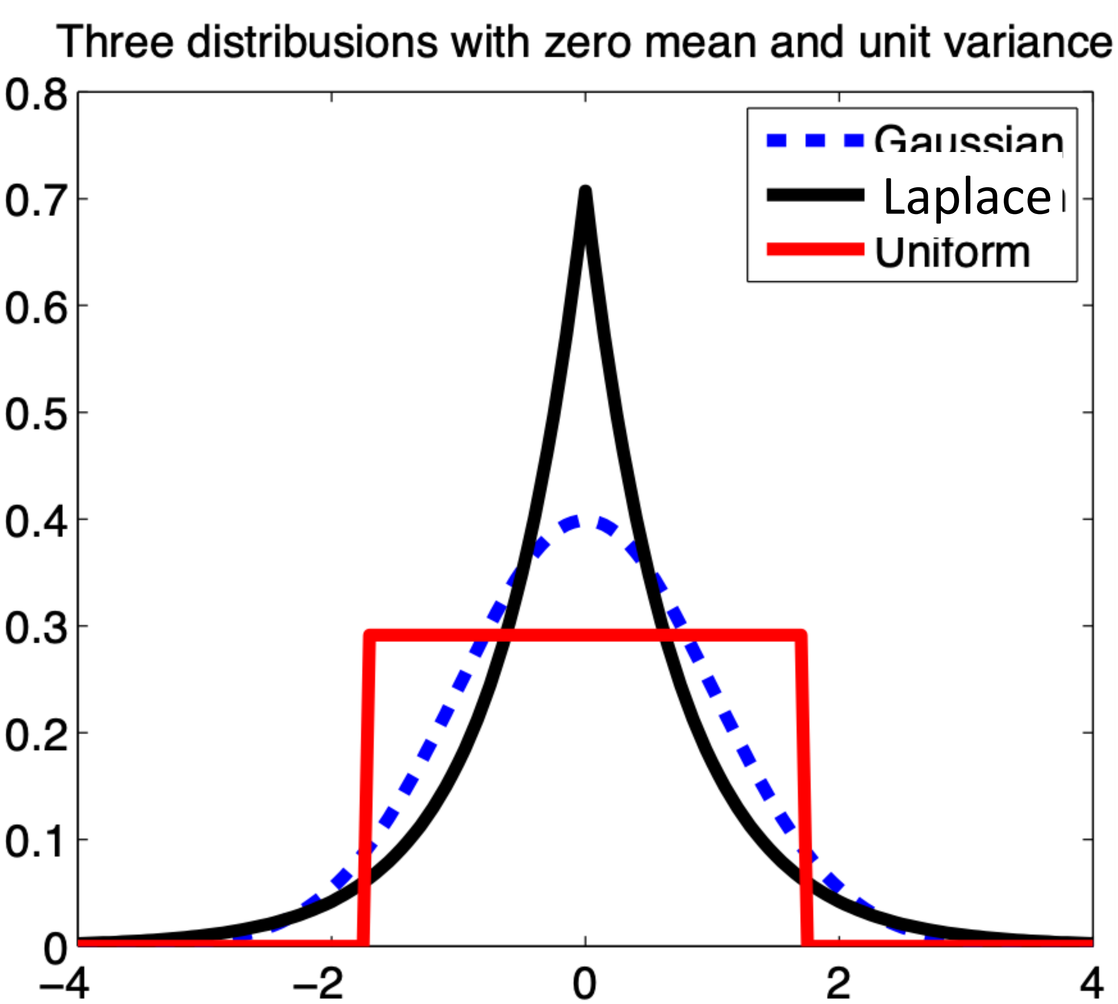
- $X \perp\!\!\!\perp Y$ iff $p(X,Y) = p(X)p(Y)$
    - or $p(X|Y) = P(X)$: Y not informative to X
- $X \perp\!\!\!\perp Y \mid Z$ iff $p(X,Y|Z) = p(X|Z)p(Y|Z)$
    - or, $p(X|Y,Z) = p(X|Z)$: given Z, Y not informative to X
- Divide & conquer, remove irrelevant info...
- By construction, regression residual is uncorrelated (but not necessarily independent !) from the predictor

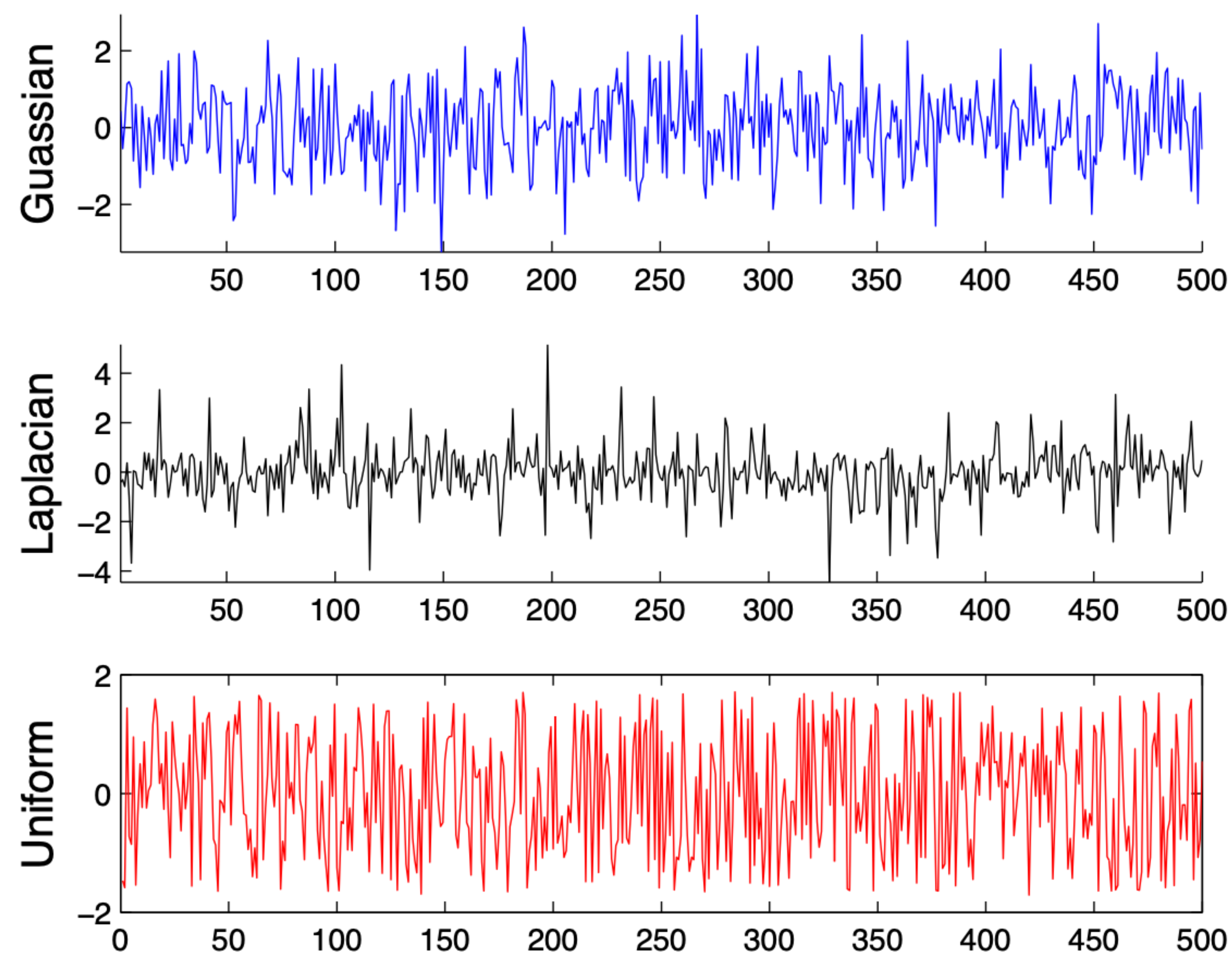




Uncorrelatedness: $E[XY] = E[X]E[Y]$

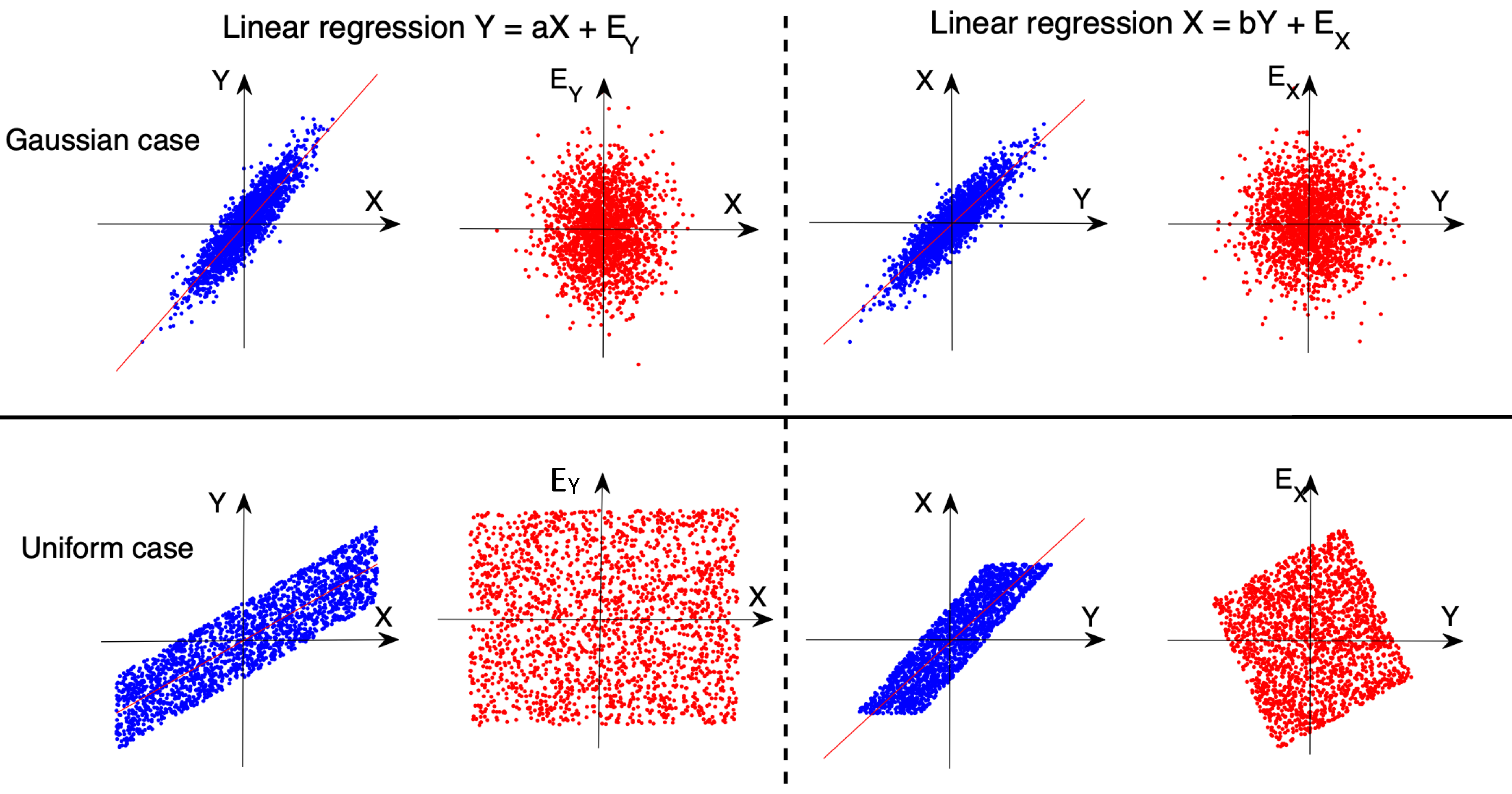
Gaussian vs. Non-Gaussian Distributions





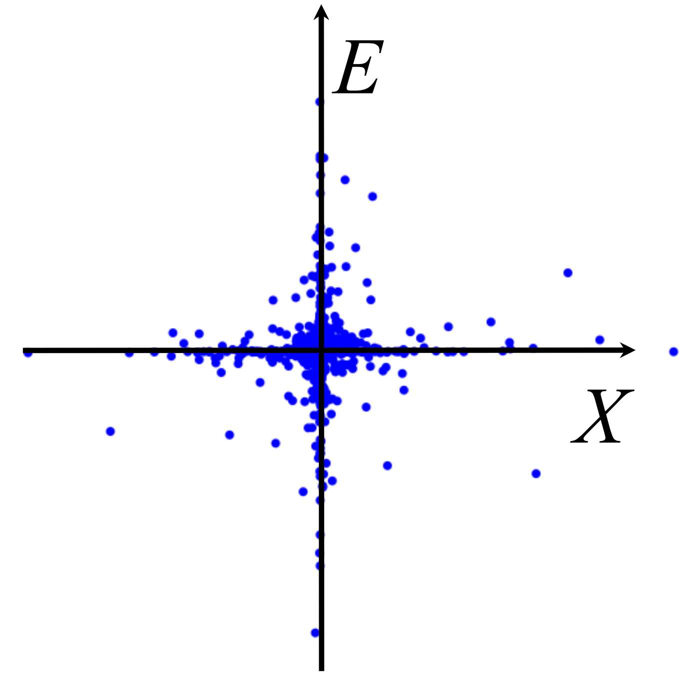
Causal Asymmetry the Linear Case: Illustration

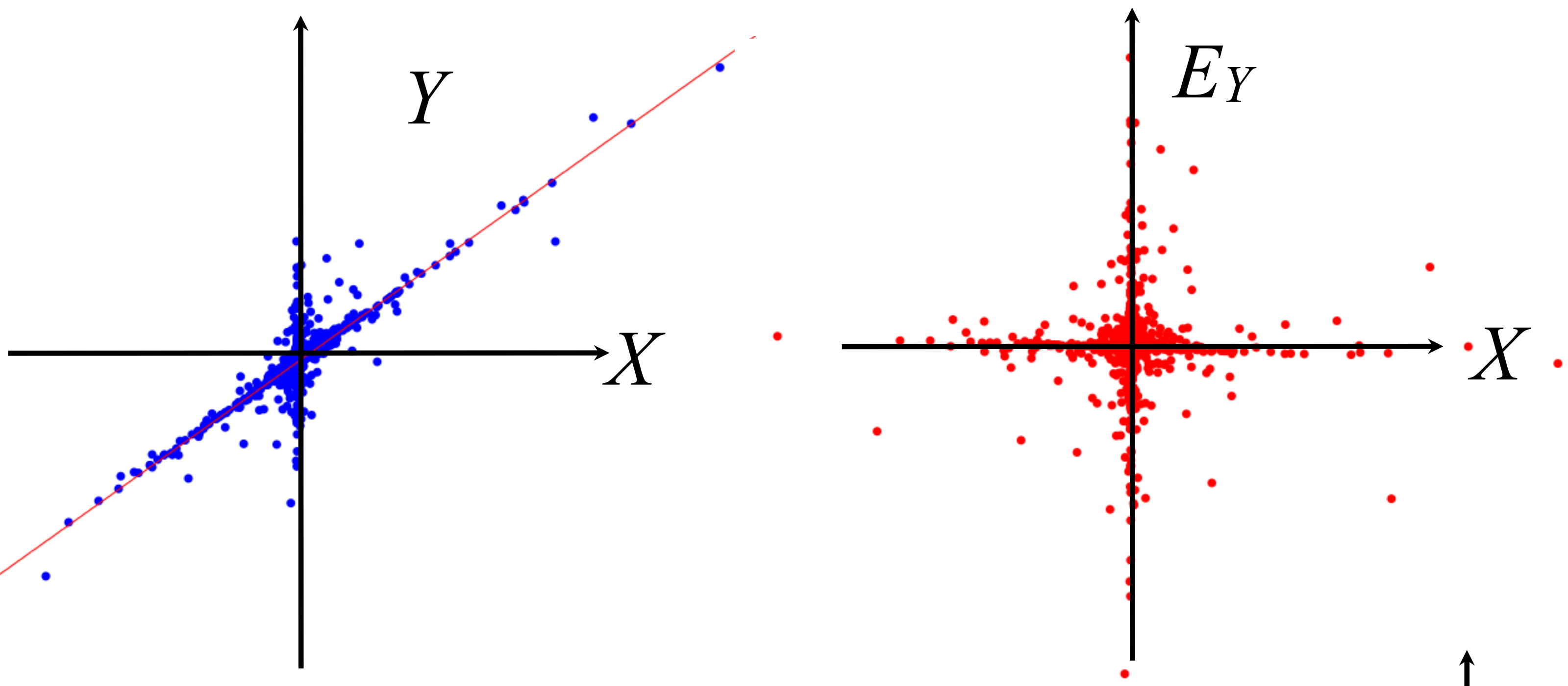
Data generated by $Y = aX + E$ (i.e., $X \to Y$):

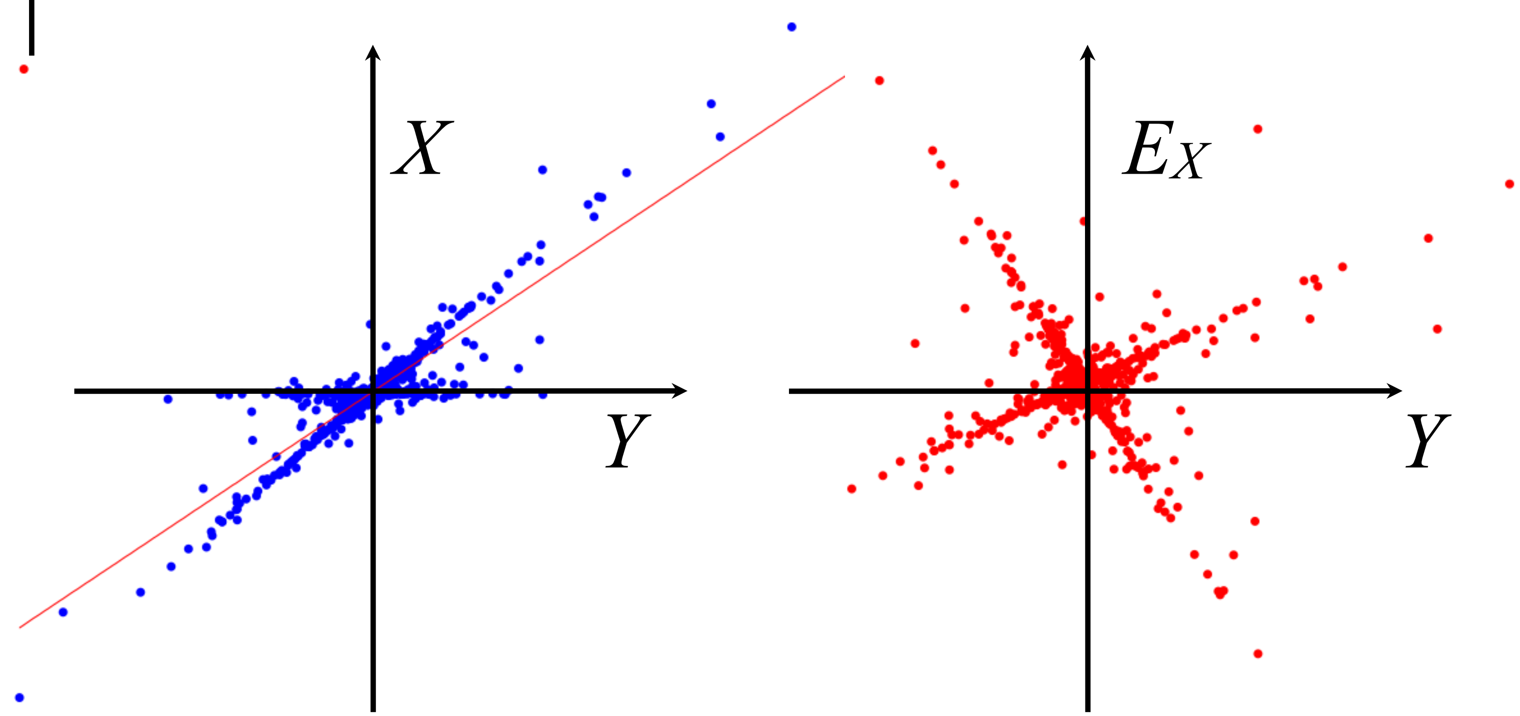


Super-Gaussian Case

Data generated by $Y = aX + E$ $(X \rightarrow Y)$:







More Generally, LiNGAM Model

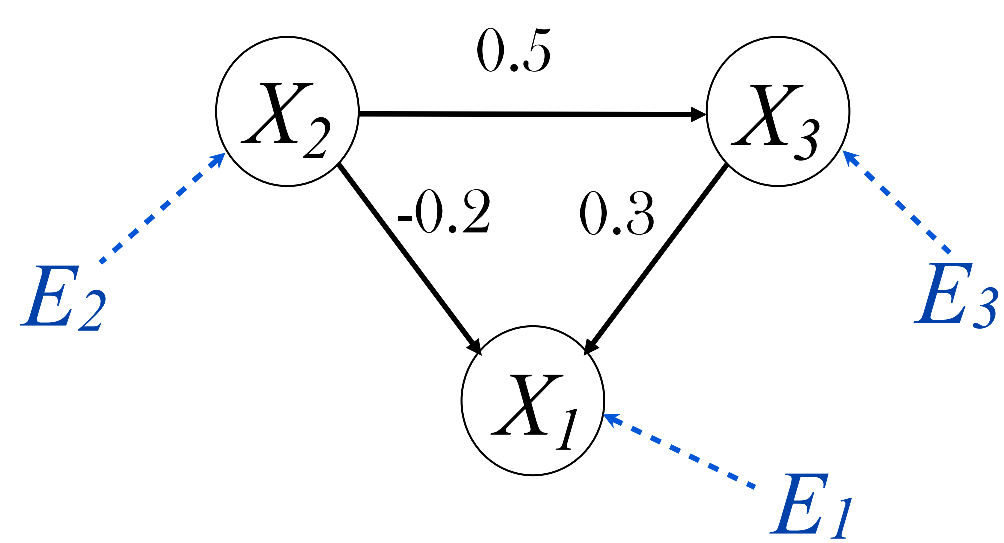
- Linear, non-Gaussian, acyclic causal model (LiNGAM) (Shimizu et al., 2006):

$$X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E}$$

  - Disturbances (errors) E_i are non-Gaussian (or at most one is Gaussian) and mutually independent

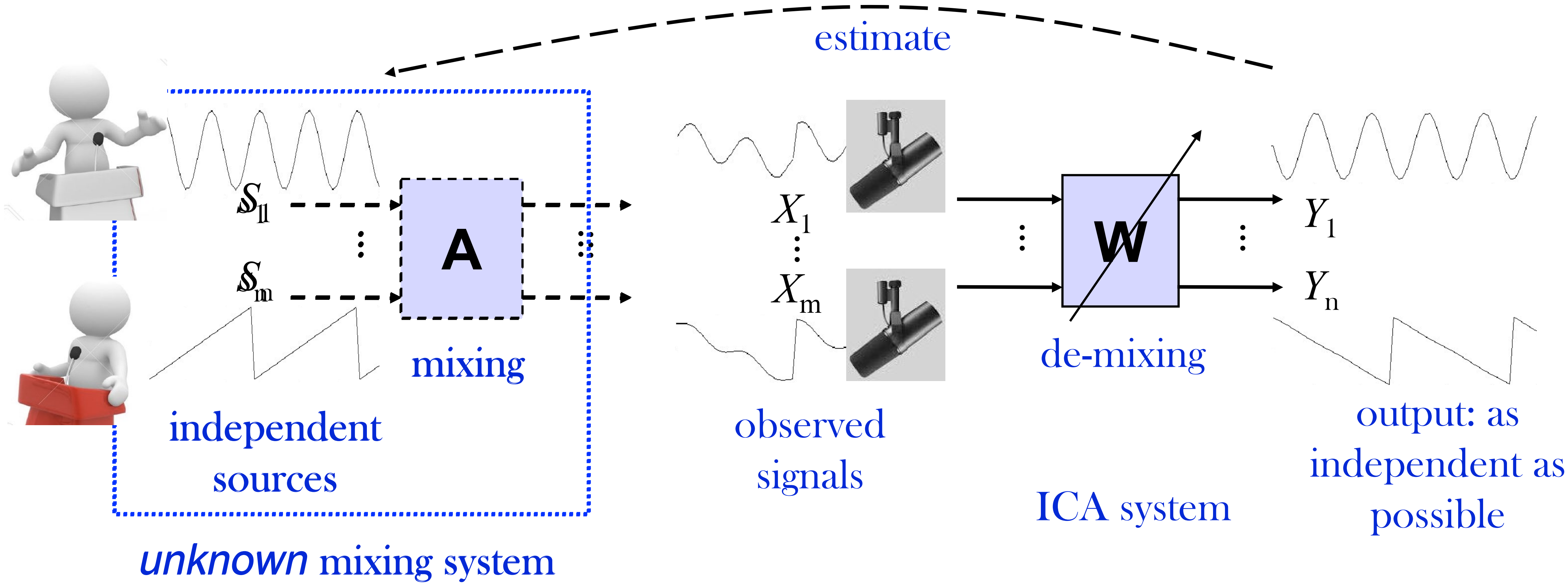
  - Example:

$$\begin{aligned} X_2 &= E_2, \\ X_3 &= 0.5 X_2 + E_3, \\ X_1 &= -0.2 X_2 + 0.3 X_3 + E_1. \end{aligned}$$



Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.

Independent Component Analysis



$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$$

$$\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$$

$$\begin{array}{c} X_1 \\ X_2 \end{array} \begin{bmatrix} .5 & .3 & 1.1 & -0.3 & \dots \\ .8 & -.7 & .3 & .5 & \dots \end{bmatrix} = \overset{A}{\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}} \cdot \begin{bmatrix} ? & ? & ? & ? & \dots \\ ? & ? & ? & ? & \dots \end{bmatrix} \begin{array}{c} s_1 \\ s_2 \end{array}$$

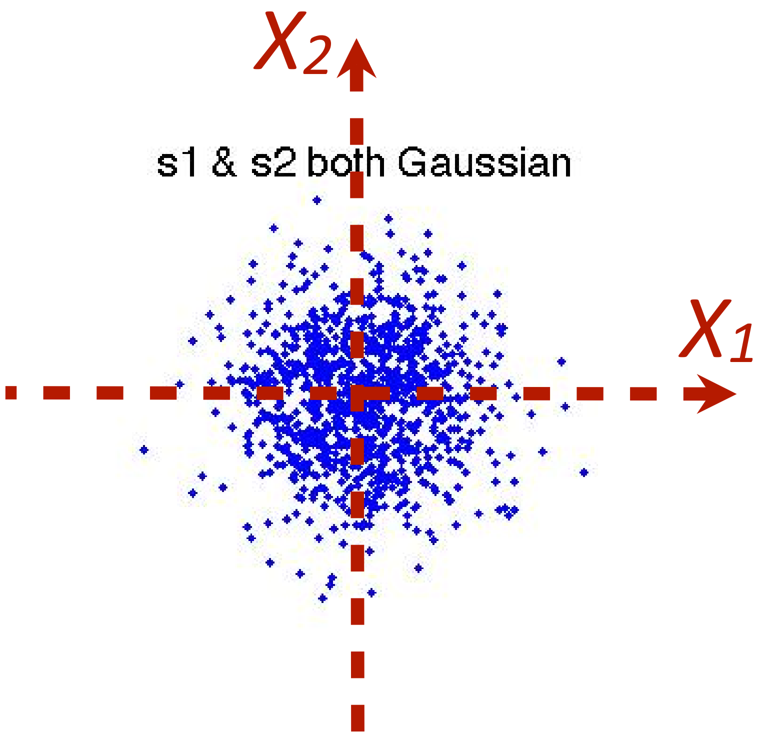
- Assumptions in ICA
  - At most one of S_i is Gaussian
  - #Source <= # Sensor, and **A** is of full column rank

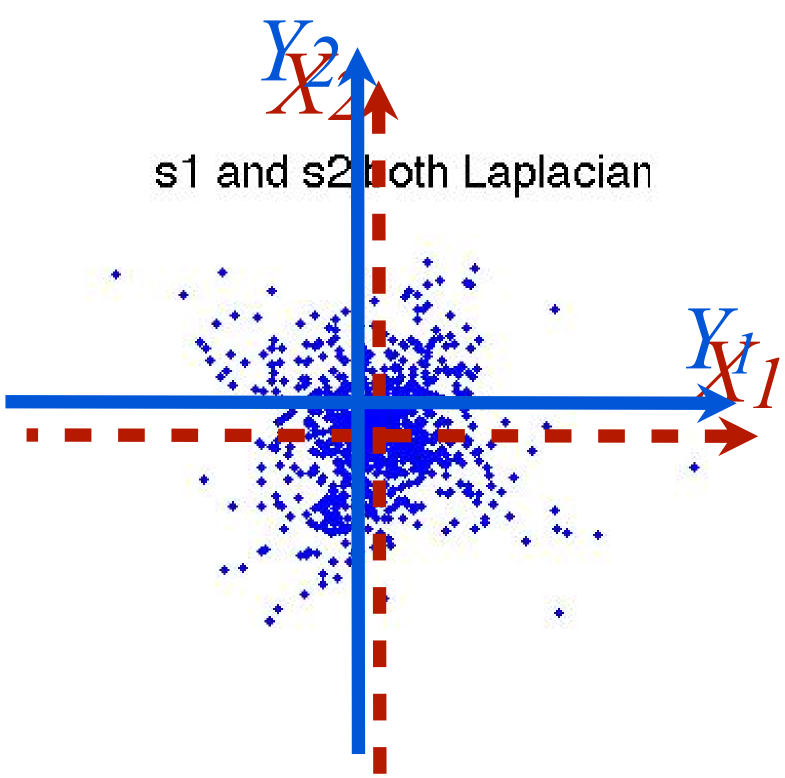
Then A can be estimated up to column scale and permutation indeterminacies

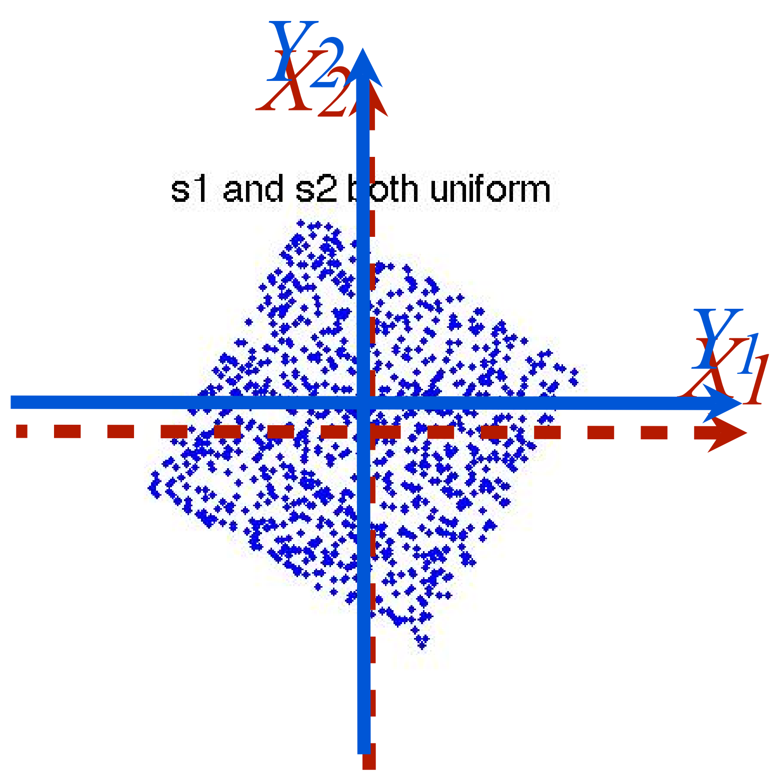
Hyvärinen et al., Independent Component Analysis, 2001

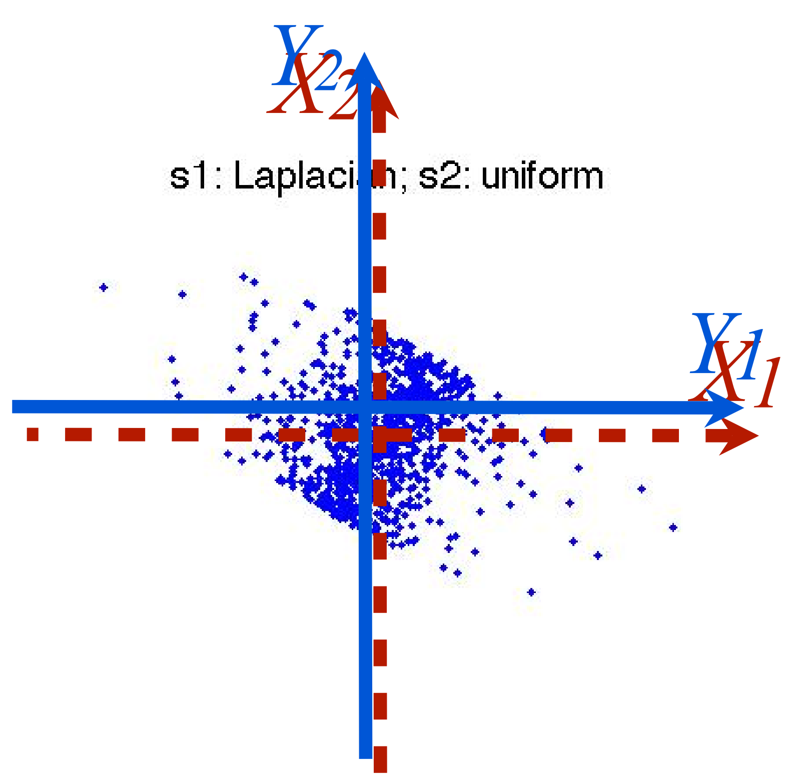
Intuition: Why ICA works?

- (After preprocessing) ICA aims to find a rotation transformation **Y = W·X** to making Y_i independent
  - By maximum likelihood log $p(X|\boldsymbol{A})$, mutual information $MI(Y_1,\ldots,Y_m)$ minimization, infomax...

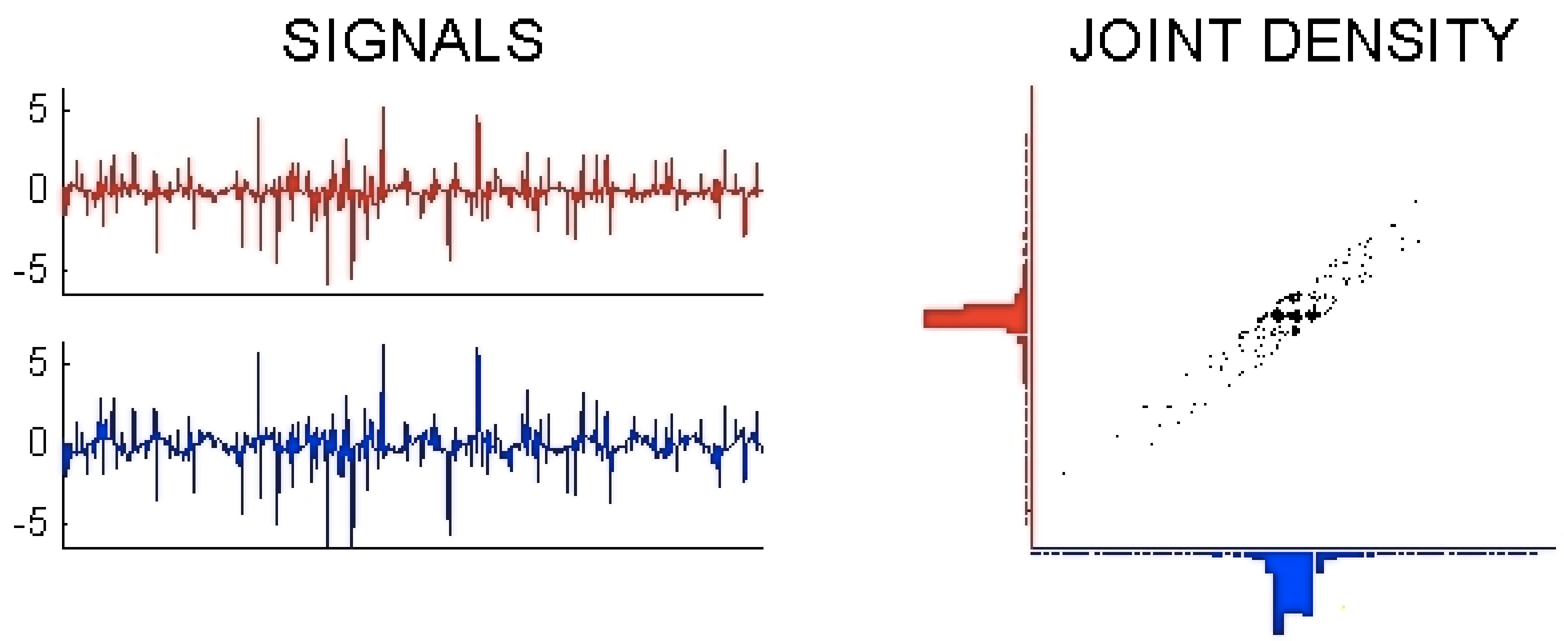




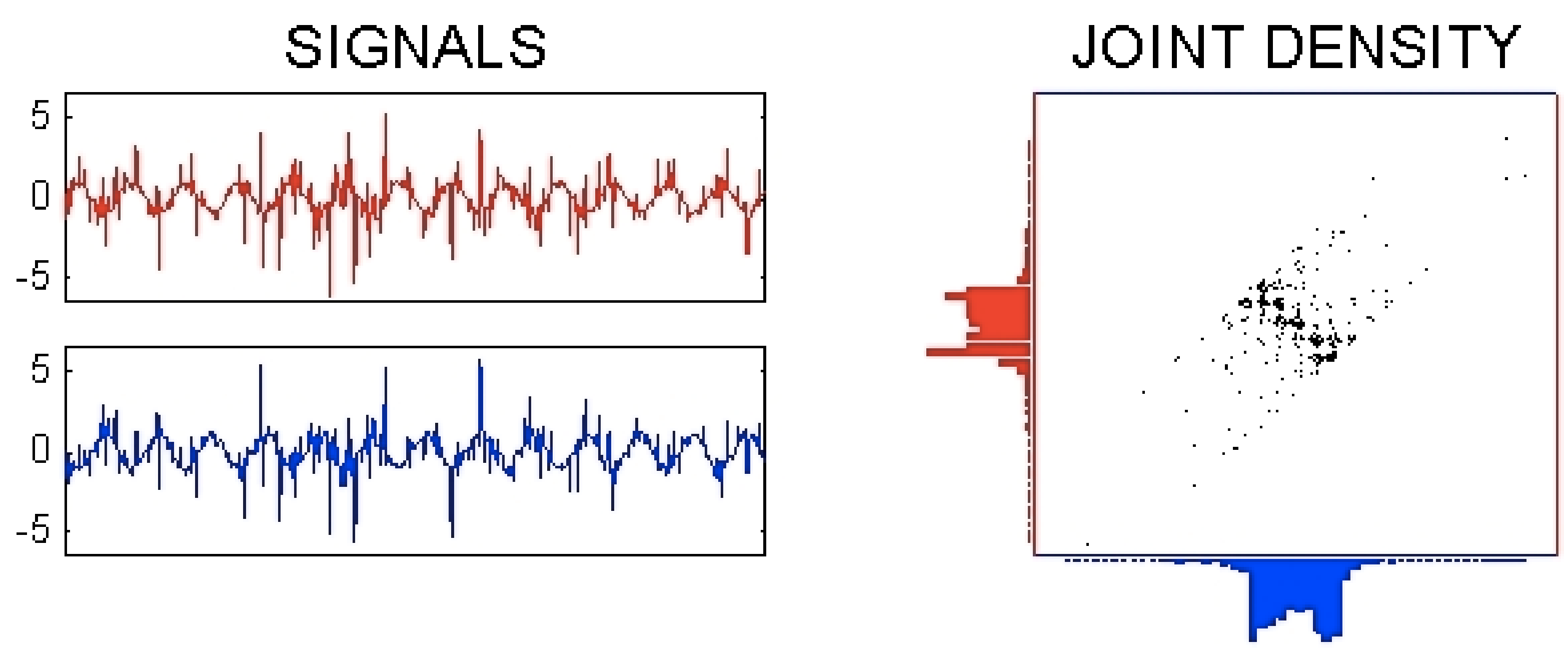




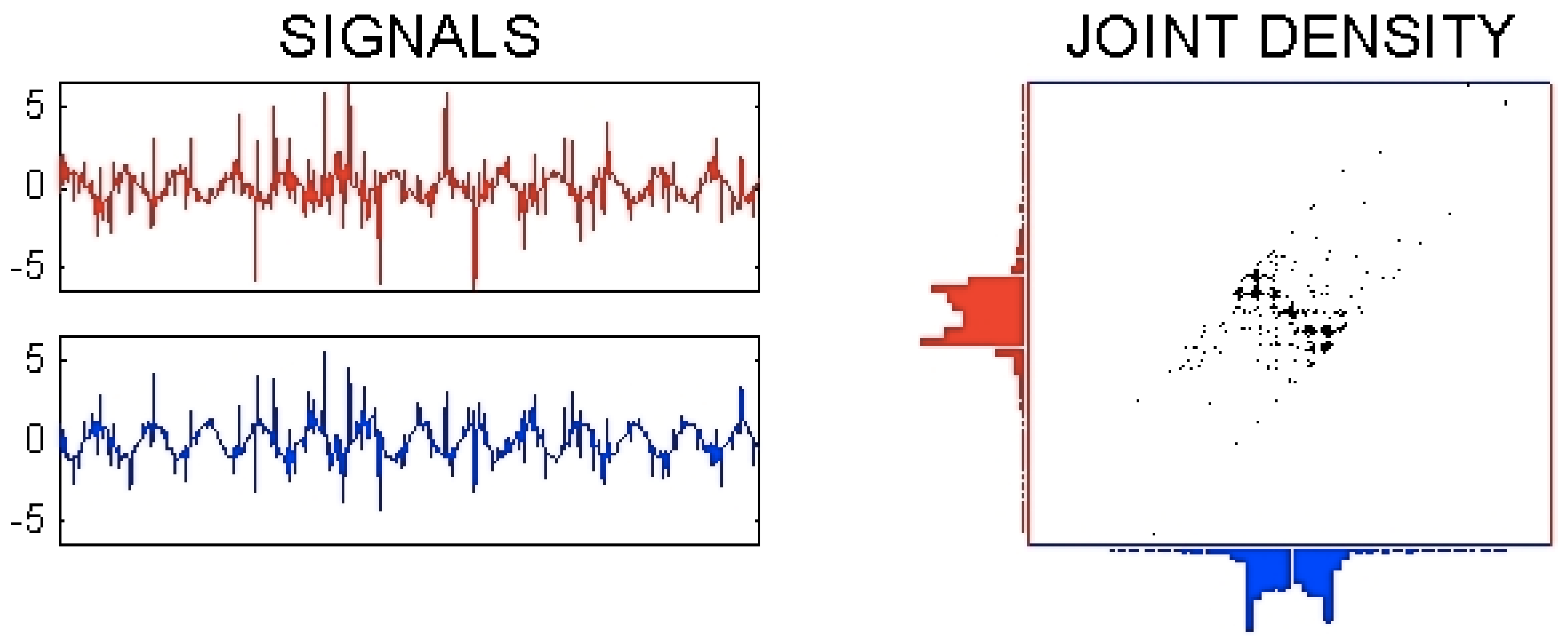
A Demo of the ICA Procedure



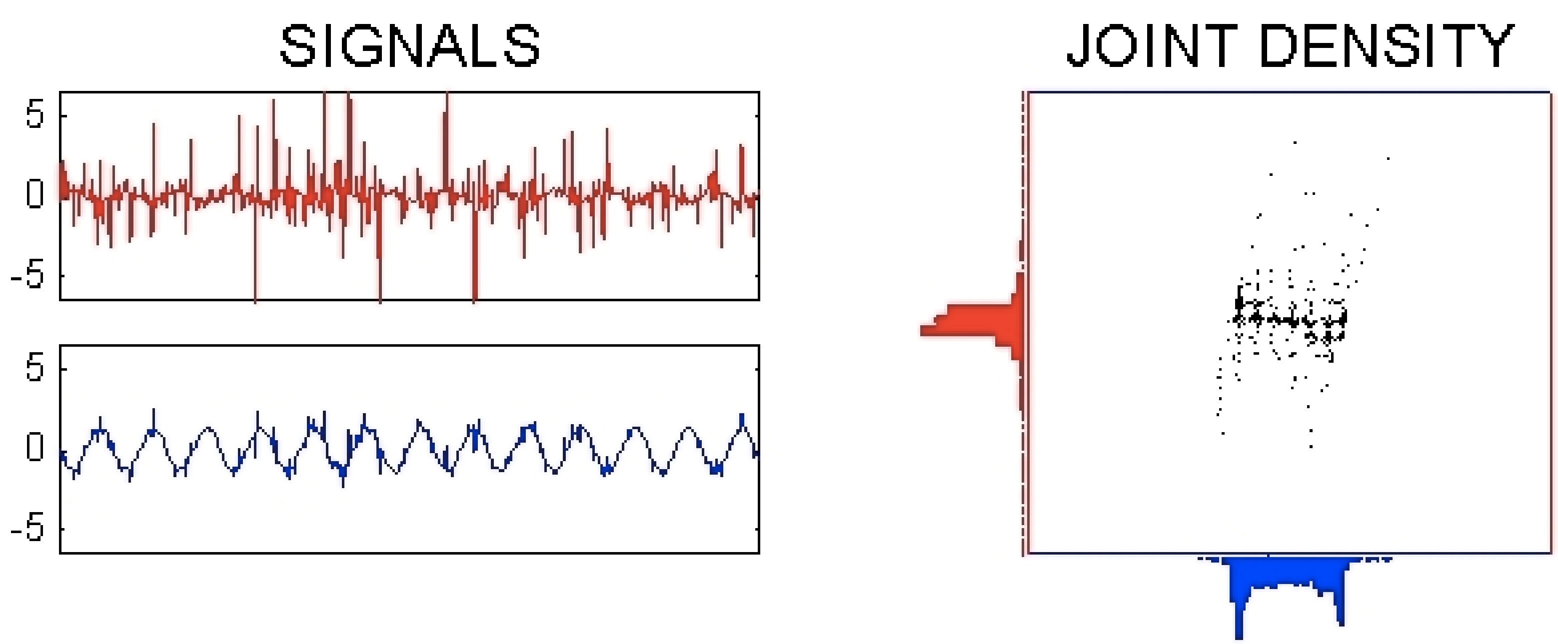
Input signals and density



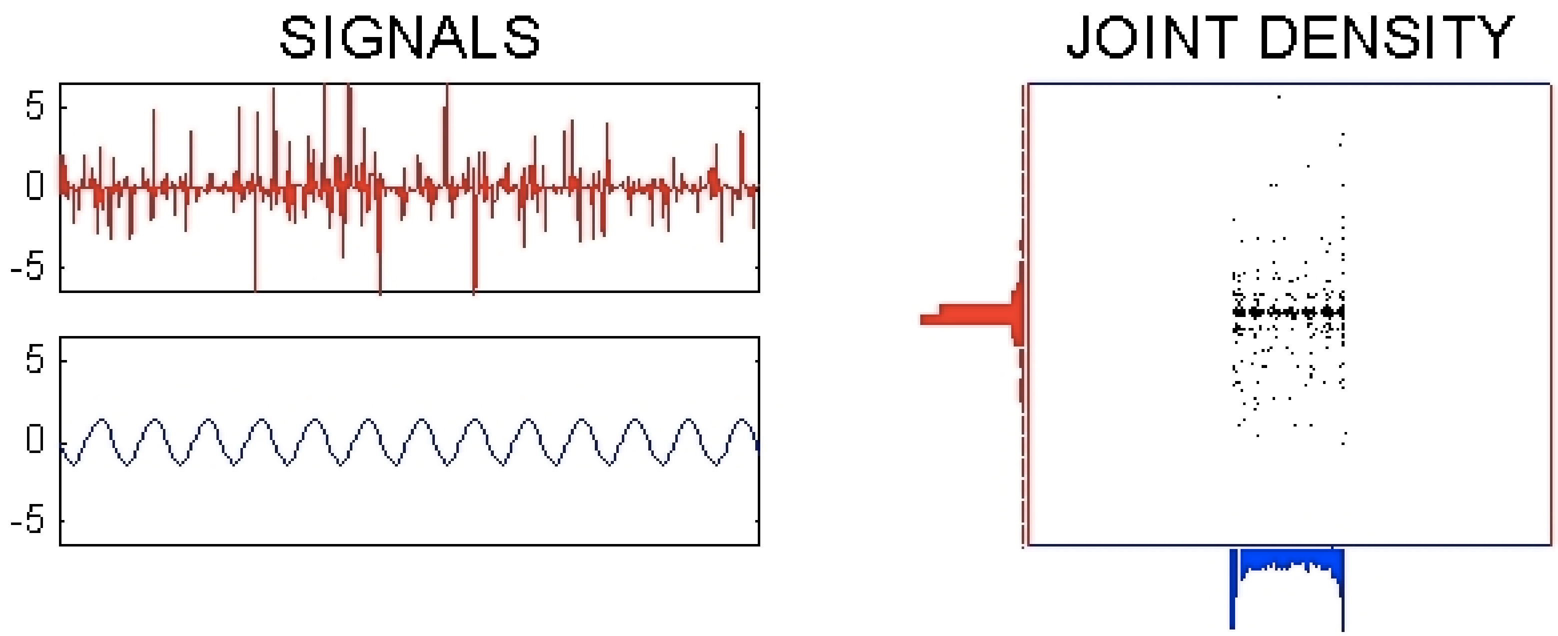
Whitened signals and density



Separated signals after 1 step of FastICA



Separated signals after 3 steps of FastICA



Separated signals after 5 steps of FastICA

LiNGAM Analysis by ICA

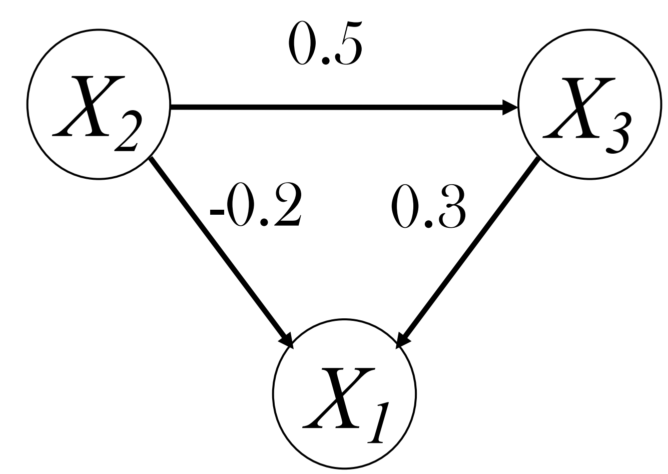
- LiNGAM: $X_i = \sum_{j:\ \text{parents of } i} b_{ij} X_j + E_i \quad or \quad \mathbf{X} = \mathbf{B}\mathbf{X} + \mathbf{E} \Rightarrow$ **E = (I-B)X**
  - **B** has special structure: acyclic relations
- ICA: **Y = WX**
- **B** can be seen from **W** by permutation and re-scaling
- Faithfulness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

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

W

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

So we have the causal relation:



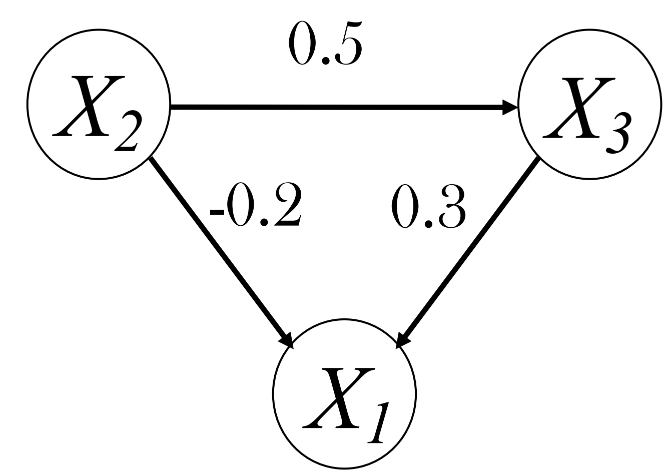
LiNGAM Analysis by ICA

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

$$\begin{bmatrix} E_1 \\ E_3 \\ E_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.2 & -0.3 & 1 \end{bmatrix}}_{\mathbf{W}} \cdot \begin{bmatrix} X_2 \\ X_3 \\ X_1 \end{bmatrix}$$

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

So we have the causal relation:



Question 1. How to find W?

Question 2. How to see B from W?

LiNGAM Analysis by ICA

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

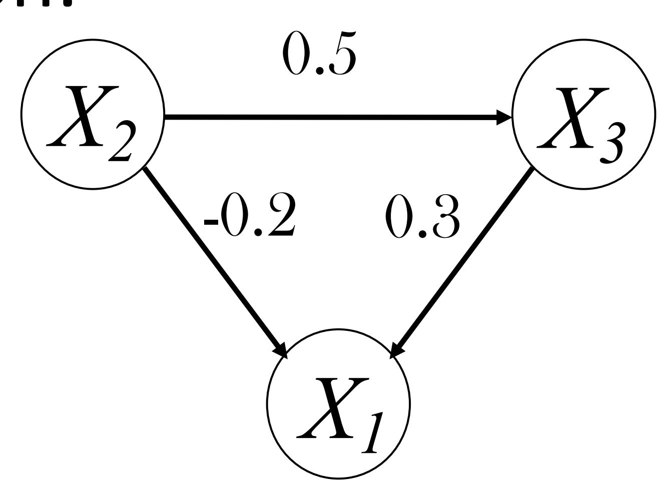
$$\Leftrightarrow \begin{cases} X_2 = E_1 \\ X_3 = 0.5X_2 + E_3 \\ X_1 = -0.2X_2 + 0.3X_3 + E_2 \end{cases}$$

1. First permute the rows of $\mathbf{W}$ 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 **Y = WX** and

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

- Can we find the causal model?

1. First permute the rows of $\mathbf{W}$ 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. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

$$\ddot{W} = \begin{pmatrix} 1.5 & 0 & 0 & 0 \\ 1.5 & 3 & 0 & 0 \\ 0.6 & -0.4 & 2 & 0 \\ 0 & 0.2 & 0 & 0.5 \end{pmatrix},$$

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

- ICA gives **Y = WX** 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 $\mathbf{W}$ 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. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

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

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

- ICA gives **Y = WX** 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 $\mathbf{W}$ 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. $\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

- Can we find the causal model?

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

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

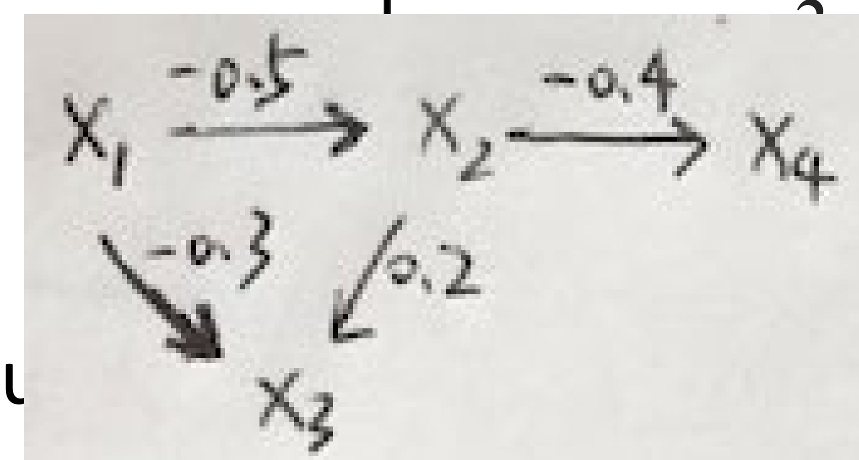
- ICA gives **Y = WX** and

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

1. First permute the rows of **W** to make all diagonal entries non-zero, yielding $\ddot{\mathbf{W}}$.

Then divide each row of $\ddot{\mathbf{W}}$ its diagonal entry, giving $\ddot{\mathbf{W}}'$.

$\hat{\mathbf{B}} = \mathbf{I} - \ddot{\mathbf{W}}'$.

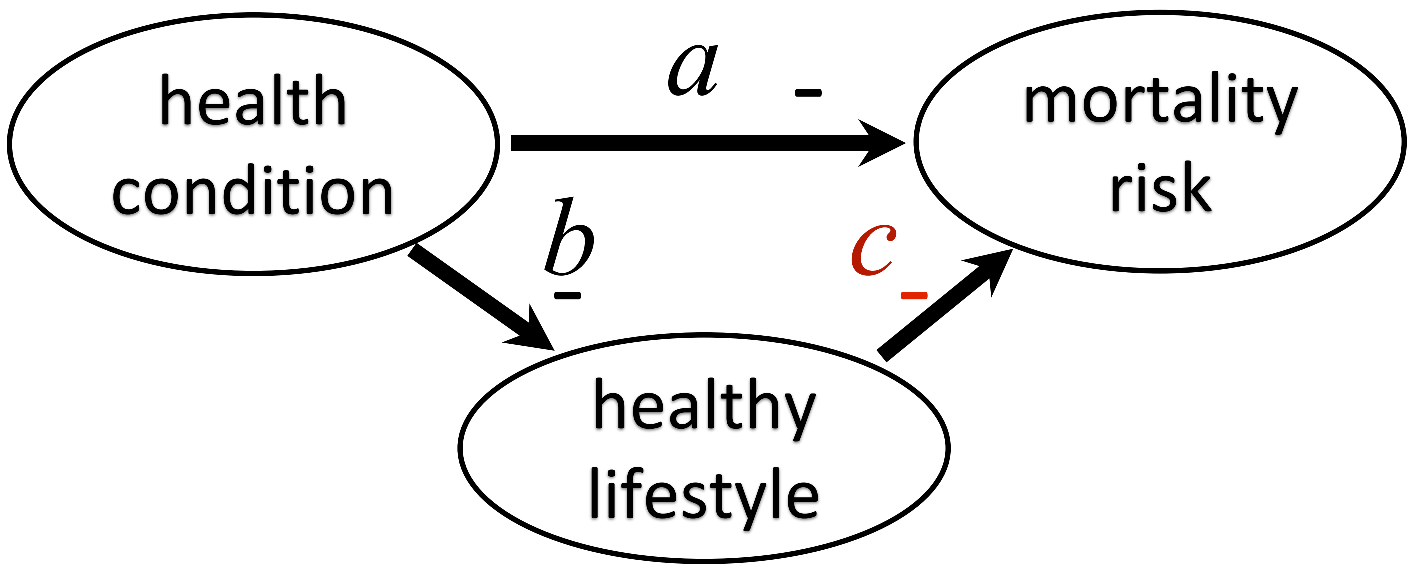


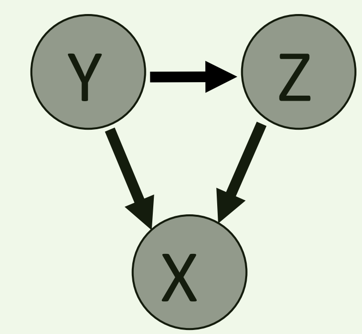
- Can we find the cau

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

Faithfulness Assumption Needed?

- One might find independence between health condition & risk of mortality. Why?





Possible to have $Y \perp\!\!\!\perp Z \mid X$?

- E.g., if $a = -bc$, then *health_condition* ⫫ *mortality_risk*, which cannot by seen from the graph!

- No faithfulness assumption is needed in LiNGAM
    - Minimality (a zero coefficient corresponds to edge absence) is sufficient

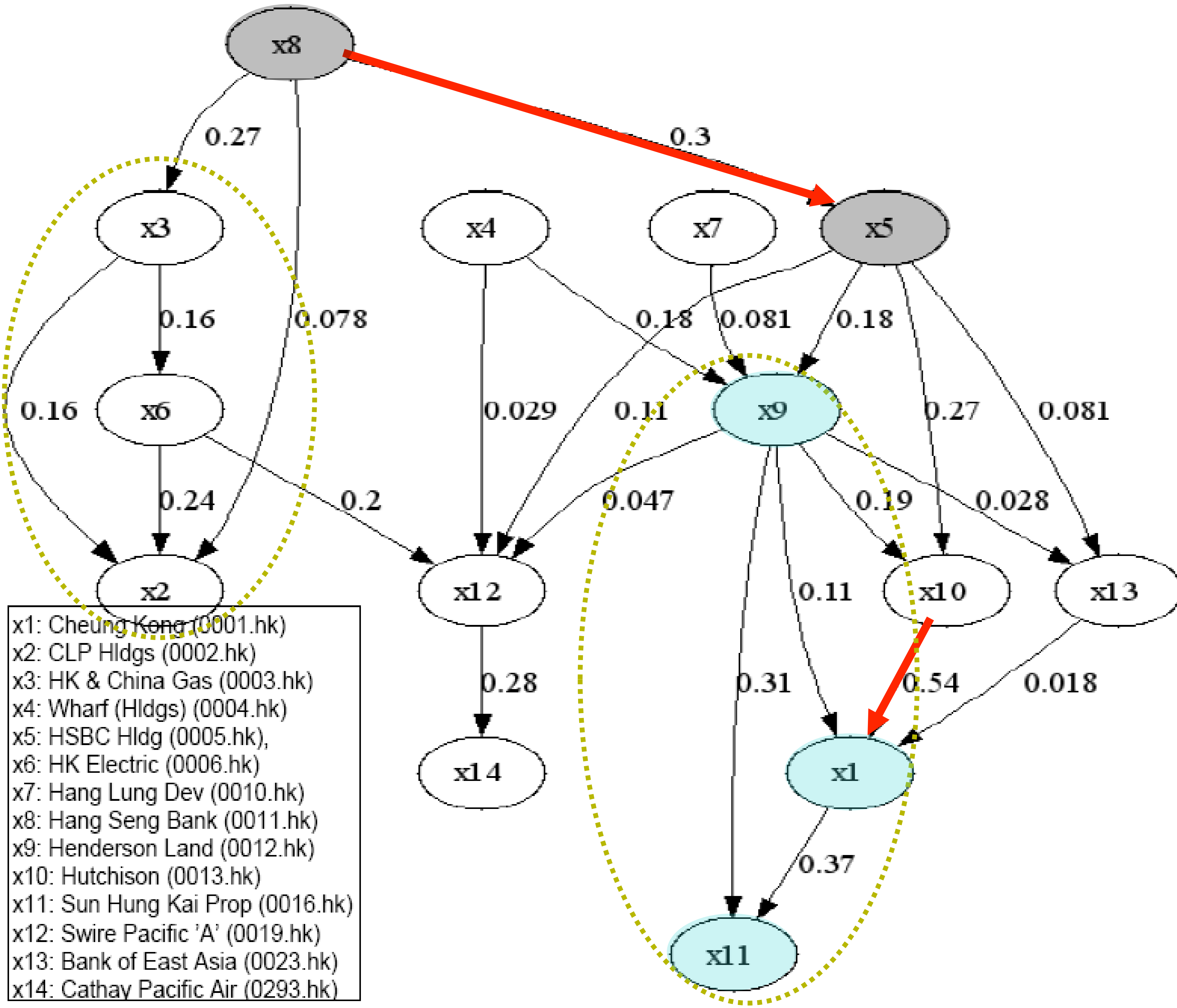
*

Some Estimation Methods for LiNGAM

- ICA-LiNGAM
- ICA with Sparse Connections
- DirectLiNGAM…

Shimizu et al. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030.
Zhang et al. (2006) ICA with sparse connections: Revisited. Lecture Notes in Computer Science, 5441:195–202, 2009
Shimizu, et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12:1225–1248.

Application: Causal diagram in HK Stock Market (Zhang & Chan, 2006)



1. Ownership relation: x5 owns 60% of x8; x1 holds 50% of x10.
2. Stocks belonging to the same subindex tend to be connected.
3. Large bank companies (x5 and x8) are the cause of many stocks.
4. Stocks in Property Index (x1, x9, x11) depend on many stocks, while they hardly influence others.

LiNGAM-based methods by causal-learn

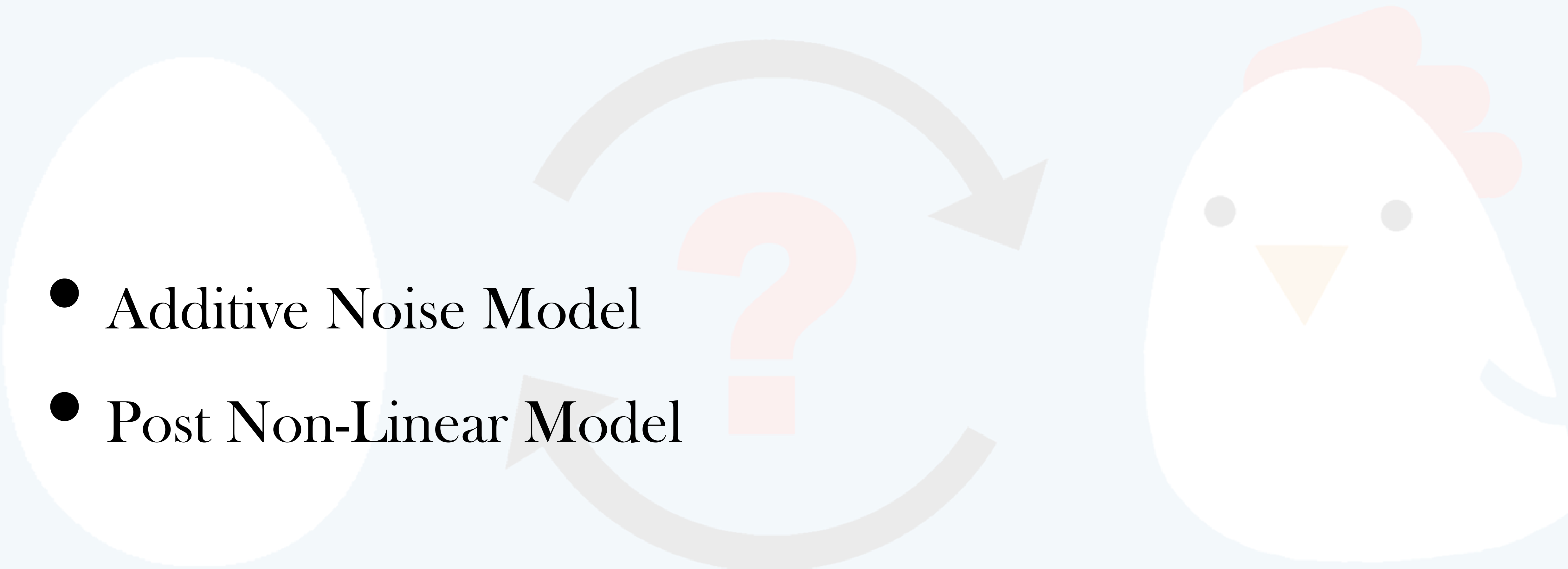
- ICA-based LiNGAM: Linear Non-Gaussian
- DirectLiNGAM: Linear Non-Gaussian
- VAR-LiNGAM:Time series
- RCD: Hidden confounders
- CAM-UV: Nonlinear additive noise

LiNGAM-based methods by causal-learn

```
from causallearn.search.FCMBased import lingam
model = lingam.ICALiNGAM(random_state, max_iter)
model.fit(X)

print(model.causal_order_)
print(model.adjacency_matrix_)
```

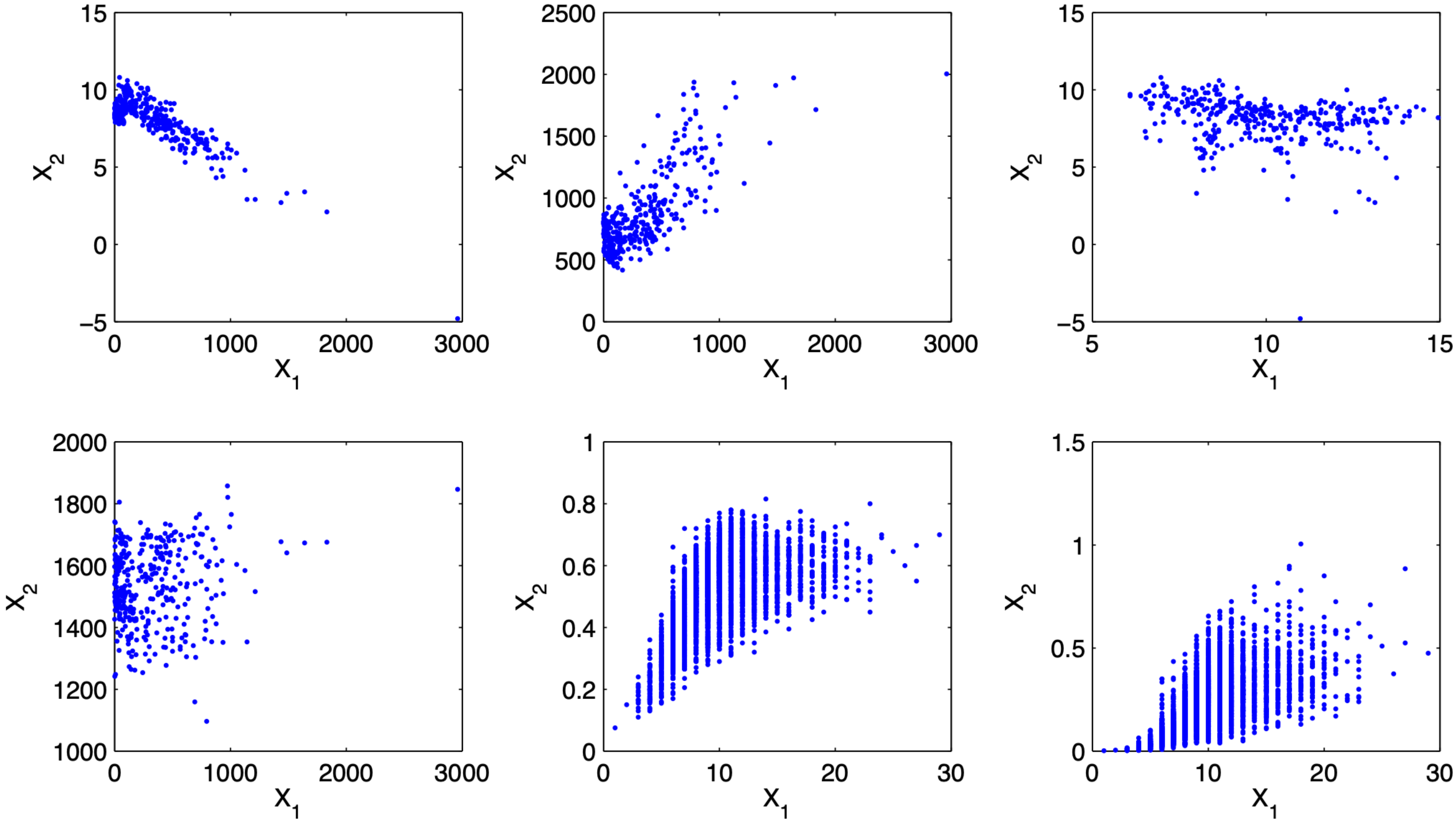
- We have seen the linear non-Gaussian case.
- How about nonlinearity?

From MECs to DAGs (2)

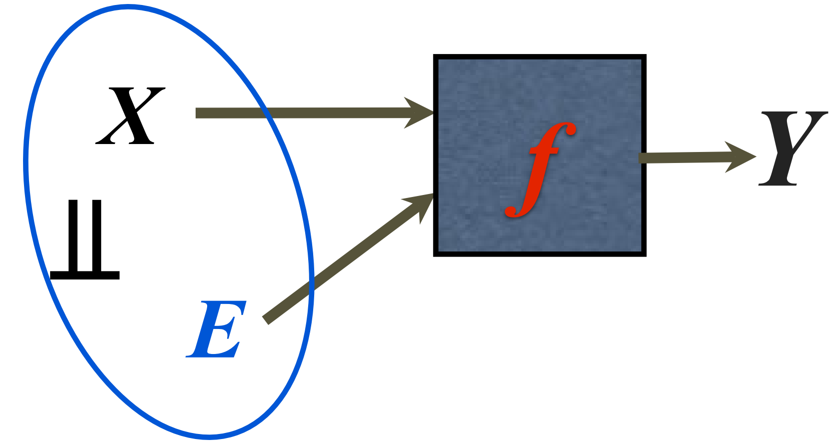


- Additive Noise Model
- Post Non-Linear Model

Some Real Data Sets



Functional Causal Models



- Effect generated from cause with **independent noise** (Pearl et al.):

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

  - A way to encode the intuition "the generating process for *X* is 'independent' from that generates *Y* from *X*"

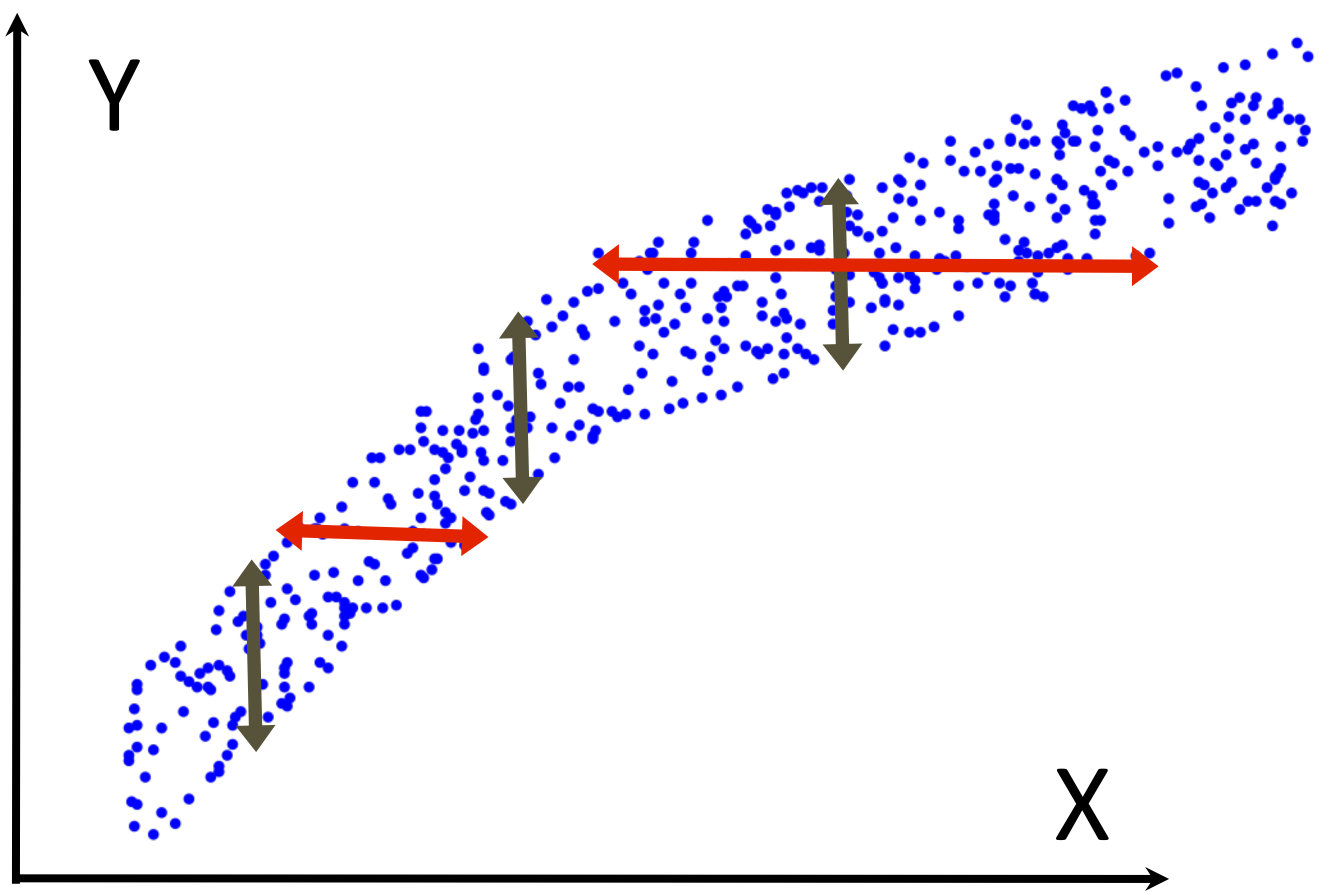
$P(Y|X)$

$P(X) \rightarrow X \rightarrow Y$

- :-(Without constraints on *f*, one can find independent noise for both directions (Darmois, 1951; Zhang et al., 2015)
  - Given any X_1 and X2, E' := conditional CDF of $X_2 \mid X_1$ is always independent from X_1 and $X2 = f(X_1, E')$
- :-) Structural constraints on *f* imply asymmetry

Causal Asymmetry with Nonlinear Additive Noise: Illustration

$$Y = f(X) + E \text{ with } E \perp\!\!\!\perp X$$



(Hoyer et al., 2009)

Additive Noise Models by causal-learn

```
from causallearn.search.FCMBased.ANM.ANM import ANM
anm = ANM()
p_value_foward, p_value_backward = anm.cause_or_effect(data_x, data_y)
```

Parameters

data_x: input data (n, 1).

data_y: output data (n, 1).

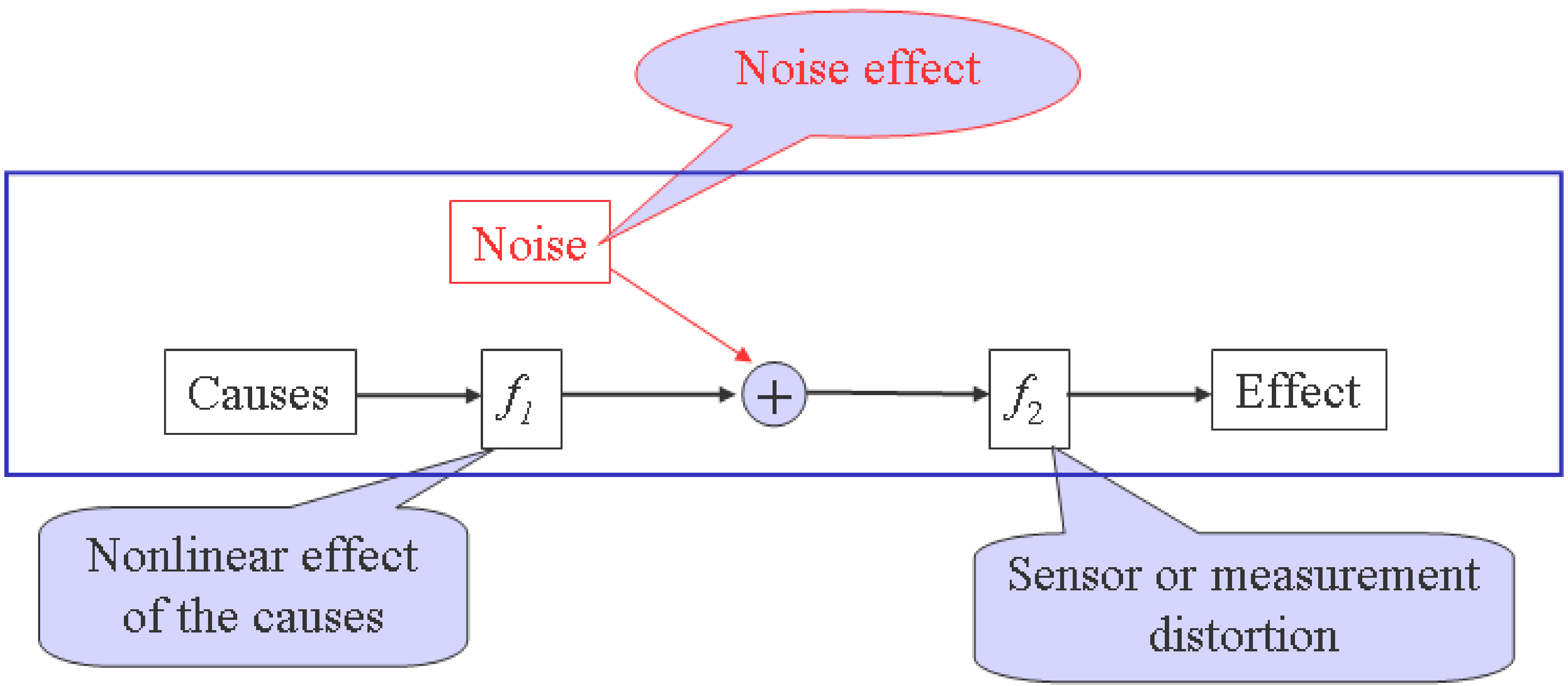
Returns

pval_forward: p value in the x->y direction.

pval_backward: p value in the y->x direction.

Three effects usually encountered in a causal model (Zhang & Chan, 2006; Zhang & Hyvärinen, '09a)

- Without prior knowledge, the assumed model is expected to be
  - general enough: adapt to approximate the true generating process
  - identifiable: asymmetry in causes and effects



- Represented by <u>post-nonlinear causal model with inner additive noise</u>

PNL Causal Model

pa_i: parents (causes) of x_i

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

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

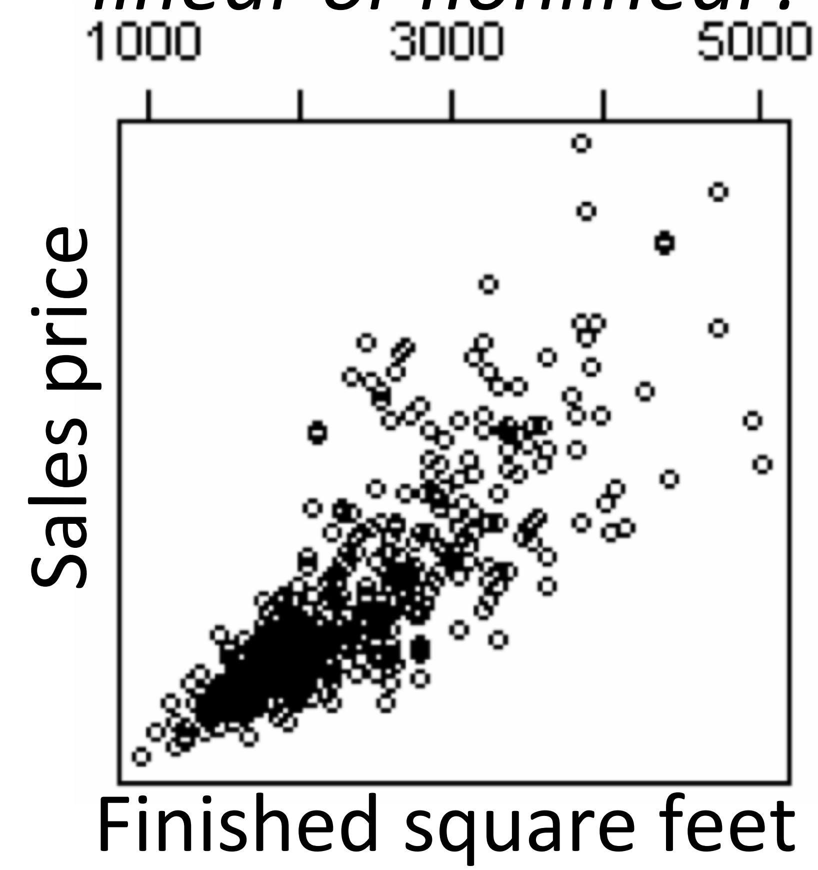
$f_{i,1}$: not necessarily invertible

e_i: noise/disturbance: independent from pa_i

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

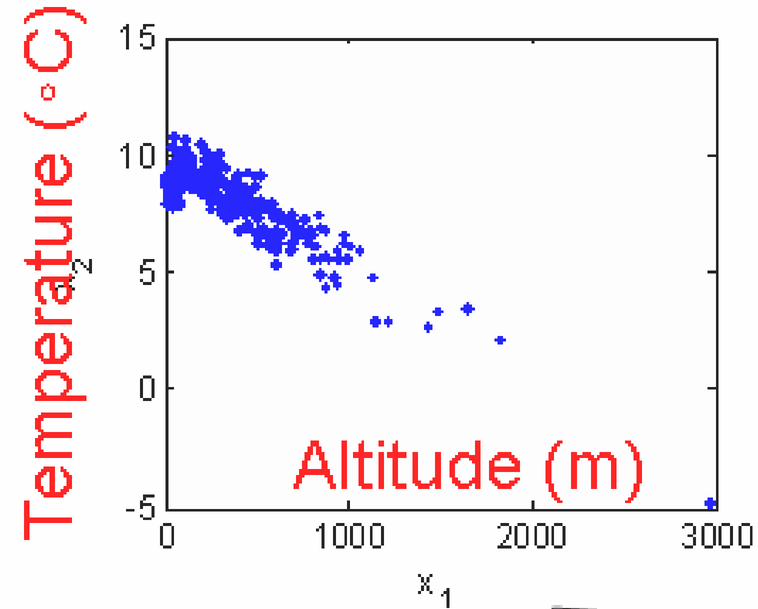
$$Y = X \cdot E = \exp\big(\log(X) + \log(E)\big)$$

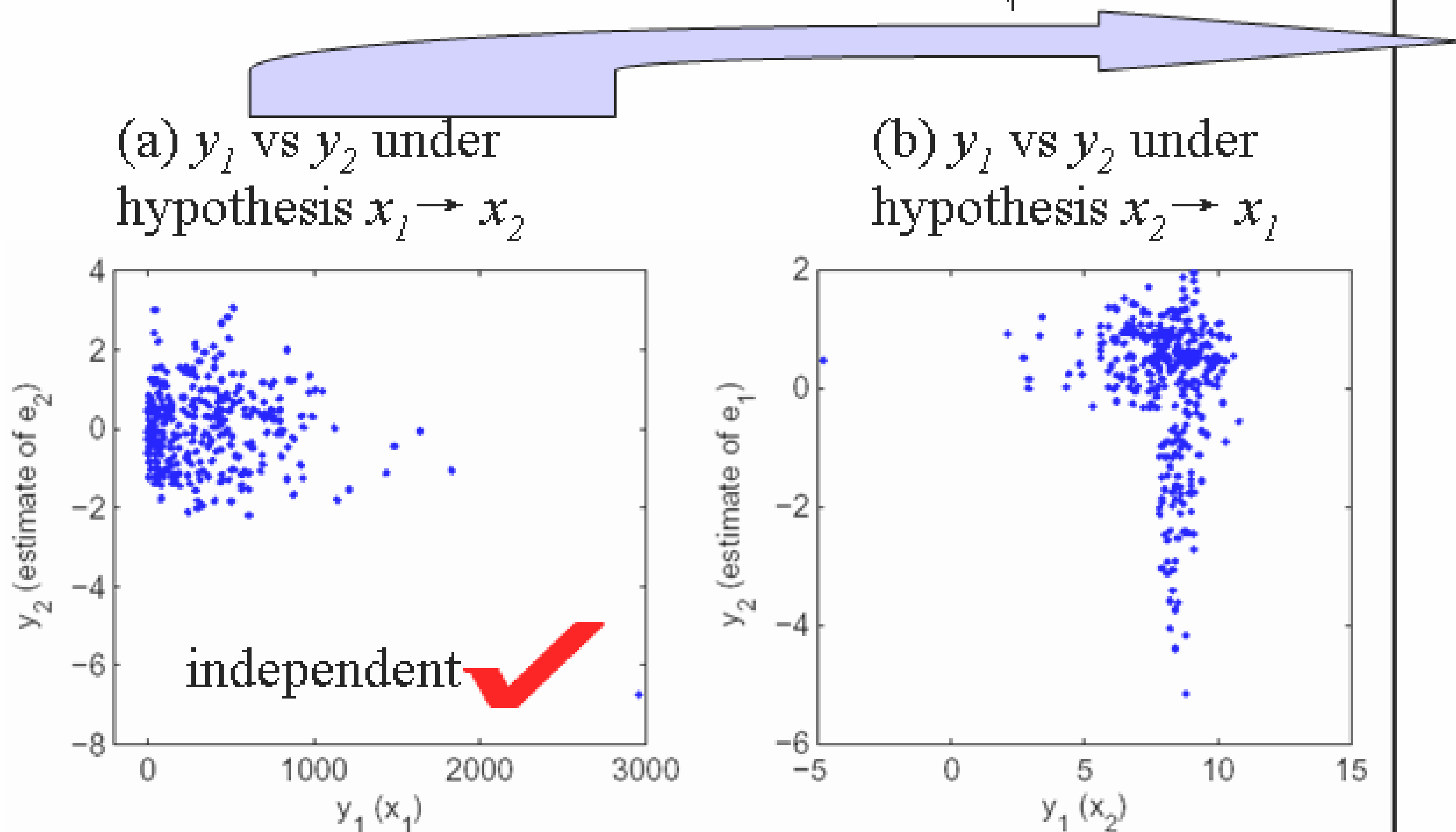
linear or nonlinear?

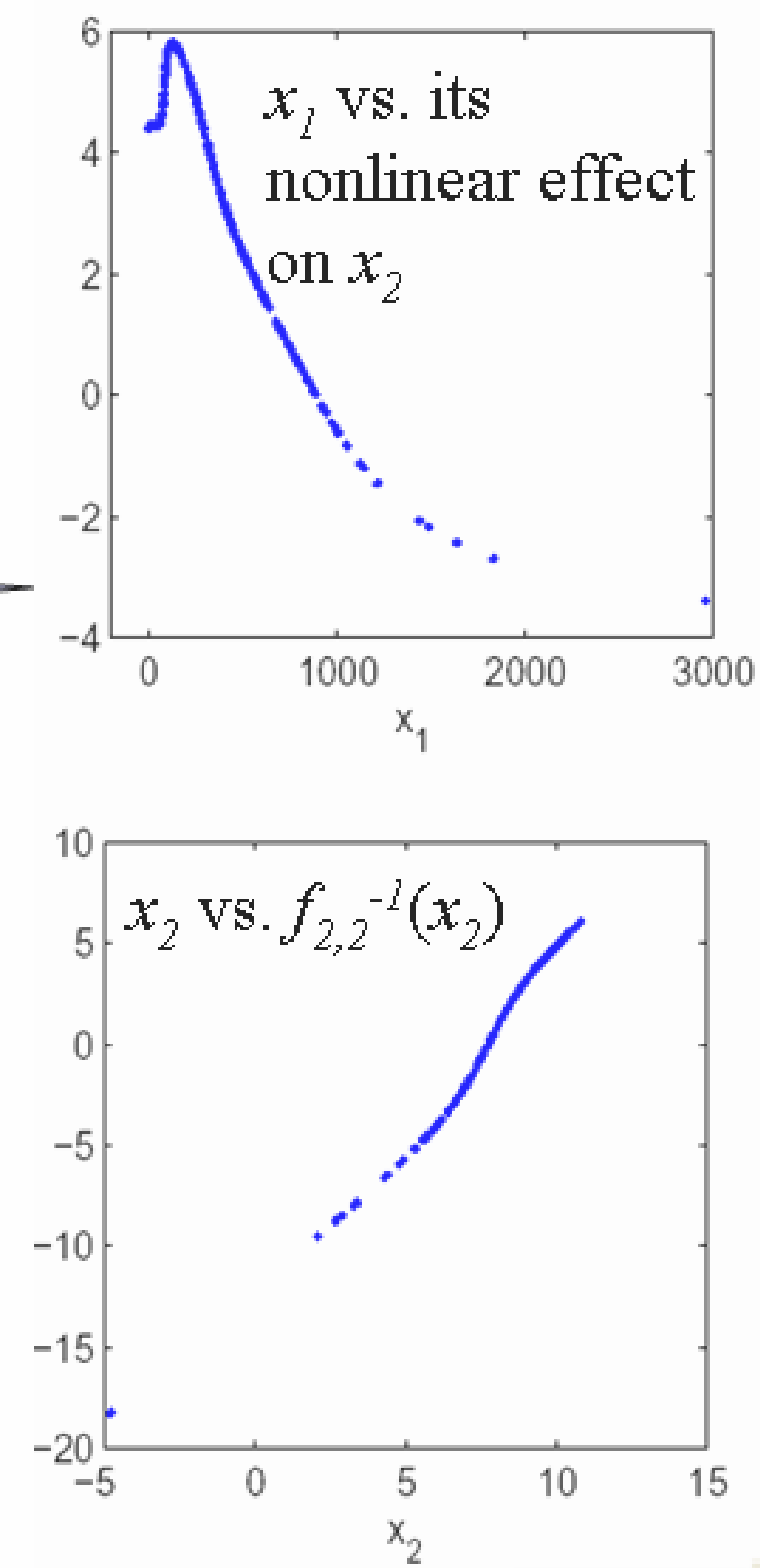


Data Set 1

with PNL Model



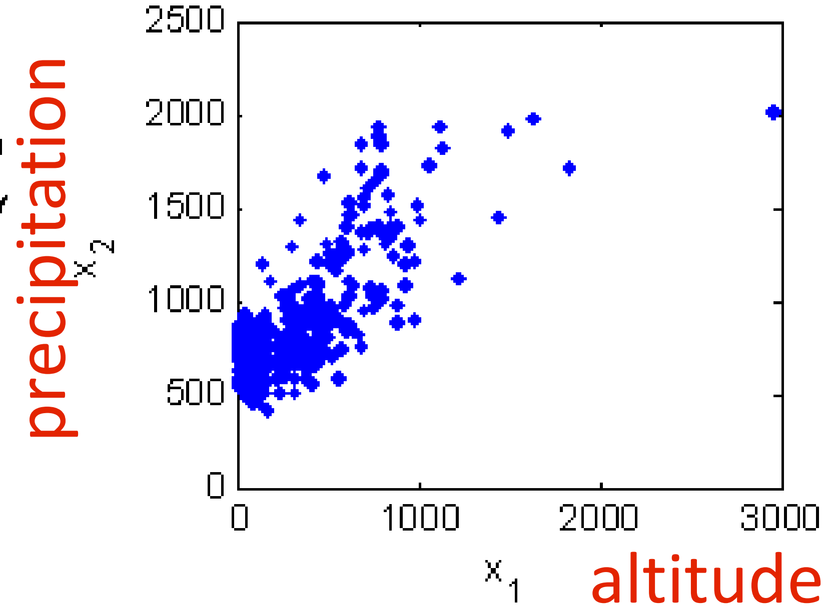


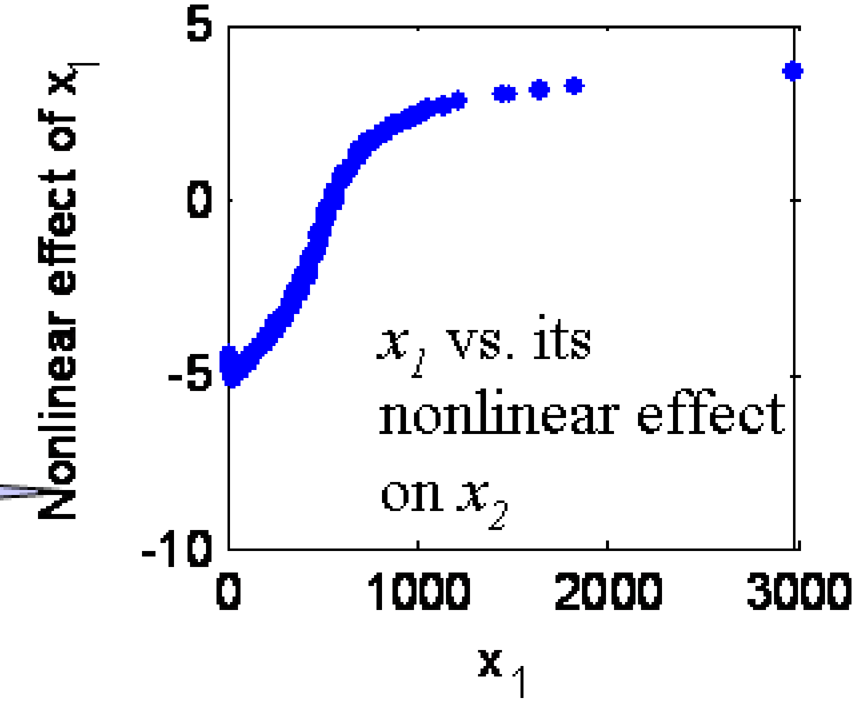


Independence test results on y_1 and y_2 with different assumed causal relations

Data Set	$x_1 \rightarrow x_2$ assumed		$x_2 \rightarrow x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}

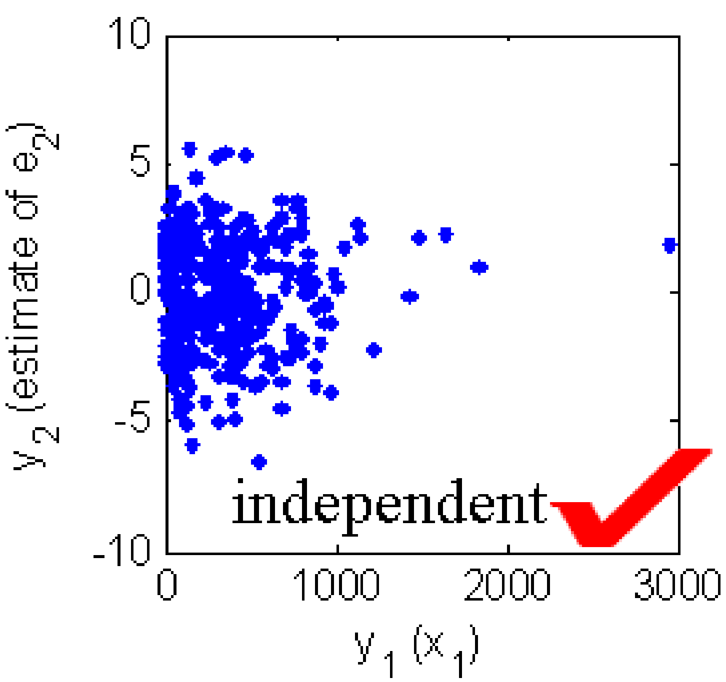
Data Set 2

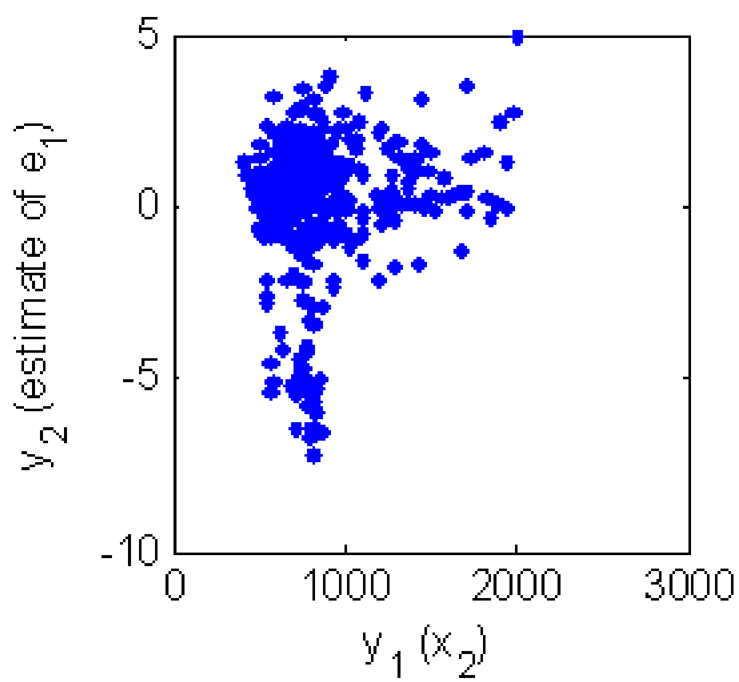


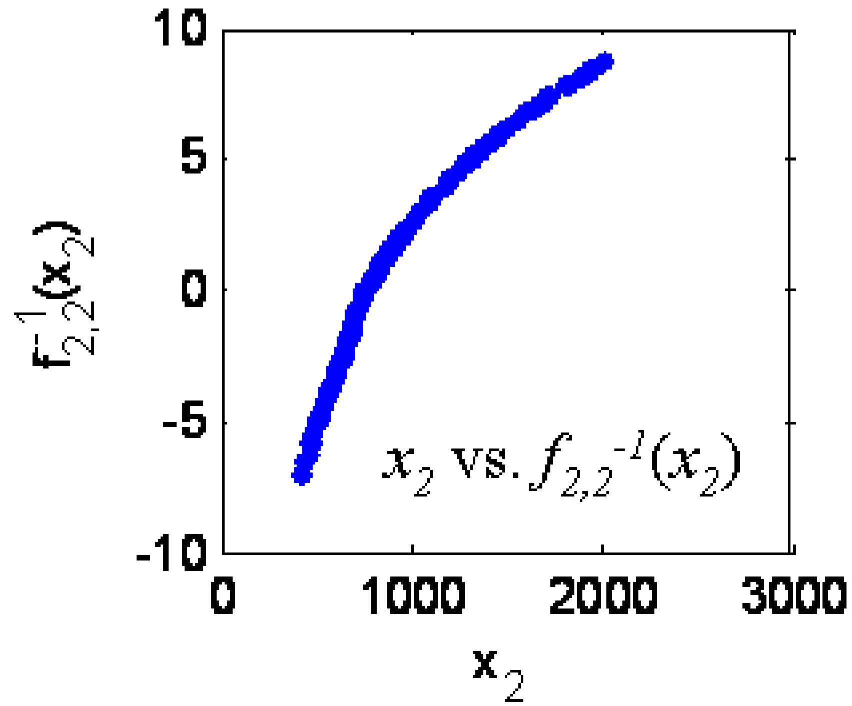


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

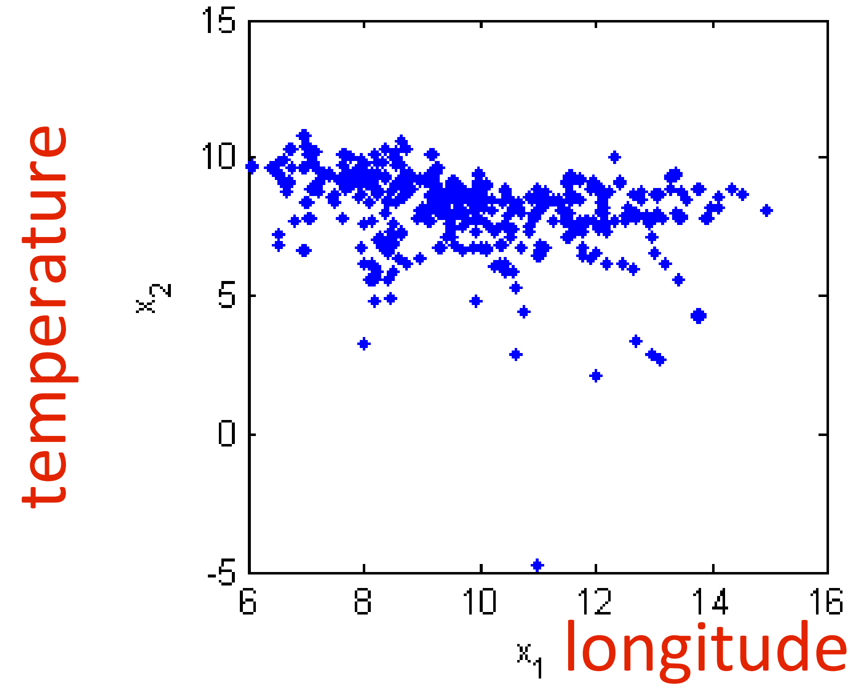
(b) y_1 vs y_2 under hypothesis $x_2 \rightarrow x_1$





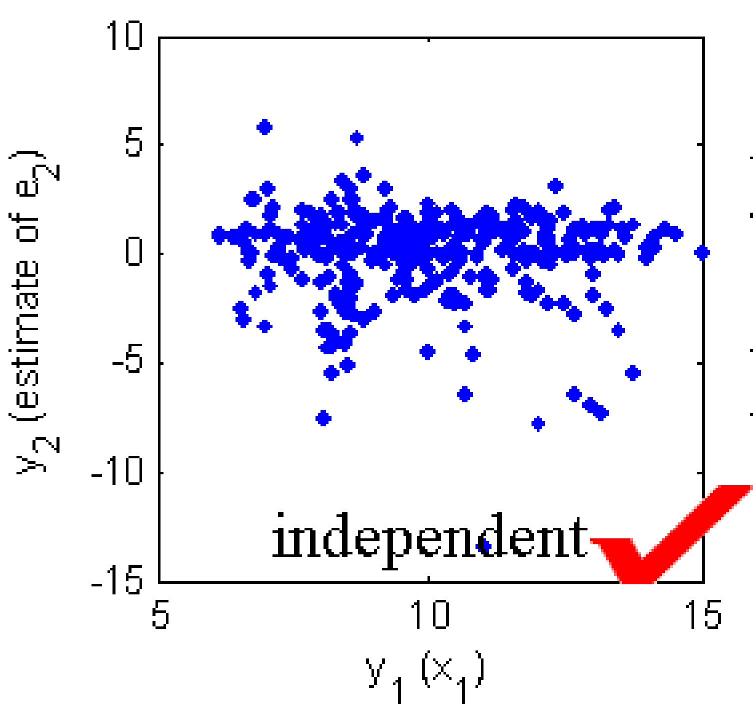


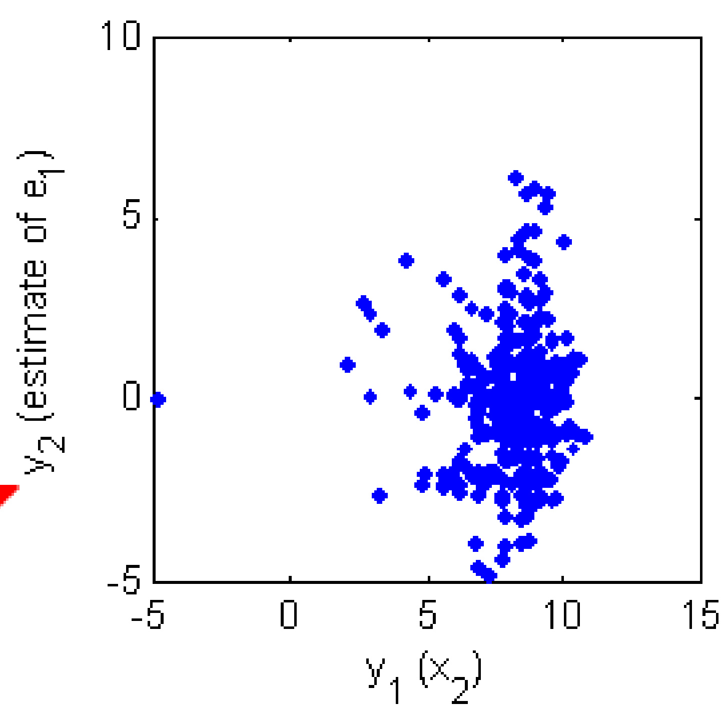
Data Set 3

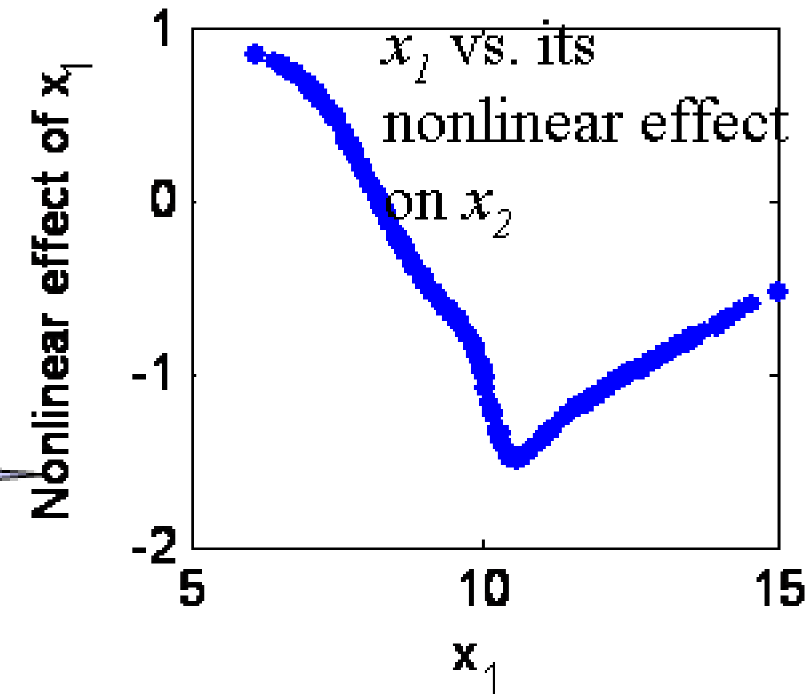


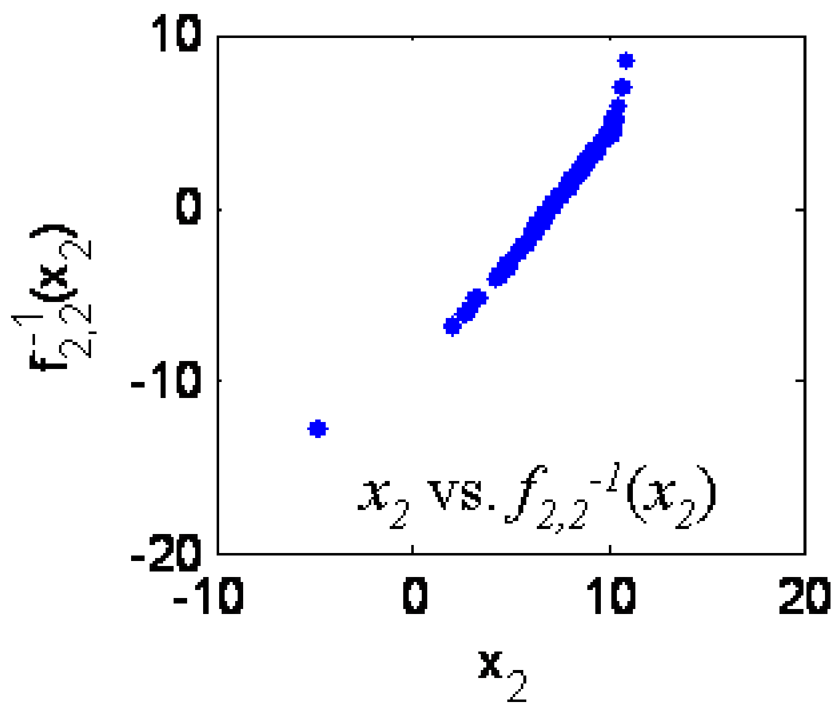
(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

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

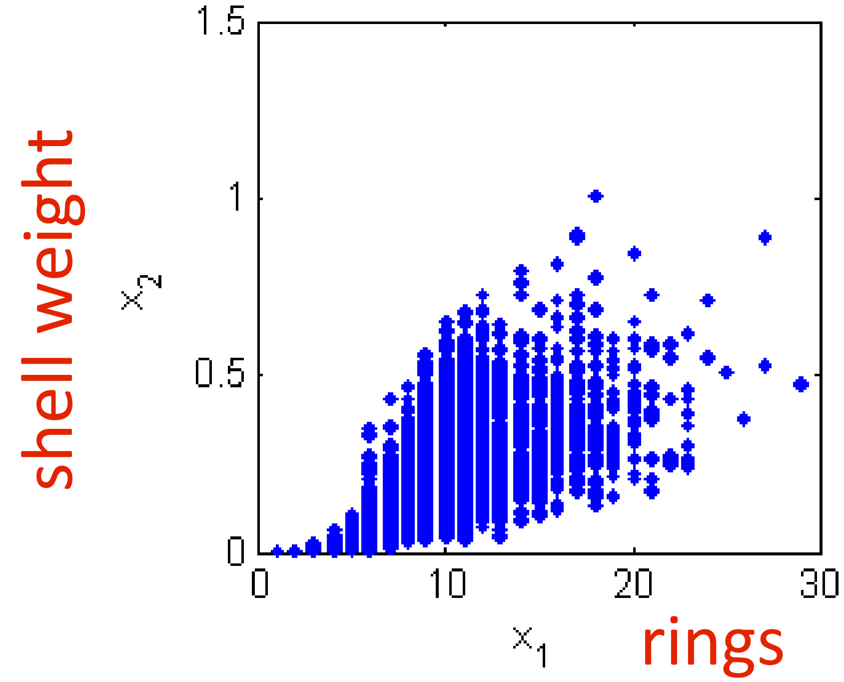


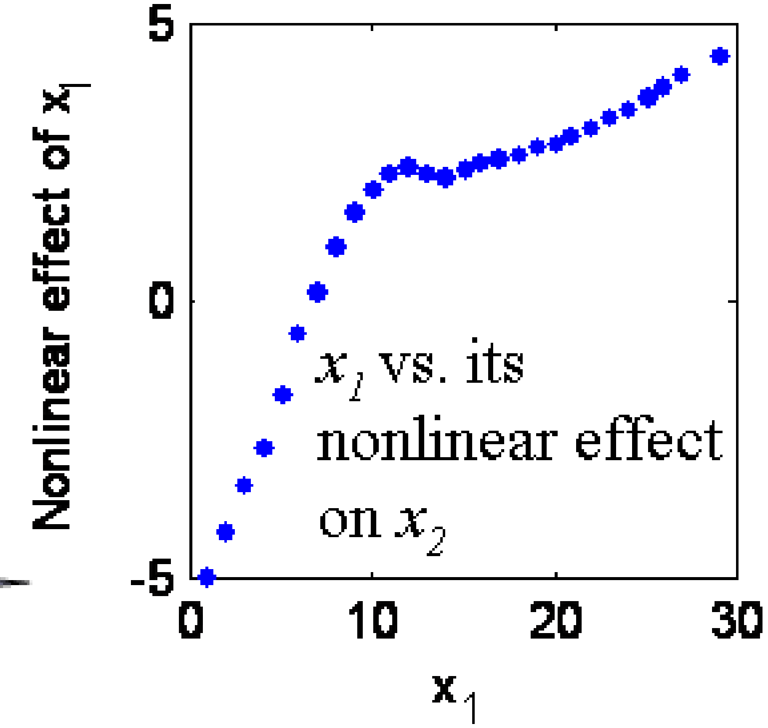






Data Set 6

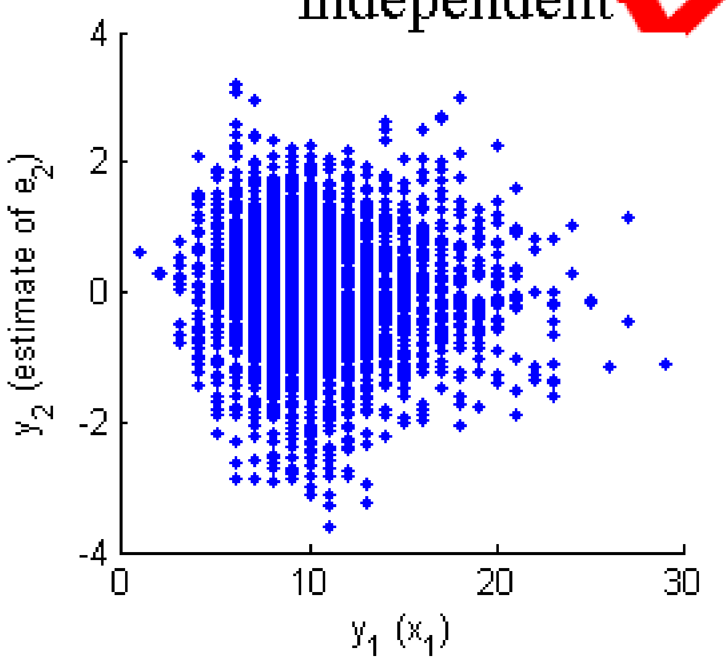


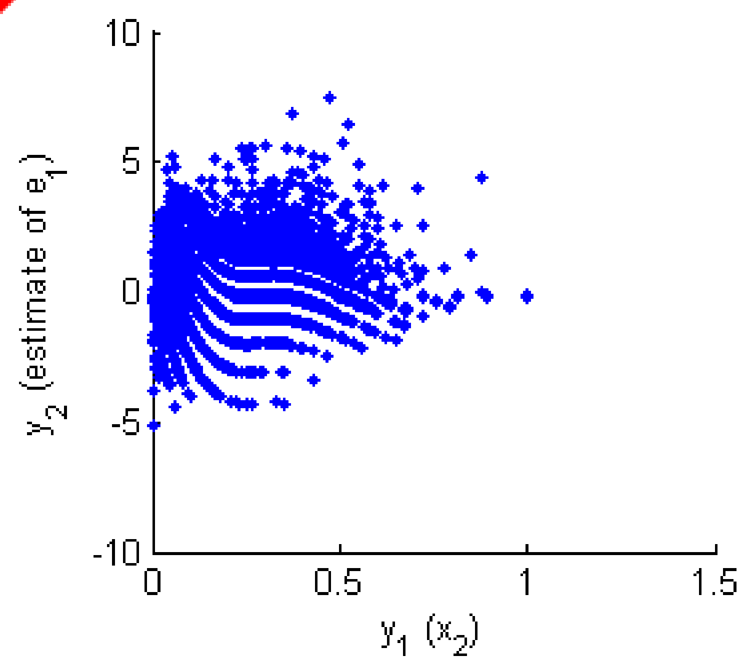


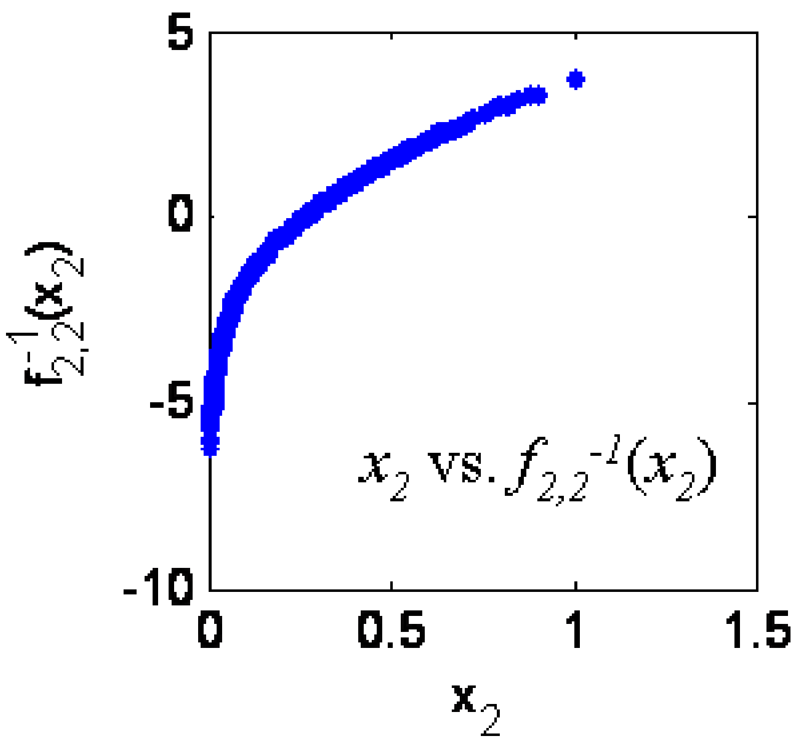
(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

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

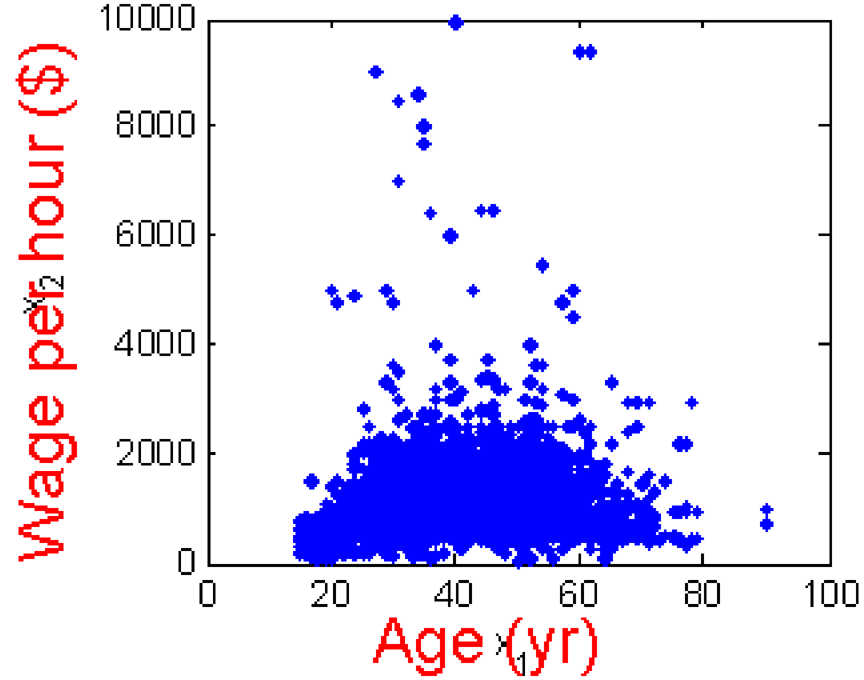
independent

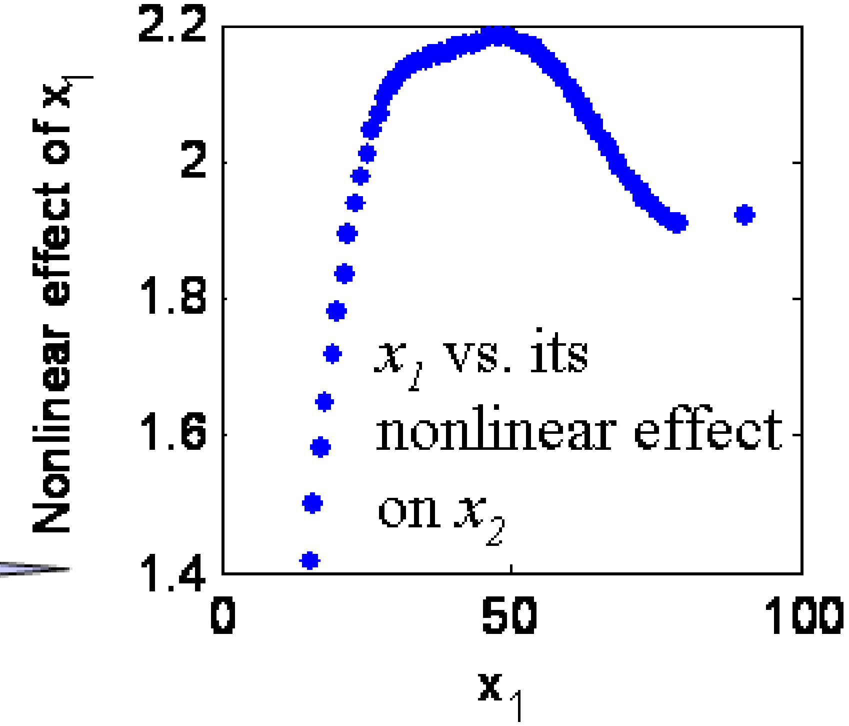






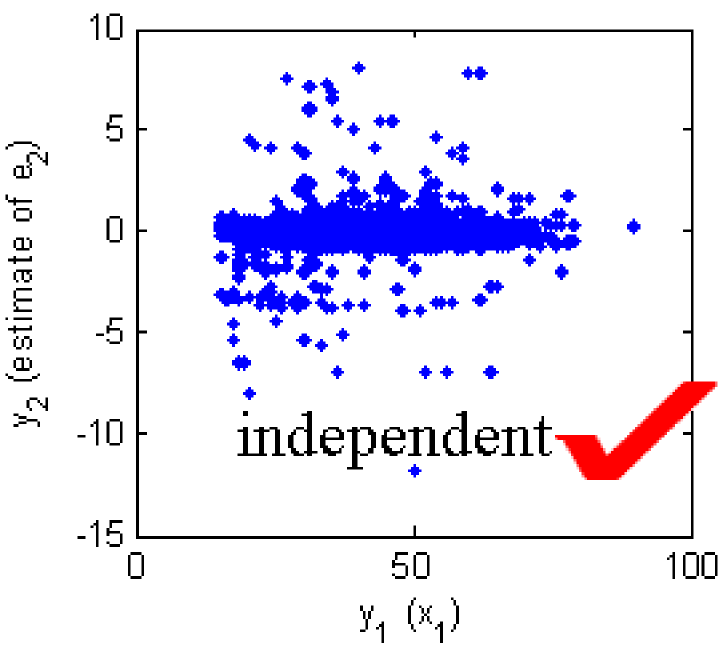
Data Set 8

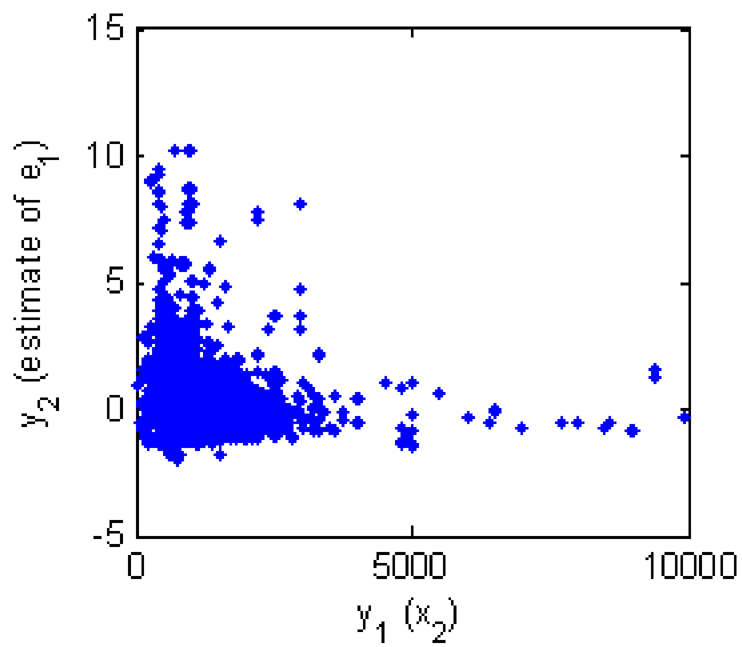


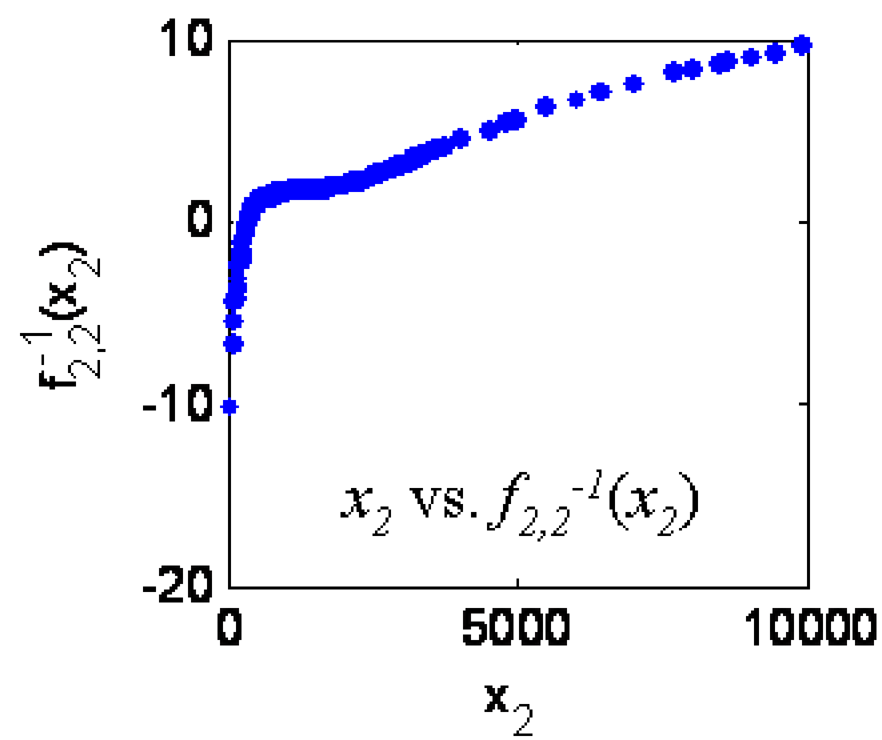


(a) y_1 vs y_2 under hypothesis $x_1 \rightarrow x_2$

(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

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

- **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

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

- 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 $v \to +\infty$

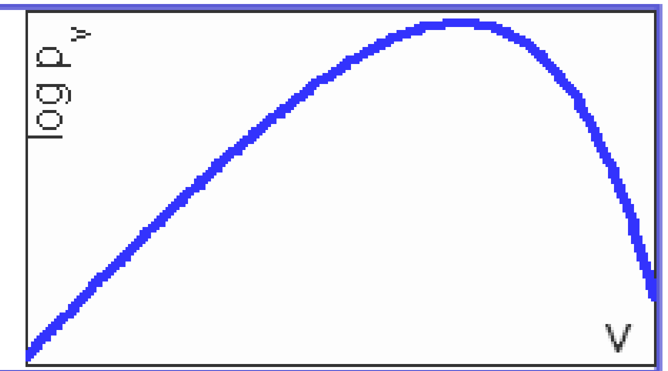
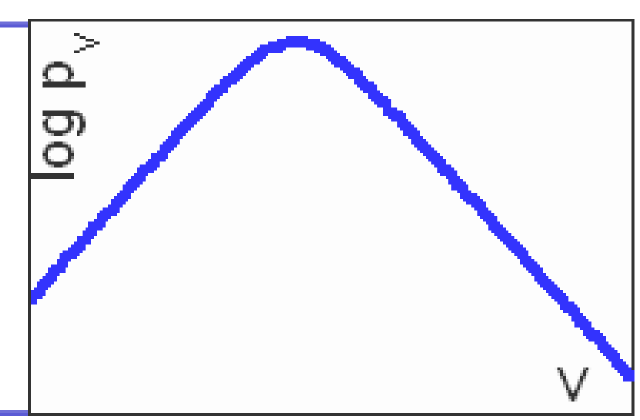


Table 1: All situations in which the PNL causal model is not identifiable.

	p_{e_2}	p_{t_1} $(t_1 = g_2^{-1}(x_1))$	$h = f_1 \circ g_2$	Remark
I	Gaussian	Gaussian	linear	h_1 also linear
II	log-mix-lin-exp	log-mix-lin-exp	linear	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	one-sided asymptotically exponential (but not log-mix-lin-exp)	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	—
IV	log-mix-lin-exp	generalized mixture of two exponentials	Same as above	—
V	generalized mixture of two exponentials	two-sided asymptotically exponential	Same as above	—

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

$(\log p_v)' \to c_1 \ (c_1 \neq 0)$,
as $v \to -\infty$ and
$(\log p_v)' \to c_2 \ (c_2 \neq 0)$,
as $v \to +\infty$



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 $v \to +\infty$

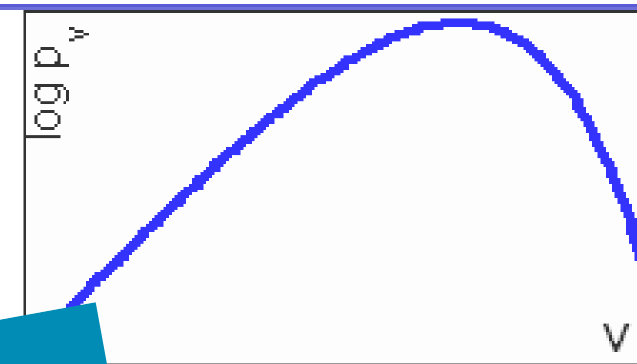


Table 1: All situations ... not identifiable.

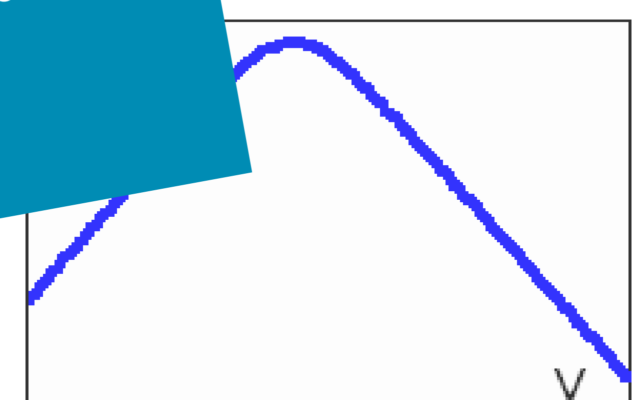
	p_{e_2}	...	Remark
I	Gaussian	...	h_1 also linear
II	log-mix-lin-exp	...	h_1 strictly monotonic, and $h_1' \to$ ...0, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	...	—
IV	log-mix-lin-exp	...	
V	generalized mixture of two exponentials	...	

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

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.

$\to c_2\ (c_2 \neq 0)$,
as $v \to +\infty$



Post-nonlinear Models by causal-learn

```
from causallearn.search.FCMBased.PNL.PNL import PNL
pnl = PNL()
p_value_foward, p_value_backward = pnl.cause_or_effect(data_x, data_y)
```

Parameters

data_x: input data (n, 1), n is the sample size.

data_y: output data (n, 1), n is the sample size.

Returns

pval_forward: p value in the x->y direction.

pval_backward: p value in the y->x direction.

Take-Home Message: Causal Discovery with Nonlinear Functional Causal Models

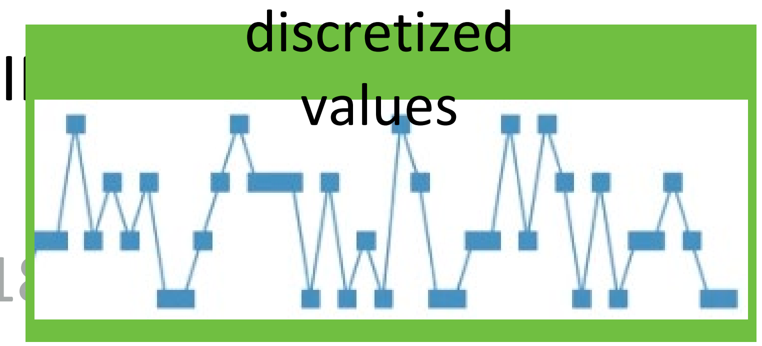
- Functional causal models naturally describe the causal processes
- Can we use them to distinguish cause from effect?
- Certain types of constraints on f are needed to guarantee the identifiability of the causal direction
- Nonlinearities are encountered frequently and should be considered
- Trade-off of generality & identifiability
- Limitation: more than one noise term? large-scale problems?

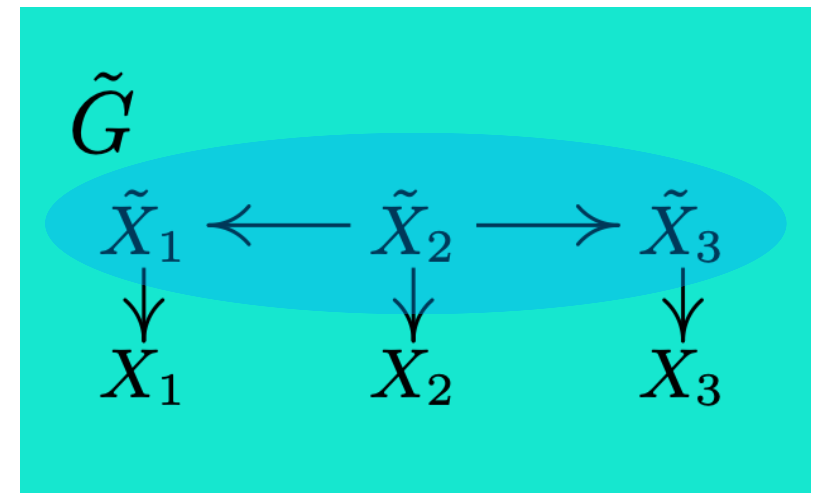
More Practical Issues

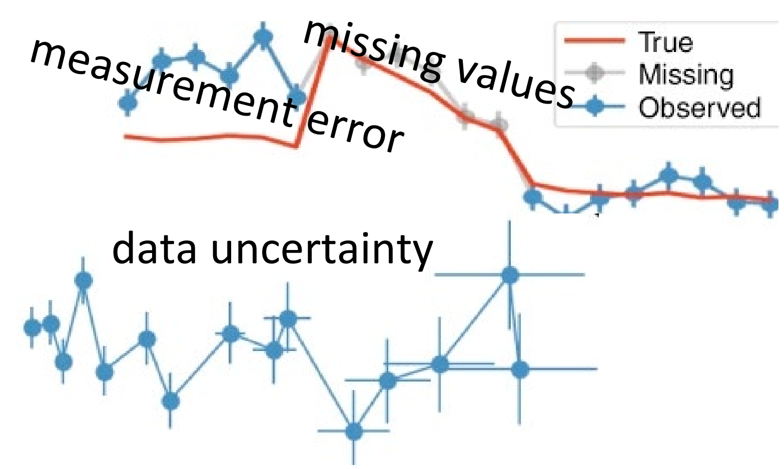
- Nonlinear Relations
- Measurement Error
- Selection Bias
- Missing data
- Nonstationarity
- …

Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NI UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'1
- **Measurement error** (Zhang et al., UAI'18; PSA'18)

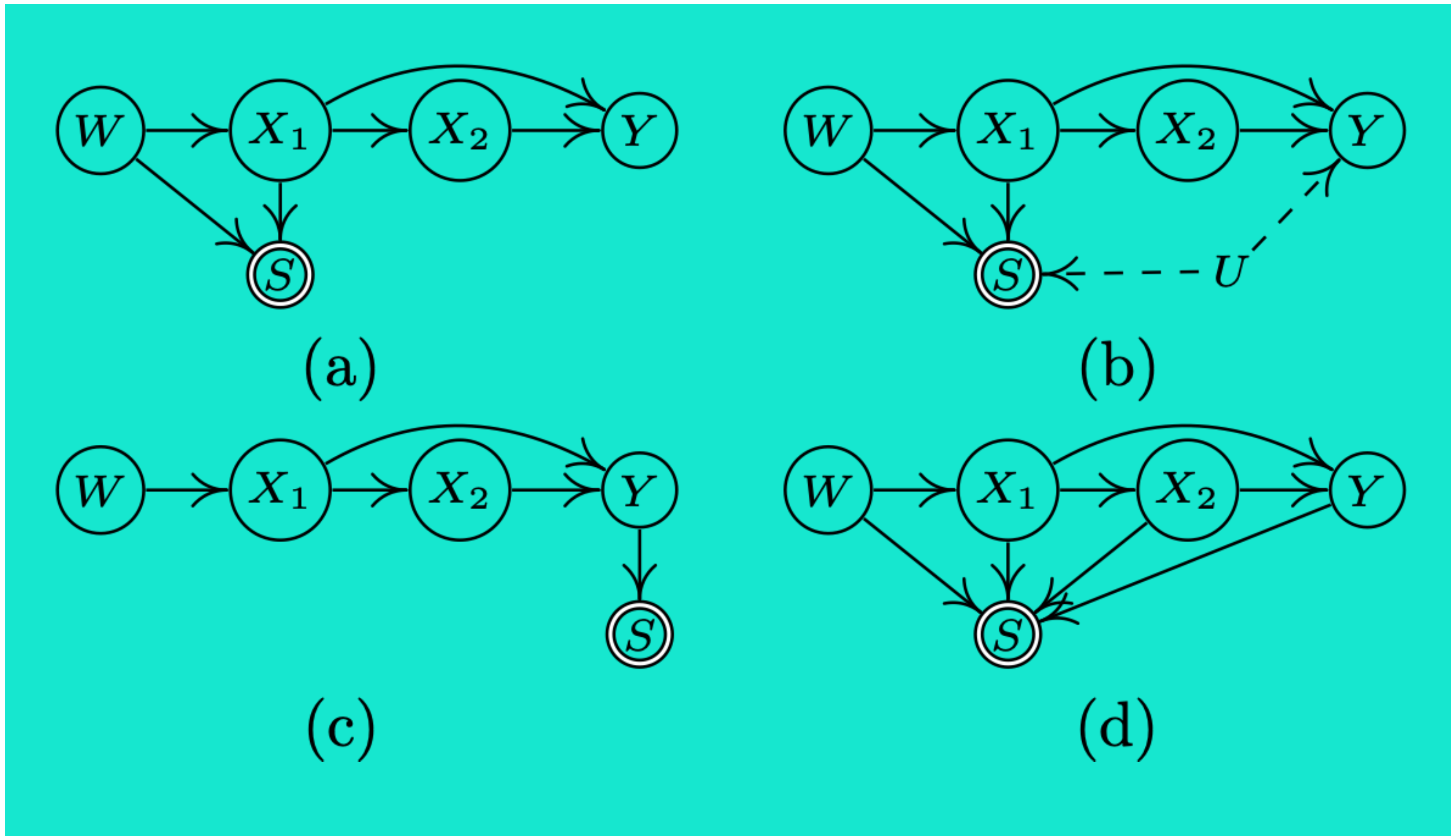






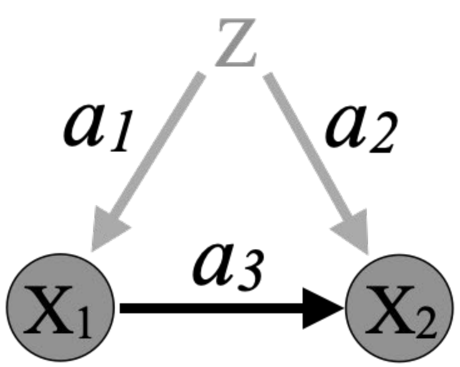
Practical Issues in Causal Discovery…

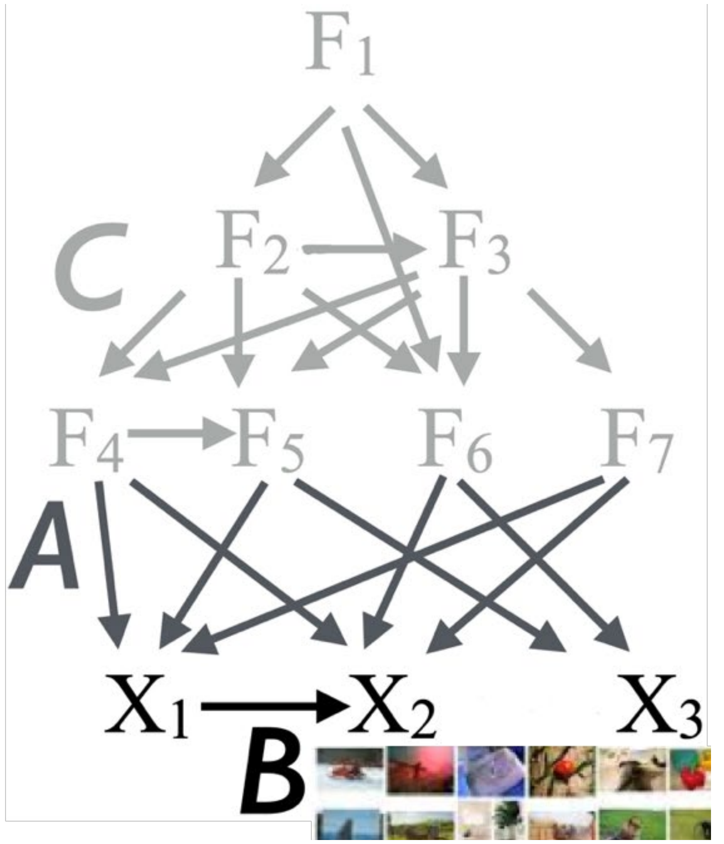
- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- **Selection bias** (Zhang et al., UAI'16)



Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- **Confounding** SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19); latent causal representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'20; Cai et al., NeurIPS'19; Adams et al., NeurIPS'21)





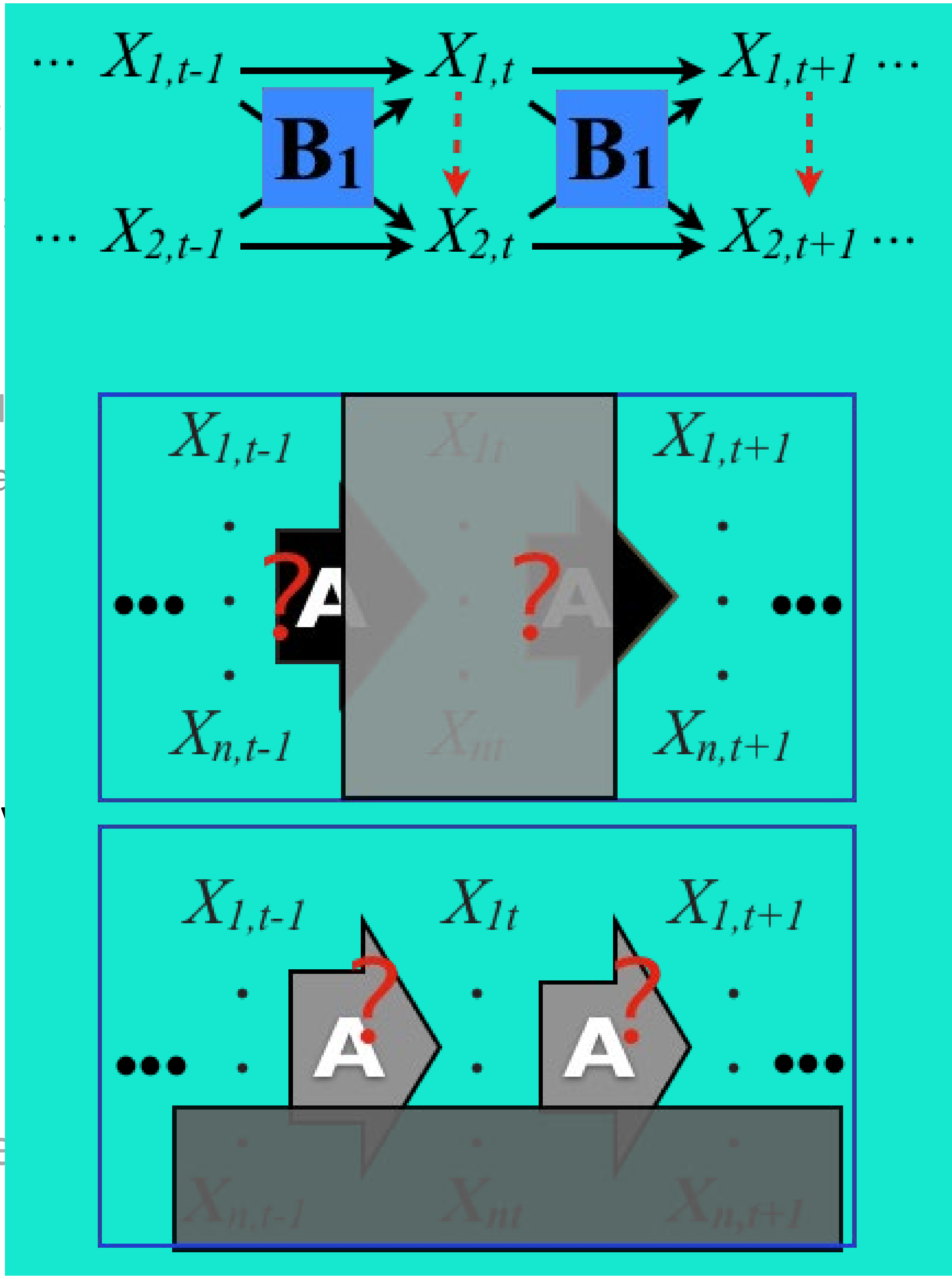
Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et al., KDD'18; Cai et al., NIPS'18)
- Measurement error (Zhang et al., UAI'18; PSA'18)
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS'19; Ding et al., NIPS'19); latent causal representation learning (Silva et al., JMLR'06; Xie et al., NeurIPS'20; Cai et al., NeurIPS'19; Adams et al., NeurIPS'21)
- **Missing values (Tu et al., AISTATS'19)**

X1	X2	X3	X4	X5	X6
-9.4653403e-01	6.6703495e-01	8.2886922e-01	-1.3695521e+00	-3.2675465e-02	1.8634806e-01
-9.4895568e-01			-4.6381657e-01	-1.8280031e+00	
	5.1435422e-01	6.7338326e-01	4.3403559e-01	9.4535076e-01	7.5164028e-01
7.2489037e-01		5.1325341e-01	8.3567780e-01	2.9825903e-01	7.7796018e-02
		-1.3440612e+00			-7.3325009e-01
1.3261794e+00	-6.1971037e-01	-1.0498756e-01	1.4171149e+00	1.6251026e+00	3.7478050e-01
-2.1128404e+00	1.3359744e-02	-2.0209600e+00	-1.7172659e+00	-2.4746799e+00	-2.8026586e+00
1.5453163e+00	-5.3986972e-01	4.5157367e-01	1.5566262e+00	9.3882105e-01	-4.3382982e-01
6.5974086e-02	5.5826895e-01	6.5247930e-01	-5.7895322e-01	5.0062743e-01	1.0183537e+00
8.9772858e-01	2.6752870e-01	-4.9204975e-01	7.7933358e-02	8.3467624e-01	9.2744311e-01
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

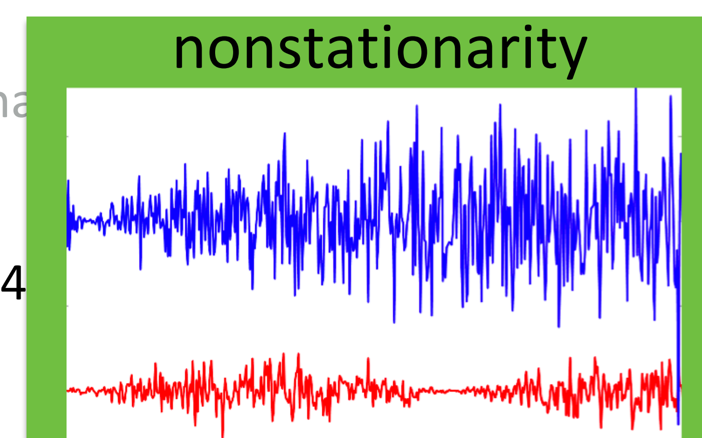
Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP'06; Hoyer et al., NIPS'08; Zhang & Hyvärinen, UAI'09; Huang et al., KDD'18)
- Categorical variables or mixed cases (Huang et
- Measurement error (Zhang et al., UAI'18; PSA'
- Selection bias (Zhang et al., UAI'16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NI representation learning (Silva et al., JMLR'06; Xie et a NeurIPS'21)
- Missing values (Tu et al., AISTATS'19)
- **Causality in time series**
    - Time-delayed + instantaneous relations (Hy ECML'09; Hyvarinen et al., JMLR'10)
    - Subsampling / temporally aggregation (Dan ICML'15 & UAI'17)
    - From partially observable time series (Geige



Practical Issues in Causal Discovery…

- Nonlinearities (Zhang & Chan, ICONIP’06; Hoyer et al., NIPS’08; Zhang & Hyvärinen, UAI’09; Huang et al., KDD’18)
- Categorical variables or mixed cases (Huang et al., KDD’18; Cai et al., NIPS’18)
- Measurement error (Zhang et al., UAI’18; PSA’18)
- Selection bias (Spirtes 1995; Zhang et al., UAI’16)
- Confounding (SGS 1993; Zhang et al., 2018c; Cai et al., NIPS’19; Ding et al., NIPS’19); latent causal representation learning (Silva et al., JMLR’06; Xie et al., NeurIPS’20; Cai et al., NeurIPS’19; Adams et al., NeurIPS’21)
- Missing values (Tu et al., AISTATS’19)
- Causality in time series
  - Time-delayed + instantaneous relations (Hyvarinen ICML’08; Zha… al., JMLR’10)
  - Subsampling / temporally aggregation (Danks & Plis, NIPS WS’14
  - From partially observable time series (Geiger et al., ICML’15)
- **Nonstationary/heterogeneous data** (Zhang et al., IJCAI’17; Huang et al, ICDM’17, Ghassami et al., NIPS’18; Huang et al., ICML’19 & NIPS’19; Huang et al., JMLR’20)



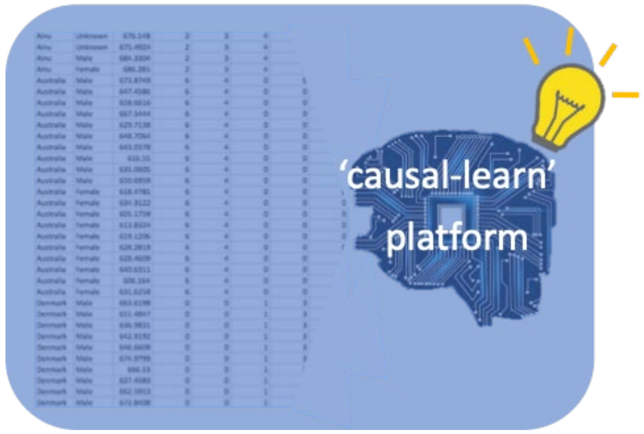
Summary: Practical Issues in Causal Discovery

- Latent confounders, cycles, nonlinearities (and even mixed data types), measurement error, selection bias, missing values, nonstationarity…
- Don't worry—look into the problems
- Learning latent confounders and their relations!

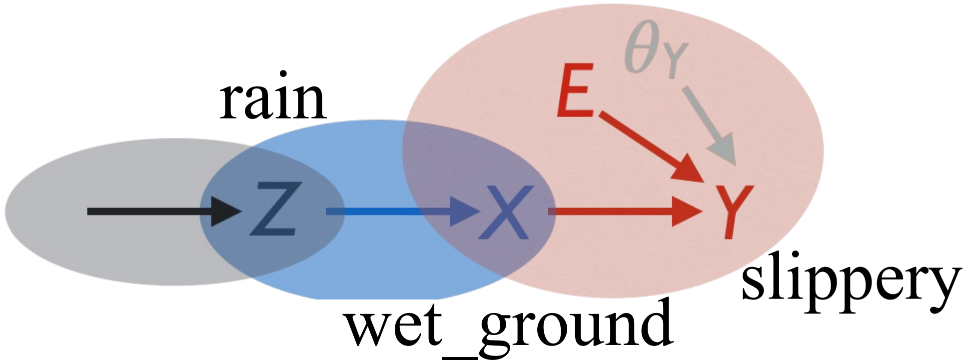
Causal Representation Learning: Recovery of the Hidden World

- Why causal/disentangled representations ?
- How?
  - IID case
    - Linear-Gaussian case
    - Linear, non-Gaussian case
    - Nonlinear case
  - From multiple distributions
  - With temporal information

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

<table>
<tr><th>i.i.d. data?</th><th>Parametric constraints?</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 rowspan="2">Non-I, but I.D.</td><td rowspan="2">No/Yes</td><td>No</td><td>(Extended) regression</td></tr>
<tr><td>Yes</td><td>Latent temporal causal processes identifiable!</td></tr>
<tr><td rowspan="4">I., but non-I.D.</td><td>No</td><td rowspan="2">No</td><td>More informative than MEC (CD-NOD)</td></tr>
<tr><td>Yes</td><td>May have unique identifiability</td></tr>
<tr><td>No</td><td rowspan="2">Yes</td><td>Changing subspace identifiable</td></tr>
<tr><td>Yes</td><td>Variables in changing relations identifiable</td></tr>
</table>

A Problem in Psychology: Finding Underlying Mental Conditions?

- 50 questions for big 5 personality test

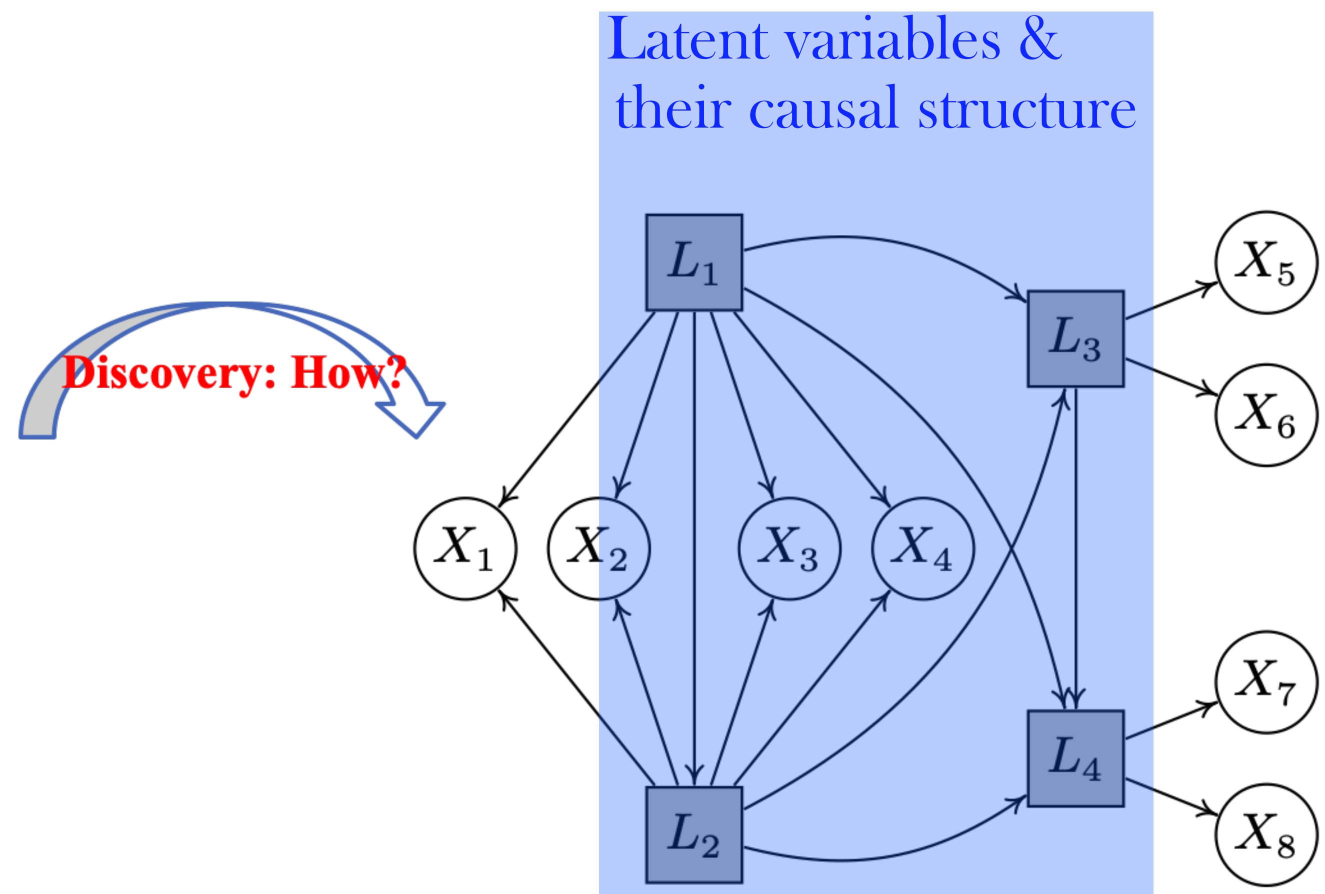
race	age	engnat	gender	hand	source	country	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	A1	A2	A3	A4	A5
3	53	1	1	1	1	US	4	2	5	2	5	1	4	3	5	1	1	5	2	5	1	1	1	1	1	1	1	5	1	5	2
13	46	1	2	1	1	US	2	2	3	3	3	3	1	5	1	5	2	3	4	2	3	4	3	2	2	4	1	3	3	4	4
1	14	2	2	1	1	PK	5	1	1	4	5	1	1	5	5	1	5	1	5	5	5	5	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	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

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

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	9.1	7.0	1.0	4.6
3.8	1.7	6.4	7.3	9.0	6.7	0.8	4.3
4.1	2.8	6.5	6.9	9.3	6.7	2.7	4.6
...	...	...	...	...	...	...	...



- Find latent variables L_i and their causal relations ?
- Rank deficiency or GIN helps solve the problem

Outline

- Why causal/disentangled representations ?
- How?
  - IID case
    - **Linear-Gaussian case**
    - Linear, non-Gaussian case
    - Nonlinear case
  - From multiple distributions
  - With temporal information

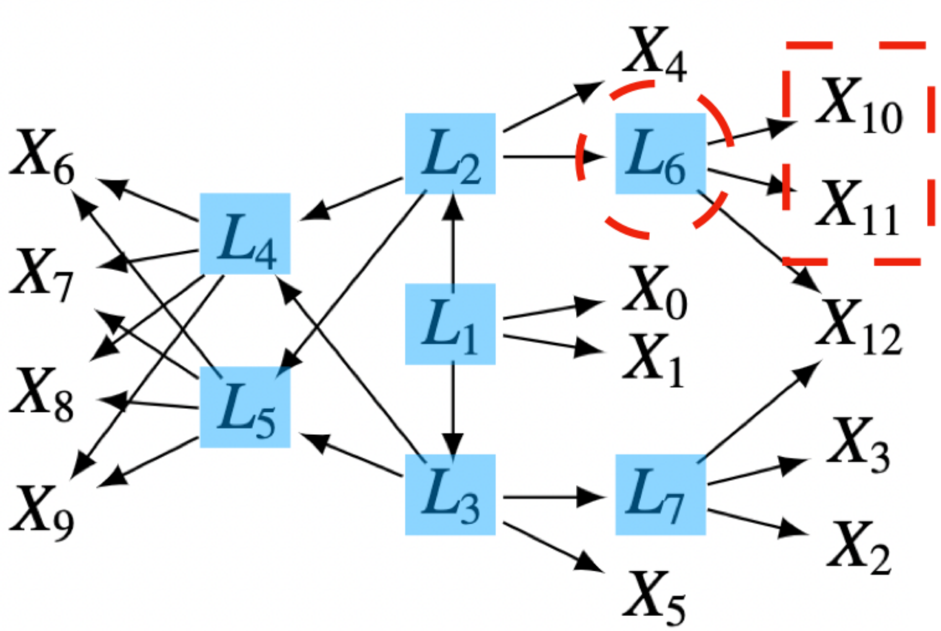
Linear, Gaussian Case: With Rank Deficiency Constrains

Basic idea:

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

foundation of this method

- $\text{rank}(\Sigma_{X_A,X_B})$, 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 $X_B = \mathbf{X} \backslash X_A$ $\text{rank}(\Sigma_{X_A,X_B}) = 1$ 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

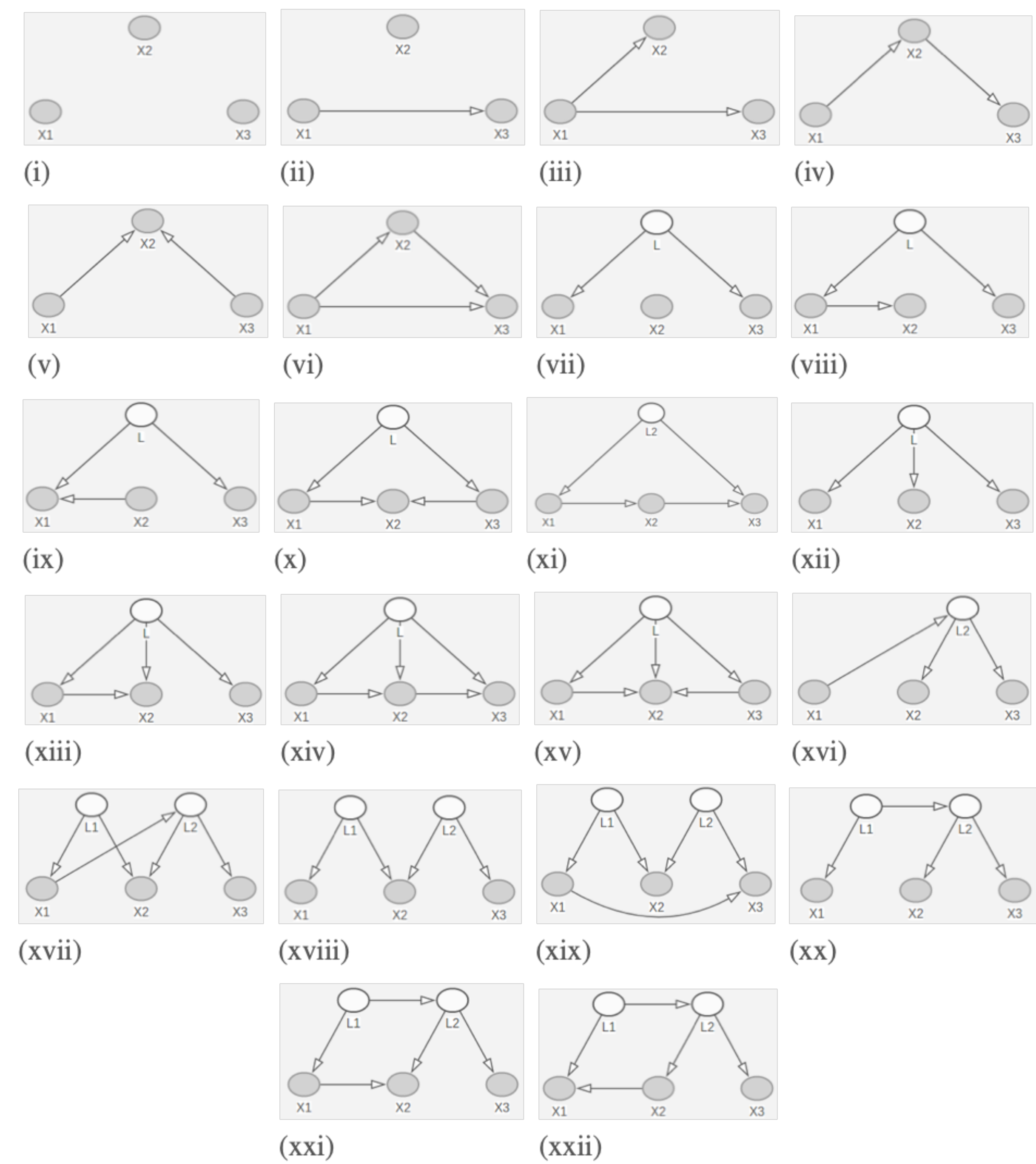
- *Huang, Low, Xie, Glymour, Zhang, "Latent Hierarchical Causal Structure Discovery with Rank Constraints, NeurIPS 2022*

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
- Estimation is generally difficult

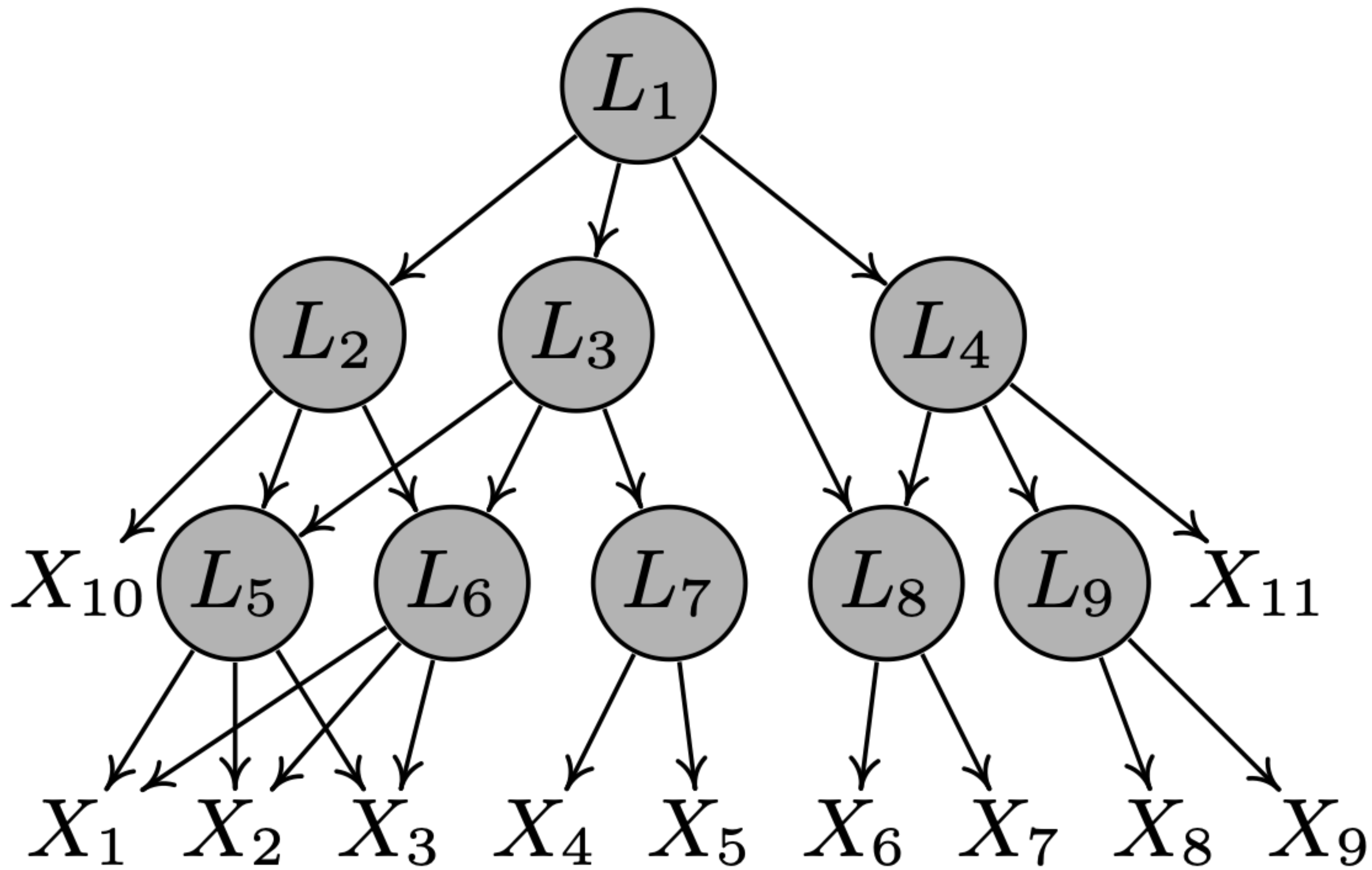
Identifiable graphs with only 3 measured variables



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

Estimating Latent Hierarchical Structure

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

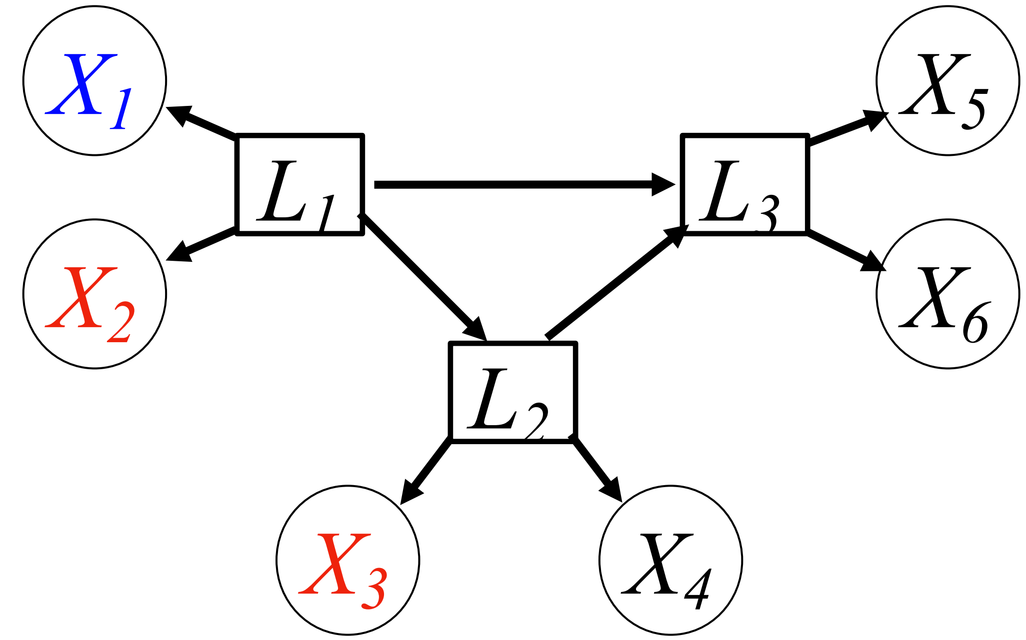


- *Xie, Huang Chen, He, Geng, Zhang, "Estimation of Linear Non-Gaussian Latent Hierarchical Structure," ICML 2022*
- *Huang, Low, Xie, Glymour, Zhang, "Latent Hierarchical Causal Structure Discovery with Rank Constraints, NeurIPS 2022*
- *Adams, Hansen, Zhang, "Identification of Partially Observed Linear Causal Models: Graphical Conditions for the Non-Gaussian and Heterogeneous Cases," NeurIPS 2021*

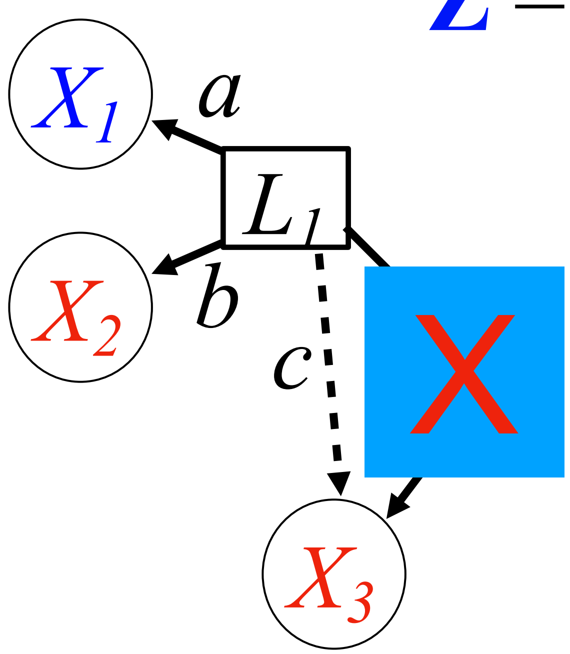
Outline

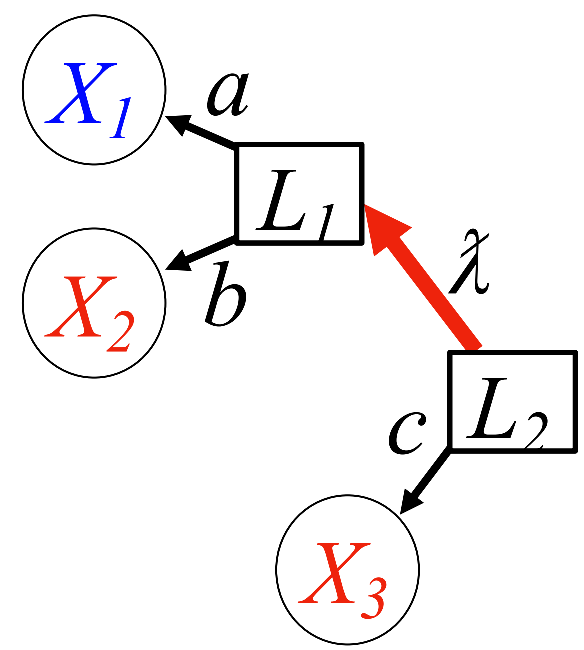
- Why?
- How?
  - IID case
    - Linear-Gaussian case
    - **Linear, non-Gaussian case**
    - Nonlinear case
  - From multiple distributions
  - With temporal information

Generalized Independent Noise Condition (GIN)



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





$$
\begin{aligned}
& c \cdot X_2 - b \cdot X_3 \\
=& c(bL1 + E_2) - b(cL_1 + E_3) \\
=& cE_2 - bE_3,
\end{aligned}
$$

independent from L_1 and from X_1,

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

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: GIN

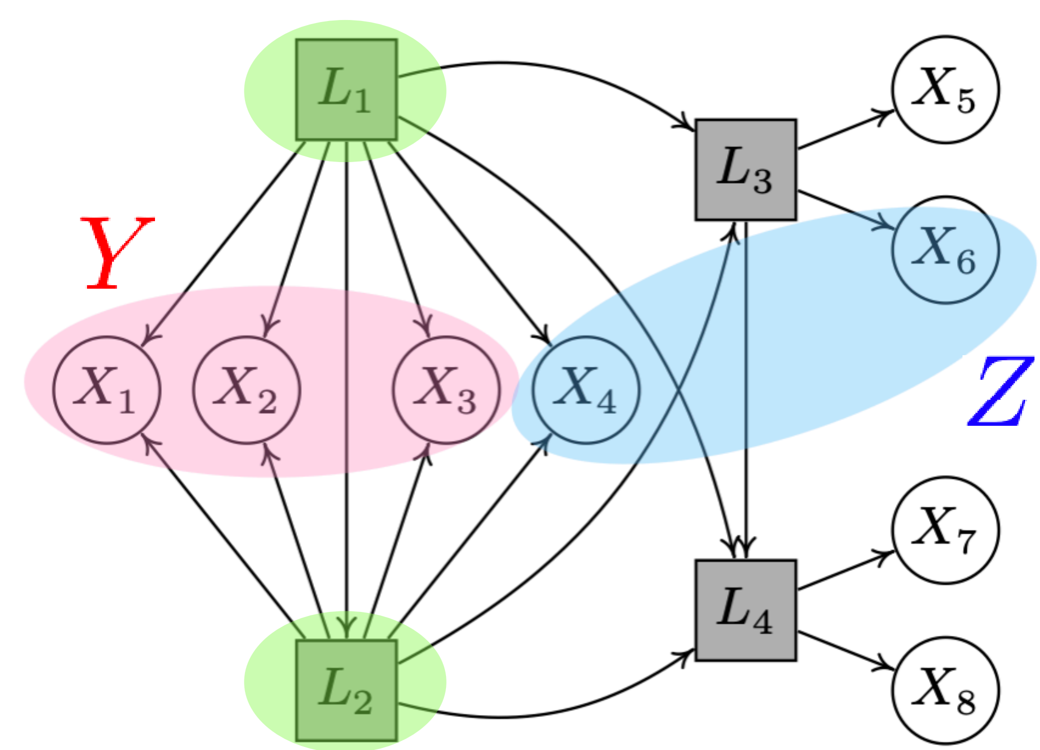
- Generalized Independent Noise (GIN) Condition:

  (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)

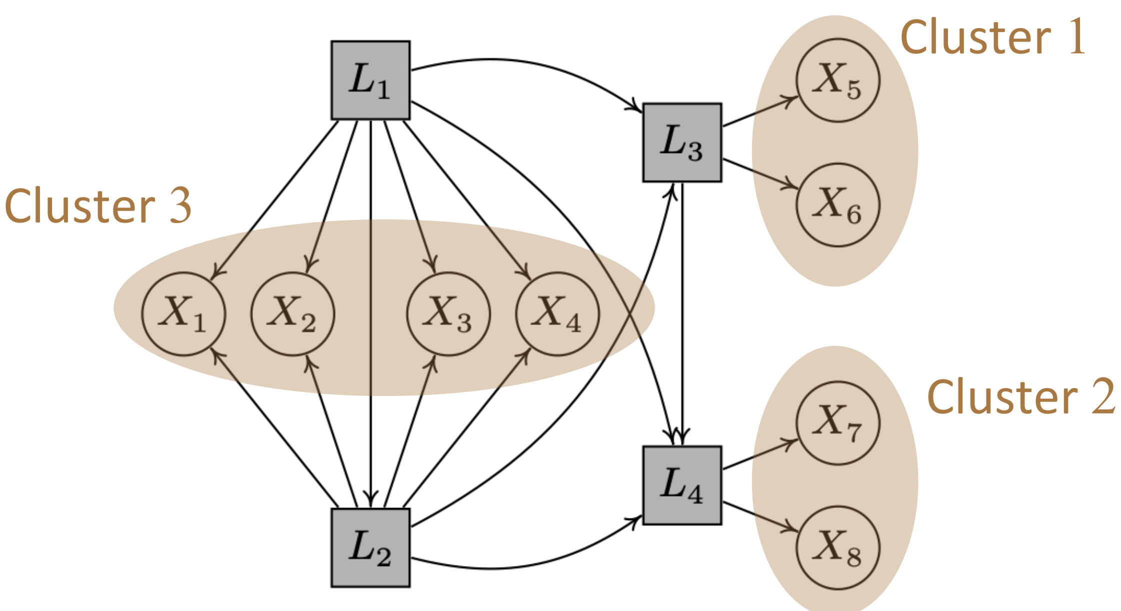


X_i: observed variables
L_i: latent variables

GIN Condition for Estimating Linear Non-Gaussian Latent Graphs

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

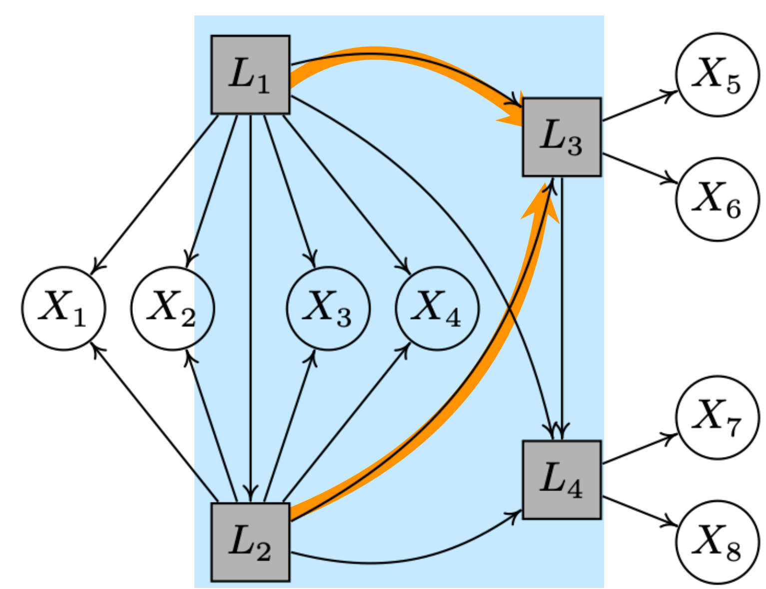
Step 1: find ***causal clusters***



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

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

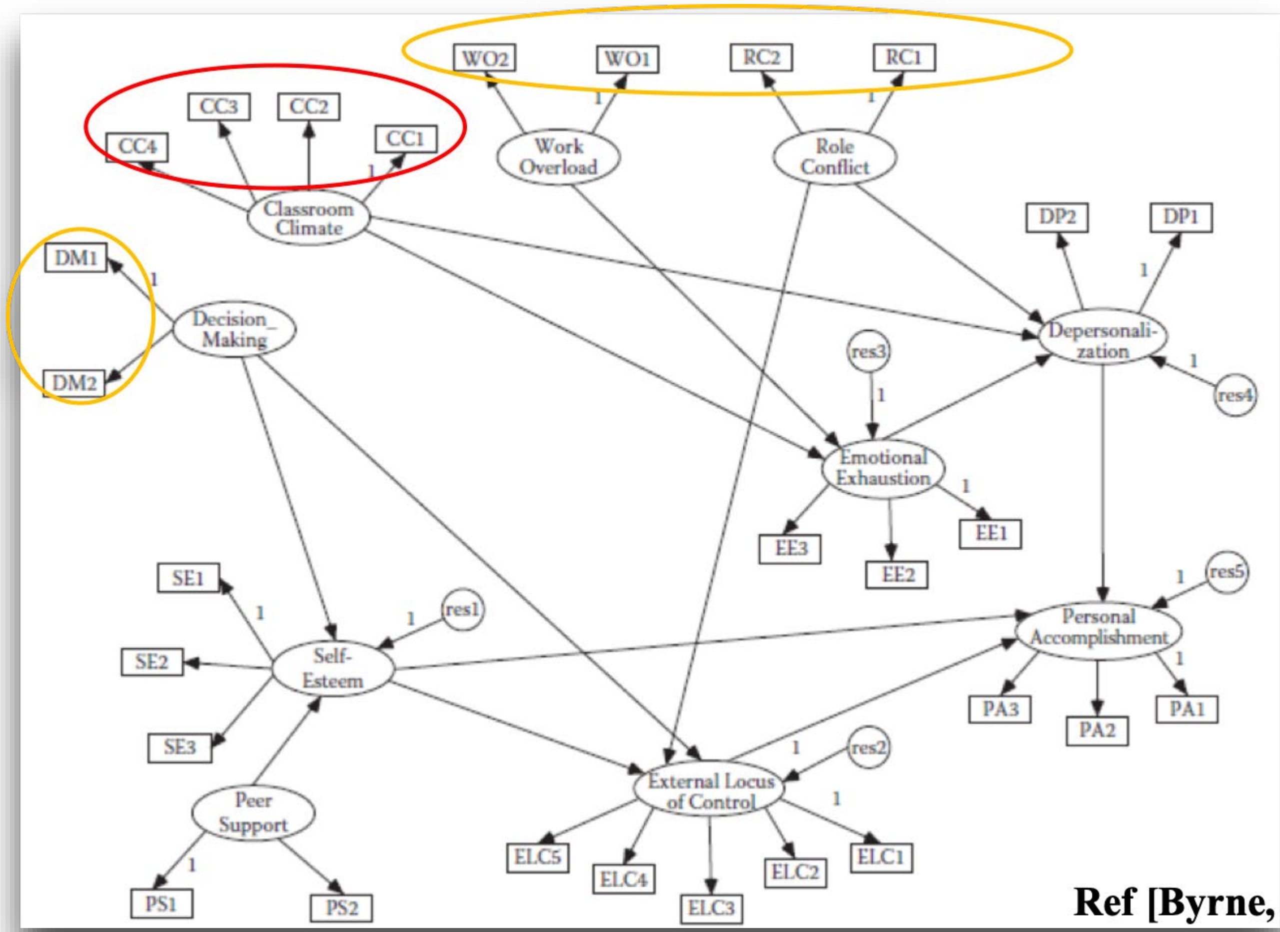
- Contains 28 measured variables
- Discovered clusters and causal order of the latent variables:

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*

Outline

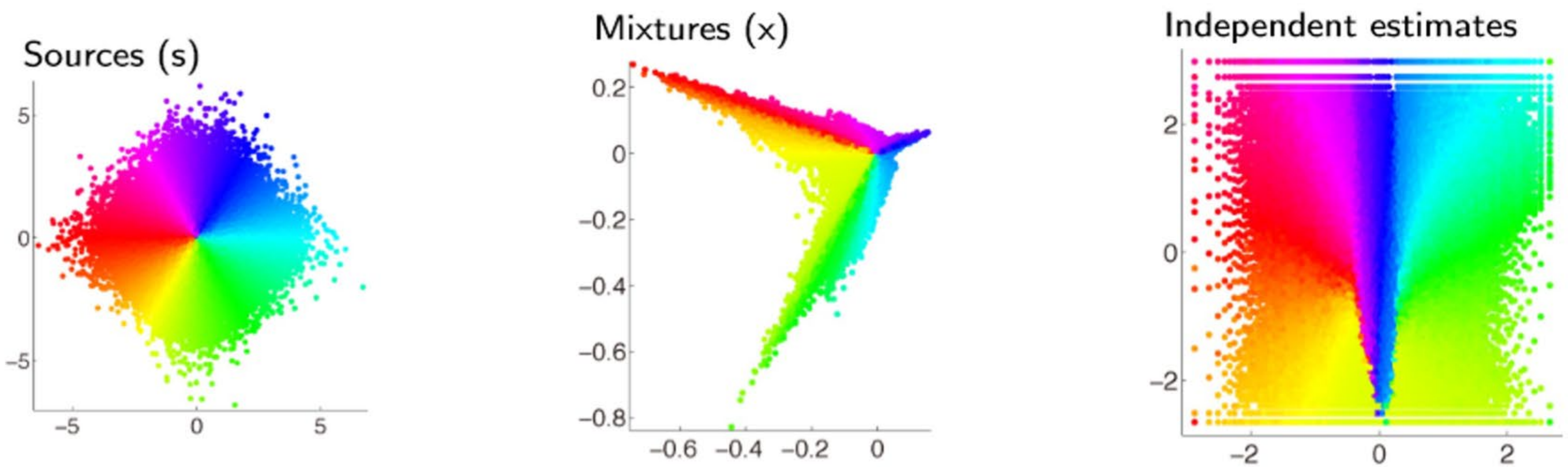
- Why?
- How?
  - IID case
    - Linear-Gaussian case
    - Linear, non-Gaussian case
    - **Nonlinear case**
  - From multiple distributions
  - With temporal information

Identifiability of nonlinear ICA: challenge

Is nonlinear ICA identifiable?

$$x = f(s)$$

No, it's ill-posed without further assumptions

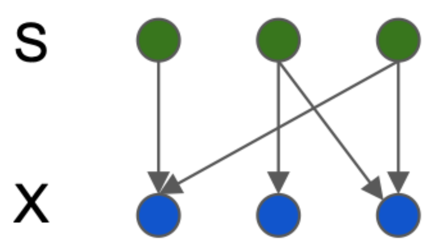


Identifiability of nonlinear ICA: auxiliary variables

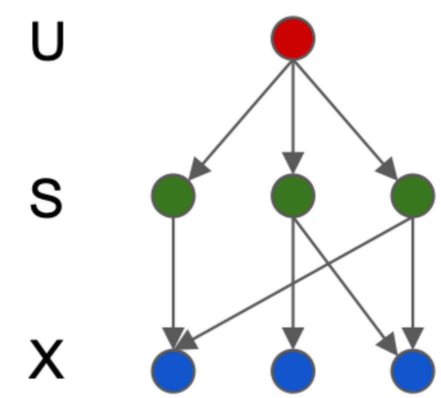
Independence alone is too weak → Conditional independence is strong enough

S1, S2, … , SN are marginally independent

S1, S2, … , SN are conditionally independent given an auxiliary variable U (e.g., domain index)



$$x = f(s)$$

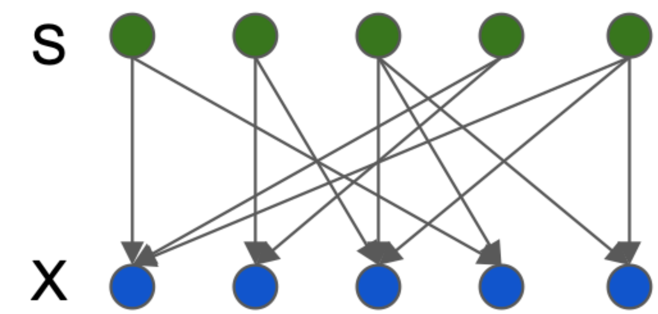


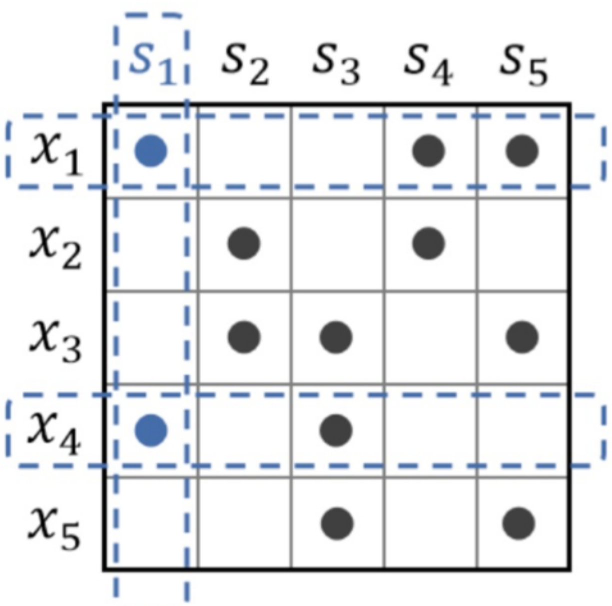
[Hyvarinen et al., Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning, AISTAT 2019]

Identifiability of nonlinear ICA: structural sparsity

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

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





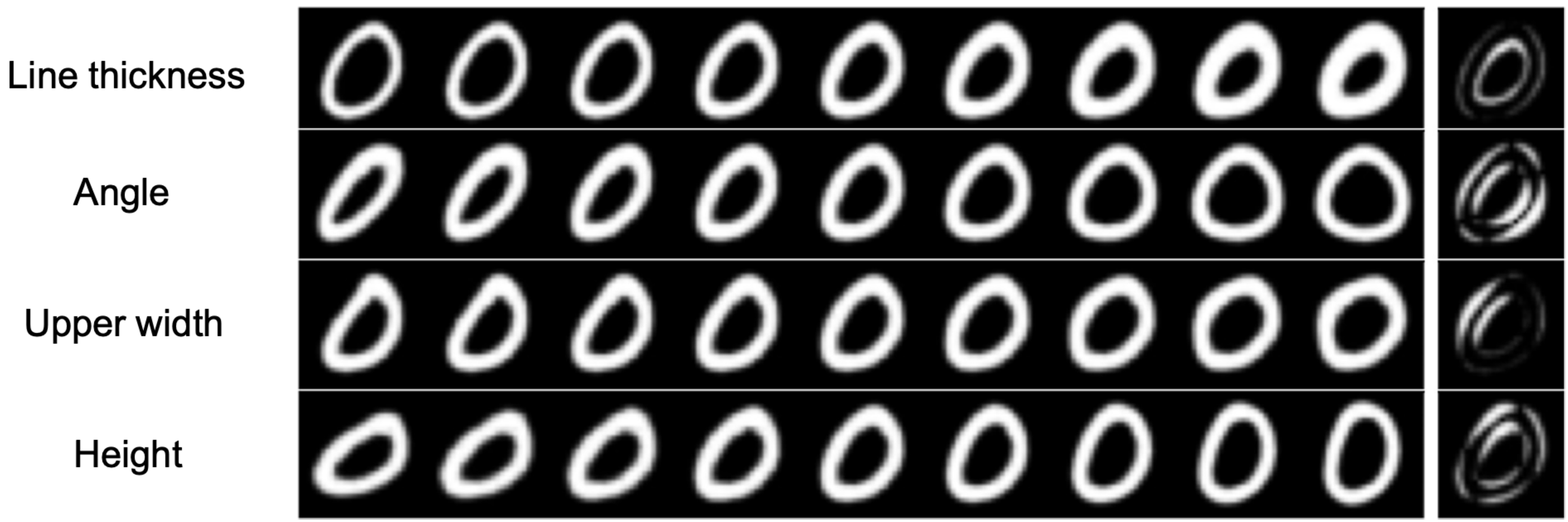
Graphically, for every latent source **s_i**, there exists a set of observed variable(s) such that the intersection of their/its parent(s) is **s_i**

Example: for **s_1**, there exists **x_1** and **x_4** such that the intersection of their parents is **s_1**

Failure: two sources influence the same set of observed variables

[Zheng et al., On the Identifiability of Nonlinear ICA: Sparsity and Beyond, NeurIPS 2022]

Identifiability of nonlinear ICA: real-world images



Identification results on EMNIST

Each row represents an identified source with its value varying

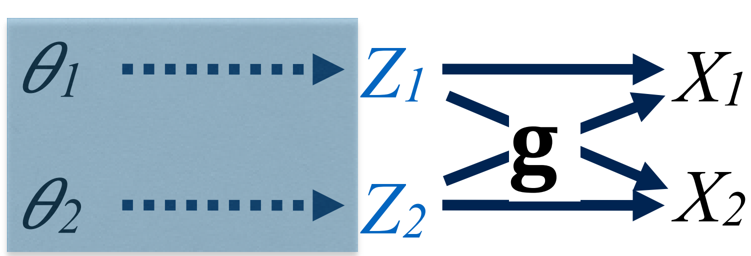
Outline

- Why?
- How?
  - IID case
    - Linear-Gaussian case
    - Linear, non-Gaussian case
    - Nonlinear case
  - **From multiple distributions**
  - With temporal information

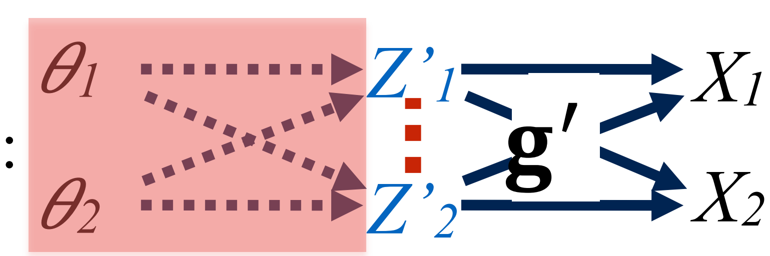
Nonlinear ICA with Multiple Domains

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

- Nonlinear ICA: observed variables follow $\mathbf{X} = \mathbf{g}(\mathbf{Z})$, in which Z_i are mutually independent
  - Solutions to nonlinear ICA high non-unique
  - If the dstr of each Z_i change across multiple domains, generally their are identifiable (up to component-wise transformations)
  - Why?



for $\mathbf{Z}' = h(\mathbf{Z})$:



- *Hyvärinen, Pajunen, Nonlinear independent component analysis: Existence and uniqueness results. Neural networks, 1999.*
- *Hyvarinen, Sasaki, Turner, "Nonlinear ICA using auxiliary variables and generalized contrastive learning," In The 22nd International Conference on Artificial Intelligence and Statistics, 2019.*

Partial Identifiability for Domain Adaptation

Lingjing Kong [1] **Shaoan Xie** [1] **Weiran Yao** [1] **Yujia Zheng** [1] **Guangyi Chen** [2 1] **Petar Stojanov** [3]
Victor Akinwande [1] **Kun Zhang** [2 1]

Abstract

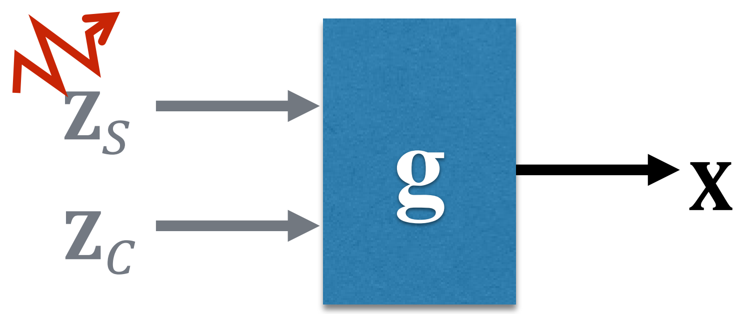
Unsupervised domain adaptation is critical to many real-world applications where label information is unavailable in the target domain. In general, without further assumptions, the joint distribution of the features and the label is not identifiable in the target domain. To address this issue, we rely on a property of minimal changes of causal mechanisms across domains to minimize unnecessary influences of domain shift. To encode this property, we first formulate the data generating process using a latent variable model with two partitioned latent subspaces: invariant components whose distributions stay the same across domains, and sparse changing components that vary across domains. We further constrain the domain shift to have a restrictive influence on the changing components. Under mild conditions, we show that the latent variables are partially identifiable, from

domain indices $\mathbf{u}$, the training (source domain) data follows multiple joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_1}$, $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_2}$, ..., $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_M}$,[1] and the test (target domain) data follows the joint distribution $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^{\mathcal{T}}}$, where $p_{\mathbf{x},\mathbf{y}|\mathbf{u}}$ may vary across $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$. During training, for each i-th source domain, we are given labeled observations $(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})_{k=1}^{m_i}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_i}$, and target domain unlabeled instances $(\mathbf{x}_k^{\mathcal{T}})_{k=1}^{m_{\mathcal{T}}}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^{\mathcal{T}}}$. The main goal of domain adaptation is to make use of the available observed information, to construct a predictor that will have optimal performance in the target domain.

It is apparent that without further assumptions, this objective is ill-posed. Namely, since the only available observations in the target domain are from the marginal distribution $p_{\mathbf{x}|\mathbf{u}^{\mathcal{T}}}$, the data may correspond to infinitely many joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^{\mathcal{T}}}$. This mandates making additional assumptions on the relationship between the source and the target domain distributions, with the hope to be able to reconstruct (identify) the joint distribution in the target domain $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^{\mathcal{T}}}$. Typically, these assumptions entail some measure of sim-

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

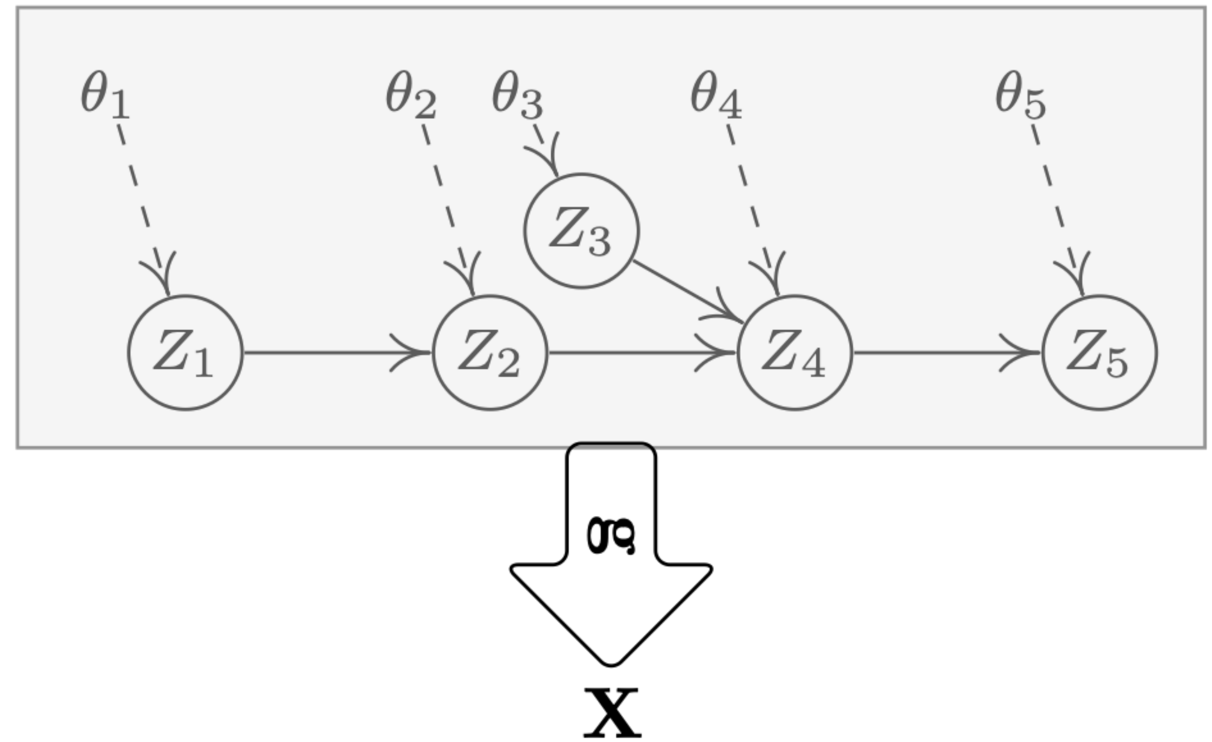
Models	→ Art	→ Clipart	→ Product	→ Realworld	Avg
Source Only (He et al., 2016)	64.58±0.68	52.32±0.63	77.63±0.23	80.70±0.81	68.81
DANN (Ganin et al., 2016)	64.26±0.59	58.01±1.55	76.44±0.47	78.80±0.49	69.38
DANN+BSP (Chen et al., 2019)	66.10±0.27	61.03±0.39	78.13±0.31	79.92±0.13	71.29
DAN (Long et al., 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*

Finding Hidden Variables With Changing Relations

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



- With sparsity of the graph over the estimated variables $\tilde{Z}_i$, with a suitable permutation over them, $\tilde{Z}_i$ is a function of Z_i and all Z_j that are adjacent to Z_i and all the other neighbors of Z_i in the Markov network
- Recovered DAG and the original DAG have the same topology
- θ_i can be recovered up to component-wise invertible transformations; so roughly speaking, Z_i can be recovered

- *Ongoing work*

Outline

- Why?
- How?
  - IID case
    - Linear-Gaussian case
    - Linear, non-Gaussian case
    - Nonlinear case
  - From multiple distributions
  - **With temporal information**

Temporally Disentangled Representation Learning

Weiran Yao
CMU
weiran@cmu.edu

Guangyi Chen
CMU & MBZUAI
guangyichen1994@gmail.com

Kun Zhang
CMU & MBZUAI
kunz1@cmu.edu

Abstract

Recently in the field of unsupervised representation learning, strong identifiability results for disentanglement of causally-related latent variables have been established by exploiting certain side information, such as class labels, in addition to independence. However, most existing work is constrained by functional form assumptions such as independent sources or further with linear transitions, and distribution assumptions such as stationary, exponential family distribution. It is unknown whether the underlying latent variables and their causal relations are identifiable if they have arbitrary, nonparametric causal influences in between. In this work, we establish the identifiability theories of nonparametric latent causal processes from their nonlinear mixtures under fixed temporal causal influences and analyze how distribution changes can further benefit the disentanglement. We propose **TDRL**, a principled framework to recover time-delayed latent causal vari-

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

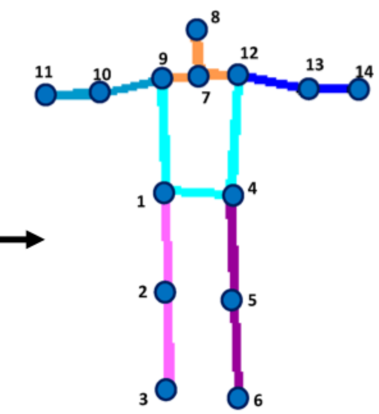
$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t$

Latent processes

Latent temporal causal processes z_{it} can be recovered if they follow

- completely **nonparametric** model; or furthermore,
- **non-stationary** noise; or
- **non-stationary** causal influence, or
- **Parametric** constraints

Causal Skeleton Recovery



Recovered latent processes

$$\underbrace{\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t)}_{\text{Nonlinear mixing}}, \quad \underbrace{z_{it} = f_i\left(\{z_{j,t-\tau} \mid z_{j,t-\tau} \in \mathbf{Pa}(z_{it})\}, \epsilon_{it}\right)}_{\text{Stationary nonparametric transition}} \; with \; \underbrace{\epsilon_{it} \sim p_{\epsilon_i}}_{\text{Stationary noise}} .$$

- *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*

Comparisons

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

Table 1: Attributes of nonlinear ICA theories for time-series. A check denotes that a method has an attribute or can be applied to a setting, whereas a cross denotes the opposite. † indicates our approach.

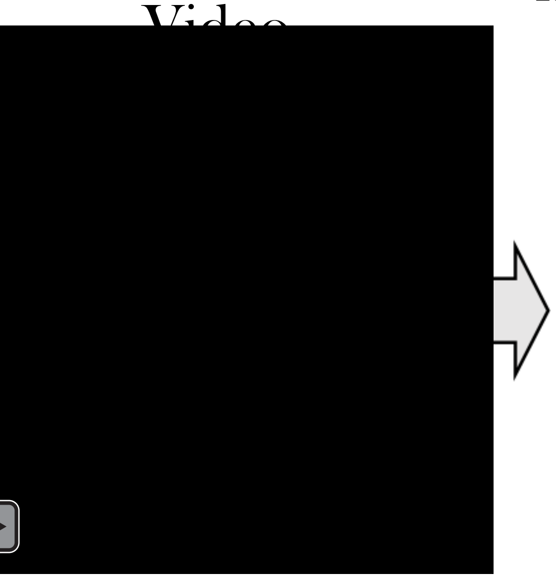
Theory	Time-varying Relation	Causally-related Process	Partitioned Subspace	Nonparametric Transition	Applicable to Stationary Environment
PCL	✗	✗	✗	✓	✓
GCL	✓	✗	✗	✓	✓
HM-NLICA	✗	✗	✗	✗	✗
SlowVAE	✗	✗	✗	✗	✓
SNICA	✓	✓	✗	✗	✗
i-VAE	✓	✗	✗	✗	✗
LEAP	✗	✓	✗	✓	✗
TDRL †	✓	✓	✓	✓	✓

- *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*

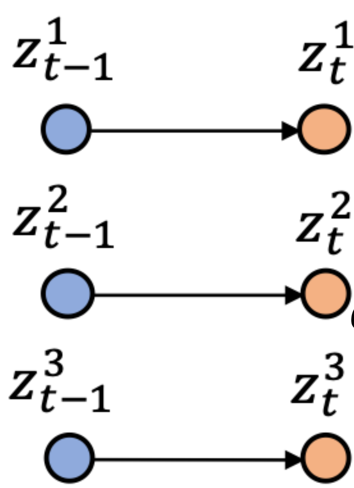
Results on Video Data

- For easy interpretation, consider two simple video data sets
  - KiTTiMask: a video dataset of binary pedestrian masks
  - Mass-spring system: a video dataset with ball movement and invisible springs

KiTTiMask Video



Learned latent processes



Interpretation

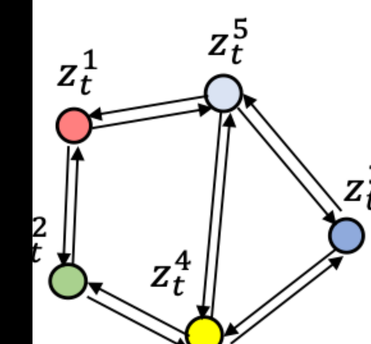
(Movement in a direction)

(Movement in an orthogonal direction)

(Size)

Mass-spring

Learned latent processes



Interpretation

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

- *Yao, Chen, Zhang, "Learning Latent Causal Dynamics," NeurIPS 2022*
- *Yao, Sun, Ho, Sun, Zhang, "Learning Temporally causal latent processes from general temporal data," ICLR 2022*

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

Summary

- Essential to learn hidden causal variables in many cases!
- Possible to achieve even in the IID case
- Benefit from distribution changes and temporal information
- Future work
  - Efficient procedure?
  - Necessary and sufficient identifiability conditions?
  - Changing relations among hidden variables?