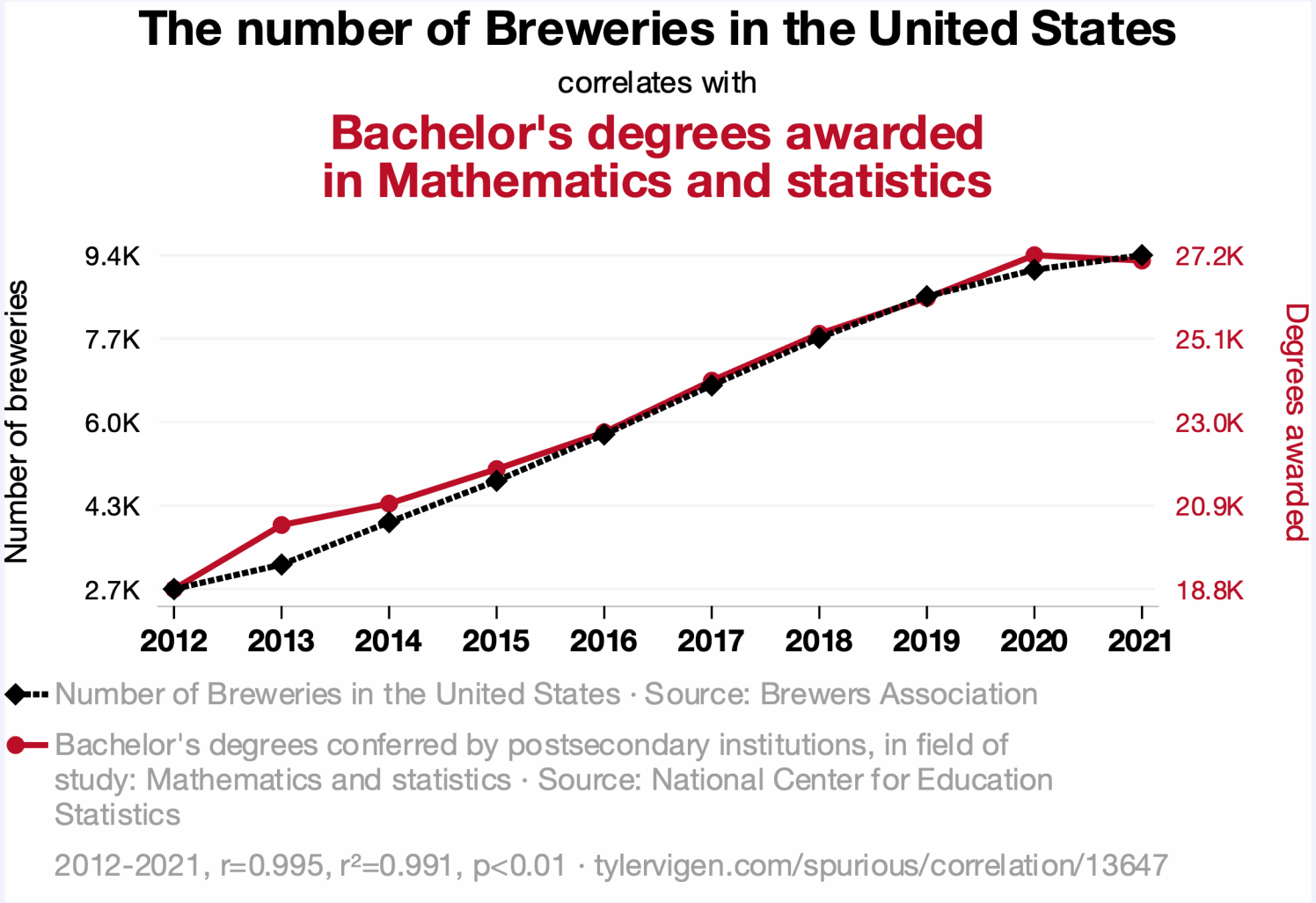
Causal Discovery in a Non-Ideal World

Tutorial website
with slides
https://eda.rg.cispa.io/events/cdrw25sdm/

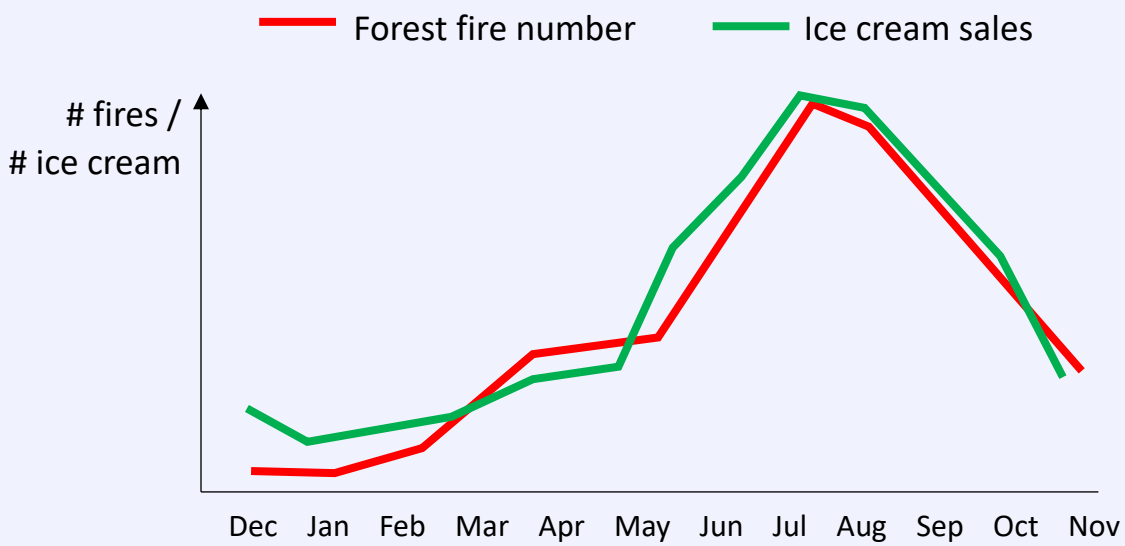
Lénaïg Cornanguer
David Kaltenpoth
Sarah Mameche
Jilles Vreeken

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

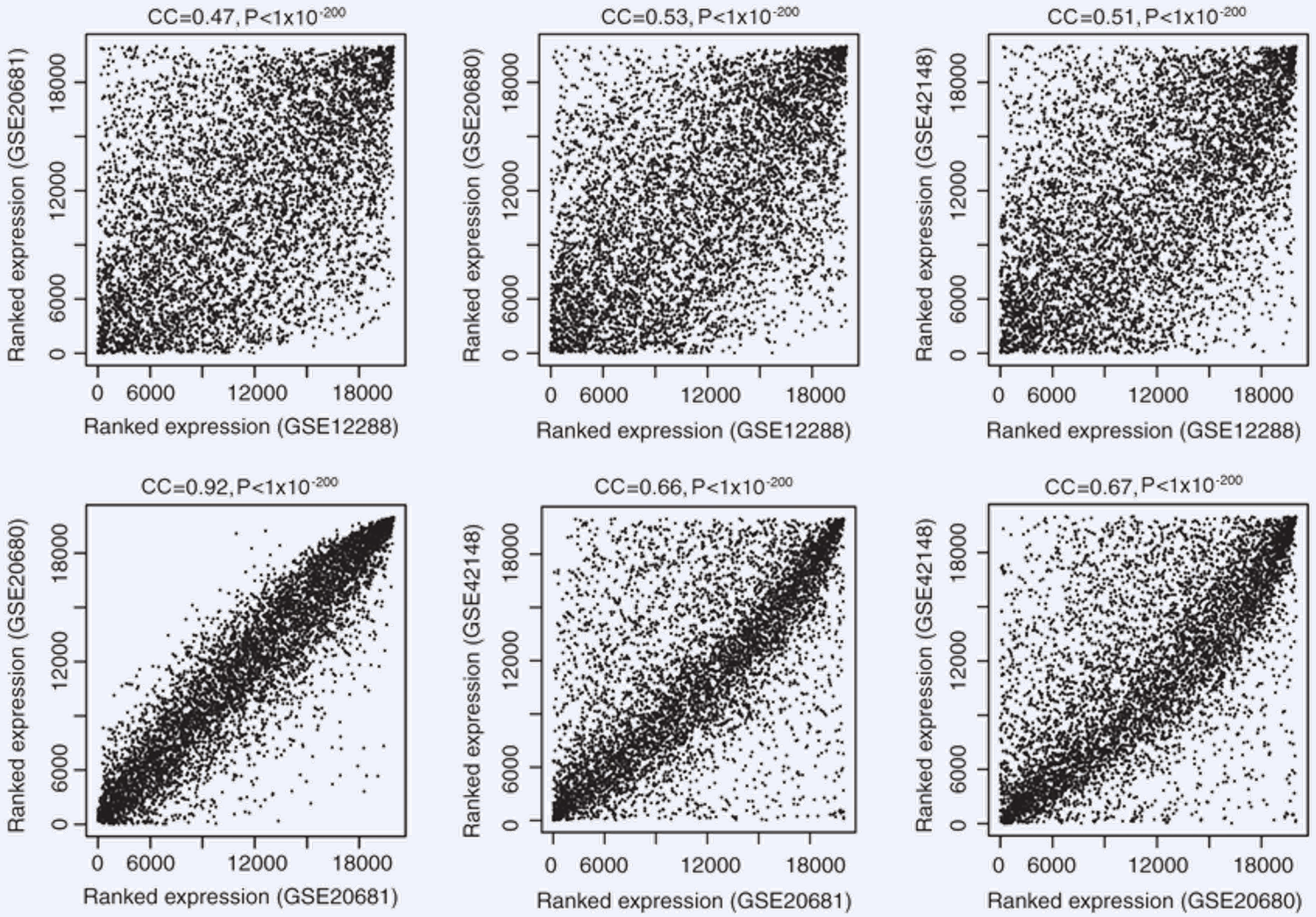
Correlation vs. causation



Correlation vs. causation



Correlation vs. causation



Correlation analysis of genes in four datasets from the Gene Expression Omnibus database. CC, correlation coefficient.

(Shilin Yan, "Integrative analysis of promising molecular biomarkers and pathways for coronary artery disease using WGCNA and MetaDE methods", 2018.)

Why causality

Decision making

Why causality

(Beery, Van Horn, and Perona, 2018)

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Why causality

(Beery, Van Horn, and Perona, 2018)

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Building models able to understand phenomena instead of predicting them at the cost of energy, data, and material resources (precious metals, water…)

How (not to discover) causality

Simpson's paradox

Condition	Treatment A	Treatment B
All	78% (273/350)	**83%** (289/350)

Recovery rate for kidney stones in function of the stone size and the treatment

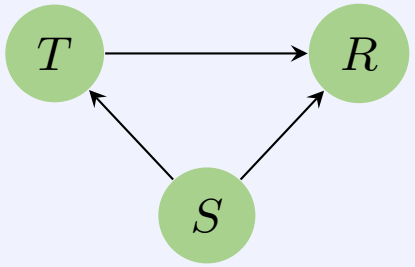
$\boldsymbol{T}$ Treatment
$\boldsymbol{R}$ Recovery

How (not to discover) causality

Simpson's paradox

Condition	Treatment A	Treatment B
Small stones	**93%** (81/87)	87% (234/270)
Large stones	**79%** (192/263)	69% (55/80)
All	78% (273/350)	**83%** (289/350)

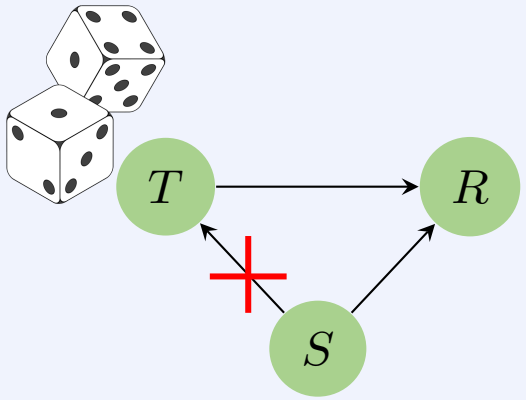
Recovery rate for kidney stones in function of the stone size and the treatment



T Treatment
R Recovery
S Size of stones

How (to discover) causality

Randomized controlled trials (RCT)



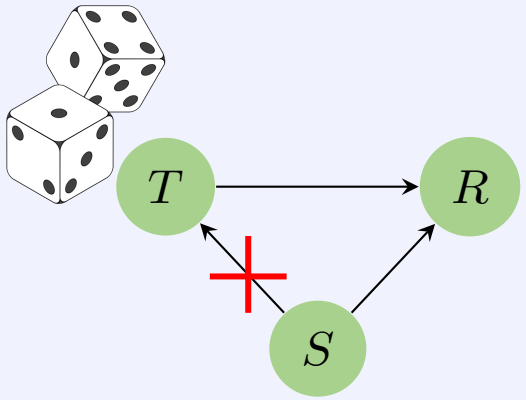
T Treatment
R Recovery
S Size of stones

$$P(R|T = t, S)$$

Observational distribution

How (to discover) causality

Randomized controlled trials (RCT)



T Treatment
R Recovery
S Size of stones

$$P(R|T=t,S) \quad \rightarrow \quad P(R|do(T=t),S)$$

Observational distribution

Interventional distribution

Intervention on variable X: "breaking" any incoming arrow of X and setting X to a particular value x

Do-operator: mathematical representation of an intervention

Intervention

Intervention

How (to discover) causality

Interventions are not always possible/ethical

(… although interventions exist in the wild)

Causal discovery from observational data

How (to discover) causality

Interventions are not always possible/ethical

(… although interventions exist in the wild)

Causal discovery from observational data

in a non-ideal world

Question of Interest

How to formalize causality?

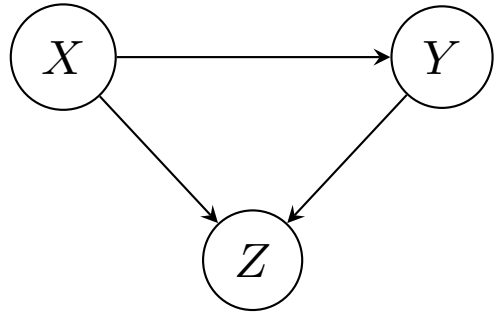
Causal model

Functional causal model (FCM)

$$X := \varepsilon_X$$
$$Y := f_Y(X, \varepsilon_Y)$$
$$Z := f_Z(X, Y, \varepsilon_Z)$$

with noise terms $\boldsymbol{\varepsilon}$ jointly independent

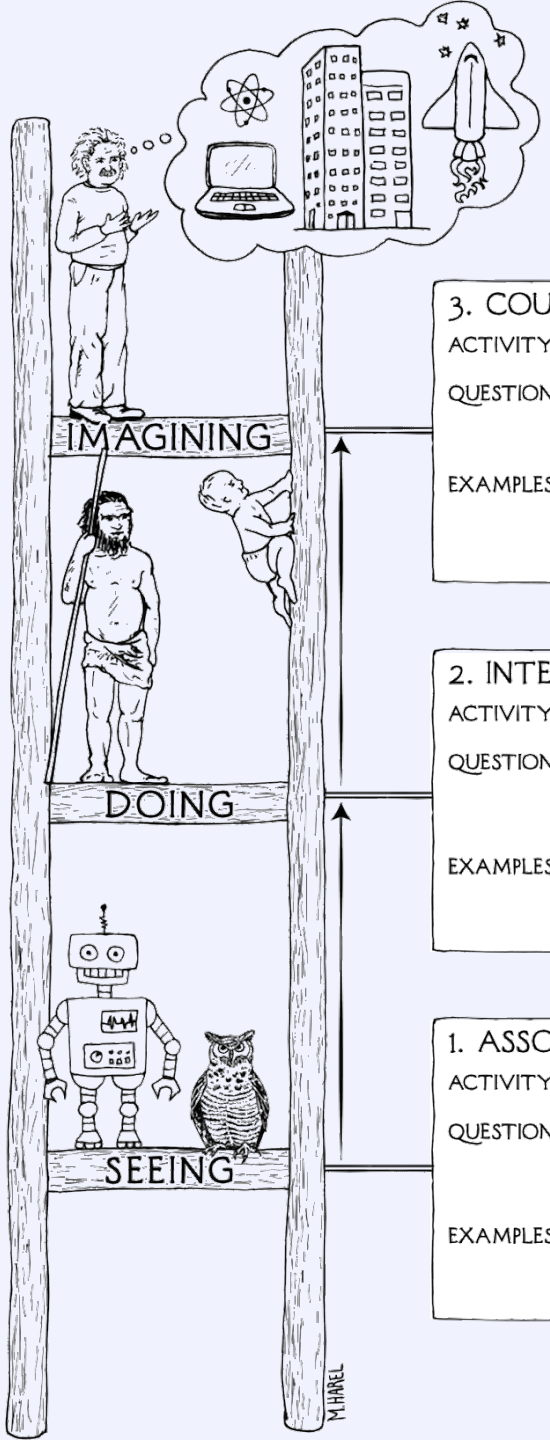
Directed acyclic graph (DAG)



Joint probability distribution

$$P(X, Y, Z)$$

Causality ladder



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

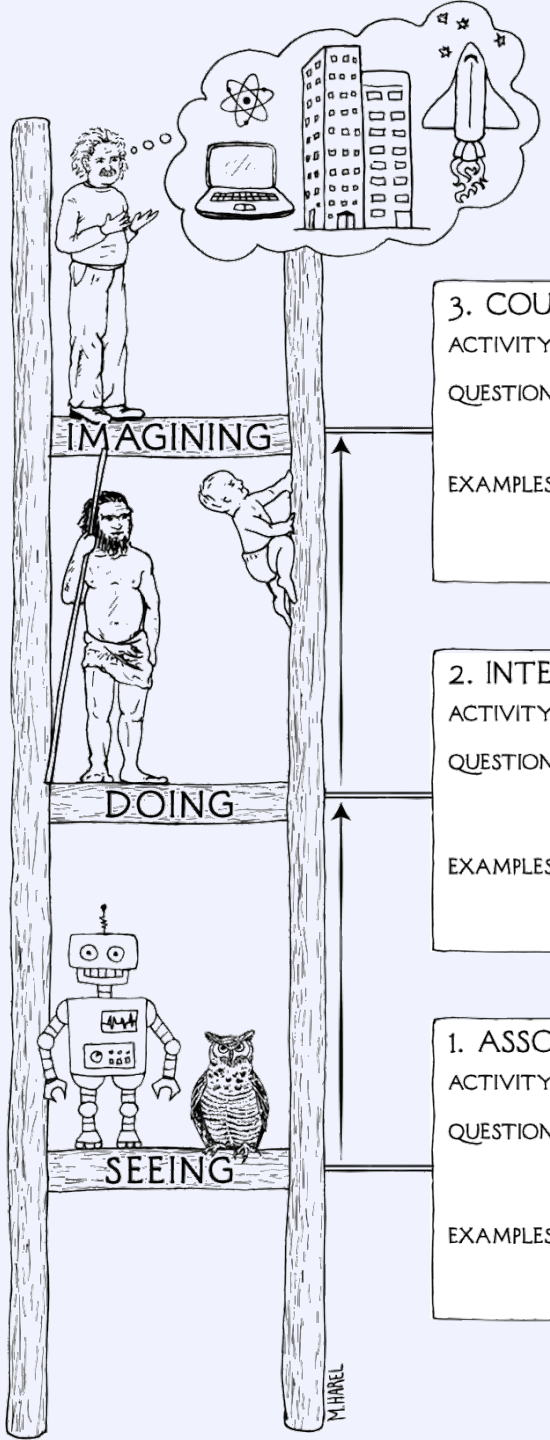
ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

(Pearl, 2000)

Causality ladder



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

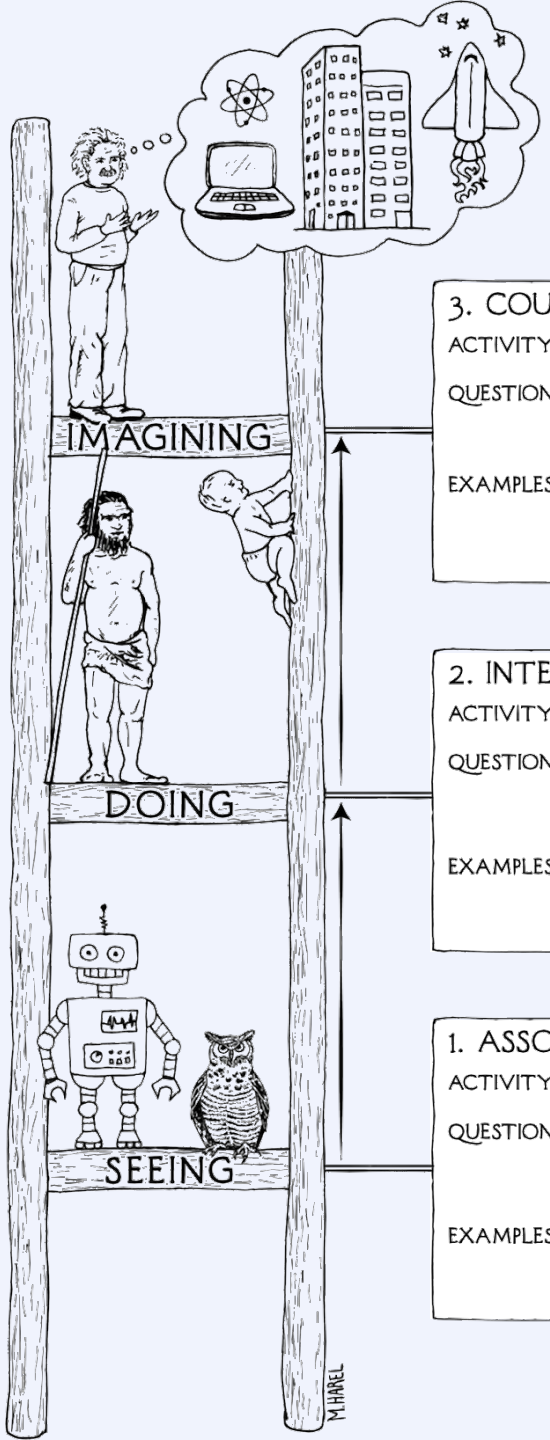
data

↓

prediction, anomaly detection

(Pearl, 2000)

Causality ladder



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

causal graph (DAG)

↓

decision making, root cause analysis

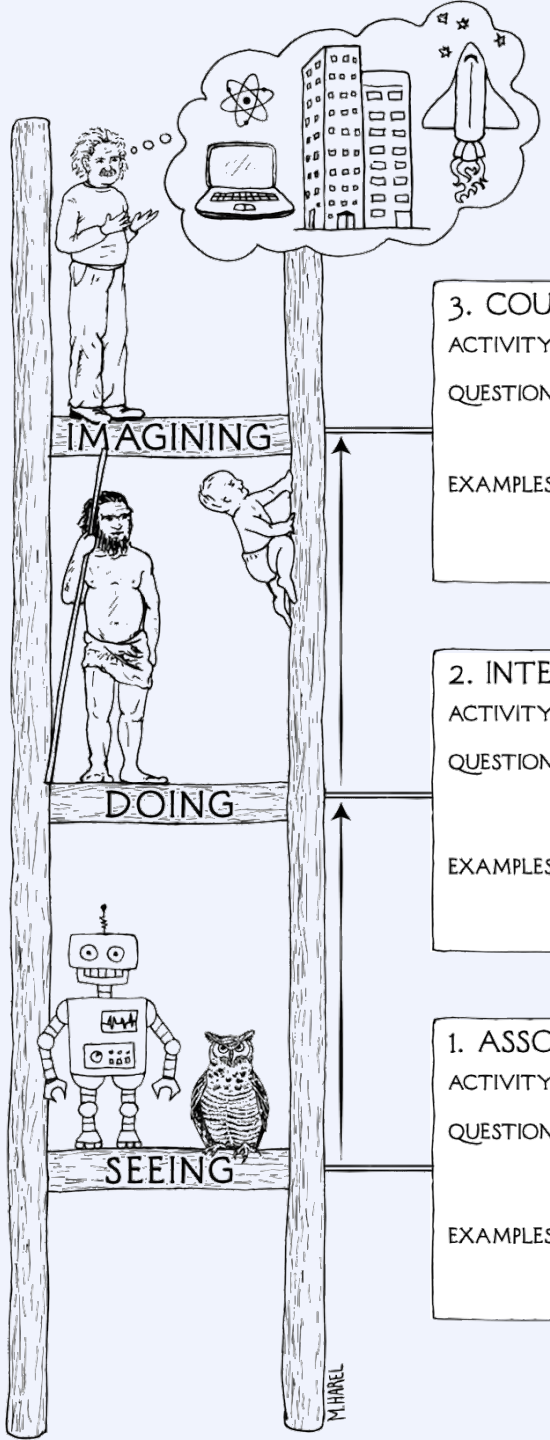
data

↓

prediction, anomaly detection

(Pearl, 2000)

Causality ladder



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?* (What would Y be if I do X? How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?* (How are the variables related? How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

equations (FCM)

↓

fairness, counterfactual reasoning, individual level

causal graph (DAG)

↓

decision making, root cause analysis

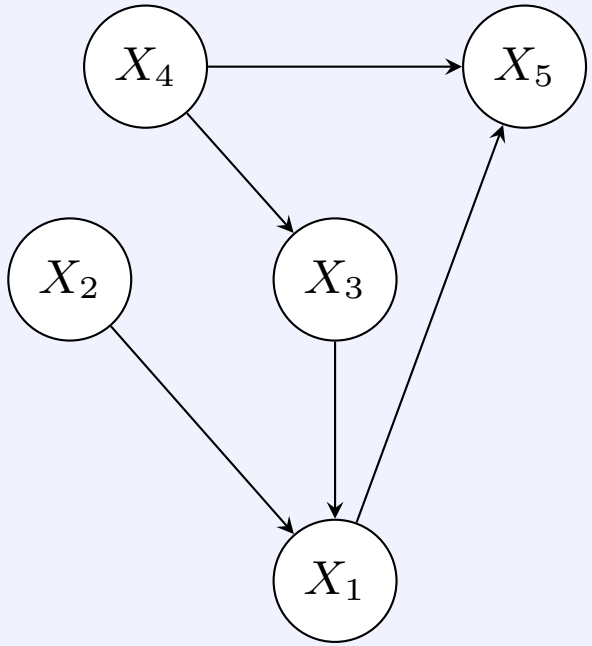
data

↓

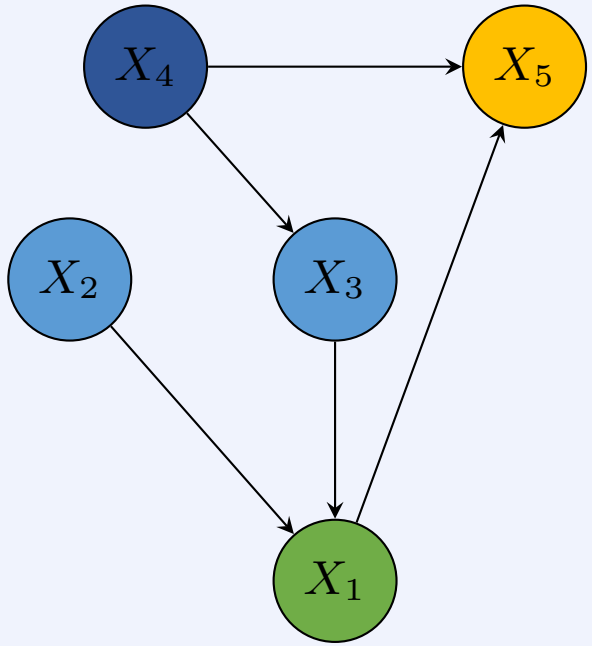
prediction, anomaly detection

(Pearl, 2000)

Directed Acyclic Graph (DAG)



Directed Acyclic Graph (DAG)

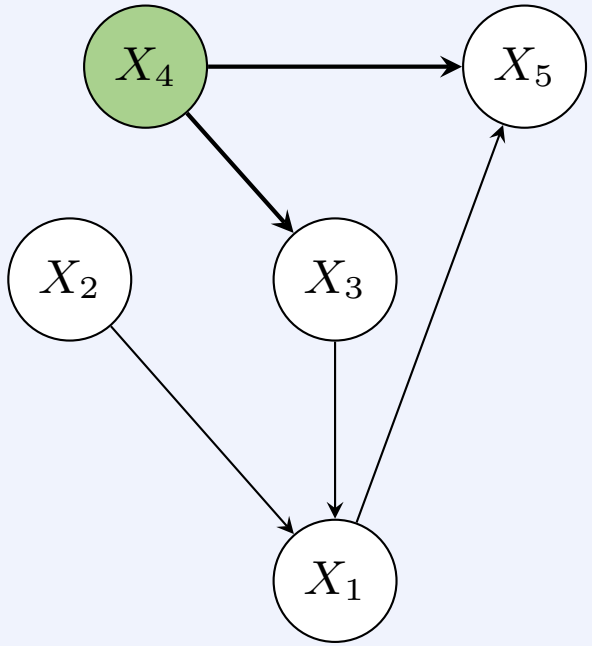


parents $pa(X_1) = \{X_2, X_3\}$

children

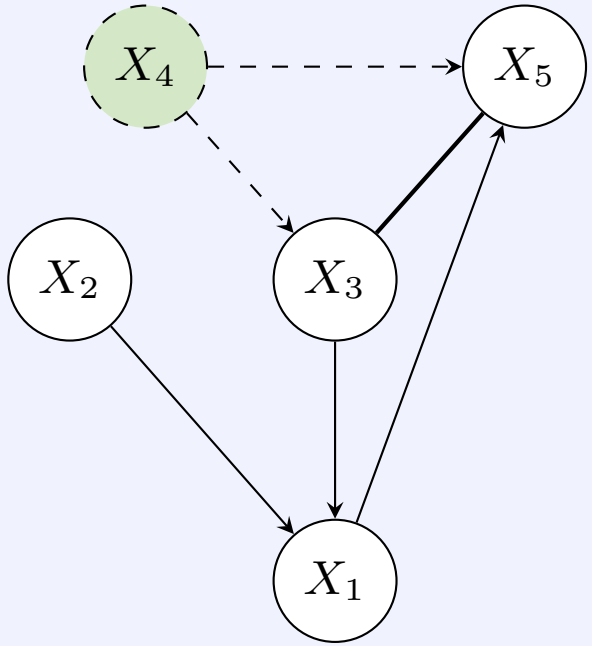
descendants and ancestors

Directed Acyclic Graph (DAG)



confounder

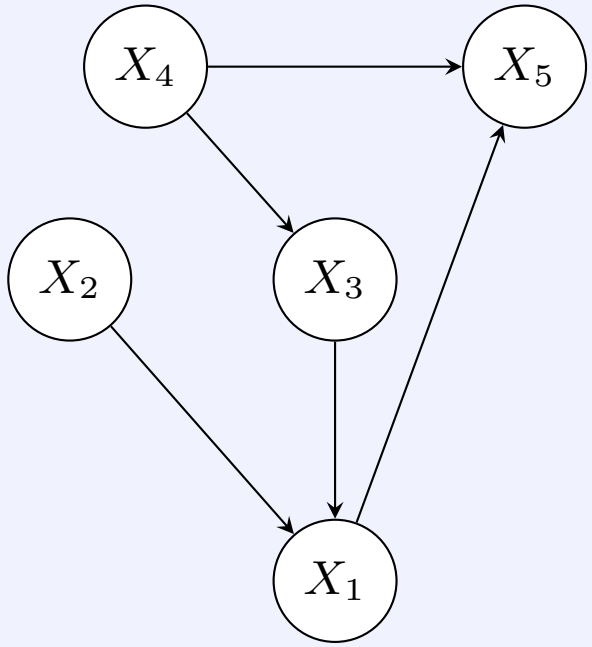
Directed Acyclic Graph (DAG)



hidden

confounder

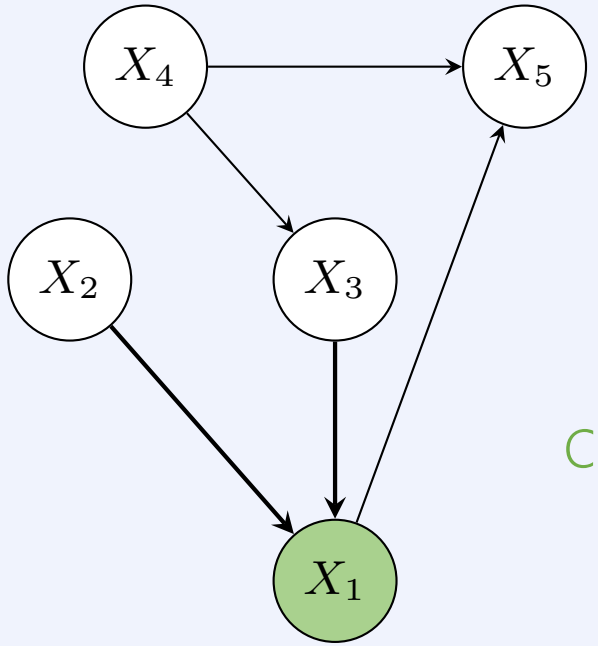
Assumption



Sufficiency

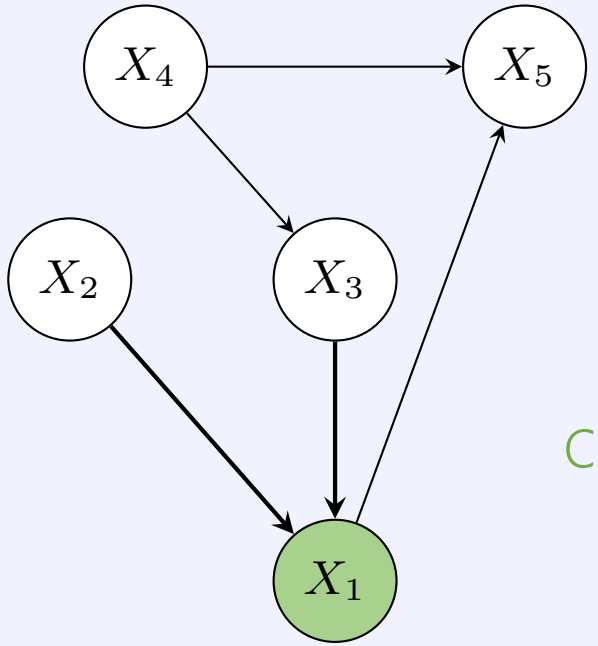
All confounders of the observed variables are also observed

Directed Acyclic Graph (DAG)

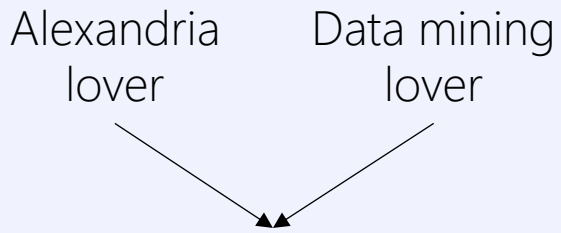


collider

Directed Acyclic Graph (DAG)



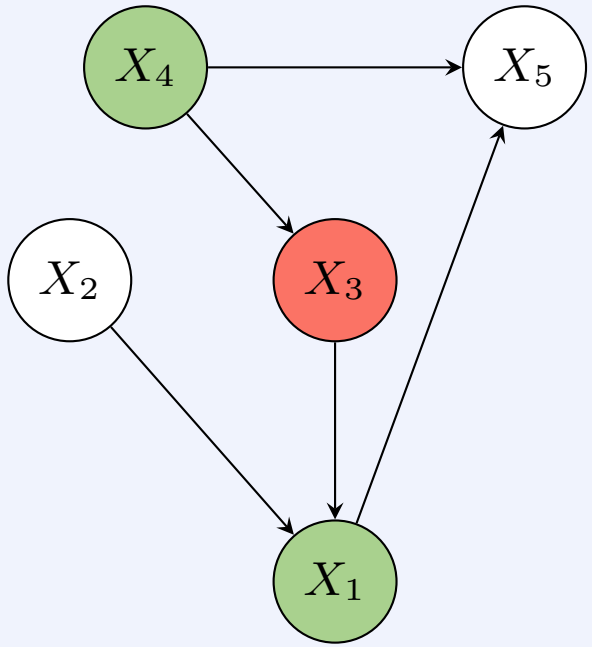
collider



selection bias

d-separation

X_i and X_j are d-separated by S if all paths between X_i and X_j are blocked by S

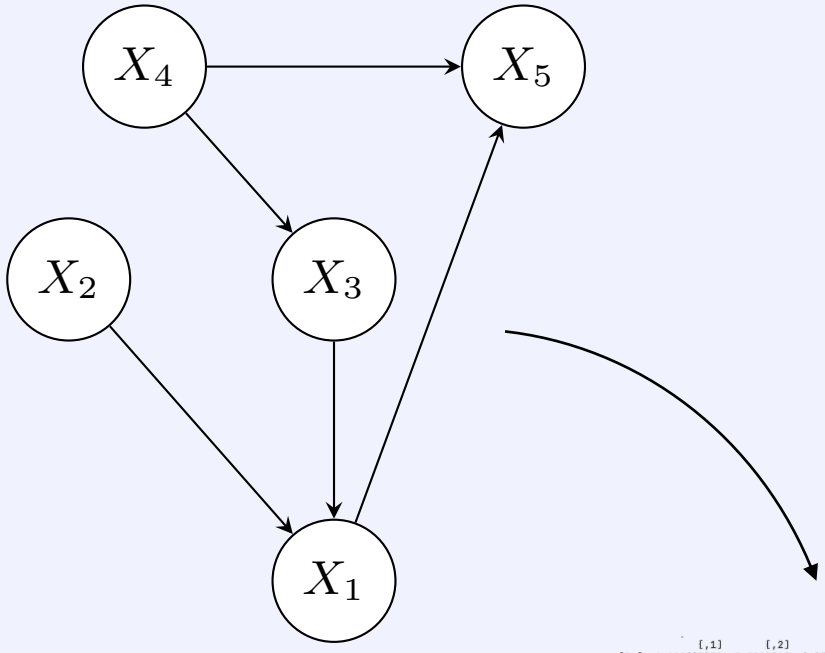


X_4 and X_1 are d-sep. by $\{X_3\}$

Rules

- ● … → ● → … ● blocks a path
- ● … ← ● → … ● blocks a path
- ● … → ● ← … ● blocks a path

Assumption



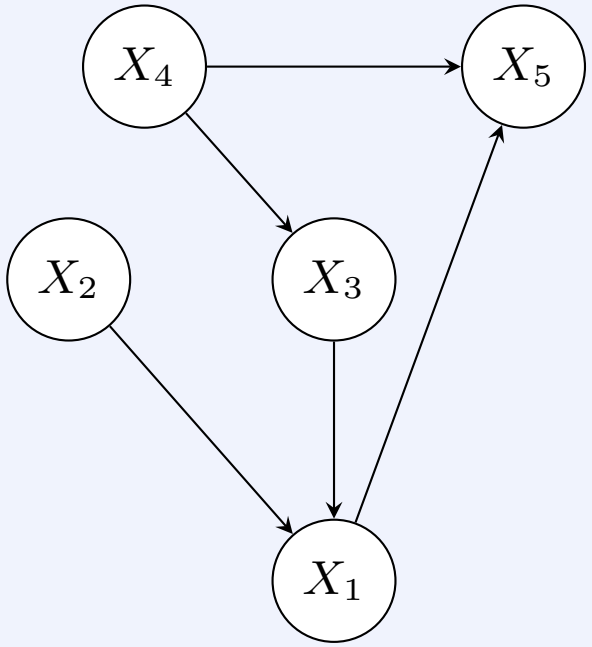
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-1.111280231	-2.0140969	0.38660442	0.05846330	0.5205943	2.40911019
[2,]	1.417817353	-0.3615978	-0.19247032	0.66140629	-0.1432120	-0.11834670
[3,]	-0.570567540	-1.6432378	-0.01828731	0.63376433	1.0635629	1.38168120
[4,]	-0.266227679	0.3599688	-0.24996129	-0.71842864	1.3101086	-0.12842456
[5,]	-0.427288260	-0.2755770	0.18415136	-0.38498679	0.7879288	0.17226519
[6,]	1.071823011	-2.2669731	-0.12903350	1.20311317	-0.9858127	-0.80158289
[7,]	0.837535622	1.1515241	1.59051510	0.38925330	0.1345126	-0.67640590
[8,]	-0.390993411	-1.1961786	-0.39611883	-0.03885206	0.6040686	-1.48233781
[9,]	0.362079425	-0.1536282	-0.07836638	0.35483976	-0.7917826	1.03274031
[10,]	0.458338530	-0.0165398	-2.03619702	-0.52135067	-0.4390771	1.20154780
[11,]	0.501343446	0.2389414	0.29264235	2.22713490	-1.0410120	-0.89328211
[12,]	-1.415642964	-0.1702699	2.38358494	-0.81265492	-0.6158825	1.26850073
[13,]	-0.046928402	-0.3022692	1.13007307	0.42498056	-0.1353464	-0.32156204
[14,]	-0.102232153	1.2782075	0.04981187	-0.20025751	-0.3551035	0.96481313
[15,]	1.341928249	0.1602453	-2.00424050	0.73607678	-0.7738258	-1.23018988
[16,]	0.379343237	0.8455179	0.38334824	-1.10415371	1.3109847	0.51595299
[17,]	0.992962014	-0.1822972	-0.62581816	-0.24508326	-1.0401618	-0.40046472
[18,]	0.148449812	1.8961460	-1.80999444	1.15871379	-0.4712393	-0.11946830
[19,]	0.343098853	-0.8892800	-0.99248067	1.25076084	-1.3800977	-0.49034137
[20,]	-0.694376265	1.0474346	-1.18596211	0.58955030	-0.1164544	-0.60899072
[21,]	-0.228495189	-0.2954567	-0.71869073	-0.45818747	-0.1463725	0.10861868
[22,]	0.452582822	1.2291624	1.93100711	1.28179874	0.5874635	-1.11419976
[23,]	0.935567535	-0.2807363	-2.28854793	-0.80001996	0.2223043	0.34980701
[24,]	0.894893812	1.6273959	0.49487719	0.83645987	1.2652432	-0.56321515
[25,]	0.007212357	-1.5697742	1.94262455	-1.32507779	0.5770311	-0.27249976
[26,]	-1.662708965	0.1443786	1.40188962	0.86200639	0.6357342	0.55804169
[27,]	-1.198312728	0.1429584	-0.47443741	0.48876408	0.6965944	-0.94838542

Markov condition

each node is independent of its non-descendants given its parents

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P\left(X_i | pa(X_i)\right)$$

Assumption



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-1.111280231	-2.0140969	0.38660442	0.05846330	0.5205943	2.40911019
[2,]	1.417817353	-0.3615978	-0.19247032	0.66140629	-0.1432120	-0.11834670
[3,]	-0.570567540	-1.6432378	-0.01828731	0.63376433	1.0635629	1.38168120
[4,]	-0.266227679	0.3599688	-0.24996129	-0.71842864	1.3101086	-0.12842456
[5,]	-0.427288260	-0.2755770	0.18415136	-0.38498679	0.7879288	0.17226519
[6,]	1.071823011	-2.2669731	-0.12903350	1.20311317	-0.9858127	-0.80158289
[7,]	0.837535622	1.1515241	1.59051510	0.38925330	0.1345126	-0.67640590
[8,]	-0.390993411	-1.1961786	-0.39611883	-0.03885206	0.6040686	-1.48233781
[9,]	0.362079425	-0.1536282	-0.07836638	0.35483976	-0.7917826	1.03274031
[10,]	0.458338530	-0.0165398	-2.03619702	-0.52135067	-0.4390771	1.20154780
[11,]	0.501343446	0.2389414	0.29264235	2.22713490	-1.0410120	-0.89328211
[12,]	-1.415642964	-0.1702699	2.38358494	-0.81265492	-0.6158825	1.26850073
[13,]	-0.046928402	-0.3022692	1.13007307	0.42498056	-0.1353464	-0.32156204
[14,]	-0.102232153	1.2782075	0.04981187	-0.20025751	-0.3551035	0.96481313
[15,]	1.341928249	0.1602453	-2.00424050	0.73607678	-0.7738258	-1.23018988
[16,]	0.379343237	0.8455179	0.38334824	-1.10415371	1.3109847	0.51595299
[17,]	0.992962014	-0.1822972	-0.62581816	-0.24508326	-1.0401618	-0.40046472
[18,]	0.148449812	1.8961460	-1.80999444	1.15871379	-0.4712393	-0.11946830
[19,]	0.343098853	-0.8892800	-0.99248067	1.25076084	-1.3800977	-0.49034137
[20,]	-0.694376265	1.0474346	-1.18596211	0.58955030	-0.1164544	-0.60899072
[21,]	-0.228495189	-0.2954567	-0.71869073	-0.45818747	-0.1463725	0.10861868
[22,]	0.452582822	1.2291624	1.93100711	1.28179874	0.5874635	-1.11419976
[23,]	0.935567535	-0.2807363	-2.28854793	-0.80001996	0.2223043	0.34980701
[24,]	0.894893812	1.6273959	0.49487719	0.83645987	1.2652432	-0.56321515
[25,]	0.007212357	-1.5697742	1.94262455	-1.32507779	0.5770311	-0.27249976
[26,]	-1.662708965	0.1443786	1.40188962	0.86200639	0.6357342	0.55804169
[27,]	-1.198312728	0.1429584	-0.47443741	0.48876408	0.6945944	-0.94838542

Faithfulness

all and only the conditional independence relationships present in the probability distribution are due to d-separation in the causal graph

Question of Interest

How to discover causality in an ideal world?

Causal discovery

Constraint-based methods

FCM-based methods

Score-based methods

Continuous optimization-based methods

Constraint-based methods

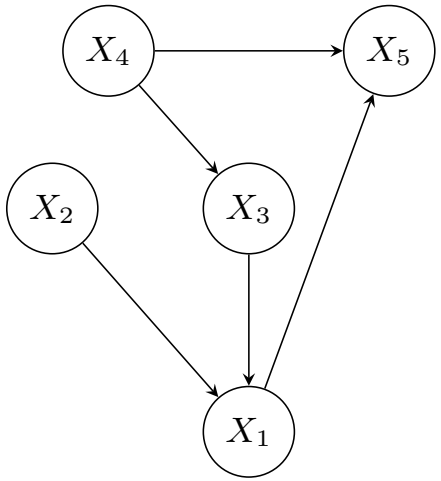
Observational distribution

$$P(X_1, X_2, ...)$$

Conditional independence tests

$$X_2 \perp\!\!\!\perp X_5 \mid X_1, X_4$$
$$X_4 \perp\!\!\!\perp X_1 \mid X_3$$
$$X_2 \perp\!\!\!\perp X_4$$
$$...$$

Graphical causal model



PC algorithm[1]

Find all (conditional) independence in data

Select the DAG(s) having the corresponding d-separation

[1](Spirtes, Glymour, Scheines, 2000)

Constraint-based methods

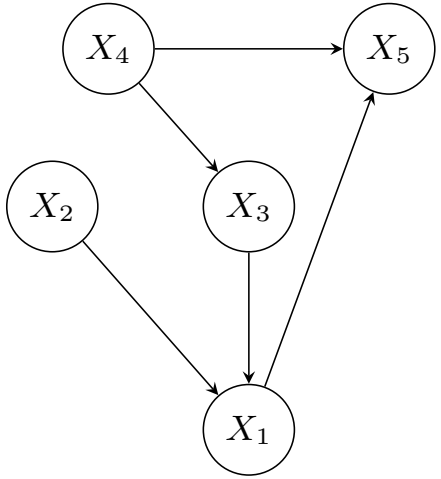
Observational distribution

$$P(X_1, X_2, ...)$$

Conditional independence tests

$$X_2 \perp\!\!\!\perp X_5 \mid X_1, X_4$$
$$X_4 \perp\!\!\!\perp X_1 \mid X_3$$
$$X_2 \perp\!\!\!\perp X_4$$
$$...$$

Graphical causal model



PC algorithm[1]

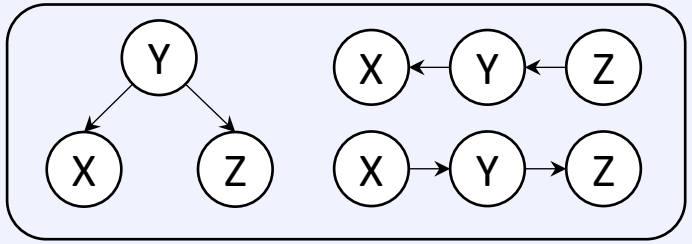
Find all (conditional) independence in data

Select the DAG(s) having the corresponding d-separation

Limitations:

- Computational efficiency
- Partially oriented graph
- Multiple testing problem
- Noise sensitivity

Markov Equivalence Class (MEC) for $\mathbf{X \perp\!\!\!\perp Z \mid Y}$



[1](Spirtes, Glymour, Scheines, 2000)

Causal discovery

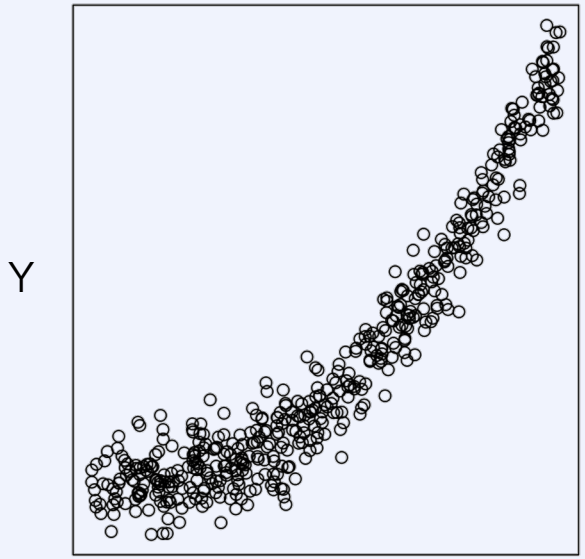
Constraint-based methods

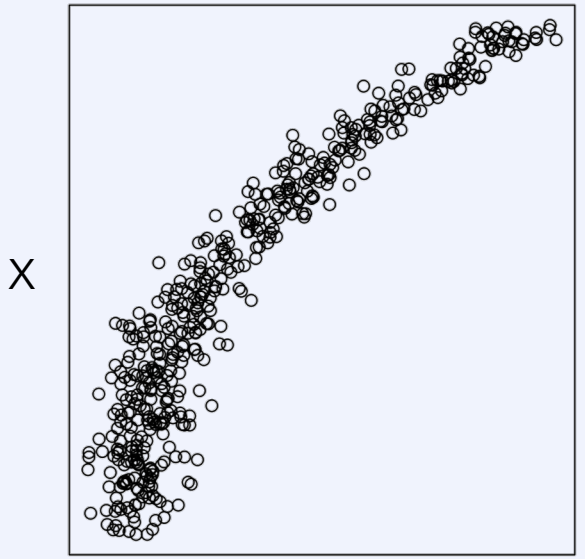
FCM-based methods

Score-based methods

Continuous optimization-based methods

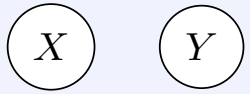
Functional causal model-based methods



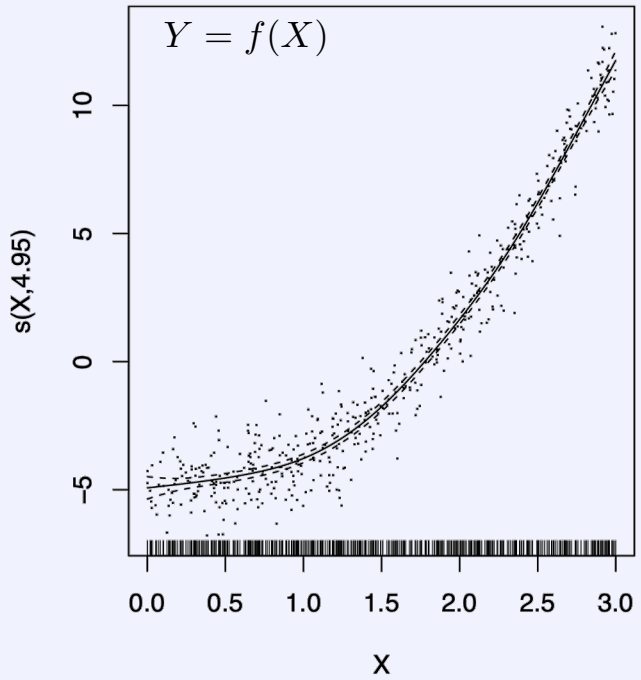


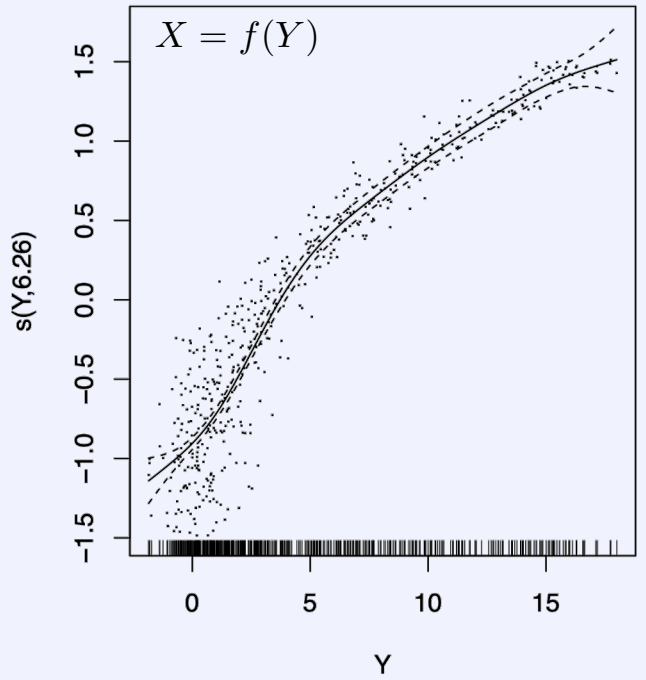
$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y$

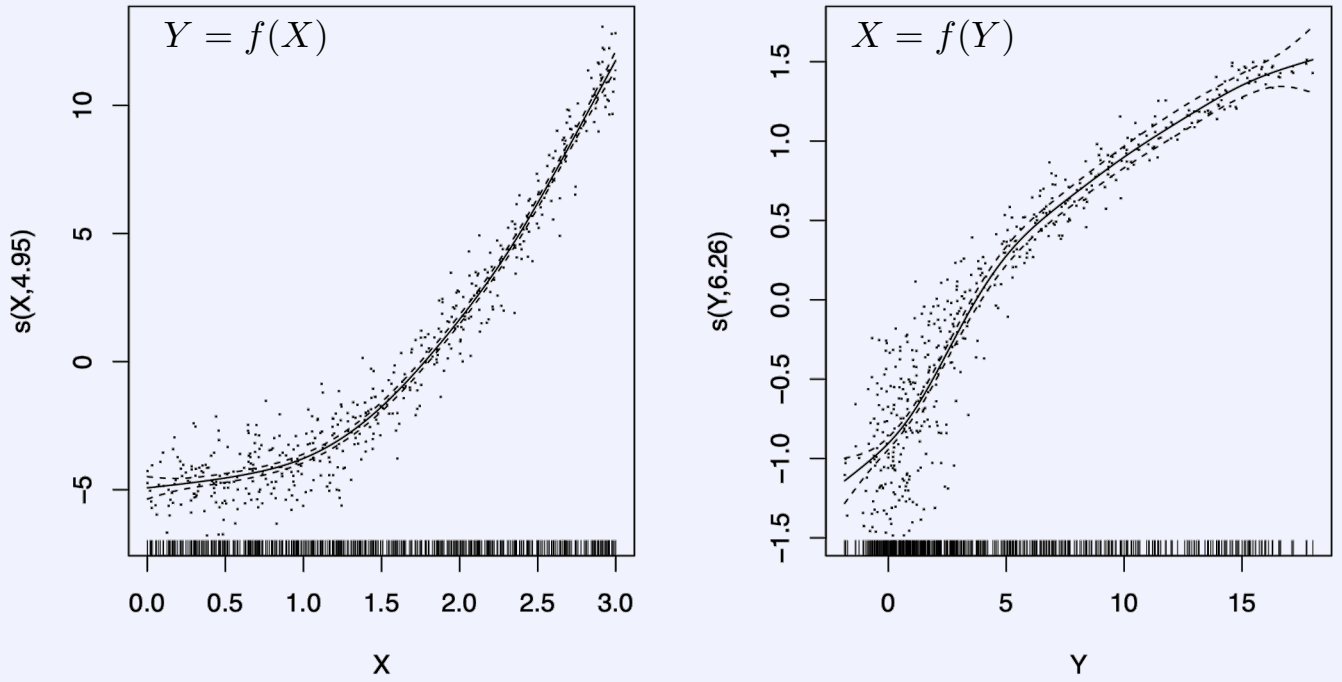


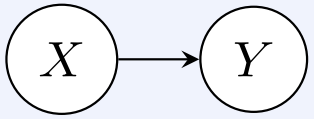
Functional causal model-based methods





Functional causal model-based methods



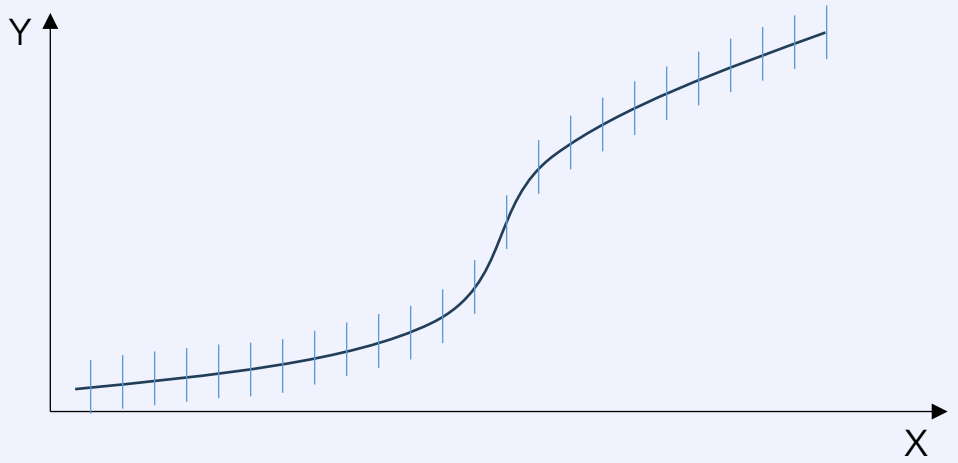


Functional causal model-based methods

Additive Noise Model (ANM)

$$Y = f(X) + \varepsilon_Y$$

with ε_Y independent of X



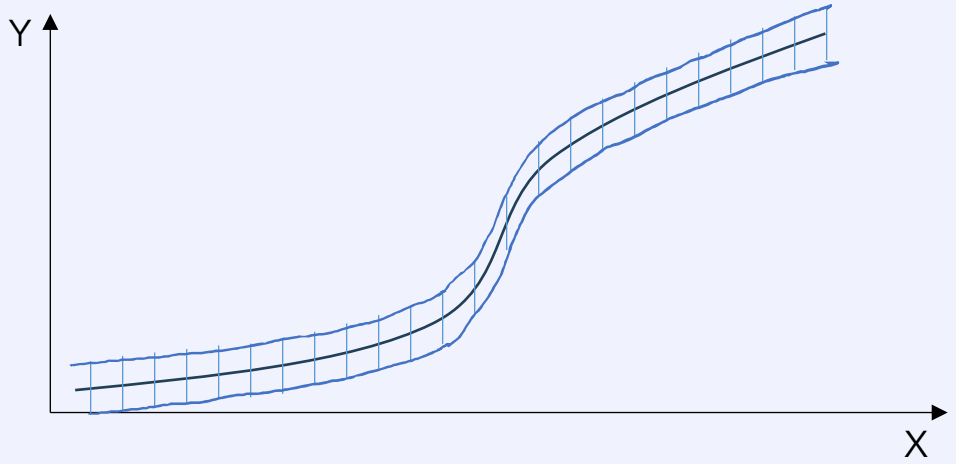
(Hoyer et al. 2009)

Functional causal model-based methods

Additive Noise Model (ANM)

$$Y = f(X) + \varepsilon_Y$$

with ε_Y independent of X



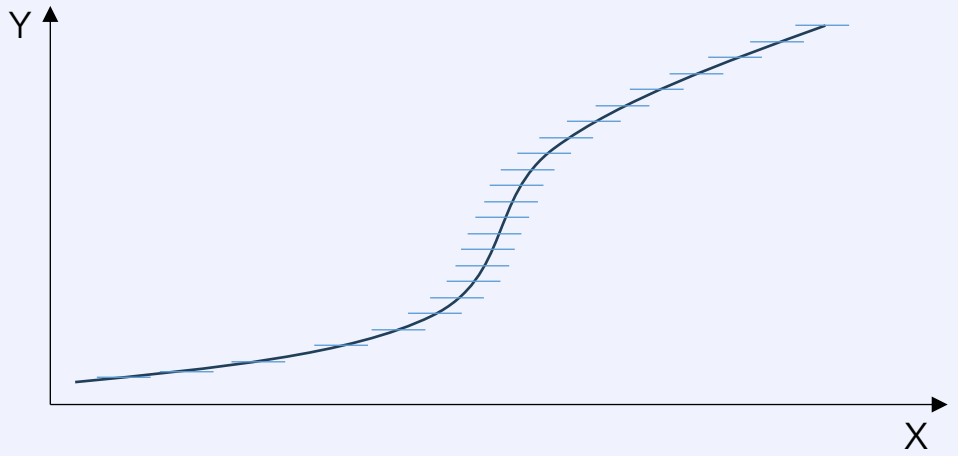
(Hoyer et al. 2009)

Functional causal model-based methods

Additive Noise Model (ANM)

$$X = f(Y) + \varepsilon_X$$

with ε_X independent of Y



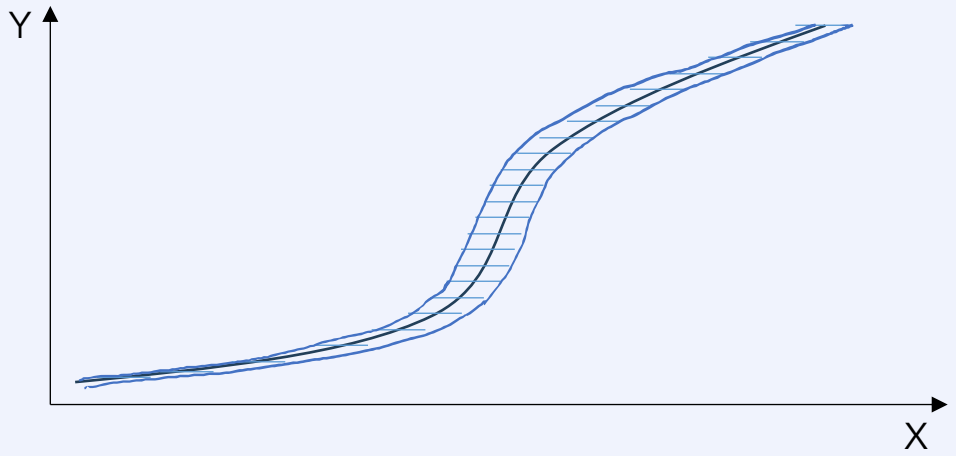
(Hoyer et al. 2009)

Functional causal model-based methods

Additive Noise Model (ANM)

$$X = f(Y) + \varepsilon_X$$

with ε_X independent of Y



(Hoyer et al. 2009)

Functional causal model-based methods

Linear Non-Gaussian Acyclic Model (LiNGAM)

$$Y = aX + \varepsilon_Y$$

with ε_Y non-Gaussian

ICA-based solutions

(Shimizu et al. 2006)

Causal discovery

Constraint-based methods

FCM-based methods

Score-based methods

Continuous optimization-based methods

Score-based methods

Observational distribution

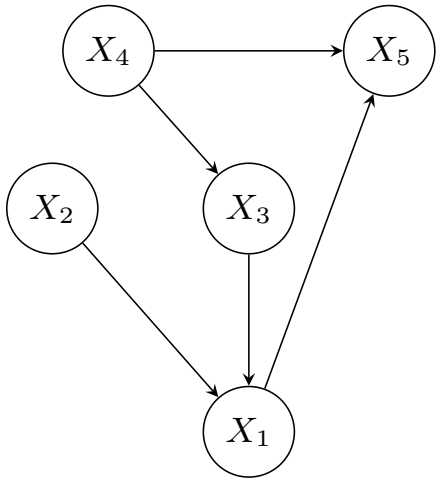
$$P(X_1, X_2, ...)$$

→

Search over the DAG space & model selection based on a score

→

Graphical causal model

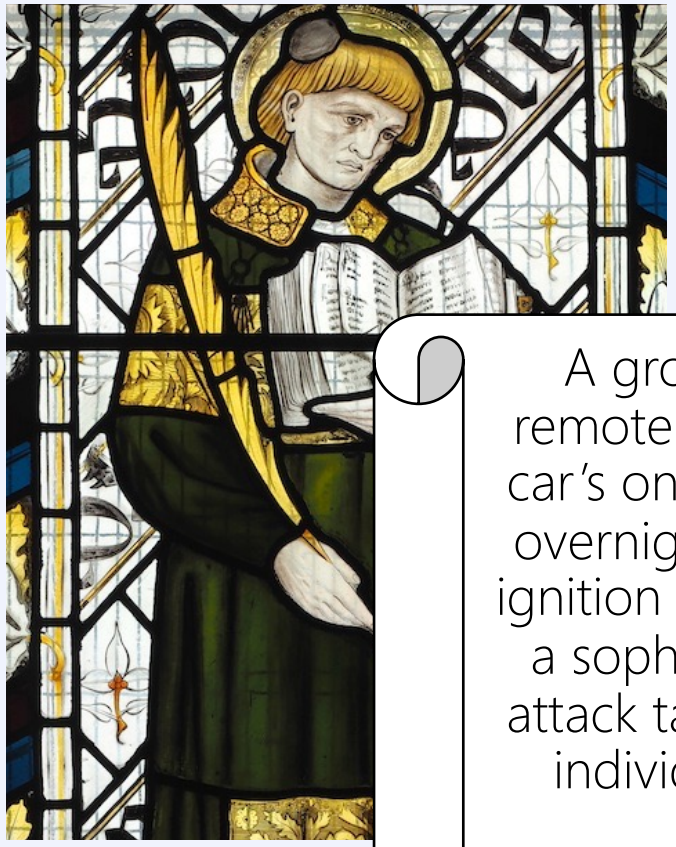


GES[1]
Bayesian scoring criterion

GLOBE[2]
MDL score (information-theoretic causal discovery)

[1](Chickering. 2002) [2](Mian, Marx, Vreeken. 2021)

Occam's razor



Why won't my car start?

A group of hackers remotely accessed your car's onboard computer overnight, disabling the ignition system as part of a sophisticated cyber-attack targeting random individuals to create chaos.

The battery is dead.

Occam's razor

The simplest
explanation
that fits the data
is usually the correct one

Occam's razor

The simplest
causal model
that fits the data
is usually the correct one

Information-theoretic causal discovery

Algorithmic information theory (AIT)

AIT principally studies measures of irreducible information content of strings

Information-theoretic causal discovery

Algorithmic information theory (AIT)

AIT principally studies measures of irreducible information content of strings

Kolmogorov complexity

Length of the shortest program for a universal Turing Machine that generates it and halts

i.e. length of its shortest lossless description in bits

Causal discovery

Markov condition

each node is independent of its non-descendants given its parents

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P\left(X_i | pa(X_i)\right)$$

Information-theoretic causal discovery

Algorithmic Markov condition

the joint complexity over all nodes is given by the sum of the complexities of each individual node given the optimal compression of its parents

$$K(x_1, x_2, ..., x_n) \stackrel{+}{=} \sum_{i=1}^{n} K\left(x_i | pa_i^*\right)$$

→ gives us the Markov equivalence class

(Janzing and Schölkopf. 2010)

Assumption

Algorithmic independence of conditionals

A causal hypothesis is only acceptable if the shortest description of the joint distribution is given by the concatenation of the shortest descriptions of the Markov kernels

$$K(P(X_1, X_2, ..., X_n)) \stackrel{+}{=} \sum_{i=1}^{n} K(P\left(X_i | pa(X_i)\right))$$

Assumption

Algorithmic independence of conditionals

A causal hypothesis is only acceptable if the shortest description of the joint distribution is given by the concatenation of the shortest descriptions of the Markov kernels

$$K(P(X_1, X_2, ..., X_n)) \stackrel{+}{=} \sum_{i=1}^{n} K(P\left(X_i | pa(X_i)\right))$$

$$I_A(P_{X_1}|P_{pa(X_1)}; ...; P_{X_n}|P_{pa(X_n)}) \stackrel{+}{=} 0$$

Information-theoretic causal discovery

Applying the principle of independent mechanisms,

if $X \to Y$

$$K(P_X) + K(P_{Y|X}) \overset{+}{\leq} K(P_Y) + K(P_{X|Y})$$

because

$$I_A(P_X; P_{Y|X}) \overset{+}{\leq} 0$$

→ gives us the edge orientation

Information-theoretic causal discovery

Algorithmic causality

gives us the causal structure

and the edge orientation

in theory

Kolmogorov complexity is not computable in practice

Information-theoretic causal discovery

Minimum description length (MDL)

$$L(D, M^*) = L(D|M^*) + L(M^*)$$

global length (# of bits needed)

length of describing the data encoded with the best causal model

length of describing the best causal model

$$M^* = \underset{M \in \mathcal{M}}{\operatorname{argmin}}\ L(D|M) + L(M)$$

Causal discovery

Constraint-based methods

FCM-based methods

Score-based methods

Continuous optimization-based methods

Continuous optimization-based methods

Replace discrete search over DAG with continuous optimization

Some strategies:

- DAG encoding into an adjacency matrix
- Acyclicity constraints
- Sparsity constraints

[1](Zheng et al. 2018)

Continuous optimization-based methods

Replace discrete search over DAG with continuous optimization

Some strategies:

- DAG encoding into an adjacency matrix
- Acyclicity constraints
- Sparsity constraints

NOTEARS[1]

Optimize the fit of a linear model while enforcing acyclicity and DAG sparsity

Limitations:

- No consistency/identifiability guarantee ← Not causal
- Orient the edges by choosing the variable with lowest variance as parent

[1](Zheng et al. 2018)

Question of Interest

What if we live in a non-ideal world?

Causal discovery in a non-ideal world

Common Biases

Confounding, Selection, and Measurement Errors

Heterogeneous Data Sources

Simpson's Paradox and Beyond

Time Series

Non-IIDness, Delayed Effects, and Non-Stationarity