ETH zürich

Causality for Machine Learning

author: Bernhard Schölkopf
presenter: Shanglun Feng
27.04.2022

Part I : Causality

- concepts & foundations

Questions

What is *Causality* in your mind?

How to formulate causality in mathematical language?

What questions can causality answer more than statistics?

How to examine causality as a discipline?

Definition

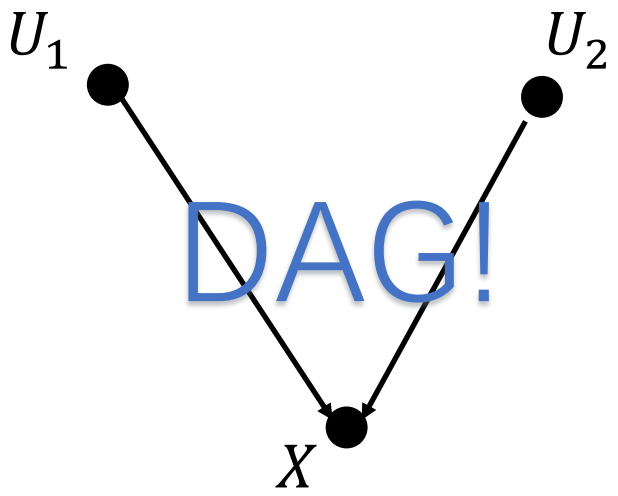
Structural Causal Model , a SCM consists of:

- A set of stochastic unexplained variables U, assumed to be **jointly independent**.
- A set of endogenous variables X
- A sets of functions f that assigns each variable in X a value based on the values of the other variables in the model $X_i = f_i(\mathrm{PA_i}, U_i)$.

$$U = \{U_1, U_2\}$$
$$X = \{X\}$$
$$F = \{f_X(U_1, U_2) = 2U_1 + 3U_2\}$$
$$X = 2U_1 + 3U_2$$

Graphical Causal Model , a GCM consist of:

- A set of nodes representing the variables in X and U.
- A set of edges between the nodes representing the functions in f.



Probabilistic Properties

Causal Markov condition conditioned on its parents, each X_j is independent of its non-descendants.

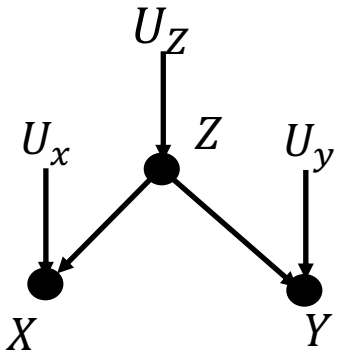
Causal factorization $p(X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i|\mathrm{PA_i})$

Entangled factorization $p(X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i|X_{i+1}, \cdots, X_n)$

EG:

graphical model	causal factorization
$X \longrightarrow Y$	$p(X)p(Y\|X)$
$Y \rightarrow X$	$p(Y)p(X\|Y)$
$X \leftarrow Z \rightarrow Y$	$\int \mathrm{d}z p(Z)p(X\|Z)p(Y\|Z)$

SCM, Graphical Model & Causal Factorization



$$Z = U_Z$$
$$X = 2Z + U_x$$
$$Y = -Z + U_y$$

$$p(X, Y, Z) = p(X|Z)p(Y|Z)p(Z)$$
$$(X \perp Y|Z)$$
$$\{p(X|Z), p(Y|Z), p(Z)\}$$
$$p(X|Y, Z)$$

Dependencies on Graph

IDEA: predict patterns of dependencies in the data, based only on the structure of the model' s graph.

d-separation:

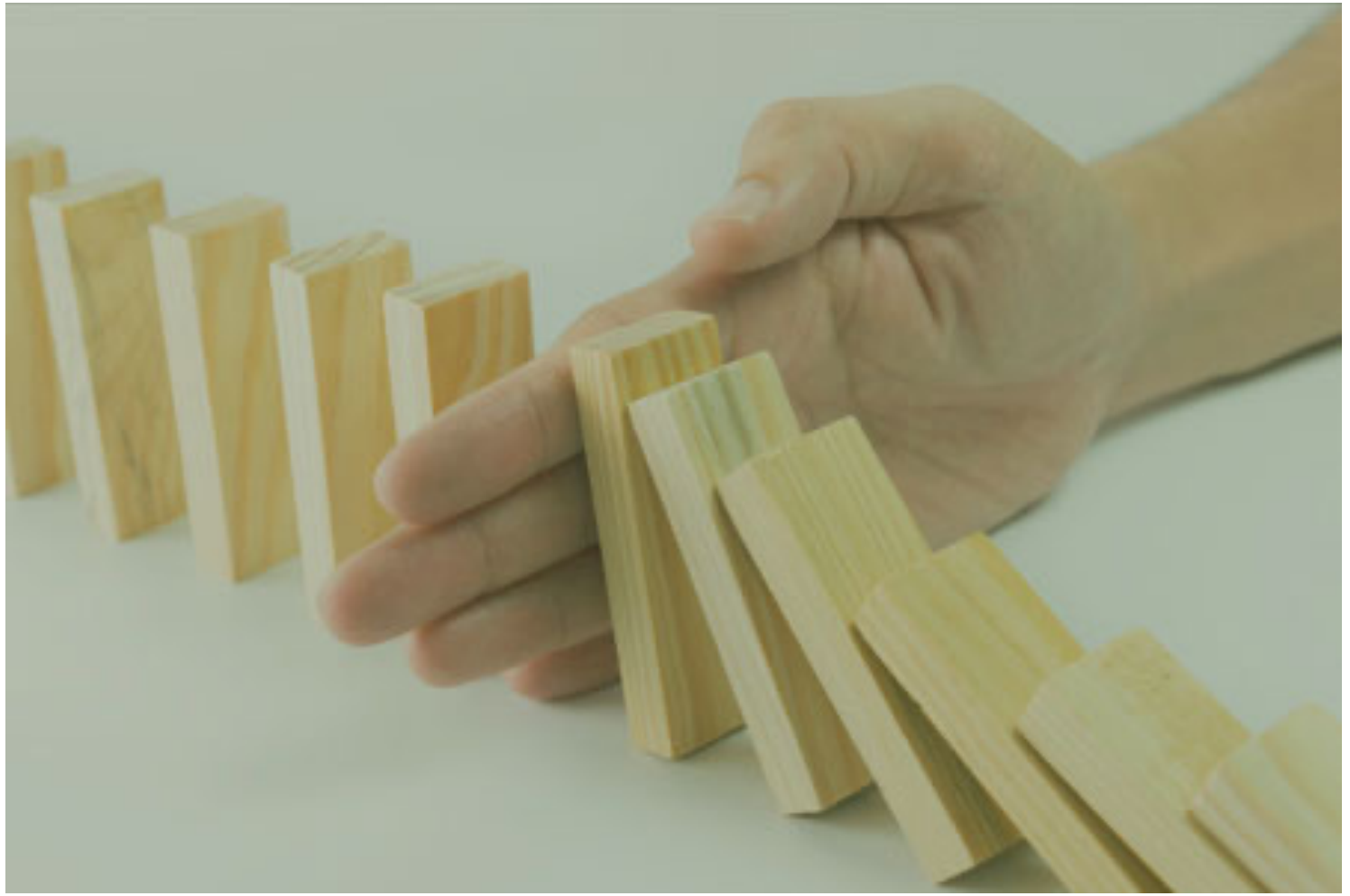
A path p is *blocked* by a set of nodes Z if and only if

- p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on)
- p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z

If Z blocks every path between two nodes X and Y, then X and Y are d-separated

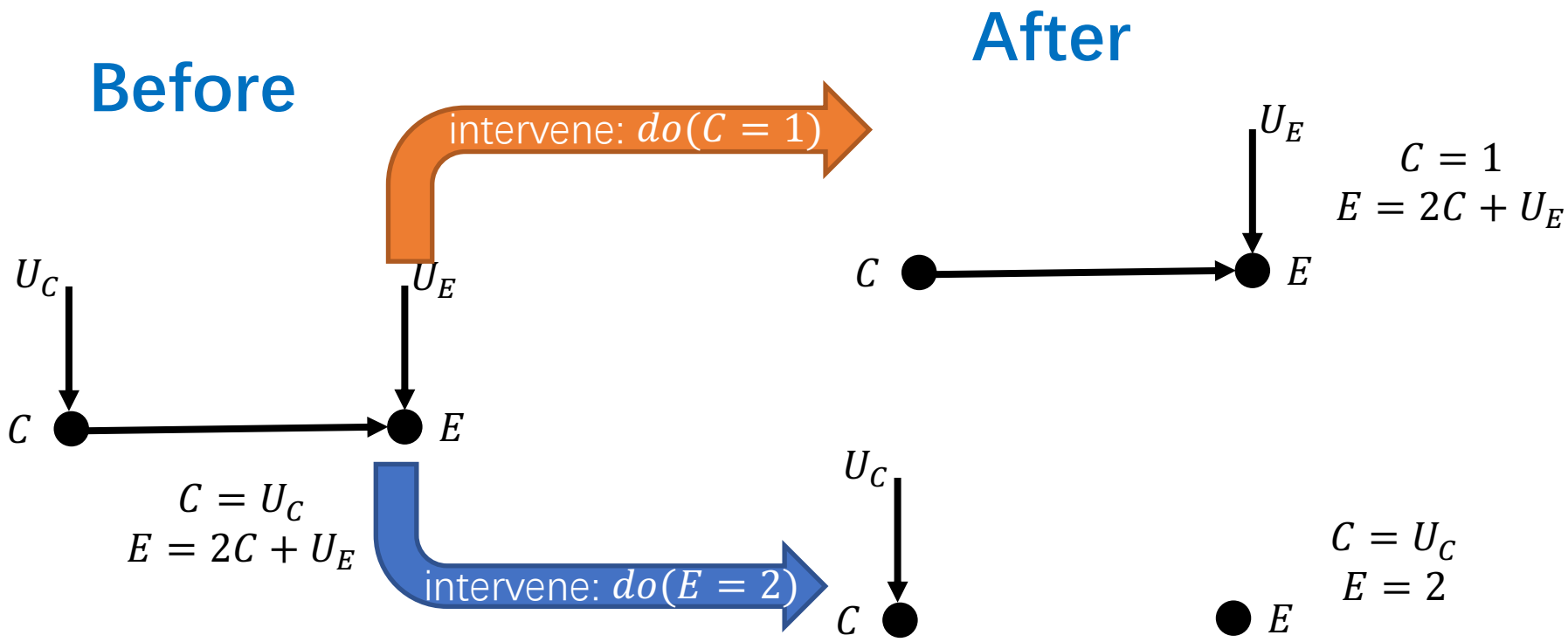
Theorem: If X and Y are d-separated by Z, X and Y are independent conditional on Z

Intervention: Intuition



Intervention

IDEA: change parts of the data-generating process.



Intervention: Uncover Causality

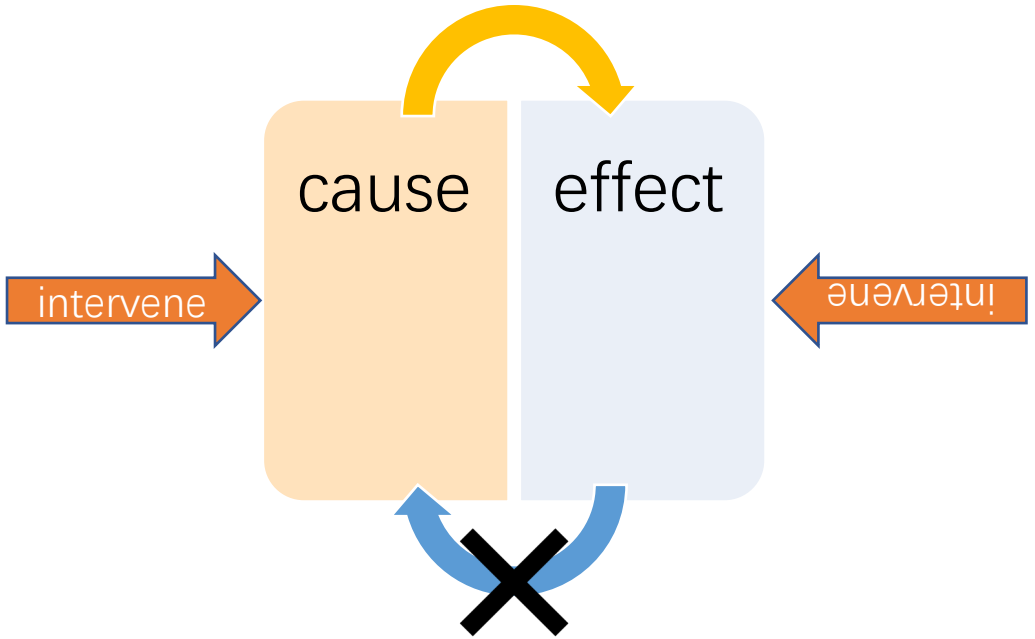
SCM: $C := U_C,\ E := 2C + U_E,\ U_C, U_E \sim \mathcal{N}(0,1)$

original distribution $C \sim \mathcal{N}(0,1), E \sim \mathcal{N}(0,5)$

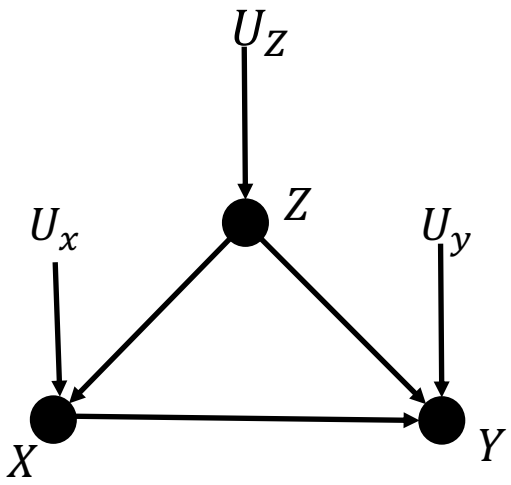
intervene $do(C = 1) : C = 1,\ E \sim \mathcal{N}(2,1)$

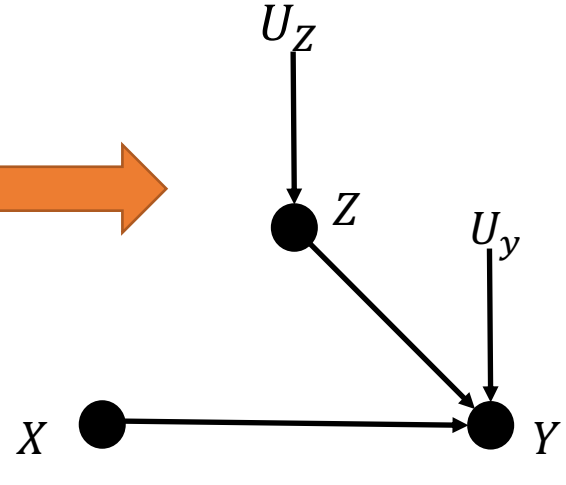
intervene $do(E = 2) : C \sim \mathcal{N}(0,1), E = 2$

asymmetry!



Intervention: Calculation





$$\Pr[Y = y|do(X = x)]$$

$$\Pr[Y = y|X = x] = \sum_z \Pr[Y = y|X = x, Z = z] \Pr[Z = z|X = x]$$

Causal Effect Rule Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by.

$$\Pr[Y = y|do(X = x)] = \sum_z \Pr[Y = y|X = x, Z = z] \Pr[Z = z]$$

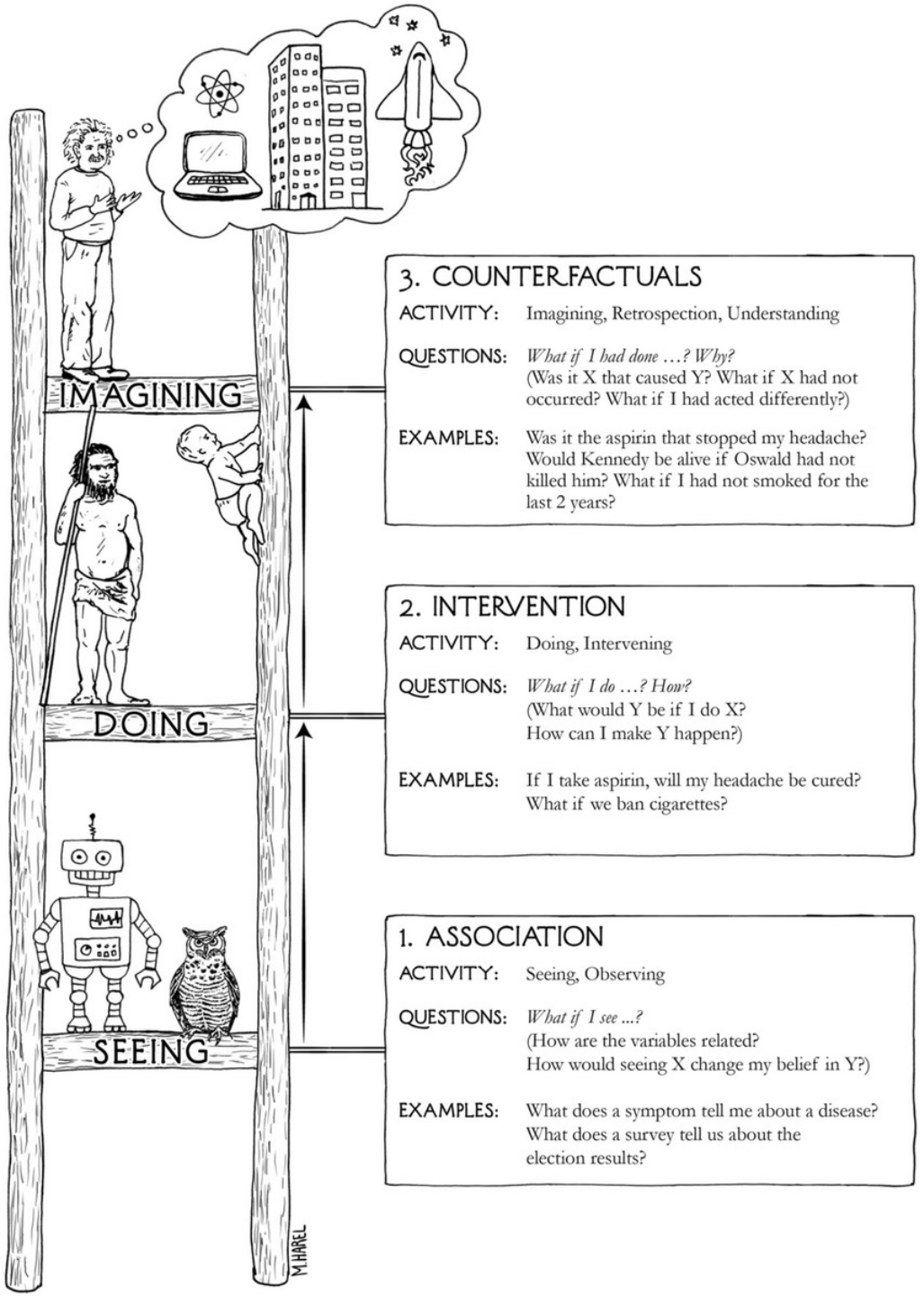
debias!

statistical learning

$p(X_1, X_2, \cdots, X_p)$

causal learning

$\{p(X_j|PA_j)\}$



Independent Causal Mechanisms

Independent Causal Mechanisms Principle: The causal generative process of a system' s variables is comosed of autonomous modules that do not **inform** or **influence** each other.

Corollary

- **physical aspect** changing(intervening upon) one mechanism $p(X_i|\mathrm{PA_i})$ does not change the other mechanisms $p(X_j|PA_j)$
- **information theoretic aspect** knowing some other mechanisms $p(X_i|\mathrm{PA_i})$ does not give us information about a mechanism $p\left(X_j\middle|PA_j\right)$

ICM: Corollary

unexplained variables : U_1, U_2: $\Omega = \{0,1\}$

causal mechanism 1: $f_1(X, U_1) = U_1 e^X + (1 - U_1)X^2$

causal mechanism 2: $f_2(Y, U_2) = U_2 e^{-Y^2} + (1 - U_2)Y$

$$\boldsymbol{U_1 \not\perp U_2 \Longrightarrow f_1 \not\perp f_2}$$

ICM: Modularity

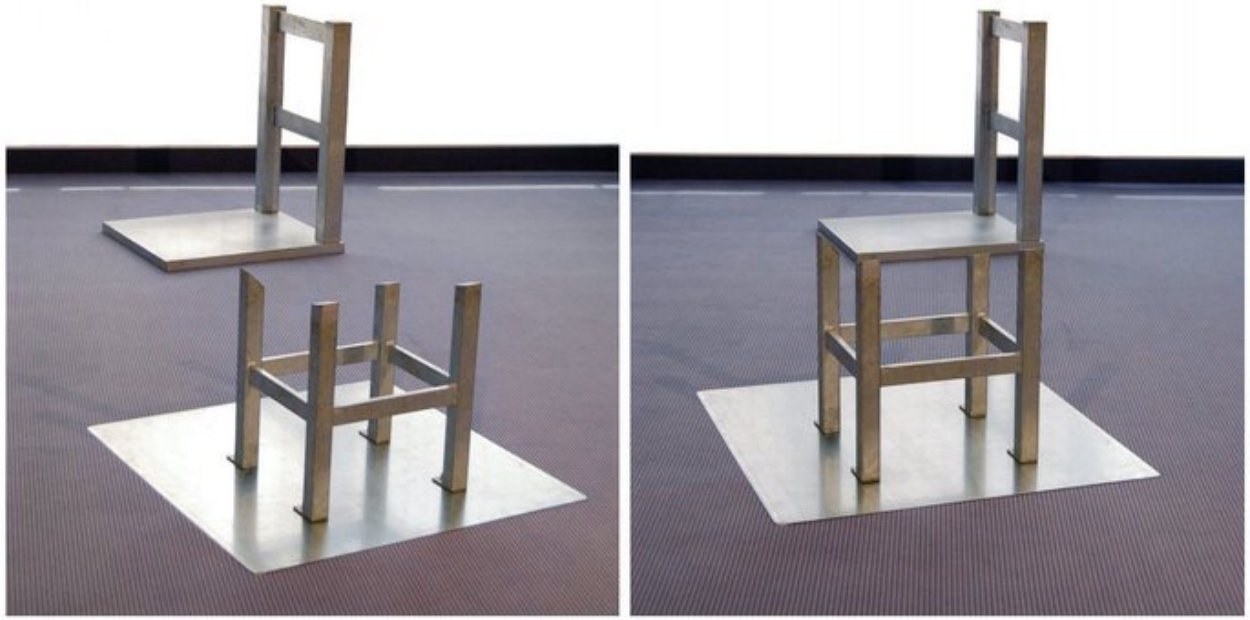
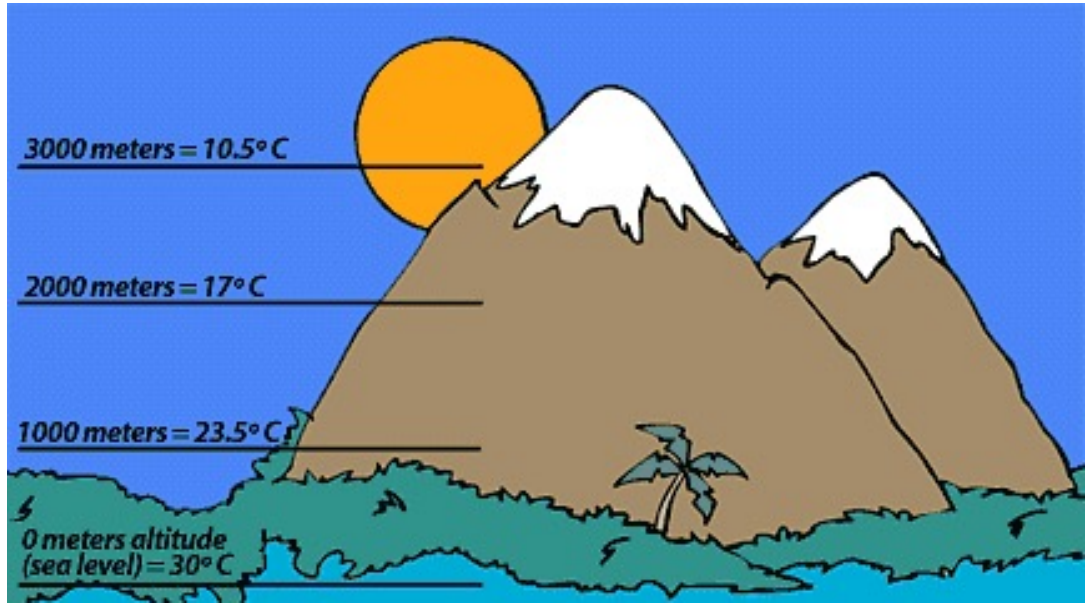


Figure 1: *Beuchet chair*, made up of two separate objects that appear as a chair when viewed from a special vantage point violating the independence between object and perceptual process. (Image courtesy of Markus Elsholz, reprinted from Peters et al. (2017).)

causal mechanism 1: perceptual process
causal mechanism 2: object

ICM : Example

Background



altitude A
temperature T

modeling

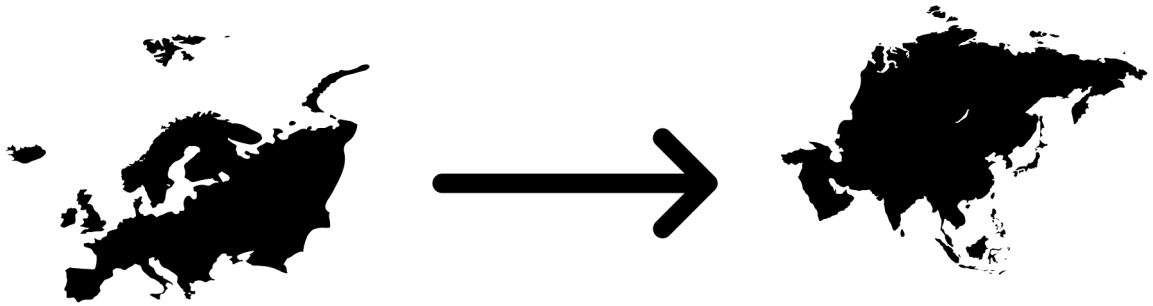
causal factorization $p(A)p(T|A)$

entangled factorization $p(T)p(A|T)$

.

ICM EG: Invariance

$p(A)$: chosen research object
$p(T|A)$: physical mechanism



What change? $p(A)$
What unchange? $p(T|A)$

Corollary:

$p_{EURO}(T) \neq p_{ASIA}(T)$
$p_{EURO}(A|T) \neq p_{ASIA}(A|T)$
$p_{EURO}(A,T) \neq p_{ASIA}(A,T)$

ICM: Related Concepts

generalization

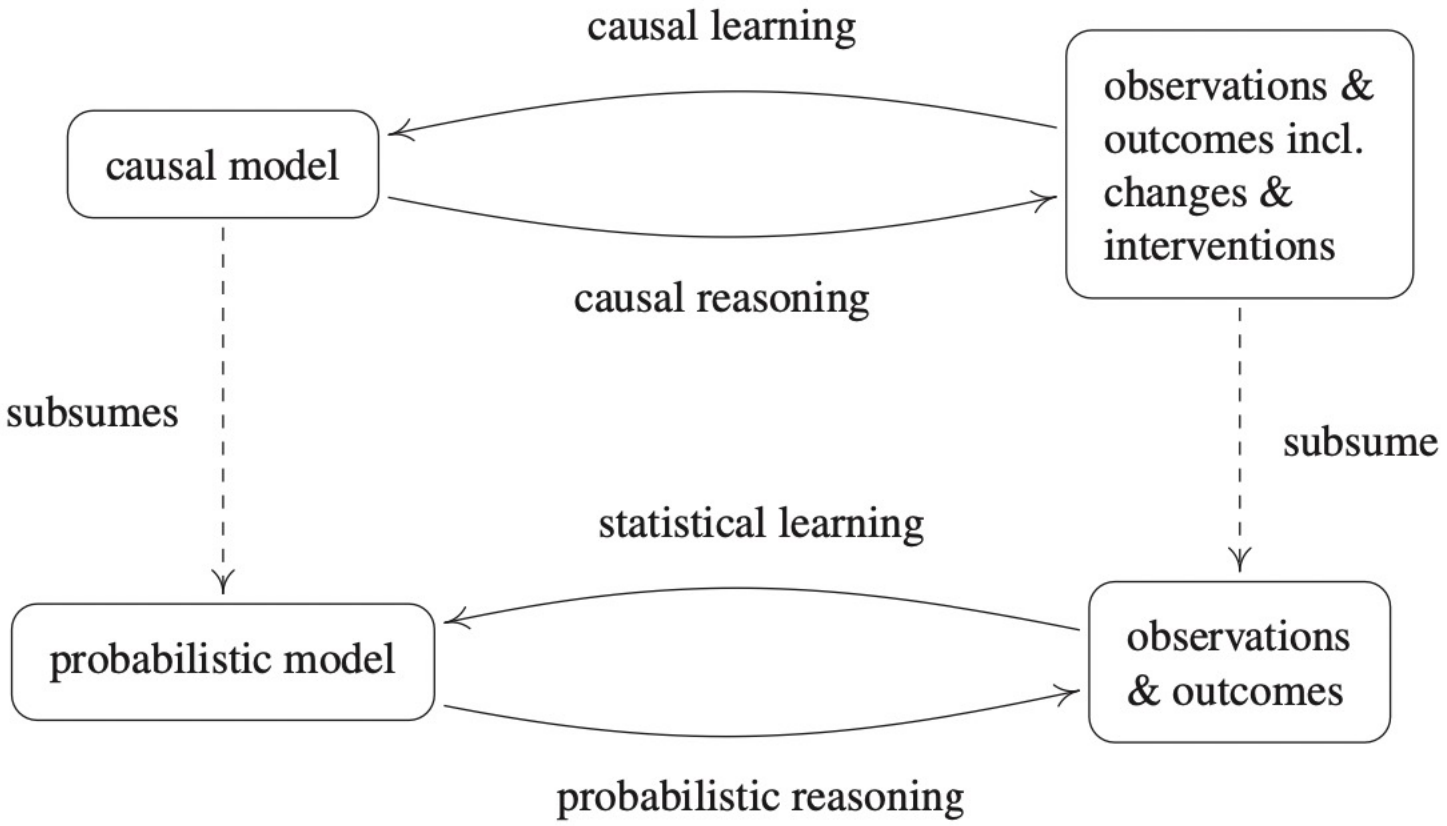
robustness

autonomy of systems

modularity

invariance

Issues for Causality



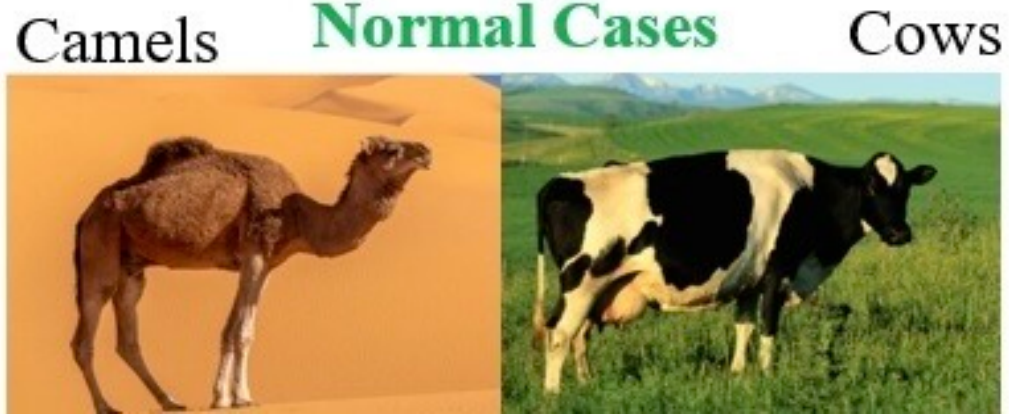
Levels of Causal Modeling

Model	Predict in IID setting	Predict under distri. shift/intervention	Obtain physical insight
Mechanistic/Physical	✓	✓	✓
Causal	✓	✓	✗
Statistical	✓	✗	✗

Part II : Causality for Machine Learning

- applications & frontiers

Why statistical learning is not enough?



training set

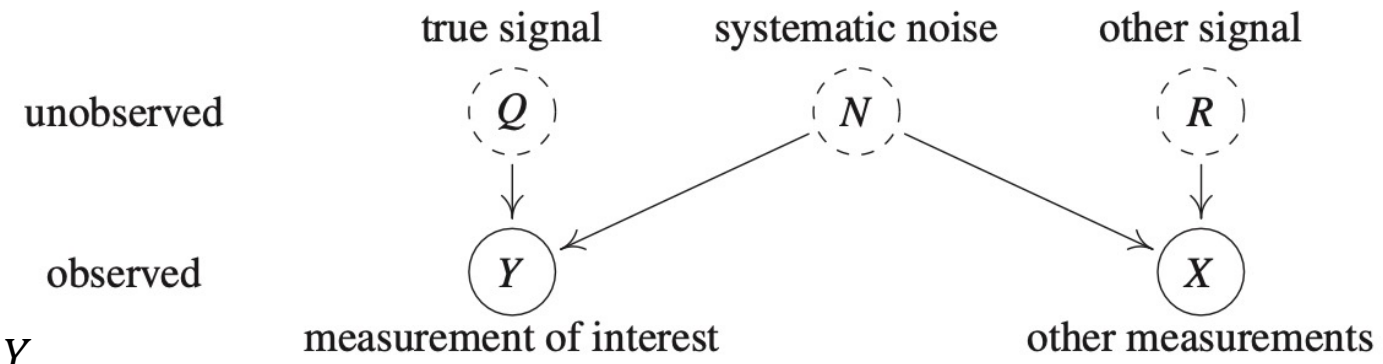
test set

circumstance?

distribution?

IDEA: a classifier incorporating causality could be made ***invariant*** with respect to certain kinds of changes

Half-Sibling Regression



Half-Siblings: $X \leftrightarrow Y$

goal: reconstruct the unobserved signal Q

IDEA: everything in Y that can be explained by X must be due to the systematic noise N and should therefore be removed.

estimation: $\hat{Q} := Y - \mathbb{E}[Y|X]$

MSE: $\mathbb{E}[\left((Q - \mathbb{E}Q) - \hat{Q}\right)^2] \leq \mathbb{E}[\left((Q - \mathbb{E}Q) - (Y - \mathbb{E}Y)\right)^2]$

estimator $\hat{Q}$ is better than estimator Y!

Half-Sibling Regression : Specialization

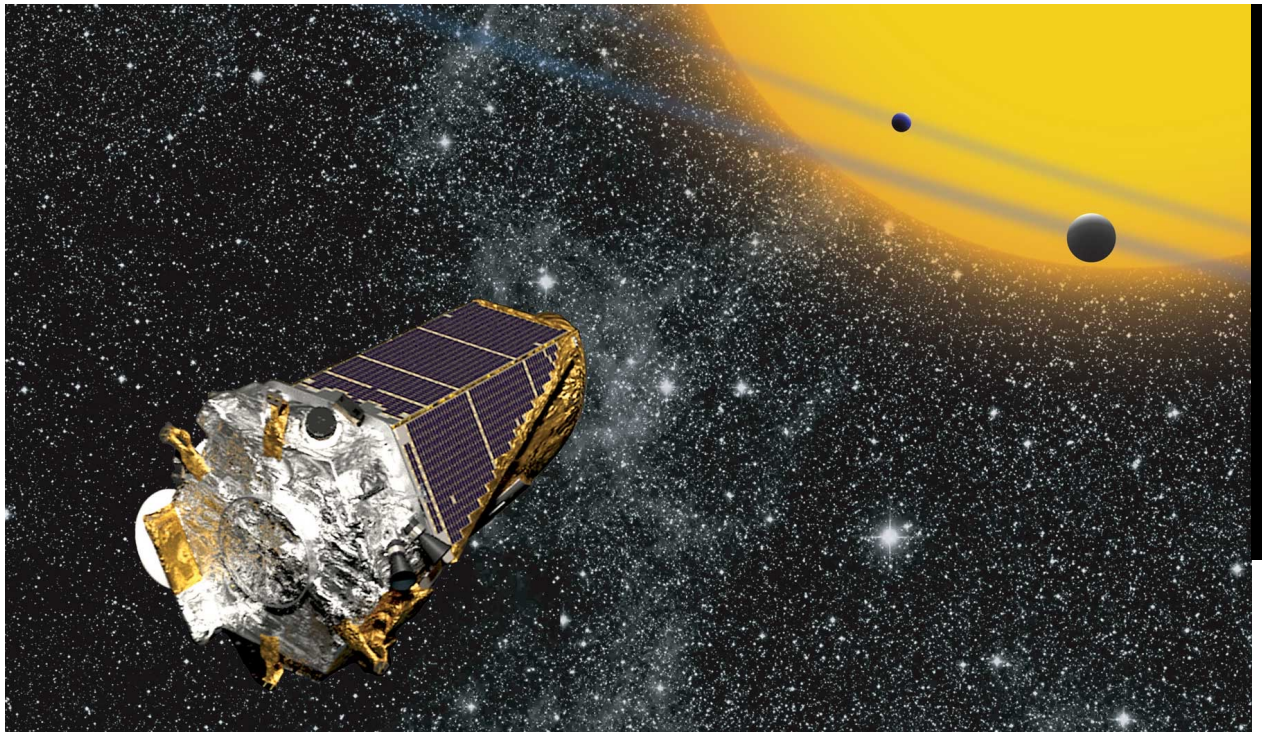
additive noise model : $Y = Q + f(N)$

MSE: $\mathbb{E}\left[\left((Q - \mathbb{E}Q) - \hat{Q}\right)^2\right] = \mathbb{E}[\mathrm{var}[f(N)|X]\,]$

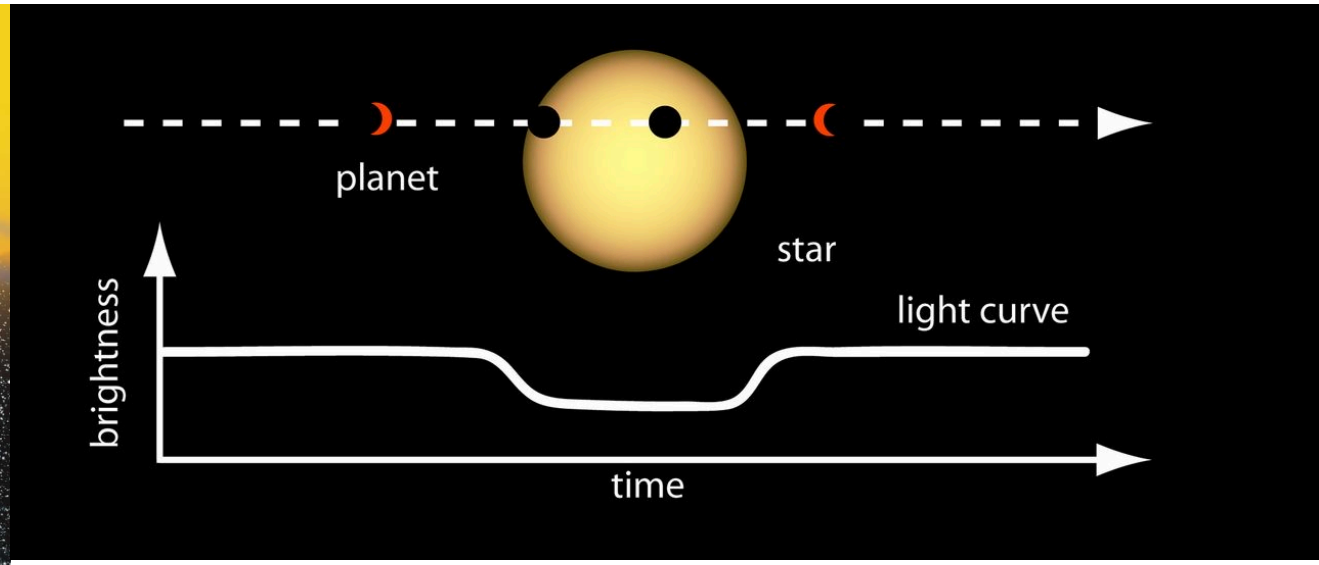
If $f(N) = \Psi(X)$:

$\hat{Q}$ recover completely Q!

HSR : Exoplanet Detection



Kepler Space Observatory

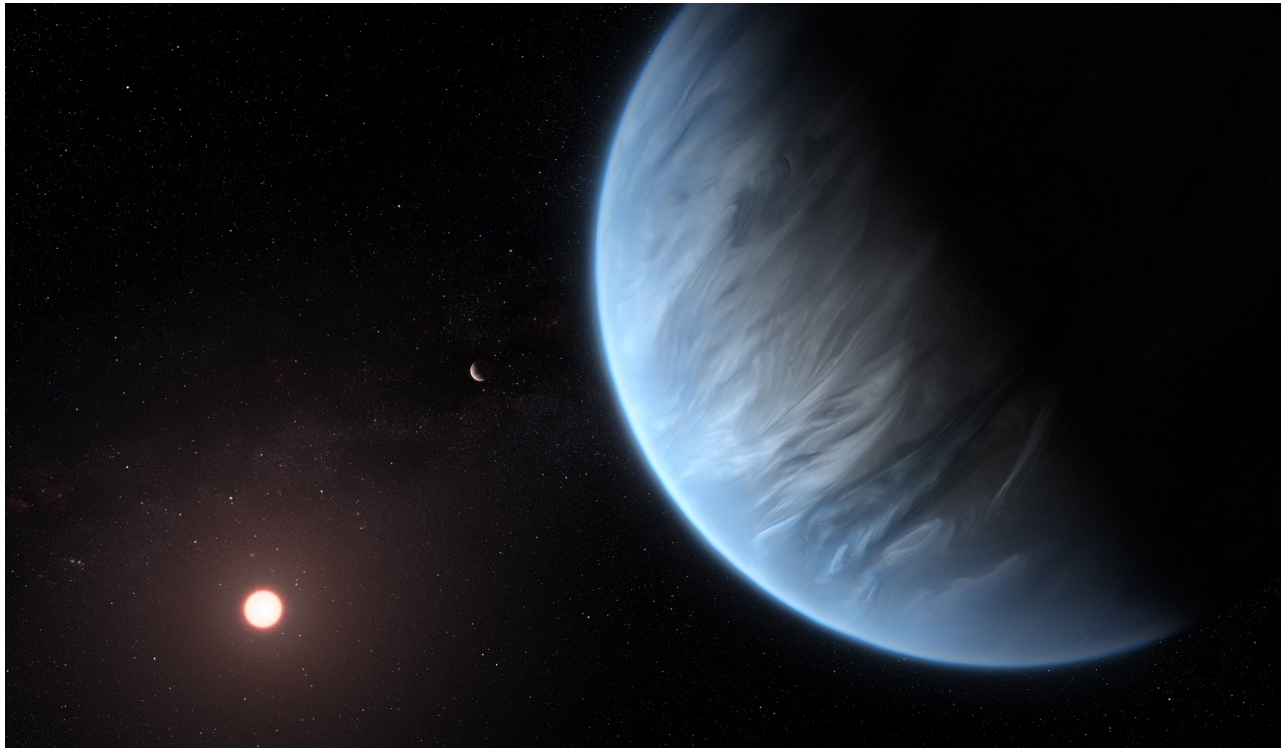


task: detect events where a planet partially occludes its host star, causing a slight decrease in brightness

problem: systematic noise N that is due to the telescope and that make signal from possible planets hard to detect.

HSR : Exoplanet Detection

IDEA: noise structure is often shared across stars



K2-18b

the first discovery for an exoplanet in the habitable zone!!!

Semi-Supervised Learning

goal : estimate $P_{Y|\mathbf{X}}$

supervised learning : $(\mathbf{X_1}, Y_1), \cdots, (\mathbf{X_n}, Y_n) \sim P_{\mathbf{X},Y}$, i.i.d

semi-supervised learning : m additional unlabeled data, $\mathbf{X}_{n+1}, \cdots, \mathbf{X}_{n+m} \sim P_{\mathbf{X}}$

IDEA: additional data points provide information about $P_{\mathbf{X}}$ which itself tells us something about $P_{Y|\mathbf{X}}$

assumptions:

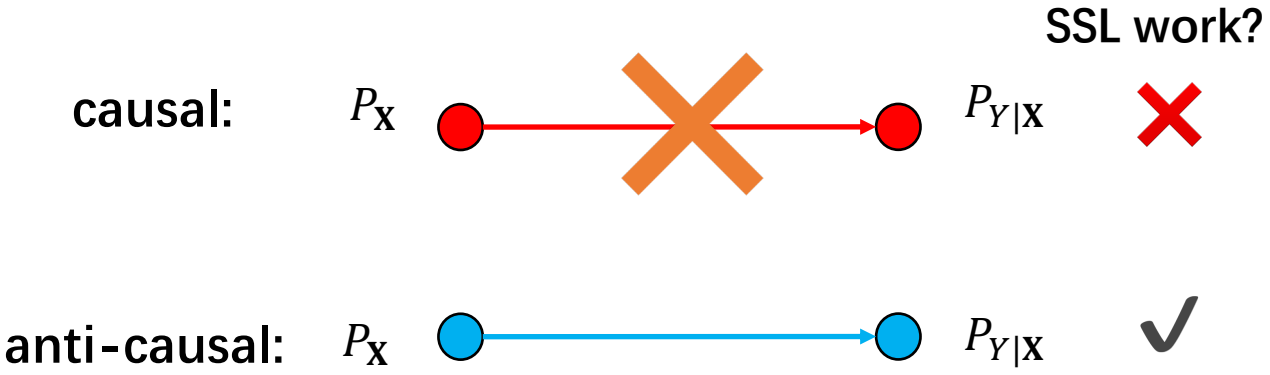
- *cluster* assumption: points lying in the same cluster of $P_{\mathbf{X}}$ have the same or a similar Y
- *low-density separation* assumption: the decision boundary of a classifier should lie in a region where is $P_{\mathbf{X}}$ small

SSL & Causality

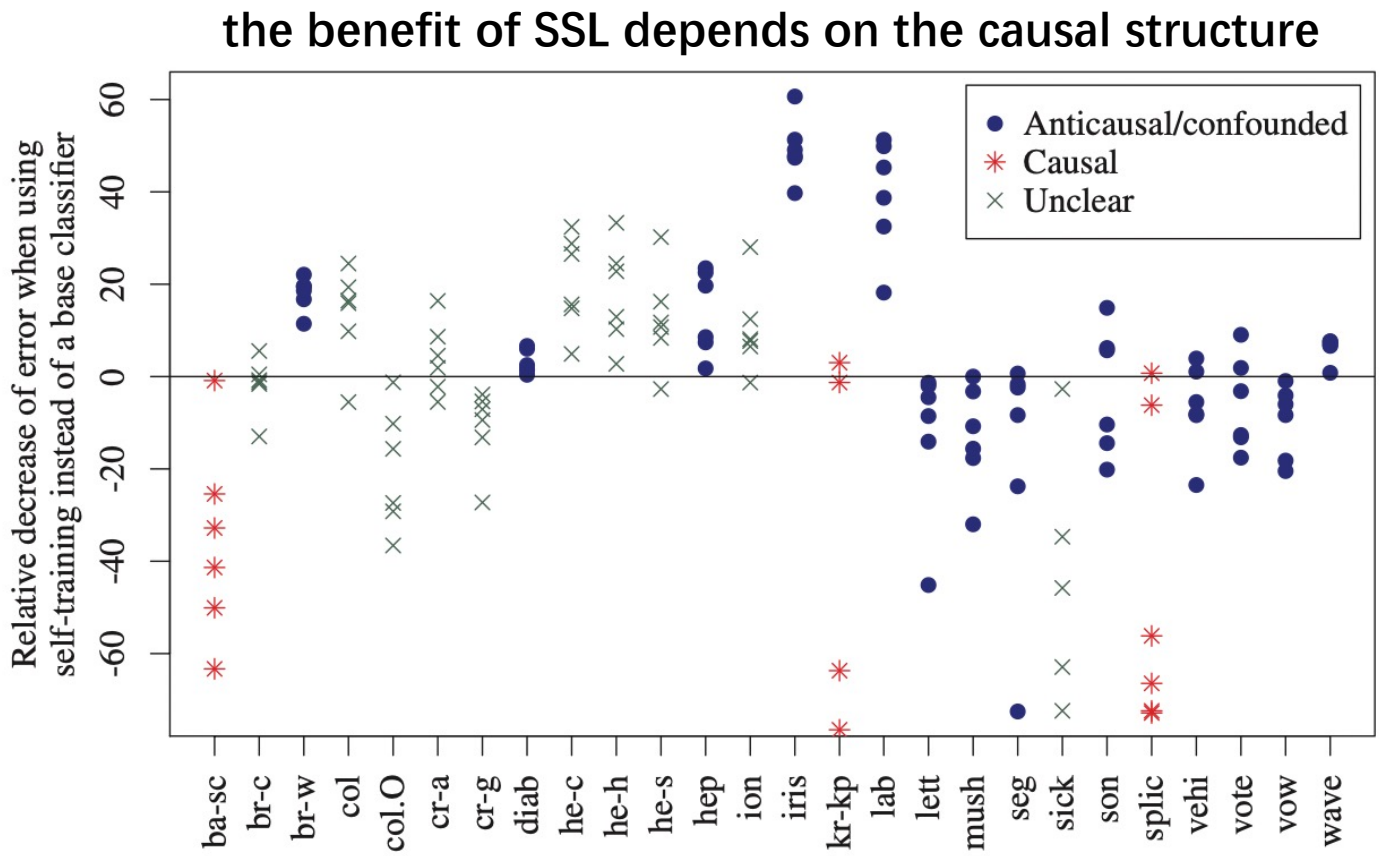
causal ML problem : predict effect from cause

anticausal ML problem : predict cause from effect

IDEA: according to *Independent Causal Mechanisms*, P_{cause} and $P_{effect|cause}$ do not contain information about one another.



SSL & Causality: Results



column: a benchmark data set from UCI
SSL method: self-training
measure: (err(base)-err(self-train))/error(base)

Theoretical Proof: whenever P_{cause} and $P_{effect|cause}$ are independent, then SSL indeed outperforms supervised learing in the causal direction but not in the causal direction.

Q&A