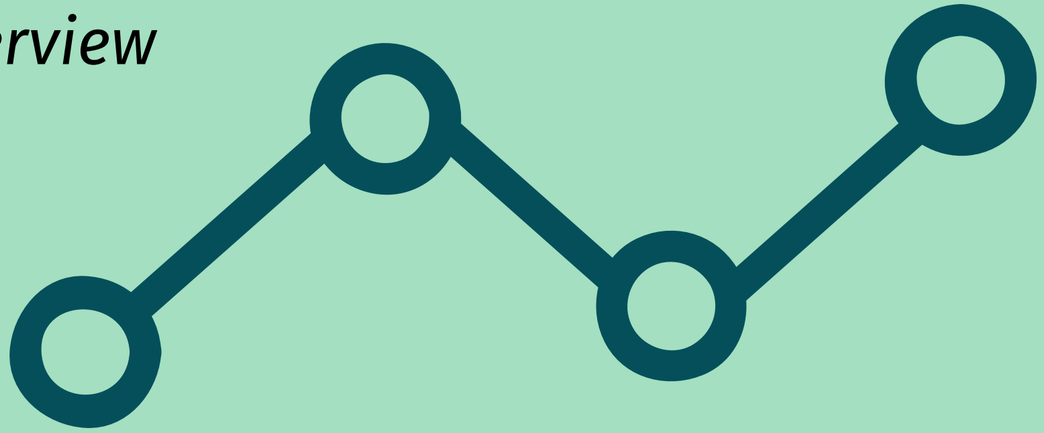
Group 3

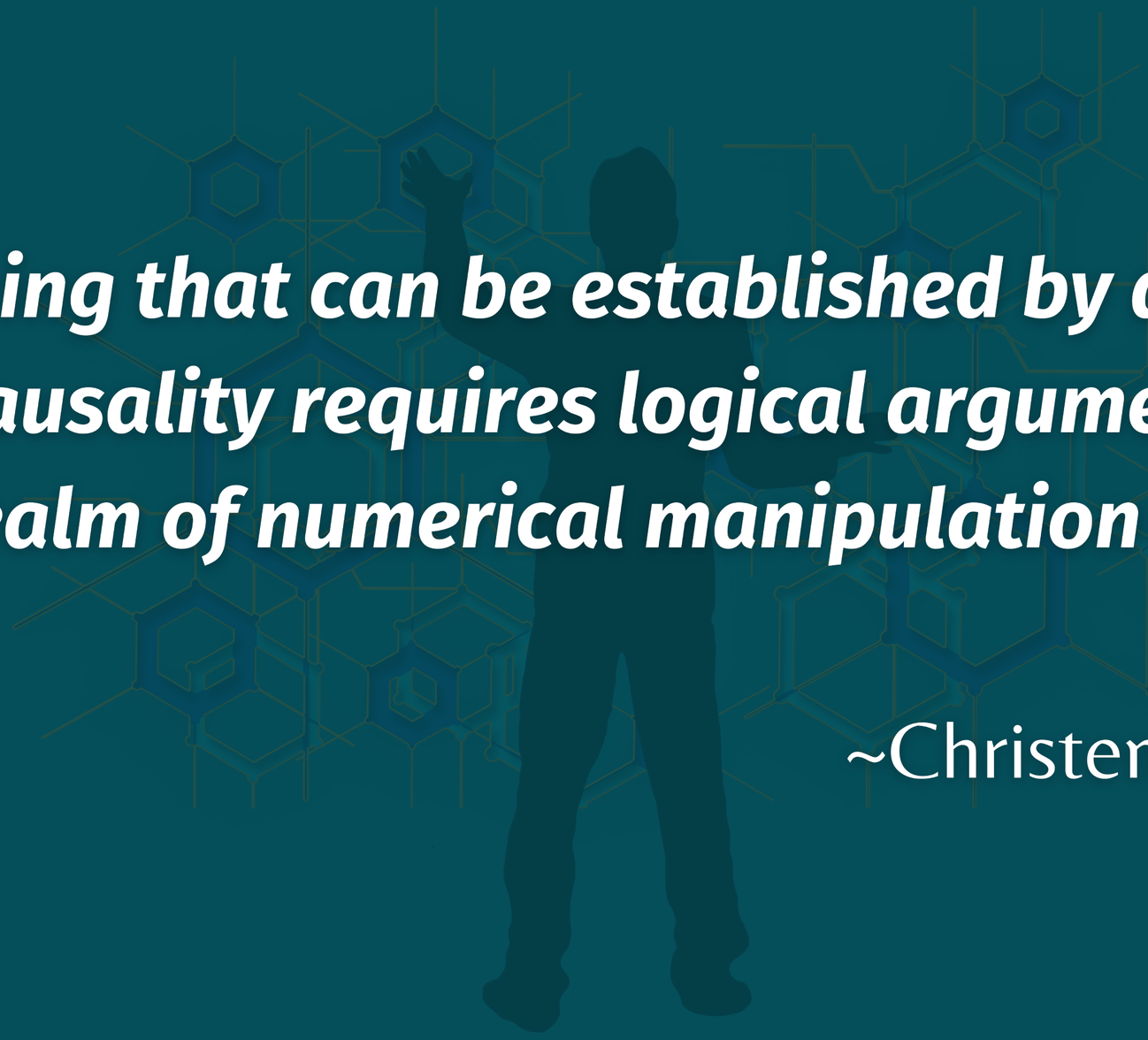
Causal Inference

An Overview



“Causality is not something that can be established by data analysis. Establishing causality requires logical arguments that go beyond the realm of numerical manipulation ”

~Christenen

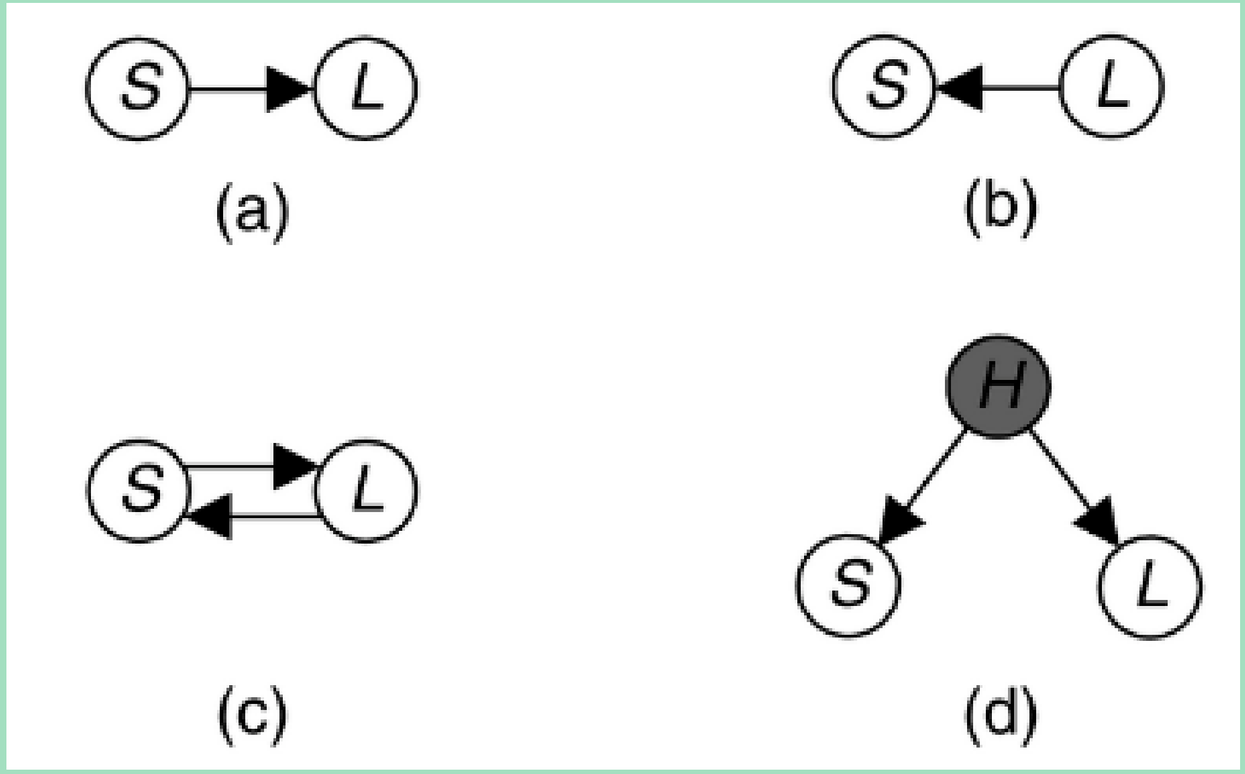


What is Causation?

- **Causation** refers to the relationship between two **events** or **states**, where one event or state is the result or **consequence** of the other.

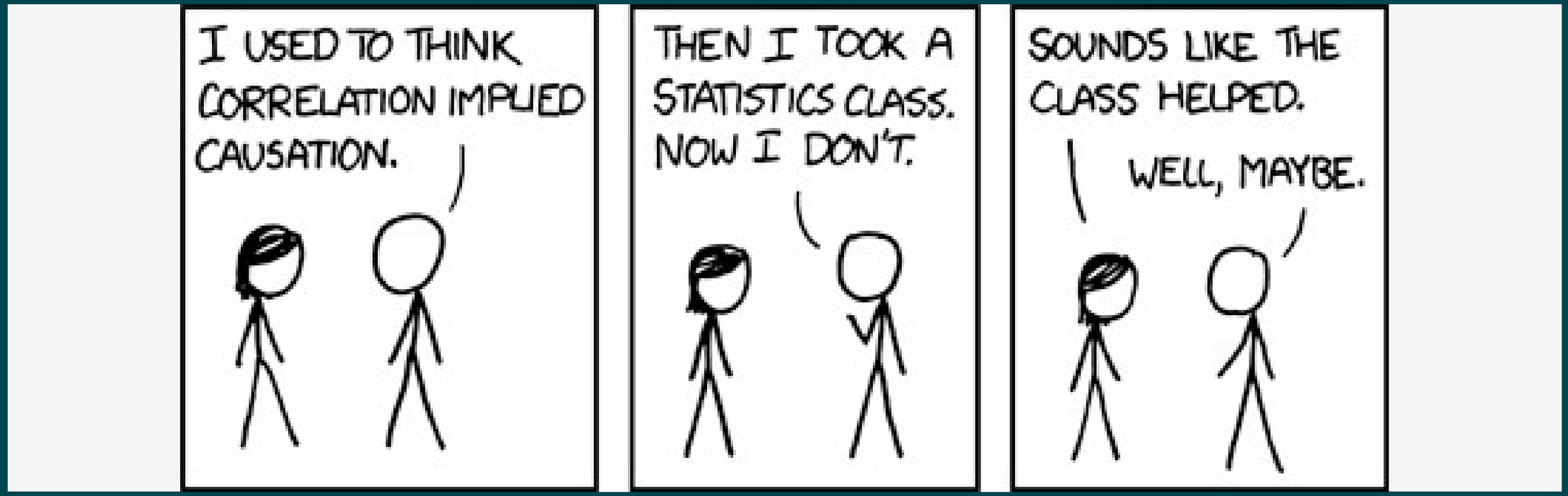
- We say we manipulate X when we force X to take some value, and we say **X causes Y** if there is some manipulation of X that leads to a change in the **probability distribution** of Y .

To find an **underlying relation**, we have to manipulate one variable and determine if it affected the values seen by the other variable.



All possible relationships (dependencies) between the variables S (Smoking) and L(Lung Cancer). Note: H here represents the hidden variable.

Causation vs. correlation dichotomy

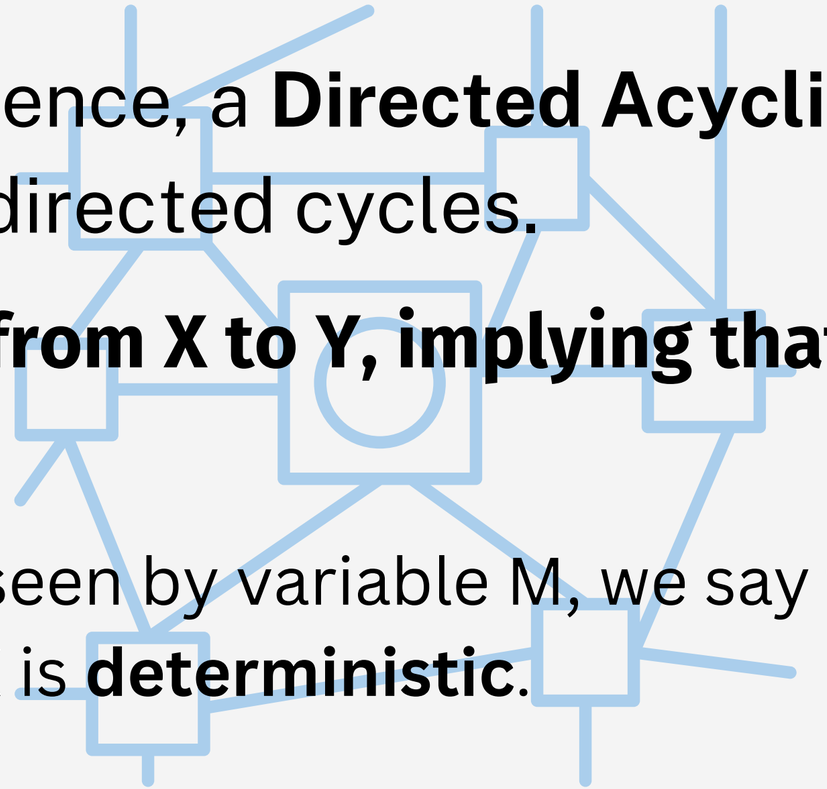


The xkcd comic plays around causation vs. correlation dichotomy and underscores the importance of the background knowledge when performing causal inference.

Original: https://xkcd.com/552/

What is Causal DAG?

In Mathematics, and Computer Science, a **Directed Acyclic Graph** is a directed graph with no directed cycles.



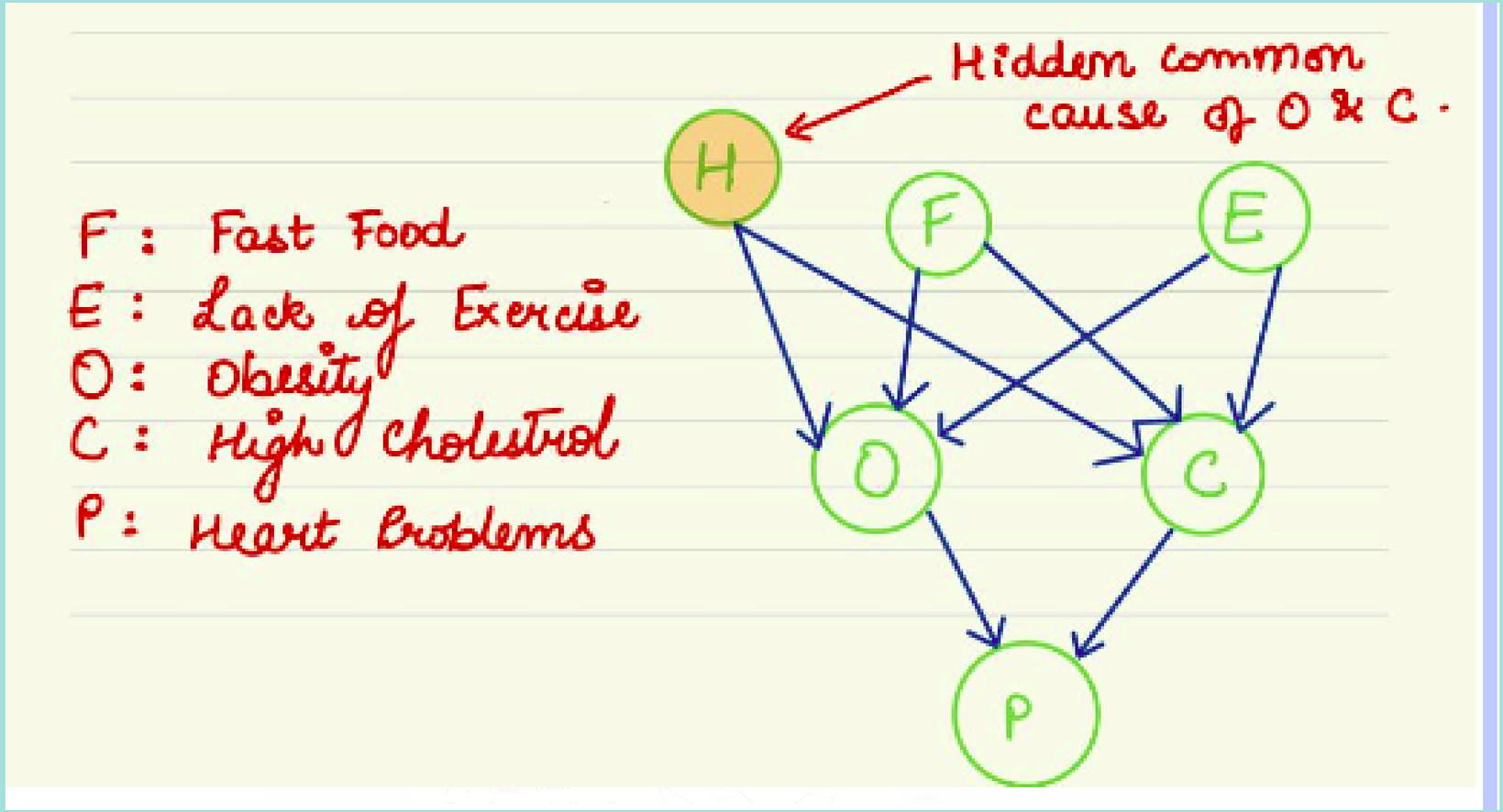
A Causal DAG is one with an edge from X to Y, implying that X is the cause of Y.

- Suppose X's manipulation is seen by variable M, we say that the relation between M and X is **deterministic**.

- This means that there exists a **one-to-one relationship** between M and X, i.e., for specific values of M, we have unique values of X.

What are hidden variables ?

Hidden variables are variables that are not **directly observable** or **measurable** in a given situation, but can affect the outcome of an experiment or observation.



Causal Networks

A **Causal Network** is a Causal DAG that follows the causal **Markov assumption**.

*The **probability distribution P** will satisfy Markov condition with a **DAG** if the probability of **each** variable/node in the DAG is independent of its **nondescendents conditional on its parents.***

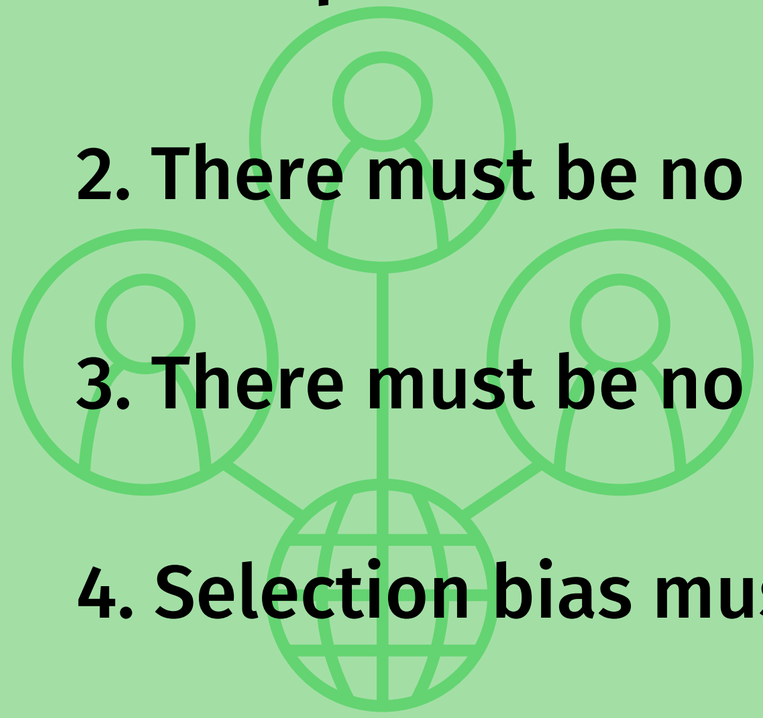
The Markov Assumption

1.If X causes Y , we must draw an edge from X to Y unless all causal paths from X to Y are mediated by observed variables.

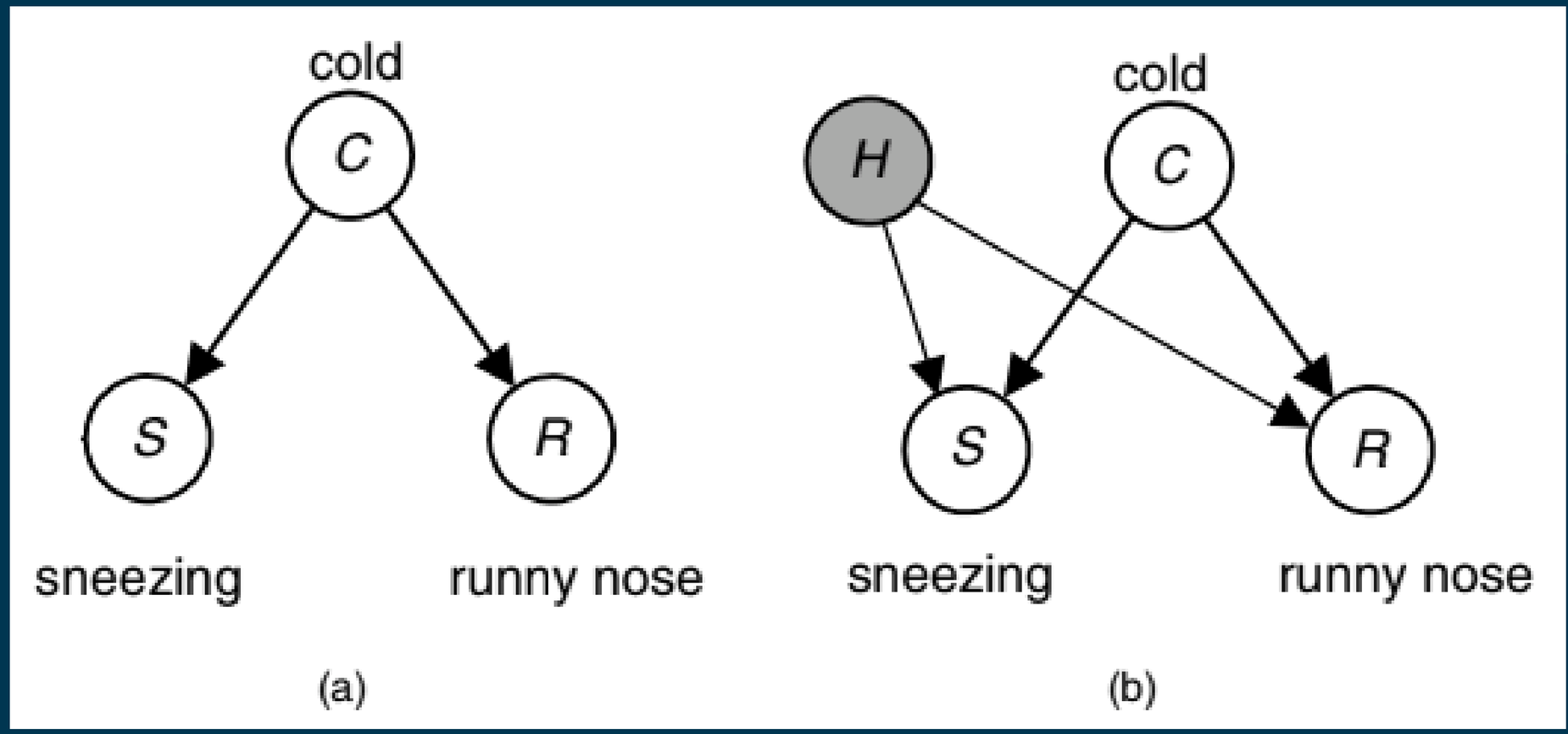
2. There must be no causal feedback loops.

3. There must be no hidden common causes.

4. Selection bias must not be present.



The causal Markov assumption would not hold for the DAG in (a) if there is a hidden common cause as depicted in (b).



In other words, *if* X is a **direct cause** of Y, **there exists no variable W** in the set such that if we knew the values of W, then **change in X no longer changes Y.**

In any causal network, no causal feedback loops are present, no **hidden variables** are present, and all hidden variables are accounted for .

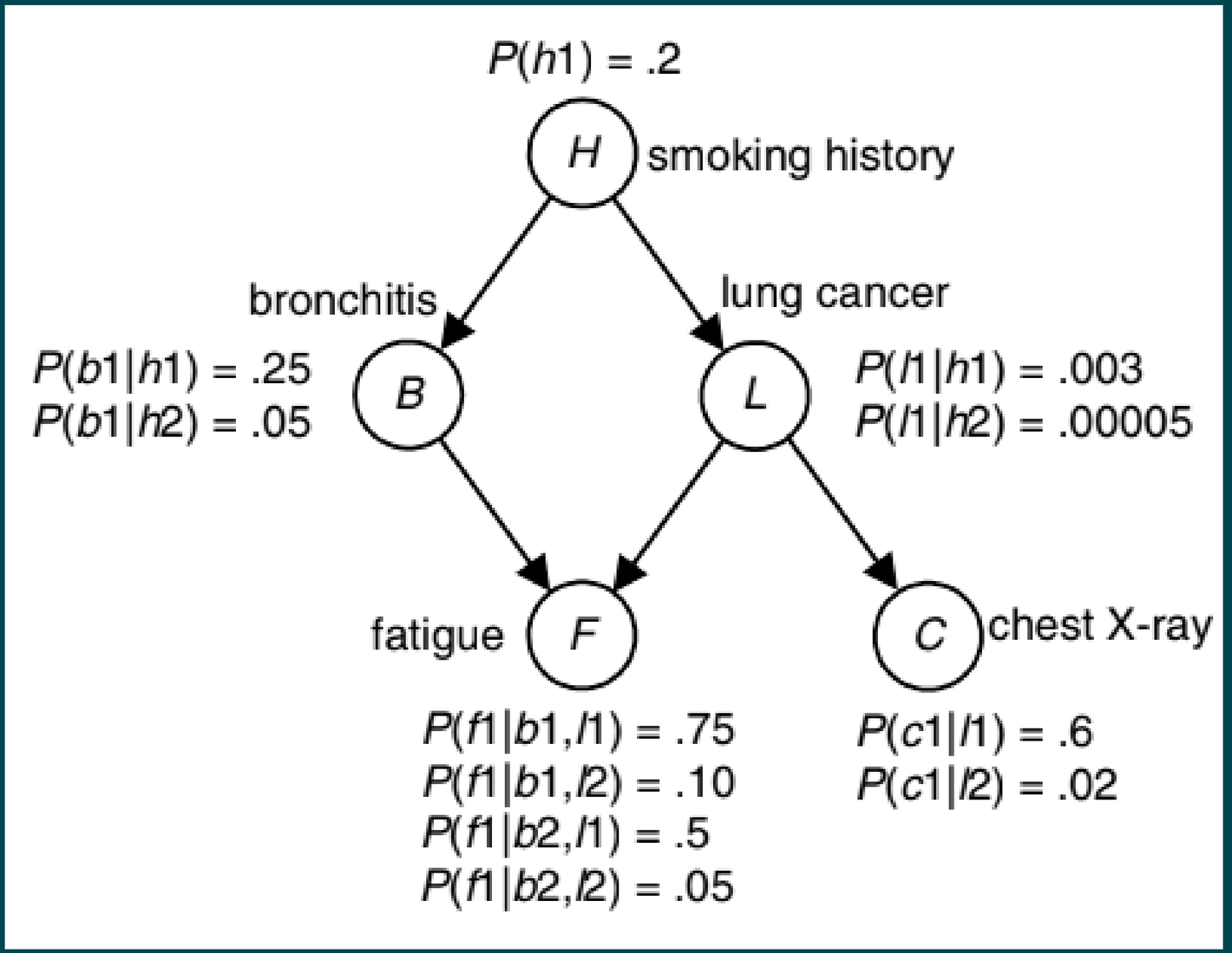
A Notation for Markov Assumption

We will use the notation **I(X, Y|Z)** to denote that **X is independent of Y conditional on Z**, saying that if Z is given, then **no matter how many times we change the probability distribution of Y, there will be no change in the probability distribution of X** (X is independent of Y).

An Example for Causal Network

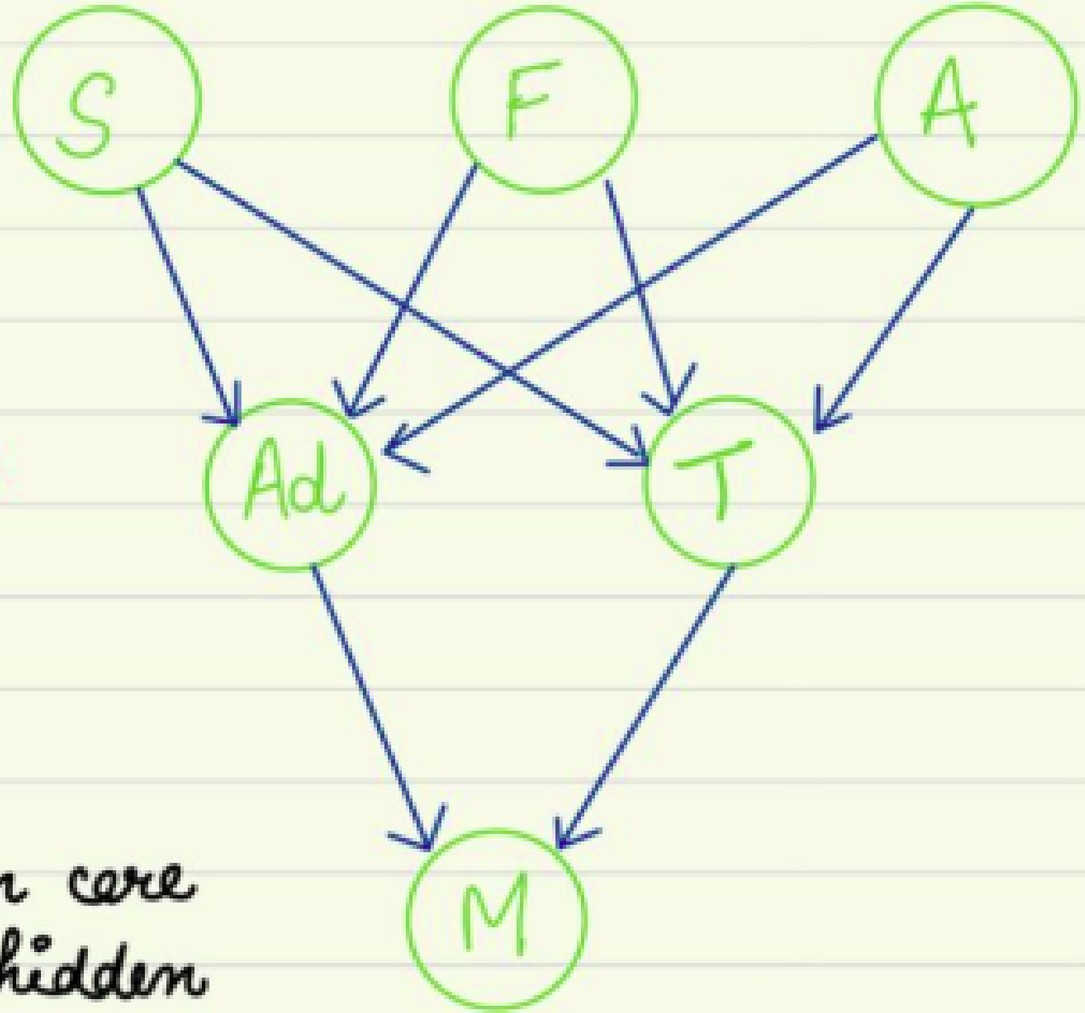
Note that the following hold for this network

- I(B, {L, C}|H)
- I(F, {H, C}|{L, B})
- I(L, B|H)
- I(C, {H, B, F}|L)



Shopping Complex simulation

S : Shops
F : Food stalls
A : Arcades
Ad : Adult Customers
T : Teenager Customers
M : Money earned by the complex



Note :- It has been taken care that there are no hidden variables in the above DAG.

Explanation

- By using the causal DAG, we deduce if there are possible causation relations between two variables present in the problem. In the example above, the dependency set includes {(S,Ad),(S,T),(F,Ad),(F,T),(A,Ad),(A,T),(Ad,M),(T,M)}.

- As this represents a causal network, it will be following the Markov condition giving us the independency set {(Ad,{T, M}|{S, F, A}),(T,{Ad, M}|{S, F, A}),(M,{S, F, A}|{Ad, T})}.

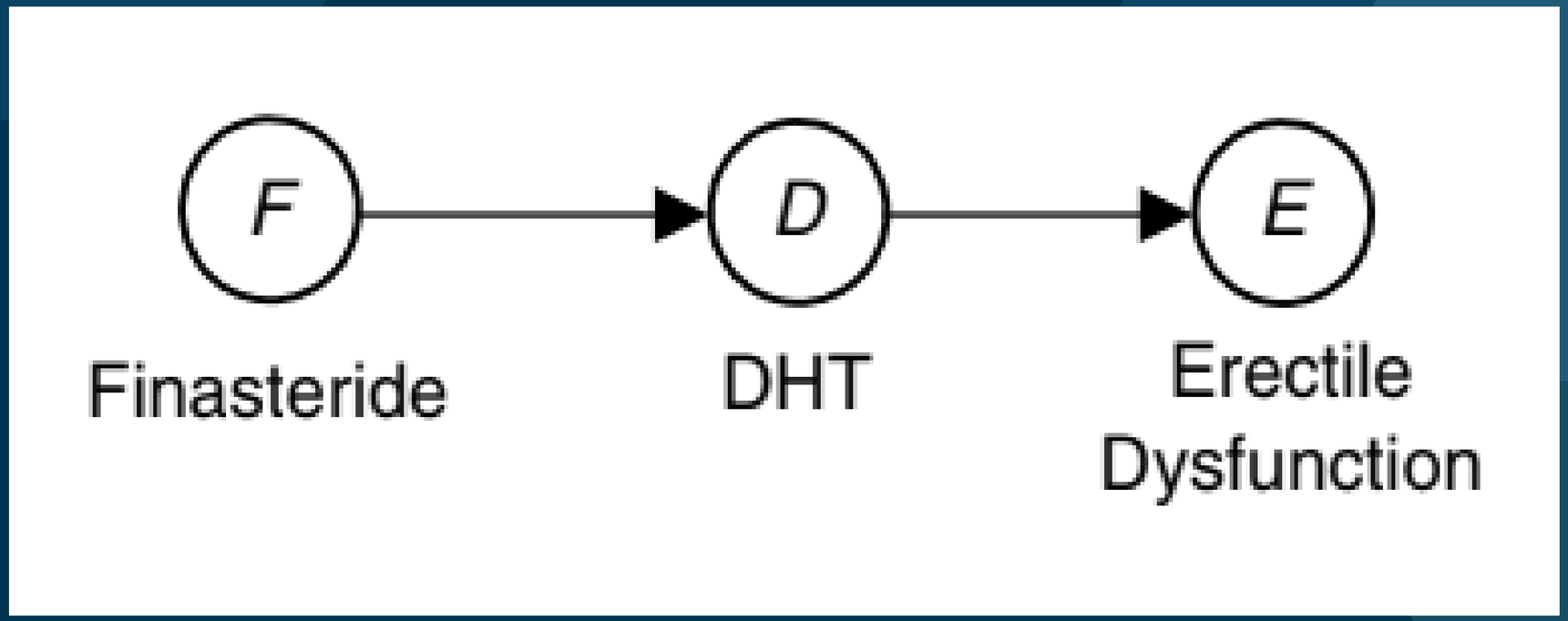
- (Ad,{T, M}|{S, F, A}) mean that Ad is independent of the variables in {T, M} provided values in the variable set {S, F, A}.

Causal Faithfulness Assumption

We say we are making the causal faithfulness assumption for a Causal DAG G, and P is the observed probability distribution,

1. (G, P) satisfies the Markov condition,
2. All conditional independencies in the observed distribution P are entailed by the Markov condition in G,

The Markov condition does not entail I(F, E) for the causal DAG. It only entails I(F, E|D). When a probability distribution has a conditional independency that is not entailed by the Markov condition, the faithfulness assumption does not hold

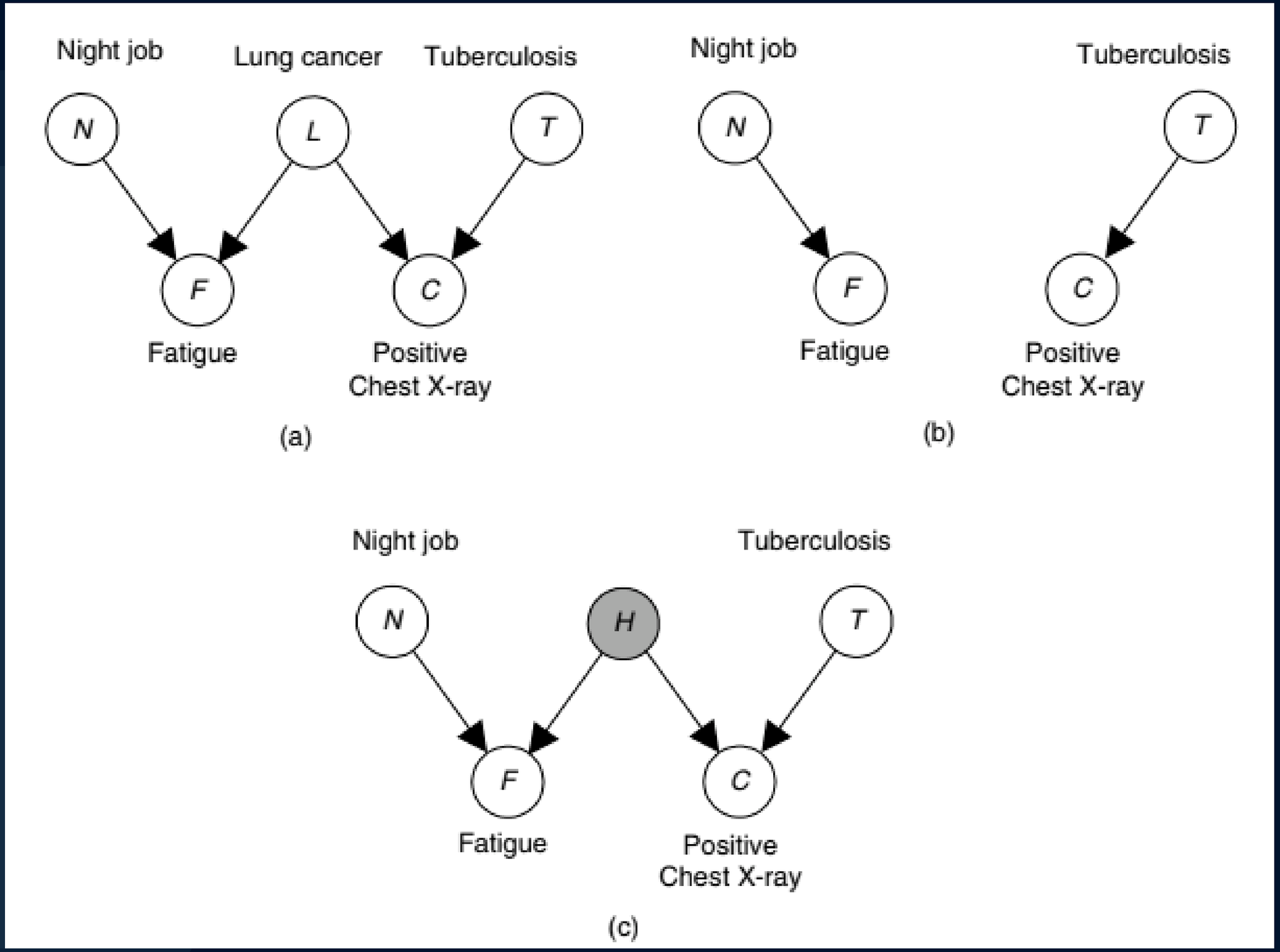


Causal Embedded Faithfulness Assumption

We say we are making the causal embedded faithfulness assumption for a Causal DAG G, and P is the observed probability distribution,

Suppose we have a probability distribution P of the variables in a set V, V is a subset of W, and G is a DAG whose set of nodes is W. Then P is embedded faithfully in W if all and only the conditional independencies in P are entailed by the Markov condition applied to W and restricted to the nodes in V.

An Example


Night job
Lung cancer
Tuberculosis
N
L
T
F
C
Fatigue
Positive
Chest X-ray
(a)
Night job
Tuberculosis
N
T
F
C
Fatigue
Positive
Chest X-ray
(b)
Night job
Tuberculosis
N
H
T
F
C
Fatigue
Positive
Chest X-ray
(c)

Learning Causal Influences

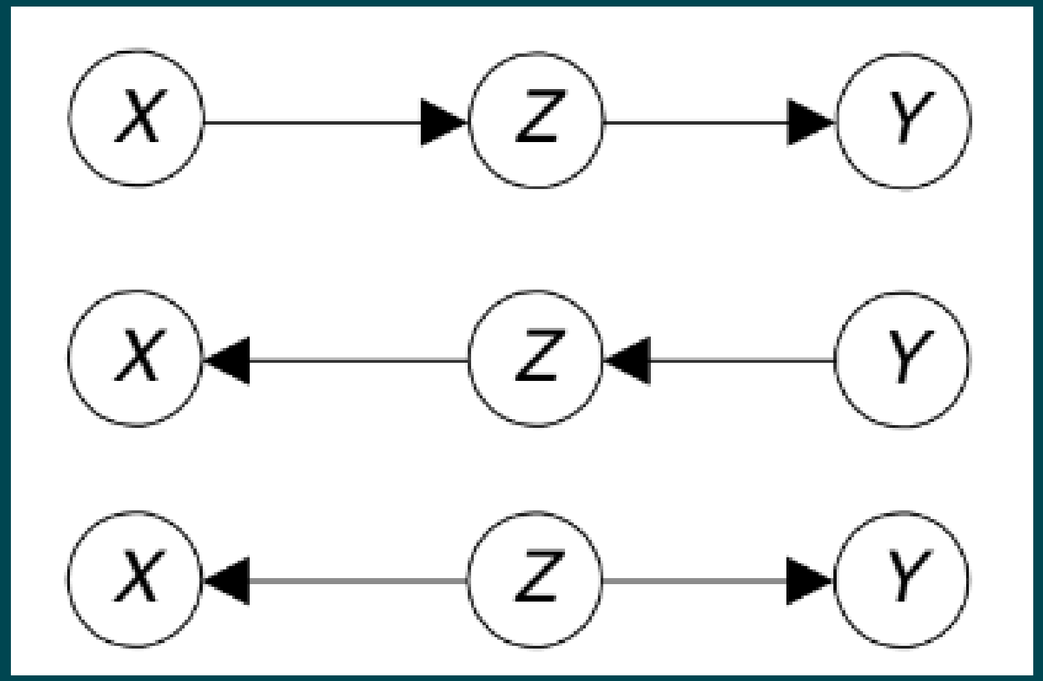
We show how **causal influences** can be learned from data if we make either the causal faithfulness or the **causal embedded faithfulness** assumption.

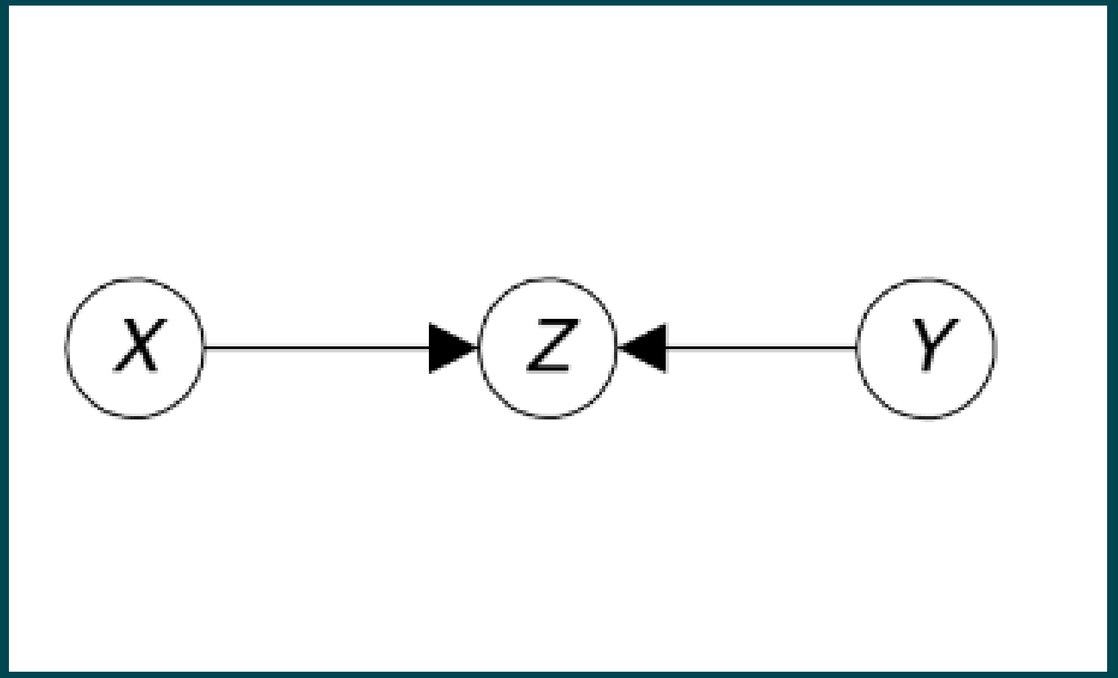
Making the Causal Faithfulness Assumption

*We assume here that the variables satisfy the **causal faithfulness assumption** and we know the conditional independencies among the variables. Given these assumptions, we present a sequence of examples showing how causal influences can be learned.*

Example 1

Suppose V is our set of observed variables, V = {X, Y, Z}, and our set of conditional independencies is {I (X, Y)}.



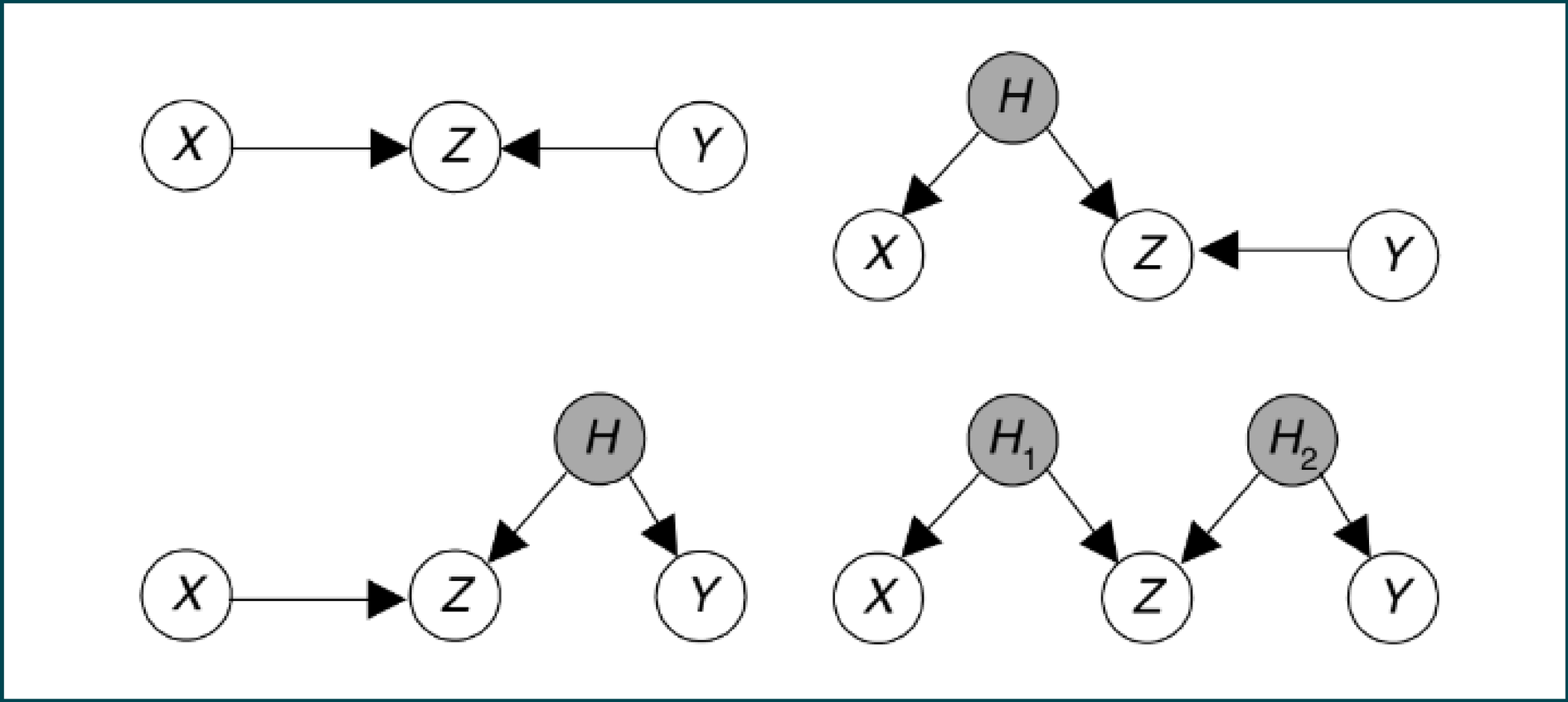


Assuming Only Causal Embedded Faithful-ness

Previously, we mentioned that the most problematic assumption in the ***causal faithfulness assumption*** *is that there must be no hidden common causes, and we eliminated that problem with the* ***causal embedded faithfulness assumption****.*

Example 2

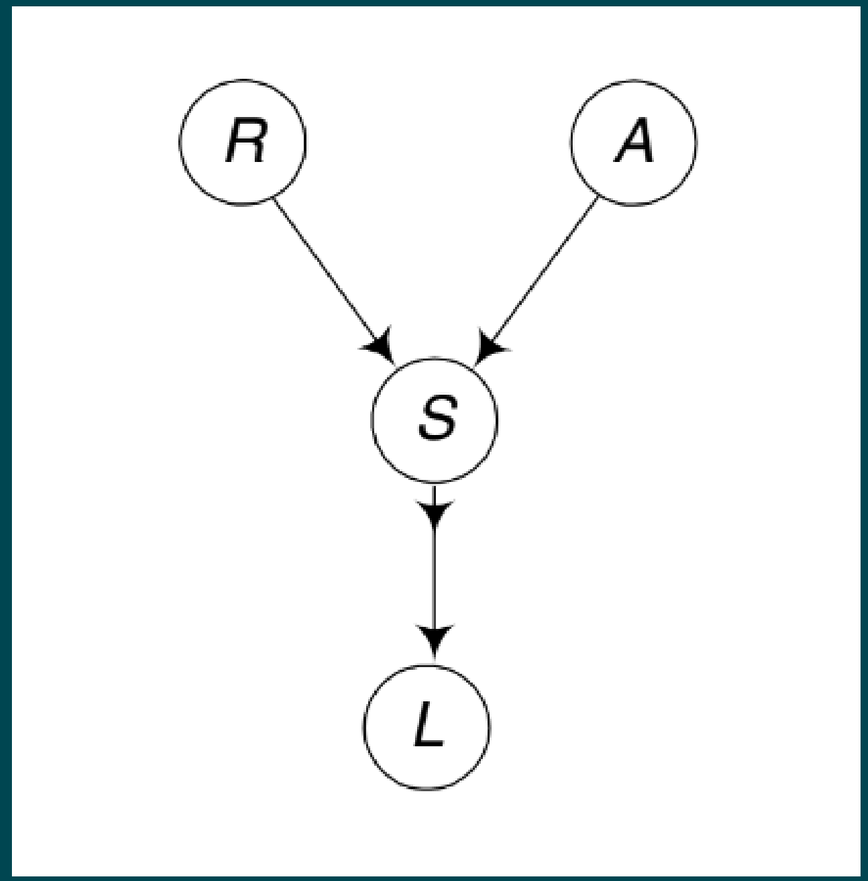
In Example 1 we had V = {X, Y, Z} and the set of conditional independencies {I (X, Y)}.



Example 3

Suppose we have these variables
R: Parent's smoking history
A: Alcohol consumption
S: Smoking behavior

L: Lung Cancer
Suppose further we learn the following conditional independencies from data:
{I (R, A), I (L, {R, A}| S)}.

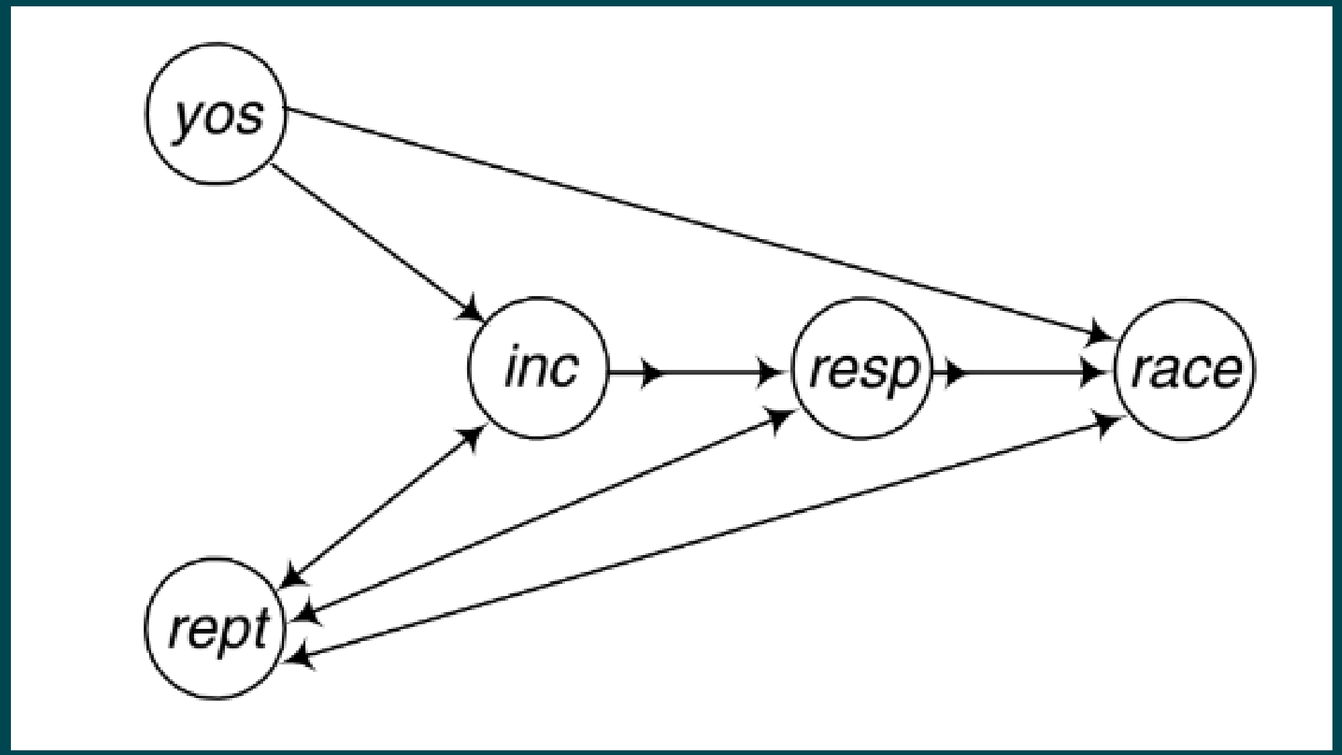


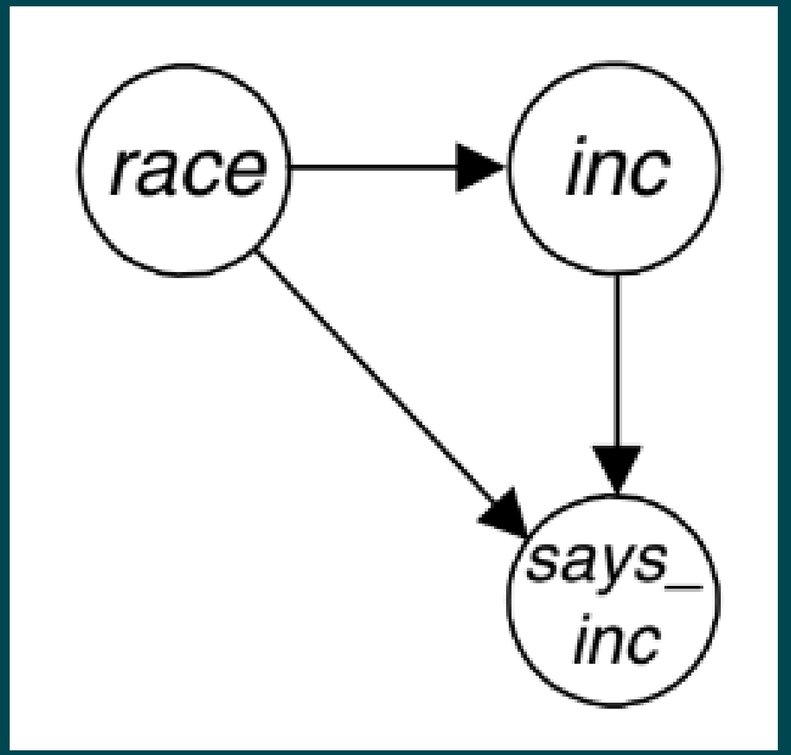
Example 4

Scarville et al. provide a data base obtained from a survey in 1996 of experiences of racial harassment and discrimination of military personnel in the United States Armed Forces. Surveys were distributed to 73,496 members of the U.S. Army, Navy, Marine Corps, Air Force and Coast Guard.

Variable	*What the Variable Represents*
race	*Respondent's race/ethnicity*
yos	*Respondent's years of military service*
inc	*Did respondent experience a racial incident?*
rept	*Was incident reported to military personnel?*
resp	*Did respondent hold military responsible for incident?*

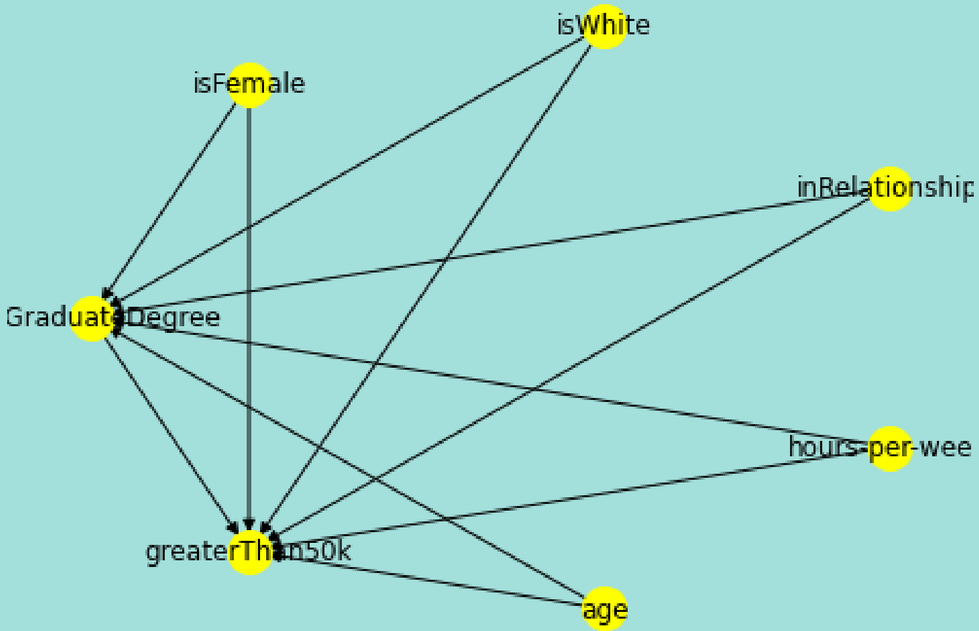
Example 4





Example 5 - Learning Causation From Data

Suppose we want to observe the causation, whether the data 'greaterThan50' (salary greater than 50k) is affected by fact that is person is graduated or not 'isGraduated'.
Several other factors were also taken into consideration while collecting the data and on applying a Causal Model, the following relations can be developed using the Census Income Data Set.



https://archive.ics.uci.edu/ml/datasets/census+income

A Detailed Analysis for Data based approach

```
[5] df.head()
```

	age	hours-per-week	hasGraduateDegree	inRelationship	isWhite	isFemale	greaterThan50k
0	39	40	0	0	1	0	0
1	50	13	0	1	1	0	0
2	38	40	0	0	1	0	0
3	53	40	0	1	0	0	0
5	37	40	1	0	1	1	0

A Detailed Analysis for Data based approach

```
df.describe()
```

	age	hours-per-week	hasGraduateDegree	inRelationship	isWhite	isFemale	greaterThan50k
count	29170.000000	29170.000000	29170.000000	29170.000000	29170.000000	29170.000000	29170.000000
mean	38.655674	40.447755	0.052348	0.406616	0.878334	0.331916	0.245835
std	13.722408	12.417203	0.222732	0.491211	0.326905	0.470909	0.430588
min	17.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	28.000000	40.000000	0.000000	0.000000	1.000000	0.000000	0.000000
50%	37.000000	40.000000	0.000000	0.000000	1.000000	0.000000	0.000000
75%	48.000000	45.000000	0.000000	1.000000	1.000000	1.000000	0.000000
max	90.000000	99.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Analysis

We have used 'causalinference' library for this demonstration, which is a software package that implements various statistical and econometric methods used in the field variously known as Causal Inference, Program Evaluation, or Treatment Effect Analysis.

We have tried to establish relationships between variables by analysing how the 'treatment' criterion affects the main variable and the subsidiary variables.

we have done this by changing the 'treatment' condition in the model and studying it's effects on the other variables
the following slides show these results in detail

Treatment 1

hasGraduateDegree

```
Treatment: hasGraduateDegree
X0: age
X1: hours-per-week
X2: inRelationship
X3: isWhite
X4: isFemale

Summary Statistics

                    Controls (N_c=27643)        Treated (N_t=1527)
      Variable         Mean         S.d.         Mean         S.d.      Raw-diff
-------------------------------------------------------------------------------
             Y        0.228        0.420        0.567        0.496         0.339

                    Controls (N_c=27643)        Treated (N_t=1527)
      Variable         Mean         S.d.         Mean         S.d.      Nor-diff
-------------------------------------------------------------------------------
            X0       38.341       13.794       44.351       10.894         0.484
            X1       40.242       12.399       44.169       12.156         0.320
            X2        0.401        0.490        0.506        0.500         0.211
            X3        0.875        0.331        0.937        0.243         0.214
            X4        0.332        0.471        0.324        0.468        -0.019
```

Having a graduate degree implies you have a 56.7% chance of earning more than 50,000/year

compared to a 22.8% chance if you don't have a graduate degree

More observations -
A person having a graduate degree is also older, works more hours per week, has a slightly higher chance of being in a relationship and being white

Treatment 2

inRelationship

```
Treatment: inRelationship
X0: age
X1: hours-per-week
X2: hasGraduateDegree
X3: isWhite
X4: isFemale

Summary Statistics

                    Controls (N_c=17309)       Treated (N_t=11861)
      Variable          Mean         S.d.          Mean         S.d.     Raw-diff
-------------------------------------------------------------------------------
             Y         0.101        0.301         0.457        0.498        0.357

                    Controls (N_c=17309)       Treated (N_t=11861)
      Variable          Mean         S.d.          Mean         S.d.     Nor-diff
-------------------------------------------------------------------------------
            X0        35.022       13.617        43.958       12.048        0.695
            X1        37.867       12.211        44.214       11.735        0.530
            X2         0.044        0.204         0.065        0.247        0.095
            X3         0.842        0.365         0.932        0.252        0.288
            X4         0.559        0.496         0.000        0.009       -1.593
```

Being in a relationship implies you have a 45.7% chance of earning more than 50,000/year

compared to a 10.1% chance if you are not in a relationship

More observations -
A person in a relationship is also older, works more hours per week and has a 0% chance of being a female (according to the dataset)

Treatment 3

isWhite

```
Treatment: isWhite
X0: age
X1: hours-per-week
X2: hasGraduateDegree
X3: inRelationship
X4: isFemale

Summary Statistics

                    Controls (N_c=3549)       Treated (N_t=25621)
      Variable         Mean         S.d.         Mean         S.d.     Raw-diff
--------------------------------------------------------------------------------
             Y        0.132        0.338        0.262        0.440        0.130

                    Controls (N_c=3549)       Treated (N_t=25621)
      Variable         Mean         S.d.         Mean         S.d.     Nor-diff
--------------------------------------------------------------------------------
            X0       37.546       13.027       38.809       13.809        0.094
            X1       38.521       10.678       40.715       12.616        0.188
            X2        0.027        0.162        0.056        0.230        0.145
            X3        0.228        0.419        0.431        0.495        0.444
            X4        0.483        0.500        0.311        0.463       -0.357
```

Being "White" implies you have a 26.2% chance of earning more than 50,000/year

compared to a 10.1% chance otherwise

More observations -
A "white" person has a significantly higher chance of being in a relationship and is less likely to be a female, with negligible differences in other variables.

Treatment 4

isFemale

```
Treatment: isFemale
X0: age
X1: hours-per-week
X2: hasGraduateDegree
X3: inRelationship
X4: isWhite

Summary Statistics

                    Controls (N_c=19488)        Treated (N_t=9682)
      Variable          Mean         S.d.          Mean         S.d.     Raw-diff
--------------------------------------------------------------------------------
             Y         0.313        0.464         0.111        0.314       -0.202

                    Controls (N_c=19488)        Treated (N_t=9682)
      Variable          Mean         S.d.          Mean         S.d.     Nor-diff
--------------------------------------------------------------------------------
            X0        39.588       13.441        36.779       14.086       -0.204
            X1        42.495       12.212        36.326       11.787       -0.514
            X2         0.053        0.224         0.051        0.220       -0.009
            X3         0.609        0.488         0.000        0.010       -1.763
            X4         0.906        0.292         0.823        0.382       -0.244
```

Being a female implies you have a 11.1% chance of earning more than 50,000/year

compared to a 31.3% chance otherwise

More observations -
A Female has a lower age, works significantly less hours per week and has a 0% chance of being in a relationship (according to the dataset)

Relevant Code

1. Random Walk Github Repository
   https://github.com/Causal-Inferance/Causal-Network-and-Random-Walk
2. Causal Inference using Propensity Score:
   https://github.com/Causal-Inferance/Causal-Inferance-propensity-score
3. Propensity Score Dataset:
   https://github.com/Causal-Inferance/propensity-score-dataset
4. Causal Networks and their explanation:
   https://github.com/Causal-Inferance/Learning-Causal-Inferance
5. GES Search Code (Greedy Equivalence Search):
   https://github.com/Causal-Inferance/Causal-Inferance-GES

Refrences

1. Neapolitan, R.E., Jiang, X. (2006). A Tutorial on Learning Causal Influence. In: Holmes, D.E., Jain, L.C. (eds) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol 194. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33486-6_2
2. Huang, B., Zhang, K., Lin, Y., Schölkopf, B., & Glymour, C. (2018, July). Generalized score functions for causal discovery. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1551-1560).
3. Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov), 507-554.
4. Talebi, S. (2022) *Causal effects via propensity scores*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/propensity-score-5c29c480130c (Accessed: December 4, 2022).
5. *Welcome to causal-learn's documentation!* (no date) *causal*. Available at: https://causal-learn.readthedocs.io/en/latest/index.html (Accessed: December 4, 2022).
6. *UCI Machine Learning Repository: Census Income Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/census+income (Accessed: December 4, 2022).

NetworkX documentation NetworkX. Available at: https://networkx.org/ (Accessed: December 4, 2022)

Group 3

Harsh Gujarathi
2020A7PS1712G

Yash Khanna
2020A7PS1713G

Dhruv Rohira
2020A7PS1725G

Krishanu Shan
2020A7PS1728G

Satyam Bansal
2020A7PS0171G

Vaibhav Jaiswal
2020A7PS1379G

Thank You!