


Why Causality


OLS and The basic Terms.


Yongwoo Jeong


Job Postings Examples (Look at Your Competitors in Overseas)


LinkedIn 1 <https://www.linkedin.com/jobs/view/3645689495>





 Home


 My Network


 Jobs


 Messaging


 Notifications


 Me

 For Business

 Post a job for free











Data Scientist / Machine Learning Specialist (Application Development | Level 2 (CAD))


Soho Square Solutions · Montreal, QC (On-site) · 6 days ago · 135 applicants


 Contract · Mid-Senior level

 201-500 employees · Business Consulting and Services

 See how you compare to 135 applicants. [Reactivate Premium](#)

 Actively recruiting

 Easy Apply

 Saved

3+ years of experience with design and implementation of machine learning, predictive analysis, data science, knowledge bases, recommendation systems, information retrieval.

Strong understanding of the foundational concepts and applied experience in Machine Learning (ideally, a combination of excellent academic research and high-impact commercial projects).

In depth understanding of common Machine Learning algorithms (e.g., for classification, regression, and clustering).

In depth knowledge of advanced statistical theories, methodologies, and inference tools.


Proven track record in some of the advanced topics such as Bayesian inference, hierarchical models, deep learning, Gaussian processes, and causal inference.


Practical experience in preparing data for Machine Learning integrating with big-data platforms and high-performance computing ecosystems.


Ability to work with global, cross-functional teams


Excellent oral and written communication skills.


LinkedIn 2 <https://www.linkedin.com/jobs/view/3641743230>





 Home


 My Network


 Jobs


 Messaging


 Notifications


 Me

 For Business

 Post a job for free











Latam Data Scientist


Khan Academy · São Paulo, São Paulo, Brazil (On-site) · 1 month ago · 14 applicants


 Full-time · Entry level

 51-200 employees · E-Learning Providers

 See how you compare to 14 applicants. [Reactivate Premium](#)

 Your profile matches this job

 Apply

 Saved


In this role you will:

- Identify and measure success of our district team's initiatives through target setting, forecasting, monitoring, and ongoing analysis of core outcome metrics
- Perform advanced exploratory analyses on large sets of data to extract insights on teacher and learner behavior and inform decisions on investments and initiatives
- Collaborate with Engineering teams to improve tools and datasets and work self-sufficiently with data pipelines (e.g. ETLs) on an as-needed basis
- Use your expertise in experimentation (i.e. AB testing, causal inference) to measure the impact of various programs and interventions
- Contribute to the development of analytics tools and data applications to enhance data access for the district team and Latam leadership, including the development of Latam dashboards to monitor the annual plan goals and objectives (KPIs and targets)
- Collaborate with efficacy researchers to provide curated datasets and/or dashboards for measuring learner outcomes.
- Develop and maintain data dashboards to support the implementation in district partnerships and monitor the adoption of communities of practice with weekly or monthly updates of targets attainment
- Own and manage the prioritization roadmap for data initiatives and projects to support reach, engagement, and resource allocation throughout Latam

More...


This is Their Mindset

It might impact overseas sales in the future!




Principal Data Scientist

Cabify
Madrid, Community of Madrid, Spain (On-site)

 Actively recruiting


1 month ago · **13 applicants**


- Experience publishing in peer-reviewed scientific journals.
- Experience with state-of-the-art methods in topics like deep learning models, time series analysis, bayesian statistics **causal inference, etc.**
- Experience with geo data modeling, analysis and visualization.
- Experience/knowledge in behavioral economics.
- Experience in mentoring/teaching.




Senior Data Scientist

Fever
Madrid, Community of Madrid, Spain (Remote)

 2 school alumni work here


1 month ago ·  Easy Apply

- Have significant expertise in technical areas such as Bayesian modeling **causal inference, real-world recommendation systems, ...**
- Have relevant work experience in marketing or internet tech companies
- Identify yourself as both a scientist and a hacker
- Have experience with current ML and scientific computing frameworks such as Pytorch, JAX, etc.



Senior Data Scientist

InspHire
Tel Aviv-Yafo, Tel Aviv District, Israel (On-site)

 Actively recruiting

1 day ago

You Should Have

- 3+ years of experience as a data scientist
- Experience with modern data tools and frameworks (ex: Pandas, Numpy, XGBoost, Scikit-Learn, etc)
- Deep understanding of modern machine learning algorithms and statistics
- Being able to own your models from development to production deployment
- Excellent communication skills to explain results to product managers, data analysts, and engineers
- Experience with Big-Data tools like Spark – Advantage
- **Experience with Causal Inference – Advantage**
- Passion and motivation to build a growth machine

Desired Skills and Experience

Statistics, Data Tools, Data Science, Development, Communications, Spark, Pandas, NumPy, scikit-learn, XGBoost, Machine Learning Algorithms, Production Deployment, Big Data Tools, **Causal Inference**

Steps

With API Examples

- Basic Linear Regression and OLS [with Jupyter Lab code](#)
- Basic Terms for OLS [with Good Notes](#)
- Why we can't use OLS for Causality [with Jupyter Lab code](#)
- DAG Basics and Basic Terms
- Causal Discovery and API [with Jupyter Lab code](#)
- Causal Inference and API [with Jupyter Lab code](#)
- More Causal Inference APIs

Basic Terms

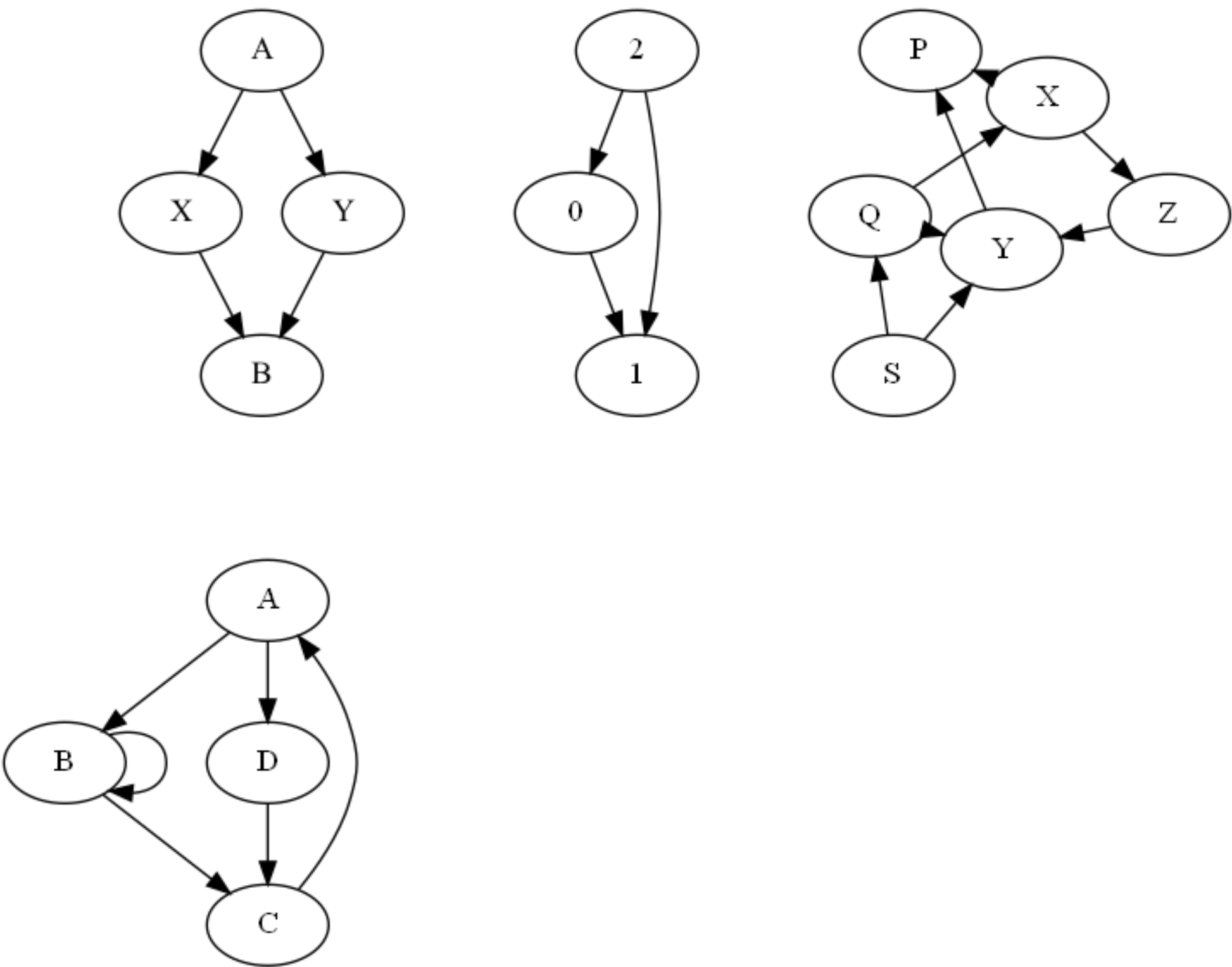
...

- OLS, Durbin-Watson, F-statistics, AIC, BIC, R-squared, etc : **Good note**
- Endogenous vs. Exogenous
- Instrumental Variables
- Counterfactual
- Chain, fork, collider and confounder
- Intervention or do-operation vs. conditional prob.
- D-separation, front-door and back-door adjustments
- ATT, ATE, ATC, CATE, ITE
- Meta Learners

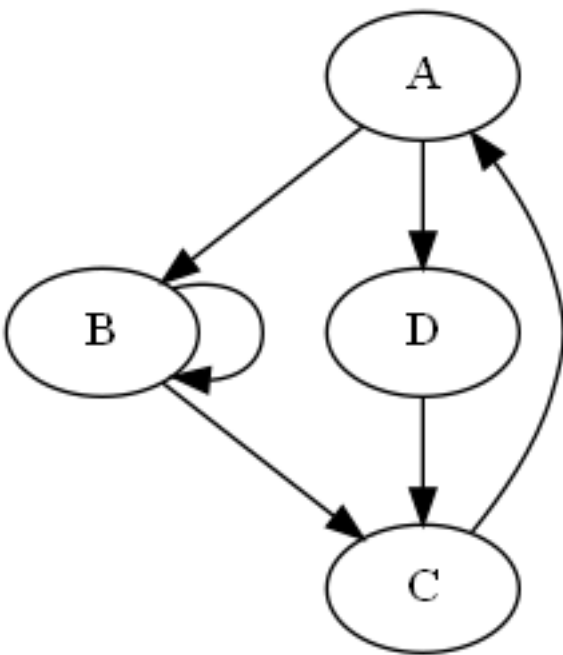
DAG and Conditional (in) dependence in Graph

Directed Acyclic Graph

DAG: Directed Acyclic Graph



DCG: Directed Cyclic Graph



DAG and Conditional (in) dependence in Graph

Directed Acyclic Graph

- Chain



- Fork

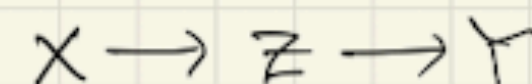


Z: confounder

- Collider

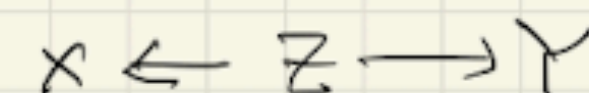


1. Chain

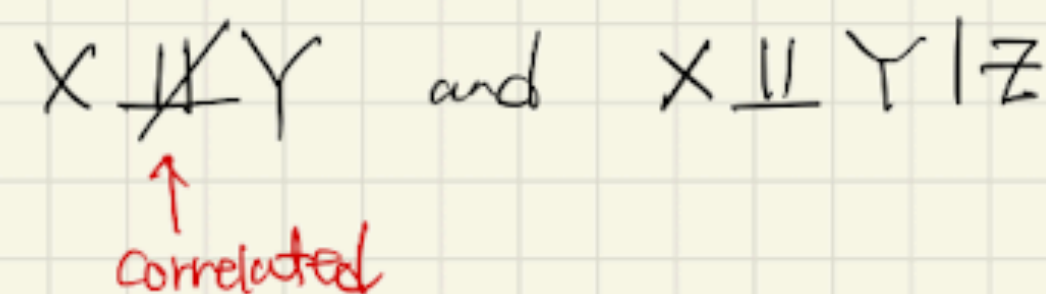


X is independent Y, conditional on Z ($X \perp\!\!\!\perp Y | Z$)

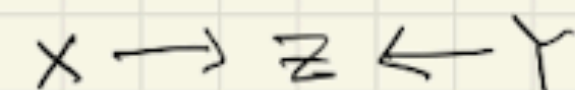
2. Fork



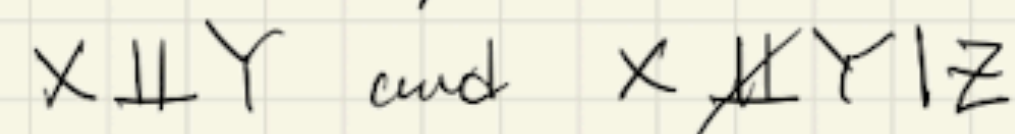
X and Y are dependent but become independent conditional on Z



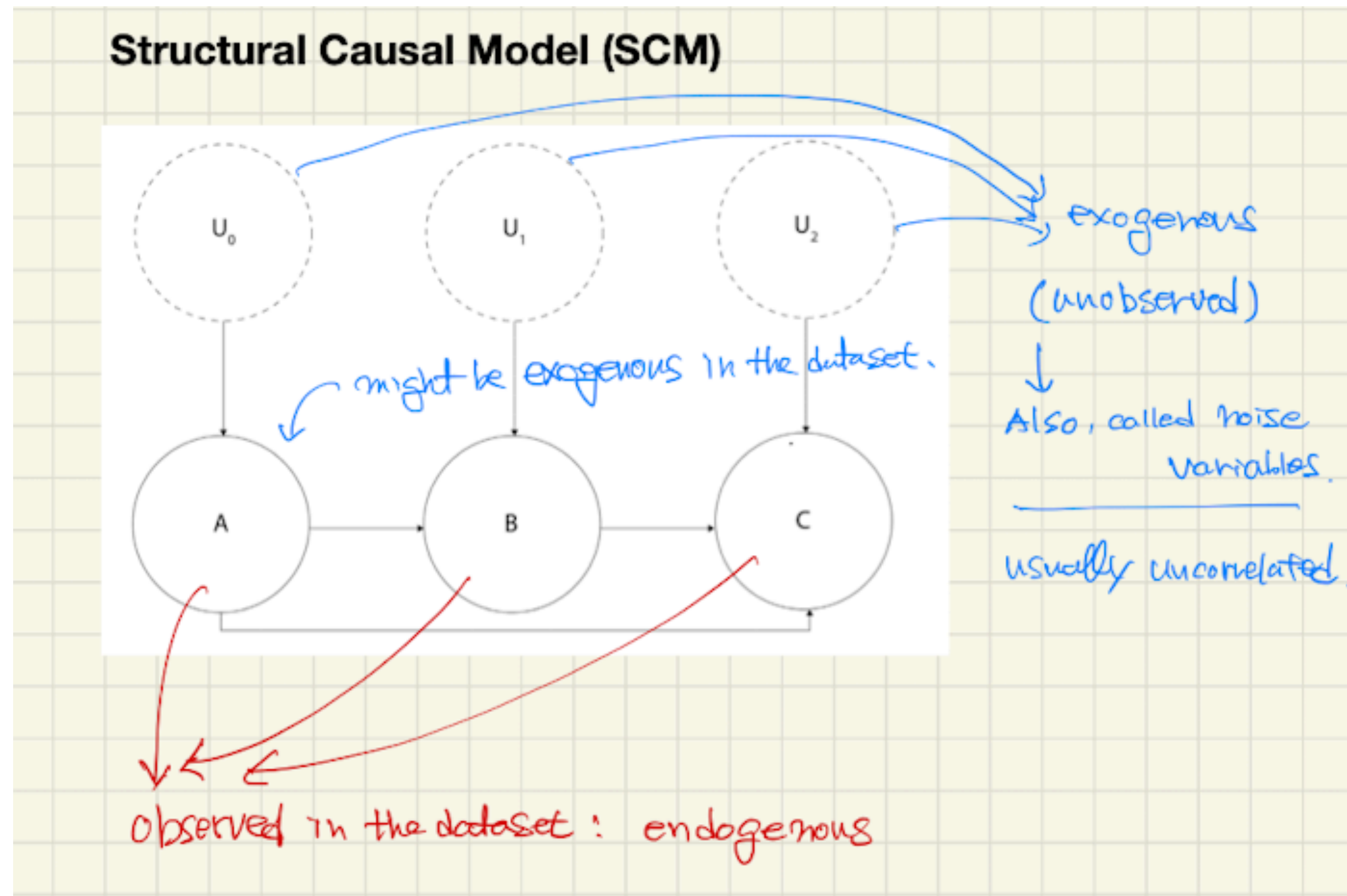
3. Collider



If Z is a collider, X and Y are unconditionally independent but become dependent conditional on Z



Endogenous vs. Exogenous vs. Instrumental Variable



Exogenous: 어떤 변수에도 영향을 안받는 변수, 하지만 실제로 그렇게 보일 뿐 일 수 있음.
데이터 셋에서만 그렇게 보일 뿐 실제로 데이터 밖의 어떤 다른 변수에 영향을 받을 수 있음.
(예, 날씨)

Endogenous: 어떤 변수에서 영향을 받는 변수 (예, 강수량, 습도, 차량 통행량, 등등)

Instrumental Variable: 어떤 영향성의 원인 X, 그 결과 Y가 존재한다고 할 때
그 X의 원인이 될 수도 있다고 믿어지는 변수를 말함.

이 자체는 Exogenous 함. 줄여서 IV라고 칭함.

좌측 그림에서 B가 X에 해당되고 C가 Y에 해당된다고 가정할 때 A는 IV에 해당됨.

이 SCM에서 A는 C에 영향을 B를 통해서만 영향을 준다고 할 수가 있다.

IV는 중간단계를 거쳐서 오기 때문에 결과 단계인 C에 직접적인 영향이 없을 때 IV로 선택된다.

Confounder와 IV를 혼동하지 말 것!!!

Why We Can't USE OLS for Causality

- Humidity after raining in the desert. Does the humidity cause the rain?
- Confounder issue in OLS : **Good Node Example**
- Counter-factual Issue due to lack of a time-machine (or time-stone)
 - Randomly chosen 300 of 1000 patients are given a new drug and recovered in 3 days. Is this drug effective? Really? How do we minimize this doubt?
- **Code Examples (Association, Intervention, Counterfactual)**

Intervention vs. Conditional Prob.

- Conditional Prob:
 - $P(A|B,C)$: 데이터 셋에서 B, C가 있을 때 A를 찾아 그 카운트를 세어 확률을 계산, 데이터셋 변경이 없음.
- Intervention:
 - $P(A|B=b, C=c)$: 데이터 셋에 모든 B를 b로 대체하고 C는 c로 대체하여 데이터셋을 변경한 다음에 A가 될 확률을 계산, 데이터셋 변경이 있음.

do-Operation

do-Operation (Intervention)

before do

```
graph LR; X((X)) --> Z((Z)); X((X)) --> Y((Y)); Z((Z)) --> Y((Y));
```

$$X = f_X(X, Z)$$
$$Z = f_Z(Z)$$
$$Y = f_Y(X, Z)$$

After do

```
graph LR; X((X)) --> Y((Y)); Z((Z)) --> Y((Y));
```

$$X = x' \rightarrow \text{Replace all } X \text{ with } x'$$
$$Z = f_Z(Z)$$
$$Y = f_Y(x', Z)$$

$$E(Y=y \mid X=x'') - E(Y=y \mid X=x')$$

→ Effect of Intervention

Before do-Operation

After do-Operation

X

X

0.1

0.1

1.2

0.1

3.1

0.1

11

0.1

34

...

0.1

22

...

0.2

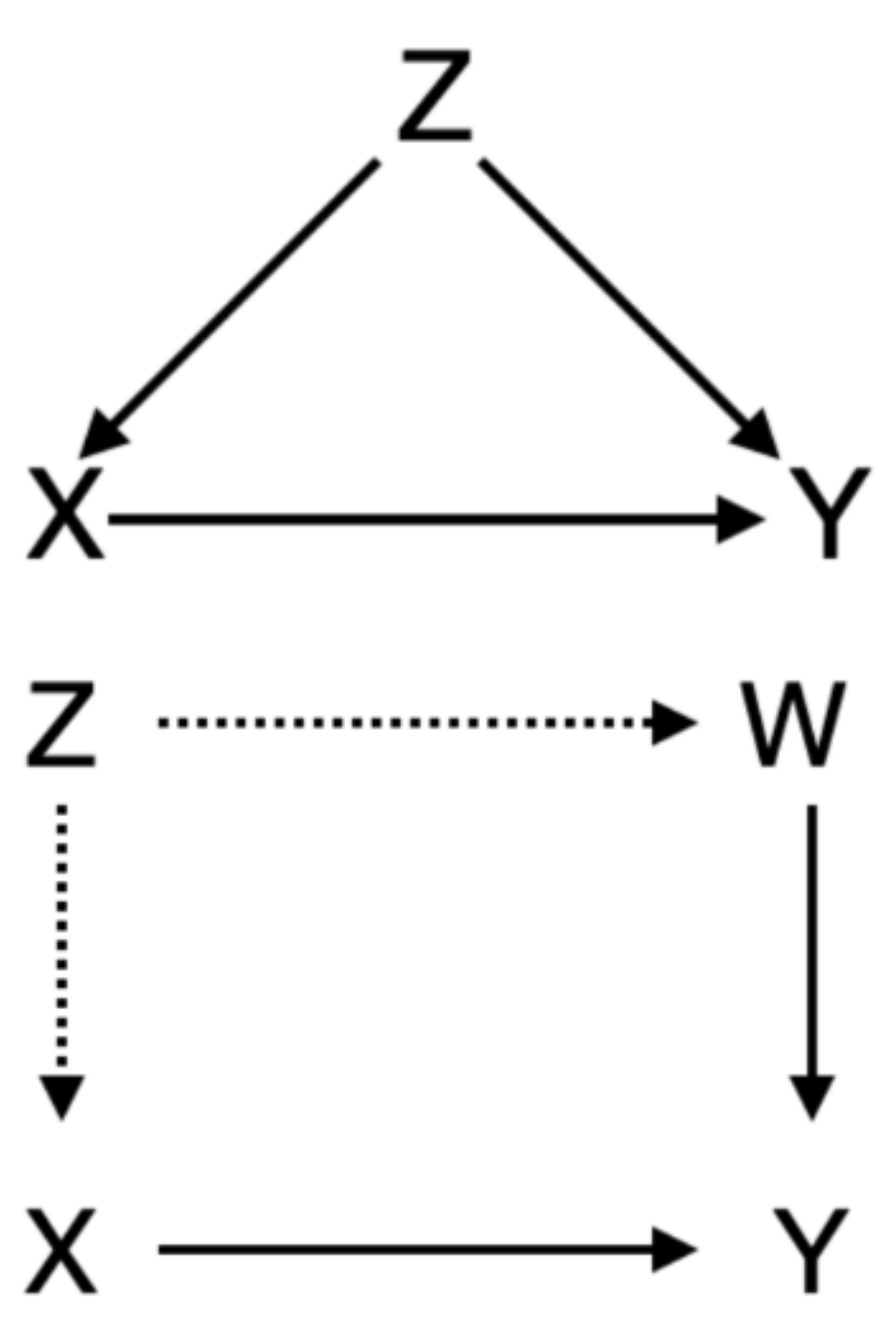
0.1

Years ago:
Did it manually

Now:
Done automatically in API

Backdoor/Frontdoor Adjustments

Backdoor



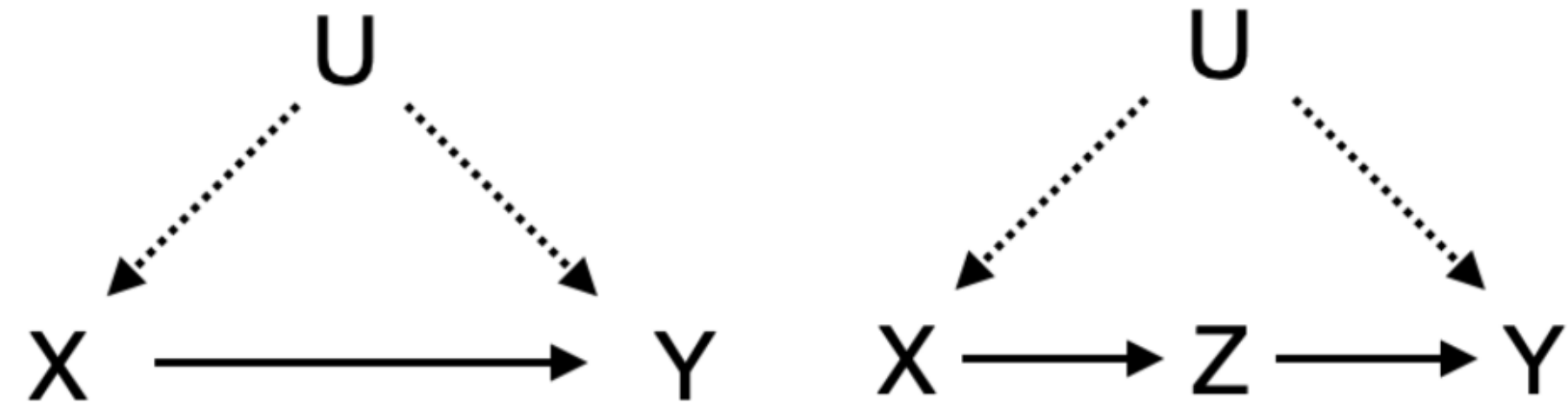
$$P(Y = y | do(X = x)) = \sum_z P(Y = y | (X = x, Z = z)P(Z = z)$$

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | (X = x, W = w)P(W = w)$$

Back-Door Criterion의 예;
위: 관찰된 Z 사건이 존재할 때,
아래: 관찰된 W 사건이 존재하지만 Z 사건이 존재한다고 의심이 될 때

Backdoor/Frontdoor Adjustments

Frontdoor



Genotype (니코틴을 애초 부터 좋아하는 DNA)은 담배 회사가 주장한 가설일 뿐이고 데이터에 없으므로 점선으로 되어 있다.

$$P(Z = z | do(X = x)) = P(Z = z | X = x)$$

$$P(Y = y | do(Z = z)) = \sum_z P(Y = y | Z = z, X = x)P(X = x)$$

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | do(Z = z))P(Z = z | do(X = x))$$

$$P(Y = y | do(X = x)) = \sum_z \sum_{x'} P(Y = y | Z = z, X = x')P(X = x')P(Z = z | X = x)$$

Definitions of Treatment Effects

ATT: Average Treatment Effect on the Treated

관측된 데이터셋에서 실제로 치료를 받은 환자 대상

ATC: Average Treatment Effect on the Control

관측된 데이터셋에서 치료를 받지 않은 환자 대상

$$ATE = \frac{1}{N} \sum_i \tau_i \quad \text{where } \tau_i = Y_i^1 - Y_i^0$$

관측된 데이터셋에서 실제로 치료를 받은 환자가
치료를 받지 않은 환자 보다 받은 평균 효과의 크기

$$ATT = \frac{1}{N_{T=1}} \sum_{i_{T=1}} \tau_i \quad "$$

$$ATC = \frac{1}{N_{T=0}} \sum_{i_{T=0}} \tau_i \quad "$$

Definitions of Treatment Effects

Name	Gender	Treatment	Outcome with Treatment	Outcome without Treatment	Individual Treatment Effect
Joe	Male	Yes	10	11	-1
Mike	Male	No	8	7	1
Ashley	Female	Yes	5	12	-7
Emily	Female	No	7	8	-1

ATT = $\sum \text{Joe, Ashley} = (-1-7)/2 = -4 = 4 \text{ days shorter to recover.}$

ATC = $\sum \text{Mike, Emily} = (1-1)/2 = 0 = 0 \text{ days shorter to recover.}$

CATE: Conditional Average Treatment Effect.

↳ (e.g. male only or female only...)

$(1-1)/2 = 0$ ←

↳ $(-7-1)/2 = -4$

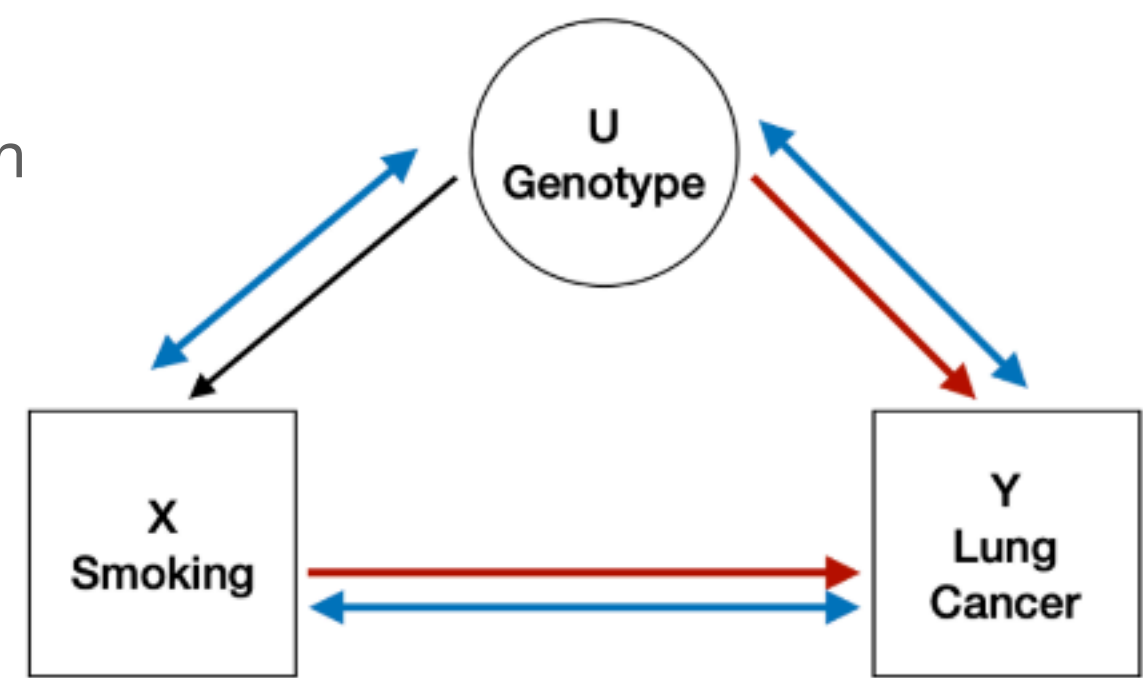
ITE: Individual Treatment Effect, (e.g. Emily only)

Example of Correlation vs. Causal Inference vs. Causal Discovery

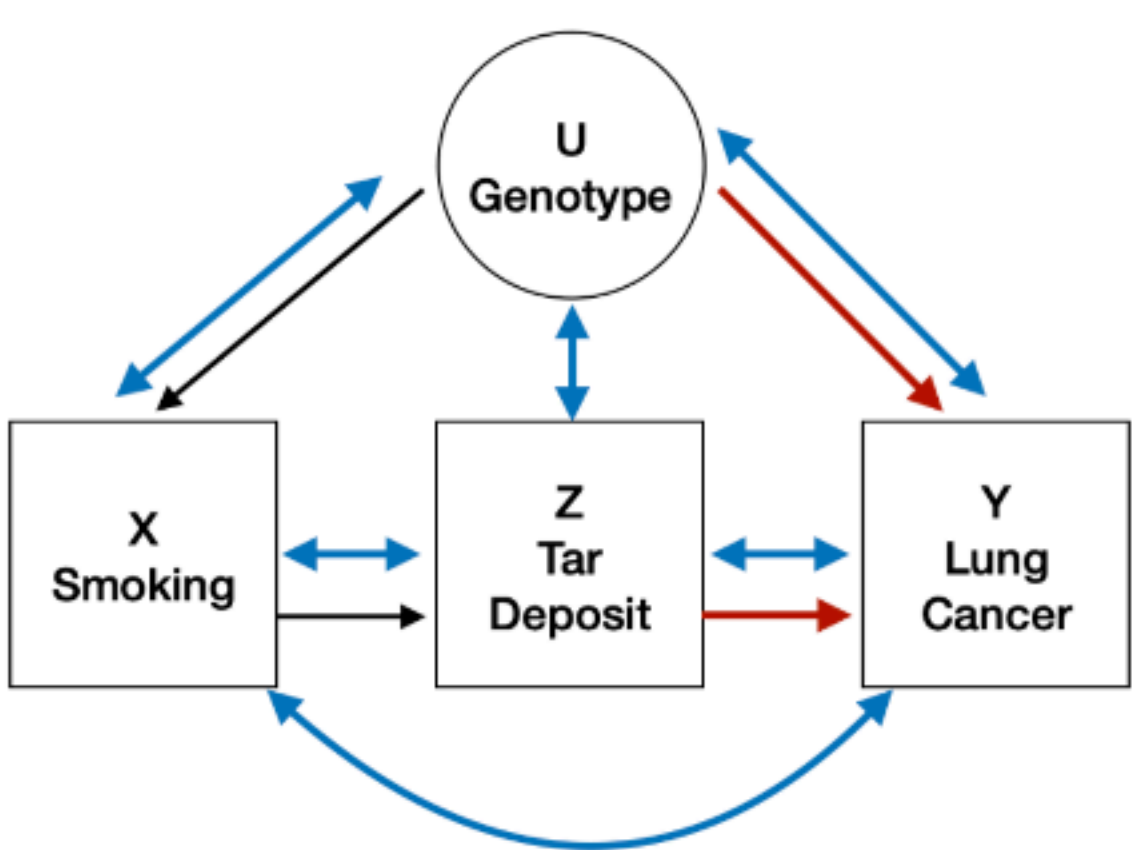
On Lung Cancer Issue

20세기 중반에 담배회사가 주장한 담배 무해론

DAG: Directed Acyclic Graph



이를 반박한 담배 유해론



- Causal Discovery
- Causal Inference
- Causal Inference

Correlation/OLS

OLS vs. Causal Discovery vs. Causal Inference

Step by Step

- **OLS**

- Null hypothesis
- Collect evidences to reject it
- If less than 0.05 or 0.01, then reject the null hypothesis and suggest an alternative hypothesis
- Or accept the null hypothesis.

- **Causal Discovery**

- Get Adjacency matrix
- Build the DAG then suggest the cause
- Use ML or causal inference to prove it.

- **Causal Inference**

- Front-door/back-door adjustments for d-separation/de-confounding
- Suggest a hypothesis
- Use multiple refute method to prove it
- If greater than 0.05 or 0.01, then accept the suggested hypothesis
- Or re-calculate the hypothesis.

Pros and Cons of Causal Discovery

- Pros:
 - It can find all conditional independence, then constructs a DAG compatible with this independence.
 - Causal discovery empirically distinguishes colliders from chains/forks
- Cons:
 - **It can't distinguish chains from forks.**
 - **It only learns the underlying DAG up to a certain equivalent class.**
- Causal Inference를 위한 가이드는 주지만 현재는 그 자체로 완벽한 솔루션은 아닐 수 없음. (앞으로는 업그레이드 될 수도?)
- **이 결과로 주어진 DAG으로 Causal Inference로 결과를 내야 완벽해짐.**