Class 6: Directed Acyclic Graphs (DAG)

Professor Song Yao
Olin Business School

Customer Analytics

What is DAG?

DAG (Directed Acyclic Graph) is a type of graph used to represent causal relationships between variables

- Visualization of Assumptions: Explicitly represent causal relationships, making assumptions transparent and easier to evaluate.
- Identification of Confounders: Systematically identify confounders.
- Bias Detection: Help highlight potential sources of bias, such as confounding, selection bias, or collider bias, enabling their mitigation.
- Communication: DAGs distill complex causal relationships into clear representations for communication.

Basics of DAG

Key Elements in a DAG:

- **Nodes**: Variables in the system (e.g., price discount, ads, sales).
- **Edges**: Arrows represent causal (occassionally correlational) relationships between the variables.
- **Directed**: The edges (arrows) in the graph indicate the direction of causality
- **Acyclic**: No cycles, hence rendering it impossible to start with one node, follow the directed edges, and return to the same node.

Example:

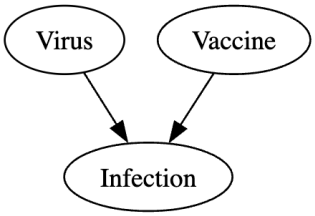
- **Nodes**: Virus, Vaccine, Infection
- **Edges**: The virus spread increases infections; vaccinations reduces infections
  - Virus ⟶ Infection; Vaccine ⟶ Infection
  - Note that the edges can have positive or negative effect

An Example

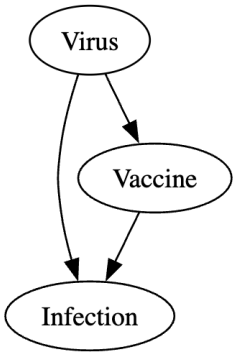
- **Nodes (variables)**: Virus, Vaccine, Infection

```
import graphviz as gr
```

```
#initial graph
g1 = gr.Digraph()

#add edges between nodes
g1.edge("Virus", "Infection")
g1.edge("Vaccine", "Infection")

g1
```




```
# add another obvious edge
g1.edge("Virus", "Vaccine")

g1
```

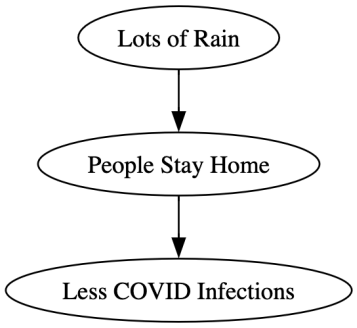


Common Structures of DAGs: Single path with intermediary

A causes B, B causes C, but A does not directly cause C

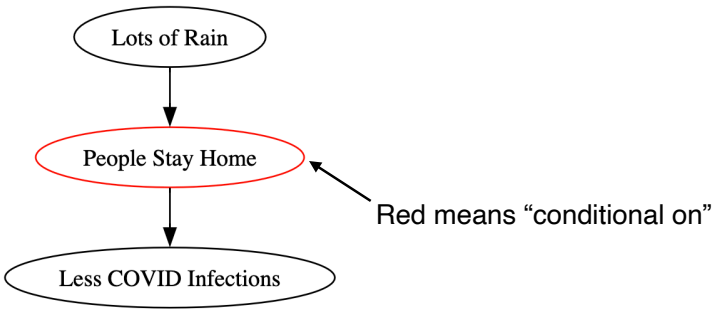
- I did some research on COVID and its impacts on customer spending
  - When a place rained a lot for several weeks, less COVID cases
- But the causal relationship is only *indirect*.
  - The rain only worked because people stayed home more time
  - The *stay-in-home* lowered COVID infections.

```
g = gr.Digraph()
g.edge("Lots of Rain", "People Stay Home")
g.edge("People Stay Home", "Less COVID Infections")
g
```



If we condition on the intermediatory—"block the dependence"

```
g.node("People Stay Home", color="red")
g
```



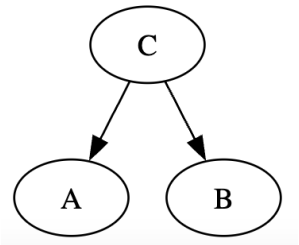
- Conditional on (aka. controlling) the level of how much "people stay home," amount of rain provide no additional info about COVID infections:

$$\text{COVID\_Infections} \not\perp \text{Rain},$$
$$\text{COVID\_Infections} \perp \text{Rain} \mid \text{People\_Stay\_Home}$$

Common Structures of DAGs: Fork Structure

The same variable causes two other variables

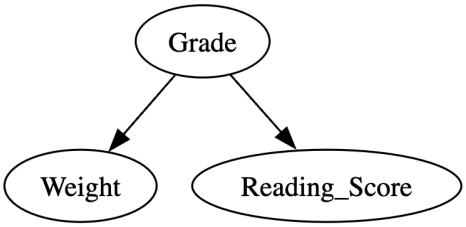
```
g = gr.Digraph()
g.edge("C", "A")
g.edge("C", "B")
g
```



- This structure is one of the most dangerous in causal inference, due to a "**backdoor path**"
  - *Backdoor path* refers to association through an alternative route instead of a direct edge.
  - A and B have no relationship, but there is a dependence (correlation) between A and B because C causes both.

A more concrete and familiar example

Heavier kids have higher reading scores.
Body weights improve reading??

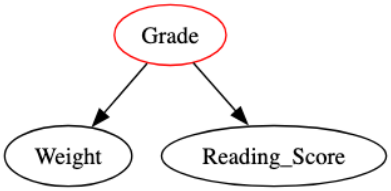


- Grade is the common factor affecting both weights and reading scores
- There is a "backdoor path" between weights and reading scores
- Conditional on (controling) grade will block the backdoor path

$$Weights \not\perp Reading$$
$$Weights \perp Reading | Grade$$

A more concrete and familiar example

Heavier kids have higher reading scores.
Body weights improve reading??



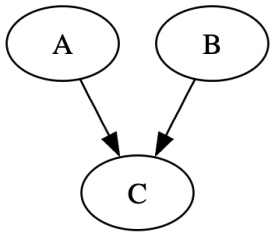
- Grade is the common factor affecting both weights and reading scores
- There is a "backdoor path" between weights and reading scores
- Conditional on (controlling) grade will block the backdoor path

$$Weights \not\perp Reading$$
$$Weights \perp Reading | Grade$$

Common Structures of DAGs: Collider Structure

Two variables causes the same variables

```
g = gr.Digraph()
g.edge("A", "C")
g.edge("B", "C")
g
```



- This structure is called collider, another dangerous structure leading to a backdoor path
  - A and B have no causal relationship; but there is a dependence (correlation) between A and B because they both cause C.

An example of a collider

- Taking "Customer Analytics" and social skills are *independent*
- If we condition on "job opportunities" during analysis, we may see a spurious association between "Customer Analytics" and social skills.
  - People who are hired are more likely to have taken this course
  - People who are hired are more likely to have good social skills

$$Cust\_Analytics \perp Social\_Skills$$
$$Cust\_Analytics \not\perp Social\_Skills | Hired = 1$$

Use DAG for Bias Diagnosis

$$E[Y|T=1] - E[Y|T=0] = \underbrace{E[Y_1 - Y_0|T=1]}_{ATT} + \underbrace{E[Y_0|T=1] - E[Y_0|T=0]}_{BIAS}$$

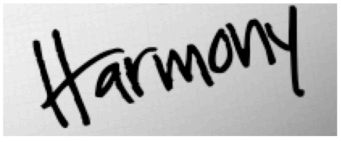
Summary of Different Types of Biases

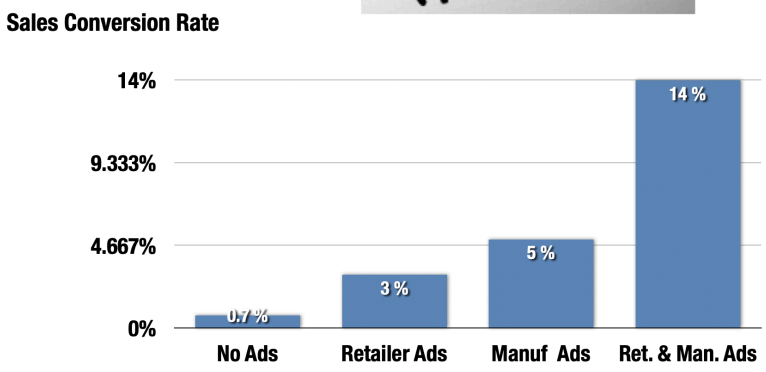
Type of Bias	Definition	Example
Confounding Bias	Arises when a third variable (Z) influences both the treatment (X) and the outcome (Y), creating a spurious association.	Effect of ads (X) on car purchase (Y): Purchase interest (Z) affects both X and Y.
Selection Bias	Arises from using a non-representative sample	Effect of loyalty program membership (X) on spending (Y): Sample includes only urban customers (Z).
Collider Bias	Occurs by conditioning on a variable caused by both treatment and outcome, creating a spurious association between the two.	Effect of ads (X) on product quality (Y): Conditioning on sales (Z), influenced by both X and Y.
Overadjustment Bias	Arises from controlling for a mediator, blocking part of the causal effect.	Effect of course (X) on salary (Y): Controlling for "FAANG" jobs (Z), which is a mediator.

Confounding Bias: The most common bias we encounter in Data Science

Definition: A common factor (confounder) affecting both treatment and outcome

- Recall the Harmony example:

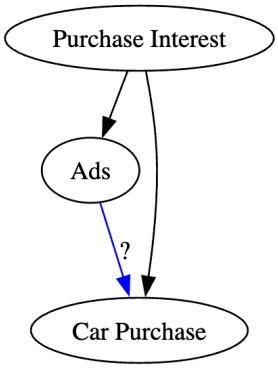




In this setting, confounding bias may overstate the effectiveness of ads

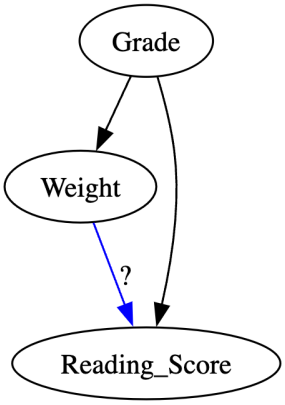
- Treatment: auto ads exposure
- Outcome: auto purchase
- Common factor (confounder): customers' interests in car buying

```
g = gr.Digraph()
g.edge("Ads", "Car Purchase", label="?", color="blue")
g.edge("Purchase Interest", "Car Purchase")
g.edge("Purchase Interest", "Ads")
g
```

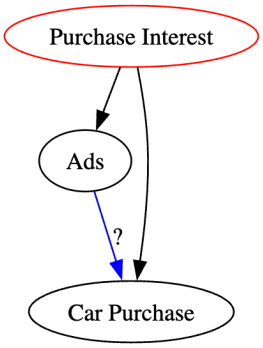


Another old example, where confounding bias almost completely messes up the causality

- Treatment: Weight of a student
- Outcome: Reading score
- Common factor (confounder): The student's age/grade



Fork Structure at the Top Is the Culprit of Confounding Bias!

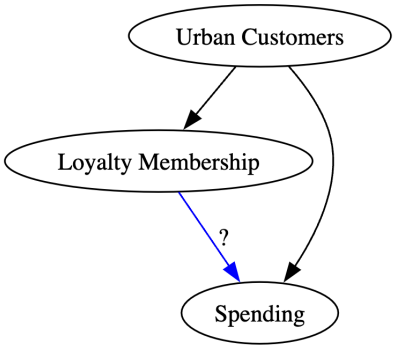


- We are interested in the blue edge with the question mark
- But because of the fork structure at the top, we have a backdoor path
  - Even if **Ads** and **Car Purchase** have little or no effect, they may appear to be associated as they are both affected by **Interest**
  - The effect of Ads on Purchase may be inflated due to the backdoor
  - Solution? Conditional on (control) **Purchase Interest** to block the backdoor.
    - Interest may be difficult to directly measure/observe. But we can control its proxies (e.g., income, age, education, and so on).

(Sample) Selection Bias: The sample used in the study is not representative of the population

Does Hilton's Loyalty Program increase travelers' spending for American customers?

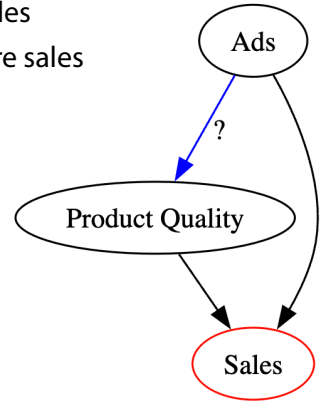
- Treatment: Loyalty program participation
- Outcome: Spending
- Sample selection: Only use US urban customers (or the sample disproportionally contains more urban customers)
  - In this situation, urban customers may have different membership sign-up and spending patterns from the overall population, thus causing bias.



Collider Bias: Over-control—Conditioning on a variable influenced by both treatment and outcome

Does spending more on ads improve product quality?

- Treatment: Ads spending
- Outcome: Product quality
- Both ads and product quality affect sales. If we control sales while investigating ads' effect on product quality, we may get a collider bias
  - More ads, more sales
  - Better quality, more sales



Overadjustment Bias: Over-control—Conditioning on an intermediary variable in your analysis

Does taking "Customer Analytics" improve your salary?

- Treatment: Taking the Customer Analytics course
- Outcome: Salary
- Intermediary variable that should not have been controlled: Getting a job at FAANG (Meta, Amazon, Apple, Nvidia, Google)
  - Taking the course improves your salary
  - Taking the course improves your chance of getting into FAANG
  - Getting into FAANG improves your salary