



IIC 2433 Minería de Datos

<https://github.com/marcelomendoza/IIC2433>

Cierre de la clase 9 – Causalidad



Causalidad:

¿Qué establece la condición causal de Markov?

¿Qué implica la condición causal de Markov?

Cadenas, forks y colliders:


Dada una cadena $A \rightarrow B \rightarrow C$ ¿Cómo garantizo independencia condicional entre A y C?

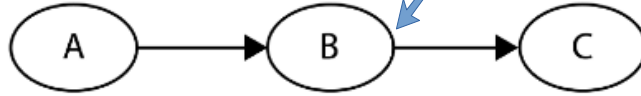
Dado un fork $A \leftarrow B \rightarrow C$ ¿Cómo garantizo independencia condicional entre A y C?

Dado un collider $A \rightarrow B \leftarrow C$ ¿Cómo garantizo independencia condicional entre A y C?

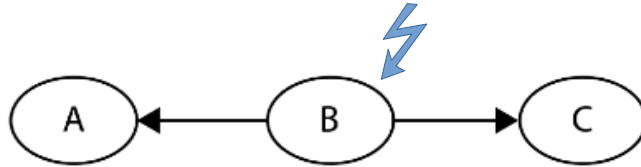
Cierre de la clase 9 – Causalidad


Debo bloquear los caminos entre A y C

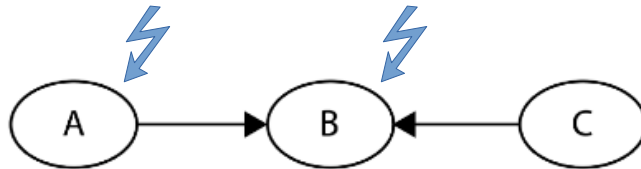
Chain 



Fork 



Collider 



Regresión
sobre C

Bloqueas B

Chain 

	coef	std err	t	P> t	[0.025	0.975]
const	-0.0086	0.007	-1.320	0.187	-0.021	0.004
A	-0.0238	0.033	-0.729	0.466	-0.088	0.040
B	1.0217	0.032	31.645	0.000	0.958	1.085

Fork 

	coef	std err	t	P> t	[0.025	0.975]
const	0.0077	0.006	1.241	0.215	-0.004	0.020
A	0.0090	0.031	0.292	0.770	-0.052	0.070
B	0.9938	0.032	31.372	0.000	0.932	1.056

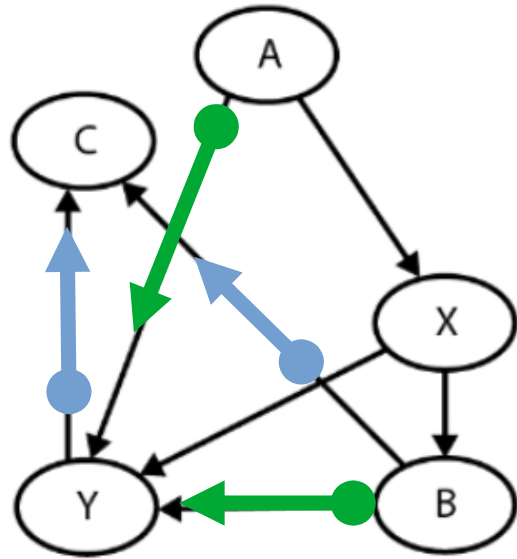
Collider 

	coef	std err	t	P> t	[0.025	0.975]
const	2.082e-17	1.35e-17	1.536	0.125	5.77e-18	4.74e-17
A	-1.0000	2.01e-17	-4.97e+16	0.000	-1.000	-1.000
B	1.0000	1.39e-17	7.19e+16	0.000	1.000	1.000

Aquí hay dependencia del efecto B
pero también de la otra causa del
efecto B (A) - UC - M. Mendoza -

Cierre de la clase 9 – Actividad formativa

¿Qué ocurre al hacer regresión sobre B?



```
NOISE_LEVEL = .2
N_SAMPLES = 1000

# Generate the data
a = np.random.randn(N_SAMPLES)
x = a + NOISE_LEVEL*np.random.randn(N_SAMPLES)
b = x + NOISE_LEVEL*np.random.randn(N_SAMPLES)
y = a + x + b + NOISE_LEVEL*np.random.randn(N_SAMPLES)
c = y + b + NOISE_LEVEL*np.random.randn(N_SAMPLES)
```

A y C son las variables con **mayor influencia** sobre B (más que X!)

	coef	std err	t	P> t	[0.025	0.975]
const	-0.0092	0.004	-2.539	0.011	-0.016	-0.002
A	-0.3377	0.023	-14.536	0.000	-0.383	-0.292
X	-0.0573	0.032	-1.813	0.070	-0.119	0.005
C	0.3370	0.015	22.352	0.000	0.307	0.367
Y	0.0145	0.026	0.565	0.572	-0.036	0.065

En estructuras superpuestas la relación de dependencia condicional es muy compleja.

- Intervenciones -

Intervenir para conocer la causa y el efecto



d -separación



Diremos que dos variables en G están d -separadas si todos los caminos entre ellas están bloqueados.

¿Cuándo se bloquea un camino entre dos variables?

Se bloquea el camino cuando hay un collider en el camino entre ellas y no podemos controlar el consecuente o si hay un fork o cadena que contiene una variable en el medio que podemos controlar.

Variables que podemos controlar: \mathcal{I}

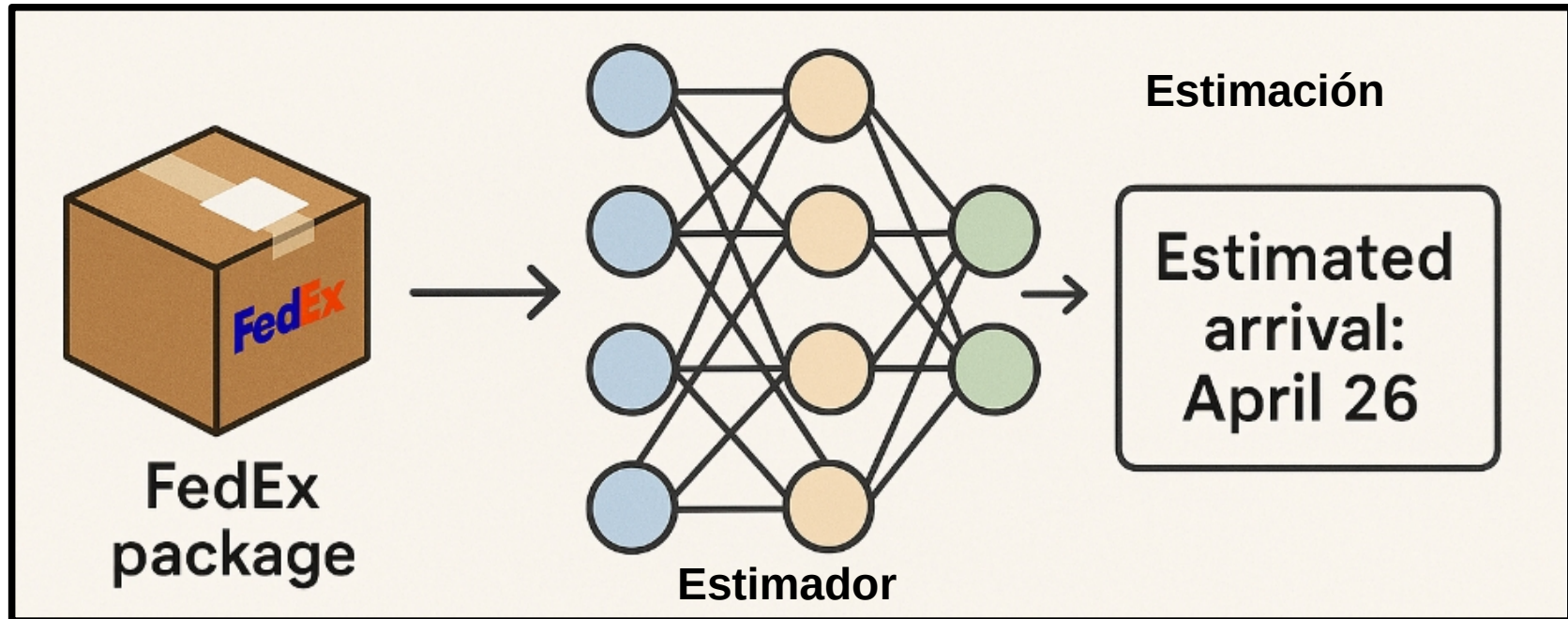
Condiciones para bloqueo entre i y k :

- Existen $i \leftarrow j \rightarrow k$ o bien $i \rightarrow j \rightarrow k$ tal que $j \in \mathcal{I}$
- o bien existe $i \rightarrow j \leftarrow k$ tal que $j \notin \mathcal{I}$

La d -separación nos permite controlar el flujo de información en G .

La d es de direccional.

Estimador, estimación y estimando



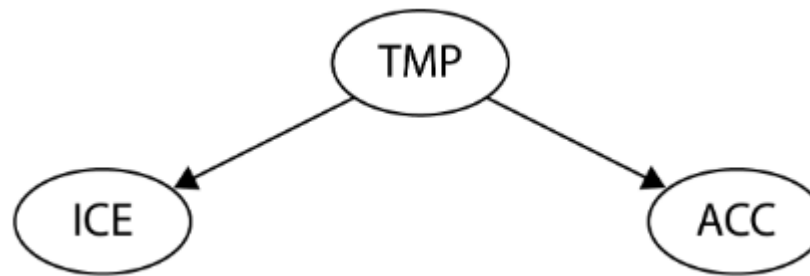
Estimando: La cantidad en la cual estamos interesados.

¿Cuán probable es que llegue el 26 de abril?

Confounding

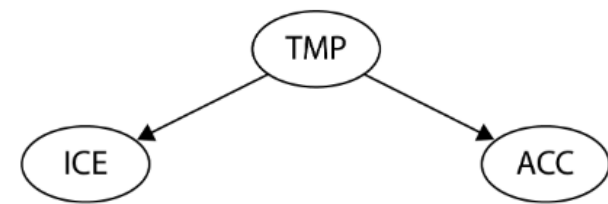
Correlación no es causalidad: la temperatura afecta el consumo de helados y también el aumento de accidentes por ahogamiento. Por tanto, consumo de helados y accidentes por ahogamiento están correlacionados.

Sin embargo, ICE y ACC no tienen relación causal, hay un fork entre ellas.



Decimos que ICE y ACC tienen un **confounding**.

Deconfounding



Queremos dilucidar cual es el efecto causal de ICE sobre ACC.

$$ACC \sim ICE$$

Lo que nos interesa es dilucidar el efecto en ACC si **intervenimos** ICE. Usaremos notación **do** para representar una intervención:

$$P(ACC = acc | do(ICE = ice))$$

Para observar el efecto sobre ACC debemos controlar sobre TMP:

$$P(ACC | do(ICE)) = \sum_{tmp} P(ACC | ICE, TMP) P(TMP)$$

porque TMP es un ascendiente sobre ICE.



La regla del efecto causal

Dado G y un conjunto de variables Pa que son ascendientes de X , el efecto causal de X sobre Y está dado por:

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

En el ejemplo, controlar por TMP bloquea los caminos no causales entre ICE y ACC. Esto nos permite obtener un **estimando** a partir del modelo, para instancias de Y condicionadas a X y Pa .



El **estimando** es el efecto causal promedio en la población.

El criterio **back-door**

Necesitamos una metodología que nos permita obtener **estimandos** en modelos causales.

El criterio **back-door** se basa en bloquear caminos espúreos entre los nodos intervenidos y los nodos de salida. Al mismo tiempo, nos queremos asegurar de no alterar los caminos directos y de no crear nuevos caminos espúreos.

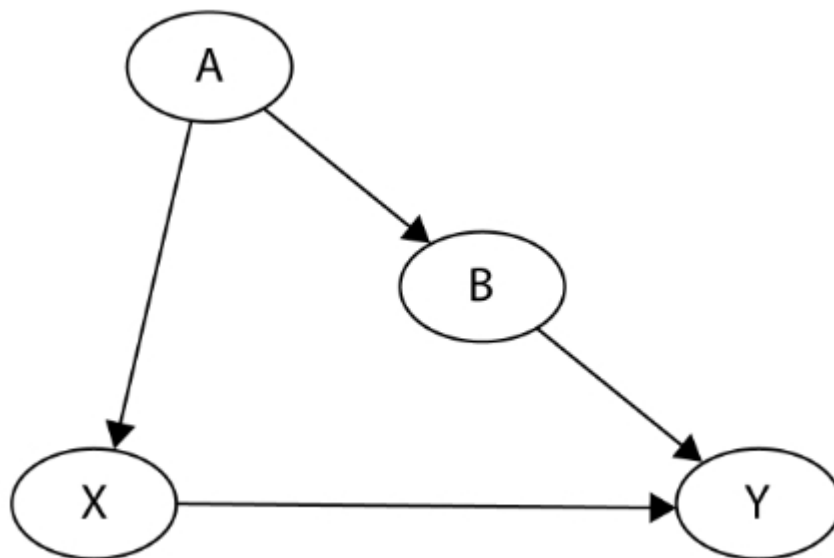
Definición. Dado G , vamos a estudiar la relación causa efecto entre X e Y . Diremos que un conjunto de variables \mathcal{Z} satisface el criterio **back-door** si ningún nodo de \mathcal{Z} es descendiente de X y \mathcal{Z} bloquea todos los caminos entre X e Y que apuntan a X .

En el ejemplo $ACC \sim ICE$, bloqueamos todos los caminos entre ambas variables. En este caso, $\mathcal{Z} = TMP$, ya que TMP no es descendiente de X y TMP bloquea todos los caminos entre ICE y ACC que apuntan hacia ICE . Notar que aquí ICE es la potencial causa.

El criterio **back-door**

... entendiendo el concepto

Sea G :



¿Cuáles nodos debemos controlar para estimar el efecto causal entre desde X a Y ?

Debemos bloquear todos los caminos que llegan a X . Hay tres formas de hacerlo, las cuales son equivalentes:

- i) Controlar A .
- ii) Controlar B .
- lii) Controlar A y B .

El criterio **back-door**

... entendiendo el concepto

Entonces, los estimandos equivalentes, en particular para las opciones i) y ii) son:

$$P(Y = y|do(X = x)) = \sum_a P(Y = y|X = x, A = a)P(A = a) = \sum_b P(Y = y|X = x, B = b)P(B = b)$$

Es decir, basta con observar A ó B para estimar la relación causa – efecto desde X hacia Y . El operador do indica que observamos A (ó B) e intervenimos X (do) y con esto observamos la causalidad en Y .

El criterio **front-door**

No siempre el criterio back-door entregará **estimandos**. Veremos porqué. En estos casos, un criterio complementario, **front-door**, podría entregar estimandos para análisis de relaciones causales.

Vamos a trabajar con un ejemplo que presupone una relación causal entre pérdida de **memoria** espacial y uso de **GPS**. Vamos a suponer que hay un supuesto confounder, **motivación**, esto es, las personas quieren minimizar el esfuerzo en recordar datos innecesarios.

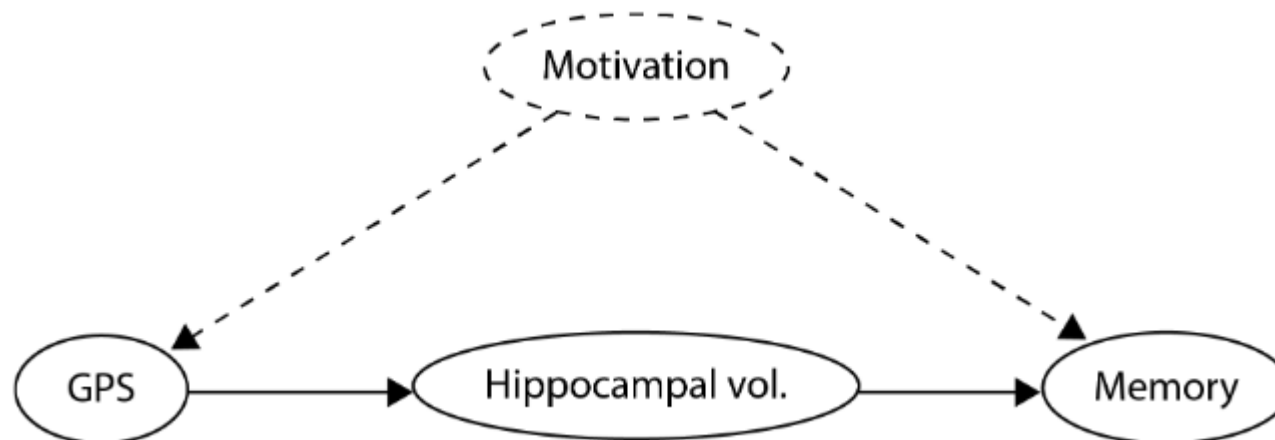
El problema es que no podemos controlar **motivación**, lo que impide aplicar back-door. Para abordar este problema, vamos a usar una variable mediadora entre *GPS* y *memoria*.

Definición. Diremos que la variable Z es **mediadora** entre X e Y cuando existe al menos un camino desde X a Y que pasa por Z . Diremos que Z media de forma **completa** cuando el camino es único.

El criterio **front-door**

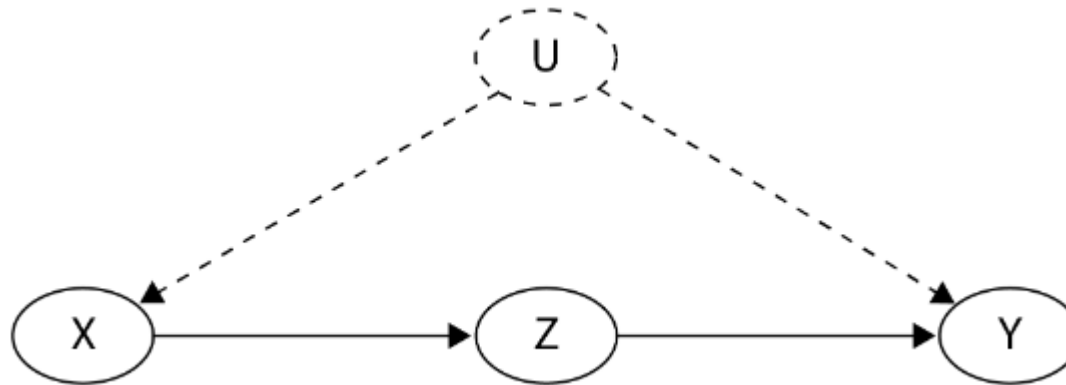
Un estudio mostró que los conductores clase A en Londres tenían un aumento en el **hipocampo**, debido al esfuerzo de entrenamiento que requerían para aprobar el examen de conducción. El hipocampo es el responsable de manejar los recuerdos, en especial, la memoria espacial.

Podemos hipotetizar que *GPS* impacta negativamente el volumen del *hipocampo*, lo que a su vez tiene efectos en la *memoria*. Es decir, hipocampo es un mediador entre GPS y memoria.



El criterio **front-door**

El criterio **front-door** se basa en la estrategia dividir para conquistar. Se analiza el efecto causal entre X y Z , y luego entre Z e Y . Volvamos al ejemplo anterior:



De X a Z , el collider $U \rightarrow Y \leftarrow Z$ bloque el camino hacia X . Luego:

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

De Z a Y , X bloquea el camino hacia Z . Luego:


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

El criterio **front-door**

Luego, combinamos ambos análisis:

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

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


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

Reemplazando tenemos:

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

Esta es la fórmula **front-door**.

El criterio **front-door**

En síntesis, decimos que las variables \mathcal{Z} satisfacen el criterio front-door, dado G , para un par de variables X e Y , si:

- \mathcal{Z} intercepta todos los caminos desde X a Y
- No hay caminos back-door abiertos desde X a
- Todos los caminos back-door desde \mathcal{Z} a Y están bloqueados por X .

Implicancia: Cuando no hay un camino causal entre X e Y , no podemos ajustar un sólo modelo para predecir Y desde X . Lo que debemos hacer es ajustar dos modelos, uno desde X a Z y otro desde $Z + X$ a Y . La combinación de ambos explicará el efecto causal $X \rightarrow Y$.

Hay más técnicas para deconfounding, como las variables instrumentales. Las veremos en la próxima clase.