

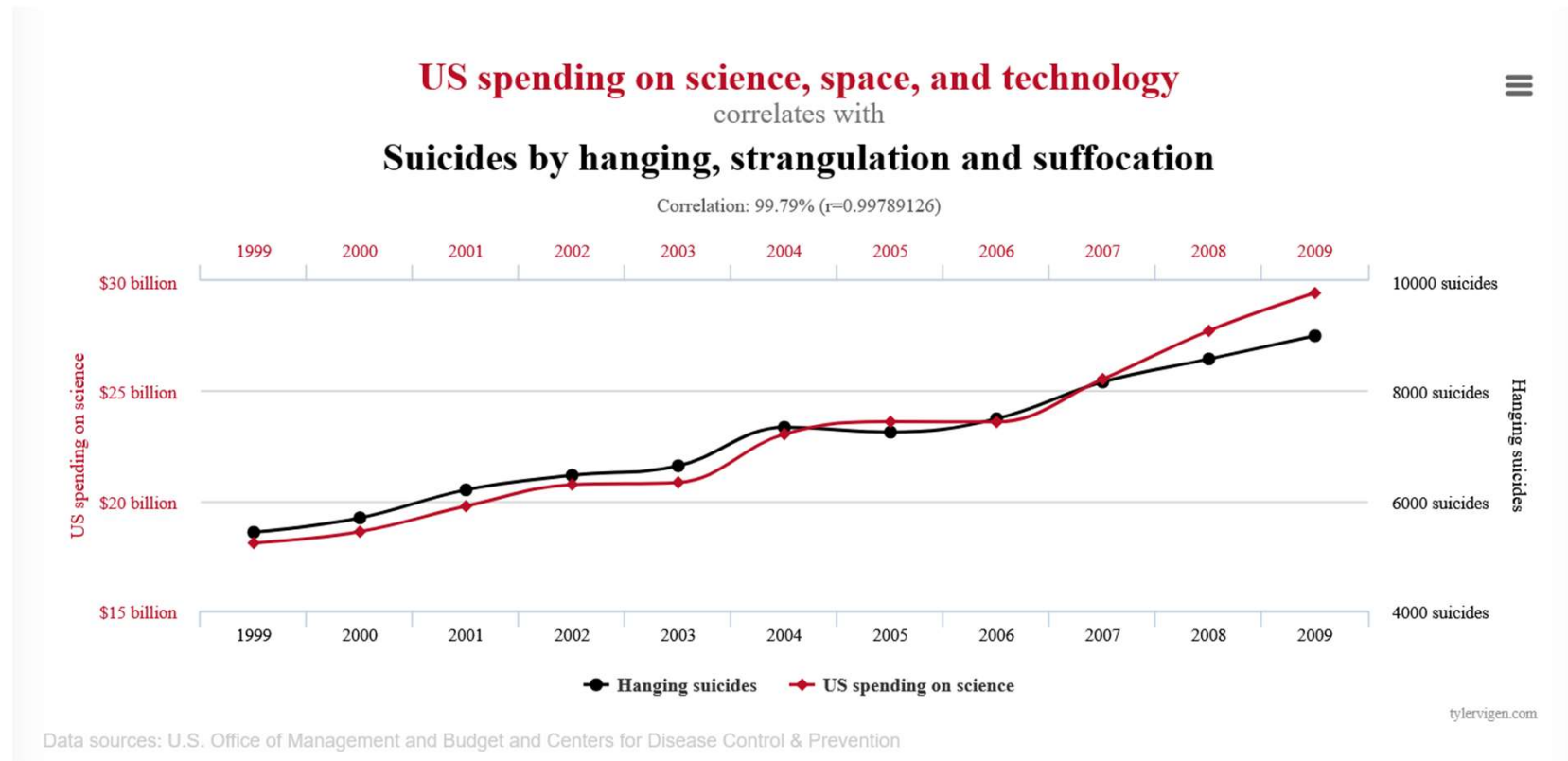


DAG fantastici... ...e come utilizzarli

Metodi Quantitativi per le Biotecnologie
Corso di laurea magistrale in Biotecnologie per le Biorisorse e lo Sviluppo
Ecosostenibile (A.A. 2023-2024)
Università di Verona

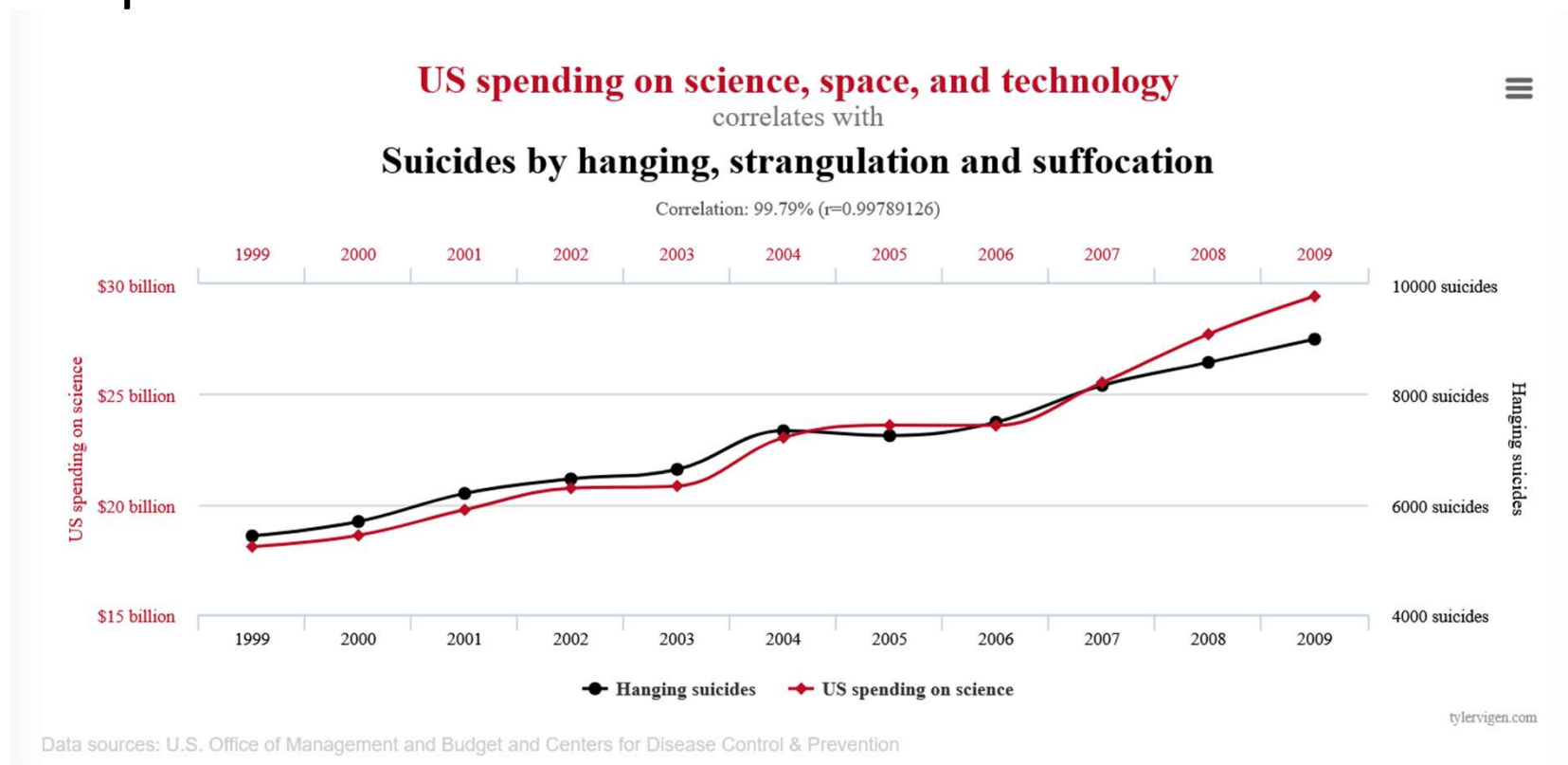
Matteo Migliorini (matteo.migliorini@univr.it)- Assegnista

Problema...



... quindi investire in scienza aumenta i casi di suicidio ???... O viceversa ???...

... Risposta: nessuno dei due



Obiettivi della lezione

- Cogliere la differenza tra inferenza causale (causa-effetto) e statistica
- Cosa sono i DAG e perché sono utili per fare inferenza causale
- Riconoscere le principali strutture nei DAG e i relativi flussi associativi
- Distinguere tra effetti diretti e totali di una variabile; interpretare i coefficienti di regressione
- Applicare queste nozioni utilizzando i pacchetti R ``dagitty`` e ``ggdag``



Inferenza causale vs. statistica

Inferenza Causale vs. statistica

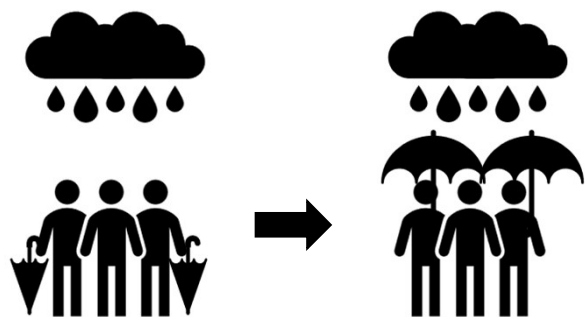
- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$$A \rightarrow B \quad = \quad B \leftarrow A \quad = \quad A \text{ ind. } B$$

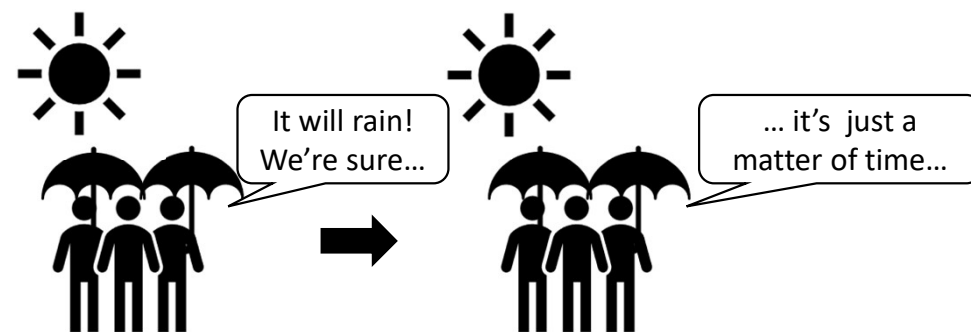
Inferenza Causale (causa-effetto)

- Def: “It is the reasoning to the conclusion that something is, or is likely to be, the cause of something else”
- Studia I rapporti di causa-effetto e le associazioni che scaturiscono da questi
- Rapporto di casusa-effetto: considerando due fenomeni A e B, A è causa di B se modificando A viene a modificarsi B, ma viceversa la modifica di B non influenza A.

Inferenza Causale (causa-effetto) - esempio



La pioggia fa aprire gli
ombrelli alle persone...



... non viceversa...

Inferenza Causale vs. statistica

- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$$A \rightarrow B \quad = \quad A \leftarrow B \quad = \quad A \text{ ind. } B$$

- L'inferenza causale presuppone che una variabile sia responsabile della variazione di un'altra variabile.

$$A := B \quad \neq \quad B := A$$

Inferenza Causale vs. statistica

- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$$\begin{array}{ccccc} A \rightarrow B & = & A \leftarrow B & = & A \text{ ind. } B \\ \text{A causa B} & & \text{B causa A} & & \text{A e B sono indipendenti (associazione spuria)} \end{array}$$

- L'inferenza causale presuppone che una variabile sia responsabile della variazione di un'altra variabile.

$$A := B \quad \neq \quad B := A$$

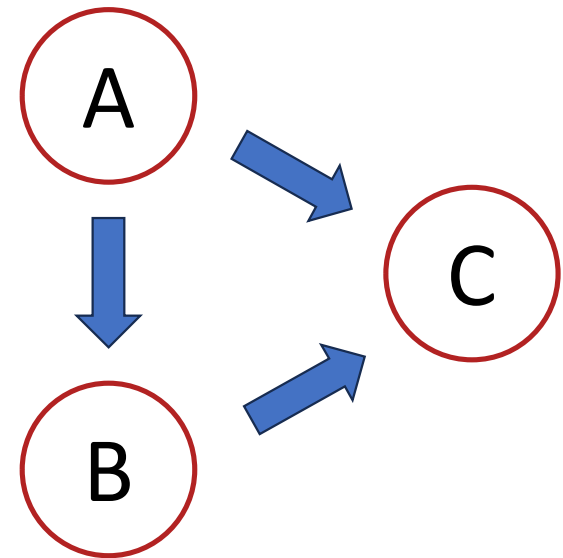
A è causato da una qualche funzione di B



Cosa sono i DAG?

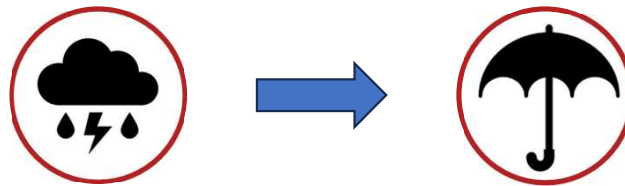
Grafi Aciclici Diretti – DAG

- Grafo: struttura orientata definita da vertici (cerchi) e archi (frecce)
- Diretto (o orientato): gli archi hanno una direzionalità
- Aciclico: selezionando un qualsiasi vertice, non è possibile ritornarvi seguendo l'orientazione degli archi




Grafi Aciclici Diretti – DAG

- Riprendendo l'esempio di prima in cui la pioggia causa l'apertura degli ombrelli, il DAG è il seguente:



- In genere il DAG si costruisce sulla base delle conoscenze scientifiche, non sui dati.



In che modo ci aiutano con
l'inferenza causale?

Elemental confounders (EC)

- I DAG ci consentono di vedere come fluisce l'associazione tra variabili e di capire come bloccare le associazioni spurie.

- Prima però dobbiamo capire come l'associazione tra le variabili *fluisce* all'interno del DAG. Ci sono quattro strutture base a tre nodi (elemental confounders) che compongono un DAG:

- Pipe $X \rightarrow Z \rightarrow Y$

- Fork $X \leftarrow Z \rightarrow Y$

- Collider $X \rightarrow Z \leftarrow Y$

- Descendant $Z \rightarrow X \rightarrow Y$

NB: X è la variabile indipendente; Y è la nostra variabile dipendente; Z è un'altra variabile indipendente misurata che possiamo o meno inserire all'interno della regressione.

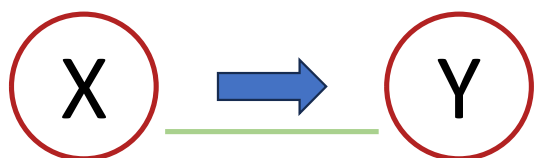
Elemental confounders (EC)

- Per comprendere il flusso associativo è utile ragionare in termini di regressione lineare multivariabile.

$$Y = \alpha + \beta_k x_k$$

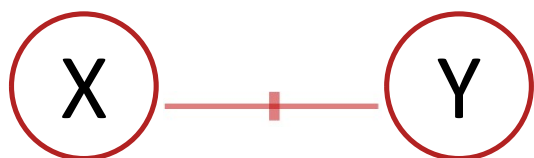
- Spiegando Y in funzione di X e aggiungendo o meno una terza variabile di controllo Z scopriremo che i flussi associativi possono essere aperti o bloccati in base alla struttura che collega i tre nodi
- Possiamo spiegare i flussi associativi considerando la significatività dei coefficienti di X e di Z

EC – base con due nodi



- X causa Y
- coefficiente di X significativo

$Y \not\perp\!\!\!\perp X$ [x and y are associated]



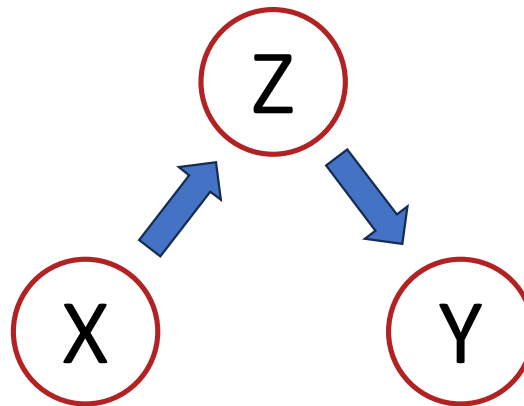
- Non ci sono rapporti di causa-effetto tra X e Y
- Coefficiente di X non significativo

$Y \perp\!\!\!\perp X$ [x and y are not associated]

*a eccezione di associazioni spurie (improbabili, ma esistono)

EC - Pipe

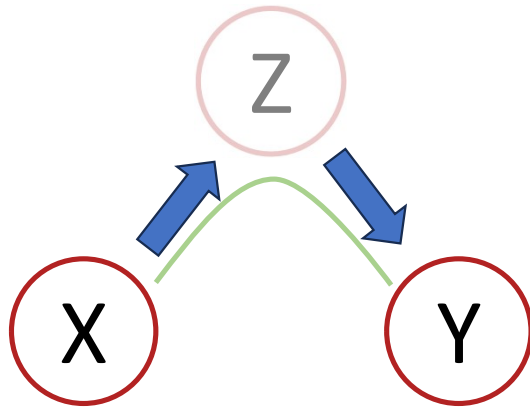
- PIPE: X è una causa di Y, mediata da Z



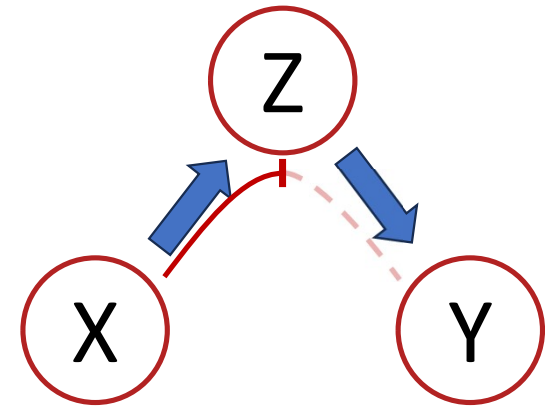
- Per capire come fluisce l'associazione, facciamo due regressioni lineari: nella prima Y è spiegato solo da X, nella seconda è spiegato sia da X che da Z:
 1. $Y \sim X$ → Il coefficiente di X è significativo (associazione aperta)
 2. $Y \sim X + Z$ → Il coefficiente di X non è significativo (associazione chiusa)

EC - Pipe

- Indipendenze condizionali della PIPE:



$Y \sim X$ (non stratifico/condiziono per Z)
Flusso associativo aperto (linea verde)



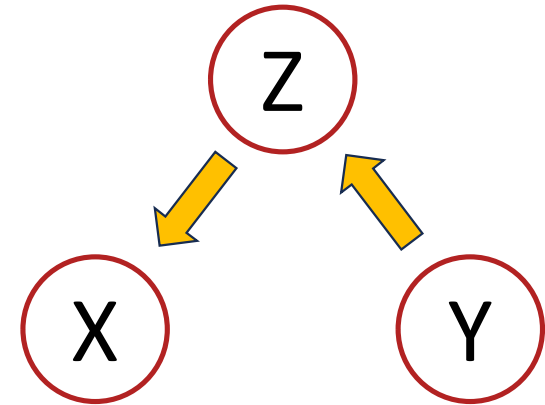
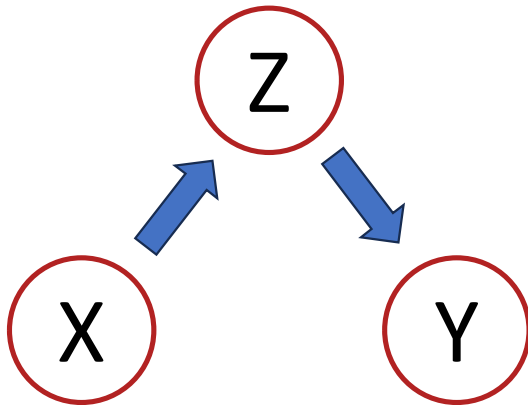
$Y \sim X$ (stratifico/condiziono per Z)
Flusso associativo chiuso (linea rossa)

$Y \not\perp X$ [x and y are associated]

$Y \perp X|Z$ [x and y are not associated, conditional on Z]

EC - Pipe

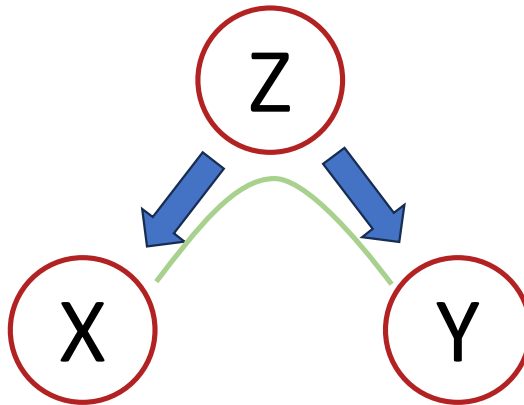
- E se ho pipe con direzionalità opposte?



- **ATTENZIONE:** statisticamente parlando queste due strutture sono identiche!
- Le indipendenze condizionali sono le medesime: vi è associazione tra le variabili X e Y a meno che non stratifichi per la variabile Z

EC - Fork

- FORK: Z è una causa comune sia di X che di Y

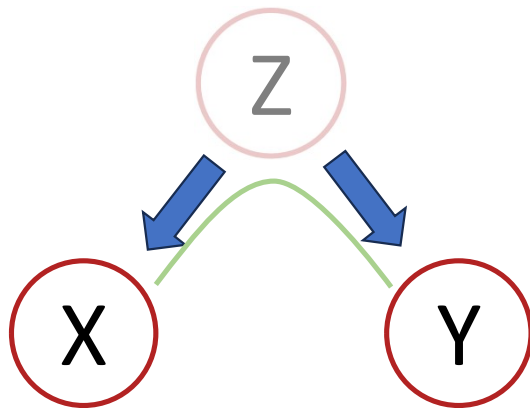


$Y \not\perp\!\!\!\perp X$ [x and y are associated]

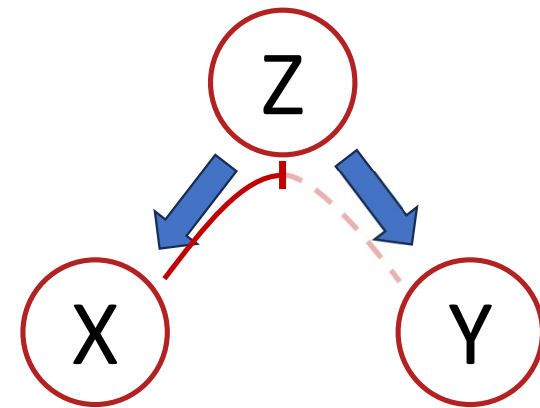
$Y \perp\!\!\!\perp X|Z$ [x and y are not associated, conditional on Z]

EC - Fork

- Indipendenze condizionali della fork:



$Y \sim X$ (non stratifico/condiziono per Z)
Flusso associativo aperto (linea verde)



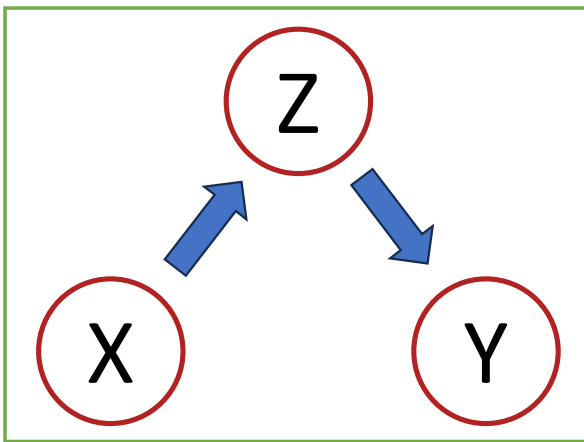
$Y \sim X + Z$ (stratifico/condiziono per Z)
Flusso associativo chiuso (linea rossa)

$Y \not\perp\!\!\!\perp X$ [x and y are associated]

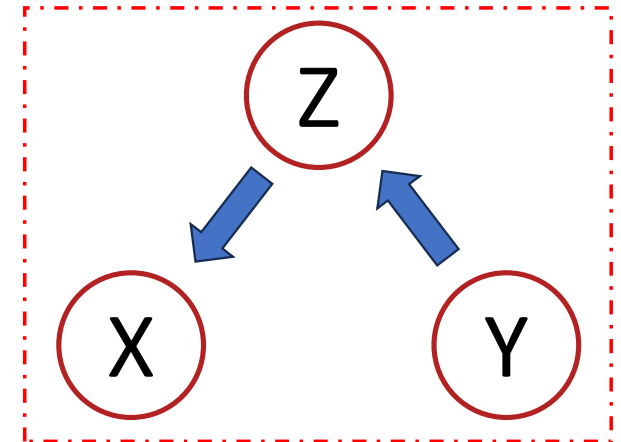
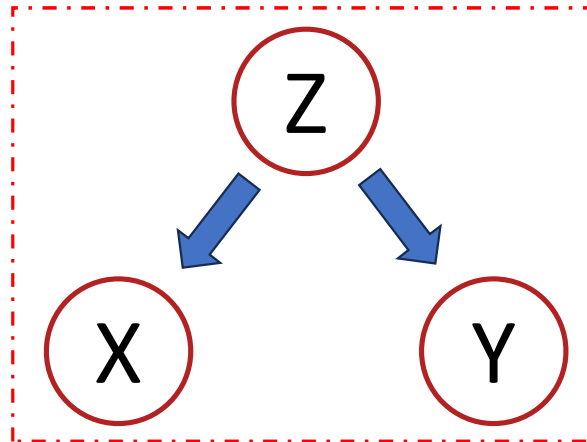
$Y \perp\!\!\!\perp X | Z$ [x and y are not associated, conditional on Z]

EC – Forks & Pipes

- FORK e PIPE hanno le stesse indipendenze condizionali !!!
- Senza conoscenze pregresse sono pertanto indistinguibili !!!



Unico caso in cui Y è una conseguenza (mediata da Z) di X

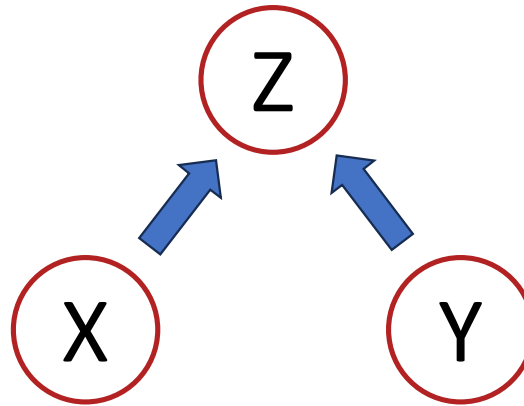


$Y \not\perp\!\!\!\perp X$ [x and y are associated]

$Y \perp\!\!\!\perp X|Z$ [x and y are not associated, conditional on Z]

EC - Collider

- COLLIDER: X e Y sono entrambi cause di Z

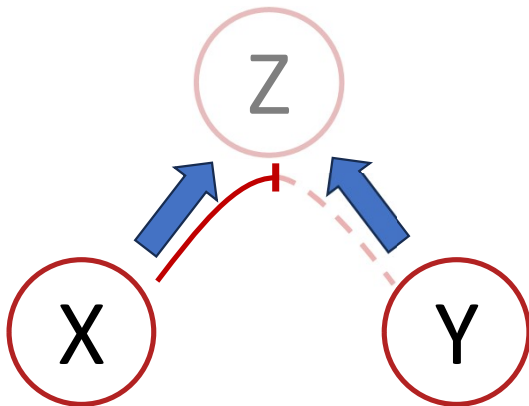


$Y \perp\!\!\!\perp X$ [x and y are not associated]

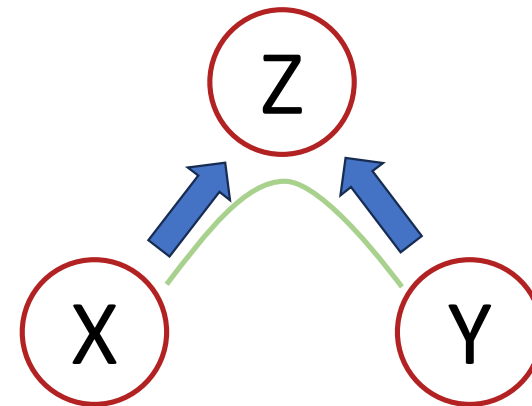
$Y \not\perp\!\!\!\perp X|Z$ [x and y are associated, conditional on Z]

EC - Collider

- COLLIDER: X e Y sono entrambi cause di Z



$Y \sim X$ (non stratifico/condiziono per Z)
Flusso associativo chiuso



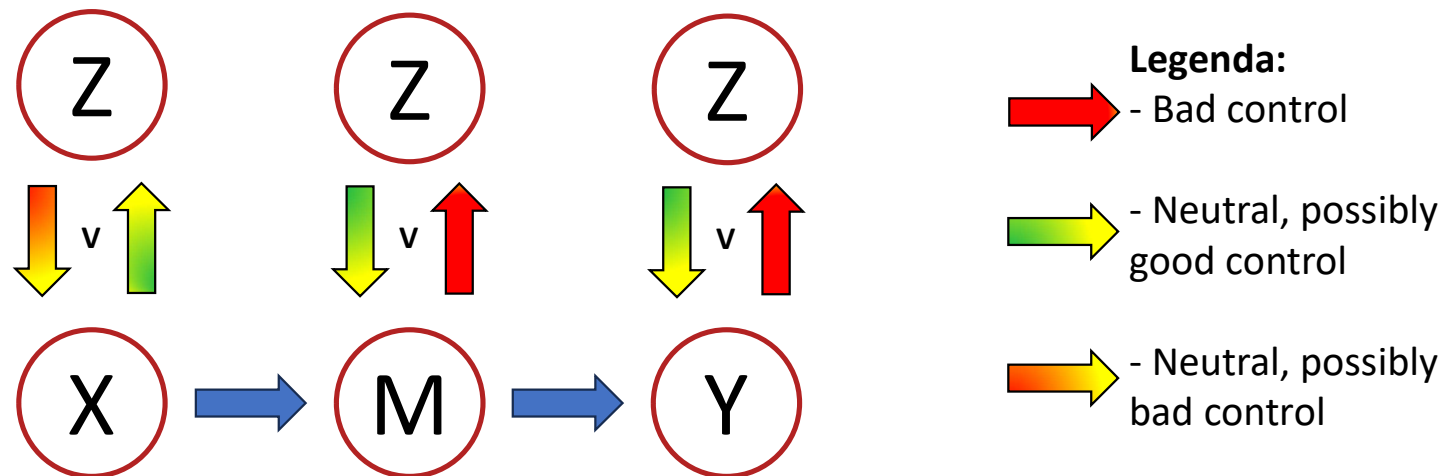
$Y \sim X+Z$ (stratifico/condiziono per Z)
Flusso associativo aperto

$Y \perp\!\!\!\perp X$ [x and y are not associated]

$Y \not\perp\!\!\!\perp X|Z$ [x and y are associated, conditional on Z]

EC: Descendant

- DESCENDANT: quando una variabile è causa o conseguenza diretta di una variabile di interesse.
- In genere il loro comportamento è ambiguo; potrebbero aumentare la precisione, aumentare il bias o essere neutrali.
- Possono essere usati come proxy di una variabile non nota.





Ora proviamo su R



Dal flusso di associazione alla causalità

È importante aggiungere le giuste variabili di controllo

- Abbiamo visto come aggiungere o meno variabili di controllo può aprire o chiudere canali associativi.
- Questo ha un grosso impatto nella stima dei coefficienti di regressione e nella stima degli effetti reali sulla variabile di interesse



Esperimenti randomizzati

- Per capire come una variabile impatti un'altra ci avvaliamo di esperimenti randomizzati. Lo scopo è intervenire su una variabile X e vedere l'effetto che provoca su una variabile di interesse Y .



- Il Do-Calculus è una branca della matematica che consente di fare lo stesso direttamente dai dati. In simboli:

$$P(Y|Do(X))$$

Do-Calculus

- Gli esperimenti randomici consentono randomizzare una variabile in analisi, mantenendo costanti le altre, in modo tale da vedere l'influenza di questa su di una variabile di interesse nella popolazione
- In un certo senso il Do-Calculus randomizza la variabile in analisi usando la matematica.
- Consente quindi di vedere l'influenza che ha una variabile anche se dipendente da altre variabili.

Backdoor path

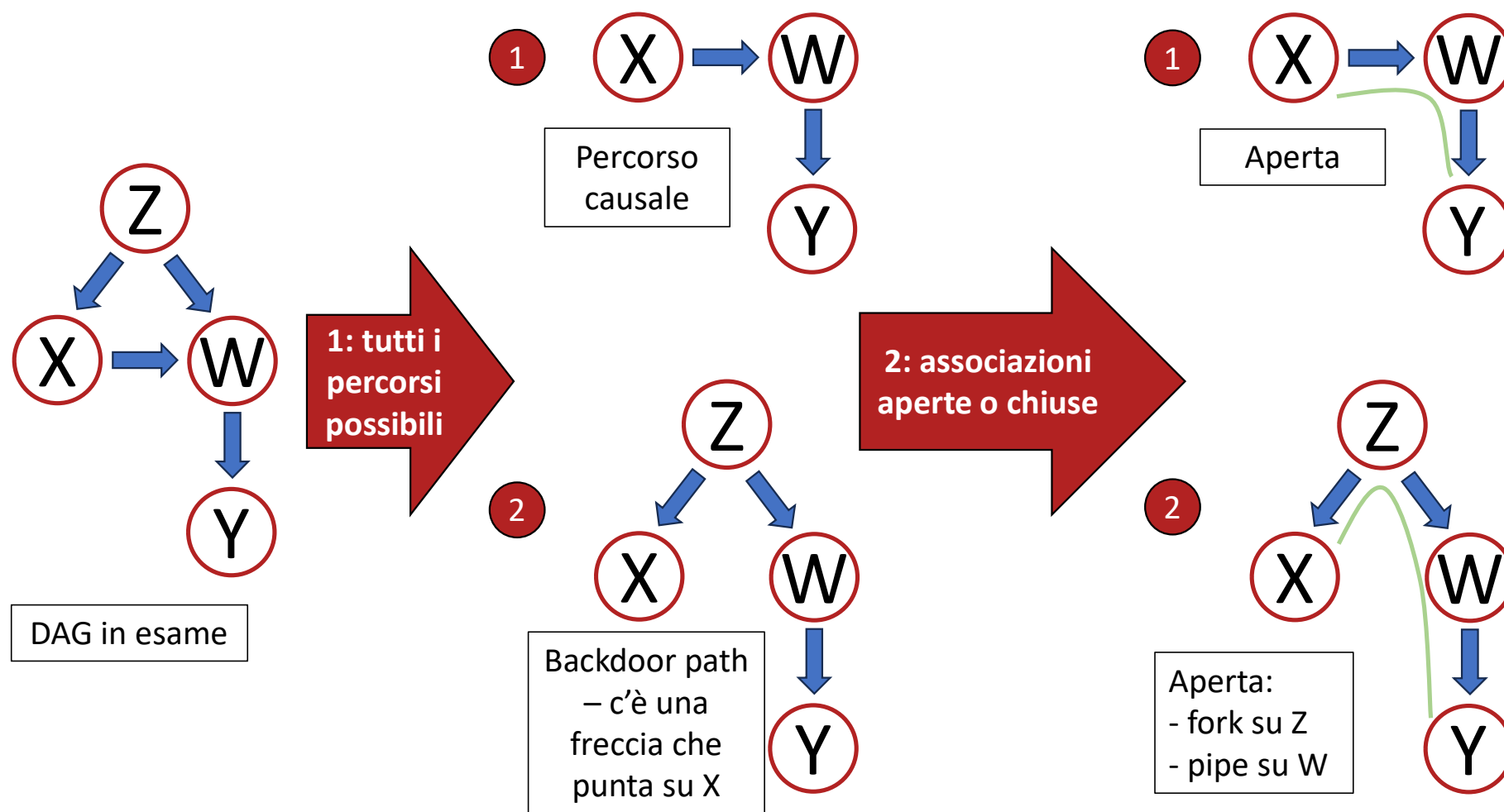
- Uno degli sviluppi del Do-Calculus è il backdoor path. È un metodo intuitivo per rimuovere le influenze delle altre variabili.
- Lo scopo è rimuovere tutte le «frecce» che puntano in X.



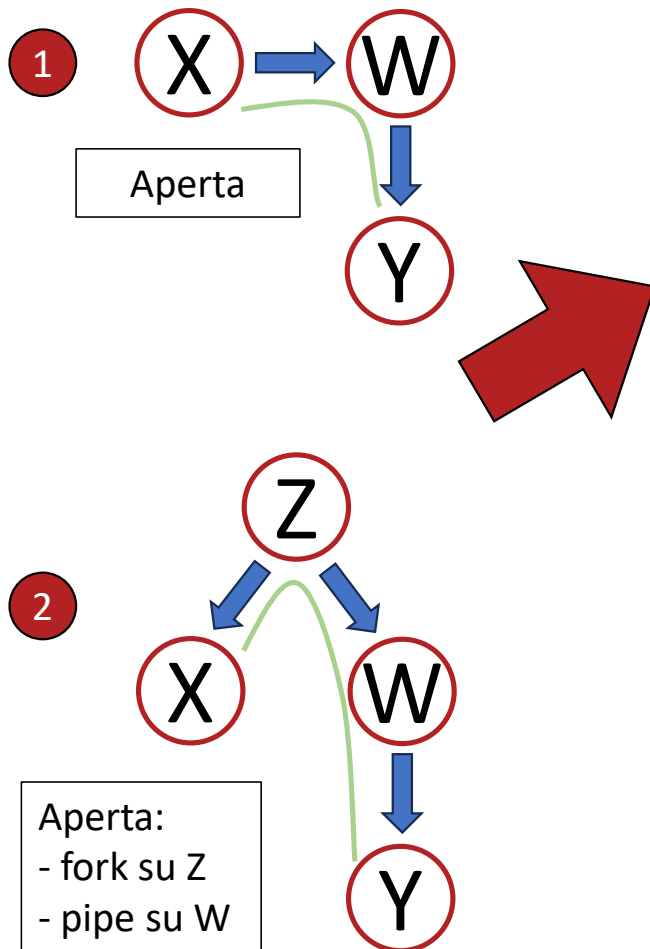
Backdoor path – how to

1. Devo elencare tutti i possibili percorsi con cui X (treatment) si collega ad Y (outcome)
2. Tutti i percorsi che hanno una freccia che punta in X sono backdoor paths (non-causali). Per ognuno di essi valuto se il flusso associativo è aperto oppure no e valuto quali sono le possibili variabili che posso utilizzare per chiuderlo
3. Scelgo quali variabili utilizzare per chiudere tutti i backdoor paths. Il set di variabili selezionate si chiama **adjustment set** (sono possibili più adjustment set).

Backdoor path - esempio



Backdoor path - esempio



- Posso chiudere il backdoor path in due modi:
- Stratificando per W: chiudo entrambi i percorsi
 - Stratificando per Z: chiudo solo il secondo



Considerando che voglio chiudere solo il backdoor path, il mio **adjustment set** è:

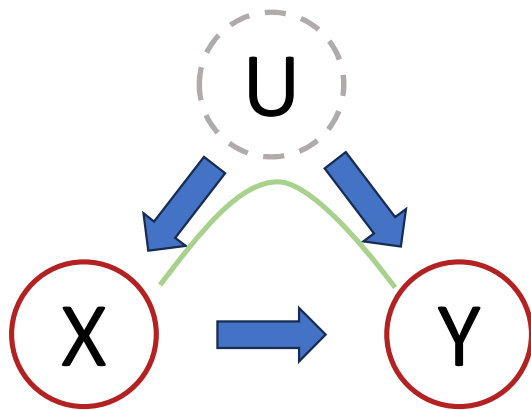
$\{ Z \}$



E se ho variabili non misurate?

Variabili non osservate

- Spesso ci troviamo a lavorare con misurazioni irrealizzabili o dati non disponibili. Questa mancanza di dati può influenzare i risultati dello studio lasciando aperti canali associativi non causali.
- Analizzando il DAG si può capire come risultati possano essere influenzati dalle variabili mancanti
- Alle volte gli effetti causali non possono essere stimati senza bias; saperlo è utile.

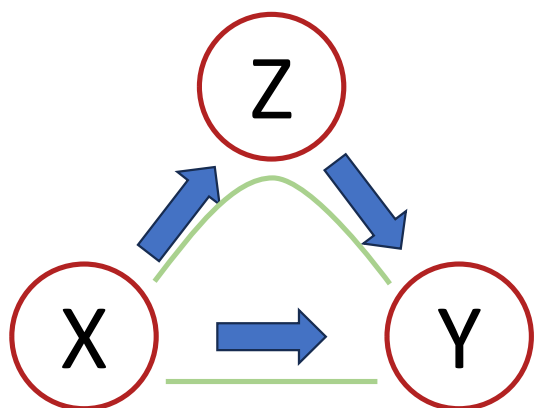


La variabile U apre un flusso associativo tra X e Y; questo genera un bias sulla stima dell'influenza di X su Y; non potendo includere U nella regressione (non ci è noto il suo valore) non possiamo stimare correttamente l'effetto di X su Y.

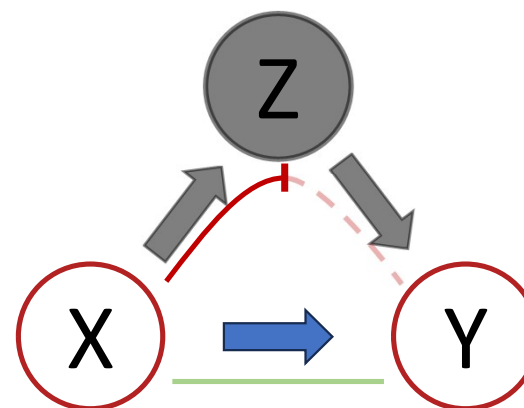
Effetti totali e diretti

- Ok, ora sappiamo capire dal DAG cosa comprende e cosa no se vogliamo una misura corretta... Ma una misura di cosa?
- In genere potremmo essere interessati a 2 particolari effetti della nostra variabile di interesse:

Effetti totali



Effetti diretti



Effetti totali e diretti – table 2 fallacy

- Ma quindi se in base alle variabili che inserisco i coefficienti cambiano, hanno tutti lo stesso significato e la stessa importanza?
- Risposta: NO. Hanno significati diversi in base alle altre variabili utilizzate. Alcuni possono avere anche nessun significato; il loro unico scopo è bloccare le associazioni non causali.
- Quindi anche nei paper attenzione alla «Table 2 Fallacy»

Effetti totali e diretti – table 2 fallacy

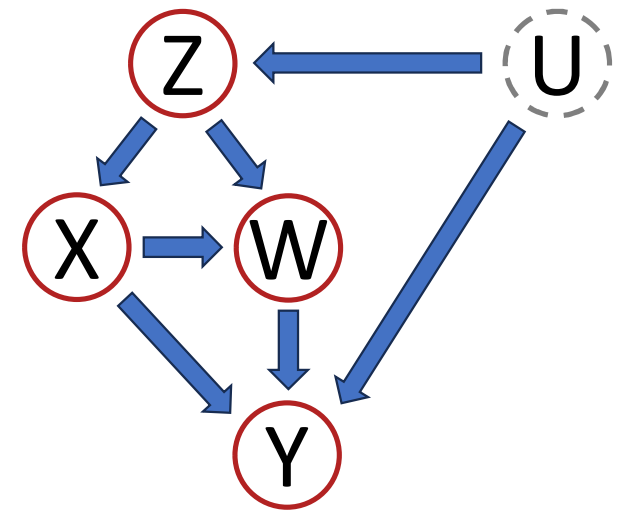
- È quindi importante capire cosa sto facendo e cosa sto misurando con la mia regressione!
- A maggior ragione è importante comunicare efficacemente e in maniera trasparente i risultati.
- Usando un semplice DAG per rappresentare il nostro modello causale viene data a tutti la possibilità di comprendere le nostre analisi. Inoltre, le rende criticabili (W il progresso scientifico)!

Come costruire un DAG

Un modo efficace per costruire un DAG è farlo in maniera progressiva.

In ordine si definiscono:

1. Trattamento (X) e variabile di interesse (Y)
2. Altre cause che influenzano Y
3. Altri effetti che influenzano le altre variabili o connessioni tra le variabili inserite
4. Altre cause o effetti non osservati/osservabili: in questo caso è utile ragionare a cause comuni di coppie di variabili



Cosa abbiamo imparato?

- Associazione \neq causalità
- Cos'è un DAG, come costruirlo e come utilizzare R per:
 - Visualizzarli e identificare i pattern principali
 - Testare se è possibile determinare gli effetti di una variabile sulla variabile di interesse
 - Trovare gli adjustment-set
- Dato il DAG è possibile analizzare con trasparenza i dati, comunicando a tutti in maniera chiara la nostra struttura causale.

Conclusioni – take home messages

- È raccomandabile fare scienza prima di statistica (think before you regress) – associazione non vuol dire causalità e i coefficienti senza buone variabili di controllo potrebbero essere biased.
- I DAG sono uno strumento efficace per descrivere modelli e ipotizzare i flussi di associazione tra le variabili
- È raccomandabile esporre chiaramente i modelli utilizzati, di modo che questi possano essere criticabili per il progresso scientifico

Bibliografia e materiali

- [A crash course in good and bad controls, article](#)
- [Table 2 fallacy, article](#)
- Videolezioni*
 - [Videolezione Elemental Confounders](#)
 - [Videolezione Backdoork Criterion e table2F](#)
- [ggdag R functions, blog](#)
- [Collider bias e covid, article](#)
- [Smoke and covid, article](#)

*fanno parte del corso «Statistical Rethinking» del Prof. Mc Elreath. Le lezioni utilizzano metodi di regressione bayesiani, ma i concetti sui DAG sono gli stessi