Causalité 1 – Introduction à la théorie de la causalité selon Judea Pearl

Séminaire TALia du 28/10/2022

1

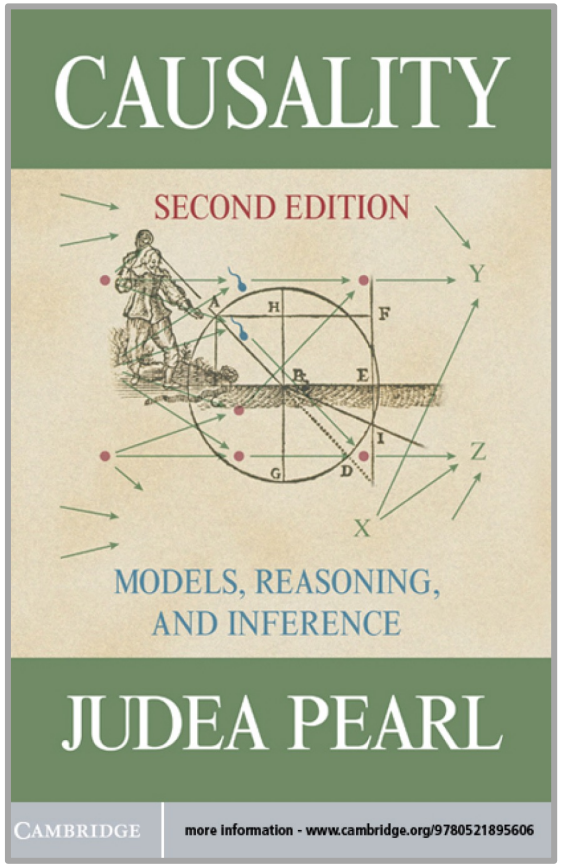
Biometrika (1995), **82**, 4, *pp.* 669–710
Printed in Great Britain

Causal diagrams for empirical research (With Discussions)

BY JUDEA PEARL

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, California 90024, U.S.A.

2



3

Causal AI: Towards Explainable, Generalizable, and Trustworthy Decision-Making

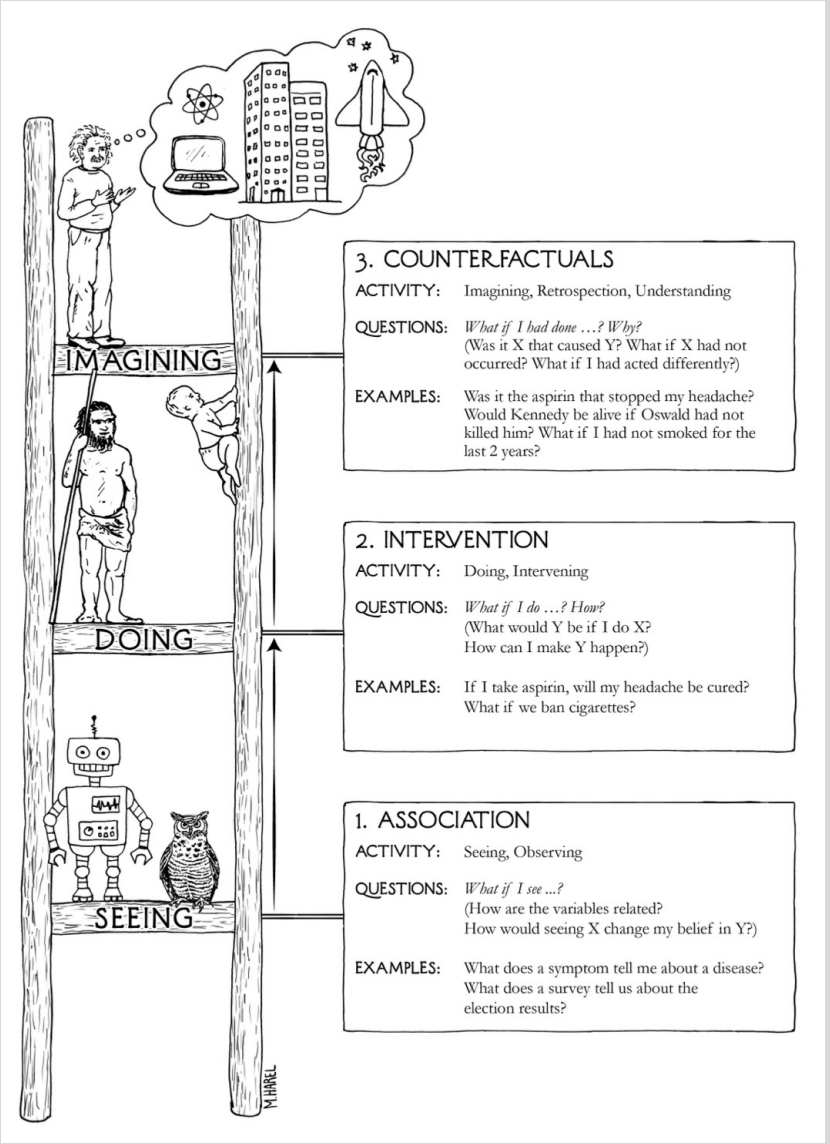
Adèle H. Ribeiro & Elias Barenboim
Causal AI Lab, Columbia University
https://adele.github.io/ | https://causalai.net/

July 29th, 2022

LXMLS

Plan

1. Les promesses de la causalité
2. Les modèles et les graphes causaux
3. Observations versus interventions
4. La notion de d-séparation
5. La fusion entre un modèle causal avec des données observationnelles
6. La notion d' « identifiabilité »
7. Deux critères graphiques
8. Les trois règles du do-calculus



Les promesses de la causalité : une IA robuste et interprétable

- Fournit **un langage précis**, au-delà des probabilités, pour
  - Décrire des **dépendances causales** entre variables aléatoires
  - Formuler la notion d'**intervention** sur un système

- Fournit **une théorie** qui permet de :
  - **Combiner** un savoir à priori causal avec des données d'observation (= « data fusion »)
  - **Identifier** les situations dans lesquelles l'effet d'une intervention peut être prédit à partir d'observations passives
  - **Proposer** des observations complémentaires lorsqu'une situation n'est pas identifiable

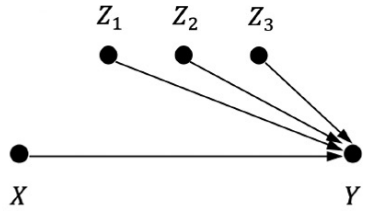
- Accroit les capacités de **généralisation** en transposant des prédictions d'un domaine à un autre

- Améliore l'**interprétabilité** par la compréhension des **mécanismes génératifs** des données

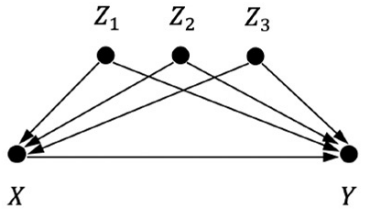
- Améliore l'**équité** des modèles prédictifs en permettant de démêler les mécanismes de discrimination

Exemples de questions auxquelles on souhaite répondre

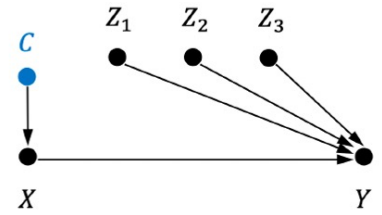
(a)



(b)



(c)

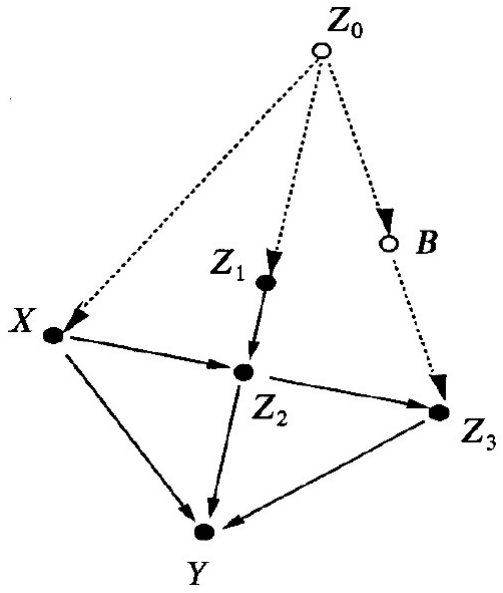


X: treatment is applied
Y: change in health condition

Z_1: age
Z_2: gender
Z_3: is diabetic

} « confounders »

C: random coin flip



- X : quantité de fertilisant
- Z_0 : population de vers l'année précédente
- Z_1 : population de vers avant le traitement
- Z_2 : population de vers après le traitement
- Z_3 : population de vers à la fin de la saison
- B : population d'oiseaux
- Y : rendement de la récolte.

Question : peut on prédire l'effet sur une variable Y d'une **intervention** sur une variable X par simple observation lorsqu'elles sont toutes deux soumises à un réseau complexe d'interactions avec d'autres variables, **observées** ou non **observées** ?

On souhaite faire ces prédictions par **observation passive** du système sans intervenir et sans recourir à un RCT (pour des raisons financières ou éthiques).

Les Structural Causal Models (SCM)

$$\mathscr{M} := \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_{XY}, U_{XZ}, U_{YZ}\} \\ \boldsymbol{\epsilon} = \{\epsilon_X, \epsilon_Y, \epsilon_Z\} \\ \mathscr{F} = \begin{cases} X = f_X(U_{XZ}, \epsilon_X) \\ Y = f_Y(X, U_{ZY}, \epsilon_Y) \\ Z = f_Z(U_{XZ}, U_{ZY}, \epsilon_Z) \end{cases} \\ P_{\mathbf{U}} \\ P_{\epsilon_X} P_{\epsilon_Y} P_{\epsilon_Z} \end{cases}$$

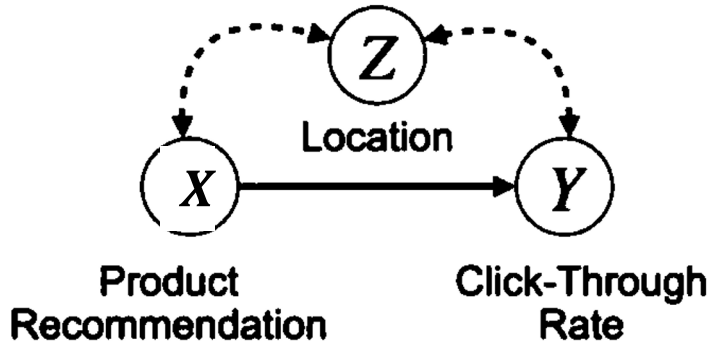
effets

causes

$\mathrm{PA}(Z)$

- Les V_i sont des **variables observées**
- Les U_i sont des **« *confounders* » non-observées** distribués selon une distribution conjointe $P_{\mathbf{U}}$
- Chaque $f_i \in \mathcal{F}$ est une fonction qui représentent les **dépendances causales** d'une variable observée V_i
- Chaque ϵ_i représente un **bruit** spécifique à V_i de distribution P_{ϵ_i}

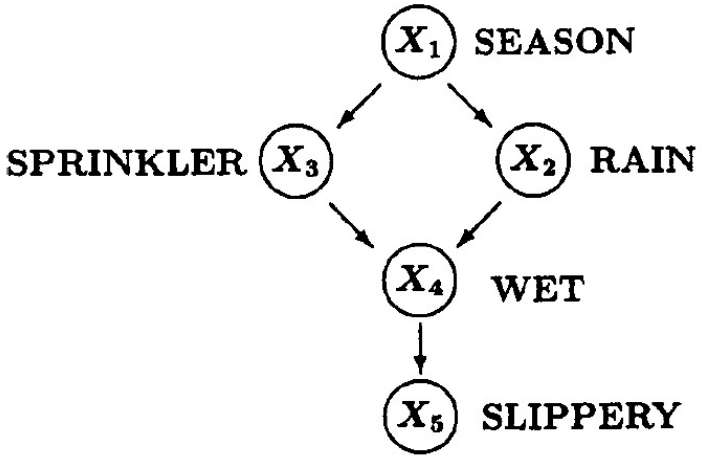
Ici les liens représentent une association **déterministe** !



<u>Définition</u> : un SCM est dit **Markovien** si G est un DAG et si les variables de bruit ϵ_i sont indépendantes.

Chaque **SCM** $\mathcal{M}$ définit un **graphe causal** $G(\mathcal{M})$ et induit une **distribution** de probabilité conjointe sur les variables $\mathbf{V}$ et $\mathbf{U}$ qui est **markovienne** par rapport à ce graphe.

La notion de distribution markovienne par rapport à G



Ici les liens représentent une association **stochastique** !

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$$

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\mathrm{pa}_i)$$

Définition (*Markov Compatibility*) : Si P se factorise selon le DAG G alors on dit que P est Markov p.r. à G.

Théorème (*Parental Markov Condition*) : une distribution est markovienne p.r. à G si chaque variable X_i, lorsqu'elle est conditionnée sur ses parents $\mathrm{PA}(X_i)$, est indépendante de ses **non-descendants** dans G.

Théorème (*Observational Equivalence*) : deux DAG G_1 et G_2 sont équivalents, au sens où toute distribution P Markov p.r. à G_1 l'est aussi p.r. à G_2, si G_1 et G_2, ont le même squelette et les mêmes v-structures.

La notion d'intervention (1)

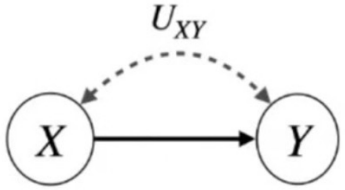
$$\mathscr{M} := \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}\} \\ \boldsymbol{\epsilon} = \{\epsilon_X, \epsilon_Y\} \\ \mathscr{F} = \begin{cases} X = f_X(U_{XZ}, \epsilon_X) \\ Y = f_Y(X, U_{XZ}, \epsilon_Y) \end{cases} \\ P(\mathbf{U}) \\ P(\epsilon_X)P(\epsilon_Y) \end{cases} \quad \xrightarrow{\text{do}(X = x)} \quad \mathscr{M}_x := \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}\} \\ \boldsymbol{\epsilon} = \{\epsilon_X, \epsilon_Y\} \\ \mathscr{F} = \begin{cases} X = x \\ Y = f_Y(x, U_{XZ}, \epsilon_Y) \end{cases} \\ P(\mathbf{U}) \\ P(\epsilon_X)P(\epsilon_Y) \end{cases}$$

Intervenir sur une variable X revient à **amputer** le modèle $\mathcal{M}$ des liens entre X et ses parents $\text{PA}(X)$.

Observation que $X = x$

$$P_{\mathcal{M}}(Y|X = x) \neq P_{\mathcal{M}}(Y)$$

X et Y sont **corrélées**.

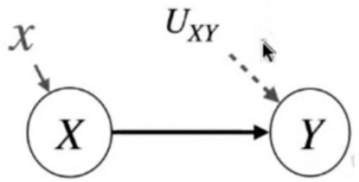


$\neq$

Intervention $\text{do}(X = x)$

$$P_{\mathcal{M}}(Y|\text{do}(X = x)) := P_{\mathcal{M}_x}(Y)$$

X **est une cause** de Y.



La notion d'intervention (2)

$$p(\mathbf{x}|\,\mathrm{do}(x_i = a)) := \begin{cases} \dfrac{p(\mathbf{x})}{p(x_i|\,\mathrm{pa}_i)} = \prod_{j \neq i} p(x_j|\,\mathrm{pa}_j) & \text{if } x_i = a, \\ 0 & \text{if } x_i \neq a. \end{cases}$$

avec $$\frac{p(\mathbf{x})}{p(x_i|\,\mathrm{pa}_i)} = \frac{p(\mathbf{x})}{p(x_i, \mathrm{pa}_i)/p(\mathrm{pa}_i)} = p(\mathbf{x}|x_i, \mathrm{pa}_i)\, p(\mathrm{pa}_i)$$

Intervenir sur une variable X revient à **amputer** le modèle $\mathcal{M}$ des liens entre X et ses parents $\mathrm{PA}(X)$.

La notion d'ajustement pour les causes directes

Distribution **après** intervention

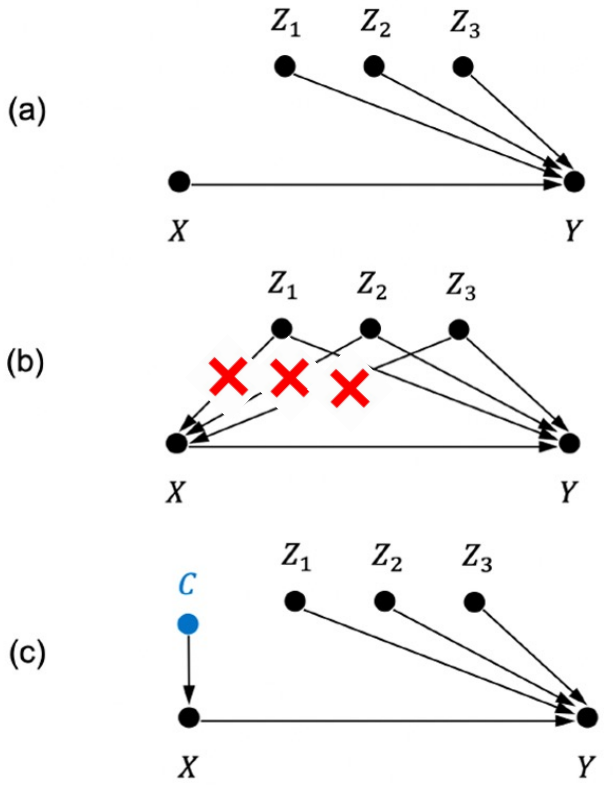
Distributions **avant** intervention

$$p(y|\,\mathrm{do}(x_i = a)) = \sum_{\mathrm{pa}_i} p(\mathrm{pa}_i)\, p(y|x_i = a, \mathrm{pa}_i)$$

Conclusion : Si on on peut mesurer l'**effet** Y, la **cause** X_i et tous ses **parents** PA_i dans G alors on peut **identifier** l'effet causal $P(y|\mathrm{do}(x))$.

Question : Qu'en est-il dans des situations plus complexes où l'on ne peut pas mesurer les parents PA_i des variables instrumentales ?

Relation ente l'opérateur $\mathbf{do}(X = x)$ et RCT



X: treatment is applied

Y: change in health condition

Z_1: age
Z_2: gender
Z_3: is diabetic

} « confounders »

C: random coin flip

$$p(y|\mathrm{do}(X = a)) = \sum_{c \in \{0,1\}} p(c)\, \underbrace{p(y|x = a, c)}_{= p(y|x = a)} = p(y|x = a)$$

Effet de l'administration du traitement

Observation de la corrélation entre le traitement et l'état

Intervenir sur une variable X revient à **amputer** le modèle $\mathcal{M}$ des liens entre X et ses parents $\mathrm{PA}(X)$.

Soumettre l'administration d'un traitement au hasard garantit que l'observation de son effet revient à tester son action non perturbée par des « confounders ».

Question : que faire lorsque RCT est Impossible pour des raisons éthiques ou financières ?

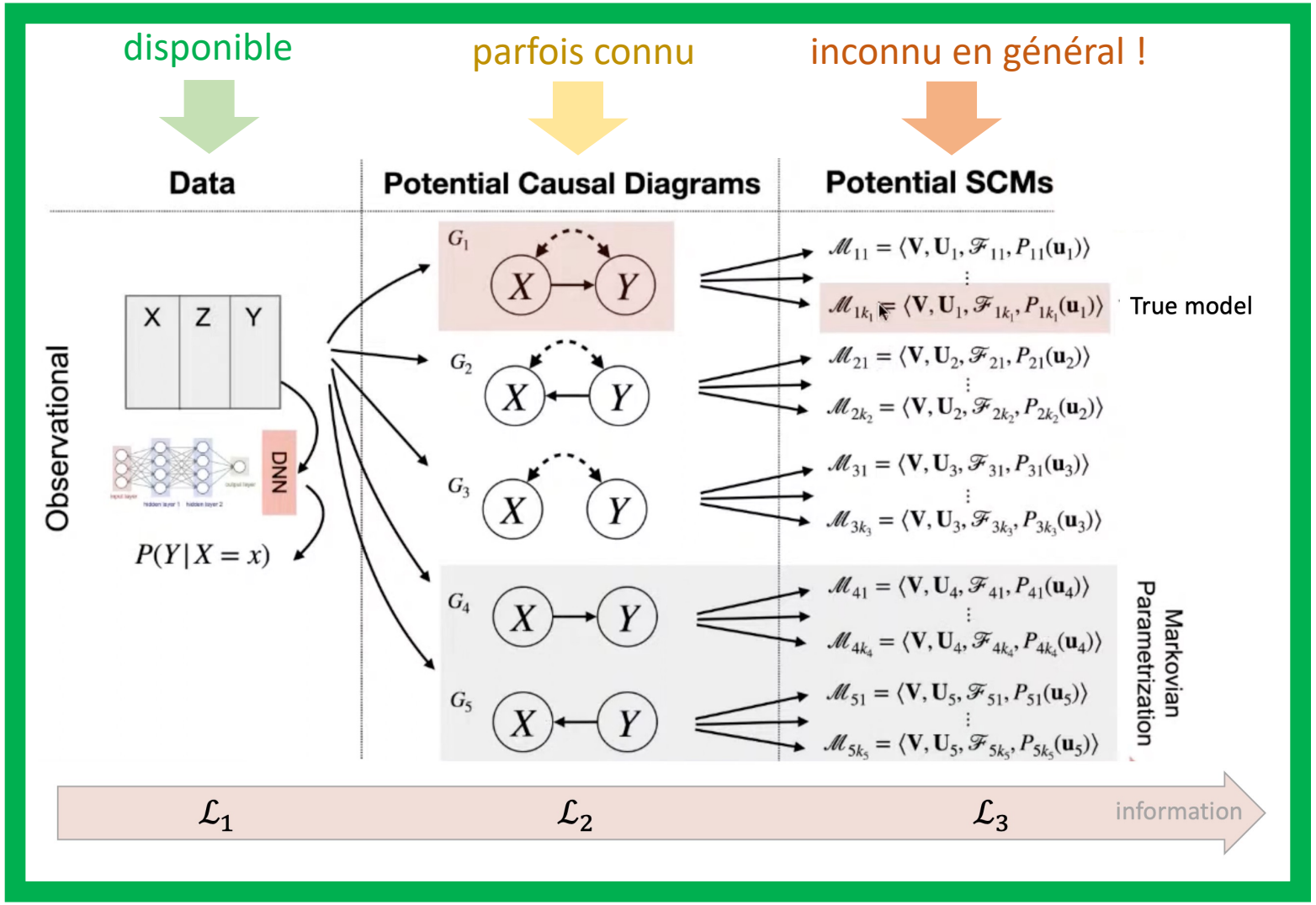
Le concept d'identifiabilité d'un effet causal

Définition : un effet causal $p(Y|\mathrm{do}(x))$ est dit **identifiable** si la connaissance du graphe $G(\mathcal{M})$, associé au modèle causal $\mathcal{M}$ qui génère les données, permet d'exprimer cet effet à l'aide d'une distribution $p(\mathbf{V})$ sur les seules variables observées $\mathbf{V}$.

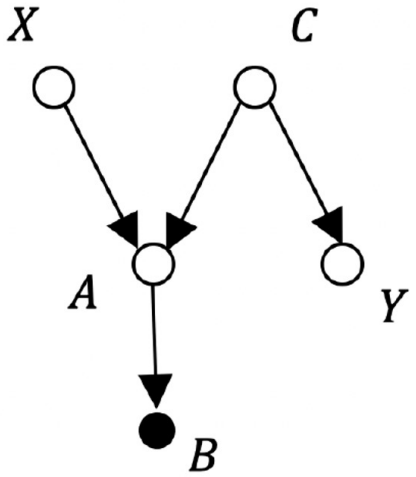
Question : peut on **fusionner** les deux sources d'information que sont :

1. les **données** (= échantillons de $p(\mathbf{V})$)
2. le **graphe causal** $G(\mathcal{M})$ qui spécifie quelles variables déterminent quelles autres.

pour faire des prédictions causales ?



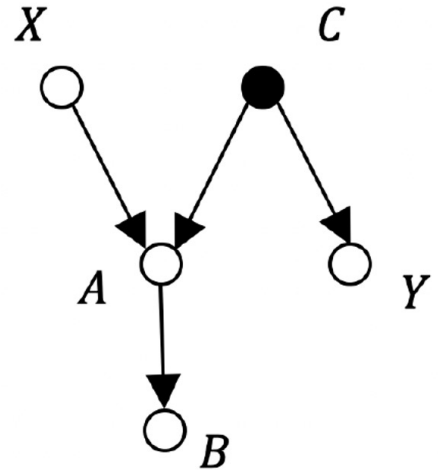
La d-séparation dans un DAG (1)



(a)

$\gamma = (X \rightarrow A \leftarrow C \rightarrow Y)$ n'est pas bloqué par B car B est un descendant du « collider » A.

$(X \not\perp\!\!\!\perp Y|B)_G$



(b)

$\gamma = (X \rightarrow A \leftarrow C \rightarrow Y)$ est bloqué par C car C est une « fork ».

$(X \perp\!\!\!\perp Y|C)_G$

La d-séparation dans un DAG (2)

Définition : Si une distribution P est telle que X est indépendant de Y étant donné $\mathbf{Z}$, on écrit $(X \perp\!\!\!\perp Y|\mathbf{Z})_P$.

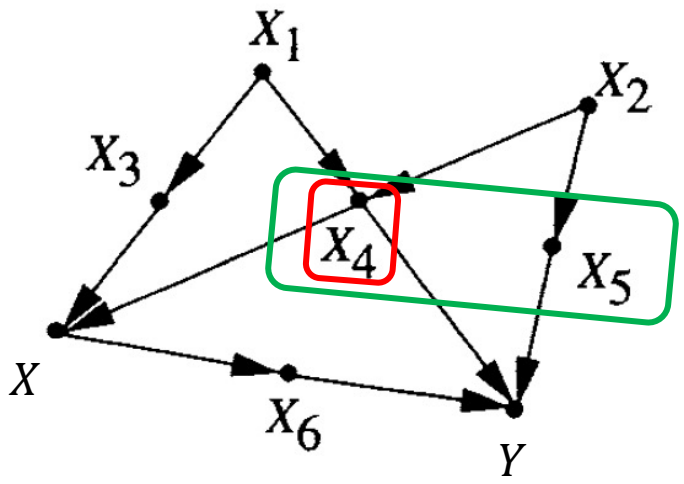
relations d'indépendance — structure de factorisation

Théorème : $(X \perp\!\!\!\perp Y|\mathbf{Z})_G \Rightarrow (X \perp\!\!\!\perp Y|\mathbf{Z})_P$ pour toute distribution P Markov p.r. au DAG G.
L' « inverse » est vrai aussi.

Remarque : c'est **l'absence de liens** entre deux nœuds qui est porteuse d'information. Exemple : une distribution sans relation d'indépendance est associée à un graphe G (dont le squelette est le graphe) complet.

Intuition pour le critère de $d-$ séparation : la transmission d'information peut-être bloquée, soit par la mesure d'une variable médiatrice, soit par l'absence de mesure : « *explaining away phenomenon* ».

Critère graphique n°1 : la « back-door rule »



Définition : soit un DAG G associé à un modèle causal $\mathcal{M}$. Parmi les variables notons $\mathbf{V}$ celles qui sont observables. Un ensemble $\mathbf{Z} \subseteq \mathbf{V}$ satisfait la **back-door rule** par rapport une cause X et un effet Y si :

1. Aucun nœud dans $\mathbf{Z}$ n'est un descendant de la cause X,
2. $\mathbf{Z}$ bloque tous les chemins entrants dans la cause X (« *back-door path* »).

Théorème (*back-door adjustment*) : Si un ensemble de variables $\mathbf{Z}$ satisfait la back-door rule par rapport a X et Y alors l'effet causal $P(Y|\mathrm{do}(X))$ est **identifiable** au moyen de :

$$p(y|\mathrm{do}(X=x)) = \sum_{\mathbf{z}} p(y|x,\mathbf{z})\,p(\mathbf{z})$$

Permet d'optimiser le choix des variables à observer selon critères de coût ou de variabilité.

$$\mathbf{Z} = \{X_4, X_5\}$$

$$\begin{aligned}
\gamma_1 &= (X \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \leftarrow X_2 \rightarrow X_5 \rightarrow Y)\\
\gamma_2 &= (X \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \rightarrow Y)\\
\gamma_3 &= (X \leftarrow X_4 \leftarrow X_2 \rightarrow X_5 \rightarrow Y)\\
\gamma_4 &= (X \leftarrow X_4 \rightarrow Y)
\end{aligned}$$

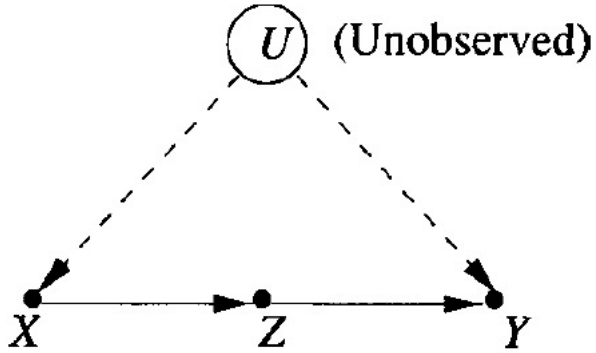
$$\mathbf{Z} = \{X_4\}$$

$$\gamma_1 = (X \leftarrow X_3 \leftarrow X_1 \rightarrow X_4 \leftarrow X_2 \rightarrow X_5 \rightarrow Y)$$

Critère graphique n°2 : la « front-door rule »

Définition : soit un DAG G associé à un modèle causal. Un ensemble $\mathbf{Z}$ de variables observées vérifie la **front-door rule** par rapport une cause X et un effet Y si :

1. $\mathbf{Z}$ intercepte tous les chemins dirigés de X vers Y.
2. Il n'y a aucun chemin entrant de X vers Z non-bloqué (par ∅).
3. Tous les chemins entrants de $\mathbf{Z}$ vers Y sont bloqués par X.



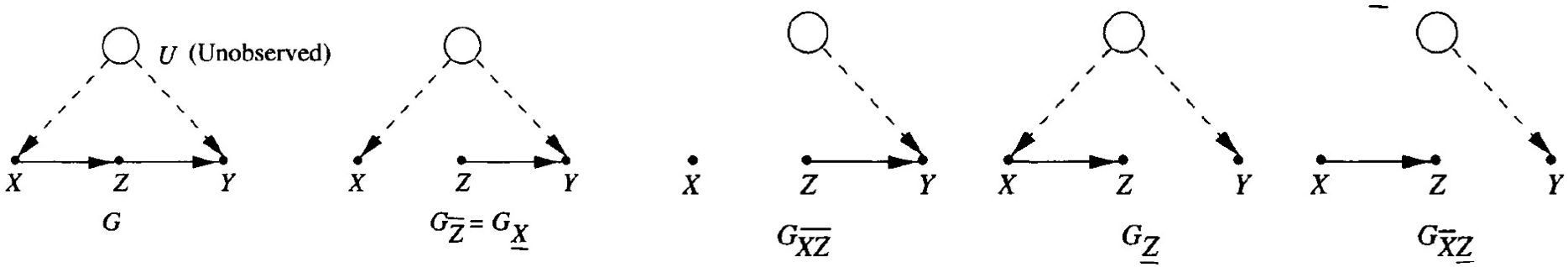
Ici $\mathbf{Z}$ est un descendant de la cause X !

Théorème (*front-door adjustment*) : Si un ensemble de variables $\mathbf{Z}$ satisfait la front-door rule par rapport a X et Y alors l'effet causal $P(Y|\mathrm{do}(x = X))$ est **identifiable** au moyen de :

$$p(y|\mathrm{do}(X = x)) = \sum_{\mathbf{z}} p(\mathbf{z}|x) \sum_{x'} p(y|x', \mathbf{z}) p(x')$$

Trois règles algébriques : le do-calculus

Notations



Règle n°1 [supprimer une observation]

Généralise le cas où $X = \emptyset$

$$p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathbf{z}, \mathbf{w})) = p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathbf{w})) \text{ when } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\, \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}}}}.$$

Règle n°2 [supprimer une intervention]

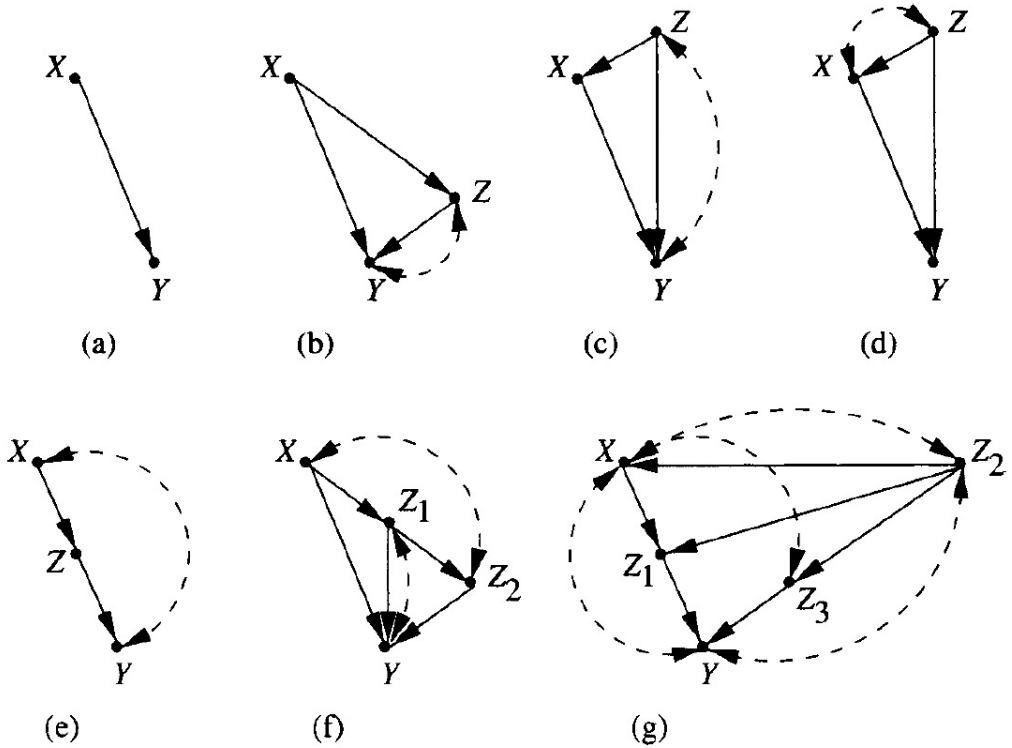
$$p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathrm{do}(\mathbf{z}), \mathbf{w})) = p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathbf{w})) \text{ when } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\, \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}} .$$

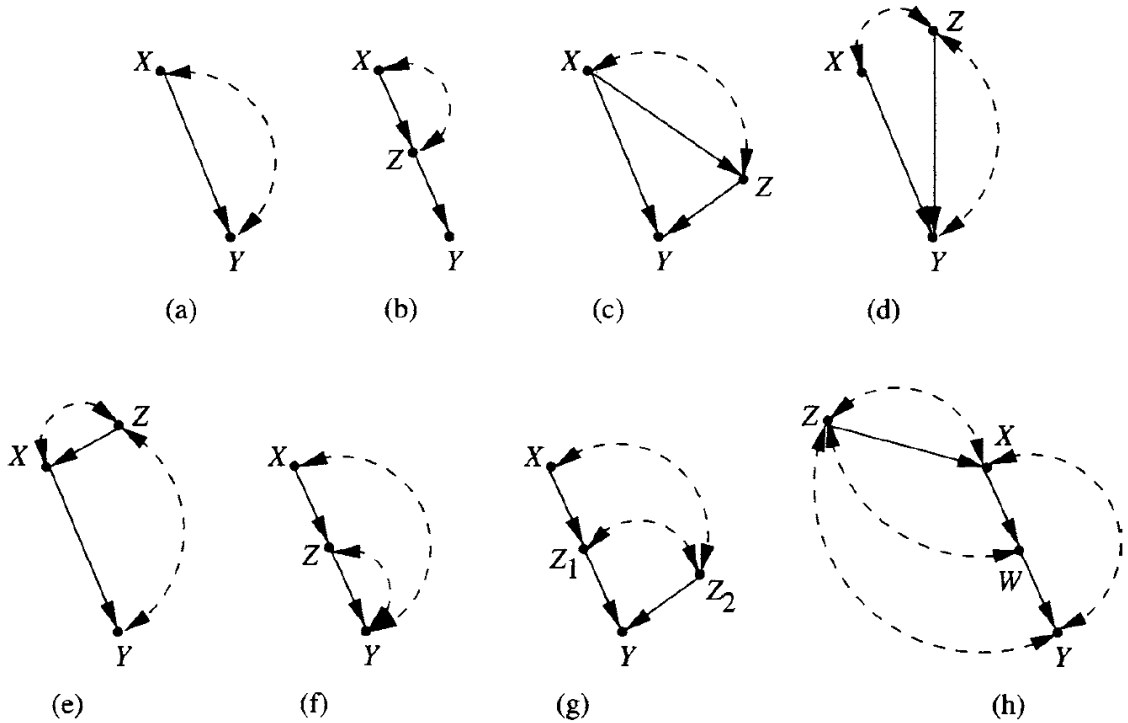
Règle n°3 [remplacer une intervention par une observation]

$$p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathrm{do}(\mathbf{z}), \mathbf{w})) = p(\mathbf{y}|\mathrm{do}(\mathbf{x}), \mathbf{z}, \mathbf{w})) \text{ when } (\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\, \mathbf{X}, \mathbf{W})_{G_{\overline{\mathbf{X}},\overline{\mathbf{Z(W)}}}}$$

Ces règles sont **complètes :** elles permettent de déduire tous les graphes identifiables !

Considérations générales sur l'identifiabilité d'un graphe

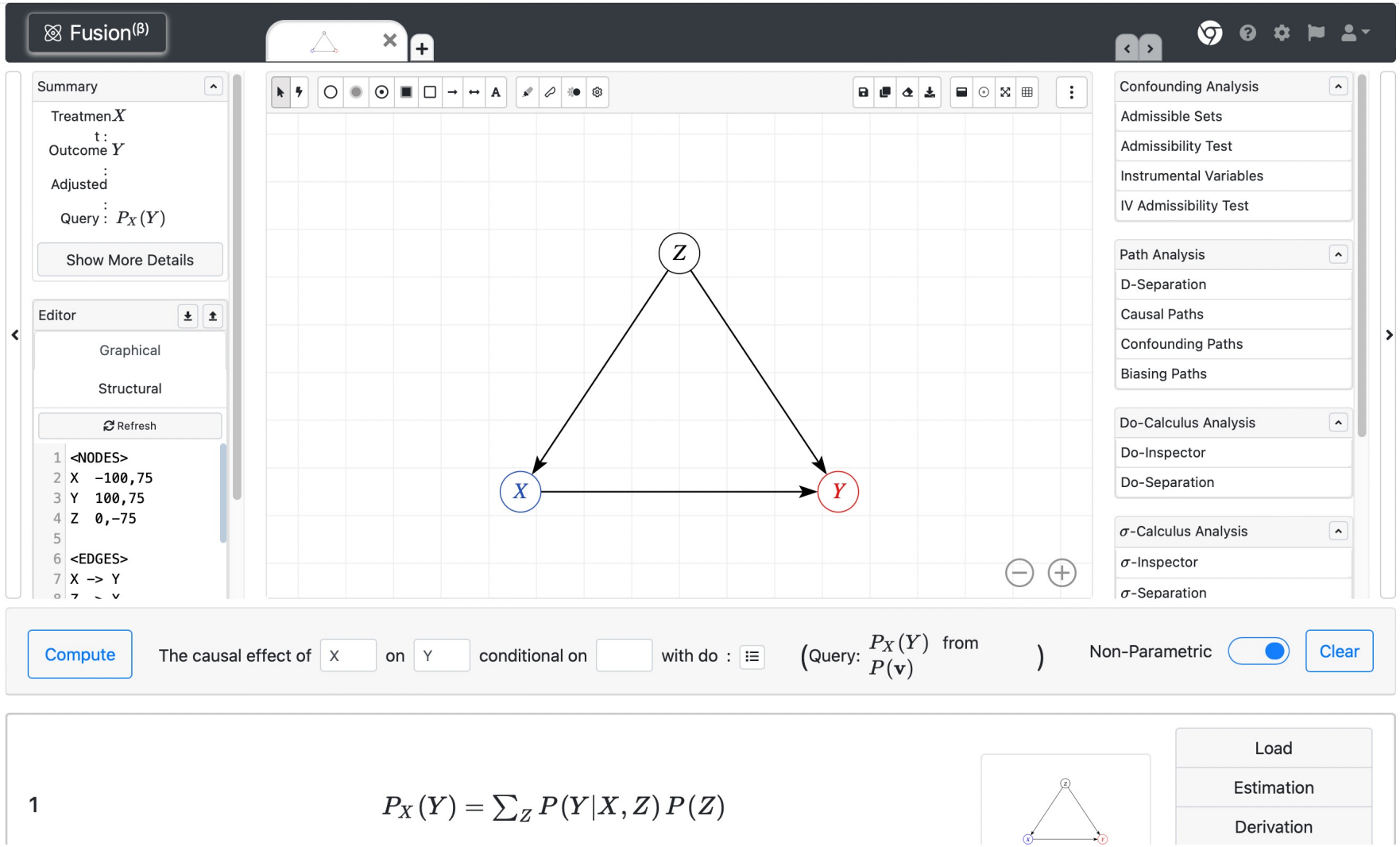




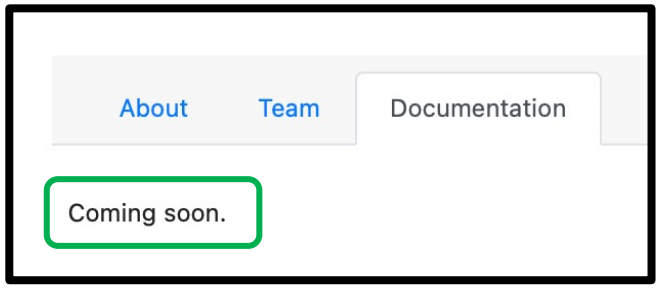
- Ce sont des graphes **identifiables** maximaux au sens où n'importe quel arc supplémentaire les rendrait non-identifiable.
- Supprimer un lien ne peut qu'augment les chances d'identifiabilité.

- Ces graphes sont tous **non-identifiables**
- Une condition suffisante est l'existence d'un **arc de confusion** entre la cause X et l'un de ces descendant sur un chemin de X à Y.

Algorithme général d'identification (gID)



https://causalfusion.net/login



Autres questions, relaxation des hypothèses, alternatives…

- **Identification partielle** d'un diagramme causal à partir d'observation (chapitre 2 dans Causality de JP)
- Quid des **relations cycliques** plutôt que des DAG ?
- Réduction d'une intervention à une ensemble d'observations et d'**interventions de substitution**
- Application à des systèmes avec un **très grand nombre de variables** (TAL !)
- Application à des **systèmes partiellement compris** (on ne connait pas entièrement le diagramme causal ou on fait l'hypothèse que le diagramme causal fait partie d'un ensemble de diagrammes)
- **Transportabilité** d'un modèle causal (dans le laboratoire) à un autre modèle causal (dans la réalité)
- La causalité « à l'ancienne », sans quitter le monde des probabilités, l'approche « ***potential outcome*** », les notion d' « ***ignorability*** » etc…

- Remarque historique : les statisticiens « *old school* », (= disciples de Fisher & Co), rejettent, dogmatiquement selon J. Pearl, l'idée que l'on puisse utiliser une autre source d'information que les seules données et insistent pour n'utiliser que des modèles probabilistes classiques. J. Pearl de son côté considère ce point de vue comme « extrémiste » et contradictoire avec la pratique scientifique. Le prix à payer un point de vue plus « orthodoxe » est une formulation non naturelle des questions causales.
- Mon point de vue : l'approche de J. Pearl est une métaphore de la méthode scientifique familière en physique p.ex.

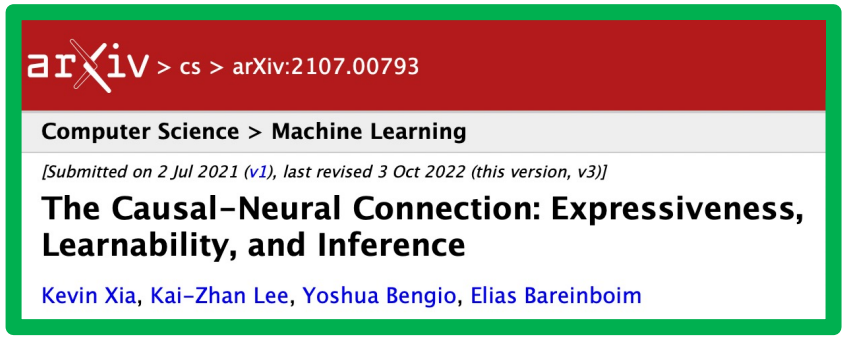
Teaser : causalité dans le TAL et le Deep Learning

- Question 1 : comment utiliser la théorie de JP dans lorsque la cause, l'effet ou les variables de confusion sont des variables textuelles ?
- Question 2 : comment tirer partie de l'analyse causale pour améliorer des modèles de TAL en allant au-delà des corrélations ?
- Question 3 : comment incorporer un modèle causal dans un modèle de Deep Learning ?
- Question 4 : quelle réduction dimensionnelle à effectuer sur une variable textuelle ?

Autres séminaires TALia liés à la causalité :

- n°35 – « *Right for the Wrong Reason* » (Jean-Louis)
- n°53 – Interprétabilité des modèles de NLP (Pirmin)

Causalité ∩ TAL



Causalité ∩ Deep Learning

Causalité ∩ interprétabilité

Pourquoi cette théorie est-elle si peu connue des data scientists ?

1. Des **raisons culturelles** : le mot causalité est longtemps resté tabou dans la communauté des statisticiens restés scrupuleusement fidèles aux fondateurs de la disciplines (Fisher et Pearson).

2. Les disciplines comme les sciences économiques, la médicine, l'agronomie ont développés un riche ensemble de modèles causaux. Mais dans les **situations courantes** où la relation causale est directe : $X \longrightarrow Y$ l'intervention $P(Y|\mathrm{do}(X = x))$ revient à une observation $P(Y|X = x)$.

3. Peu de chercheurs, de statisticiens, de data scientists peuvent se permettre de maîtriser un usage judicieux des règles du do-calculus. Mais… la situation pourrait changer avec l'avènement de **CausalFusion**, disponible en ligne !

4. Les ouvrages et les articles de Pearl sont **excessivement verbeux** (n'omet aucun argument échangé avec ses adversaires). Écrit dans un **style non conventionel**, à mi-chemin entre philosophie des sciences et mathématiques, (mais sans rédaction détaillée et systématique des démonstrations de théorèmes) Pearl est difficile, sinon fatigant à lire.