UE Inférence de réseaux

Introduction à la causalité

Olivier Goudet

Université d'Angers

15 janvier 2026

Organisation générale du cours

4 séances de 2 heures de CM :

1. Cours 1 - S. Aubourg
   - Introduction aux réseaux de gènes.
2. Cours 2 - O. Goudet
   - Introduction à la causalité.
   - Notions d'indépendance entre différentes variables.
   - Graphes non dirigés.
3. Cours 3 - O. Goudet
   - Graphes dirigés.
   - Causalité paire à paire.
4. Cours 4 - O. Goudet
   - Méthodes d'inférence de réseaux utilisées en bioinformatique.

Objectif général du cours

- Présenter la notion d'inférence de graphes non dirigés et dirigés à partir de données.
- C'est en lien avec le cours que vous avez eu cette année :
  - Régression linéaire en Grande Dimension - Fabien Panloup
- Mettre en perspective certaines méthodes utilisées dans la littérature pour l'inférence de réseaux biologiques.
- Application en pratique de ces méthodes sur un **challenge** avec des données génomiques simulées (4 séances de TP de 2 heures). Plate-forme CodaLab.

Section 1

Introduction au domaine de la causalité en *machine learning*

Buts des modèles causaux

- Enjeu dans de nombreux domaines : comprendre le processus génératif des données avec des modèles causaux.
- Va au delà de la prévision et de la simple notion de dépendance statistique : **corrélation n'implique pas causalité !**
- Un domaine relativement nouveau dans la communauté *Machine learning*.
- Travaux précurseurs de Judea Pearl à la fin des années 90.
- Livre Causality (J. Pearl 2009).
- Depuis 2022 une conférence internationale spécialisée sur la causalité en *machine learning* : *Causal Learning and Reasoning (CLeaR)*.

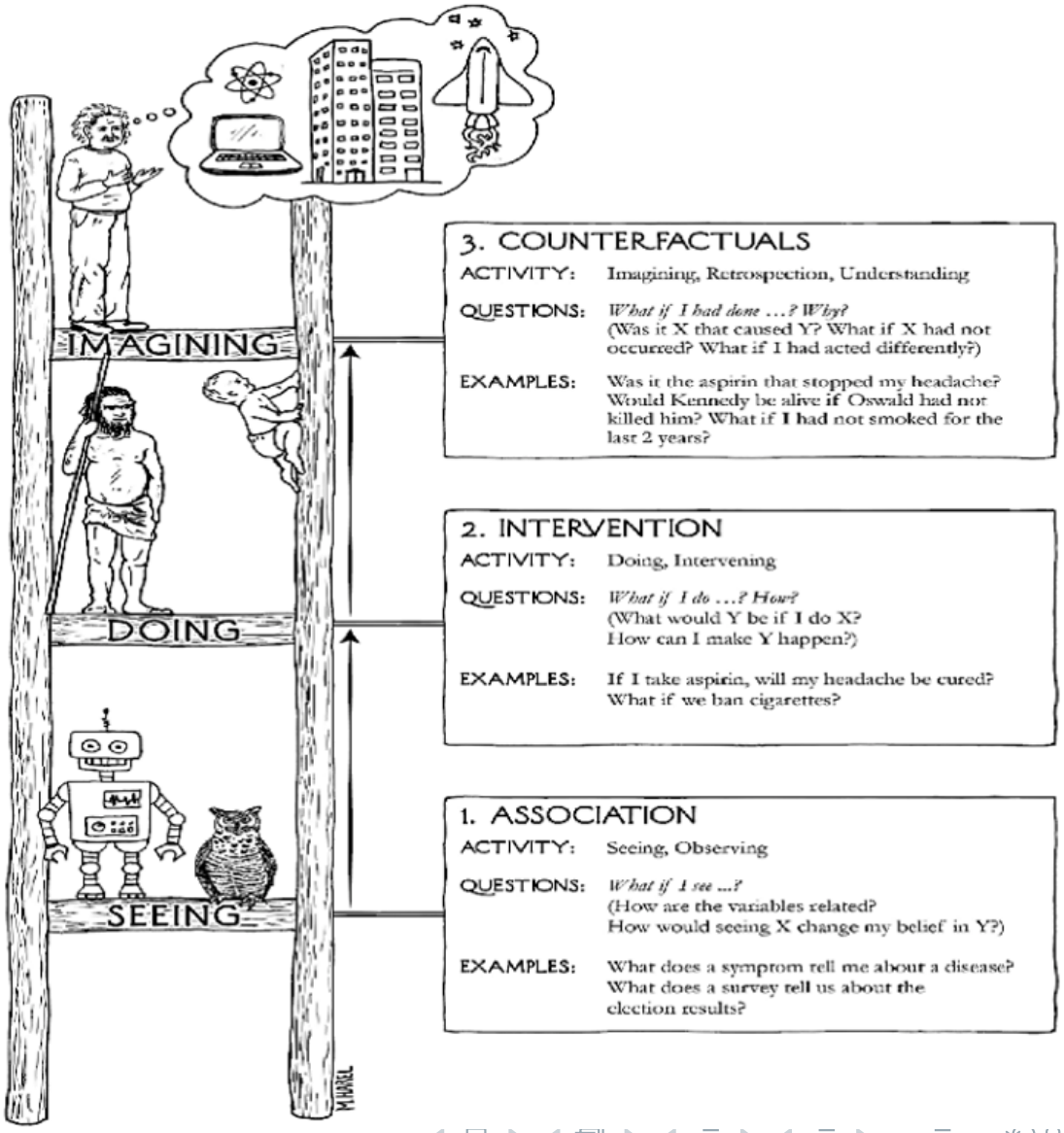
Échelle de causalité (J. Pearl)

The Ladder of Causality

"Actual" Causality

"Causality-in-mean"

Statistics



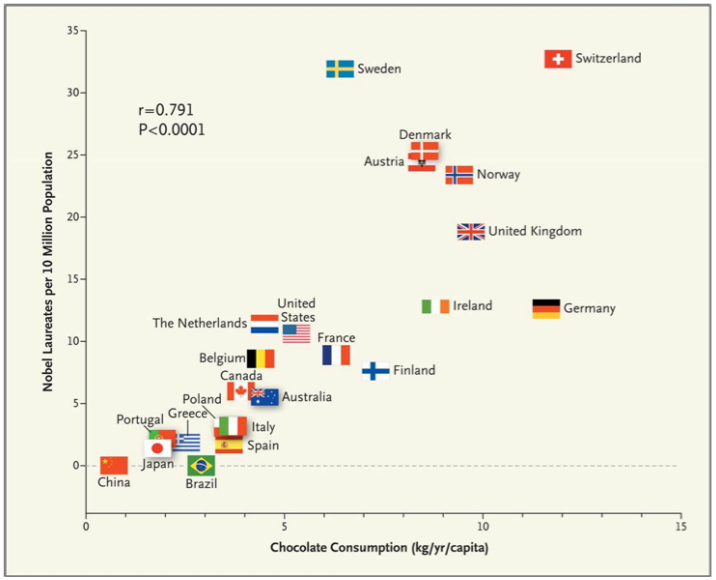
Modèle causal en biologie

- En biologie : **réseau de co-expression (GCN)** $\rightarrow$ **réseau de régulation (GRN)**
- Permet de mieux expliquer des phénomènes.
- Permet de réaliser des actions qui vont avoir des impacts sur des variables cibles.
- Est-ce que modifier le fonctionnement d'un gène donné va avoir un impact sur la résistance de la plante à des éléments pathogènes ?

Utilité en pratique des modèles causaux

- Baser des **recommandations** simplement sur des statistiques sans connaître le graphe causal peut mener à des erreurs.
- On va voir quelques exemples où des prescriptions qui ne s'appuieraient pas sur la **connaissance du "vrai" graphe causal sous-jacent** peuvent mener à des interprétations erronées.
- Certains cas sont évoqués depuis plus de cent ans dans la littérature. Ils étaient vus comme des **paradoxes** car les modèles causaux étaient mal connus à l'époque.

Exemple 1 (I.Guyon) : faut-il distribuer du chocolat aux chercheurs pour obtenir plus de prix Nobel ?



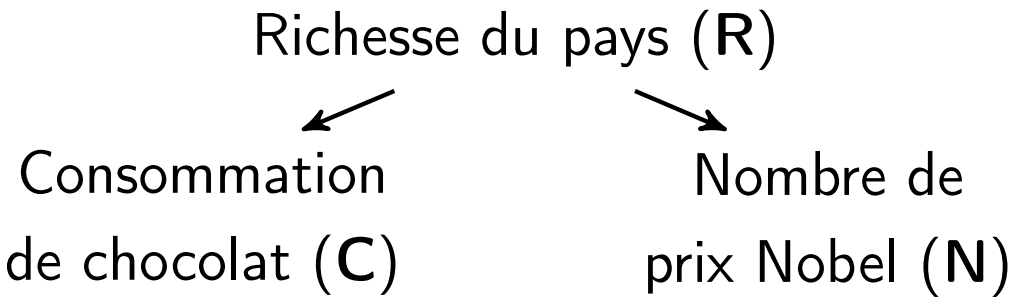
F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

- Corrélation positive et très significative entre la consommation de chocolat (en abscisses) et le nombre de prix Nobel (en ordonnées) pour cet ensemble de pays.
- Est-ce qu'il existe une relation causale entre ces deux variables ?
- Sinon quelle explication proposez vous ?

Première explication possible

- Ici on observe que les deux variables **C** et **N** sont dépendantes à partir des données observées.
- Il peut s'agir d'une **corrélation fallacieuse** (*spurious*) qui apparaît parfois quand on dispose d'un nombre limités de données d'observation, qui s'alignent juste par l'effet du hasard.
- Voir des exemples amusants ici : `https://www.tylervigen.com/spurious-correlations`

Une explication causale possible



- Si cette corrélation existe vraiment (et se vérifie avec d'autres données), les deux variables observées sont peut être toutes les deux la cause d'une variable commune cachée, par exemple la richesse du pays (**R**).
- On aurait dans ce cas, d'après le graphe de causalité ci-dessus la relation $C \not\perp\!\!\!\perp N$ (les variables C et N ne sont pas indépendantes). Par contre, connaissant la richesse du pays, ces deux variables sont peut être indépendantes : $C \perp\!\!\!\perp N | R$.

Pour le vérifier : effectuer une expérience randomisée !

Faire un test avec des chercheurs pendant 10 ans :

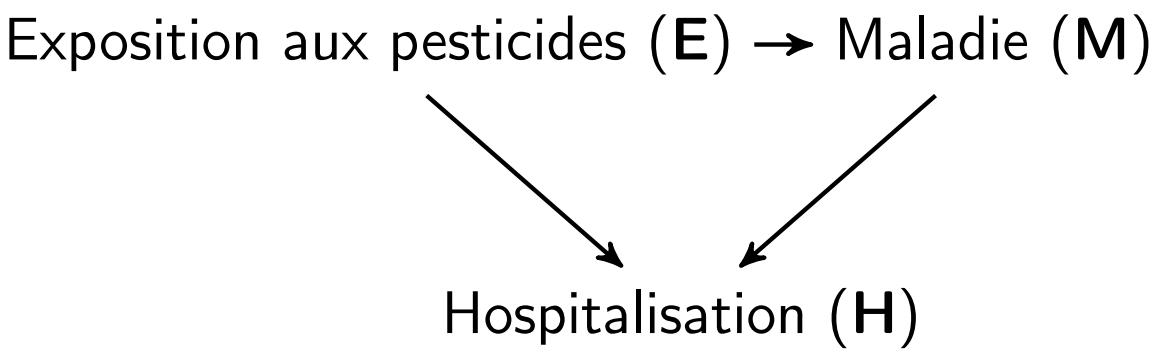
- Choisir aléatoirement la moitié d'entre eux et leur donner du chocolat.
- Donner des pommes aux autres.
- Comparer les nombres de prix Nobel obtenus dans les deux groupes.

Exemple 2 (A. Aussem) : paradoxe de Berkson (1946)

	Exposé		Non exposé	
	Malade	non malade	Malade	non malade
Hospitalisé	22	600	8	200
Non Hospitalisé	8	400	12	800

- Parmi les personnes hospitalisées, la prévalence de la maladie est de 3.5% (22/622) parmi les personnes exposées au risque et de 3.8% (8/208) parmi les patients non-exposés.
- En regardant uniquement les personnes à l'hôpital, on pourrait conclure que le fait d'être exposé aux pesticides réduit en fait le risque d'avoir un cancer...
- Or, dans le groupe entier la prévalence d'un cancer est de 2.9% $((22+8)/(22+8+600+400))$ parmi les personnes exposées aux pesticides et de 1.9% parmi les personnes non-exposées.
- Comment expliquer qu'on peut avoir des conclusions complètement différentes en regardant ces deux groupes ?

Explication causale possible de ce phénomène

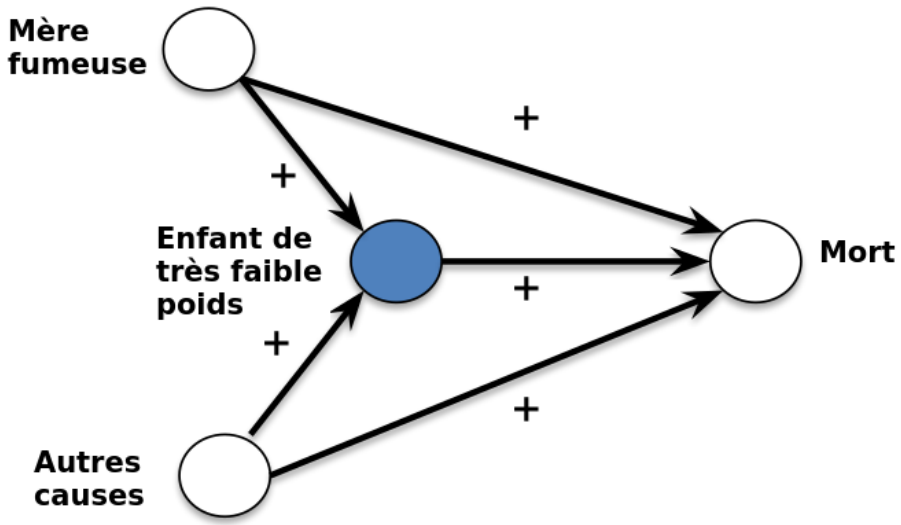


- Conditionnellement au fait d'être hospitalisé (**H**) les variables Maladie (**M**) et Exposition aux pesticides (**E**) deviennent dépendantes négativement.
- Alors que dans la population générale (si on ne conditionne pas sur **H**), il peut y avoir un impact "positif" de **E** sur **M** (dans le sens que augmenter **E** augmente **M**).

Exemple 3 (A. Aussem) : paradoxe du poids des naissances (1967)

- Les enfants de mères fumeuses sont plus susceptibles de donner naissance à un enfant de très faible poids.
- Les enfants de très faible poids ont un taux de mortalité beaucoup plus important que les autres.
- Contrairement à ce qu'on pourrait penser, les enfants de très faible poids, mais nés de mères fumeuses ont un taux de mortalité plus bas que les autres enfants de très faible poids.
- Conclusion : avoir une mère fumeuse est bénéfique pour la santé de l'enfant !

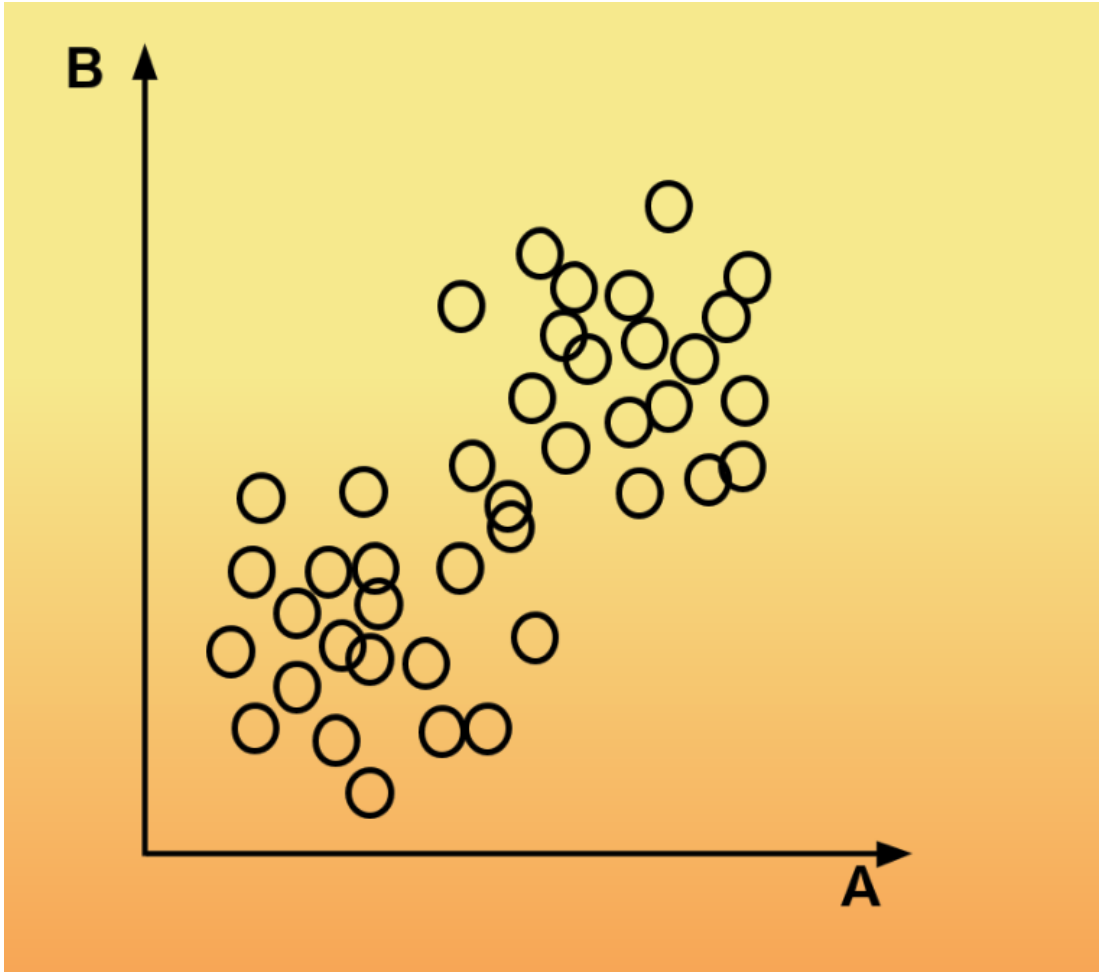
Explication causale possible de ce phénomène



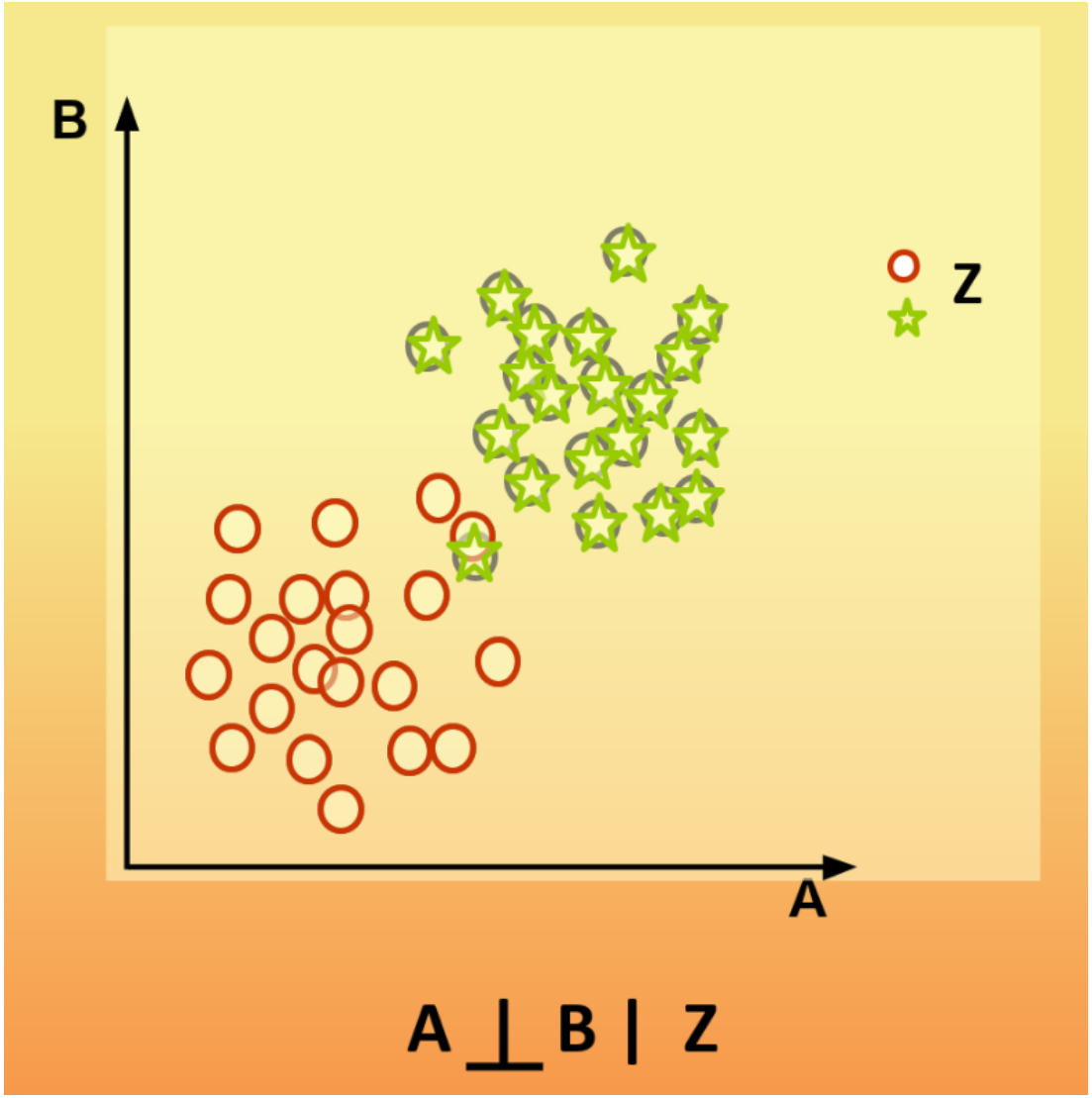
- Fumer peut être nuisible dans le sens où cela contribue au risque de sous-poids de l'enfant, mais d'autres causes de sous-poids peuvent être encore plus néfastes sur le taux de survie de l'enfant (ex : graves problèmes génétiques).
- Si on considère un enfant de faible poids, savoir que sa mère est fumeuse réduit en fait la probabilité qu'une autre cause soit présente.

Exemple 4 (I. Guyon) : paradoxe de Simpson (1899)

Est-ce que vous voyez un lien de dépendance statistique entre A et B ?
Est-ce qu'il existe un lien de cause à effet entre A et B ?

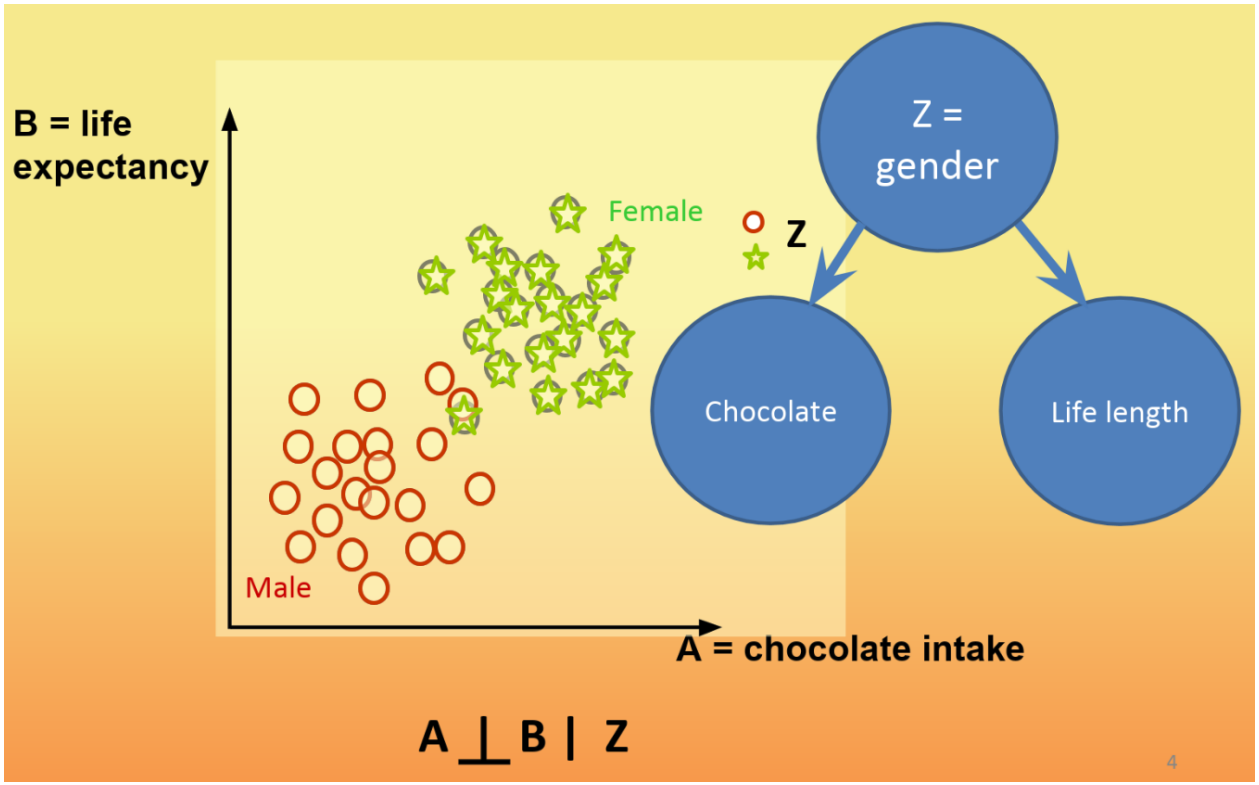


Paradoxe de Simpson



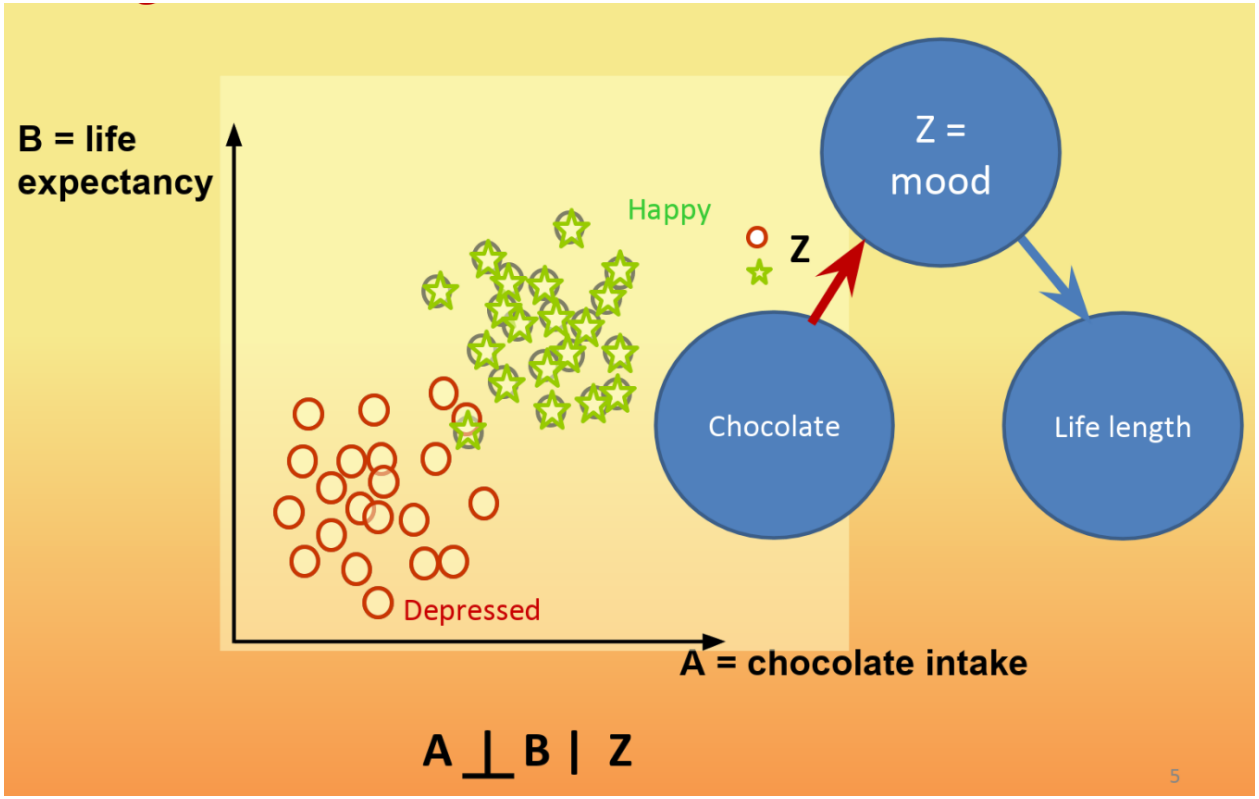
Paradoxe de Simpson

Exemple de cas où il n'y a pas de relation de cause à effet entre A et B.



Paradoxe de Simpson

Exemple de cas où il y a effectivement une relation de cause à effet entre A et B.



Plusieurs modèles possibles

- Ils existent plusieurs modèles causaux explicatifs cohérents avec les tests d'indépendance statistiques effectués : $A \not\perp\!\!\!\perp B$ et $A \perp\!\!\!\perp B | Z$.
- On verra plus tard cette notion de classes d'équivalence de Markov: $A \to Z \to B$, $B \to Z \to A$ et $A \leftarrow Z \to B$.
- On verra aussi qu'il existe quand même un moyen de retrouver le vrai graphe causal en cherchant le modèle le plus "simple" qui permet d'expliquer les données (Rasoir d'Ockham).
- **Rasoir d'Ockham** : "Les hypothèses suffisantes les plus simples doivent être préférées" (principe heuristique fondamental en science).

Comment inférer un graphe causal ?

- La voie royale :
  - Interventions et essais randomisés.
  - Ce qui est fait pour tester les effets des médicaments par exemple.
- Mais dans beaucoup de domaines, les interventions sont :
  - impossibles (climat)
  - non éthique (convaincre les gens de fumer)
  - trop chères (économie)
  - longues à mettre en place expérimentalement (biologie)

Deux principaux problèmes dans le domaine de la causalité

Si on ne s'appuie pas sur des interventions, il y a deux grands problèmes en causalité explorés dans la littérature scientifique :

1. **Problème 1** (*Causal discovery* ou *Structure learning*): Il s'agit de trouver le graphe causal représentant les relations de cause à effet entre des variables observées à partir de données d'observation. Parfois on se limite à la découverte d'une classe d'équivalence de graphe. On donne aussi souvent en pratique des scores de confiance à chaque relation causale que l'on peut inférer.

2. **Problème 2** (*Do-calculus*): Etant donné des données d'observation et un graphe causal qui relie des variables observées, trouver la distribution d'intervention d'une variable aléatoire X_j en réponse à la modification d'une autre variable X_i par une intervention externe (manipulation) en contrôlant les effets des autres variables observées $\mathbf{X}_{\setminus ij} := X_{\{1,\dots,d\}\setminus i,j}$:

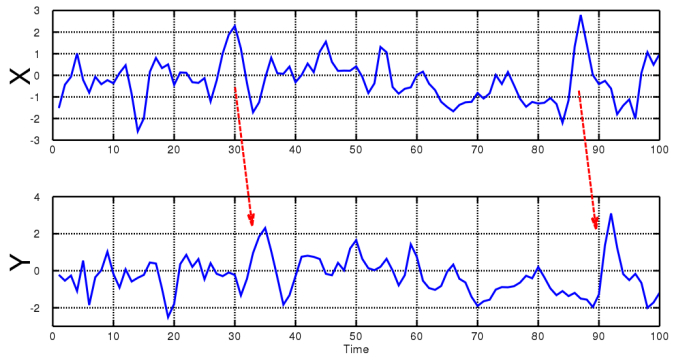
$$P_{X_j|\mathrm{do}(X_i=x,\mathbf{X}_{\setminus ij}=\mathbf{c})} \tag{1}$$

Inférence de la structure d'un graphe causal à partir de données

- Apprendre un graphe de causalité sans avoir besoin de faire de nouvelles expériences.
- Modèles fondées sur des données déjà acquises.
- Construire un modèle plausible qui explique la façon dont les données ont pu être générées.
  $\rightarrow$ **Sélection de modèles.**
- Inférence d'un graphe causal.
  $\rightarrow$ Tâche de "rétro-ingénierie".

Différents types de données disponibles pour l'inférence d'un graphe causal (1/2)

- **Cas 1 :** des séries temporelles $[x_{1,t}, x_{2,t}, ..., x_{d,t}]$. cf. notion de causalité au sens de Granger. On ne parlera pas de ce cas ici.



- **Cas 2 :** des données d'intervention. On sait qu'une ou plusieurs variables précises ont été modifiées et on a mesuré leurs effets sur les autres variables étudiées.
- **Cas 3 :** des données observées sans indications de temps mais supposées échantillonnées de façon iid suivant une distribution jointe inconnue.

Différents types de données disponibles pour l'inférence d'un graphe causal (2/2)

Types de variables observées :

- **Cas A :** variables continues - par exemple, la mesure de l'expression d'un gène relevé sur une puce à ADN.
- **Cas B :** variables discrètes - par exemple, le contexte de l'expérience (1 jour, 0 nuit)

Cas étudié dans ce cours

- Certaines méthodes de la littérature combinent ces différents types de données: Cas 1, 2 et 3, mais aussi Cas A et B (données mixtes).
- Dans ce cours, on considère que l'on n'a pas d'information de temps, qu'il n'y a pas d'interventions connues et que toutes les variables sont continues (Cas 3A).
- C'est le cas qui nous intéresse dans ce cours pour l'étude des matrices d'expression des gènes.

Inférence d'un graphe causal de régulation génétique à partir de données d'observation continues

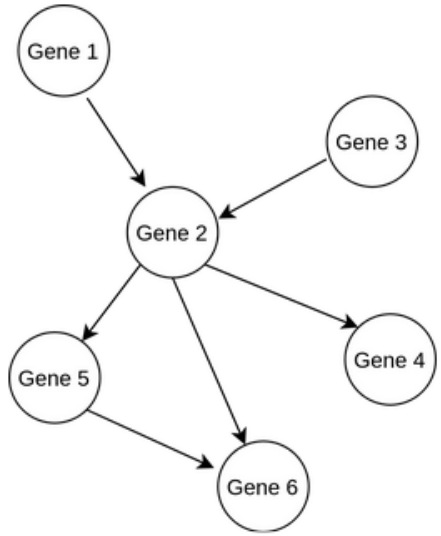
Matrice d'expression du génome de la plante *Arabidopsis thaliana*

26 374 gènes

1042 Expériences

	Gène 1	Gène 2	Gène 3	Gène 4	Gène 5	Gène 6	...
Expérience 1	1,13	1,49	0,78	1,71	2,11	5,09	
Expérience 2	1,10	0,58	1,25	0,37	1,04	4,56	
Expérience 3	2,13	1,40	0,40	1,56	2,03	5,48	
Expérience 4	0,70	0,78	0,85	0,62	1,51	4,25	
Expérience 5	1,52	1,28	0,88	0,85	1,35	5,55	
Expérience 6	2,01	2,04	-0,04	0,10	3,14	4,27	
Expérience 7	0,74	0,54	1,45	0,20	2,52	5,92	
Expérience 8	0,23	0,85	1,61	0,15	2,14	6,04	
Expérience 9	1,64	1,64	1,65	1,52	3,53	6,25	
Expérience 10	1,10	1,37	1,28	0,97	3,02	5,71	
Expérience 11	0,77	0,92	0,83	0,73	3,14	6,16	
Expérience 12	1,39	1,73	0,93	1,24	3,49	5,88	
Expérience 13	1,41	1,65	1,32	1,18	3,88	7,02	
Expérience 14	1,57	2,25	0,61	0,80	1,74	4,80	
Expérience 15	0,98	1,67	0,42	0,13	1,96	4,64	
Expérience 16	2,23	1,71	0,83	0,85	1,91	7,39	
...							

Exemple de graphe de régulation génétique inféré



Données simulées et données réelles

Pendant les TP, on va exploiter deux types de données :

- **Des données simulées** : on connaît le "vrai" graphe qui a généré les données. Le but est de le retrouver.
  - $\rightarrow$ Permet de comparer des méthodes d'inférence de réseau avec un score global.
- **Des données réelles** : on ne connaît rien du vrai graphe (ou très peu de choses). On cherche à inférer un graphe qui permettra de suggérer des relations causales entre les gènes qui seront à valider par des expériences.
  - Base de mesures d'expression des gènes relevées pour un grand nombre d'expériences menées au cours des dernières années à l'IRHS.
  - Suggestion de relations entre les gènes (pour faire de nouvelles découvertes ?...)

En plus des données...

En pratique, tenir compte des connaissances du domaine peut aussi être utile.

- Par exemple : restreindre la recherche des causes de certains gènes parmi une liste connue de facteurs de transcription. Cas des challenges Dream4 (2009) et Dream5 (2010).
- Restreindre la recherche des mécanismes causaux à certaines classes de fonctions ou de processus si on a des a priori dessus.
  - Exemple : on sait que certains mécanismes physiologiques correspondent à des effets linéaires ou bien se déclenchent au delà d'un certain seuil.