

# Causal Inference & Paradoxes

---

Alexandre Aussem

LIRIS UMR 5205 CNRS

**Data Mining & Machine Learning (DM2L) Group**

Université Claude Bernard Lyon 1

[perso.univ-lyon1.fr/alexandre.aussem](http://perso.univ-lyon1.fr/alexandre.aussem)

[alexandre.aussem@liris.cnrs.fr](mailto:alexandre.aussem@liris.cnrs.fr)



# Outline of the talk

---

- Bayesian networks: From probability to causality
  - Manipulation theorem to estimate the **effect of external interventions**
  - **Confounding**: fundamental impediments to the elucidation of causal inferences from observational data
  - Elucidation of some well-known controversies :
    - The selection bias or **Berkson's paradox** (1946),
    - The **birth-weight paradox** (1967)
    - The **Simpson's paradox** (1899)
    - The old debate on the relation between **smoking and lung cancer** (1964),
    - **Sex discrimination**: The « reverse regression controversy » between sex and salary which occupied the social science in the 1970s
  - Rules of « **do calculus** »
  - Case study: **effect of the pesticides on agricultural yields**
  - Unbiased estimates despite **selection bias** and **missing data**
-

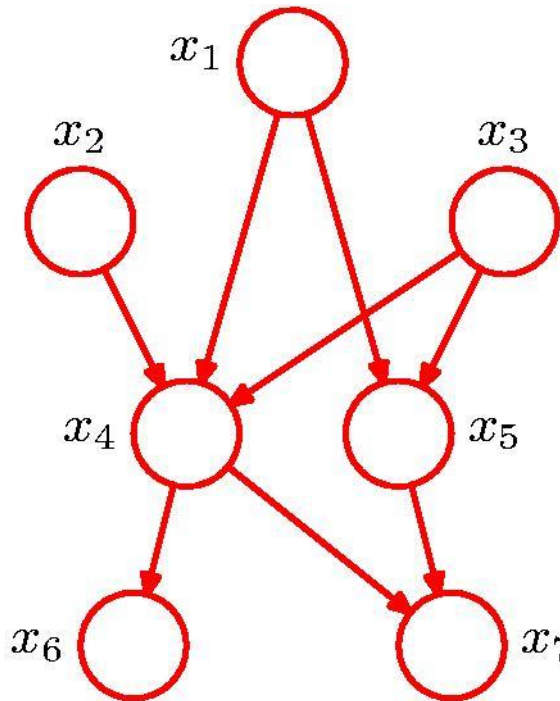
# Cause-effect relationships

---

- The central aim of many studies in the physical, behavioral, social, and biological sciences is the **elucidation of cause-effect relationships** among variables or events, e.g., risk factor exposure on disease occurrence, advertising campaign on benefits, treatment on recovery rate, etc.
  - However, the appropriate **methodology for extracting such relationships** from data has been fiercely debated.
  - **Graphical models** provide clear semantics for causal claims, and non-trivial causal phenomena, **paradoxes and controversies** in causal analysis that long were regarded as **metaphysical** can now be understood, exemplified, analyzed and solved using **elementary mathematics**.
  - Most of the material presented here is borrowed from **Judea Pearl's** books and papers.
-

# Bayesian Networks

$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



**General Factorization :**  $p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$

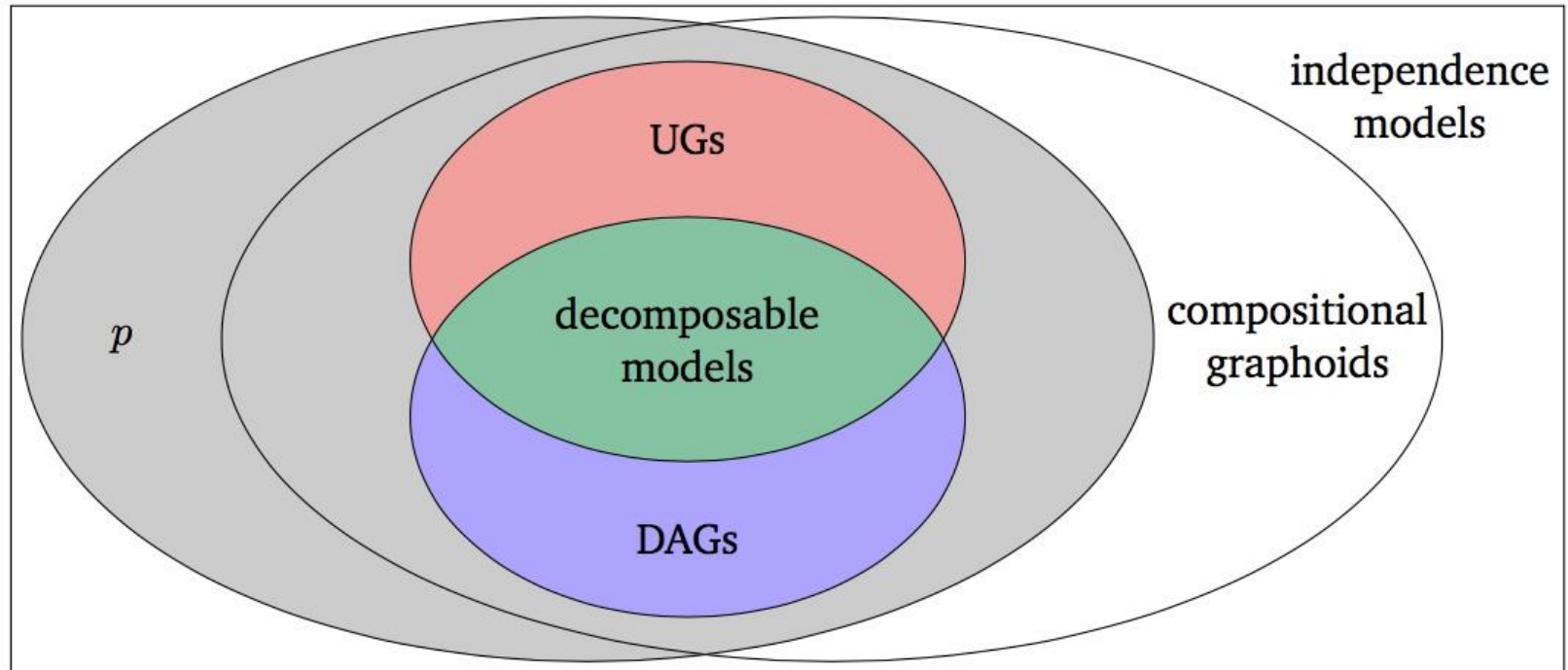
Conditional distribution for (binary) node  $x_7$  :

| $x_4$ | $x_5$ | 1   | 0   |
|-------|-------|-----|-----|
| 0     | 0     | 0.4 | 0.6 |
| 0     | 1     | 0.1 | 0.9 |
| 1     | 0     | 0.7 | 0.3 |
| 1     | 1     | 0.6 | 0.4 |

**Corollary (Markov condition) :** every node given its parents is independent on its non-descendants nodes. Other independencies are entailed (**d-separation** criterion).

# Independence models

---

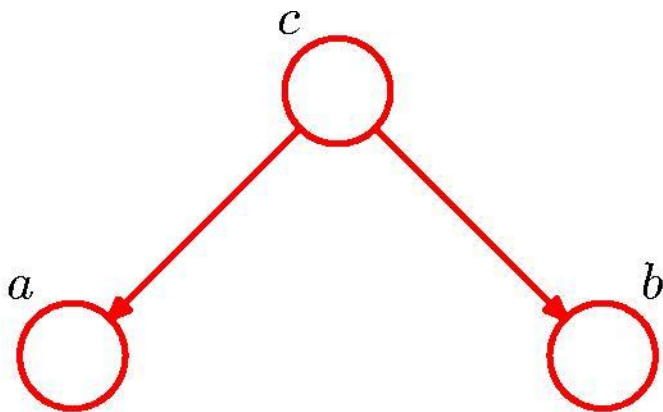


Overlapping between probabilistic independence models ( $p$ ), independence models based on u-separation (UG-faithful), and d-separation (DAG-faithful).

---

# Conditional Independence: Example 1

---

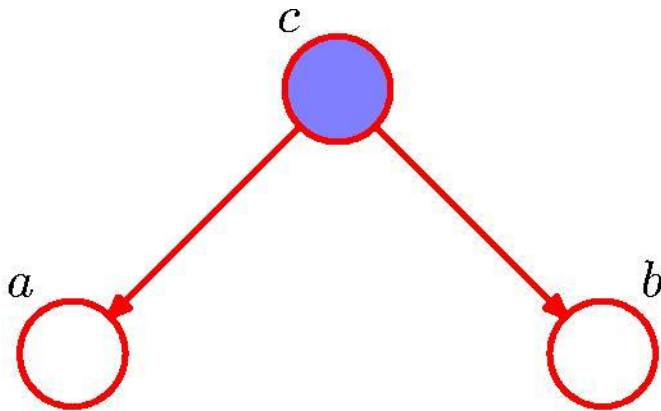


$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

$$p(a, b) = \sum_c p(a|c)p(b|c)p(c)$$

$$a \not\perp b \mid \emptyset$$

# Conditional Independence: Example 1

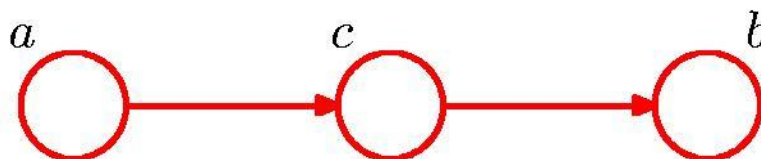


$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

$$a \perp\!\!\!\perp b \mid c$$

---

# Conditional Independence: Example 2



$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

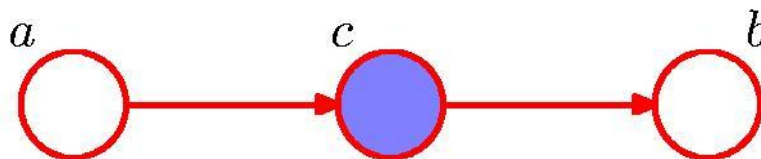
$$p(a, b) = p(a) \sum_c p(c|a)p(b|c) = p(a)p(b|a)$$

$$a \not\perp b \mid \emptyset$$

---

# Conditional Independence: Example 2

---



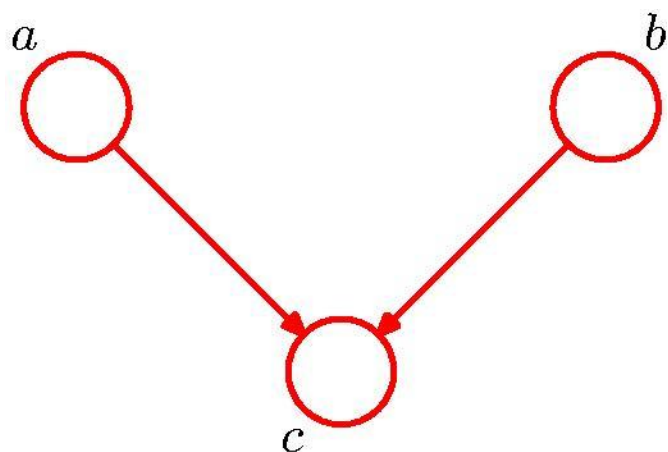
$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

$$a \perp\!\!\!\perp b \mid c$$

---

# Conditional Independence: Example 3

---



$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

$$p(a, b) = p(a)p(b)$$

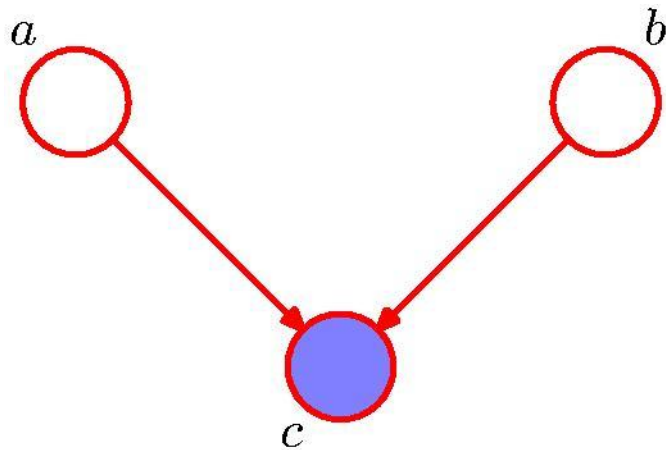
$$a \perp\!\!\!\perp b \mid \emptyset$$

Note: this is the opposite of Example 1, with  $c$  unobserved.

---

# Conditional Independence: Example 3

---



$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a, b)}{p(c)} \end{aligned}$$

$$a \not\perp b \mid c$$

Compared to the previous examples, the opposite is observed:  
Two **independent variables become dependent** given a third variable!

---

# Limits of Bayesian Networks

---

- Two given DAGs are **observationally equivalent** if *every* probability distribution that is compatible (or faithful) with one of the DAGs is also compatible with the other (same conditional independences encoded).
  - **Theorem:** Two DAGs are observationally equivalent if and only if they have the same skeletons and the **same sets of v-structures**, that is, two converging arrows whose tails are not connected by an arrow.
  - Observational equivalence places a limit on our ability to infer directionality from probabilities alone.
  - Networks that are observationally equivalent **cannot be distinguished without** resorting to **manipulative experimentation** or **human knowledge**
-

---

# **Causal Bayesian Networks**

---

# Graphs as Models of Interventions

---

- **Causal models**, unlike probabilistic models, can serve to **predict the effect of interventions**. This added feature requires that the joint distribution  $P$  be supplemented with a **causal diagram** - that is, a DAG that identifies causal connections.
- The causal diagram may represent the **investigator's understanding** of the major causal influences among measurable quantities in the domain.
- Each child-parent family in a DAG  $G$  represents a deterministic function:

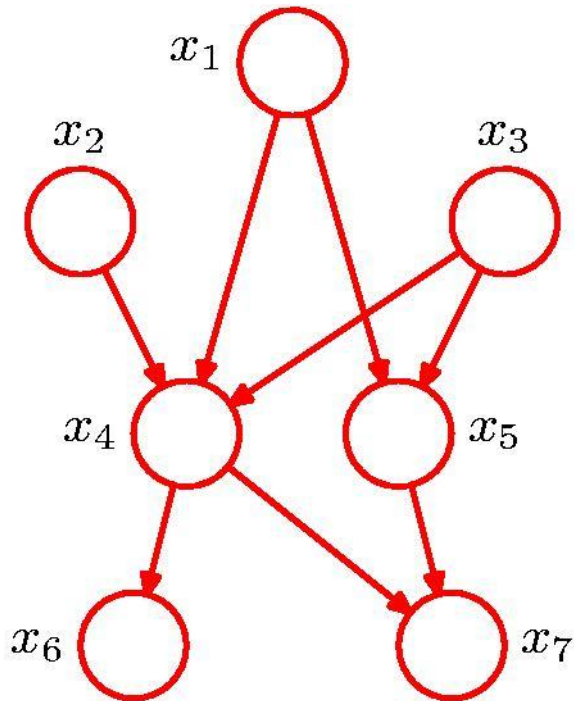
$$x_i = f_i(pa_i, \epsilon_i), \quad i = 1, \dots, n$$

where  $pa_i$  are the parents of variable  $x_i$  in  $G$ ; the  $\epsilon_i$  ( $i=1, \dots, n$ ) are mutually **independent**, arbitrarily distributed random disturbances.

- The equality signs in structural equations convey the **asymmetrical relation** of "is determined by".
-

# Causal Bayesian Networks

$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



## General Factorization

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k | \text{pa}_k)$$

Now **supplemented** with causal assumptions

$$x_i = f_i(\text{pa}_i, \epsilon_i), \quad i = 1, \dots, n$$

# Finding causal relationships

---

- For finding causal relationships, the gold standard are **randomized controlled trials** initially developed in the context of agricultural research (Fisher, 1926).
- Problem: Not always feasible for ethical, financial or other reasons.

We are left with two problems:

- **Problem 1 (Causal Structure)**: Given observational data, find the DAG representing the causal structure, or, if this is not possible, give a class of DAGs to which the true DAG belongs.
  - **Problem 2 (Interventional Distribution)**: Given observational data, find the interventional distribution of a random variable  $Y$  after some other random variable  $X$  was set to a certain value by external intervention to make quantitative predictions on the effect of interventions.
-

# Manipulation theorem

---

- The **manipulation theorem** (Spirtes et al. 1993) states that given an external intervention on a variable  $X$  in a causal graph, we can derive the posterior probability distribution over the entire graph by simply modifying the conditional probability distribution of  $X$ .
  - Intervention amounts to **removing all edges that are coming into**  $X$ . Nothing else in the graph needs to be modified, as the causal structure of the system remains unchanged.
  - Thus, intervention can be expressed in a **simple truncated factorization** formula.
-

# The do(.) operator

---

- Interventions are defined through a new mathematical operator called **do**( $X=x$ ), which simulates physical interventions by deleting the probability factor corresponding to variable  $X$  in the joint factorization, while keeping the rest unchanged elsewhere with  $X$  fixed to  $x$ .
- 
- The causal effect of  $X$  on  $Y$  is denoted  $P(y|\mathbf{do}(X=x))$ . It is termed an **interventional distribution** and should not be confused from the **observational distribution**  $P(y|x)$ .
- Interventions can be expressed as a simple **truncated factorization** formula:

$$P(x_1, \dots, x_n \mid \mathbf{do}(x_i = x'_i)) = \begin{cases} \prod_{j \neq i} P(x_j \mid \mathbf{pa}_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases}$$

# The do(.) operator

---

$$P(x_1, \dots, x_n \mid \mathbf{do}(x_i = x'_i)) = \begin{cases} \prod_{j \neq i} P(x_j \mid \mathbf{pa}_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases}$$

Can be rewritten as:

$$P(x_1, \dots, x_n \mid \mathbf{do}(x_i = x'_i)) = \begin{cases} P(x_1, \dots, x_n \mid x_i, \mathbf{pa}_j) P(\mathbf{pa}_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases}$$

Summing over all variables except  $x_i$  and  $y$  leads to the result called **adjustment for direct causes**:

$$P(y \mid \mathbf{do}(x_i = x'_i)) = \sum_{\mathbf{pa}_i} P(y \mid x'_i, \mathbf{pa}_i) P(\mathbf{pa}_i)$$

In compact form:

$$P(y \mid \mathbf{do}(x)) = \sum_{\mathbf{pa}_x} P(y \mid x, \mathbf{pa}_x) P(\mathbf{pa}_x)$$

# Controlling confounding bias

---

$$P(y \mid \text{do}(x)) = \sum_{\text{pa}_x} P(y \mid x, \text{pa}_x) P(\text{pa}_x)$$

- We ***adjust*** our measurements for possible variations of the **parents of  $X$**  in the causal DAG  $G$ , they are acting as “covariates” or « **confounders** ».
  - **Adjustment for the direct parents** amounts to **partitioning the population** into groups that are homogeneous relative to  $\text{pa}_x$  assessing the effect of  $X$  on  $Y$  in each homogeneous group, and then averaging the results.
  - This expression requires all the parents to be *observed*. Are other **variables appropriate for adjustment**?
  - What criterion should one use to decide **which variables are appropriate for adjustment**?
-

# Back-Door adjustment

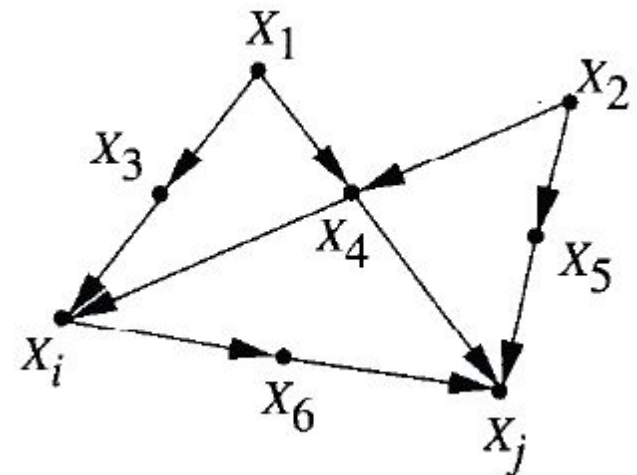
More generally, a set of variables  $\mathbf{Z}$  satisfies the **back-door criterion** relative to  $(X, Y)$  in a DAG  $G$  iff,

- No node in  $\mathbf{Z}$  is a descendant of  $X$ , and
- $\mathbf{Z}$  blocks every path between  $X$  and  $Y$  that contains an arrow into  $X$ .

**Theorem** – *If a set of variables  $\mathbf{Z}$  satisfies the back-door criterion relative to  $(X, Y)$ , then the causal effect of  $X$  on  $Y$  is **identifiable** and is given by the formula,*

$$P(y \mid do(x)) = \sum_{\mathbf{z}} P(y \mid x, \mathbf{z})P(\mathbf{z})$$

Example:



- The sets  $\mathbf{Z} = \{X_3, X_4\}$  and  $\mathbf{Z} = \{X_4, X_5\}$  meet the back-door criterion relative to  $(X_i, X_j)$
- But  $\mathbf{Z} = \{X_4\}$  does not !

---

# **Paradoxes & Controversies**

---

# Berkson's paradox

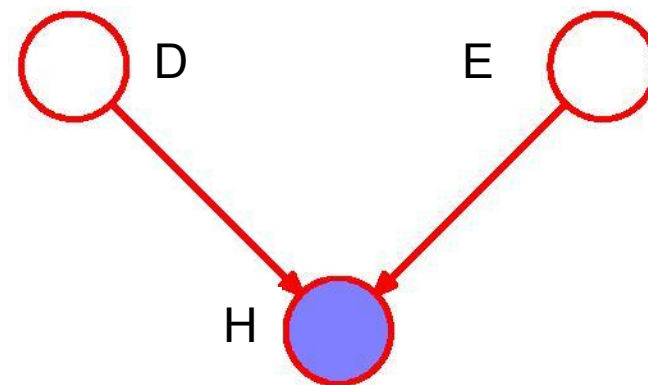
---

- Berkson's paradox is a result in **conditional probability** (not related to causality) which is counterintuitive for some people: given **two independent events**, if you only consider **outcomes where at least one event occurs**, then they become **negatively dependent**.
- **Example:** Berkson's original illustration involves a retrospective study examining a risk factor for a disease in a statistical sample. Because samples are taken from a hospital in-patient population, rather than from the general public, this can result in a spurious negative association between the disease and the risk factor

# Berkson's paradox

---

|       | $E^+$ |       | $E^-$ |       |
|-------|-------|-------|-------|-------|
|       | $D^+$ | $D^-$ | $D^+$ | $D^-$ |
| $H^+$ | 800   | 600   | 400   | 200   |
| $H^-$ | 200   | 400   | 600   | 800   |



- The prevalence of the disease (**D**) is 50% among exposed (**E**) and unexposed.
  - 70% are hospitalized (**H**) among exposed patients (30% among non exposed)
  - 60% are hospitalized among diseased patients (40% among non diseased).
  - **Within those hospitalized**, the prevalence of the disease is 57% among exposed and 66% among unexposed patients.
-

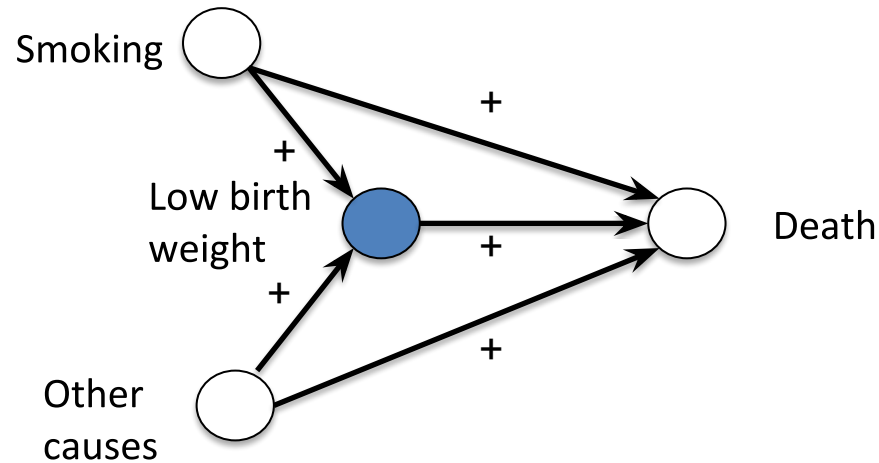
# Birth weight paradox

---

- The birth-weight paradox concerns the relationship between the birth weight and mortality. **Children of smoking mothers are more likely to be of low birth weight and low birth weight children have a significantly higher mortality rate than others** (it is in fact 100-fold higher)
- **Contrary to expectations**, low birth weight babies of smoking mothers have a lower child mortality than low birth weight babies of nonsmokers. **Having a smoking mother might be beneficial to one's health!**
- Like the Berkson's paradox, it is counterintuitive as it involves two independent events that become negatively dependent, having observed a third event.

# Birth weight paradox

---



- Smoking may be harmful in that it contributes to low birth weight, but other causes (not measured) of low birth weight are generally more harmful.
  - Consider a low weight baby, **finding that the mother smokes reduces the likelihood that those other causes are present.**
-

# Simpson's paradox

---

**C** : taking a certain drug or treatment

**E** : recovery

**F** : gender

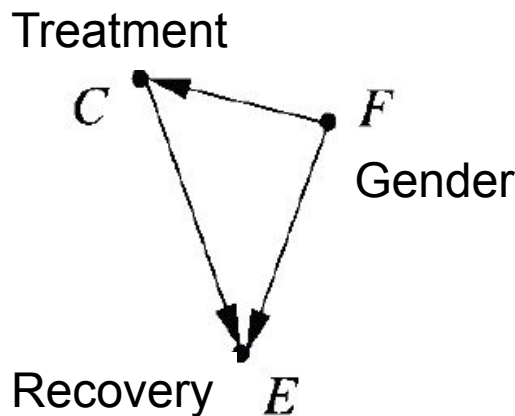
Under a causal interpretation the drug seems to be **harmful** to both males and females yet **beneficial** to the population as a whole !

| <b>Males</b>          | $E$ | $\bar{E}$ | Tot. | Recovery rate |
|-----------------------|-----|-----------|------|---------------|
| Drug ( $C$ )          | 18  | 12        | 30   | 60%           |
| No Drug ( $\bar{C}$ ) | 7   | 3         | 10   | 70%           |
|                       | 25  | 15        | 40   |               |
| <b>Females</b>        | $E$ | $\bar{E}$ | Tot. | Recovery rate |
| Drug ( $C$ )          | 2   | 8         | 10   | 20%           |
| No Drug ( $\bar{C}$ ) | 9   | 21        | 30   | 30%           |
|                       | 11  | 29        | 40   |               |
| <b>Combined</b>       | $E$ | $\bar{E}$ | Tot. | Recovery rate |
| Drug ( $C$ )          | 20  | 20        | 40   | 50%           |
| No Drug ( $\bar{C}$ ) | 16  | 24        | 40   | 40%           |
|                       | 36  | 44        | 80   |               |

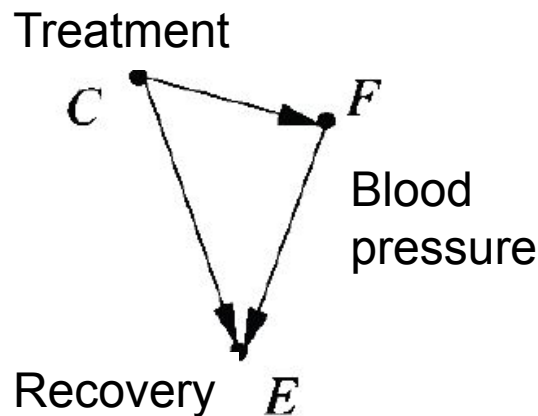
# Simpson's paradox

---

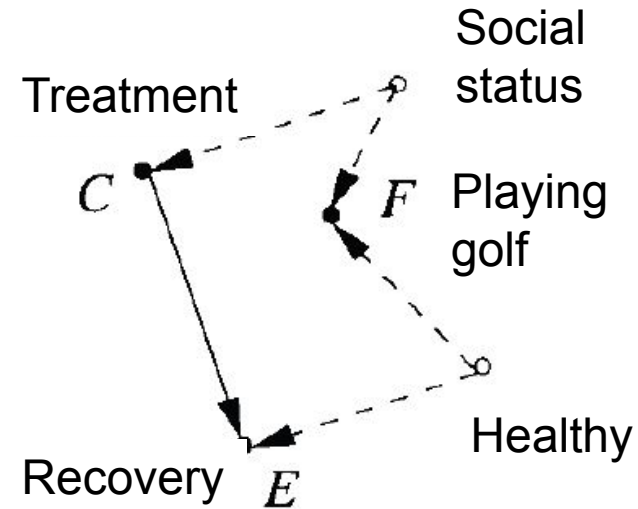
Three causal models capable of generating the data Model  
(a) dictates use of the **gender-specific tables**,  
whereas (b) and (c) dictates use of the **combined table**.



(a)



(b)



(c)

# Simpson's paradox

---

As **F** connotes gender, the correct answer is the gender specific table, i.e.

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

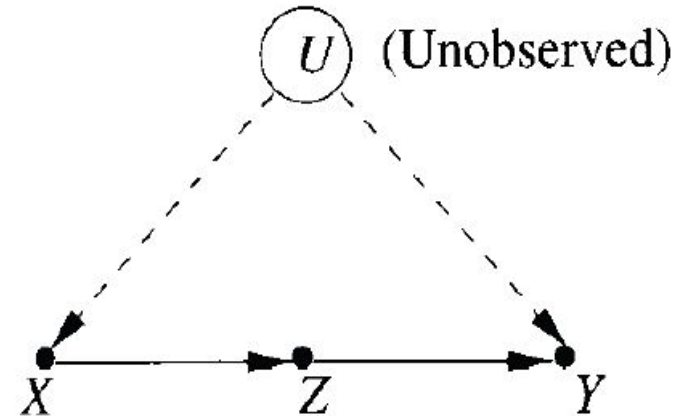
- **Conclusion:** every question related to the effect of actions must be decided by causal considerations; statistical information alone is insufficient.
  - The question of choosing the correct table on which to base our decision is a special case of the **covariate selection problem**.
-

# Front-Door adjustment

---

A set of variables  $Z$  is said to satisfy the **front-door criterion** relative to  $(X, Y)$  if

- $Z$  intercepts all directed paths from  $X$  to  $Y$ ;
- there is no back-door path from  $X$  to  $Z$ ;
- all back-door paths from  $Z$  to  $Y$  are blocked by  $X$ .



**Theorem** : If  $Z$  satisfies the front-door criterion relative to  $(X, Y)$  and if  $P(x, z) > 0$ , then the causal effect of  $X$  on  $Y$  is **identifiable** and is given by the formula:

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

If  $Z$  were **not observed**, the causal effect of  $X$  on  $Y$  would **not be identifiable**!

---

# Smoking and Lung Cancer

---

|                |                    | $P(x, z)$<br>Group Size<br>(% of Population) | $P(Y = 1 \mid x, z)$<br>% of Cancer Cases<br>in Group |
|----------------|--------------------|----------------------------------------------|-------------------------------------------------------|
| $X = 0, Z = 0$ | Nonsmokers, No Tar | 47.5                                         | 10                                                    |
| $X = 1, Z = 0$ | Smokers, No Tar    | 2.5                                          | 90                                                    |
| $X = 0, Z = 1$ | Nonsmokers, Tar    | 2.5                                          | 5                                                     |
| $X = 1, Z = 1$ | Smokers, Tar       | 47.5                                         | 85                                                    |

- Old debate on the relation between **smoking**,  $X$ , and **lung cancer**,  $Y$ .
  - If we ban smoking, will the rate of cancer cases be roughly the same as the one we find today among non smokers in the population ?
  - **Controlled experiments** could answer the question but they are **illegal** to conduct.
-

# Smoking and Lung Cancer

---

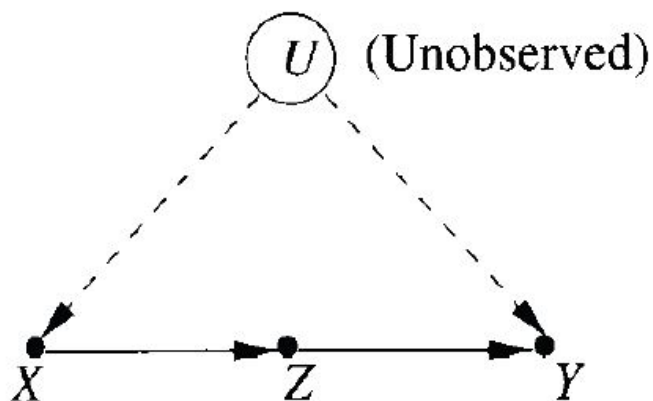
|                |                    | $P(x, z)$<br>Group Size<br>(% of Population) | $P(Y = 1 \mid x, z)$<br>% of Cancer Cases<br>in Group |
|----------------|--------------------|----------------------------------------------|-------------------------------------------------------|
| $X = 0, Z = 0$ | Nonsmokers, No Tar | 47.5                                         | 10                                                    |
| $X = 1, Z = 0$ | Smokers, No Tar    | 2.5                                          | 90                                                    |
| $X = 0, Z = 1$ | Nonsmokers, Tar    | 2.5                                          | 5                                                     |
| $X = 1, Z = 1$ | Smokers, Tar       | 47.5                                         | 85                                                    |

The tobacco industry has managed to forestall antismoking legislation (1964) by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of **carcinogenic genotype**,  $U$  (unknown), that involves **inborn craving for nicotine**.

---

# Smoking and Cancer

|                |                    | $P(x, z)$<br>Group Size<br>(% of Population) | $P(Y = 1   x, z)$<br>% of Cancer Cases<br>in Group |
|----------------|--------------------|----------------------------------------------|----------------------------------------------------|
| $X = 0, Z = 0$ | Nonsmokers, No Tar | 47.5                                         | 10                                                 |
| $X = 1, Z = 0$ | Smokers, No Tar    | 2.5                                          | 90                                                 |
| $X = 0, Z = 1$ | Nonsmokers, Tar    | 2.5                                          | 5                                                  |
| $X = 1, Z = 1$ | Smokers, Tar       | 47.5                                         | 85                                                 |



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

# Numerical application

---

- **Crude analysis:**

$$P(X = 1) = 0.5; P(Z = 1) = 0.5; P(Y = 1) = 0.475$$

$$P(Y = 1 \mid X = 0) = (0.1 \times 0.475 + 0.05 \times 0.025)/0.5 = 0.0975$$

$$P(Y = 1 \mid X = 1) = (0.9 \times 0.025 + 0.85 \times 0.475)/0.5 = 0.8525$$

- These results seem to prove that **smoking is a major contributor to lung cancer**.
  - However, the tobacco industry might argue that the table tells a different story - that smoking actually decreases one's risk of lung cancer...
-

# Numerical application

---

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

$$\begin{aligned} P(Y = 1 \mid do(X = 1)) &= 0.05 \times (0.1 \times 0.5 + 0.9 \times 0.5) \\ &\quad + 0.95 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\ &= 0.4525 \end{aligned}$$

$$\begin{aligned} P(Y = 1 \mid do(X = 0)) &= 0.95 \times (0.1 \times 0.5 + 0.9 \times 0.5) \\ &\quad + 0.05 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\ &= 0.4975 \end{aligned}$$

Contrary to expectation, the data prove **smoking** to be somewhat **beneficial to one's health** !

---

# Discrimination controversy

---

- Another example involves a controversy called « reverse regression », which occupied the social science literature in the 1970s.
  - Should we, in **salary discrimination cases**, compare **salaries of equally qualified men and women** or instead compare **qualifications of equally paid men and women**?
  - Remarkably, the **two choices may lead to opposite conclusions**. It turns out that men earns a higher salary than equally qualified women and, ***simultaneously***, men are more qualified than equally paid women.
  - The moral is that all conclusions are extremely sensitive to which variables we choose to hold constant when we are comparing groups.
-

# Discrimination controversy

---

- **Men earns a higher salary than equally qualified women** reads:

$$\sum_Q P(S|Male, Q)P(Q) > \sum_Q P(S|Female, Q)P(Q)$$

- **Men are more qualified than equally paid women** reads:

$$\sum_S P(Q|Male, S)P(S) > \sum_S P(Q|Female, S)P(S)$$

- The question we seek to answer: **does sex *directly* influence salary ?** Which is the court definition of discrimination, and reads:

$$P(S|\mathbf{do}(Male)) > P(S|\mathbf{do}(Female))$$

---

# Discrimination controversy

---

|   |    | S1 | S2 |     |
|---|----|----|----|-----|
| F | Q1 | 20 | 20 | 50% |
| H | Q1 | 20 | 40 | 66% |
| F | Q2 | 30 | 20 | 40% |
| H | Q2 | 20 | 30 | 60% |

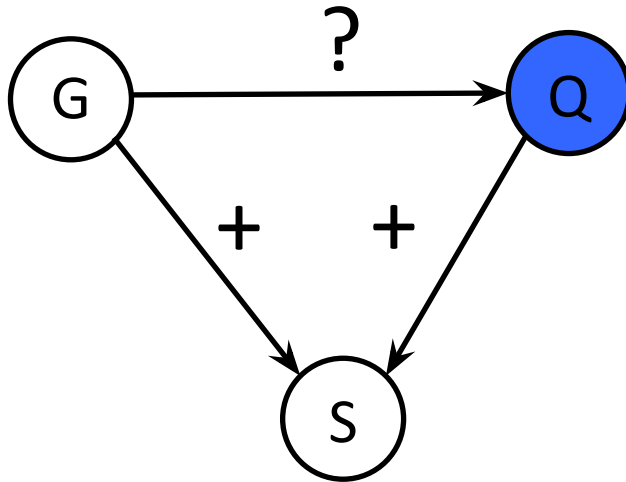
|   |    |    |     |
|---|----|----|-----|
| F | 50 | 40 | 44% |
| H | 40 | 70 | 62% |

# Discrimination controversy

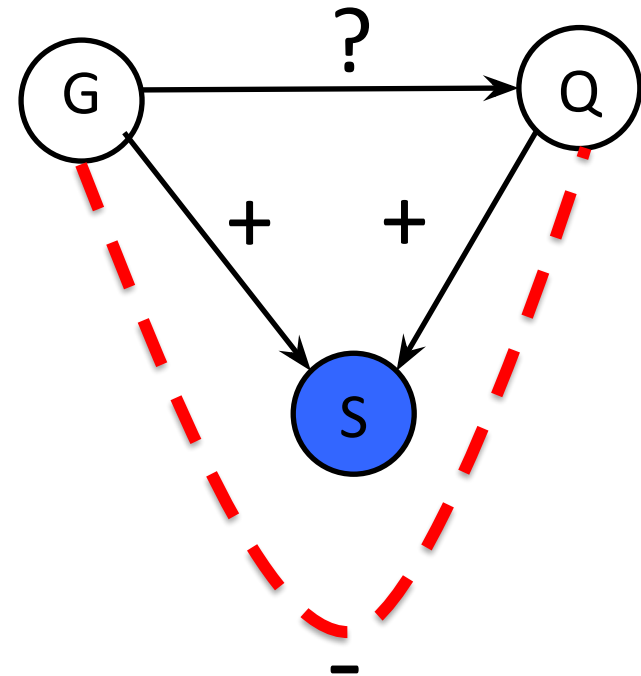
---

Let  $G=1$  for men,  $G=0$  for women,  $Q=1$  for high qualification and  $Q=0$  for low qualification. Suppose two direct effects are positive (hence sex discrimination on salary). Conditioned on  $S$ ,  $G$  and  $Q$  become negatively correlated via the open path in dotted lines.

Men earns a higher salary than equally qualified women



Men are more qualified than equally paid women



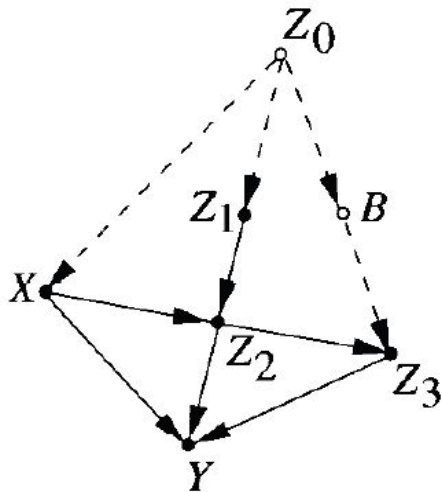
# The Rules of do-calculus

---

- When a query is given in the form of a do-expression, for example  $P(y|\mathbf{do}(x),z)$ , its **identifiability** can be decided systematically using an **algebraic procedure** known as the do-calculus.
  - The **do-calculus** was developed by **J. Pearl in 1995** to facilitate the identification of causal effects in non-parametric models.
  - It consists of **three inference rules** that permits to map interventional and observational distributions whenever certain conditions hold in the causal diagram  $G$ .
  - The do-calculus was shown to be *complete* (Tian and Pearl 2002a; Huang and Valtorta 2006; Shpitser and Pearl 2006; Bareinboim and Pearl 2012a).
-

# Causal graphs: illustration

---



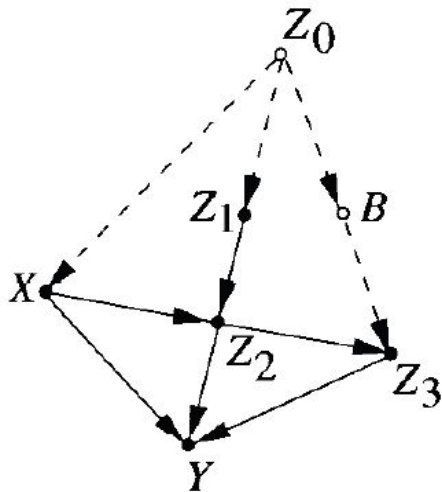
- We wish to assess the **total effect of the fumigants  $X$  on yields  $Y$** .
- The causal diagram represents the investigator's understanding of the major causal influences among measurable quantities in the domain.
- $Z_1$ ,  $Z_2$ ,  $Z_3$  represent the **eelworm population** *before* treatment, *after* treatment, and *at the end* of the season, respectively.
- $Z_0$  represents *last year's* eelworm population.
- $B$  is the population of birds and other predators.

Unmeasured quantities are designated by hollow circles and dashed lines.

---

# The Rules of do-calculus

---



- **Using the do-calculus**, one can establish that the total effect of  $X$  on  $Y$  can be estimated consistently from the observed distribution of  $X, Z_1, Z_2, Z_3$ , and  $Y$ .
- These conclusions are obtained by **performing a sequence of symbolic derivations** (the 3 inference rules).

$$P(y \mid \mathbf{do}(x)) = \sum_{z_1, z_2, z_3} P(y \mid z_2, z_3, x) P(z_2 \mid z_1, x) \\ \times \sum_{x'} P(z_3 \mid z_1, z_2, x') P(z_1, x')$$

---

---

# **Confounding & Selection bias**

---

# Confounding & Selection bias

---

- The biases arising from confounding and selection are fundamentally different, though both constitute threats to the validity of causal inferences.
    - The **confounding bias** is the result of treatment  $X$  and outcome  $Y$  being affected by common ancestral variables,
    - The **selection bias** is due to treatment  $X$  or outcome  $Y$  (or ancestors) affecting the inclusion of the subject in the sample.
  - In both cases, we have extraneous “flow” of information between treatment and outcome, which falls under the rubric of “spurious correlation,” since it is not what we seek to estimate.
  - What are the conditions for **recoverability of interventional distributions** for when selection and confounding biases are both present?
-

# Confounding with latent variables

---

- Some relevant confounders are difficult to measure in many real-world applications (e.g., intention, mood, DNA mutation), which leads to the need of modelling explicitly **latent variables** that affect more than one observed variable in the system (Semi- Markovian models).
  - In such models, identifiability is *not always achievable*.
  - **Causal Effects Identifiability:** Let be  $V$  the set of *observable* variables,  $U$  is the set of *unobservable* variables. The causal effect of an action,  $\text{do}(X = x)$  is said to be identifiable from  $P$  in  $G$  if  $P(y|\text{do}(x))$  is uniquely computable from  $P(v)$ .
  - The evaluation of identifiability goes through a **non-trivial algebraic process**, namely the *do-calculus*.
-

# Confounding : risks and pitfalls

---

- Researchers must weigh the benefit of reducing confounding bias carried by those covariates against the risk of **amplifying residual bias** carried by **unmeasured confounders**.
  - According to Judea Pearl, epidemiologists often adjust for wrong sets of covariate (usually *Sex* and *Age* but other covariates are missing) .
  - Is the prevailing practice in epidemiology misguided?
-

# Selection bias

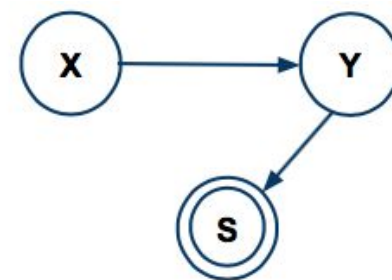
---

- Another major challenge that needs to be addressed when evaluating the effect of interventions is the problem of selection bias, caused by **preferential exclusion of samples** from the data.
  - Selection bias is a **major obstacle to valid causal and statistical inferences**; it can hardly be detected in either experimental or observational studies.
  - **Example**: in a typical study of the effect of training program on earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less.
-

# Selection bias

---

- To illuminate the nature of this bias, consider a variable  $S$  affected by both  $X$  (treatment) and  $Y$  (outcome), indicating entry into the data pool.
- Such preferential selection to the pool **amounts to conditioning on  $S$** , which creates spurious association between  $X$  and  $Y$ .
- Our assumption about the selection mechanism are embodied in an **augmented causal graph  $G_s$** .



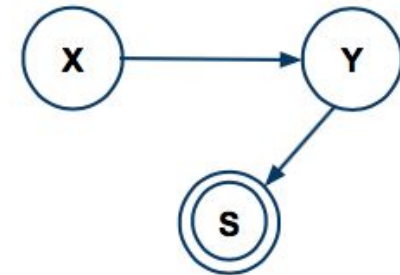
- **Illustration** : Effect of training program on earnings
  - $S$  represents the selection mechanism.  $S=1$  indicates presence in the sample, and  $S=0$  exclusion.
-

# Recoverability

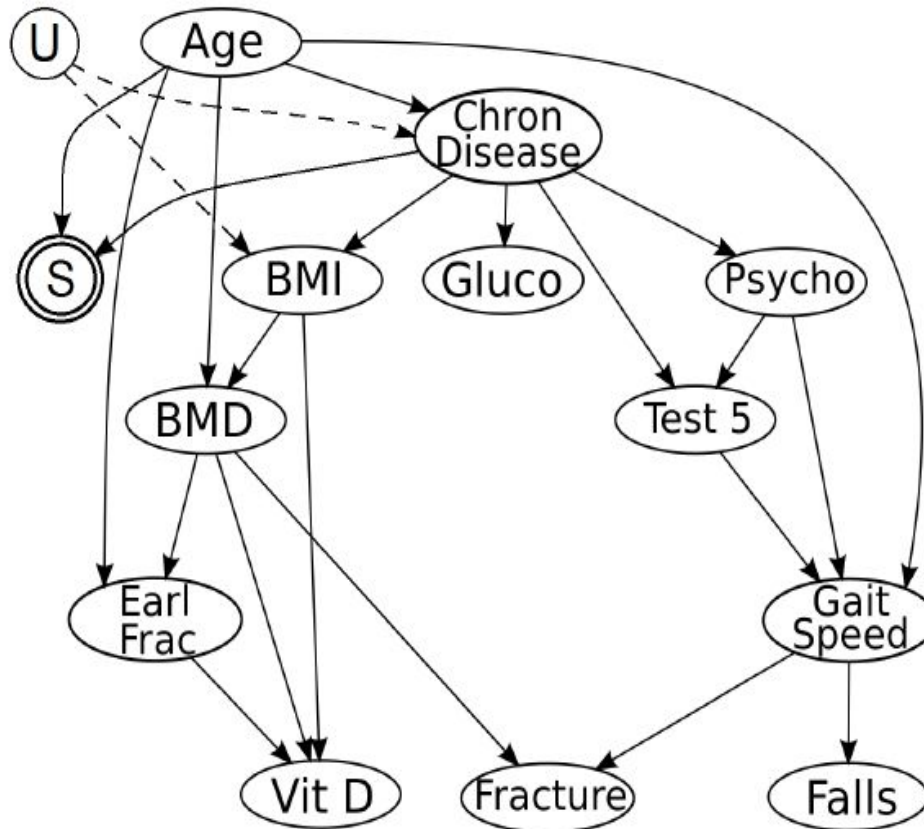
---

- Under what conditions  $P(y|\mathbf{do}(x))$  can be recovered from data drawn from  $P(y, x|S = 1)$ ?
- **Recoverability from Selection Bias:** Given a causal graph  $G_s$  augmented with  $S$ ,  $P(y|\mathbf{do}(x))$  is said to be recoverable from selection biased data in  $G_s$  if  $P(y|\mathbf{do}(x))$  is expressible in terms of the distribution under selection bias  $P(v|S = 1)$ .

- In this example,  $P(y|\mathbf{do}(x))$  is *not* recoverable



# Osteoporotic fracture risk assessment



- Prospective cohort study with 7500 elderly osteoporotic women followed-up during 4 years.
- A *plausible* causal BN was learned from a combination of **non-experimental data** and qualitative assumptions that are deemed likely by health experts.
- Inclusion of a **selection mechanism** and an **unobserved confounder**.
- 
- We seek to estimate the strength of the **causal effect of psychotropic drugs** on the risk of **hip fracture**:

$$P(\text{Fracture} | \text{do}(\text{Psycho})) = ?$$

---

# Missing data

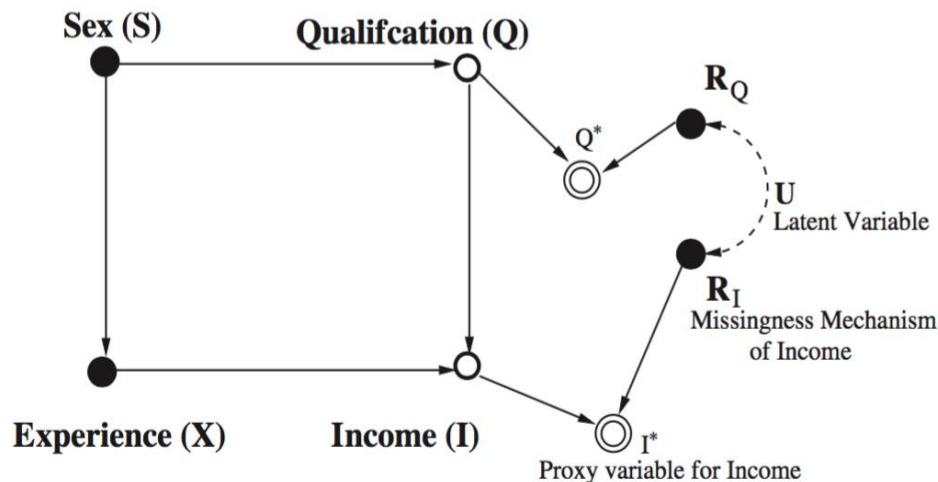
---

# Missing data

---

- All branches of experimental science are plagued by missing data
  - The “missing data” problem arises when values for one or more variables are missing from recorded observations
  - Occurs often in social science, epidemiology, biology and survival data analysis etc.
  - Caused by varied factors such as high cost involved in measuring variables, failure of sensors, reluctance of respondents in answering certain questions
  - **Improper handling of missing data** can bias outcomes and potentially **distort the conclusions** drawn from a study.
-

# Misingness mechanism : *m*-graph

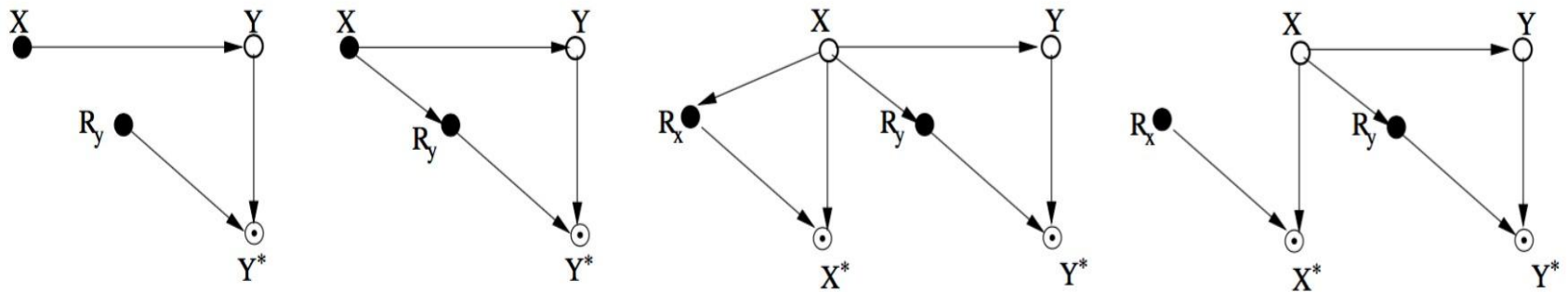


- Associated with every partially observed variable  $V_j \in V_{\text{miss}}$  are **two other variables**  $R_j$  and  $V_j^*$
- $V_j^*$  is a proxy variable that is actually observed.
- $R_j$  represents the status of the causal mechanism responsible for the missingness of  $V_j^*$

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases}$$

Observed and partially missing variables are represented by full and hollow circles respectively.

# Missingness mechanisms



- $V_j^*$  is a proxy variable that is actually observed, and  $R_i$  represents the status of the **causal mechanism responsible for the missingness** of  $V_j$ .
- Data that are: (a) MCAR, (b) MAR, (c) & (d) MNAR. Hollow and solid circles denote partially and fully observed variables respectively

# Recoverability with missing data

---

Let  $V_{\text{obs}}, V_{\text{miss}}$  be the set of observed and missing variables

- **Recoverability from Data Missingness Bias:** Given a causal graph  $G$  augmented with the missingness variables  $R$ ,  $P(y|\mathbf{do}(x))$  is said to be recoverable in  $G$  if  $P(y|\mathbf{do}(x))$  is expressible in terms of the distributions  $P(V_{\text{obs}}, V_{\text{miss}} | R = 0)$ .

# Conclusions

---

- Testing for cause and effect is difficult, discovering cause effect is even more difficult.
  - But, once the **causal diagram** is provided (both from *expert knowledge* and data), identification of causal effects is straightforward using the *do-calculus* rules.
  - Many **paradoxes** and **controversies** in social and medical sciences can be illustrated and understood by simple graphical means.
  - The **data missingness** and **selection mechanisms** can easily be represented in the diagram for **bias correction** purposes.
  - **Inference of causal relationships from massive data sets** is still a challenge but may eventually lead to new discoveries (e.g. cancer)
-

# References

---

- J. Pearl. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press, 2009.
  - J. Pearl "Understanding Simpson's Paradox", UCLA Cognitive Systems Laboratory, Tech. Rep. R-414, 2013.
  - J. Pearl "Do-Calculus Revisited" UCLA Cognitive Systems Laboratory, Conference on Uncertainty in Artificial Intelligence (UAI) 2012.
  - S. Lauritzen. *Graphical Models*. Clarendon Press: Oxford, 1996.
  - J. Pearl "Myth, confusion, and science in causal analysis", UCLA Cognitive Systems Laboratory, Tech. Rep. R-348, 2009.
  - J. Pearl "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates", Proceedings of UAI, 417-424, 2012.
  - E. Bareinboim and J. Tian. "Recovering Causal Effects from Selection Bias", Proc. AAAI 2015.
  - J.A. Myers et al. « Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. of Epidemiology* 2011.
  - A. Goldberger. "Reverse regression and salary discrimination". *The Journal of Human Resources* 1984.
  - J. Berkson. "Limitations of the application of fourfold table analysis to hospital data ». *Biometrics Bulletin* 1946.
  - P. Spirtes, et al. *Causation, Prediction and Search*, MIT Press, 1993.
  - A.P. Dawid: *Fundamentals of Statistical Causality*. Research Report No. 279, University College London, 2007.
  - M. Kalisch et P. Bühlmann. *Causality: A selective review, Quality Technology & Quantitative*, 2014
  - K. Mohan et al. "Missing Data as a Causal Inference Problem". NIPS, 2013
-

---

Thank you for your attention, any question ?

---