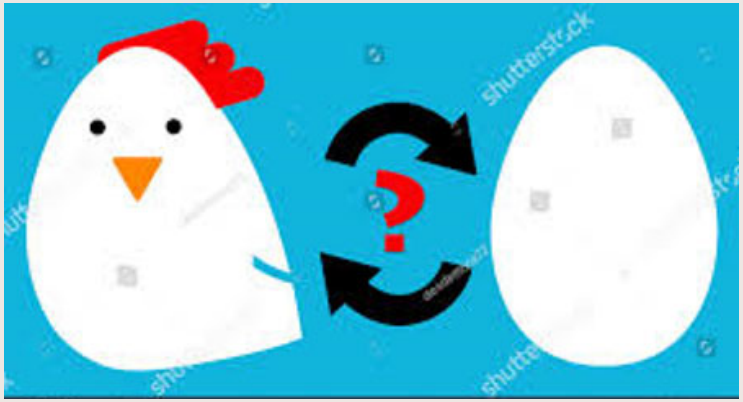
A little journey through Causality

Philippe Bastien

L'Oréal R&I (philippe.bastien@loreal.com)



Hélio SPIR

L'ORÉAL Recherche & Innovation

The causal revolution

Yoshua Bengio *(Turing Award 2018)* *"**deep learning is blind to cause and effect**. Deep learning needs to start asking **why things happen**."*

What are the most important statistical ideas of the past 50 years?*

Andrew Gelman† and Aki Vehtari‡

29 Nov 2020

nature COMMUNICATIONS

Q&A

https://doi.org/10.1038/s41467-021-25743-9 OPEN

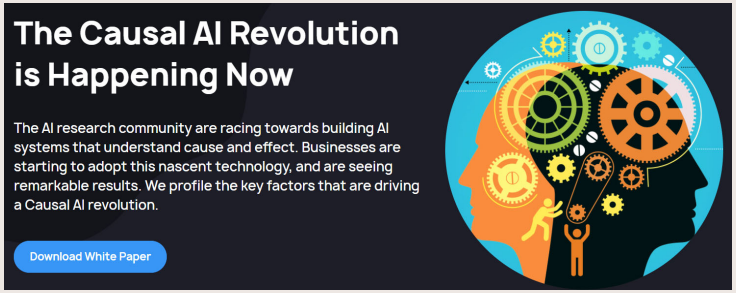
Causality in digital medicine

Causal inference and the data-fusion problem

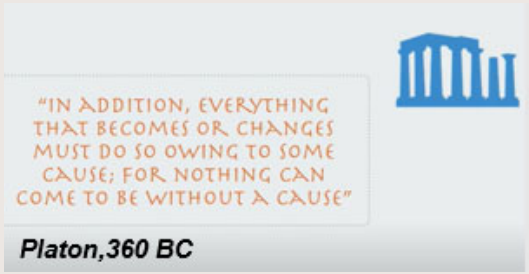
Elias Bareinboim[a,b,1] and Judea Pearl[a]

Transportability of Causal Effects: Completeness Results*

Elias Bareinboim and Judea Pearl
Cognitive Systems Laboratory
Department of Computer Science
University of California, Los Angeles



Current AI approaches often result both in false positives, identifying drivers that are not predictive; and in false negatives, failing to identify predictive drivers. In contrast, Causal AI technology is specifically designed to uncover the true causal relationships in data.



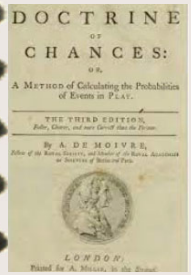
Causality a new concept ?

Surprisingly today, causality can be seen as an emerging field in science.

People may be surprised to hear that causality has been anathema to statisticians for the longest time. Asking the questions in terms of causation until recently could even be considered **unscientific**.

Among many others:

*"**Considerations of causality should be treated as they always have been treated in statistics, preferably not at all**…" (Speed, 1990).*


DOCTRINE
OF
CHANCES:
OR,
A METHOD of Calculating the Probabilities of Events in PLAY.
THE THIRD EDITION,
By A. DE MOIVRE,
LONDON:

Three main reasons

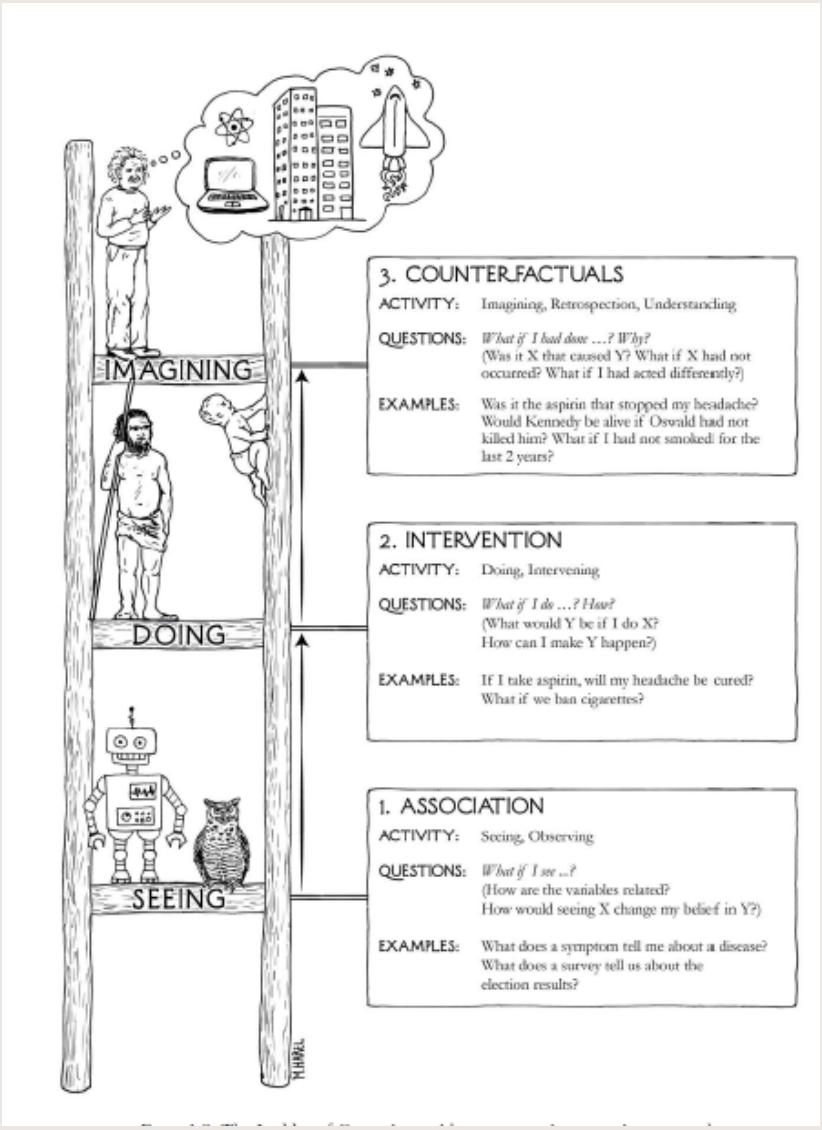
This is due to various reasons, in particular:

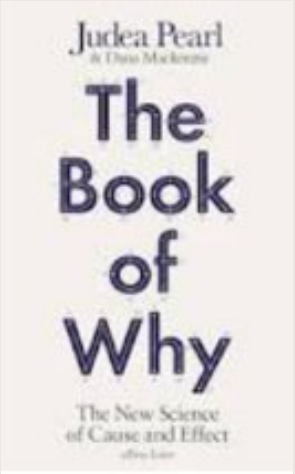
- There was **no mathematical language** associated with causality. In particular, the probability does not integrate the notion of causality.

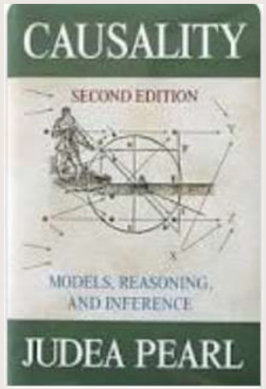
- The use of **algebraic equations** introduced by Copernicus in the 17th century in astronomy and then generalized to all sciences, has been a brake on the expression of causal relations due to their **symmetry**.

- The notion of causality, pre-existing to the development of probability theory (Abraham de Moivre's Doctrine of Chance, 1718) and statistics, almost disappeared as a specific concept at the end of the 19th century, with the appearance of the notion of **correlation**.

The ladder of causation







Statistics and Causality


Ronald Aylmer Fisher – ...

Ronald Fisher (1922):
« **The object of statistical methods is the reduction of data** »

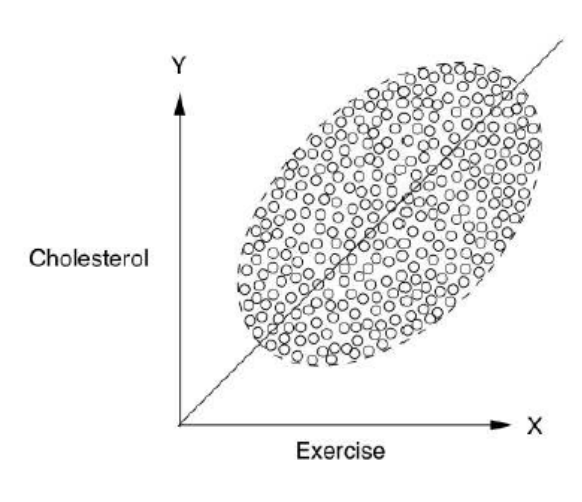
Traditional methods are oriented towards inference: finding a parsimonious mathematical description of the joint distribution of a set of variables of interest.

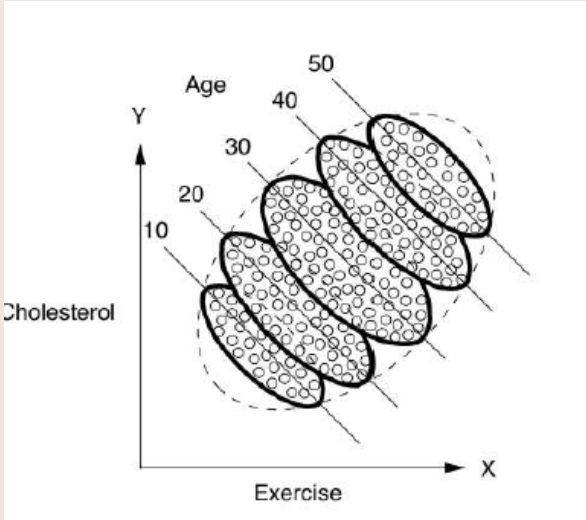
This is a **description** of the data and **not the process responsible for the data**. **Yet the fondamental question** at the core of statistical inference **is causal**: do change in one variable cause change in another? and how much change do they cause ?

Causation is an enrichment of Statistics to uncover part of the world that traditional methods cannot approached.

Simpson's paradox

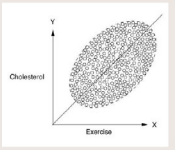
(Edward Simpson 1951)

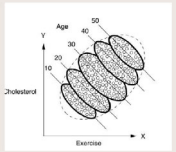




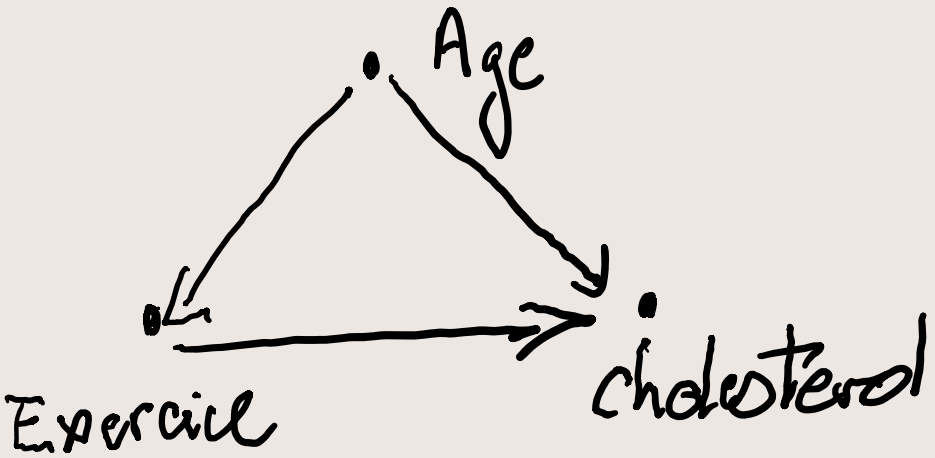
Simpson's paradox alerts us to cases where at least one of the statistical trends in the aggregated data or the partitioned data **cannot represent the causal effects**.

Simpson's paradox

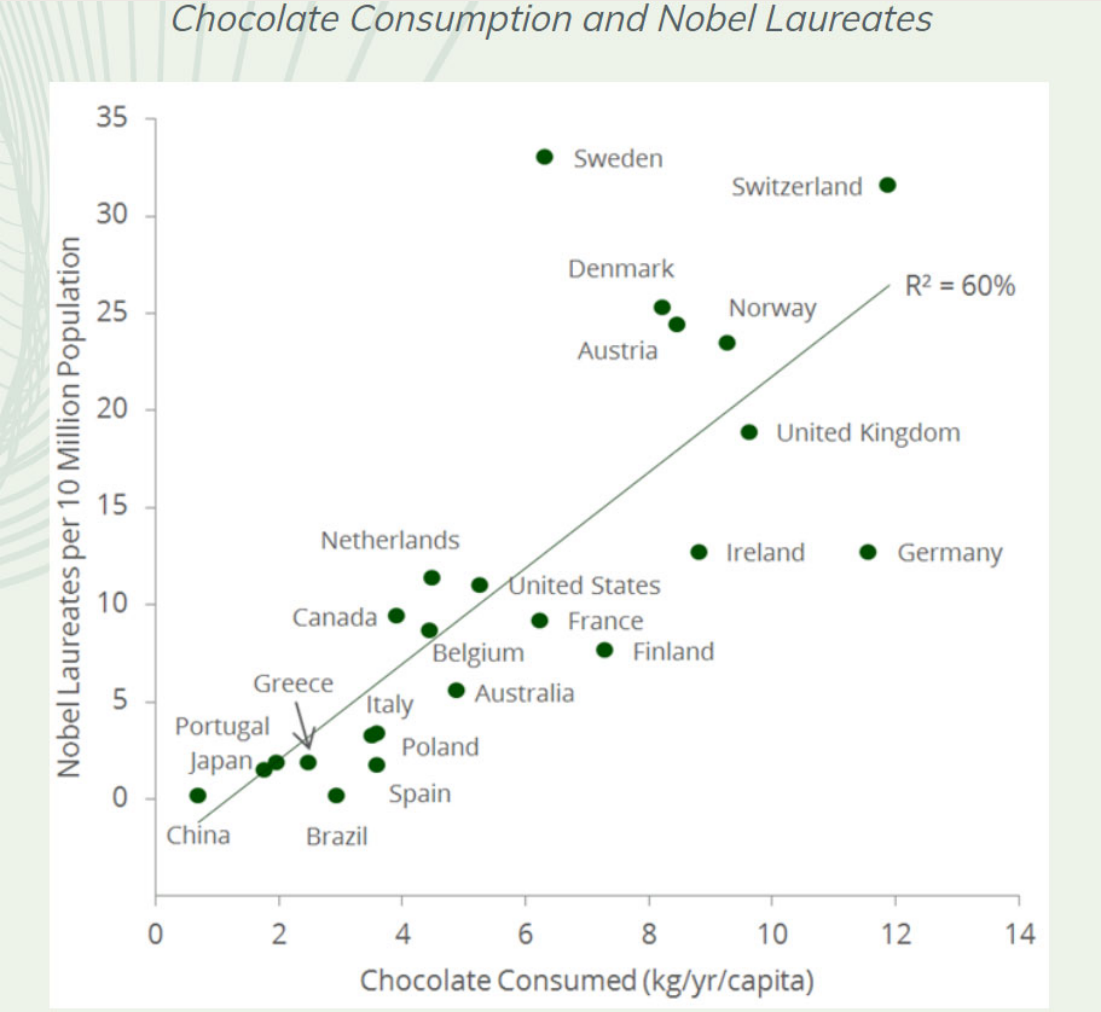


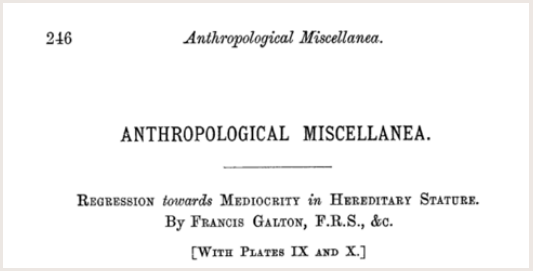


It is a **paradox only for statisticians or people** who are trained in "conventional" methodology **without causal lenses**.



Spurious association





246 *Anthropological Miscellanea.*

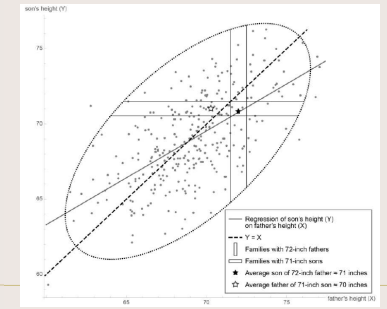
ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.
By FRANCIS GALTON, F.R.S., &c.
[WITH PLATES IX AND X.]

Francis Galton

As an irony of history, Francis Galton, statistician, anthropologist, meteorologist, sociologist,…, inventor of fingerprints, cousin of Charles Darwin, in 1888 seeking a genetic explanation (causality) for the phenomenon of regression towards the mean has put in evidence the notion of **correlation** (co-related) which is devoid of causality.

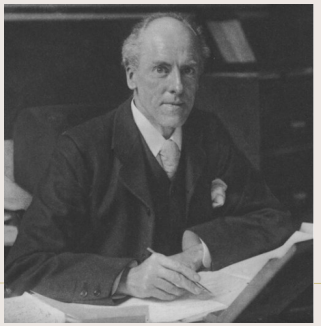
He observed that tall parents have tall children but rather shorter and conversely tall children have tall but rather shorter parents. He called it "**regression towards mediocrity**"



co-relation

In 1888, Francis Galton measured the length of a person's forearm and the size of that person's head and asked to what degree one of these quantities can predict the other.

He made the following discovery: If you plot one quantity against the other and scale the two axes properly, then **the slope of the best-fit line has some nice mathematical properties**. The slope is 1 only when one quantity can predict the other precisely; it is zero whenever the prediction is no better than a random guess; and, most remarkably the slope is the same no matter if you plot X against Y or Y against X.



Karl Pearson (1900)

Galton's discovery dazzled one of his disciples, Karl Pearson, now considered to be one of the founder of modern statistics. Pearson (the father of Chi², PCA,…) mathematically formulated the correlation coefficient.

For Pearson there was a category broader than causation, namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine, and sociology in large parts into the field of mathematical treatment.

It is therefore not necessary for Pearson to speak specifically of causality, correlation encompasses this notion.

Judea Pearl and the new causal revolution

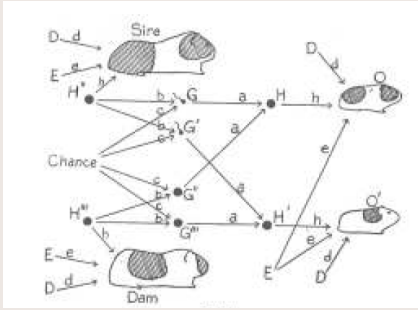
- Potential outcomes : Neyman (1923) & Rubin (1974)

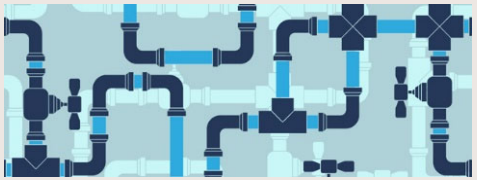
- Fisher : Randomization in experimental design (Statistical Methods for Research Workers 1925)

- Structural Causal Models (SCM) with Directed Acyclic Graphs (DAG) : Sewall Wright (1920)

- Structural Equation Modeling (SEM): Haavelmo & Wold (1960)

- Do calculus / SCM with DAG : Judea Pearl (1990)

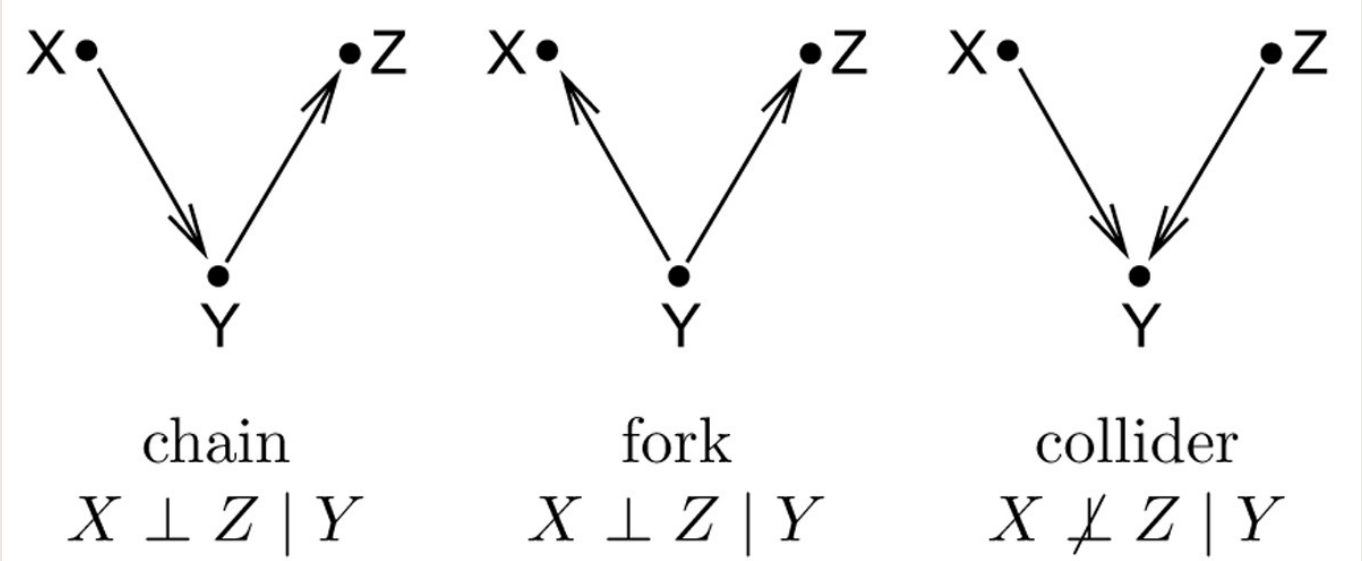




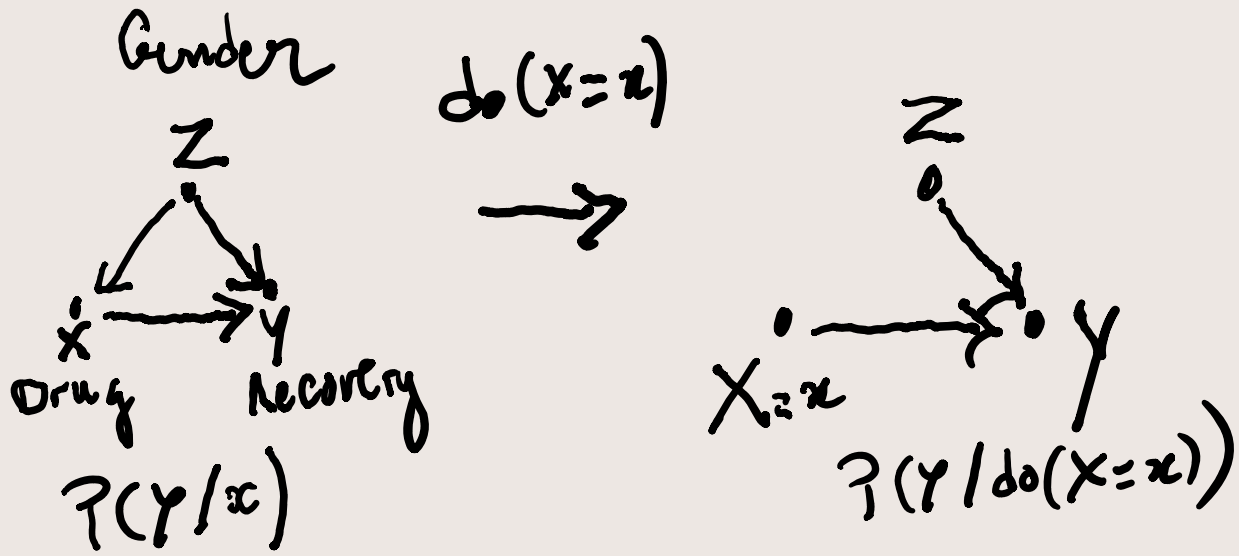
Directed Acyclic Graph

The DAG will allow us to **visually determine the causal relationships** and determine if they are **identifiable** i.e. if we can **estimate them from observable data without having to intervene** (e.g. randomized trial) and how.

only 3 types of connection



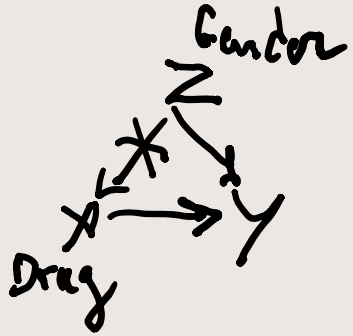
Intervention in a causal diagram and arrow deletion



Spurious correlation / Paradox when P(Y=y/do(X=x)) ≠ P(Y=y/ X=x)

How effective is the drug ?
The back-door criterion

Causal Effect = P(Y=1 / do(X=1)) - P(Y=1 / do(X=0))



P(Y=y / do(X=x)) = ?

P(Y=y / do(X=x)) = $\sum_z \boldsymbol{P(Y = y/X = x, Z = z)P(Z = z)}$

The adjustment formula instruct us to ignore the aggregated population data P(Y=1/X=1) and P(Y=1/X=0) from which we might falsely conclude on the overall drug effect.

The BC identifies which variables in a causal diagram are **deconfounders.** It allows to make predictions about the results of an intervention **without** performing it.

Propensity Score

Given a DAG G in which a set of variables Pa are designated as the parents of X, the causal effect of X on Y is given by :

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

Where z range over all the combinations of values that the variables in Pa can take. If we multiply and divide by P(X=x/Pa=z) we get:

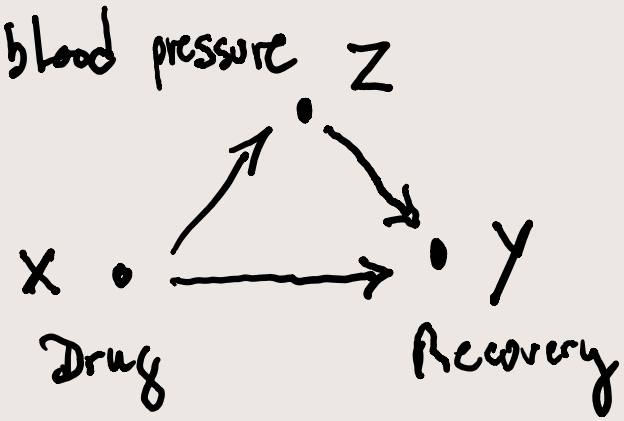
$$P(Y=y \,/\, do(X=x)) = \sum_z \frac{P(X=x,Y=y,Pa=z)}{P(X=x \,/Pa=z)}$$

$\boldsymbol{P(X = x \,/Pa = z)}$ is known as **the "Propensity Score"**

the "Propensity Score" weights the distribution of non-experimental data to correct for the bias associated with Z

Conditioning or not conditioning

We must not condition on all variables !!



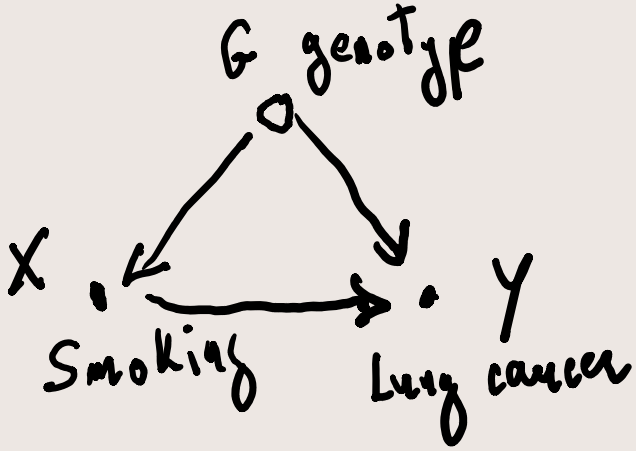
Here X must be treated as a randomized treatment (no arrows entering X, X has no parents)

P(Y=y/do(X=x)) = P(Y=y/X=x)

Adjusting on Z would produce an **incorrect assessment**, **blocking** the **indirect effect** of the drug **mediated** by blood pressure

The front-door criterion

The do operator can be applied to graphical pattern that do not specify the BC. One such pattern is the FC.

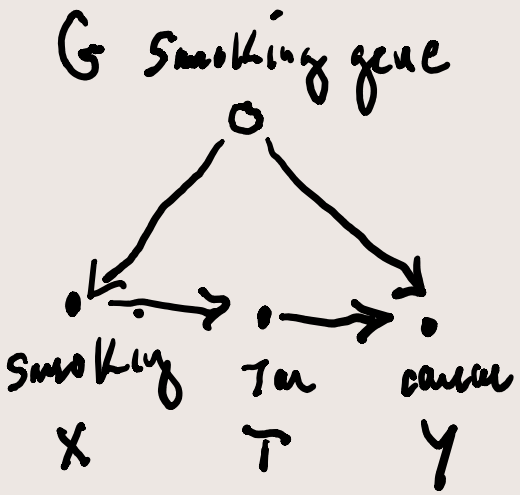


Theory of the tobacoo industry (1970)
Lung cancer could be explained by some carcinogene genotype that also induce an inborn craving for nicotine.

The graph does not satisfy the BC because **G is not observed** and hence cannot be used to block the back-door path between X and Y. Therefore the **causal effect of smoking on lung cancer is not identifiable** in this model. One can never ascertain which portion of the observed correlation between X and Y is spuriously attribuable to their common effect G, and what portion is genuinely causative.

The front-door criterion

However we can consider the next model where an additional measurement is available: the amount of tar deposit in patient lungs.



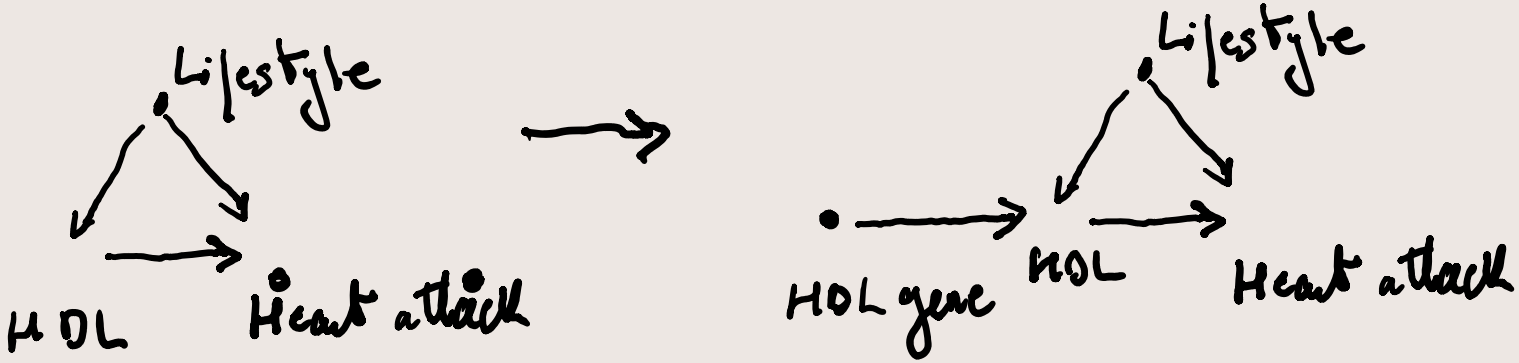
This model does not satisfy the BC but the causal effect P(Y=y/do(X=x)) is nevertheless identifiable.

Hypotheses:
-no link between smoking gene and tar deposit
-smoking leads to cancer only through the accumulation of tar deposit
-no direct path between smoking and cancer

Fisher hypothesis: A smoking gene confounds smoking behavior and lung cancer

Instrumental variables: Mendelian randomization

Lifestyle not measurable so no possible back-door



Hypothesis: HDL has a protective effect against heart attacks

Suppose there is a gene that caused people to have higher HDL levels (with no effect on LDL supposed to be the « bad » cholesterol)

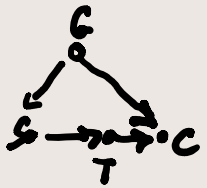
Our genes are randomized at the time of conception: **Mendelian randomization**

The do-calculus

The do-calculus uncovers all causal effects that can be identified from a given graph beyond the BC or FC. From only **3 rules** we can determine **when a "do" quantity can be reduced to a "see" quantity.**

Either we can apply the procedure and find ourselves in possession of the **causal effect without to intervene.** Otherwise, we would at least know that the assumptions imbedded in the model are **not sufficient to uncover the causal effect from observational data** and no matter how clever we are, there is no escape from running an interventional experiment to some kind.

The Front-door by the do-calculus



$P(C/do(S)) = \sum_t P(C/do(s),t)P(t/do(s))$ proba axiom

$= \sum_t P(C/do(s),do(t))P(t/do(s))$ Rule 2

T→C causal car pas de BD grâce à do(s)

$= \sum_t P(C/do(s),do(t))P(t/s)$ Rule 2

S→T causal car pas de BD grâce au collider en C

$= \sum_t P(C/do(t))P(t/s)$ Rule 3

Si on intervient sur T l'intervention sur S n'a pas d'utilité

$= \sum_{s'} \sum_t P(C/do(t),s')P(s'/do(t))P(t/s)$ proba axiom

$= \sum_{s'} \sum_t P(C/t,s')P(s'/do(t))P(t/s)$ Rule 2

Car S→T causal, une fois S fixé, T est défini

$= \sum_{s'} \sum_t P(C/t,s')P(s'))P(t/s)$ Rule 3

Car S→T causal et S avant T

Counterfactual

Had Cleopatra's nose been shorter, the whole face of the world would have changed.
BLAISE PASCAL (1669)

$\mathbf{E(Y_{X=treated}\,/X = control, Y = tumor\ recurred)}$

Two worlds: $\boldsymbol{X = treated\ and\ X = control}$ **and hence cannot be expressed as a do-expression** which means that it **cannot be estimated from interventional experiments.**

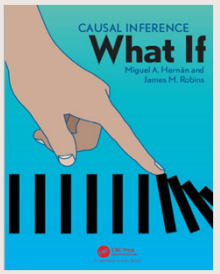
Average treatment effect of treatment on the treated

$$\text{ATT} = \text{E}(Y_{X=treated}/X = treated) - E(Y_{X=control}/X = treated)$$

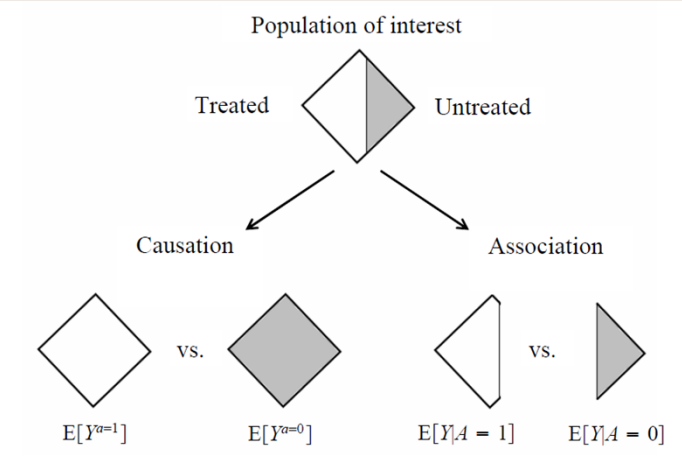
$$E(Y_{X=treated}/X = treated) = \text{E}(\text{Y}/X = treated)$$

$$E(Y_{X=control}/X = treated) = ?$$

Potential outcomes /counterfactuals



- Neyman-Rubin causal model
  - Ignorability /exchangeability



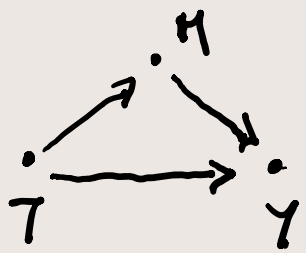
The Fundamental Problem of Causal Inference[a]

[a] Paul W Holland. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

It is impossible to observe the value of $Y_i(1)$ and $Y_i(0)$ for the same unit. Therefore, it is impossible to observe $Y_i(1) - Y_i(0)$.

- Pearl (2000) proposed an alternative to express counterfactuals through SCM/DAG.

Mediation



The **Natural direct effect** (NDE)

$= \mathrm{E}(Y_{T=1,M=M_0} - Y_{T=0,M=M_0})$

$= \mathrm{P}(Y_{M=M_0} = 1\ /do(T=1)) - \mathrm{P}(Y_{M=M_0} = 1\ /do(T=0))$

NDE measures the expected increase in Y as the treatment change from T=0 to T=1, while the mediator is set to whatever value it would have attained (for each individual) prior to the change, that is under T=0.

The **Natural indirect effect** (NIE)

$= \mathrm{E}(Y_{T=0,M_1} - Y_{T=0,M_0})$

$= \mathrm{P}(Y_{M=M_1} = 1\ /do(T=0)) - \mathrm{P}(Y_{M=M_0} = 1\ /do(T=0))$

NIE measures the expected increase in Y when the treatment is held constant, at T=0, and M changes to whatever value it would have attained (for each individual) under T=1. It captures, therefore, the portion of the effect that can be explained by mediation alone, while disabling the capacity of Y to respond to T.

Probabilities of Causation (legal and scientific probabilities)

Probability of necessity

$PN = P(Y_{X=0} = 0 \,/\, X = 1, Y = 1)$

PN stands for the probability that event y would not occured in the absence of event x, given that x and y did in fact occur.

Probability of sufficiency

$PS = P(Y_{X=1} = 1 / X = 0, Y = 0)$

PS stands for the probability that setting x would produce y in a situation where x and y are absent.

THANK YOU