Identifying Intended Effects with Causal Models

Dario Compagno (Université Paris Nanterre)

Aim

The aim of this work is to extend the framework of causal inference, in particular as it has been developed by Pearl (2009), in order to model actions and their ends.

It introduces means-ends relationships as a way to model actions based on a causal model and of its implied correlations observable in data; it defines *teleological confounding* and introduces *interference* as a way to control for it.

Motivation

Causal inference is about events. If I put a coin into a coffee machine, as a result I will obtain a cup of espresso. This is the kind of explanation offered by causal models: the coin-event works as a cause of the coffee-event. But what if I wanted to understand the reasons why somebody puts a coin into the machine, that is, for what ends some action is realized?

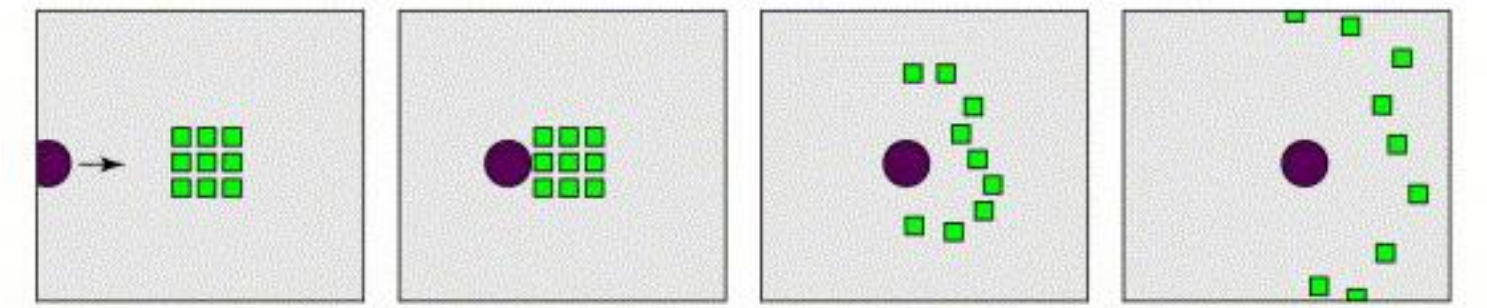
To know the effects of some event is not enough to characterize it as action. In fact, from any event descends an indefinite number of effects, and not all of them are intended. If I put a coin in the coffee-machine, I get coffee, but I also become one coin poorer, and I waste some resources (water, electricity). It would be a mistake to see all effects as reasons for action: I do not intentionally aim at becoming poorer, nor at wasting resources.

Explaining actions needs to answer a particular kind of counterfactual question: if the machine did not produce waste, would the agent keep putting coins into it? (yes); if the machine did not produce coffee, would the agent keep putting coins into it? (no). There is a need to learn how to differentiate intended and unintended effects from empirical data, starting from a causal model and an hypothesis of intentional behavior. Identifying intended effects can then be done by controlling for some of the action's effects–and not for its causes, as it is done in usual causal inference.

Plan

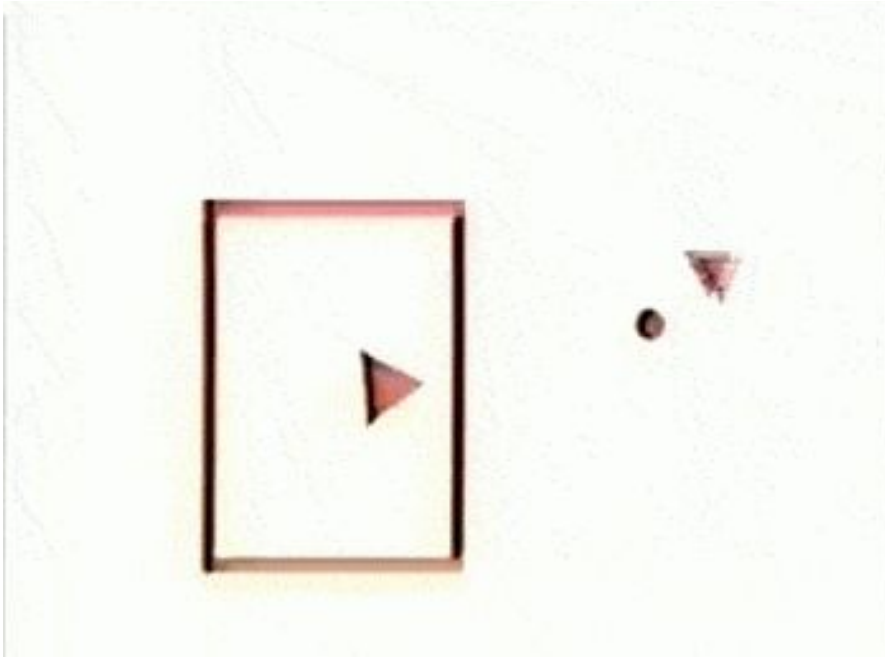
- What is teleological inference
  - Causal and teleological inferences
  - Events and actions
  - Counterfactual definition of causes and intentions
- Identifying means-ends relationships
  - The “fundamental problem” of teleological inference
- Teleological confounding
  - Three kinds of teleological confounding
- Interference
  - Controlling for teleological confounding
- Why intentions cannot be reduced to latent causal variables
  - Intentions as unobservable causes
  - Intentions as repeated measurements

What is teleological inference



Causal inference: is A the cause of B?

BJ Scholl & PD Tremoulet, “Perceptual causality and animacy”, *Trends in cognitive science*, 4/8, 2000.



Teleological inference: is B the end of doing A?

F Heider & M Simmel, “Experimental study of apparent behavior”, *The American Journal of Psychology*, 57/2, 1944 (animation).

Causal and teleological inferences both demand to go beyond correlations in data

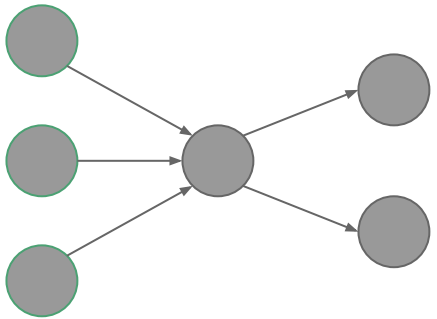
Only correlations are observable; causal models are needed to interpret data causally or teleologically.

Causal models

A causal model is a set of assumptions about how the value of a variable depends on the value of other variables (its causes).

A causal model can be represented as a non-cyclic graph in which nodes are variables, and arcs are causal links between them.

A causal model formalizes causal hypotheses and permits to validate them with observational data (i.e. without performing experiments).

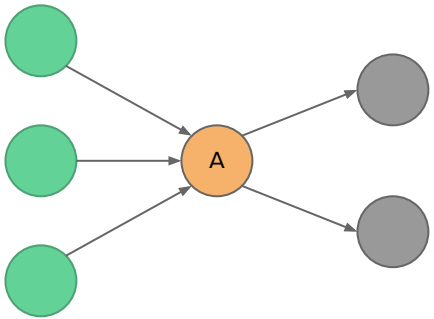


Events and actions

Events

An event is the assignment of a value to a variable which “listens” to ***all*** other variables linked to it in a causal model.

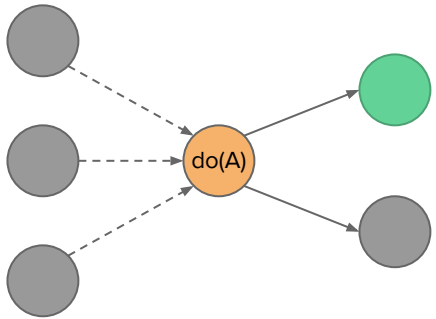
The value of *A* depends on the values of all the variables having arcs going into *A*.



Actions

An action *do(A)* is the assignment of a value to a variable which “listens” to ***some*** of the variables to which it is linked to.

The value of *do(A)* depends on the values of some of the variables having arcs coming from *do(A)*.



Actions are not events

Example: observing that the stove is on is not the same thing as turning it on.

do(A) means that the value of A is observed under specific conditions (i.e. controlling for its antecedents).

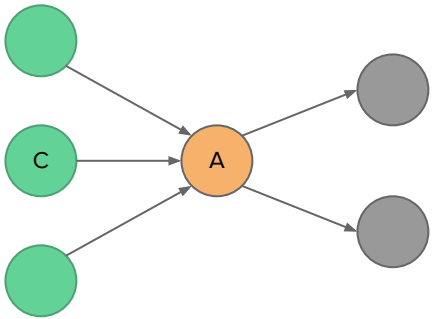
Example: I turn the stove on at random times and observe if the water boils only when the stove is on.

Counterfactual definition of causes and intentions

Counterfactual definition of causes

What are the causes of *A* ?

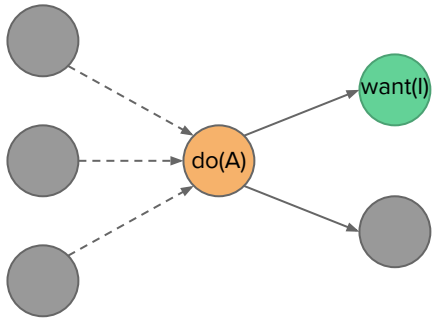
“Had *C* did not happen, then *A* would not have happened.”



Counterfactual definition of intentions

What are the ends of *do(A)* ?

“Was *I* not an intended effect of *do(A)*, then *do(A)* would not have happened.”

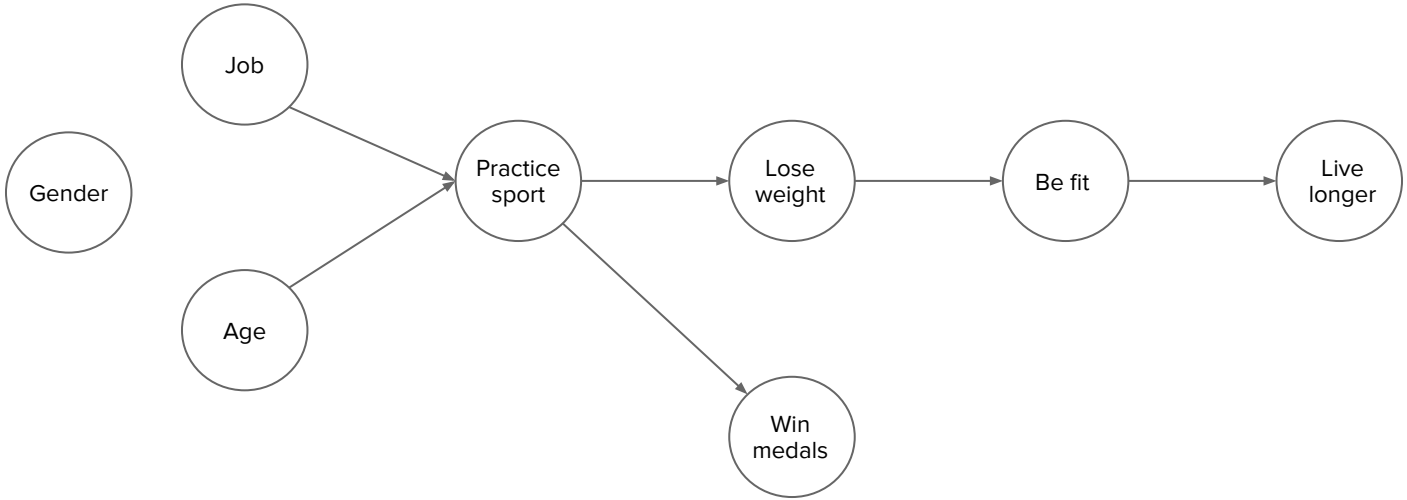


Not all the effects of an action are intended

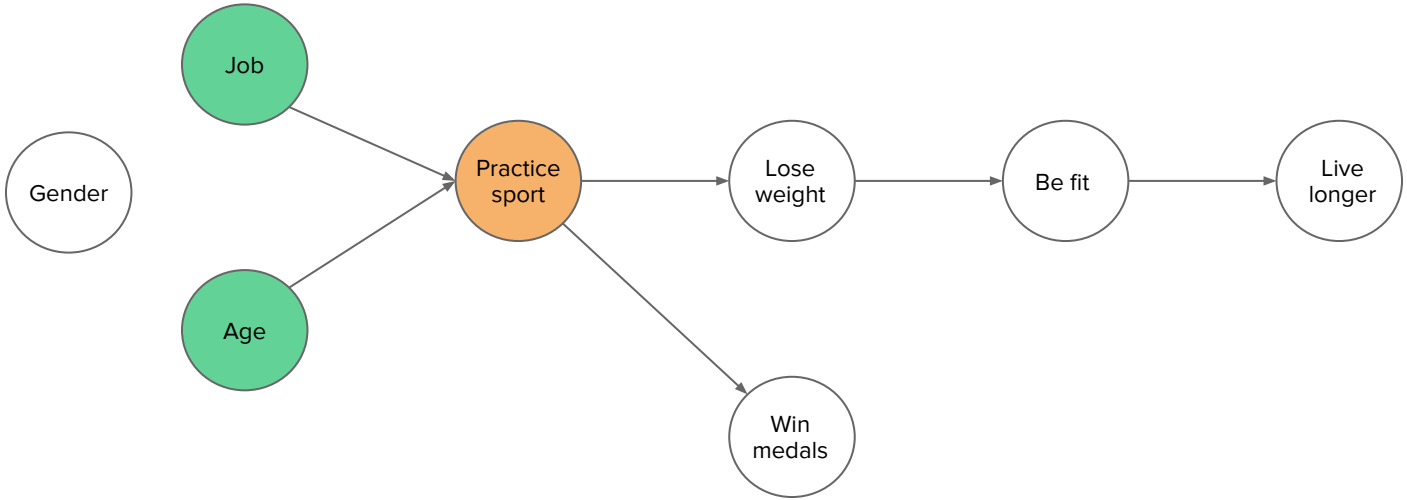
Example: planes travel fast and pollutes, but I do not take a plane in order to pollute.

want(I) means that the value of A is observed under specific conditions (i.e. controlling for its consequences).

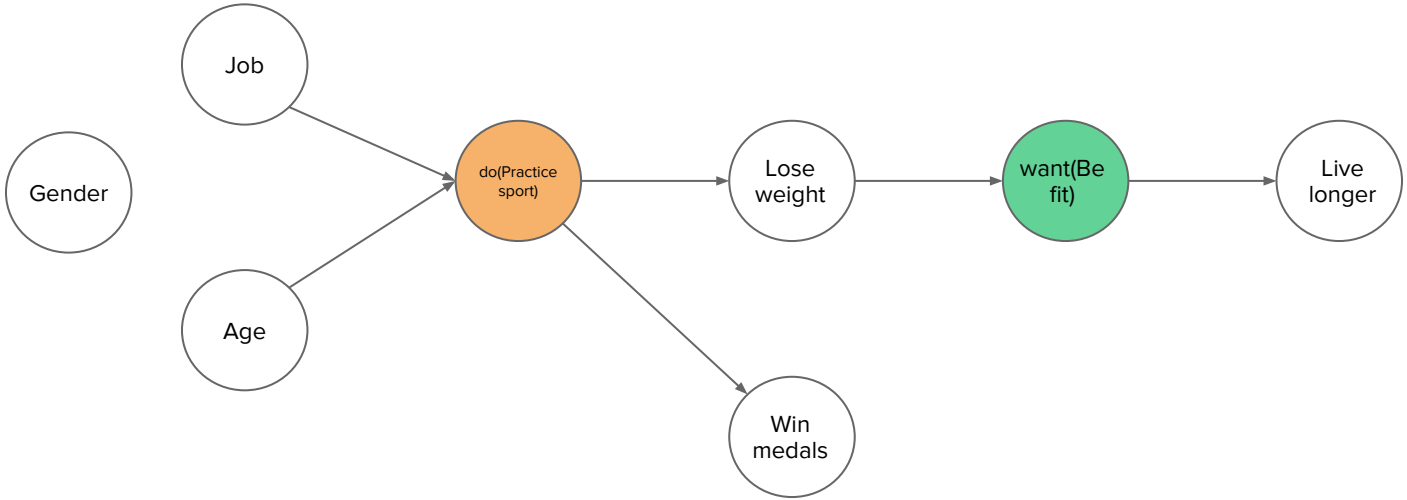
Example: I expect people to fly on fast non-polluting planes and not to fly on slow polluting planes.



Example of causal model



Which are the causes of practicing sport?
(Job and Age: old people and people without a job do not practice sport)



Which is the end of practicing sport?
(Be fit: people who do not want to be fit do not practice sport)

Causal inference and teleological inference

Causal Inference

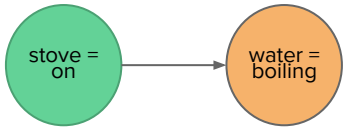
The aim of causal inference is to identify a causal model.

A causal model is a set of variables and causal relationships among them.

“The water is boiling *because* the stove is hot.”

water “listens to” stove

stove = on ➔ water = boiling



Teleological inference

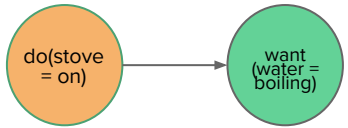
The aim of teleological inference is to identify a means-ends relationship within a causal model.

A means-ends relationship associates an intervention to some intended effects.

“The stove is hot *in order to* make the water boil.”

do(stove) “listens to” want(water)

do(stove = on) ➔ want(water = boiling)



Actions are done in order to obtain some effects, but this does not mean that causes follow from their effects!

Example: boiling water does not turn the stove on by itself, but if I had no water to boil, I would not turn the stove on.

do(stove) “listens to” want(water)
is not the same thing as
stove “listens to” water

Identifying means-ends relationships

Causal models can be identified with data, starting from combinations of values that should not be observed

Example: if I observe old people practicing sport, I conclude that my assumption Job → Practice Sport is false.

Job	Sport
0	0
1	1

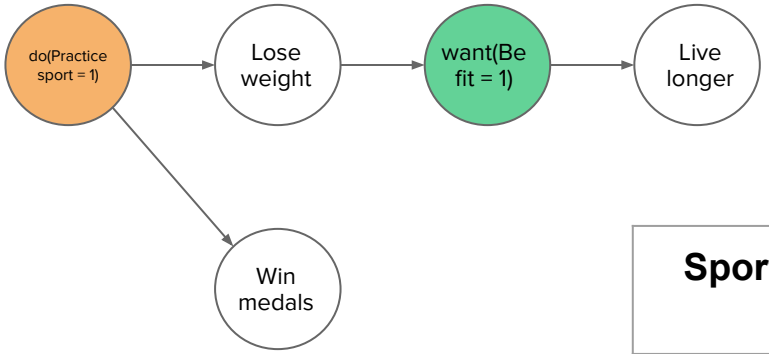
Combinations of values compatible with the causal model on the left
(If I do not have a job I do not do sport, if I have a job a do sport)

Job	Sport
0	0
1	1
1	0
0	1

The combinations of values in red are incompatible with the causal model on the left

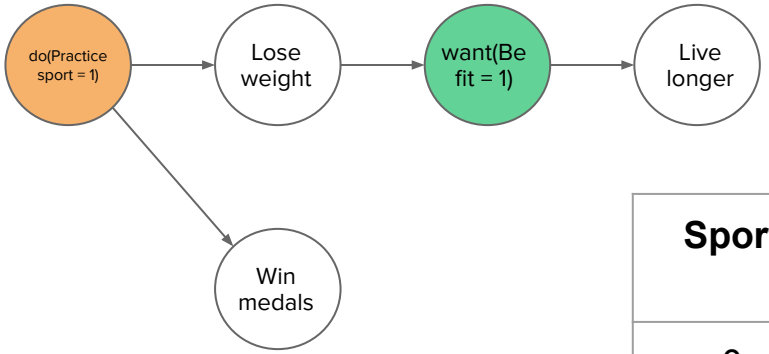
It would be great if means-ends relationships could be identified by looking for combinations of values that should not be observed

Example: if I observed that people stop practicing sport if they do not live longer by doing so, I would conclude that my hypothesis do(Practice sport) ➜ want(Be fit) is false.



Sport	Win medals	Lose weight	Be fit	Live longer
0	0	0	1	1
1	1	1	1	1
0	0	0	0	0

Combinations of values compatible with the means-ends relationship on the left (Either I am already fit and so I do not practice sport, or I practice sport in order to become fit, or I do not want to be fit and so I do not practice sport)



Sport	Win medals	Lose weight	Be fit	Live longer
0	0	0	1	1
1	1	1	1	1
0	0	0	0	0
1	1	0	0	0
1	1	1	0	0

The combinations of values in red are incompatible with the means-ends relationship on the left
(But they are also incompatible with the grounding causal model and can *never* be observed)

However, the combinations of values that should not be observed in order to identify any means-ends relationship are all incompatible with the grounding causal models

Example: if I practice sport, I cannot just get fitter and not also live longer. Becoming fitter and living longer are both effects of practicing sport; if this wasn't the case the causal assumption Be fit ➔ Live longer would simply be false.

The “fundamental problem” of teleological inference

What could be called the “fundamental problem” of teleological inference is that any event A causes **all** of its effects, but only **some** of the effects of do(A) are intended.

So it is not possible to simply observe which effect of do(A) was intended, given that all the effects of A co-occur.

Teleological confounding

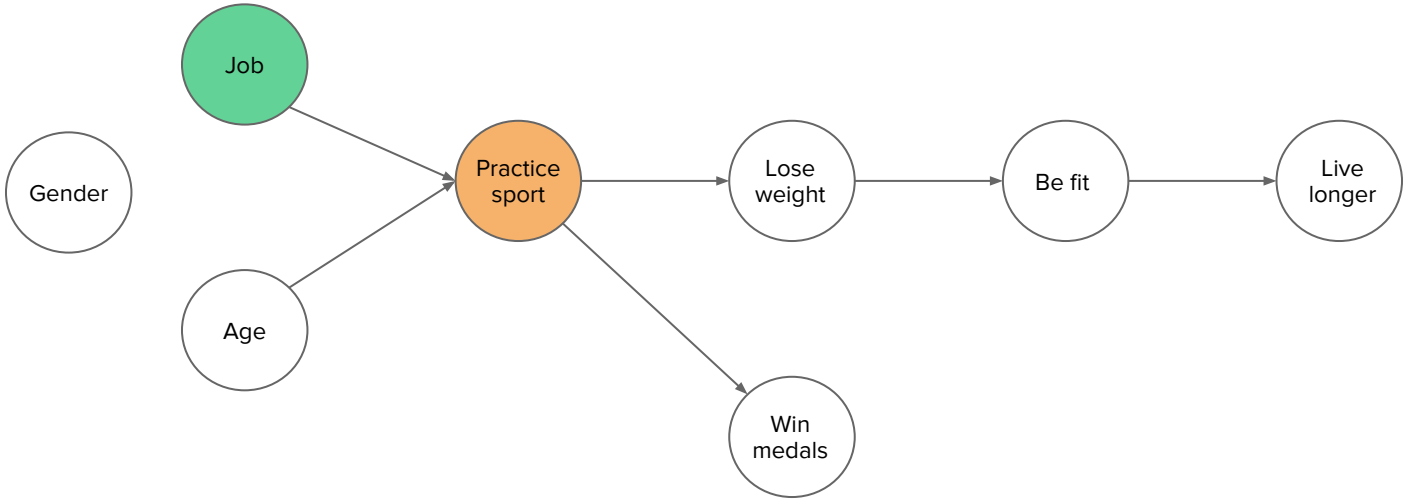
Teleological confounding

We need a way to control for those effects which co-occur with the intended one but are not intended per se–i.e. *teleological confounders*.

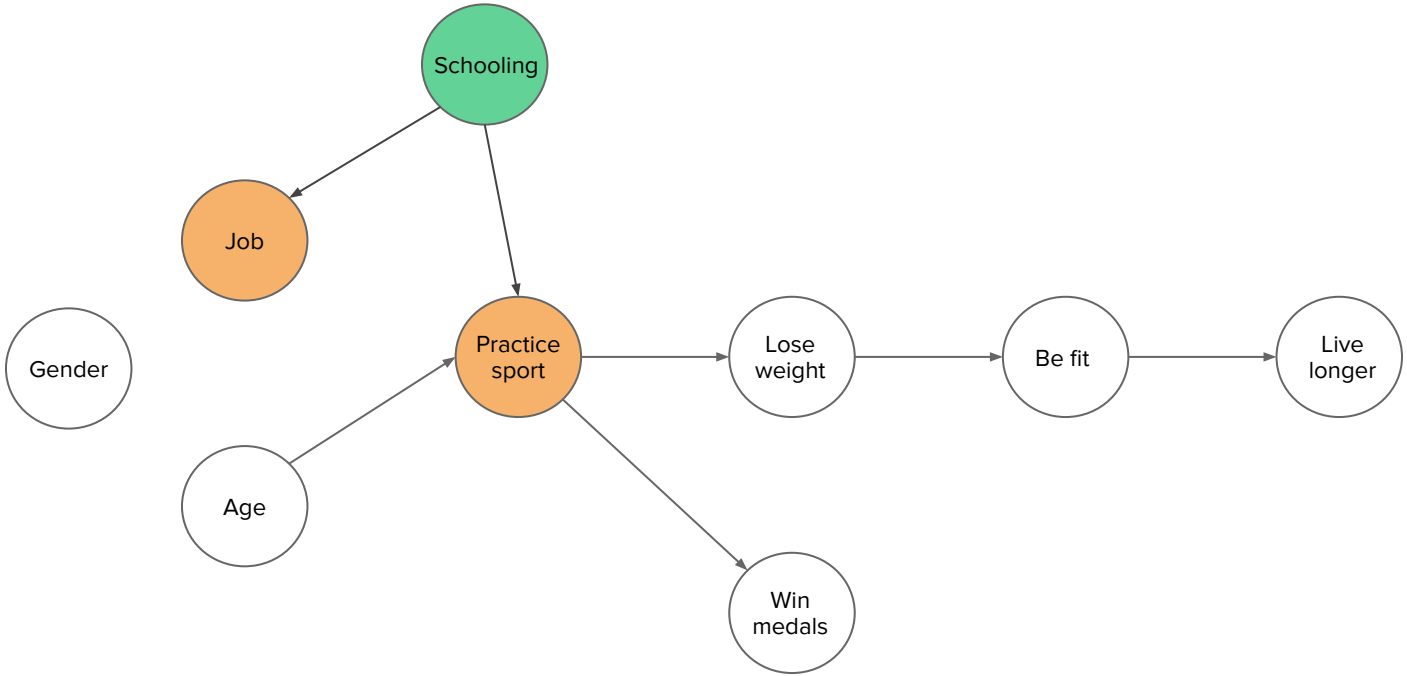
Example: planes travel fast and pollute, but I do not take a plane in order to pollute. If I could travel fast without polluting, would I do it? (Yes) If I could pollute without travelling fast, would I do it? (No) Polluting is a confounder for teleological inference, while traveling fast is the intended effect.

A causal confounder is a variable explaining differently some observable correlation (a common cause)

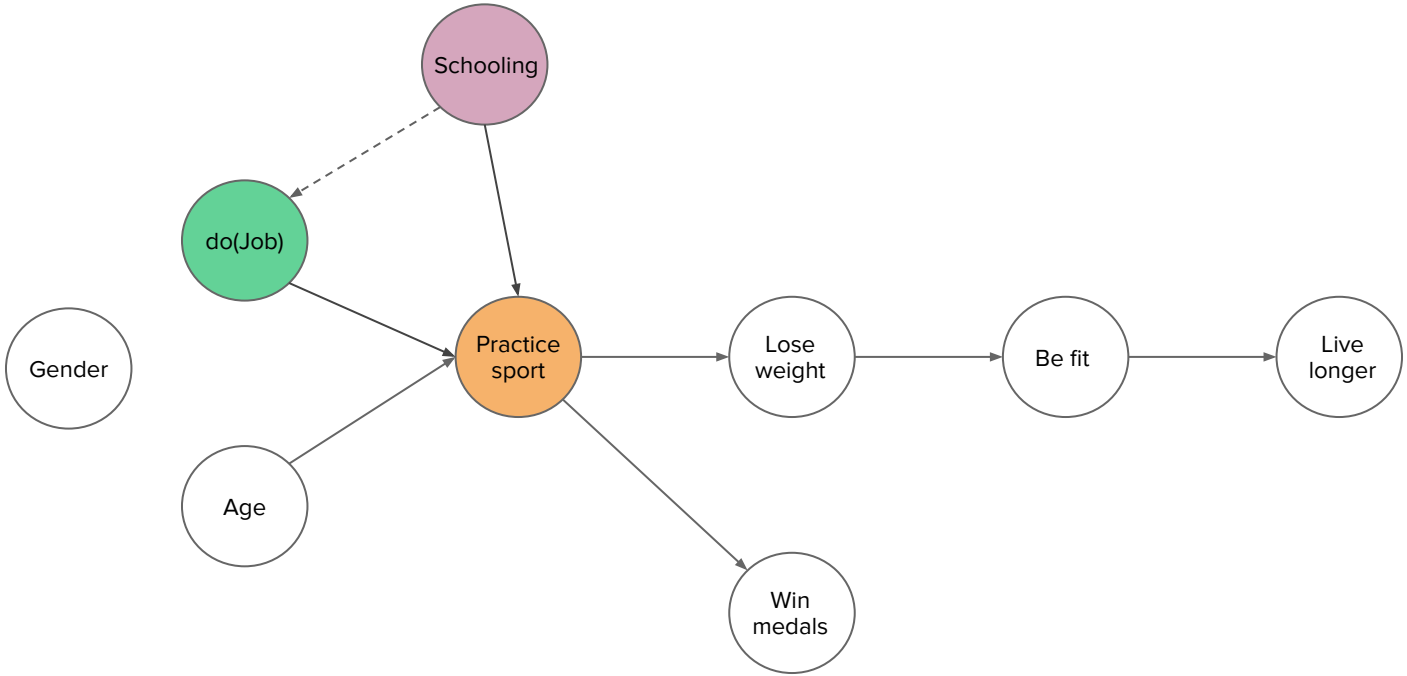
It can be controlled for easily.



Authentic causal relationship linking Job to Sport



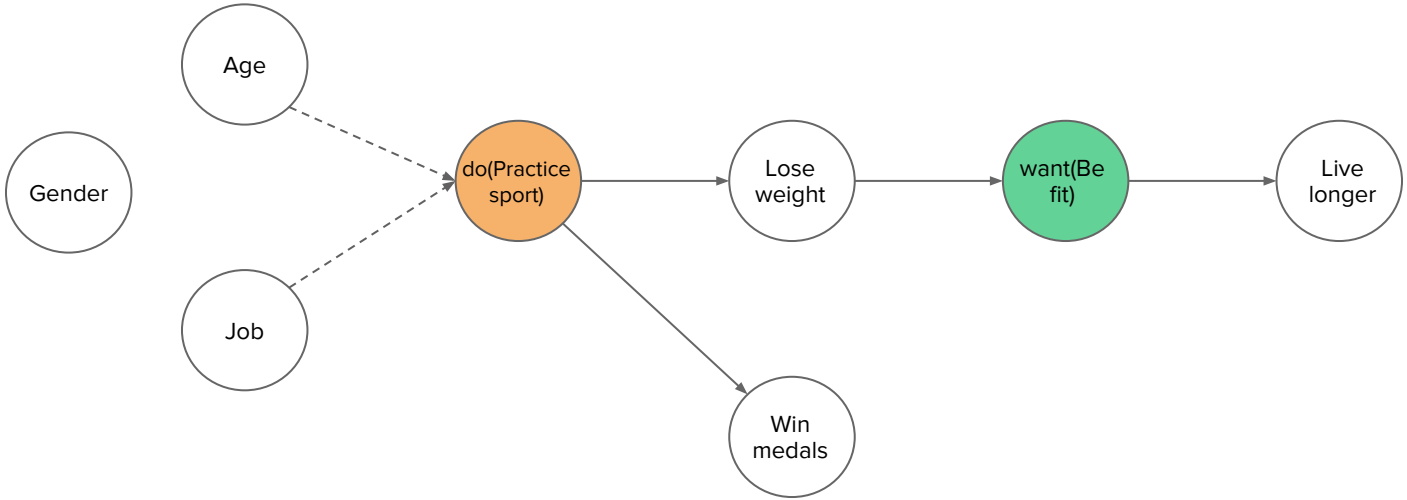
Causal confounding: here Schooling is a common cause of Job and Sport, producing a spurious correlation between the two



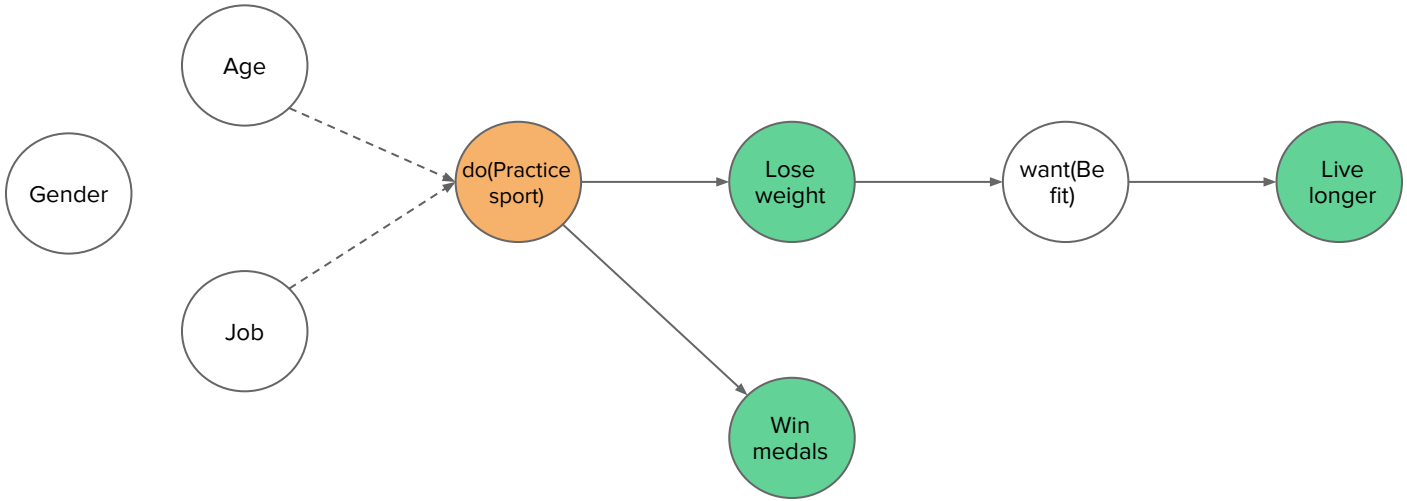
Controlling for causal confounding permits to validate the causal link going from Job to Sport (the variable Schooling is kept constant)

A teleological confounder is also a variable explaining differently some observable correlation (another effect)

It cannot be controlled for as easily.



Causal model grounding the means-ends hypothesis “Sport is done in order to Be fit”



Teleological confounding: Sport may in reality be due to Lose weight, Live longer or Win medals
(My hypothesis that want(Be fit) is the intended effect would therefore be wrong)

Three kinds of teleological confounders

1) **Further effects**

   The true intended effect is a descendent of *I*
   “They actually practice sport to live longer”

2) **Parallel effects**

   The true intended effect is a descendent of *do(A)* unrelated to *I*
   “They actually practice sport to win medals”

3) **Mediating effects**

   The true intended effect is an event mediating between *do(A)* and *I*
   “They actually practice sport to lose weight”

Interference

One cannot simply "erase" some effects of an event, but one can control for them via ***interference***

Example: all planes pollute, but I may compensate for the emissions–would people fly more? (Yes) All planes travel fast, but I could make trips artificially longer–would people still fly? (No)

Interference

Interference is a sort of intervention based on the assumption that one variable in the causal model is an action “listening to” some of its effects; the effects of such action are interfered with.

Example: if I prevent athletes from winning medals, I expect to observe a significant reduction in their sport practice **only if** I assume that practicing sport (cause) is an action oriented towards the end of winning medals (effect).

An ordinary intervention would instead prevent people from practicing sport (cause) and observe if they win less medals (effect).

Controlling for teleological confounders

1) **Further effects**

   The true intended effect is a descendent of *I*
   “Would they KEEP practicing sport even if they did not live longer?”

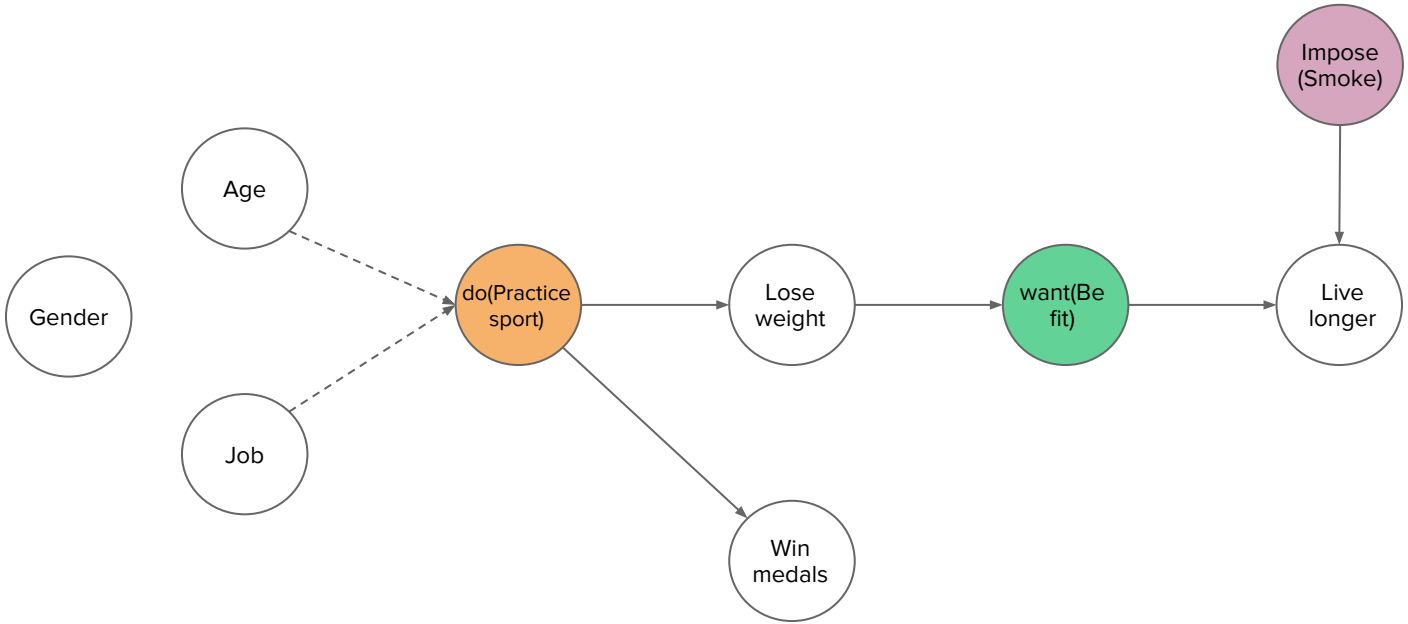
2) **Parallel effects**

   The true intended effect is a descendent of *do(A)* unrelated to *I*
   “Would they KEEP practicing sport even if they did not win medals?”

3) **Mediating effects**

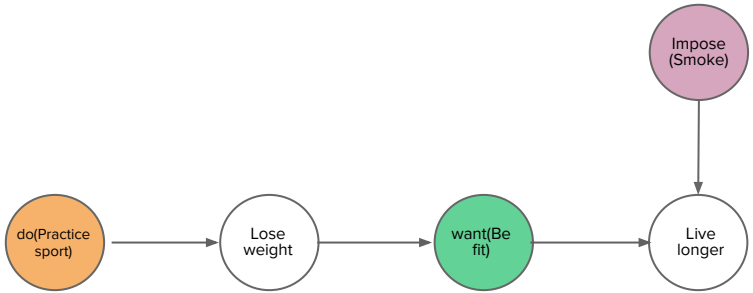
   The true intended effect is an event mediating between *do(A)* and *I*
   “Would they STOP practicing sport in case they did not become fitter?”

Controlling for further effects



Do people KEEP practicing sport even if they do not live longer by doing so?
If want(Be fit) is the correct means-ends relationship, I expect no significant change in Practice sport even if I Impose(Smoke)

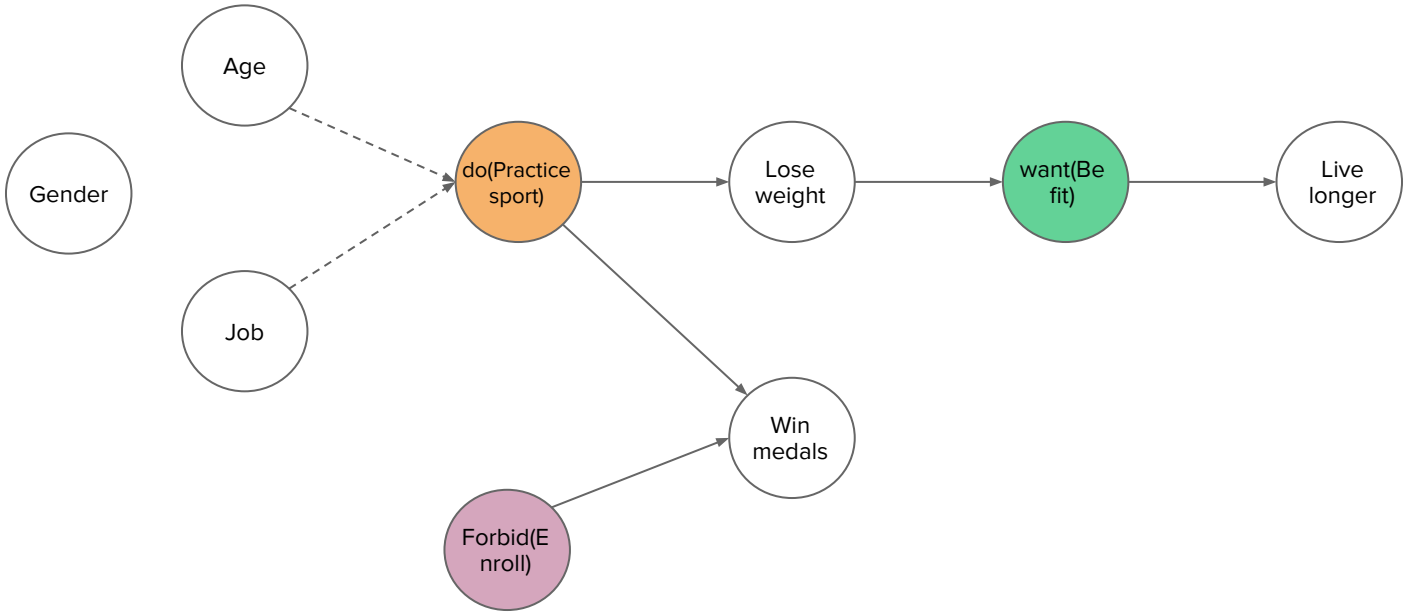
Controlling for further effects



Sport	Lose weight	Be fit	Live longer	Smoke
0	0	1	1	0
1	1	1	1	0
1	1	1	0	1
0	0	0	0	0
0	0	0	0	1

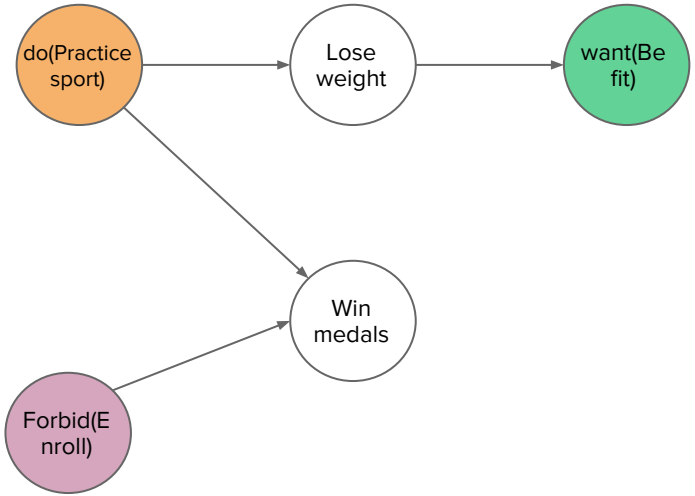
If the hypothesis want(Be fit) is correct, I must observe the line in green.
(I expect people to keep practicing sport even if they cannot live longer because of smoking.)

Controlling for parallel effects



Do people KEEP practicing sport even if they cannot win medals?
If want(Be fit) is the correct means-ends relationship, I expect no significant change in Practice sport even if I Forbid(Enroll)

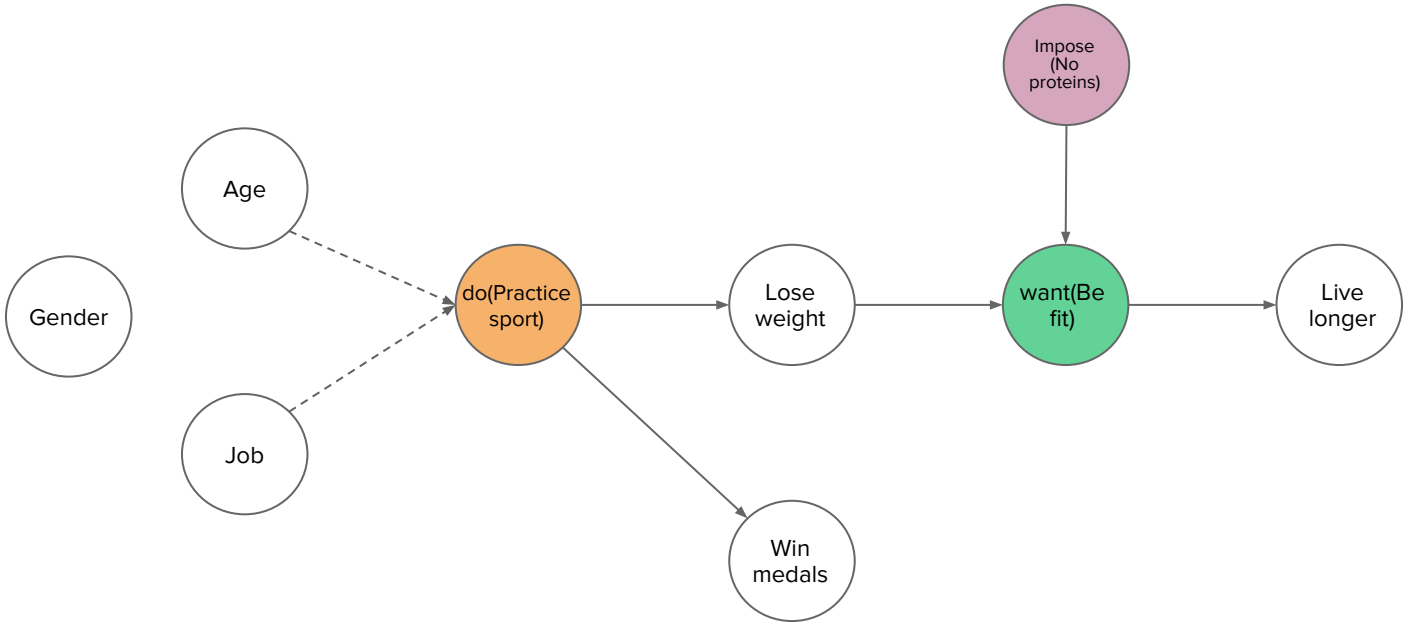
Controlling for parallel effects



Sport	Lose weight	Be fit	Win medals	Enroll
0	0	0	0	0
1	1	1	1	1
1	1	1	0	0
0	0	0	0	0
0	0	1	0	0
1	1	1	1	1

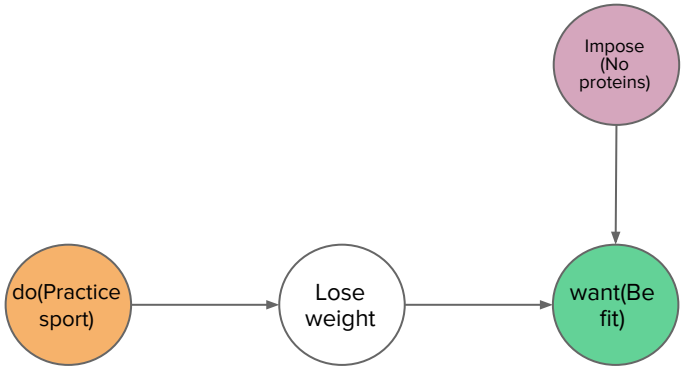
If the hypothesis want(Be fit) is correct, I must observe the line in green.
(I expect people to keep practicing sport even if they cannot win medals because they cannot enroll in competitions.)

Controlling for mediating effects



Do people STOP practicing sport if they are prevented from getting fit?
If want(Be fit) is the correct means-ends relationship, I expect a significant reduction in Practice sport if I Impose(No proteins)

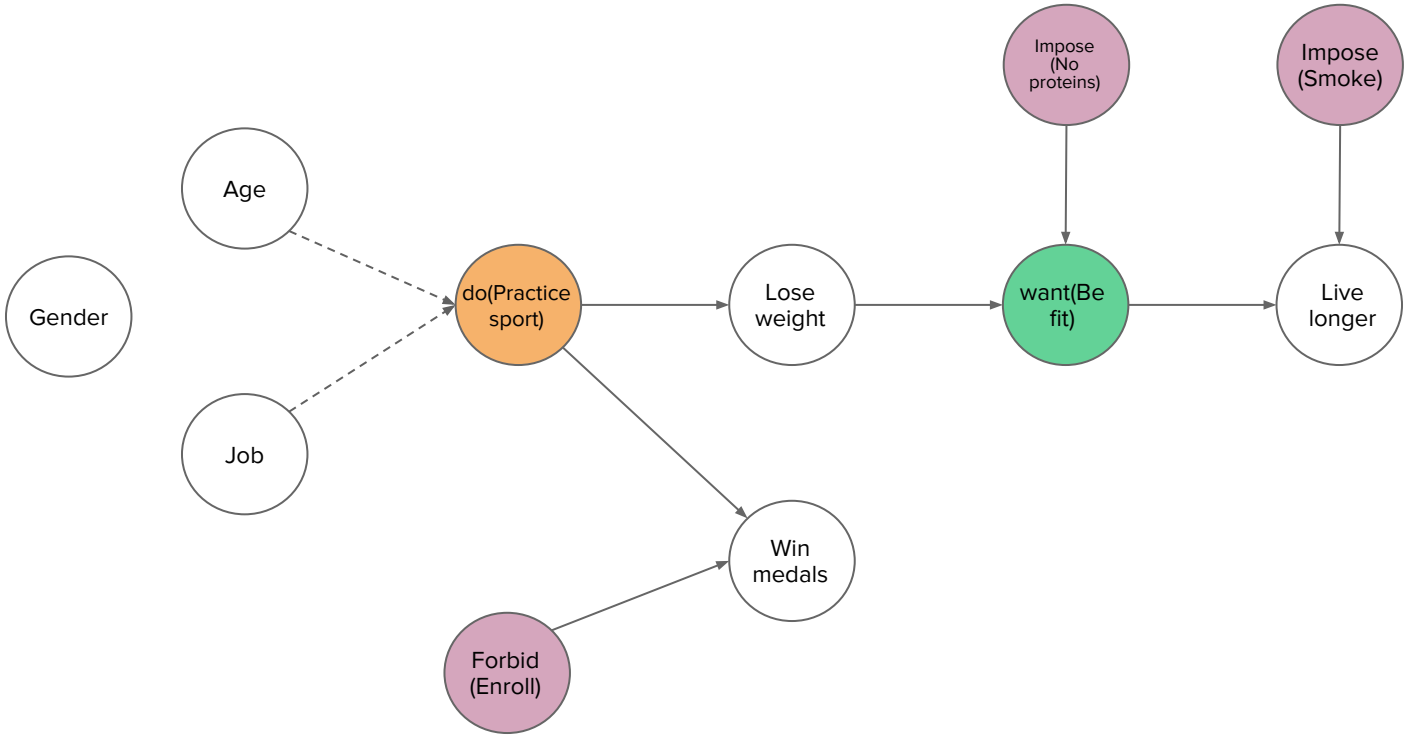
Controlling for mediating effects



Sport	Lose weight	Be fit	Proteins
0	0	0	0
1	1	1	1
1	1	0	0
0	0	0	0
0	0	1	0

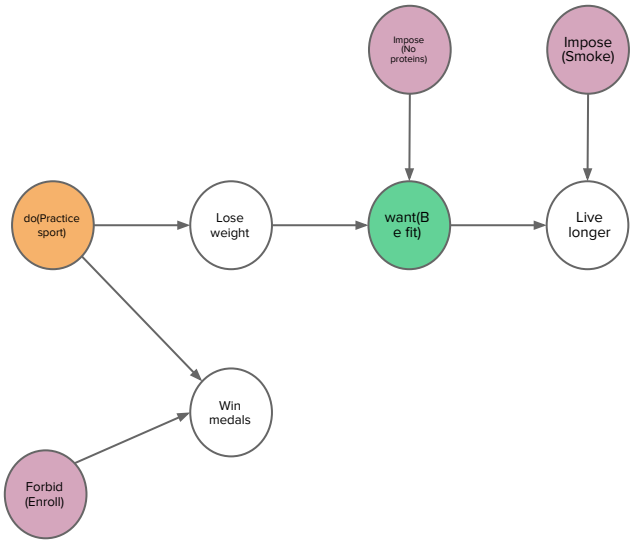
If the hypothesis want(Be fit) is correct, I must not observe the line in red.
(I expect people to stop practicing sport if they cannot become fitter because of a strict no proteins diet.)

Controlling for all teleological confounding



Controlling for all teleological confounding of the means-ends relationship do(Practice sport) ➔ want(Be fit)

Controlling for all teleological confounding



Sport	Lose weight	Be fit	Live longer	Win medals	Smoke	Enroll	Proteins
1	1	1	0	0	1	0	1
1	1	0	0				0

If the hypothesis want(Be fit) is correct, I need to observe the combination of values in green and not to observe the one in red

Why intentions cannot be reduced to latent causal variables

Can intentions be represented as causes and then estimated from data with causal inference?

Intentions could be represented either as (1) unobservable causes, or as (2) repeated measurements of observable variables.

1. Intentions cannot be represented as unobservable causes

Teleological representation of intentions

Intentions are relational descriptions of observable effects.

do(Practice sport) “listens to” want(Be fit)

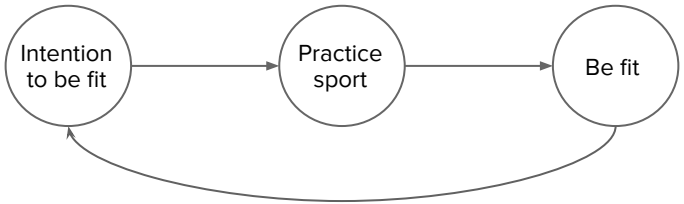
do(Practice sport) is not the same as Practice sport; want(Be fit) can be estimated from Be fit and do(Practice sport).

Intentions as unobservable causes

Intentions are unobservable causes (mental states) distinct from other events.

Practice sport “listens to” Intention to be fit

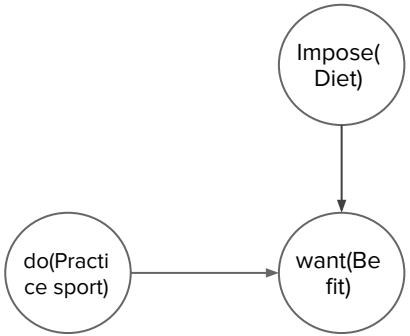
Problem: Intention to be fit still needs to “listen to” the state of Be fit (producing a cycle): in order to estimate the Intention to be fit, one needs data about Be fit.



Interference is not intervention (1)

Interference is about controlling for effects

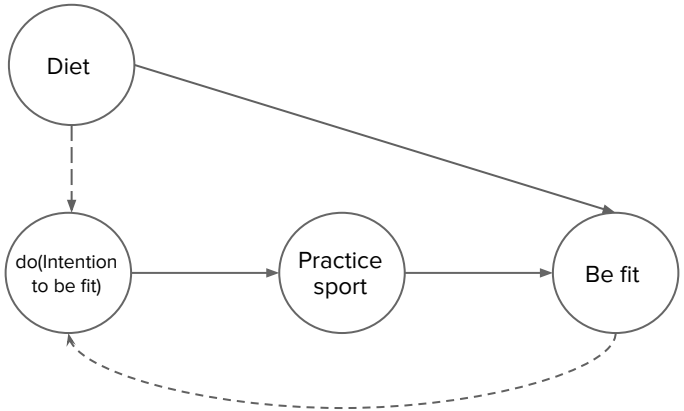
If you follow a diet which makes you fit, would you practice sport anyway? If you follow a super-caloric diet and cannot get fitter because of it, would you practice sport anyway?



Intervention is about controlling for causes

Diet impacts both the Intention to be fit and the fact of Being fit.

Problem: in order to identify this causal model, the intervention do(Intention to be fit) is needed; but this would block information from Be fit.



2. Intentions cannot be represented as repeated measurements

Teleological representation of intentions

Intentions are relational descriptions of observable effects.

do(Practice sport) “listens to” want(Be fit)

do(Practice sport) is not the same as Practice sport; want(Be fit) can be estimated from Be fit and do(Practice sport).

Intentions as repeated measurements

Intentions can be inferred by repeated measurements of observable variables.

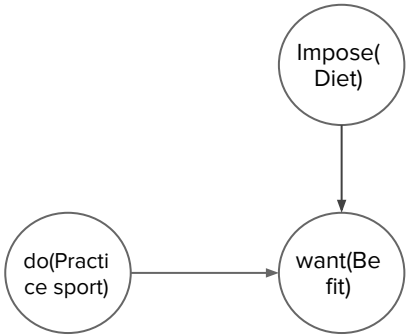
Practice sport “listens to” Be fit T_1

Problem: if I control for Be fit T_2, I expect to observe significant change in Practice sport, independently from Be fit T_1.

Interference is not intervention (2)

Interference is about controlling for effects

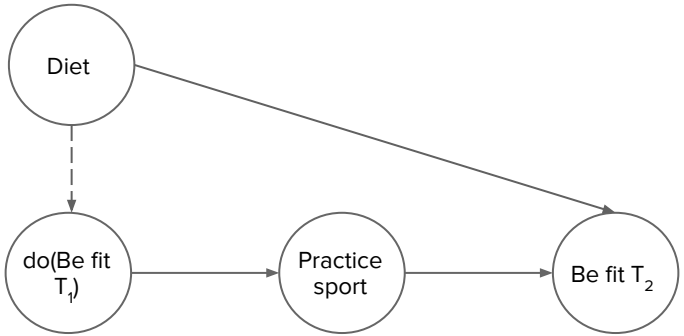
If you follow a diet which makes you fit, would you practice sport anyway? If you follow a super-caloric diet and cannot get fitter because of it, would you practice sport anyway?



Intervention is about controlling for causes

Diet impacts both Be fit T_1 and Be fit T_2.

Problem: one can only estimate in such a way if people who are already fit do not practice sport, and not if people who do not become fitter (because of a bad diet) practice sport anyway.



Reducing ends to causes is a well-known conundrum: it may look more scientific, but it mixes two different interpretive categories

Causal inference assumes an a priori category not found in the data (i.e. causality), but needed to explain it; teleological inference demands a second a priori category (i.e. intention) not found in the data and irreductible to causality.

The identification of means and ends depends on a previous identification of causal relationships

Teleological inference is by no means ‘acausal’; it is instead a further description of correlations, based on causal relationships.

Example: in order to understand the reason why you push a button, I need to know what are the effects of the button.

The aim of both causal and teleological inference is to explain observable correlation

"Nothing is hidden" (Wittgenstein): intentions should not be thought of as unreachable mental events, or as nothing at all, but instead inferred from data and hypotheses sui generis.

Thank you

dario.compagno@parisnanterre.fr