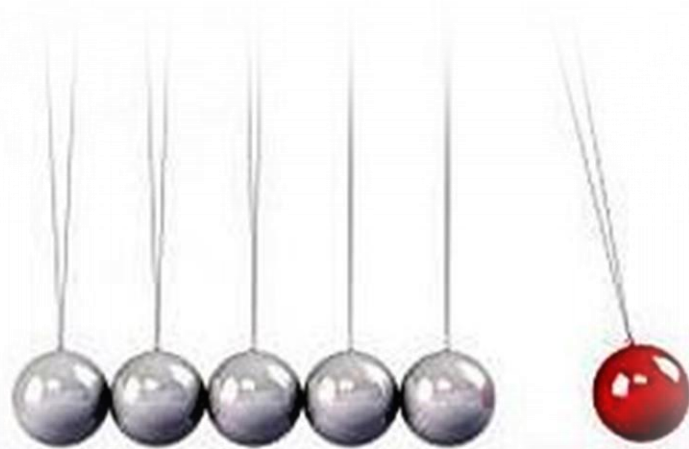




*Hunting causes of social phenomena for public policies:  
can big data studies overcome the “gold standard”?*



*SSST Thesis by Virginia Ghiara*

*Supervisor: Prof. Rosa Meo*

There are two reasons why social scientists and politicians aim to discover causal relations behind social phenomena:

1. To explain why something happened
2. To decide how to intervene to obtain a specific outcome

Several methods of causal discovery have been proposed, but one has known as the “gold standard”: the **Randomized Control Trial (RCT)**

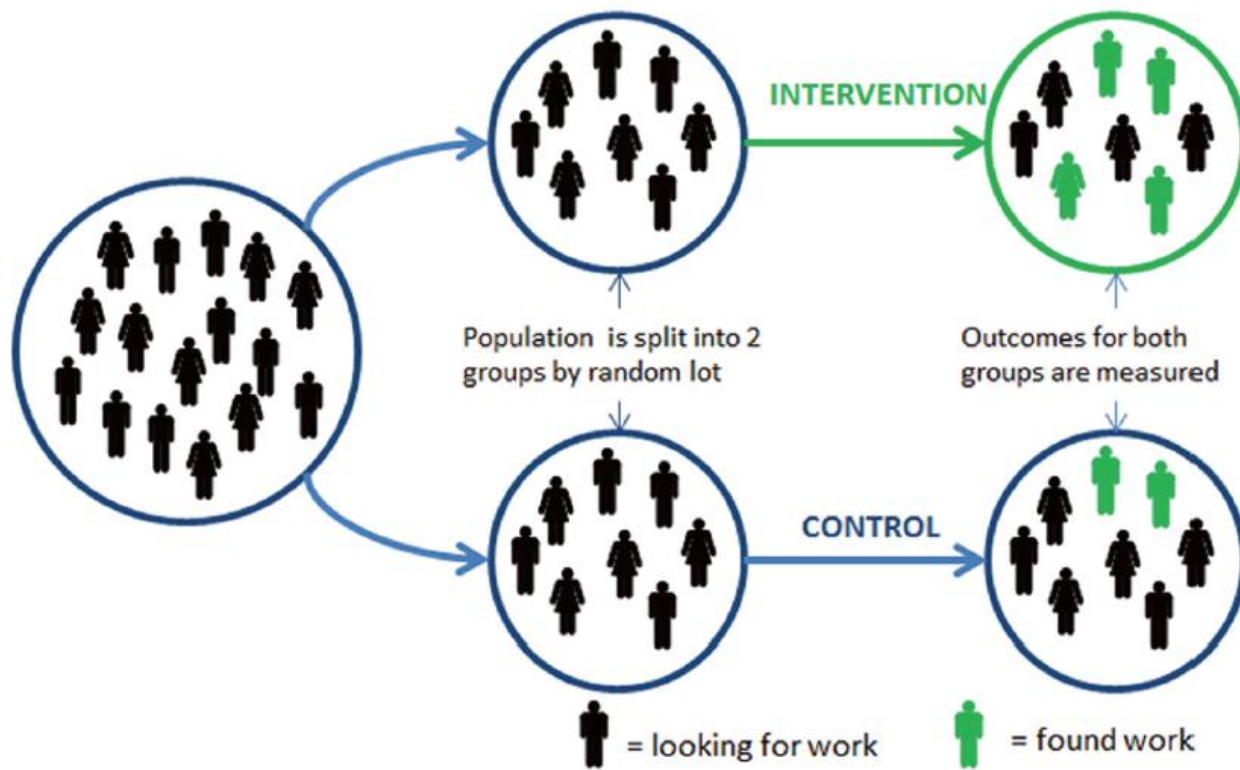


## What is the RCT?

The RCT is an experimental design which involves two groups of subjects, the experimental group and the control group, created through a random process.

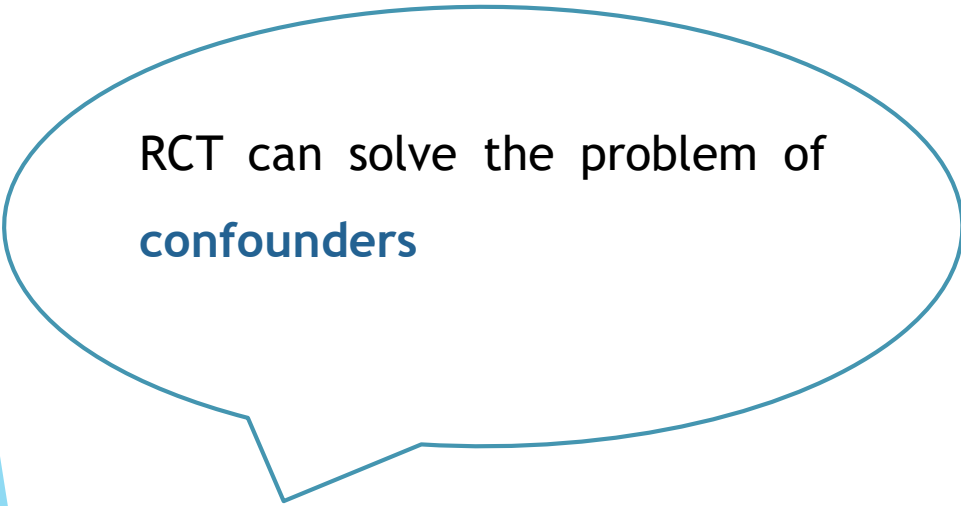
During the RCT the experimental group receives the treatment T, while the control group either remains untreated or receives only a placebo.

Therefore, if an outcome is observed in the experimental group and it is not present in the control group, scientists can establish with certainty that such result has been caused by the treatment T.

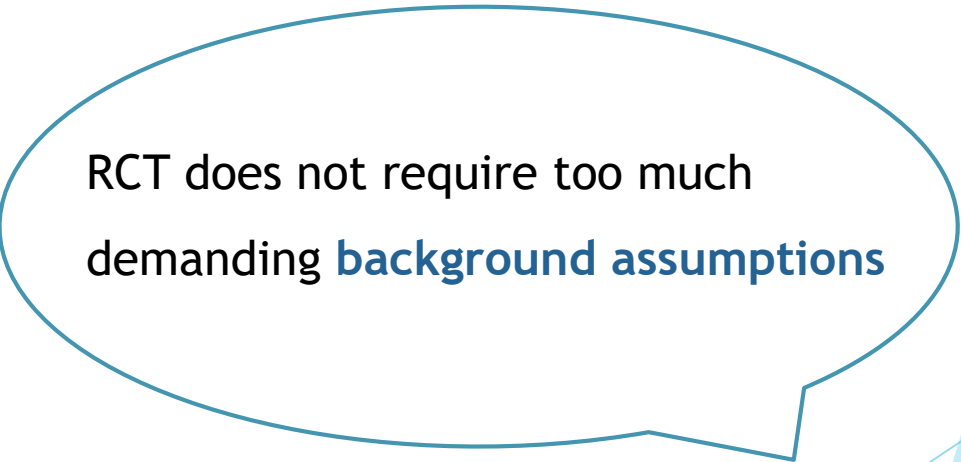


The widespread of RCT within diverse disciplines has been accompanied by the claim that it is the “gold standard” of causal inferences.

Such claim is supported by two considerations:



RCT can solve the problem of **confounders**



RCT does not require too much demanding **background assumptions**

# The failure of the “gold standard”

## Ideal vs. Real RCT

- ▶ In the real world, the wrong selection of the randomized unit can lead to disparities between the two groups which may affect the outcome  $O$  and make the result of the RCT invalid (Haynes et al. 2012).



## Ideal vs. Real RCT

- ▶ Randomization can ensure that not distinguishing factors create differences between the experimental and the control group only as the size of the groups goes to infinity, but intuitively real group cannot but have finite sizes, as a consequence the two groups may be not balanced (Reiss 2013).



## Ideal vs. Real RCT

- ▶ RCT should be, if not “full-blind”, at least “double-blind”, but a real RCT is sometimes “zero-blind” (Scriven 2008).





Wrong RCT units

Finite groups

“Zero-blind” RCT



Real RCT cannot solve the problem of confounders!

In addition... also ideal RCT has some limitations!

- ▶ RCT cannot answer **questions concerning “why”**
- ▶ RCT provides information about the **average** treatment effect, not about the effect of the treatment for particular individuals

# What about Big Data?

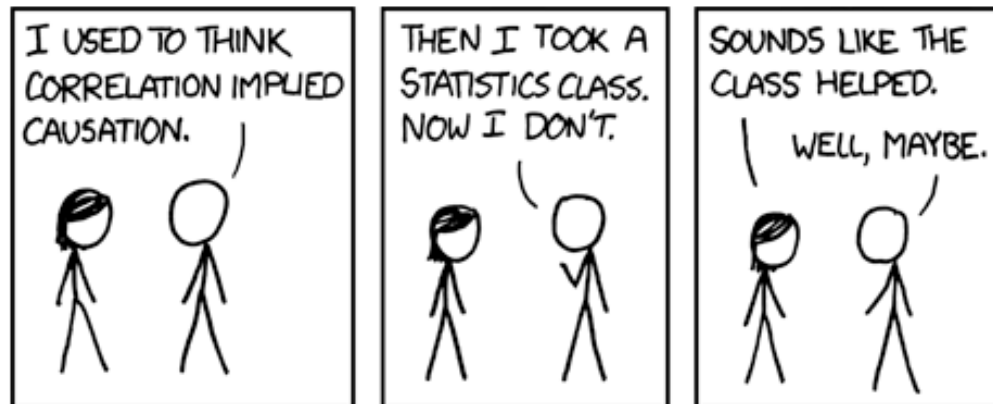
A new data deluge has shed light on the possibility to use observational rather than experimental data to find causal relations



Given the huge quantity of data now available about the social world, several correlations can be found between numerous variables.

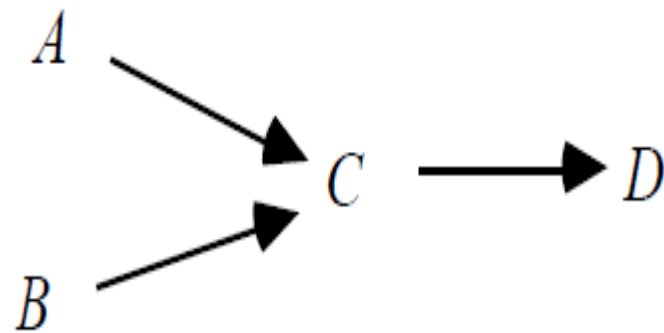
If A and B are correlated, this correlation could be explained in different ways:

- ▶ A causes B
- ▶ B causes A
- ▶ The common cause C causes A and B
- ▶ The correlation is spurious



Despite this difficulty, many algorithms have been proposed to find causal relations between data. Among them, some are based on the so-called Causal Bayesian Networks (CBN).

The causal inference methods based on the Causal Bayesian networks aim to infer causal relations from probabilistic dependencies and independencies over a set of variables  $V$ , which can be illustrated through a directed acyclic graph (DAG)



The vital centre of the inductive procedure performed by many algorithms searching for causal relationships between data is based on the Causal Markov Condition.

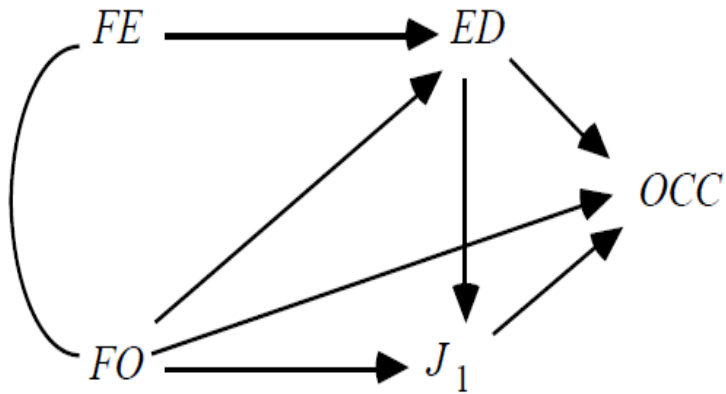
**The Causal Markov Condition** (CMC) Every variable is (conditionally) independent of its non-effect given its direct causes

$A \rightarrow B \rightarrow C$  Conditional on B, C is independent of A

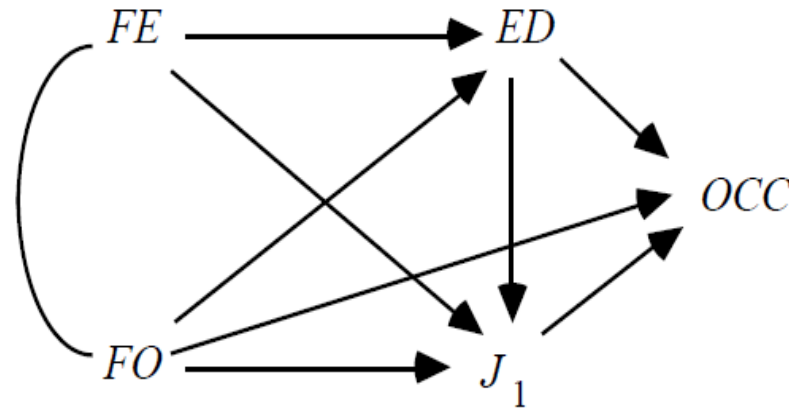
**The Faithfulness Condition** (FC) In a causal graph, no probabilistic independencies hold other than those predicted by the CMC

Several algorithms have been proposed to discover causal relations with the aid of CMC. In other words, if CMC holds, these methods should manage to find the causal relations.

Moreover, some of these algorithms have been shown to be capable to find the same causal linkages that have been found also in other ways.



Blau and Duncan's analysis (1967) of the causal relations between the role of education (*ED*), first job (*J<sub>1</sub>*), father's occupation (*FO*) and father's education (*FE*) in determining one's occupation (*OCC*).



The causal relations found by applying a BN algorithm to the same data set (the conditional independence relations found in the data at a significance level of .0001 are faithful to Blau and Duncan's directed graph).

Of course, also this approach has its problems...

“BN methods do not apply where [...] positive and negative effects of a single factor cancel”  
(Cartwright 2007, p. 12)

These algorithms cannot cope with the  
Simpson Paradox

How to deal with data sets to be integrated?

Questions concerning “why”  
are still without an answer!!



## In conclusion...

- ▶ The RCT is not a gold standard because it, as all the other methods, has important limitations
- ▶ Other methods, among which those based on big data studies might enhance the possibility to find causal relations existing in the social world, however also in these cases several problems and limitations can be found.





No “gold standard” could be found in the search for causal relations.

Therefore, the development of only one method would be counterproductive. Indeed, new approaches might not be considered and they would not be valorized, such as those dealing with the data deluge of the last years.



Thanks for your attention!

## Literature

Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge University Press, Cambridge

Eaton, C. et al. (2012). *Understanding Big Data*. McGraw-Hill, New York

Einav, L., and Levin, J. (2014). “Economics in the age of big data”. *Science*, 346(6210)

Haynes, L. et al. (2012). “Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials”. The Cabinet Office Behavioural Insights Team

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge

Reiss, J. (2013). *Philosophy of economics*. Routledge, New York

Scriven, M. (2008). “A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research”. *Journal of MultiDisciplinary Evaluation*:5 (9)

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. MIT Press, Cambridge MA, second (2000) edition

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press, Oxford