**Judea Pearl (2018) “The Book of Why: The New Science of Cause and Effect”.**

It is not about advocating for big data per se. Data is passive.

“…casual questions can never be answered from data alone. They require us to formulate a model of the process that generates the data, or at least some aspects of that process. Anytime you see a paper or study that analyze the data in a model-free way, you can be certain that the output of the study will merely summarize, and perhaps transform, but not interpret the data.” (p.351)

Proposed CAUSAL DIAGRAMS as a way to present the model. Judea is NOT discounting RCT, but there are situations when with the right observational data, causal questions can be addressed

THREE progressive ways of answering causality (slightly different from his idea on Ladder of Causation – which focused on data and human progressive level of causal thinking)

1st method of answering causality – BACK DOOR ADJUSTMENT

If we are able to *know all* the confounds and other variables and the way these variables affect X 🡪 Y, then we can calculate causality.

For instance, we control for confounds but not mediators or variables that are ‘fork’ A 🡨 B 🡪 C or colliders A 🡪 B 🡨 C. “M-bias” variables also should not be controlled for.

2nd method of answering causality – FRONT DOOR METHOD

Basic set-up for front-door method

**C**

**X** **M** **Y**

There must not be any confounding factors going to M so that X on Y can then be estimated from observational data (*see*-*X*). This diagram must be satisfied, otherwise we cannot estimate causality from observational data, that is we need to conduct an experiment (*do*-*X*).

3rd method of answering causality – *do*-CALCULUS

THREE foundational rules of the *do*-calculus

RULE 1: *P*(*Y* | *do*(*X*), *Z*, *W*) = *P*(*Y* | *do*(*X*), *Z*)

That is, if we observe that variable W is irrelevant to Y, then the probability of Y will not change.

RULE 2: *P*(*Y* | *do*(*X*), *Z*) = *P*(*Y* | *X*, *Z*)

That is, if a set of *Z* variables blocks all back-door paths from *X* to *Y*, then conditional on *Z*, *do*(*X*) is equivalent to *see*(*X*).

RULE 3: *P*(*Y* | *do*(*X*)) = *P*(*Y*)

We can remove *do*(*X*) from *P*(*Y* | *do*(*X*)) when there is no causal paths from *X* to *Y*.

“…if we cannot find a way to estimate *P*(*Y* | *do*(*X*)) from Rules 1 to 3, then a solution does not exist. In that case, we know that there is no alternative to conducting a randomized controlled trial. It further tells us what additional assumptions or experiments might make the causal effect estimable.” (p.238)

Judea’s students had over the years helped to build-up this *do*-calculus algorithm. Add to this is the concept of counterfactuals (the what-if a different route is taken). Together *do*-calculus and counterfactuals are the basis for this new science on causality. Causal diagrams carefully specified with the right assumptions are the bedrock, which is not revolutionary from structural equation modeling but SEM has been misused over the years, reducing it to regression lines and correlation coefficients.

Other considerations mentioned by Judea Pearl:

1. Instrumental variable can be used to estimate causality if and only if an appropriate variable can be found or created by an experiment
2. Mediation analysis can be used to disentangle the direct effects from the indirect effect (passing through the mediator). Counterfactuals play a role in disentangling the indirect from the direct.

Big data, while in itself cannot be used to answer causality, it has been suggested as being useful for ‘extrapolating’ data from different sources to address external validity – that is, how the data can be used to answer a question in a different setting.

Machine learning = big data

Deep learning = neural networks (after training, neural networks are left to evolve on their own and the programmer has no knowledge about the workings of these networks)