# Notes on DoWhy

DoWhy is based on a simple unifying language for causal inference. Causal inference may seem tricky, but almost all methods follow four key steps:

1. Model a causal inference problem using assumptions.
2. Identify an expression for the causal effect under these assumptions ("causal estimand").
3. Estimate the expression using statistical methods such as matching or instrumental variables.
4. Finally, verify the validity of the estimate using a variety of robustness checks.

4 Steps of causal inference according to DoWhy

1. **Model a causal problem**

DoWhy creates an underlying causal graphical model for each problem. This serves to make each causal assumption explicit. This graph need not be complete---you can provide a partial graph, representing prior knowledge about some of the variables. DoWhy automatically considers the rest of the variables as potential confounders.

1. **Identify a target estimand under the model**

Based on the causal graph, DoWhy finds all possible ways of identifying a desired causal effect based on the graphical model. It uses graph-based criteria and do-calculus to find potential ways find expressions that can identify the causal effect.

1. **Estimate causal effect based on the identified estimand**

DoWhy supports methods based on both back-door criterion and instrumental variables. It also provides a non-parametric permutation test for testing the statistical significance of obtained estimate.

Currently supported back-door criterion methods.

* ***Methods based on estimating the treatment assignment***
  + Propensity-based Stratification
  + Propensity Score Matching
  + Inverse Propensity Weighting
* ***Methods based on estimating the response surface***
  + Regression

Currently supported methods based on instrumental variables.

* Binary Instrument/Wald Estimator
* Regression discontinuity

1. **Refute the obtained estimate**

Having access to multiple refutation methods to verify a causal inference is a key benefit of using DoWhy.

DoWhy supports the following refutation methods.

* Placebo Treatment
* Irrelevant Additional Confounder
* Subset validation

**Let’s break down each of these 4 steps**

#1 and #2 – a language for describing the direction of relationships, the conditions, and the assumptions among variables. Note the limitations that might be encountered in these 2 steps: What if we don’t have sufficient knowledge to determine the nature of relationships? Is there any intervening variable(s) between cause and effect? Big data by itself cannot explain causality. Bear the limitations in mind as we proceed.

#3 -- supports the estimate of the causal relationship. Certain conditions must exist for certain methods proposed here to be valid. The suggested statistical methods are not new. There are existing methods to determine causal relationships besides RCT. Causal DAG still require expert knowledge that might not be derived from data. In traditional statistical analysis, creating a DAG or model to represent the nature of the relationships among variables has been lacking.

#4 – this is a good practice to test if the causal estimate is valid. Should be done but frequently overlooked in traditional statistical analysis.