SAND2019-8439C

Sandia National Laboratories

Using causal models to analyze imperfect data

PRESENTED BY

Lauren Hund

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

How does causal modeling work?

Recall that causal queries (counterfactuals) are outcomes from hypothetical interventions on a system.

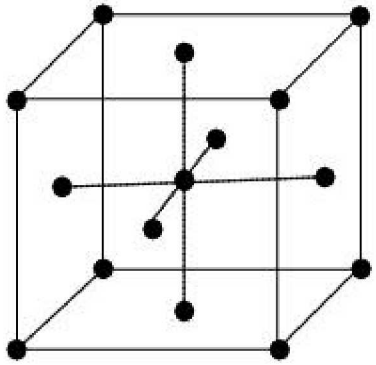
In an ideal world:

- Conduct the intervention on the system and see what happens – randomization (design of experiments, clinical trials)

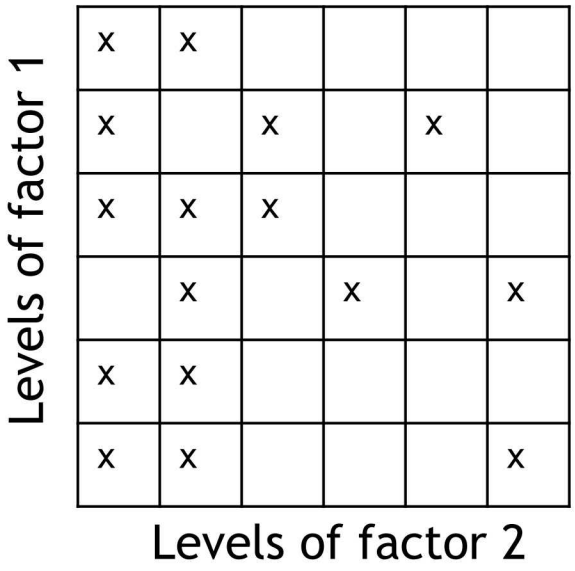
In a data-limited world:

- Analyze observational data (you don't control the design – what you see is what you get)
- Common approach to causal modeling: fake randomization.

Ideal



Actual



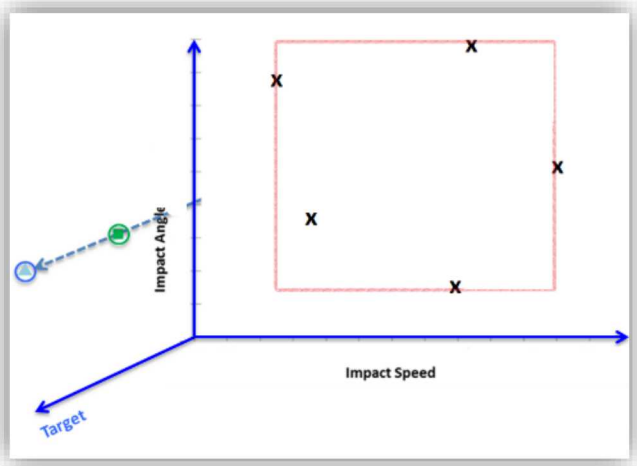
Sandia is all about counterfactuals!

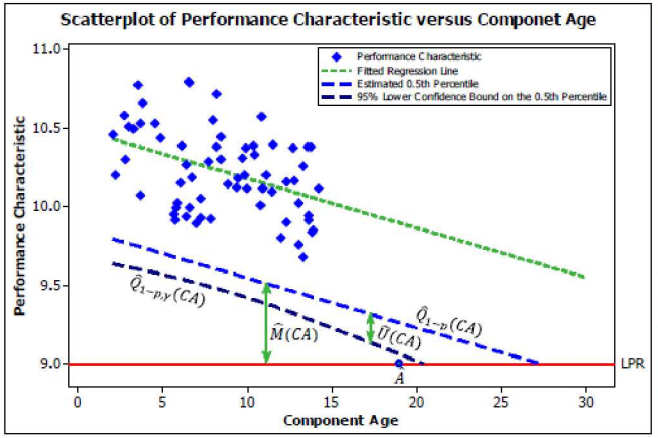
Sandia problems commonly concern "extrapolative prediction."

- Generating a predictive distribution for an unobserved outcome – prediction of a counterfactual.

Some examples:

- Nuclear weapons are the ultimate counterfactual predictions - without full-system tests, we certify weapons.
- Weapon components: How will a component perform across a variety of conditions (temperature, environments, age?)
- Computer models: Run model and predict to setting without data (counterfactual).





Dealing with data insufficiencies

Causality gives a language to talk about credibility of a prediction given less-than-ideal data.

- Much of causal inference is simply ensuring that your data analysis methods accurately reflect the "**data generating mechanism**," i.e. how your data were generated.
- Under what set of assumptions is my counterfactual prediction valid?

Structural causal modeling is one language of causality.

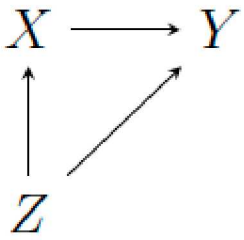
Steps for causal analysis:

1. Define a causal query.
   - Often a function of a counterfactual.
2. Determine how the collected data relates to the true underlying structural causal model.
   - Make a DAG!
3. Check if sufficient data to estimate query.
4. **Estimate the query from the data.**

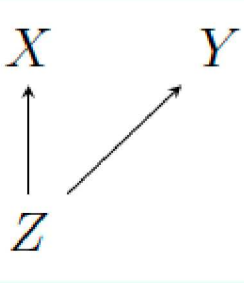
Step 1: Causal query

$$P(Y = do(X = x))$$

Step2: DAG

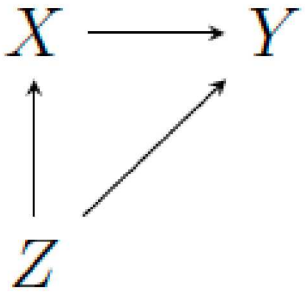


Step 3: Check criteria



Estimate QoI

In practice, we want to move from ***qualitative*** DAG model to ***quantitative*** statistical model in order to estimate a causal query.



Adjustment formula:

$$P\big(Y = do(X = x)\big) = \sum_z P(Y|X = x, Z = z)P(Z = z)$$

Unobserved counterfactual

Observed in data

Stratifying on Z, we can estimate the counterfactual of interest from the data.

- Other formulations of the adjustment formula exist, e.g. for selection variables and for the front-door criterion.

Estimate causal query

Causal inference is all about models and assumptions.

- What assumptions are you willing to make?
- Do you have enough data to fit a "good" statistical model under those assumptions?

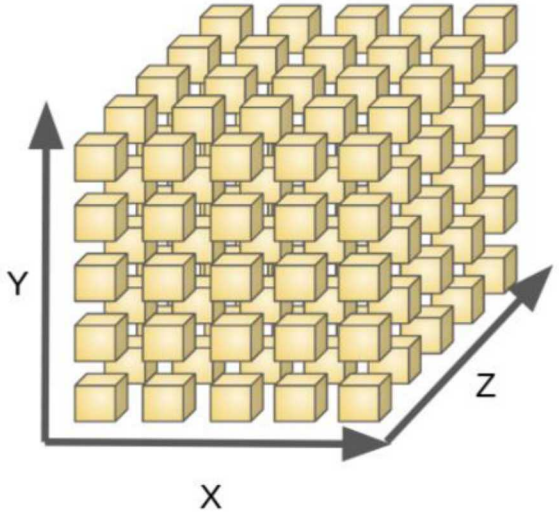
Fundamental assumptions of causality: Given a random sample from a population:

- ***Exchangeability:*** no unmeasured confounding
  - Measure enough variables?
- ***Positivity:*** enough data to estimate $P(Y|X = x, Z = z)$.
  - Have enough data?
- ***Consistency:*** no multiple versions of treatment
  - Treatment can be hypothetically manipulated in a consistent manner
  - Example: drug; counterexample: BMI.

Role of modeling and assumptions

Curse of dimensionality – use statistical models to approximate distribution of Y|X,Z.

"Art" of statistical modeling.



Picture taken from: https://medium.freecodecamp.org/the-curse-of-dimensionality-how-we-can-save-big-data-from-itself-d9fa0f872335

There is an implicit fourth assumption needed for causal estimation: correct model specification.

- Causal methods are often 'model-agnostic': how you model is separate from how you calculate causal estimands given the model.
- The 'modeling' stage is where good statistical and ML models come into play.

Thank you!

"Use of technical causal language, a good use, in our estimation, must be recognized as simply a shorthand for better versus worse analyses, as judged by the author, and not a metaphysical statement about causation per se..."

Lipton and Odegaard (2005)