UNIVERSITY OF COPENHAGEN

Targeted register analyses
PhD short course

The Causal Roadmap: introduction and motivation

Andrew Mertens
University of California, Berkeley, Division of Biostatistics

Presentation outline

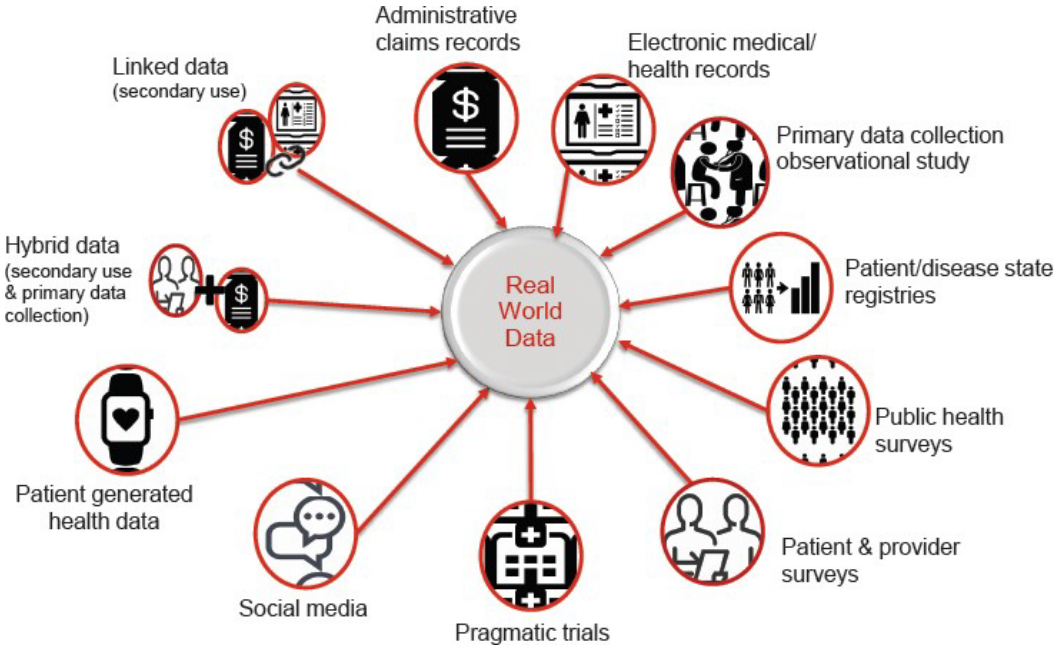
1. Motivation
2. Traditional approaches
3. Targeted Learning and The Causal Roadmap
4. Example causal questions

Presentation outline

1. Motivation
2. Traditional approaches
3. Targeted Learning and The Causal Roadmap
4. Example causal questions

A landscape of opportunities from electronic health registries and beyond

- Today's data ecosystem: Rich and diverse data sources
  - Registries
  - Electronic health records
  - Clinical trials
- Ability to link and combine
- Powerful new analytic tools
  - Machine learning
  - Computing power

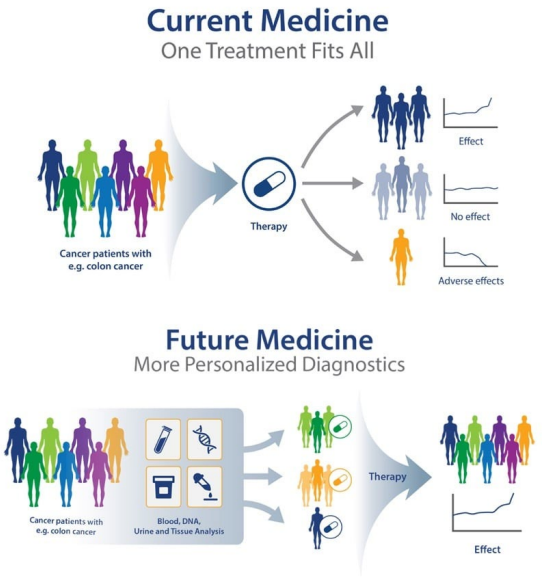


https://www.nap.edu/read/25352/chapter/7#73

The Promise:

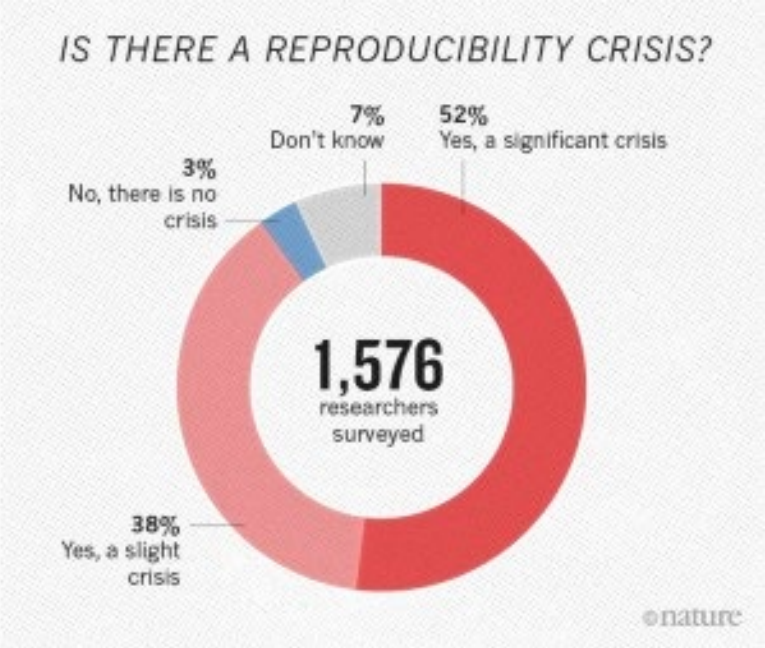
Big data and statistical advancements can provide novel insights for how best to treat patients and deliver care

- Real-world comparative effectiveness
  - Long-term cumulative effects
  - Unexpected benefits and risks
- Transport of effects
  - From trials to new populations
- Personalized medicine
  - Which types of patients respond best to which treatments
  - When to initiate/modify/intensify therapies



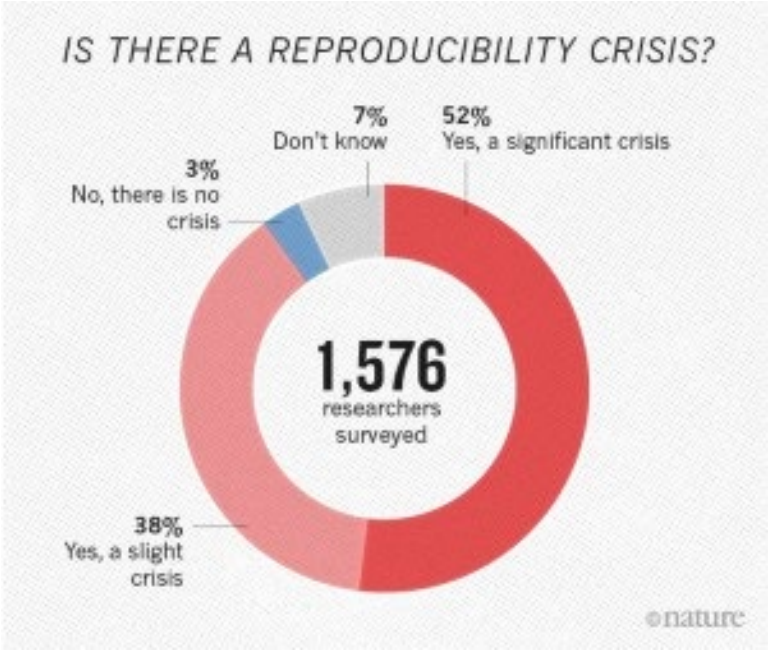
https://blog.crownbio.com/pdx-personalized-medicine

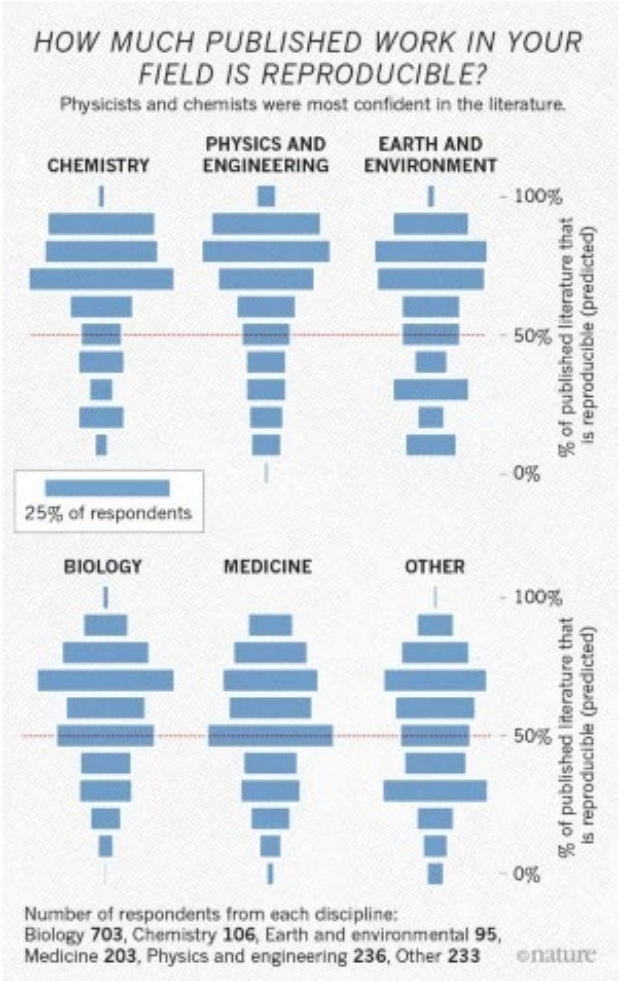
The reproducibility crisis



https://www.nature.com/articles/533452a

The reproducibility crisis

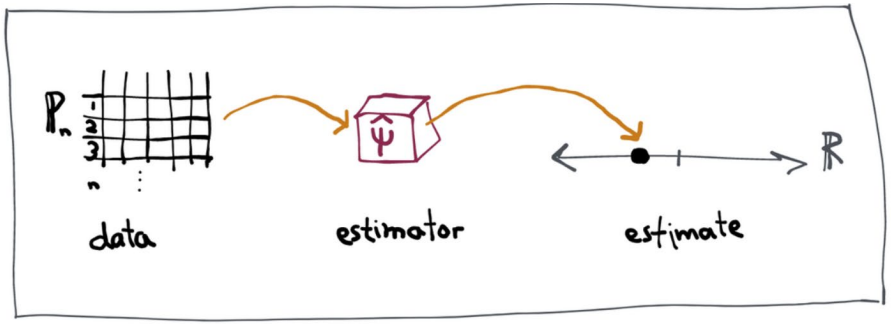




https://www.nature.com/articles/533452a

The reproducibility crisis

- Most writing about the crisis of reproducibility has concentrated on:
  - Issues of analysis pre-specification
  - Using reproducible workflows
  - P-hacking/publication bias
- Less focus on specific methodologies that avoid the bias inherent in traditional analytic procedures.



https://www.nature.com/articles/533452a

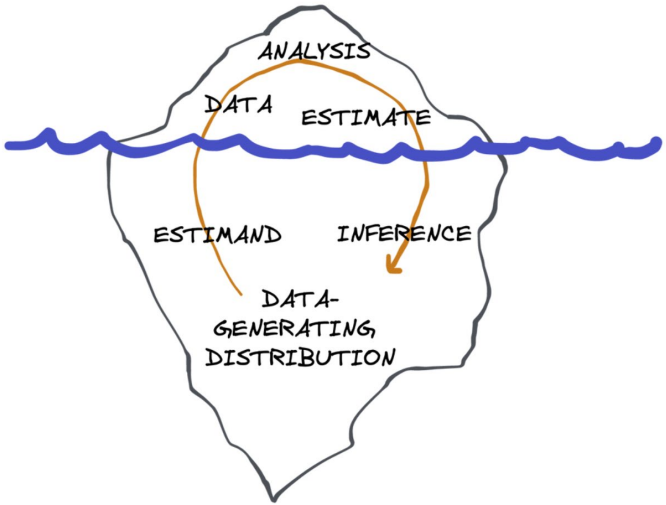
Presentation outline

1. Motivation
2. Traditional approaches
3. Targeted Learning and The Causal Roadmap
4. Example causal questions

The reproducibility crisis

"Cargo-cult statistics"

- "The ritualistic miming of statistics rather than [its] conscientious practice,"
- Overly specified statistical modeling choices to guide how scientific questions are answered
- Parametric statistical model encodes strong assumptions about the underlying data-generating process (DGP), like:
  - the outcome being linear with respect to covariates
  - errors being normally distributed with constant variance conditional on covariates
  - Interactions
- All potential sources of statistical model misspecification



https://alejandroschuler.github.io/mci

Stark, P.B. and Saltelli, A. (2018) Cargo-cult statistics and scientific crisis. *Significance*, **15** (4), 40–43.

The challenges of “real-world” data

- Classic confounding: Patients who received drug differ from those who didn’t received drug
  - In ways that may affect (or be correlated with) risk of dementia outcome

Baseline “risk factors” Treatment history; Co-morbidities
Demographics; Biomarkers; ….

Which variables should we adjust for?

A huge potential adjustment set

How to adjust for differences in risk?

We have a (large) set of adjustment variables… **now which estimator**?

Parametric outcome regression?

- Model specification?
  - Binary outcome? Maybe a logistic regression?
  - Time-to-event outcome with some right censoring? Maybe a Cox model?
  - Main terms? Interactions?
  - Polynomials/nonlinearities?....

Propensity scores?

- Probability of receiving treatment (vs. control) given adjustment variables
- Model specification?
  - Main terms? Interactions? Etc…?
- Estimator?
  - Matching? How?
  - Inverse weighting?

Many options? How to find our way?

Fully pre-specify our analysis?

- Before looking at the data, pre-specify our choices
- Adjustment variables, estimator, model specification, etc…
- Protects against “researcher degrees of freedom”
- Prevent p-hacking

Many options? How to find our way?

Fully pre-specify our analysis?

- Perils:
  - Relationships in the real world are complex
  - We don't understand them fully
  - A pre-specified parametric model may fit the data terribly
  - Bias and misleading inference

Many options? How to find our way?

Be flexible, use our knowledge, look at the data as we go?

- In practice:
  - try a bunch of models and estimators
- Choose the approach with results that “make the most sense”

Many options? How to find our way?

Be flexible, use our knowledge, look at the data as we go?

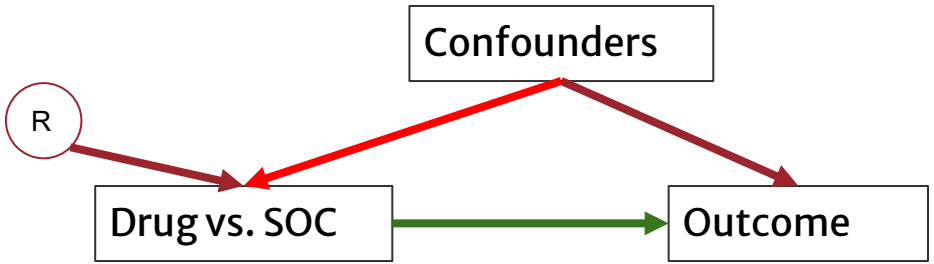
- Perils:
  - Misleading (under)estimate of uncertainty
    - Not accounting for model selection
    - P-hacking
    - Bias
  - Humans are good at creating narratives
    - Tend to confirm what we expect to find
    - Separating causation from association in observational data

Many options? How to find our way?

- Pre-specified parametric models: Model misspecification and bias
- Exploratory “common sense”: Underestimate uncertainty; Human bias

Not just an issue in observational data…

Randomized Trials as gold standard

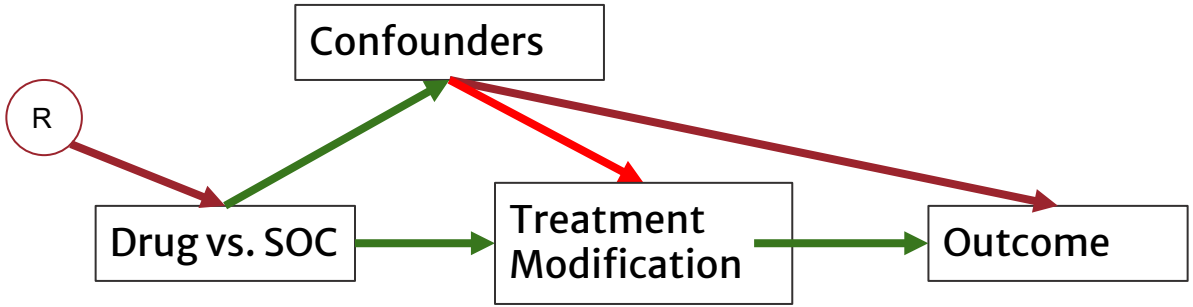


https://www.nature.com/articles/533452a

Not just an issue in observational data…

Many challenges not addressed by randomization

- Drop-ins, treatment modification, drop-outs, missing data, competing risks, selection bias etc…



- Analysis: series of complex decisions that must be pre-specified
- Requires casual knowledge; not just a statistics problem

https://www.nature.com/articles/533452a

Randomized Trials face similar challenges

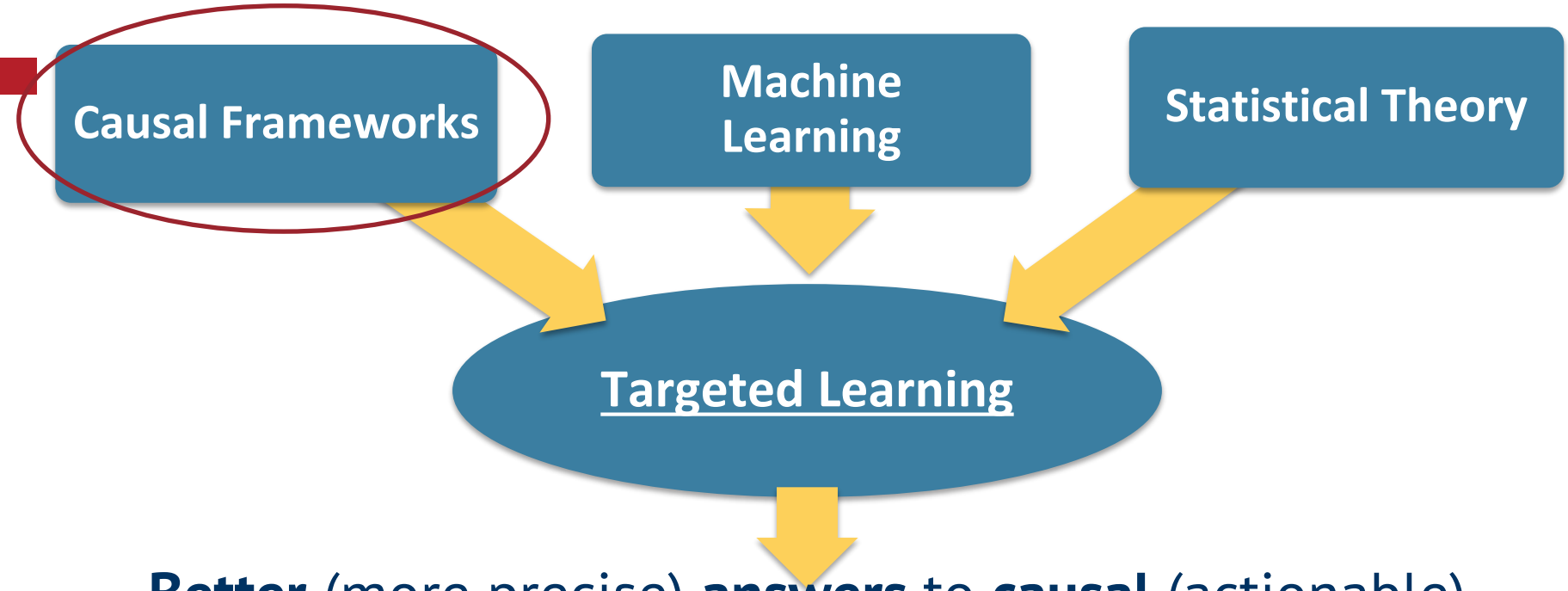
- Intention to treat analyses: Still many choices!
  - Adjustment can improve precision but must be prespecified
  - Rigid pre-specified approaches can fail to perform
- And much we cannot control… (much about trials is observational!)
  - Adherence, Patient drop out, Treatment modifications, Protocol deviations
  - Intent to Treat Analyses may not be interpretable or informative

We can **and should** use our trial data to go after more relevant questions

- Ex. Per-protocol analyses
- Ex. Treatment effect heterogeneity

Presentation outline

1. Motivation
2. Traditional approaches
3. Targeted Learning and The Causal Roadmap
4. Example causal questions



Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

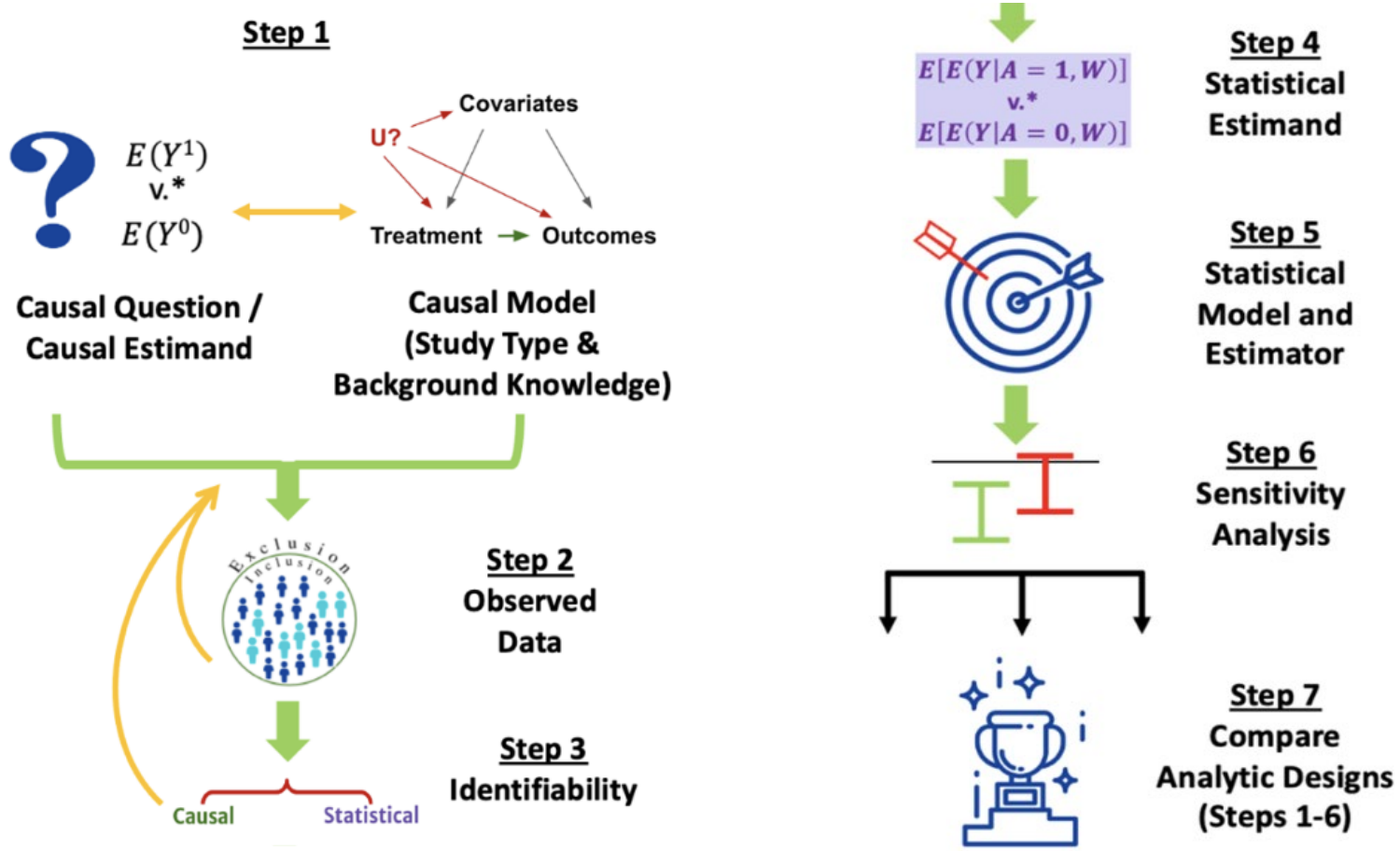
Why a Causal Roadmap is essential

Not defining a clear question or set of assumptions is by far the largest source of catastrophic errors in any data analysis.

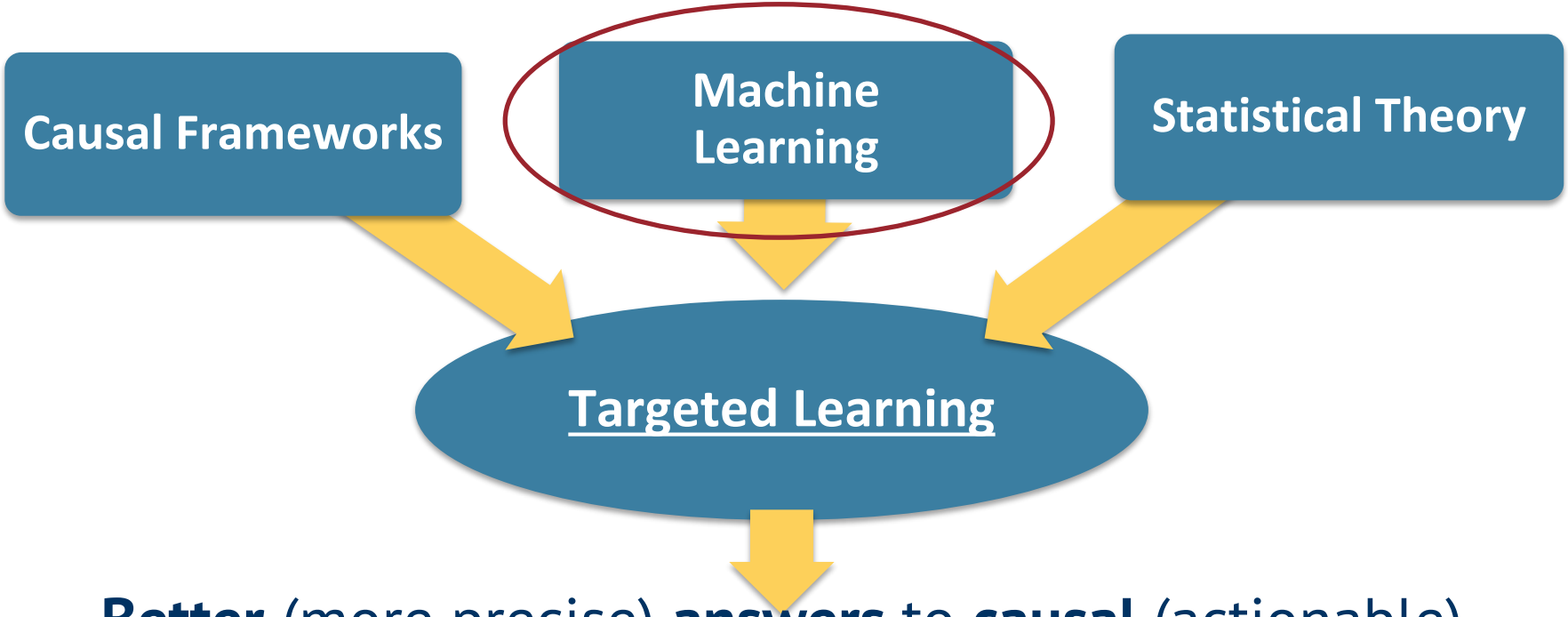
- Roadmap to guide:
  - Sharp framing of causal question (“causal estimand”)
  - Transparent articulation of causal assumptions
- Improved data and design
- Statistical estimation tailored to the causal question
- Statistical estimators that meet performance benchmarks
- Sensitivity analysis to allow for robust causal inference

eg, Petersen & van der Laan, 2014

Figure 1: The *Causal Roadmap*

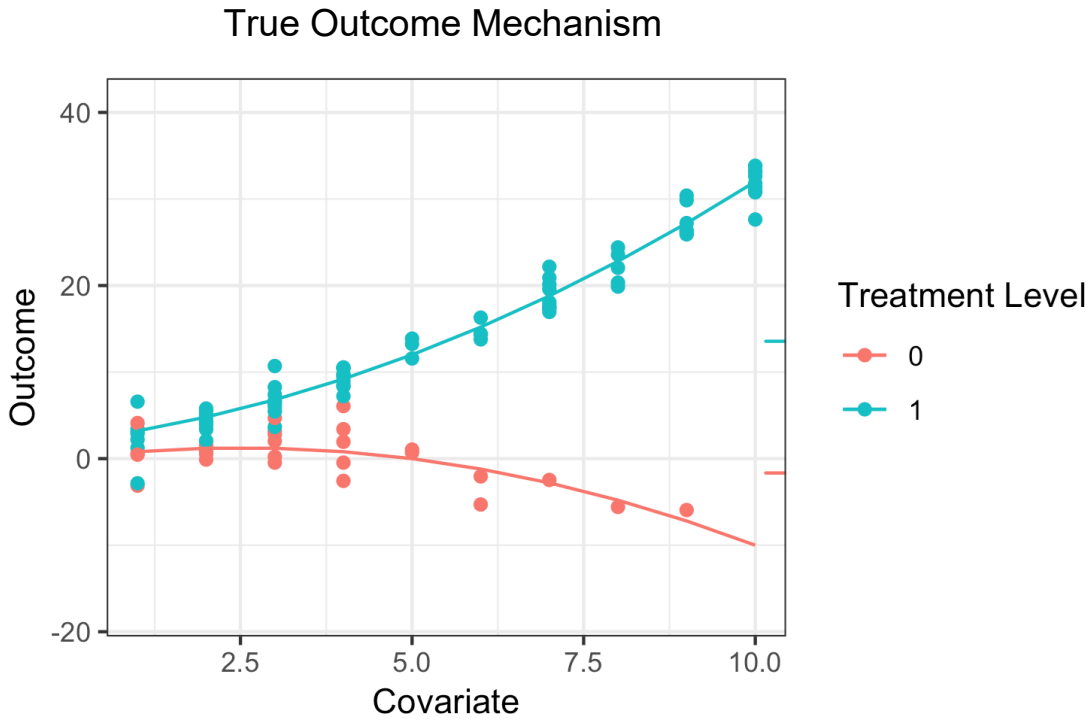


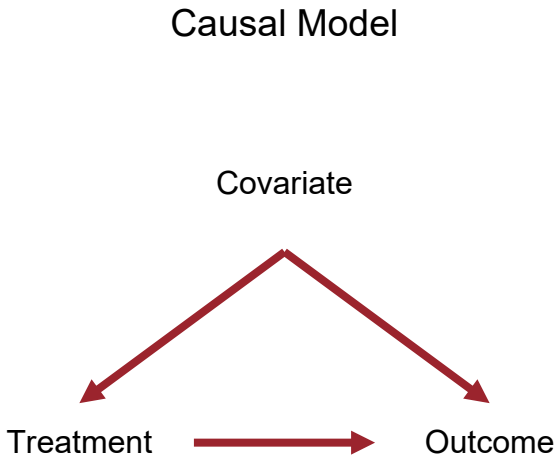
Dang et al. 2023



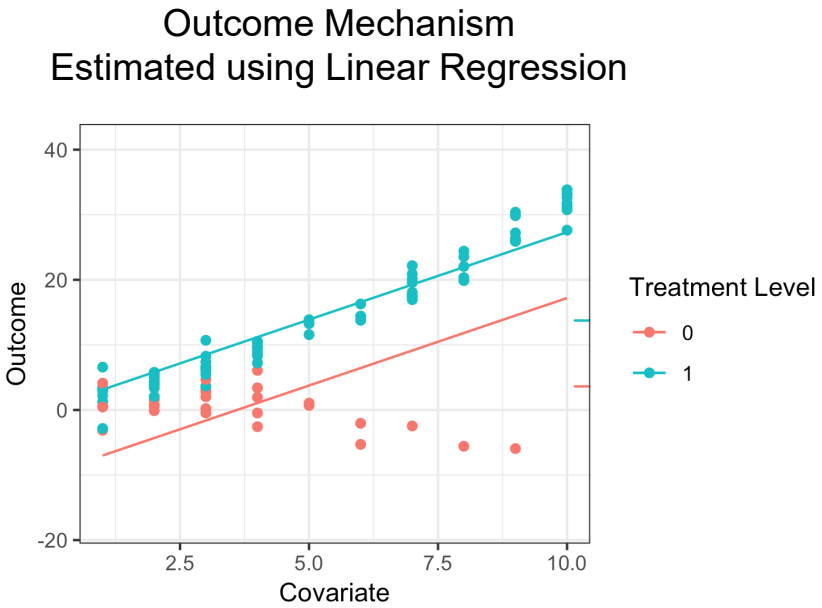
Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

Targeted Learning Schematic



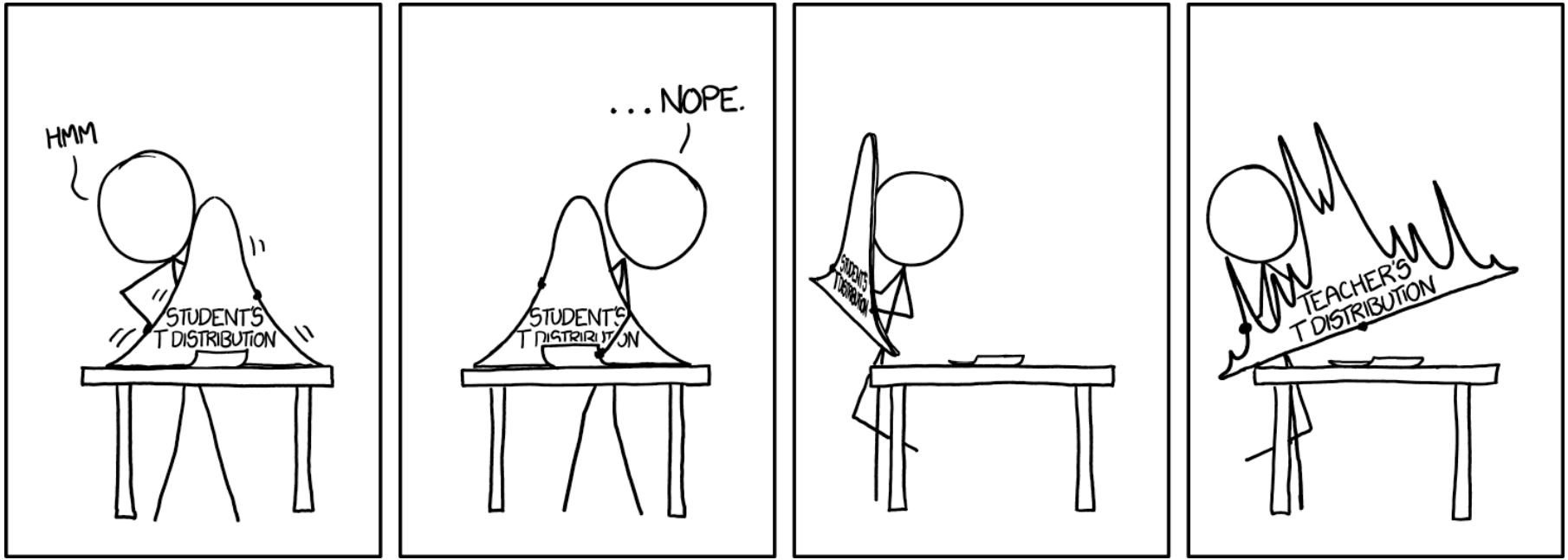


Targeted Learning Schematic



Underlying distributions

Assumptions vs. reality

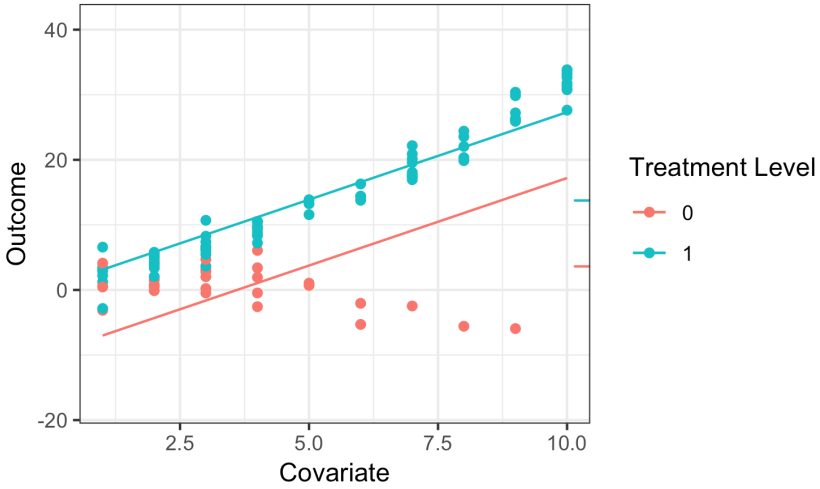


https://xkcd.com/1347/

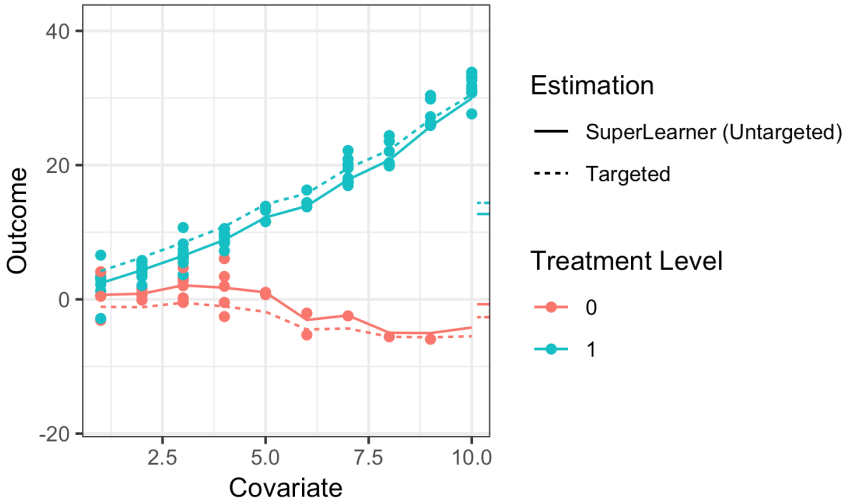
Semiparametric estimation methods like TMLE can rely on machine learning to avoid making unrealistic parametric assumptions about the underlying distribution of the data (e.g. multivariate normality).

Targeted Learning Schematic

Outcome Mechanism
Estimated using Linear Regression

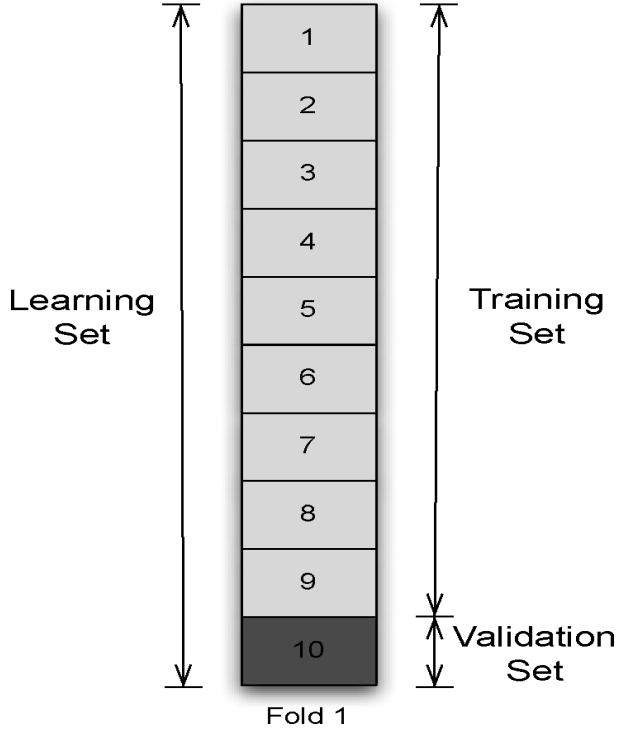


Outcome Mechanism
Estimated using SuperLearner and
TMLE

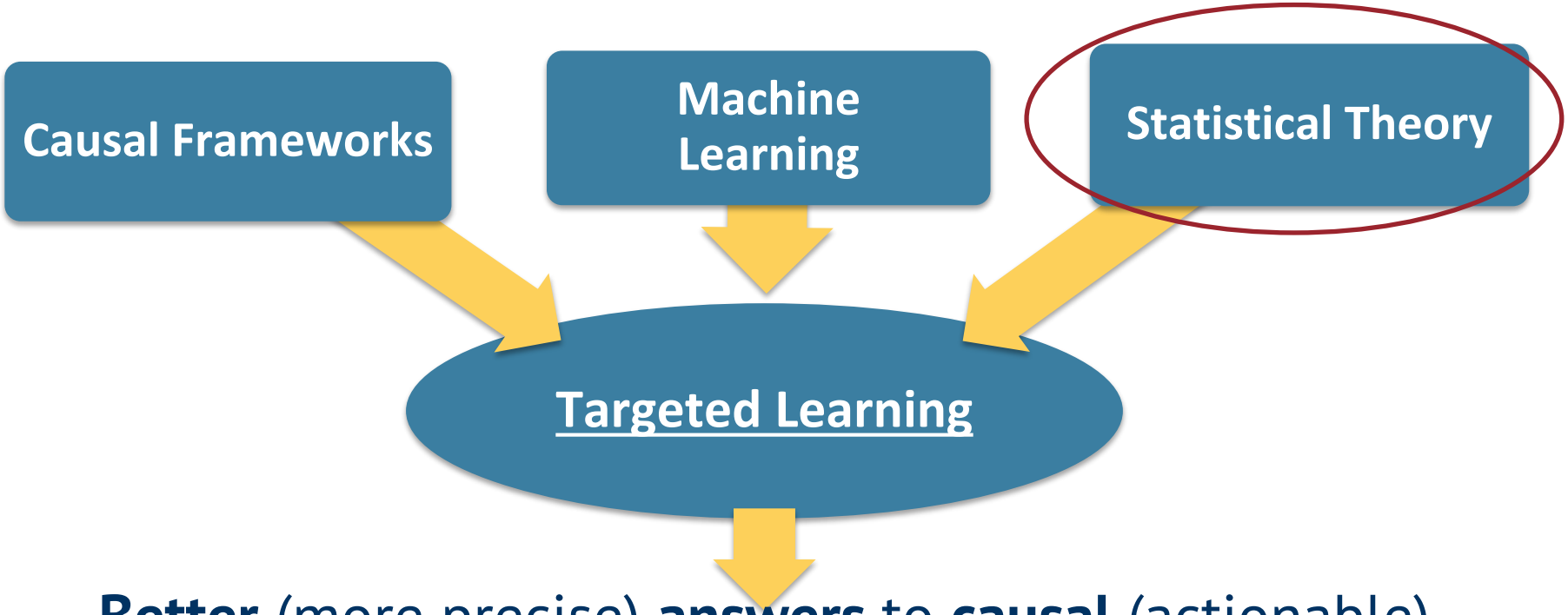


Super Learning: Ensemble Machine Learning

- "Competition" of algorithms
  - Parametric models
  - Data-adaptive (ex. Random forest, Neural nets)
- Best "team" wins
  - Convex combination of algorithms
- Performance judged on independent data
  - V-fold cross validation (Internal data splits) to avoid overfitting
  - Seek to minimize a specified loss function, for example, the mean squared error (MSE)
- Also called stacking, stacked generalizations, and weighted ensembling

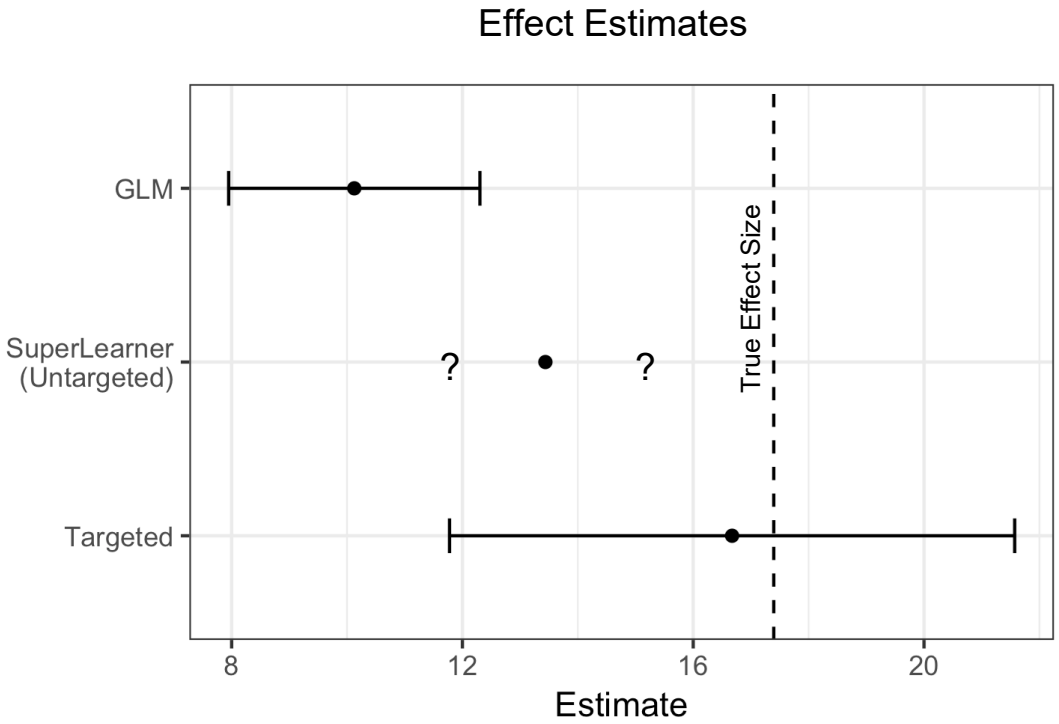


Van der Laan, Polley, 2007



Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

Targeted Maximim Likelidood Estimation:
removing bias AND robust inference



GLM did not learn the correct outcome mechanism, so its estimate is very biased

SuperLearner does a better job of estimating the outcome mechanism, but does not allow valid inference

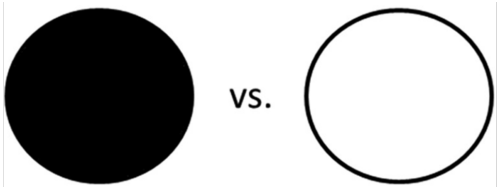
TMLE combines good outcome mechanism estimation with targeting to get valid inference

Presentation outline

1. Motivation
2. Traditional approaches
3. Targeted Learning and The Causal Roadmap
4. Example causal questions

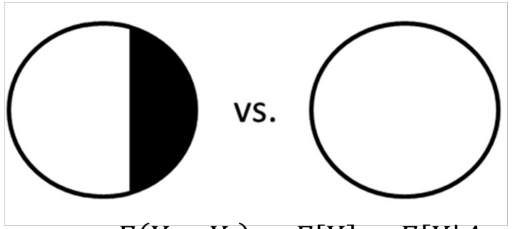
Targeted learning to directly estimate the parameters we care about

Average Treatment Effect



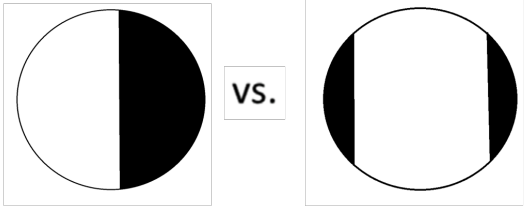
$$\varphi = E(Y_1 - Y_0) = E[Y|A = 1, W] - E[Y|A = 0, W]$$

Population Intervention Impact

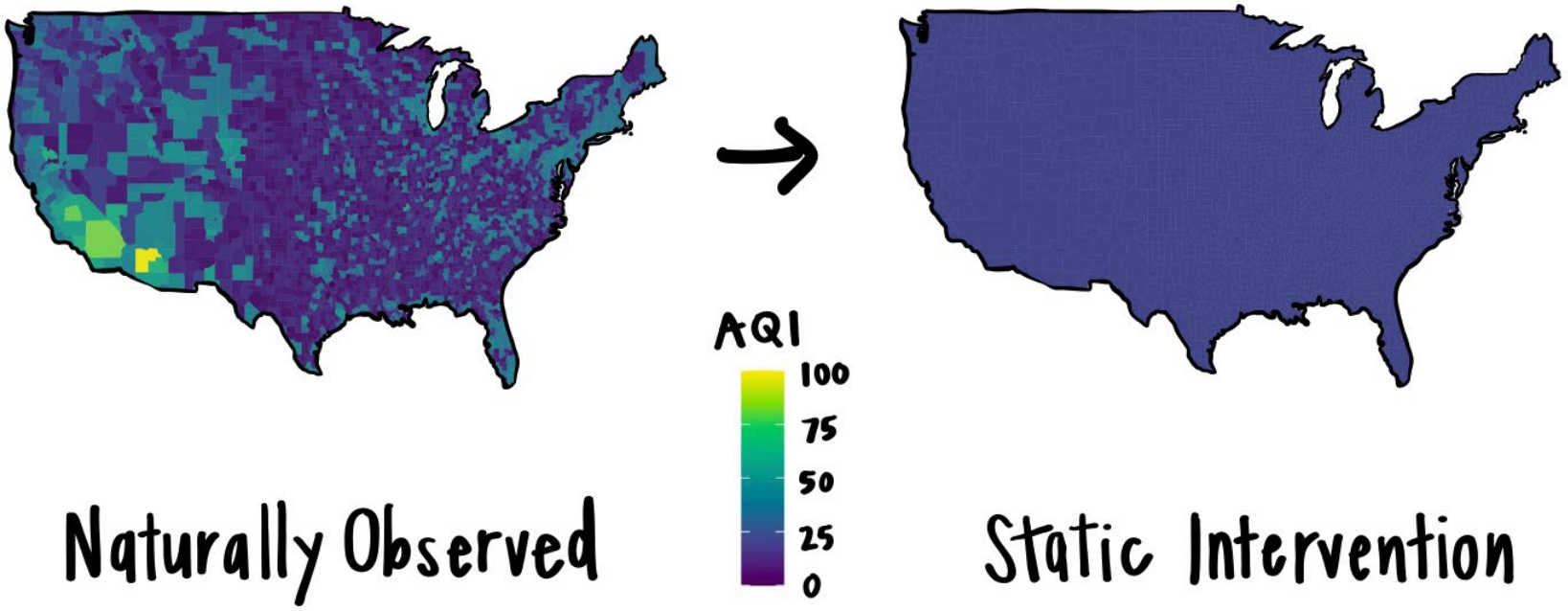


$$\varphi_{PI} = E(Y - Y_0) = E[Y] - E[Y|A = 0, W]$$

Optimal Dynamic Treatment Rule



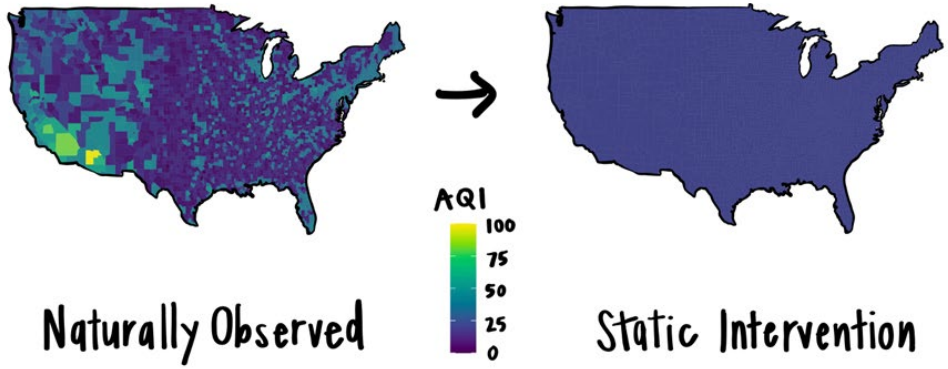
$$\varphi_{ODTR} = E\left(Y - Y_{d_{opt(W)}}\right) = E[Y] - E[Y|A = d_{opt(W)}, W]$$

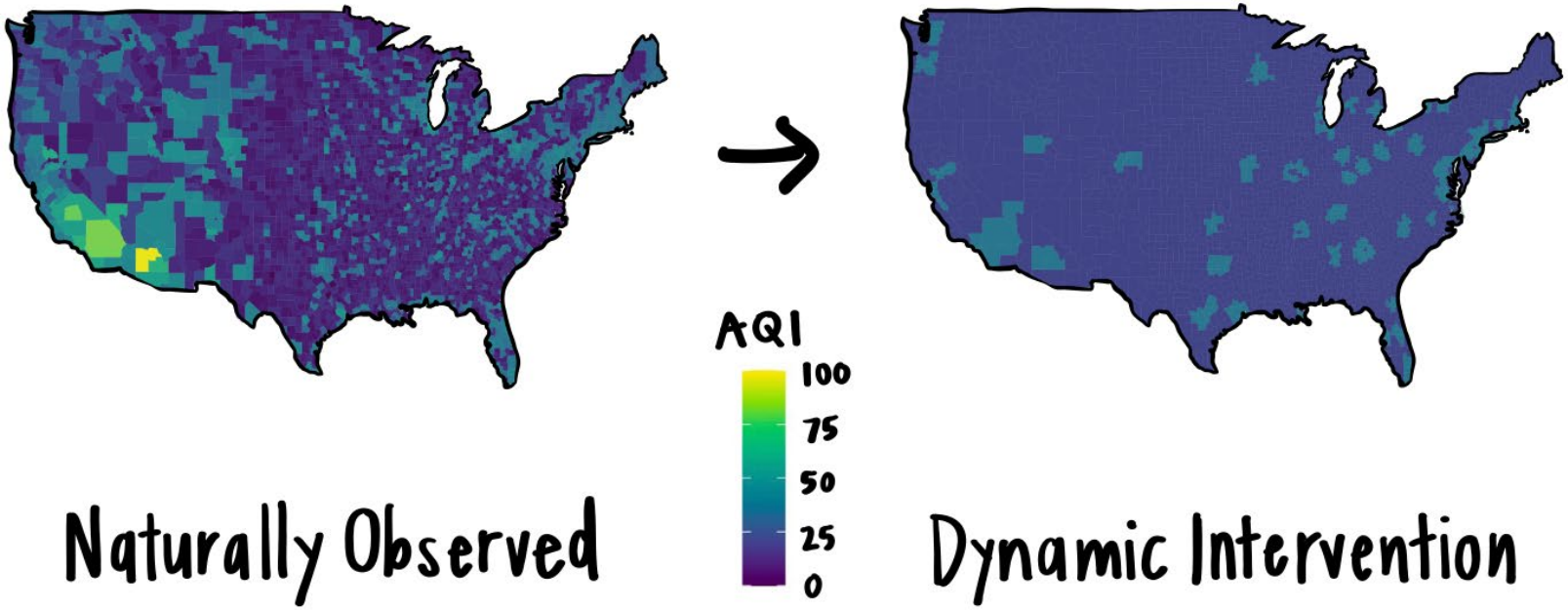


https://www.khstats.com/blog/lmtp/lmtp

Many causal questions are not static interventions

- Issues:
  - Positivity violations
    - Theoretical
    - Practical
  - Unrealistic Interventions



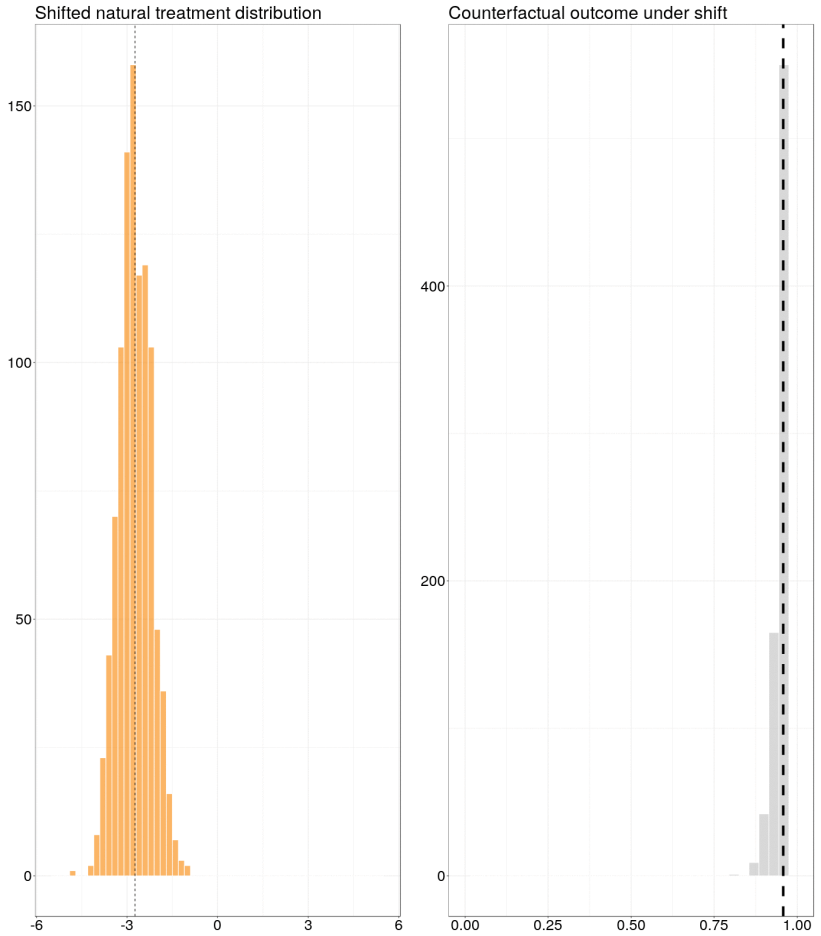


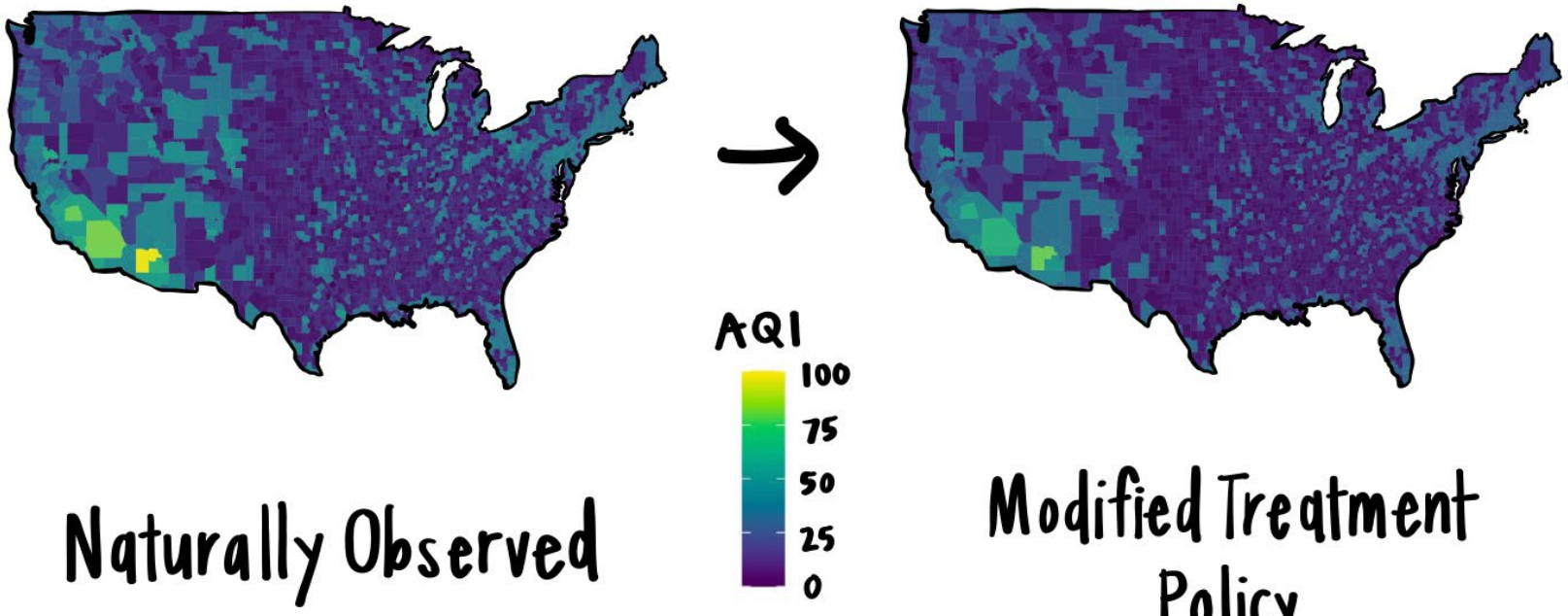
https://www.khstats.com/blog/lmtp/lmtp

Stochastic interventions

- Present a relatively simple yet extremely flexible manner by which *realistic* causal effects (and contrasts thereof) may be defined.

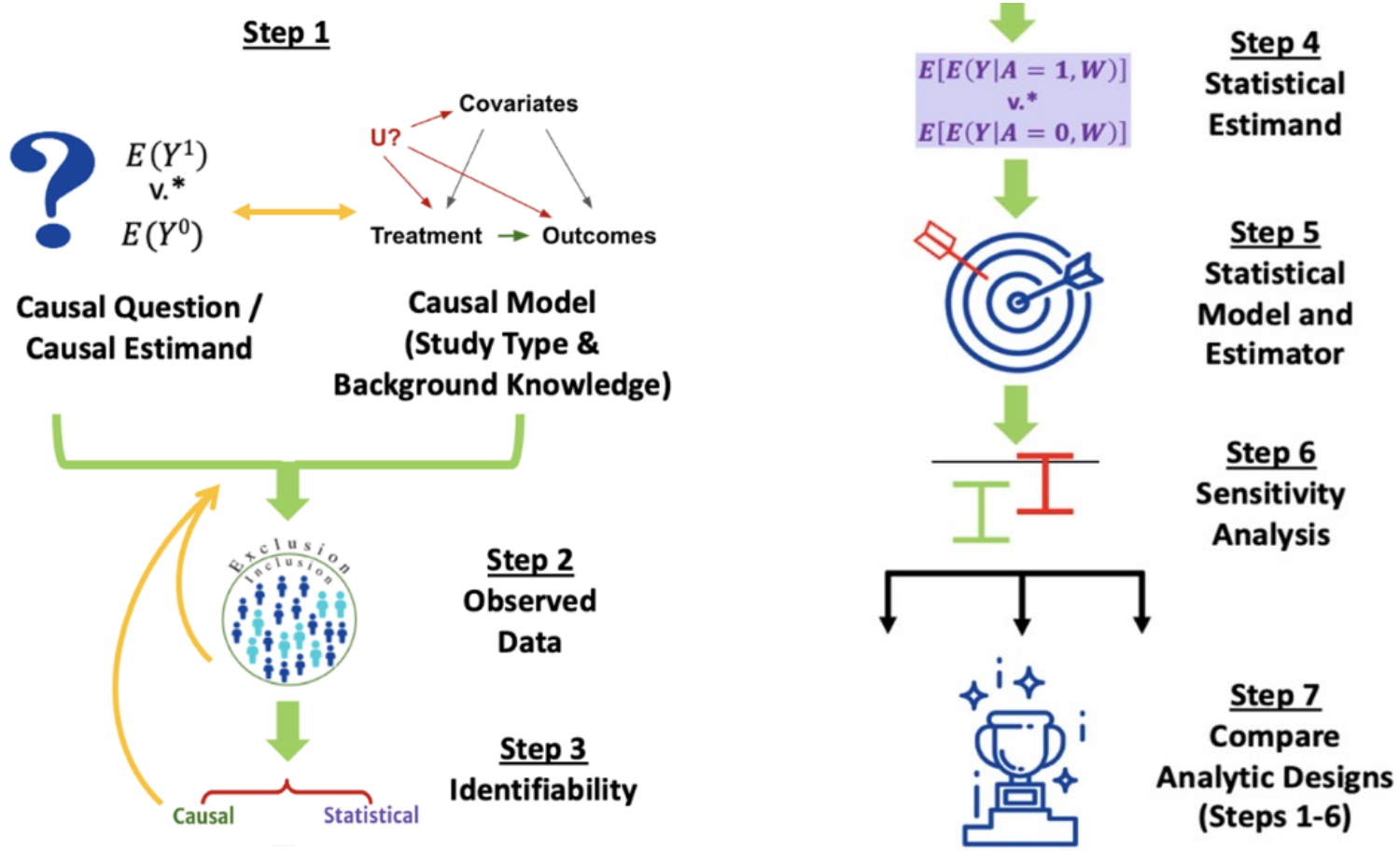
- Allows for estimating the effect of shifting a distribution of a variable
- Rather than contrasting two specific levels





https://www.khstats.com/blog/lmtp/lmtp

Figure 1: The *Causal Roadmap*



Dang et al. 2023