Targeted Learning: Minimizing Estimation Bias in Observational Studies

Mark van der Laan
Division of Biostatistics, UC Berkeley

July 18, 2018, Washington
Workshop on Real World Evidence, National Academy of Science

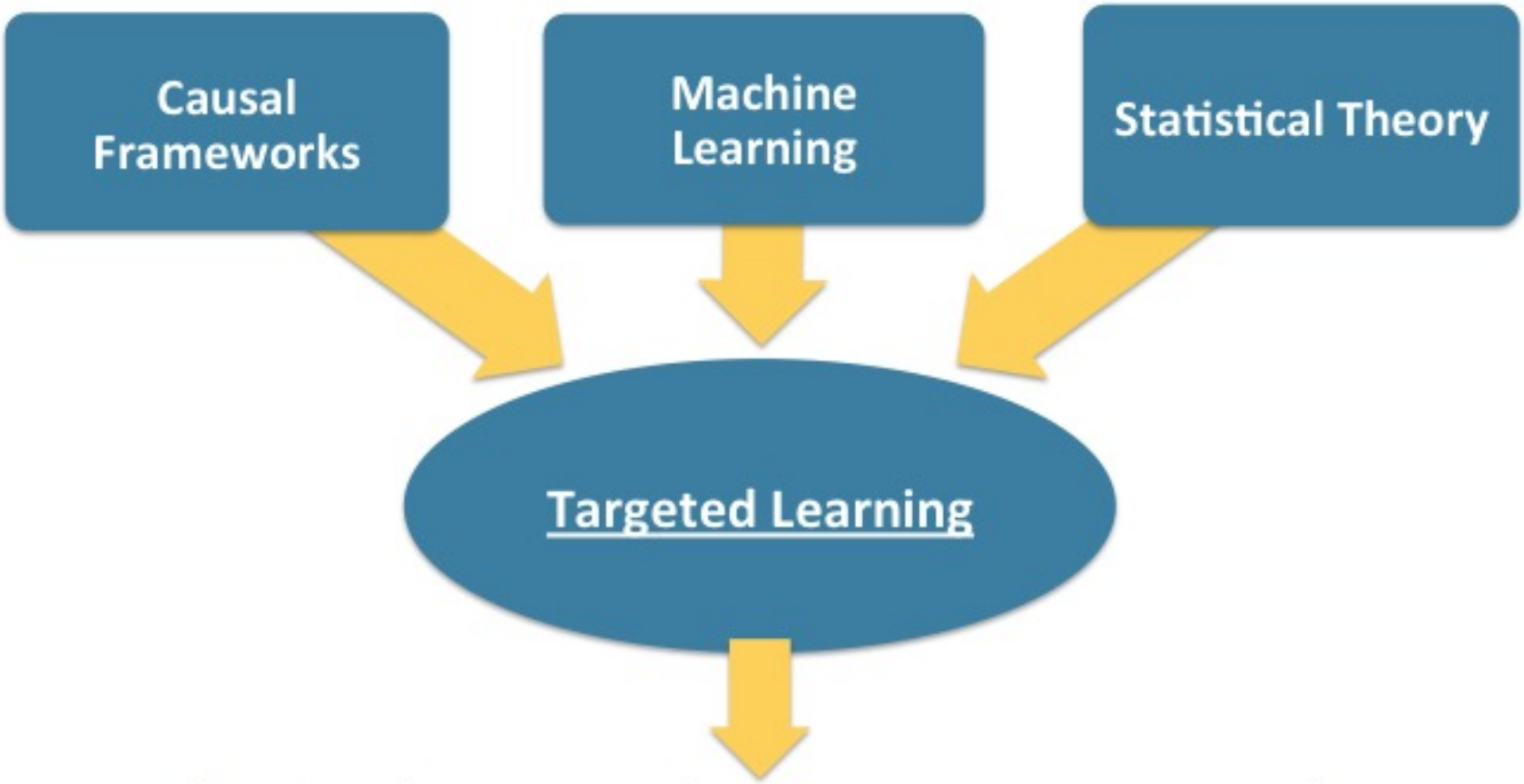
Roadmap for Causal Inference

- Define **Causal Quantity** and **realistic** causal model for underlying desired **full data** (*time ordering, outcome, treatment/censoring nodes, intervention specific counterfactuals, relations*, etc).
- Define and represent observed data as **missing/censored data/biased sample** on desired full data.
- Establish **identification** (i.e, estimand) of causal quantity from data probability distribution, under **non-testable** (e.g. no unmeasured confounders, MAR, CAR) assumptions.
- Commit to **target estimand** and **realistic statistical model** for data distribution.
- Develop a priori specified **estimator and inference** for target estimand: Targeted machine Learning.
- **Interpret results**, possibly with **sensitivity analysis**, concerning discrepancy between causal quantity and estimand.

Targeted Learning (TL)

is the subfield of statistics concerned with development of (targeted ML) estimators of the data distribution based on observed data with corresponding plug-in estimates and **confidence intervals** for the desired estimand, **based on realistic statistical models**.

By necessity, TL involves highly data adaptive estimation (e.g., super learning).

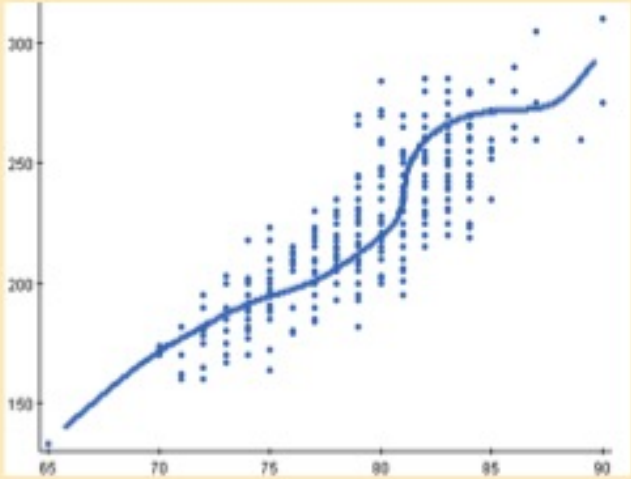


Better (more precise) **answers** to **causal** (actionable) **questions** with **accurate quantification of uncertainty** (signal from noise)

Targeted Learning

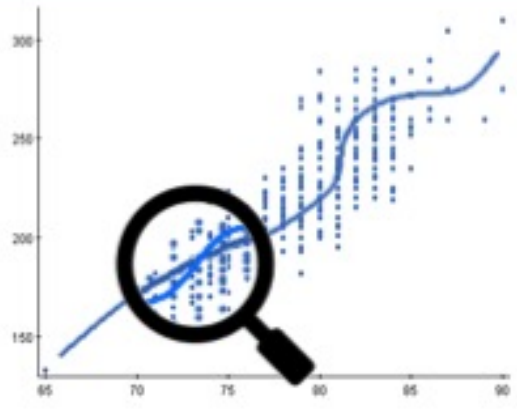
1. Super Learning fit of Stochastic system

- Set up a "competition"
  - Between models/algorithms
- Judge it honestly
  - Internal data splits
- Encourage teamwork
  - Learn from data how to build best team



2. Targeting (TMLE)

- Focus on causal question
  - One aspect of the data—the target
- Update Super learning estimate of stochastic system to give best answer for the causal query



Highly Adaptive Lasso (HAL)

- This is a machine learning algorithm that estimates functionals (e.g outcome regression and propensity score) by approximating them with linear model in many indicator basis functions.
- Guaranteed to converge to truth at rate faster than $n^{-1/4}$.
- When used in super-learner library, TMLE is guaranteed **consistent, (double robust) asymptotically normal and efficient**: one only needs to assume *strong positivity assumption*.

Use outcome data to learn propensity score/censoring mechanism

- We fit the treatment and censoring mechanism so that it results in best estimate of target estimand; Collaborative TMLE combined with outcome adaptive HAL-TMLE (ideas from Susan Shortreed, Ashkan Ertefaie).
- Both variable selection as well as bias-variance trade-off is based on outcome data.
- Better than methods ignoring outcome data.

Simulation HAL-C-TMLE Kang Shafer 2007

We follow the simulation from Kang, Shafer, 2007. The original pre-treatment covariates $(Z_{i1}, \cdots, Z_{i4})$ are generated from uncorrelated standard normal distribution.
With the pre-treatment covariates, the treatment indicator is then generated from a Bernoulli distribution with:

$$P(A_i = 1|Z_i) = \text{Expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4})$$

Only transformed covariates W are provided:

$$\begin{aligned} W_{i1} &= \exp(Z_{i1}/2) \\ W_{i2} &= z_{i2}/(1 + \exp(Z_{i1}) + 10 \\ W_{i3} &= (Z_{i1}Z_{i3}/25 + 0.6)^3 \\ W_{i4} &= (Z_2 + Z_4 + 20)^2. \end{aligned}$$

The continuous potential outcomes are generated by the linear combination of the pre-treatment covariates, and does not rely on the treatment A. More specific, the potential outcome is generated by:

$$Y_i^{(1)} = Y_i^{(0)} = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + \epsilon$$
$$\epsilon \sim N(0, 1)$$

Thus the ATE, equals 0

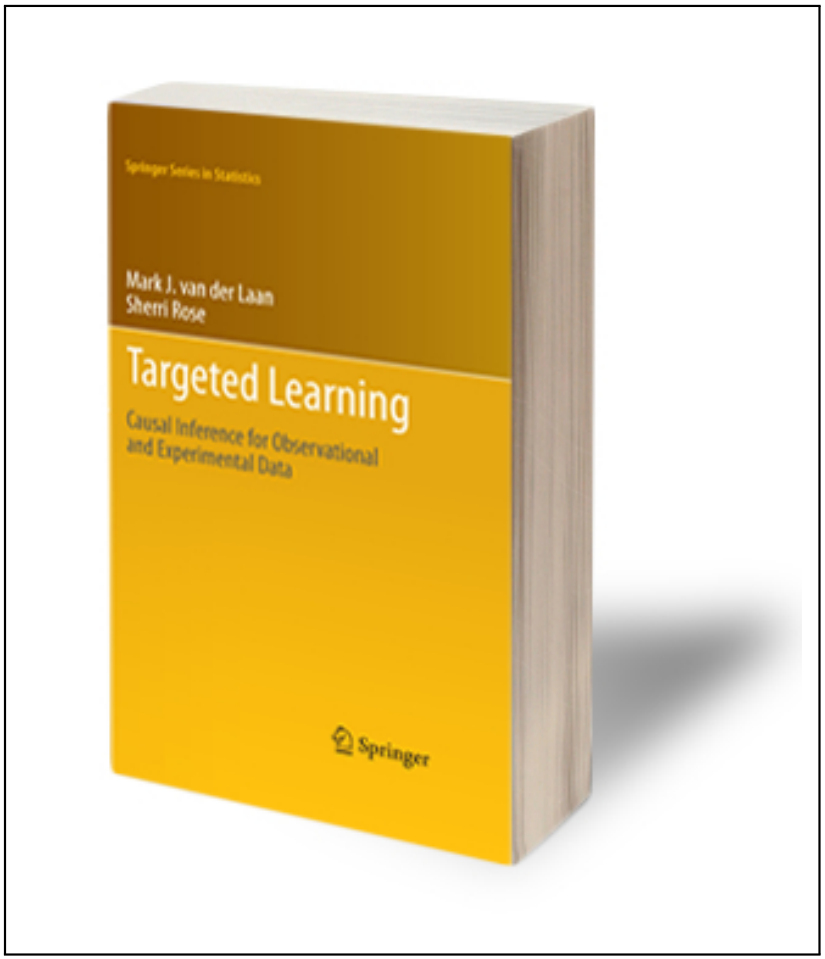
Simulation Results

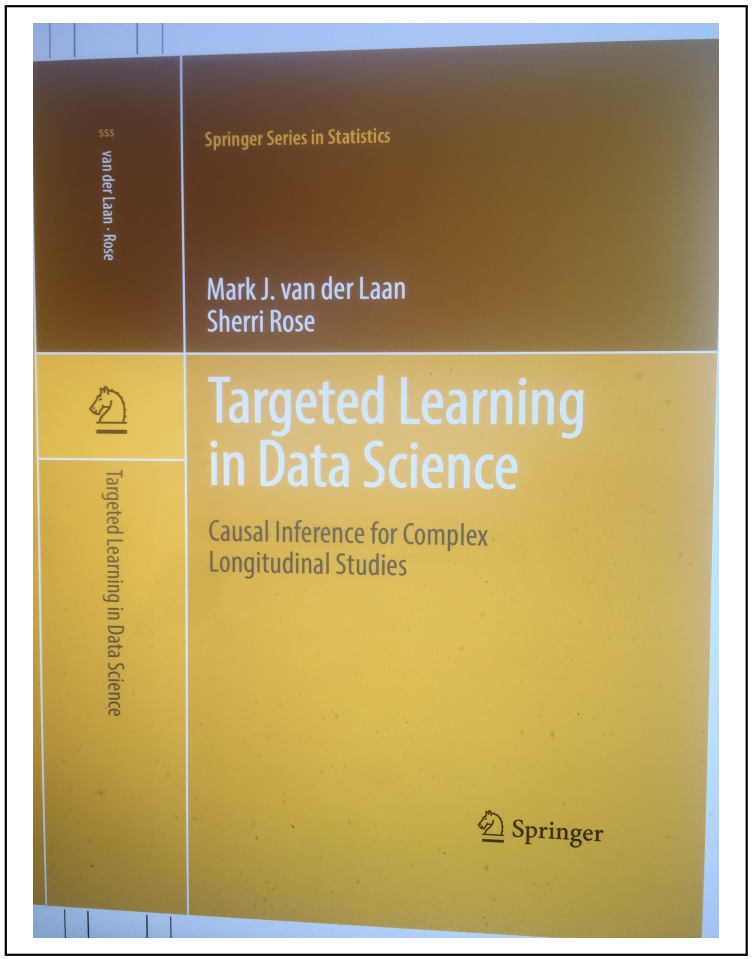
We use main-term linear regression to create biased outcome regression for TMLE/C-TMLE.

	TMLE-HAL	CTMLE-HAL	CTMLE-OHAL	oracle
N=500	7.06	6.34	**2.94**	3.64
N=1000	4.42	3.55	**1.40**	1.70
N=2000	2.94	1.85	**0.83**	0.87

Table: This table reports MSE for each estimator across 200 replications with different sample size. We can see CTMLE-HAL improves TMLE-HAL. CTMLE-OHAL further improves CTMLE-HAL with outcome adaptive regularization. It even achieves better performance than the oracle estimator, which is defined as TMLE with true PS.

Targeted Learning (targetedlearningbook.com)





van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.