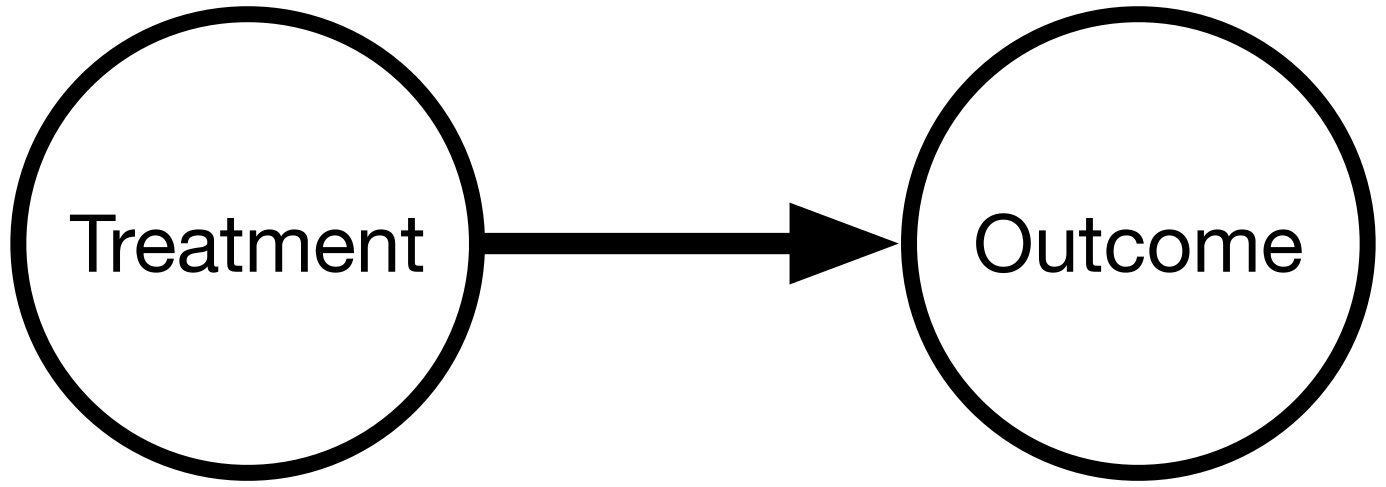
How and Why to Evaluate Causal Inference Methods Using Experimental Data

Amanda Gentzel, Purva Pruthi, and David Jensen
Knowledge Discovery Laboratory
College of Information and Computer Science
University of Massachusetts Amherst

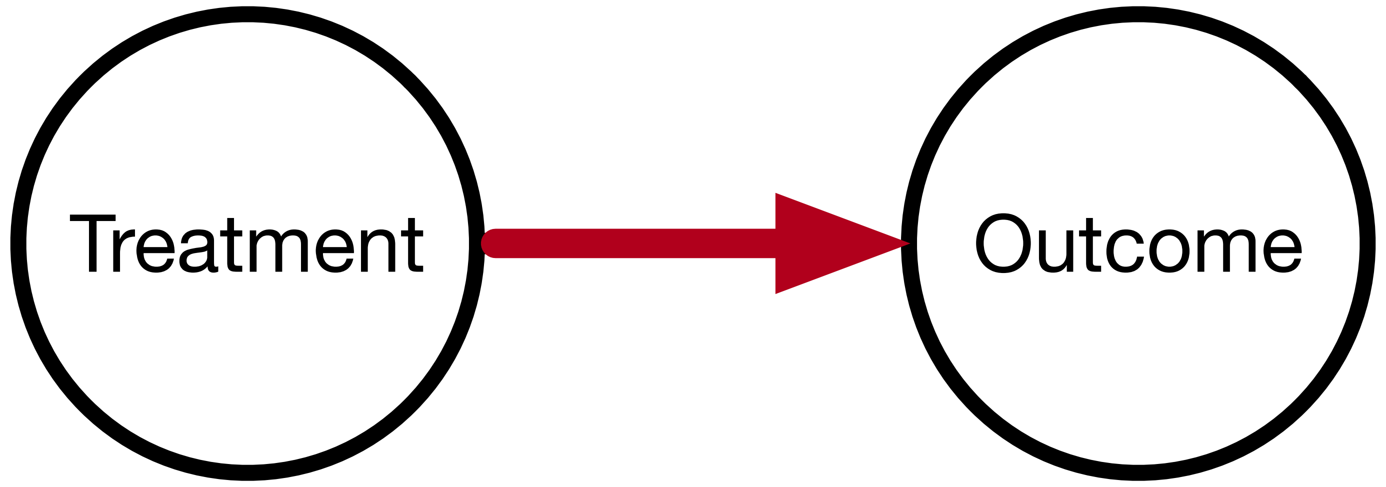
Causal modeling from observational data

- Data from observing a system, rather than experimenting



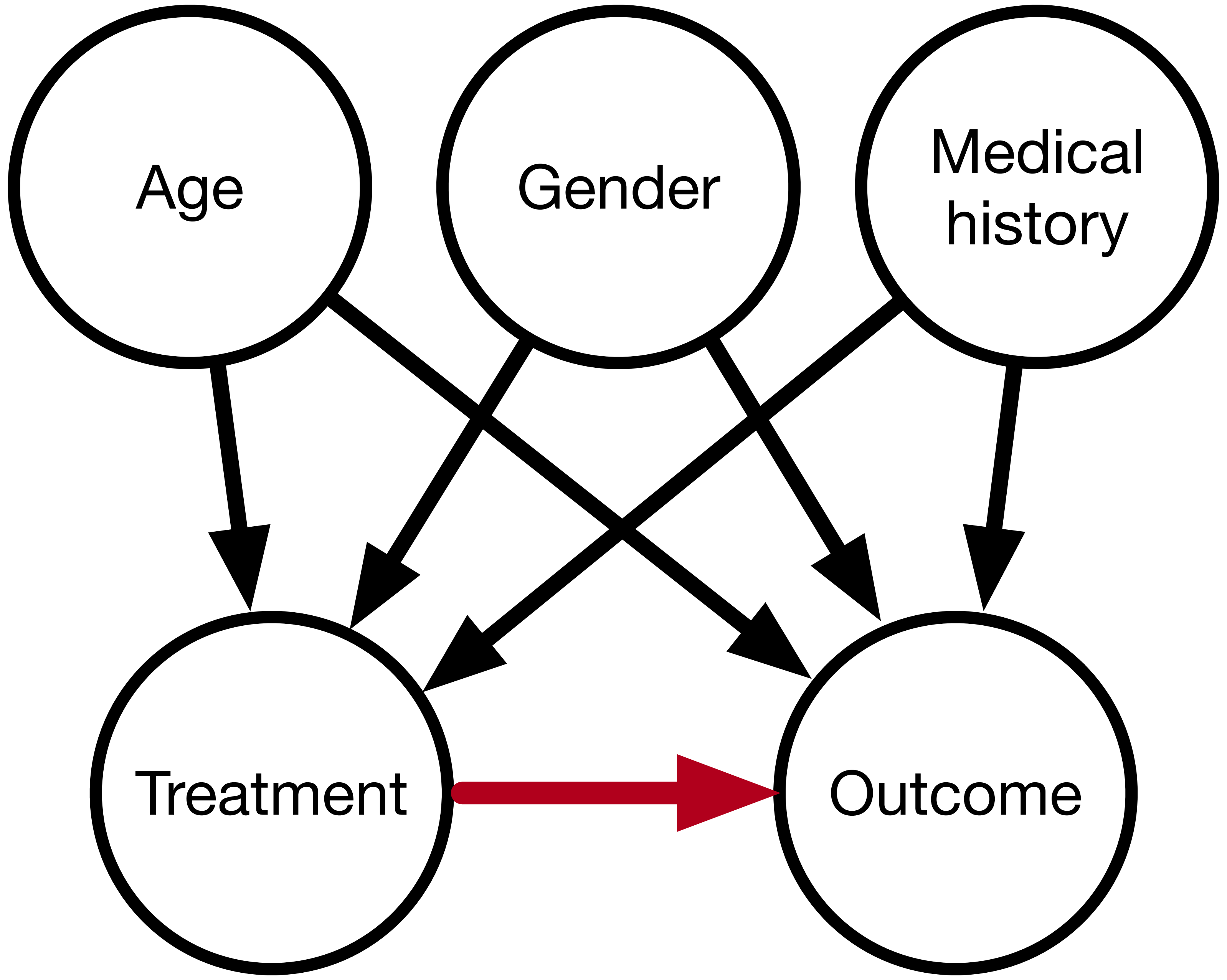
Causal modeling from observational data

- Data from observing a system, rather than experimenting



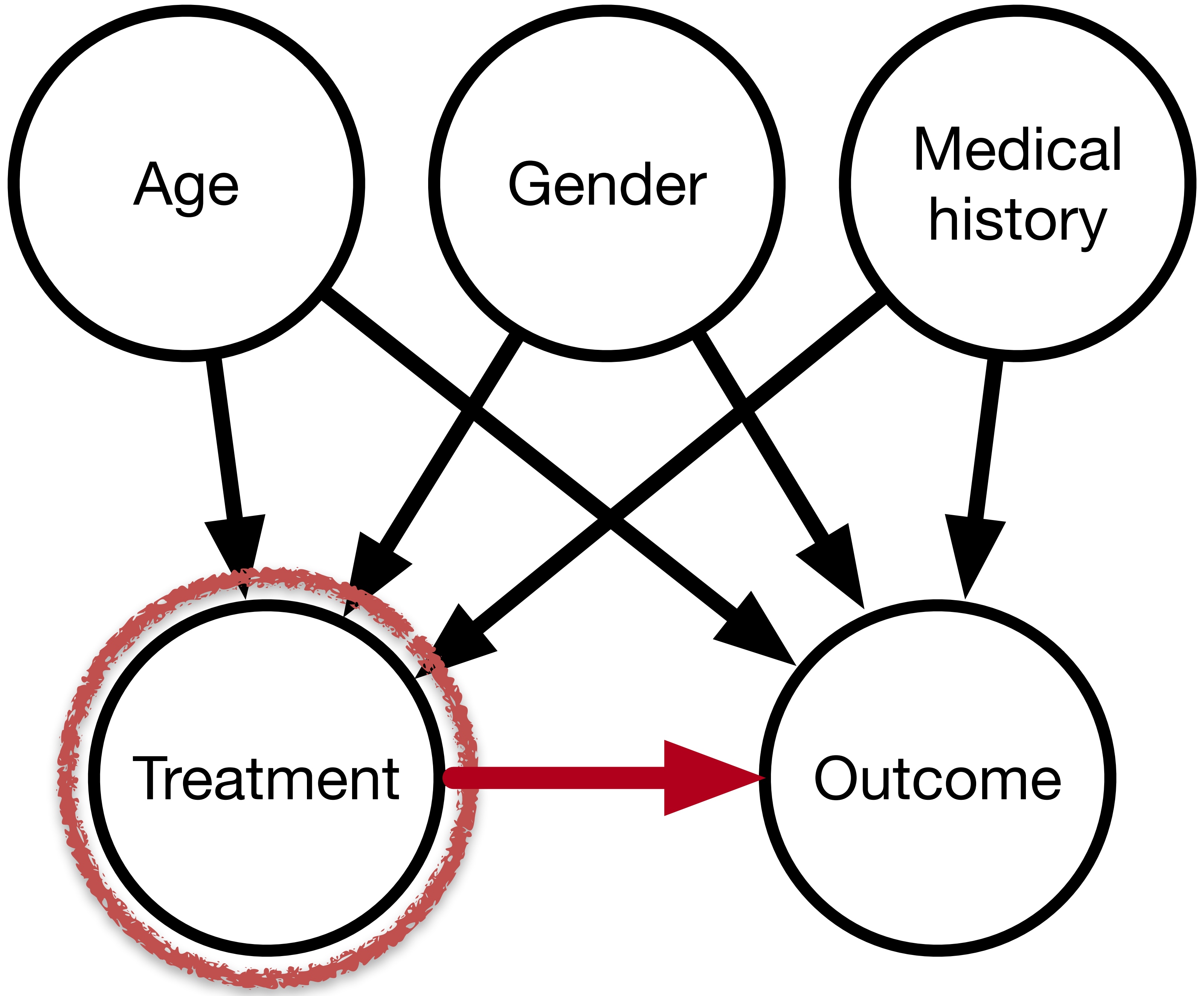
Causal modeling from observational data

- Data from observing a system, rather than experimenting



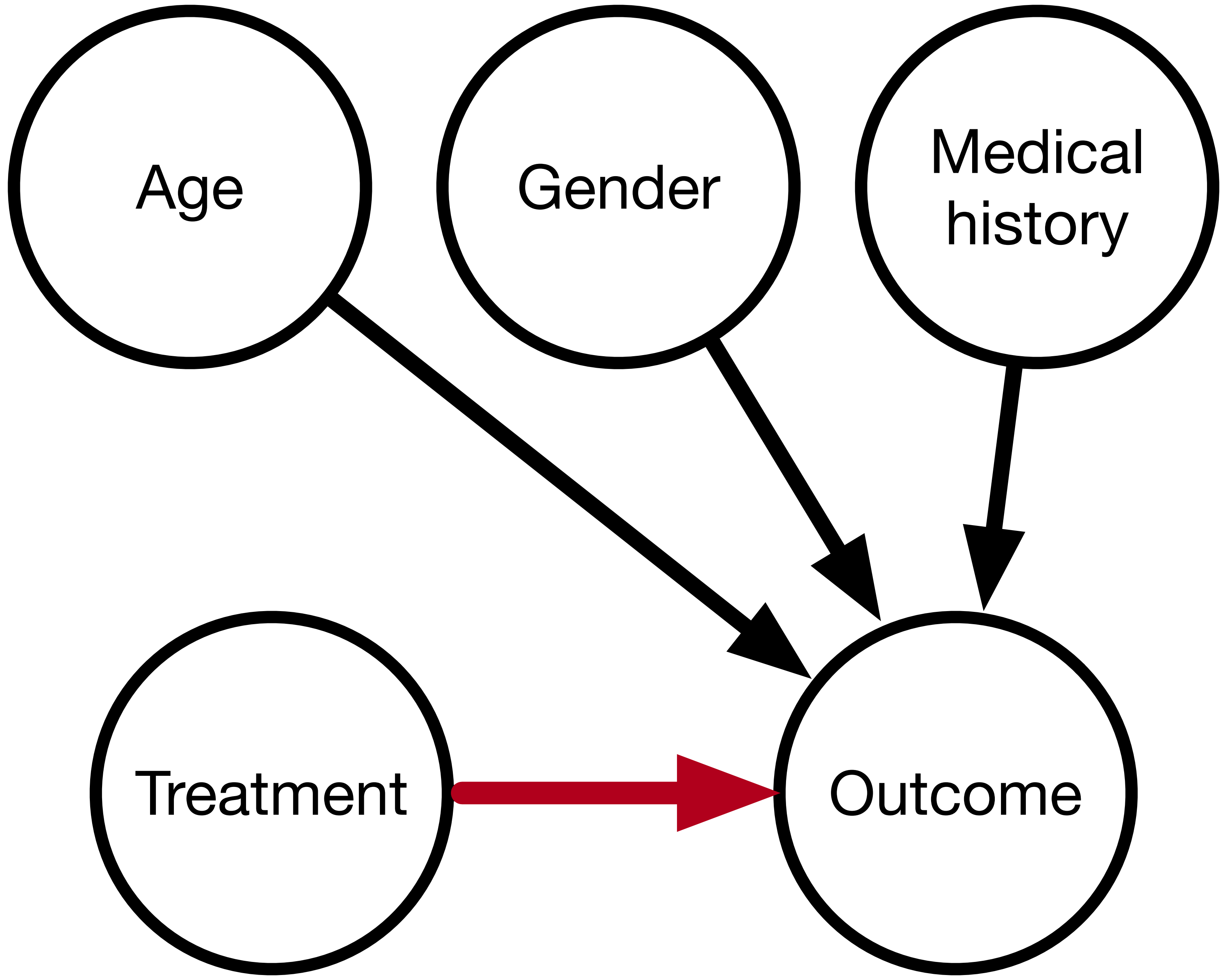
Causal modeling from observational data

- Data from observing a system, rather than experimenting



Causal modeling from observational data

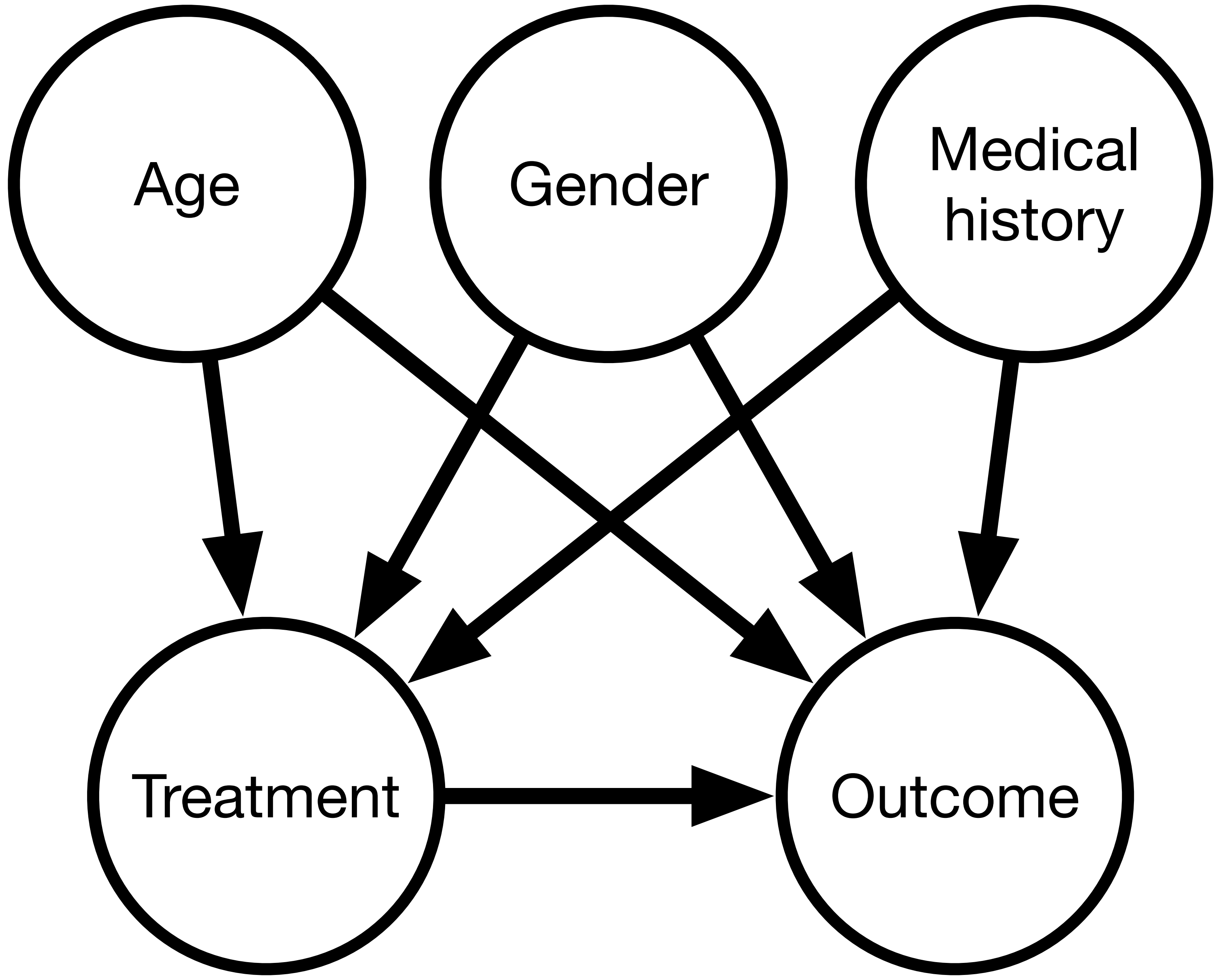
- Data from observing a system, rather than experimenting



Causal inference from observational data

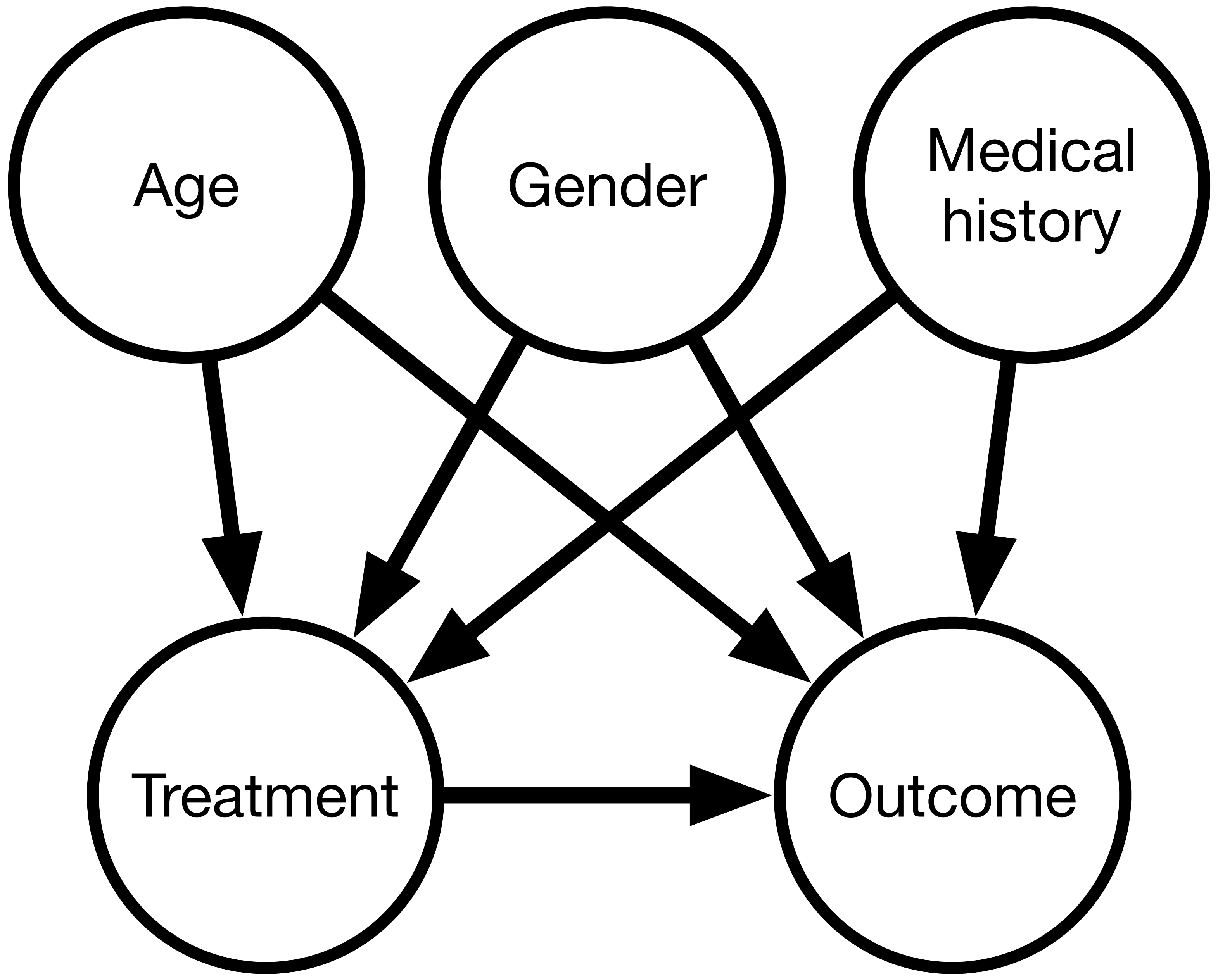
Causal inference from observational data

Data generating model

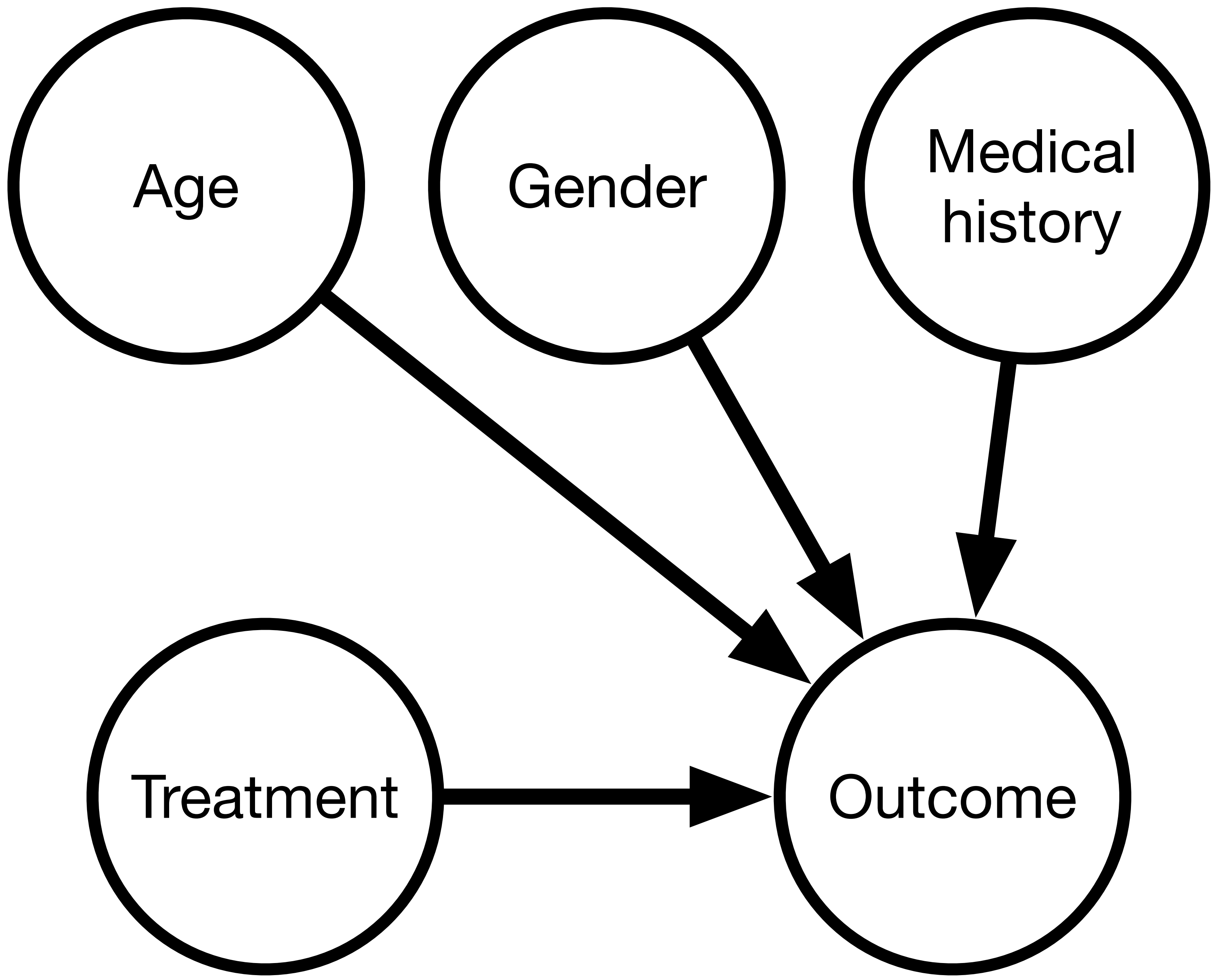


Causal inference from observational data

Data generating model



Inference model



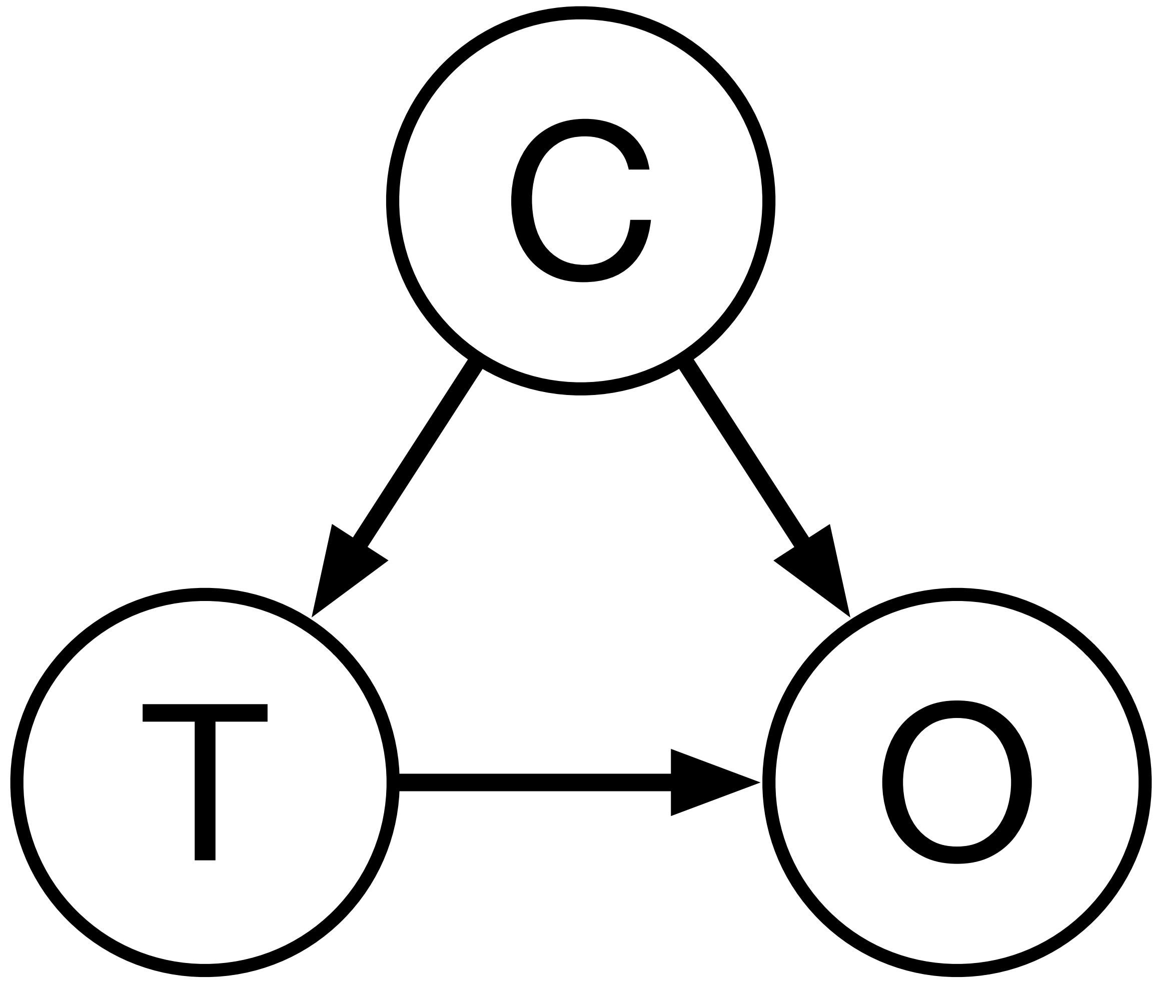
What data do we need?

Observational

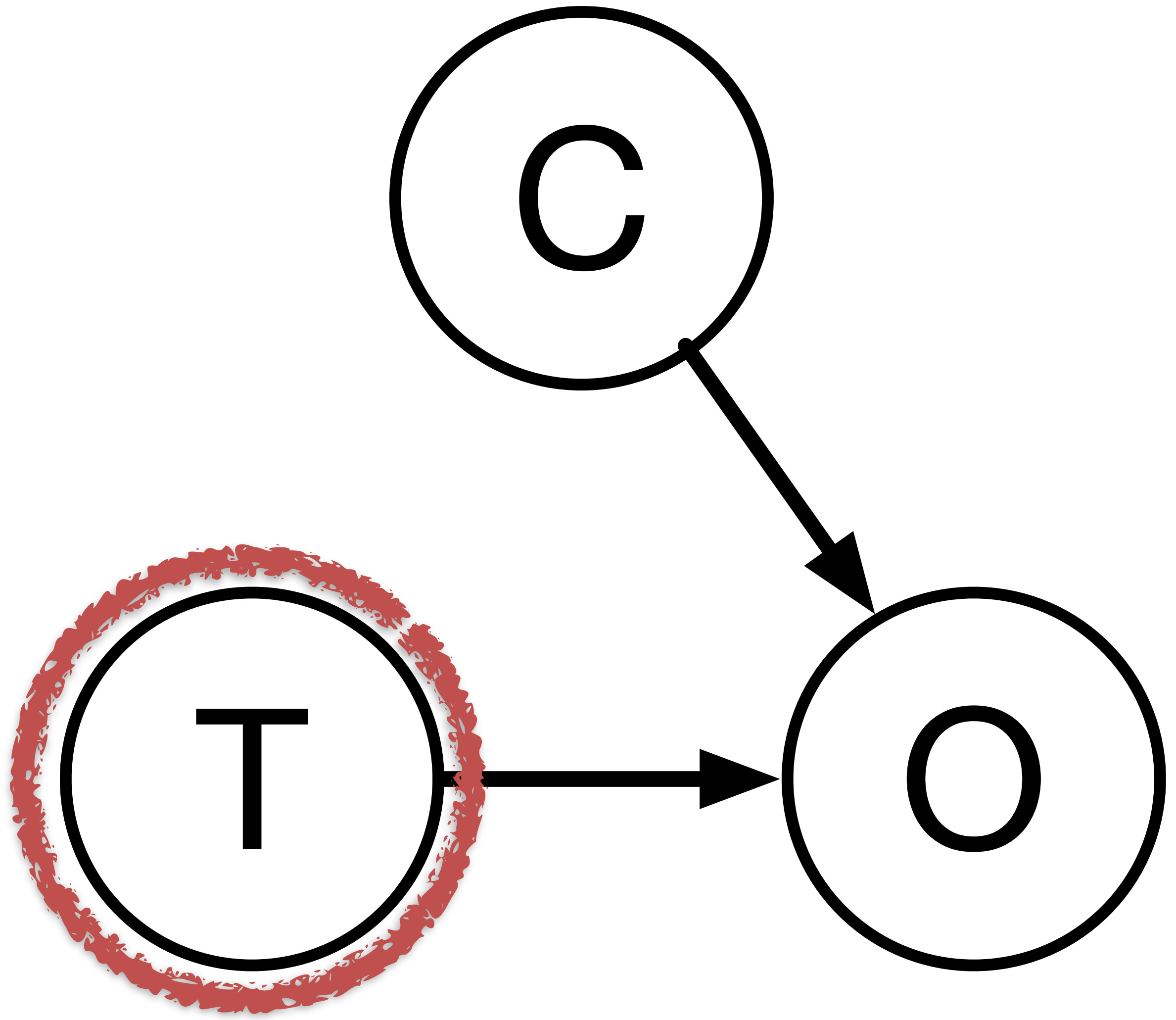
Experimental

What data do we need?

Observational



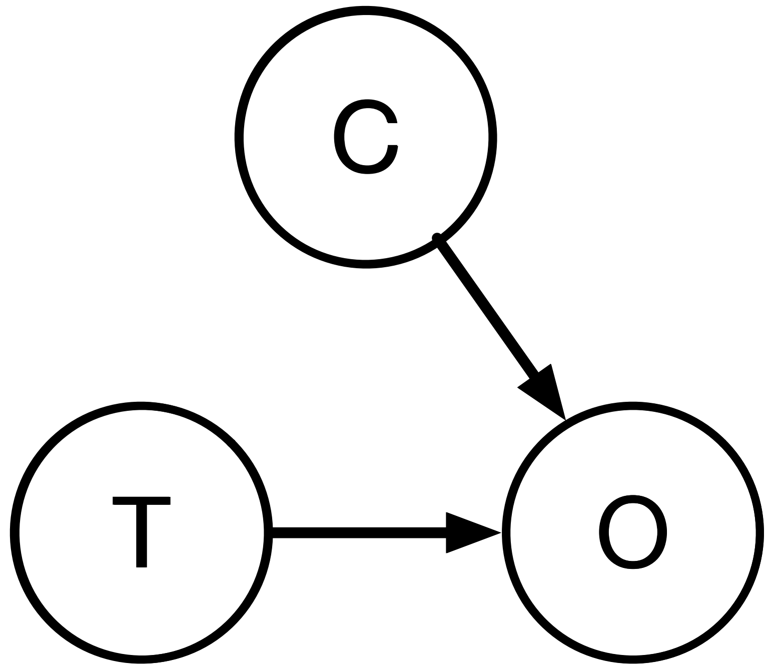
Experimental



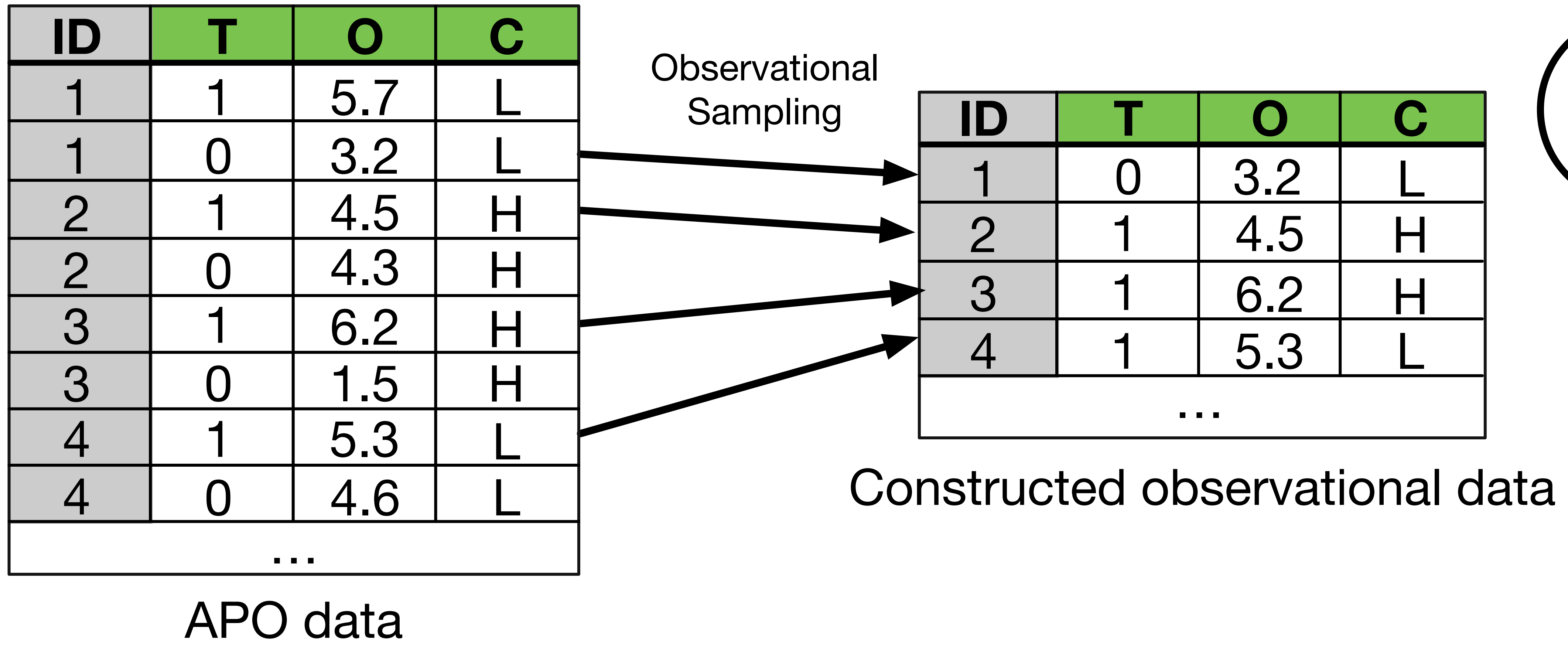
Observational Sampling from APO data (OSAPO)

ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

APO data



Observational Sampling from APO data (OSAPO)



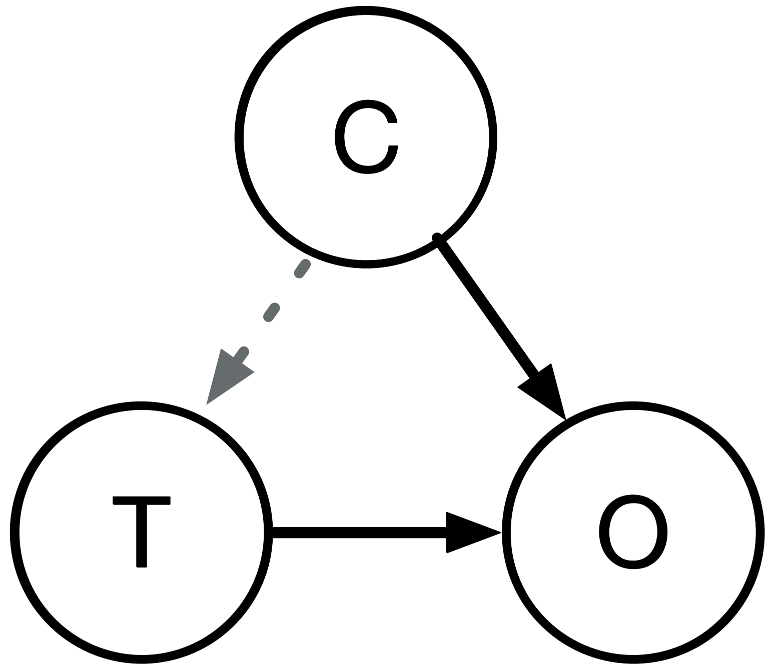
ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

APO data

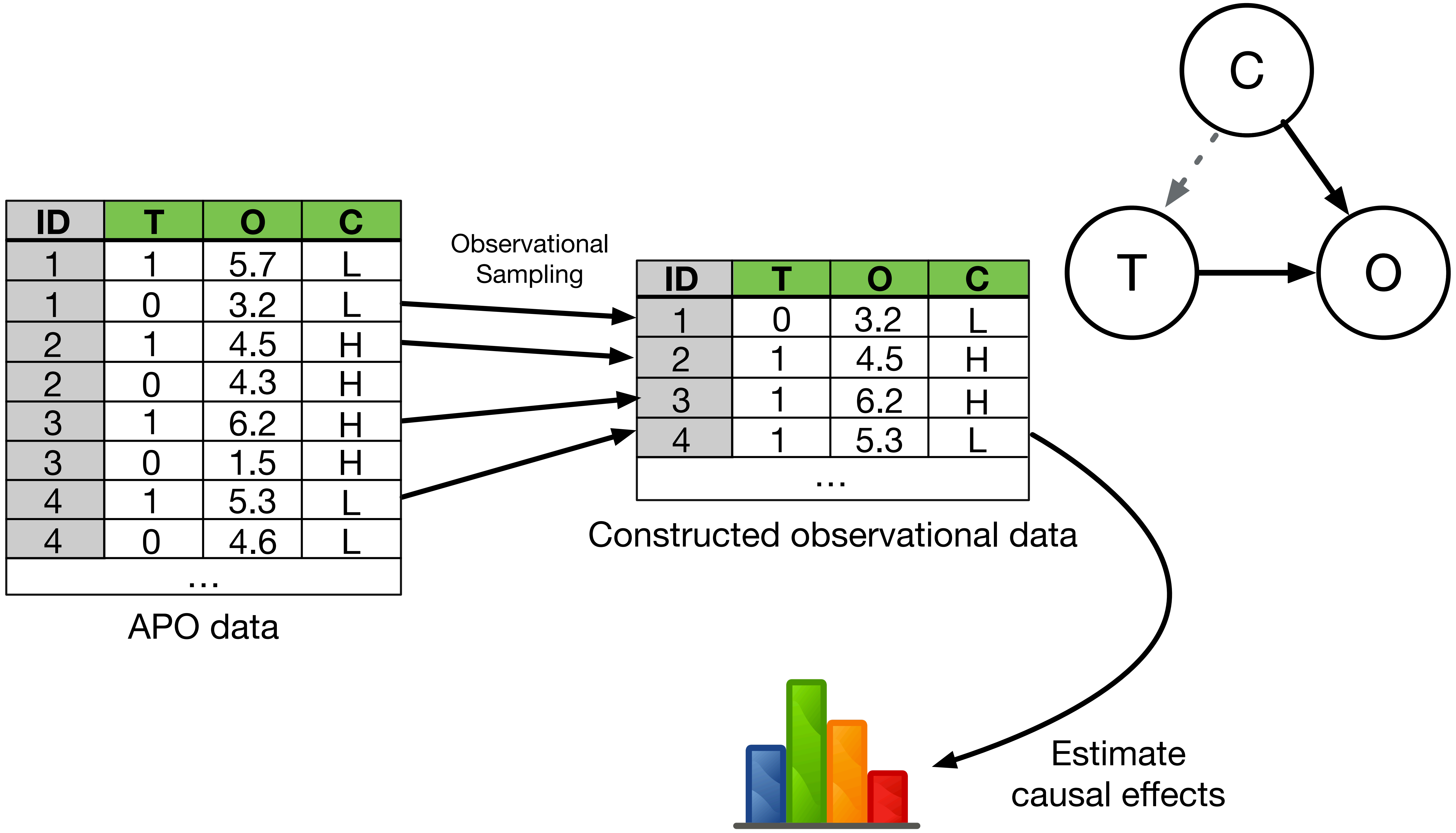
Observational Sampling

ID	T	O	C
1	0	3.2	L
2	1	4.5	H
3	1	6.2	H
4	1	5.3	L
…			

Constructed observational data



Observational Sampling from APO data (OSAPO)



ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

APO data

Observational Sampling

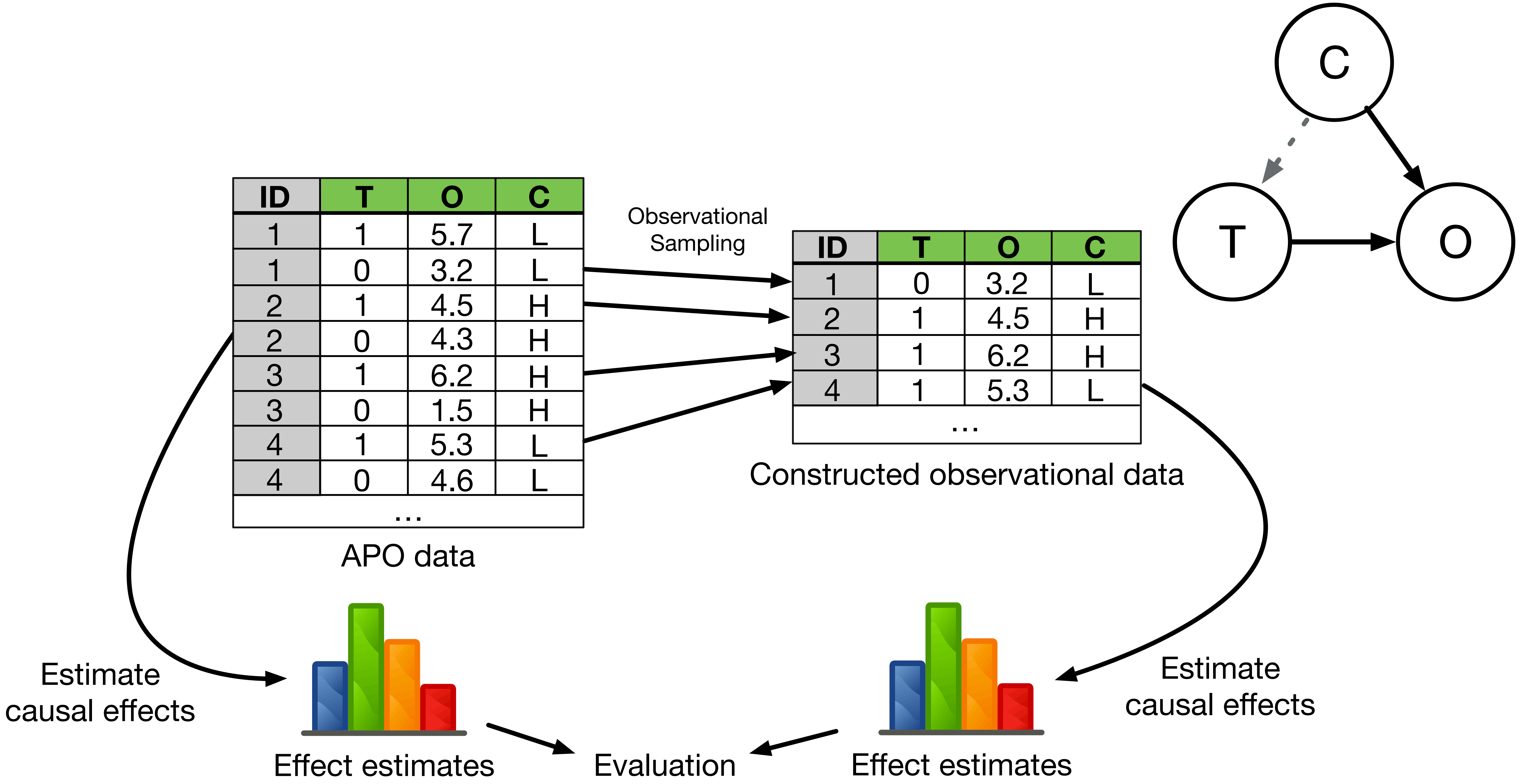
ID	T	O	C
1	0	3.2	L
2	1	4.5	H
3	1	6.2	H
4	1	5.3	L
…			

Constructed observational data

Estimate causal effects

Effect estimates

Observational Sampling from APO data (OSAPO)



ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

APO data

Observational Sampling

ID	T	O	C
1	0	3.2	L
2	1	4.5	H
3	1	6.2	H
4	1	5.3	L
…			

Constructed observational data

Estimate causal effects

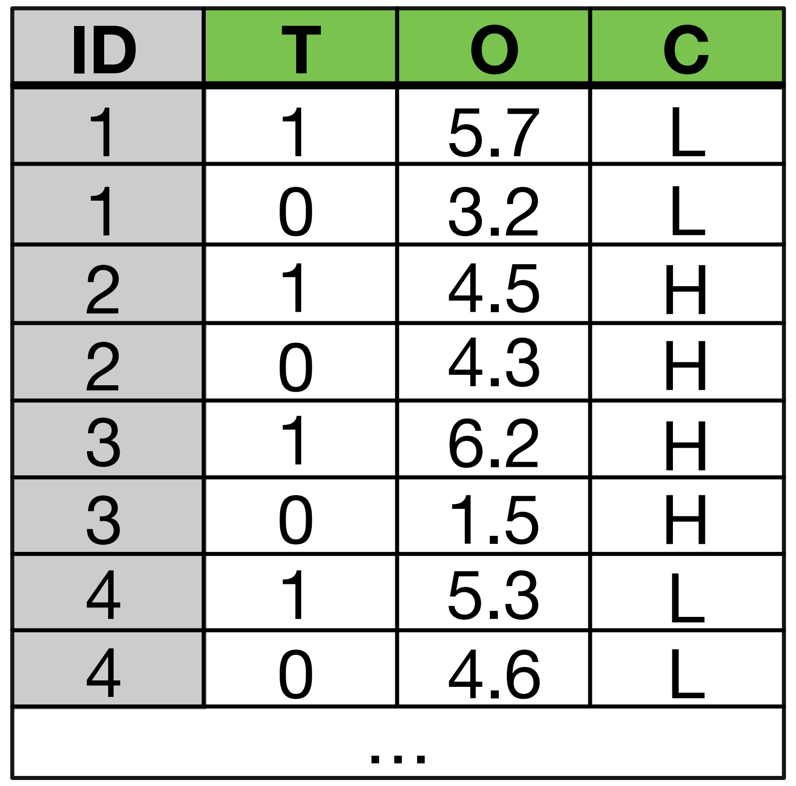
Effect estimates

Evaluation

Effect estimates

Estimate causal effects

Observational sampling from RCT data (OSRCT)



ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

APO data

Observational sampling from RCT data (OSRCT)

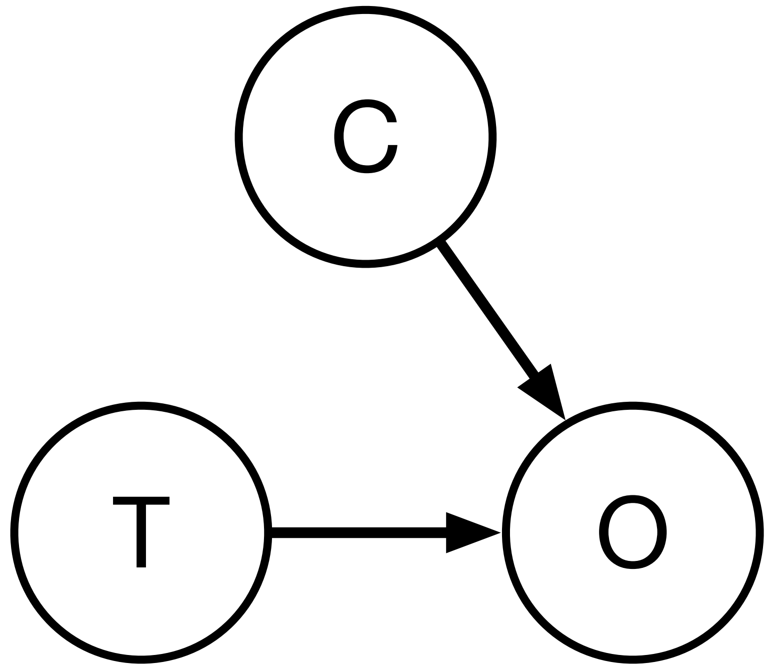
ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

RCT data

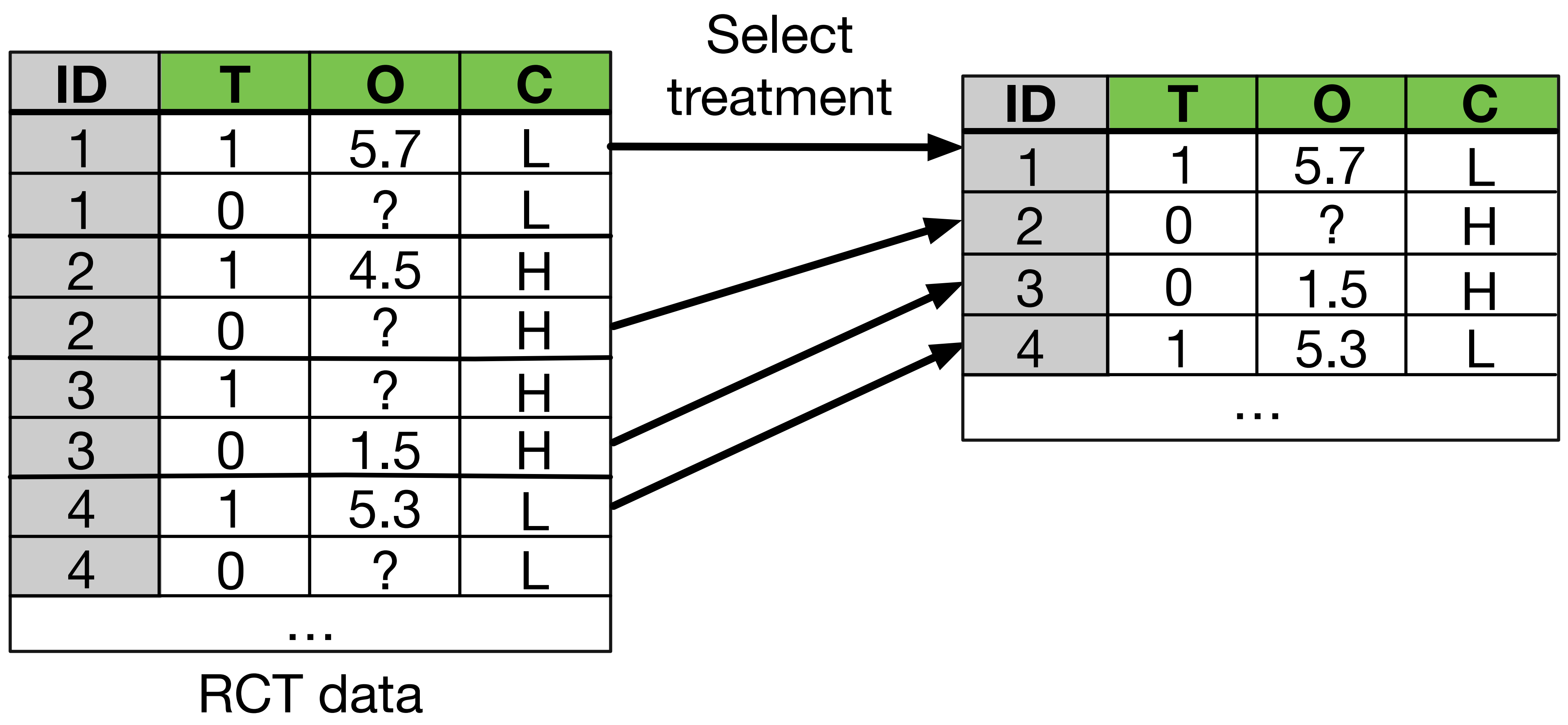
Observational sampling from RCT data (OSRCT)

ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
...			

RCT data



Observational sampling from RCT data (OSRCT)

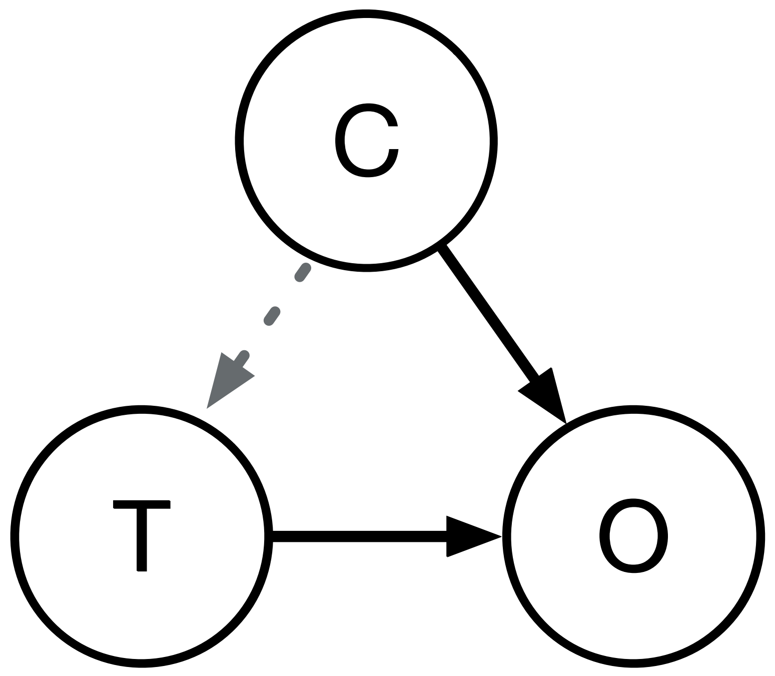


ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

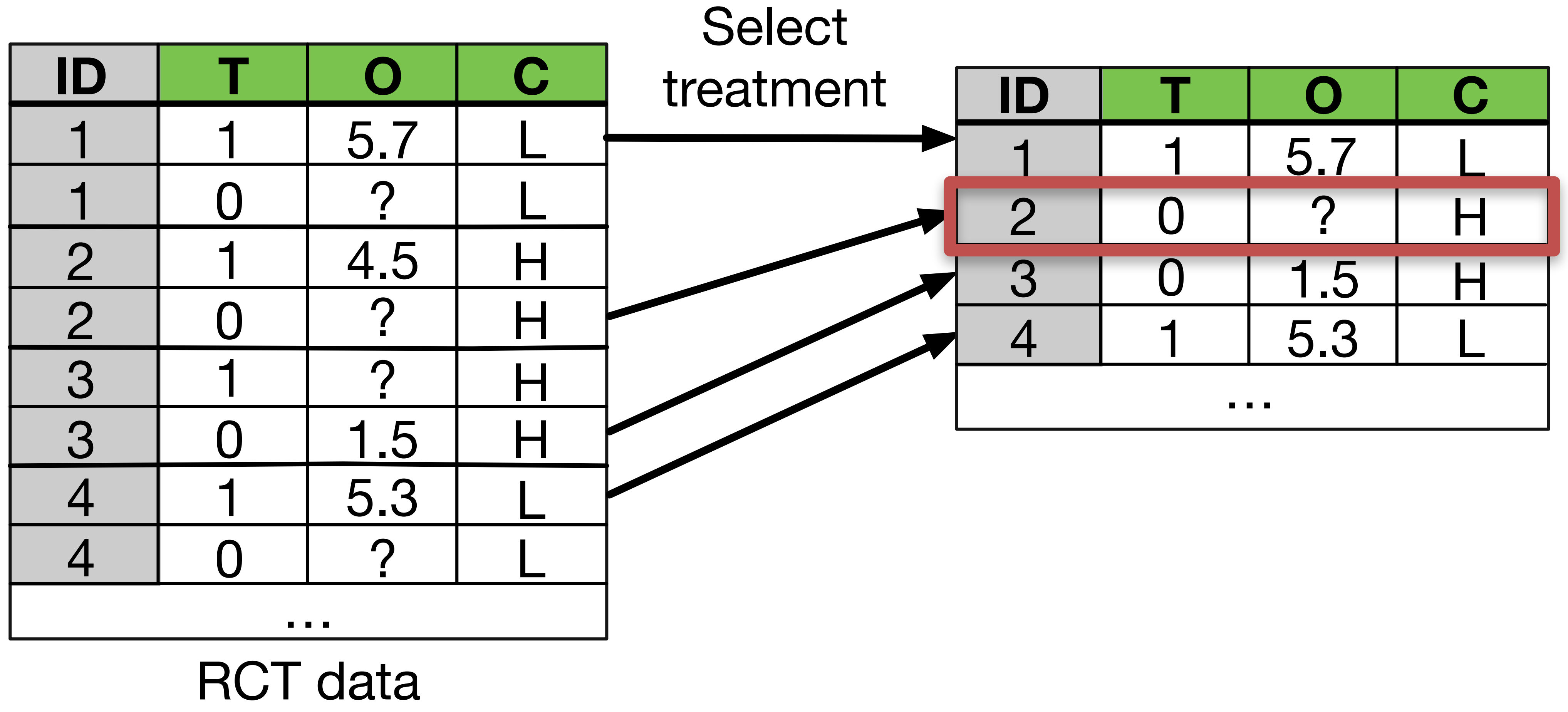
RCT data

Select treatment

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	5.3	L
…			



Observational sampling from RCT data (OSRCT)

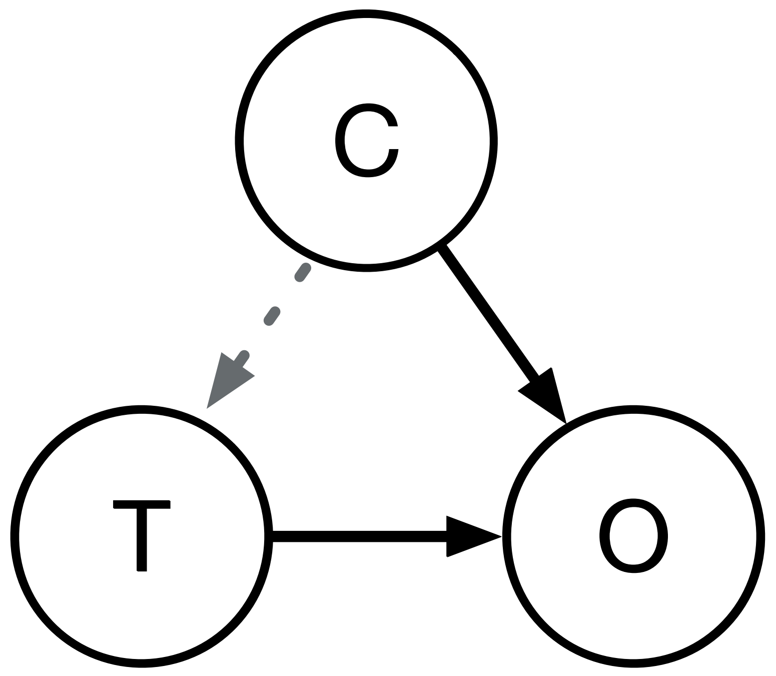


ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

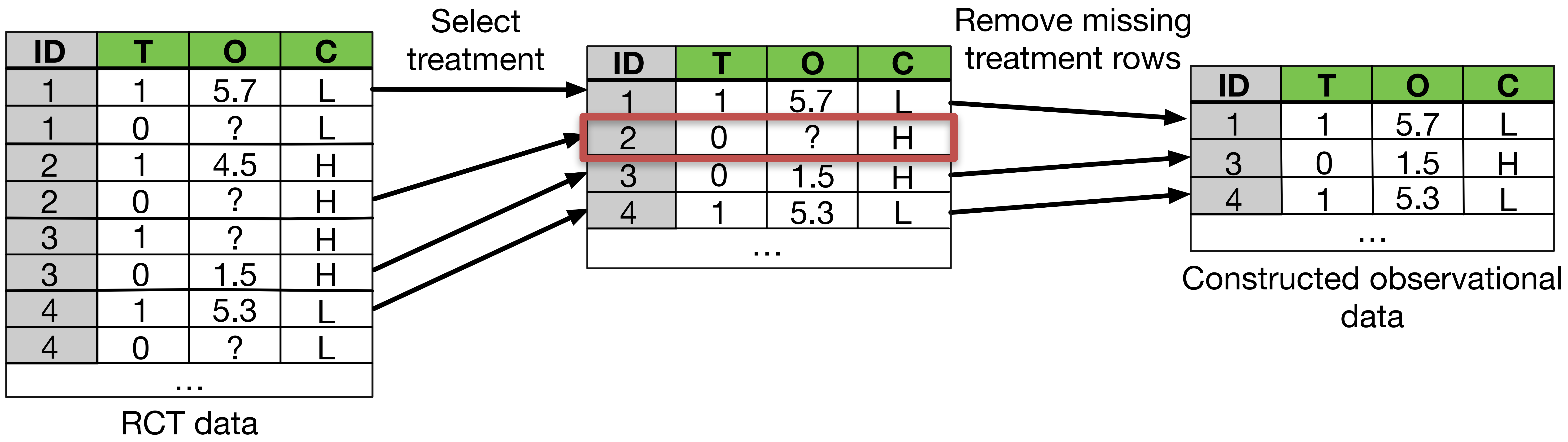
RCT data

Select treatment

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	5.3	L
…			



Observational sampling from RCT data (OSRCT)



ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

RCT data

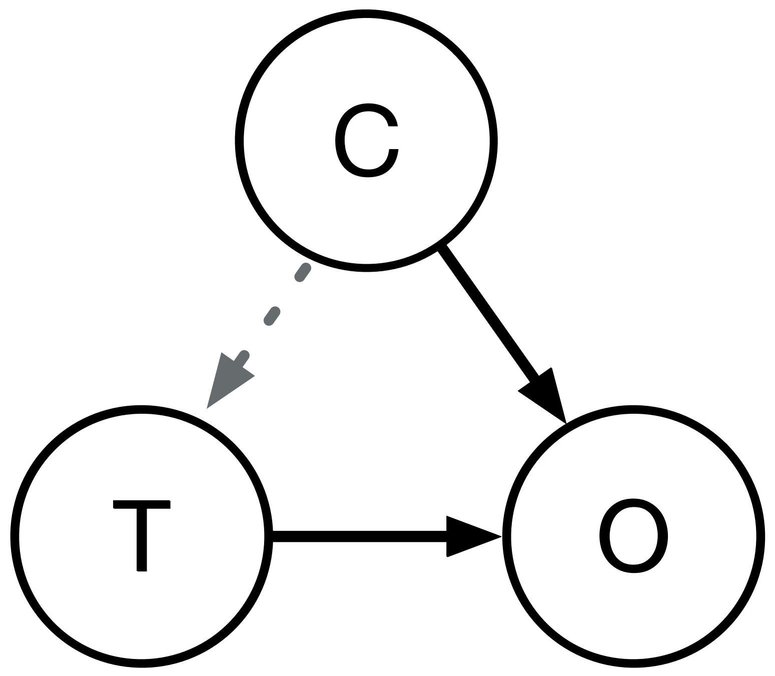
Select treatment

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	5.3	L
…			

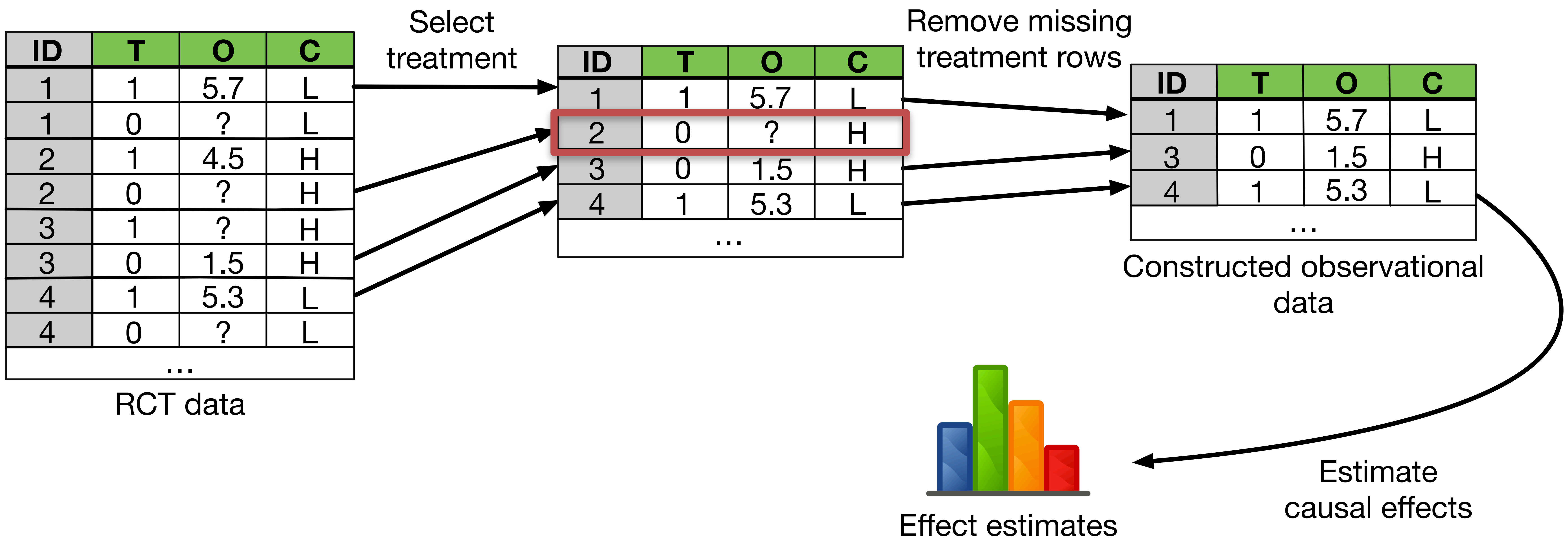
Remove missing treatment rows

ID	T	O	C
1	1	5.7	L
3	0	1.5	H
4	1	5.3	L
…			

Constructed observational data



Observational sampling from RCT data (OSRCT)



ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

RCT data

Select treatment

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	5.3	L
…			

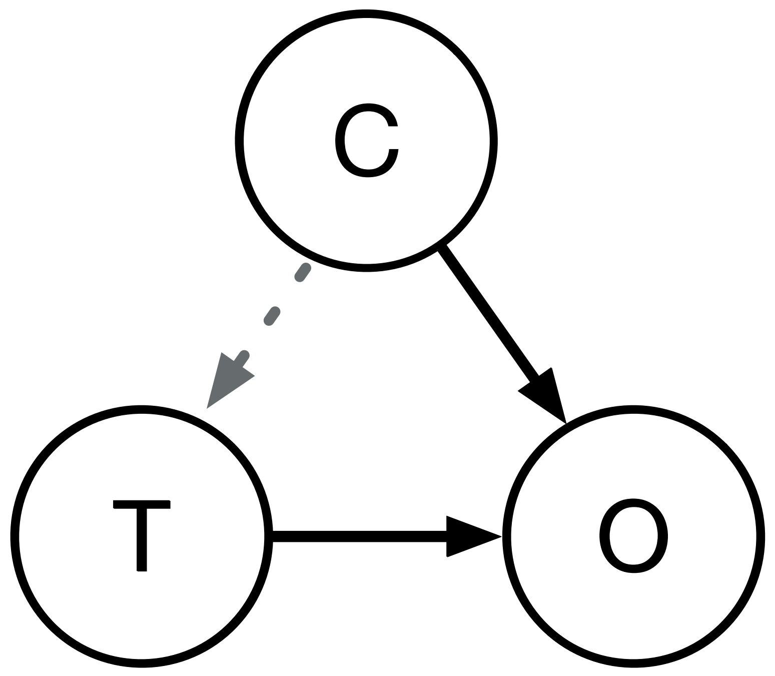
Remove missing treatment rows

ID	T	O	C
1	1	5.7	L
3	0	1.5	H
4	1	5.3	L
…			

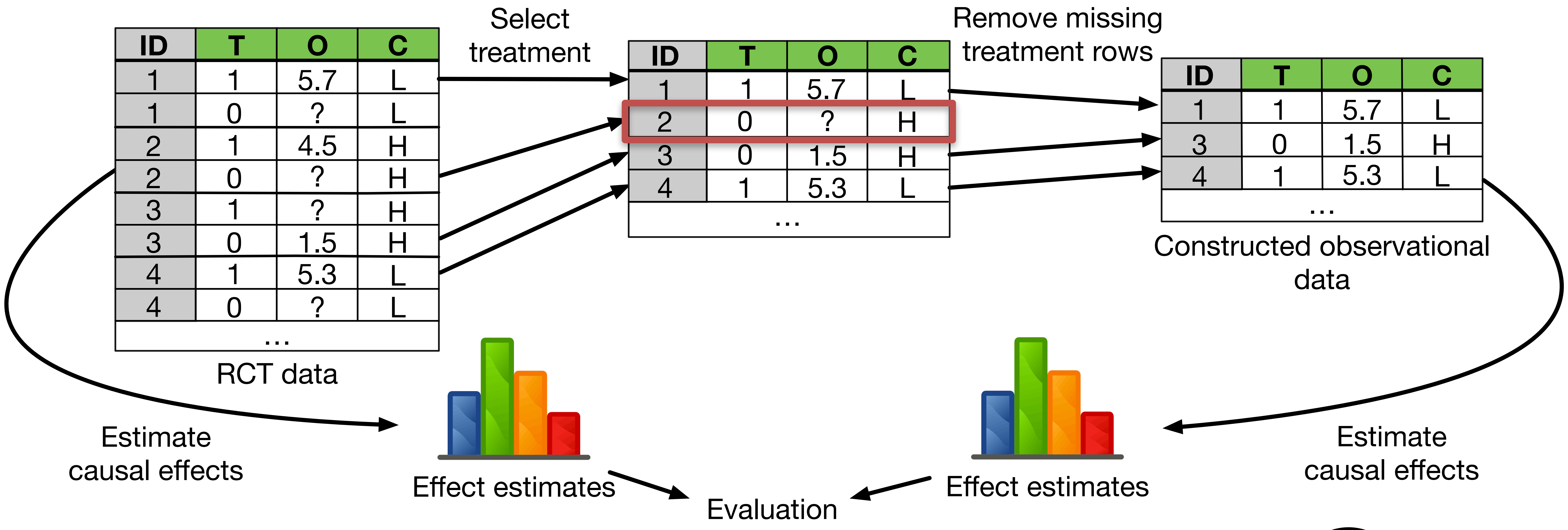
Constructed observational data

Estimate causal effects

Effect estimates



Observational sampling from RCT data (OSRCT)



ID	T	O	C
1	1	5.7	L
1	0	?	L
2	1	4.5	H
2	0	?	H
3	1	?	H
3	0	1.5	H
4	1	5.3	L
4	0	?	L
…			

RCT data

Select treatment

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	5.3	L
…			

Remove missing treatment rows

ID	T	O	C
1	1	5.7	L
3	0	1.5	H
4	1	5.3	L
…			

Constructed observational data

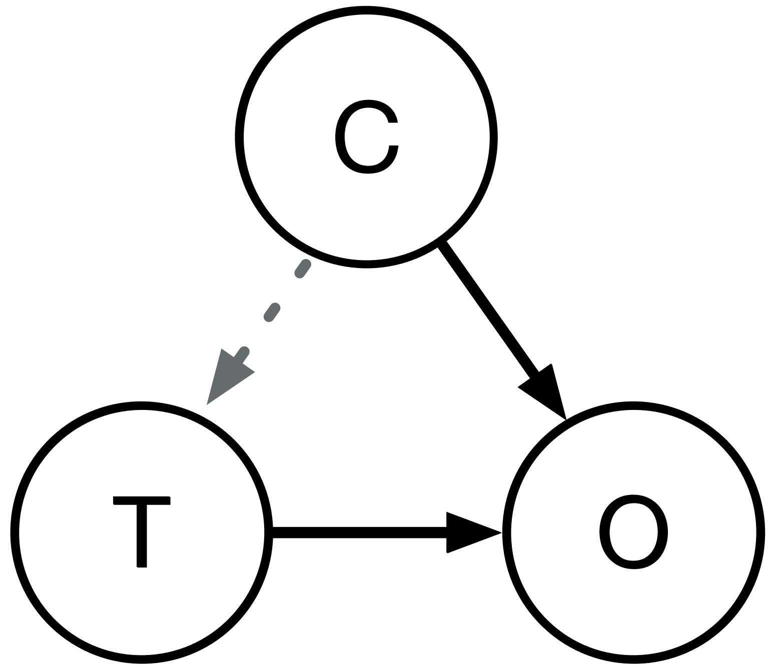
Estimate causal effects

Effect estimates

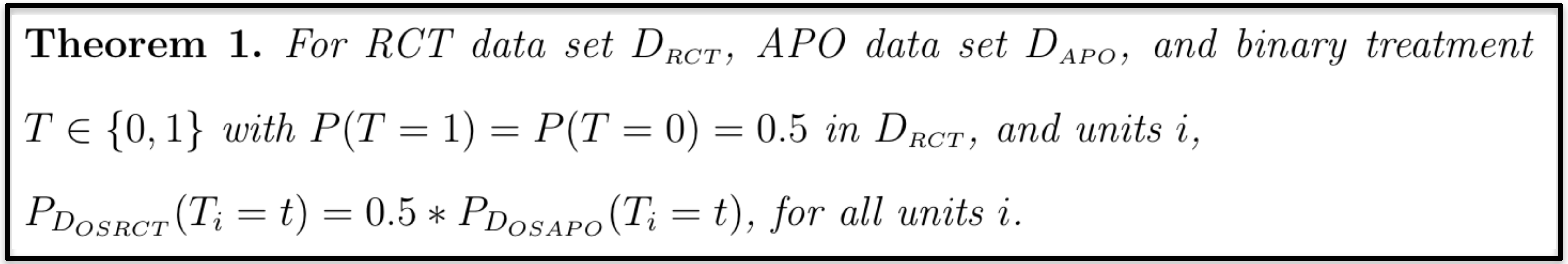
Evaluation

Effect estimates

Estimate causal effects



Using RCT data

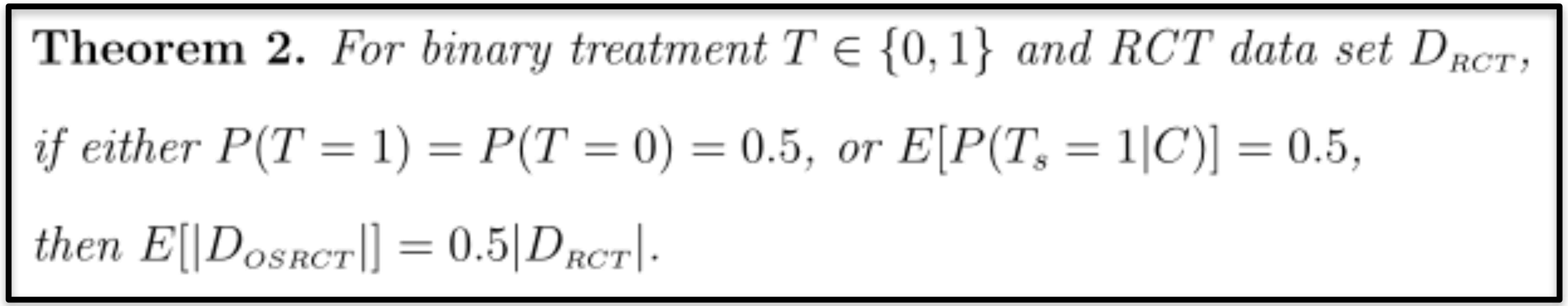


Theorem 1. *For RCT data set* D_{RCT}, *APO data set* D_{APO}, *and binary treatment* $T \in \{0, 1\}$ *with* $P(T = 1) = P(T = 0) = 0.5$ *in* D_{RCT}, *and units* i,

$P_{D_{OSRCT}}(T_i = t) = 0.5 * P_{D_{OSAPO}}(T_i = t)$, *for all units* i.

Observational sampling of RCT data is equivalent to observational sampling of APO data

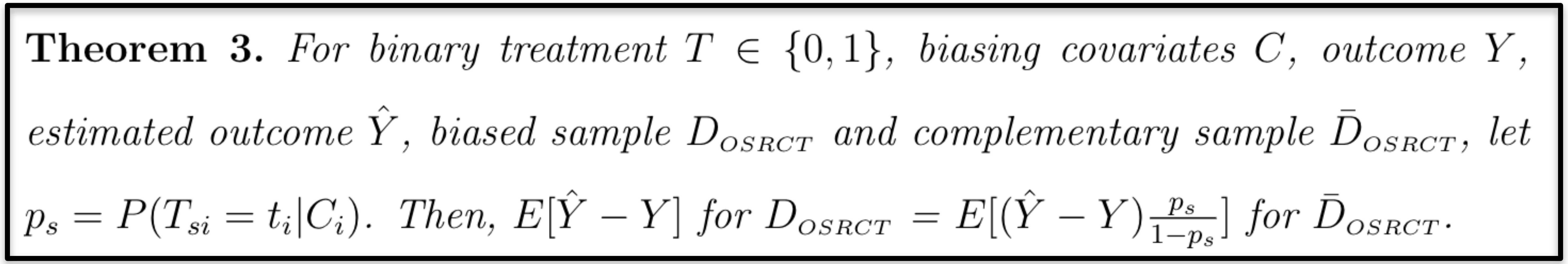
Other features of RCT data



Theorem 2. *For binary treatment* $T \in \{0, 1\}$ *and RCT data set* D_{RCT}*, if either* $P(T = 1) = P(T = 0) = 0.5,$ *or* $E[P(T_s = 1|C)] = 0.5,$ *then* $E[|D_{OSRCT}|] = 0.5|D_{RCT}|.$

In most cases, regardless of biasing strength, the sub-sampled data will be half the size of the original data

Other features of RCT data



Theorem 3. *For binary treatment* $T \in \{0,1\}$, *biasing covariates* C, *outcome* Y, *estimated outcome* $\hat{Y}$, *biased sample* D_{OSRCT} *and complementary sample* $\bar{D}_{\mathrm{OSRCT}}$, *let* $p_s = P(T_{si} = t_i | C_i)$. *Then,* $E[\hat{Y} - Y]$ *for* $D_{\mathrm{OSRCT}} = E[(\hat{Y} - Y)\frac{p_s}{1-p_s}]$ *for* $\bar{D}_{\mathrm{OSRCT}}$.

The data rejected during sub-sampling can be reweighted according to the biasing probability and used as a held-out test set

Data sources

- APO data (3)
    - computational systems
- RCT data (15)
    - publicly available RCTs
- Synthetic-response data (10)
    - ACIC 2016 Competition, IBM Causal Inference Benchmarking Framework
- Simulators (9)
    - Neuropathic Pain Simulator, Nemo, Whynot

Data sources

- APO data (3)
  - computational systems
- RCT data (15)
  - publicly available RCTs
- Synthetic-response data (10)
  - ACIC 2016 Competition, IBM Causal Inference Benchmarking Framework
- Simulators (9)
  - Neuropathic Pain Simulator, Nemo, Whynot

37 data sets

Algorithms

Focus on modeling treatment

- Propensity score matching (PSM)
- Inverse probability of treatment weighting (IPTW)
- Causal forests (CF)

Algorithms

Focus on modeling treatment

- Propensity score matching (PSM)
- Inverse probability of treatment weighting (IPTW)
- Causal forests (CF)

Focus on modeling outcome

- Outcome regression (OR)
- Bayesian additive regression trees (BART)

Algorithms

Focus on modeling treatment

- Propensity score matching (PSM)
- Inverse probability of treatment weighting (IPTW)
- Causal forests (CF)

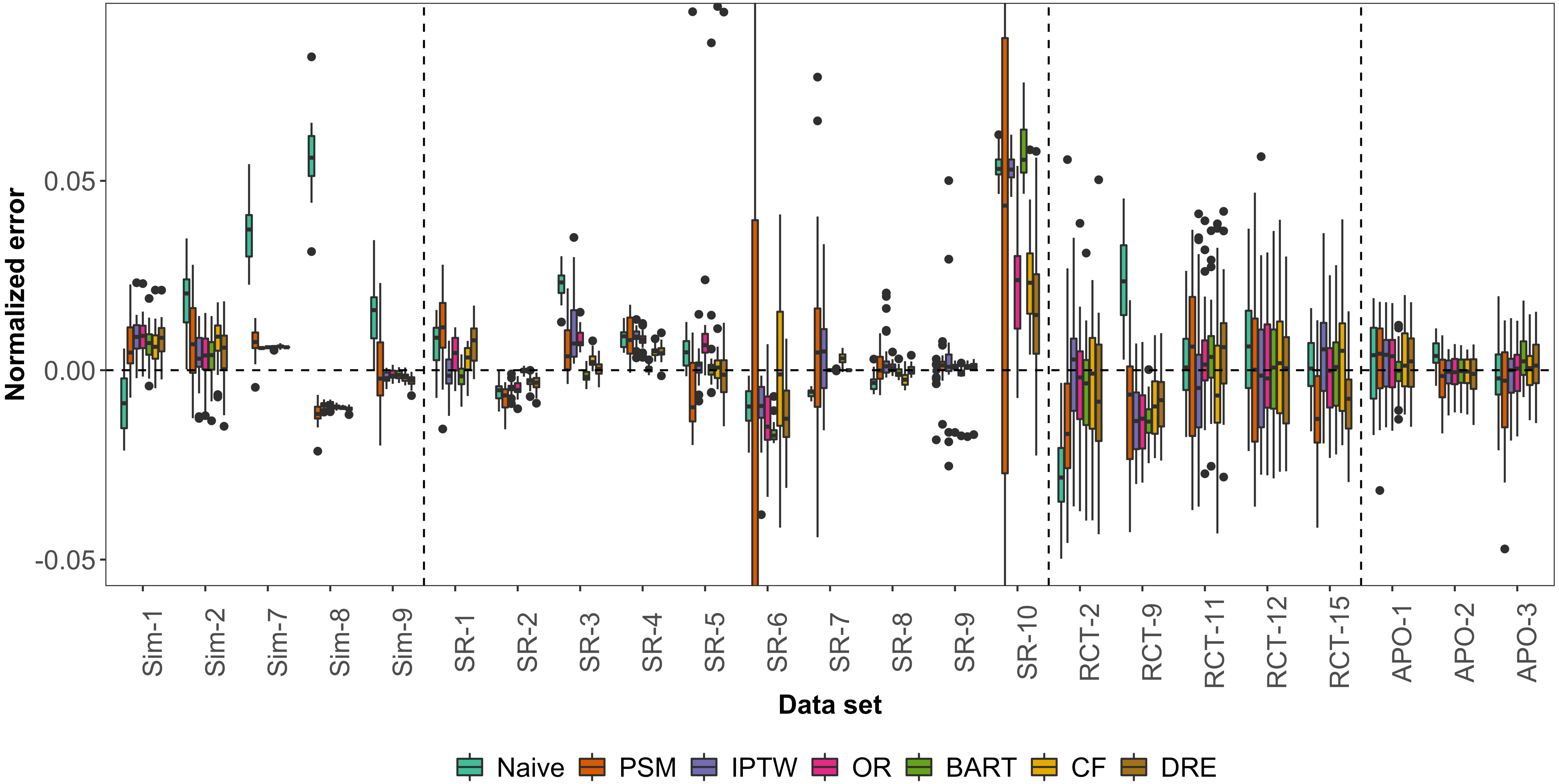
Focus on modeling outcome

- Outcome regression (OR)
- Bayesian additive regression trees (BART)

Combination of treatment and outcome estimation

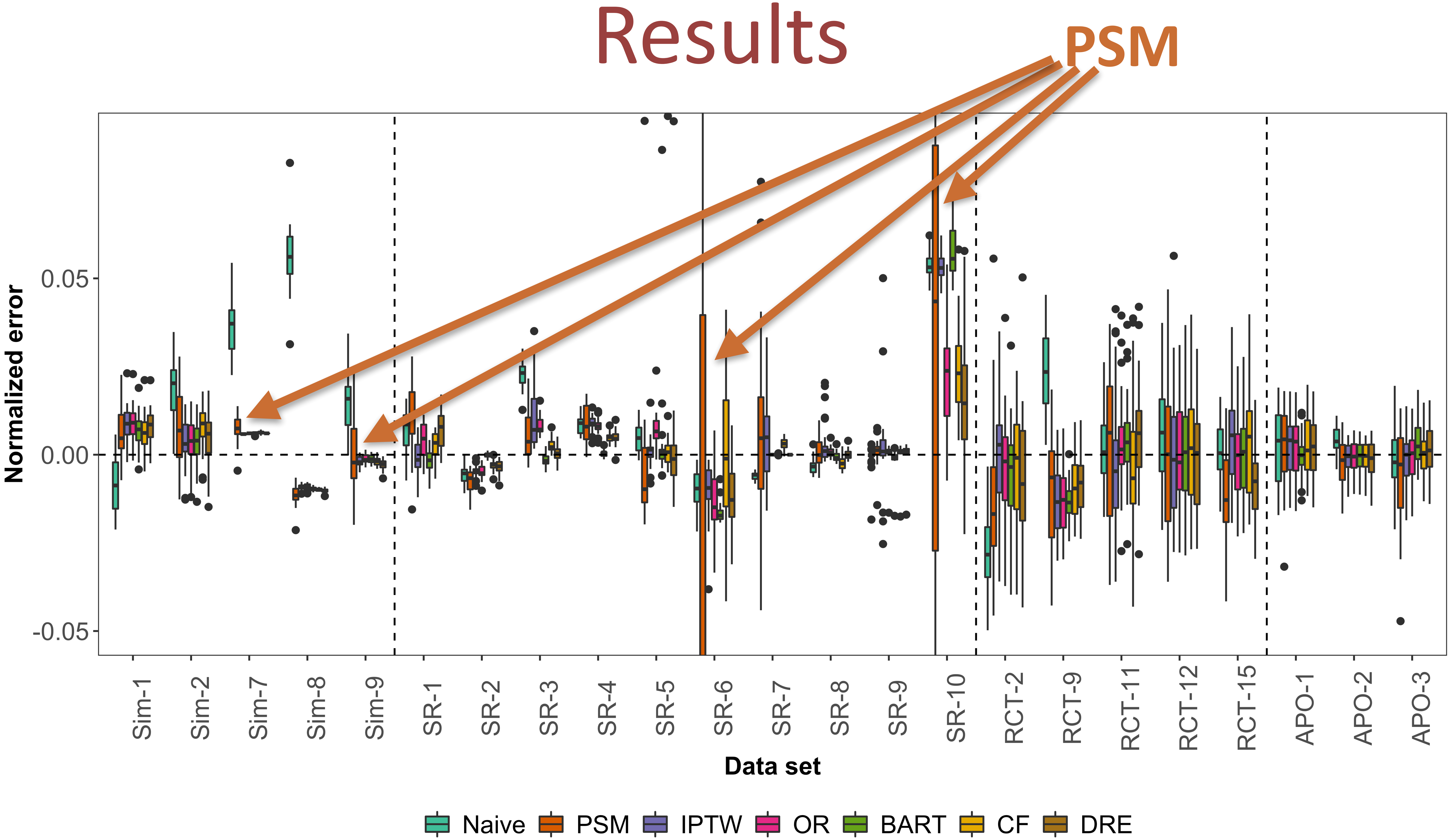
- Doubly-robust estimation (DR)
- Neural network method (NN)

Results



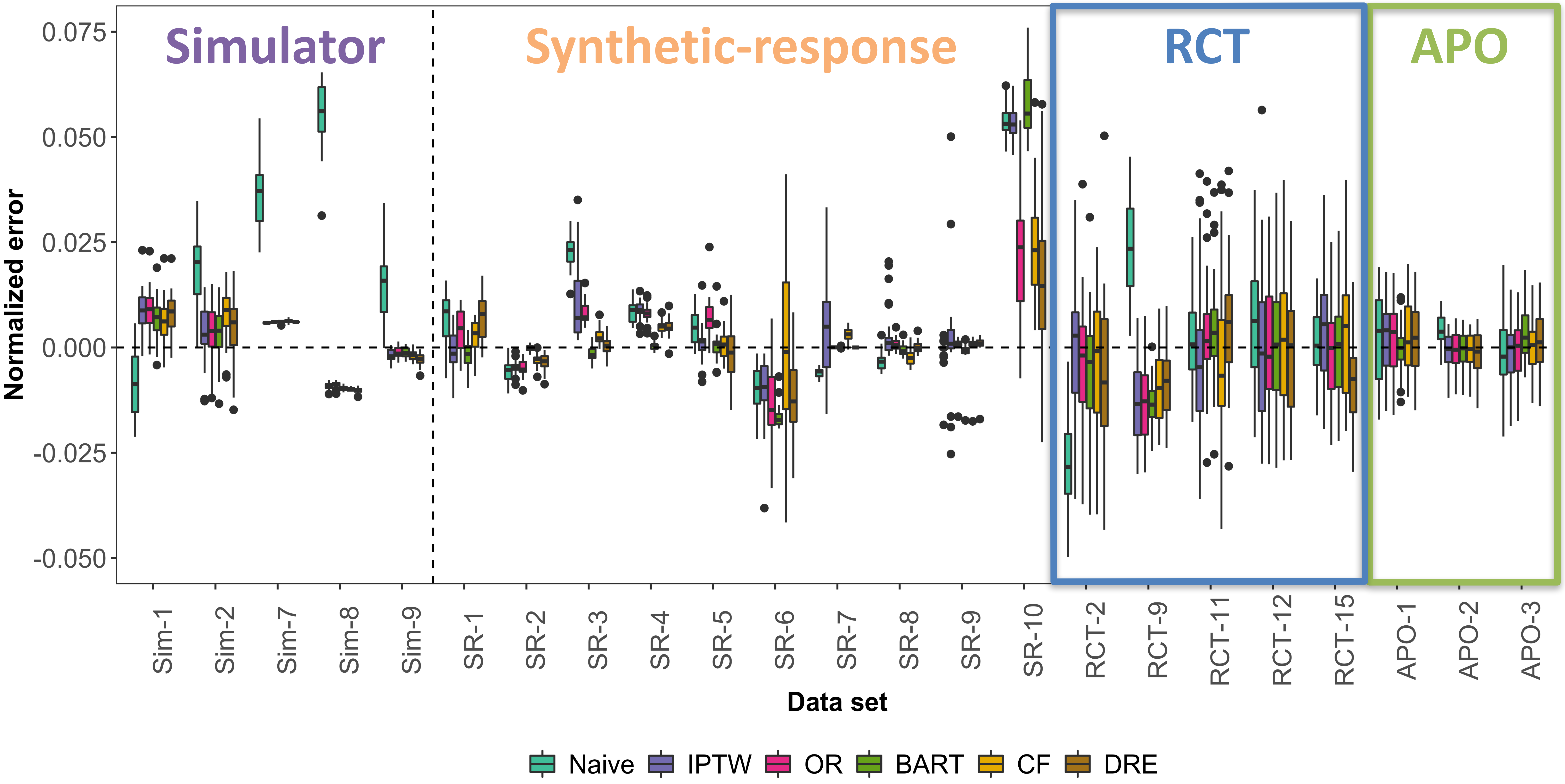
Normalized error
0.05
0.00
-0.05
Sim-1
Sim-2
Sim-7
Sim-8
Sim-9
SR-1
SR-2
SR-3
SR-4
SR-5
SR-6
SR-7
SR-8
SR-9
SR-10
RCT-2
RCT-9
RCT-11
RCT-12
RCT-15
APO-1
APO-2
APO-3
Data set
Naive
PSM
IPTW
OR
BART
CF
DRE

Results



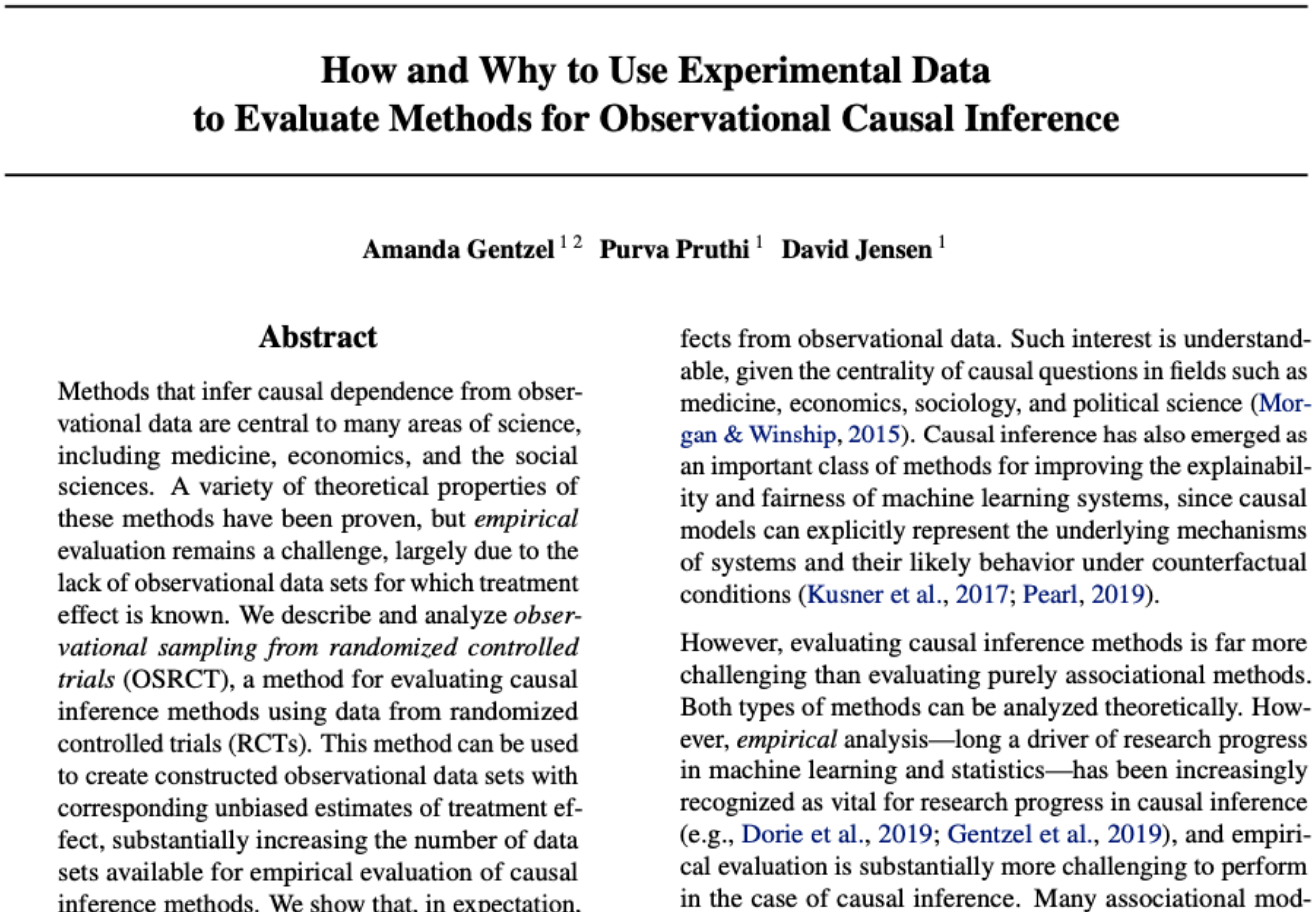
Propensity score matching sometimes performs poorly

Results



Simulator
Synthetic-response
RCT
APO
Normalized error
0.075
0.050
0.025
0.000
-0.025
-0.050
Sim-1
Sim-2
Sim-7
Sim-8
Sim-9
SR-1
SR-2
SR-3
SR-4
SR-5
SR-6
SR-7
SR-8
SR-9
SR-10
RCT-2
RCT-9
RCT-11
RCT-12
RCT-15
APO-1
APO-2
APO-3
Data set
Naive
IPTW
OR
BART
CF
DRE

For more details, please see our paper!



How and Why to Use Experimental Data to Evaluate Methods for Observational Causal Inference

Amanda Gentzel [1 2] Purva Pruthi [1] David Jensen [1]

Abstract

Methods that infer causal dependence from observational data are central to many areas of science, including medicine, economics, and the social sciences. A variety of theoretical properties of these methods have been proven, but *empirical* evaluation remains a challenge, largely due to the lack of observational data sets for which treatment effect is known. We describe and analyze *observational sampling from randomized controlled trials* (OSRCT), a method for evaluating causal inference methods using data from randomized controlled trials (RCTs). This method can be used to create constructed observational data sets with corresponding unbiased estimates of treatment effect, substantially increasing the number of data sets available for empirical evaluation of causal inference methods. We show that, in expectation,

fects from observational data. Such interest is understandable, given the centrality of causal questions in fields such as medicine, economics, sociology, and political science (Morgan & Winship, 2015). Causal inference has also emerged as an important class of methods for improving the explainability and fairness of machine learning systems, since causal models can explicitly represent the underlying mechanisms of systems and their likely behavior under counterfactual conditions (Kusner et al., 2017; Pearl, 2019).

However, evaluating causal inference methods is far more challenging than evaluating purely associational methods. Both types of methods can be analyzed theoretically. However, *empirical* analysis—long a driver of research progress in machine learning and statistics—has been increasingly recognized as vital for research progress in causal inference (e.g., Dorie et al., 2019; Gentzel et al., 2019), and empirical evaluation is substantially more challenging to perform in the case of causal inference. Many associational mod-