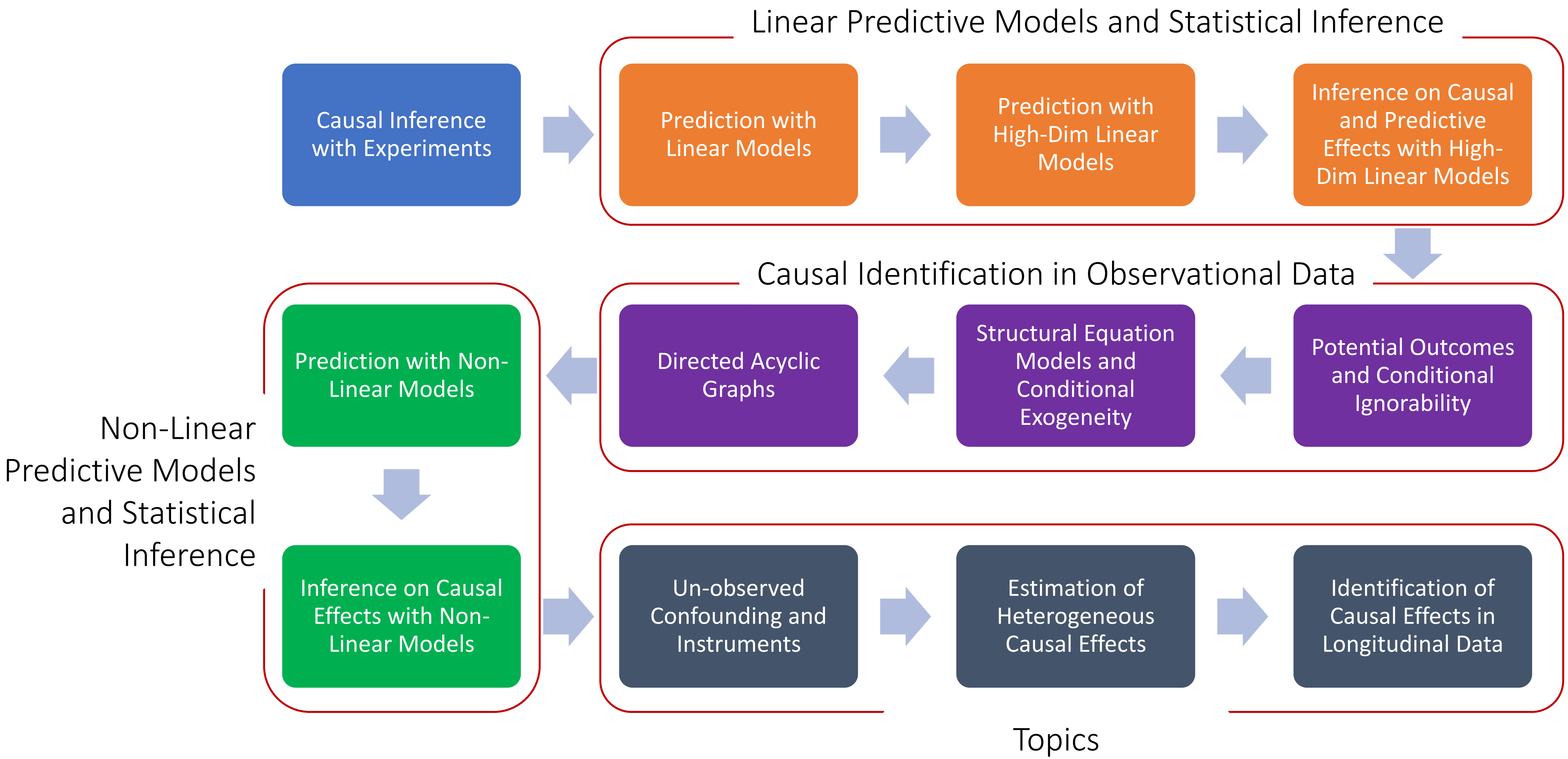
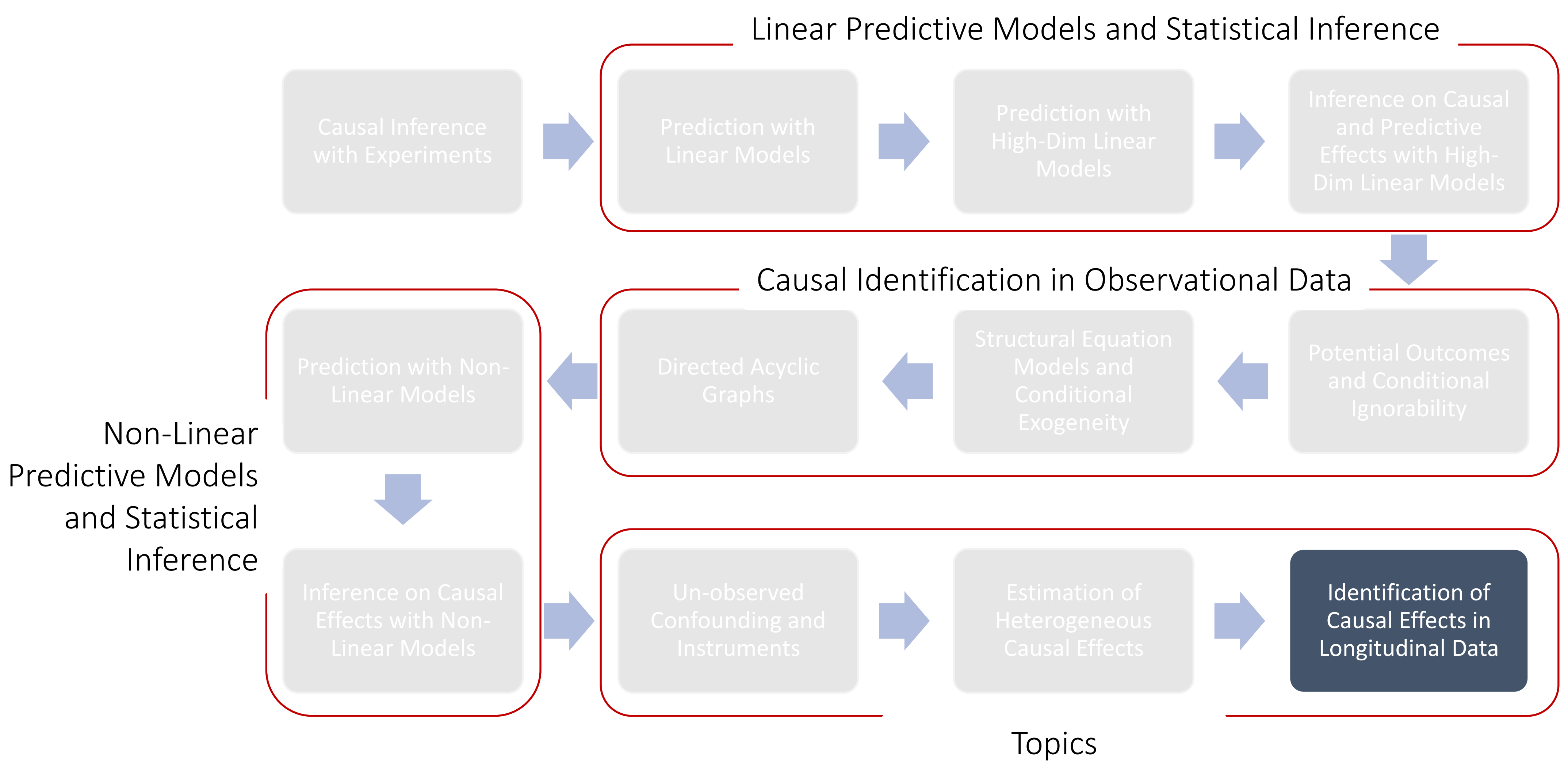
MS&E 228: Advanced Topics

Vasilis Syrgkanis

MS&E, Stanford


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Non-Linear Predictive Models and Statistical Inference
Prediction with Non-Linear Models
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Identification of Causal Effects in Longitudinal Data
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Potential Outcomes and Conditional Ignorability
Structural Equation Models and Conditional Exogeneity
Directed Acyclic Graphs
Prediction with Non-Linear Models
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Estimation of Heterogeneous Causal Effects
Identification of Causal Effects in Longitudinal Data
Topics

Longitudinal Data

- Longitudinal data refer to datasets where we have multiple observations of the same unit over time

They present a separate set of new problems

- **Correlation:** samples stemming from each unit (or sometimes cluster of units from the same "site") are correlated
- **Censoring/Missingness:** units typically drop out at (potentially not) random times before the end of the study
- **Dynamic treatments:** units are treated with multiple treatments over time in a manner that is adaptive and auto-correlated

Longitudinal Data

- Longitudinal data refer to datasets where we have multiple observations of the same unit over time

They present a separate set but also opportunities for causal identification strategies

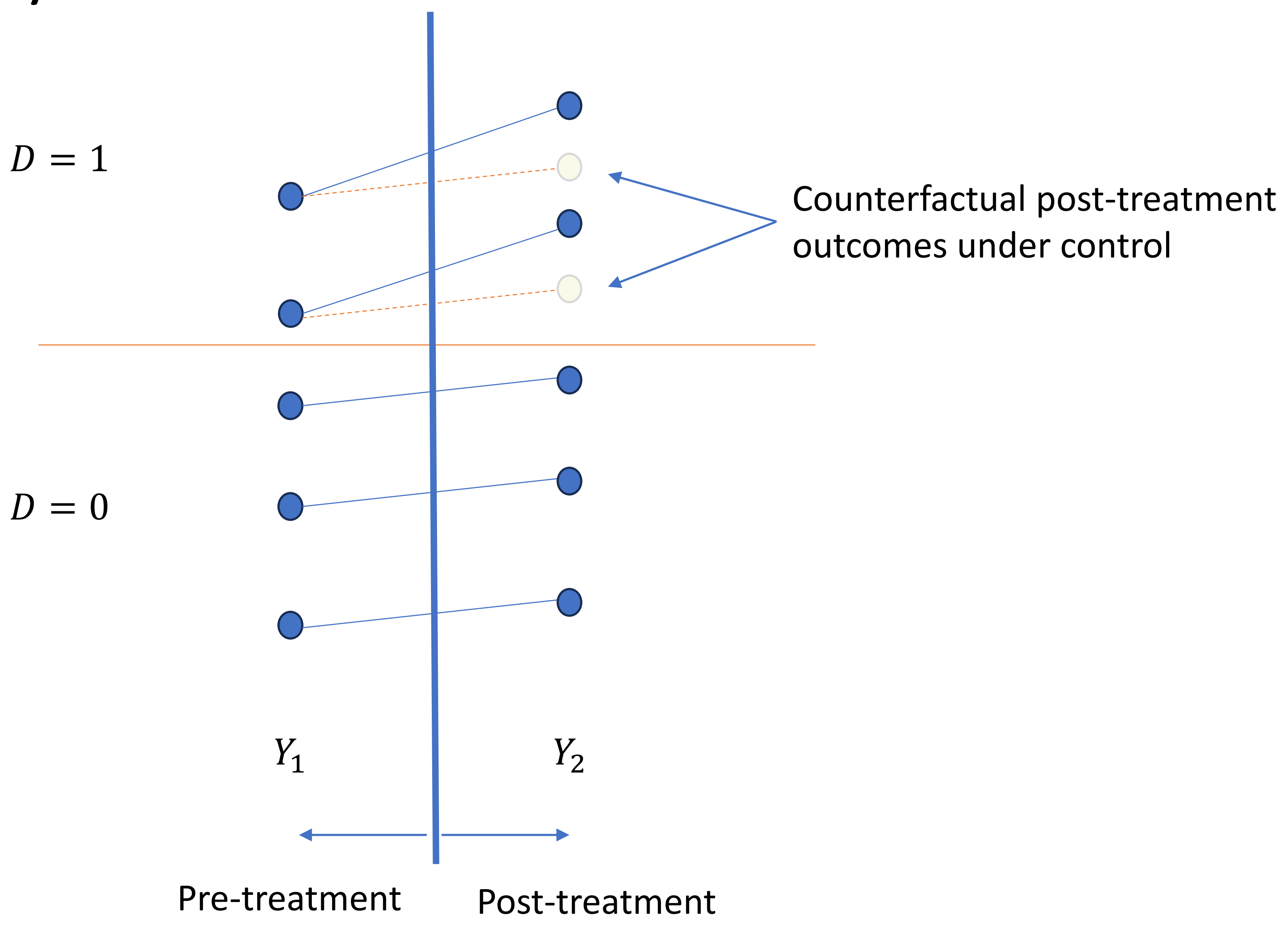
- **Differences-in-Differences:** assumption of treatment not being chosen based on potential "growth" trajectory of unit under control
- **Synthetic Controls:** assumption that trajectory of a treated unit under control, can be well approximate as some combination of trajectories of un-treated units

Differences-in-Differences

Differences-in-Differences

- Suppose that we have two period observations per unit Y_1, Y_2
- Units were treated in the second period
- D denotes whether someone was treated in the second period.
- Goal. Estimate the average treatment effect of the treated population
$$E[Y_2(1) - Y_2(0) \mid D = 1]$$
- Treatment was assigned based on un-observed confounders
- **Parallel trends:** willing to assume that treatment was not assigned based on the potential growth that units would have under control
$$Y_2(0) - Y_1(0) \perp\!\!\!\perp D$$

Visually



$D = 1$
Counterfactual post-treatment outcomes under control
$D = 0$
Y_1
Y_2
Pre-treatment
Post-treatment

Differences-in-Differences

- **Parallel trends:** willing to assume that treatment was not assigned based on the potential growth that units would have under control
$$Y_2(0) - Y_1(0) \perp\!\!\!\perp D$$
- **No anticipation:** treated units do not anticipate their treatment
$$E[Y_1(0) \mid D = 1] = E[\, Y_1(1) \mid D = 1\,]$$

- Then we can identify the ATT!

Identification of ATT

- Note that if we look at the differences in outcomes, i.e. $\tilde{Y} = Y_2 - Y_1$
- Then they satisfy one-sided exogeneity

$$\tilde{Y}(0) \perp\!\!\!\perp D$$

- We can identify mean counterfactual outcomes for treated units:

$$E\left[\tilde{Y}(0) \mid D = 1\right] = E\left[\tilde{Y}(0) \mid D = 0\right] = E\left[\tilde{Y} \mid D = 0\right]$$

- We can identify ATT for these transformed outcomes:

$$E\left[\tilde{Y}(1) - \tilde{Y}(0) \mid D = 1\right] = E\left[\tilde{Y} - \tilde{Y}(0) \mid D = 1\right] = E[\tilde{Y} \mid D = 1] - E\left[\tilde{Y} \mid D = 0\right]$$

- No-anticipation $\Rightarrow$ ATT of transformed outcomes = ATT of second period outcomes:

$$E\left[\tilde{Y}(1) - \tilde{Y}(0) \mid D = 1\right] = E\left[Y_2(1) - \cancel{Y_1(1)} - \left(Y_2(0) - \cancel{Y_1(0)}\right) \mid D = 1\right]$$

- No-anticipation assumption, $E[Y_1(1) - Y_1(0) \mid D = 1] = 0$

$$E\left[\tilde{Y}(1) - \tilde{Y}(0) \mid D = 1\right] = E[Y_2(1) - Y_2(0) \mid D = 1]$$

Conditional Parallel Trends

- Parallel trends can be quite strong of an assumption
- Much more plausible once we much on observable characteristics
- **Conditional Parallel Trends:** conditional on X, treatment was not assigned based on potential growth of unit under control

$$Y_2(0) - Y_1(0) \perp\!\!\!\perp D \mid X$$

- Then we can also identify ATT!

ATT under Conditional Parallel Trends

- Note that if we look at the differences in outcomes, i.e. $\tilde{Y} = Y_2 - Y_1$
- Then they satisfy one-sided conditional exogeneity

$$\tilde{Y}(0) \perp\!\!\!\perp D \mid X$$

- We can identify mean counterfactual outcomes for treated units:

$$\begin{aligned} E\left[\tilde{Y}(0) \mid D = 1\right] &= E\left[E\left[\tilde{Y}(0) \mid D = 1, X\right] \mid D = 1\right] \\ &= E\left[E\left[\tilde{Y}(0) \mid D = 0, X\right] \mid D = 1\right] = E\left[E\left[\tilde{Y} \mid D = 0, X\right] \mid D = 1\right] \end{aligned}$$

- We can identify ATT for these transformed outcomes:

$$E\left[\tilde{Y}(1) - \tilde{Y}(0) \mid D = 1\right] = E\left[\tilde{Y} - E\left[\tilde{Y} \mid D = 0, X\right] \mid D = 1\right]$$

- No-anticipation ⇒ ATT of transformed outcomes = ATT of second period outcomes:

$$\begin{aligned} E[Y_2(1) - Y_2(0) \mid D = 1] &= E\left[\tilde{Y}(1) - \tilde{Y}(0) \mid D = 1\right] \\ &= E\left[\tilde{Y} - E\left[\tilde{Y} \mid D = 0, X\right] \mid D = 1\right] \end{aligned}$$

Debiased ML Estimation

- We want to estimate the parameter

$$\theta = E\left[\,\tilde{Y} - g(0, X) \mid D = 1\,\right]$$

$$g(0, X) = E\left[\,\tilde{Y} \mid D, X\,\right]$$

- This is again a moment problem with nuisance functions
- Letting $\pi = \Pr(D = 1)$

$$E\left[\left(\tilde{Y} - g(0, X) - \theta\right)\frac{D}{\pi}\right] = 0$$

- We need to add a debiasing correction term!

Debiased ML for ATT

- Letting $\pi = \Pr(D = 1)$

$$E\left[\left(\tilde{Y} - g(0, X) - \theta\right)\frac{D}{\pi}\right] = 0$$

- There are two nuisance quantities g, π.
- Moment already orthogonal to π.
- We need to add a debiasing correction term for g!

$$E\left[\left(\tilde{Y} - g(0, X) - \theta\right)\frac{D}{\pi}\right] + E\left[a(D, X)\left(\tilde{Y} - g(D, X)\right)\right]$$

Debiased ML for ATT

- We need to add a debiasing correction term for g!

$$E\left[\left(\tilde{Y} - g(0, X) - \theta\right)\frac{D}{\pi}\right] + E\left[a(D, X)\left(\tilde{Y} - g(D, X)\right)\right]$$

- Function $a(D, X)$ is the Riesz representer of the functional

$$E\left[-g(0, X)\frac{D}{\pi}\right] = E\left[-g(0, X)\ \frac{\Pr(\ D = 1 \mid X\ )}{\pi}\right]$$

- Let $p(X) = \Pr(\ D = 1 \mid X\ )$. Need to find a function $a(D, X)$ such that

$$\forall g\colon\ E\left[-g(0, X)\frac{p(X)}{\pi}\right] = E[a(D, X) g(D, X)]$$

- This again is an inverse propensity

$$a(D, X) = -\frac{1 - D}{1 - p(X)}\frac{p(X)}{\pi}$$

Debiased Moment

- The debiased moment is

$$E\left[\left(\tilde{Y} - g(0,X) - \theta\right)\frac{D}{\pi} - \frac{1-D}{1-p(X)}\frac{p(X)}{\pi}\left(\tilde{Y} - g(D,X)\right)\right] = 0$$

- Can be simplified to

$$E\left[\left(\tilde{Y} - g(0,X) - \theta\right)\frac{D}{\pi} - \frac{1-D}{1-p(X)}\frac{p(X)}{\pi}\left(\tilde{Y} - g(0,X)\right)\right] = 0$$

- Can be simplified to

$$E\left[\frac{D-p(X)}{1-p(X)}\frac{1}{\pi}\left(\tilde{Y} - g(0,X)\right) - \theta\frac{D}{\pi}\right] = 0$$

Doubly Robust Estimate

- The estimate solves the empirical version of the moment

$$E_n\left[\frac{D-\hat{p}(X)}{1-\hat{p}(X)}\frac{1}{\pi}\left(\tilde{Y}-\hat{g}(0,X)\right)-\hat{\theta}\,\frac{D}{\hat{\pi}}\right]=0$$

- Equivalently

$$\hat{\theta}=\frac{1}{\sum_i D_i}\sum_i \frac{D_i-\hat{p}(X_i)}{1-\hat{p}(X_i)}\left(\widetilde{Y_i}-\hat{g}(0,X_i)\right)$$

- Enjoys similar double robustness properties as the double robust estimate of the ATE

Inference with Doubly Robust Algorithm

- Moment is Neyman orthogonal

$$m(Z;\theta,g) := \underbrace{\frac{D-p(X)}{1-p(X)}\frac{1}{\pi}\left(\tilde{Y}-g(0,X)\right)}_{\hat{Y}(g,p,\pi)} - \theta\frac{D}{\pi}$$

- If product of RMSEs of propensity and regression model goes down at rate faster than $n^{1/2}$, plus regularity conditions

$$\sqrt{n}(\hat{\theta}-\theta_0) = \sqrt{n}\left(E_n[\hat{Y}(\hat{g},\hat{p},\hat{\pi})] - \theta_0\frac{E_n[D/\hat{\pi}]}{E_n[D/\hat{\pi}]}\right) \approx \sqrt{n}\frac{E_n[\hat{Y}(g_0,a_0,\pi_0)-\theta_0 D/\pi_0]}{E_n[D/\pi_0]} \approx \sqrt{n}\frac{E_n[\hat{Y}(g_0,a_0,\pi_0)-\theta_0 D/\pi_0]}{E[D/\pi_0]}$$

- Consequently, it is *asymptotically normal*

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) \sim_a N(0,V), \qquad V := \frac{E\left[\left(\hat{Y}(g_0,a_0,\pi_0)-\theta_0 D/\pi_0\right)^2\right]}{E[D/\pi_0]^2}$$

- *Confidence intervals* for any projection based on estimate of variance are asymptotically valid

$$\ell'\theta \in \left[\ell'\hat{\theta} \pm c\sqrt{\frac{\ell'\hat{V}\ell}{n}}\right], \qquad \hat{V} = \frac{E_n\left(\hat{Y}(\hat{g},\hat{p},\hat{\pi})-\hat{\theta}\frac{D}{\hat{\pi}}\right)^2}{E_n[D/\hat{\pi}]^2}$$

Data with Multiple Periods

- If we have data with multiple periods where different units were treated at different periods then it is an active area of research
- Typical empirical practice is what is known as two-way-fixed-effects
- Can be problematic when there is heterogeneity of the effect across periods and cannot incorporate relaxed conditional parallel trends

One Approach

- Consider explicitly different segments of the population
- Consider subset of data that were treated at period t_0
- For every subsequent period $t \in [t_0, \ldots, K]$
  - Define control population as units that have not yet been treated at period t
  - Define pre-intervention period as period $t_0 - 1$
  - Apply the two-period approach outlined in previous slides
- This provides a separate estimate of the effect of the treatment at period t_0 on each of the subsequent periods t
- Variables that relate to properties of the units prior to $t_0 - 1$ can be used as control variables X, to make conditional parallel trends more plausible

Sanity Checks: Pre-Trends

- Repeat the above calculation but viewing $t_0 - 2$ as the "treatment" period and t_0 as the "reference" period
- No statistically significant treatment effect should be identified
- If it is identified, then signal that parallel trends assumption does not hold

Longitudinal Data

- Longitudinal data refer to datasets where we have multiple observations of the same unit over time

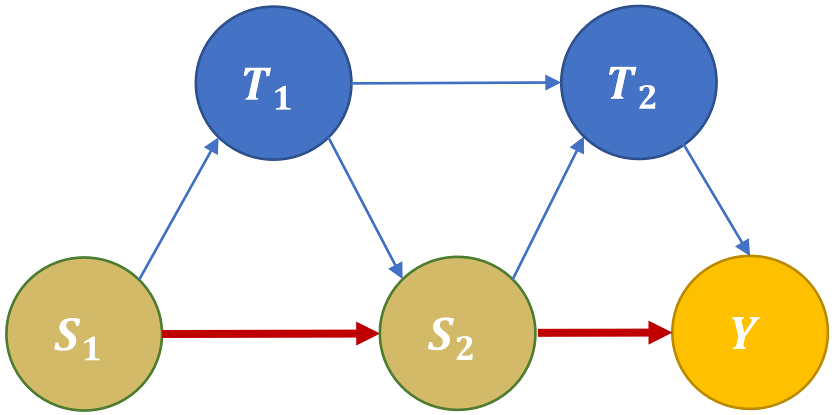
They present a separate set of new problems

- **Correlation:** samples stemming from each unit (or sometimes cluster of units from the same "site") are correlated
- **Censoring/Missingness:** units typically drop out at (potentially not) random times before the end of the study
- **Dynamic treatments:** units are treated with multiple treatments over time in a manner that is adaptive and auto-correlated

Dynamic Treatment Regime

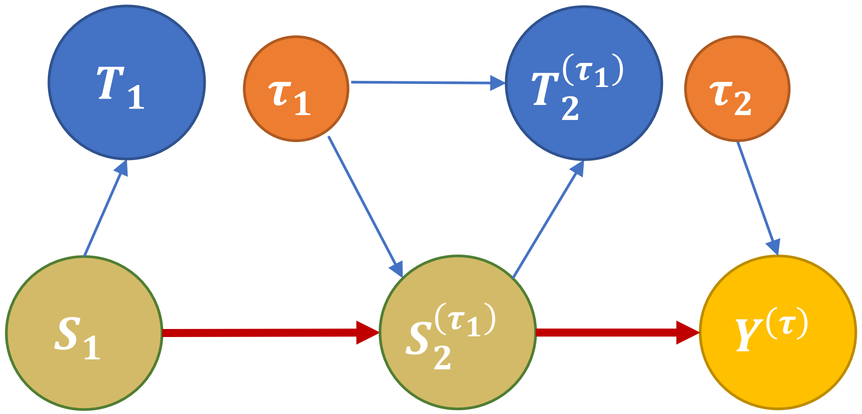
Dynamic Treatment Regime

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



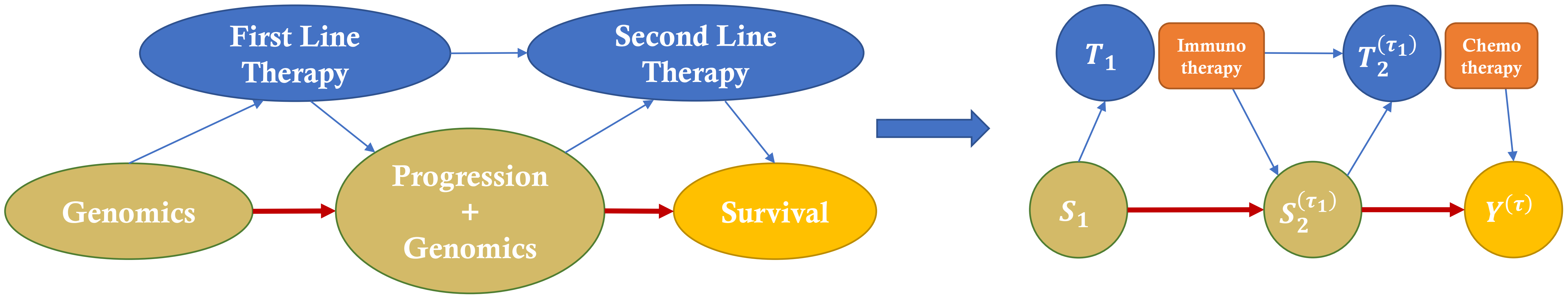
- Treatments are offered in an adaptive manner, in response to previous period controls
- The surrogate – control feedback precludes viewing this as a one-shot treatment problem
- Setting is known as the dynamic treatment regime [Robins'94,'04, Chakraborty-Murphy'14]

Examples

- Healthcare: Patients treated over time and adaptively

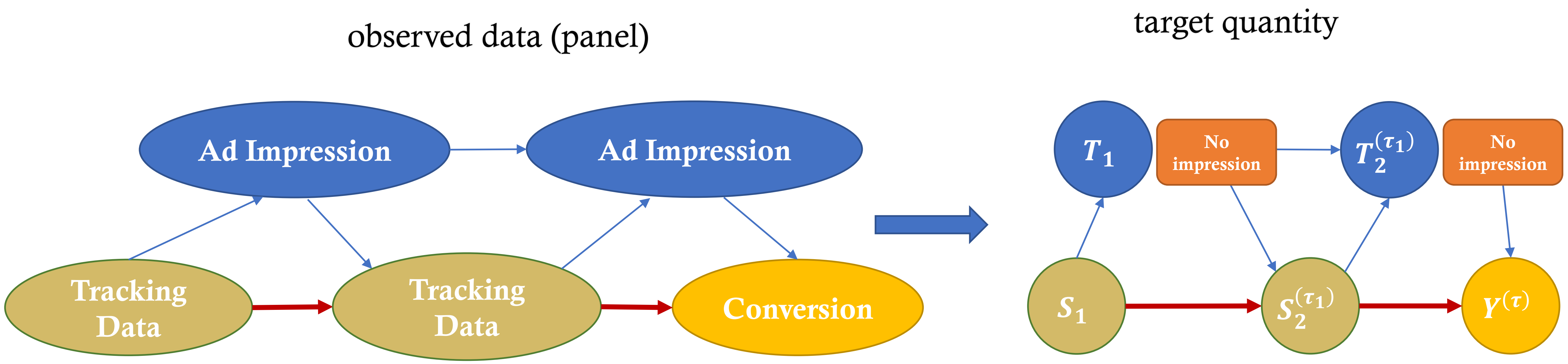
observed data (panel)

target quantity



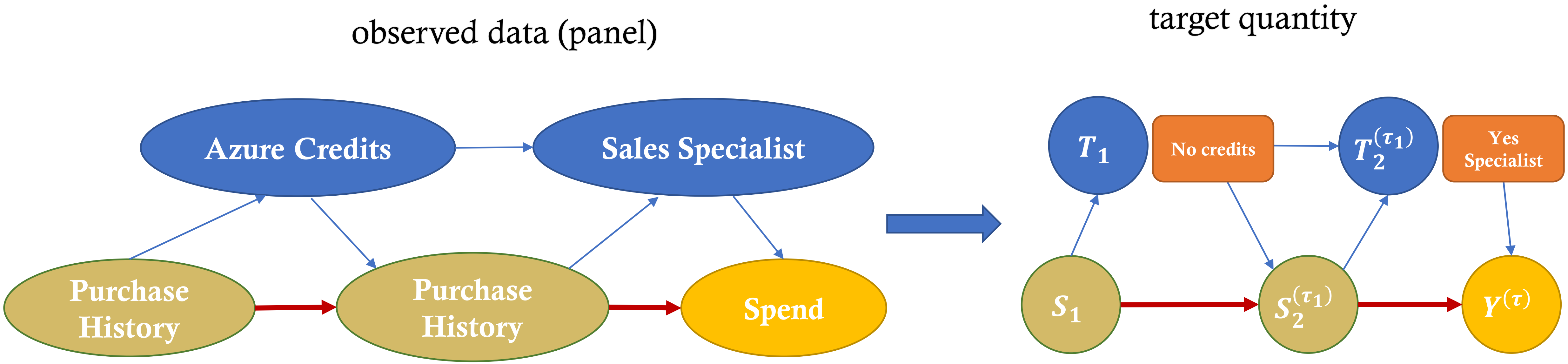
Examples

- Digital Marketing: Web users shown ads multiple times



Examples

- Business-to-Business (B2B) Operations: Customers offered multiple discount/support interventions over time



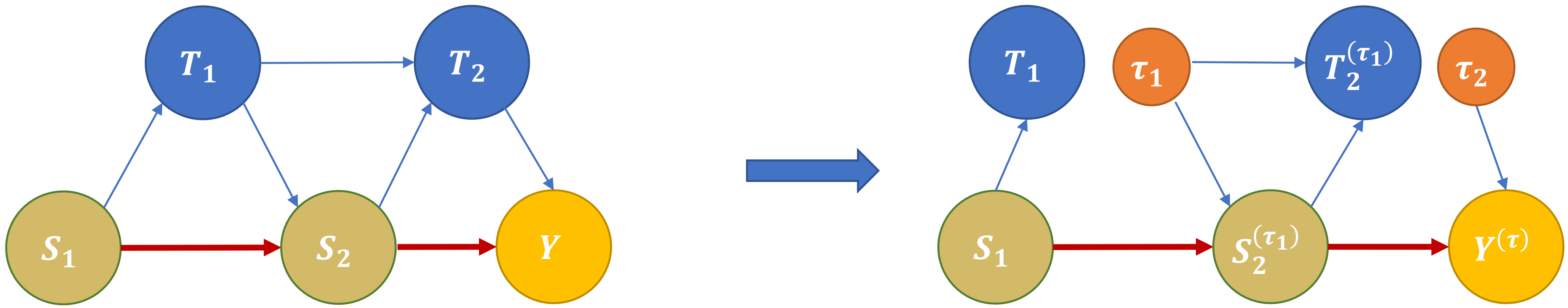
Identification in the Dynamic Treatment Regime

Dynamic Treatment Regime

observed data (panel)

target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

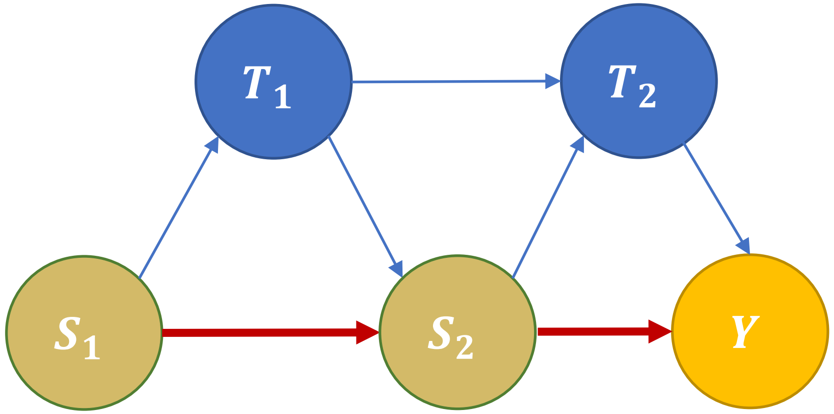
$$\theta \coloneqq E\left[Y^{(\tau)}\right]$$



- Treatments are offered in an adaptive manner, in response to previous period controls
- The surrogate – control feedback precludes viewing this as a one-shot treatment problem
- Setting is known as the dynamic treatment regime [Robins'94,'04, Chakraborty-Murphy'14]

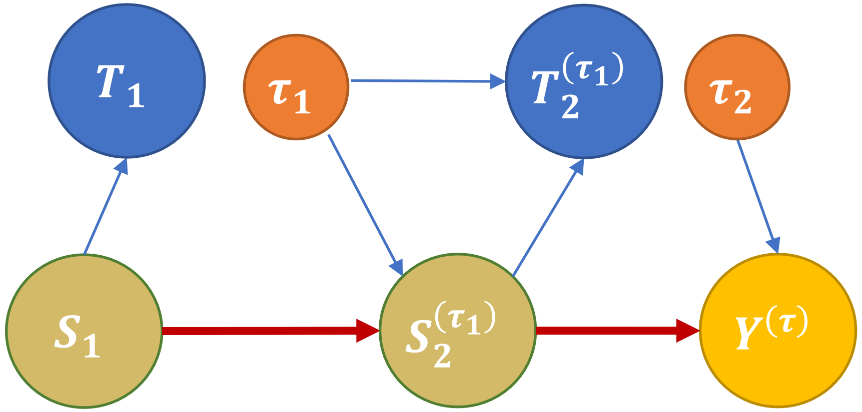
Identification in Dynamic Treatment Regime

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



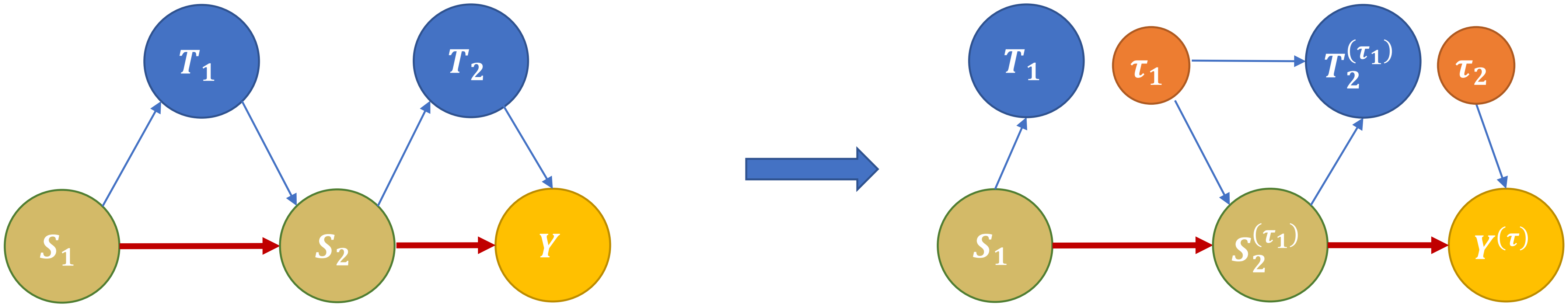
- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!

Identification in Dynamic Treatment Regime

observed data (panel)

target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



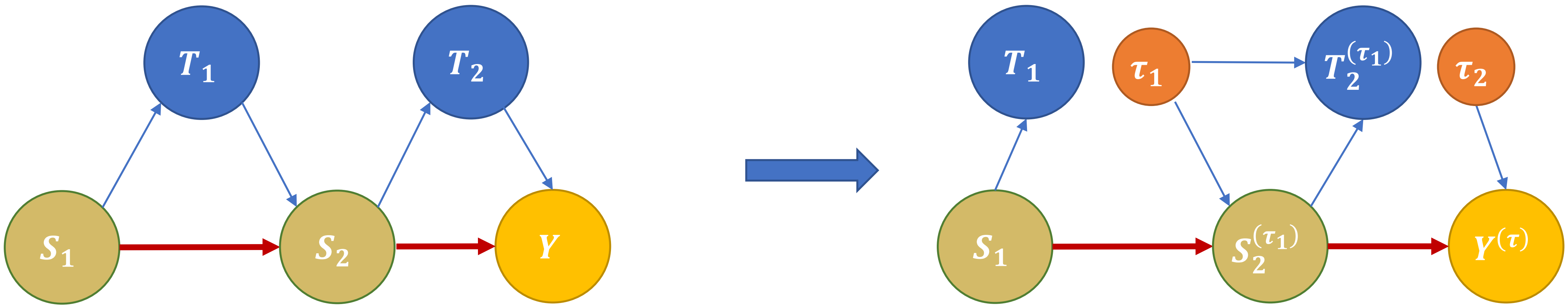
- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!

Identification in Dynamic Treatment Regime

observed data (panel)

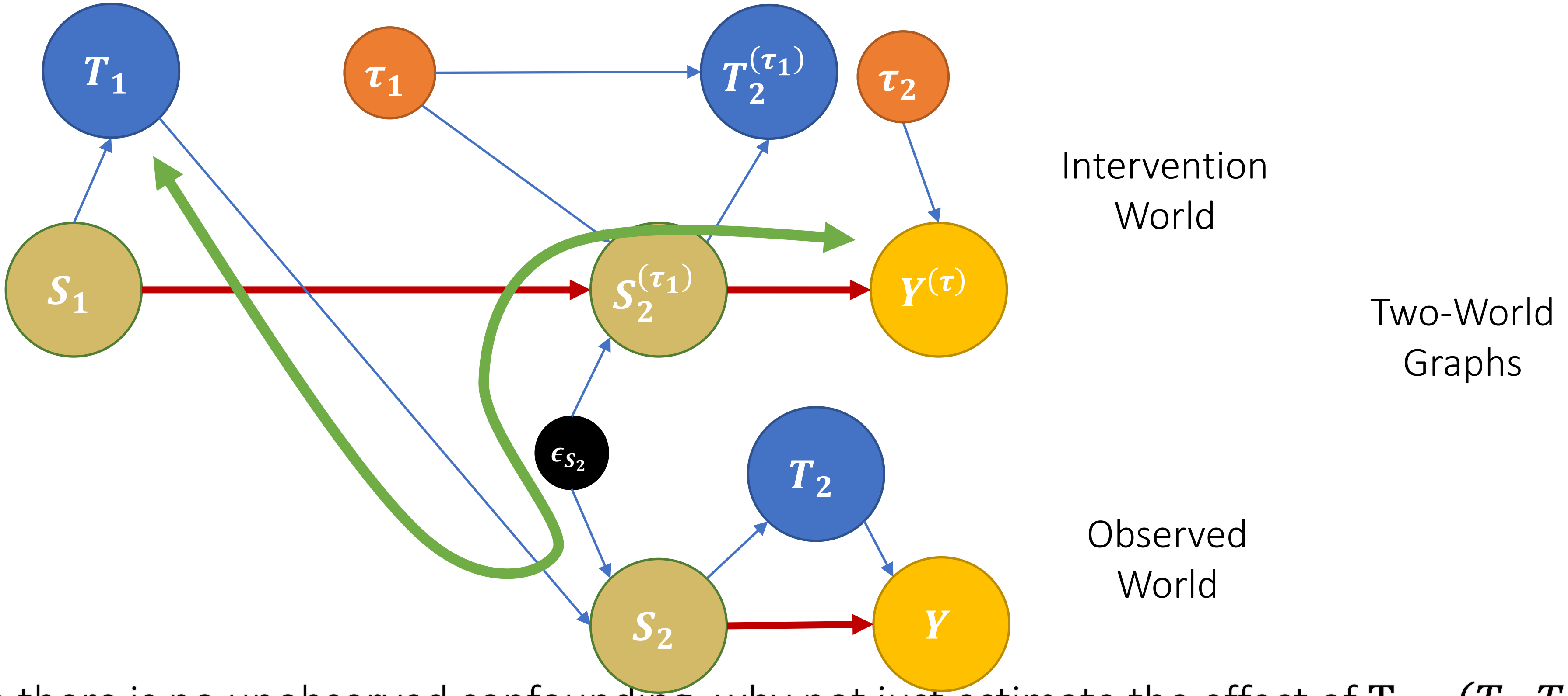
target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!
- Can also be understood using D-separation that $Y^{(\tau)} \not\perp\!\!\!\perp T_1 \mid S_1, S_2$

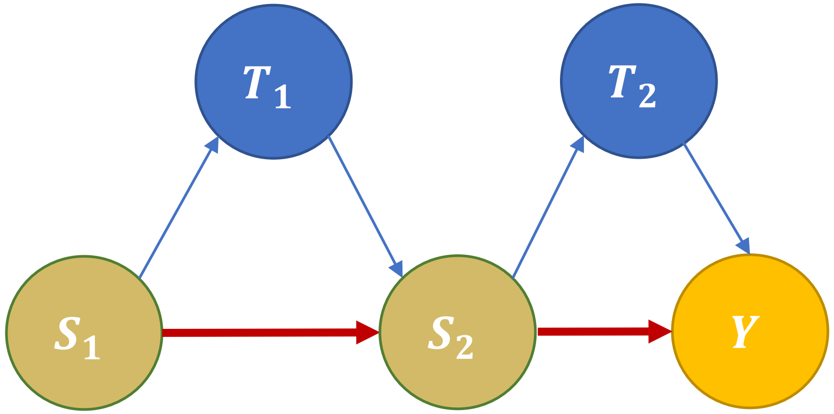
Identification in Dynamic Treatment Regime



- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!
- Can also be understood using D-separation that $Y^{(\tau)} \not\perp\!\!\!\perp T_1 \mid S_1, S_2$

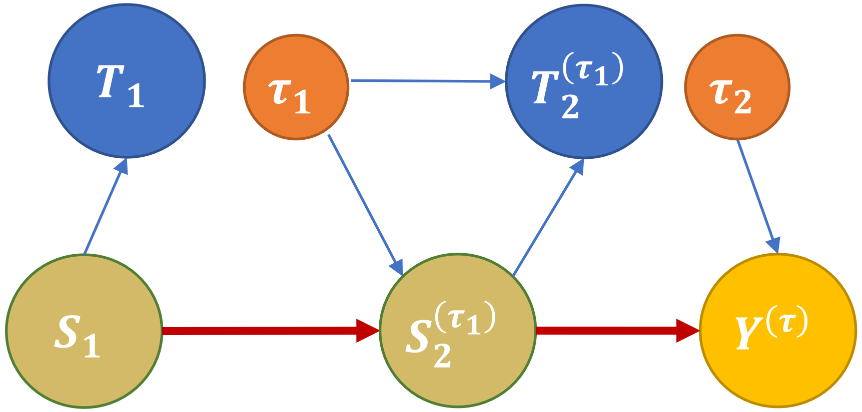
Identification in Dynamic Treatment Regime

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



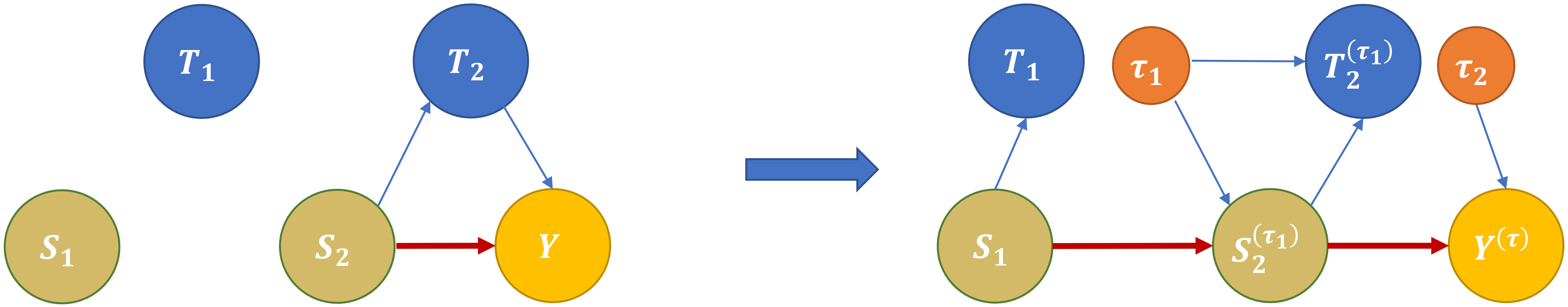
- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!

Identification in Dynamic Treatment Regime

observed data (panel)

target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



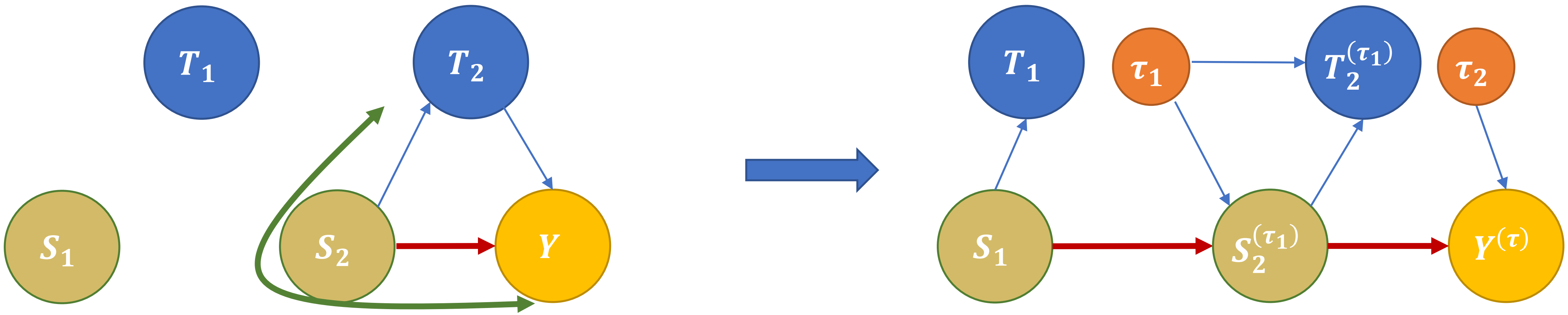
- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!

Identification in Dynamic Treatment Regime

observed data (panel)

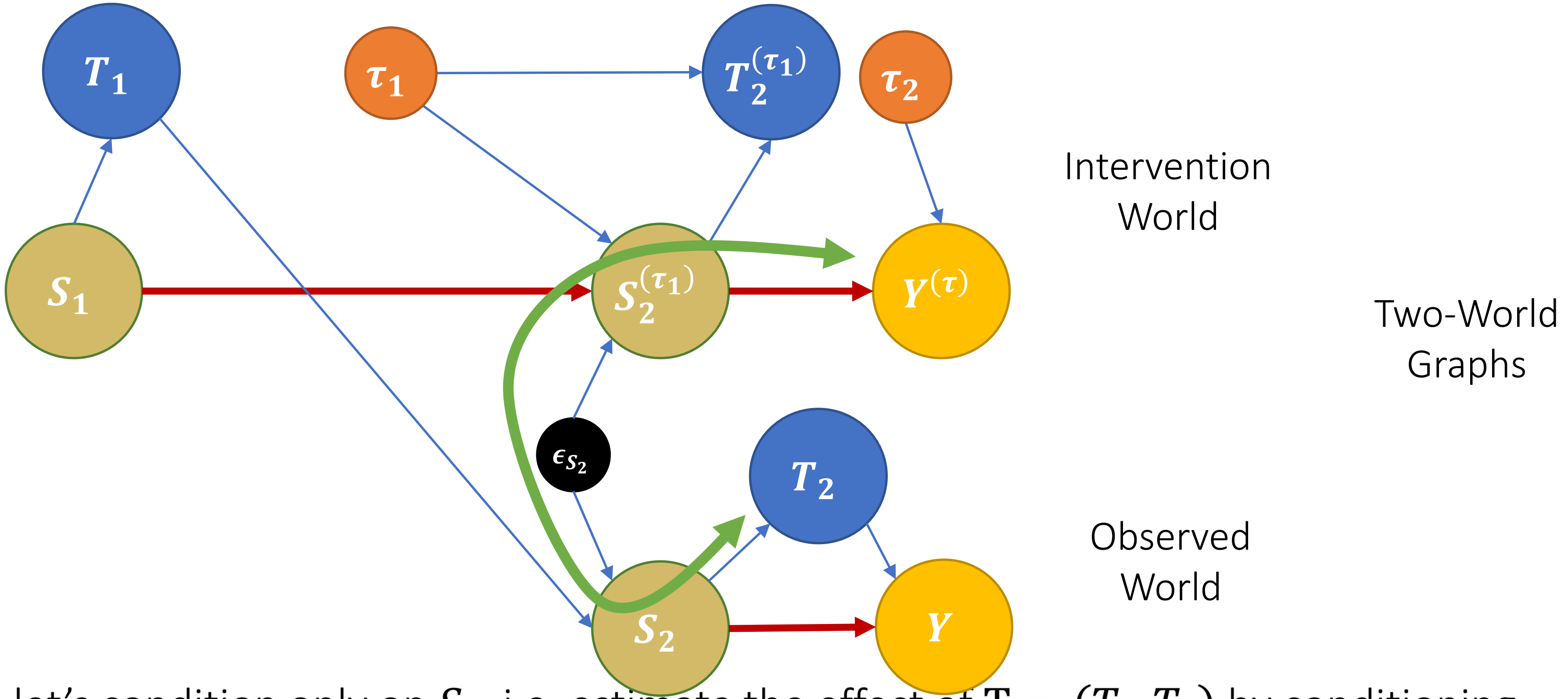
target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!

Identification in Dynamic Treatment Regime

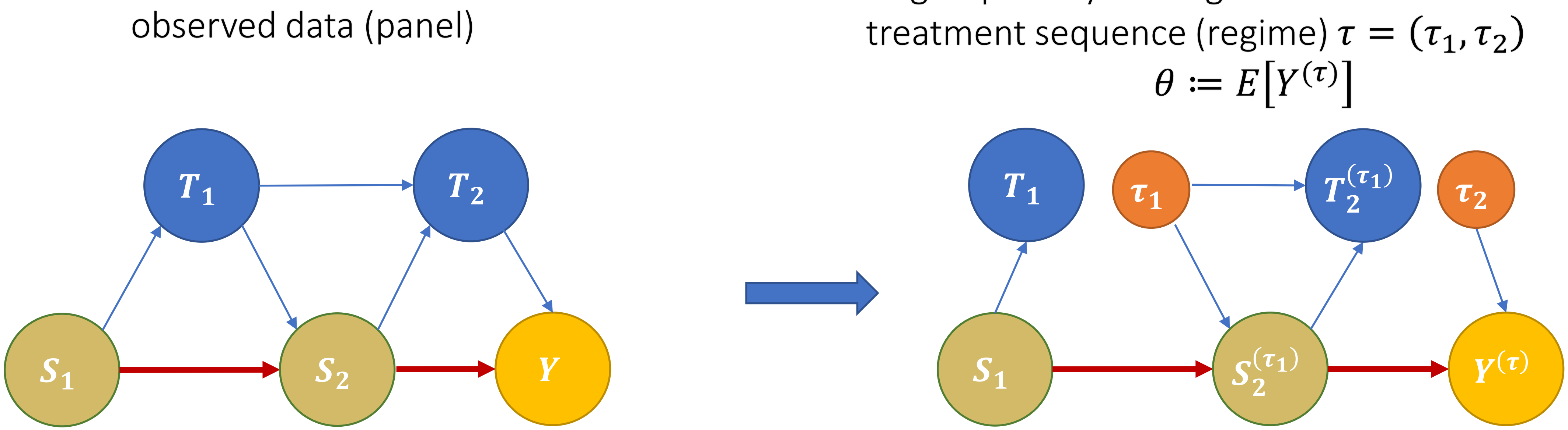


- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!
- Can also be understood using D-separation that $Y^{(\tau)} \not\perp\!\!\!\perp T_2 \mid S_1$

Dynamic Treatment Regime

target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

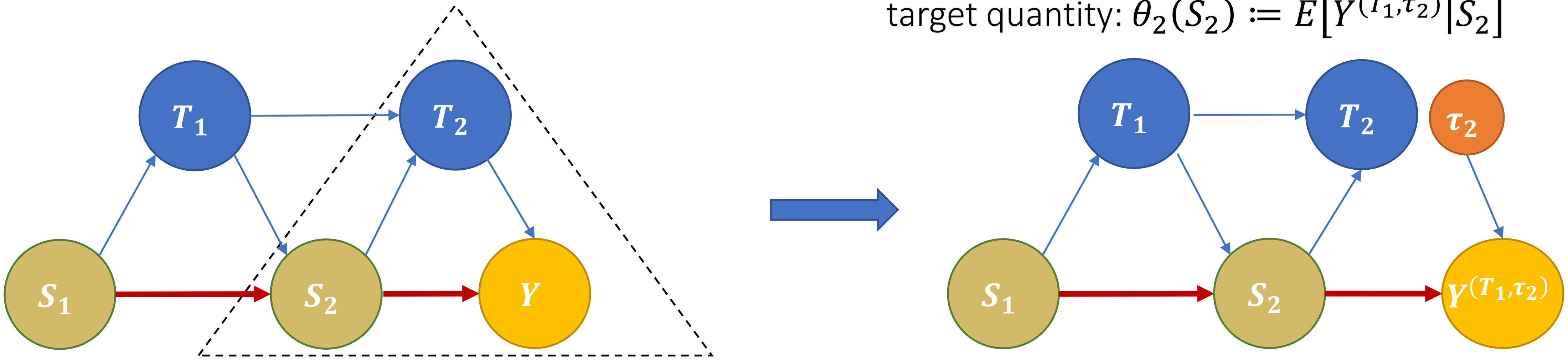
$$\theta := E\left[Y^{(\tau)}\right]$$



- We cannot identify this expected potential outcome simply by conditioning!
- Let's take it one step at a time
- What can we identify by conditioning?

Identification of Last Period Intervention

target quantity: $\theta_2(S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} | S_2\right]$



We have conditional ignorability $Y^{(T_1,\tau_2)} \perp\!\!\!\perp T_2 \mid S_2$

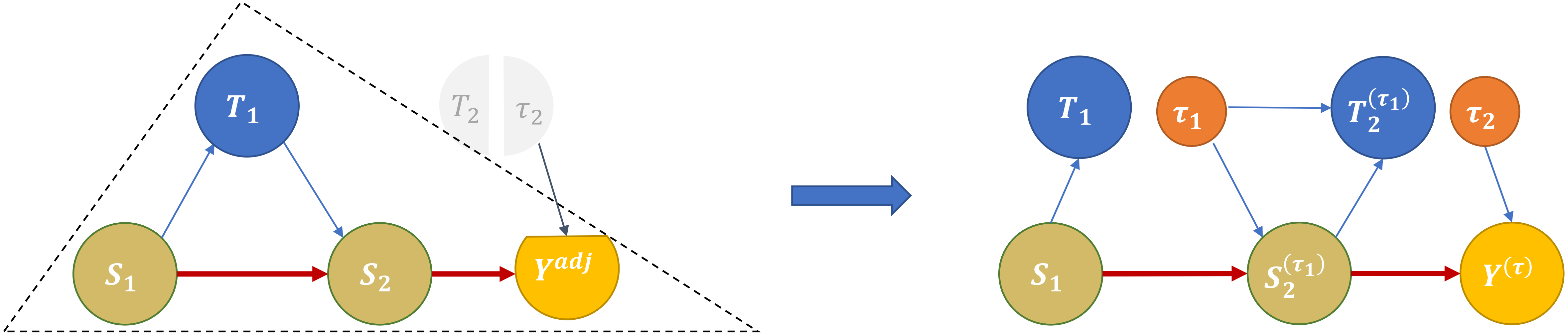
Target quantity $\theta_2(S_2)$ can be identified by conditioning

$$\begin{aligned}\theta_2(S_2) &= E\left[Y^{(T_1,\tau_2)} \middle| S_2\right] \\ &= E\left[Y^{(T_1,\tau_2)} \middle| T_2 = \tau_2, S_2\right] \\ &= E[Y | T_2 = \tau_2, S_2]\end{aligned}$$

Identification by conditioning:

- Train a predictive model $f_2(T_2, S_2)$
  $$Y \sim T_2, S_2$$
- Evaluate the model at $T_2 = \tau_2$
  $$\theta_2(S_2) = f_2(\tau_2, S_2)$$

Identification via Backwards Recursion

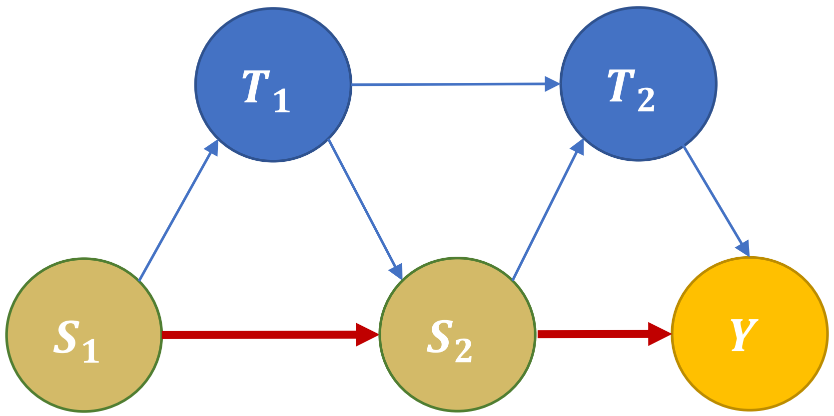


- But now we can "adjust the observed outcome" using this counterfactual model to remove the second period treatment!

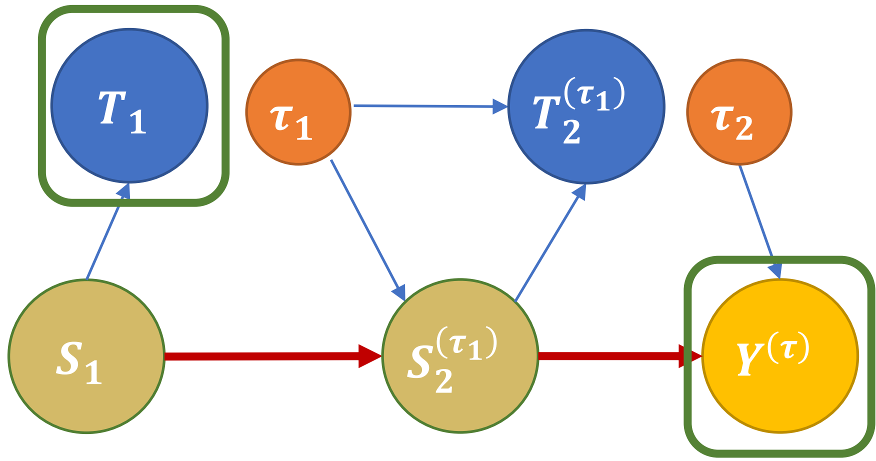
- Replace Y with $Y_{\text{adj}} \coloneqq f_2(\tau_2, S_2)$

- Estimating the target quantity $E\left[Y^{(\tau_1,\tau_2)}\right]$ is the same as estimating the effect of T_1 on Y_{adj}
- We can estimate this by conditioning on S_1

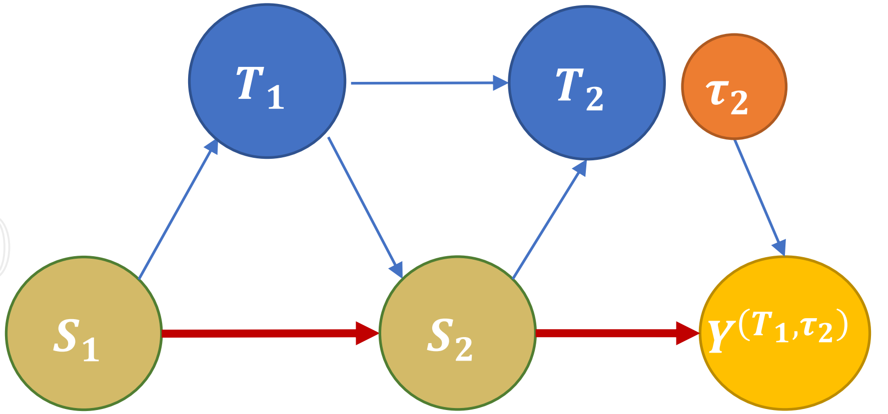
Identification via Backwards Recursion



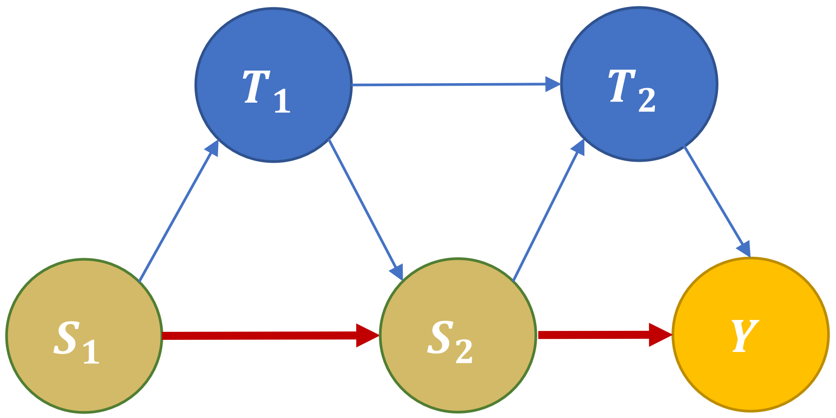
$$
\begin{aligned}
E\left[Y^{(\tau_1,\tau_2)}\right] &= E\left[E\left[Y^{(\tau_1,\tau_2)}|S_1\right]\right] \\
&= E\left[E\left[Y^{(T_1,\tau_2)}|T_1=\tau_1,S_1\right]\right] \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)}|S_2,T_1,S_1\right]|T_1=\tau_1,S_1\right]\right] && \text{(tower law)} \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)}|S_2\right]|T_1=\tau_1,S_1\right]\right] && \left(Y^{(T_1,\tau_2)} \perp\!\!\!\perp S_1,T_1|S_2\right) \\
&= E\left[E\left[\theta_2(S_2)|T_1=\tau_1,S_1\right]\right]
\end{aligned}
$$

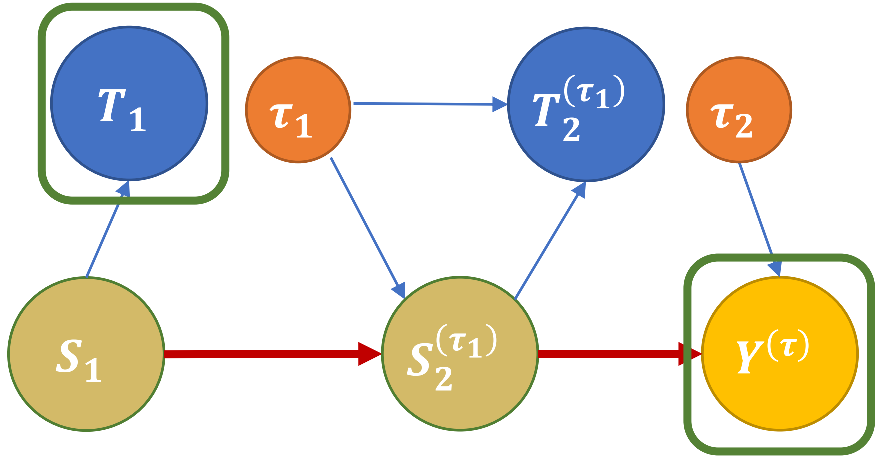


$$\boxed{\left(Y^{(\tau)} \perp\!\!\!\perp T_1|S_1\right)}$$



Identification via Backwards Recursion



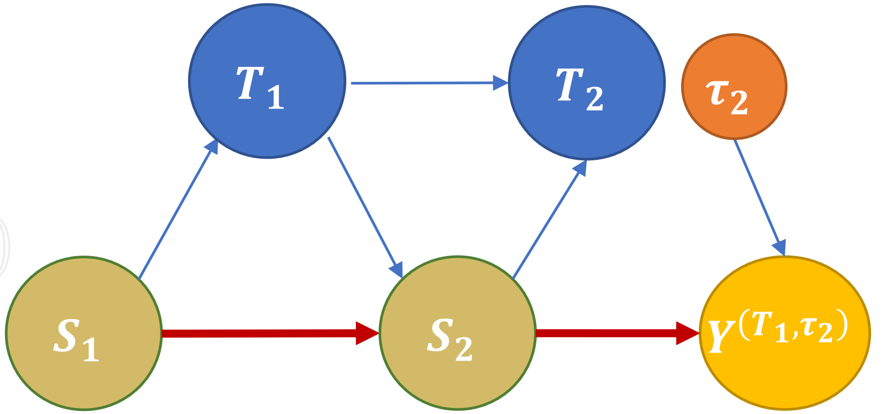


$$
\begin{aligned}
E\left[Y^{(\tau_1,\tau_2)}\right] &= E\left[E\left[Y^{(\tau_1,\tau_2)}\middle|S_1\right]\right] \\
&= E\left[E\left[Y^{(T_1,\tau_2)}\middle|T_1=\tau_1,S_1\right]\right] \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)}\middle|S_2,T_1,S_1\right]\middle|T_1=\tau_1,S_1\right]\right] \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)}\middle|S_2\right]\middle|T_1=\tau_1,S_1\right]\right] \\
&= E\left[E\left[\theta_2(S_2)\middle|T_1=\tau_1,S_1\right]\right]
\end{aligned}
$$

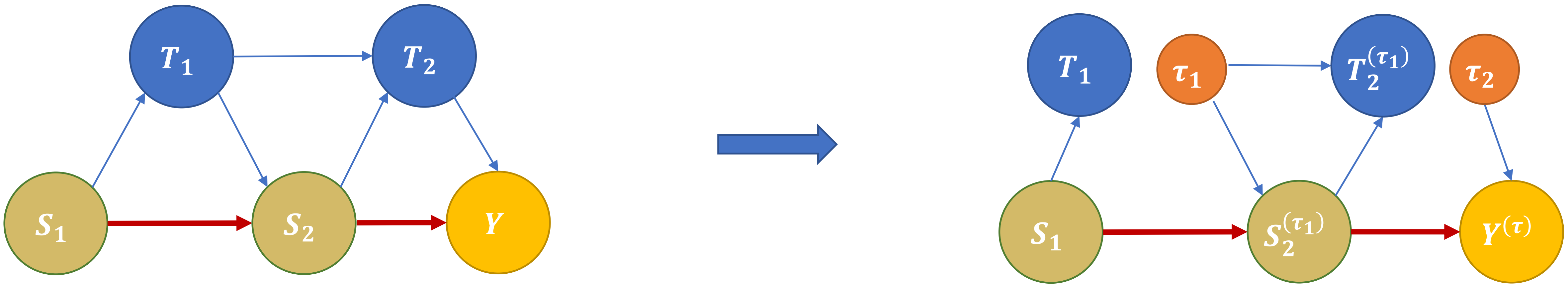
$\boxed{\left(Y^{(\tau)} \perp\!\!\!\perp T_1|S_1\right)}$

(tower law)

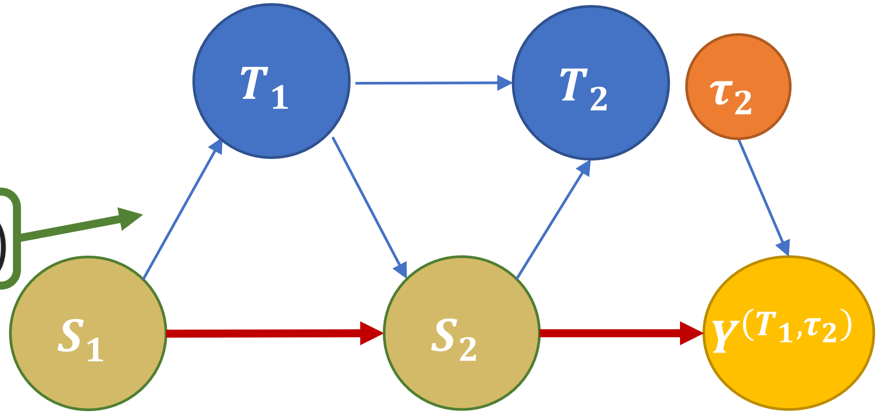
$\left(Y^{(T_1,\tau_2)} \perp\!\!\!\perp S_1,T_1|S_2\right)$



Identification via Backwards Recursion



$$
\begin{aligned}
E\left[Y^{(\tau_1,\tau_2)}\right] &= E\left[E\left[Y^{(\tau_1,\tau_2)} \middle| S_1\right]\right] \\
&= E\left[E\left[Y^{(T_1,\tau_2)} \middle| T_1 = \tau_1, S_1\right]\right] && \left(Y^{(\tau)} \perp\!\!\!\perp T_1 \middle| S_1\right) \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)} \middle| S_2, T_1, S_1\right] \middle| T_1 = \tau_1, S_1\right]\right] && \text{(tower law)} \\
&= E\left[E\left[E\left[Y^{(T_1,\tau_2)} \middle| S_2\right] \middle| T_1 = \tau_1, S_1\right]\right] && \boxed{\left(Y^{(T_1,\tau_2)} \perp\!\!\!\perp S_1, T_1 \middle| S_2\right)} \\
&= E\left[E\left[\theta_2(S_2) \middle| T_1 = \tau_1, S_1\right]\right] && (\text{definition of } \theta_2)
\end{aligned}
$$



Identification Process

- Estimate predictive model $f_2: Y \sim T_2, S_2$
- Adjust outcome for second period treatment

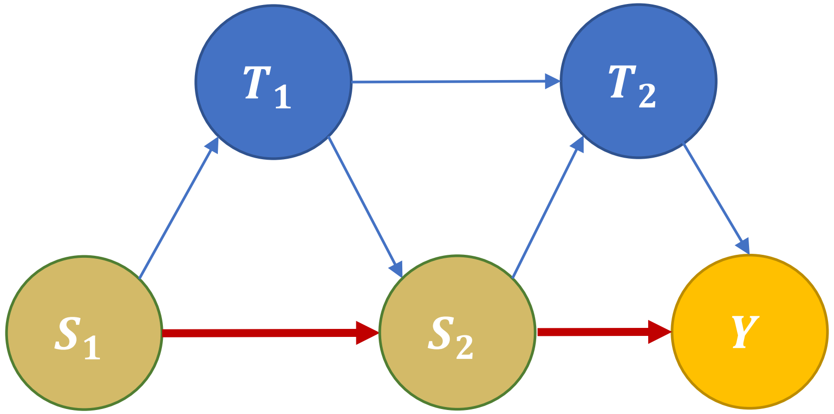
$$Y_{adj} \leftarrow f_2(\tau_2, S_2)$$

- Estimate predictive model $f_1: Y_{adj} \sim T_1, S_1$
- Adjust outcome for first period treatment

$$Y_{adj} \leftarrow f_1(\tau_1, S_1)$$

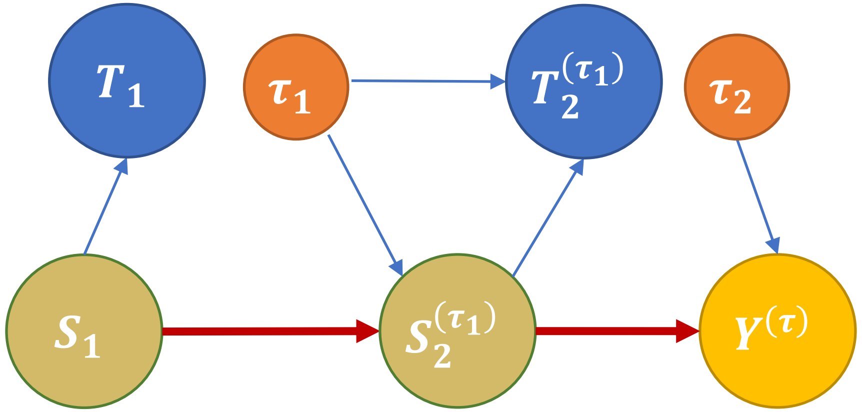
- Return target quantity: $\theta = E\left[Y_{adj}\right]$

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



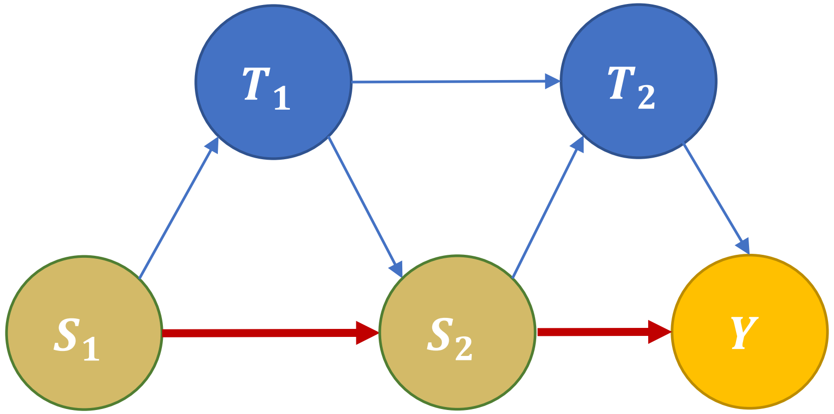
Moment Based Framework

$$\theta = E[f_1(\tau_1, S_1)]$$

$$f_1(T_1, S_1) := E[f_2(\tau_2, S_2) | T_1, S_1]$$
$$f_2(T_2, S_2) := E[Y | T_2, S_2]$$

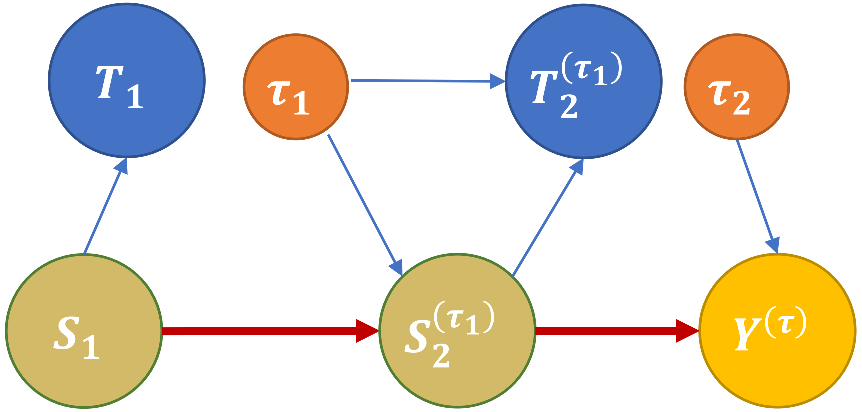
These set of equations (and their generalization to many periods) are known as the g-formula

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



Estimation in the Dynamic Treatment Regime

G-computation

Moment function: $m(Z;\theta,f)$

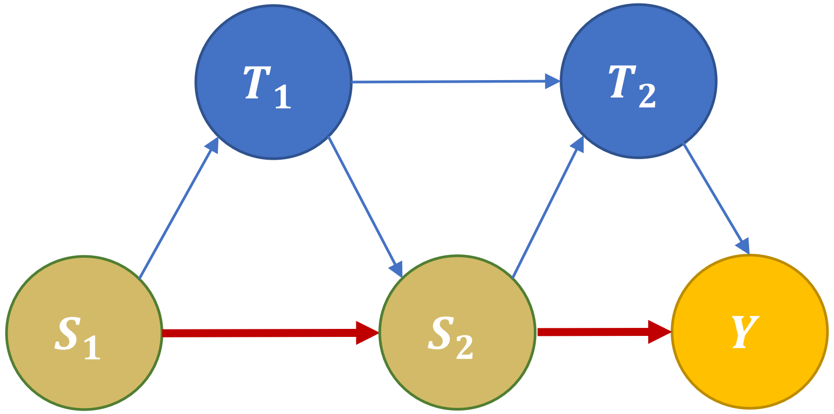
$$E\left[\boxed{\theta - f_1(\tau_1, S_1)}\right] = 0$$

$$f_1(T_1, S_1) := E\left[f_2(\tau_2, S_2) \mid T_1, S_1\right]$$
$$f_2(T_2, S_2) := E\left[Y \mid T_2, S_2\right]$$

Nuisance functions

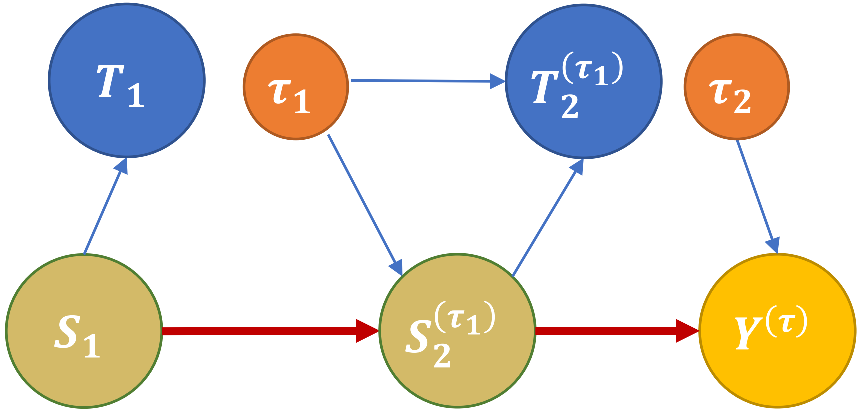
- Using parametric models to estimate these functions (e.g. linear or logistic regression) and plug them in the formula is known as the "parametric g computation"

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



ML Based Estimation

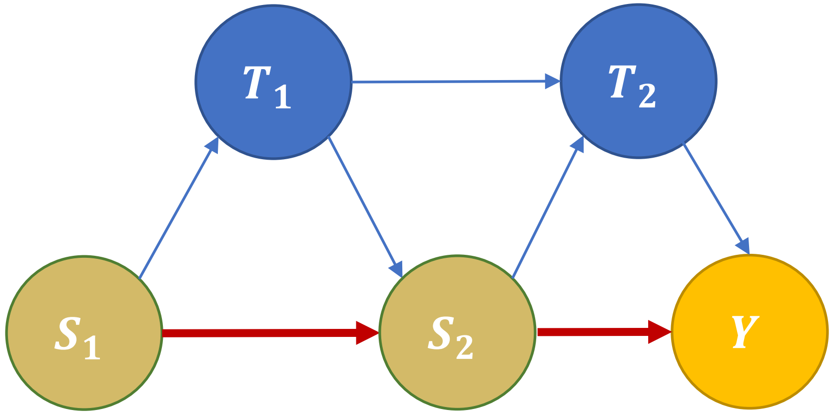
Moment function: $m(Z;\theta,f)$

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)}\right] = 0$$

$$\boxed{\begin{array}{l} f_1(T_1, S_1) \\ f_2(T_2, S_2) \end{array}} \begin{array}{l} := E[f_2(\tau_2, S_2)|T_1, S_1] \\ := E[Y|T_2, S_2] \end{array}$$

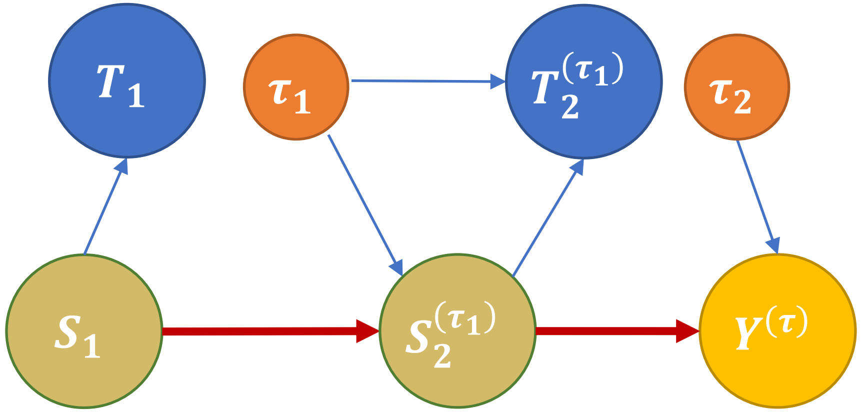
- If we use ML for these predictive problems, then we wont be able to construct confidence intervals for θ because moment is not Neyman orthogonal
- We can apply the Inverse Propensity based debiasing idea

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta := E\left[Y^{(\tau)}\right]$$



Debiasing Moments and Neyman Orthogonality

Original moment

Inverse Propensity

Residual of the Regression Problem that defines nuisance function f_1

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)} + \underbrace{\boxed{\frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}}\left(\boxed{f_2(\tau_2, S_2) - f_1(T_1, S_1)}\right)}_{\text{Debiasing correction for nuisance function } f_1}\right] = 0$$

$$\boxed{\begin{aligned} f_1(T_1, S_1) &:= \\ f_2(T_2, S_2) &:= \end{aligned}}\begin{aligned} &E[f_2(\tau_2, S_2) | T_1, S_1] \\ &E[Y | T_2, S_2] \end{aligned}$$

Nuisance functions

Debiasing Moments and Neyman Orthogonality

We are still left with this nuisance, that we have not "debiased"

Original moment

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)} + \underbrace{\frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}\left(\boxed{f_2(\tau_2, S_2)} - f_1(T_1, S_1)\right)}_{\text{Debiasing correction for nuisance function } f_1}\right] = 0$$

$$\boxed{\begin{aligned} f_1(T_1, S_1) &:= \\ f_2(T_2, S_2) &:= \end{aligned}} \begin{aligned} &E[f_2(\tau_2, S_2) | T_1, S_1] \\ &E[Y | T_2, S_2] \end{aligned}$$

Nuisance functions

Debiasing Moments and Neyman Orthogonality

Original moment

We are still left with this nuisance, that we have not “debiased”

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)} + \frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}\left(\boxed{f_2(\tau_2, S_2)} - f_1(T_1, S_1)\right)\right] = 0$$

$$\underbrace{+ \frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)} \frac{1\{T_2 = \tau_2\}}{\Pr(T_2 = \tau_2 | S_2)}\left(y - f_2(T_2, S_2)\right)}_{\text{Debiasing correction for nuisance function } f_2}$$

$$f_1(T_1, S_1) \coloneqq E[f_2(\tau_2, S_2) | T_1, S_1]$$
$$f_2(T_2, S_2) \coloneqq E[Y | T_2, S_2]$$

Nuisance functions

Debiasing Moments and Neyman Orthogonality

We are still left with this nuisance, that we have not "debiased"

Original moment

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)} + \frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}\left(\boxed{f_2(\tau_2, S_2)} - f_1(T_1, S_1)\right)\right.$$

IPS term already multiplying f_2

$$\left. + \boxed{\frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}} \boxed{\frac{1\{T_2 = \tau_2\}}{\Pr(T_2 = \tau_2 | S_2)}} \boxed{\left(y - f_2(T_2, S_2)\right)}\right] = 0$$

New IPS term for second period treatment introduced to transform $f_2(T_2, S_2) \to f_2(\tau_2, S_2)$

Residual of the regression problem that defines f_2

$$\boxed{\begin{aligned} f_1(T_1, S_1) \\ f_2(T_2, S_2) \end{aligned}} \begin{aligned} &:= E[f_2(\tau_2, S_2) | T_1, S_1] \\ &:= E[Y | T_2, S_2] \end{aligned}$$

Nuisance functions

Debiasing Moments and Neyman Orthogonality

Original moment

$$E\left[\boxed{\theta - f_1(\tau_1, S_1)} + \frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)}\left(f_2(\tau_2, S_2) - f_1(T_1, S_1)\right)\right] = 0$$

$$+ \frac{1\{T_1 = \tau_1\}}{\Pr(T_1 = \tau_1 | S_1)} \frac{1\{T_2 = \tau_2\}}{\Pr(T_2 = \tau_2 | S_2)}\left(y - f_2(T_2, S_2)\right)$$

$$\boxed{\begin{aligned} f_1(T_1, S_1) \\ f_2(T_2, S_2) \end{aligned}} \begin{aligned} &\coloneqq E[f_2(\tau_2, S_2) | T_1, S_1] \\ &\coloneqq E[Y | T_2, S_2] \end{aligned}$$

Nuisance functions

- This moment now satisfies Neyman orthogonality with respect to all the nuisance functions
- We need to also estimate the propensity functions via generic ML classification

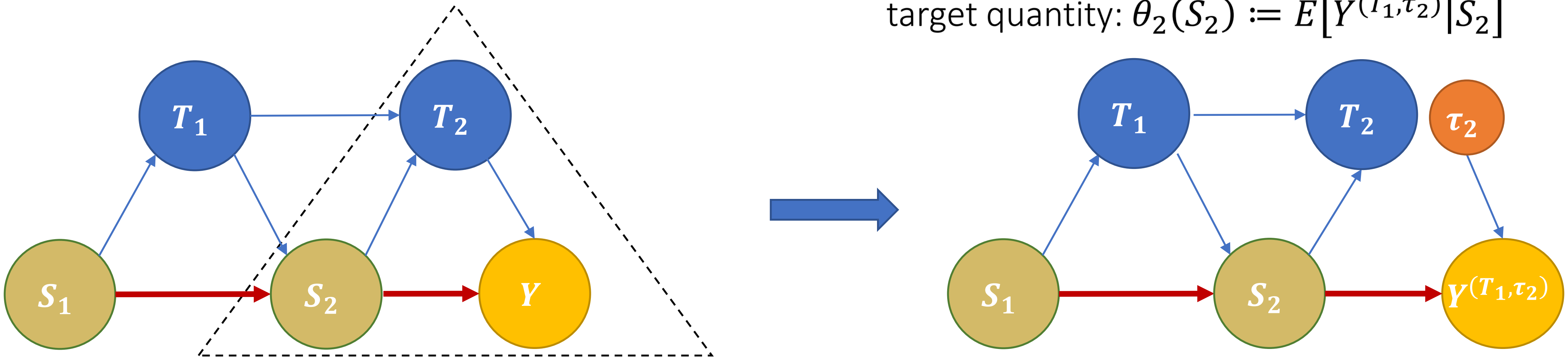
Continuous Treatments and Alternative Approach to Identification and Estimation

What if we have continuous treatments

- The propensity based approach to de-biasing does not apply
- What is the analogue of the “Residual-on-Residual” or “Partialling-Out” (FWL) approach for the dynamic treatment regime, which is also applicable to continuous treatments?

Identification via "Instantaneous" or "Blip" Effects

target quantity: $\theta_2(S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} \middle| S_2\right]$



- Instead of estimating the outcome $Y^{(T_1,\tau_2)}$, we will estimate the "effect" of T_2

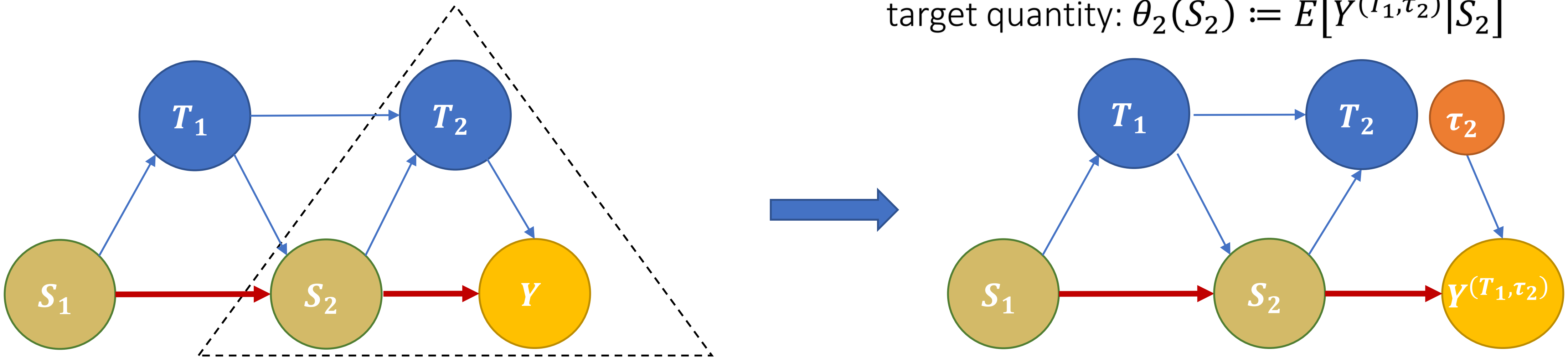
$$\alpha_2(\tau_2, S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} - Y^{(T_1,0)} \middle| S_2\right]$$

- Adjust outcome by subtracting effect of observed treatment and adding effect of target treatment

$$Y_{\text{adj}} \coloneqq Y - \alpha_2(T_2, S_2) + \alpha_2(\tau_2, S_2)$$

Identification via “Instantaneous” or “Blip” Effects

target quantity: $\theta_2(S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} | S_2\right]$



- Instead of estimating the outcome $Y^{(T_1,\tau_2)}$, we will estimate the “effect” of T_2

$$\alpha_2(\tau_2, S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} - Y^{(T_1,0)} \middle| S_2\right]$$

- Adjust outcome by subtracting effect of observed treatment and adding effect of target treatment

$$Y_{\text{adj}} \coloneqq Y - \alpha_2(T_2, S_2) + \alpha_2(\tau_2, S_2)$$

- If “effect” is assumed to have a simple parametric form, then this approach leverages this simplicity!

Repeat in a backwards manner

Identification Process

Estimate the "effect" $\alpha_2(\tau_2, S_2)$

$$\alpha_2(\tau_2, S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} - Y^{(T_1,0)} \middle| S_2\right]$$

Adjust outcome for second period treatment

$$Y_{\text{adj}} \leftarrow Y - \alpha_2(T_2, S_2) + \alpha_2(\tau_2, S_2)$$

Estimate the effect $\alpha_1(\tau_1, S_1)$

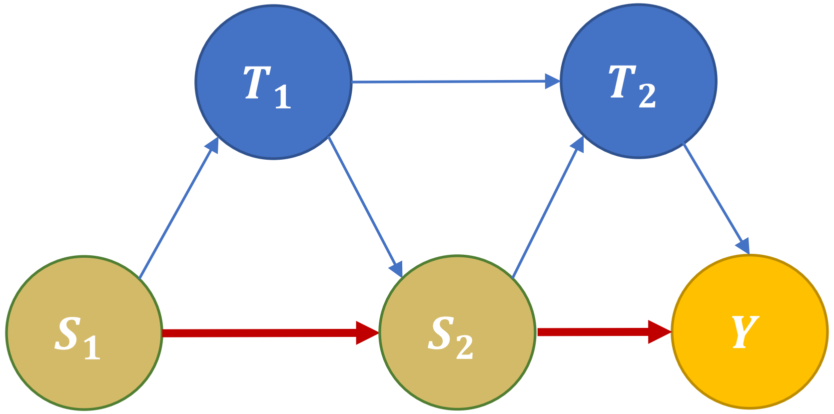
$$E\left[Y^{(\tau_1,\tau_2)} - Y^{(0,\tau_2)} \middle| S_1\right] \approx E\left[Y_{\text{adj}}^{(\tau_1)} - Y_{\text{adj}}^{(0)} \;\middle|\; S_1\right]$$

Adjust outcome for first period treatment

$$Y_{\text{adj}} \leftarrow Y_{\text{adj}} - \alpha_1(T_1, S_1) + \alpha_1(\tau_1, S_1)$$

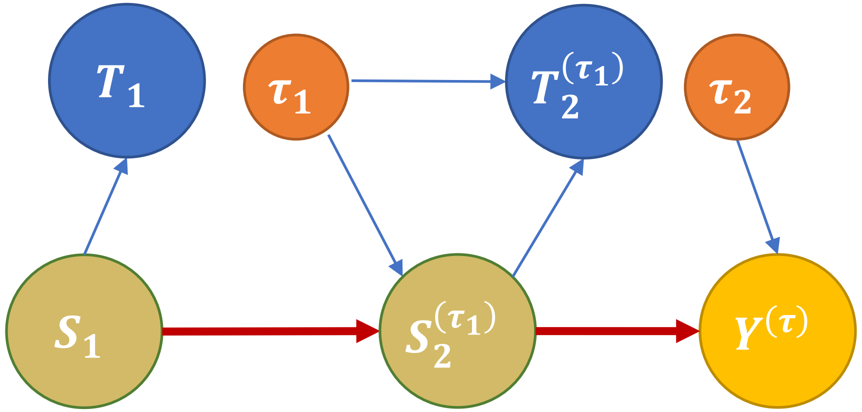
Return target quantity: $\theta = E\left[Y_{\text{adj}}\right]$

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta \coloneqq E\left[Y^{(\tau)}\right]$$

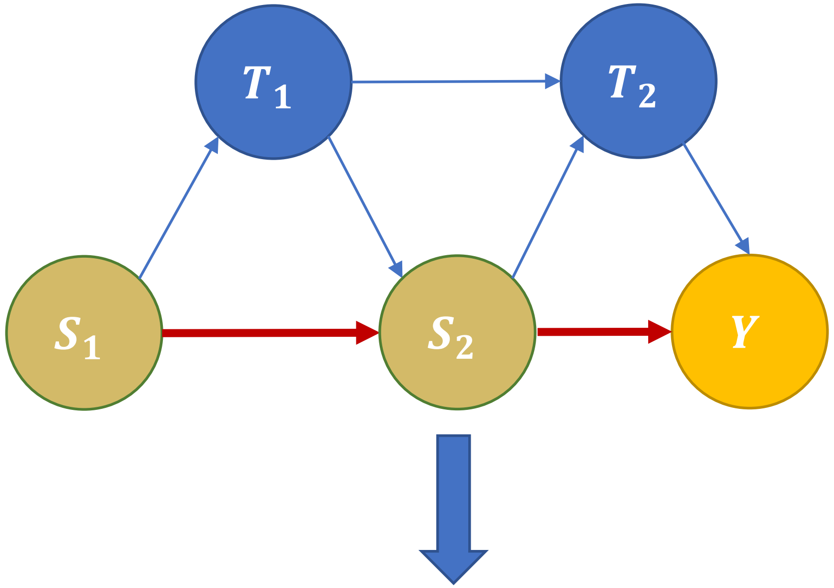


Partial Linearity: Linear Effects

- If we assume that these effects are partially linear:

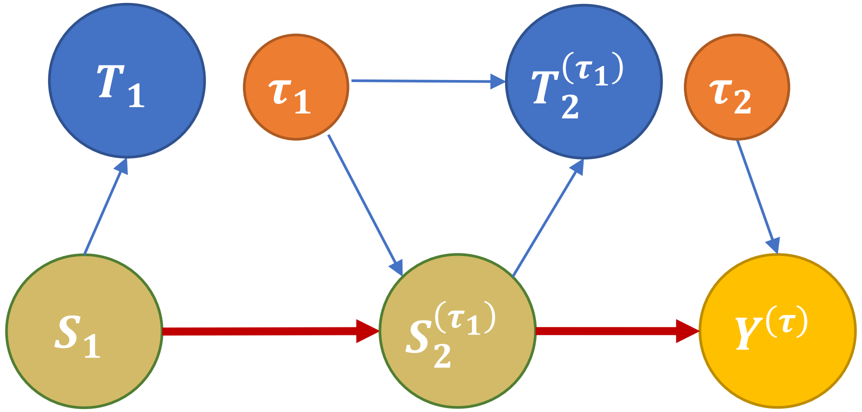
$$a_2(\tau_2, S_2) = \delta_2 \tau_2$$
$$a_1(\tau_1, S_1) \coloneqq \delta_1 \tau_1$$

observed data (panel)



target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

$$\theta \coloneqq E\left[Y^{(\tau)}\right]$$

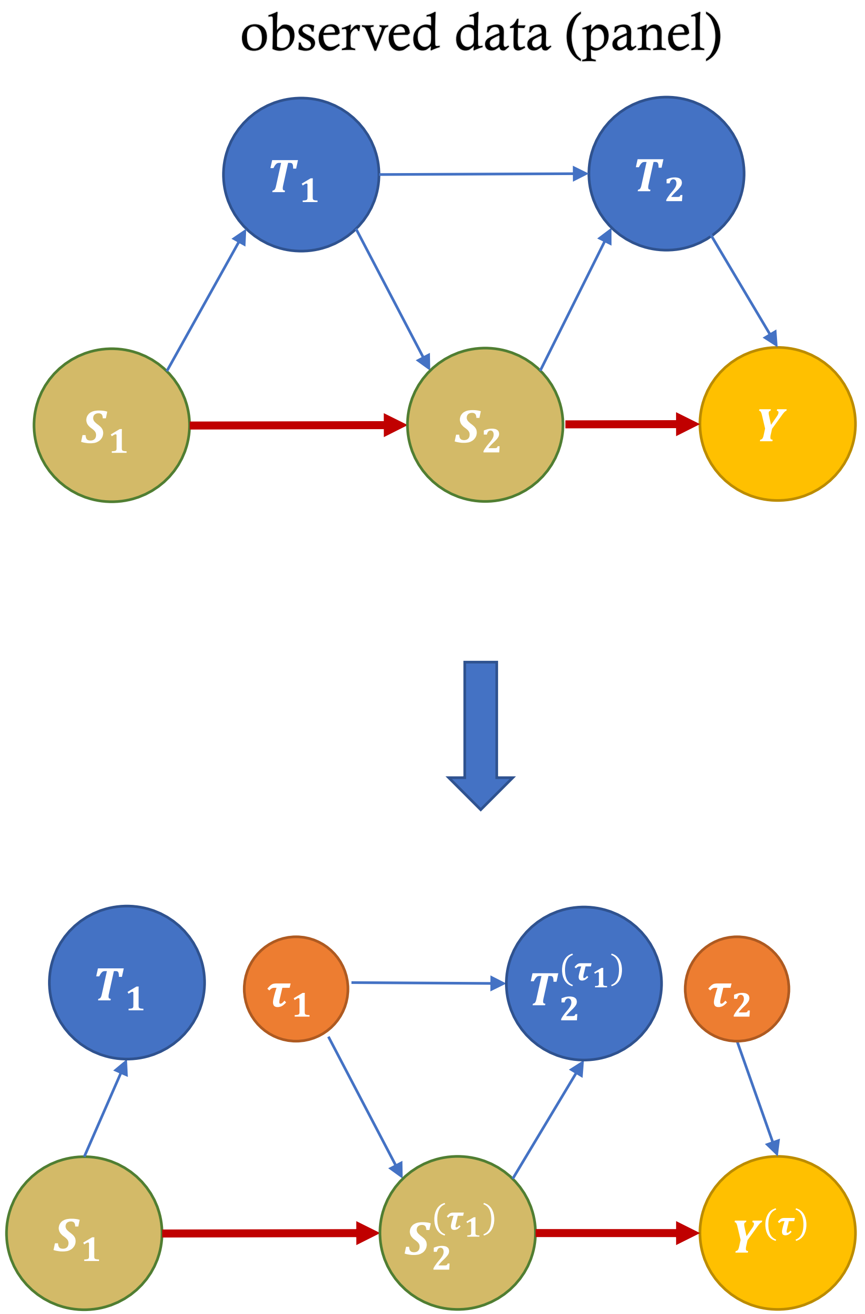


Identification Process

- If we assume that these effects are partially linear:

$$a_2(\tau_2, S_2) = \delta_2 \tau_2$$
$$a_1(\tau_1, S_1) \coloneqq \delta_1 \tau_1$$

- We can estimate these effects via partialling out!



Dynamic DML (Partialling Out)

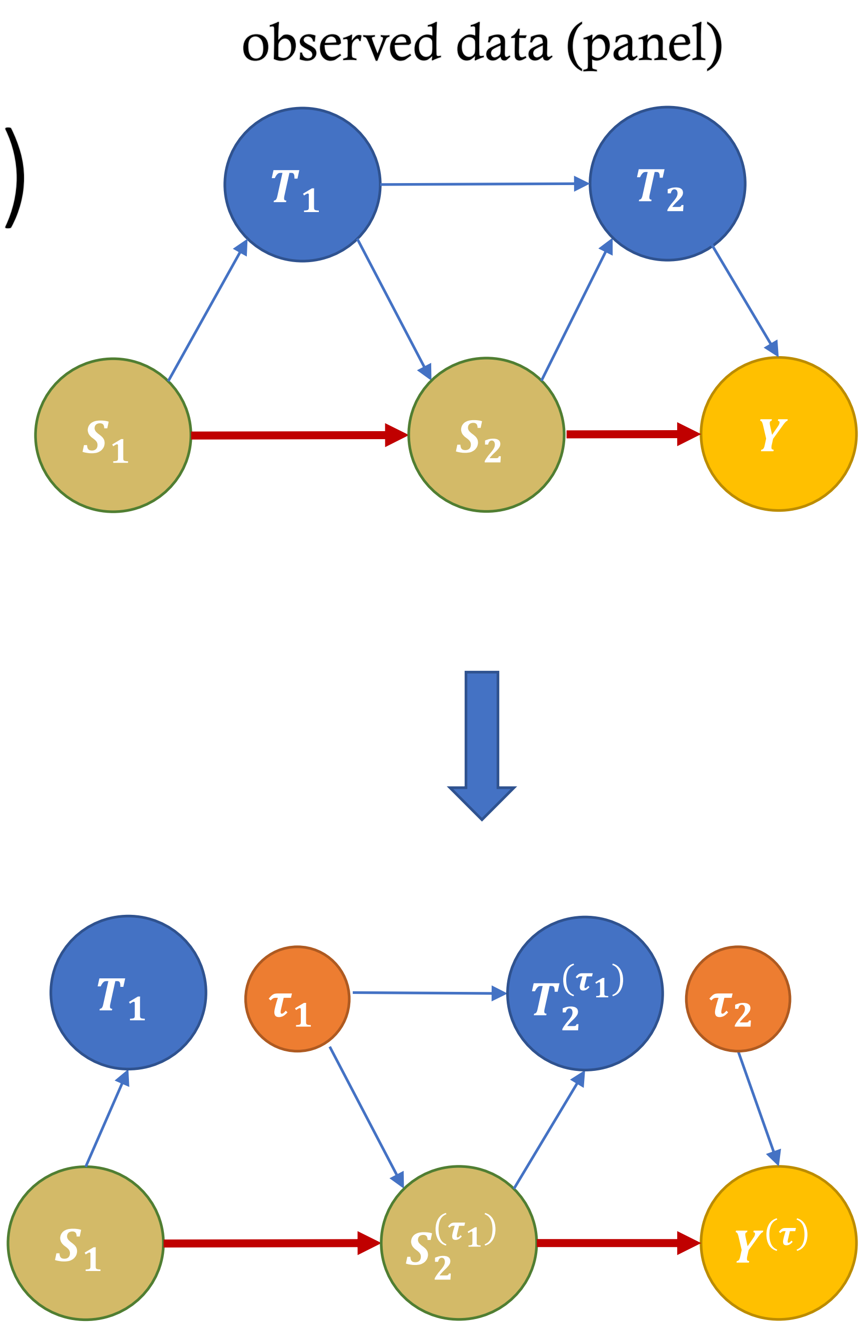
- Construct residuals
  $\tilde{Y}_2 := Y - E[Y|S_2], \qquad \tilde{T}_2 = T_2 - E[\,T_2 \mid S_2\,]$
- Run OLS: $\tilde{Y}_2 \sim \tilde{T}_2$ to estimate δ_2

- Adjust outcome $Y_{adj} = Y - \delta_2 T_2 + \delta_2 \tau_2$

- Construct residuals:
  $\tilde{Y}_1 := Y_{adj} - E\left[Y_{adj}\middle|S_1\right], \qquad \tilde{T}_1 := T_1 - E[T_1|S_1]$
- Run OLS: $\tilde{Y}_1 \sim \tilde{T}_1$ to estimate δ_1

- Adjust outcome $Y_{adj} \leftarrow Y_{adj} - \delta_1 T_1 + \delta_1 \tau_1$

- Return: $\theta := E\left[Y_{adj}\right]$



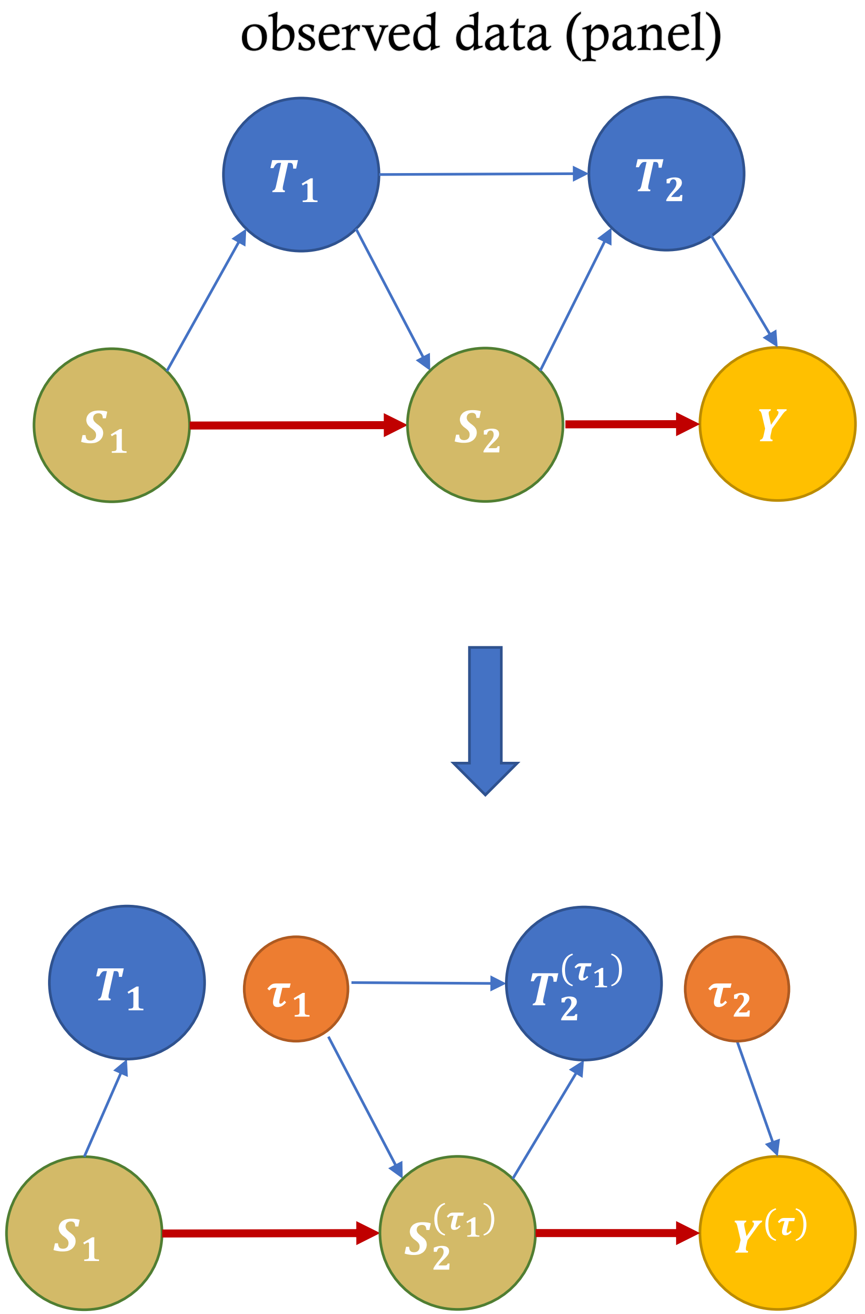
Moment Based Framework

$$\theta \coloneqq E[Y - \delta_2 T_2 + \delta_2 \tau_2 - \delta_1 T_1 + \delta_1 \tau_1] = 0$$

$$E\left[\left(\tilde{Y}_2 - \delta_2 \tilde{T}_2\right)\tilde{T}_2\right] = 0$$
$$E\left[\left(\tilde{Y}_1 - \delta_1 \tilde{T}_1\right)\tilde{T}_1\right] = 0$$

$$\boxed{\begin{array}{c} E[Y|S_2], \qquad \mathrm{E}[\mathrm{T}_2|\mathrm{S}_2] \\ E[Y - \delta_2 T_2 + \delta_2 \tau_2 | S_1], \qquad \mathrm{E}[\mathrm{T}_1|\mathrm{S}_1] \end{array}}$$

Nuisance functions



Moment Based Framework

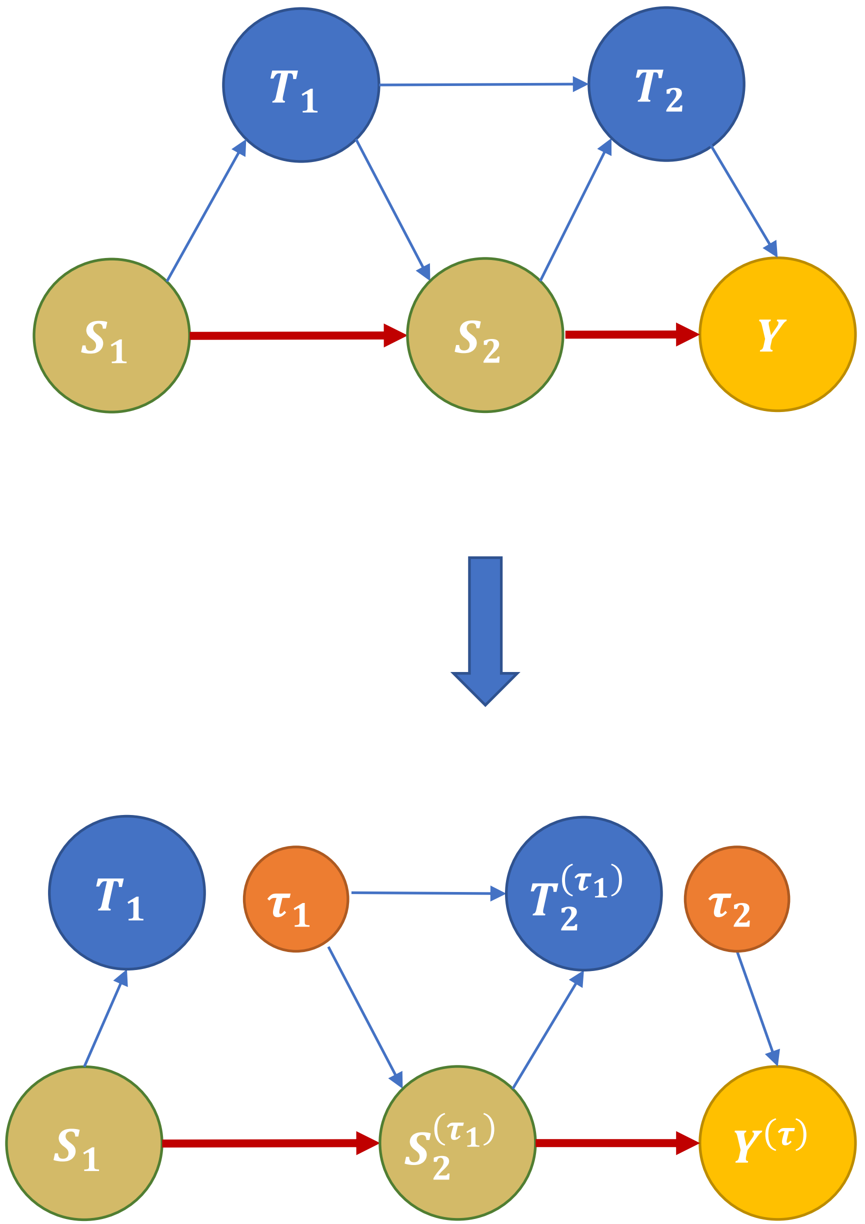
$$\theta := E[Y - \delta_2 T_2 + \delta_2 \tau_2 - \delta_1 T_1 + \delta_1 \tau_1] = 0$$

$$E\left[\left(\tilde{Y}_2 - \delta_2 \tilde{T}_2\right)\tilde{T}_2\right] = 0$$
$$E\left[\left(\tilde{Y}_1 - \delta_1 \tilde{T}_1\right)\tilde{T}_1\right] = 0$$

$$\boxed{\begin{array}{cc} E[Y|S_2], & \mathrm{E}[\mathrm{T}_2|\mathrm{S}_2] \\ E[Y - \delta_2 T_2 + \delta_2 \tau_2 | S_1], & \mathrm{E}[\mathrm{T}_1|\mathrm{S}_1] \end{array}}$$

For the same reason why the Residual-on-Residual moment was Neyman orthogonal, this estimation process is also orthogonal and we can use ML for these problems

observed data (panel)



An Application from Operations Management

Return on Investment at Microsoft

Estimating Long-Term Returns on Investment

- Companies frequently deploys new discount or customer support programs
- Which of these programs ("investments") are more successful than others?
- Success is a long-term objective: what is the effect of the program on the two-year customer journey (e.g., effect on two-year revenue)
- We cannot wait two years to evaluate a program
- Main Question. Can we construct estimates of the values of these programs with short-term data, e.g. after 6 months?

Long-Term Effects from Short-Term Surrogates

- Suppose that there are many short-term signals S that are indicative of a customer's long-term reward Y (e.g. the next 6-month purchase patterns of a customer could be indicative of their long-term spend)
- Suppose that investment program T affects long-term rewards if and only if it affects these short-term signals

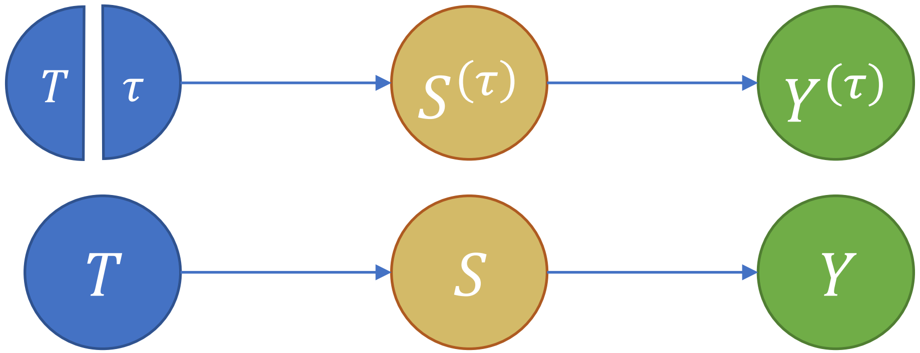
- We will call these short-term signals S surrogates

Causal Inference with Surrogates 101

- Since long-term effect goes only through surrogates:

  expected effect on long-term reward = effect on projected long-term reward based on surrogates

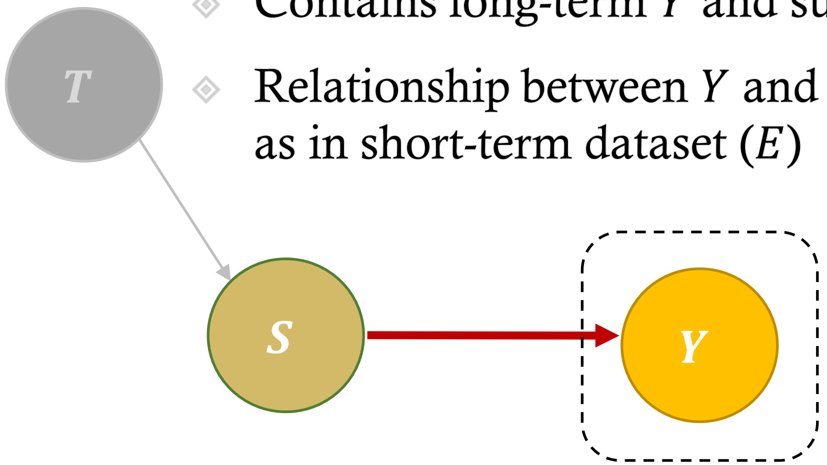
$$\underbrace{E\left[Y^{(\tau)}\right]}_{\substack{\text{Average reward}\\\text{if intervene and}\\\text{set investment}=\tau}} = \underbrace{E\left[Y^{(\tau)}\middle|T=\tau\right]}_{\substack{\text{Independence in}\\\text{counterfactual}\\\text{graph}}} = \underbrace{E[Y|T=\tau]}_{\substack{\text{Average reward of}\\\text{samples that received}\\\text{investment}=\tau\text{ in data}}} = \underbrace{E[E[Y|T=\tau,S]|T=\tau]}_{\substack{\text{Tower Law of}\\\text{Expectations}}} = E[\underbrace{E[Y|S]}_{\substack{\text{Forecasted}\\\text{reward from}\\\text{surrogates}}}|T=\tau]$$



Causal Inference with Surrogates 101

historical/long-term (O)

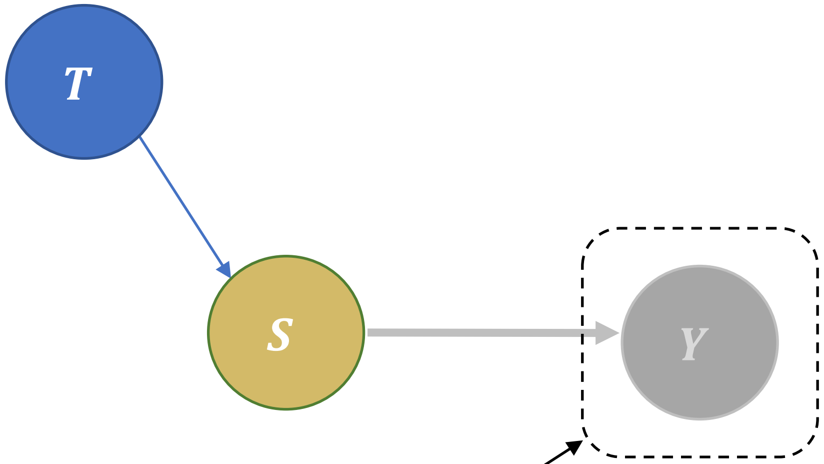
- Contains long-term Y and surrogates S
- Relationship between Y and S the same as in short-term dataset (E)



$g(S) = E_o[Y|S]$

1. Estimate $g(S) \coloneqq E[Y|S]$ (surrogate index) from (O) by regressing $Y \sim S$

recent/short-term (E)

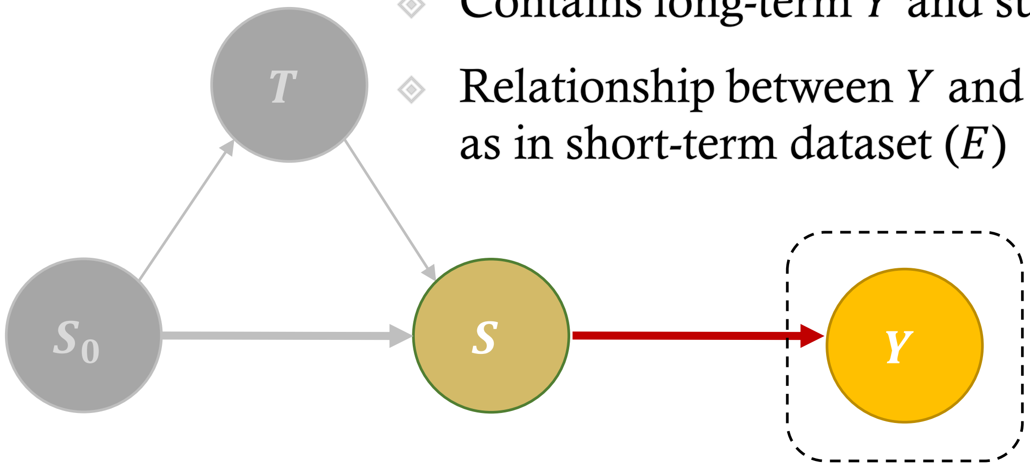


2. Impute expected long-term outcomes in (E)
3. Regress $g(S) \sim T$ to estimate effect of T on Y from (E)

Causal Inference with Surrogates 101

historical/long-term (O)

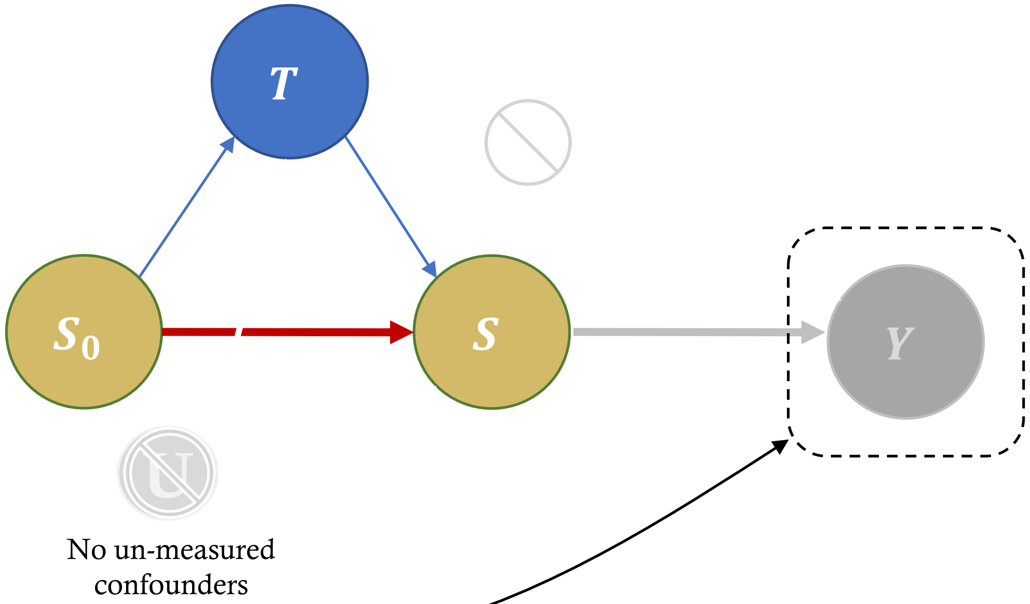
- Contains long-term Y and surrogates S
- Relationship between Y and S the same as in short-term dataset (E)



$g(S) = E_o[Y|S]$

1. Estimate $g(S) \coloneqq E[Y|S]$ (surrogate index) from (O) by regressing $Y \sim S$

recent/short-term (E)



2. Impute expected long-term outcomes in (E)
3. Regress $g(S) \sim T$ to estimate effect of T on Y from (E)

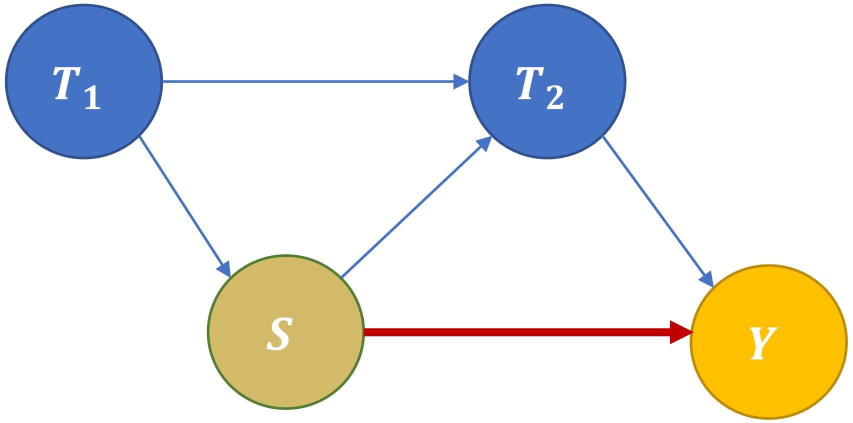
Key Assumptions

- Long-term effect only goes through surrogates
- Expected relationship between surrogates and long-term reward is the same long-term setting (O) and in short-term setting (E)

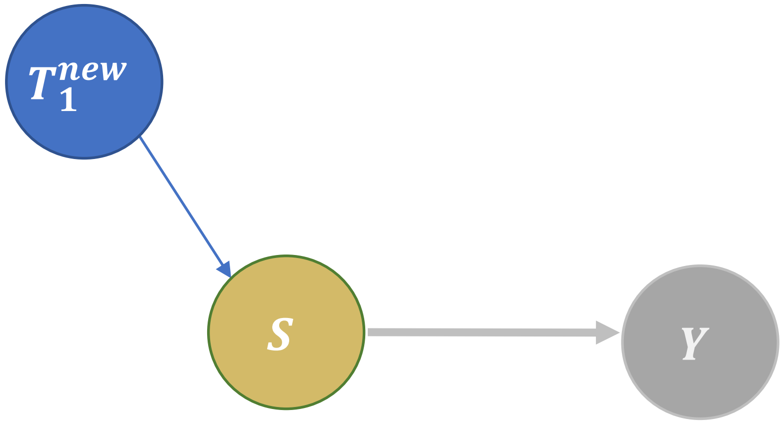
Key Assumptions can be Easily Violated

Investment policies are dynamic and change

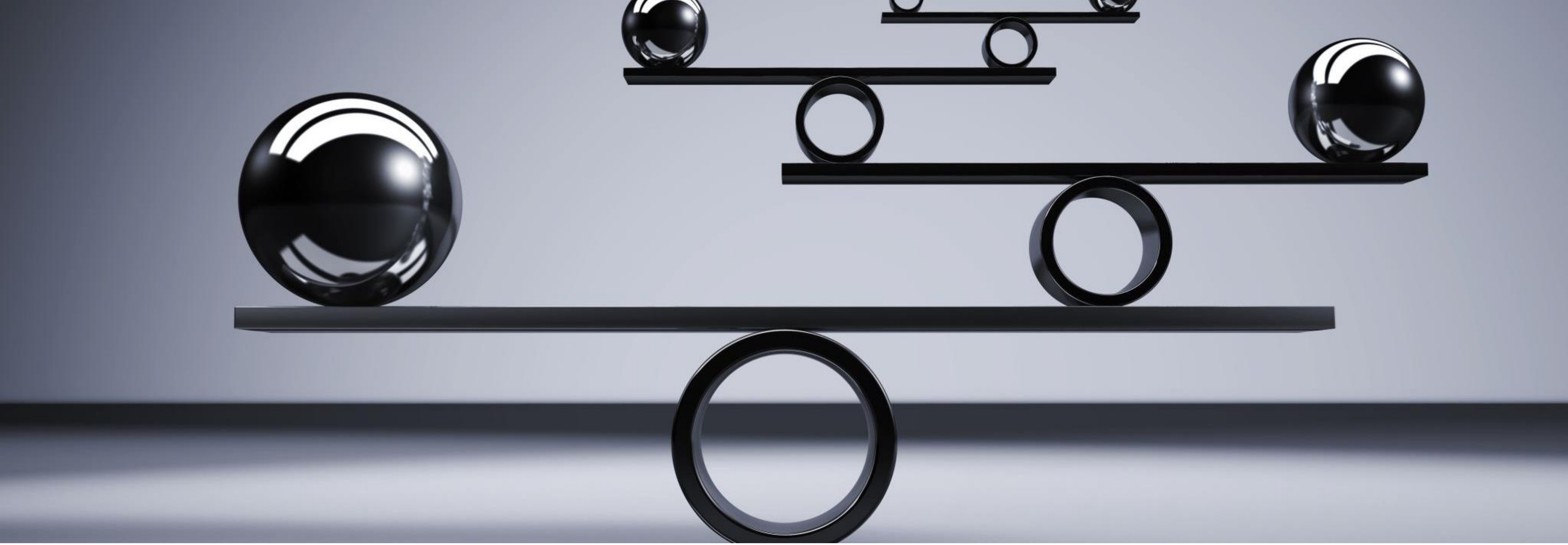
historical/long-term (O)



recent/short-term (E)



- We deployed older/deprecated investments
- In a potentially long-term highly auto-correlated manner
- Investments are potentially adaptive
- Investment policies change

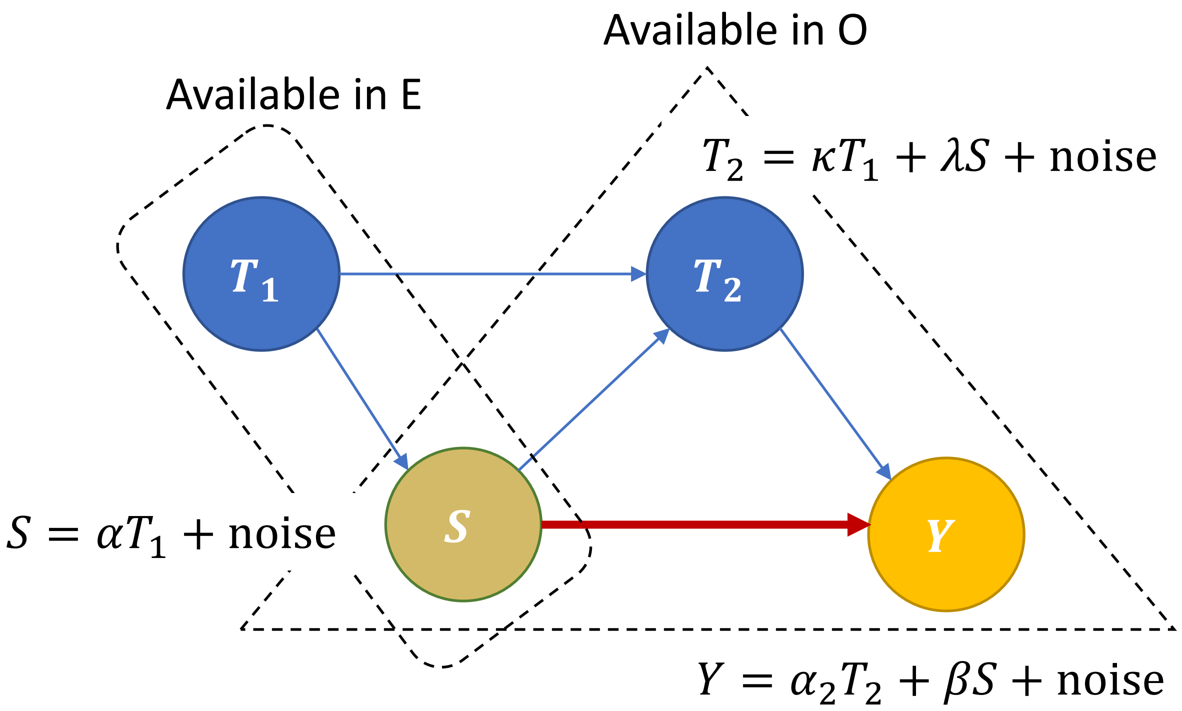


Bias of Vanilla Surrogate

Illustrative Example

Illustrative Example

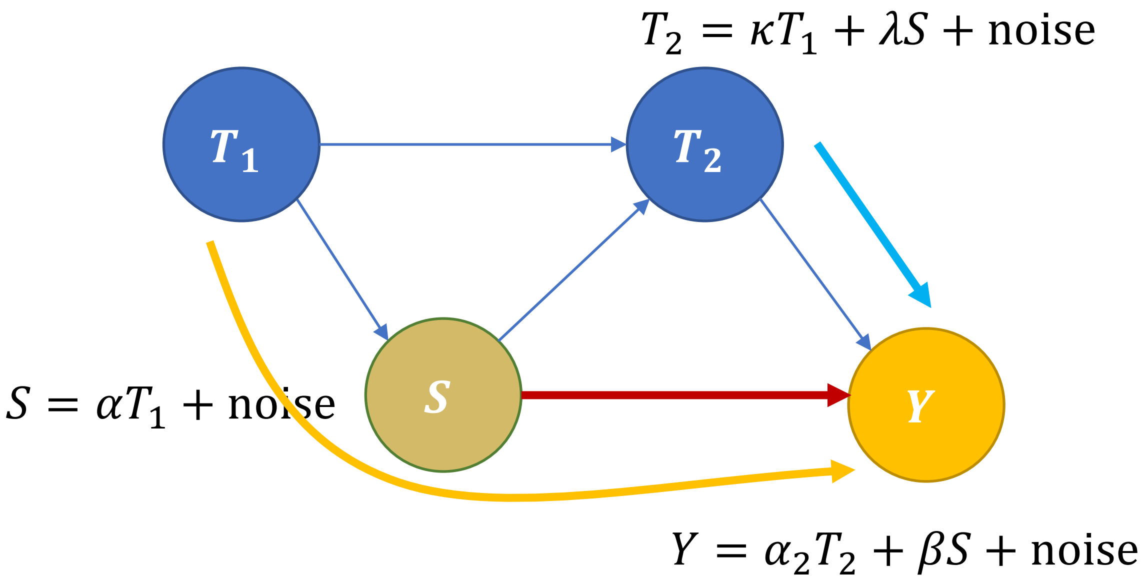
- Want the effect of T_1 on Y, with no future treatments



Bias of Vanilla Surrogate

- What do we want to estimate?

$$Y = \underset{\substack{\parallel \\ \boldsymbol{\theta_0}}}{\boldsymbol{\beta\,\alpha}}\, T_1 + \alpha_2\, T_2 + noise$$



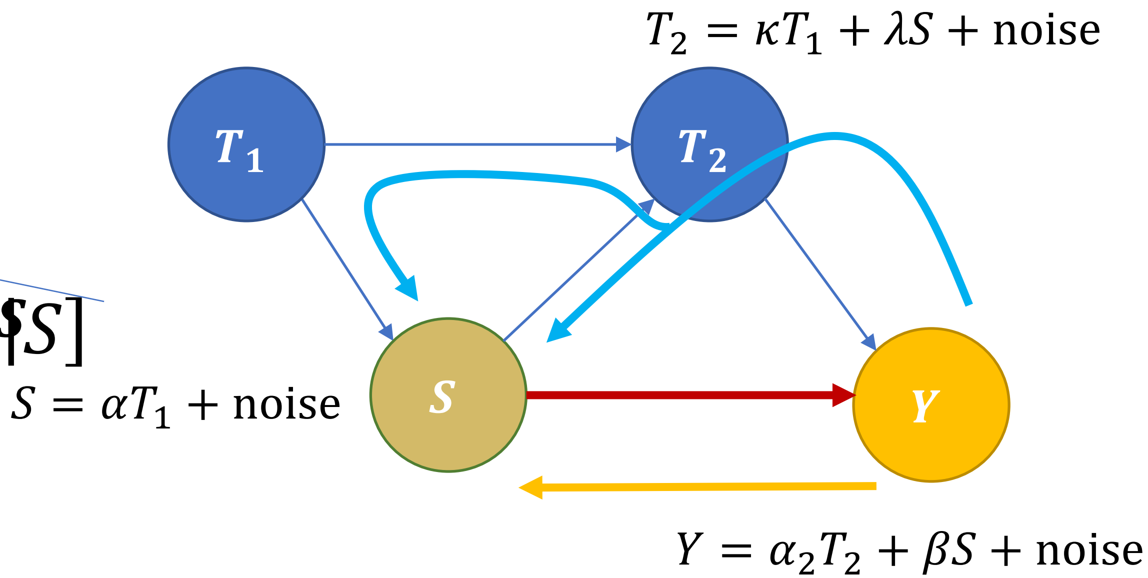
Bias of Vanilla Surrogate

- What do we want to estimate?

$$Y = \beta \underbrace{\alpha}_{\theta_0} T_1 + \alpha_2 T_2 + noise$$

- What does the surrogate approach estimate?

$$g_0(S) \coloneqq E[Y|S] = \beta S + \alpha_2 \overbrace{E[T_2|S]}^{\kappa E[T_1|S] + \lambda S}$$



Bias of Vanilla Surrogate

- What do we want to estimate?

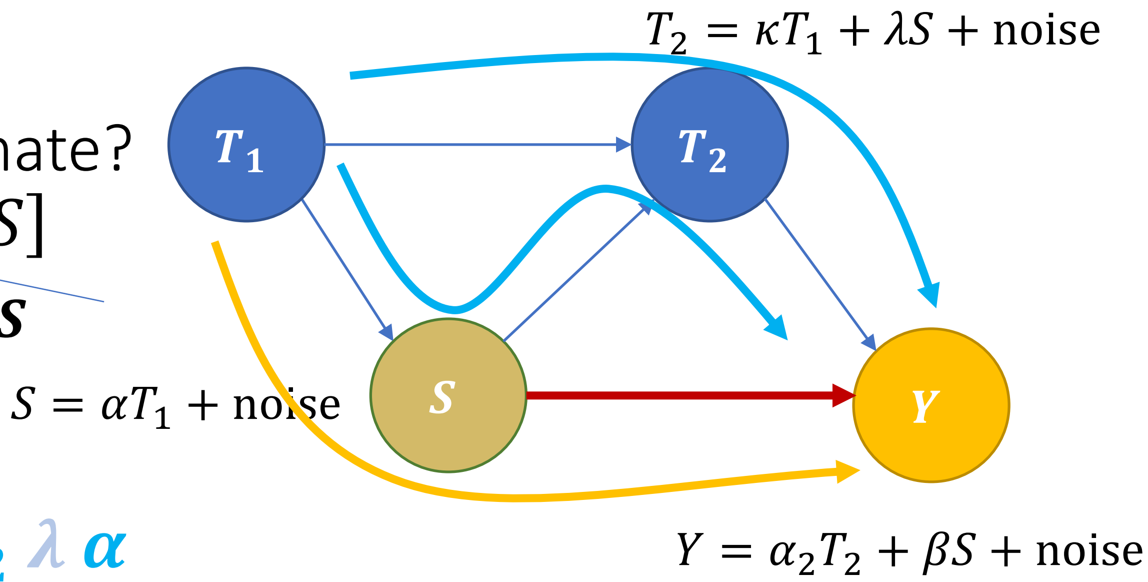
$$Y = \underset{\substack{\| \\ \boldsymbol{\theta_0}}}{\boldsymbol{\beta}\,\boldsymbol{\alpha}}\, T_1 + \alpha_2\, T_2 + noise$$

- What does the surrogate approach estimate?

$$g_0(S) \coloneqq E[Y|S] = \boldsymbol{\beta}\, S + \boldsymbol{\alpha_2}\, \underbrace{E[T_2|S]}_{\kappa E[T_1|S] + \lambda \boldsymbol{S}}$$

- The effect of T_1 on $g_0(S)$ is:

$$\theta_* \coloneqq \boldsymbol{\beta}\,\boldsymbol{\alpha} + \boldsymbol{\alpha_2}\,\boldsymbol{\kappa}\,\frac{E[E[T_1|S]\cdot T_1]}{E[T_1^2]} + \boldsymbol{\alpha_2}\,\boldsymbol{\lambda}\,\boldsymbol{\alpha}$$



Bias of Vanilla Surrogate

- What do we want to estimate?

$$Y = \beta\, \underbrace{\alpha}_{\theta_0}\, T_1 + \alpha_2\, T_2 + noise$$

- What does the surrogate approach estimate?

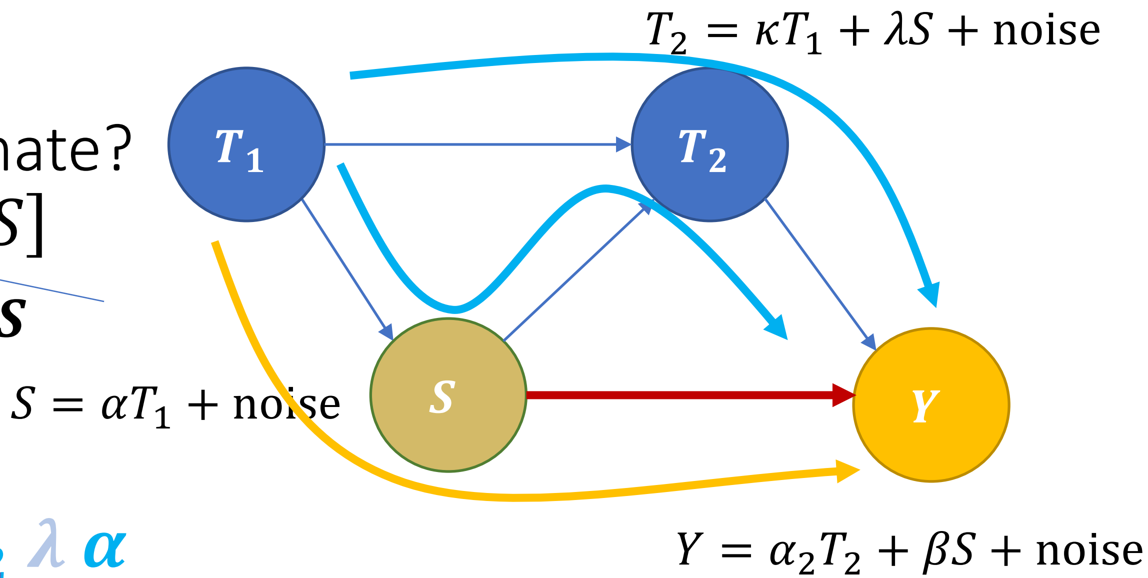
$$g_0(S) \coloneqq E[Y|S] = \beta\, S + \alpha_2\, \underbrace{E[T_2|S]}_{\kappa E[T_1|S] + \lambda \boldsymbol{S}}$$

- The effect of T_1 on $g_0(S)$ is:

$$\theta_* \coloneqq \beta\, \alpha + \underbrace{\alpha_2\, \kappa \frac{E[E[T_1|S] \cdot T_1]}{E[T_1^2]}} + \underbrace{\alpha_2\, \lambda\, \alpha}$$

Bias from the fact that policy is auto-correlated violating the "effect only through surrogates" assumption

Bias from the fact that policy is adaptive on past surrogates violating "relationship of S and Y unchanged" assumption



Prevalence of Such Violation

- Marketing departments frequently deploy new marketing campaigns
  - In historical data, older campaigns were deployed
  - Due to targeting, if a customer received an ad, most probably will receive one in the future
  - Ad display is adaptive to signals of customer behavior
- Pharmaceutical companies frequently deploy new drugs
  - Can we get estimates of long-term effect from short-term trials
  - Use short-term signals of patient response that correlated with long-term survival
  - In historical long-term clinical data, patients are treated with multiple treatments over time
  - Treatments are typically adaptive to signals of patient trajectory

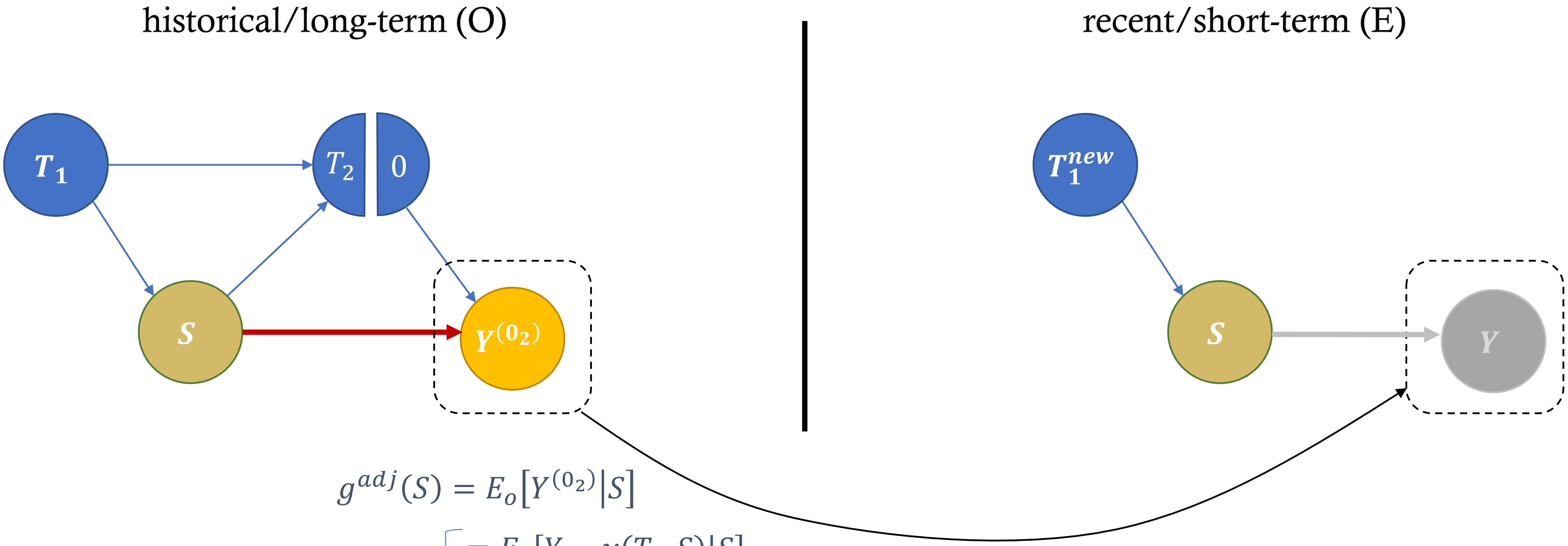
Key Idea 1

Dynamic Adjustment of Surrogate Index

Battocchi, Dillon, Hei, Lewis, Oprescu, Syrgkanis, Estimating the Long-Term Effects of Novel Treatments, NeurIPS’21

https://arxiv.org/abs/2103.08390

Dynamically Adjusted Surrogate Index



$$g^{adj}(S) = E_o\left[Y^{(0_2)} \middle| S\right]$$

$$= E_o[Y - \underbrace{\gamma(T_2, S)}_{\text{"blip" effect of second treatment conditional on surrogate}} | S]$$

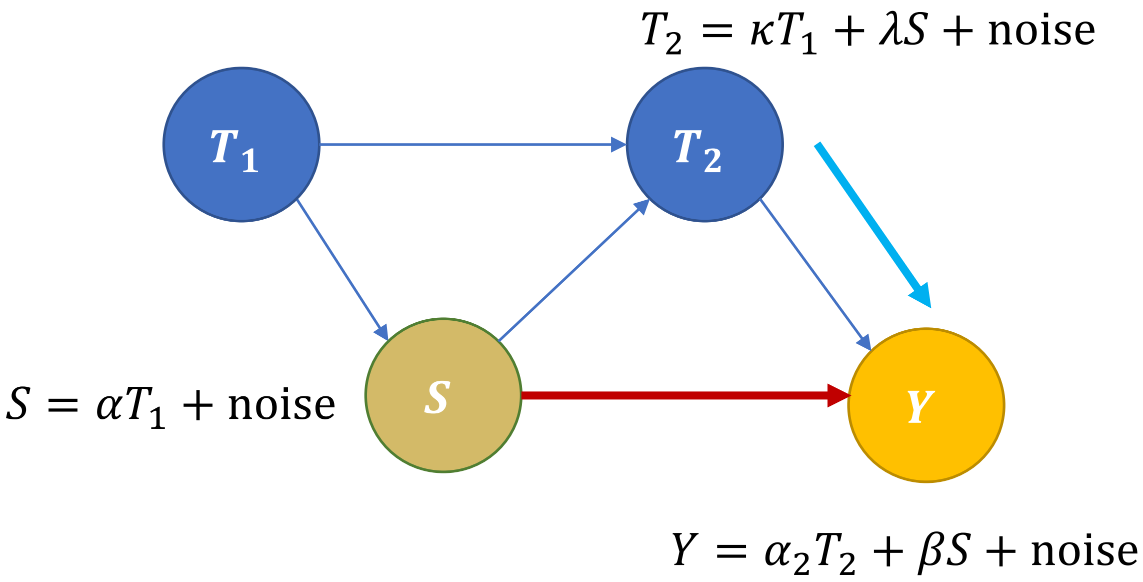
Identification argument

effect of T_2 controlling for S

Illustrative Example: Dynamically Adjusted Index

- Estimate effect α_2 of T_2 on Y, controlling for S

$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$



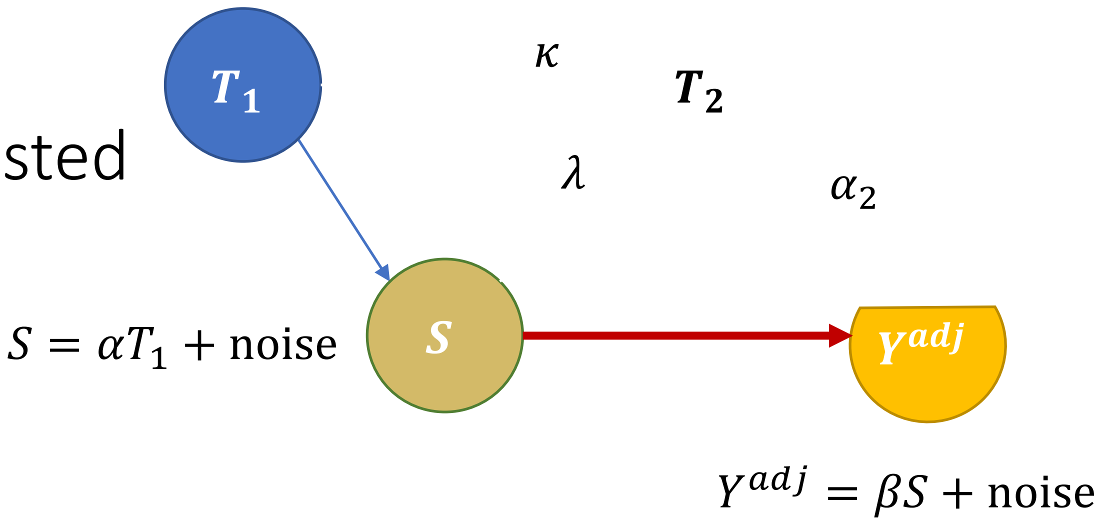
Illustrative Example: Dynamically Adjusted Index

- Estimate effect α_2 of T_2 on Y, controlling for S

$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$

- Subtract that effect to create the adjusted outcome

$$Y^{adj} := Y - \alpha_2 T_2$$



Illustrative Example: Dynamically Adjusted Index

- Estimate effect α_2 of T_2 on Y, controlling for S

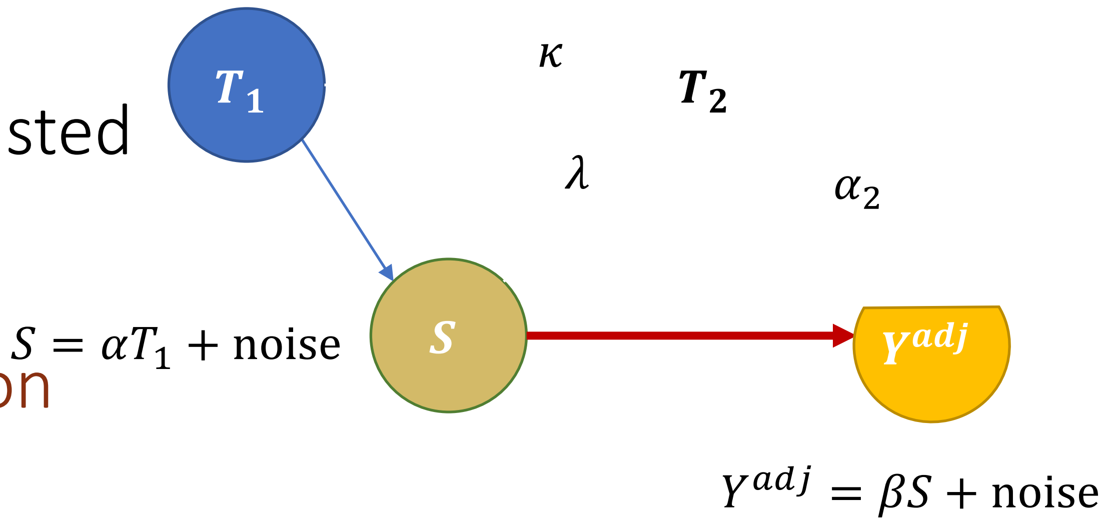
$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$

- Subtract that effect to create the adjusted outcome

$$Y^{adj} := Y - \alpha_2 T_2$$

- Estimate dynamically adjusted index on long-term data

$$g_0^{adj}(S) := E\left[Y^{adj} \middle| S\right] = \beta\, S$$



Illustrative Example: Dynamically Adjusted Index

- Estimate effect α_2 of T_2 on Y, controlling for S

$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$

- Subtract that effect to create the adjusted outcome

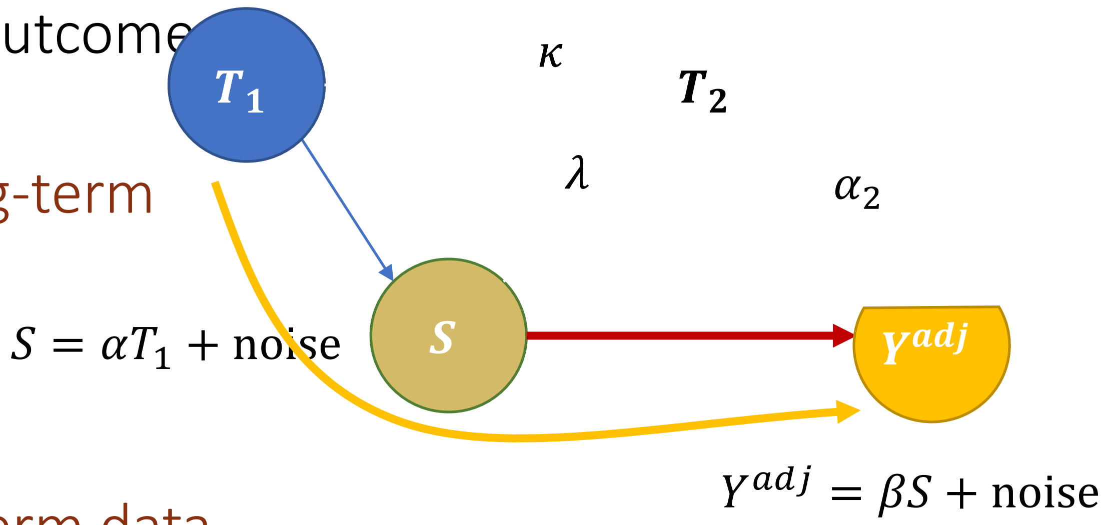
$$Y^{adj} := Y - \alpha_2 T_2$$

- Estimate dynamically adjusted index on long-term data

$$g_0^{adj}(S) := E\left[Y^{adj} \middle| S\right] = \beta\, S$$

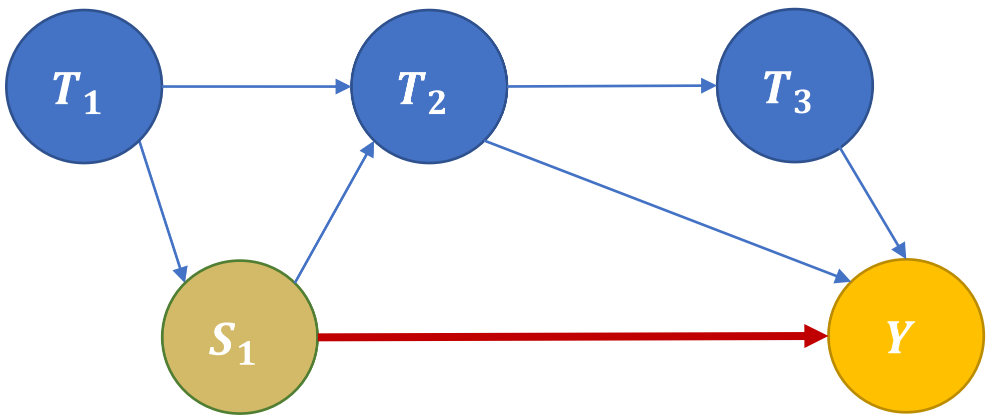
- Estimate effect of T_1 on $g_0^{adj}(S)$ on short-term data

$$\theta_0 := \frac{E\left[g_0^{adj}(S) \cdot T_1\right]}{E\left[T_1^2\right]} = \beta \cdot \alpha$$

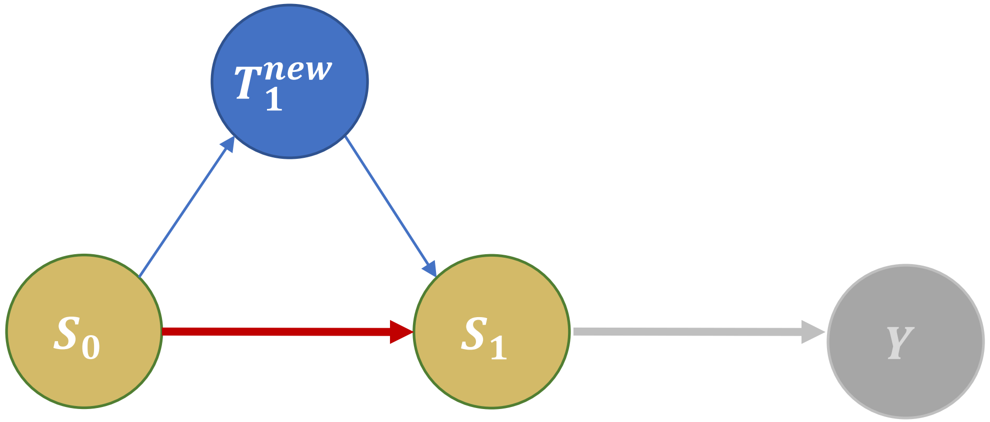


Beyond Two Periods

historical/long-term (O)



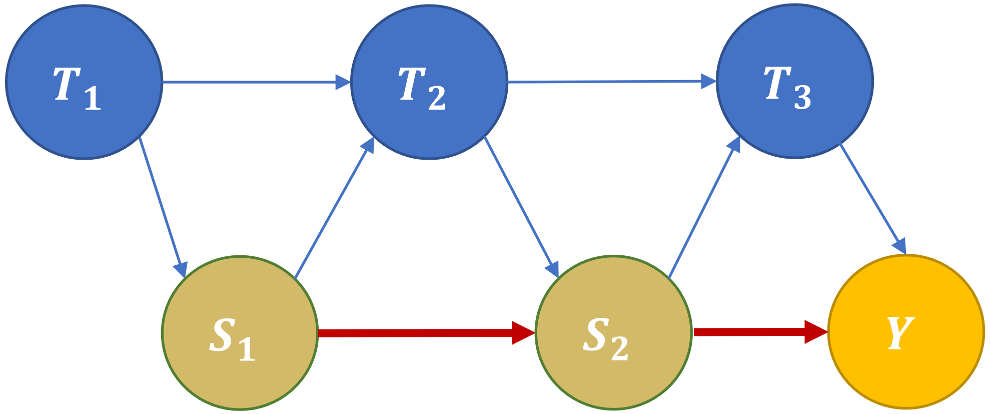
recent/short-term (E)



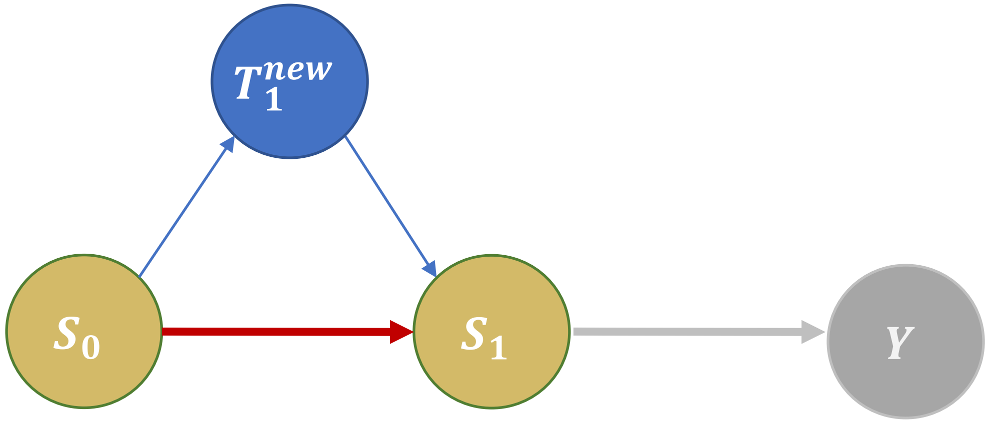
- Multiple treatments are offered after the surrogate variable
- Easy: estimate their effect controlling for S_1
- Wrong!

Beyond Two Periods

historical/long-term (O)

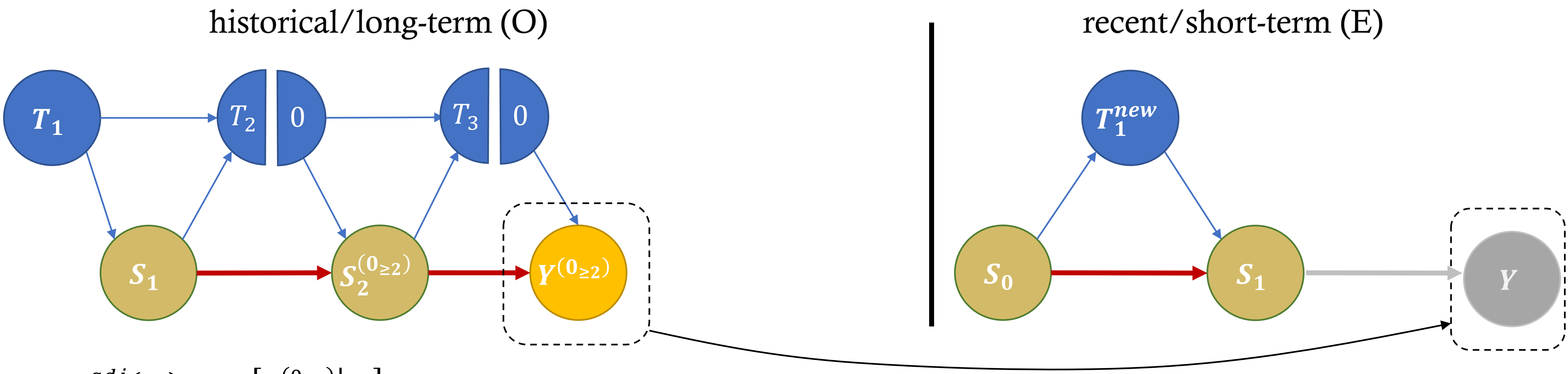


recent/short-term (E)



- Treatments are offered in an adaptive manner, in response to previous period surrogate/state
- The surrogate – treatment feedback precludes viewing this as a one-shot treatment problem
- Setting is known as the dynamic treatment regime [Robins'94,'04, Chakraborty-Murphy'14]

Beyond Two Periods: Target Quantity



$$g^{adj}(S_1) = E_o\left[Y^{(0_{\geq 2})} \middle| S_1\right]$$

$$= E_o\left[Y - \sum_{t\geq 2} \underbrace{\gamma_t(T_t, S_{t-1})}_{} \middle| S_1\right]$$

"blip" effect of treatment at period t

- What effect do we subtract from Y?
- Do we estimate effect of T_2, T_3 controlling for S_1, S_2? Wrong
- Do we estimate effect of T_2, T_3 controlling for S_1? Wrong