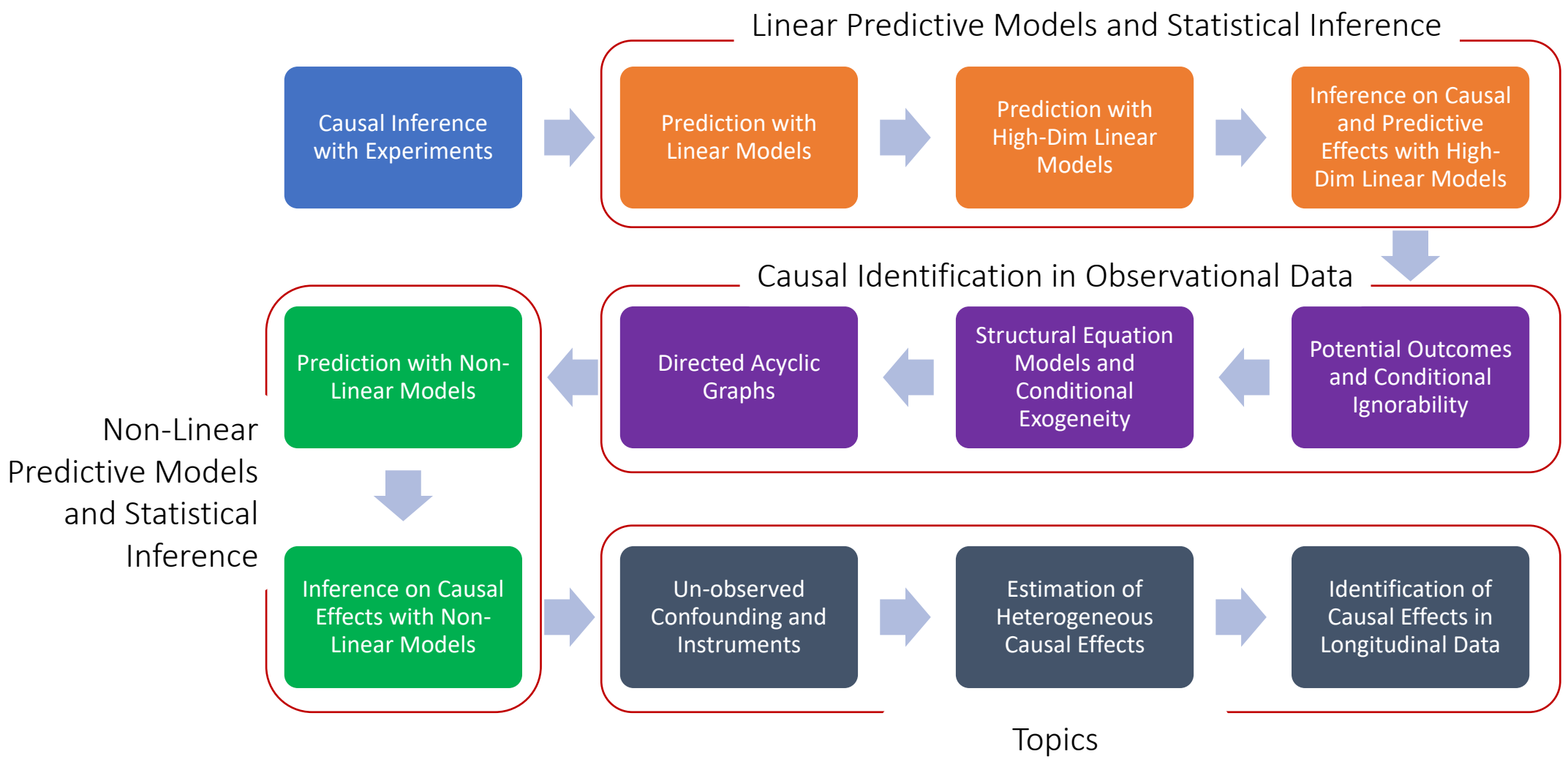
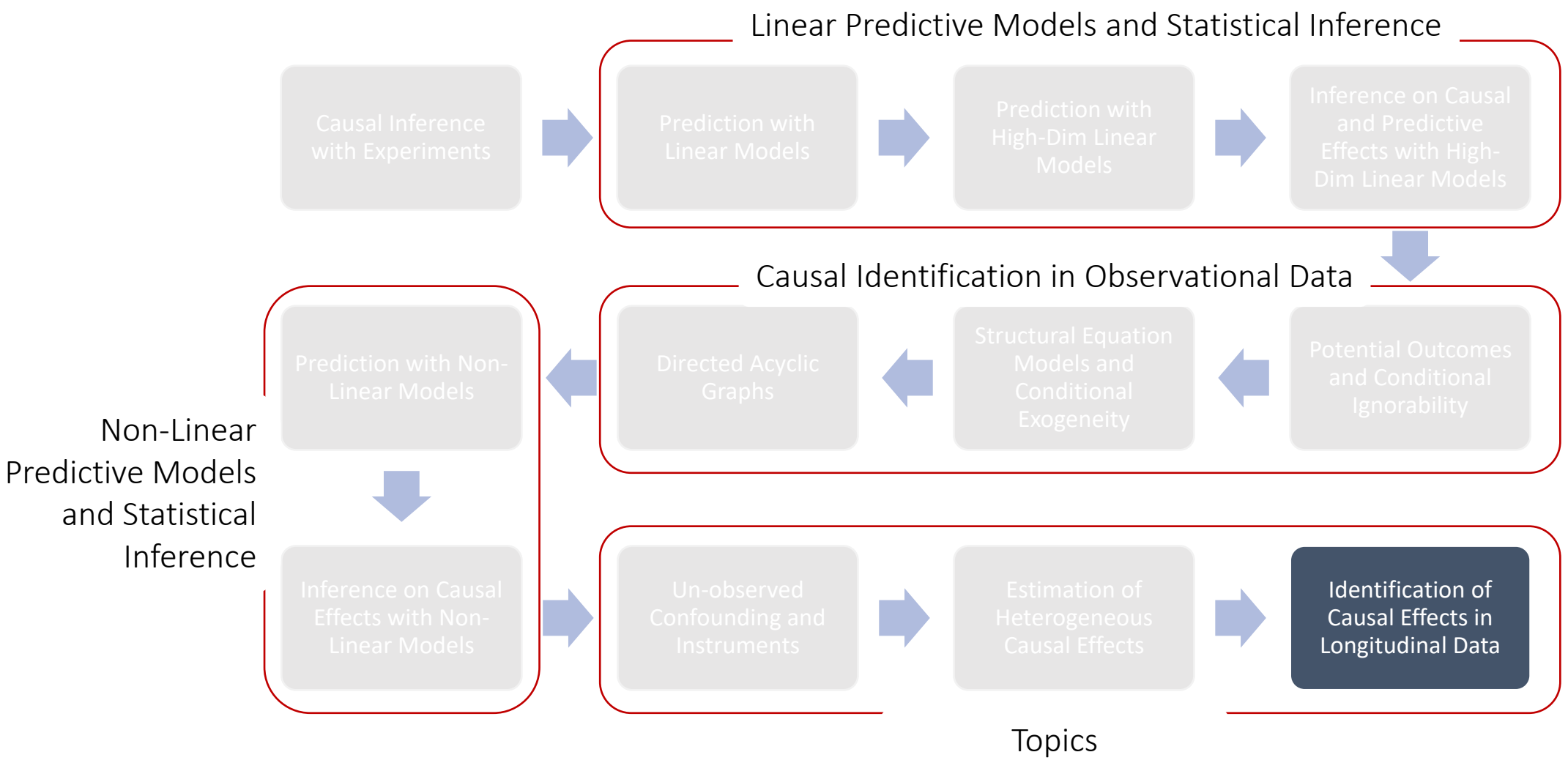
MS&E 228: Topics on Longitudinal Data and Causal Machine Learning

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 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

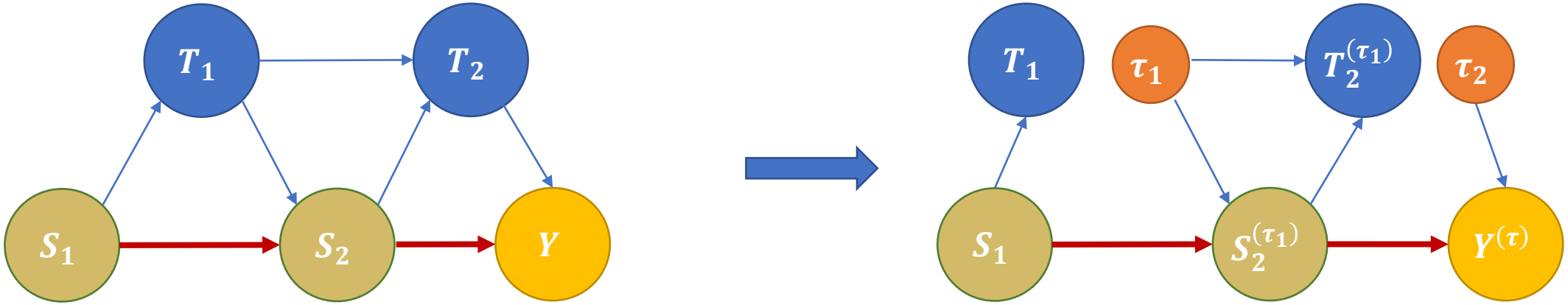
Dynamic Treatment Regime

Dynamic Treatment Regime

observed data (panel)

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

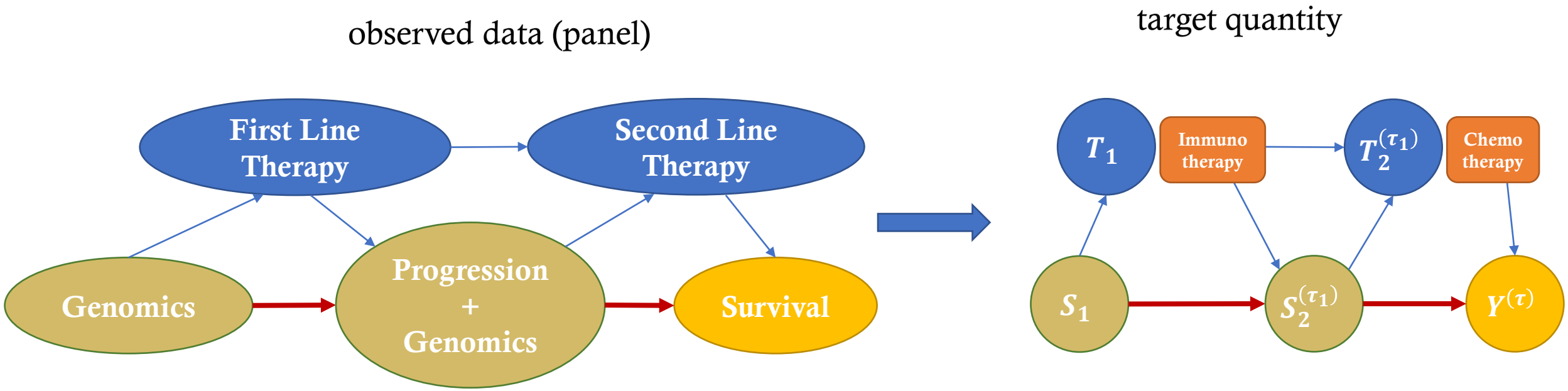
$$\theta := E\left[Y^{(\tau)}\right]$$



- 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]

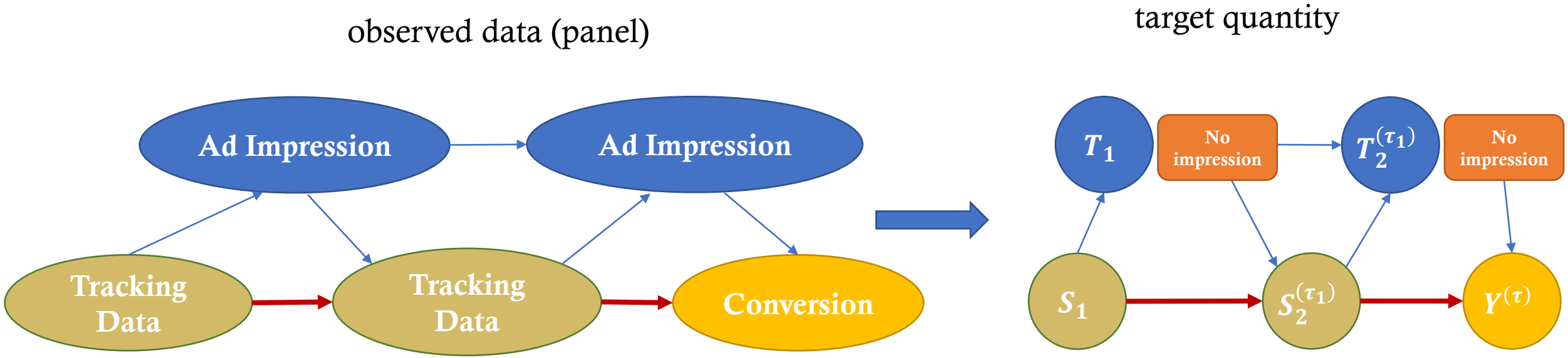
Examples

- Healthcare: Patients treated over time and adaptively



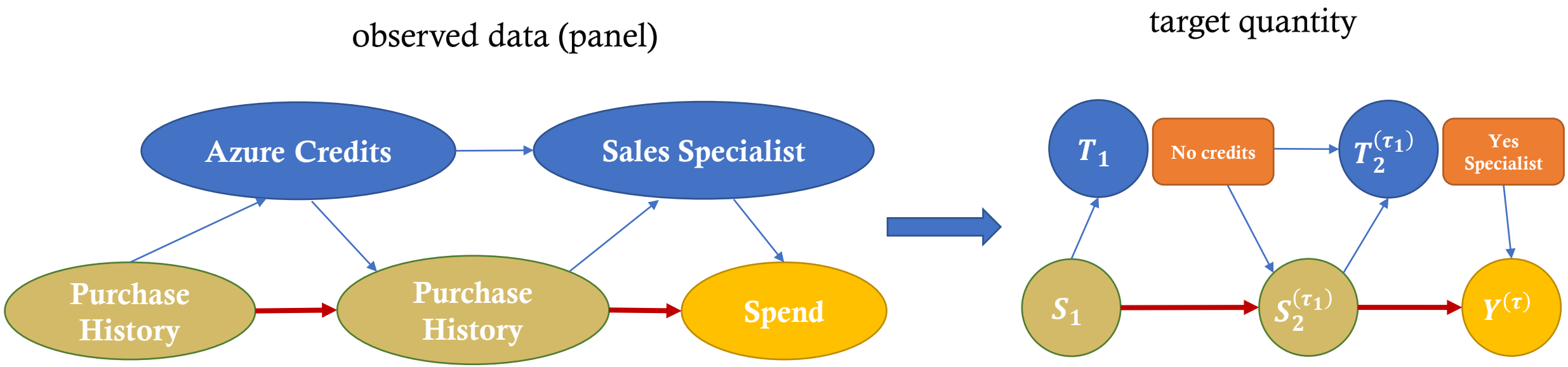
Examples

- Digital Marketing: Web users shown ads multiple times



Examples

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



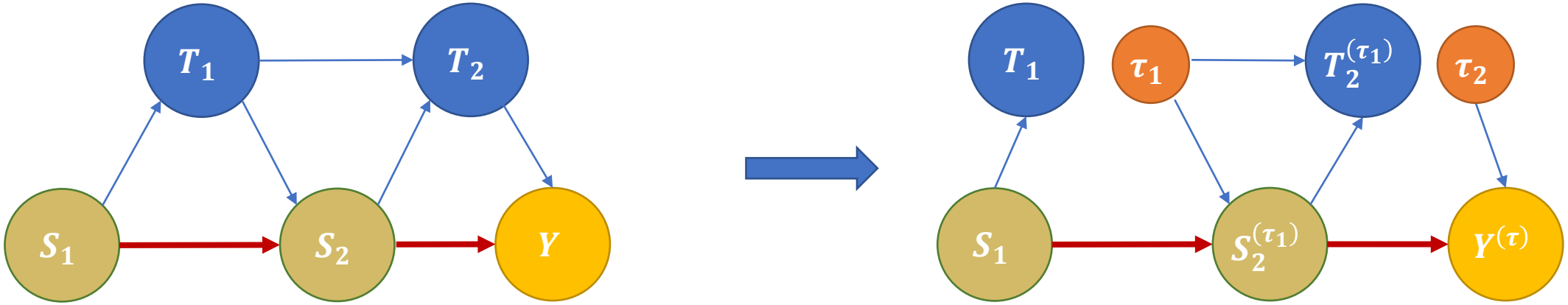
Identification in the Dynamic Treatment Regime

Dynamic Treatment Regime

observed data (panel)

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

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



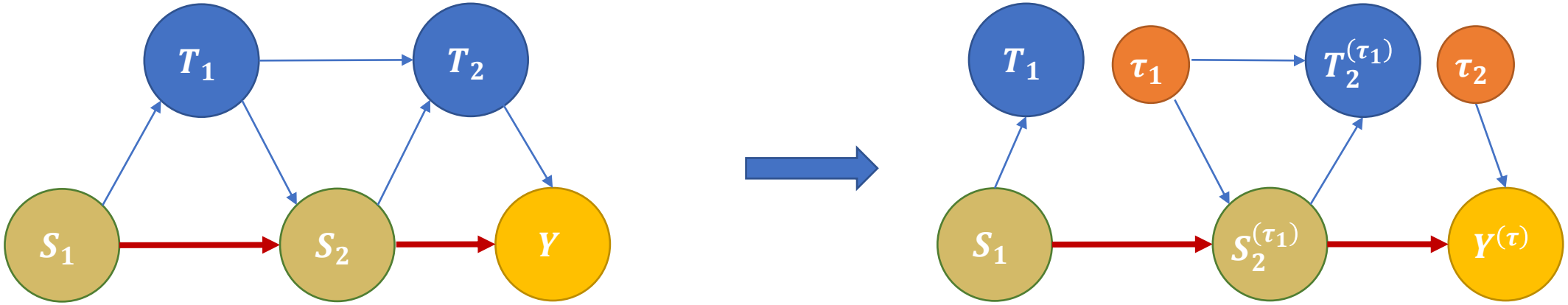
- Treatments are offered in an adaptive manner, in response to previous period controls
- The surrogate – control feedback precludes viewing this as a one-shot treatment problem
- Setting is known as the dynamic treatment regime [Robins'94,'04, Chakraborty-Murphy'14]

Identification in Dynamic Treatment Regime

observed data (panel)

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

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



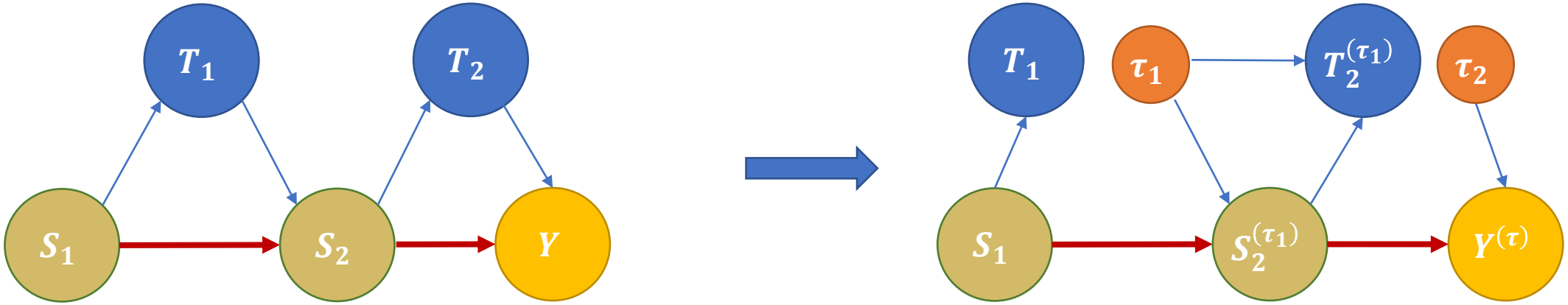
- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!

Identification in Dynamic Treatment Regime

observed data (panel)

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

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



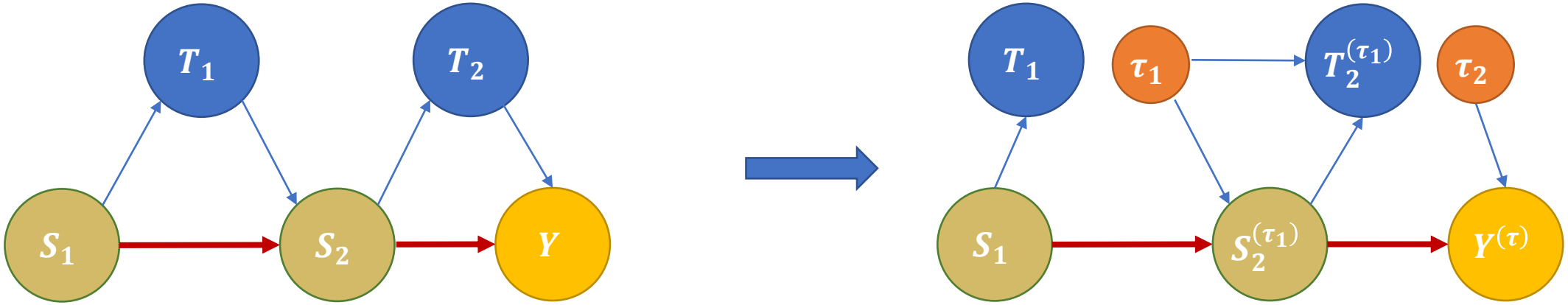
- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!

Identification in Dynamic Treatment Regime

observed data (panel)

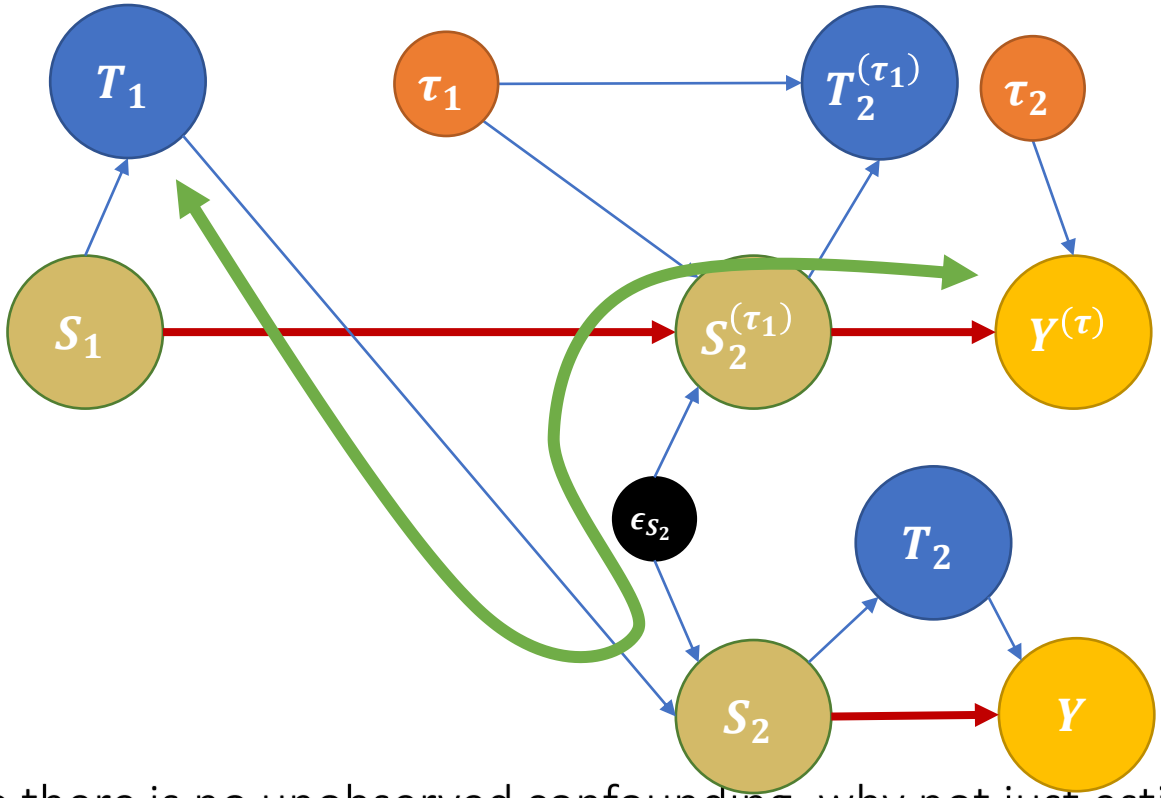
target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

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



- 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 $Y^{(\tau)} \perp\!\!\!\not\perp T_1 \mid S_1, S_2$

Identification in Dynamic Treatment Regime



Intervention World

Two-World Graphs

Observed World

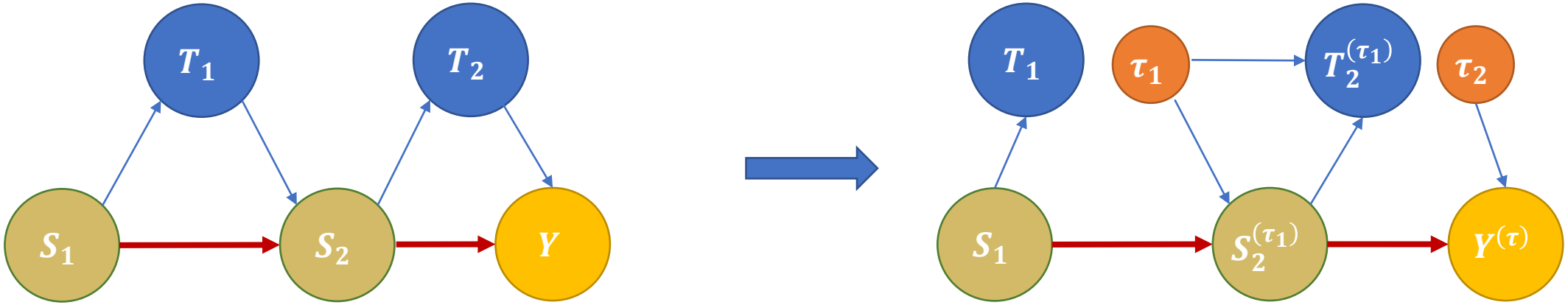
- Since there is no unobserved confounding, why not just estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on $S = (S_1, S_2)$
- Wrong: conditioning on S_2 blocks all the effect from T_1!
- Can also be understood using D-separation that $Y^{(\tau)} \perp\!\!\!\not\perp T_1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

Identification in Dynamic Treatment Regime

observed data (panel)

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

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



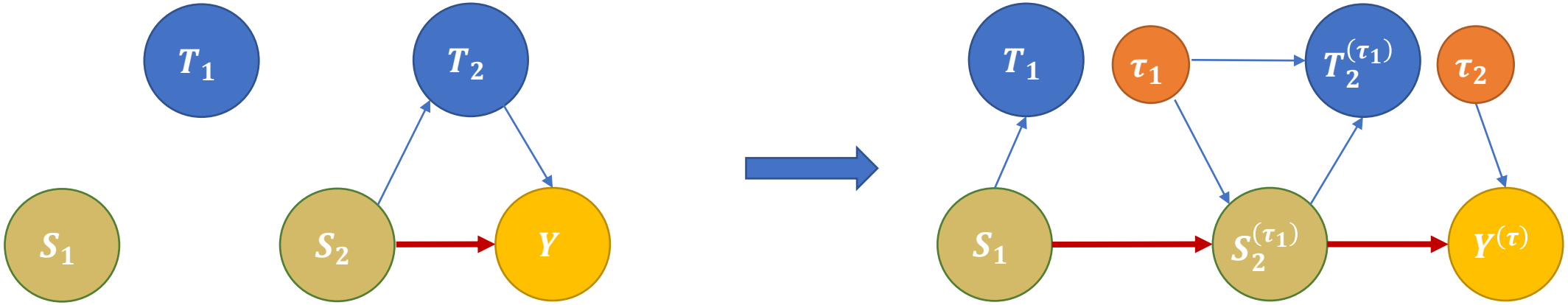
- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!

Identification in Dynamic Treatment Regime

observed data (panel)

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

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



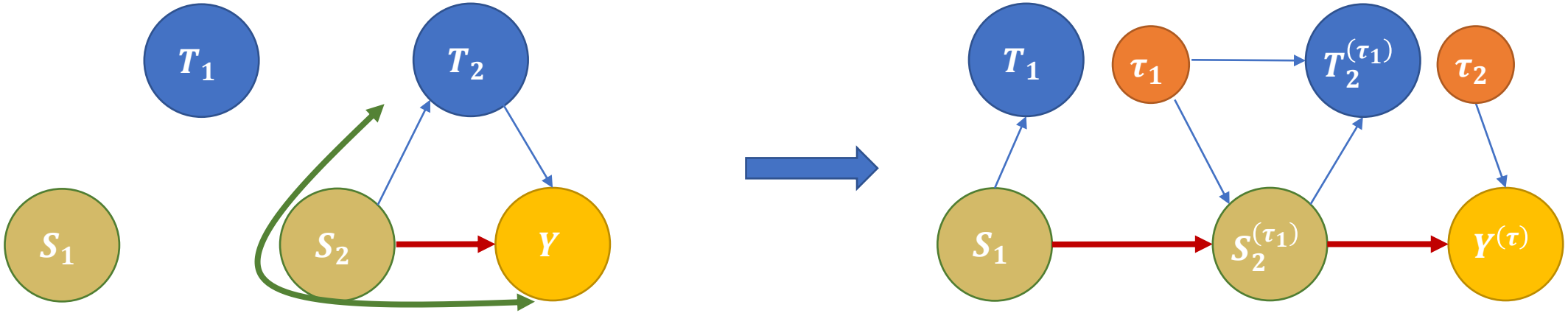
- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!

Identification in Dynamic Treatment Regime

observed data (panel)

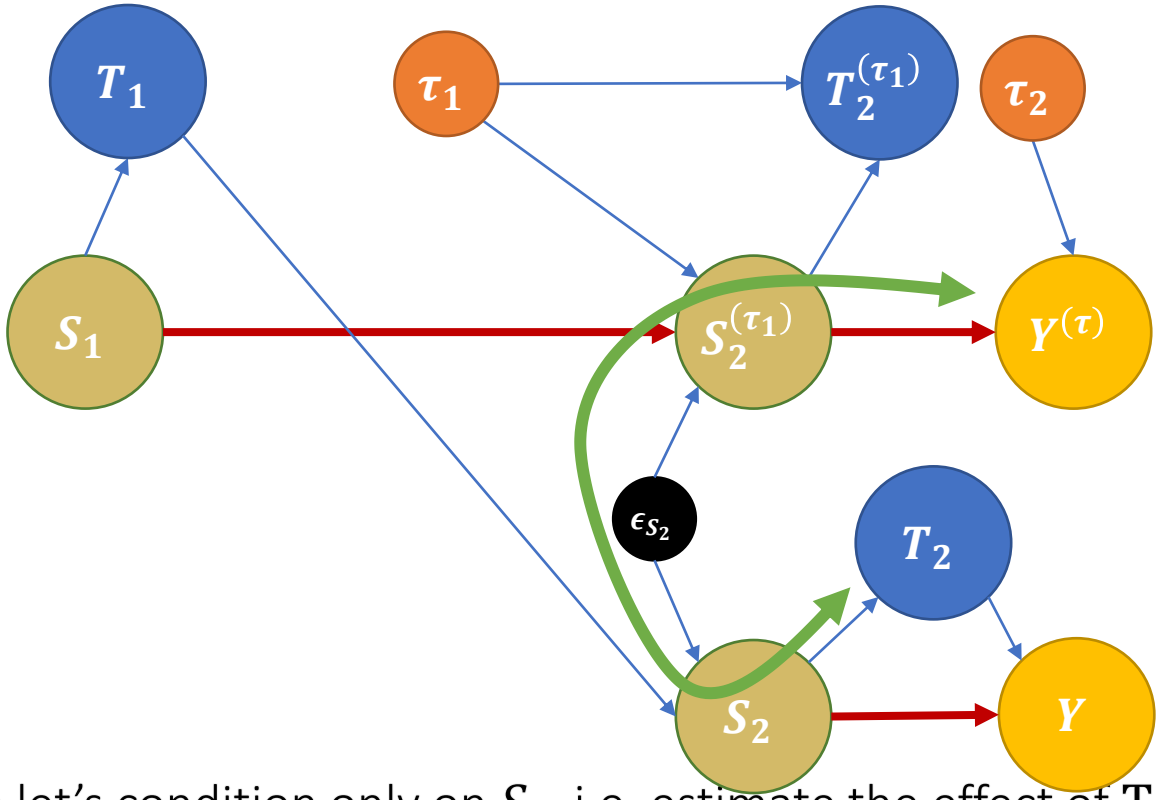
target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

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



- 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



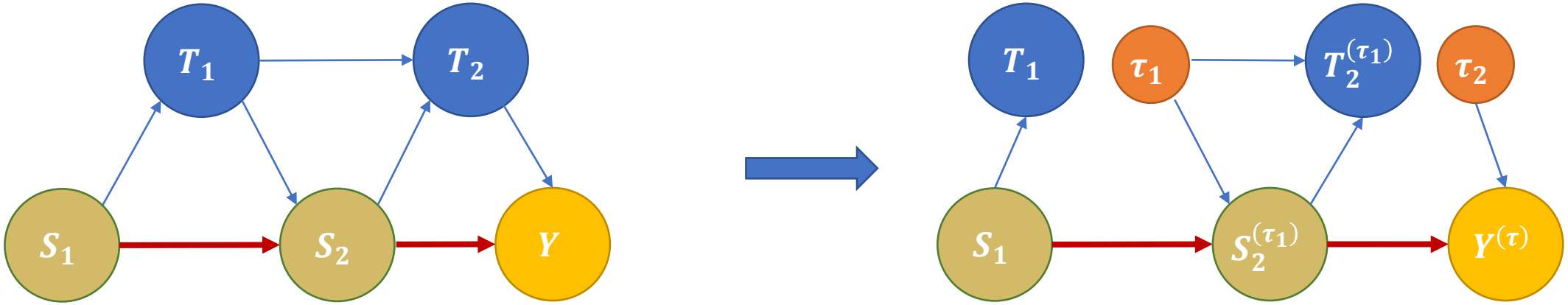
- Then let's condition only on S_1, i.e. estimate the effect of $\mathrm{T} = (T_1, T_2)$ by conditioning, conditioning on S_1
- Wrong: conditioning only on S_1 leaves "unobserved confounding" between T_2 and Y!
- Can also be understood using D-separation that $Y^{(\tau)} \not\perp\!\!\!\perp T_2 \mid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

Dynamic Treatment Regime

observed data (panel)

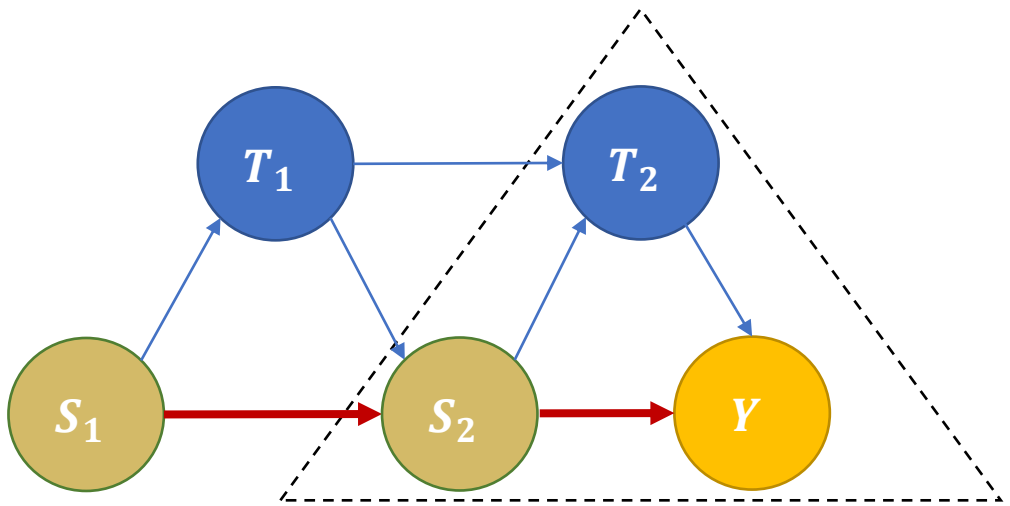
target quantity: average outcome under a static treatment sequence (regime) $\tau = (\tau_1, \tau_2)$

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

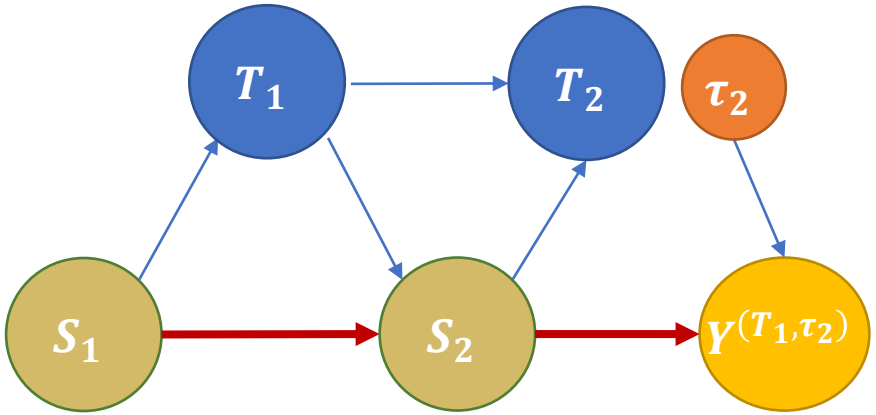


- We cannot identify this expected potential outcome simply by conditioning!
- Let's take it one step at a time
- What can we identify by conditioning?

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

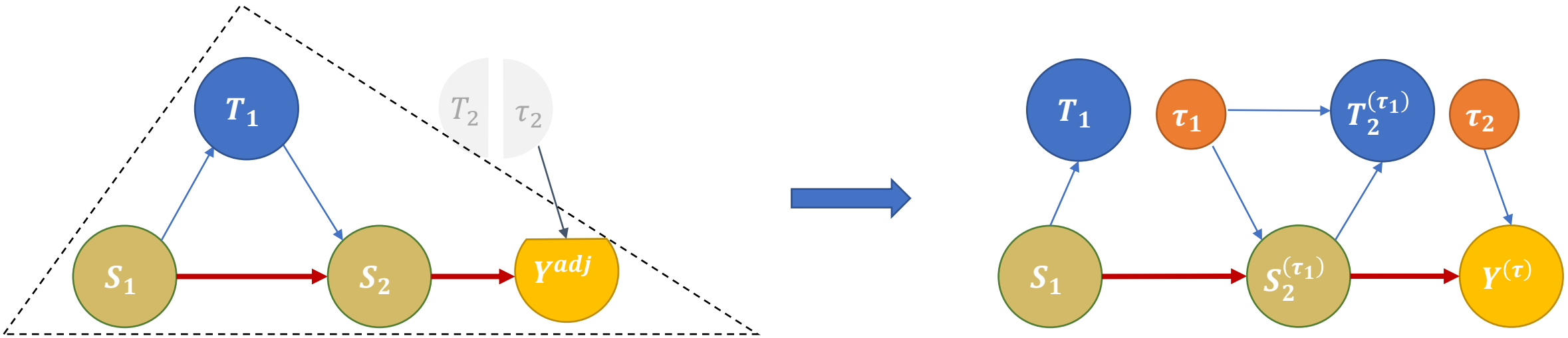
Target quantity $\theta_2(S_2)$ can be identified by conditioning

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

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

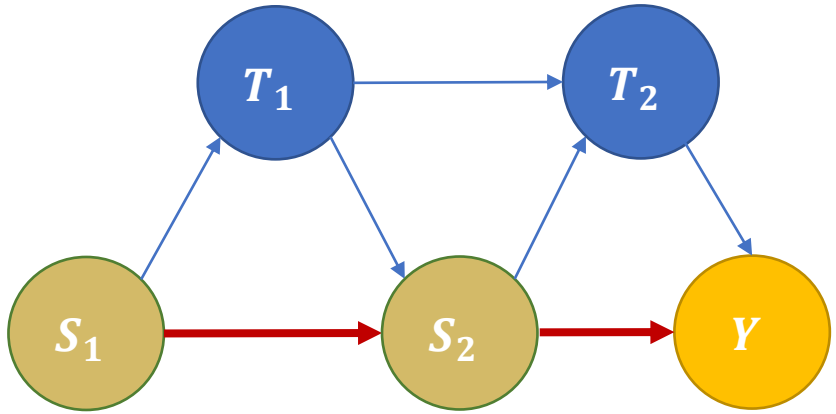


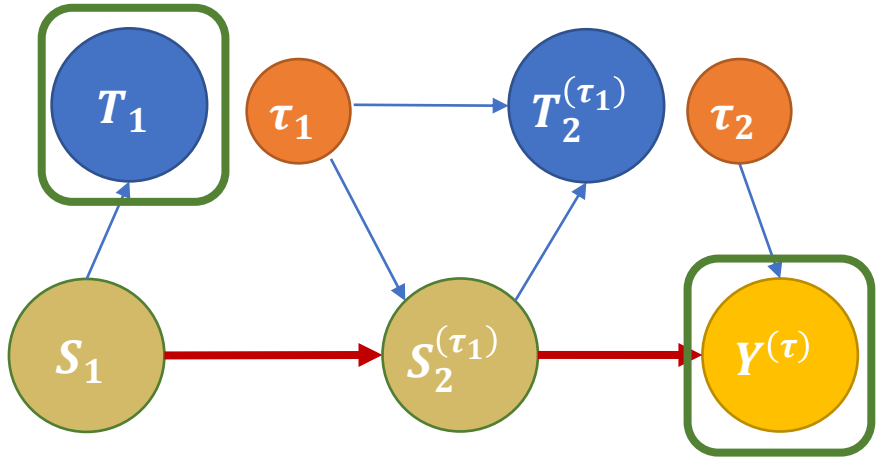
- But now we can "adjust the observed outcome" using this counterfactual model to remove the second period treatment!

- 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

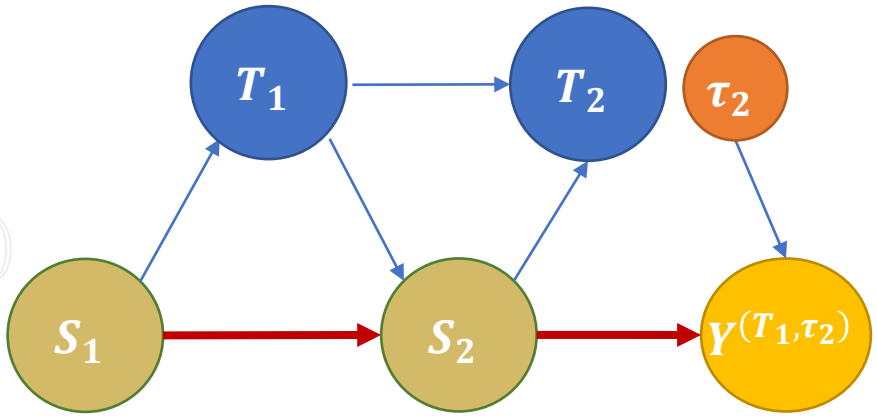
Identification via Backwards Recursion



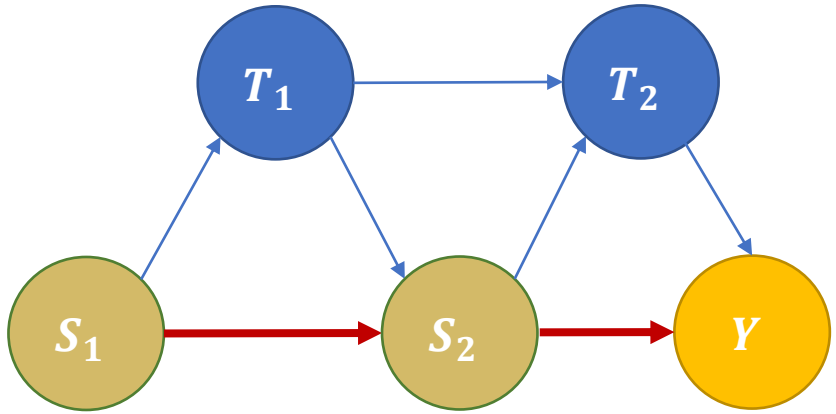


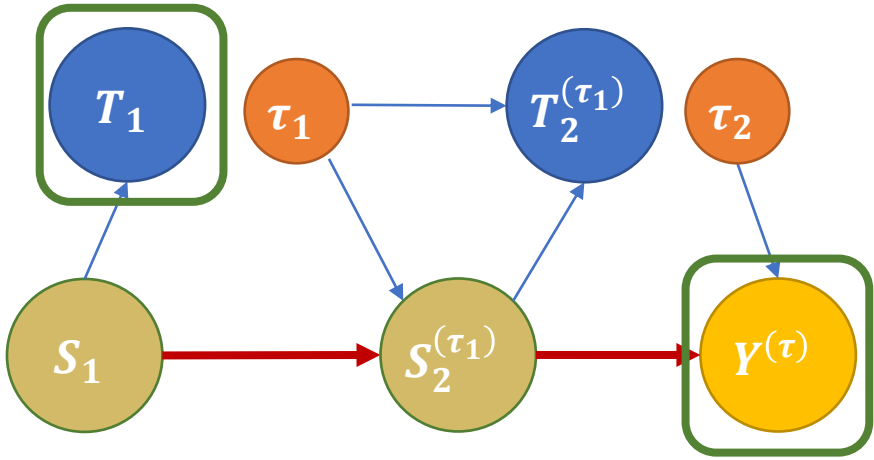
$$E\left[Y^{(\tau_1,\tau_2)}\right] = E\left[E\left[Y^{(\tau_1,\tau_2)}\middle|S_1\right]\right]$$
$$= E\left[E\left[Y^{(T_1,\tau_2)}\middle|T_1=\tau_1,S_1\right]\right]$$
$$= E\left[E\left[E\left[Y^{(T_1,\tau_2)}\middle|S_2,T_1,S_1\right]\middle|T_1=\tau_1,S_1\right]\right] \quad \text{(tower law)}$$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

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



Identification via Backwards Recursion

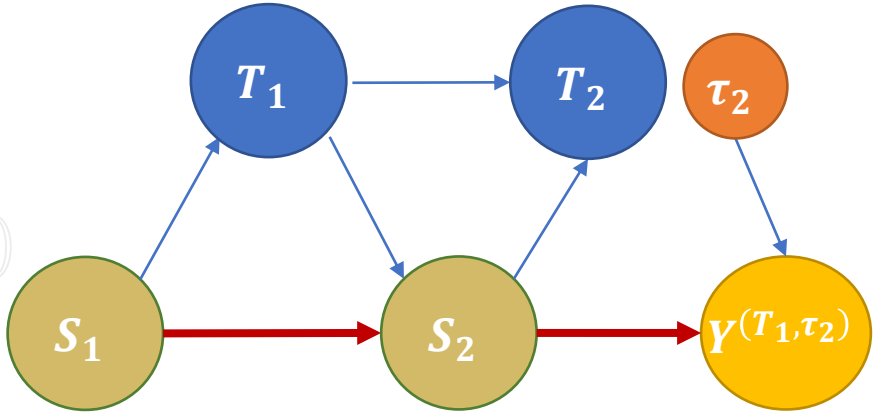




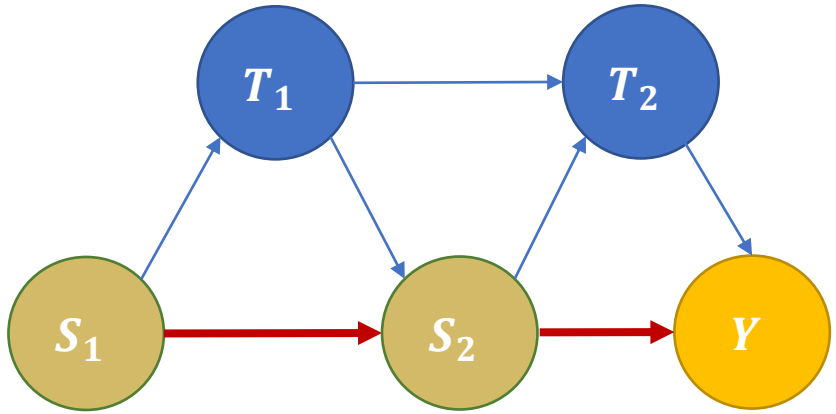
$$E\left[Y^{(\tau_1,\tau_2)}\right] = E\left[E\left[Y^{(\tau_1,\tau_2)}\middle|S_1\right]\right]$$
$$= E\left[E\left[Y^{(T_1,\tau_2)}\middle|T_1=\tau_1,S_1\right]\right]$$
$$= E\left[E\left[E\left[Y^{(T_1,\tau_2)}\middle|S_2,T_1,S_1\right]\middle|T_1=\tau_1,S_1\right]\right]$$
$$= E\left[E\left[E\left[Y^{(T_1,\tau_2)}\middle|S_2\right]\middle|T_1=\tau_1,S_1\right]\right] \qquad \left(Y^{(T_1,\tau_2)} \perp\!\!\!\perp S_1,T_1\middle|S_2\right)$$
$$= E\left[E\left[\theta_2(S_2)\middle|T_1=\tau_1,S_1\right]\right]$$

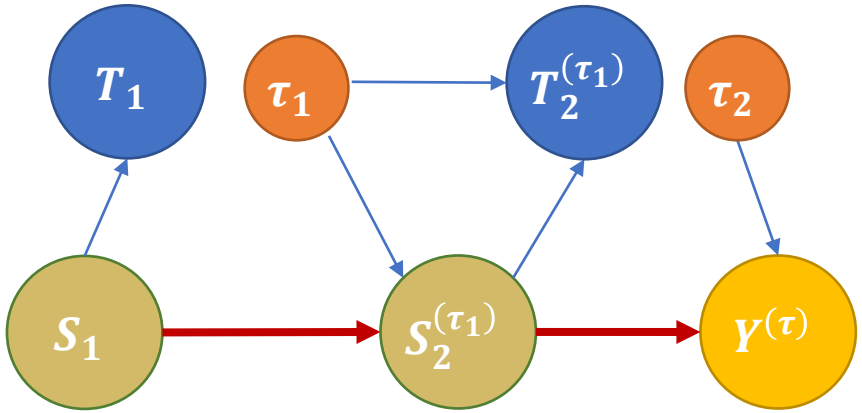
$\left(Y^{(\tau)} \perp\!\!\!\perp T_1\middle|S_1\right)$

(tower law)

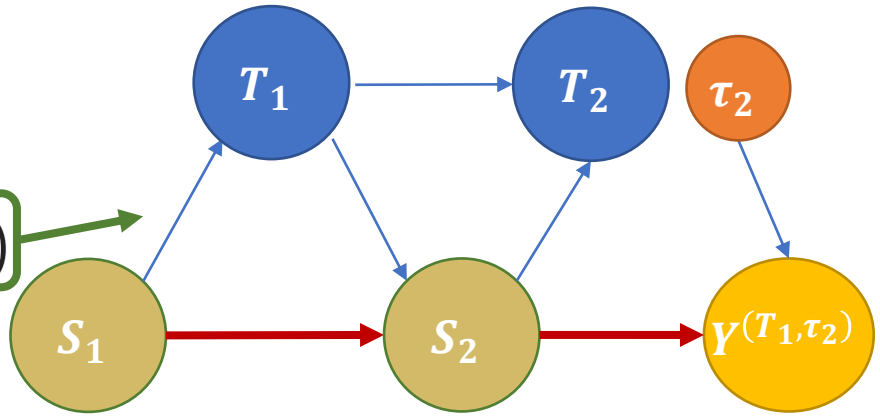


Identification via Backwards Recursion





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



Identification Process

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

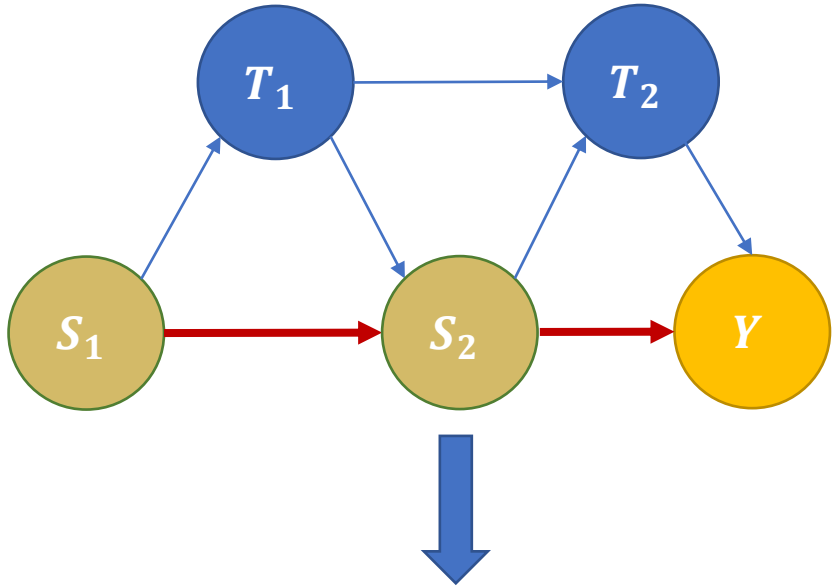
$$Y_{adj} \leftarrow f_2(\tau_2, S_2)$$

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

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

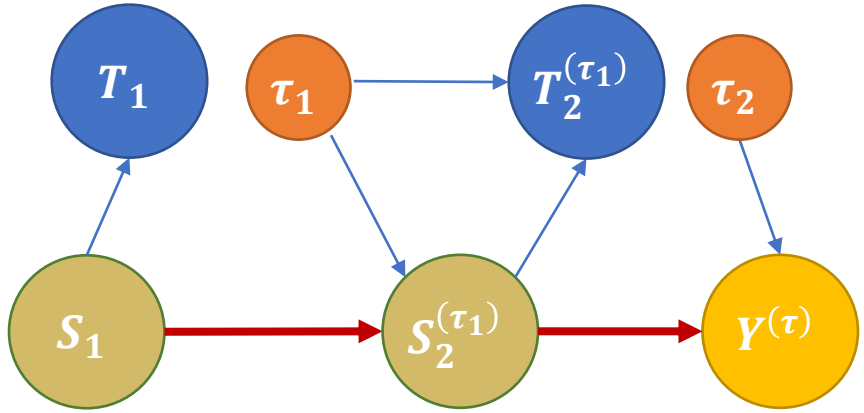
- Return target quantity: $\theta = E[Y_{adj}]$

observed data (panel)



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

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

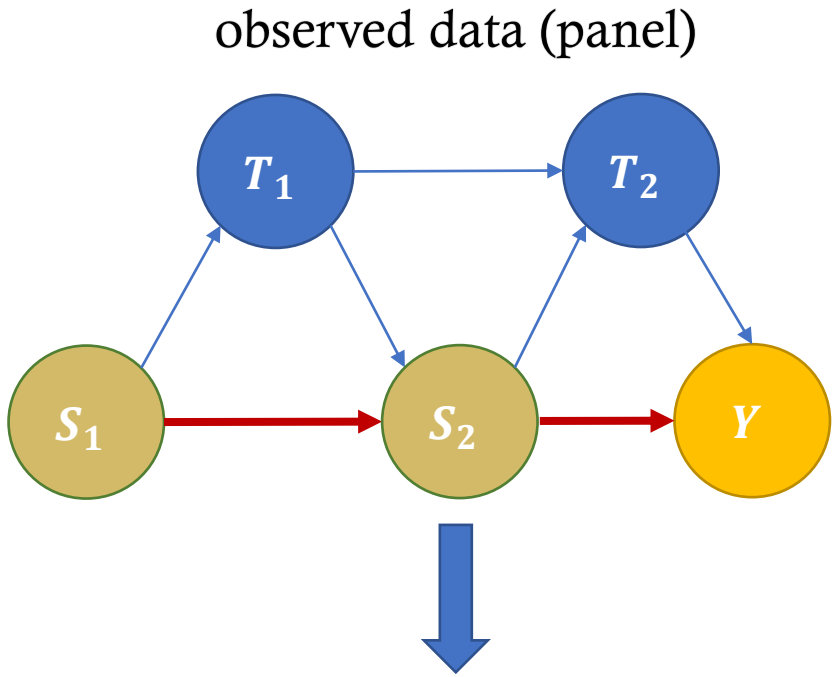


Moment Based Framework

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

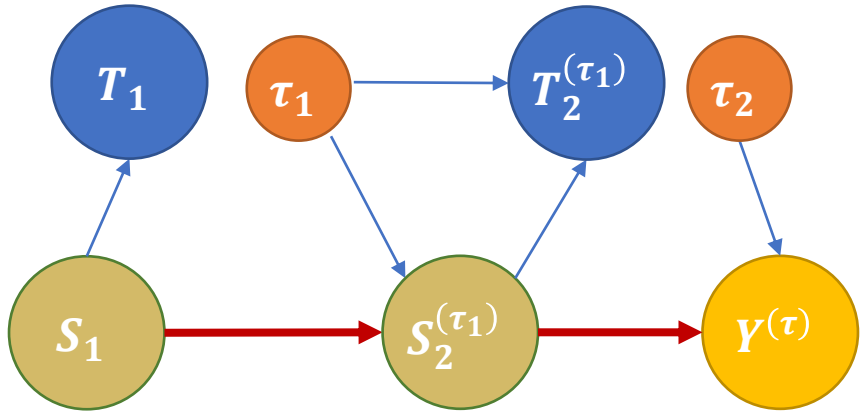
$$f_1(T_1, S_1) := E[f_2(\tau_2, S_2)|T_1, S_1]$$
$$f_2(T_2, S_2) := E[Y|T_2, S_2]$$

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



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

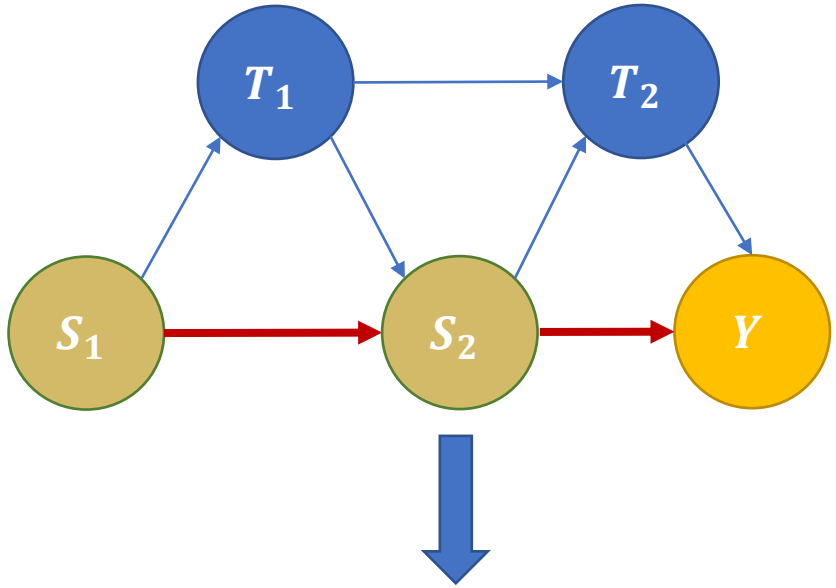
$$\theta := E[Y^{(\tau)}]$$



Estimation in the Dynamic Treatment Regime

G-computation

observed data (panel)



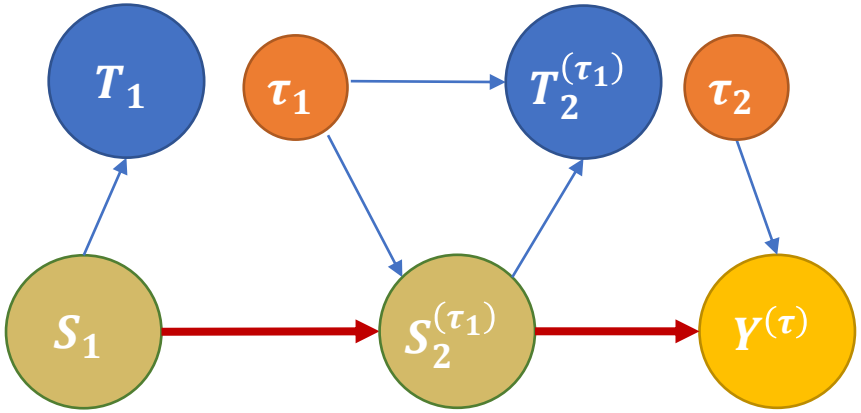
Moment function: $m(Z;\theta,f)$

$$E[\theta - f_1(\tau_1, S_1)] = 0$$

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

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

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



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”

ML Based Estimation

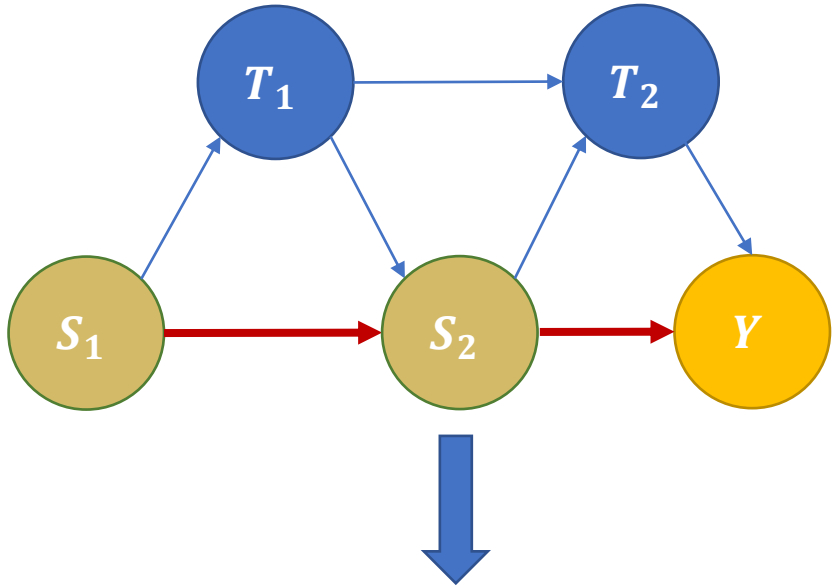
Moment function: $m(Z;\theta,f)$

$$E[\theta - f_1(\tau_1, S_1)] = 0$$

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

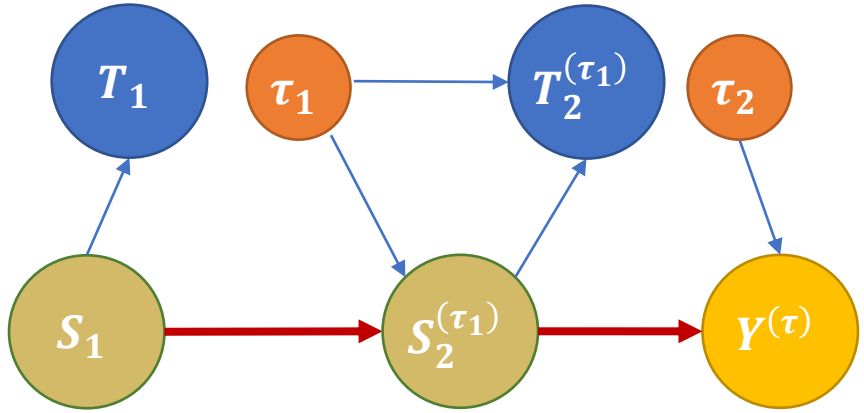
- If we use ML for these predictive problems, then we wont be able to construct confidence intervals for θ because moment is not Neyman orthogonal
- We can apply the Inverse Propensity based debiasing idea

observed data (panel)



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

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



Debiasing Moments and Neyman Orthogonality

Original moment

Inverse Propensity

Residual of the Regression Problem that defines nuisance function f_1

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

Debiasing correction for nuisance function f_1

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

Nuisance functions

Debiasing Moments and Neyman Orthogonality

Original moment

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

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

Debiasing correction for nuisance function f_1

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

Nuisance functions

Debiasing Moments and Neyman Orthogonality

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

Original moment

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

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

Debiasing correction for nuisance function f_2

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

Nuisance functions

Debiasing Moments and Neyman Orthogonality

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

Original moment

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

IPS term already multiplying f_2

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

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

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

Nuisance functions

Debiasing Moments and Neyman Orthogonality

Original moment

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

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

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

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

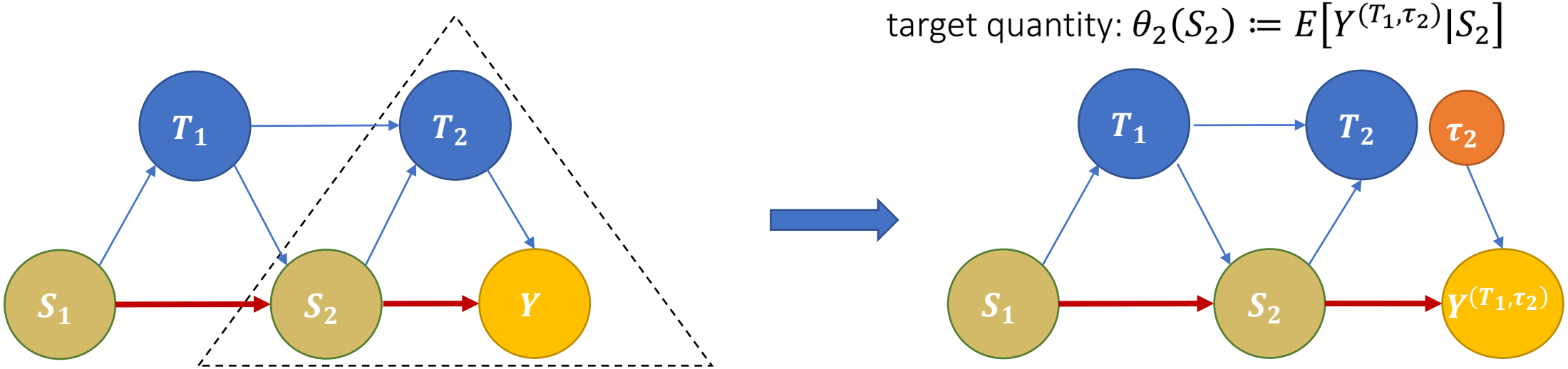
Continuous Treatments and Alternative Approach to Identification and Estimation

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



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

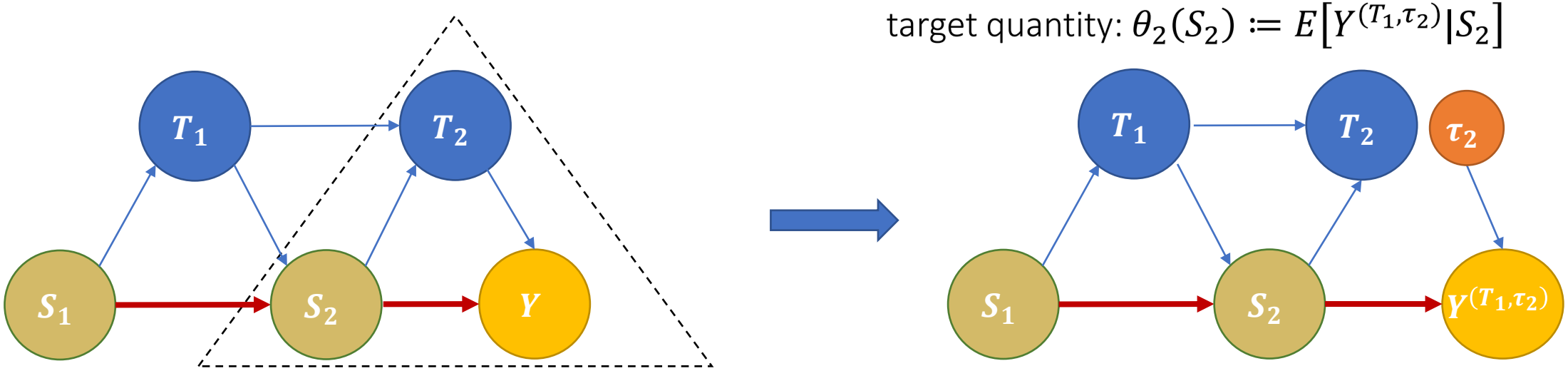
$$\alpha_2(\tau_2, S_2) \coloneqq E\left[Y^{(T_1,\tau_2)} - Y^{(T_1,0)}\middle|S_2\right]$$

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



- 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

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

- 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" $\alpha_2(\tau_2, S_2)$

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

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 $\alpha_1(\tau_1, S_1)$

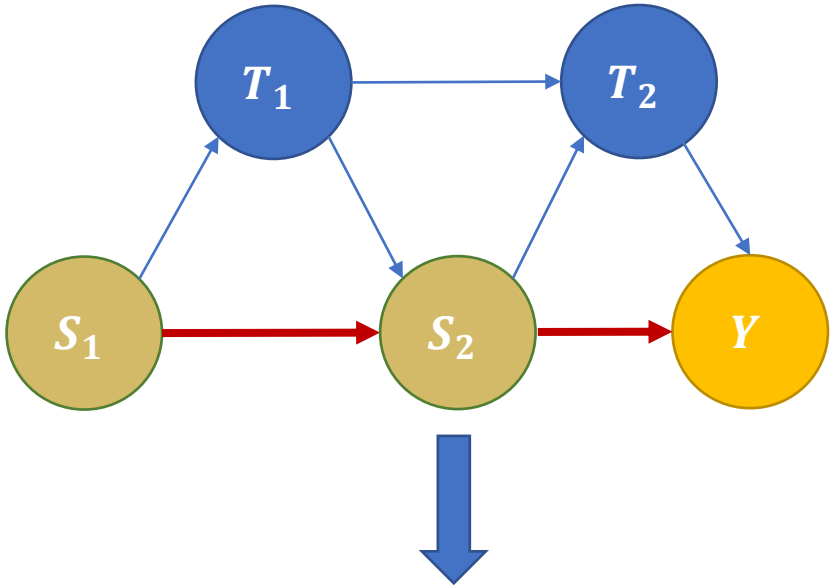
$$E\left[Y^{(\tau_1,\tau_2)} - Y^{(0,\tau_2)} \middle| S_1\right] \approx E\left[Y_{\text{adj}}^{(\tau_1)} - Y_{\text{adj}}^{(0)} \,\middle|\, S_1\right]$$

Adjust outcome for first period treatment

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

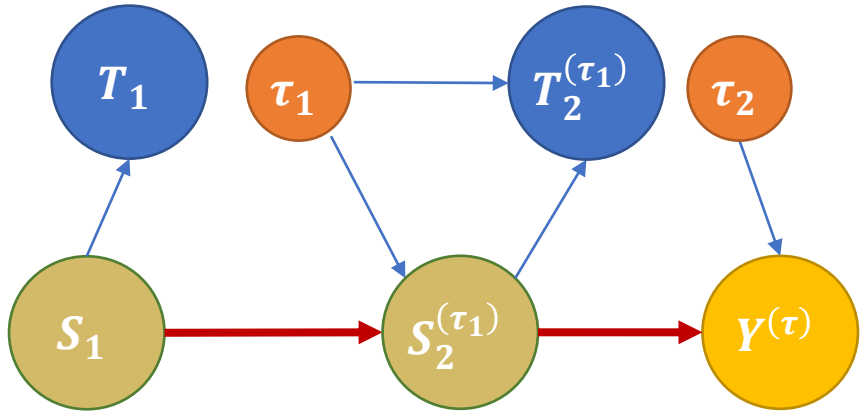
Return target quantity: $\theta = E[Y_{\text{adj}}]$

observed data (panel)



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

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

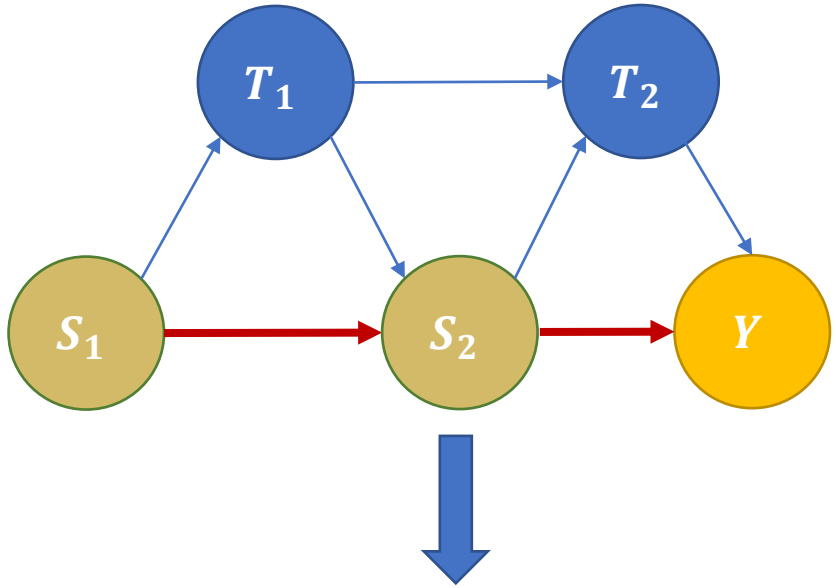


Partial Linearity: Linear Effects

- If we assume that these effects are partially linear:

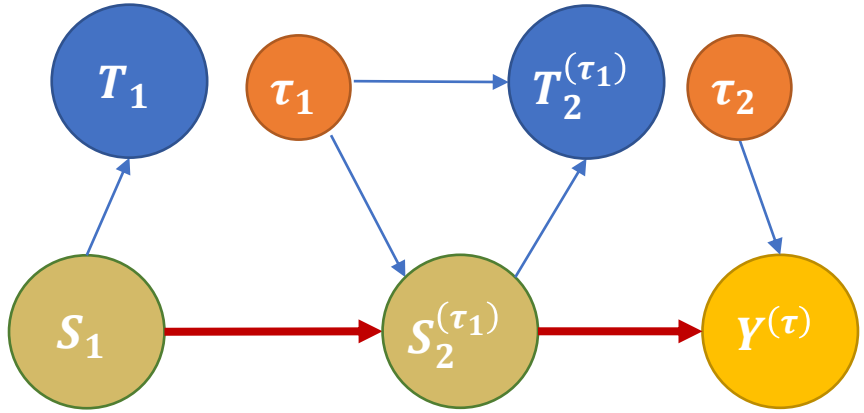
$$a_2(\tau_2, S_2) = \delta_2 \tau_2$$
$$a_1(\tau_1, S_1) := \delta_1 \tau_1$$

observed data (panel)



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

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

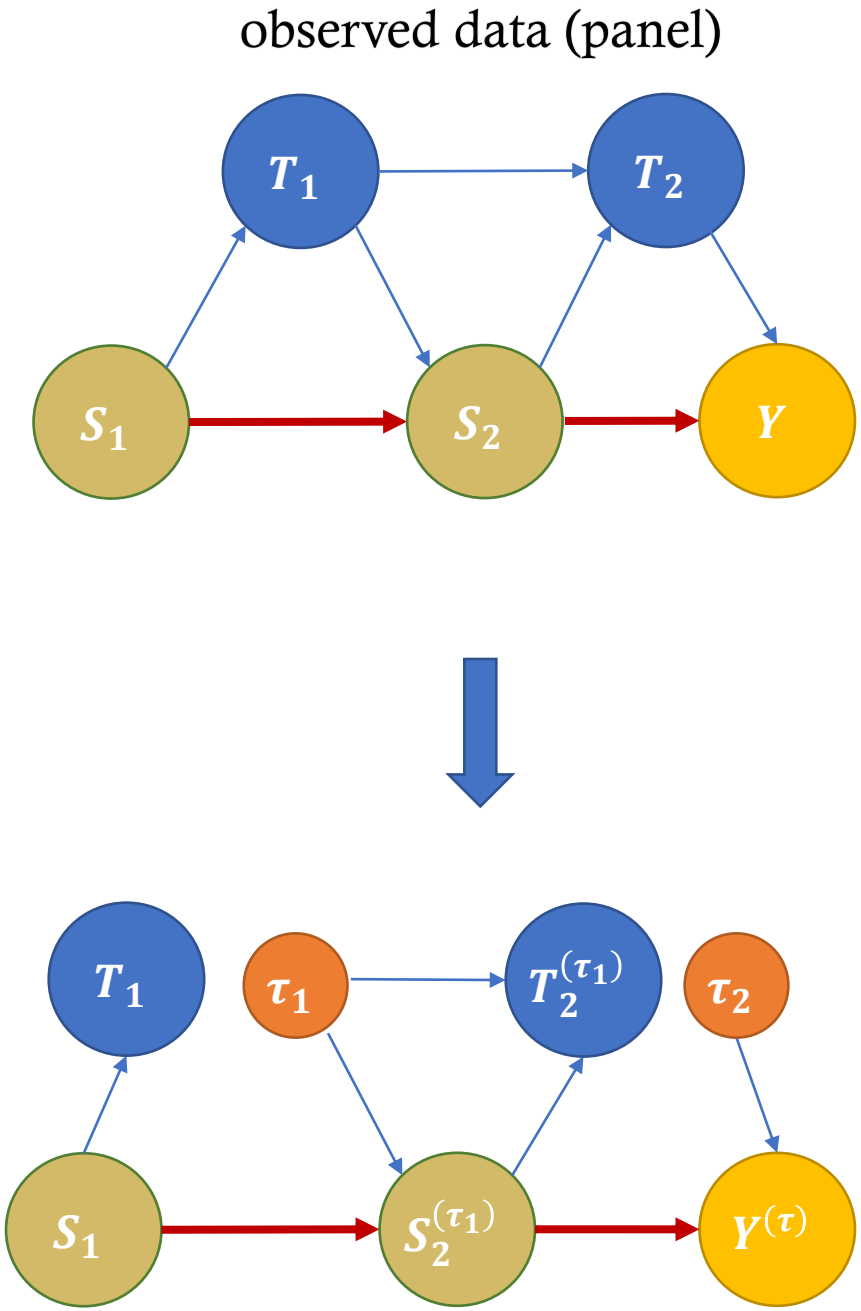


Identification Process

- 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)

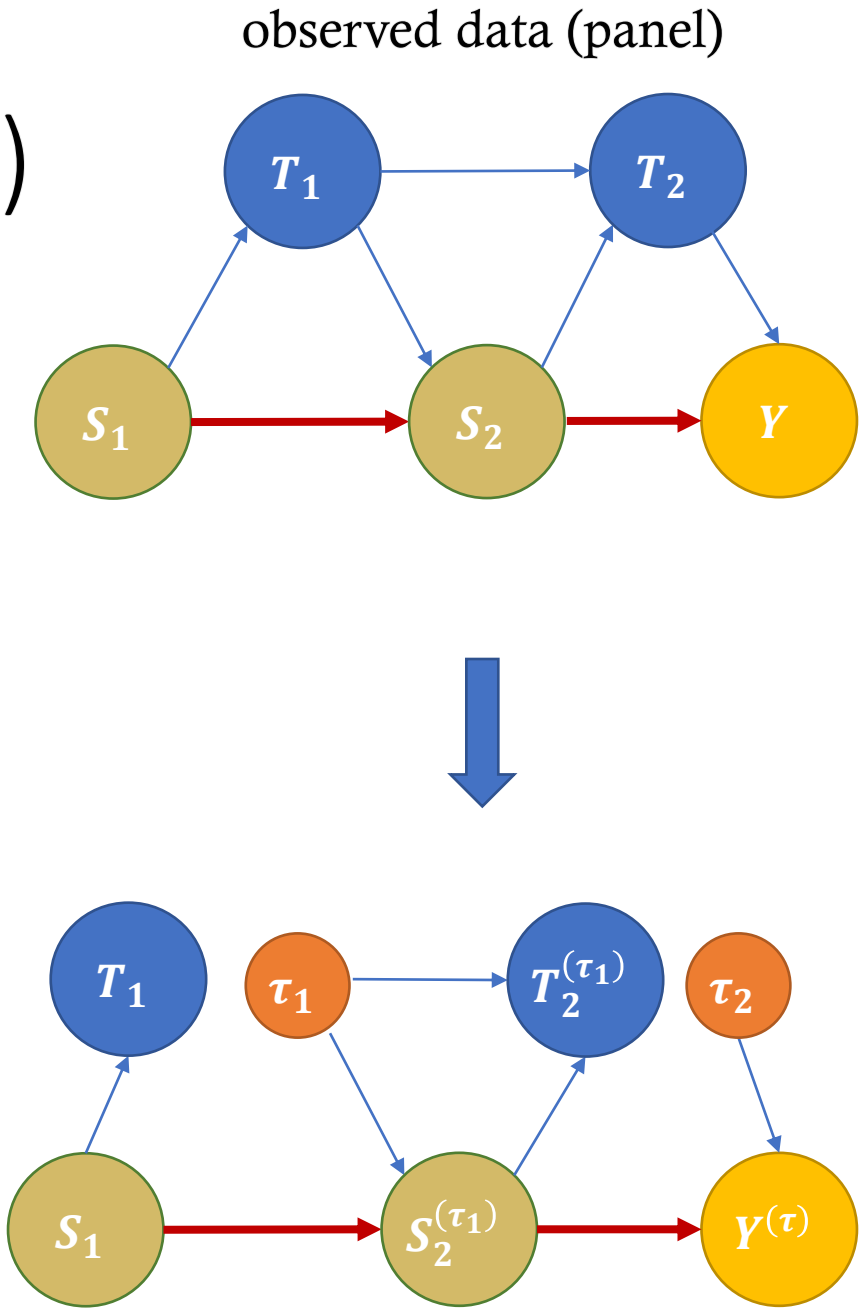
- Construct residuals
  $\tilde{Y}_2 := Y - E[Y|S_2], \qquad \tilde{T}_2 = T_2 - E[T_2 \mid S_2]$
- Run OLS: $\tilde{Y}_2 \sim \tilde{T}_2$ to estimate δ_2

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

- Construct residuals:
  $\tilde{Y}_1 := Y_{adj} - E[Y_{adj}|S_1], \qquad \tilde{T}_1 := T_1 - E[T_1|S_1]$
- 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: $\theta := E[Y_{adj}]$



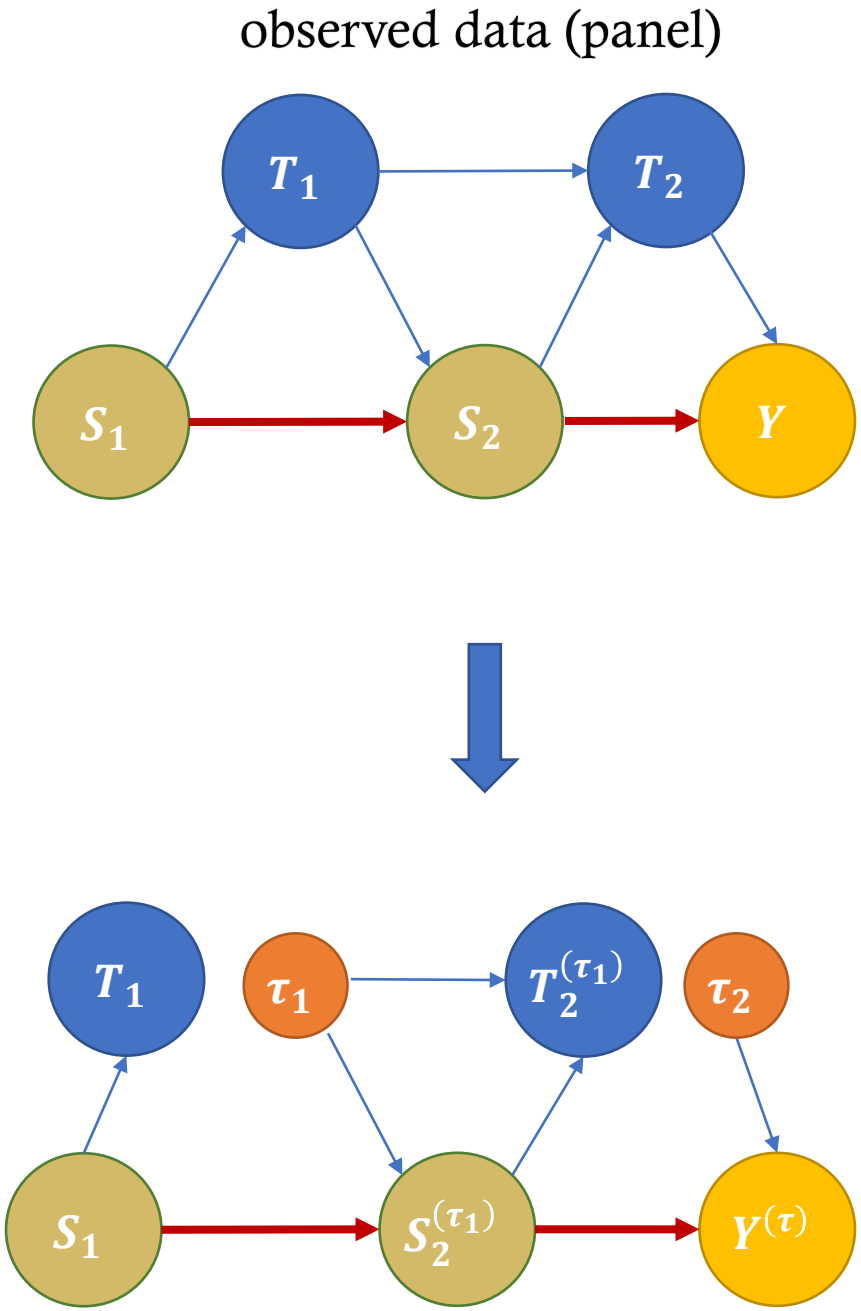
Moment Based Framework

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

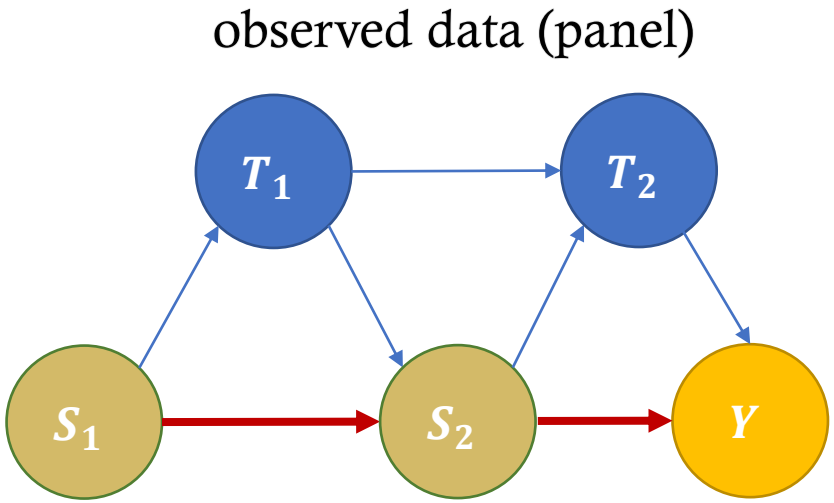
$$E\left[\left(\tilde{Y}_2 - \delta_2 \tilde{T}_2\right)\tilde{T}_2\right] = 0$$
$$E\left[\left(\tilde{Y}_1 - \delta_1 \tilde{T}_1\right)\tilde{T}_1\right] = 0$$

$E[Y|S_2]$, $\mathrm{E}[\mathrm{T}_2|\mathrm{S}_2]$
$E[Y - \delta_2 T_2 + \delta_2 \tau_2|S_1]$, $\mathrm{E}[\mathrm{T}_1|\mathrm{S}_1]$

Nuisance functions



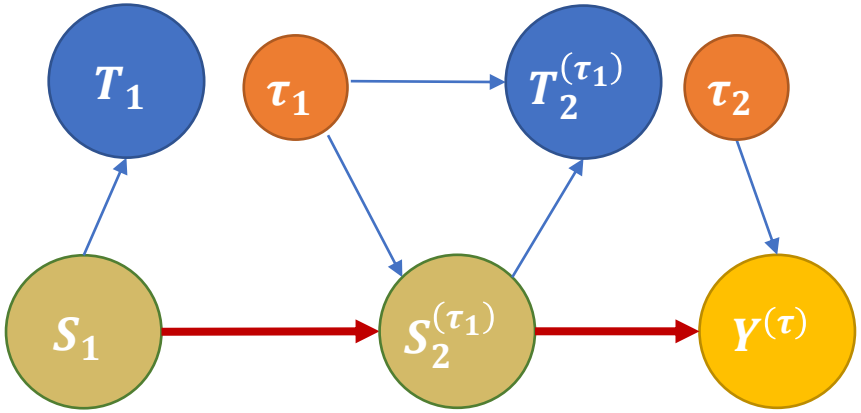
Moment Based Framework



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

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

$$E[Y|S_2], \quad \mathrm{E}[\mathrm{T}_2|\mathrm{S}_2]$$
$$E[Y - \delta_2 T_2 + \delta_2 \tau_2|S_1], \quad \mathrm{E}[\mathrm{T}_1|\mathrm{S}_1]$$



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

Long-Term Causal Inference from Short-Term Data with Surrogates

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

$$E\left[Y^{(\tau)}\right] = E\left[Y^{(\tau)} \middle| T = \tau\right] = E[Y|T = \tau] = E[E[Y|T = \tau, S]|T = \tau] = E[E[Y|S]|T = \tau]$$

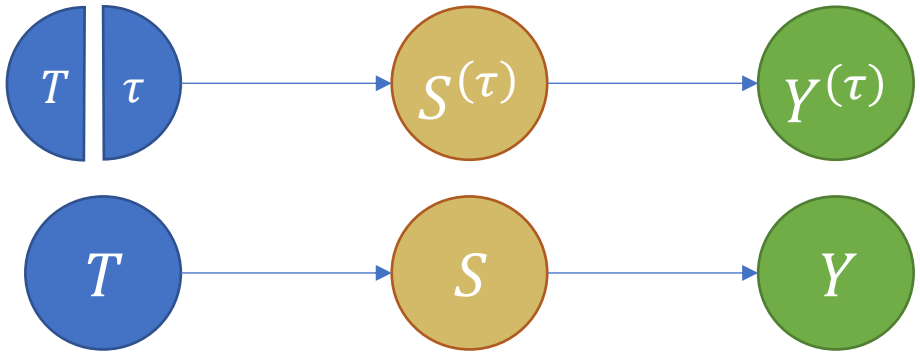
Average reward if intervene and set investment= τ

Independence in counterfactual graph

Average reward of samples that received investment=τ in data

Tower Law of Expectations

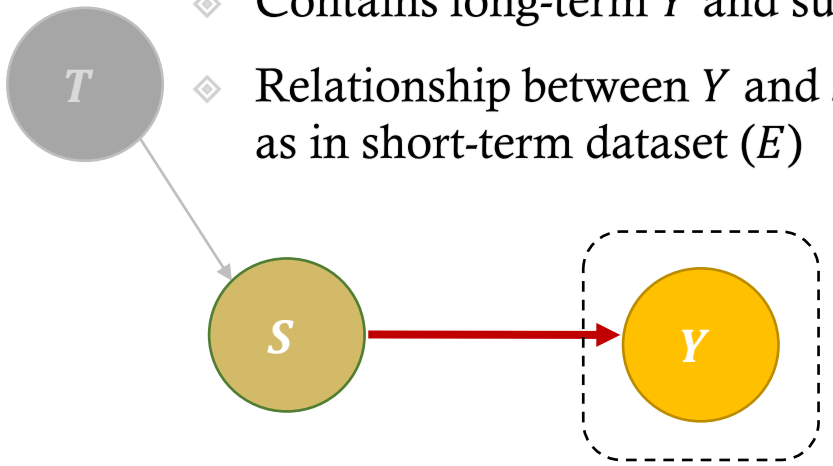
Forecasted reward from surrogates



Causal Inference with Surrogates 101

historical/long-term (O)

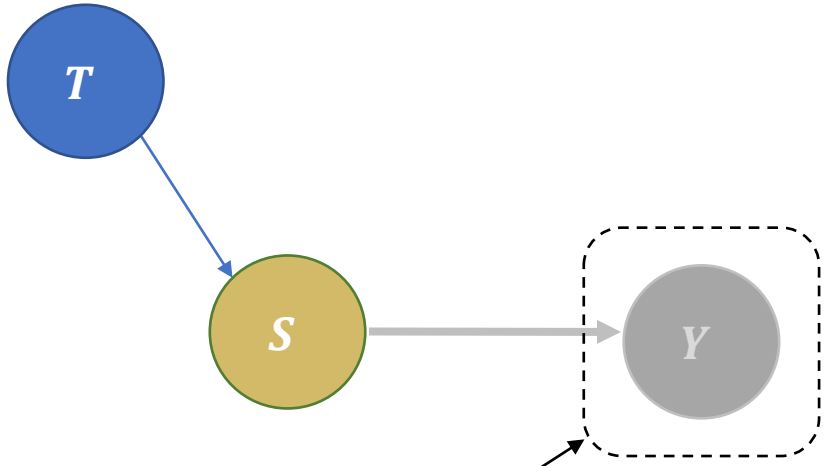
- Contains long-term Y and surrogates S
- Relationship between Y and S the same as in short-term dataset (E)



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

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

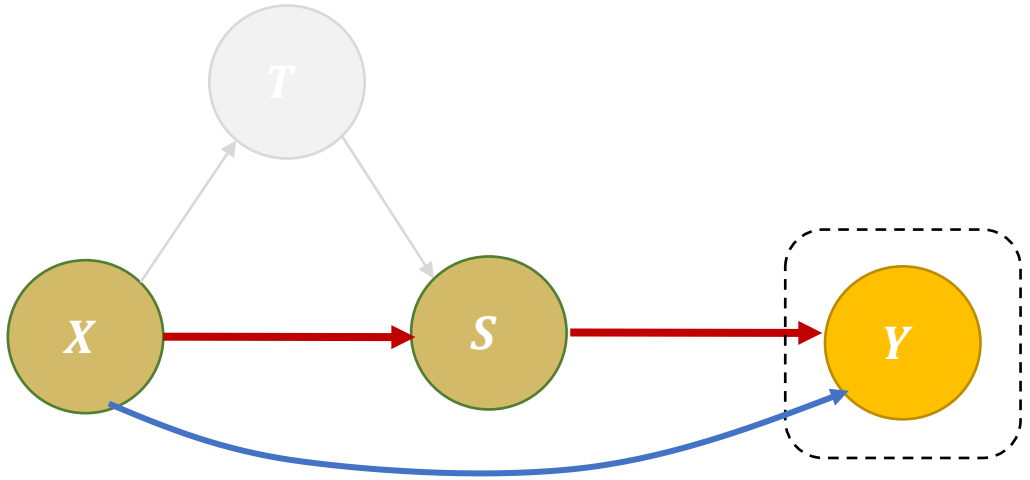
recent/short-term (E)



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

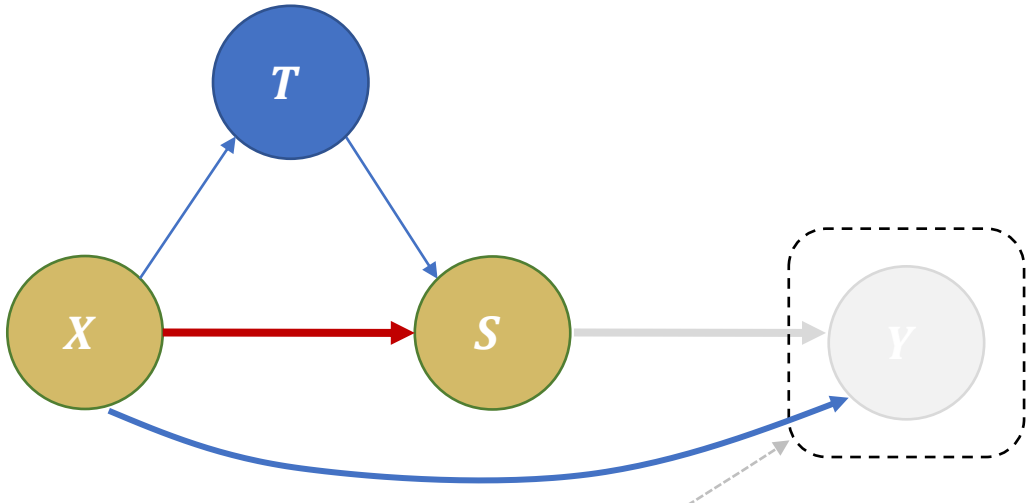
long-term data (ℓ)



$h(S, X) = E_{\ell}[Y|S, X]$

1. Estimate $h(S, X) \coloneqq E[Y|S, X]$ (surrogate index) by regressing $Y \sim S, X$

short-term data (s)



2. Impute expected long-term outcomes
3. Estimate effect of T on imputed outcomes

(Athey et al., 2020)

Causal Inference with Surrogates 102

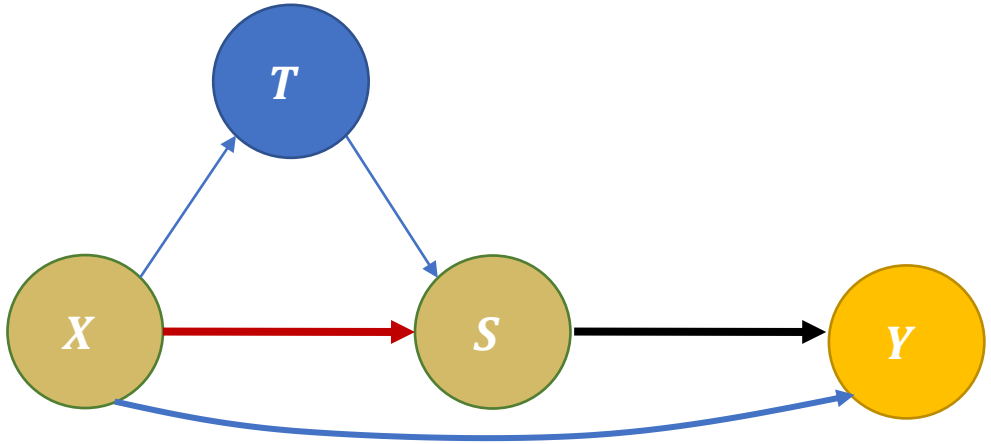
$$\textit{Reminder: } \theta_0 := E\left[Y^{(\tau)}\right] = E\left[E[Y|T=\tau, X]\right], \text{ (g–formula)}$$

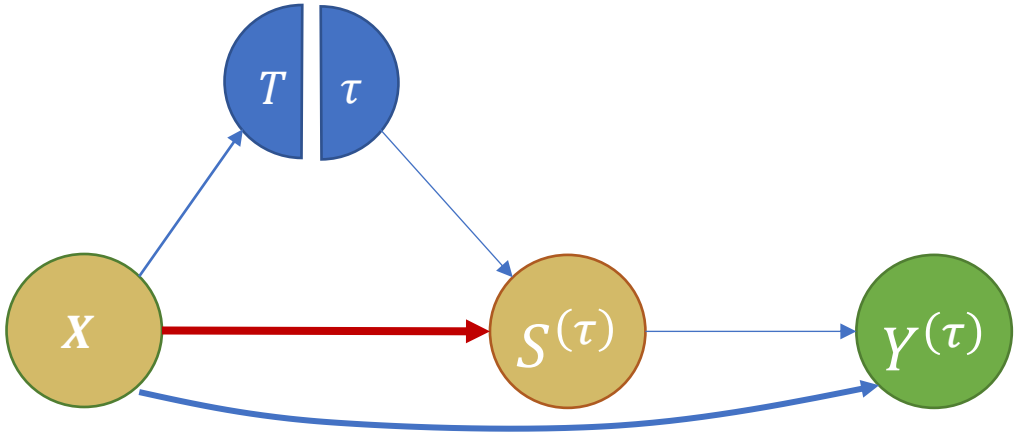
- Surrogacy assumption implies:

$$E[Y|T=\tau, X] = E[\, E[Y|S, X] \mid T=\tau, X\,]$$

- Surrogacy based g-formula

$$\theta_0 = E\left[E[h(S,X)|T=\tau, X]\right]$$





Nested Regression Estimand

- Surrogacy based g-formula

$$\theta_0 = E_s[g(\tau, X)]$$
$$g(T, X) = E_s[h(S, X)|T, X]$$
$$h(S, X) = E_\ell[Y|S, X]$$

- g regression function is trained on the outputs of the h regression function
- Surrogate Index Model h is trained on long-term data distribution
- Target parameter is averaged over short-term data distribution

Automatic Debiased Machine Learning

- Surrogacy based g-formula

$$\theta_0 = E_s[g(\tau, X)]$$
$$g(T, X) = E_s[h(S, X)|T, X]$$
$$h(S, X) = E_\ell[Y|S, X]$$

- Automatic debiased formula

$$\theta_0 = E_s\left[g(\tau, X) + a_1(T, X)\left(h(S, X) - g(T, X)\right)\right]$$

Recursive Automatic Debiasing

- Surrogacy based g-formula

$$\theta_0 = E_s[g(\tau, X)]$$
$$g(T,X) = E_s[h(S,X)|T,X]$$
$$h(S,X) = E_\ell[Y|S,X]$$

- Automatic debiased formula

$$\theta_0 = E_s\left[g(\tau,X) + a_1(T,X)\left(h(S,X) - g(D,X)\right)\right] + E_\ell\left[a_2(S,X)\left(Y - h(S,X)\right)\right]$$

Inverse Propensity ratio

$$\frac{1\{T=\tau\}}{\Pr(T=\tau \mid X)}$$

Surrogate Score(S,X) * $E[\,a_1(T,X) \mid S,X\,]$

Surrogate score: (density ratio of S,X between ℓ and s "environment")

$$\frac{p(S,X \mid E=s)}{p(S,X \mid E=\ell)} = \frac{\Pr(E=s \mid S,X)\Pr(E=\ell)}{\Pr(E=\ell \mid S,X)\Pr(E=s)}$$

Multiplied by

$$E[\,a_1(T,X) \mid S,X\,] = \frac{\Pr(T=t \mid S,X)}{\Pr(T=t \mid X)}$$

An Application from Operations Management

Return on Investment at Microsoft using Surrogates and Dynamic Treatment Effects

Estimating Long-Term Returns on Investment

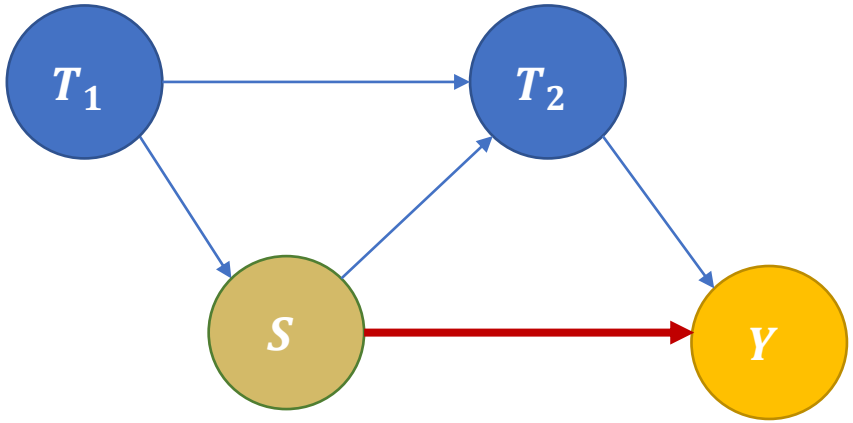
- Companies frequently deploys new discount or customer support programs
- Which of these programs ("investments") are more successful than others?
- Success is a long-term objective: what is the effect of the program on the two-year customer journey (e.g., effect on two-year revenue)
- We cannot wait 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?

Key Assumptions

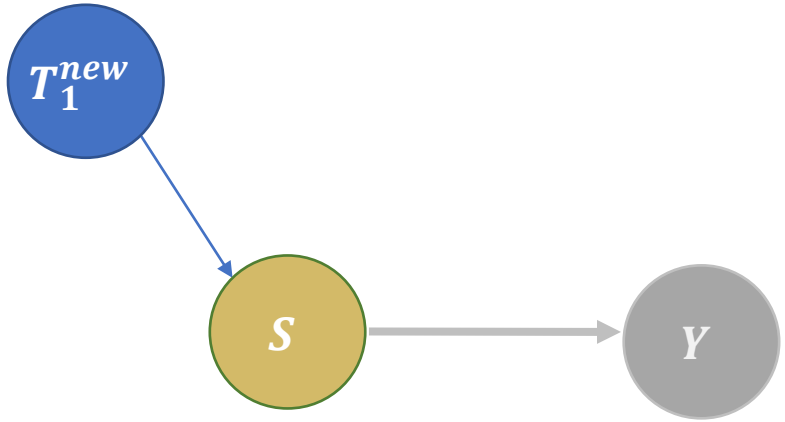
- Long-term effect only goes through surrogates
- Expected relationship between surrogates and long-term reward is the same long-term setting (O) and in short-term setting (E)

Key Assumptions can be Easily Violated
Investment policies are dynamic and change

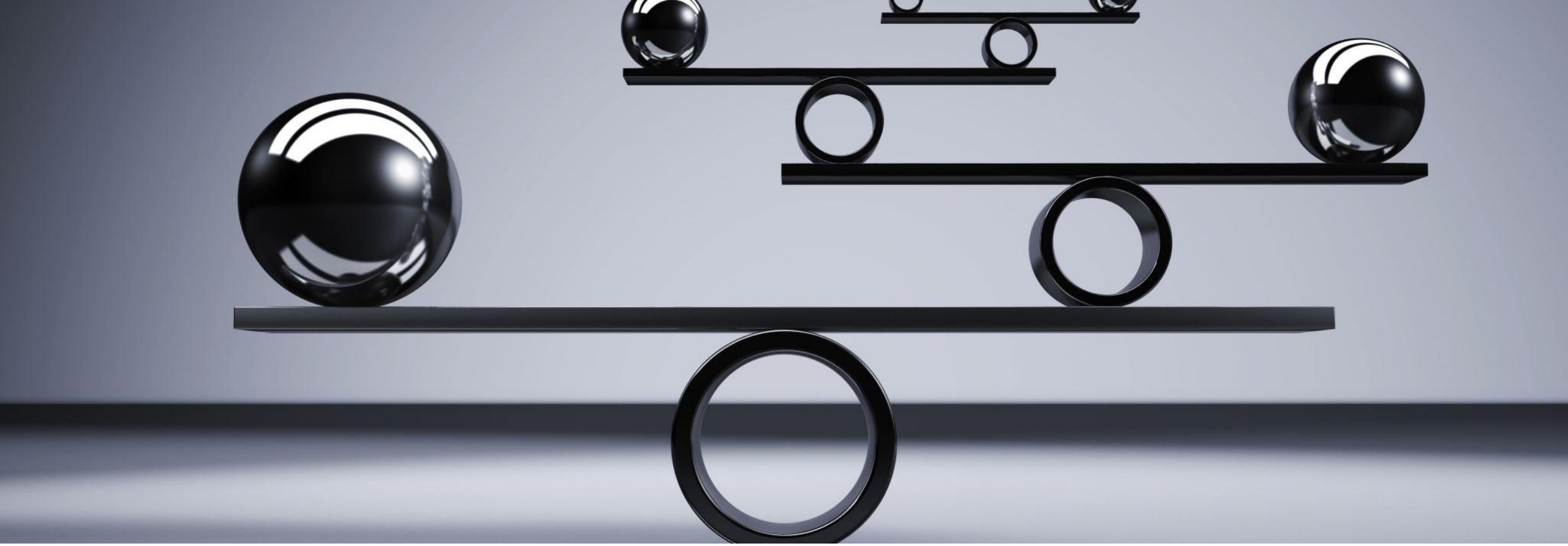
historical/long-term (O)



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

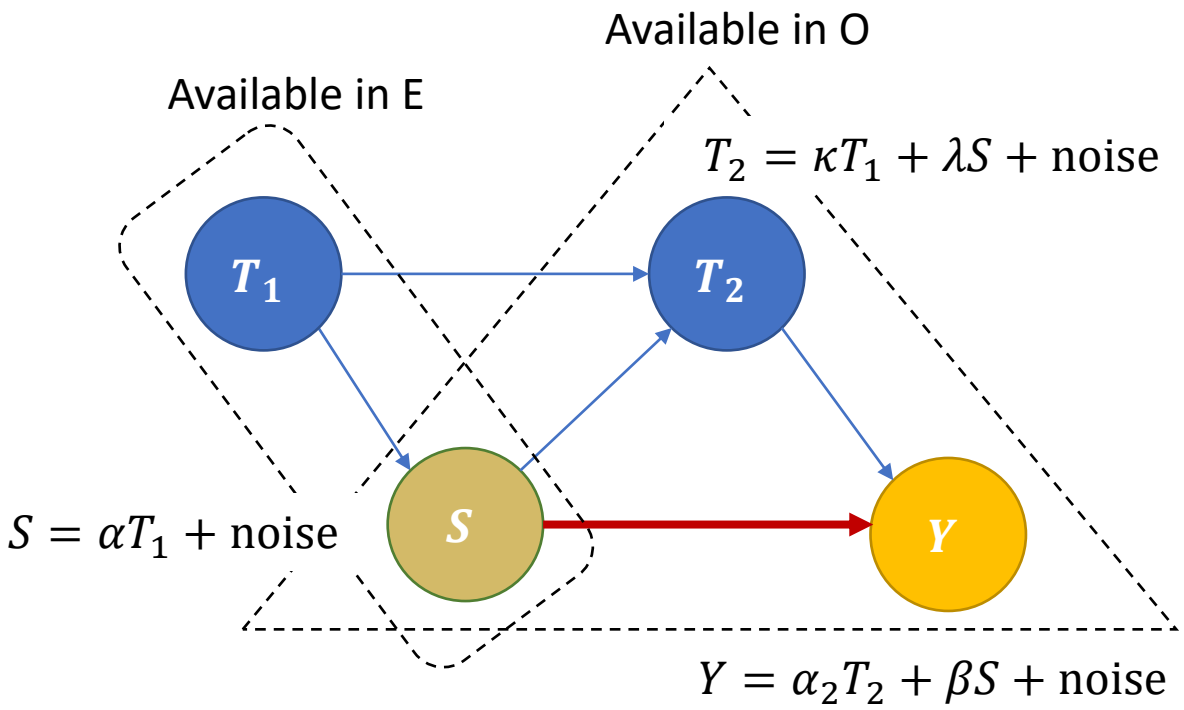


Bias of Vanilla Surrogate

Illustrative Example

Illustrative Example

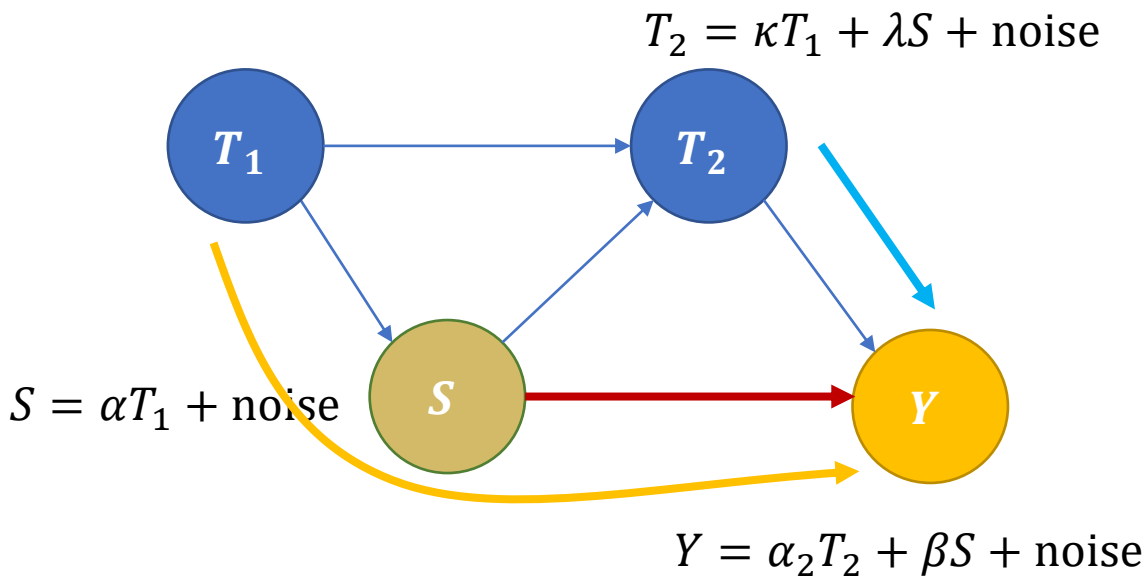
- Want the effect of T_1 on Y, with no future treatments



Bias of Vanilla Surrogate

- What do we want to estimate?

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



Bias of Vanilla Surrogate

- What do we want to estimate?

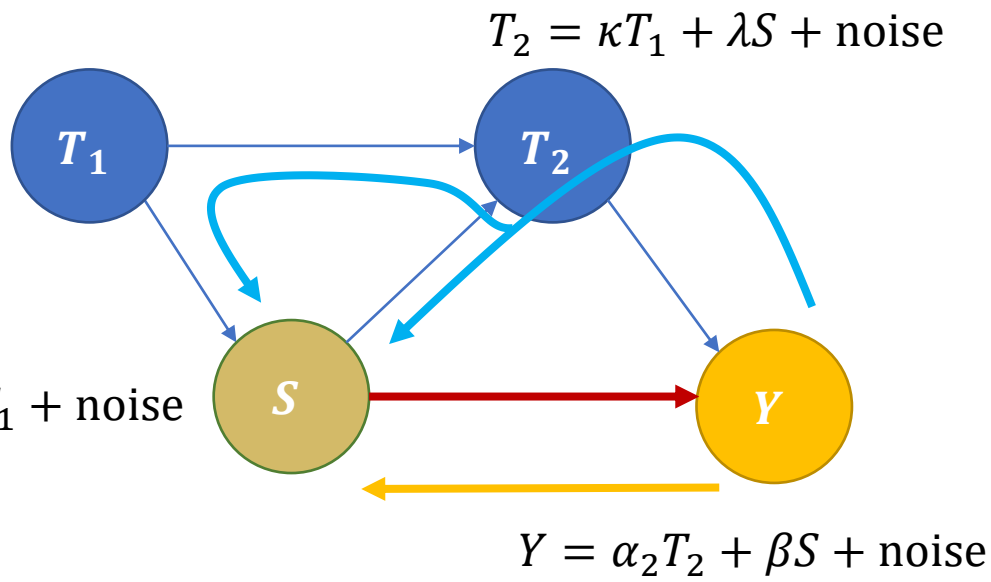
$$Y = \boldsymbol{\beta}\,\boldsymbol{\alpha}\,T_1 + \alpha_2\,T_2 + noise$$

$$\boldsymbol{\beta}\,\boldsymbol{\alpha} \;\;!!\;\; \boldsymbol{\theta_0}$$

- What does the surrogate approach estimate?

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

$$\boldsymbol{\kappa}E[T_1|S] + \boldsymbol{\lambda}\boldsymbol{S}$$



Bias of Vanilla Surrogate

- What do we want to estimate?

$$Y = \boldsymbol{\beta}\,\boldsymbol{\alpha}\,T_1 + \alpha_2\,T_2 + noise$$

$$\boldsymbol{\beta}\,\boldsymbol{\alpha} \;\;!!\;\; \boldsymbol{\theta_0}$$

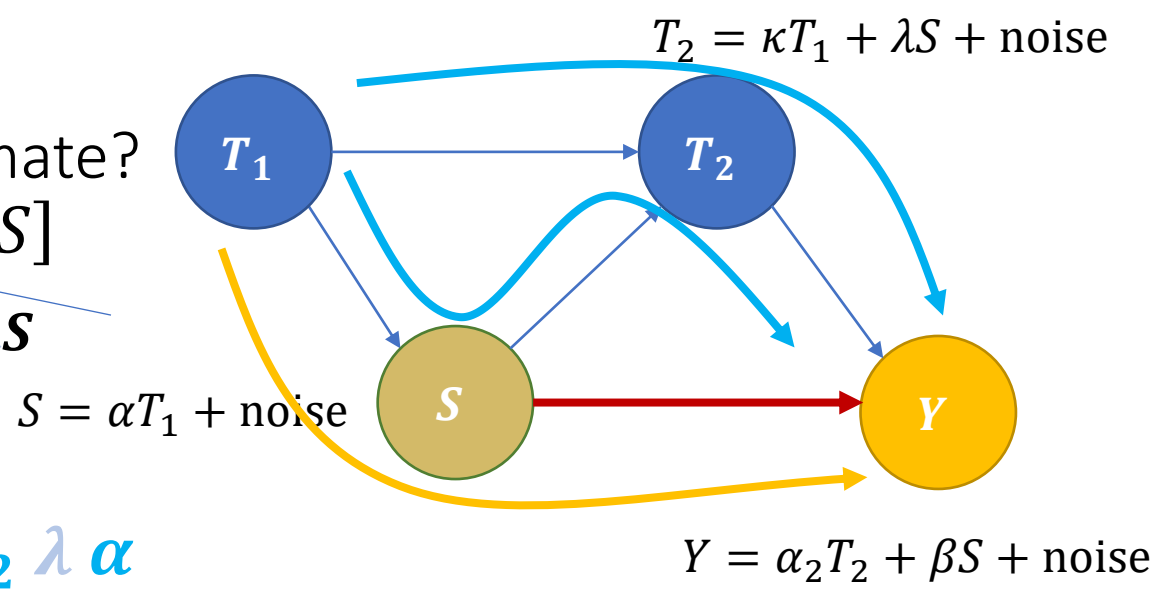
- What does the surrogate approach estimate?

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

$$\boldsymbol{\kappa} E[T_1|S] + \boldsymbol{\lambda} \boldsymbol{S}$$

- 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?

$$Y = \boldsymbol{\beta}\,\boldsymbol{\alpha}\,T_1 + \alpha_2\,T_2 + noise$$

$$\boldsymbol{\beta}\,\boldsymbol{\alpha} \;\;\text{!!}\;\; \boldsymbol{\theta_0}$$

- What does the surrogate approach estimate?

$$g_0(S) := E[Y|S] = \boldsymbol{\beta}\,S + \boldsymbol{\alpha_2}\,E[T_2|S]$$

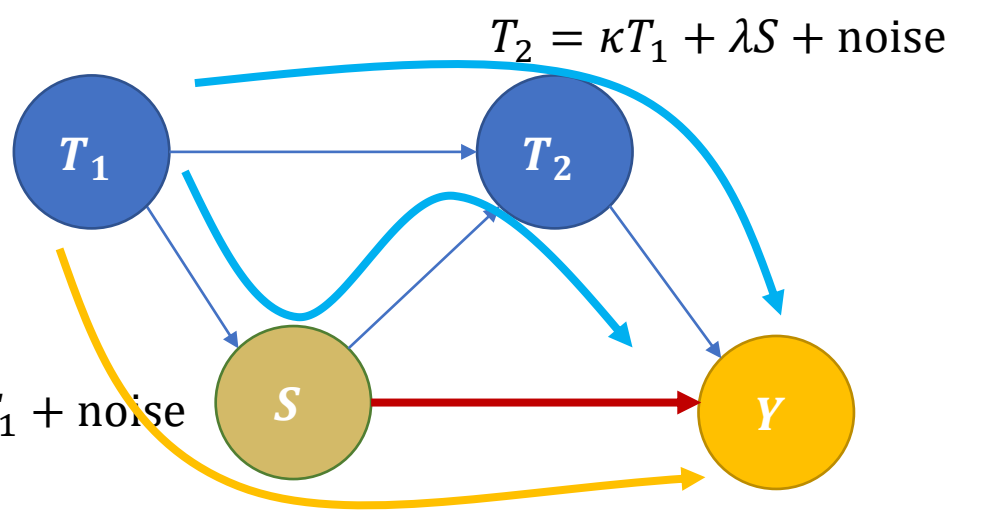
$$\boldsymbol{\alpha_2}\,E[T_2|S] = \boldsymbol{\alpha_2}\,\boldsymbol{\kappa}E[T_1|S] + \boldsymbol{\alpha_2}\,\boldsymbol{\lambda}\boldsymbol{S}$$

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

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

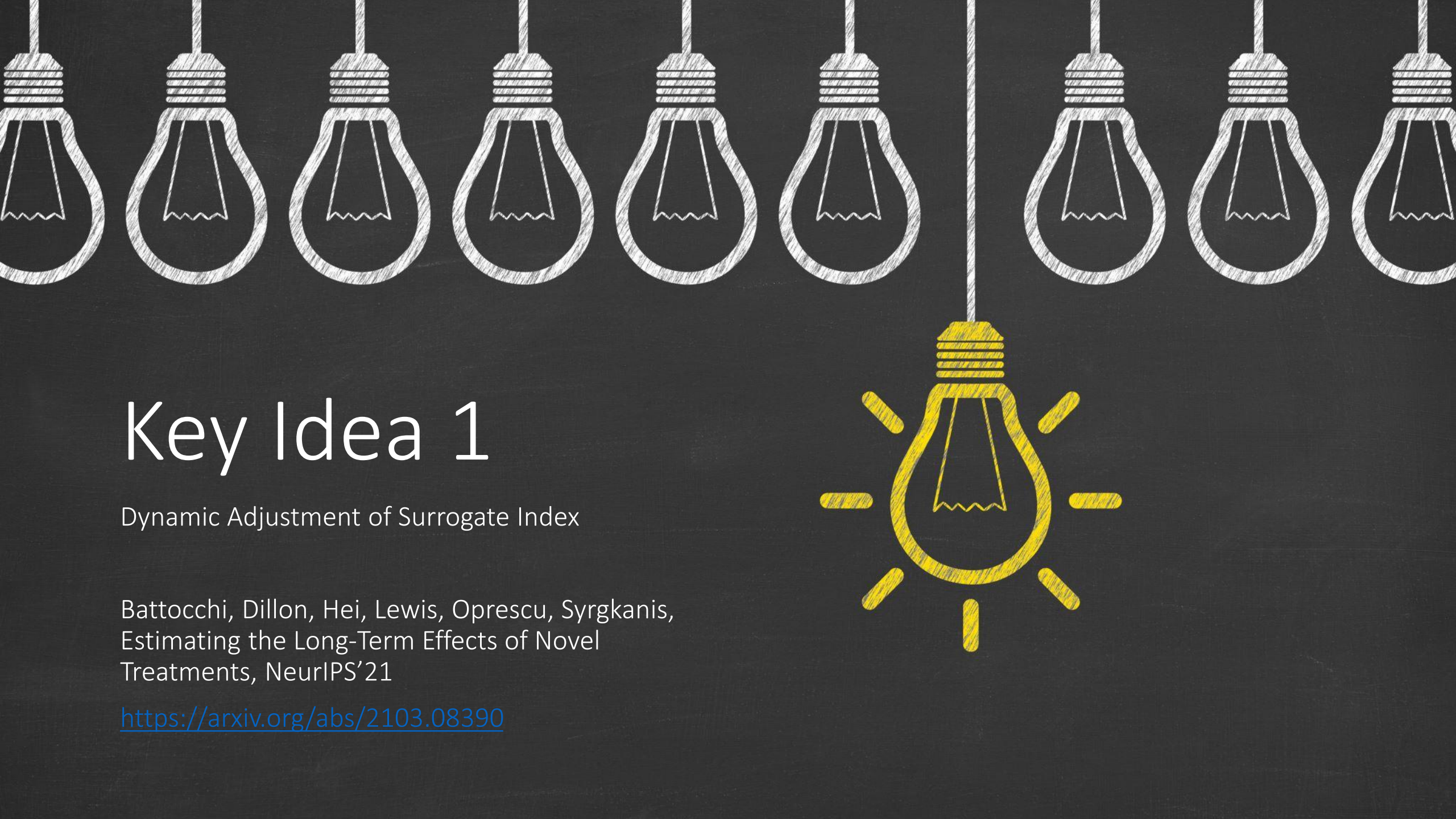
Bias from the fact that policy is auto-correlated violating the "effect only through surrogates" assumption

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



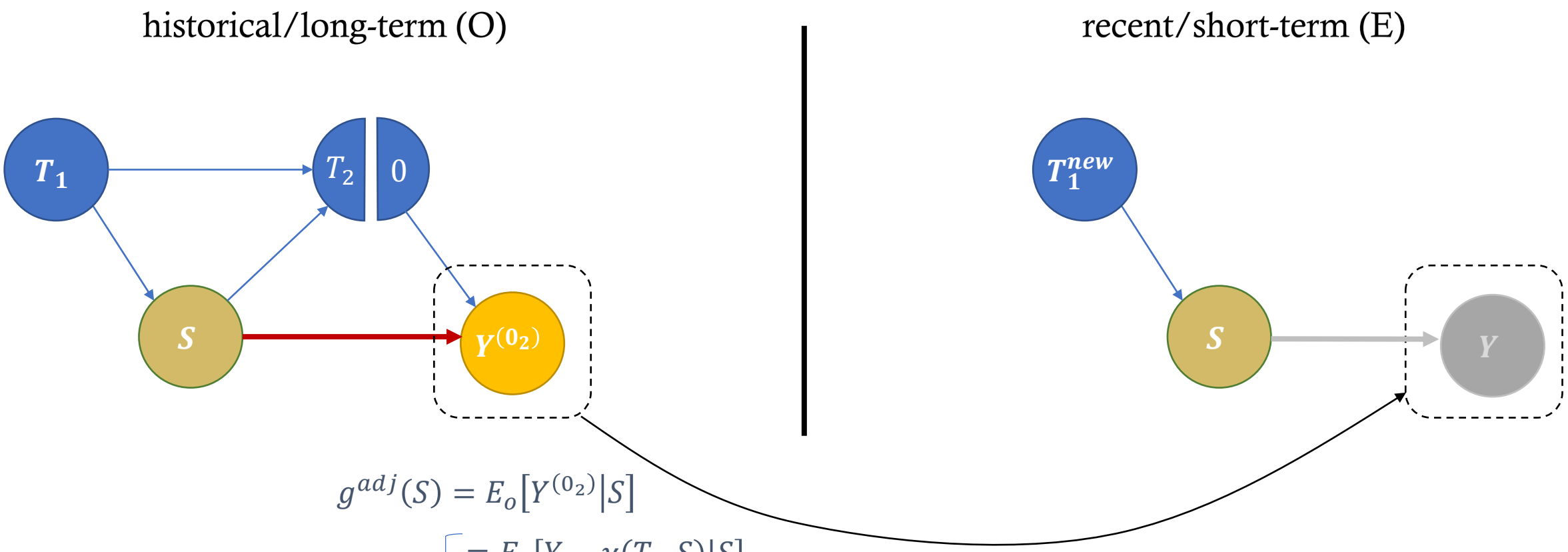
Key Idea 1

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



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

$$= E_o[Y - \gamma(T_2, S)|S]$$

Identification argument

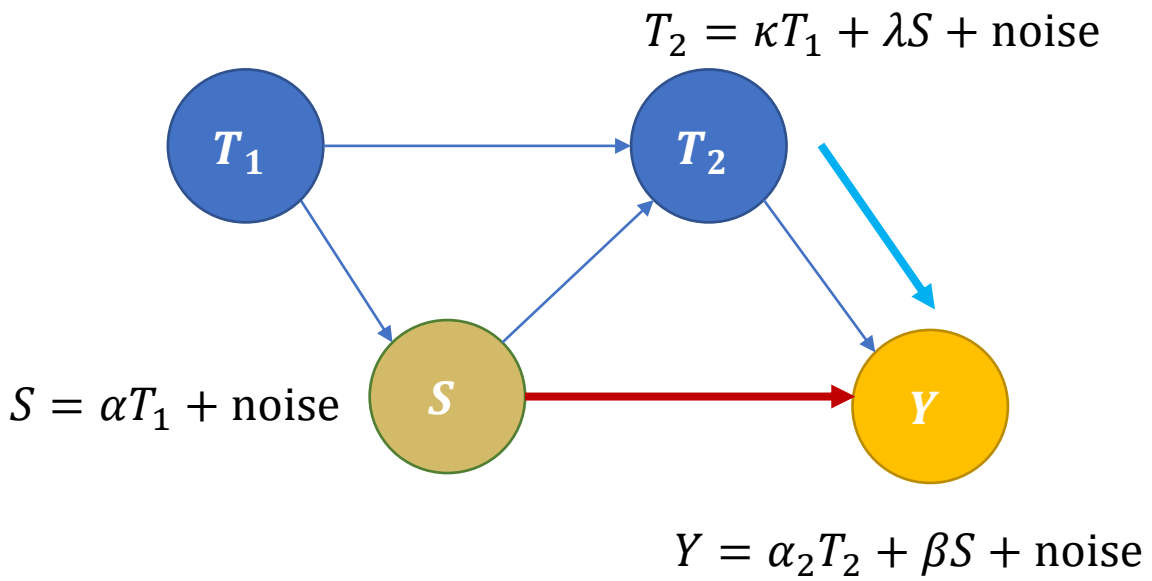
"blip" effect of second treatment conditional on surrogate

⟺ effect of T_2 controlling for S

Illustrative Example: Dynamically Adjusted Index

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

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



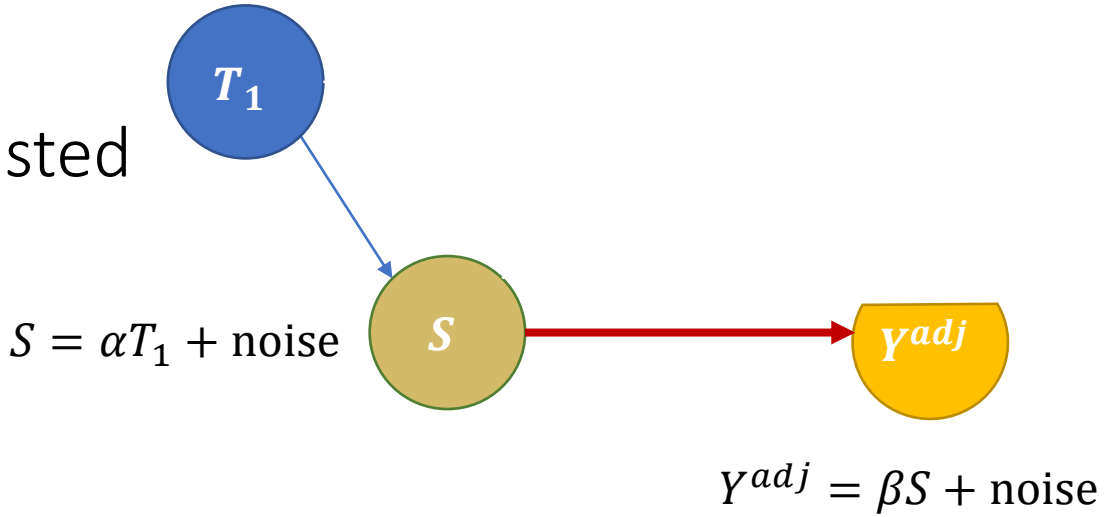
Illustrative Example: Dynamically Adjusted Index

- 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

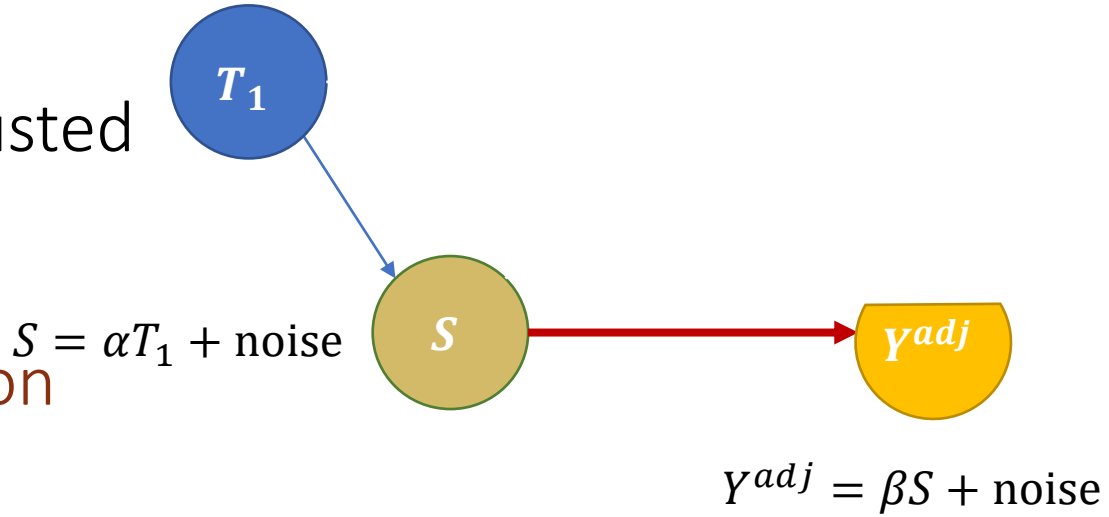
$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$

- 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

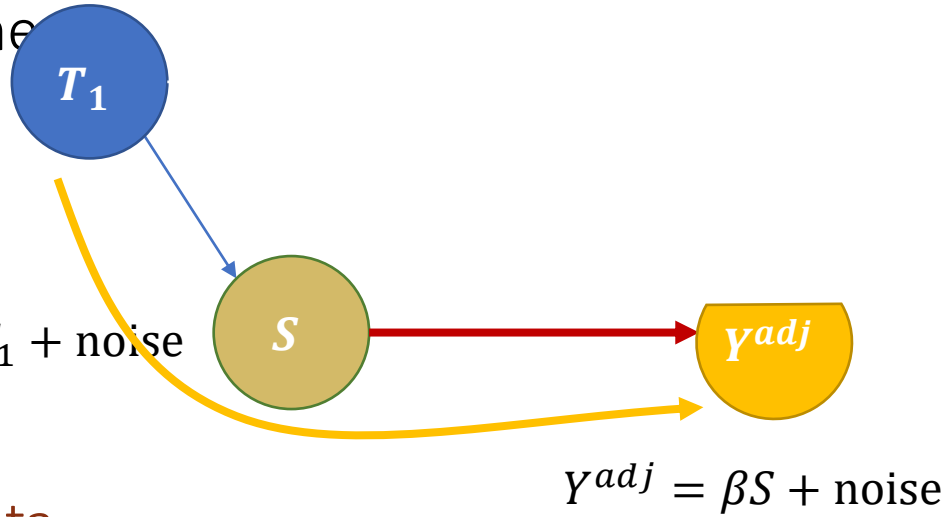
$$E[Y|T_2, S] = \alpha_2 T_2 + \beta S$$

- Subtract that effect to create the adjusted outcome

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

- Estimate dynamically adjusted index on long-term data

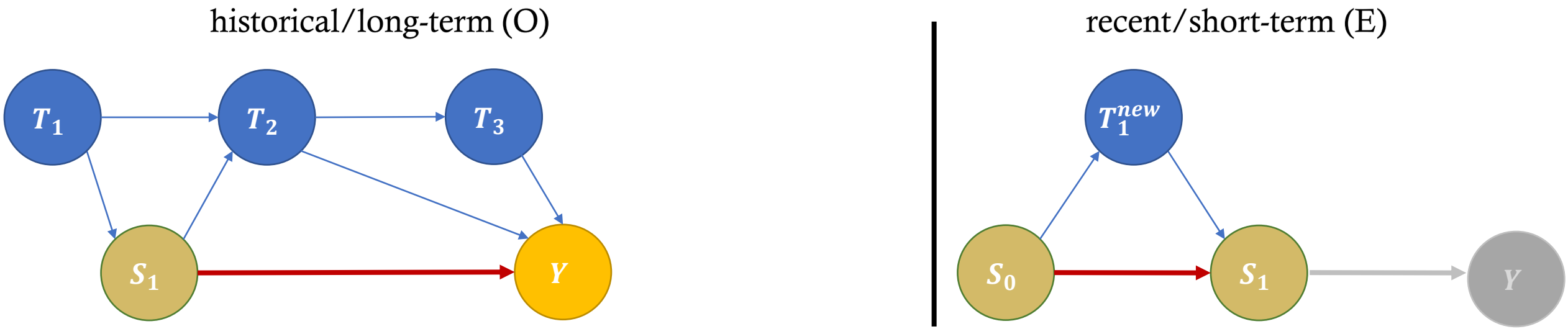
$$g_0^{adj}(S) := E\left[Y^{adj}\middle|S\right] = \beta\, S$$



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

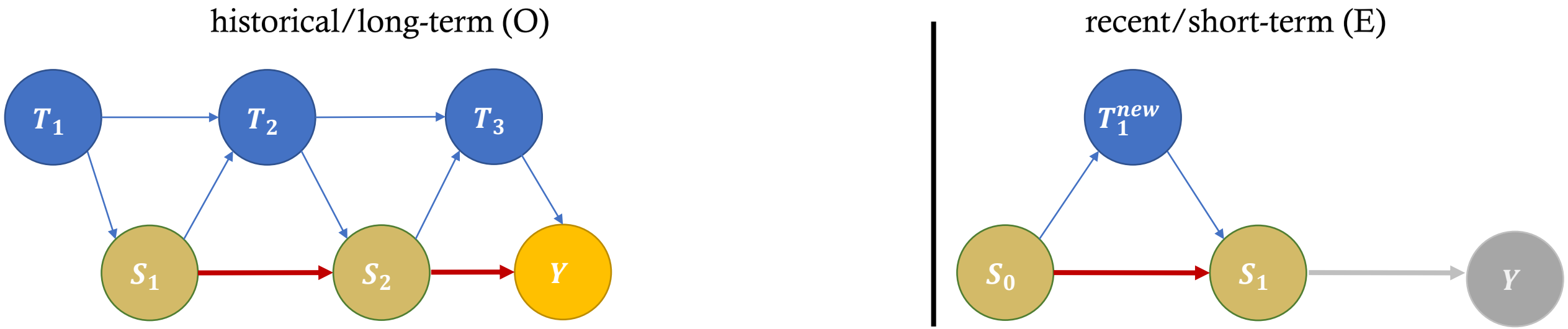
$$\theta_0 := \frac{E\left[g_0^{adj}(S) \cdot T_1\right]}{E\left[T_1^2\right]} = \beta \cdot \alpha$$

Beyond Two Periods



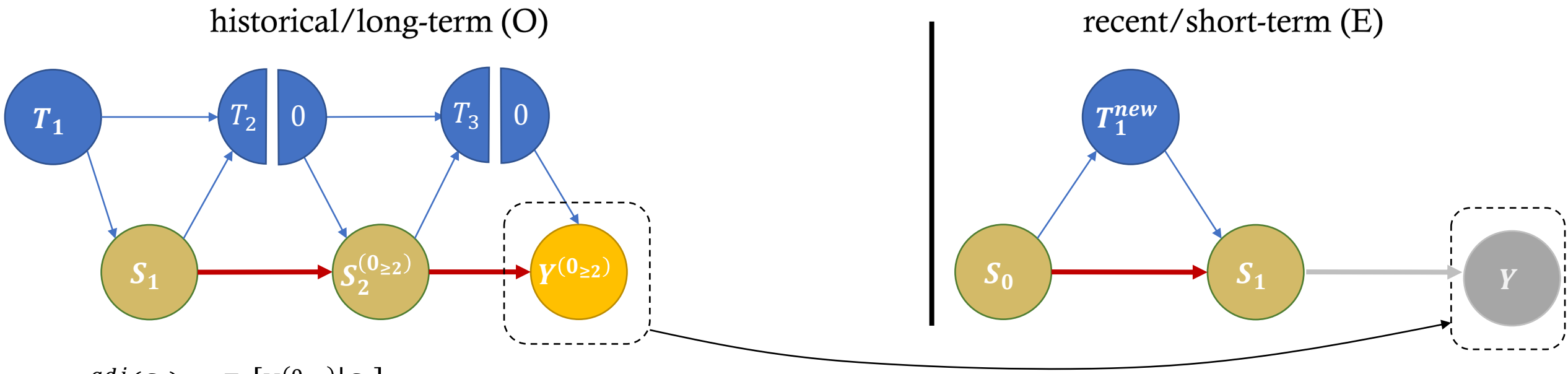
- Multiple treatments are offered after the surrogate variable
- Easy: estimate their effect controlling for S_1
- Wrong!

Beyond Two Periods



- 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



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

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

"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