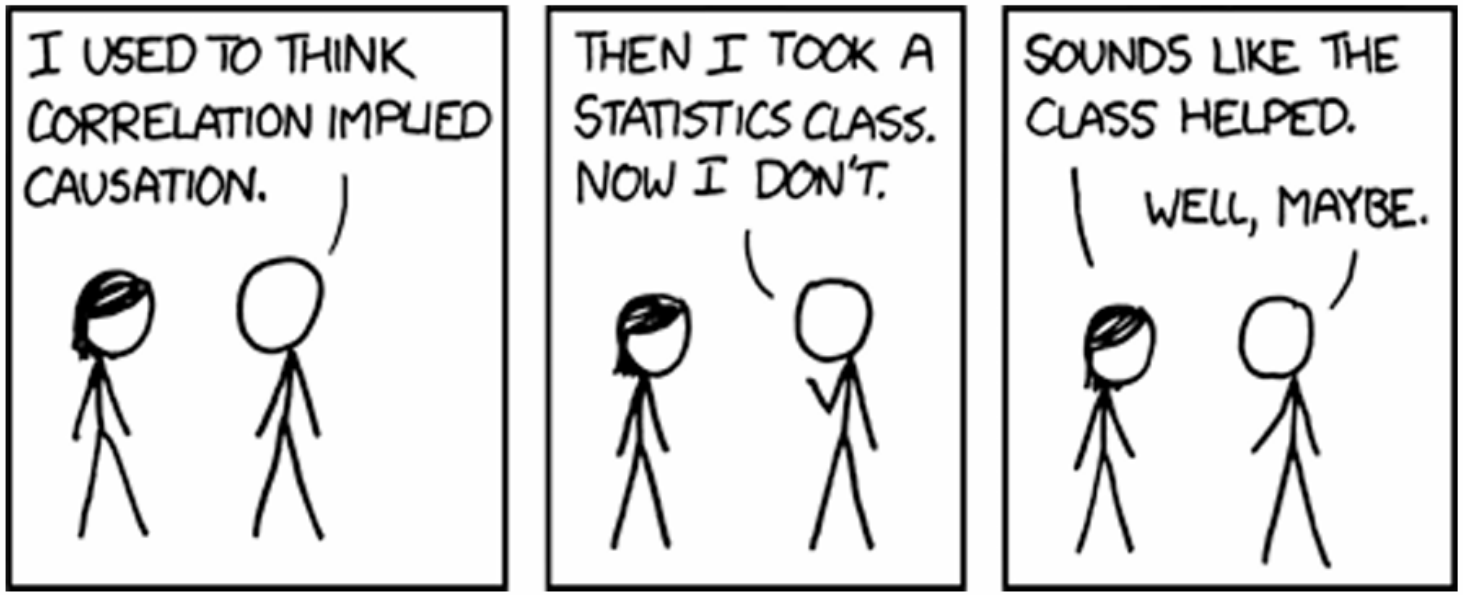
Causal Learning



Inferring the effect of one thing on another

Did X cause Y?

Golden standard of causal inference are randomized experiments

But we can't always run an experiment

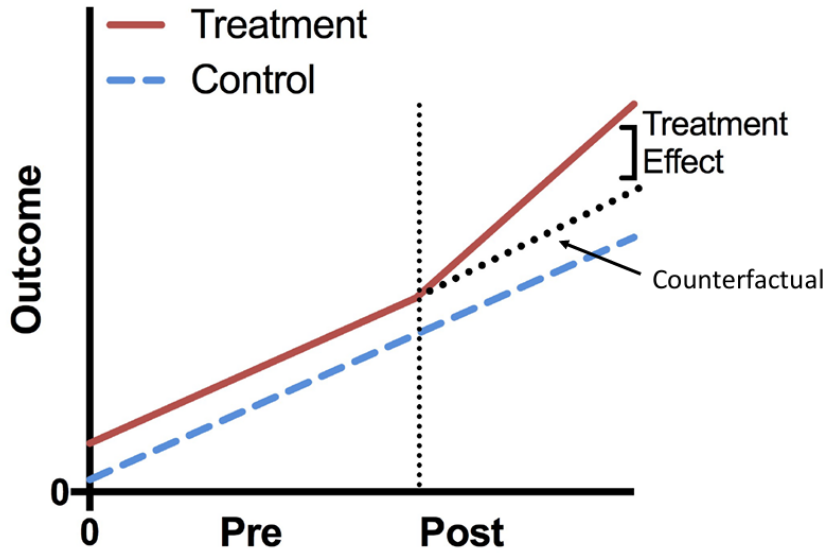
Enter quasi-experiments!

Data Setting: Pre/Post Treatment/Control

Challenge: Randomized control design is infeasible
Solution: Quasi-experimental methods

Methods for causal inference

- Difference-In-Differences (DID)
- Synthetic Control



$$Y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 (TREAT_i)(POST_t) + u_{it}$$

Parallel Trends Assumption
Treatment unit would have been parallel to the control unit in the absence of the treatment

What if it is violated?

Synthetic Control Methods to the rescue!

Main idea

- Whereas DID essentially require equal weights on control units, SC allows the weights on control units to vary (and allow zero weights)
- Additional flexibility to match controls to treatment
- Use weighted average of control units to create a synthetic version of treatment unit

Synthetic Control Method Overview

Using pre-treatment data, select weights **w** to minimize

$$\sum_{t=1}^{T_1} \left[y_{treat,t} - y'_{control,t} w \right]^2$$

s.t. $\sum_{j=1}^{N_{co}} w_j = 1$ and $w_j \geq 0,\ j = 1, \ldots, N_{co}$

Note: need long enough time series

Relaxing restrictions -> more flexible methods

Actual Sales

$$ATT = \frac{1}{T_2} \sum_{T_1+1}^{T} (y_{treat,t} - \hat{y}^0_{treat,t})$$

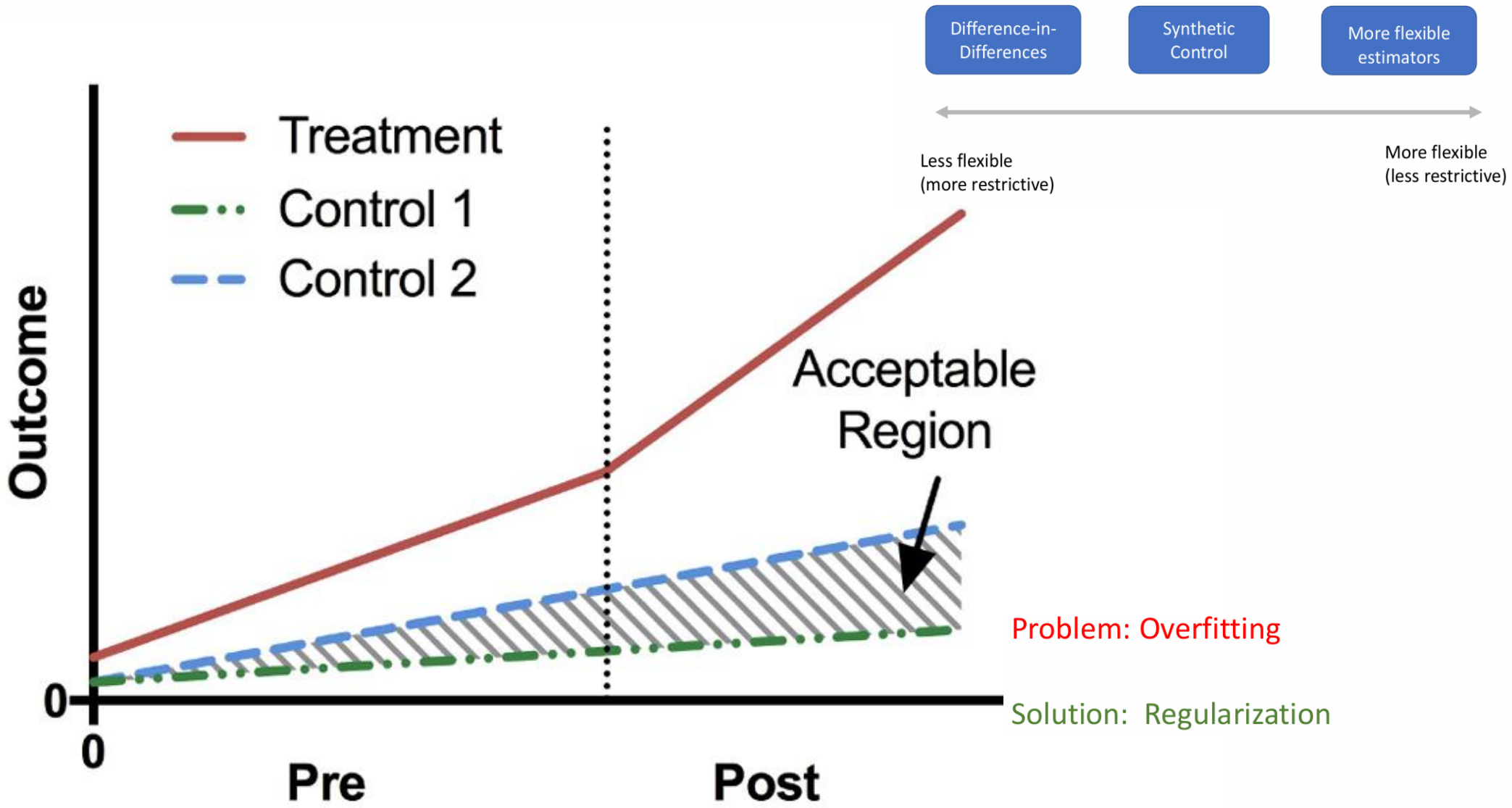
Counterfactual

$$\hat{y}^0_{treat,t} = \hat{w}_1 y_{control1,t} + \ldots + \hat{w}_{N\,co} y_{controlN_{co},t}$$

How to quantify uncertainty when using synthetic control method? (Li 2020, *JASA*)

- Previously, researchers had to rely on placebo tests that have strong symmetry assumptions
  - Strong symmetry assumption treatment and controls have similar data variation (variance)
  - No confidence bounds, hypothesis testing or (general) p-values
- Li (2020) develops formal inference theory for the synthetic control method and
- Proves that subsampling procedure can be used to calculate confidence intervals, conduct hypothesis testing and obtain p-values

What if we need even more flexible methods?



Problem: Overfitting

Solution: Regularization

Data characteristics of research problem guide the choice of method

- number of pre/post time periods
- number of treatment and control units
- treatment unit within or outside range of control units

Goal: Use the simplest (i.e. most restrictive) method that has parallel pre-trends without overfitting

Factor Model: Flexible & Frequentist method

It is popular in treatments effects as well as effects in marketing events. It quantifies uncertainty through hypothesis testing and confidence intervals, allows treatment and control unit outcomes to have different distributions (unequal error variance).

Factor Model Approach Overview

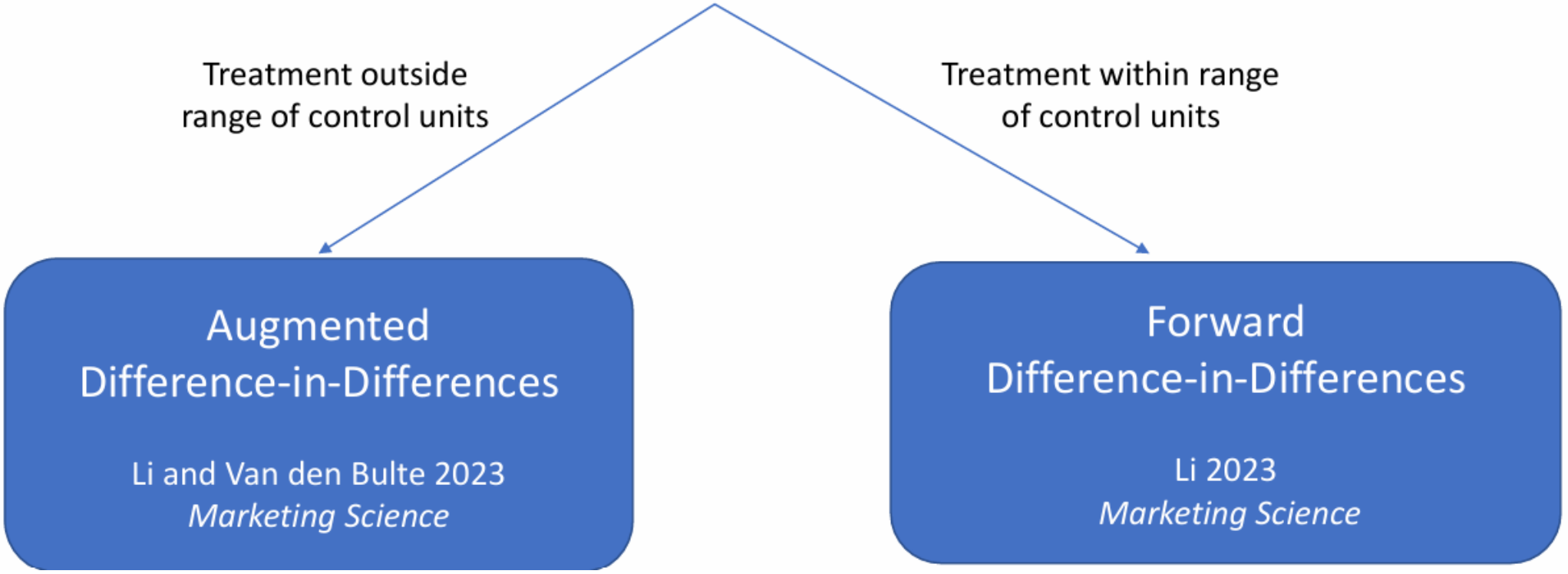
(Gobillon and Magnac 2016, Chan and Kwok 2016, Xu 2017)

- Similar in spirit but mechanically different to synthetic control
  1) Project the control units onto a lower dimensional factor space (largest eigenvectors of control data)
  2) *Regress the treatment unit's outcome on the latent factors to recover the factor loadings (weights) without any constraints to predict the treated counterfactuals*

- Uses all the control units' information to estimate common factors without discarding information
- Can handle a large number of control units
- Can handle treatment outside range of the control units
- Implicit regularization leads to better out of sample predictions
- Easily applied to multiple treatment units because dimension reduction to obtain factors only has to be done once

- Also called interactive fixed effects model, generalized synthetic control

What if we don't have enough time periods?



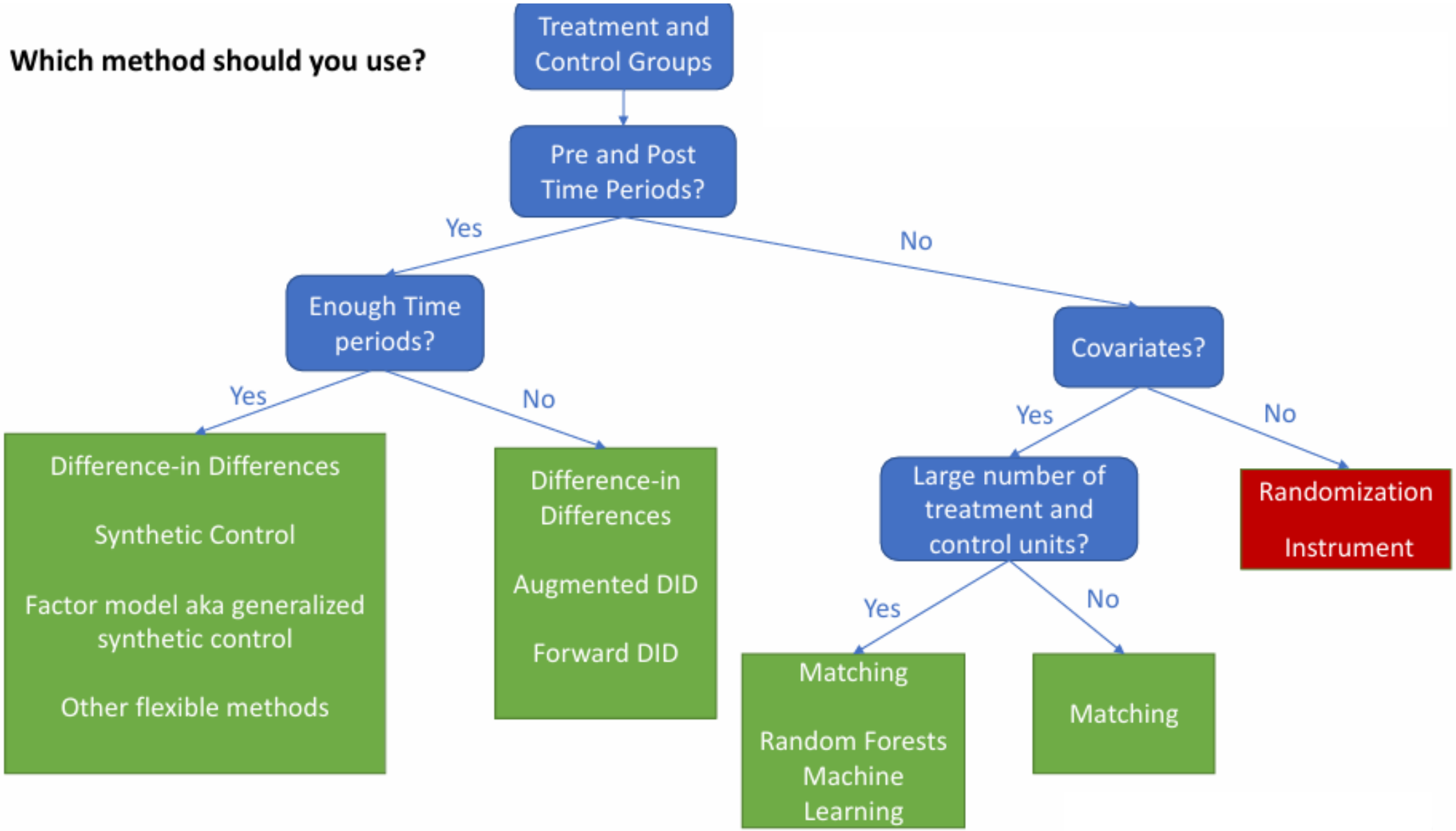
Augmented DID and Forward DID are especially useful when

- DID is not applicable so we need more flexible estimators
- We have short (although still applies to long) pre and post treatment time periods

How to arrive at more flexible estimators from synthetic control method?

1) Change the weights or add an intercept
   Doudchenko and Imbens 2016, Li 2020
   Hsiao Ching and Wan 2012
   Augmented DID Li and Van den Bulte 2023, *Mkt Sci*
   Synthetic DID (introduce time weights) Arkhangelsky et al 2021 *AER*
2) Dimension reduction via factor model aka generalized synthetic control
   Xu 2017, Inference - Li and Sonnier 2023, *JMR*
3) Choose a relevant subset of control units
   Shi and Huang 2022, Li 2023
4) Bayesian
   Kim, Lee and Gupta 2020 *JMR*, Pang, Liu and Xu 2022

Which method should you use?



Final points

- Causal inference is fundamental to science and knowledge generation
- DID is most popular, widely used quasi-experimental method
- Synthetic control method started the renaissance in more flexible estimators (good for small to moderate # treatment, large enough time periods)
- When to use which method is guided by data characteristics / satisfying corresponding identifying assumption
- Expanding the in causal inference toolkit available to marketing scholars (factor model, Bayesian, Augmented DID, Forward DID) allows researchers to answer previously unanswerable questions!