Synthetic Interventions

Dennis Shen

Joint work with

Anish Agarwal

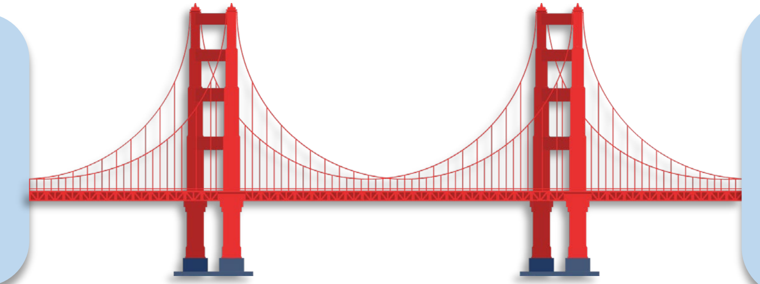
Devavrat Shah

Bridging causal inference & machine learning

Synthetic controls

(what if no intervention occurred?)

Core idea in econometrics



Matrix/tensor completion

(impute missing data in a matrix/tensor)

Core idea in EE/CS/Stats/ML

1. Synthetic Interventions

[A. Agarwal, D. Shah, D. Shen]

(synthetic controls ⟵ tensor completion)

2. Causal Matrix Completion

[A. Agarwal, M. Dahleh, D. Shah, D. Shen]

(synthetic controls ⟶ matrix completion)

Clinical trial study w. Alzheimer's Therapeutics company

2 year study

1000+ subjects

4 therapies (1 placebo)

Alzheimer's clinical trial study

Inconclusive

Average treatment effect for all 3 therapies was insignificant

Costly

Total cost of trial: $500M - $1B USD (cost of recruiting one patient: $5k – $100k USD)
Ethical concerns of testing on human subjects

Question we set out to answer

Maybe therapies were effective for subset of patients?

↓

Can we estimate ADAS-COG score for each patient under each therapy?

A question:

Can we estimate ADAS-COG score for each patient under each therapy?

A framework :

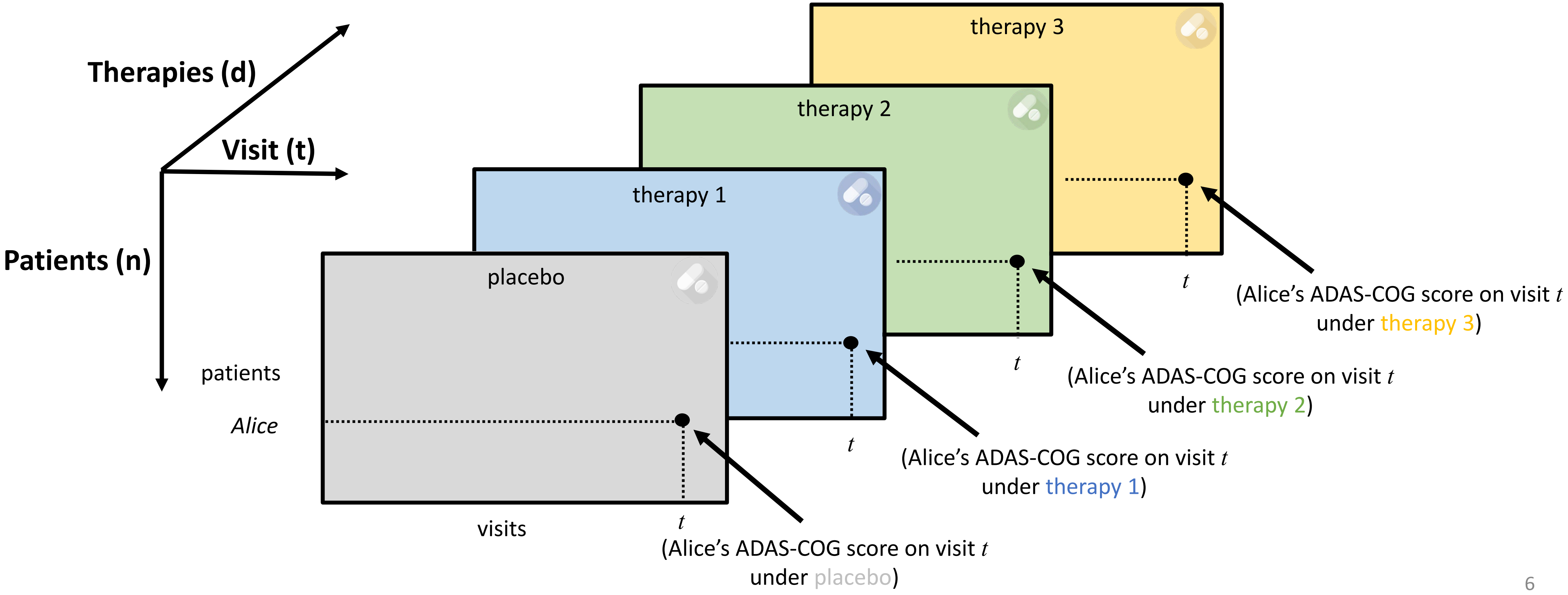
Causal inference as tensor completion

Potential outcomes: a tensor viewpoint

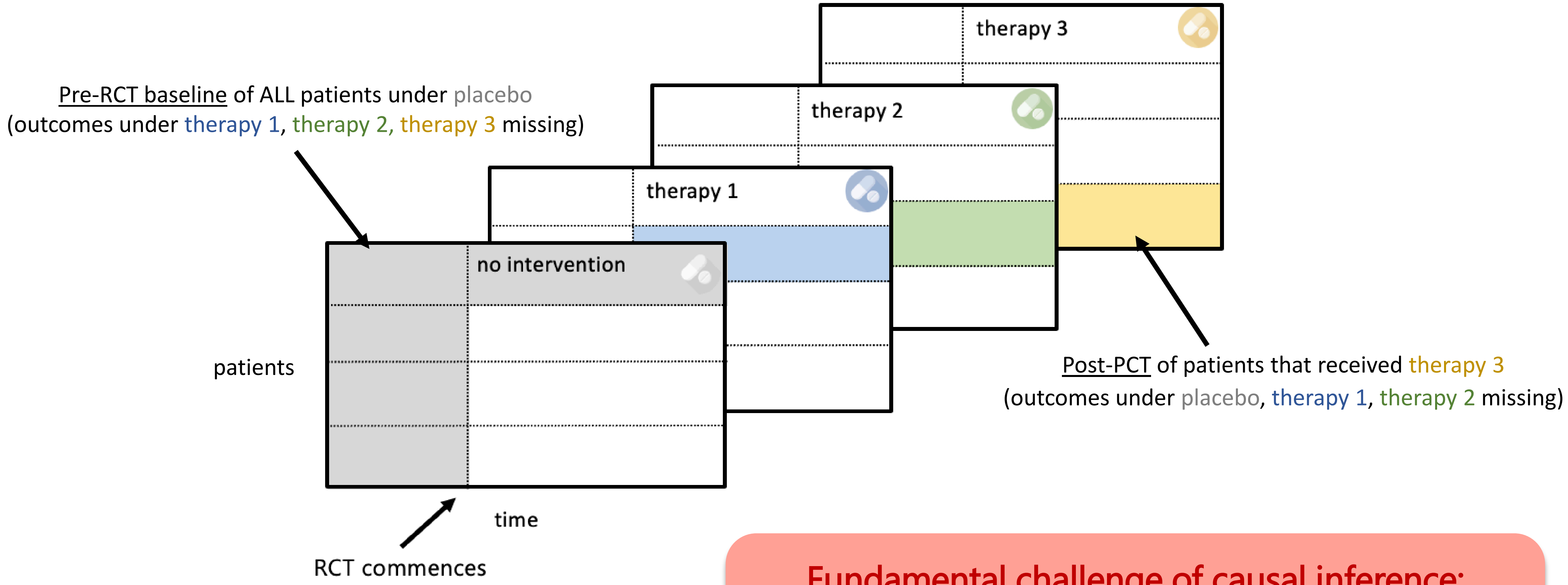
Potential outcomes: $Y_{nt}^{(d)}$

<u>What if</u> patient n on visit t had been assigned therapy d

[Neyman'23, Rubin'74]



What data we had from



Fundamental challenge of causal inference:
only observe one outcome
want to know all potential outcomes

RCTs – randomization but no personalization

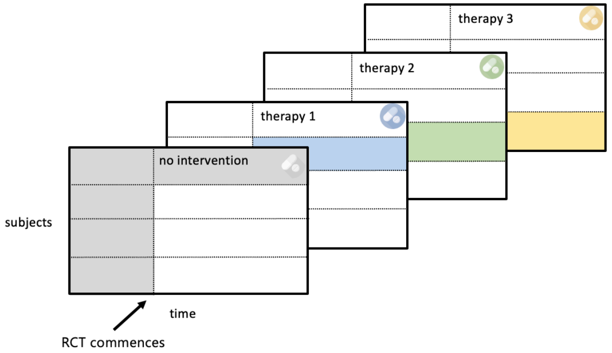
What RCT estimates

Average treatment effect

Avg( ) Therapy 1

-

Avg( ) placebo



Why are RCTs beloved?

Explicit randomization

What RCT cannot estimate

Individual treatment effect

-

Alice under therapy 1

Alice under placebo

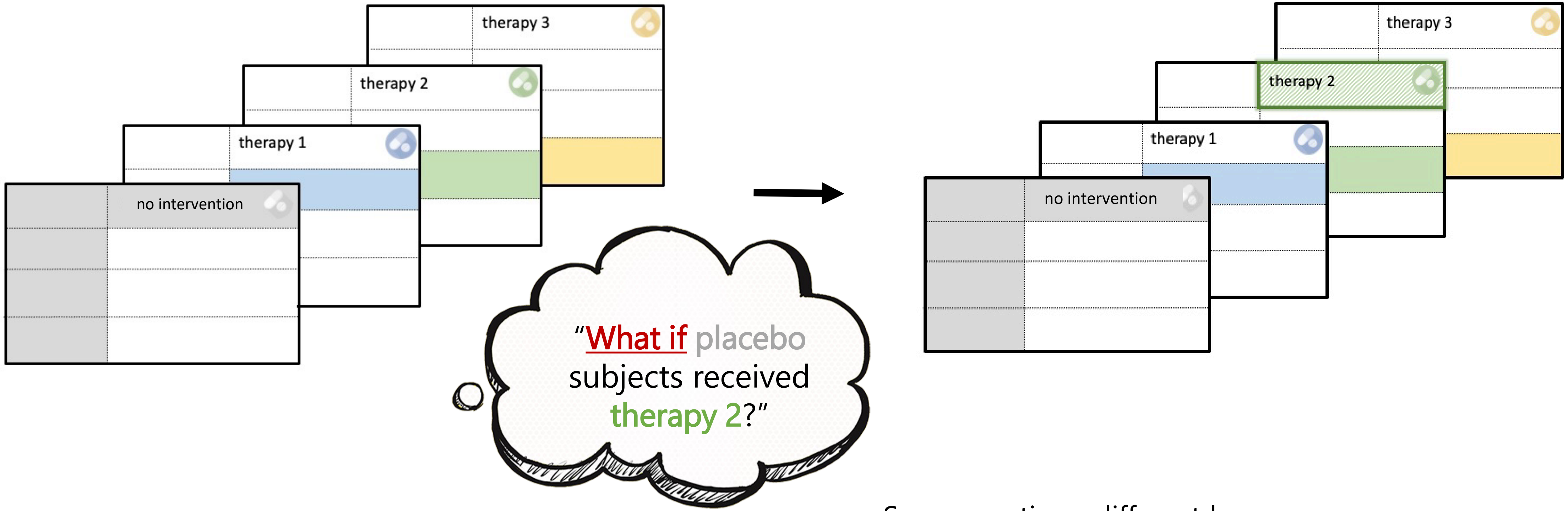
Why? Can only observe Alice under ONE intervention

Limitation of RCTs

What works best on average may not work best for each individual

Randomization but NO personalization

Counterfactual estimation = Tensor Completion



"causal inference is a missing data problem"

vis-à-vis

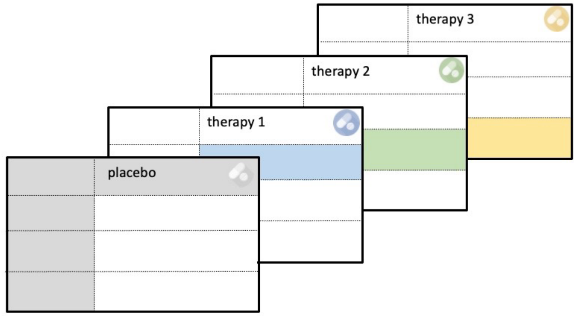
"tensor completion is a missing data problem"

Same questions, different language

Causal Inference	Tensor Completion
causal estimand	error metric (norm)
confounded data	missing not at random data
observational & experimental studies	sparsity patterns

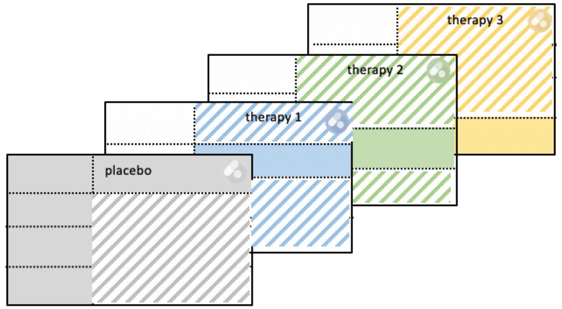
Alzheimer's clinical trial study w. TauRx Therapeutics

A question:
Can we estimate ADAS-COG score for each patient under each therapy?

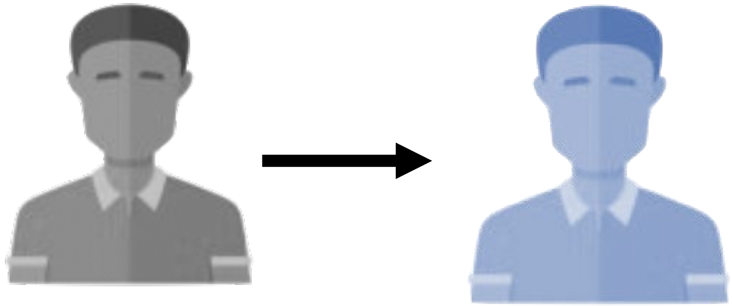


A framework:
Causal inference as tensor completion

An answer:
Synthetic interventions (SI)

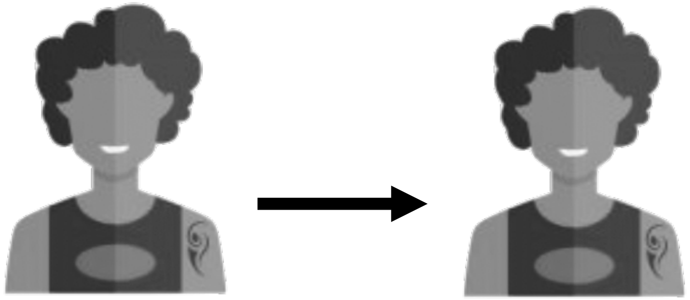


Bob and Alice: counterfactuals of interest



Suppose Bob received therapy 1 after RCT
(under placebo prior to RCT)

What if
Bob remained under placebo?
Bob received therapy 2?
Bob received therapy 3?



Suppose Alice remained under placebo after RCT
(under placebo prior to RCT)

What if
Alice received therapy 1?
Alice received therapy 2?
Alice received therapy 3?

A partial answer: synthetic controls (SC) [Abadie et al ‘03, ‘10]

Estimates counterfactuals in absence of intervention

“What if Bob received therapy?”

“What if Alice remained under placebo?”

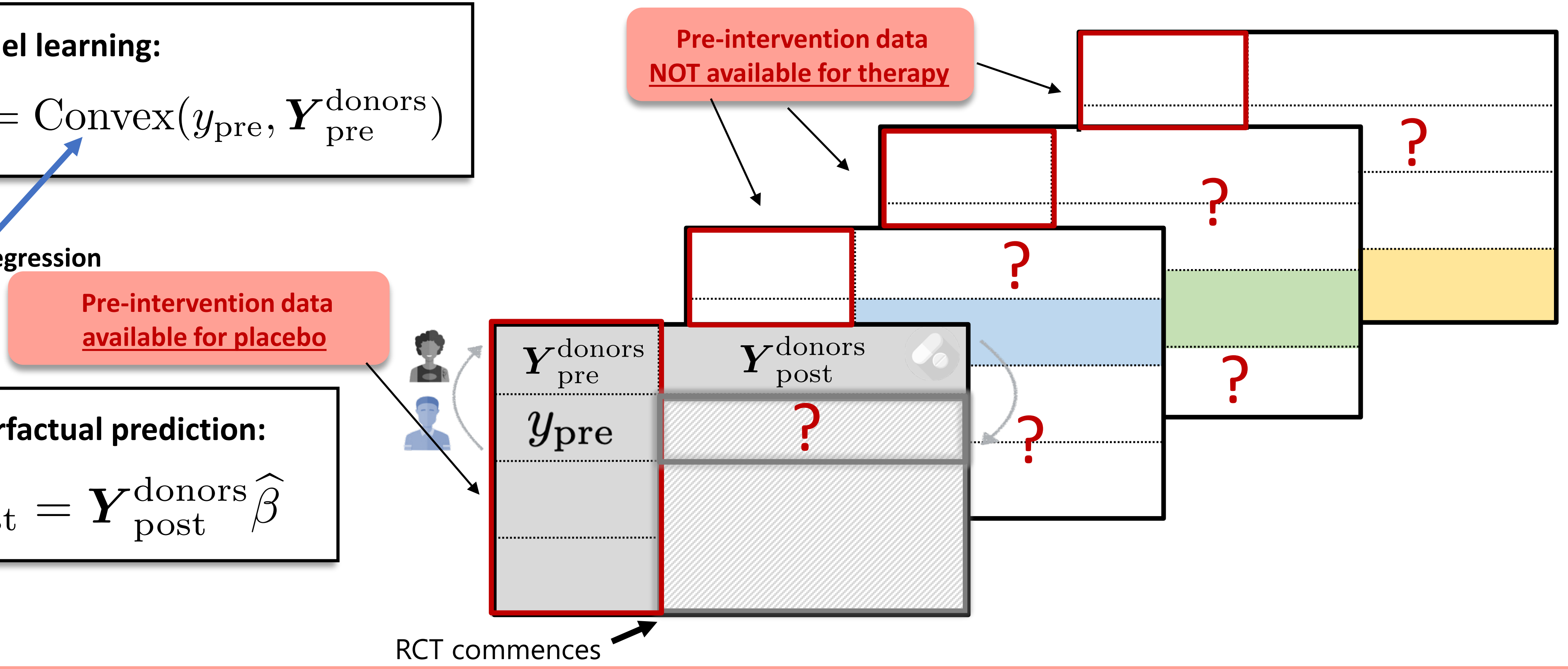
Model learning:

$$\widehat{\beta} = \mathrm{Convex}(y_{\text{pre}}, \boldsymbol{Y}^{\text{donors}}_{\text{pre}})$$

Convex regression

Counterfactual prediction:

$$\widehat{y}_{\text{post}} = \boldsymbol{Y}^{\text{donors}}_{\text{post}}\,\widehat{\beta}$$



“Results on estimation with multiple interventions are absent in the literature” [Abadie’20]

Synthetic controls – widely used but only a partial answer

SC: Absence of intervention

What if Bob remained under placebo?

Estimate counterfactual if policy did not occur:

[Pfizer, Merck, Medidata, Flatiron Health (acquired by Roche for ~$2Bn)….]

- Tax legislation [Abadie et al'10]
- Police reform [Rydberg'18]
- Brexit [Opatrny'19]

⋮

"One of the most important innovations in the policy evaluation literature in the last 15 years"

[Athey and Imbens'16]

Presence of intervention

What if Alice got therapy 1, therapy 2, therapy 3?

Necessary for clinical case study

"Results on estimation with multiple [in presence of] interventions are absent in the literature"

[Abadie'20]

Significant impact of estimating counterfactuals in presence of intervention

A full answer: synthetic interventions [Agarwal, Shah, Shen '21]

Estimates counterfactuals in **absence** & **presence** of intervention

"What if Alice received therapy 1?"

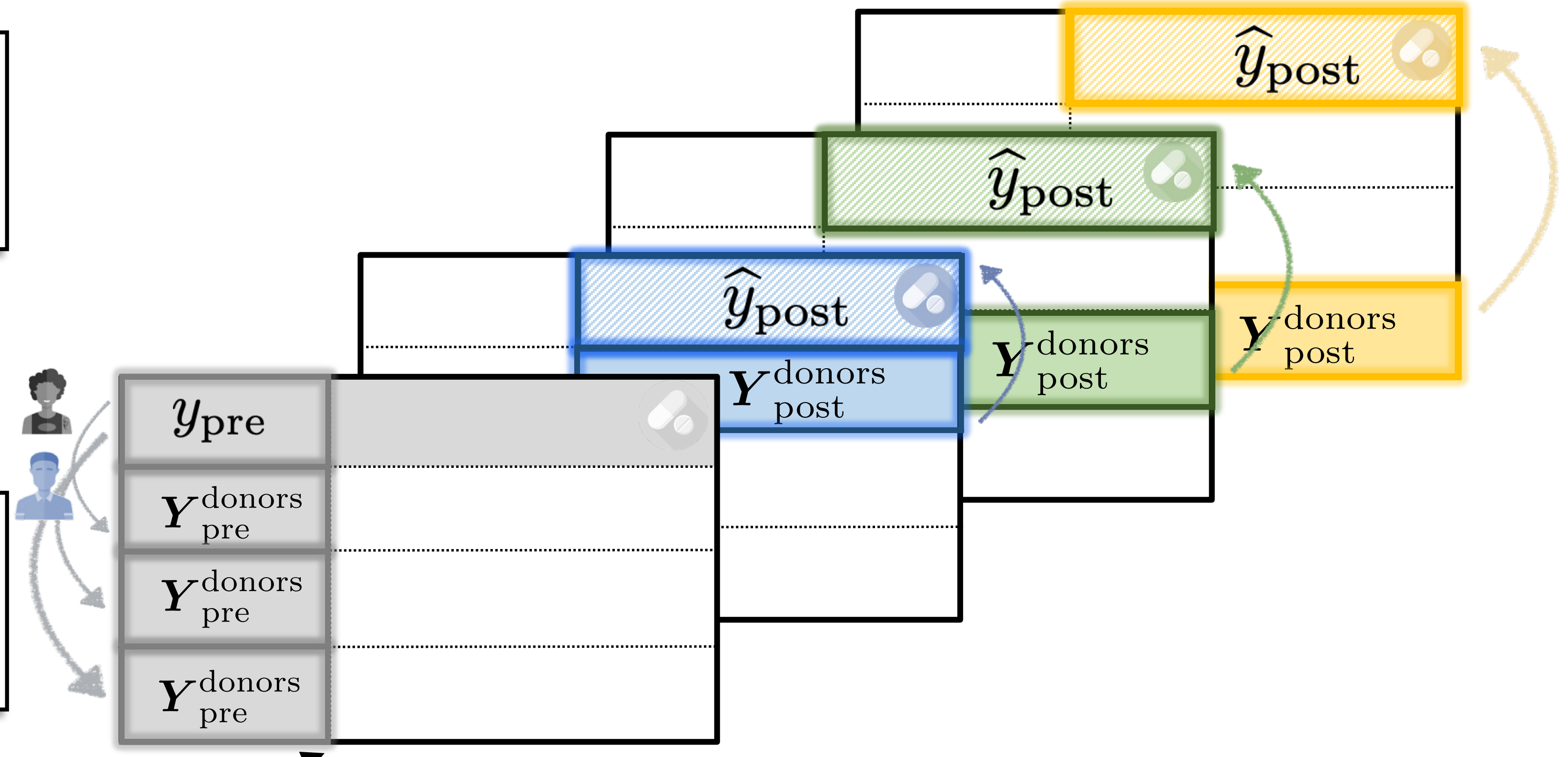
Model learning:

$$\widehat{\beta} = \mathrm{PCR}(y_{\text{pre}}, \boldsymbol{Y}^{\text{donors}}_{\text{pre}})$$

via Principal Component Regression

Counterfactual prediction:

$$\widehat{y}_{\text{post}} = \boldsymbol{Y}^{\text{donors}}_{\text{post}} \widehat{\beta}$$



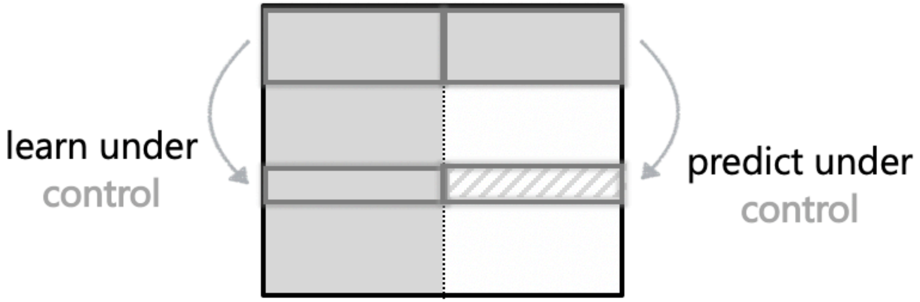
RCT commences

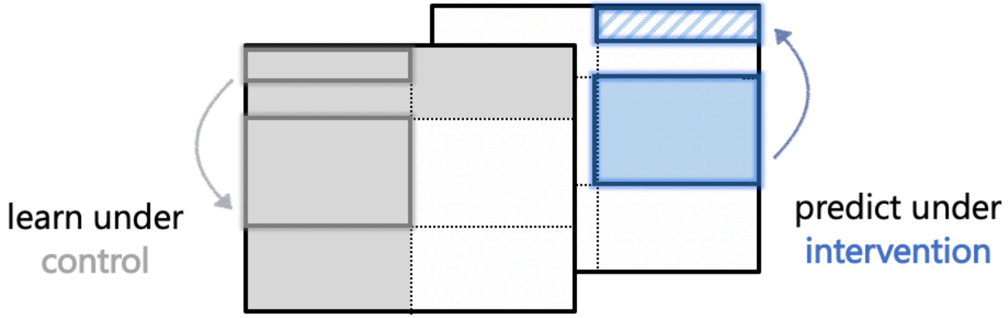
NO additional data over RCT data used!

Synthetic controls (SC)

Synthetic interventions (SI)

1. Where model is applied





When can a linear model be transferred across interventions?
i.e., transfer learning, distribution shift, causal transportability…

2. How model is learned

Convex regression

$$\widehat{\beta} = \text{Convex}(y_{\text{pre}}, \boldsymbol{Y}_{\text{pre}}^{\text{donors}})$$

Principal component regression (PCR)

$$\widehat{\beta} = \text{PCR}(y_{\text{pre}}, \boldsymbol{Y}_{\text{pre}}^{\text{donors}})$$

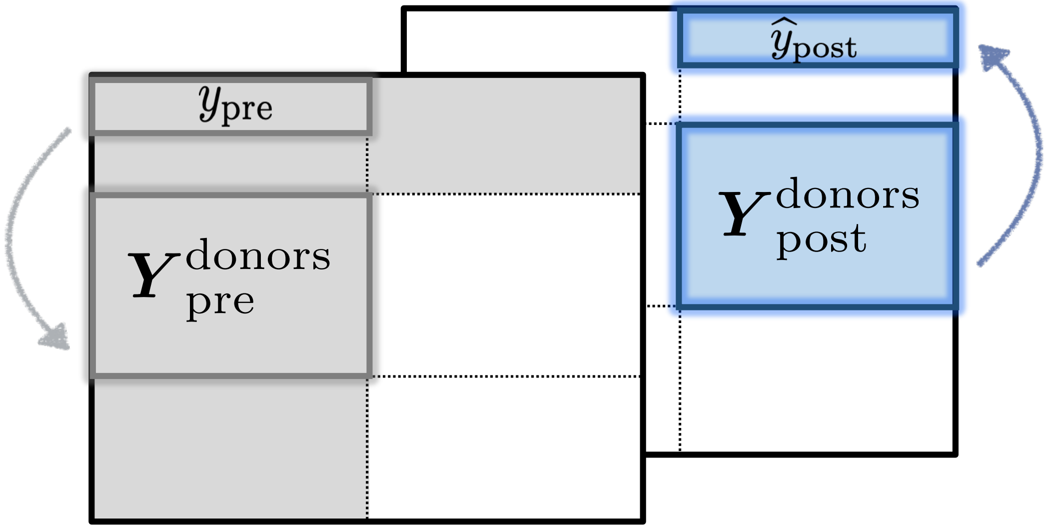
PCR is crucial to proving our formal statistical guarantees

When does synthetic interventions work?

Causal framework, statistical guarantees

Essential questions

learn under
control

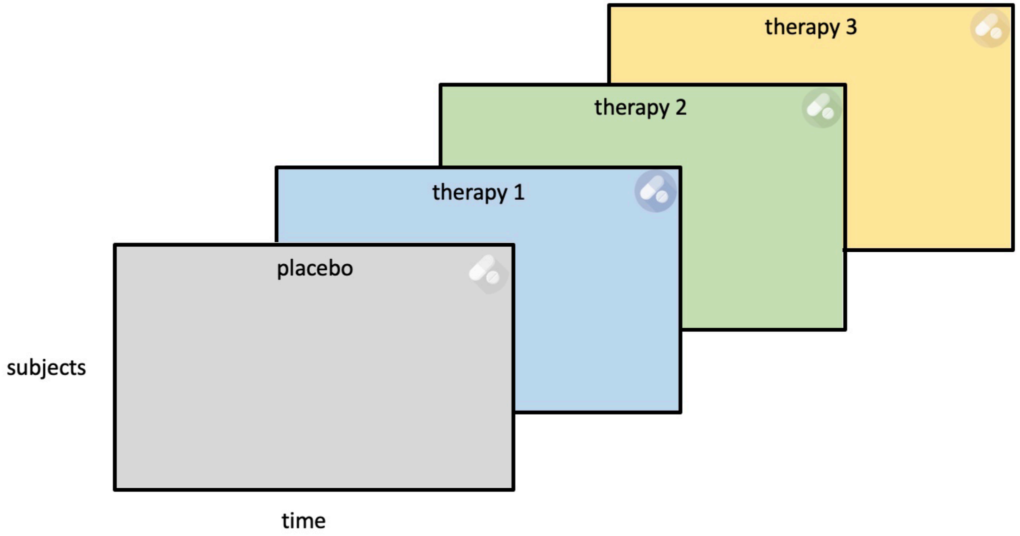


predict under
intervention

When can a linear model be transferred between different interventions?

What type of confounding is allowed in observational data?

Why linear?—*low rank tensor*



low rank

time latent factors

$$\boldsymbol{Y} = \underbrace{\sum_{\ell=1}^{r} u_\ell \otimes v_\ell \otimes w_\ell}_{\mathbb{E}[\boldsymbol{Y}]} + \boldsymbol{\varepsilon}$$

stochasticity

unit latent factors

intervention latent factors

Low rank implies linear span inclusion

Produce counterfactuals for

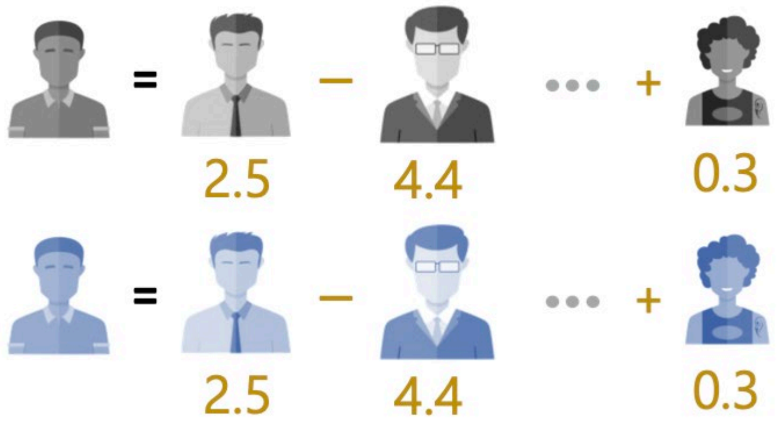
unit n under **intervention d**

We require:

$$\boxed{u_n = \sum_{j \in \mathcal{I}^{(d)}} \beta_j^* \cdot u_j}$$

$\mathcal{I}^{(d)} =$ units under intervention d

Holds w.h.p if factors sampled independently



Exists **<u>invariant</u>** linear model across time, interventions

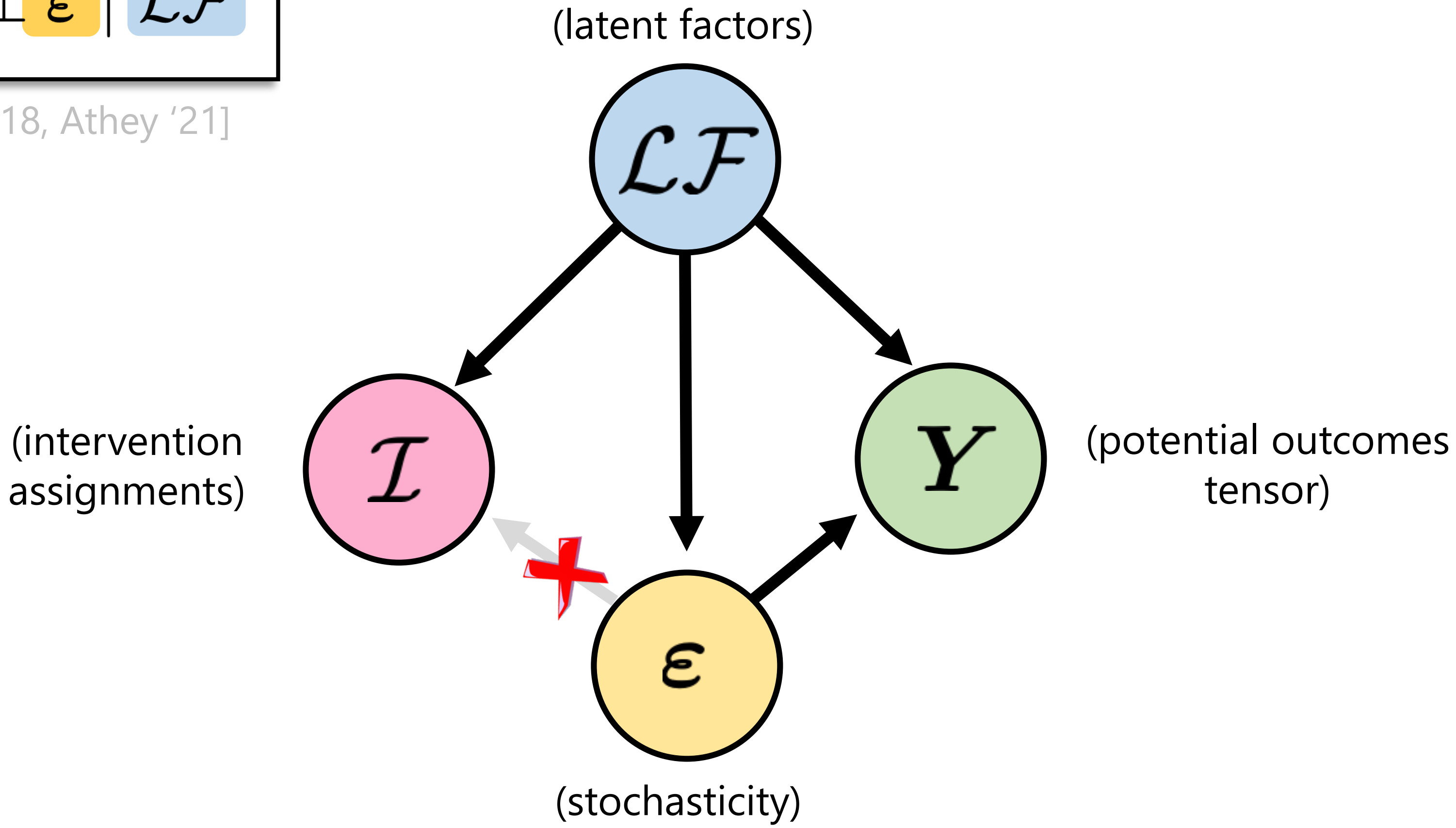
What type of confounding?—*selection on latent factors*

$$\boldsymbol{Y} = \sum_{\ell=1}^{r} u_\ell \otimes v_\ell \otimes w_\ell + \boldsymbol{\varepsilon}$$

$$+ \boxed{\mathcal{I} \perp\!\!\!\perp \boldsymbol{\varepsilon} \mid \mathcal{LF}}$$

[Kallus '18, Athey '21]

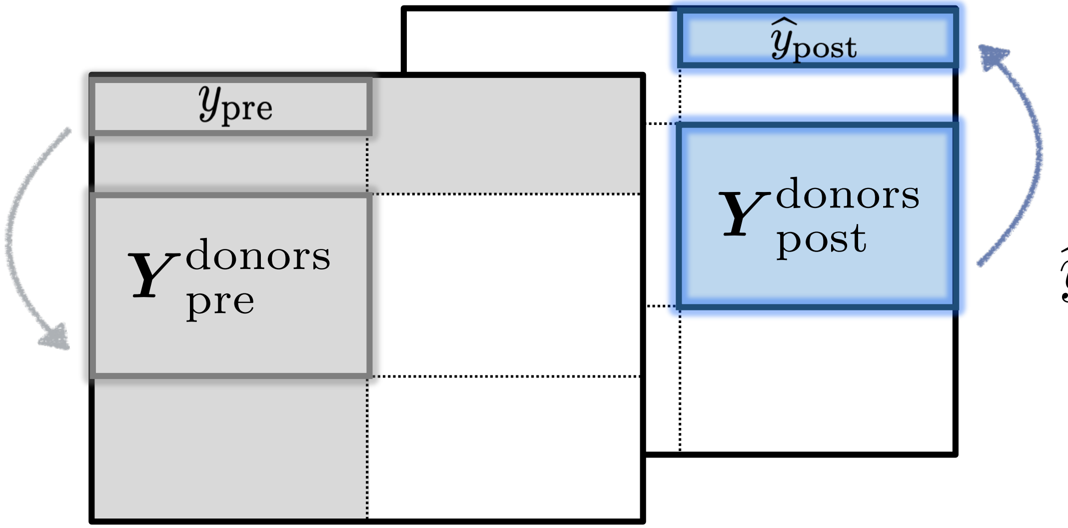
$$\Longrightarrow \quad \boxed{\mathcal{I} \perp\!\!\!\perp \boldsymbol{Y} \mid \mathcal{LF}}$$



Identification

Model learning:

$$\widehat{\beta} = \mathrm{PCR}(y_{\mathrm{pre}}, \boldsymbol{Y}^{\mathrm{donors}}_{\mathrm{pre}})$$

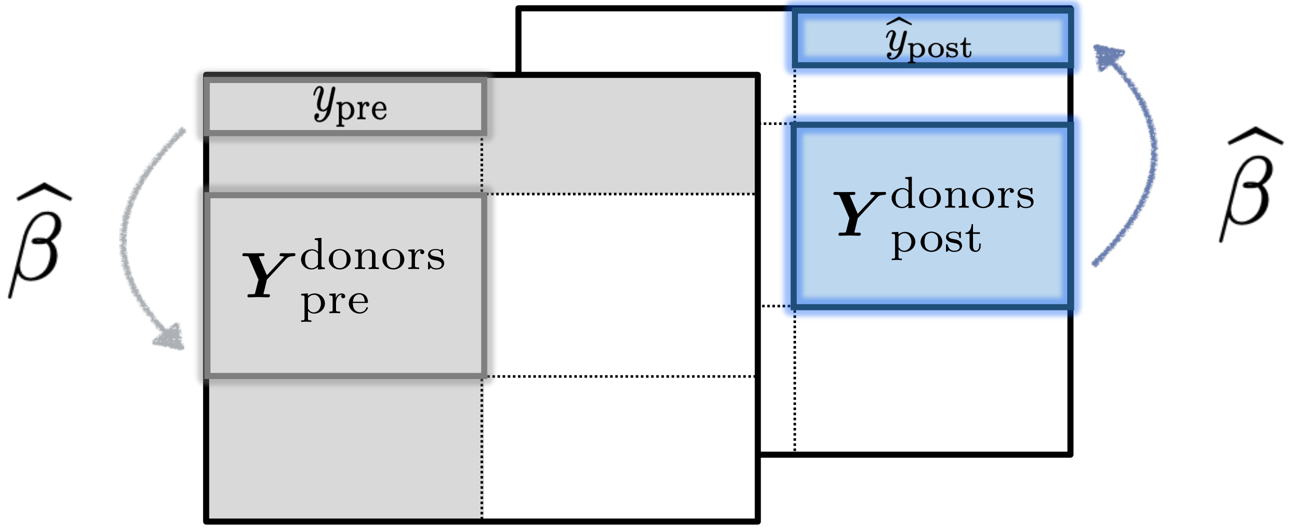


Counterfactual prediction:

$$\widehat{y}_{\mathrm{post}} = \boldsymbol{Y}^{\mathrm{donors}}_{\mathrm{post}} \widehat{\beta}$$

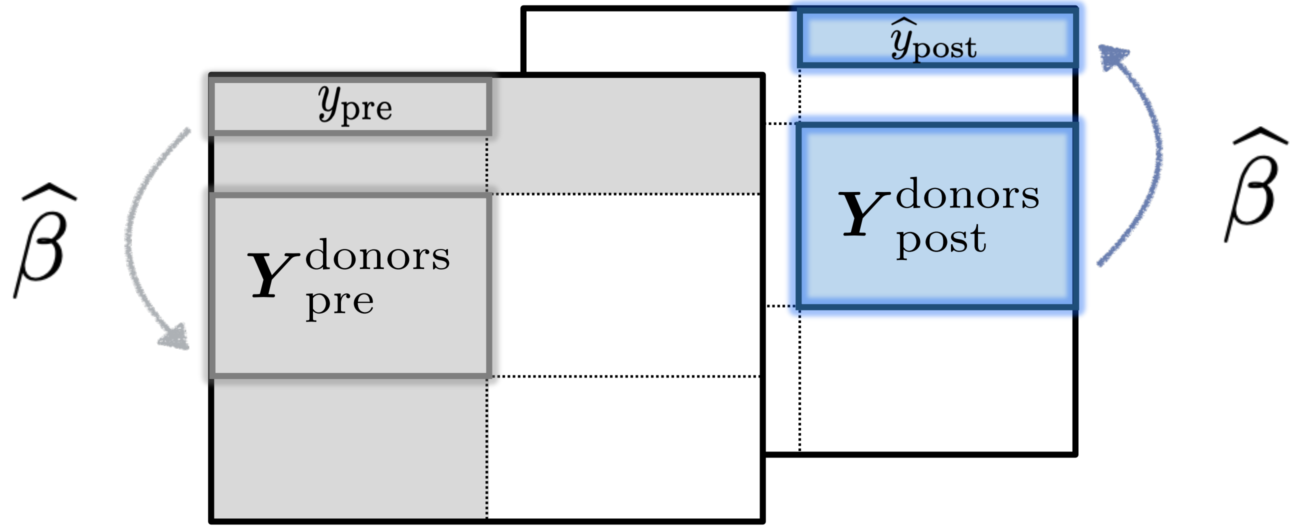
$$\boxed{\boxed{\mathbb{E}[y_{\mathrm{post}}]} = \mathbb{E}[\, \boxed{\boldsymbol{X}_{\mathrm{post}}} \,] \cdot \beta^{*}}$$

When transferrable?—*subspace inclusion*



$$\text{complexity}\Big(\underbrace{\boxed{\boldsymbol{X}_{\text{post}}}}_{\text{“test” set}}\Big) \leq \text{complexity}\Big(\underbrace{\boxed{\boldsymbol{X}_{\text{pre}}}}_{\text{“train” set}}\Big)$$

When transferrable?—*subspace inclusion*



$$\operatorname{colspan}(\boldsymbol{X}_{\text{post}}) \subseteq \operatorname{colspan}(\boldsymbol{X}_{\text{pre}})$$

"test" set

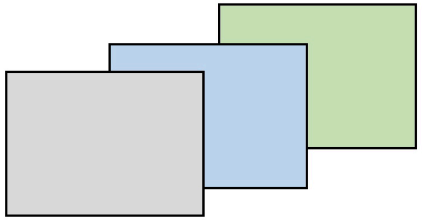
"train" set

hypothesis test

Putting it all together

When does SI work?

Low-rank tensor factor model



+

linear span inclusion

$$u_{n\ell} = \sum_{j \in \mathcal{I}^{(d)}} \beta_j^* \cdot u_{j\ell}$$

+

selection on latent factors

$$Y_{nt}^{(d)} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{LF}$$

+

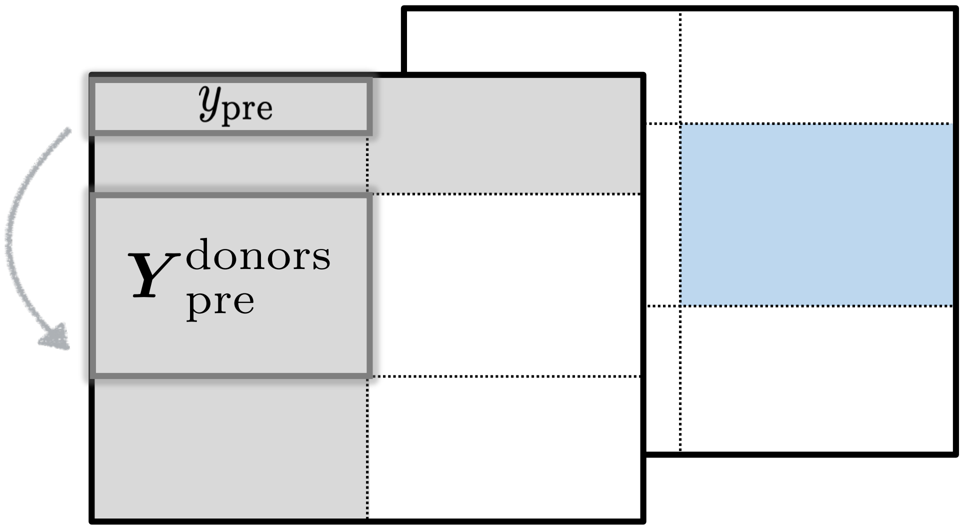
subspace inclusion

$$\text{colspan}(\mathbb{E}[\boldsymbol{Y}_{\text{pre}}^{\text{donors}}]) \supseteq \text{colspan}(\mathbb{E}[\boldsymbol{Y}_{\text{post}}^{\text{donors}}])$$

Model estimation

Model learning:

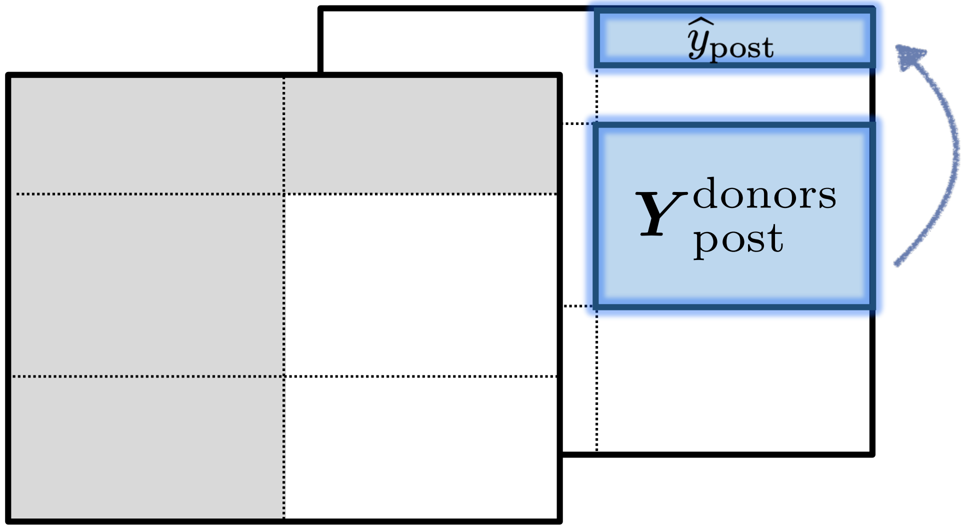
$$\widehat{\beta} = \mathrm{PCR}(y_{\text{pre}}, \boldsymbol{Y}_{\text{pre}}^{\text{donors}})$$



$$\boxed{\|\widehat{\beta} - \beta^*\|_2 = o(1)}$$

minimum norm model

Consistency

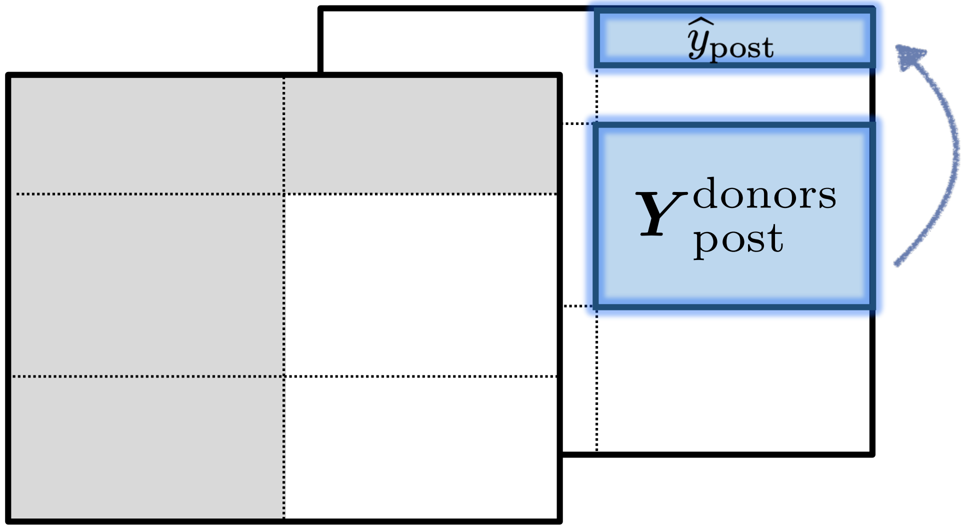


Counterfactual prediction:

$$\widehat{y}_{\text{post}} = \boldsymbol{Y}^{\text{donors}}_{\text{post}} \widehat{\beta}$$

$$\left|\text{avg}\left(\boxed{\widehat{y}_{\text{post}}}\right) - \text{avg}\left(\boxed{\mathbb{E}[y_{\text{post}}]}\right)\right| = o(1)$$

Asymptotic normality



Counterfactual prediction:

$$\widehat{y}_{\text{post}} = \boldsymbol{Y}^{\text{donors}}_{\text{post}} \widehat{\beta}$$

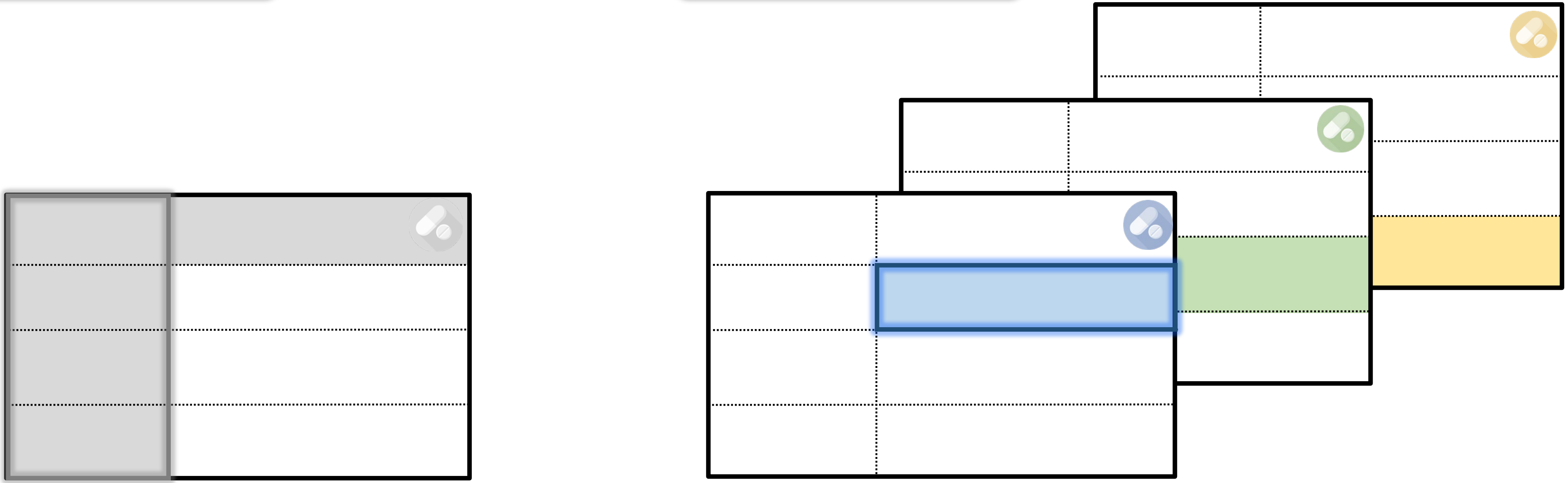
$$\text{avg}(\boxed{\widehat{y}_{\text{post}}}) \sim \mathcal{N}\left(\text{avg}(\boxed{\mathbb{E}[y_{\text{post}}]}), \sigma^2(\beta^*)\right)$$

enables confidence intervals

Validation study setup

Training data

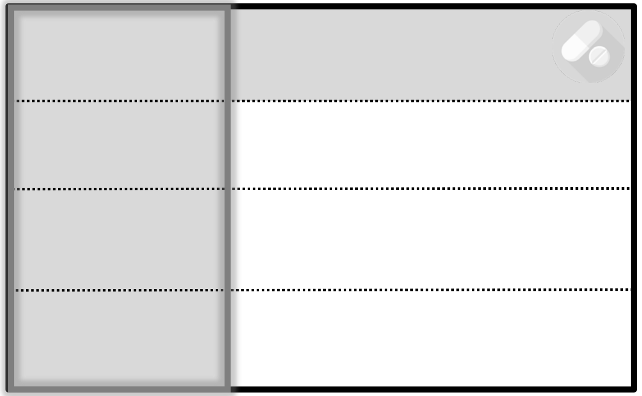
Test data

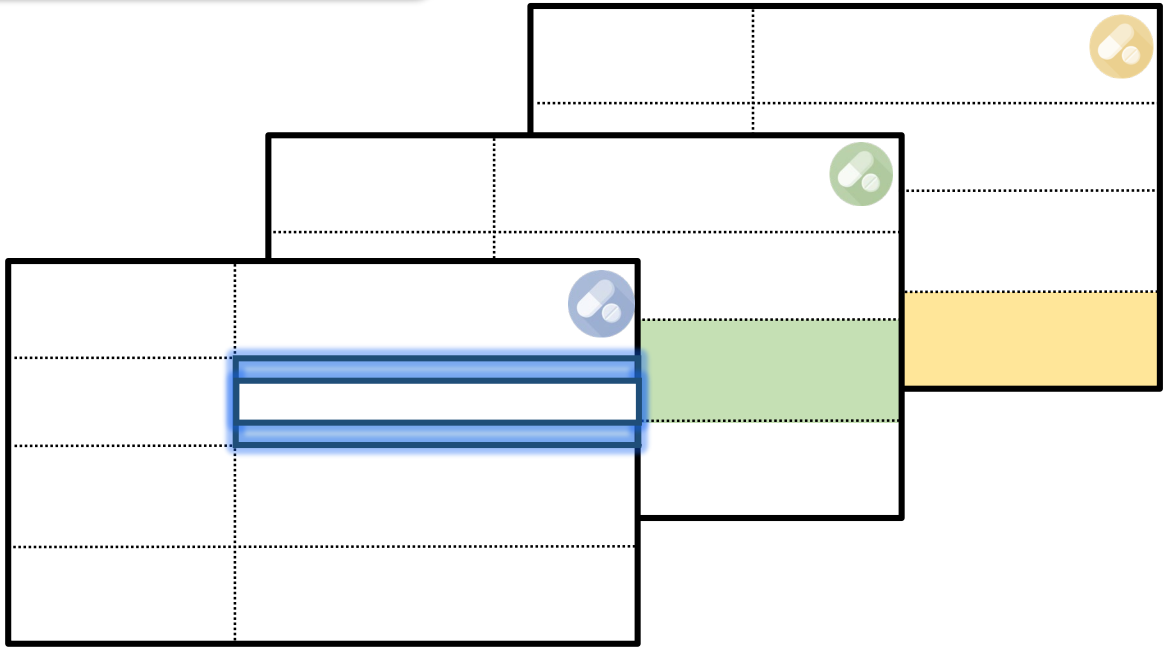


Validation study setup

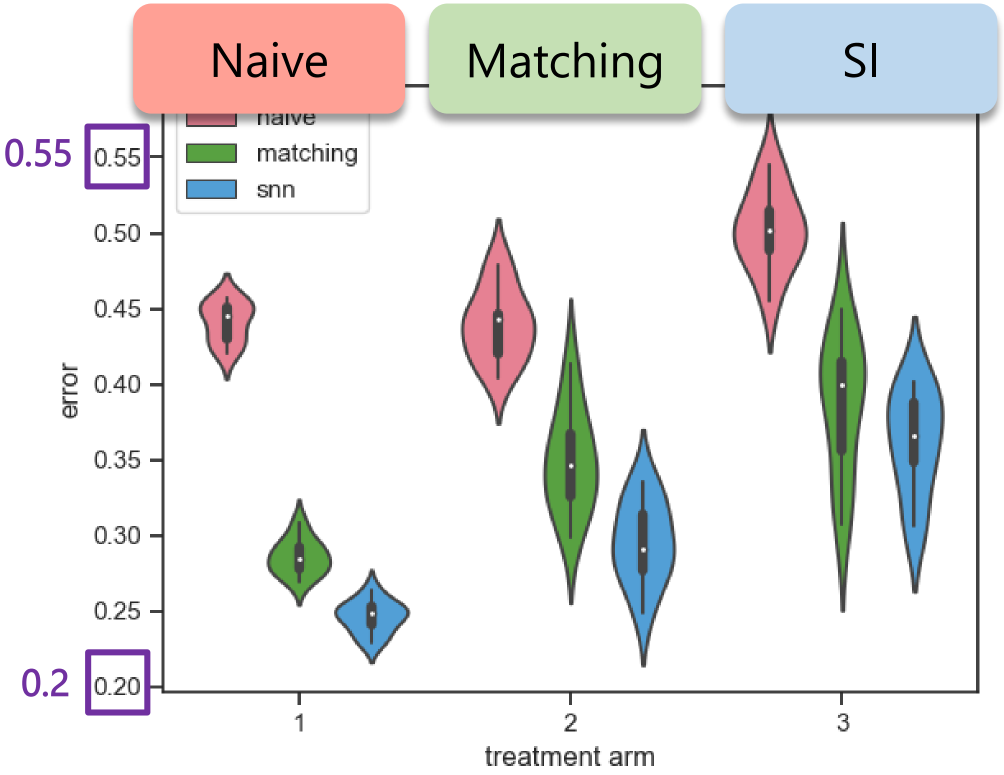
Training data

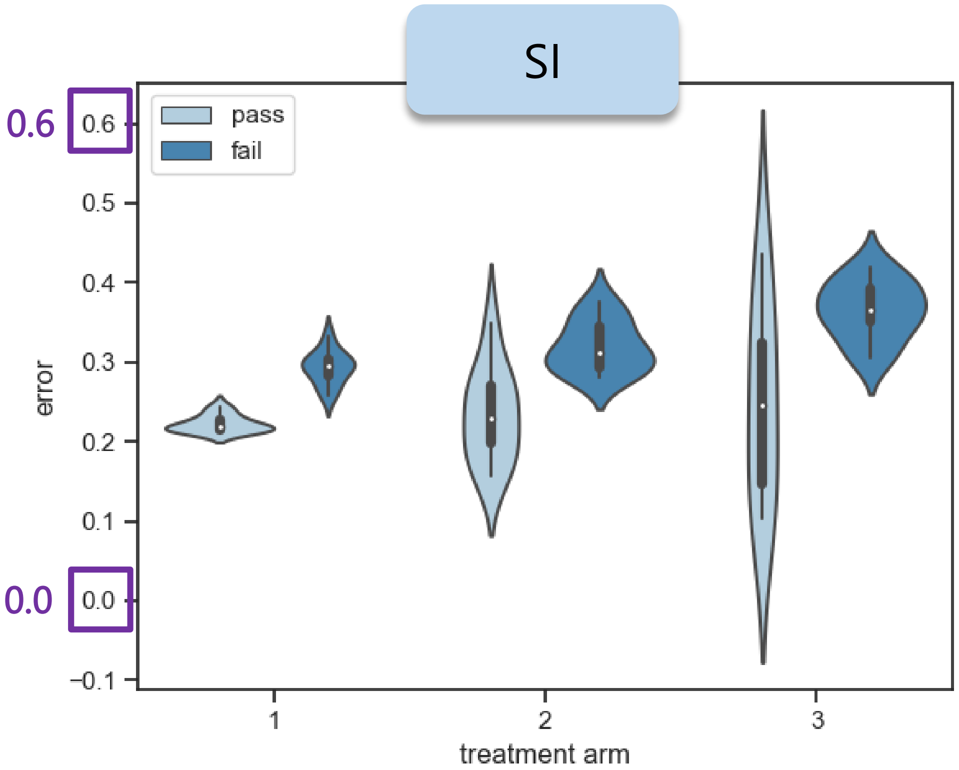
Test data



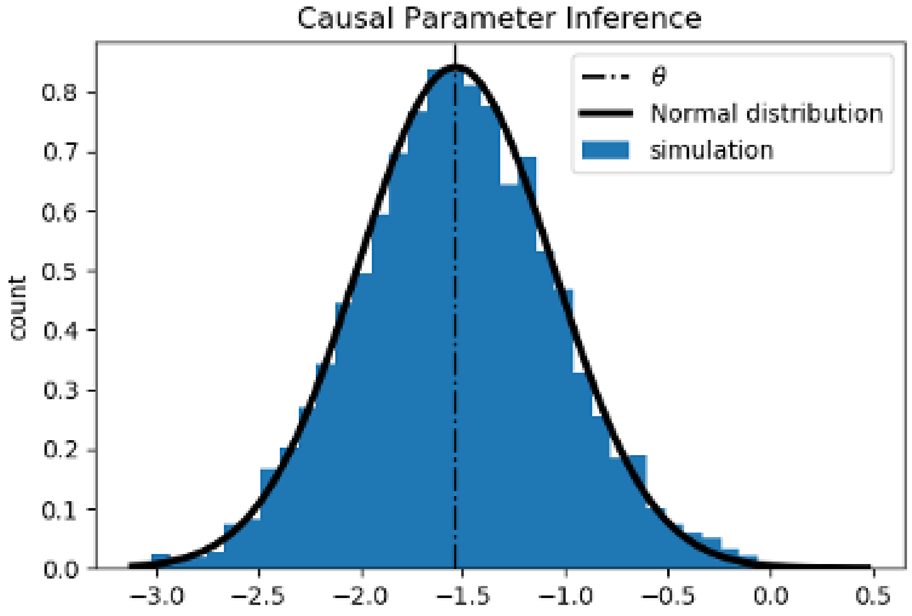


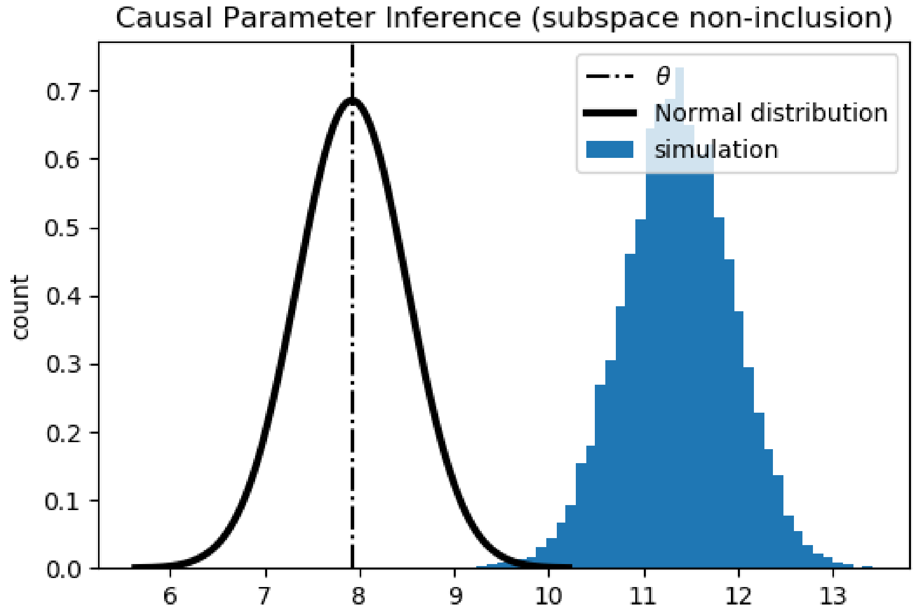
Validation study results





Importance of subspace inclusion





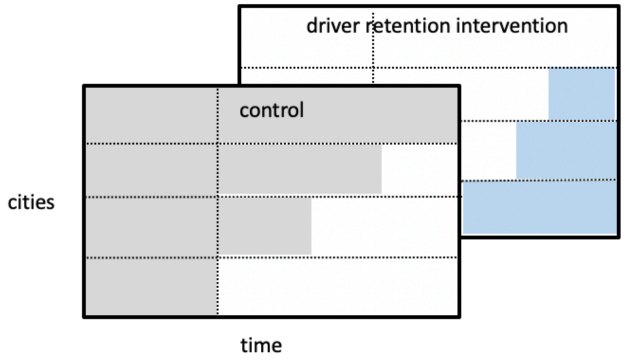
- Train and test data obey different distributions
- Subspace inclusion holds

- Train and test data obey same distribution
- Subspace inclusion fails

Additional case studies

Ride-sharing

Question:
Passenger wait time if U.S. cities adopted routing policies used in Brazil?



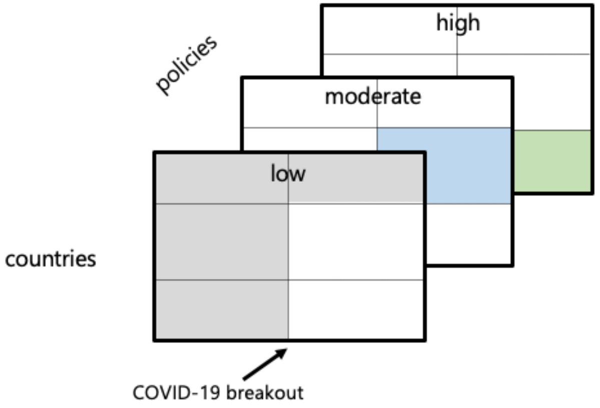
SI

Estimated answer:
Larger U.S. cities would see greatest increase in wait time

observational

COVID-19 policy evaluation

Question:
U.S. COVID-19 mortality rates if stricter social distancing adopted?



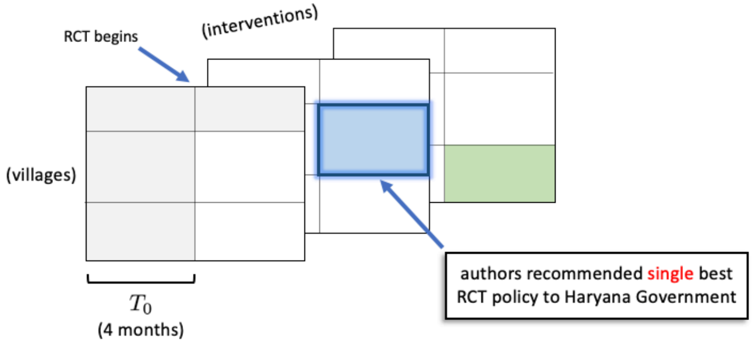
SI

Estimated answer:
> 5x drop in mortality rates if mobility restriction increased by approx. 30%

observational

Poverty Action Lab @ MIT

Question:
Uptake in childhood immunization rates if personalized policies used across villages in India ?

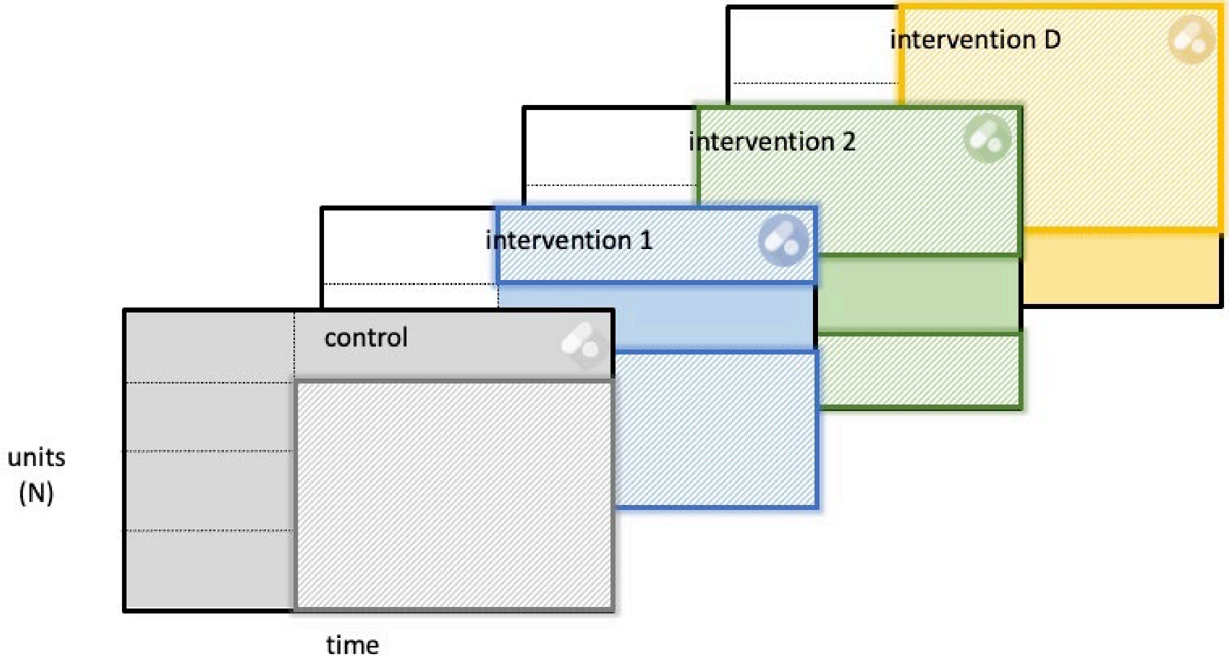


SI

Estimated answer:
~6x uptake in immunization if policy personalized to each village

limited experimental

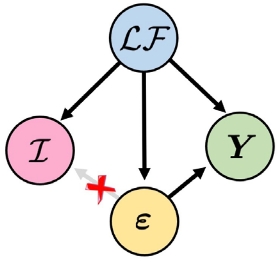
SI—key takeaway



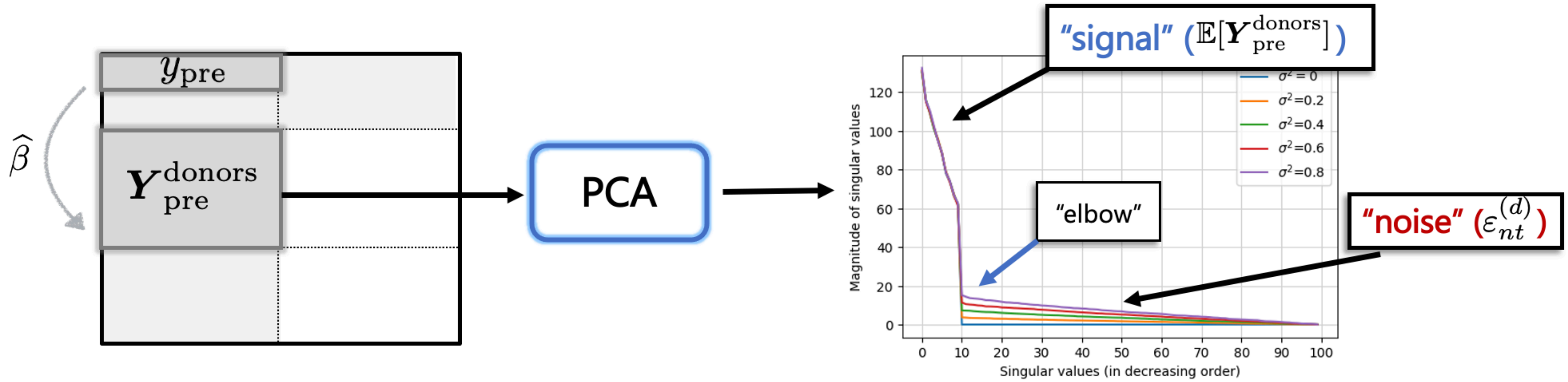
provably learns all N × D causal estimands with

(i) N × 2 observations (requires meas. under common intervention)

(ii) confounded data that respects *selection on latent factors*



Why PCR?



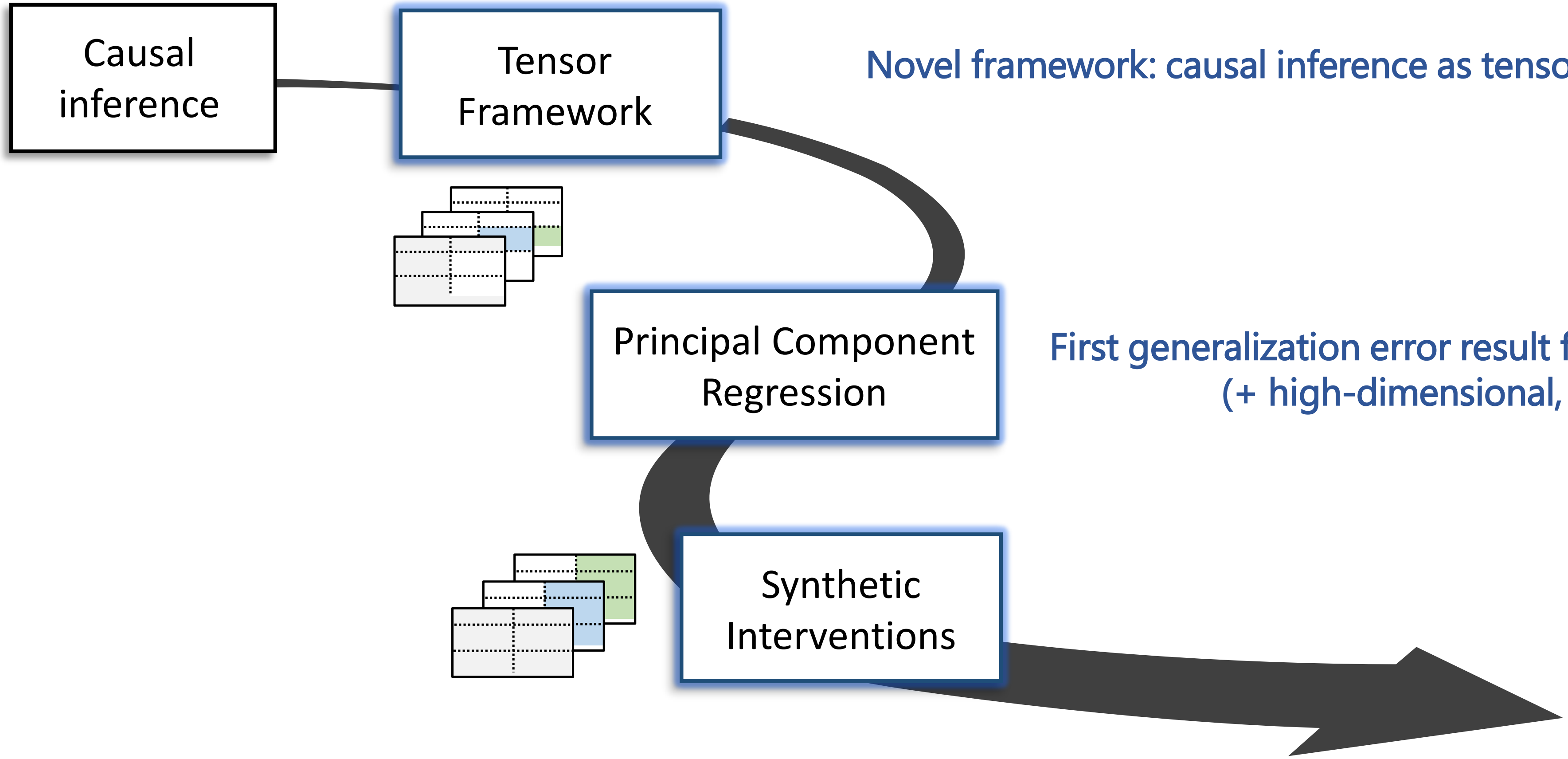
Intuition:

Low-rank signal is spectrally concentrated, noise is spectrally diffused

1. PCR enforces spectral sparsity in $\boldsymbol{Y}^{\text{donors}}_{\text{pre}}$

2. PCR implicitly de-noises $\boldsymbol{Y}^{\text{donors}}_{\text{pre}}$

Summary



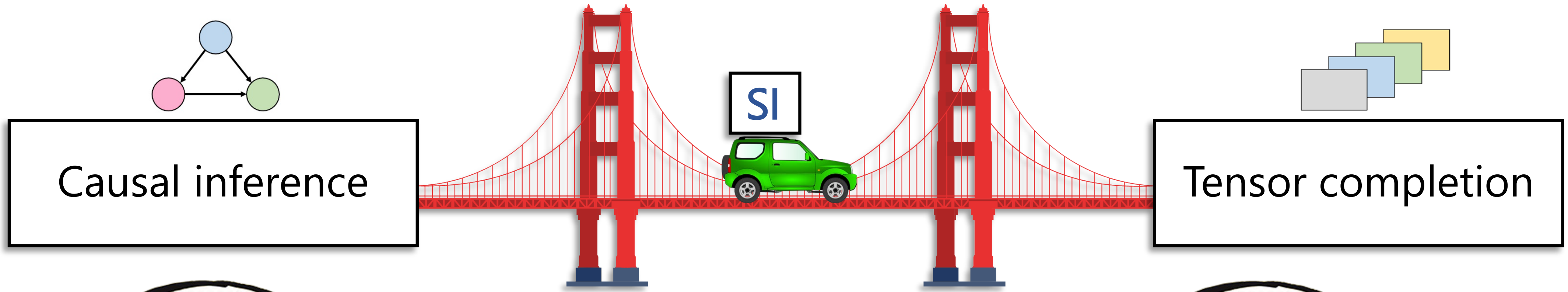
Novel framework: causal inference as tensor completion

First generalization error result for error-in-variables regression
(+ high-dimensional, fixed design setting)

Addressed unresolved problem of counterfactuals in presence of interventions
[Abadie'20]

Novel results for synthetic controls literature itself

Causal tensor completion—looking forward



modeling

What *(estimand, confounding, study)* combinations allow identification?

modeling

What *(error metric, sparsity pattern)* combinations allow completion?

algorithmic

If achievable, what are the *computational/statistical trade-offs*?

THANK YOU

dshen24@berkeley.edu

Acknowledgements:

Alberto Abadie, Abdullah Alomar, Romain Cosson, Esther Duflo, Anna Mikusheva, Rahul Singh