Synthetic Interventions

Dennis Shen

Joint work with

Anish Agarwal

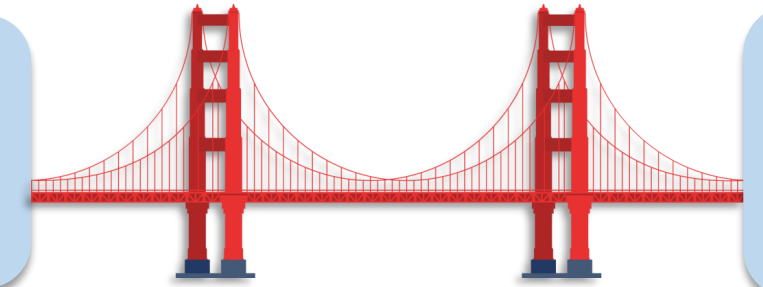
Devavrat Shah

Bridging causal inference & machine learning

Synthetic controls

(what if intervention did not occur?)

Core idea in econometrics



Matrix/tensor completion

(impute missing data in a matrix/tensor)

Core idea in EE/CS/Stats/ML

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

Is there a subpopulation of responders?

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

A question:

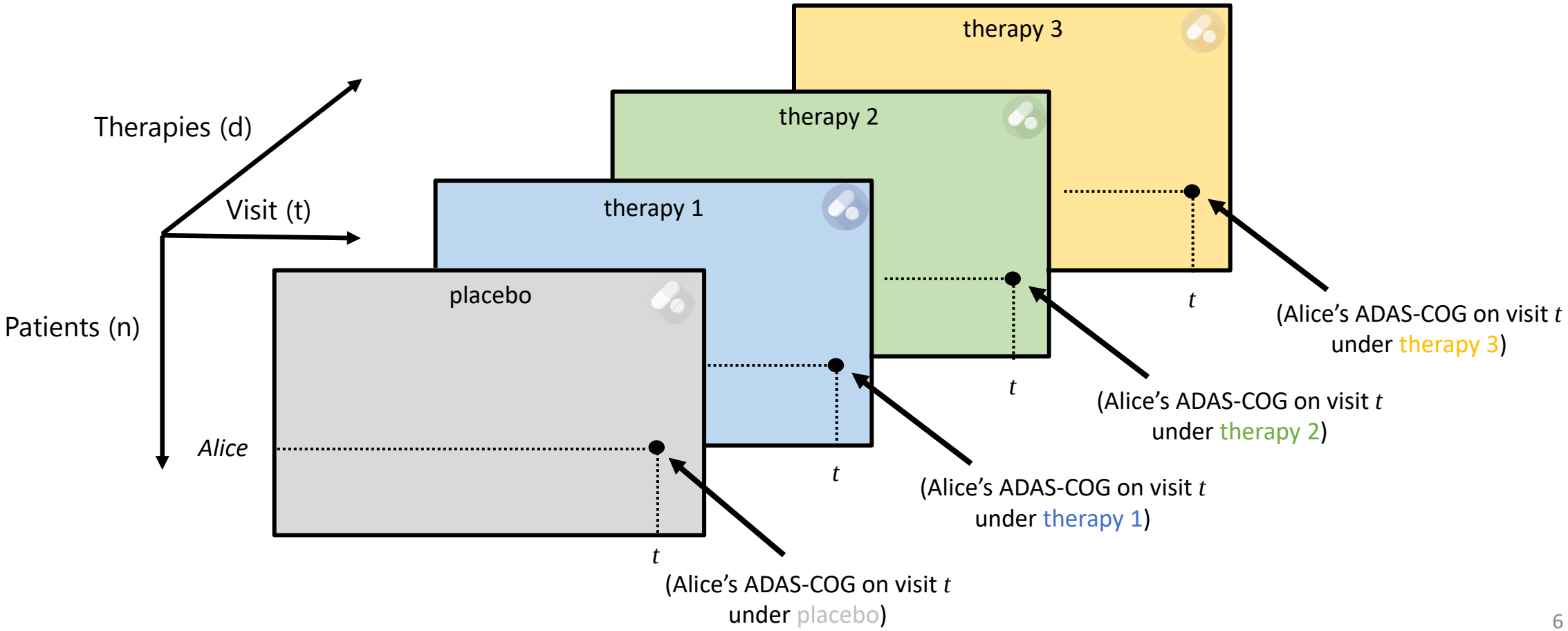
Can we estimate ADAS-COG for each patient under each therapy?

A framework :

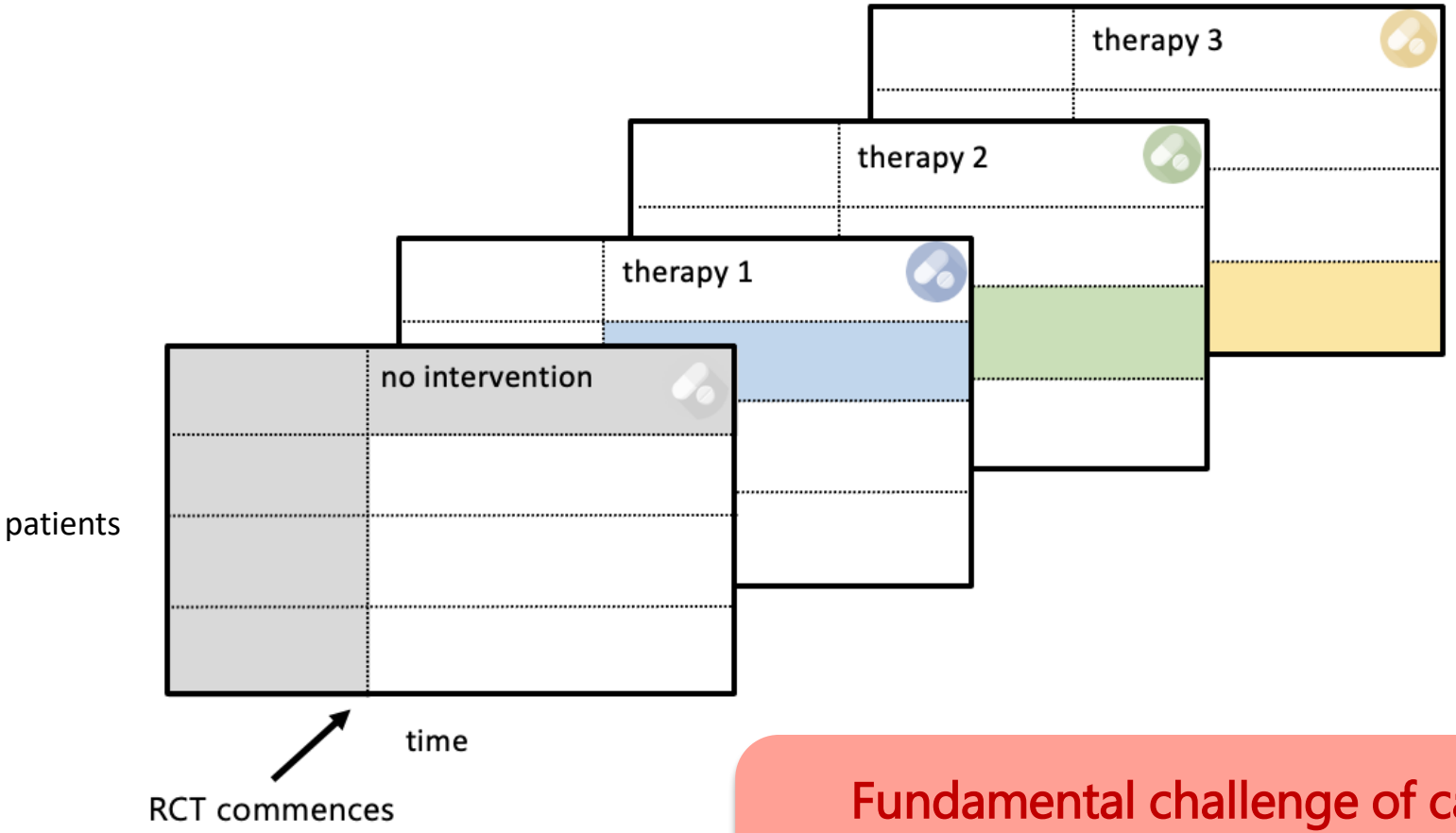
Causal inference as tensor completion

Potential outcomes tensor

$Y_{nt}^{(d)}$ = outcome if patient **n** on visit **t** was given therapy **d**



Observed tensor



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

RCTs—no confounding but no personalization

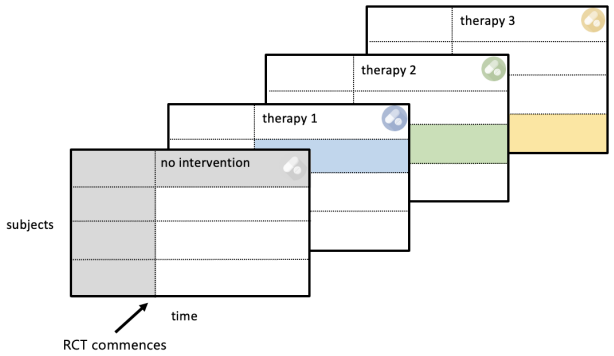
What RCTs can estimate

Average treatment effect

Avg( ) therapy 1

−

Avg( ) placebo



Why are RCTs beloved?

Explicit randomization

What RCTs cannot estimate

Individual treatment effect

Alice under therapy 1

−

Alice under placebo

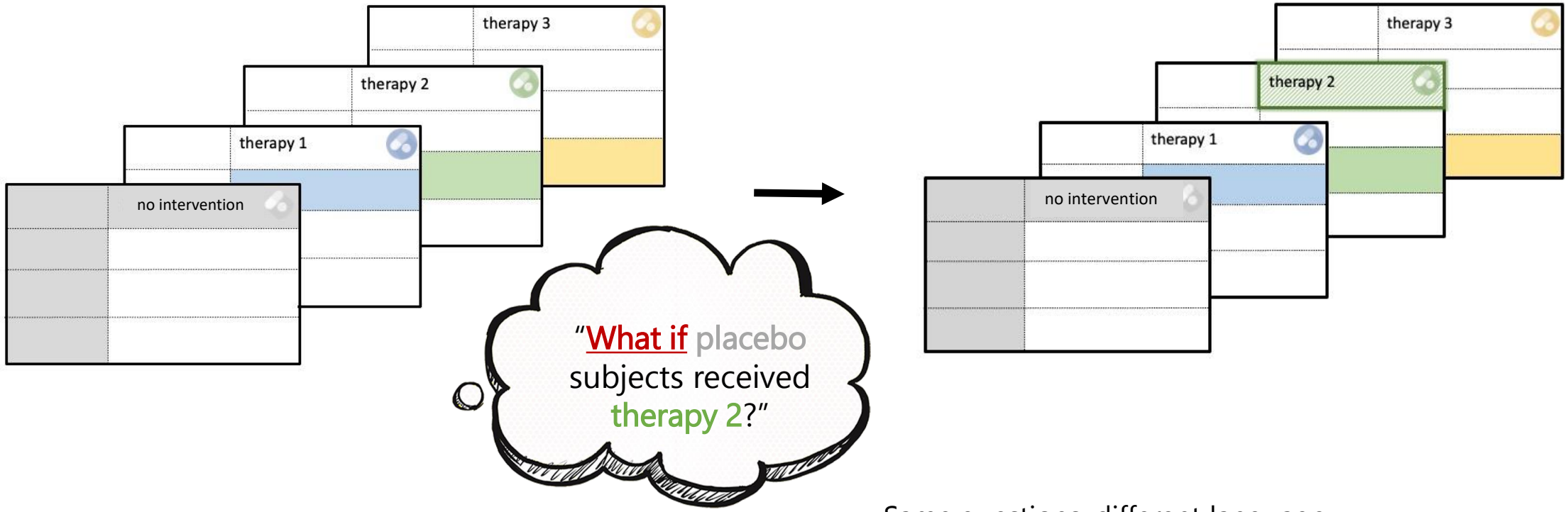
Why? Can only observe Alice under ONE treatment

Limitation of RCTs

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

Randomization but NO personalization

Counterfactual estimation = Tensor Completion



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

"causal inference is a missing data problem"

vis-à-vis

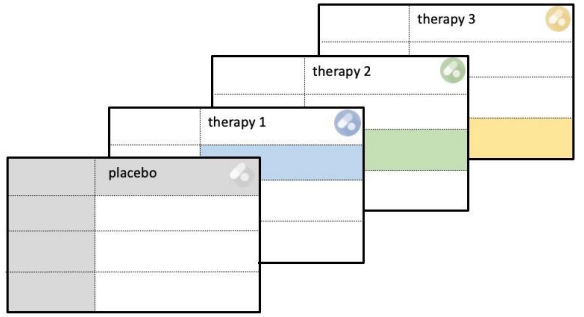
"tensor completion is a missing data problem"

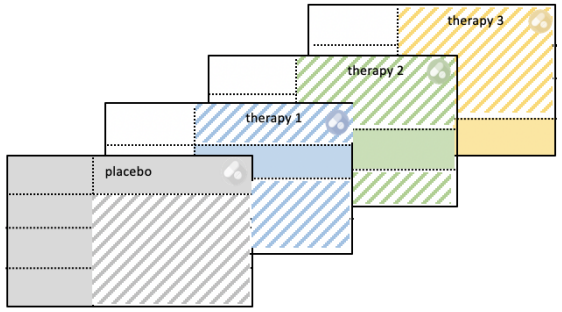
Alzheimer's clinical trial study

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

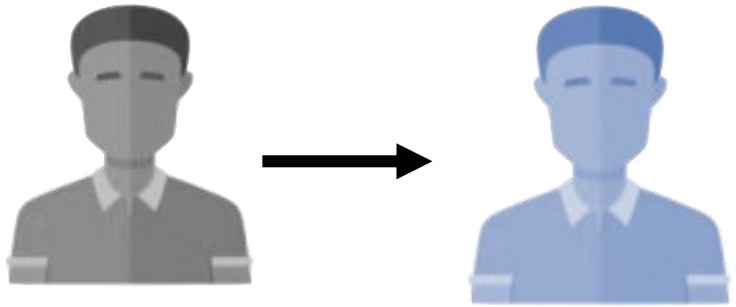
A framework:
Causal inference as tensor completion

An answer:
Synthetic interventions (SI)



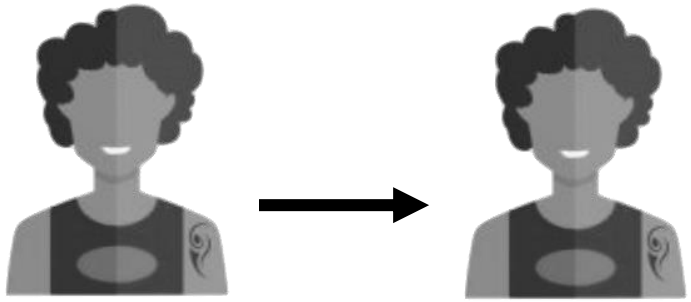


Counterfactual questions of interest



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?"

Model learning:

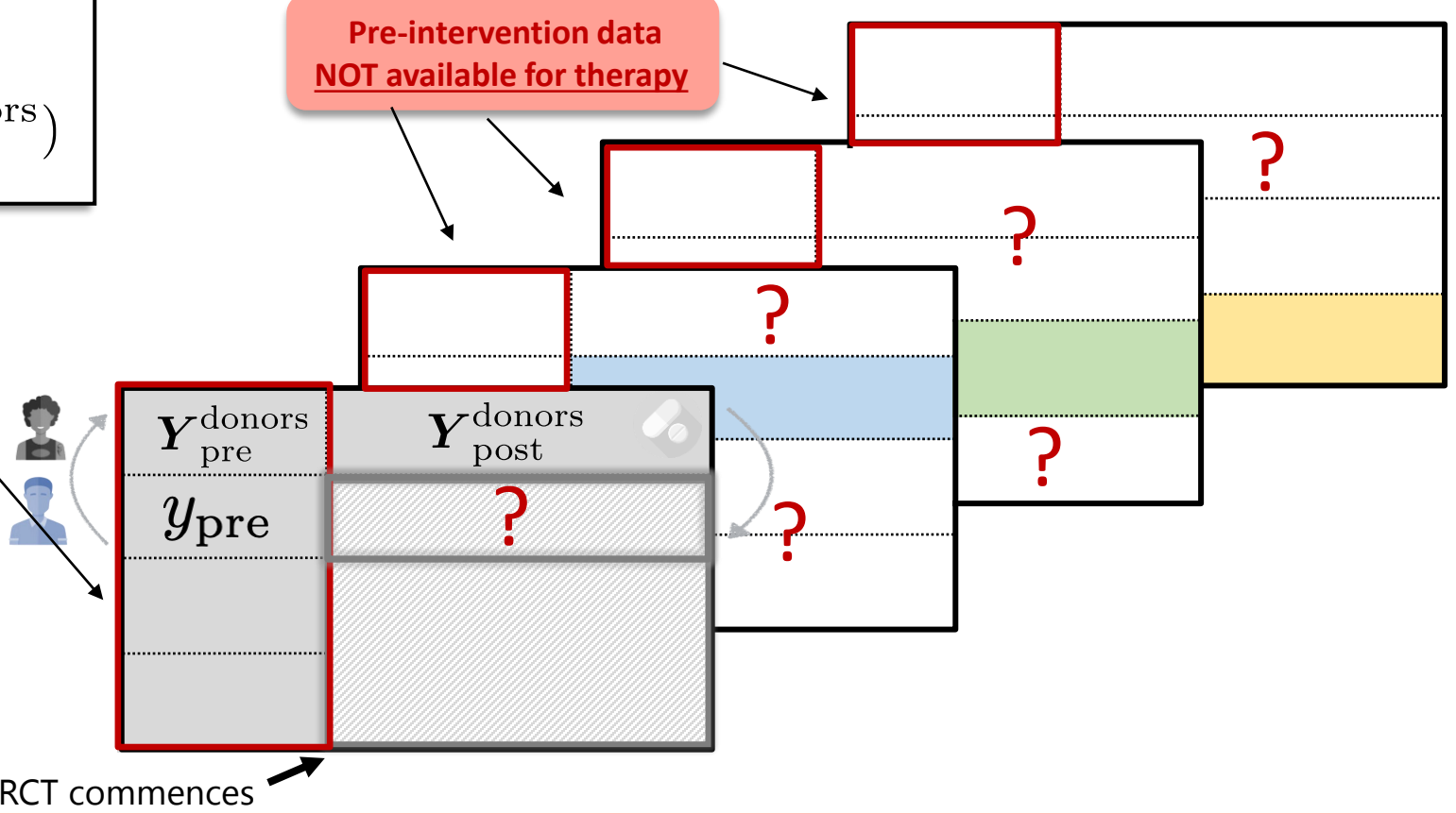
$$\widehat{\beta} = \text{Convex}(y_{\text{pre}}, \boldsymbol{Y}_{\text{pre}}^{\text{donors}})$$

Counterfactual prediction:

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

Pre-intervention data NOT available for therapy

Pre-intervention data available for placebo



"Results on estimation with multiple interventions are absent in the literature" [Abadie'20]

Synthetic controls—a partial answer

SC: Absence of intervention

What if Bob remained under placebo?

Estimate counterfactual if policy did not occur:

- [Pfizer-Wyeth, 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 trial study

“Results on estimation with multiple interventions are absent in the literature”

[Abadie‘20]

Significant impact of estimating counterfactuals in presence of intervention

Synthetic interventions

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

"What if Alice received therapy 3?"

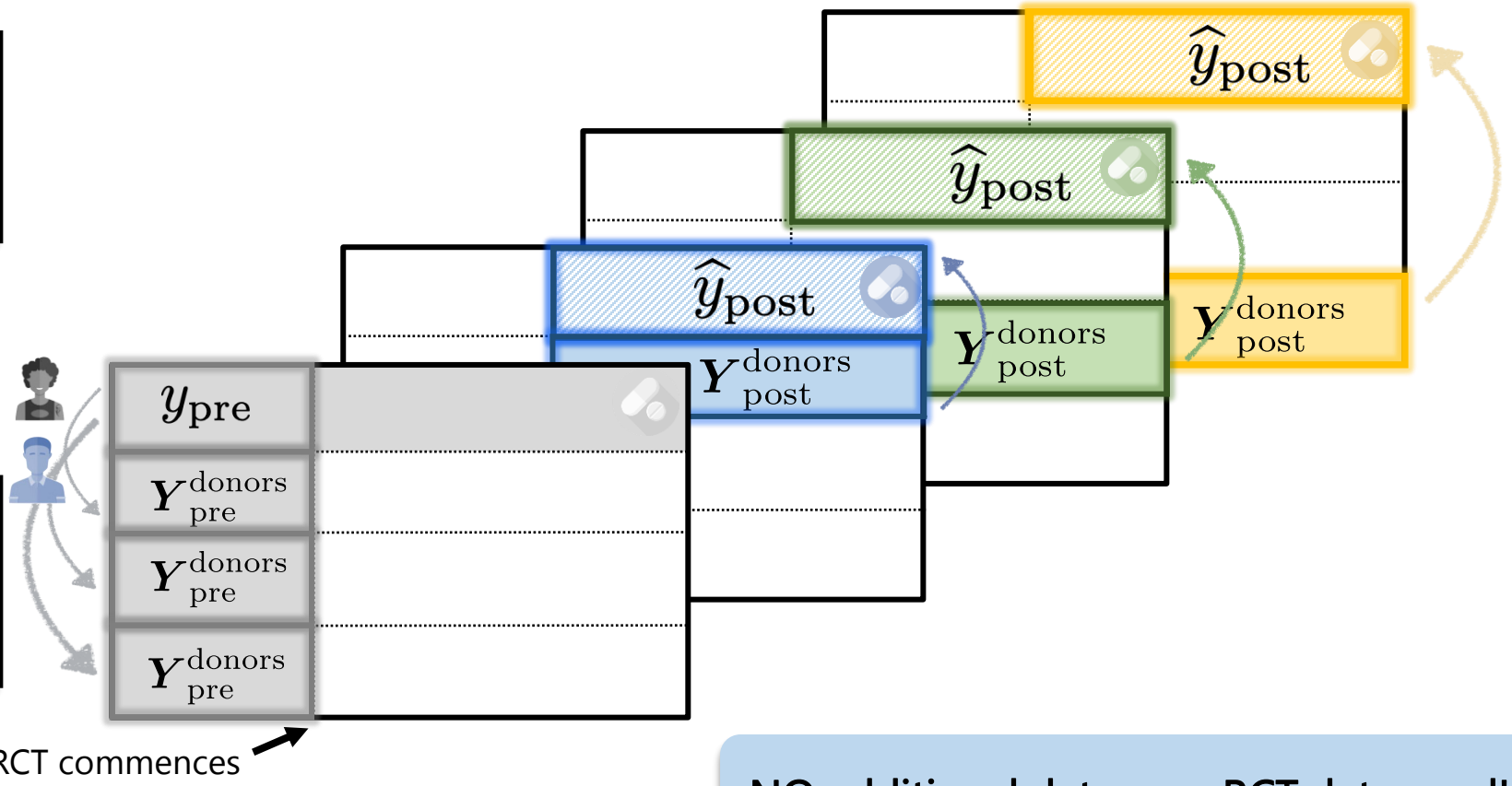
Model learning:

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

via Principal Component Regression

Counterfactual prediction:

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

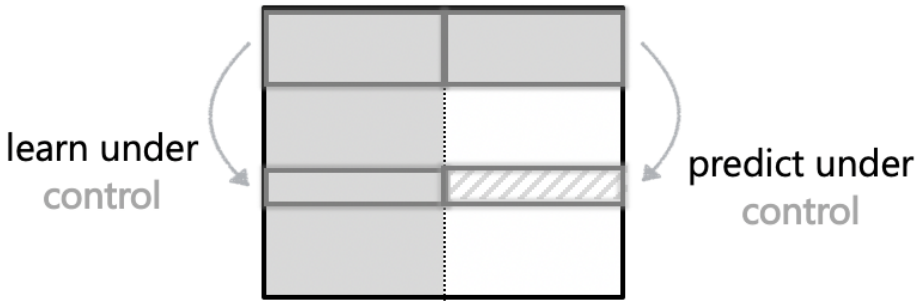


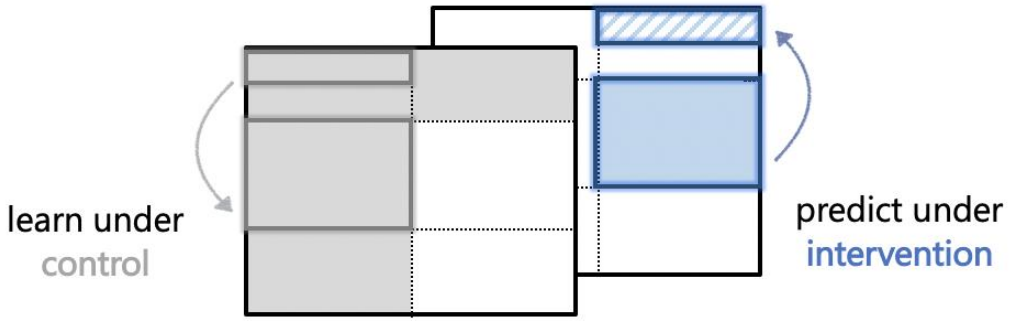
NO additional data over RCT data used!

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

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

Principal component regression (PCR)

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

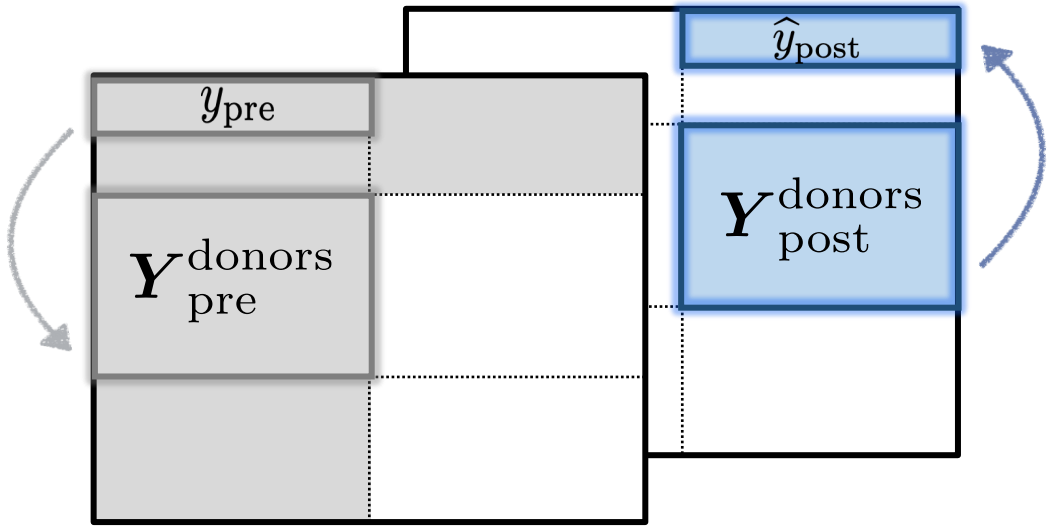
PCR is crucial to proving our formal statistical guarantees

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?

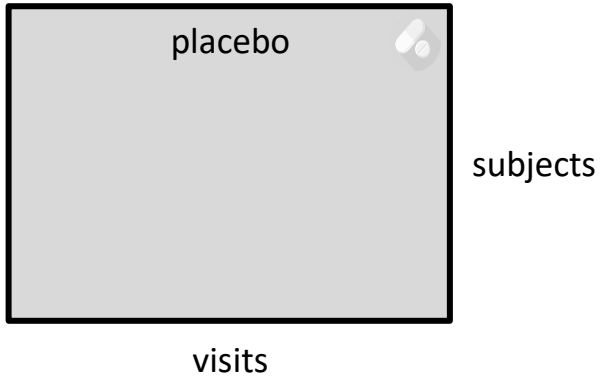
How to share info. across units with one treatment?

Matrix completion:
Low rank matrix

[Candes-Tao '10; Recht '11; Chatterjee '15, …]

Causal inference:
Factor models

[Chamberlain '84; Liang-Zeger '86;
Bai '03, '09; Pesaran '06, …]



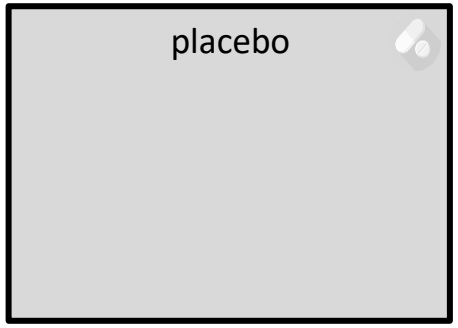
How to share info. across units with one treatment

Matrix completion:
Low rank matrix

[Candes-Tao '10; Recht '11; Chatterjee '15, …]

Causal inference:
Factor models

[Chamberlain '84; Liang-Zeger '86;
Bai '03, '09; Pesaran '06, …]



subjects

visits

low rank

time latent factors

$$Y_{it}^{(0)} = \sum_{\ell=1}^{r} u_{i\ell} v_{t\ell} + \varepsilon_{it}$$

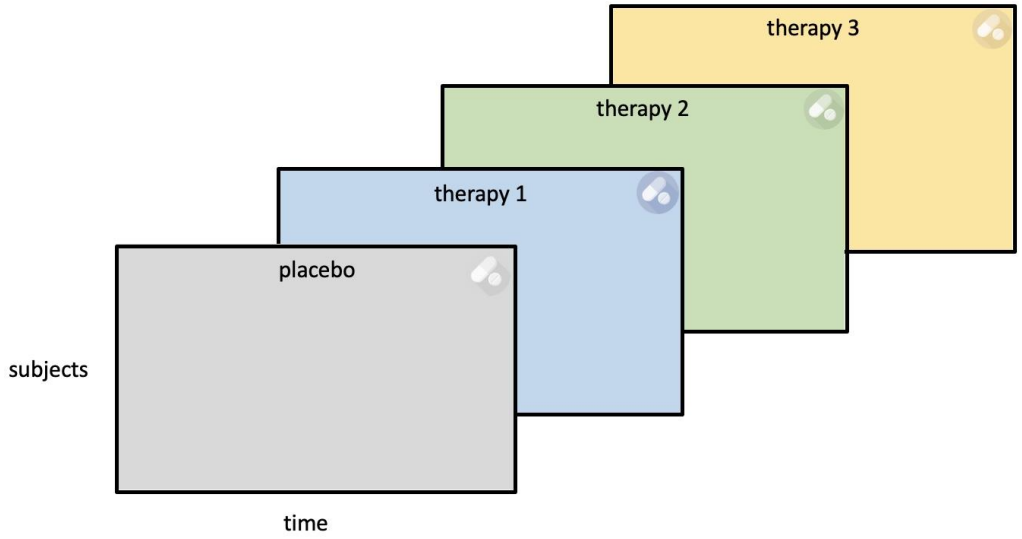
stochasticity

$\mathbb{E}[Y_{it}^{(0)}]$

unit latent factors

How to share info. across units with multiple treatments?

Low rank tensor!



$$Y_{it}^{(d)} = \sum_{\ell=1}^{r} u_{i\ell} v_{t\ell} w_{d\ell} + \varepsilon_{it}^{(d)}$$

intervention latent factors

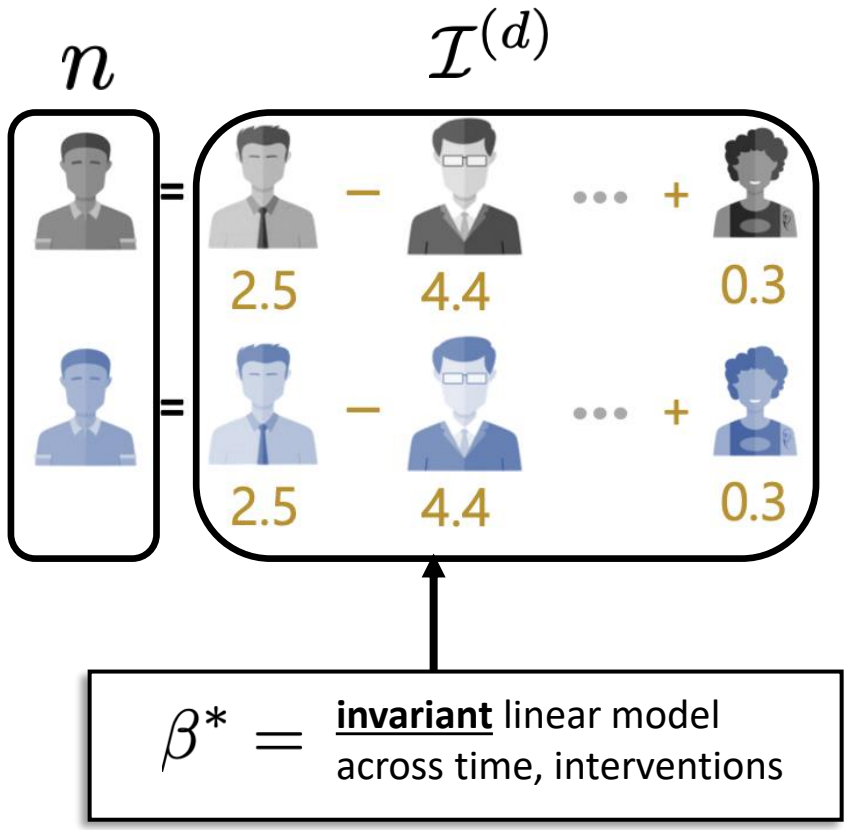
Low rank implies linear models

Produce counterfactuals for
unit n under **intervention d**

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

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

Holds w.h.p if factors sampled independently



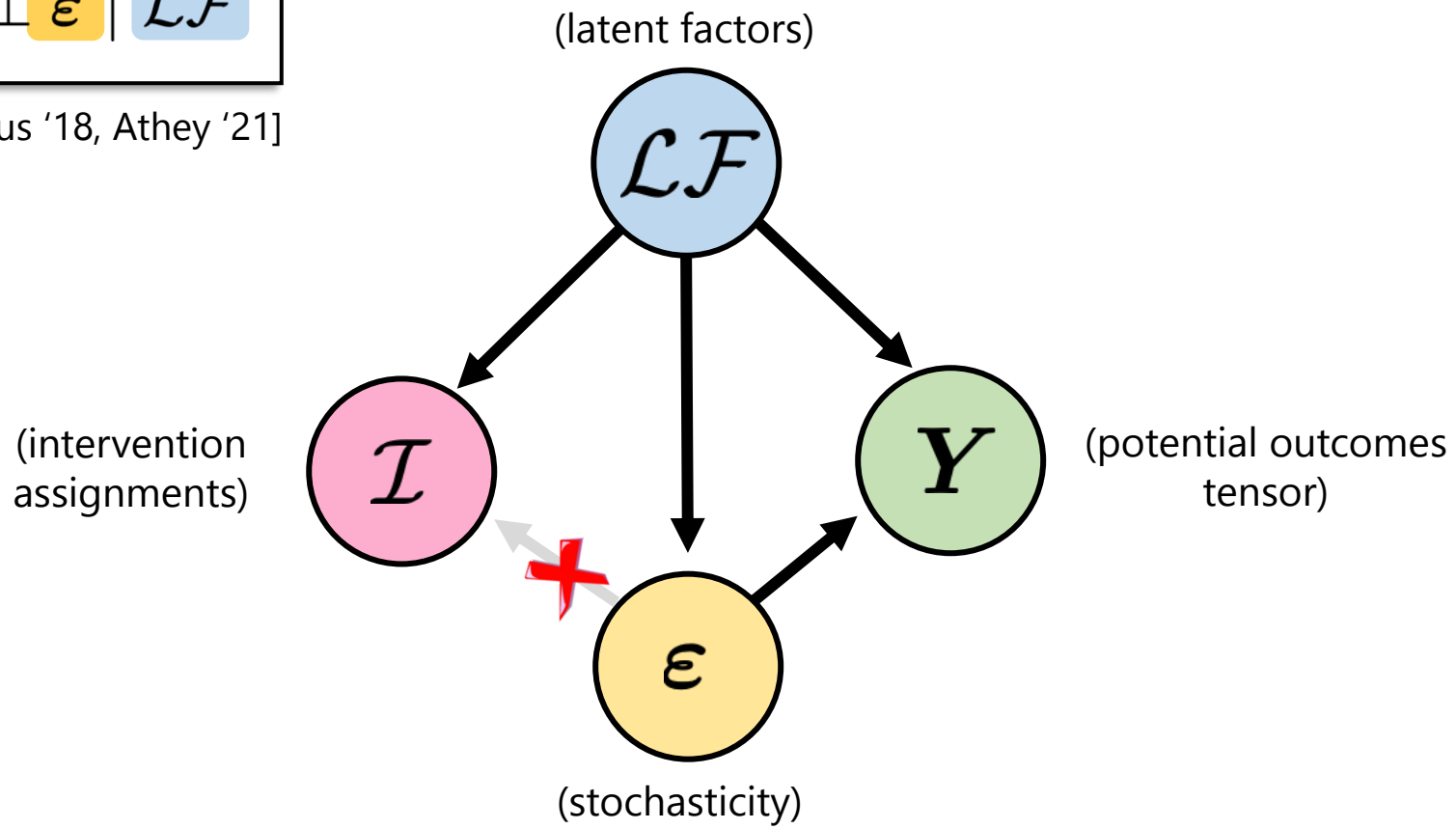
What type of confounding?—*selection on latent factors*

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

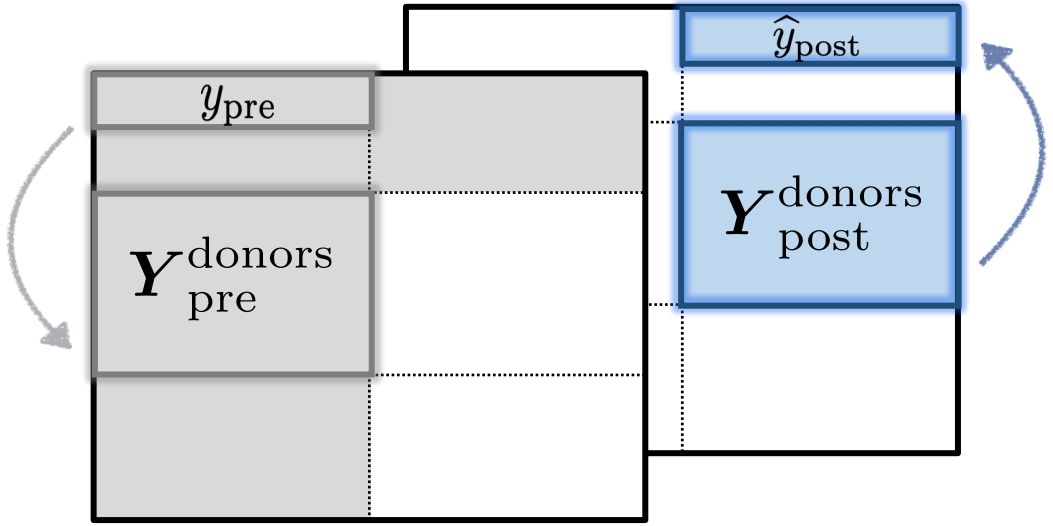
$$+ \boxed{\mathcal{I} \perp\!\!\!\perp \boldsymbol{\varepsilon} \mid \mathcal{LF}}$$

[also Kallus '18, Athey '21]

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



Identification



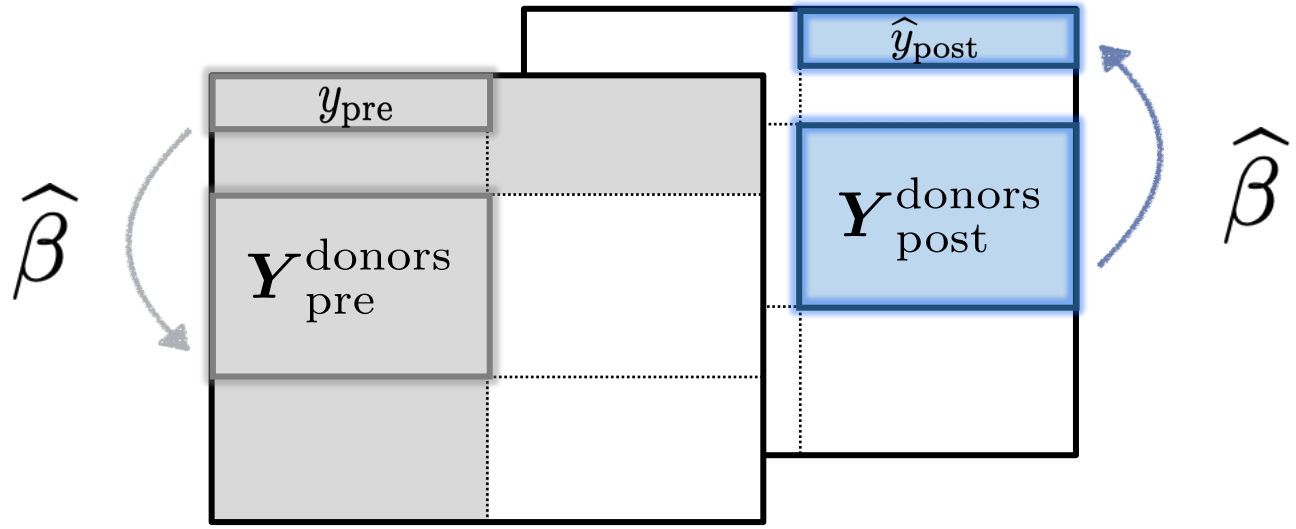
$$\mathbb{E}[y_{\text{post}}] = \mathbb{E}[\,\boldsymbol{X}_{\text{post}}\,] \cdot \beta^*$$

Philosophical object

observed

estimable

When transferrable?—*subspace inclusion*

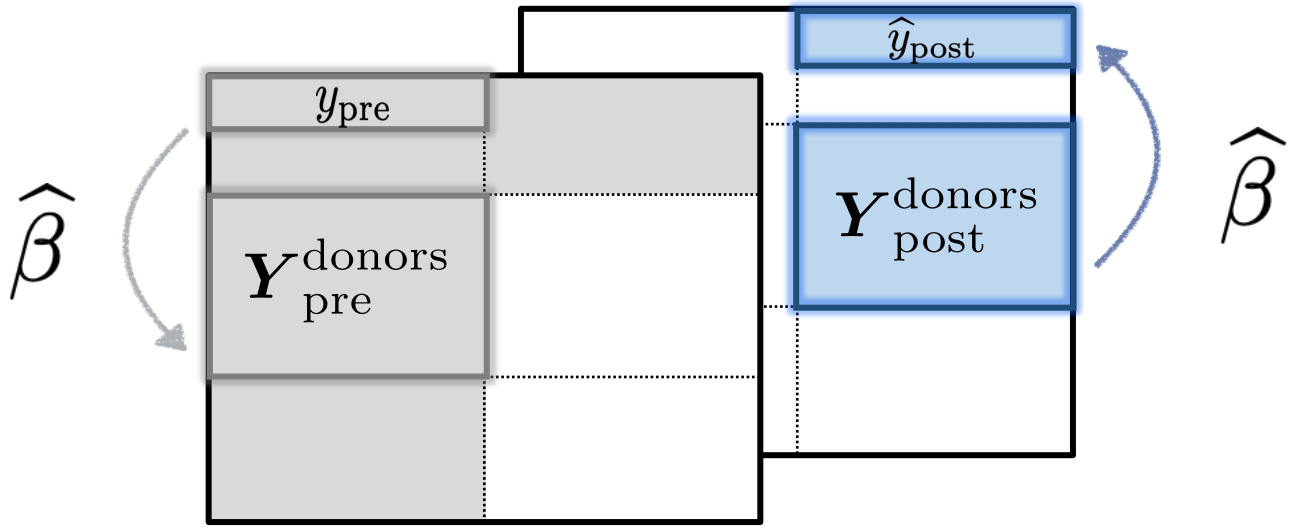


$$\text{complexity}\left(\mathbf{Y}^{\text{donors}}_{\text{post}}\right) \leq \text{complexity}\left(\mathbf{Y}^{\text{donors}}_{\text{pre}}\right)$$

"test" set

"train" set

When transferrable?—*subspace inclusion*



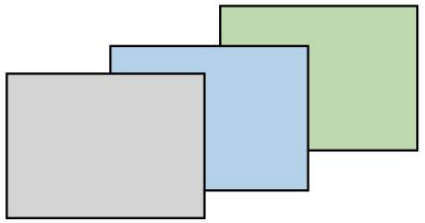
$$\text{colspan}\left(\mathbf{Y}^{\text{donors}}_{\text{post}}\right) \subseteq \text{colspan}\left(\mathbf{Y}^{\text{donors}}_{\text{pre}}\right)$$

"test" set

"train" set

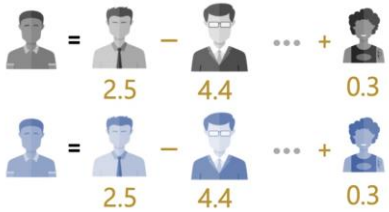
Operating assumptions

low rank tensor



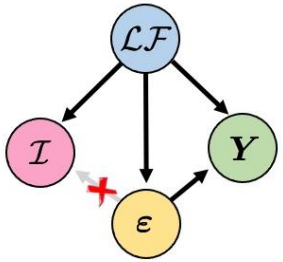
+

linear model



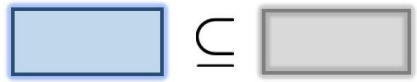
+

selection on latent factors



+

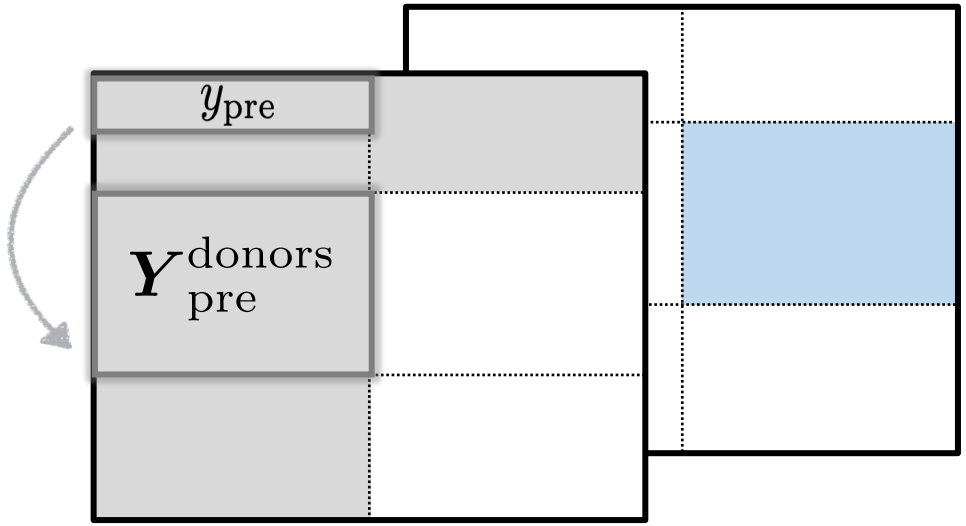
subspace inclusion



Model identification

Model learning:

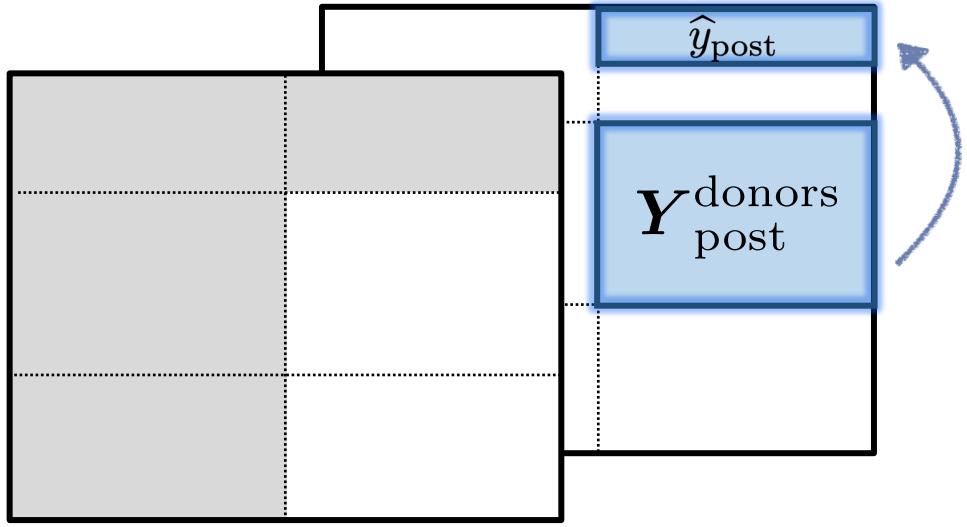
$$\widehat{\beta} = \mathrm{PCR}(y_{\mathrm{pre}}, \boldsymbol{Y}_{\mathrm{pre}}^{\mathrm{donors}})$$



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

minimum norm model

Consistency

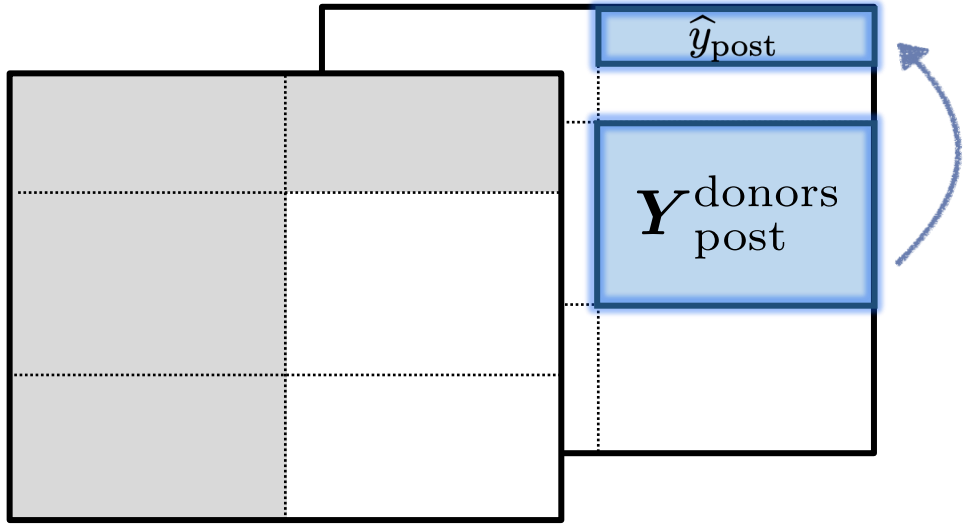


Counterfactual prediction:

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

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

Asymptotic normality



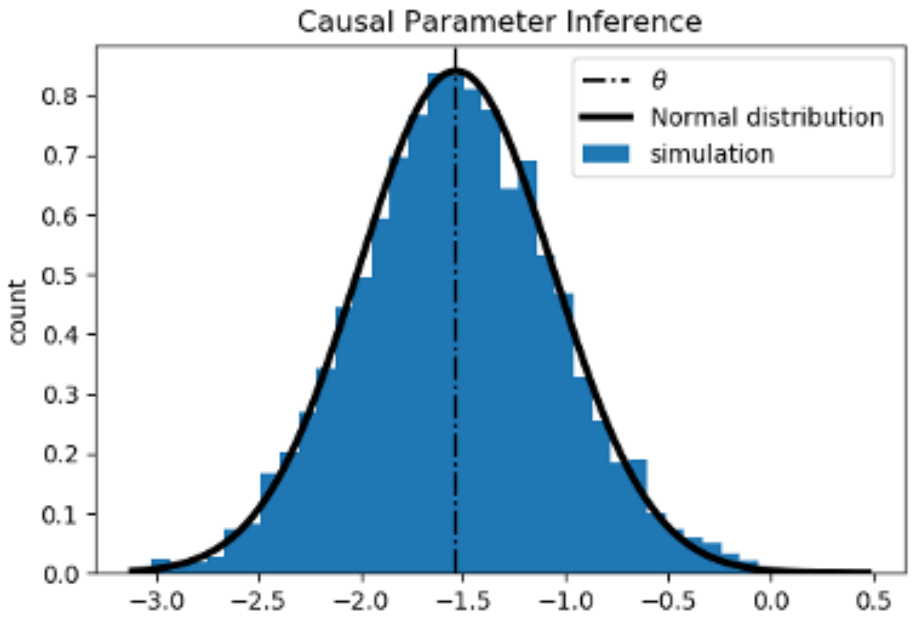
Counterfactual prediction:

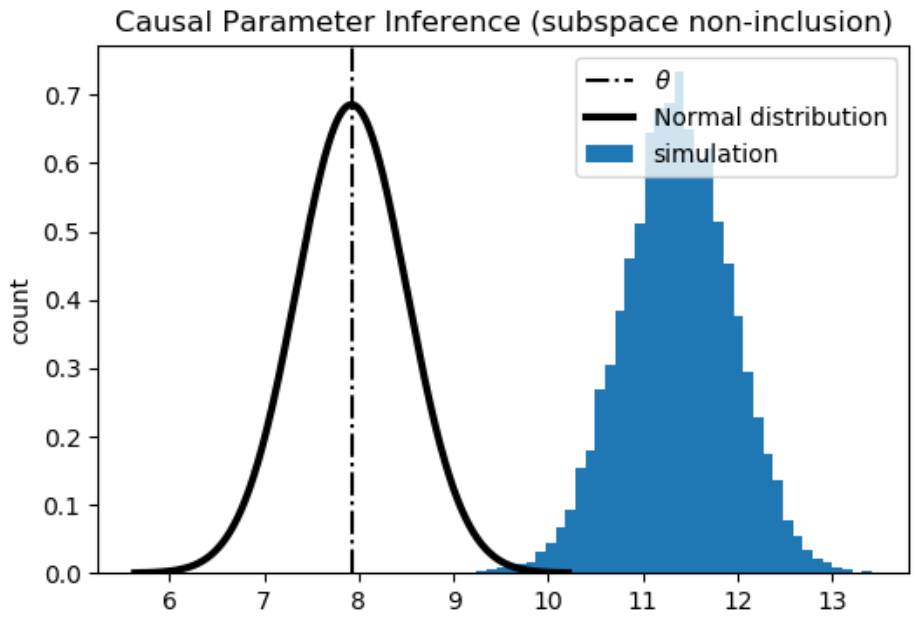
$$\widehat{y}_{\text{post}} = \boldsymbol{Y}^{\text{donors}}_{\text{post}} \widehat{\beta}$$

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

enables confidence intervals

Importance of subspace inclusion





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

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

Practitioner's Guide:
Empirical illustrations

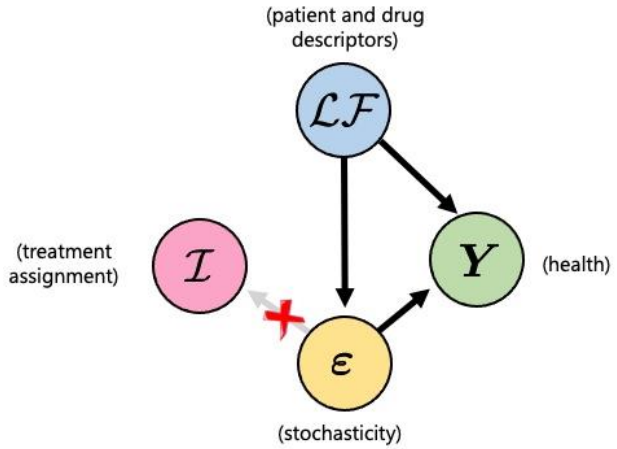
Clinical trial study

Experimental study

2 year study

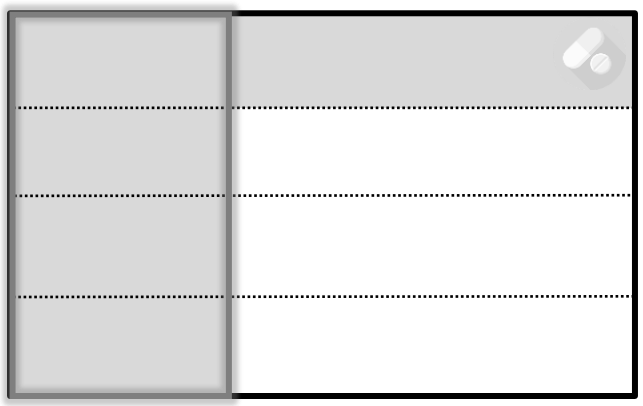
1000+ subjects

4 therapies (1 placebo)

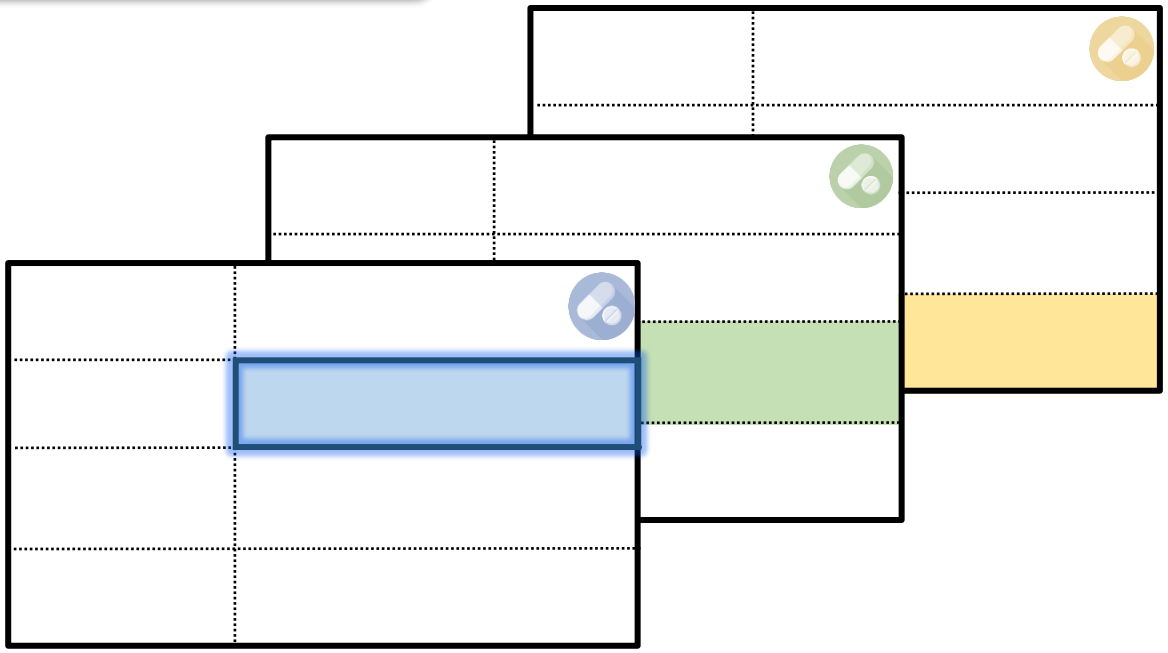


Synthetic RCT—an application

Training data

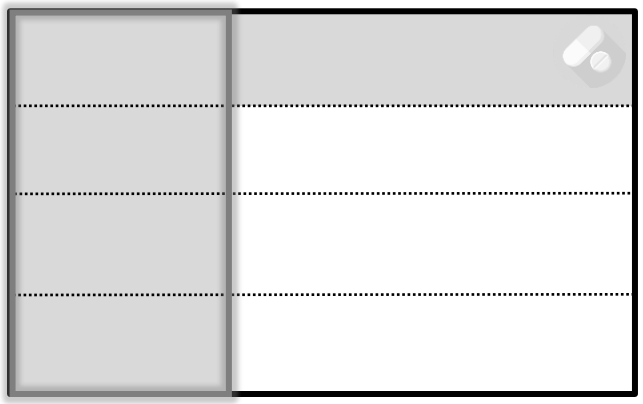


Test data

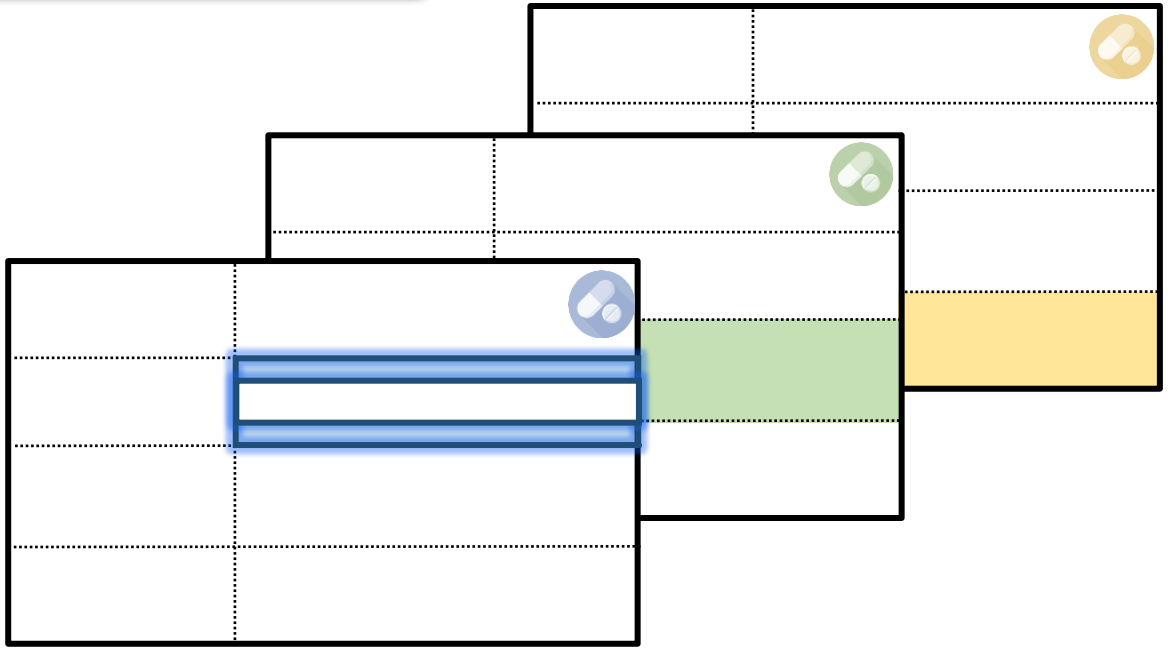


Synthetic RCT—validation setup

Training data

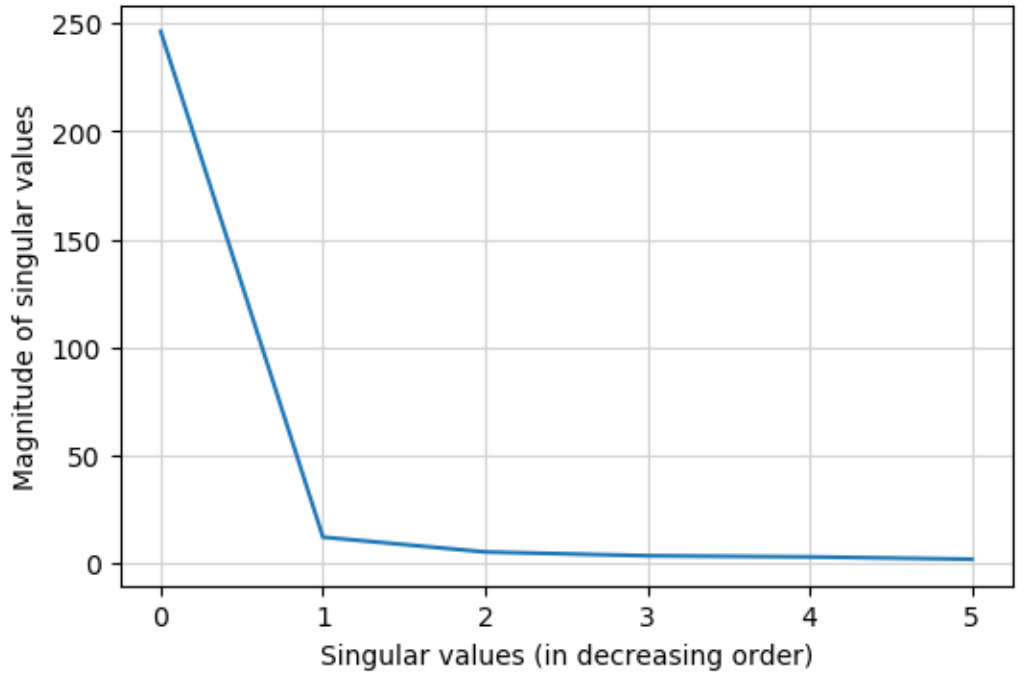


Test data

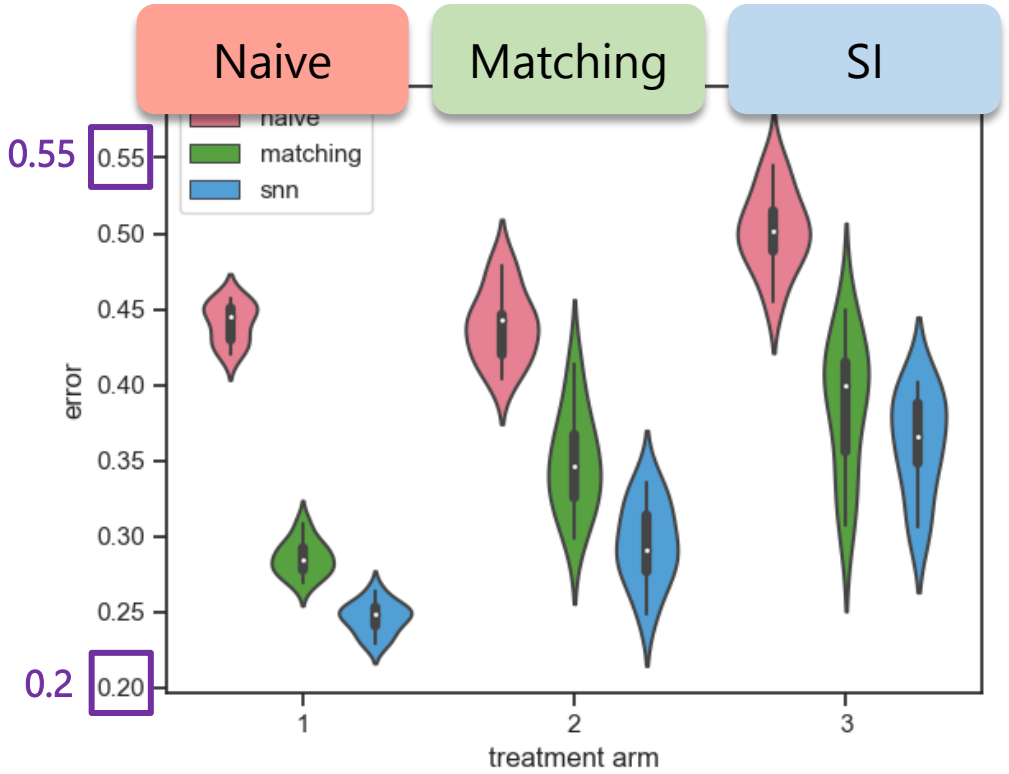


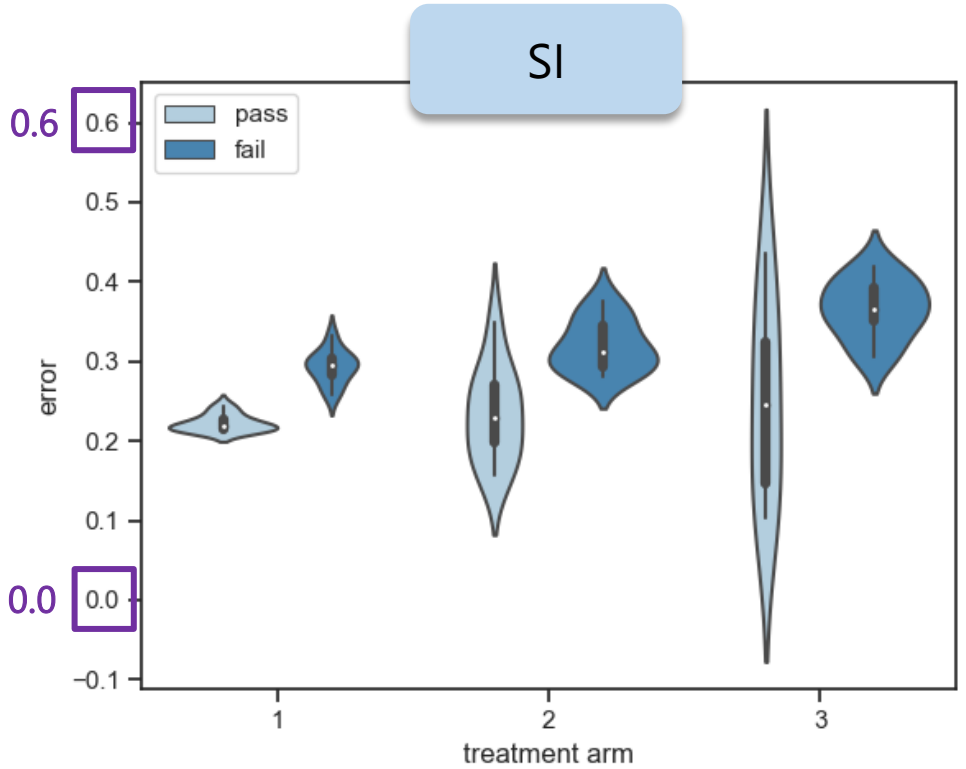
Diagnostic: look for low-rank structure!

```
import numpy as np
import matplotlib.pyplot as plt
(u, s, v) = np.linalg.svd(data, full_matrices=False)
plt.figure()
plt.plot(s)
plt.show()
```



Validation study results





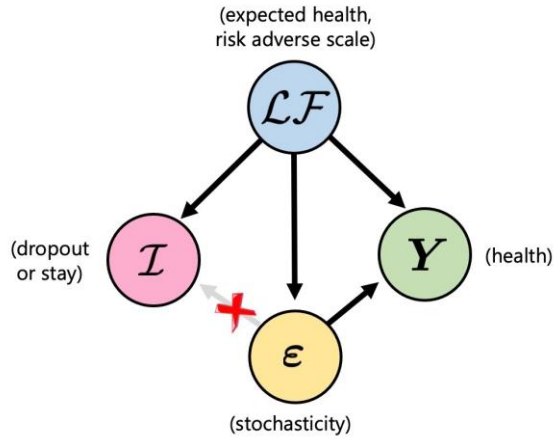
Clinical trial study II

Observational study

2 year study

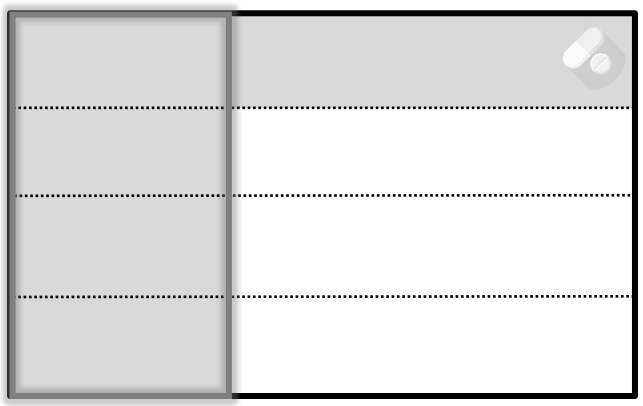
1000+ subjects

Comply or dropout

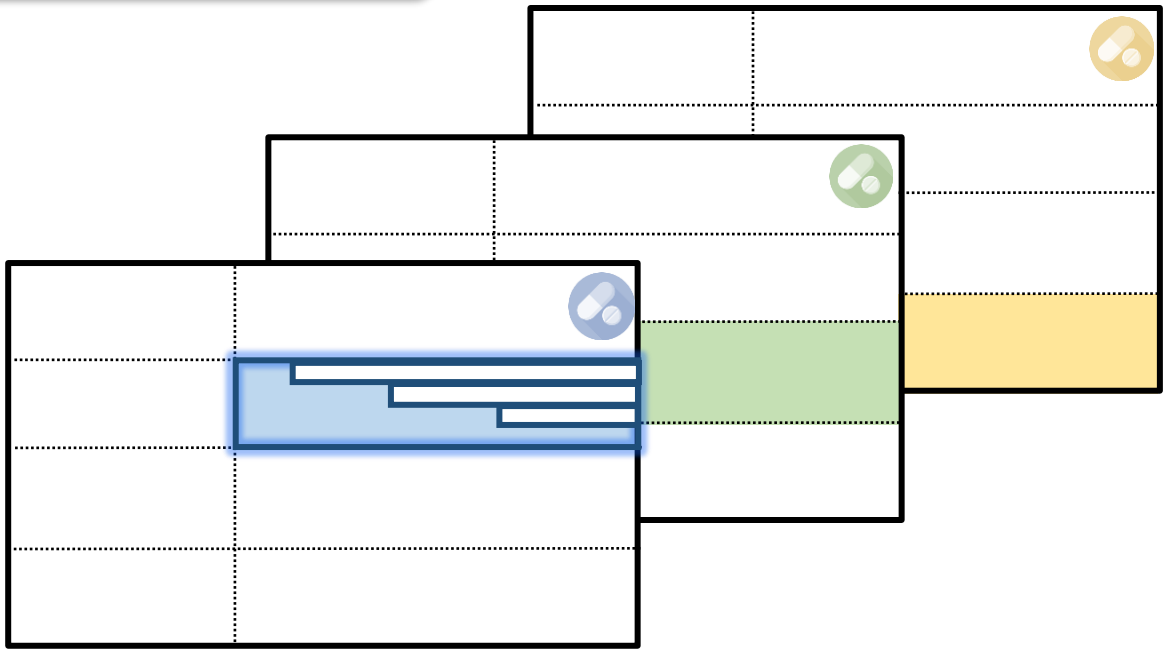


Impute dropout data

Training data

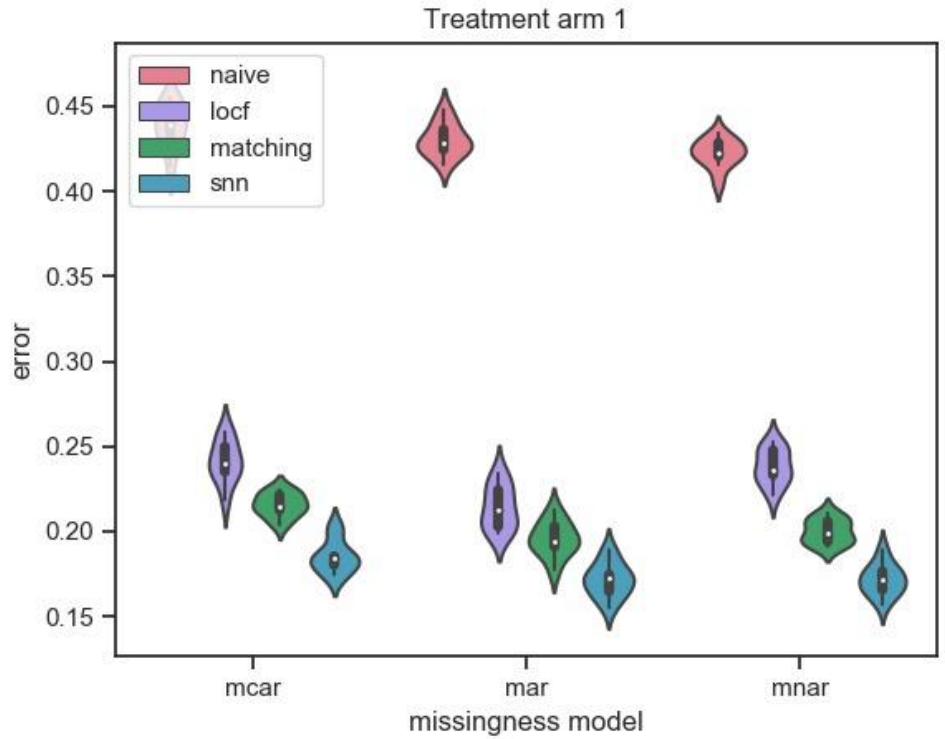


Test data

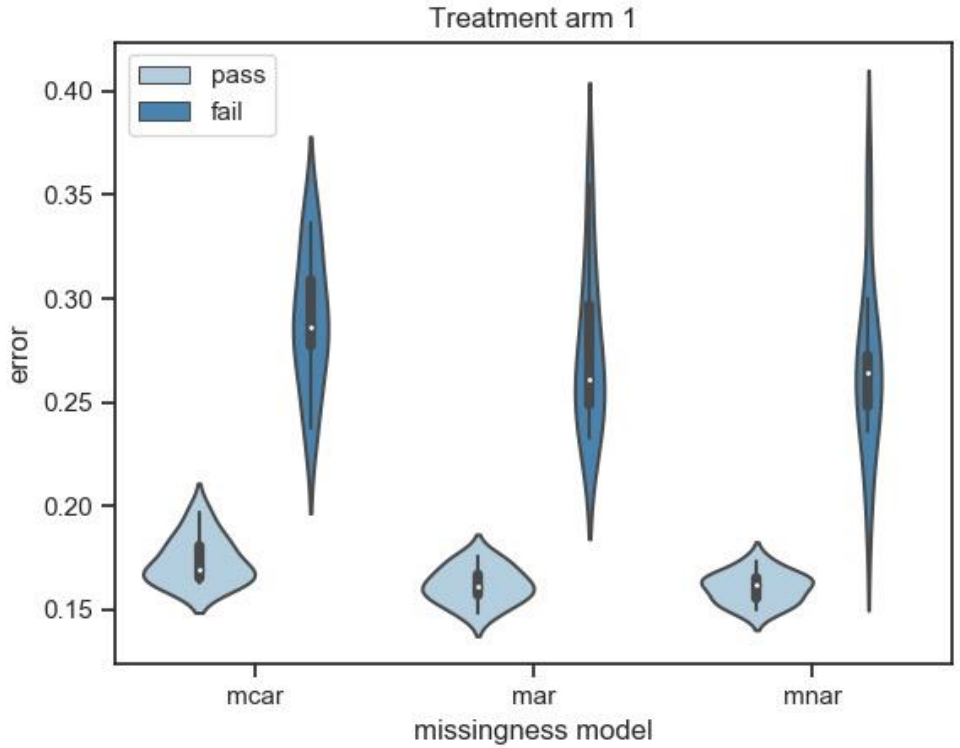


Validation study results

Naive | LOCF | Matching | SI

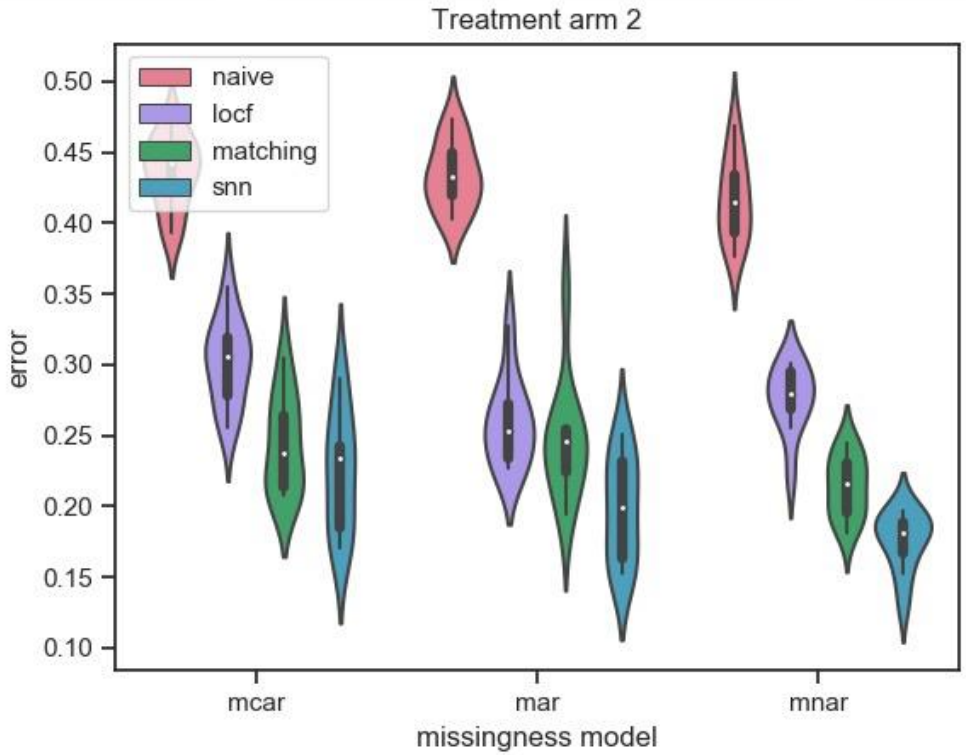


SI

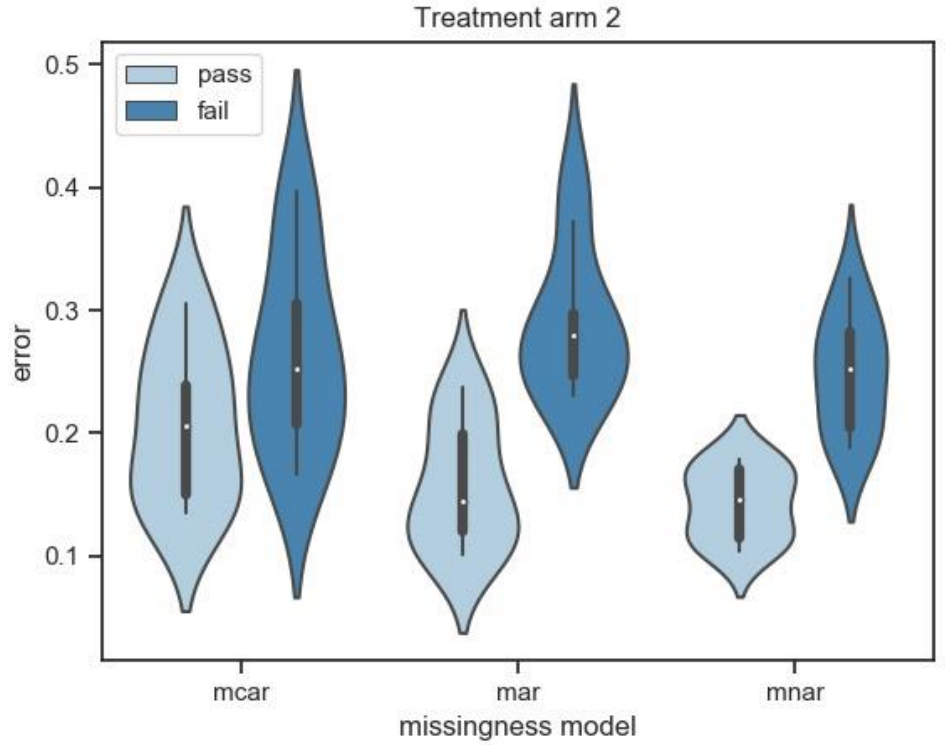


Validation study results

Naive | LOCF | Matching | SI

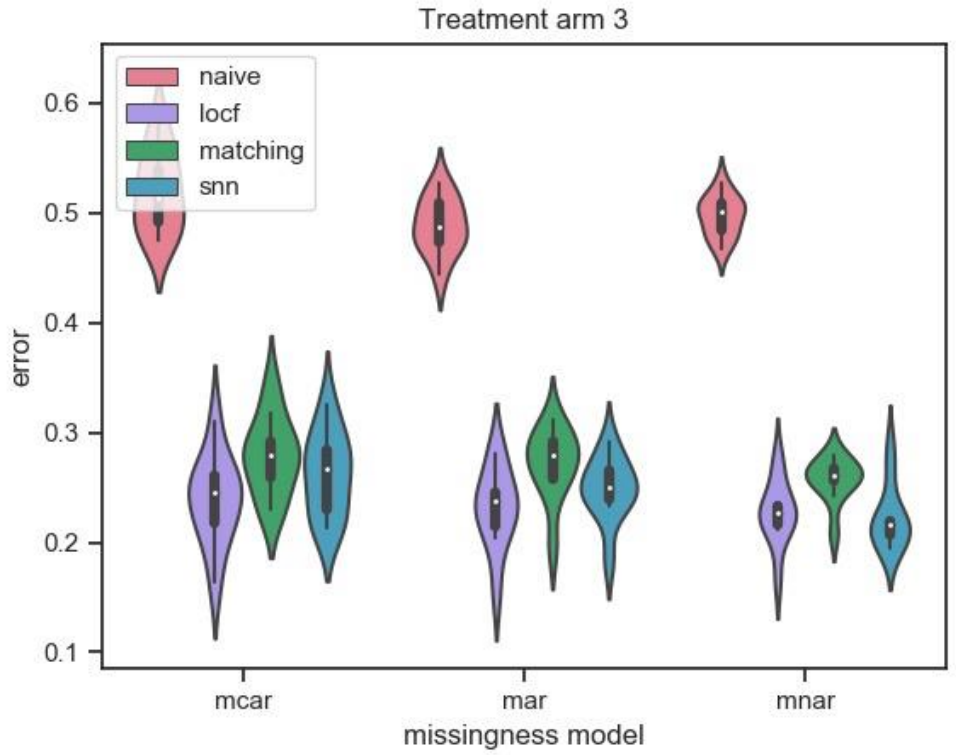


SI

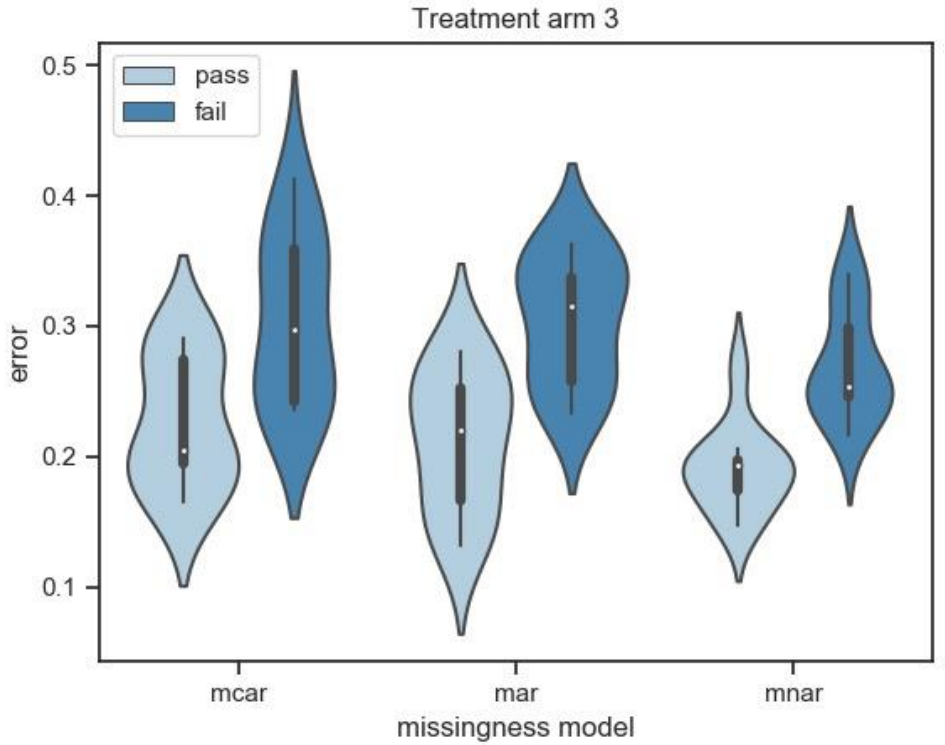


Validation study results

Naive | LOCF | Matching | SI



SI



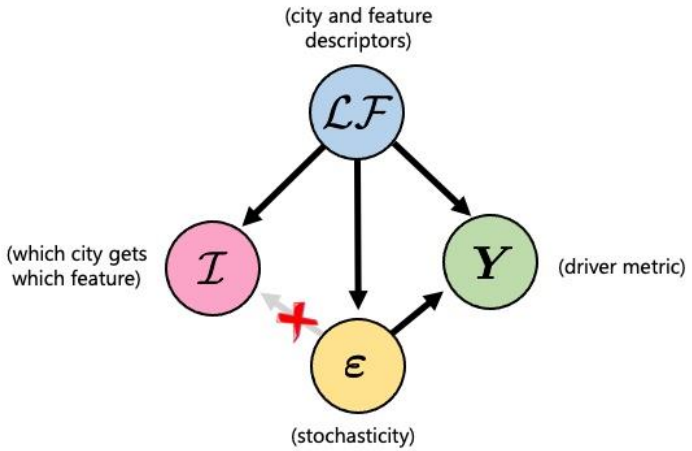
Ride sharing study

Observational study

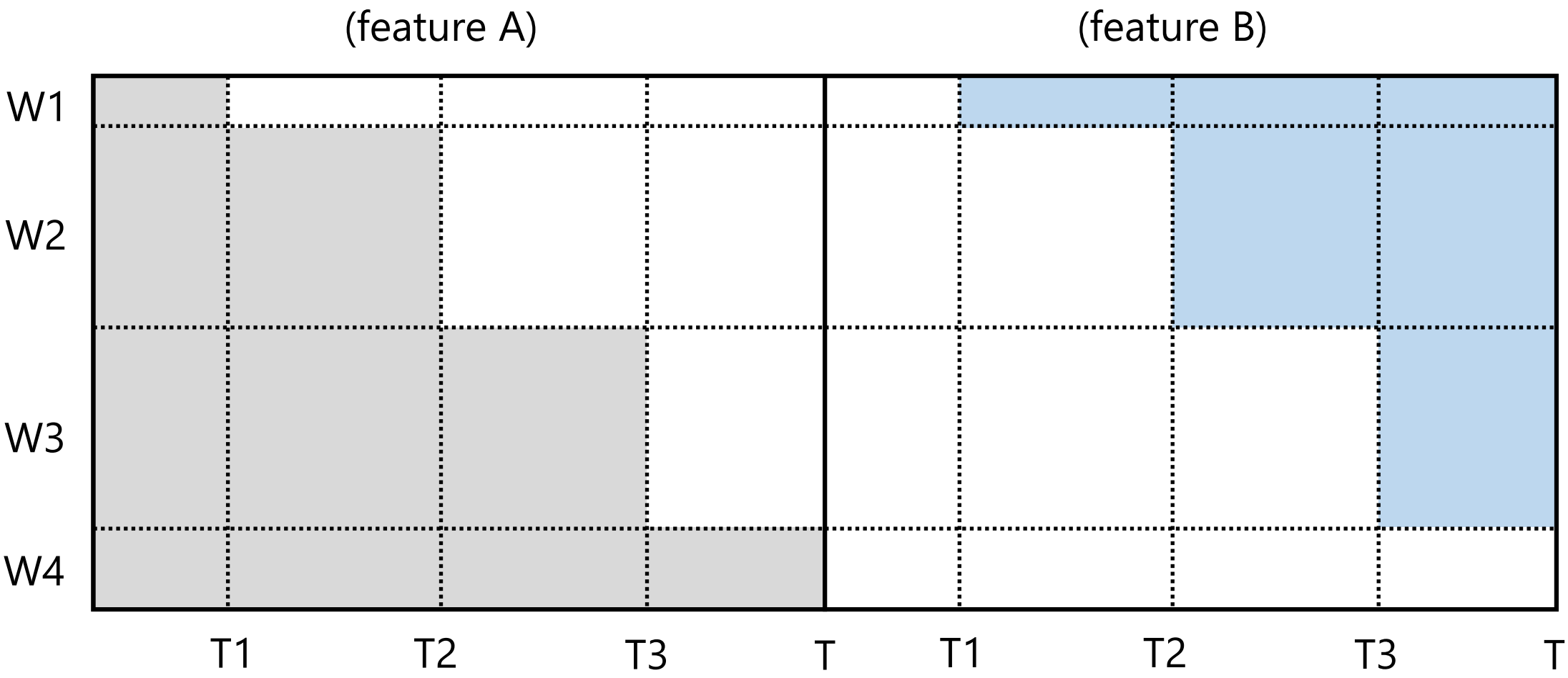
72 week study

~180 cities

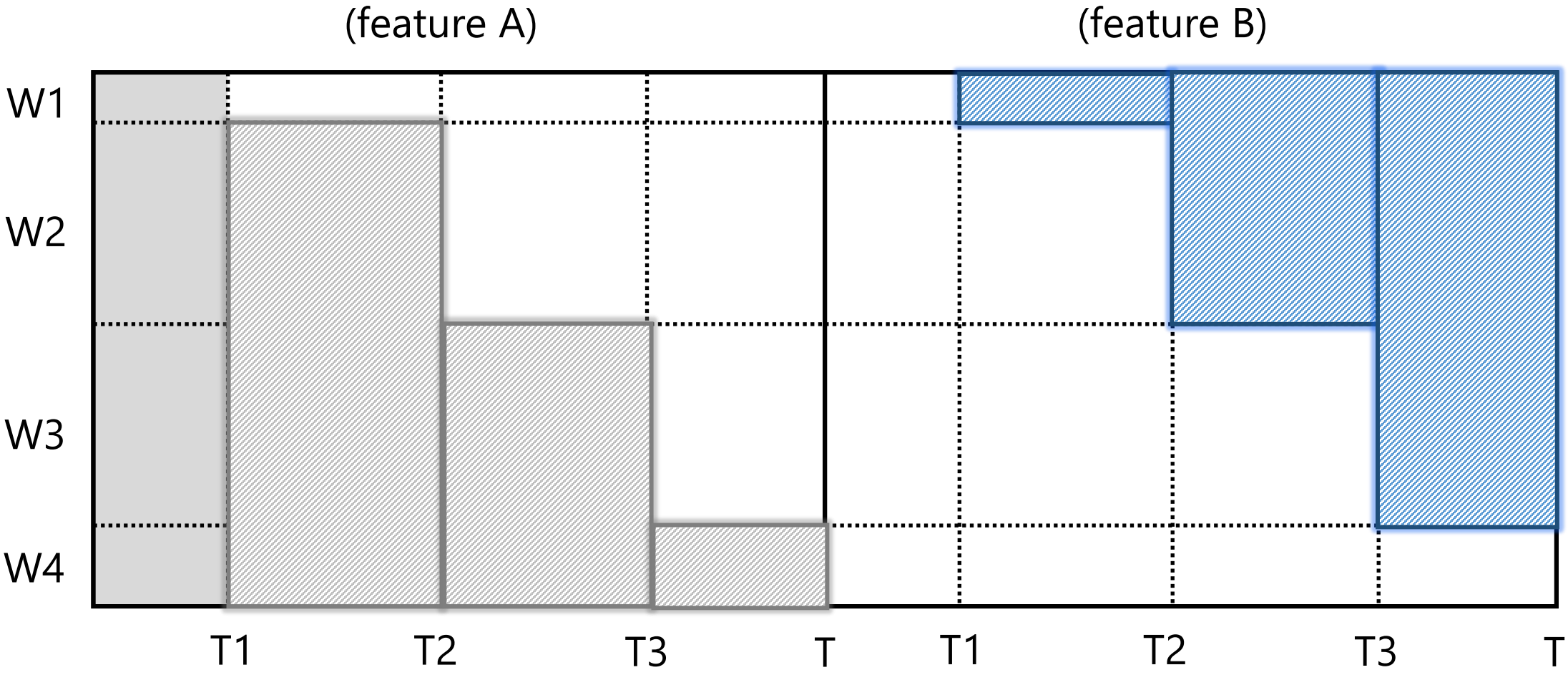
2 features



Simulate rollouts—an application

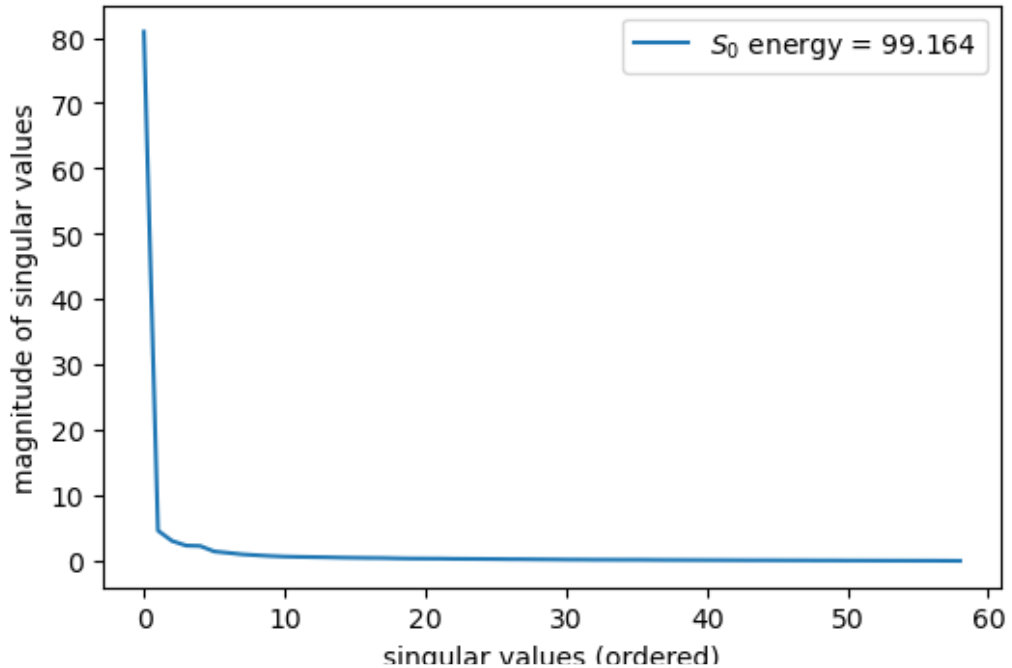


Simulate rollouts—validation setup

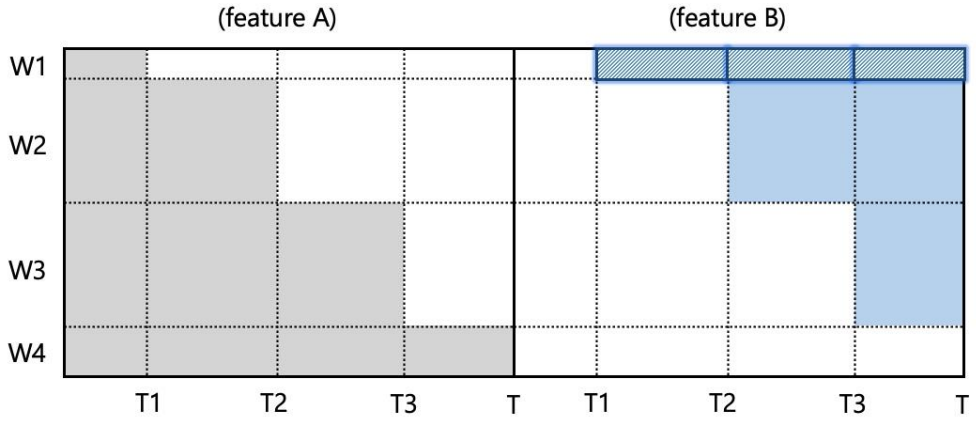


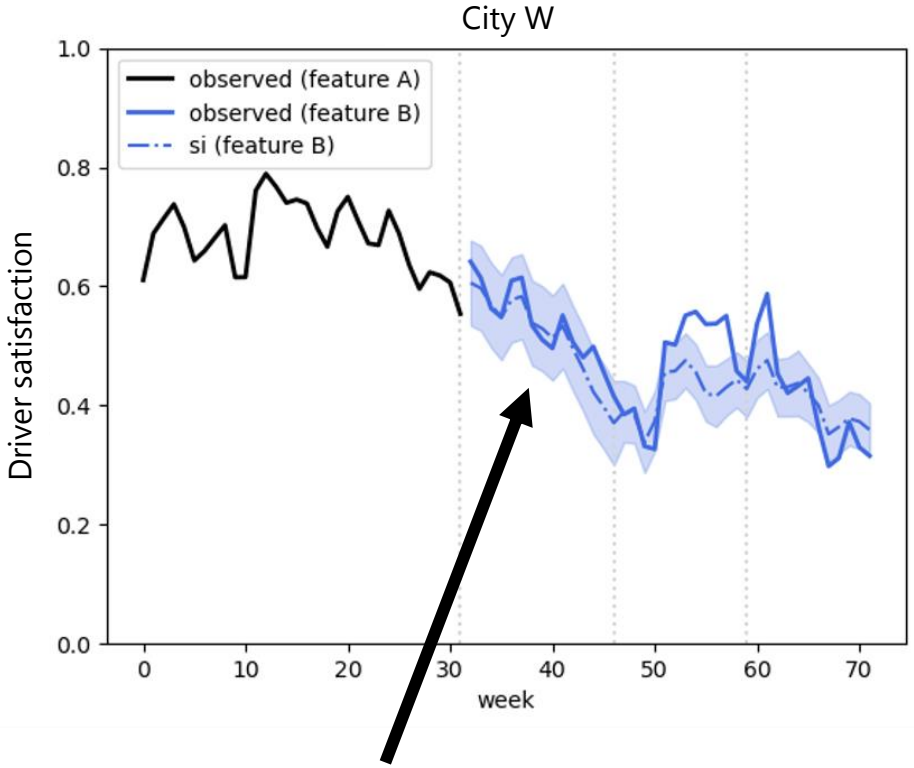
Diagnostic: look for low-rank structure!

```
import numpy as np
import matplotlib.pyplot as plt
(u, s, v) = np.linalg.svd(data, full_matrices=False)
plt.figure()
plt.plot(s)
plt.show()
```



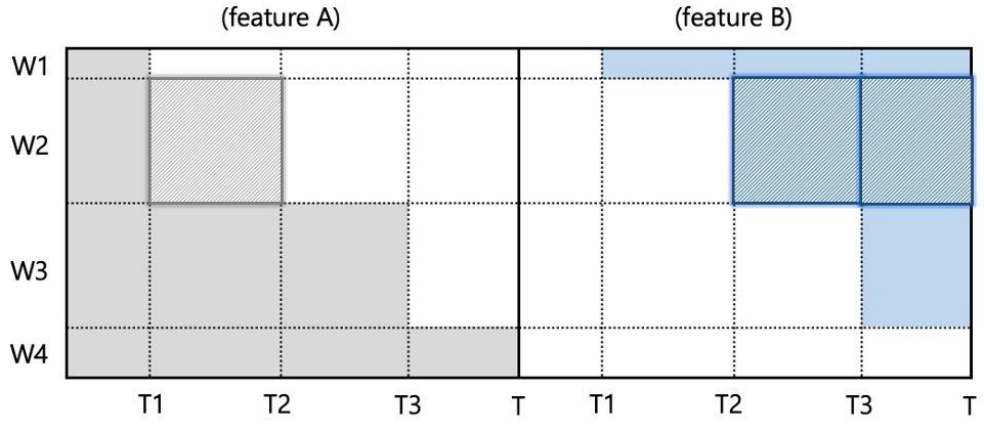
Validation study results—wave 1

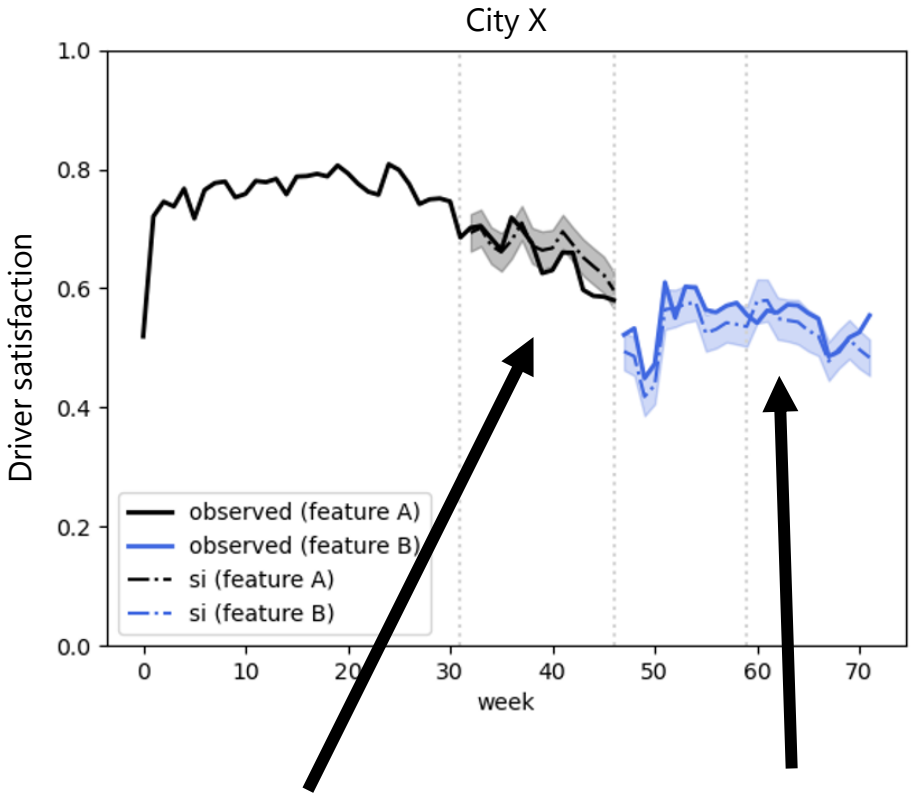




"synthetic interventions"

Validation study results—wave 2

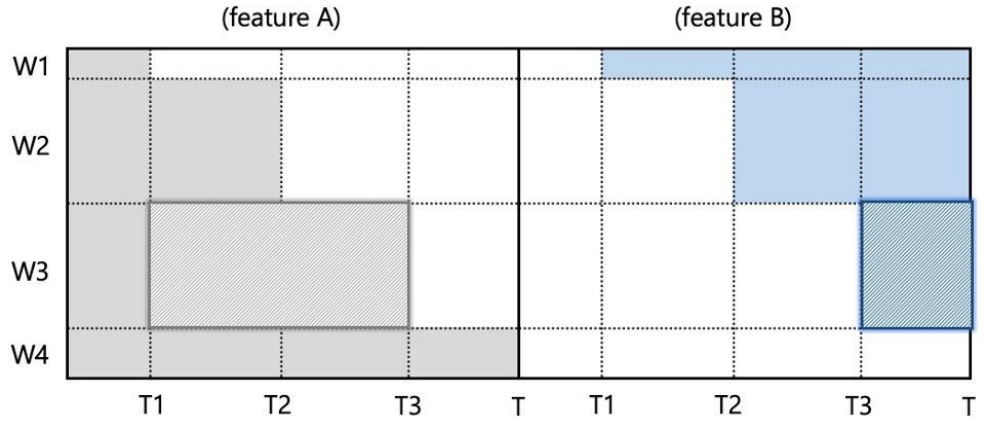


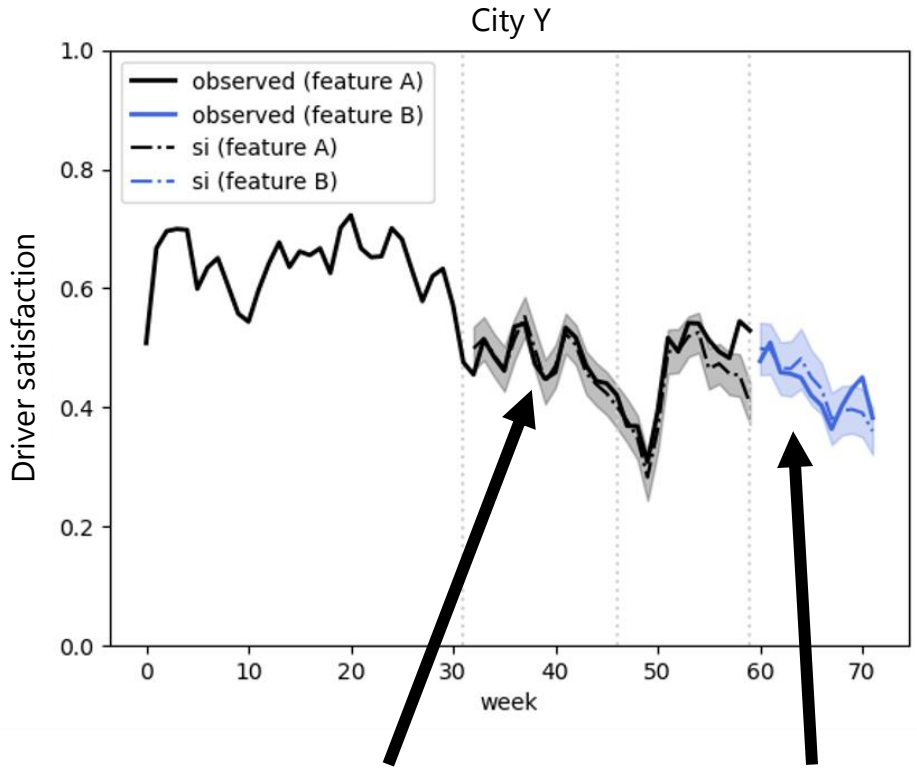


"synthetic controls"

"synthetic interventions"

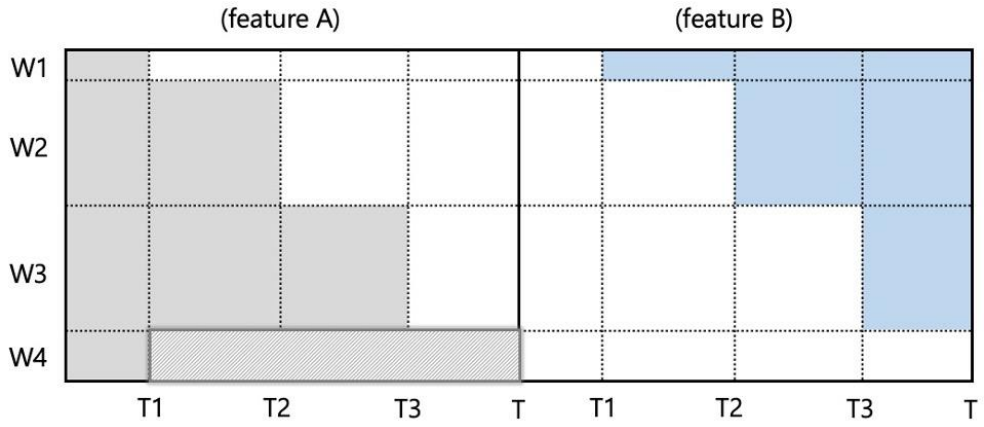
Validation study results—wave 3

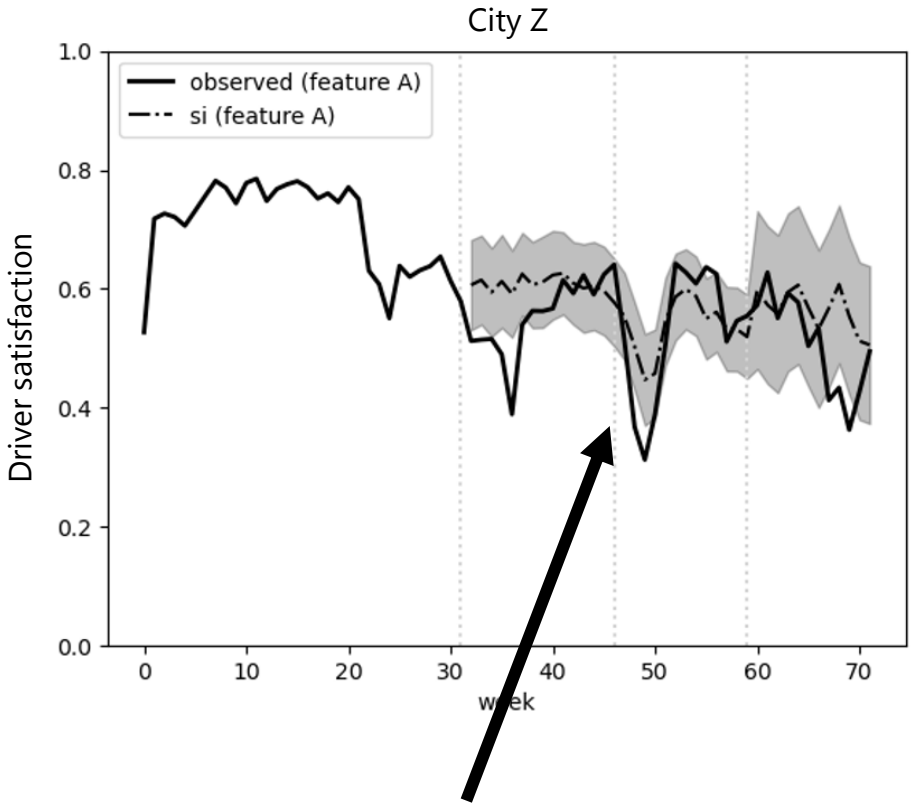




"synthetic controls" "synthetic interventions"

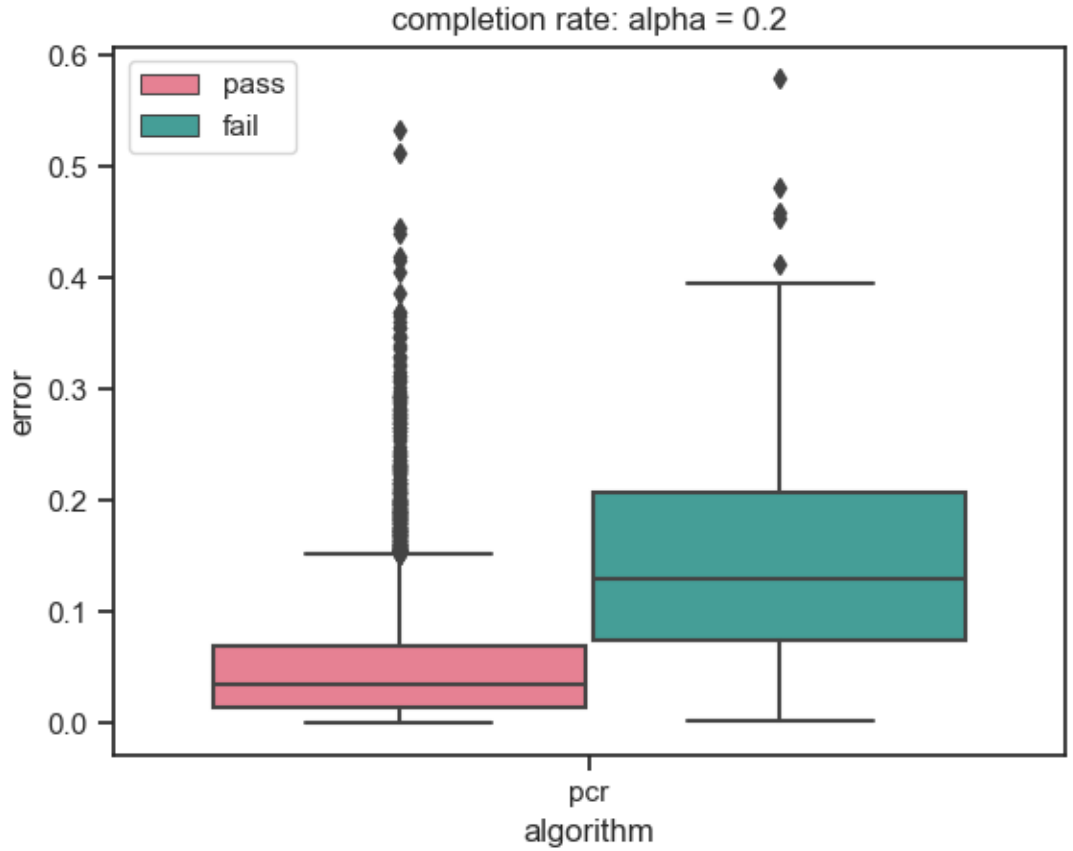
Validation study results—wave 4





"synthetic controls"

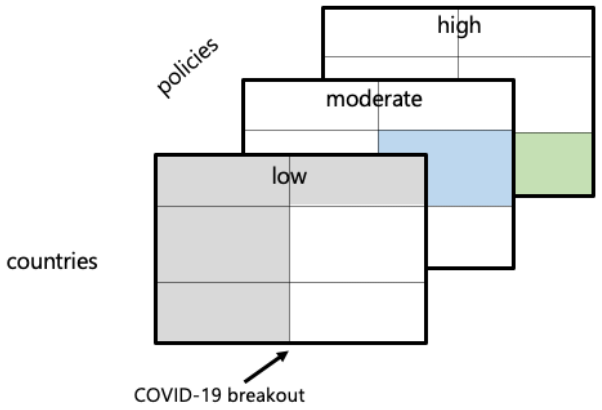
Validation study results—subspace inclusion



Additional studies

COVID-19 policy evaluation

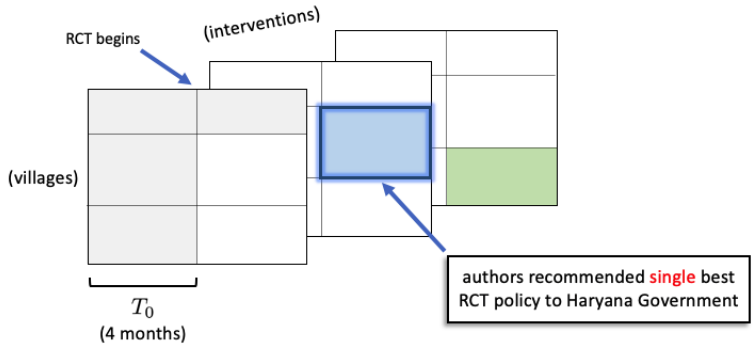
Question:
U.S. COVID-19 mortality rates if stricter social distancing adopted?



observational

Poverty Action Lab @ MIT

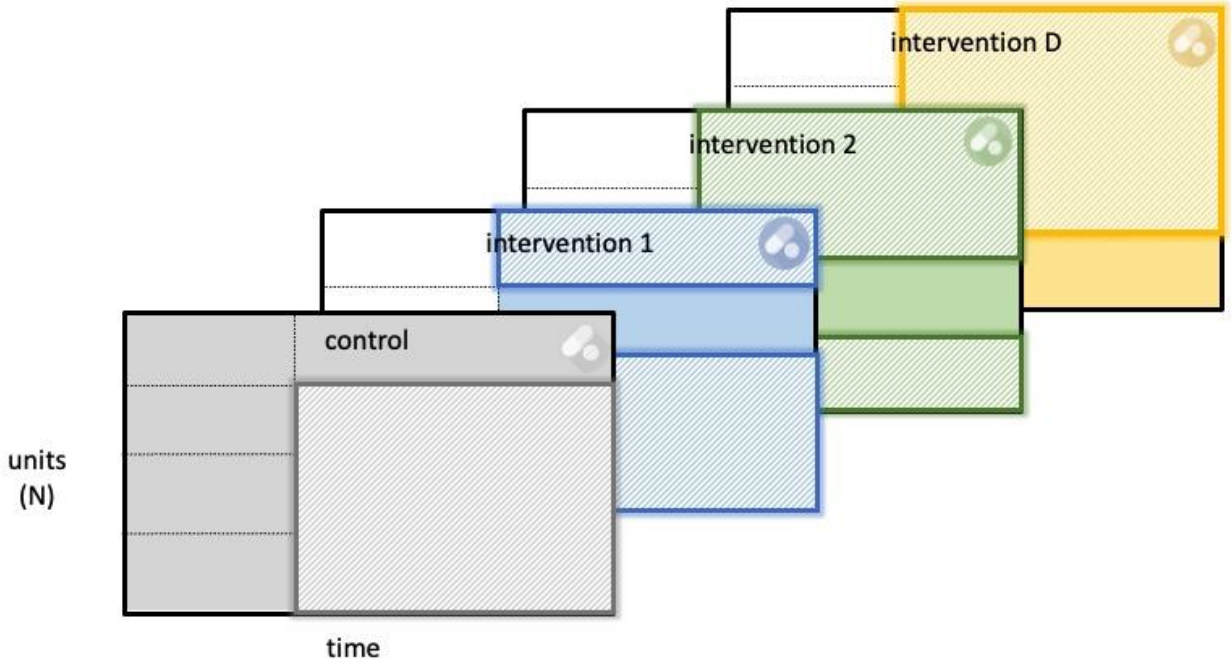
Question:
Uptake in childhood immunization rates if personalized policies used across villages in India ?



limited experimental

Block pattern ubiquitous in practice

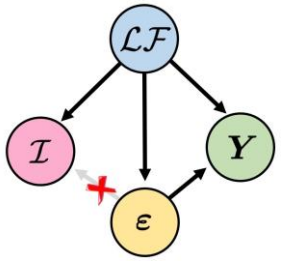
Key takeaway



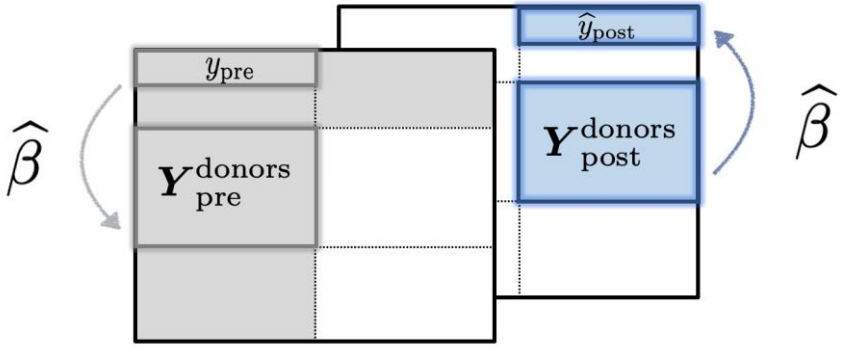
provably learns all N × D causal estimands with

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

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

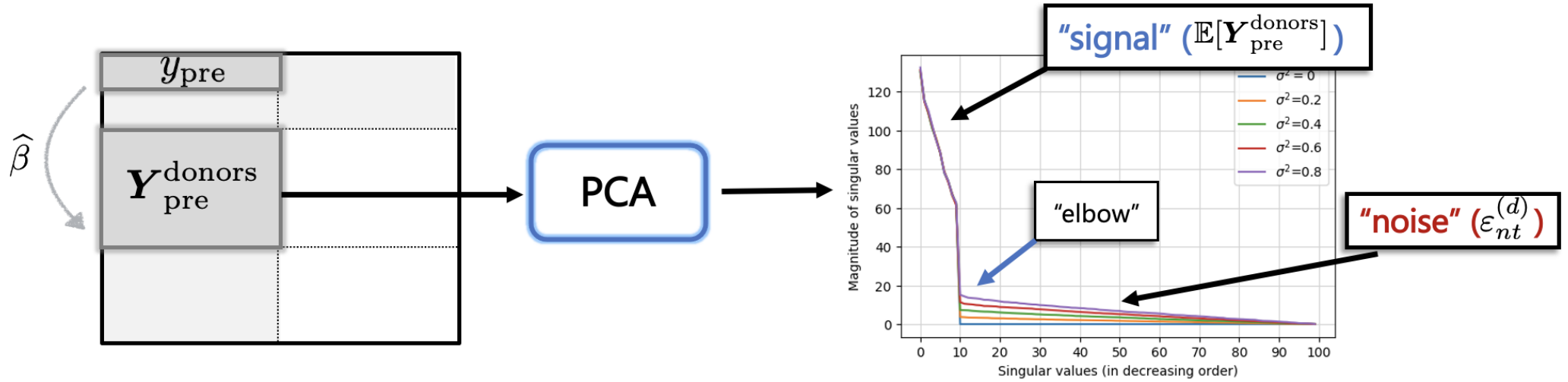


Numerous ways to estimate model parameter



- Ordinary least squares [Hsiao et al '12, Li-Bell '17]
- Convex regression [Abadie et al '03, 10, 15]
- Ridge regression ['Ben-Michael et al '21]
- Lasso regression [Carvalho '18, Chernozukhov '21]
- Elastic net regression [Doudchenko-Imbens '16]
- Non-negative weights [Li' 20]
- Fancy ML

Don't forget about PCR!



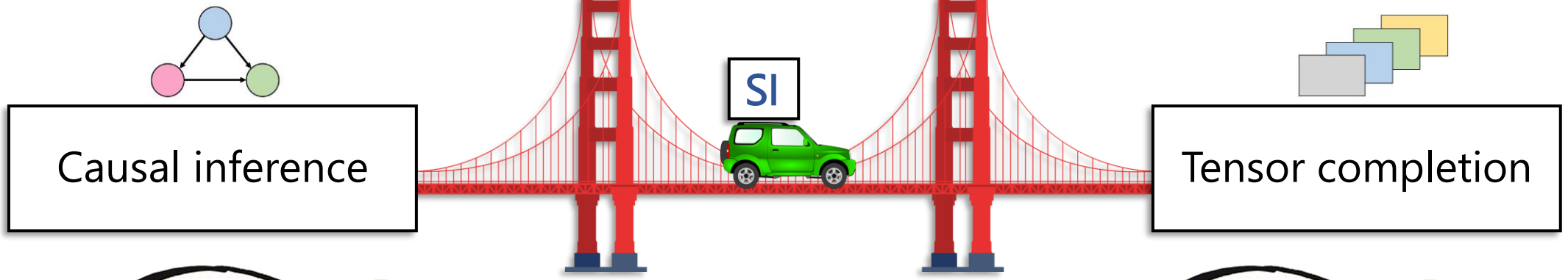
Intuition:

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

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

How to find "elbow": [Gavish-Donoho '14, Chatterjee '15]

Looking forward



modeling

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

modeling

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

algorithmic

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

THANK YOU

dshen24@berkeley.edu

https://arxiv.org/pdf/2006.07691.pdf

Acknowledgements:

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