Causal Tensor Estimation

Devavrat Shah

Alberto Abadie Anish Agarwal Dennis Shen

Massachusetts Institute of Technology

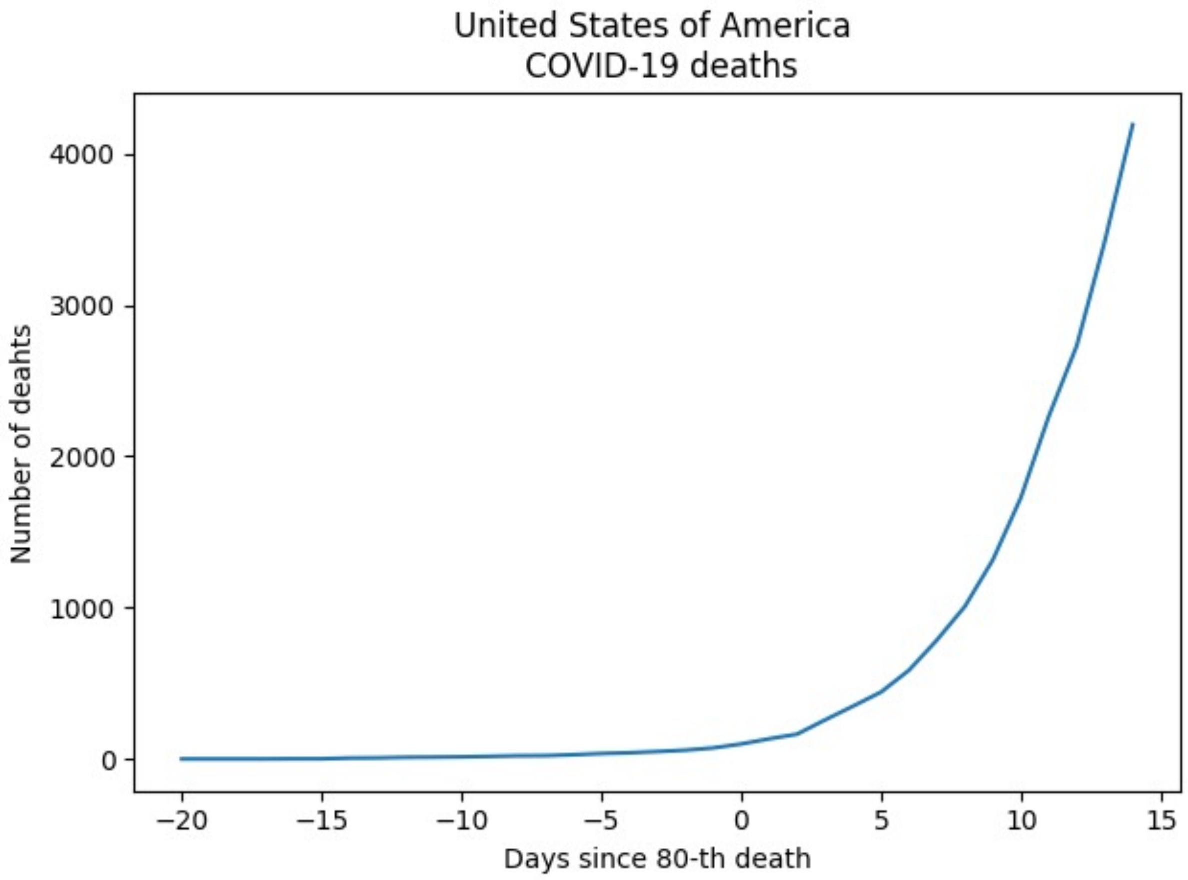
Synthetic Interventions: https://arxiv.org/abs/2006.07691

Causal Tensor Estimation: working paper

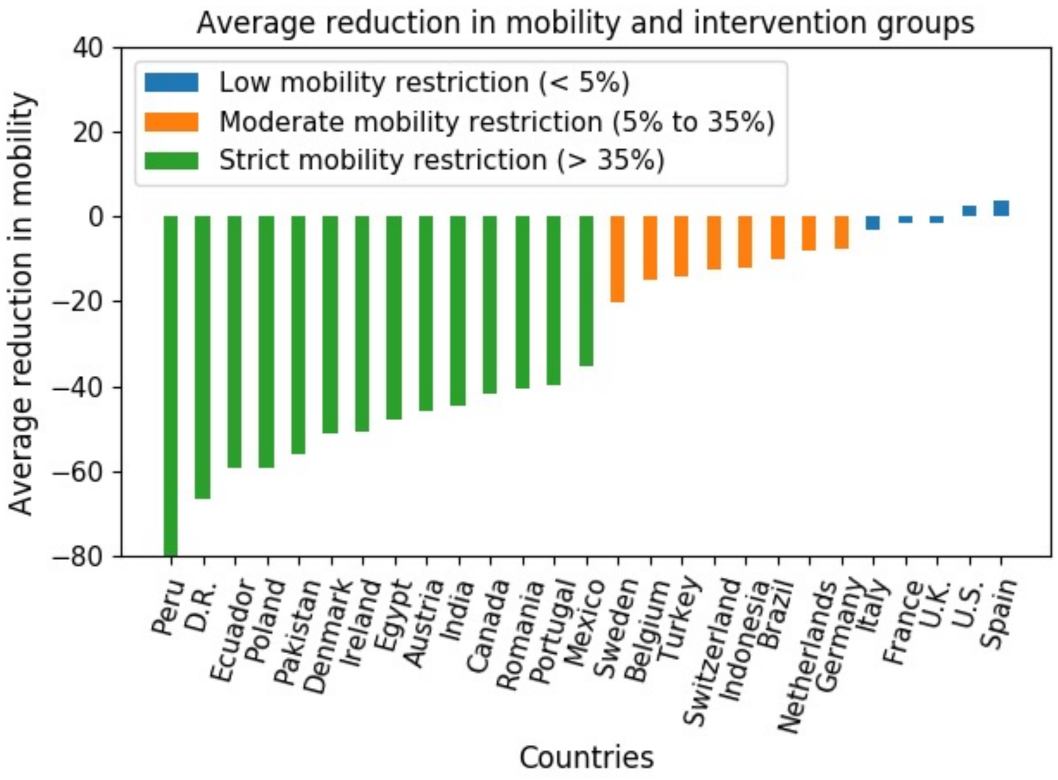
Policy Evaluation

Anish Agarwal Abdullah Alomar Arnab Sarkar Dennis Shen

United States



Looking Across the Globe



Low mobility restriction

< 5% reduction

Moderate mobility restriction

5 - 35% reduction

Severe mobility restriction

> 35% reduction

What would have happened to United States if…

- United States had experienced (through appropriate policy)
  - Low mobility restriction
    - < 5% reduction [the reality]
  - Moderate mobility restriction
    - 5-35% reduction
  - Severe mobility restriction
    - > 35% reduction

- In terms of
  - Trajectory of death counts in region of interest

It's Causal Inference

Potential Outcomes Framework [Newman '23, Rubin '74]

An individual contains many latent selves

Low

Moderate

High

$Y =$ observed outcome (health outcome)

$M^{(d)} =$ potential outcome under intervention d (health outcome under policy d)

$d^* =$ observed intervention (low restrictions), that is $Y \stackrel{\mathbb{E}}{=} M^{d^*}$

Goal: estimate $M^{(d)},\ d \neq d^*$

Fundamental Question

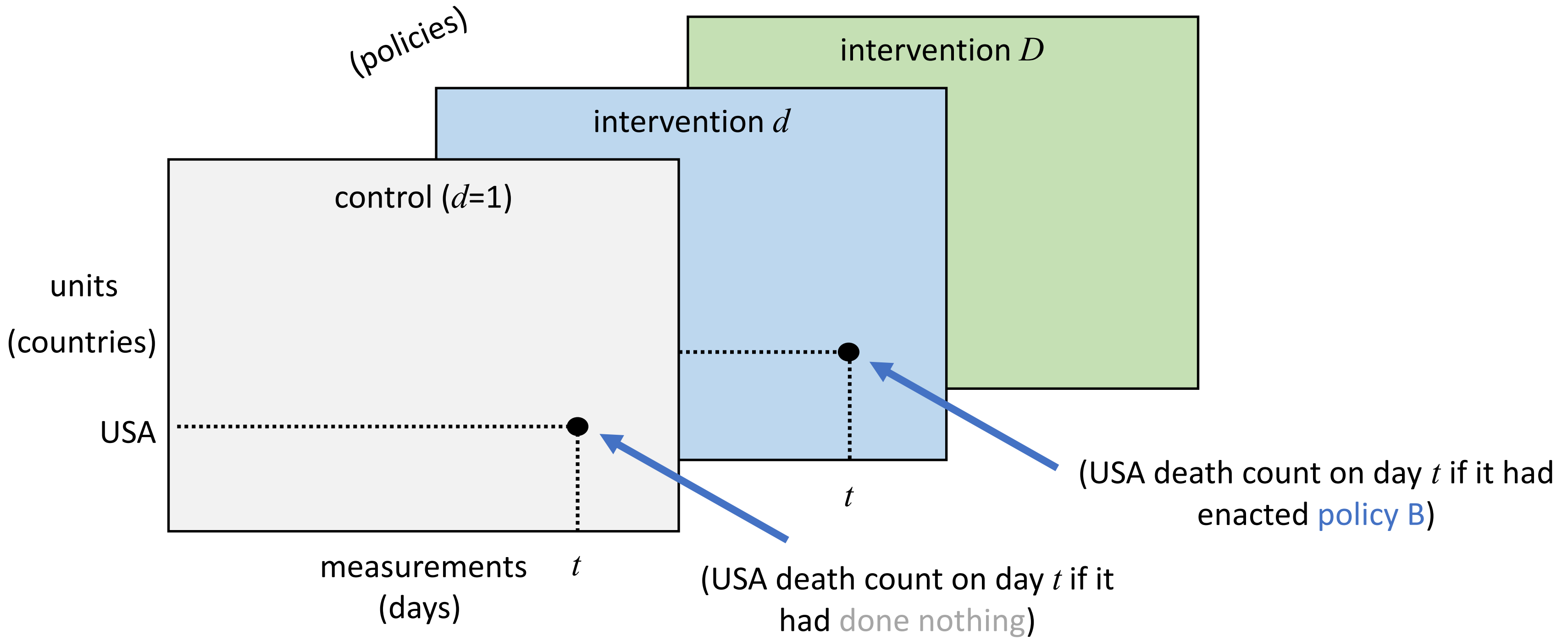
Only one outcome can be revealed
But want to know *all possible* outcomes

Let's Look At An Alternative Representation: Tensor

Encode potential outcomes into a Tensor

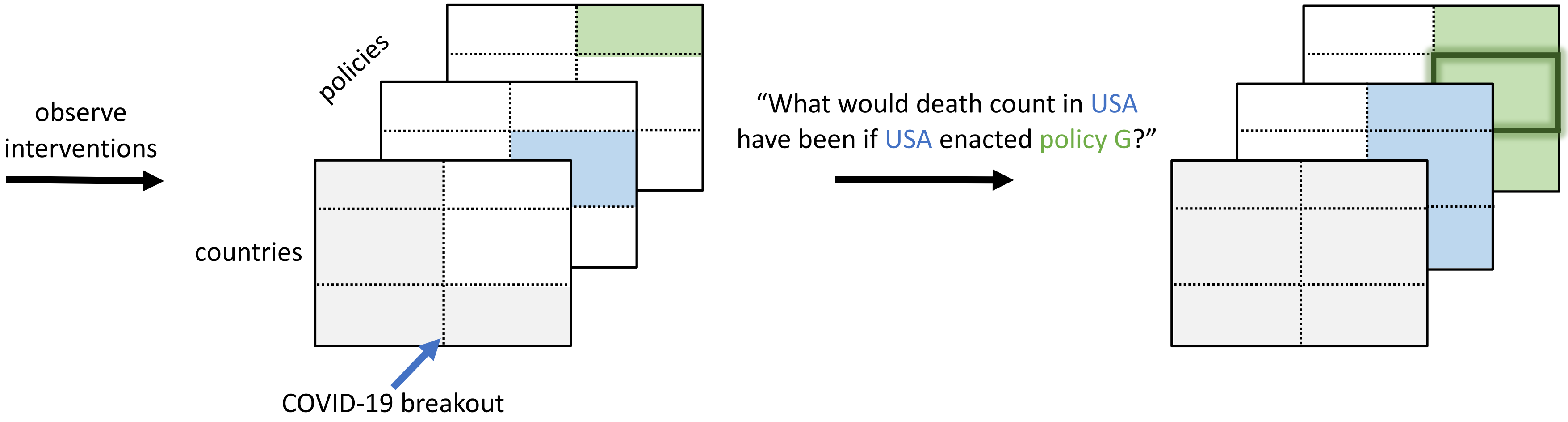
N units, T measurements, D interventions

(n, t, d)



Causal Inference = Causal Tensor Estimation

Policy Evaluation



- Causal Tensor Estimation
  - "Imputing" missing values in a Tensor
  - Potentially "confounded" observations (e.g. not missing at random)
    - The policy implemented in a country depends on the "characteristics" of the country!

What is Confounding, Why Is it a Problem

- To determine A vs B:
  - Access to 100 M + 100 W patients

- Randomized Trial
  - 50 M, 50 W receive A (similarly B)
  - Average efficacy: 5 for A and 10 for B
  - Conclusion: B is better than A

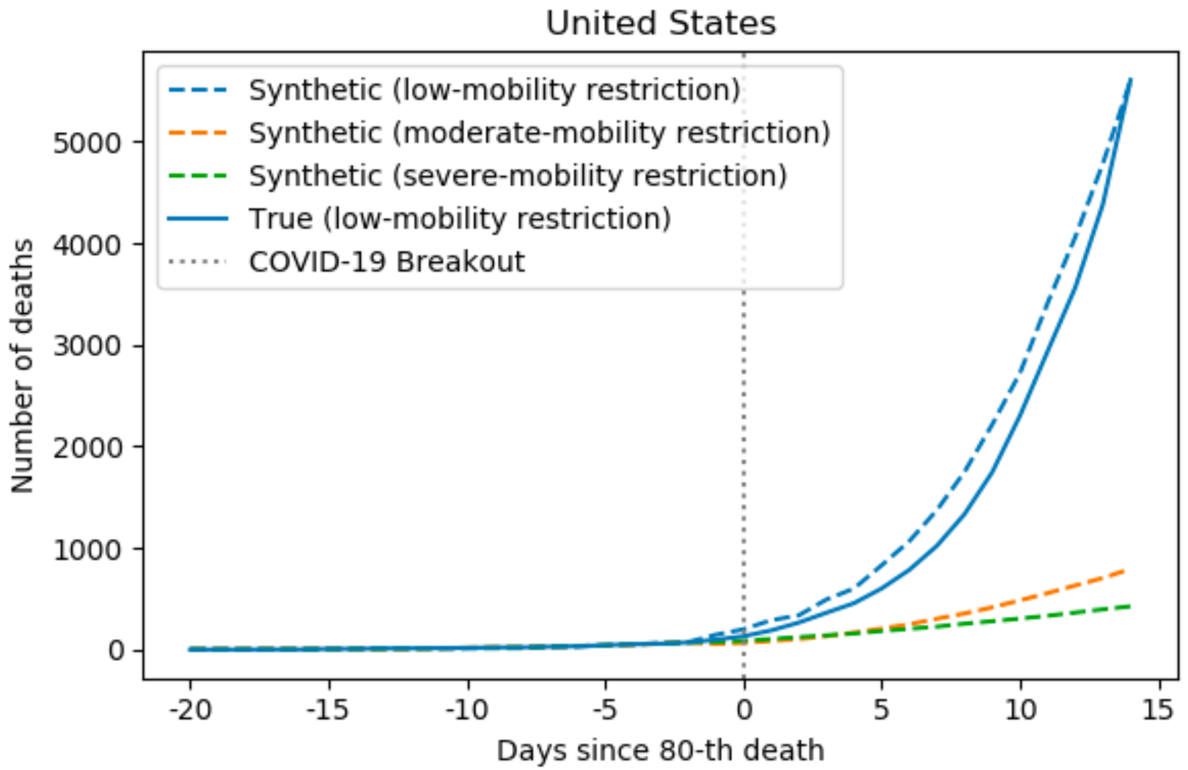
- Observational data (“confounded” selection)
  - 100 M get A, 100 W get B
  - Average efficacy: 10 for A and 0 for B
  - Conclusion: A is better than B

Ground Truth or Potential Outcome

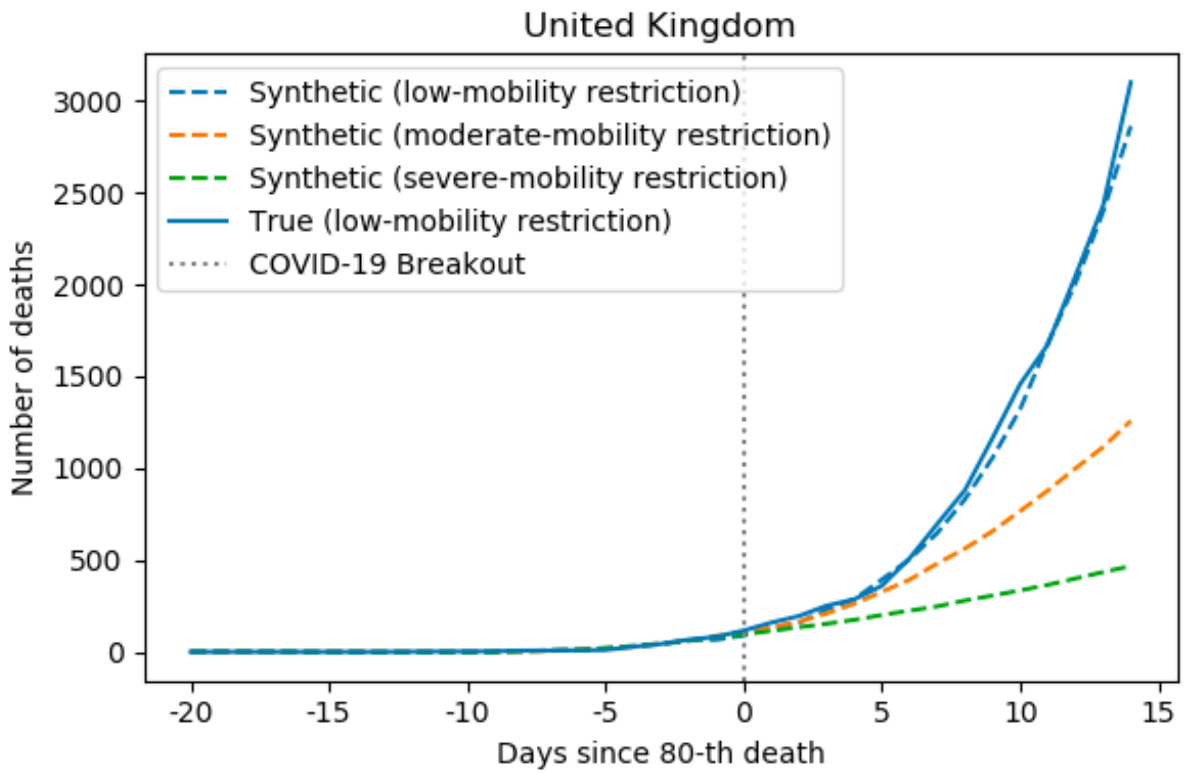
Efficacy of Two Drugs Across Gender

	Men	Women
Drug A	**10**	**0**
Drug B	**20**	**0**

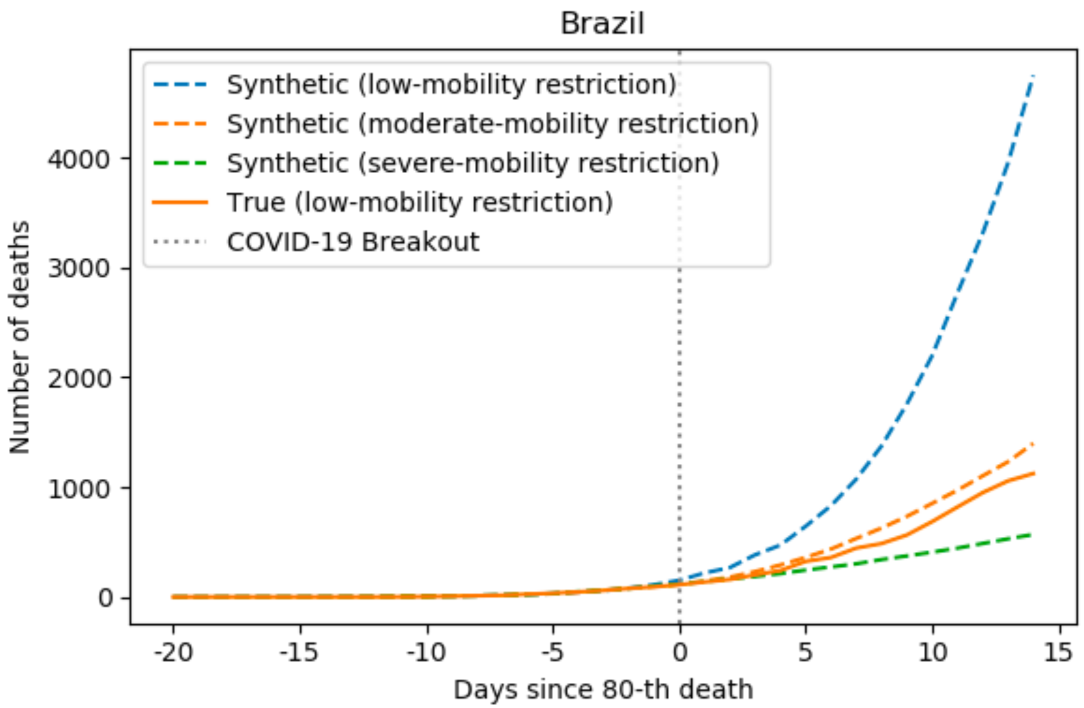
United States: Causal Tensor Estimation w Synthetic Interventions

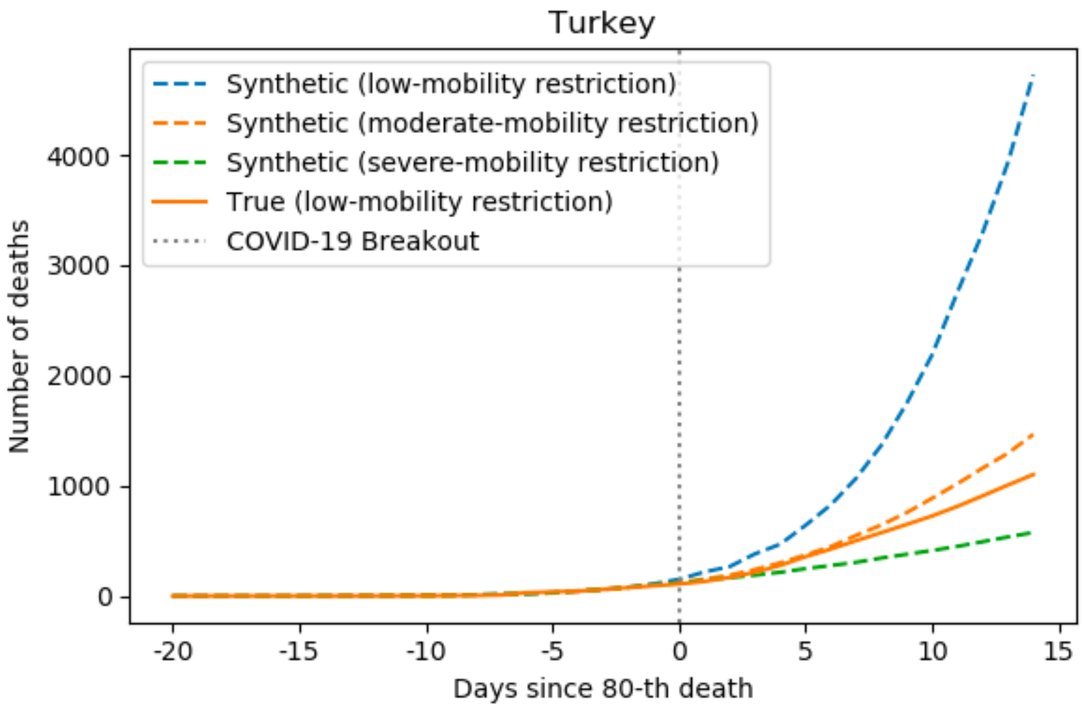


United Kingdom: Causal Tensor Estimation w Synthetic Interventions

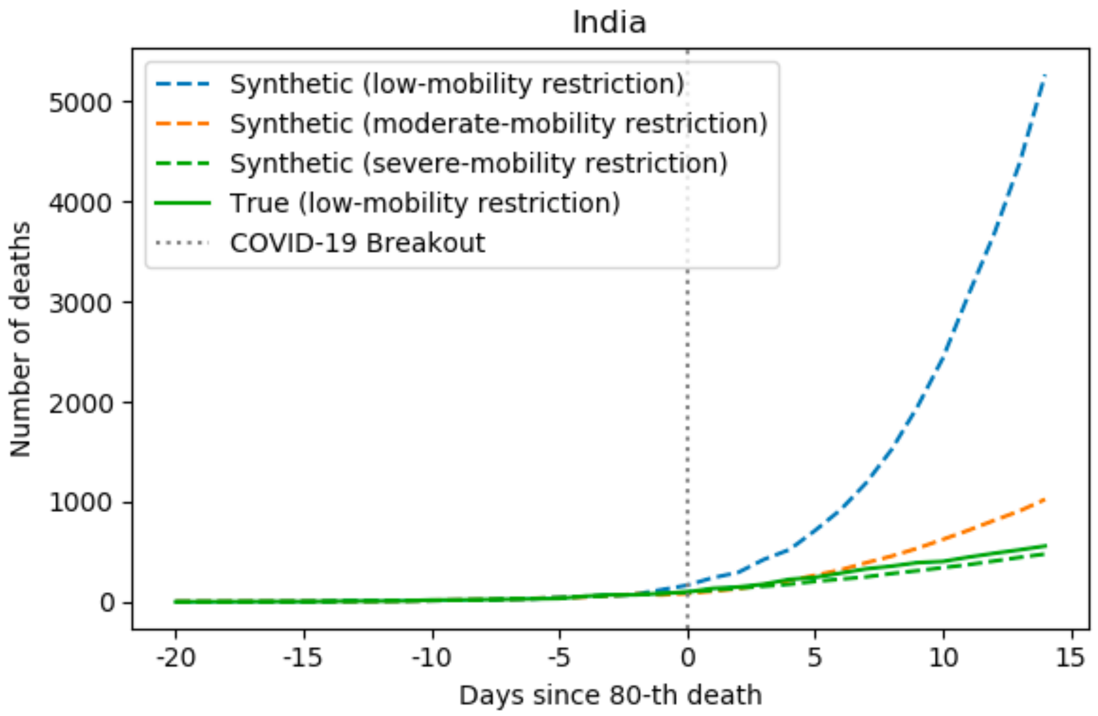


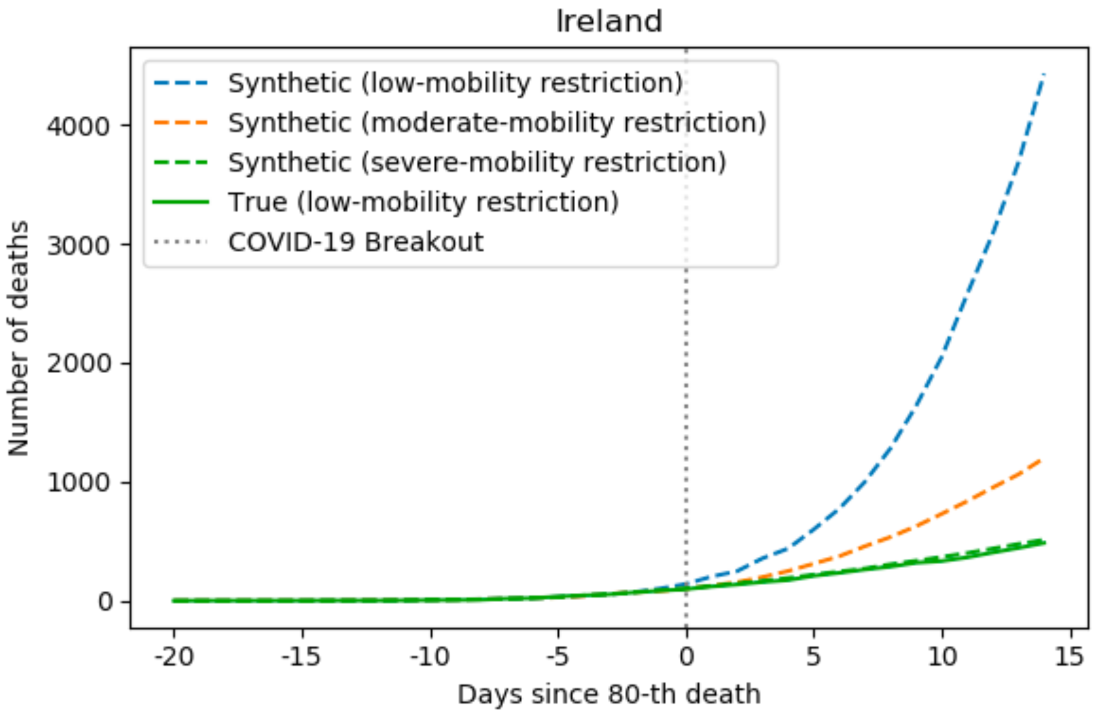
Brazil, Turkey: Causal Tensor Estimation w Synthetic Interventions





India, Ireland: Causal Tensor Estimation w Synthetic Interventions





Data Efficient Randomized Control

Anish Agarwal Vishal Misra Dennis Shen

Clinical Trial For Personalized Treatment

6 patient types
(N=6)

3 drugs
(1 placebo) (D=4)

Intervention	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
placebo	✓	✓	✓	✓	✓	✓
drug 1	✓	✓	✓	✓	✓	✓
drug 2	✓	✓	✓	✓	✓	✓
drug 3	✓	✓	✓	✓	✓	✓

Clinical Trial For Personalized Treatment

Real clinical trial: $D \times N$

Intervention	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
placebo	✓	✓	✓	✓	✓	✓
drug 1	✓	✓	✓	✓	✓	✓
drug 2	✓	✓	✓	✓	✓	✓
drug 3	✓	✓	✓	✓	✓	✓

[the reality]

Data-efficient clinical trial: $2 \times N$

Intervention	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
placebo	✓	✓	✓	✓	✓	✓
drug 1	✓	✓				
drug 2			✓	✓		
drug 3					✓	✓

[our proposal]

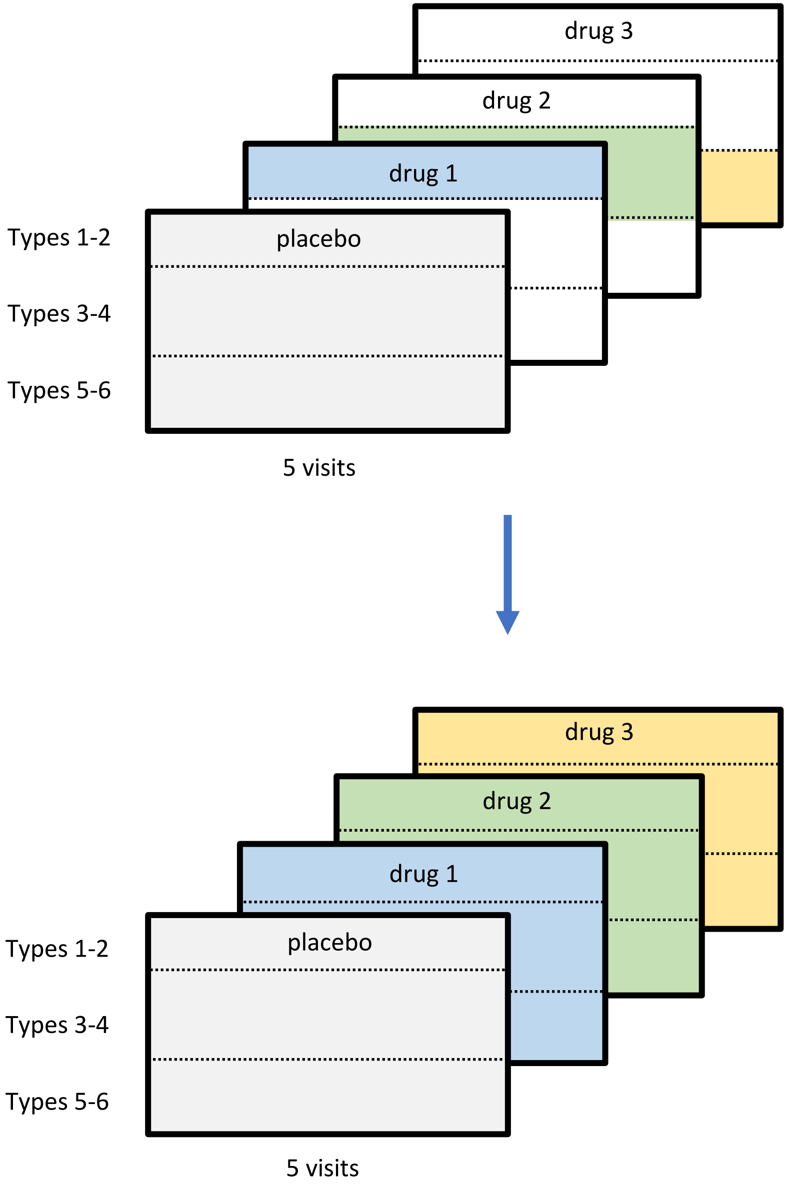
Clinical Trial For Personalized Treatment = Tensor Estimation

Real clinical trial: $D \times N$

Intervention	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
placebo	✓	✓	✓	✓	✓	✓
drug 1	✓	✓	✓	✓	✓	✓
drug 2	✓	✓	✓	✓	✓	✓
drug 3	✓	✓	✓	✓	✓	✓

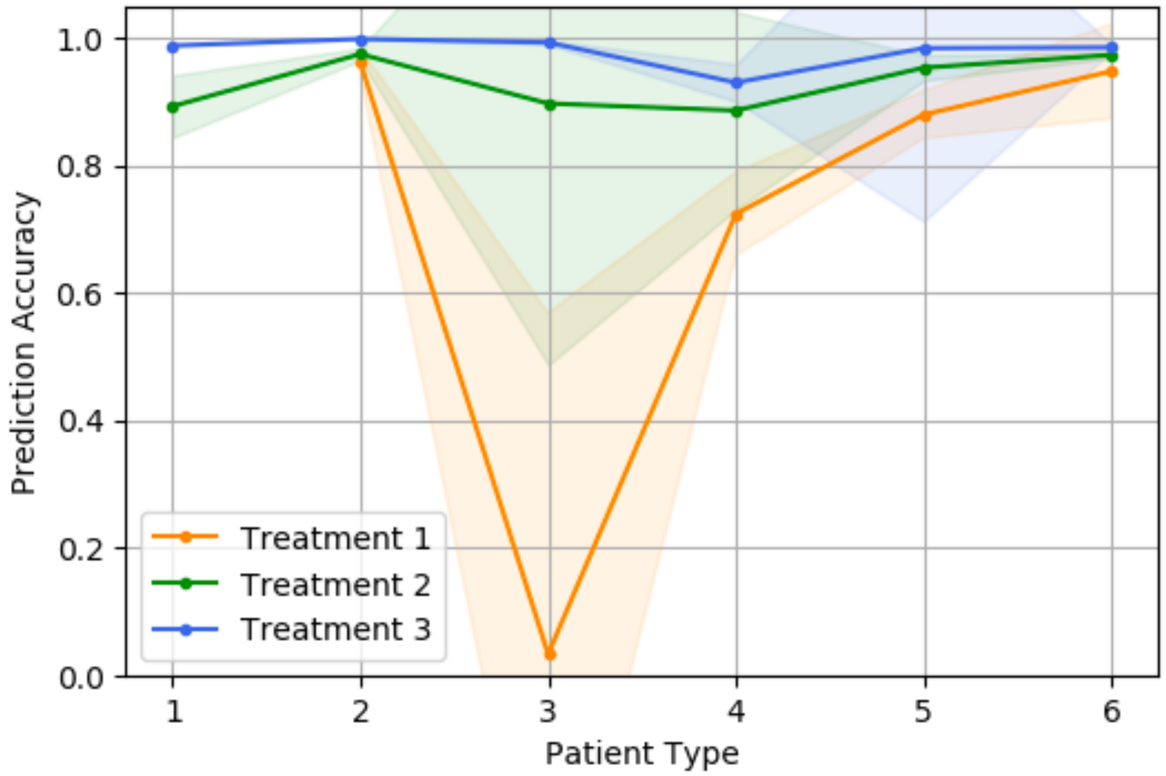
Data-efficient clinical trial: $2 \times N$

Intervention	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
placebo	✓	✓	✓	✓	✓	✓
drug 1	✓	✓				
drug 2			✓	✓		
drug 3					✓	✓



- Causal Tensor Estimation
  - Estimate outcomes for every (patient type, drug)
  - Using partial observations (no confounding)

Tensor Estimation Using Synthetic Interventions

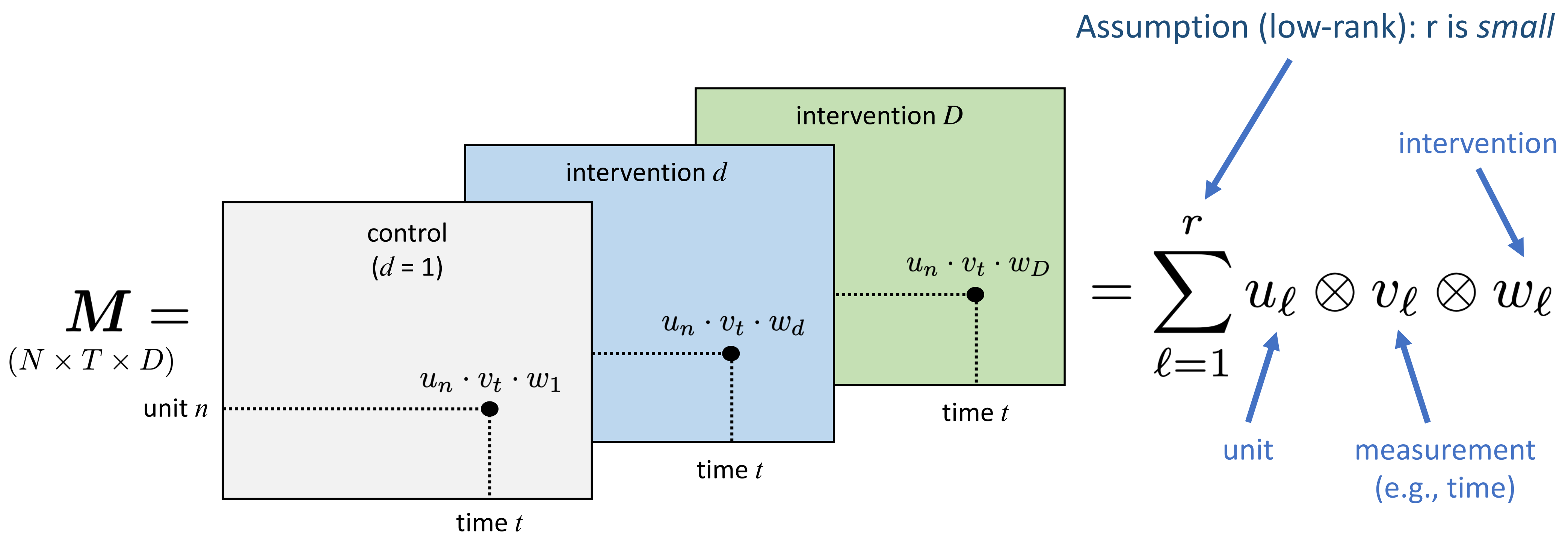


Accurately predicts outcome of 6 x 4 trials using only 6 x 2 trials

Framework: Causal Tensor Estimation

Alberto Abadie   Anish Agarwal   Dennis Shen

Potential Outcomes Tensor

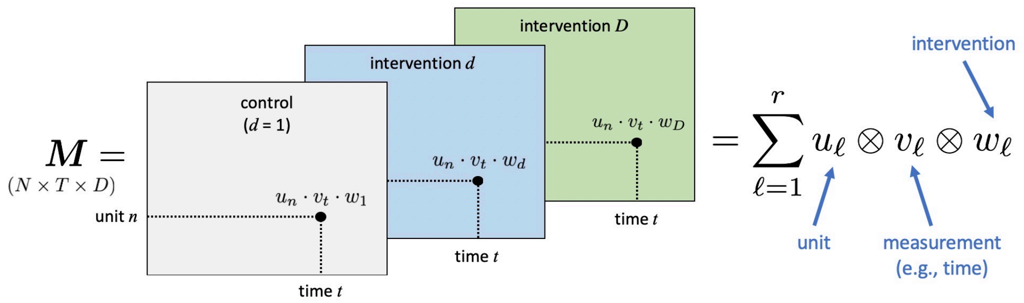


The Model

1. Sample (or given) latent unit, time, intervention factors

$$(u_n, v_t, w_d)$$

2. Sample potential outcomes tensor



3. Sample treatment assignment (determines sparsity pattern of observed tensor)

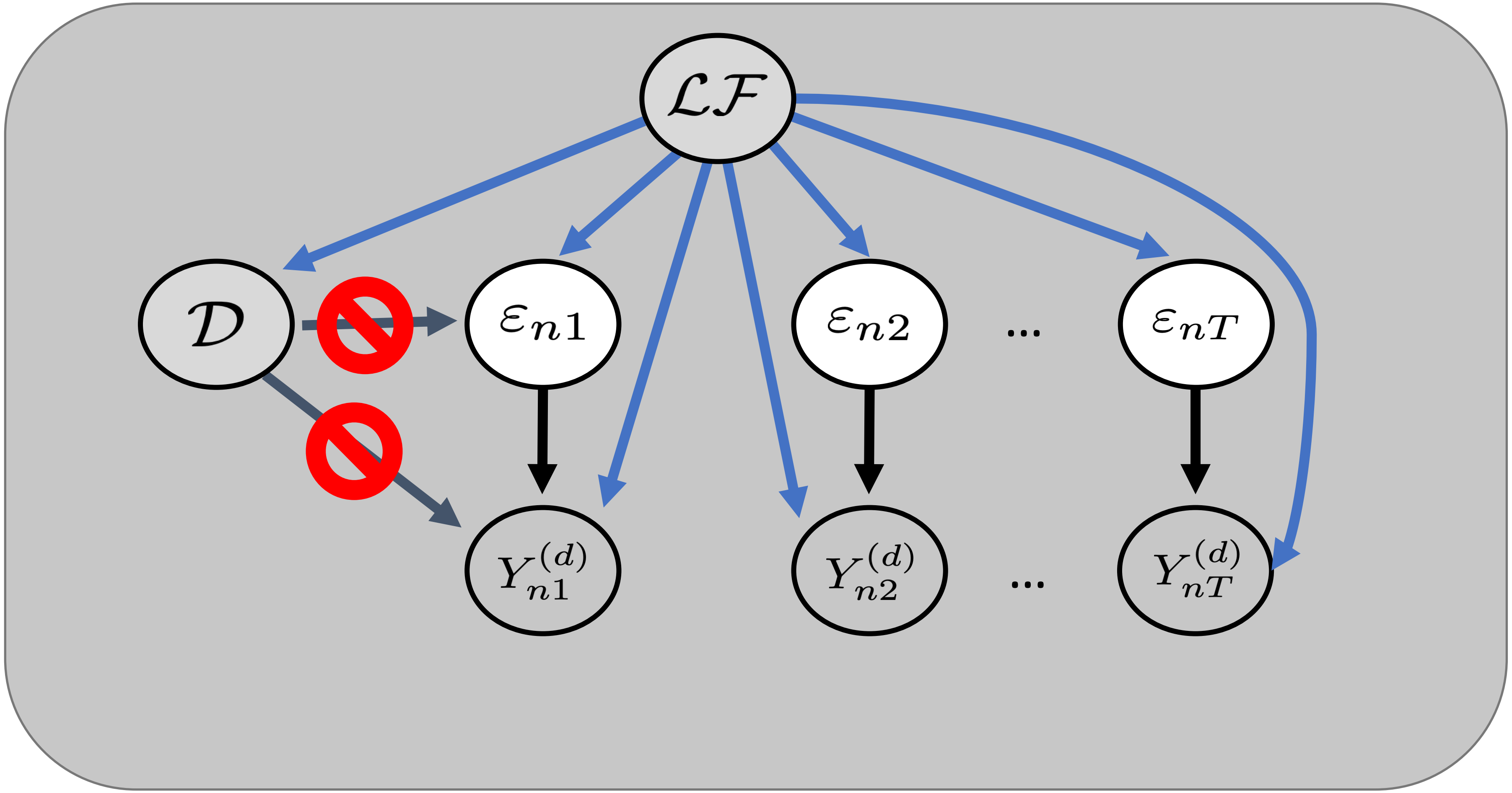
$$D(n,t) : [N] \times [T] \to 2^{[D]}$$

4. Observe noisy measurements: sampled entries of

$$Y = M + \varepsilon$$

What Type of Confounding is Allowed?

The joint distribution of latent factors (confounders, covariates), treatment assignment and observations satisfy the following Causal Structure



What Type of Confounding is Allowed?

- Recall

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

- Why is there confounding?

$$\mathcal{D} \not\perp\!\!\!\perp Y_{nt}^{(d)}$$

  - Treatment assignments correlated with latent factors (i.e. unmeasured confounders)

- Selection on Latent Factors

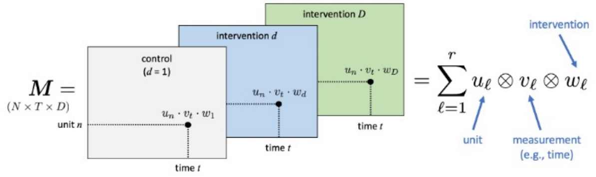
$$\mathcal{D} \perp\!\!\!\perp Y_{nt}^{(d)} \mid \mathcal{LF}$$

Causal Tensor Estimation

1. Sample (or given) latent unit, time, intervention factors

$$(u_n, v_t, w_d)$$

2. Sample potential outcomes tensor



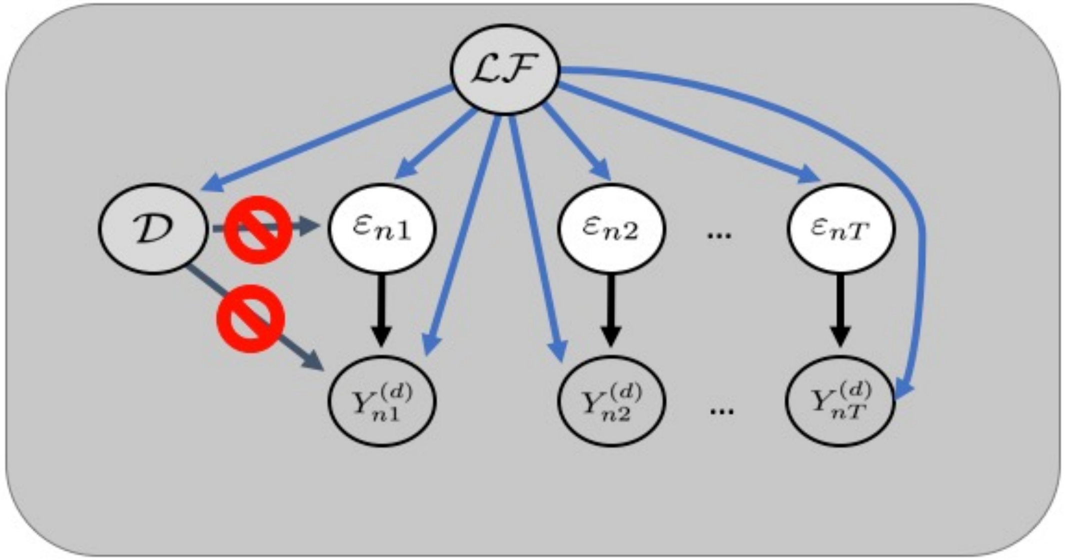
3. Sample treatment assignment (determines sparsity pattern of observed tensor)

$$D(n,t) : [N] \times [T] \to 2^{[D]}$$

4. Observe noisy measurements: sampled entries of

$$Y = M + \varepsilon$$

+



Produce an estimate $\widehat{M}$ of M so that

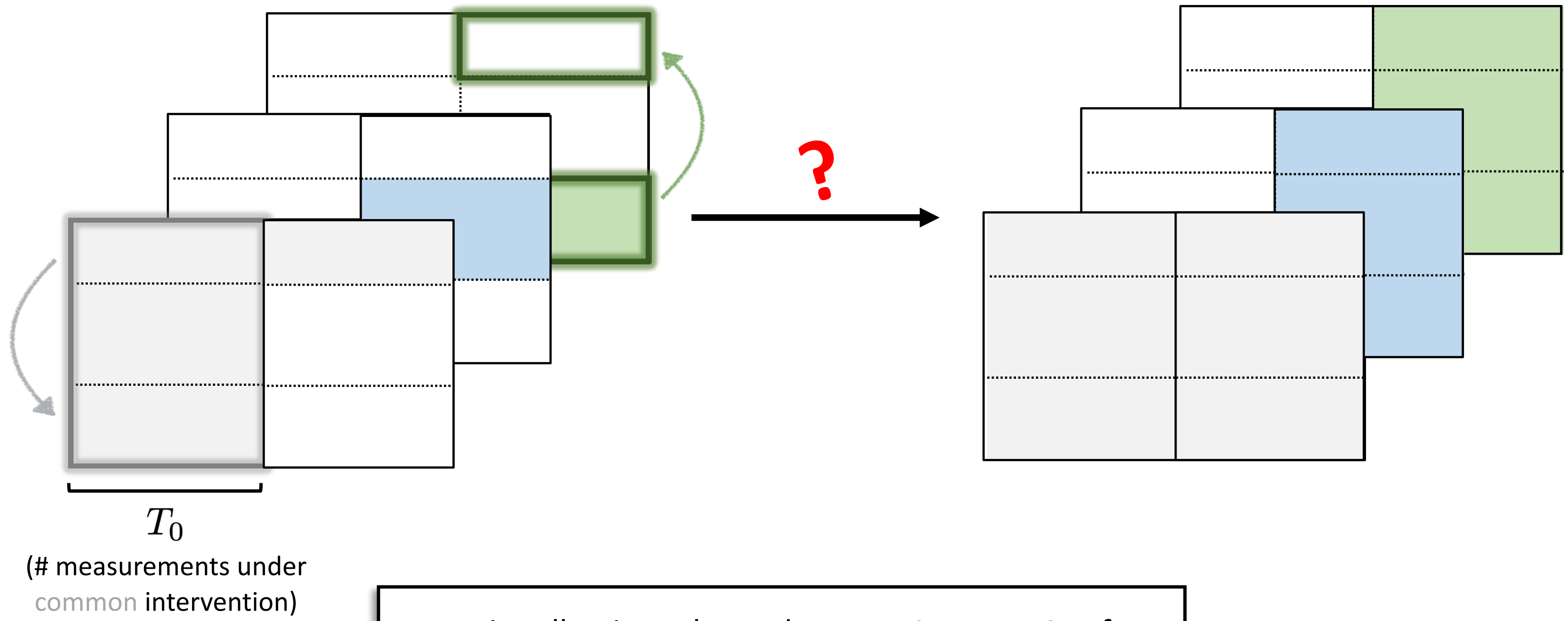
$$\widehat{M} \approx M$$

A Method: Synthetic Interventions

Anish Agarwal   Dennis Shen

Key Insight

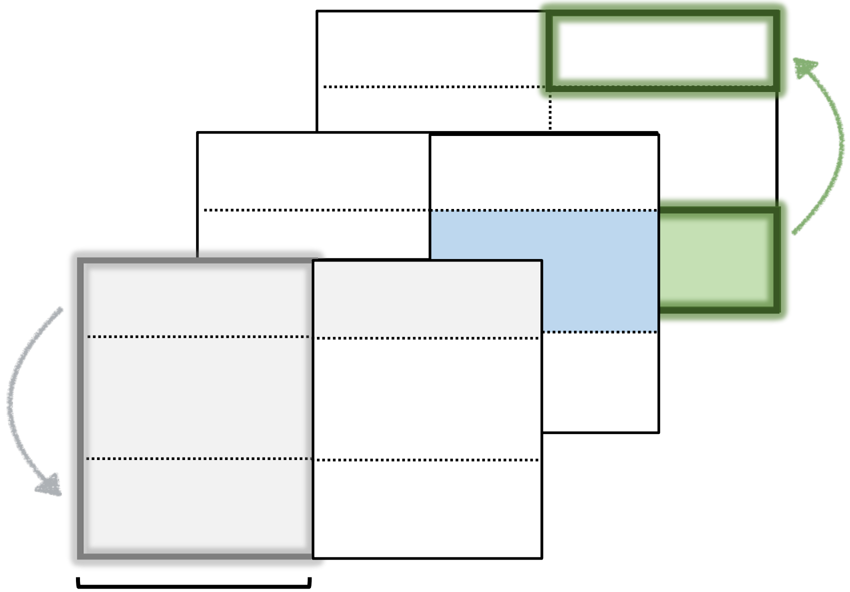
leverage data from other units
learn relationships between units



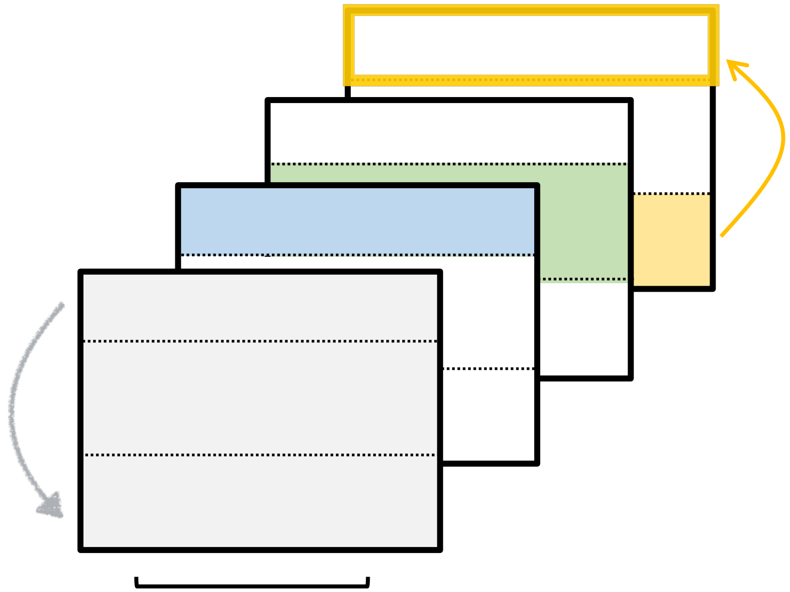
require all units to be under same intervention for some number of measurements

Key Insight

leverage data from other units
learn relationships between units



T_0
(# measurements under common intervention)



T_0
(# measurements under common intervention)

require all units to be under same intervention for some number of measurements

Synthetic Control (SC)

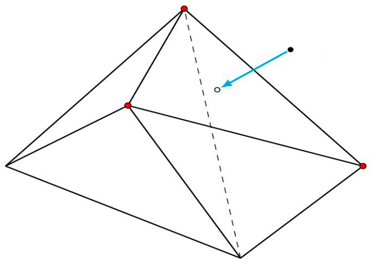
Estimates counterfactuals in absence of intervention

“What would US’s death count have been if US did nothing?”

1. Learn Model
under control

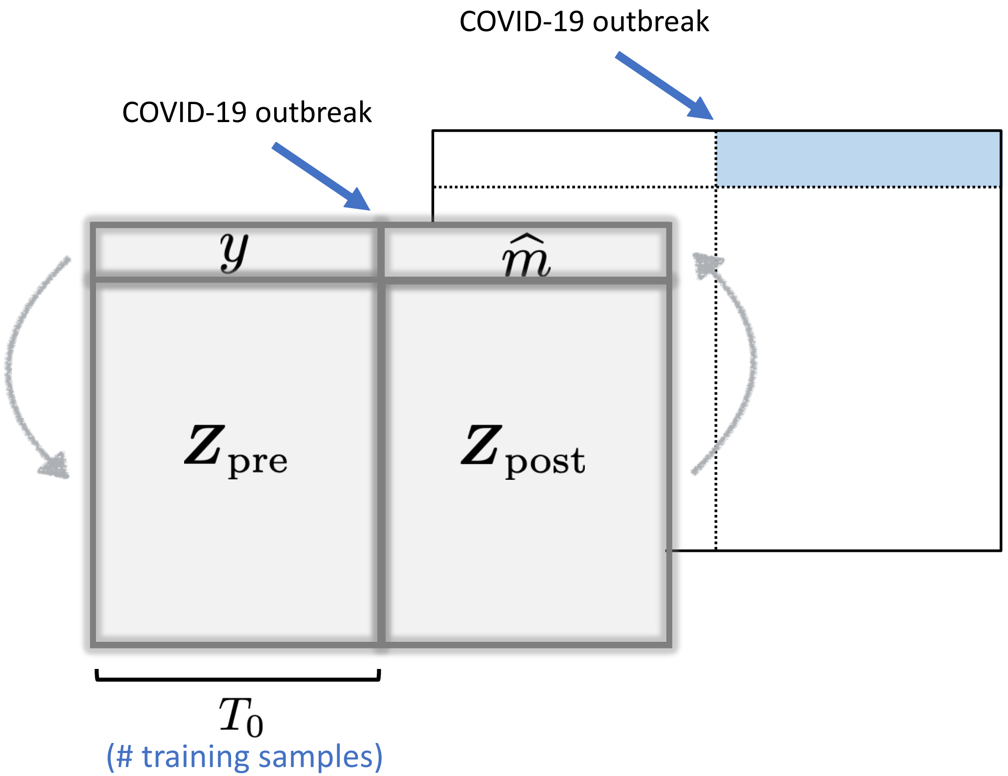
$$\widehat{\beta} = \underset{w\in\mathcal{W}}{\operatorname{argmin}} \; \|y - \boldsymbol{Z}_{\text{pre}} w\|_2^2$$

“synthetic” control



2. Predict
under control

$$\widehat{m} = \boldsymbol{Z}_{\text{post}} \widehat{\beta}$$



“And Other Countries did nothing?”

[Abadie et al ‘03, ‘10] [Abadie, A. (2020). Using synthetic controls: Feasibility, data requirements, and methodological aspects. Journal of Economic Literature]

Synthetic Interventions (SI)

Estimates counterfactuals in absence and presence of interventions

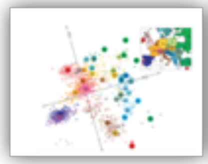
"What would US's death count have been if it enacted low-mobility restricting policies?"

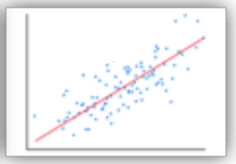
"What would US's death count have been if it enacted severe-mobility restricting policies?"

1. Learn Model
under control

$$\widehat{\beta} = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \; \|y - \boldsymbol{Z}_{\text{pre}} w\|_2^2$$

"synthetic" intervention

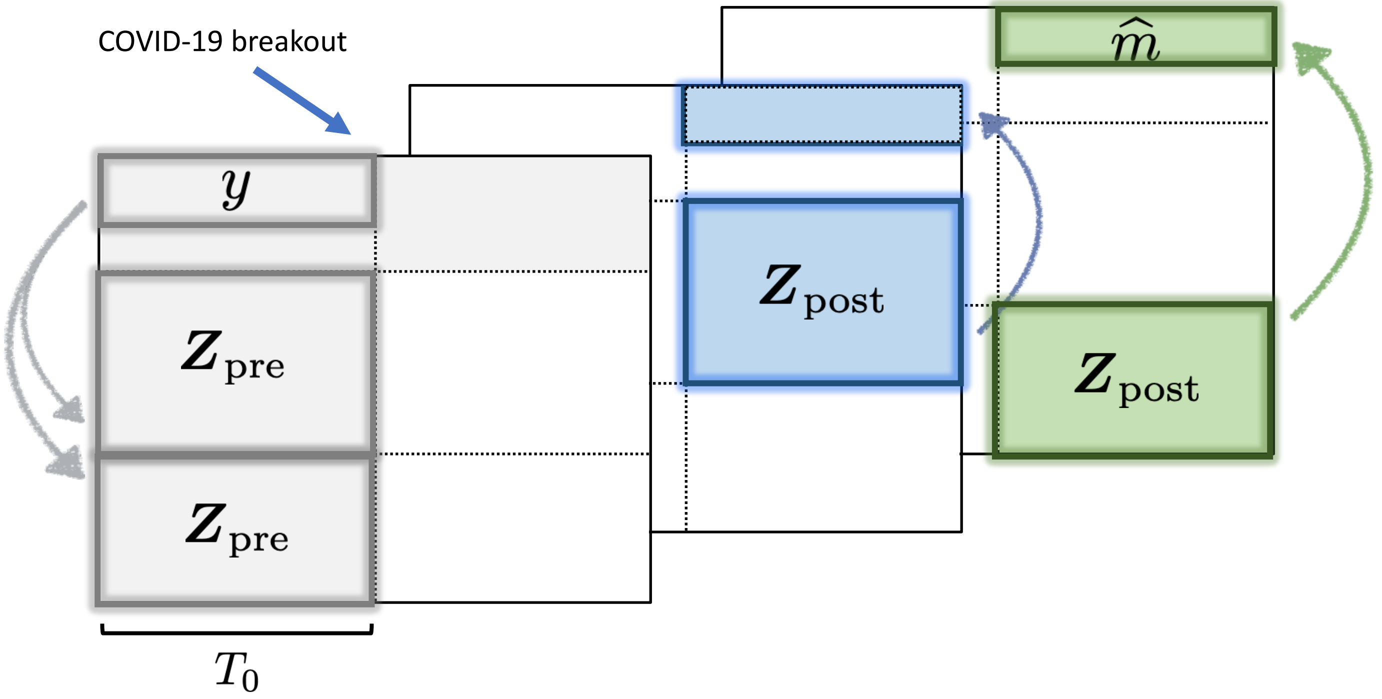




2. Predict
under intervention

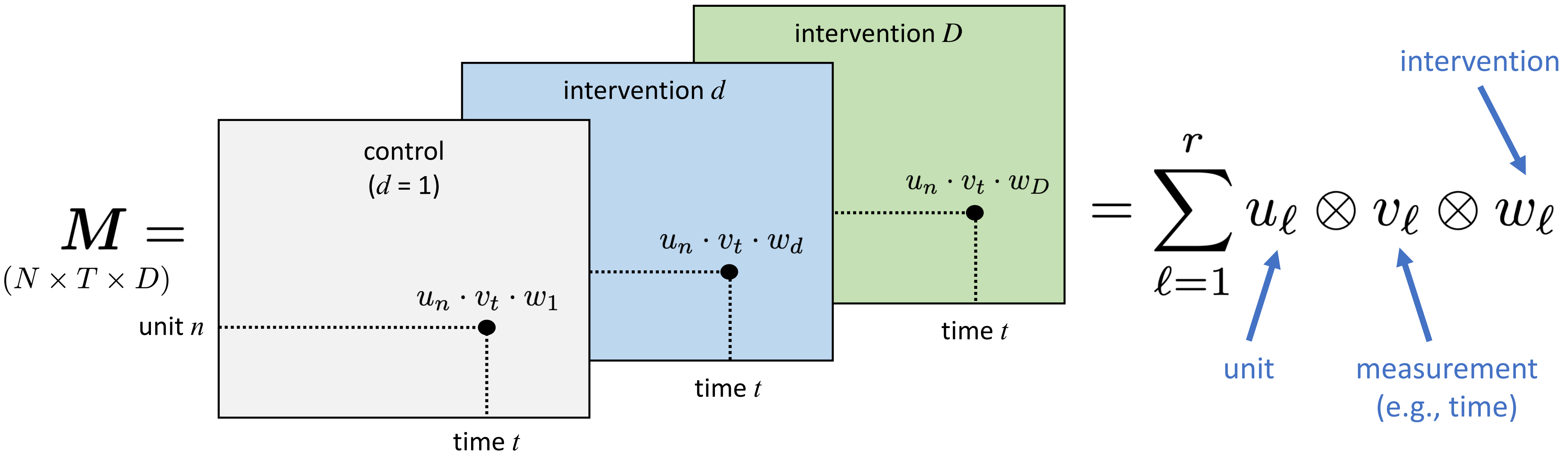
$$\widehat{m} = \boldsymbol{Z}_{\text{post}} \widehat{\beta}$$

(de-noise via matrix estimation)



Why can we transfer learned model between different interventional frameworks?

Why does SI Work?



Under intervention d

$$\boldsymbol{M}^{(d)} = \sum_{\ell=1}^{r} u_\ell \otimes (w_{d\ell} \cdot v_\ell) = \boldsymbol{U}(\boldsymbol{V}^{(d)})^T$$

- $\boldsymbol{U}$ describes an **invariant** relationship between units across interventions
- **Each** intervention d is a linear transformation of $\boldsymbol{U}$
- **SI** learns linear relationship between rows of $\boldsymbol{U}$

Why does SI Work?

(WLOG) suppose unit 1 satisfies:

$$\boxed{u_{1\ell} = \sum_{n>1} \beta_n^* \cdot u_{n\ell}}$$

(occurs w.h.p.)

(low rank = few canonical unit profiles)

$$M_{1t}^{(d)} = \sum_{\ell=1}^{r} u_{1\ell} \cdot v_{t\ell} \cdot w_{d\ell}$$ for any (t,d)

(via tensor factor model)

$$= \sum_{\ell=1}^{r} \sum_{n>1} \beta_n^* \cdot u_{n\ell} \cdot v_{t\ell} \cdot w_{d\ell}$$

(via assumption)

$$= \sum_{n>1} \beta_n^* \cdot M_{nt}^{(d)}$$

invariant across
time, interventions

SI (and thus SC) exists

Identification, Consistency

$\theta_n^{(d)}$ = individual potential outcome under every intervention averaged over post-intervention period

$$= \frac{1}{T_1} \sum_{t > T_0} \mathbb{E}[Y_{nt}^{(d)} \mid \{u_t^{(d)}, v_n : t > T_0\}]$$

$$\widehat{\theta}_n^{(d)} - \theta_n^{(d)} = \mathcal{O}_p\left(\frac{1}{T_0^{1/4}} + \frac{\|\tilde{w}^{(n,d)}\|_2}{\sqrt{T_1}} + \frac{\|\tilde{w}^{(n,d)}\|_1}{\min\{\sqrt{T_0}, \sqrt{N_d}\}} \right)$$

Normality

$$\boxed{\sqrt{T_1}(\widehat{\theta}_n^{(d)} - \theta_n^{(d)}) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \mathrm{plim} \|\tilde{w}^{(n,d)}\|_2^2\right)}$$

95% confidence interval

$$\theta_n^{(d)} \in \left[\widehat{\theta}_n^{(d)} \pm \frac{1.96\,\widehat{\sigma}\|\widehat{w}^{(n,d)}\|_2}{\sqrt{T_1}}\right]$$

Computable quantities
(with provable guarantees)

Subspace Inclusion: Hypothesis Test

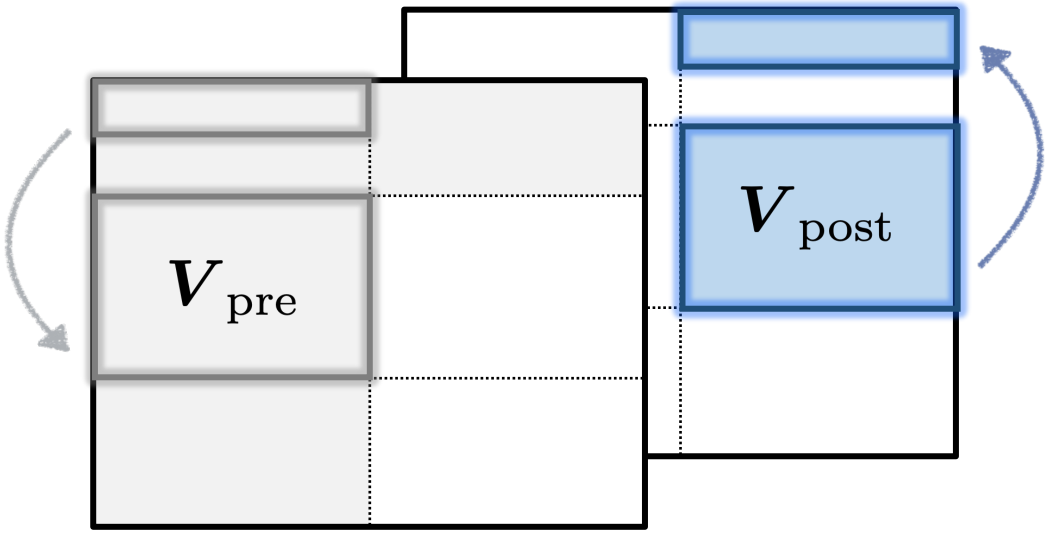
$H_0 : \mathrm{span}(\boldsymbol{V}_{\mathrm{post}}) \subseteq \mathrm{span}(\boldsymbol{V}_{\mathrm{pre}})$
$H_1 : \mathrm{span}(\boldsymbol{V}_{\mathrm{post}}) \not\subseteq \mathrm{span}(\boldsymbol{V}_{\mathrm{pre}})$

If H_0 holds:

$$\|(\boldsymbol{I} - \boldsymbol{V}_{\mathrm{pre}}\boldsymbol{V}_{\mathrm{pre}}^T)\boldsymbol{V}_{\mathrm{post}}\|_F^2 = 0$$

If H_1 holds:

$$\|(\boldsymbol{I} - \boldsymbol{V}_{\mathrm{pre}}\boldsymbol{V}_{\mathrm{pre}}^T)\boldsymbol{V}_{\mathrm{post}}\|_F^2 > 0$$



Subspace Inclusion: Hypothesis Test

$H_0 : \operatorname{span}(\boldsymbol{V}_{\text{post}}) \subseteq \operatorname{span}(\boldsymbol{V}_{\text{pre}})$

$H_1 : \operatorname{span}(\boldsymbol{V}_{\text{post}}) \not\subseteq \operatorname{span}(\boldsymbol{V}_{\text{pre}})$

Test statistic

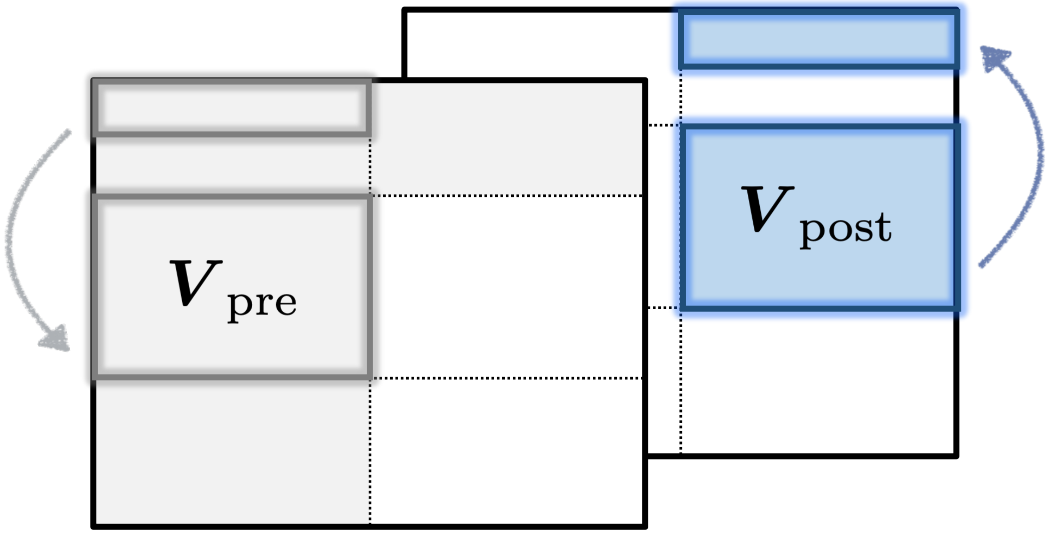
$$\widehat{\tau} = \|(\boldsymbol{I} - \widehat{\boldsymbol{V}}_{\text{pre}}\widehat{\boldsymbol{V}}_{\text{pre}}^T)\widehat{\boldsymbol{V}}_{\text{post}}\|_F^2$$

Test

For any significance level $\alpha \in (0,1)$

Retain H_0 if $\widehat{\tau} \leq \tau(\alpha)$

Reject H_0 if $\widehat{\tau} > \tau(\alpha)$



Subspace Inclusion: Type I & Type II Error Guarantees

Fix any $\alpha \in (0,1)$

Type I error: $\mathbb{P}(\widehat{\tau} > \tau(\alpha) | H_0) \leq \alpha$

Type II error: $\mathbb{P}(\widehat{\tau} \leq \tau(\alpha) | H_1) \leq \alpha$

$$\text{where } \tau(\alpha) = \mathcal{O}\left(\frac{\sqrt{\log(1/\alpha)}}{\min\{\sqrt{T_0}, \sqrt{T_1}, \sqrt{N_d}\}} \right)$$

Parting Remarks

Statistical & Computational Tradeoffs in Causal Inference

Block Sparsity

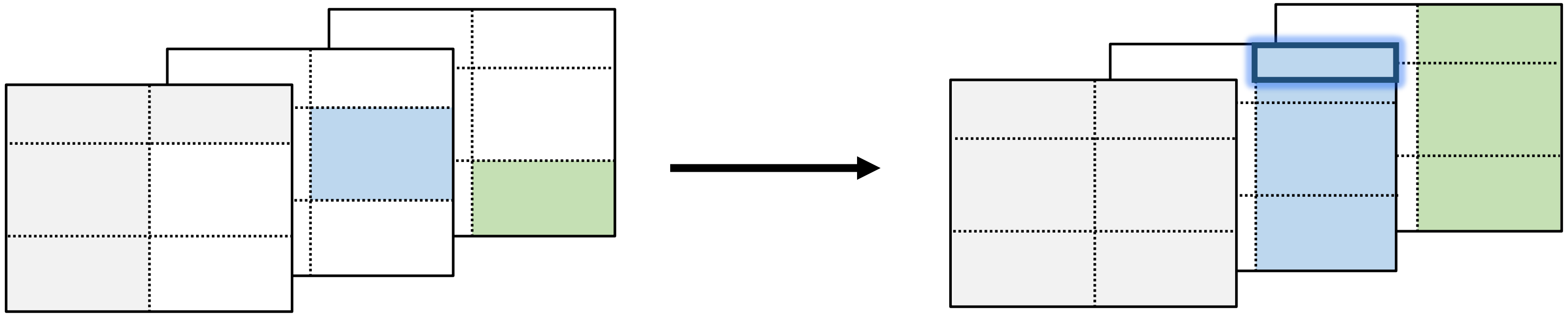
- Sample complexity:

$$\Omega(\sqrt{r} \cdot n^{3/2})$$

- Computational complexity:

$$\mathcal{O}(n^3)$$

(PCR)



$(n \times n \times n)$ $\quad$ $(n \times cn \times n), \ c \leq 1$

Start discussion about computational/statistical trade-offs
New experimental designs via sampling?

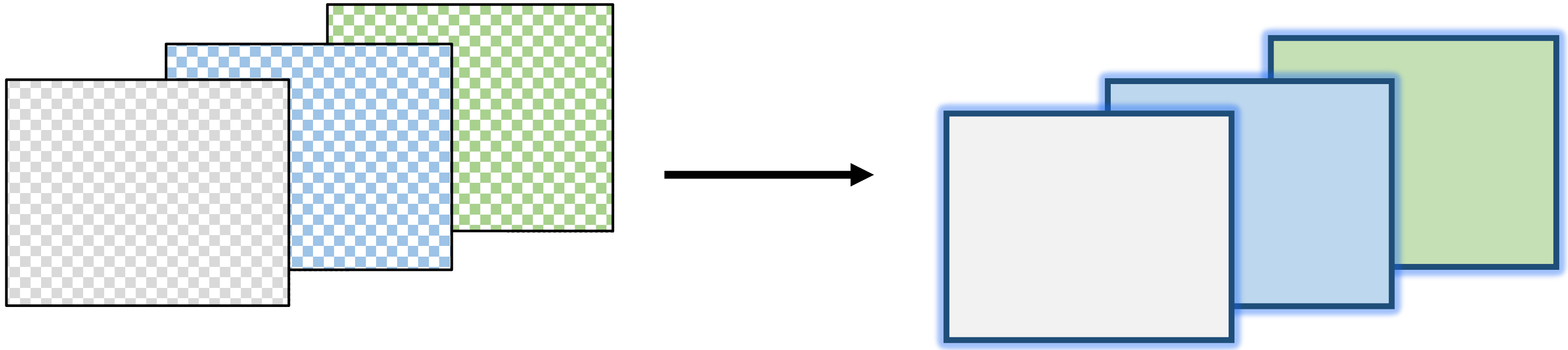
Uniform Sparsity

- Sample complexity:

$$\Omega(r \cdot n^{3/2})$$

- Computational complexity:

$$\mathcal{O}(\text{poly}(n))$$



$(n \times n \times n)$ $\quad$ $(n \times n \times n)$

Causal Tensor Estimation: A Generic Framework

- Enables novel estimation
  - Regression discontinuity design in the panel data setting
- Experiment design
  - Observational pattern in tensor to enable identification
- Computational and statistical tradeoff
  - A missing discussion in Causal inference
- Role of error metric for tensor estimation
  - What causal quantities can be identified (or not)
- Causal estimation methods
  - SI is one such method, but more is needed

Questions

+ please feel free to contact at:
devavrat@mit.edu

Appendix

Development Economics

Development Economics [Banerjee et al 2019]

Experimental Study

(unit)

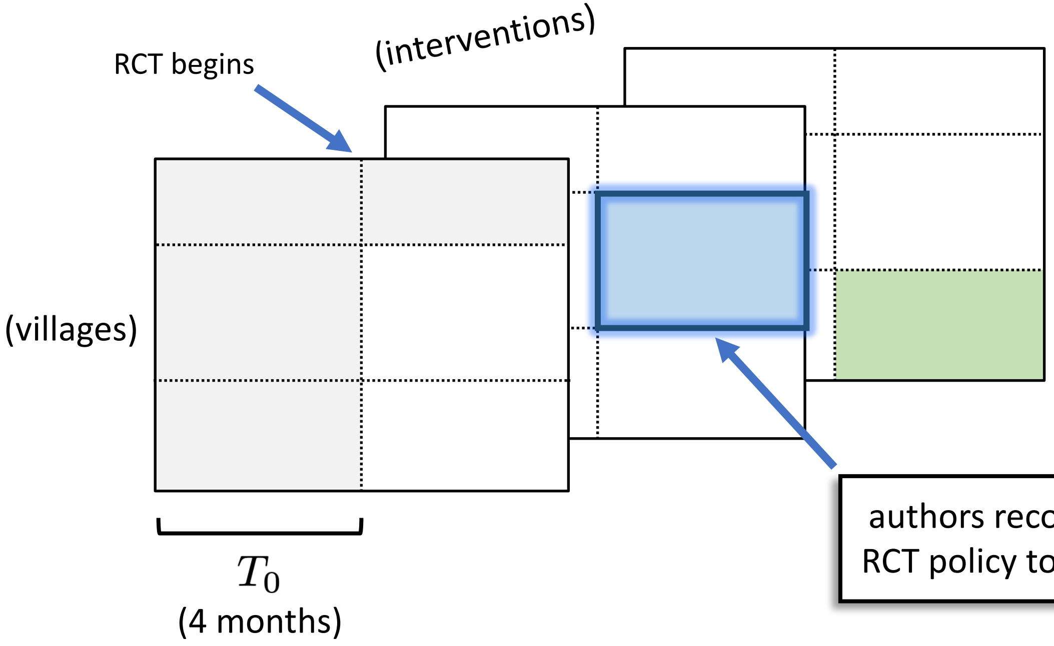
2523 villages

(intervention)

74 interventions (+1 control)

(measurement)

13 months



authors recommended single best RCT policy to Haryana Government

Would personalized policies have led to a greater uptake in immunization? If so, can we quantify?

Recreating Observed Immunization Rates

Intervention	000	001	002	010	031	032	040	050	100	101	102	200	201	202	300	301	302	400	401	402
Hyp. Test ($\alpha = 0.05$)	✓	✓	x	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	x	✓	✓	x	x	✓
R^2_{rct} (vs. RCT)	0.55	0.50	0.48	0.73	0.62	0.73	0.57	0.75	0.50	0.68	0.48	0.70	0.66	0.45	0.34	0.46	0.60	0.29	0.29	0.42

$$R^2_{\mathrm{rct}} = 1 - \frac{\text{SI error}}{\text{RCT error}}$$

heterogenous villages

(RCT estimate)

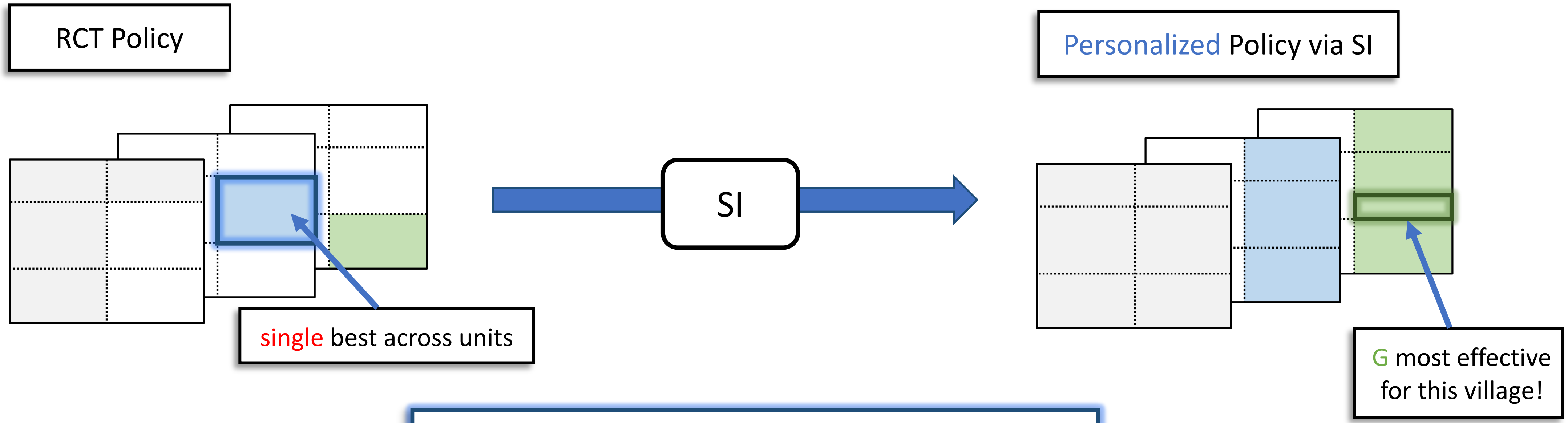
= (1/n)  + (1/n)  + … + (1/n) 

Heterogenous villages → SI is a stronger predictor than RCT estimator

Policy Recommendations

matches authors' recommended policy

Policy Recommendation Method	Avg. net increase in immunization rates (estimated)
Random policy (per village)	1.0
Best RCT policy (031)	1.3
SI's personalized policy (per village)	**2.8**



SI enables personalized policies
from same RCT data → significant gains

A/B Testing in E-commerce

Data Efficient RCTs: Web A/B Testing in Ecommerce

25 user groups

(>10k users per group)

3 interventions

(+1 control intervention)

8 days

(measuring customer engagement)

We get access to access to customer engagement trajectories of all 25 user groups under all interventions

Data Efficient RCTs: Web A/B Testing in Ecommerce

- Ideal RCT setting – experiments run

Intervention	Groups 1-8	Groups 9-16	Groups 17-25
Control	✓	✓	✓
10% Discount	✓	✓	✓
30% Discount	✓	✓	✓
50% Discount	✓	✓	✓

- Synthetic Interventions – experiments run

Intervention	Groups 1-8	Groups 9-16	Groups 17-25
Control	✓	✓	✓
10% Discount	✓		
30% Discount		✓	
50% Discount			✓

Data Efficient RCTs: Web A/B Testing in Ecommerce

- Hypothesis Test

Intervention	Metric	Projection Test
10% Discount	Subscription rate	(Pass, $\alpha = 0.05$)
30% Discount	Subscription rate	(Pass, $\alpha = 0.05$)
50% Discount	Subscription rate	(Pass, $\alpha = 0.05$)

Data Efficient RCTs: Web A/B Testing in Ecommerce

- Quantifying prediction accuracy
  - R^2 score (access to true counterfactual)

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{reg}}}$$

$$\mathrm{SS}_{\mathrm{reg}} = \sum_i \left(Y_{ni}^{(d)} - \bar{Y}_n^{(d)} \right)^2$$

$$\mathrm{SS}_{\mathrm{res}} = \sum_i \left(Y_{ni}^{(d)} - \widehat{Y}_{ni}^{(d)} \right)^2$$

- Quantifying utility over standard RCTs
  - R^2 score (using RCT as a predictor)

$$R^2_{\mathrm{rct}} = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{rct}}}$$

$$\mathrm{SS}_{\mathrm{rct}} = \sum_i \left(Y_{ni}^{(d)} - \frac{1}{|\mathcal{I}^{(d)}|} \sum_{m \in \mathcal{I}^{(d)}} Y_{mi}^{(d)} \right)^2$$

Intervention	R^2 score (True Counterfactual)	R^2 score (RCT Baseline)
10% Discount	0.76	0.98
30% Discount	0.56	0.99
50% Discount	0.75	0.98

accurate recovery of true customer engagements

heterogeneous user groups

Synthetic Interventions simulates ideal RCT experiment