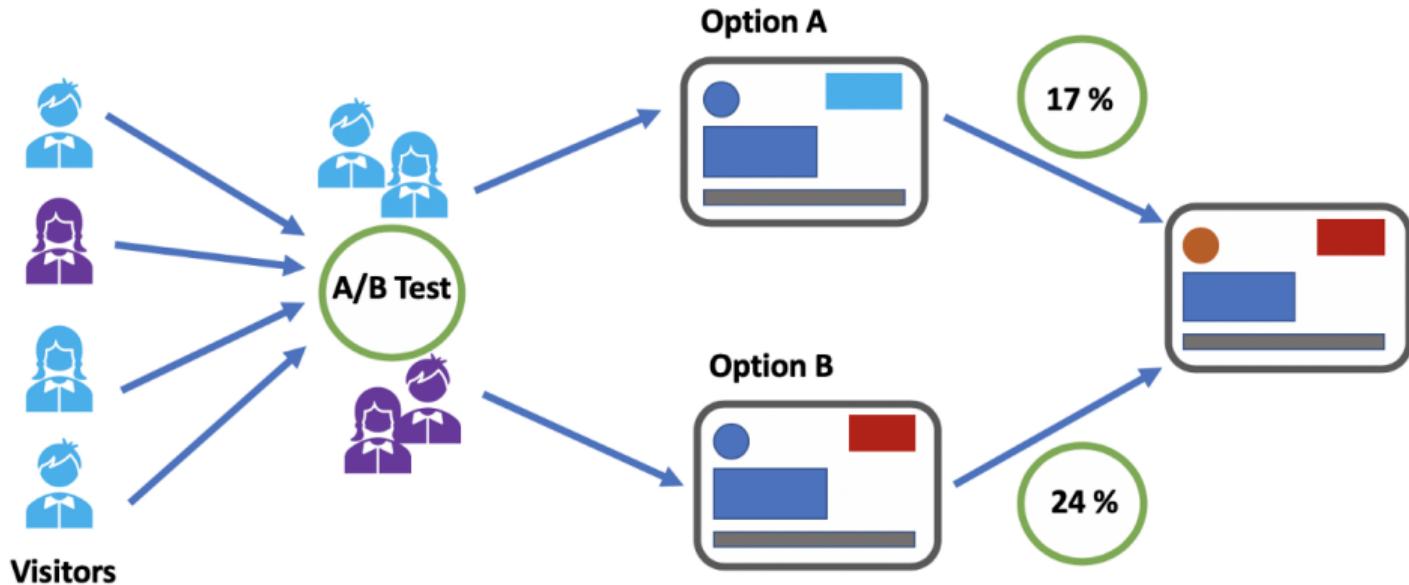


Optimal Design for Efficient A/B Testing

Chengchun Shi

Associate Professor of Data Science
London School of Economics and Political Science

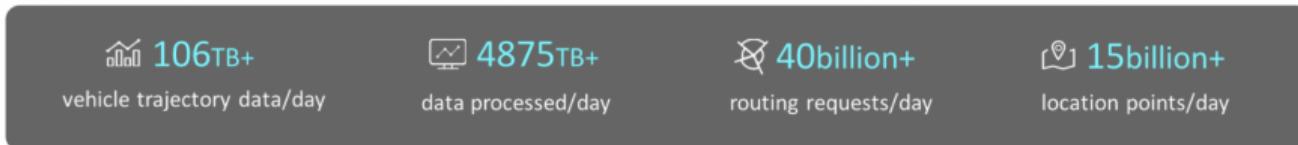
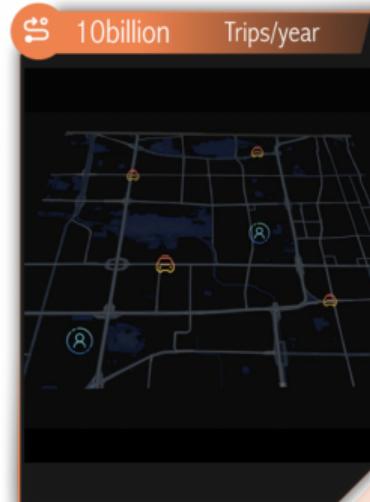
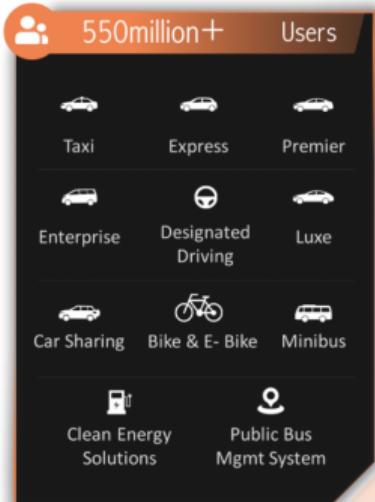
A/B Testing



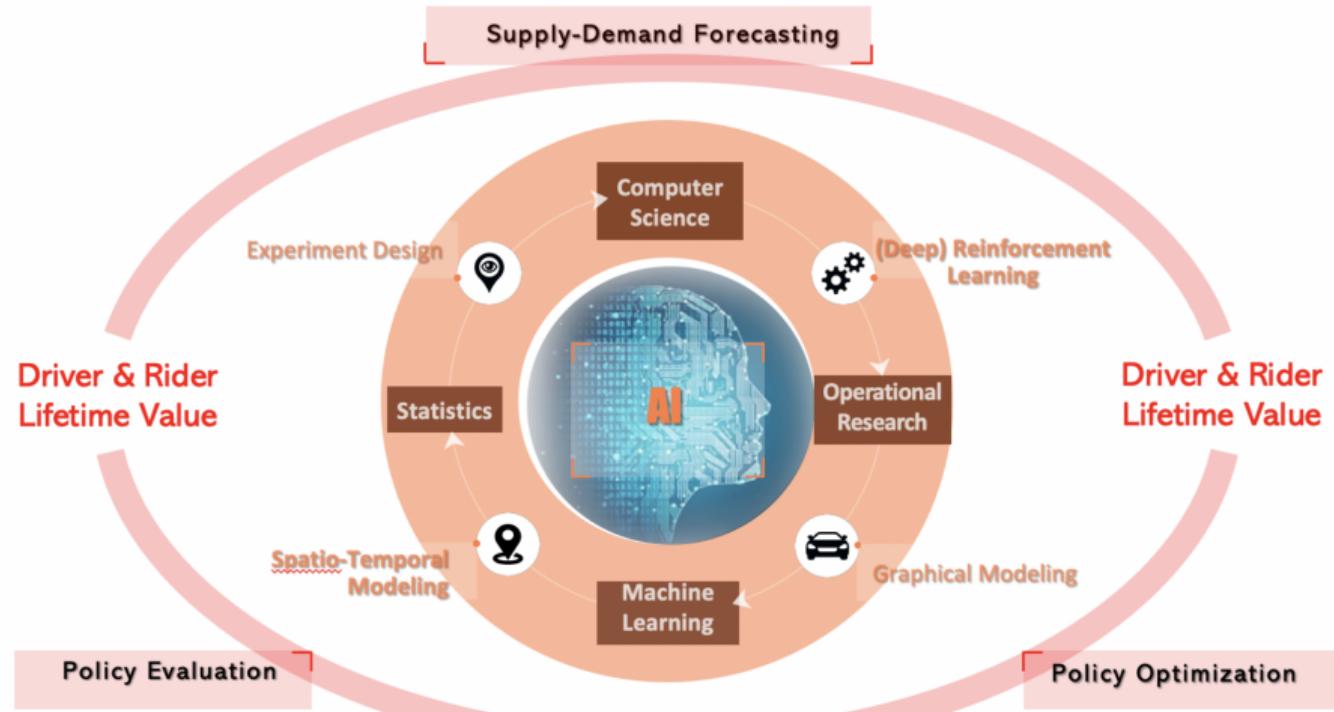
Taken from

<https://towardsdatascience.com/how-to-conduct-a-b-testing-3076074a8458>

Ridesharing

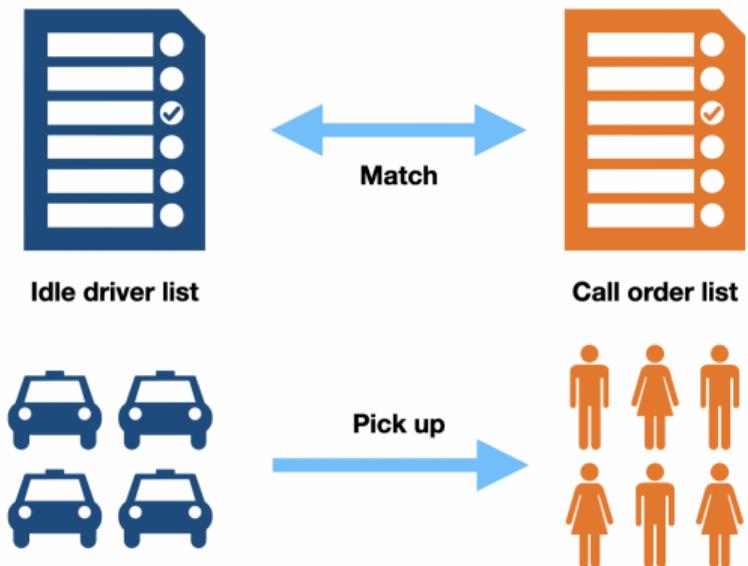


Ridesharing (Cont'd)



Policies of Interest

- Order dispatching

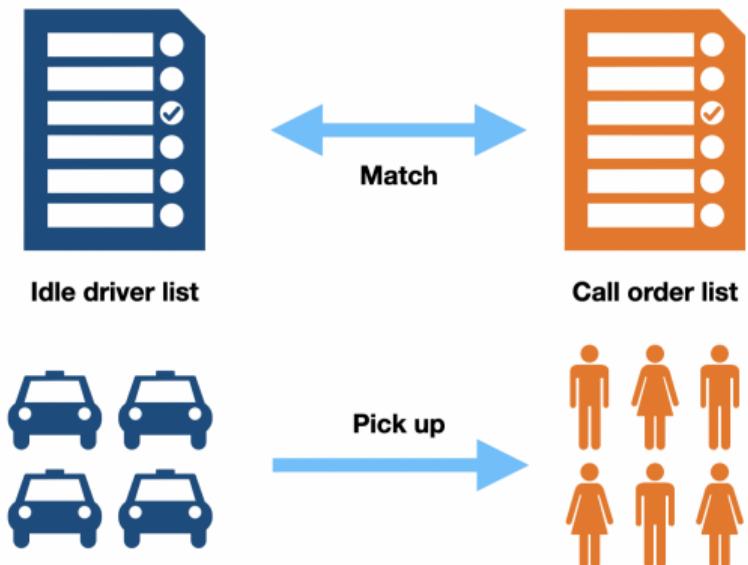


- Subsidizing



Policies of Interest

- Order dispatching



- Subsidizing



Time Series Data

- Online experiment typically lasts for **two weeks**
- **30 minutes/1 hour** as one time unit
- Data forms a **time series** $\{(Y_t, U_t) : 1 \leq t \leq T\}$
- **Observations** $Y_t \in \mathbb{R}^3$:
 1. **Outcome**: drivers' income or no. of completed orders
 2. **Supply**: no. of idle drivers
 3. **Demand**: no. of call orders
- **Treatment** $U_t \in \{1, -1\}$:
 - **New** order dispatching policy B
 - **Old** order dispatching policy A

Challenges

1. Carryover Effects:

- Past treatments influence future observations [Li et al., 2024a, Figure 2] →
- Invalidating many conventional A/B testing/causal inference methods [Shi et al., 2023].

2. Partial Observability:

- The environmental state is not fully observable →
- Leading to the violation of the Markov assumption.

3. Small Sample Size:

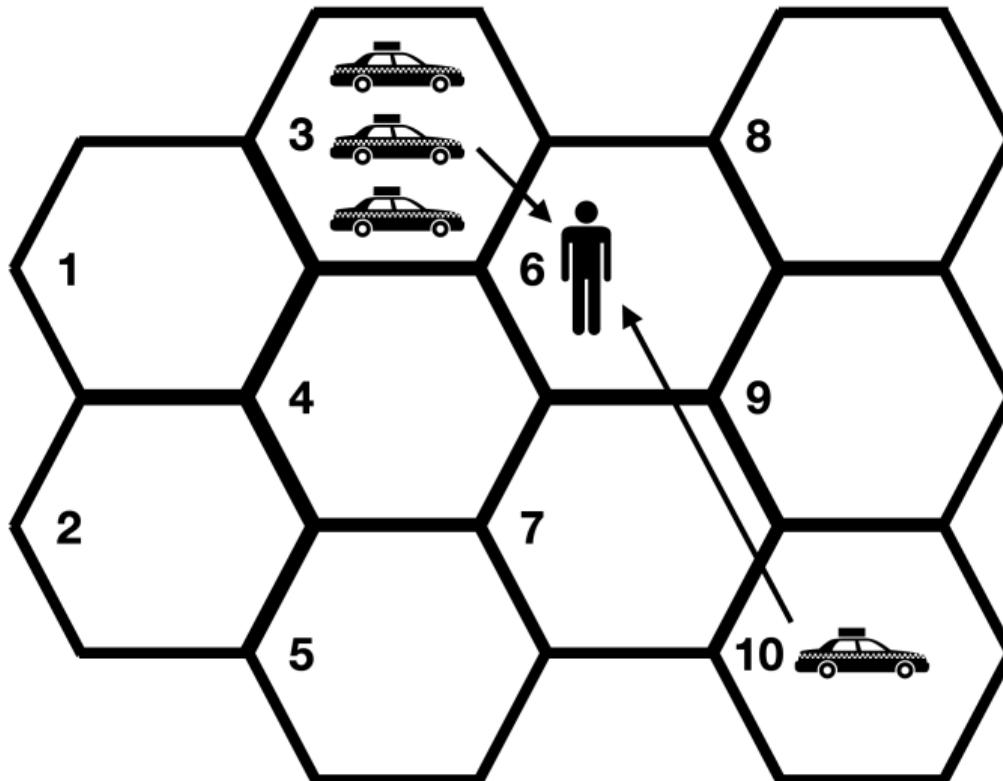
- Online experiments typically last only two weeks [Xu et al., 2018] →
- Increasing the variability of the average treatment effect (ATE) estimator.

4. Weak Signal:

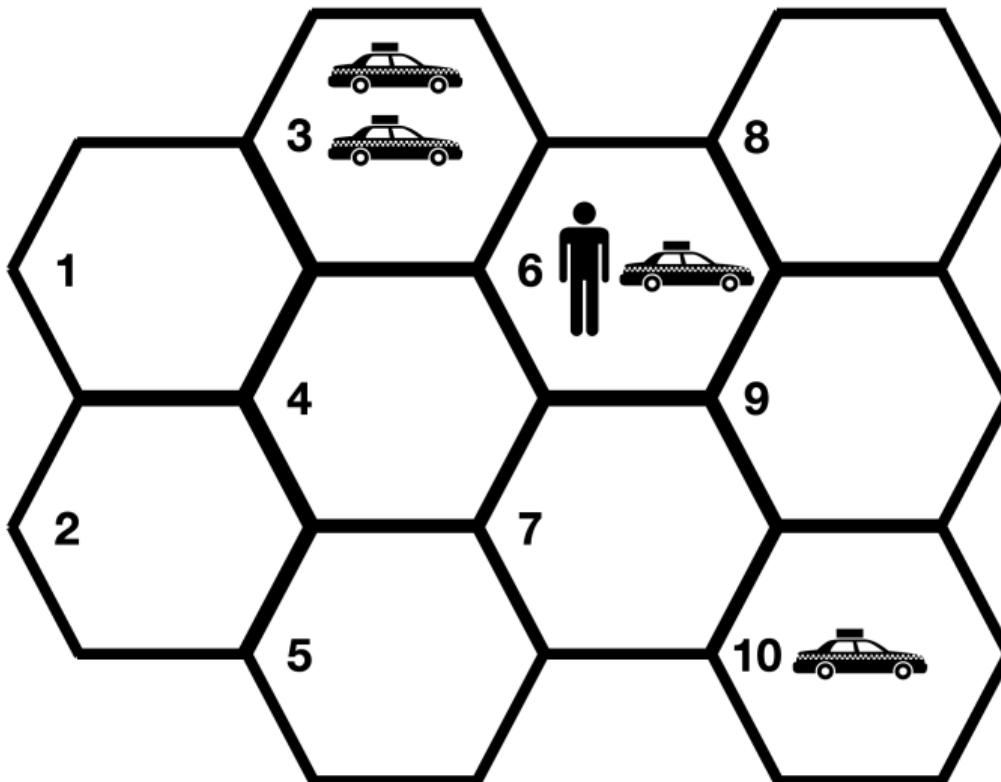
- Size of treatment effects ranges from 0.5% to 2% [Tang et al., 2019] →
- Making it challenging to distinguish between new and old policies.

To our knowledge, **no** existing method has simultaneously addressed all four challenges.

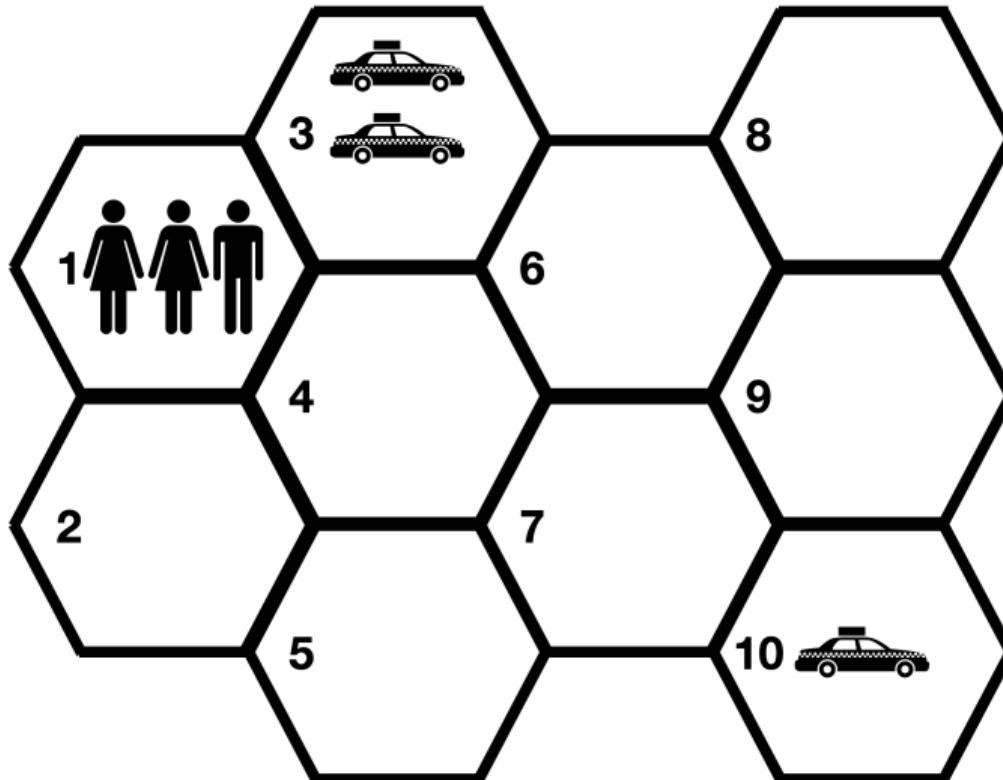
Challenge I: Carryover Effects



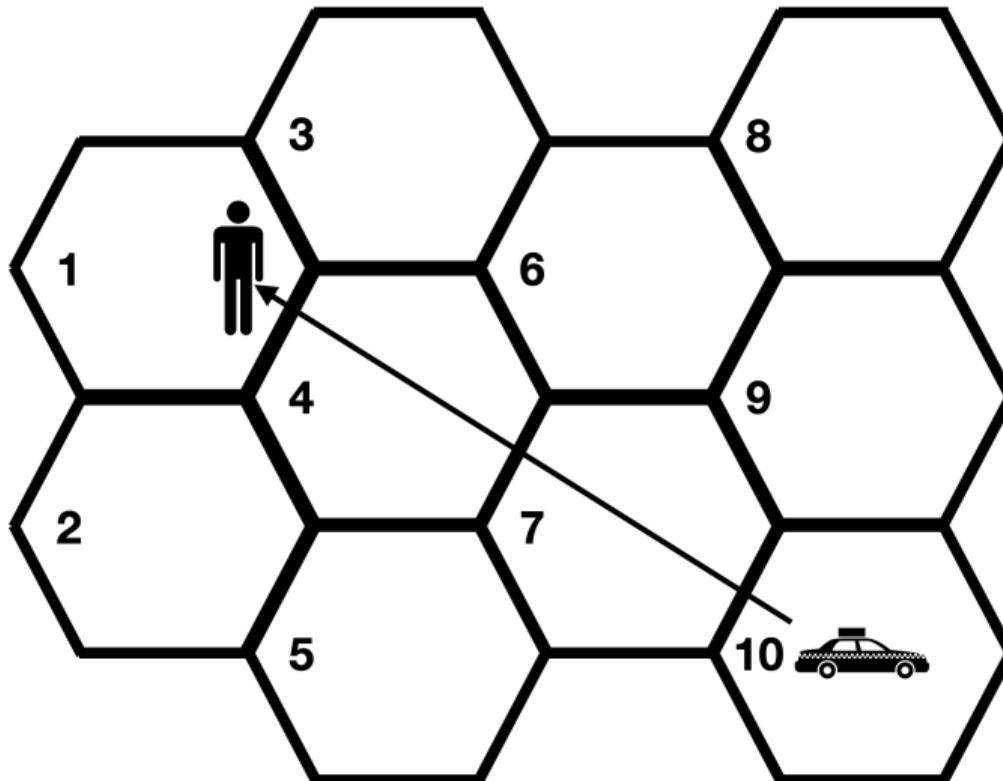
Adopting the Closest Driver Policy



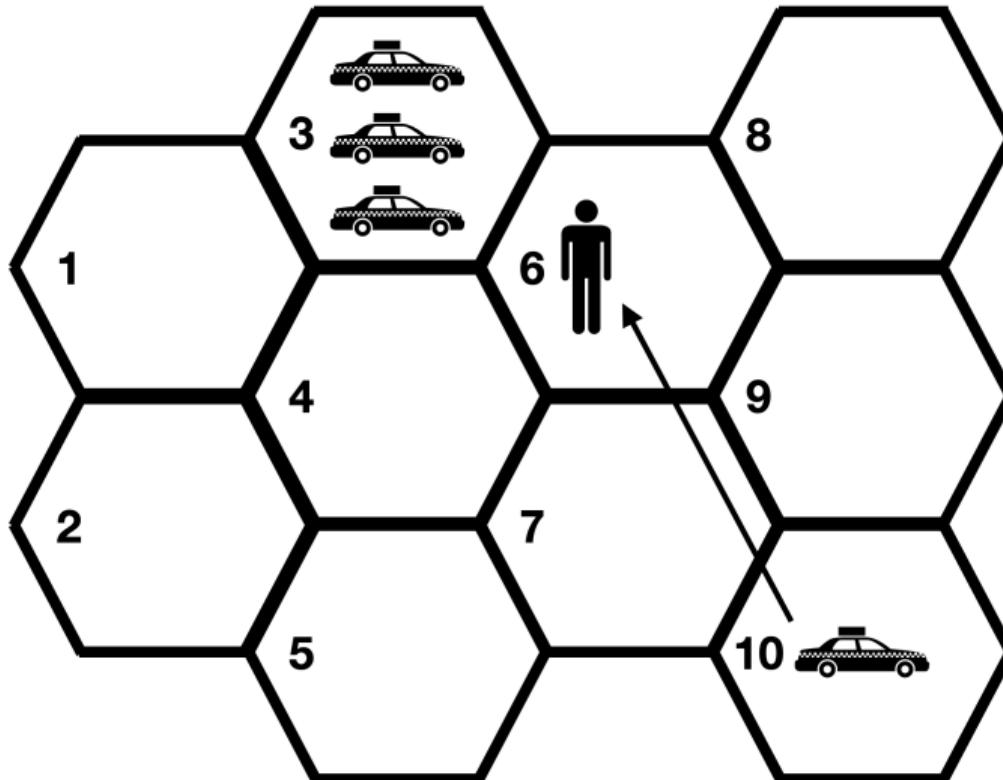
Some Time Later . . .



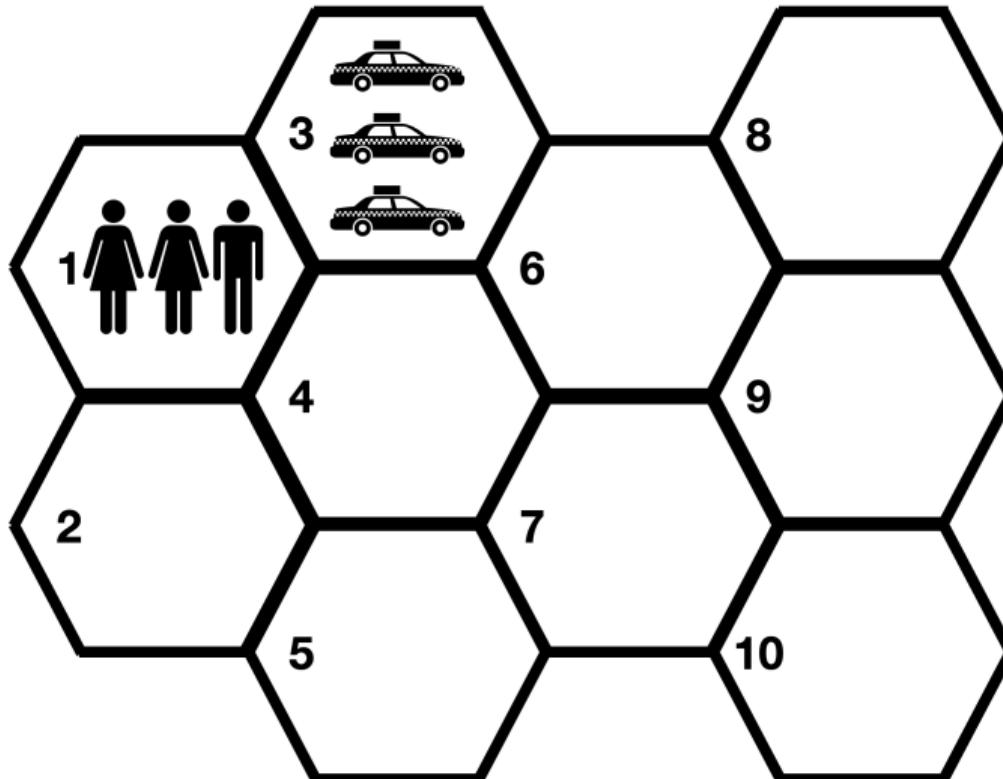
Miss One Order



Consider a Different Action



Able to Match All Orders

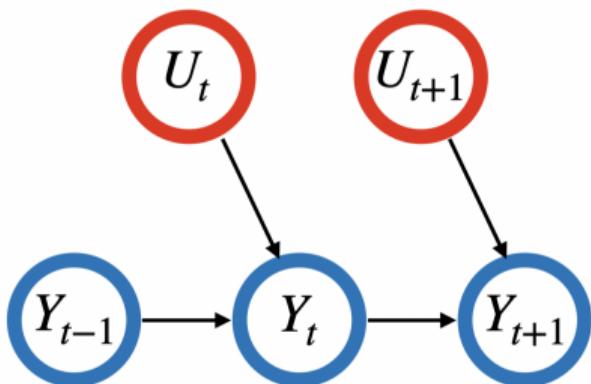


Challenge I: Carryover Effects (Cont'd)

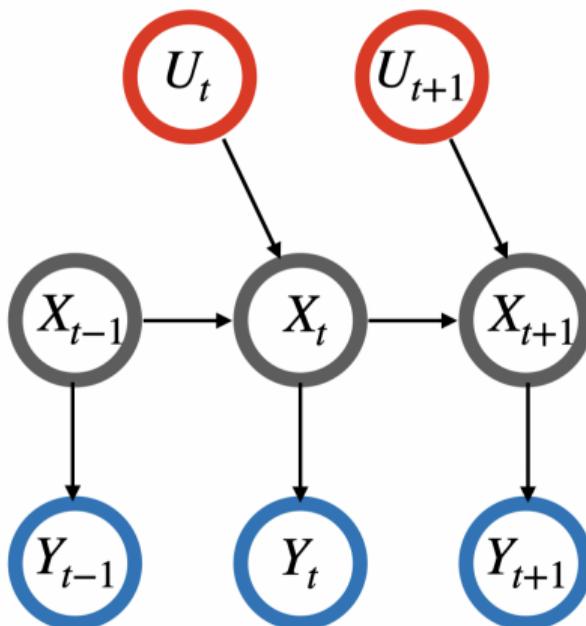
past treatments → distribution of drivers → future outcomes

Challenge II: Partial Observability

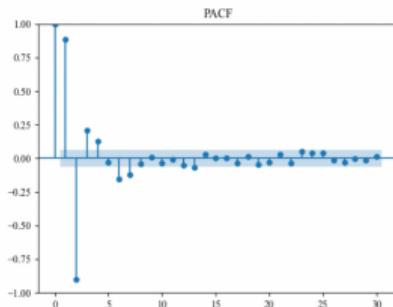
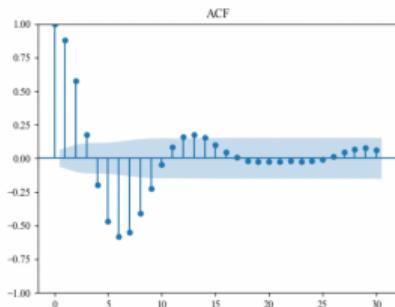
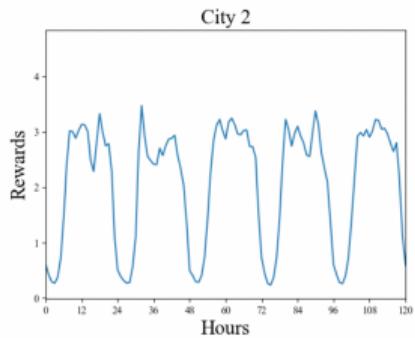
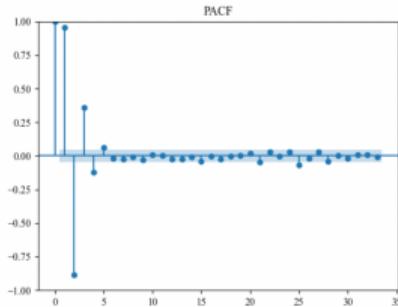
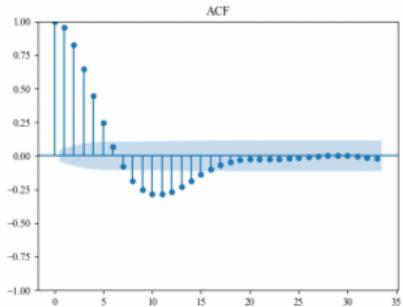
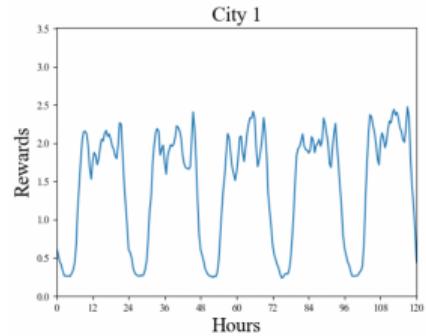
- Fully Observable
Markovian Environments



- Partially Observable
non-Markovian Environments



Challenge II: Partial Observability (Cont'd)



(a) Rewards

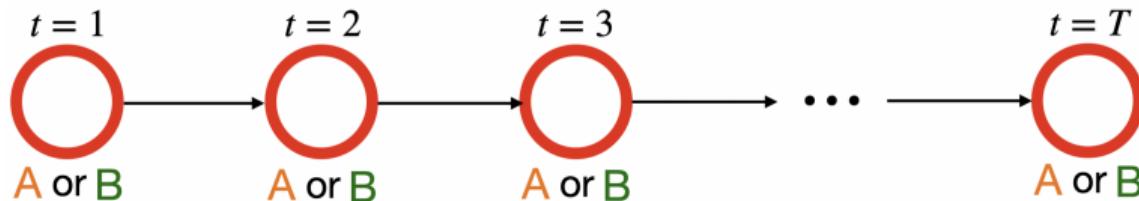
(b) ACF Plot

(c) PACF Plot

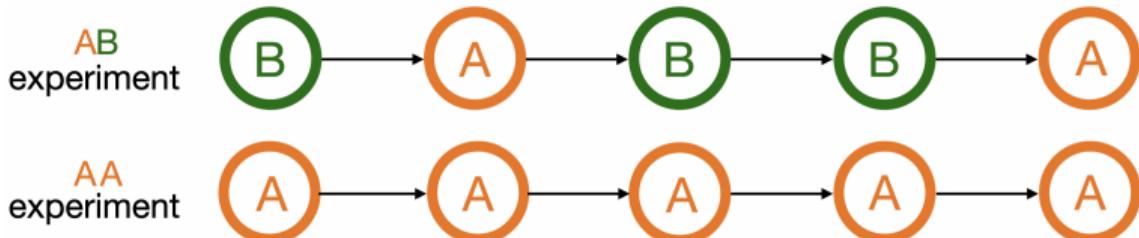
AIC indicate that the two datasets are best modeled by VARMA(2,2) and VARMA(2,1).

Challenge III & IV: Small sample & Weak Signal

- **Aim 1: Design.** Identify **optimal treatment allocation strategy** in online experiment that **minimizes MSE of ATE estimator**



- **Aim 2: Data Integration.** Combine **experimental data (A/B)** with **historical data (A/A)** to improve ATE estimation [Li et al., 2024b]



Optimal Treatment Allocation Strategies for A/B Testing in Partially Observable Environments

Joint work with Ke Sun, Linglong Kong & Hongtu Zhu

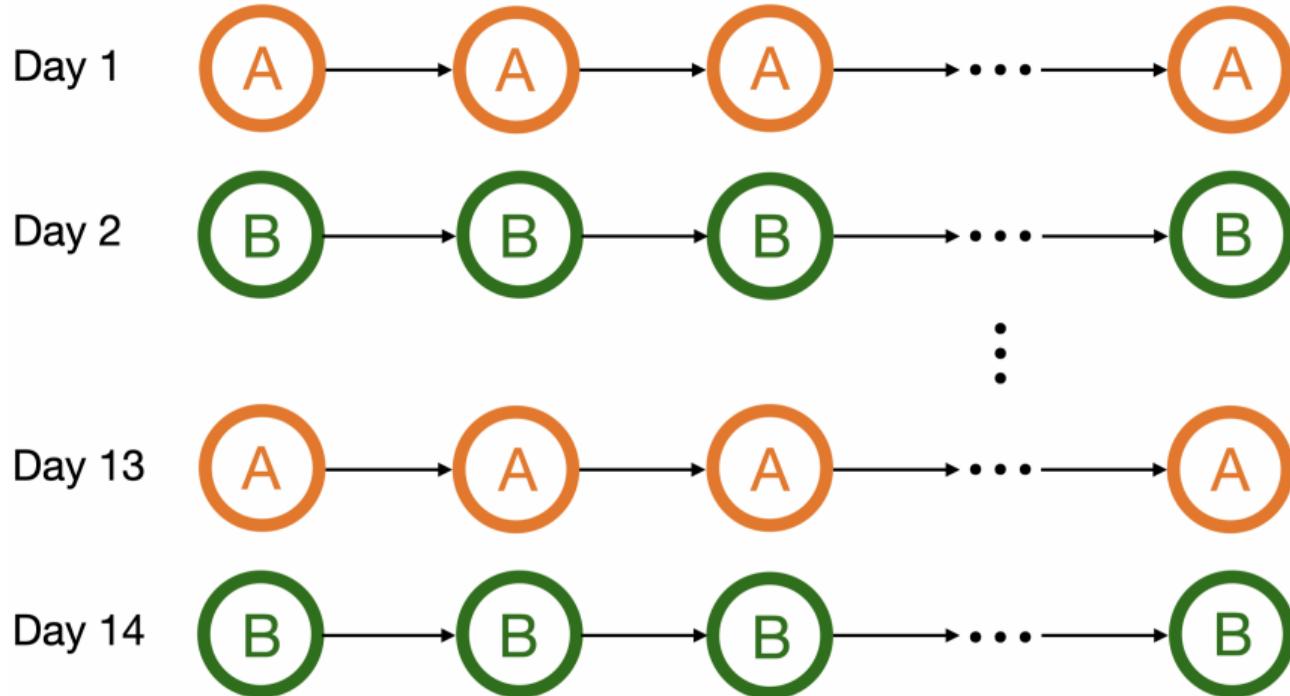
Average Treatment Effect

- Data summarized into a **time series** $\{(\mathbf{Y}_t, \mathbf{U}_t) : 1 \leq t \leq T\}$
- The first element of \mathbf{Y}_t – denoted by R_t – represents the **outcome**
- **ATE = difference in average outcome** between the **new** and **old** policy

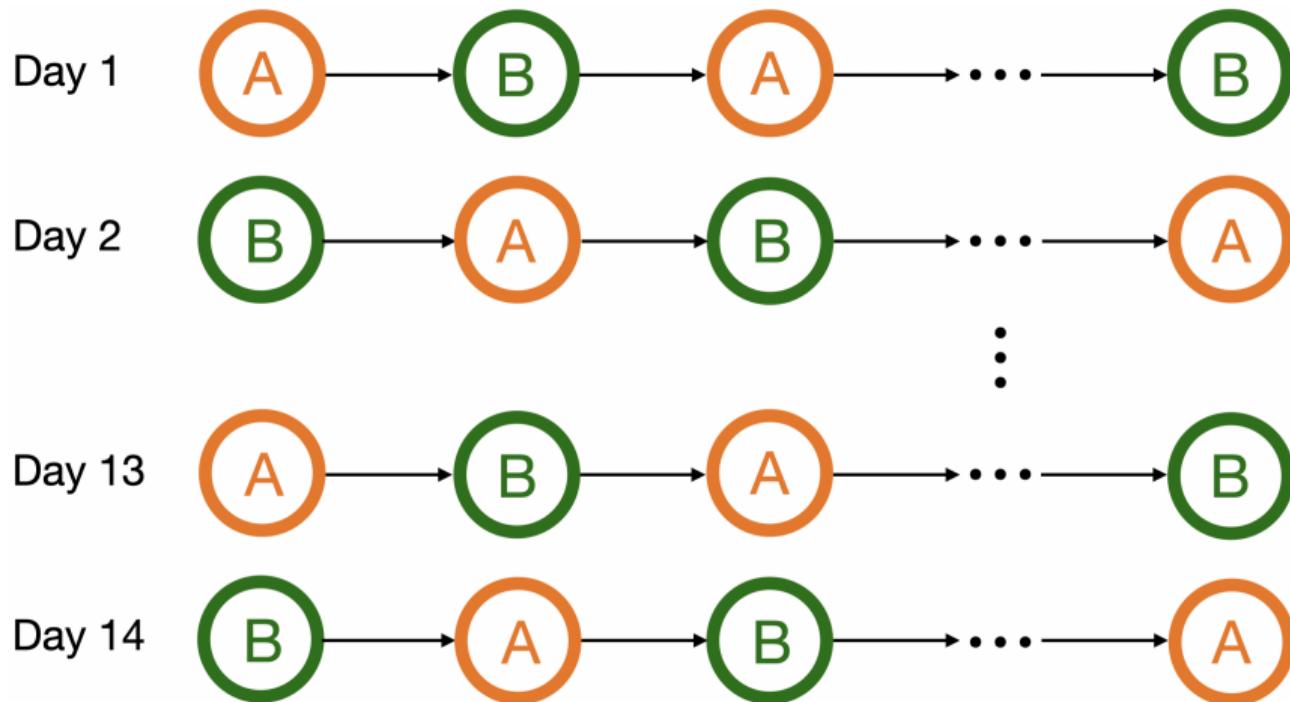
$$\lim_{T \rightarrow \infty} \left[\frac{1}{T} \sum_{t=1}^T \mathbb{E}R_t \right] - \lim_{T \rightarrow \infty} \left[\frac{1}{T} \sum_{t=1}^T \mathbb{E}R_t \right].$$

Letting $T \rightarrow \infty$ simplifies the analysis.

Alternating-day (AD) Design



Alternating-time (AT) Design



AD v.s. AT

Pros of **AD design**:

- Within each day, it is **on-policy** and avoids **distributional shift**, as opposed to **off-policy** designs (e.g., AT)
- On-policy designs are proven **optimal** in **fully observable Markovian** environments (Li et al., 2023).

Pros of **AT design**:

- Widely employed in ridesharing companies like Lyft and Didi [Chamandy, 2016, Luo et al., 2024]
- According to my industrial collaborator, AT yields **less variable ATE estimators** than AD

AD v.s. AT (Cont'd)

- Q: Why can off-policy designs, such as AT, be more efficient than AD?
- A: Due to partial observability...

A Thought Experiment

- A simple setting **without carryover effects**:

$$R_t = \beta_{-1}\mathbb{I}(U_t = -1) + \beta_1\mathbb{I}(U_t = 1) + \varepsilon_t$$

- ATE equals $\beta_1 - \beta_{-1}$ and can be estimated by

$$\widehat{\text{ATE}} = \frac{\sum_{t=1}^T R_t \mathbb{I}(U_t = 1)}{\sum_{t=1}^T \mathbb{I}(U_t = 1)} - \frac{\sum_{t=1}^T R_t \mathbb{I}(U_t = -1)}{\sum_{t=1}^T \mathbb{I}(U_t = -1)}$$

A Thought Experiment (Cont'd)

The ATE estimator's asymptotic MSE under AD and AT is proportional to

$$\lim_{t \rightarrow \infty} \frac{1}{t} \text{Var}(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \cdots + \varepsilon_t) \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{1}{t} \text{Var}(\varepsilon_1 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 + \cdots - \varepsilon_t)$$

which depends on the residual correlation:

- With **uncorrelated residuals**, both designs yield **same** MSEs
- With **positively correlated residuals**:
 - **AD assigns the same treatment** within each day, under which ATE estimator's variance inflates due to **accumulation** of these residuals
 - **AT alternates treatments** for adjacent observations, effectively **negating** these residuals, leading to more efficient ATE estimation
- With **negatively correlated residuals**, AD generally outperforms AT

When Can AT Be More Efficient than AD

Key Condition: Residuals are positively correlated

- **Rule out full observability** (Markovianity) where residuals are uncorrelated.
- Can only be met under **partial observability**.
- Suggest partial observability is more realistic, aligning with my collaborator's finding.
- **Often satisfied** in practice:

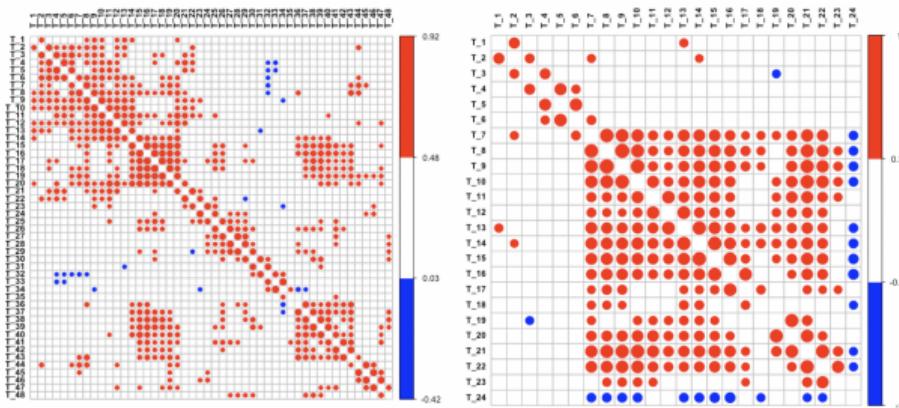


Figure: Estimated correlation coefficients between pairs of fitted outcome residuals from the two cities

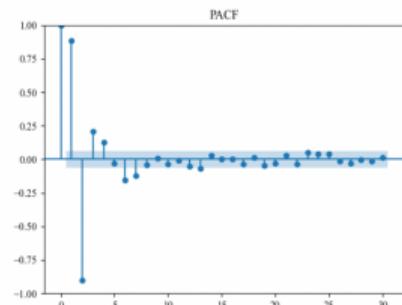
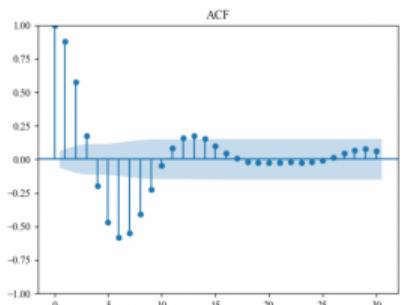
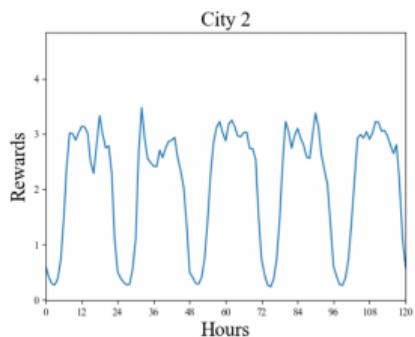
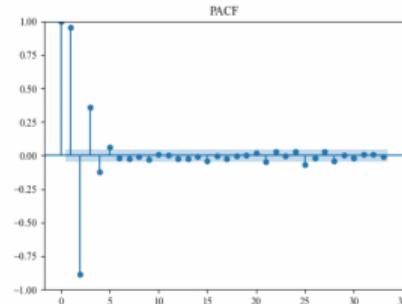
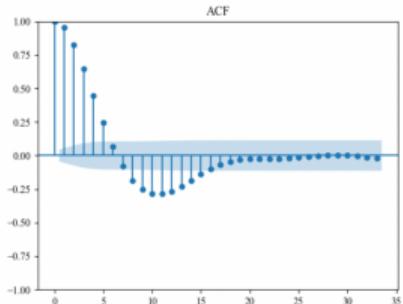
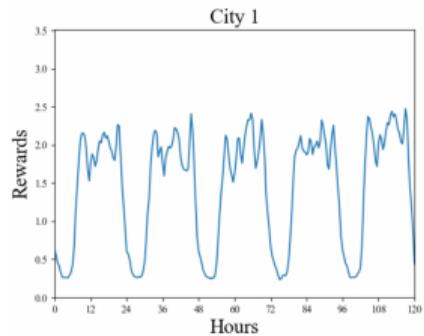
Some Motivating Questions

- Q1: Previous analysis excludes carryover effects. Can we extend the results to accommodate carryover effects?
- Q2: Previous analysis focuses on AD and AT. Can we consider more general designs?

Our Contributions

- **Methodologically**, we propose:
 1. A **controlled (V)ARMA** model → allow **carryover effects & partial observability**
 2. Two **efficiency indicators** → compare commonly used designs (AD, AT)
 3. A **reinforcement learning** (RL) algorithm → compute the **optimal design**
- **Theoretically**, we:
 1. Establish **asymptotic MSEs** of ATE estimators → compare different designs
 2. Introduce **weak signal condition** → simplify asymptotic analysis in sequential settings
 3. Prove the **optimal treatment allocation strategy** is q -dependent → form the basis of our proposed RL algorithm
- **Empirically**, we demonstrate the advantages of our proposal using:
 1. A dispatch simulator (<https://github.com/callmespring/MDPOD>)
 2. Two real datasets from ridesharing companies.

Controlled VARMA Model: Motivation

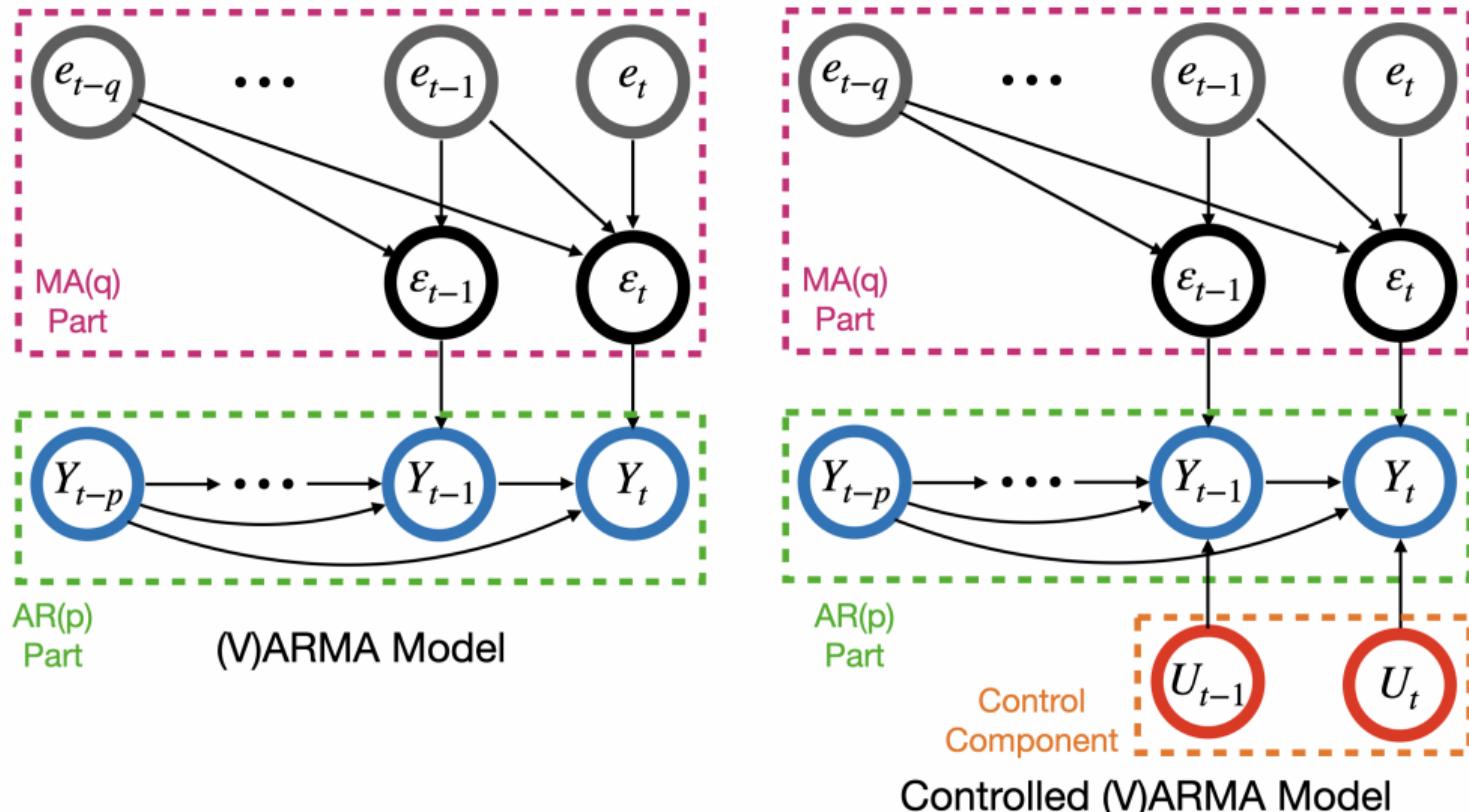


(a) Rewards

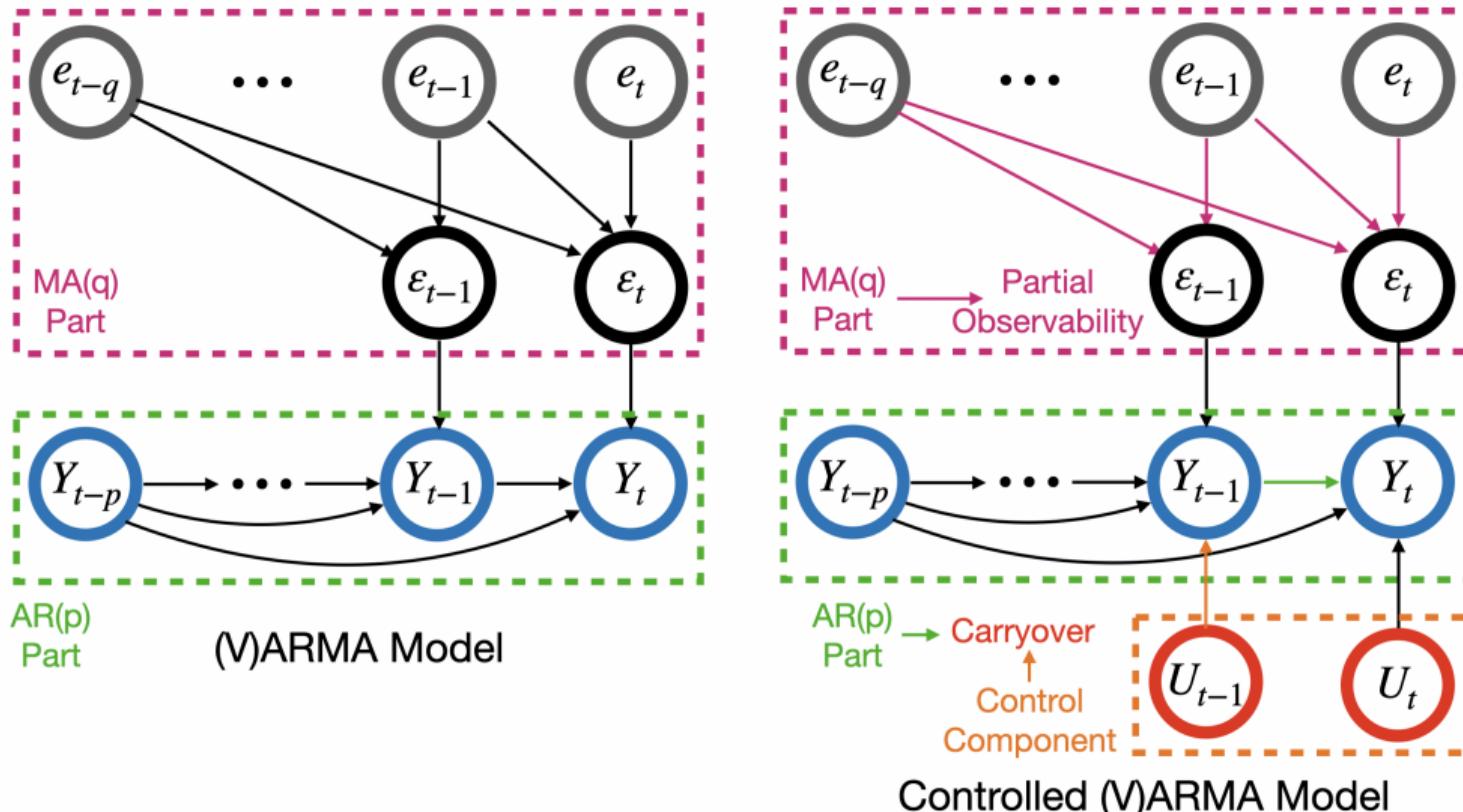
(b) ACF Plot

(c) PACF Plot

Controlled VARMA Model: Introduction



Controlled VARMA Model: Introduction



Controlled VARMA Model: Connection

- Closely related to **state space models** or **linear quadratic regulator (LQR)**
 - The latter being a rich sub-class of **partially observable MDPs**
 - Using VARMA as opposed to LQR allows to leverage asymptotic theories developed in time series to derive optimal designs
- Compared to **MDPs**
 - Both controlled VARMA and MDP accommodate **carryover effects**
 - MDPs require full observability whereas controlled VARMA allows **partial observability**

Controlled VARMA Model: Estimation

Consider a univariate controlled ARMA

$$Y_t = \mu + \underbrace{\sum_{j=1}^p a_j Y_{t-j}}_{\text{AR Part}} + \underbrace{b U_t}_{\text{Control}} + e_t + \underbrace{\sum_{j=1}^q \theta_j e_{t-j}}_{\text{MA Part}}$$

- **AR parameters** $\{a_j\}_j$ & **control parameter** b → **ATE**, equal to $2b / \sum_j a_j$
 - Partial observability → standard OLS **fails** to consistently estimate b & $\{a_j\}_j$
 - Employ **Yule-Walker estimation** (method of moments) instead
 - Similar to **IV** estimation, utilize past observations as IVs
- **MA parameters** $\{\theta_j\}_j$ → **residual correlation** → **optimal design**

Theory: Weak Signal Condition

- **Asymptotic framework:** large sample $T \rightarrow \infty$ & weak signal $\text{ATE} \rightarrow 0$
- **Empirical alignment:** size of ATE ranges from 0.5% to 2%
- **Theoretical simplification:** considerably simplifies the computation of ATE estimator's MSE in sequential settings. According to Taylor's expansion:

$$\widehat{\text{ATE}} - \text{ATE} = \frac{2\widehat{b}}{1 - \sum_j \widehat{a}_j} - \frac{2b}{1 - \sum_j a_j}$$
$$= \frac{2(\widehat{b} - b)}{1 - \sum_j a_j} + \frac{2b}{(1 - \sum_j a_j)^2} \sum_j (\widehat{a}_j - a_j) + o_p\left(\frac{1}{\sqrt{T}}\right)$$

Leading term. Easy to calculate its asymptotic variance under weak signal

Challenging to obtain the closed form of its asymptotic variance, but negligible under weak signal condition

High-order reminder

Theory: Asymptotic MSE

We focus on the class of **observation-agnostic** designs:

- \mathbf{U}_1 is randomly assigned
- The distribution of \mathbf{U}_t depends on $(\mathbf{U}_1, \dots, \mathbf{U}_{t-1})$, independent of $(\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1})$

It covers three commonly used designs:

1. Uniform random (UR) design: $\{\mathbf{U}_t\}_t$ are uniformly independently generated
2. AD: $\mathbf{U}_1 = \mathbf{U}_2 = \dots = \mathbf{U}_D = -\mathbf{U}_{D+1} = \dots = -\mathbf{U}_{2D} = \mathbf{U}_{2D+1} = \dots$
3. AT: $\mathbf{U}_1 = -\mathbf{U}_2 = \mathbf{U}_3 = -\mathbf{U}_4 = \dots = (-1)^{T-1} \mathbf{U}_T$

Theorem (Asymptotic MSE)

Given an **observation-agnostic** design, let $\xi = \lim_T \sum_{t=1}^T (\mathbb{E} \mathbf{U}_t / T)$. Under the **weak signal condition**, its ATE estimator's asymptotic MSE (after normalization) equals

$$\lim_T \frac{4}{(1 - \sum_j \mathbf{a}_j)^2 (1 - \xi)^2 T} \text{Var} \left[\sum_{t=1}^T (\mathbf{U}_t - \xi) \varepsilon_t \right].$$

Theory: Asymptotic MSE (Cont'd)

Corollary (Asymptotic MSE)

Under the **weak signal** condition, the ATE estimator's asymptotic MSE (after normalization) under **AD**, **UR** and **AT** equals

$$\text{MSE(AD)} = \frac{4\sigma^2}{(1 - \sum_j \mathbf{a}_j)^2} \left[\sum_{j=0}^q \theta_j^2 + \sum_{j_1 \neq j_2} \theta_{j_1} \theta_{j_2} \right]$$

$$\text{MSE(UR)} = \frac{4\sigma^2}{(1 - \sum_j \mathbf{a}_j)^2} \sum_{j=0}^q \theta_j^2$$

$$\text{MSE(AT)} = \frac{4\sigma^2}{(1 - \sum_j \mathbf{a}_j)^2} \left[\sum_{j=0}^q \theta_j^2 + 2 \sum_{j_1 \neq j_2} (-1)^{|j_2 - j_1|} \theta_{j_1} \theta_{j_2} \right],$$

where σ^2 denotes the variance of the white noise process.

Design: Efficiency Indicator

Define two efficiency indicators

$$\mathbf{EI}_1 = \sum_{j_1 \neq j_2} \theta_{j_1} \theta_{j_2} \quad \text{and} \quad \mathbf{EI}_2 = \sum_{j_1 \neq j_2} (-1)^{|j_2 - j_1|} \theta_{j_1} \theta_{j_2}.$$

They measure **residual correlations** and can be used to compare the three designs:

- If both \mathbf{EI}_1 and $\mathbf{EI}_2 > 0$, **UR** outperforms **AD** & **AT**
- If $\mathbf{EI}_2 < 0$ and $\mathbf{EI}_1 > \mathbf{EI}_2$, **AT** outperforms the rest
- If $\mathbf{EI}_1 < 0$ and $\mathbf{EI}_2 > \mathbf{EI}_1$, **AD** outperforms the rest

MA parameters can be estimated using historical data (even without treatment data).

Design: Optimality

Theorem (Optimal Design)

The optimal design must satisfy $\lim_T \sum_{t=1}^T (\mathbb{E} \mathbf{U}_t / T) = \mathbf{0}$. Additionally, it must minimize

$$\sum_{k=1}^q \left[\lim_T \left(\frac{1}{T} \sum_{t=1}^T \mathbb{E} \mathbf{U}_t \mathbf{U}_{t+k} \right) \underbrace{\sum_{j=k}^q \theta_j \theta_{j-k}}_{c_k} \right]$$

Objective: learn the optimal observation-agnostic design that:

- (i) **Minimizes** the above criterion
- (ii) **Maintains** a zero mean asymptotically, i.e., $\lim_T \sum_{t=1}^T (\mathbb{E} \mathbf{U}_t / T) = \mathbf{0}$

Design: An RL Approach

Solution: reformulate the minimization as an infinite-horizon average-reward RL problem

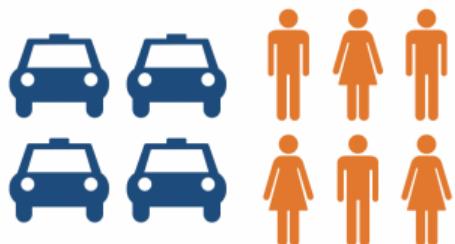
- **State S_t :** the collection of past q treatments ($\mathbf{U}_{t-q}, \mathbf{U}_{t-q+1}, \dots, \mathbf{U}_{t-1}$)
- **Action A_t :** the current treatment $\mathbf{U}_t \in \{-1, 1\}$
- **Reward R_t :** a deterministic function of state-action pair, $-\sum_{k=1}^q c_k(\mathbf{U}_t \mathbf{U}_{t-k})$

Easy to verify:

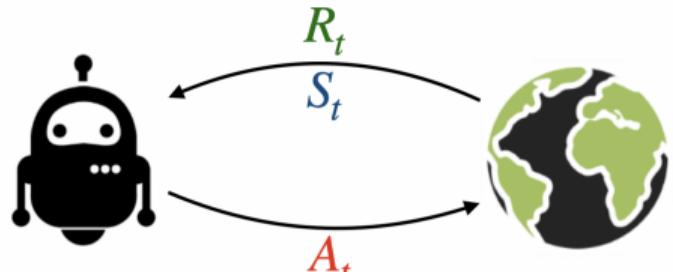
1. The minimization objective equals the negative average reward \rightarrow equivalent to **maximizing the average reward**
2. The process is an **MDP** \rightarrow there exists an optimal stationary policy maximizes the average reward \rightarrow optimal design is **q -dependent**, i.e., \mathbf{U}_t is a deterministic function of ($\mathbf{U}_{t-q}, \mathbf{U}_{t-q+1}, \dots, \mathbf{U}_{t-1}$) & this function is stationary in t
3. **Uniformly randomly** assign the first q treatments \rightarrow the resulting design maintains a zero mean and is indeed optimal

Design: An RL Approach (Cont'd)

Step 1: Retrieve Historical Data



Step 4: Online Learning of Optimal Design



Step 5: Implement the Design
Collect Additional Data

MLE

Step 2: Estimate MA Parameters

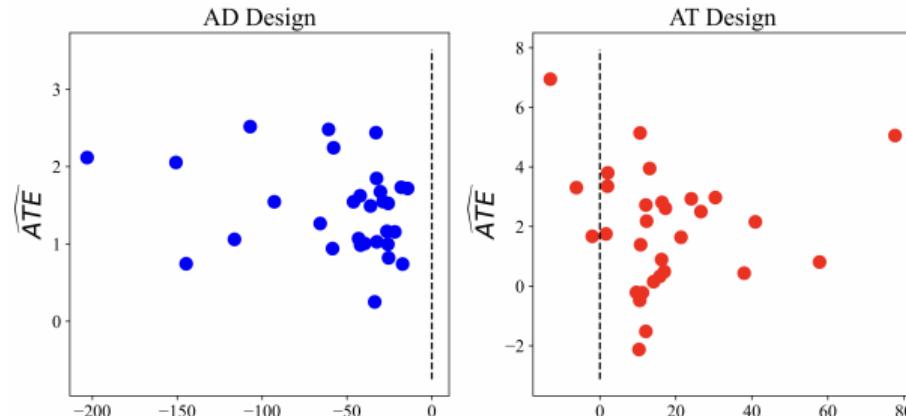
Step 3: Construct the MDP using estimated $\{C_k\}_k$

Model-based
Learning

Value
Iteration

Empirical Study: Synthetic Environments

- A 9×9 dispatch simulator
- Available at <https://github.com/callmespring/MDPOD>
- Two efficiency indicators

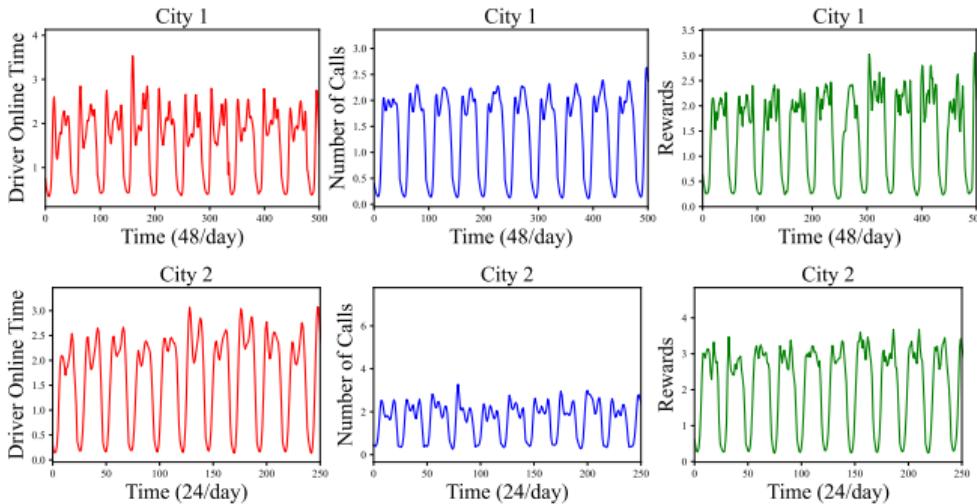


- ATE estimator's MSE under various designs

Design	AT	UR	Greedy	TMDP	NMDP	AD	Ours
MSE	8.33	2.23	1.10	0.56	0.42	0.28	0.28

Empirical Study: Real Datasets

- Data:



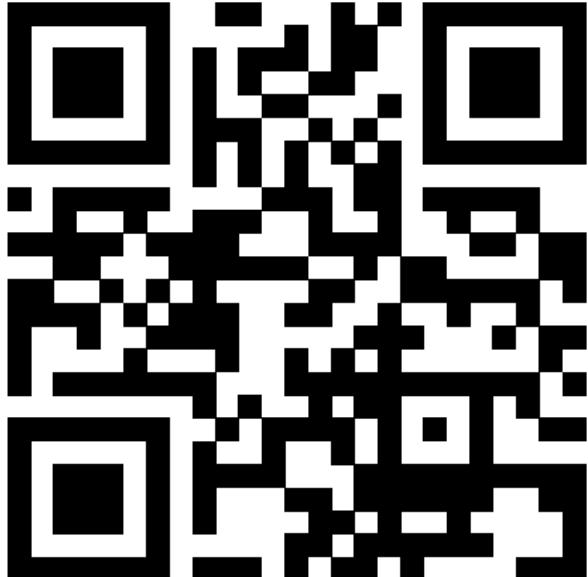
- We incorporate a **seasonal** term in our controlled VARMA model to account for seasonality. Below are MSEs of ATE estimators under different designs

City	El ₁	El ₂	AD	UR	AT	Ours
City 1	20.98	-21.11	11.98	11.63	9.72	8.24
City 2	-4.89	0.22	9.64	30.04	546.79	8.38

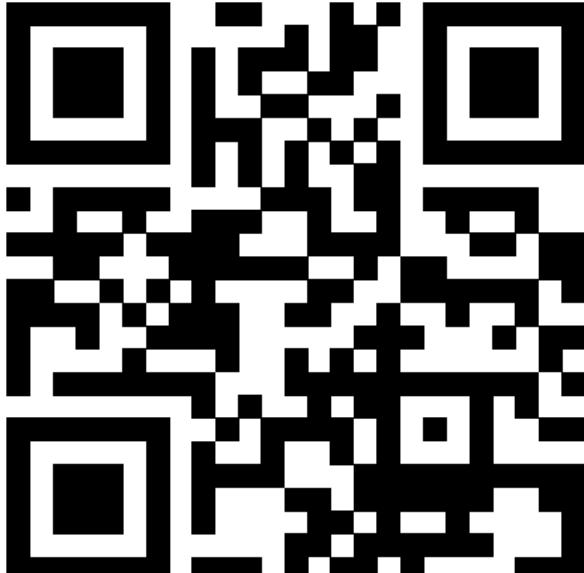
Thank You!



My RL short course



My personal website



References |

- Nicholas Chamandy. Experimentation in a ridesharing marketplace. <https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e>, 2016.
- Ting Li, Chengchun Shi, Zhaohua Lu, Yi Li, and Hongtu Zhu. Evaluating dynamic conditional quantile treatment effects with applications in ridesharing. *Journal of the American Statistical Association*, accepted, 2024a.
- Ting Li, Chengchun Shi, Qianglin Wen, Yang Sui, Yongli Qin, Chunbo Lai, and Hongtu Zhu. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2024b.
- Shikai Luo, Ying Yang, Chengchun Shi, Fang Yao, Jieping Ye, and Hongtu Zhu. Policy evaluation for temporal and/or spatial dependent experiments. *Journal of the Royal Statistical Society, Series B*, 2024.

References II

- Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 118(543):2059–2071, 2023.
- Xiaocheng Tang, Zhiwei Qin, Fan Zhang, Zhaodong Wang, Zhe Xu, Yintai Ma, Hongtu Zhu, and Jieping Ye. A deep value-network based approach for multi-driver order dispatching. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1780–1790, 2019.
- Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 905–913, 2018.