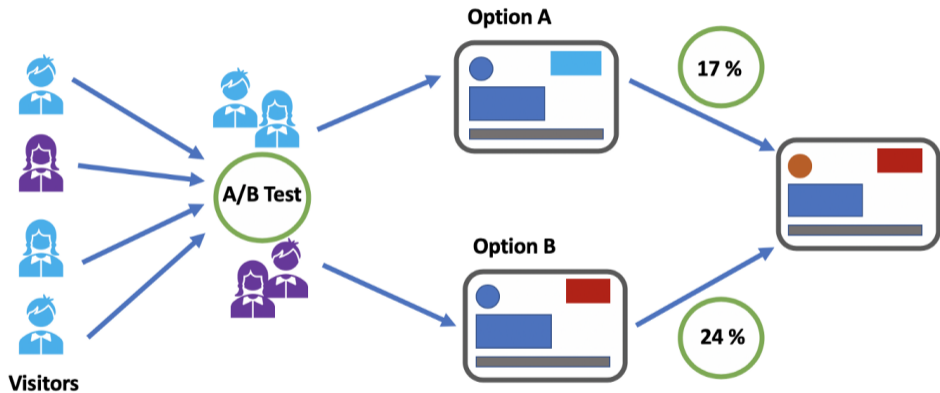


A Reinforcement Learning Framework for Dynamic Causal Effects Evaluation in A/B Testing

Chengchun Shi

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

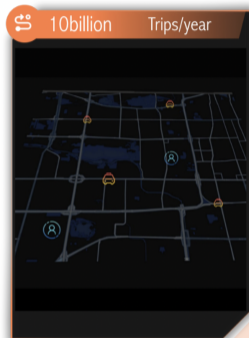
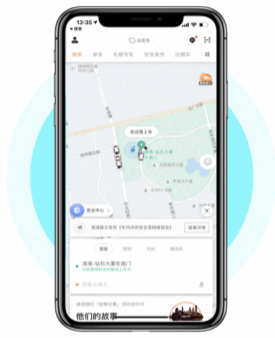
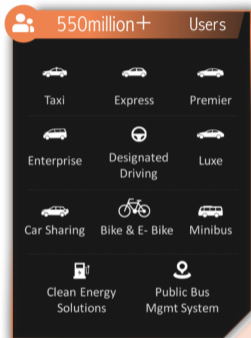
A/B Testing



Taken from

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

Ridesharing



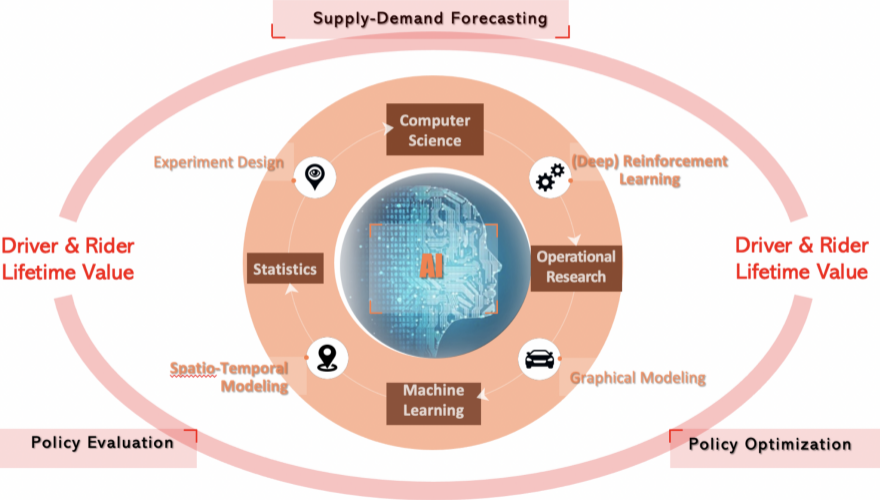
106TB+
vehicle trajectory data/day

4875TB+
data processed/day

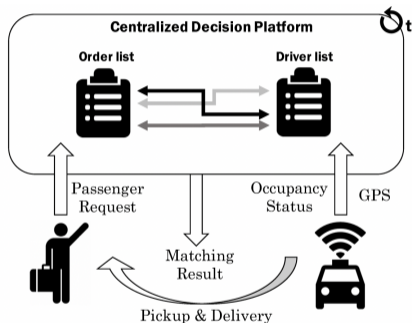
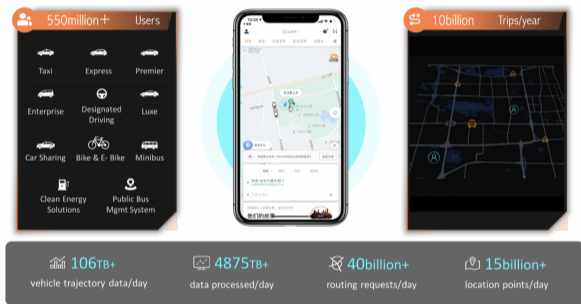
40billion+
routing requests/day

15billion+
location points/day

Applications in Ridesharing



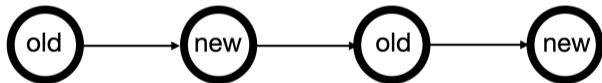
Motivation: Order Dispatch



Our project is motivated by the need for comparing the **long-term rewards** of different **order dispatching** policies in **ridesharing platforms**

Data

- Data from an **online experiment** that last for two weeks
- **30 minutes/1 hour** as one time unit
- **Time-varying variables** S_t : e.g., number of drivers (supply), number of call orders (demand)
- **Treatment** A_t : new policy v.s. old policy; adopts an alternating-time-interval (switchback) design

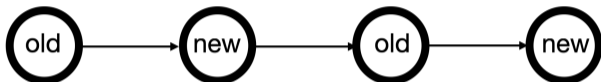


- **Outcome** R_t :
 - **Answer rate** (percentage of call orders being responded by drivers)
 - **Completion rate** (percentage of call orders being completed)
 - **Drivers' income**

Challenges

1. The existence of **carryover effects**:

- Under the alternating-time-interval (or switchback) design



- Past actions will affect future outcomes

2. The need for **early termination**:

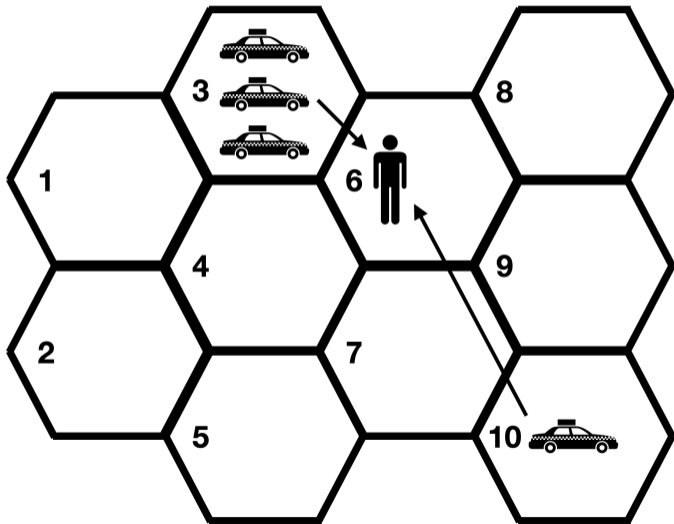
- Each experiment takes a considerable time (at most 2 weeks)
- Early termination to save time and budget

3. The need for **adaptive randomization**:

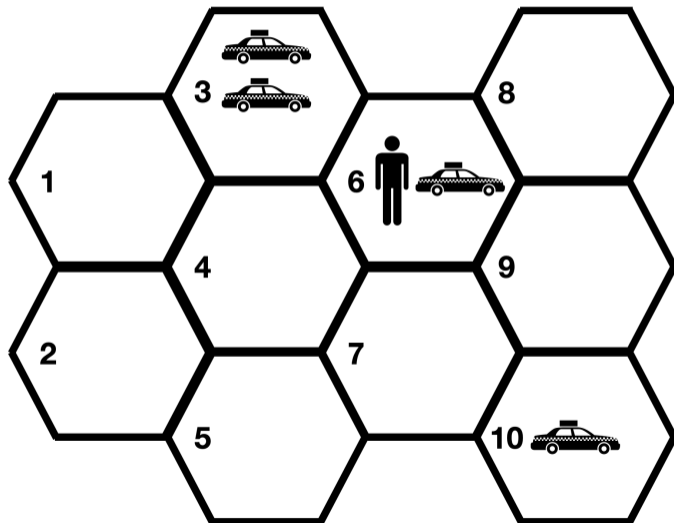
- Maximize the total reward (e.g., epsilon-greedy)
- Detect the alternative faster

To our knowledge, **no** existing test has addressed three challenges simultaneously

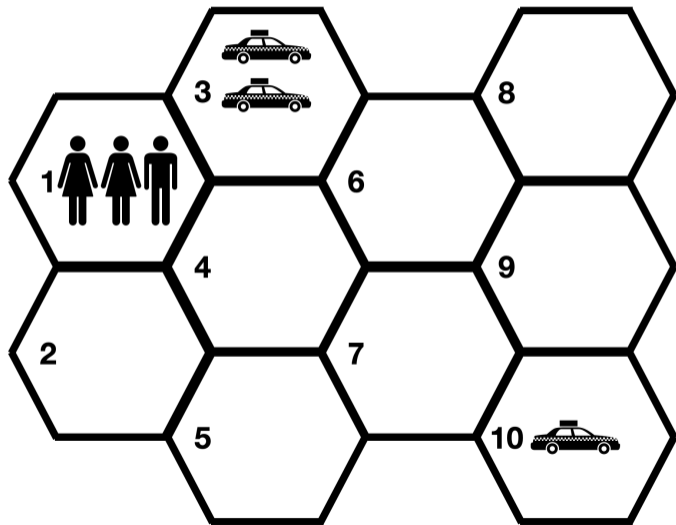
Illustration of the Carryover Effects



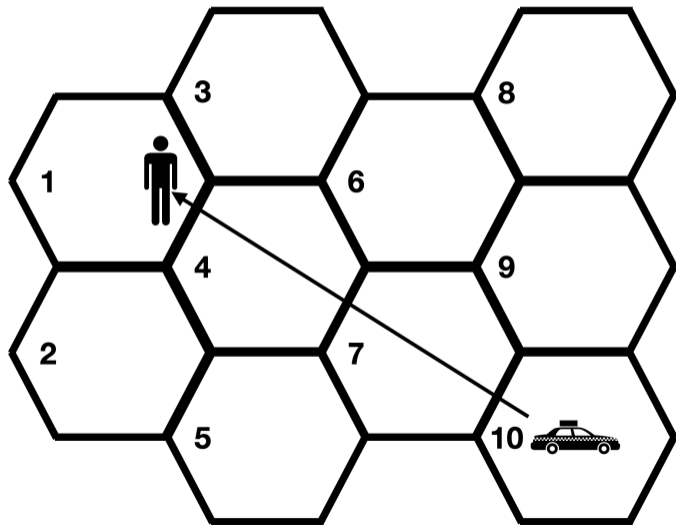
Adopting the Closest Driver Policy



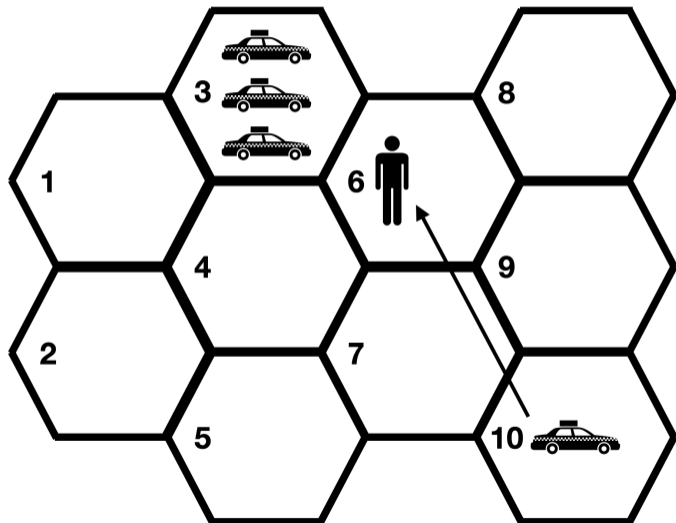
Some Time Later ...



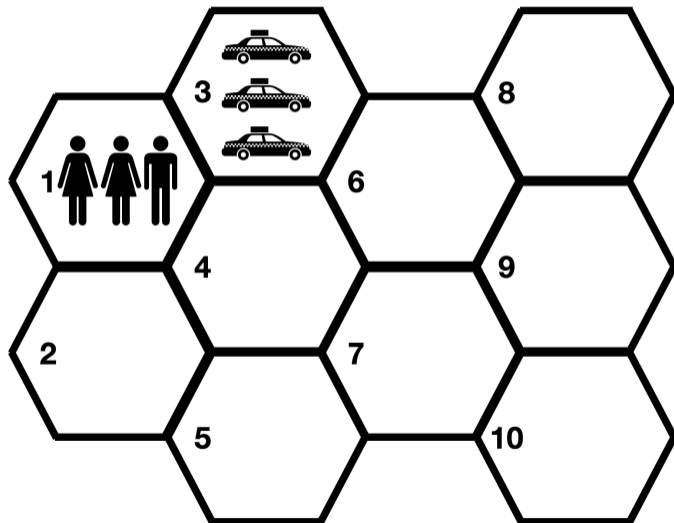
Miss One Order



Consider a Different Action



Able to Match All Orders



Existence of Carryover Effects

past actions → distribution of drivers → future rewards

Limitations of Existing A/B tests

- Most existing tests **cannot** detect carryover effects
- \mathcal{H}_0 : The old policy ($A = 0$) has larger cumulative rewards
- \mathcal{H}_1 : The new policy ($A = 1$) has larger cumulative rewards
- **Example 1.** $S_t \sim N(0, 0.25)$, $R_t = S_t + \delta A_t$
- **Example 2.** $S_t = 0.5S_{t-1} + A_{t-1} + N(0, 0.25)$, $R_t = S_t$

Example 1	t-test 0.76	DML-based test 1.00	our test 0.98
Example 2	t-test 0.04	DML-based test 0.06	our test 0.73

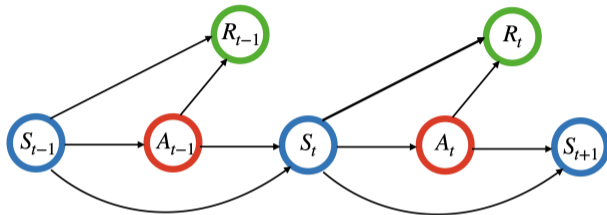
Table: Powers of t-test, DML-based test (Chernozhukov et al., 2018) and the proposed test with $T = 500$, $\delta = 0.1$

Contributions and Advances of Our Proposal

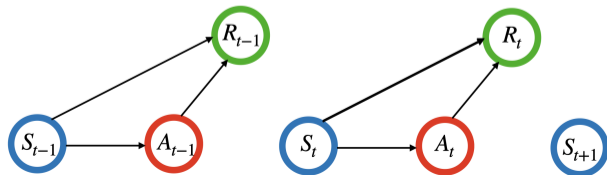
- Introduce an RL framework for A/B testing
 1. Allow to measure long-term rewards using **value function**
 2. Model carryover effects using the **dynamic system transitions** (address Challenge 1)
 3. Enable **consistent** estimation with a **single** time series
- Propose an original test procedure for comparing long-term rewards of two policies
 1. allows for **sequential monitoring** (address Challenge 2)
 2. allows for **online updating**
 3. applicable to a wide range of designs, including the **Markov** design, **alternating-time-interval** design and **adaptive** design (address Challenge 3)

An RL Framework for A/B Testing

- Allows to model the carryover effects using the **dynamic state transitions**



- A_{t-1} impacts R_t indirectly through its effect on S_t
 - S_t shall include important **mediators** between A_{t-1} and R_t
- Most existing works require the independence assumption



Contributions and Advances (Cont'd)

Algorithm	Carryover effects	Sequential monitoring	Switchback design
Two-sample t-test	X	X	✓
Classical sequential tests	X	✓	✓
Bojinov & Shephard (2019)	✓	X	X
V-learning (Lockett et al., 2020)	✓	X	X
Double RL (Kallus & Uehara, 2019)	✓	X	X
CausalRL (our proposal)	✓	✓	✓

Methodology

- Apply **temporal difference learning** with **sieve** method to evaluate value difference and provide **uncertainty quantification** (Shi et al., 2021, JRSSB)
- Adopt the **α -spending approach** (Lan & DeMets, 1983) for sequential monitoring
- Develop a **bootstrap-assisted procedure** for determining the stopping boundary
 - The numerical integration method designed for classical sequential tests is **not** applicable in adaptive design, due to the carryover effects

Theory

Theorem (Validity and Consistency)

*Under the Markov, alternating-time-interval or adaptive design, the proposed test can **control type-I error** and is **consistent** against alternatives that converge to the null at the parametric rate*

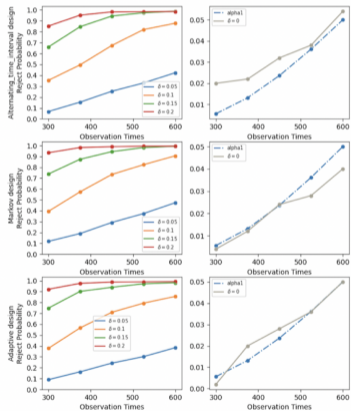
Theory (Cont'd)

Theorem (Undersmoothing and Efficiency)

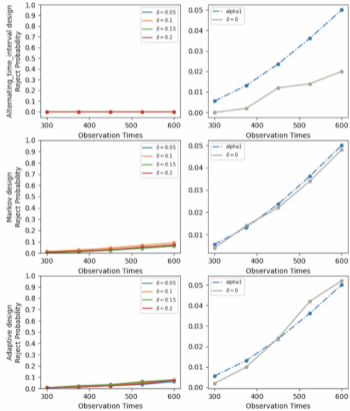
Suppose **sieve** method is used for function approximation in temporal difference learning.

1. **Undersmoothing** is not needed to guarantee that the resulting value estimator has a tractable limiting distribution.
2. The value estimator is **semiparametrically efficient**.
 - Sieve estimators of conditional expectations are **idempotent** (Shen et al., 1997)
 - The proposed test will **not** be overly sensitive to the number of basis functions
 - **Cross-validation** can be employed to select the basis functions

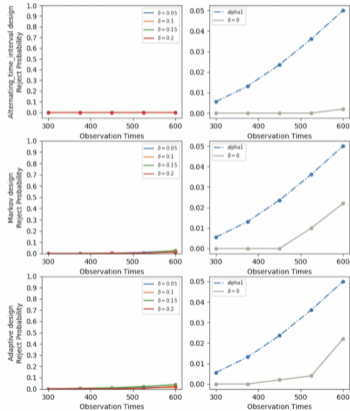
Simulation



(a) Power and size of our test



(b) Power and size of t test



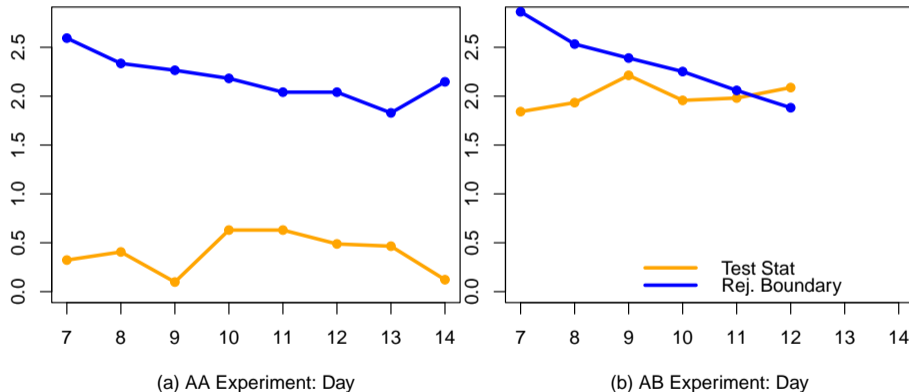
(c) Power and size of a version of the O'Brien Fleming sequential test

Application to Ridesharing Platform

- **Data:** a given city from December 3rd to 16th (two weeks)
- **30 minutes** as one time unit, sample size = **672**
- **State:**
 1. number of drivers (supply)
 2. number of requests (demand)
 3. supply and demand equilibrium metric (mediator)
- **Action:** new policy $A = 1$ v.s. old $A = 0$
- **Reward:** drivers' income
- The new policy is expected to have **better** performance

Application to Ridesharing Platform (Cont'd)

- The proposed test



- t-test: **fail** to reject \mathcal{H}_0 in A/B experiment with p-value 0.18

Thank You!

😊 Dynamic Causal Effects Evaluation in A/B Testing with a
Reinforcement Learning Framework (JASA, accepted)

Papers and softwares can be found on my personal website

`callmespring.github.io`