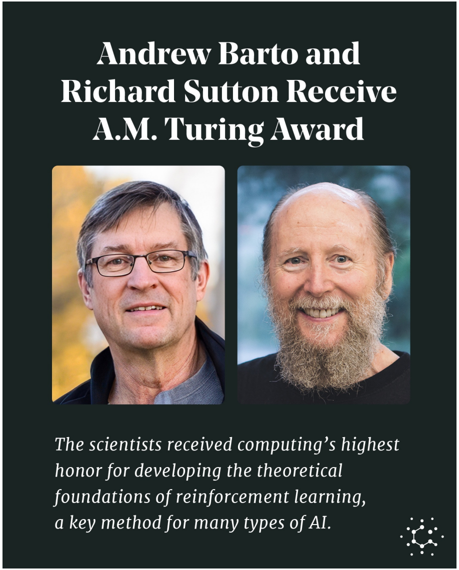
From Theory to Practice: Challenges in Real-World Reinforcement Learning

Chengchun Shi

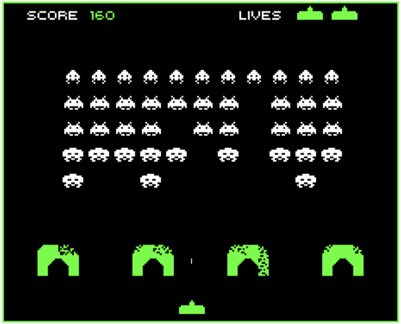
Associate Professor of Data Science

London School of Economics and Political Science

Reinforcement Learning (RL)



Developing AI with RL



Video Games

Go

RL Applications

Mobile health

Ride-sharing

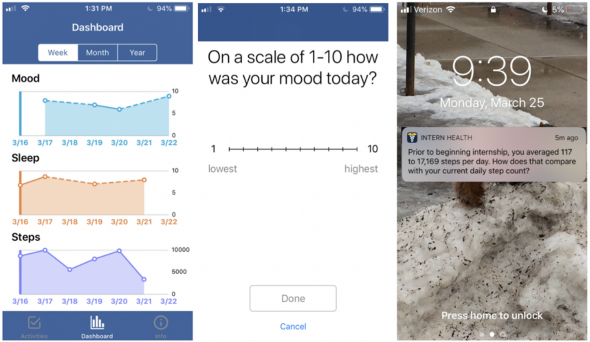
Deep brain stimulation

Psychology

Large language models

Mobile Heath (mHealth)

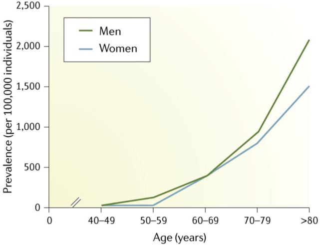
- Use of cellphones and wearable devices in healthcare
- **Data**: Intern Health Study (NeCamp et al., 2020)
- **Subject**: First-year medical interns working in stressful environments (e.g., long work hours and sleep deprivation)
- **Objective**: Promote physical and mental well-beings
- **Intervention**: Determine whether to send certain text message to a subject



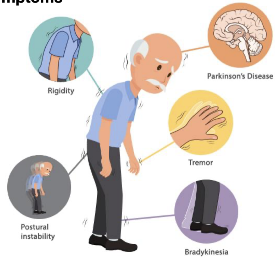
(i) App Dashboard (ii) Mood EMA (iii) Notifications

Deep Brain Stimulation

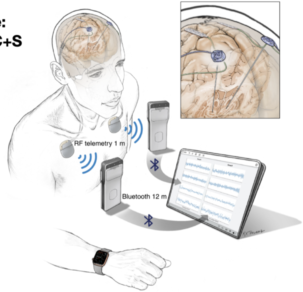
PD Prevalence



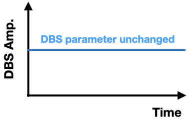
PD Symptoms



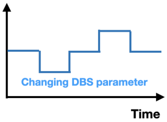
DBS device: Medtronic RC+S



Current clinical practice: Continuous DBS (cDBS)



Can we do better? Adaptive DBS (aDBS)

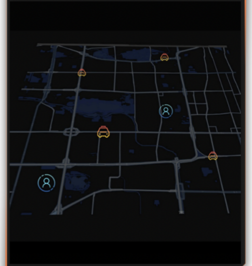


Ridesharing

550million+ Users

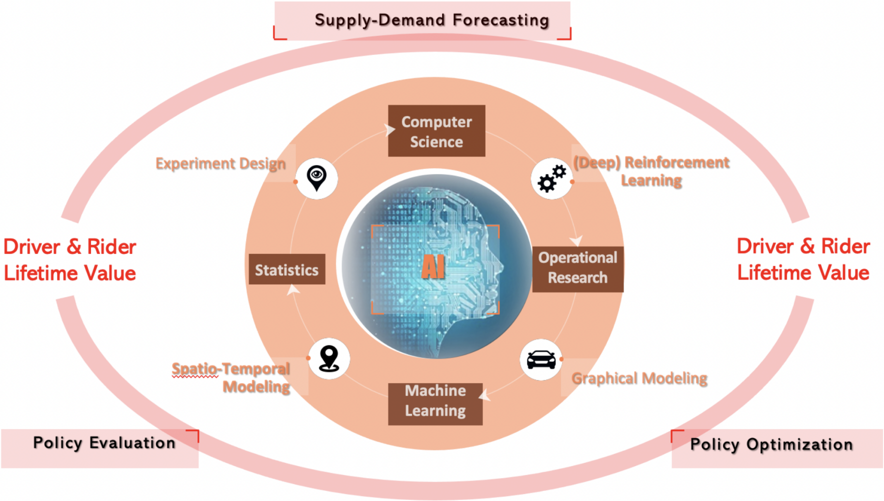
Taxi	Express	Premier
Enterprise	Designated Driving	Luxe
Car Sharing	Bike & E- Bike	Minibus
Clean Energy Solutions	Public Bus Mgmt System	

10billion Trips/year



106TB+	4875TB+	40billion+	15billion+
vehicle trajectory data/day	data processed/day	routing requests/day	location points/day

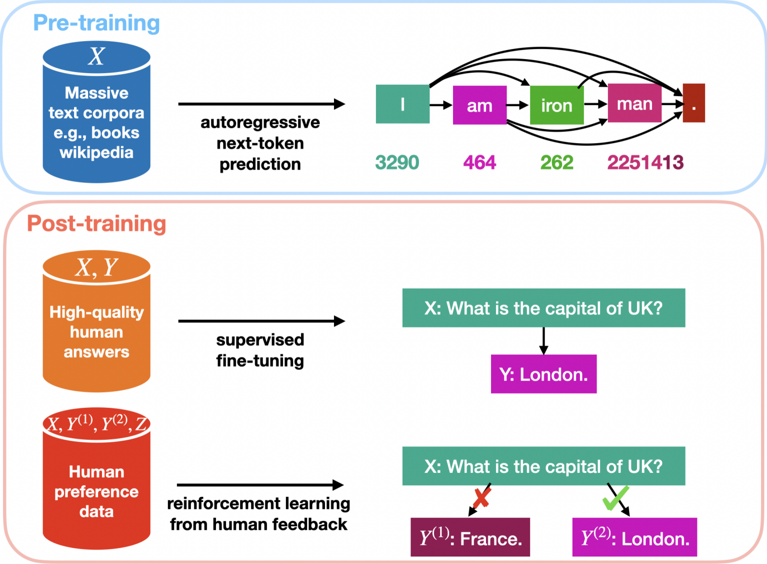
Ridesharing (Cont'd)



Large Language Models (LLM)

Note

- X: a sentence or prompt.
- Y: responses.
- Z: $Z = \mathbb{I}(Y^{(2)} \succ Y^{(1)})$ represents the resulting human feedback



Reinforcement Learning from Human Feedback

2017

Deep Reinforcement Learning from Human Preferences

Paul F Christiano
OpenAI
paul@openai.com

Jan Leike
DeepMind
leike@google.com

Tom B Brown
nottombrown@gmail.com

Miljan Martic
DeepMind
miljanm@google.com

Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com

First introduction to deep RLHF

2022

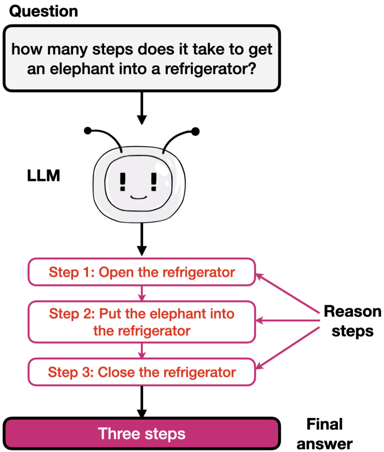
Training language models to follow instructions with human feedback

Long Ouyang* **Jeff Wu*** **Xu Jiang*** **Diogo Almeida*** **Carroll L. Wainwright***
Pamela Mishkin* **Chong Zhang** **Sandhini Agarwal** **Katarina Slama** **Alex Ray**
John Schulman **Jacob Hilton** **Fraser Kelton** **Luke Miller** **Maddie Simens**
Amanda Askell† **Peter Welinder** **Paul Christiano*†**
Jan Leike* **Ryan Lowe***

OpenAI

First successful application of RLHF to LLM

Reinforcement Learning with Verifiable Rewards



deepseek

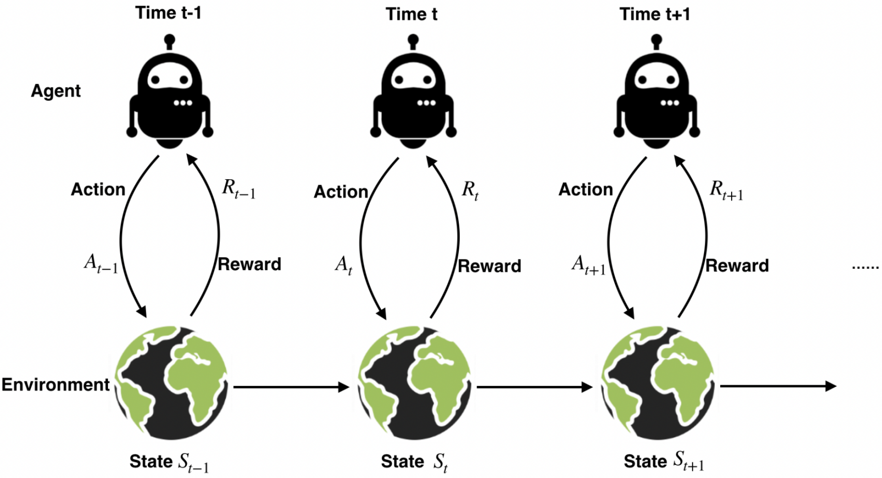
DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2*†], Peiyi Wang[1,3*†], Qihao Zhu[1,3*†], Runxin Xu[1], Junxiao Song[1]
Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

[1]DeepSeek-AI, [2]Tsinghua University, [3]Peking University

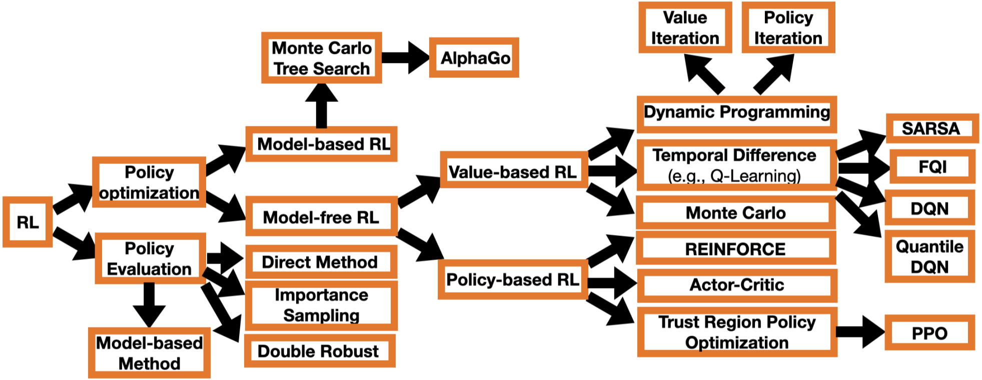
{zhihongshao,wangpeiyi,zhuqh,guoday}@deepseek.com
https://github.com/deepseek-ai/DeepSeek-Math

What Is RL?



Objective: find an optimal policy that maximizes the cumulative reward

Many RL Algorithms Were Proposed...



But far fewer have found successful applications in healthcare

Gap between Theory & Practice

- **Action** is well-defined in most applications

- So is **reward** (LLM being one exception)

- Can we identify a proper **state**?

The main challenge

Gap between Theory & Practice

- **Action** is well-defined in most applications
- So is **reward** (LLM being one exception)
- Can we identify a proper **state**?

It depends

Yes

How to identify the state

- Hypothesis testing
- State abstraction

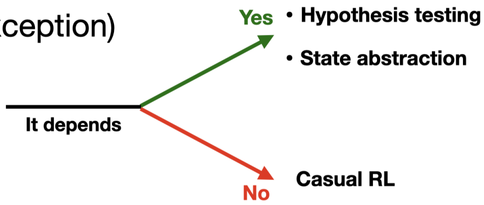
  — similar to confounder selection in causal inference

No

Casual RL

Gap between Theory & Practice

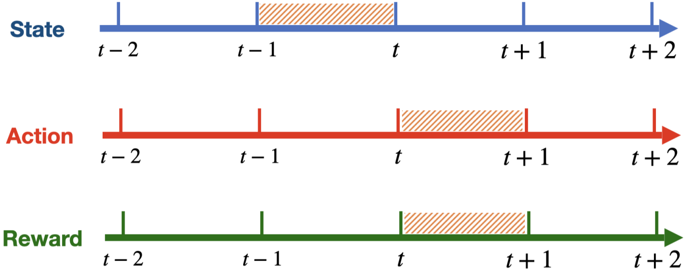
- **Action** is well-defined in most applications
- So is **reward** (LLM being one exception)
- Can we identify a proper **state**?



- **RL is inherently a causal inference problem.**
- Causal inference answers *what if* questions:
  - *What would happen under different **interventions**?*
- Similarly, RL asks *what if we adopt this **policy**?*
  - *How will it affect the expected return?*
- **Value functions** in RL is closely related to **potential outcomes** in causal inference

How to Identify the State

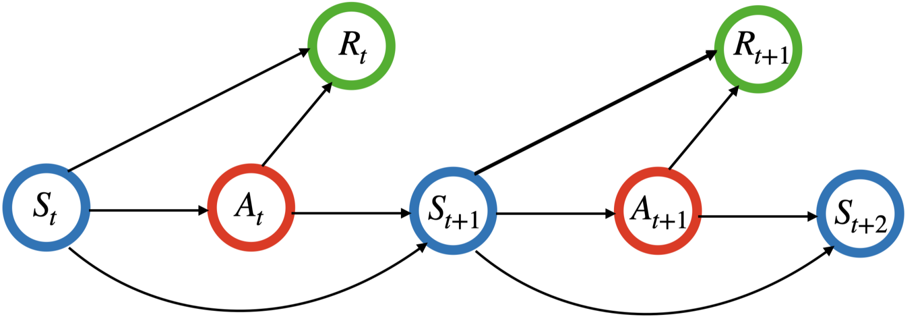
Rule 1: States be collected prior to actions and rewards



- ***Assumption 1:*** $S_t \rightarrow A_t/R_t$***, not the other way around***

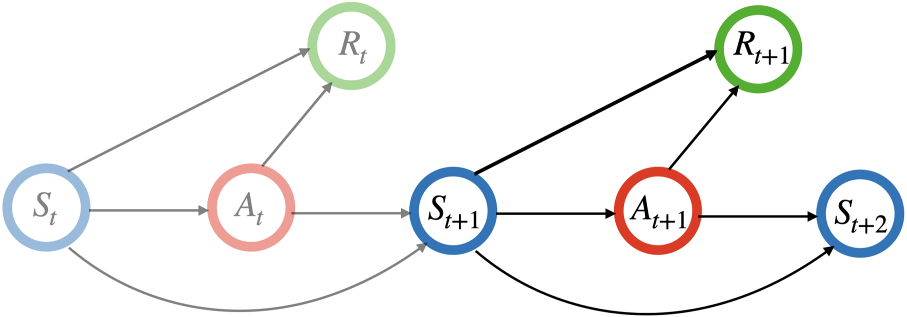
How to Identify the State (Cont'd)

Rule 2: States be chosen to make the system an MDP



How to Identify the State (Cont'd)

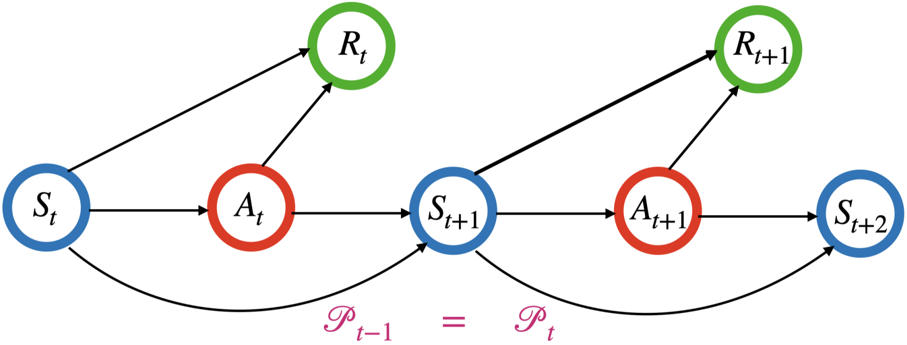
Rule 2: States be chosen to make the system an MDP



- ***Assumption 2(a): Markov assumption***

How to Identify the State (Cont'd)

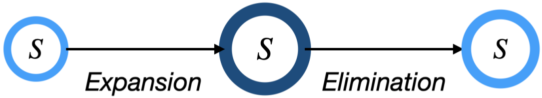
Rule 2: States be chosen to make the system an MDP



- *Assumption 2(a): Markov assumption*
- ***Assumption 2(b): Time-homogeneity assumption***

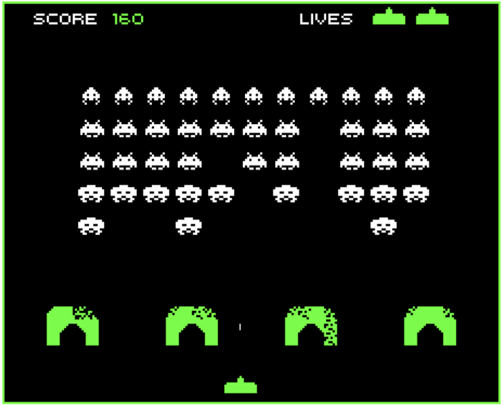
To Meet Assumptions 2(a): Markovanity

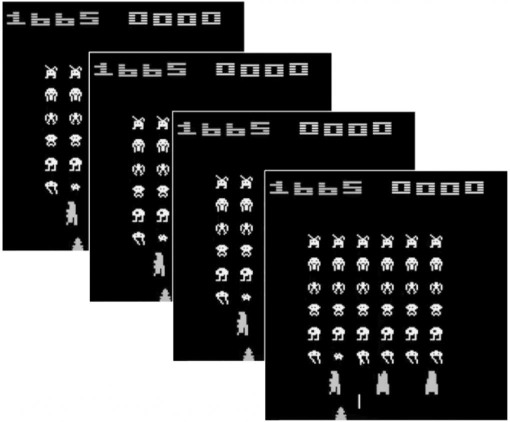
Double-E procedure: (*Expansion & Elimination*)



To Meet Assumptions 2(a): Markovanity

Double-E procedure: (*Expansion* & *Elimination*)

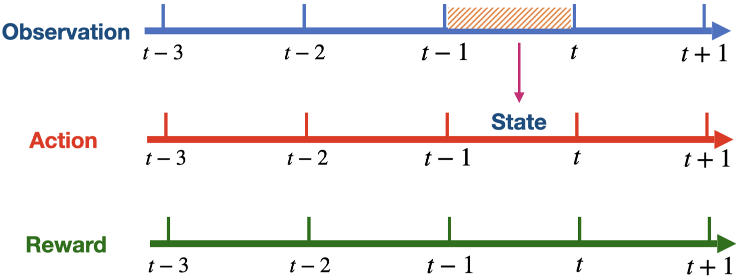




In DQN, state is a stack of 4 most recent frames (Mnih, et al., 2015, *Nature*)

To Meet Assumptions 2(a): Markovanity

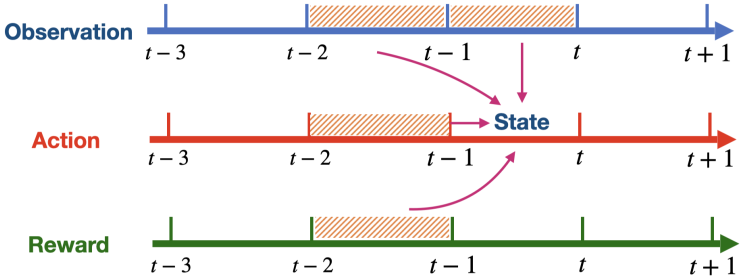
Double-E procedure: (*Expansion* & *Elimination*)



Test the Markov assumption (Chen and Hong, et al., 2012, *Econometric Theory*; Shi et al., 2020, *ICML*; Zhou et al., *2023*)

To Meet Assumptions 2(a): Markovanity

Double-E procedure: (*Expansion* & *Elimination*)

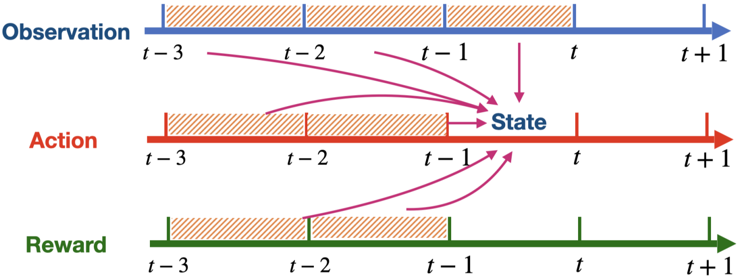


If rejected: MA does not hold

Test the Markov assumption (Chen and Hong, et al., 2012, *Econometric Theory*; Shi et al., 2020, *ICML*; Zhou et al., *2023*)

To Meet Assumptions 2(a): Markovanity

Double-E procedure: (*Expansion* & *Elimination*)

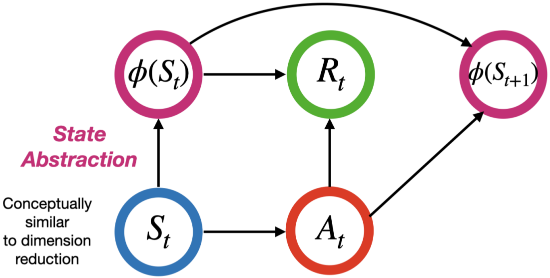


If rejected: MA does not hold

Test the Markov assumption (Chen and Hong, et al., 2012, *Econometric Theory*; Shi et al., 2020, *ICML*; Zhou et al., *2023*)

To Meet Assumptions 2(a): Markovanity

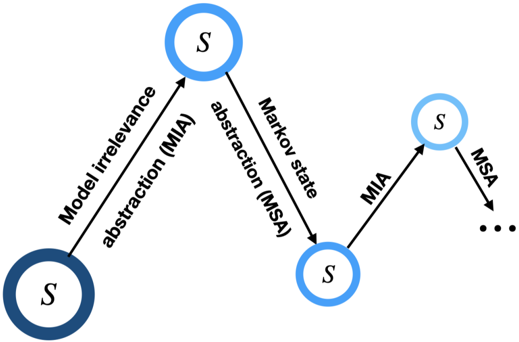
Double-E procedure: (*Expansion* & *Elimination*)



- Model irrelevance abstraction (Li et al., 2006, *AI&M*)
- Markov state abstraction (Allen et al., 2021, *NeurIPS*)

To Meet Assumptions 2(a): Markovanity

Double-E procedure: (*Expansion* & *Elimination*)



Hao et al. (2024; *Arxiv, 2406.19531*)

To Meet Assumption 2(b): Time-homogeneity

Approach 1: Include time index in the state

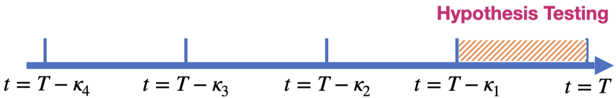
- *Day of week (e.g., Monday, Friday)*
- *Time of day (e.g., morning, afternoon)*

To Meet Assumption 2(b): Time-homogeneity

Approach 1: Include time index in the state

- *Day of week (e.g., Monday, Friday)*
- *Time of day (e.g., morning, afternoon)*

Approach 2: Change point detection



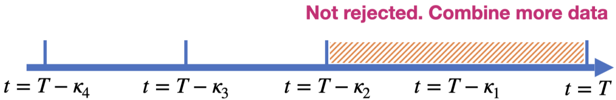
Test time-homogeneity (Padakandla, et al., 2020, *Applied Intelligence*; Alegre et al., 2021, *AAMAS*; Wang, et al., 2023, *ICML*; Li et al., 2025, *AoS*)

To Meet Assumption 2(b): Time-homogeneity

Approach 1: Include time index in the state

- *Day of week (e.g., Monday, Friday)*
- *Time of day (e.g., morning, afternoon)*

Approach 2: Change point detection



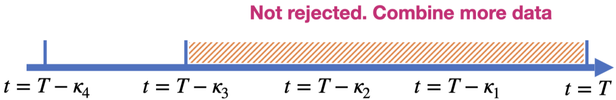
Test time-homogeneity (Padakandla, et al., 2020, *Applied Intelligence*; Alegre et al., 2021, *AAMAS*; Wang, et al., 2023, *ICML*; Li et al., 2025, *AoS*)

To Meet Assumption 2(b): Time-homogeneity

Approach 1: Include time index in the state

- *Day of week (e.g., Monday, Friday)*
- *Time of day (e.g., morning, afternoon)*

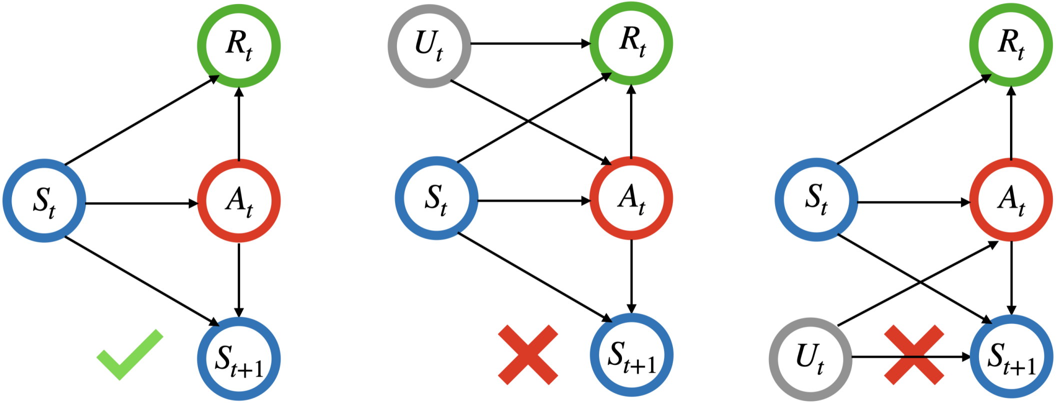
Approach 2: Change point detection



Test time-homogeneity (Padakandla, et al., 2020, *Applied Intelligence*; Alegre et al., 2021, *AAMAS*; Wang, et al., 2023, *ICML*; Li et al., 2025, *AoS*)

How to Identify the State (Cont'd)

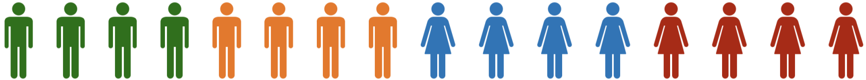
Rule 3: States be chosen to contain all confounders



- ***Assumption 3: No unmeasured confounders***

How to Identify the State (Cont'd)

Rule 4: All Subjects Possess Same Markov Transition Function



Approach 1: Include baseline information in the state

Approach 2: Clustering (Chen et al., 2025, JASA)

Approach 3: Transfer learning

Causal RL

- **Confounded POMDPs**:
  - Tennenholtz et al. (2020, *AAAI*)
  - Nair and Jiang (2021, *Arxiv*)
  - Shi et al. (2022, *ICML*)
  - Bennett and Kallus (2023, *OR*)
- **Confounded MDPs**:
  - Wang et al. (2021, *NeurIPS*)
  - Xu et al. (2023, *ICML*)
  - Shi et al. (2024, *JASA*)
  - Yu et al. (2024, *NeurIPS*)

Summary

Topics in RL		Topics in Statistics
• Offline policy optimization (Levine et al., 2022)	→	• Estimation
• Off-policy evaluation (Uehara et al., 2022)	→	• Confidence interval construction
• Non-Markovanity (Shi et al., 2020)	→	• Hypothesis testing
• Non-Stationary RL (Li et al., 2025)	→	• Changepoint detection
• Causal RL (Tennenholtz et al., 2020)	→	• Causal inference
• Behavior policy search (Hanna et al., 2017, 2024)	→	• Design of experiments
• RL from human feedback (Ouyang et al., 2022)	→	• Ranking models
• State abstraction (Li et al., 2006)	→	• Dimension reduction

Thank You!

☺My RL short course