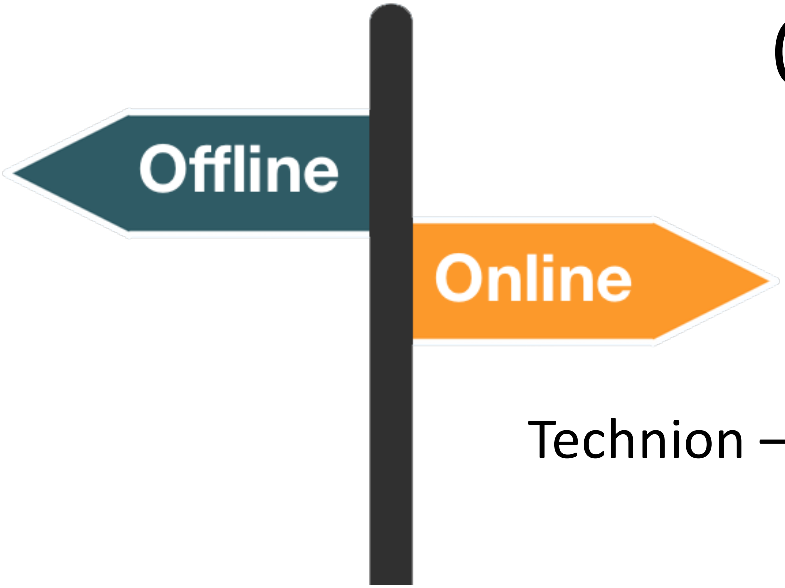
# Online Reinforcement Learning With The Help Of Confounded Offline Data



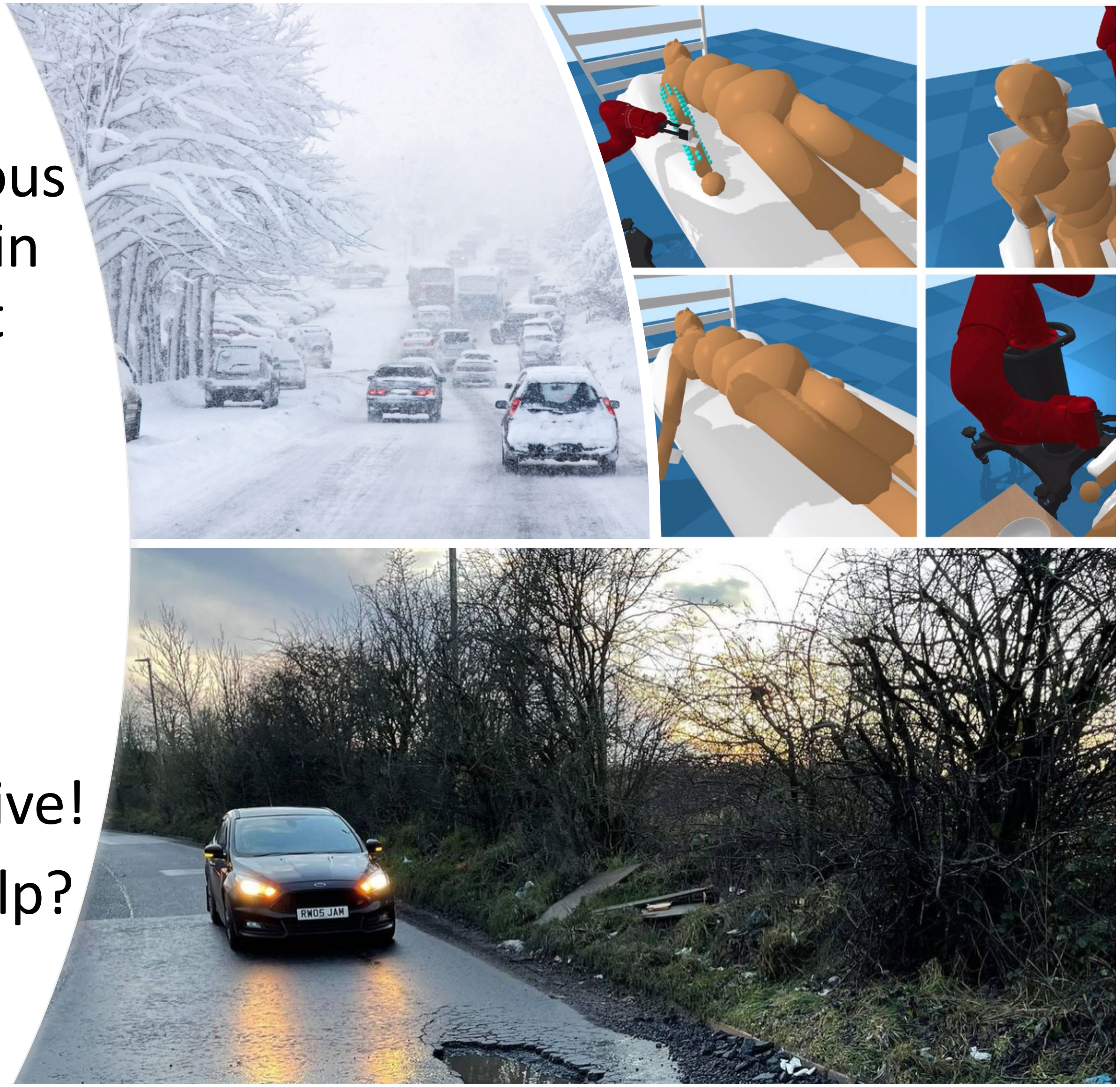

Uri Shalit
Technion – Israel Institute of Technology

Simons Institute Workshop on Learning from Interventions

February 2022

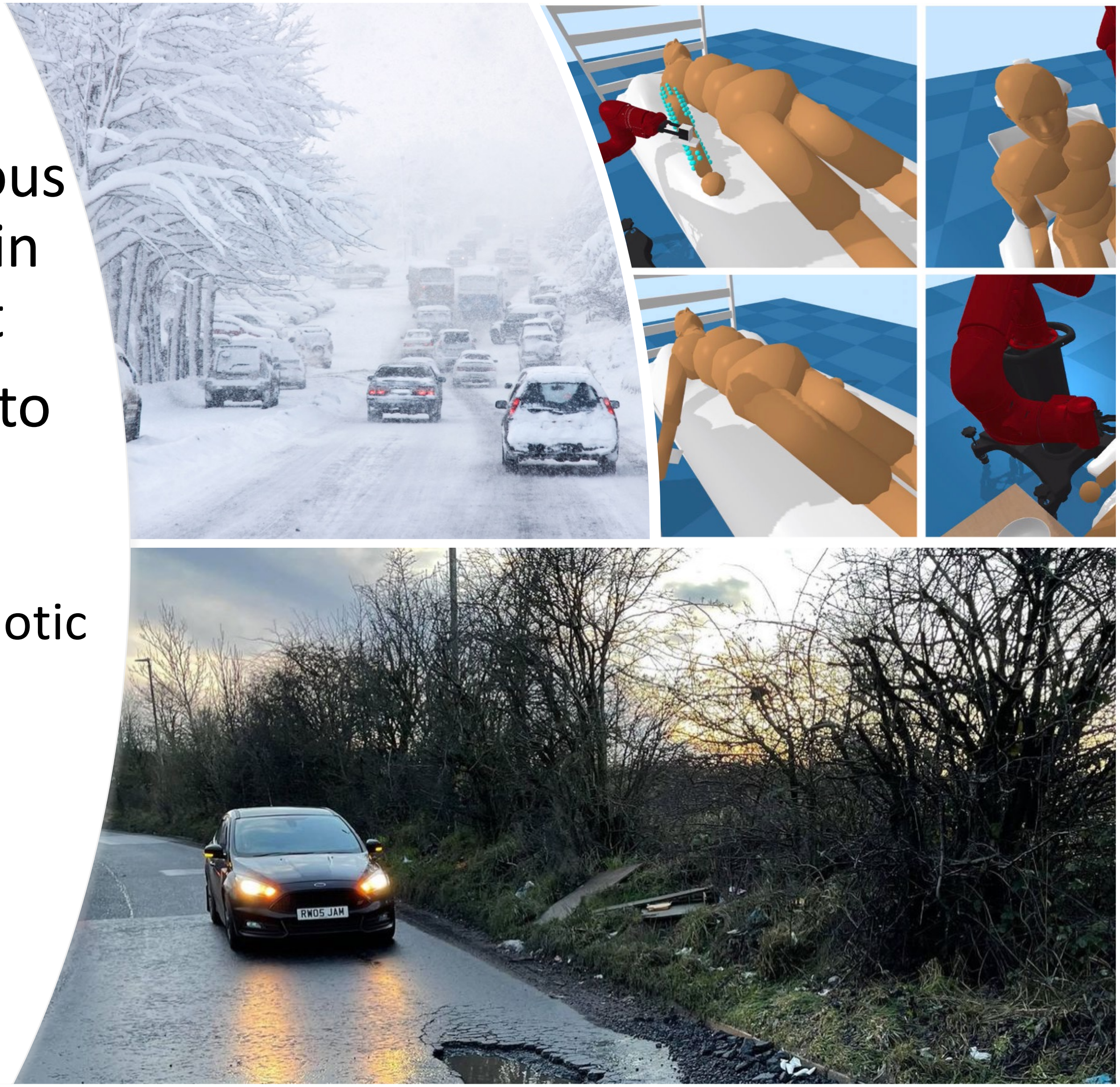
# Motivation

- Consider tasks like autonomous driving, robotics, or perhaps in the future medical treatment
- Many actions, long-term dependencies, high-dim state-spaces
- Learning online is crucial
  - Causally “optimal”
- But learning online is expensive!
- Can (massive) offline data help?
  - Save money / interactions / mistakes?



# Motivation

- Consider tasks like autonomous driving, robotics, or perhaps in the future medical treatment
- Assumption: we have access to large offline data
  - Logs from driving
  - Human demonstrations of robotic tasks
- Challenges:
  - Confounding
  - Partial observability
  - Distribution shifts



# Learning to act (intervene) with offline data

Obvious baselines

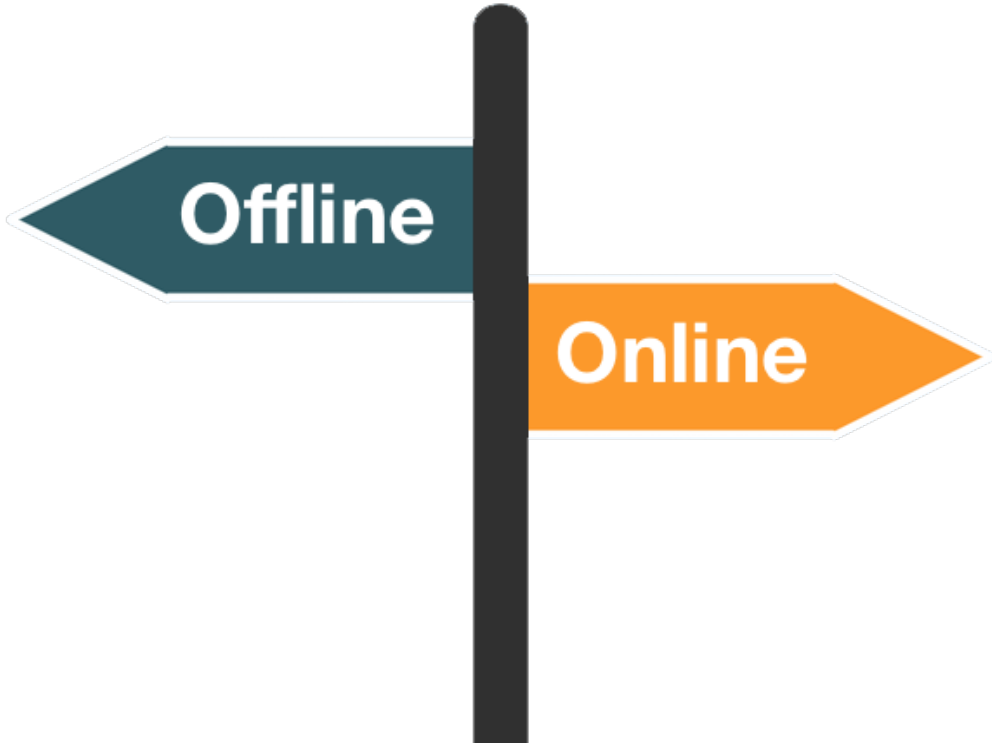
1. Don’t use offline data at all
    - Most of the bandit and RL literature
2. Don’t use online data at all
    - Off-policy RL
    - Vulnerable to hidden confounding and distribution shifts
    - Proxies might help (Tennenholtz 2020, Nair & Jiang 2021, Kallus et al. 2021, Shi et al. 2021)

- This talk: how to use merge offline & online in challenging scenarios
    - We are not the first, see e.g. Bareinboim & Pearl 2013, Zhang & Bareinboim 2017, Kallus et al. 2018 and more

# Talk outline

How to act online with the help of offline data?

- Part I: Contextual bandits with confounded offline data
- Part II: Online imitation and reinforcemnt learning with offline data from a possibly different distribution

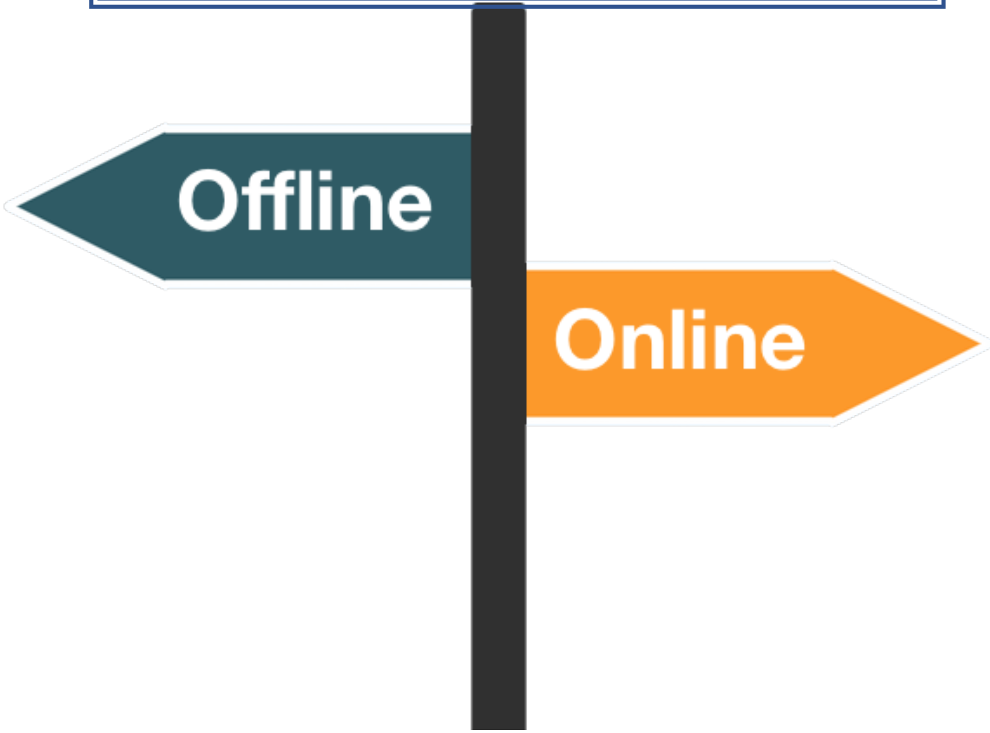


# Talk outline

“Bandits with partially observable confounded data”, Tennenholtz, S, Mannor, Efroni UAI 2021

How to act online with the help of offline data?

- **Part I: Contextual bandits with confounded offline data**
- Part II: Online imitation and reinforcemnt learning with offline data from a possibly different distribution



**Algorithm** A Linear Bandit Interaction Model

**for** $t = 1, 2, ., , ,$ **do**

Observe $x_t \sim \mu_X(\cdot)$

Take action $a_t(x_t)$ where $a \in [1, .., A]$

Receive noisy feedback $r_t = \langle x_t, w^*_{a_t} \rangle + \epsilon_t$

Suffer immediate regret $\max_a \langle x_t, w^*_a \rangle - \langle x_t, w^*_{a_t} \rangle$

**end for**

**Goal:** Minimize cumulative regret

$$\sum_t \max_a \langle x_t, w^*_a \rangle - \langle x_t, w^*_{a_t} \rangle$$

# Linear bandits

- Optimal action is context dependent
- No state
- Classic explore - exploit tradeoffs
- Goal is sub-linear regret, usually $\tilde{O}\left(\sqrt{T}\right)$ where $T$ is number or interactions / interventions / actions

---

**Algorithm** A Linear Bandit Interaction Model

---

**for** $t = 1, 2, ., , ,$ **do**

  Observe $x_t \sim \mu_X(\cdot)$

  Take action $a_t(x_t)$ where $a \in [1, .., A]$

  Receive noisy feedback $r_t = \langle x_t, w^*_{a_t} \rangle + \epsilon_t$

  Suffer immediate regret $\max_a \langle x_t, w^*_a \rangle - \langle x_t, w^*_{a_t} \rangle$

**end for**

---

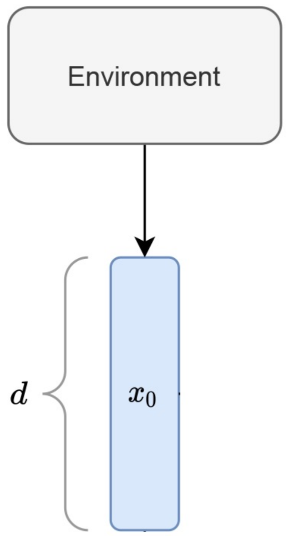
**Goal:** Minimize cumulative regret

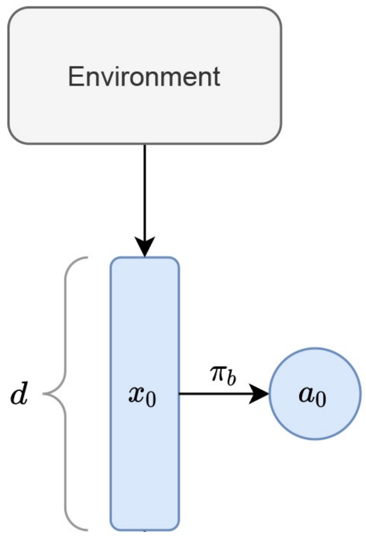
$$\sum_t \max_a \langle x_t, w^*_a \rangle - \langle x_t, w^*_{a_t} \rangle$$

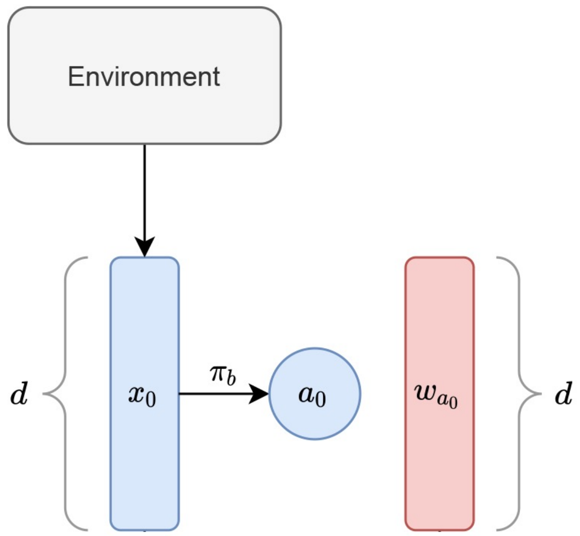
# Linear bandits

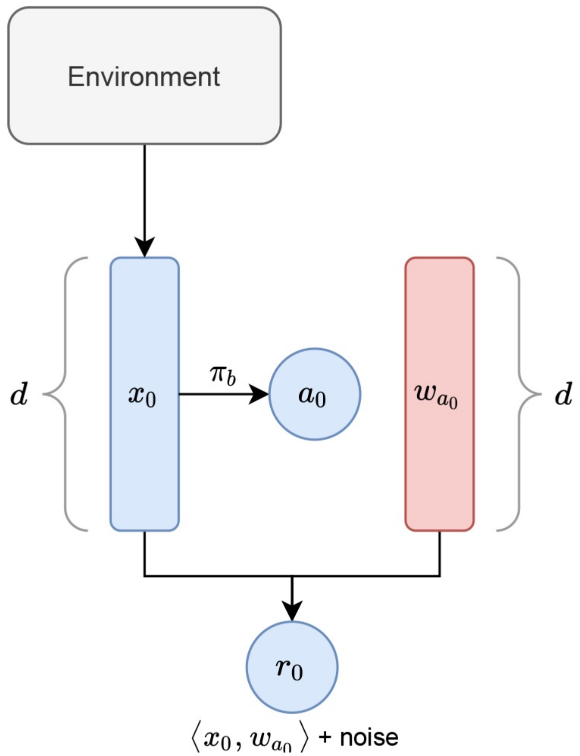
- Goal is sub-linear regret, usually $\tilde{O}\left(\sqrt{T}\right)$ where $T$ is number or interactions / interventions / actions
- Assume we have triplets of historic (context, action, reward) data
- If fully observed: can use learning from logged bandit feedback (e.g. Dudík et al. 2011, Swaminathan & Joachims 2015) to initalize online bandit
- What if the context in historic offline data is **partially observed**? E.g.:
  - Actions taken by humans
  - Not fully recorded
  - Privacy

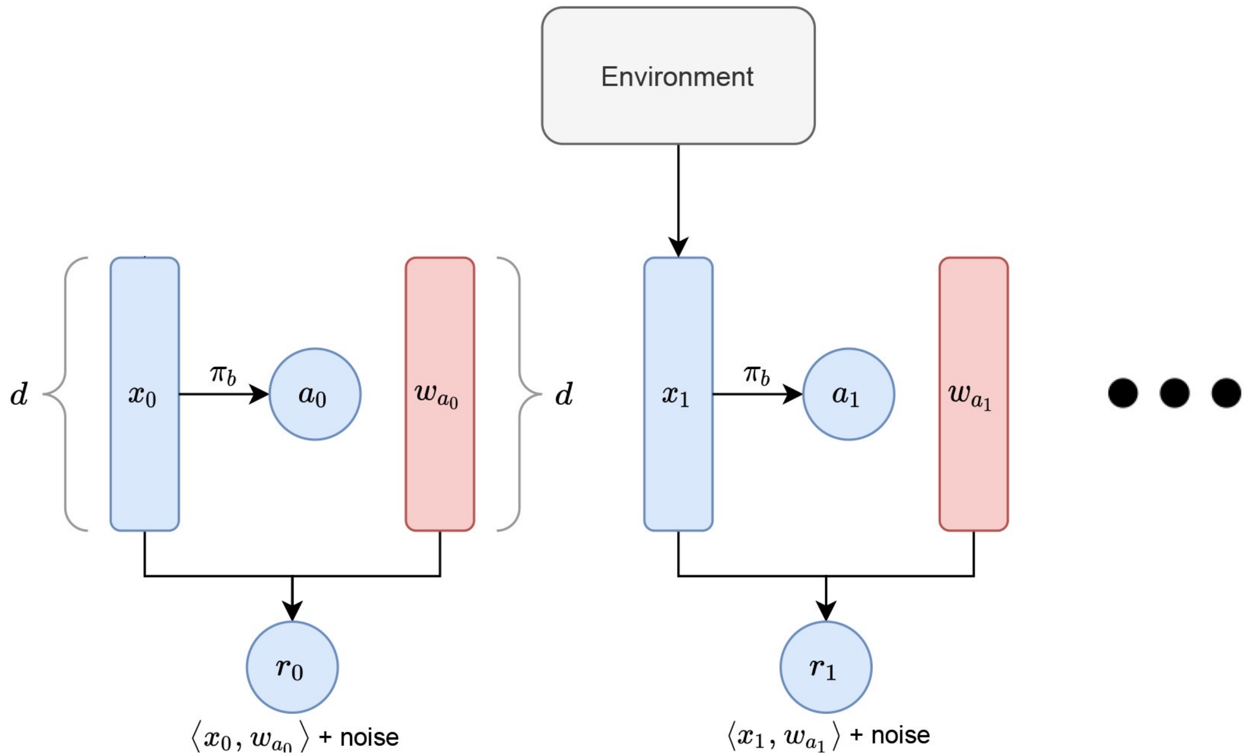

Environment

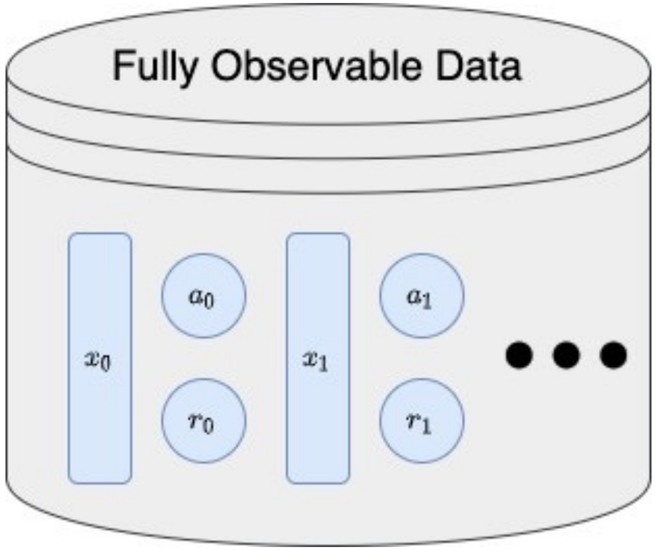

Environment
$d$
$x_0$

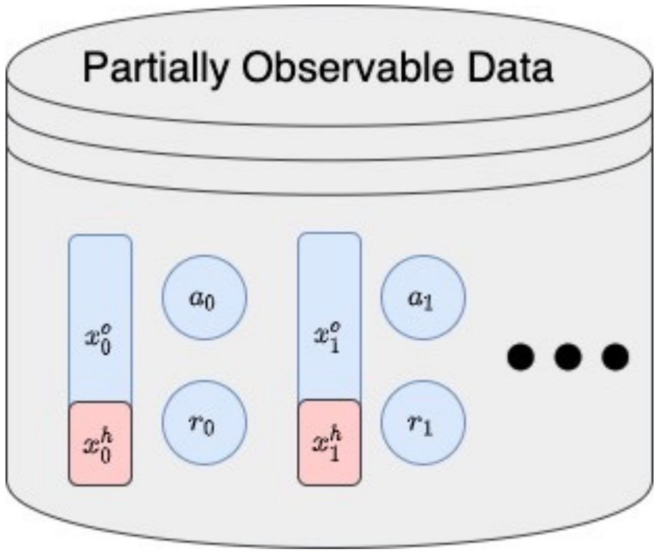

Environment
$d$
$x_0$
$\pi_b$
$a_0$


Environment
$x_0$
$\pi_b$
$a_0$
$w_{a_0}$
$d$
$d$


Environment
$d$
$x_0$
$\pi_b$
$a_0$
$w_{a_0}$
$d$
$r_0$
$\langle x_0, w_{a_0} \rangle$ + noise

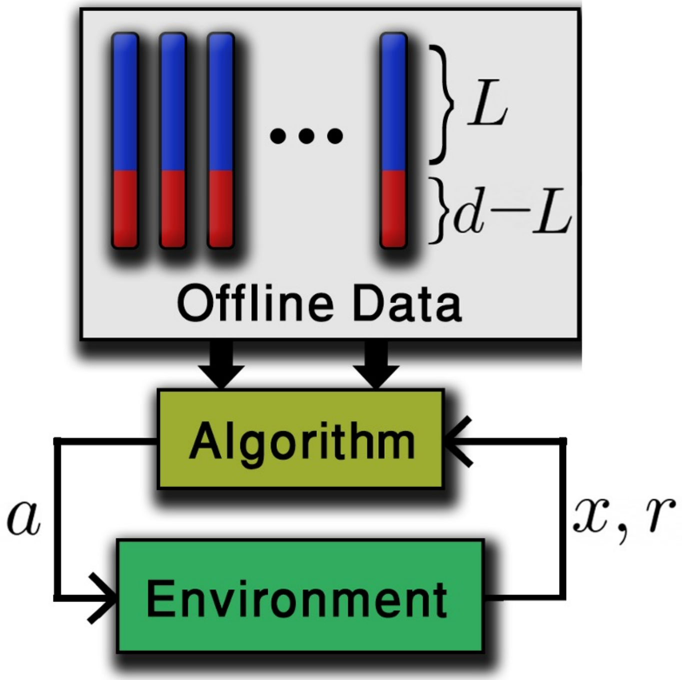

Environment
$d$
$x_0$
$\pi_b$
$a_0$
$w_{a_0}$
$d$
$x_1$
$\pi_b$
$a_1$
$w_{a_1}$
$r_0$
$\langle x_0, w_{a_0} \rangle$ + noise
$r_1$
$\langle x_1, w_{a_1} \rangle$ + noise


Fully Observable Data
$x_0$
$a_0$
$r_0$
$x_1$
$a_1$
$r_1$

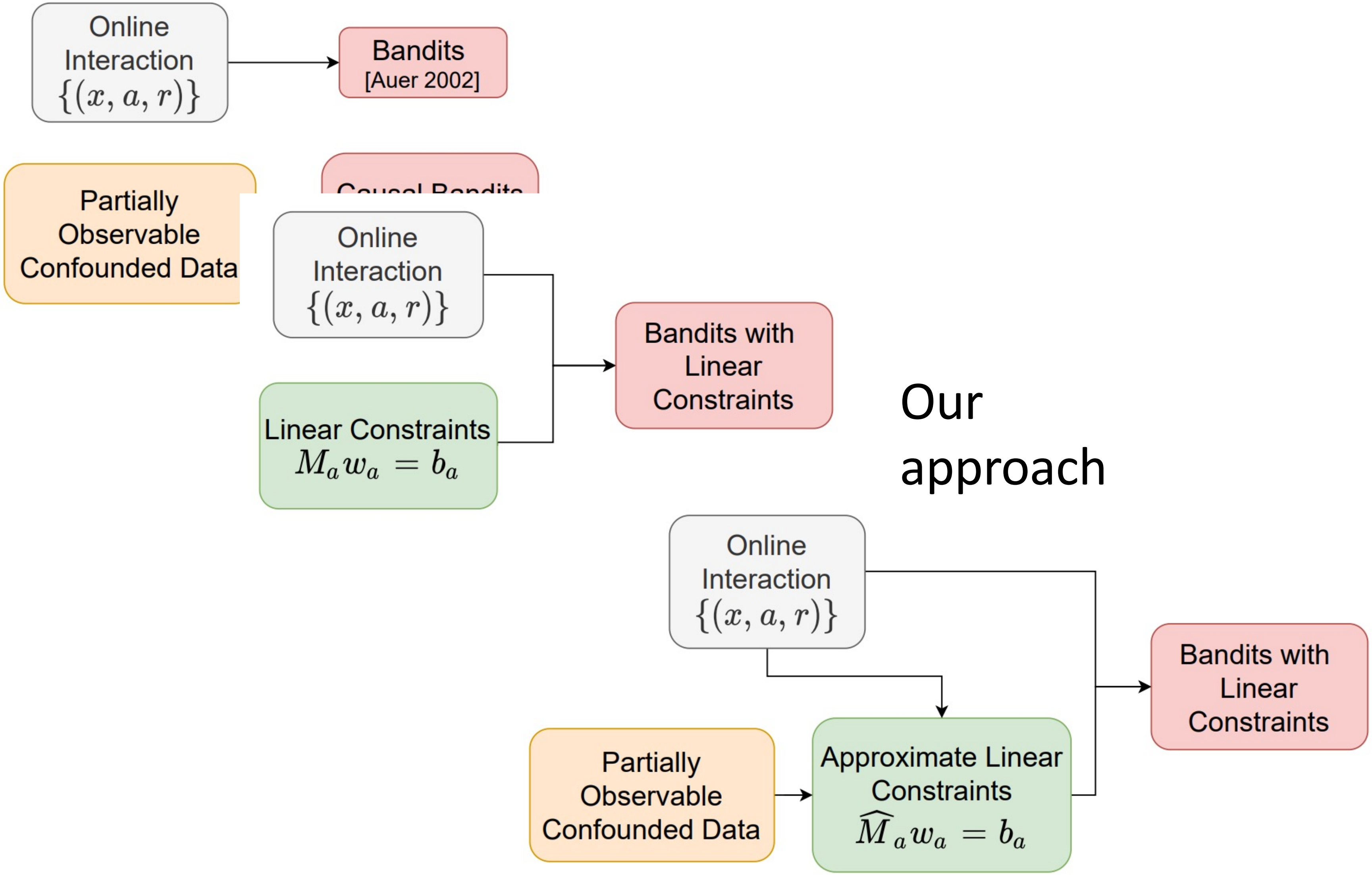

Partially Observable Data
$x_0^o$
$x_0^h$
$a_0$
$r_0$
$x_1^o$
$x_1^h$
$a_1$
$r_1$

# Learning with Partially Observable Data

- Access to partially observable offline data
- Context: $x = (x^o, x^h)$
  - $x^o \in \mathbb{R}^L$, $x^h \in \mathbb{R}^{d-L}$ denote the observed and unobserved features of the context
- Offline data was generated by an unknown, fixed behavior policy $\pi_b(a|x)$
- When online we act using the full $x$




- Without further assumptions the offline data might be almost useless
- E.g. all of the important information might be in $x_h$


Online Interaction $\{(x, a, r)\}$
Bandits [Auer 2002]
Partially Observable Confounded Data
Online Interaction $\{(x, a, r)\}$
Linear Constraints $M_a w_a = b_a$
Bandits with Linear Constraints
Our approach
Online Interaction $\{(x, a, r)\}$
Partially Observable Confounded Data
Approximate Linear Constraints $\widehat{M}_a w_a = b_a$
Bandits with Linear Constraints

# Observable consequences

**Proposition**

*Let the least-square estimator of* $\{r_n\}_{n=1}^{N}$ *be*

$$b^{LS}(a) = \left( \frac{1}{N_a} \sum_{i \in \{n : a_n = a\}} x_n^o (x_n^o)^T \right)^{-1} \left( \frac{1}{N_a} \sum_{n \in \{n : a_n = a\}} x_n^o r_n \right).$$

# Observable consequences

**Proposition**

*Let the least-square estimator of* $\{r_n\}_{n=1}^{N}$ *be*

$$b^{LS}(a) = \left( \frac{1}{N_a} \sum_{i \in \{n : a_n = a\}} x_n^o (x_n^o)^T \right)^{-1} \left( \frac{1}{N_a} \sum_{n \in \{n : a_n = a\}} x_n^o r_n \right).$$

*Define the following correlation matrices*
$R_{o,o}(a) = \mathbb{E}[x_i^o (x_i^o)^T | a, \pi_b]$, *and* $R_{o,h}(a) = \mathbb{E}[x_i^o (x_i^h)^T | a, \pi_b]$.

# Observable consequences

## Proposition

*Let the least-square estimator of* $\{r_n\}_{n=1}^N$ *be*

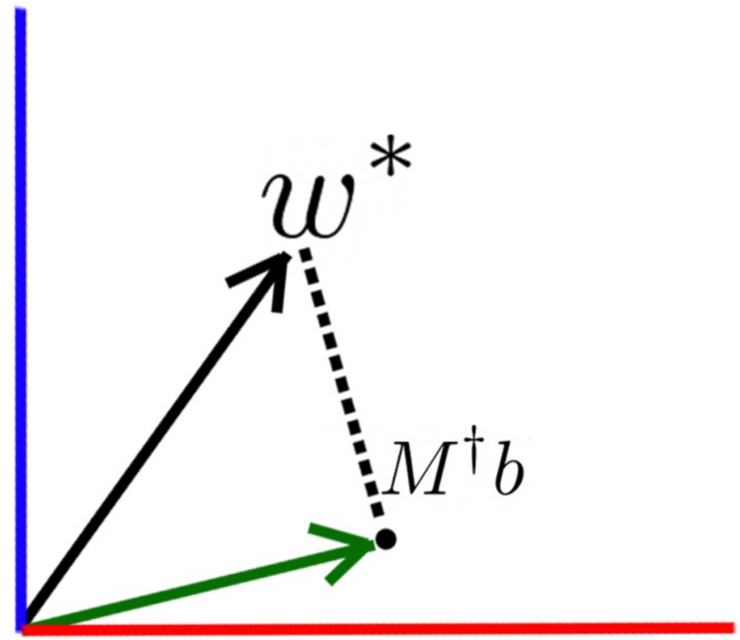
$$b^{LS}(a) = \left( \frac{1}{N_a} \sum_{i \in \{n : a_n = a\}} x_n^o (x_n^o)^T \right)^{-1} \left( \frac{1}{N_a} \sum_{n \in \{n : a_n = a\}} x_n^o r_n \right).$$

*Define the following correlation matrices* $R_{o,o}(a) = \mathbb{E}[x_i^o (x_i^o)^T | a, \pi_b]$, *and* $R_{o,h}(a) = \mathbb{E}[x_i^o (x_i^h)^T | a, \pi_b]$. *In the limit* $N \to \infty$ *and assuming* $\lambda_{\min}(R_{o,o}(a)) > 0$

$$b^{LS}(a) = \left( I_{L \times L} \quad R_{o,o}(a)^{-1} R_{o,h}(a) \right) w_a^*$$

# Observable consequences: linear constraints

- For every action $a$ we have $b^{LS}(a) = M(a)w_a^*$
  - $M(a) = \left(I_{L\times L} \quad R_{o,o}^{-1}(a)R_{o,h}(a)\right)$
  - Denote by $M(a)^\dagger$ the pseudo-inverse of $M(a)$
- At every online round, project current $\hat{w}_a$ to $M(a)^\dagger\, b_{LS}(a)$
- We prove we can reduce regret from $\mathcal{O}\left(d\sqrt{AT}\right)$ to $\mathcal{O}\left((d-L)\sqrt{AT}\right)$



# This is still not enough

- For every action $a$ we have $b^{LS}(a) = Mw_a^*$
  - $M(a) = \left(I_{L\times L} \quad R_{o,o}^{-1}(a)R_{o,h}(a)\right)$
  - Denote by $M(a)^\dagger$ the pseudo-inverse of $M(a)$
- At every online round, project current $\widehat{w}_a$ to $M(a)^\dagger\, b_{LS}(a)$
- From offline data we have:
  - $b^{LS}(a), R_{o,o}^{-1}(a)$
- Still missing $R_{o,h}(a) = \mathbb{E}\left[x^o\left(x^h\right)^\top \mid a, \pi_b\right]$
  - The covariance of hidden and observed features in offline data

Need some way to approximate

$$M(a)^{\dagger} = \left( I_{L\times L} \quad R_{o,o}^{-1}(a) R_{o,h}(a) \right)^{\dagger}$$

- We prove a result under general approximations of $R_{o,h}(a)$
- We further explore a specific assumption allowing approximation:
  during online operation we are allowed to query $\pi_b$
  - Similar to Zhang and Bareinboim (2016) notion of "intuition"
- Approximating pseudo-inverse $M(a)^{\dagger}$ only possible due to special structure of $M(a)$

**Theorem**

*Assume for every $t > 0$ we can sample $a \sim \pi_b(x)$. Then there exists a tractable algorithm such that for any $T > 0$, with probability at least $1 - \delta$, achieves regret*

$$Regret(T) \leq \widetilde{\mathcal{O}}\left((1 + f_{B_1})(d - L)\sqrt{AT}\right).$$

- $f_{B_1}$ is a factor indicating how hard it is to estimate the linear constraints
  - Relates to how well-spread $\pi_b$ is and how well conditioned and correlated are $R_{o,h}$ and $R_{o,o}$
- Worst case dependence: $f_{B_1} \leq \tilde{O}\left((L(d-L))^{1/4}\right)$
  - $d - L \sim O(d)$, $\text{Regret}(T) \leq d^{5/4}\sqrt{AT}$, worse than discarding the data
  - $d - L \sim O(1)$, $\text{Regret}(T) \leq d^{1/4}\sqrt{AT}$, improved performance

**Theorem**

*Assume for every $t > 0$ we can sample $a \sim \pi_b(x)$. Then there exists a tractable algorithm such that for any $T > 0$, with probability at least $1 - \delta$, achieves regret*

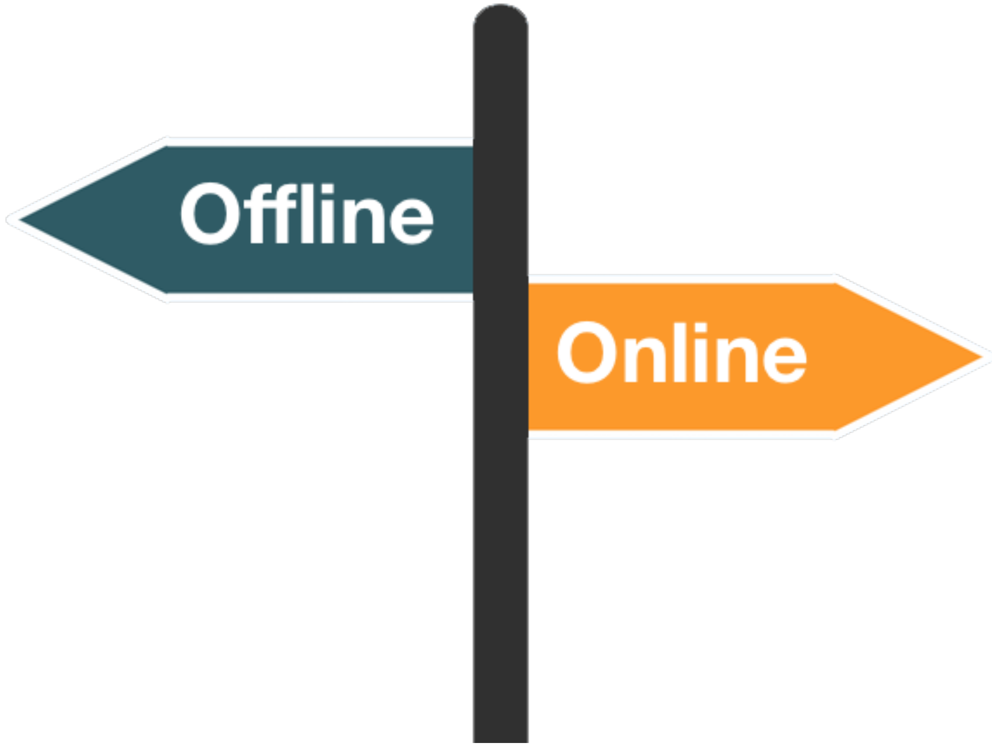
$$Regret\,(T) \leq \widetilde{\mathcal{O}}\left((1 + f_{B_1})(d - L)\sqrt{AT}\right).$$

- As usual, some assumptions about "unobservables" must be made
- Here:
  access to knowledge of behavorial policy →
  partially observed offline data can help make online learning faster

# Talk outline

How to act online with the help of offline data?

- **Part I: Contextual bandits with confounded offline data**
- Part II: Online imitation and reinforcemnt learning with offline data from a possibly different distribution

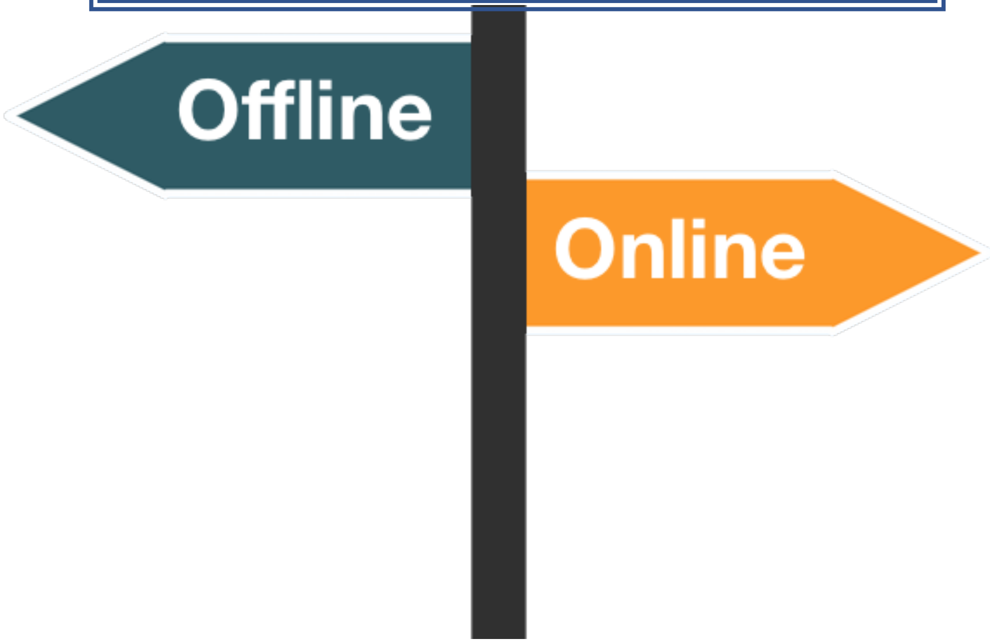


# Talk outline

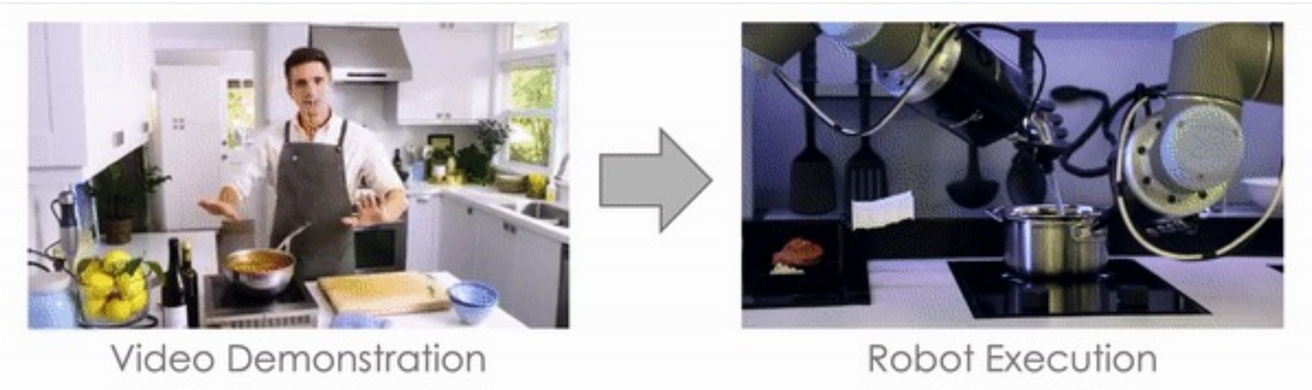
How to act online with the help of offline data?

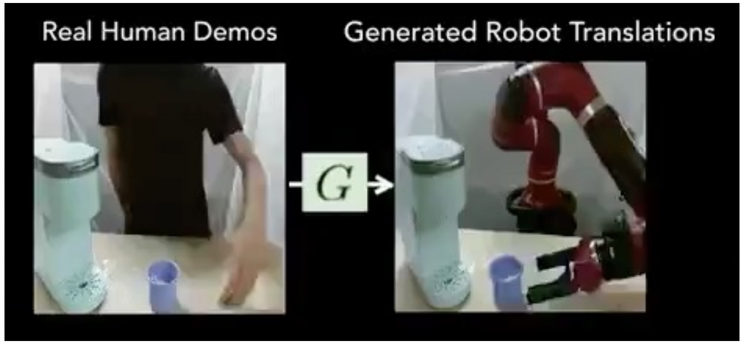
- Part I: Contextual bandits with confounded offline data
- **Part II: Online imitation and reinforcemnt learning with offline data from a possibly different distribution**

"On Covariate Shift of Latent Confounders in Imitation and Reinforcement Learning", Tennenholtz, Hallak, Dalal, Mannor, Chechik, S
ICLR 2022



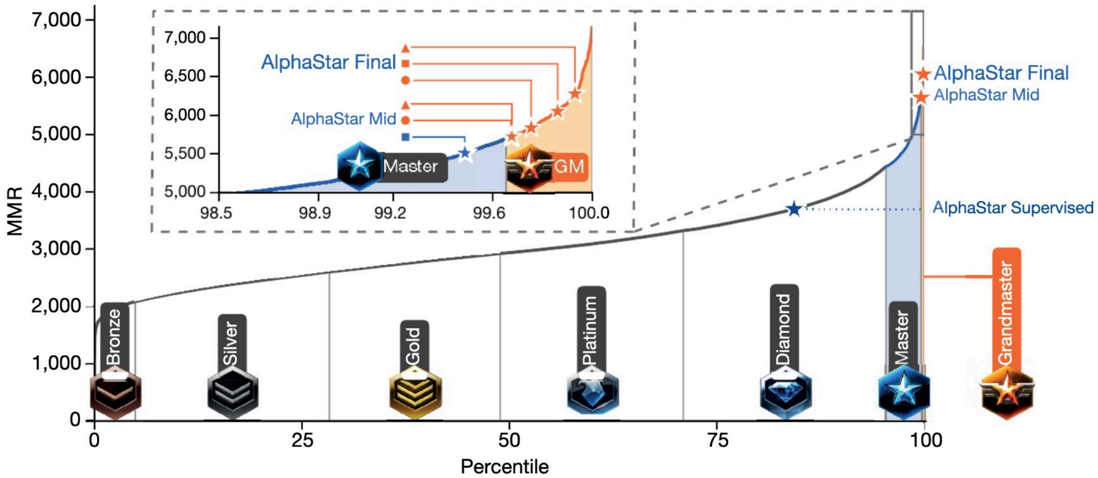
# Imitation Learning Background



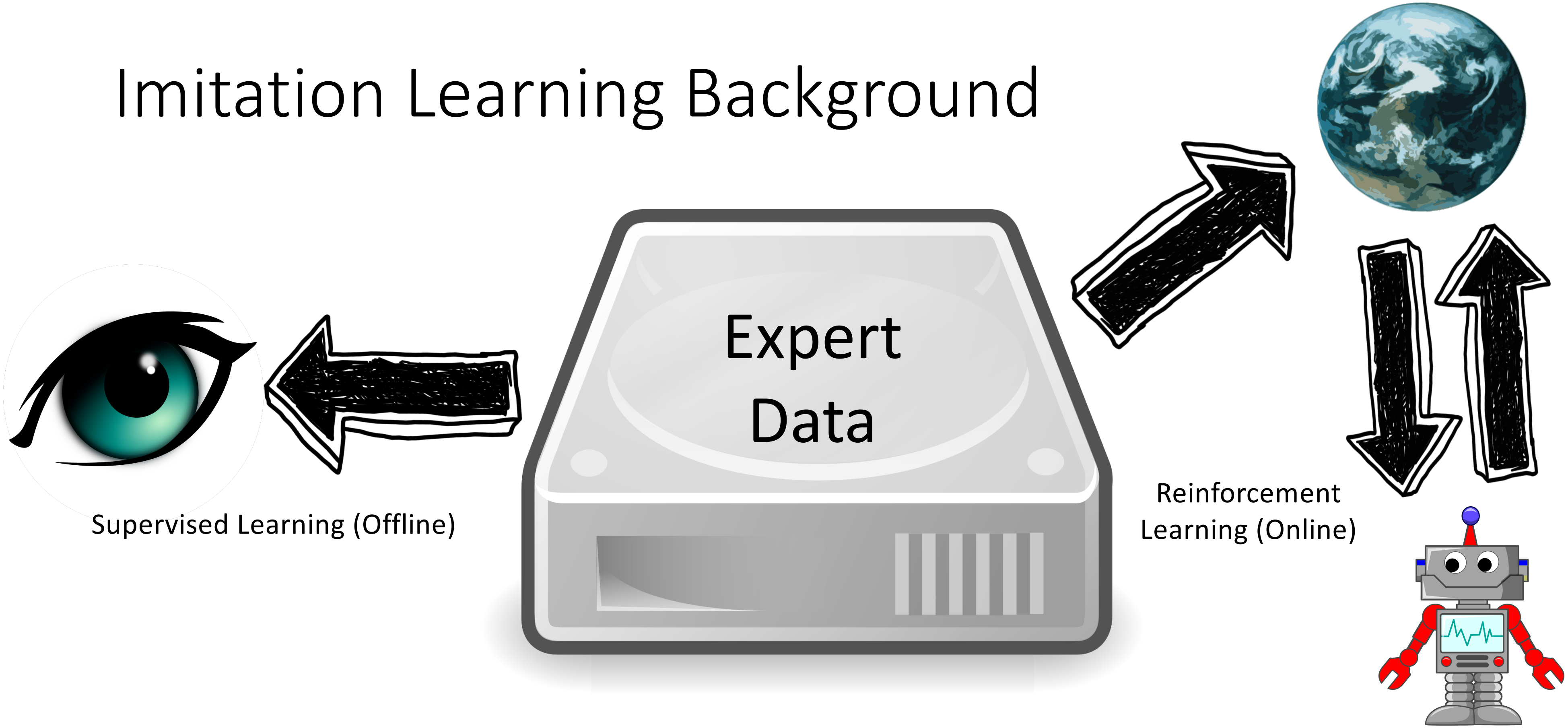




# Imitation Learning Background

Pure imitation achieved state of the art performance in StarCraft 2 and reached 70% of final alpha-star performance ("Diamond league")




# Imitation Learning Background




Behavior Cloning (Michie, Bain, & Hayes-Michie, 1990)
Offline RL (2005-today)

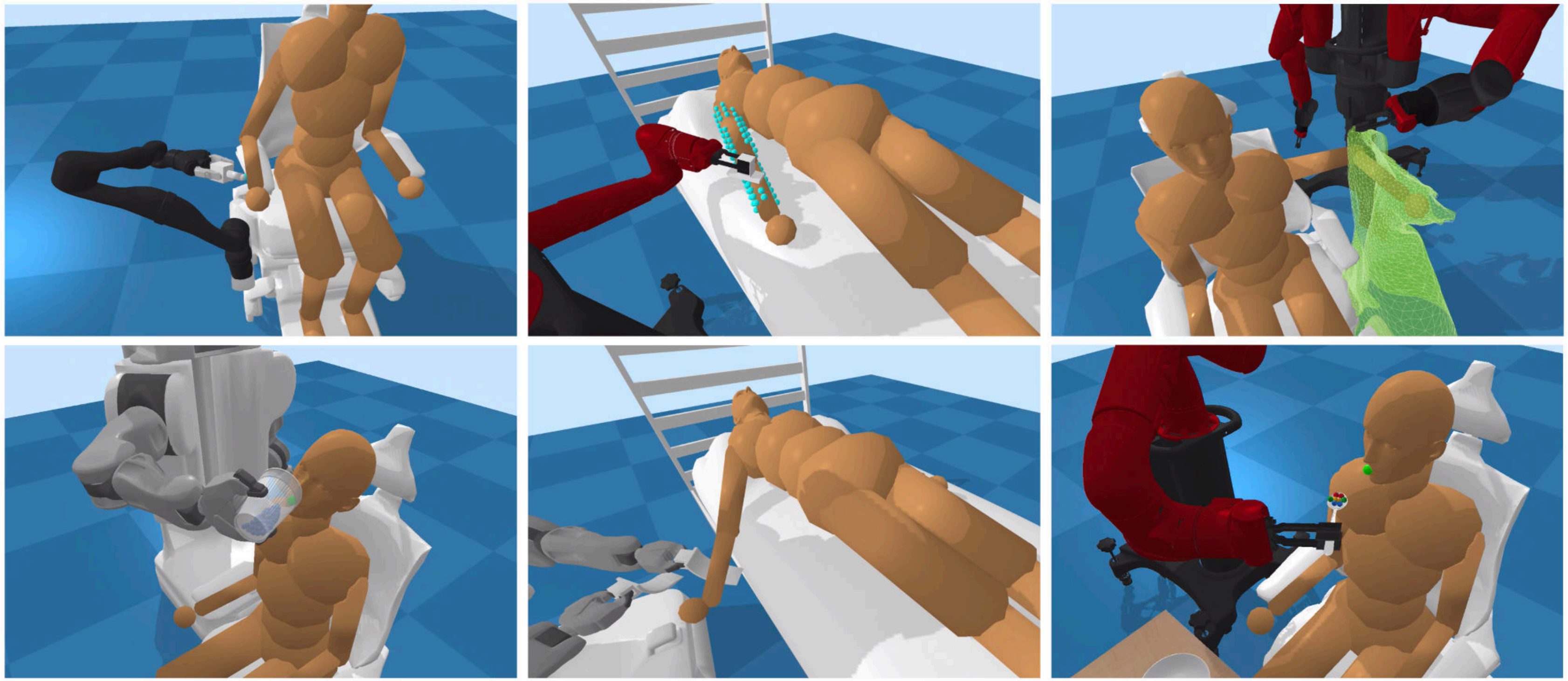
Ho & Ermon (2016), Fu et al. (2017),
Kostrikov et al. (2019), Brantley et al. (2019),

# Imitation Learning + Partial Observability

Some information was not collected in the expert dataset

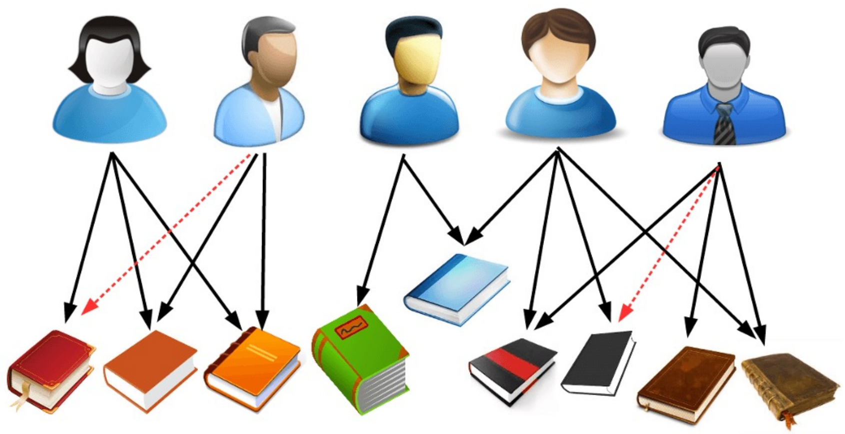
# Imitation Learning + Partial Observability

Privacy constrains (e.g., medical)



# Imitation Learning + Partial Observability

Information added with new releases
of product, e.g., recommender systems



# Setup

**Online Simulator of a Contextual MDP**

$\mathcal{X}$ – context space

$\mathcal{S}$ – state space

$\mathcal{A}$ – state space

$\rho_0(x)$ – initial context distribution

$\nu(s_0|x)$ – initial state distribution

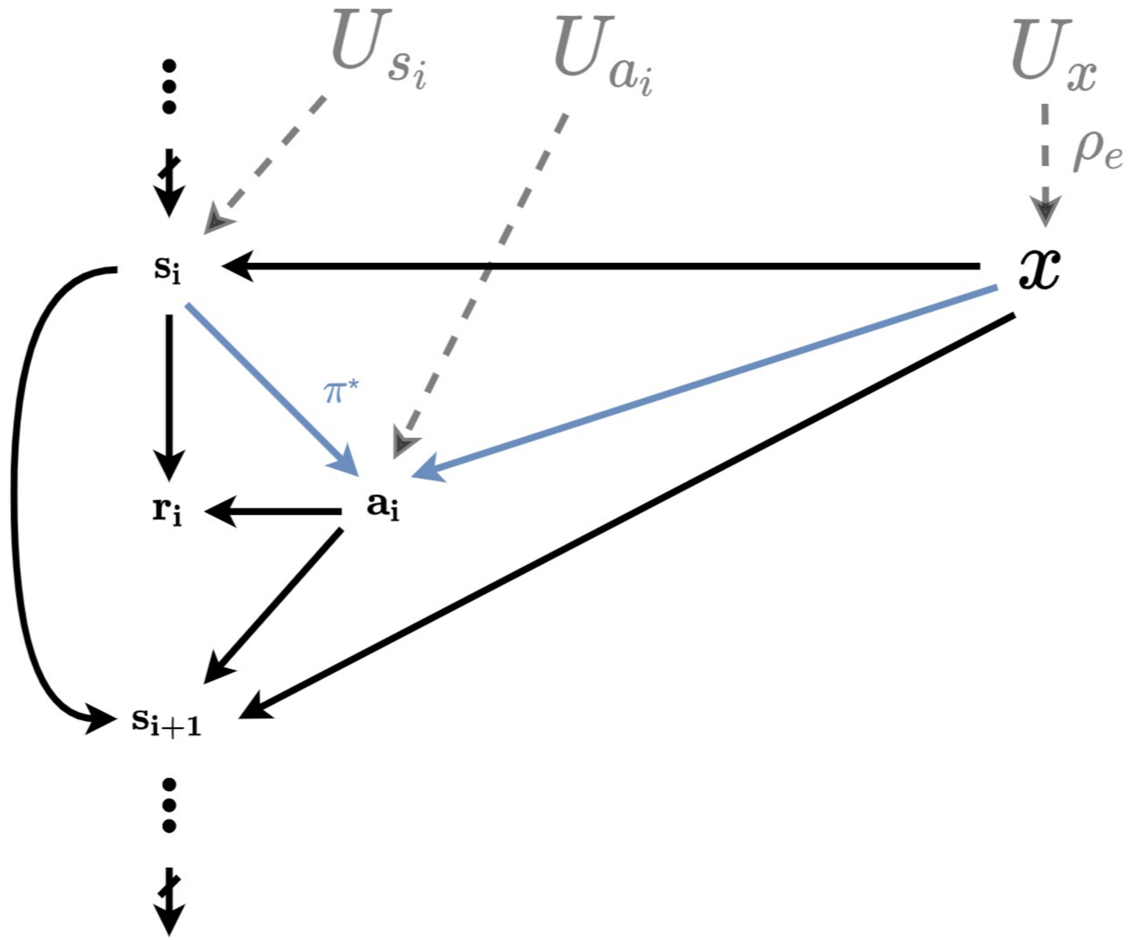
$P(s'|s,a,x)$ – transition function

$r(s,a,x)$ – reward function

$\gamma$ – discount factor

$\pi\left(s,x\right)$ – policy

Contextual MDP causal graph



# Imitation learning with partial observability

- As usual in imitation learning, we don’t see the expert’s reward
- We assume the expert performs the optimal policy $\pi^*(s, x)$
- However, we don’t see the context $x$ the expert saw, only the state and actions
- Further, we might have $\rho_e(x) \neq \rho_o(x)$, i.e. covariate shift between the expert setup and the online setup

# Setup

$$v(\pi) = \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t, x) | x \sim \rho_0, s_0 \sim \nu(\cdot \mid x), a_t \sim \pi(s_t, x_t))\right]$$

**Optimal Policy**

$$v^* = \max_{\pi} v(\pi), \qquad \pi^* \in \arg\max_{\pi} v(\pi).$$

$$\Pi^*_{\mathcal{M}} = \arg\max_{\pi} v_{\mathcal{M}}(\pi)$$

# Setup

**Expert Data**

Assume expert data of a policy $\pi^*$

$$\{(\cancel{x^i}, s_0^i, a_0^i, s_1^i, a_1^i, \ldots, s_H^i, a_H^i)\}_{i=1}^n$$

$$\{(s_0^i, a_0^i, s_1^i, a_1^i, \ldots, s_H^i, a_H^i)\}_{i=1}^n$$

$$P^*(s_0, a_0, s_1, a_1, \ldots, s_H, a_H) = \sum_x \rho_e(x)\nu(s_0|x)\left(\prod_{i=0}^{H-1} P(s_{i+1}|s_i, a_i, x)\right)\left(\prod_{i=0}^{H} \pi^*(a_i|s_i, x)\right)$$

# State-Action Frequency Distribution

- The state-action frequency distribution of policy $\pi$ given context $x$ is

$$d^\pi(s, a|x) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^\pi(s_t = s, a_t = a|x, s_0 \sim \nu(\cdot|x))$$

- The mean-context state-action frequency distribution is given by

$$d^\pi_{\rho_o}(s, a) = \mathbb{E}_{x\sim\rho_o}[d^\pi(s, a \mid x)] \quad \text{(environment)},$$
$$d^\pi_{\rho_e}(s, a) = \mathbb{E}_{x\sim\rho_e}[d^\pi(s, a \mid x)] \quad \text{(expert data)}.$$

# No Covariate Shift ( $\rho_o(x) = \rho_e(x)$ )

**Definition 1** (Ambiguity Set). *For a policy* $\pi \in \Pi$, *we define the set of all deterministic policies that match the context-free stationary distributions of* $\pi$ *by*

$$\Upsilon_\pi = \left\{ \pi' \in \Pi_{det} : d_{\rho_o}^{\pi'}(s,a) = d_{\rho_e}^{\pi}(s,a), s \in \mathcal{S}, a \in \mathcal{A} \right\}.$$

$$\left( \begin{array}{ll} d_{\rho_o}^{\pi}(s,a) = \mathbb{E}_{x \sim \rho_o}[d^\pi(s,a \mid x)] & \text{(environment)}, \\ d_{\rho_e}^{\pi}(s,a) = \mathbb{E}_{x \sim \rho_e}[d^\pi(s,a \mid x)] & \text{(expert data)}. \end{array} \right)$$

# No Covariate Shift ( $\rho_o(x) = \rho_e(x)$ )

**Theorem 1.** *[Sufficiency of $\Upsilon_{\pi^*}$] Assume $\rho_e \equiv \rho_o$. Let $\pi^* \in \Pi^*_{\mathcal{M}}$ and let $\pi_0 \in \Upsilon_{\pi^*}$. Then, $\Upsilon_{\pi^*} = \Upsilon_{\pi_0}$ and, if $\pi_0 \neq \pi^*$, there exists $r_0$ such that $\pi_0 \in \Pi^*_{\mathcal{M}_0}$ but $\pi^* \notin \Pi^*_{\mathcal{M}_0}$, where $\mathcal{M}_0 = (\mathcal{S}, \mathcal{A}, \mathcal{X}, P, r_0, \rho_o, \nu, \gamma)$.*

**In Layman's Terms:**

**Any policy in $\Upsilon_{\pi^*}$ is a candidate optimal policy, and none of them can be ruled out using state-action frequency distributions. Some might be suboptimal.**

**Algorithm 1** Confounded Imitation

**input:** Expert data with missing context $\mathcal{D}^*$ ($d^{\pi^*}_{\rho_e}$), $\lambda > 0$.
**init:** $\Upsilon = \emptyset$
**for** $n = 1, \dots$ **do**
  $L^*(\pi; g_0) := \mathbb{E}_{s,a \sim d^{\pi}_{\rho_o}(s,a)}[g_0(s,a)] - \mathbb{E}_{s,a \sim d^{\pi^*}_{\rho_e}(s,a)}[g_0(s,a)]$
  $L_i(\pi; g_i) := \mathbb{E}_{x \sim \rho_o, s,a \sim d^{\pi}(s,a|x)}[g_i(s,a,x)] - \mathbb{E}_{x \sim \rho_o, s,a \sim d^{\pi_i}(s,a|x)}[g_i(s,a,x)] \quad , i \geq 1$
  Compute $\pi_n$ by solving

$$\min_{\pi \in \Pi_{\text{det}}} \max_{|g_0| \leq \frac{1}{2}, |g_i| \leq \frac{1}{2}} \left\{ L^*(\pi; g_0(s,a)) - \lambda \min_i L_i(\pi; g_i(s,a,x)) \right\}$$

  **if** $\pi_n \in \Upsilon$ **then**
    Terminate and return $\bar{\pi}(a|s,x) = \frac{\sum_{i=1}^{n-1} d^{\pi_i}(s,a,x)}{\sum_{i=1}^{n-1} \sum_{a'} d^{\pi_i}(s,a',x)}$
  **else**
    $\Upsilon = \Upsilon \cup \{\pi_n\}$
  **end if**
**end for**

# With Covariate Shift ( $\rho_o(x) \neq \rho_e(x)$ )

- Result 1: Context Free Transition → Impossibility of Imitation

**Theorem 2.** *[Catastrophic Imitation] Assume* $|\mathcal{X}| \geq |\mathcal{A}|$ *and* $P(s'|s,a,x) = P(s'|s,a,x')$ *for all* $x, x' \in \mathcal{X}$. *Then there exist* $\pi_{e,1}, \pi_{e,2}$ *s.t.* $\{\pi_{e,1}, \pi_{e,2}\}$ *are non-identifiable, catastrophic expert policies.*

In Layman's Terms:

**If the transition is independent of the context, then the worst-case policy cannot be ruled out.**

**(observed states and actions act as proxies for context)**

# With Covariate Shift ( $\rho_o(x) \neq \rho_e(x)$)

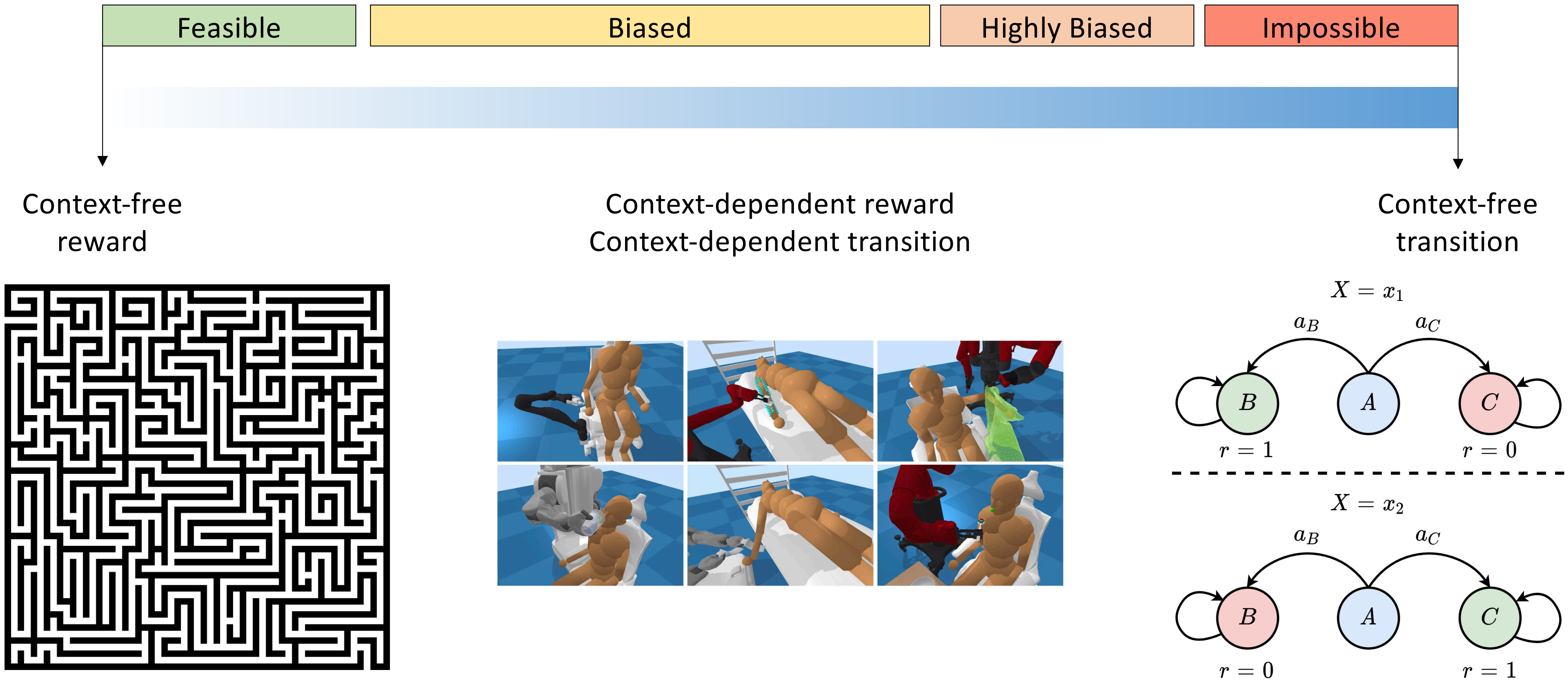
- Result 2: Context Free Rewards → Possibility of Imitation

**Theorem 3.** *[Sufficiency of Context-Free Reward] Assume* $r(s,a,x) = r(s,a,x')$ *for all* $x, x' \in \mathcal{X}$. *Then* $\Upsilon_{\pi^*} \subseteq \Pi^*_{\mathcal{M}}$.

**In Layman's Terms:**

**If the reward is independent of the context, then standard imitation techniques suffice (even if the transition function depends on the context).**

# Hardness of Confounded Imitation



# Expert Data as Side Information

- Now we further assume we have access to the true reward signal (online)
- First try:

$$\max_{\pi \in \Pi} \mathbb{E}_{x \sim \rho_o, s,a \sim d^{\pi}(s,a|x)}[r(s,a,x)] - \lambda D_f(d^{\pi}_{\rho_o}(s,a) || d^{\pi^*}_{\rho_e}(s,a))$$

- This is biased + we don't know $\rho_e$
- We show a more involved optimization problem is unbiased

$$\max_{\pi \in \Pi} \min_{\substack{g: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R} \\ \rho_s}} \mathbb{E}_{x \sim \rho_o, s,a \sim d^{\pi}(s,a|x)}[r(s,a,x) + \lambda g(s,a)] - \lambda \mathbb{E}_{s,a \sim d^{\pi^*}_{\rho_s}(s,a)}[f^*(g(s,a))].$$

$D_f$ is an $f$-divergence (e.g. KL-divergence, TV-distance, $\chi^2$-divergence)

- We propose:
  1. A provably convergent but slow algorithm based on Follow The Leader
  2. A more efficient gradient-based method over the non-convex objective

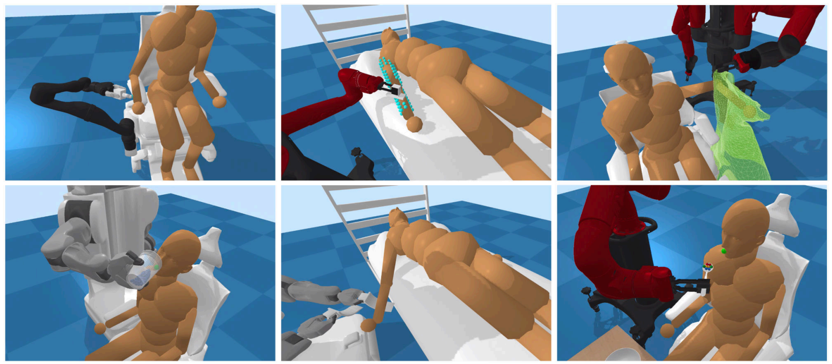
# Corrective Trajectory Sampling (CTS)

**Algorithm 3** Reinforcement Learning using Confounded Expert Data (Online Gradient Descent)

**input:** Expert data with missing context, $\lambda, B, N > 0$, policy optimization algorithm ALG-RL.
**init:** Policy $\pi^0$, bonus network $g_\theta$
**for** $k = 1, \dots$ **do**
    $\rho_s \leftarrow \arg\min_\rho D_f(d_{\rho_o}^{\pi_{k-1}}(s,a) || d_\rho^{\pi^*}(s,a))$.
    **for** $e = 1, \dots N$ **do**
        Sample batch $\{s_i, a_i\}_{i=1}^B \sim d_{\rho_o}^{\pi_{k-1}}(s,a)$.
        Sample batch $\{s_i^e, a_i^e\}_{i=1}^B \sim d_{\rho_s}^{\pi^*}(s,a)$.
        Update $g_\theta$ according to $\nabla_\theta L(\theta) = \frac{1}{B}\sum_{i=1}^B \nabla_\theta[g_\theta(s_i,a_i) - f^*(g_\theta(s_i^e,a_i^e))]$.
    **end for**
    $\pi^k \leftarrow \text{ALG-RL}(r(s,a,x) - \lambda g_\theta(s,a))$.
**end for**

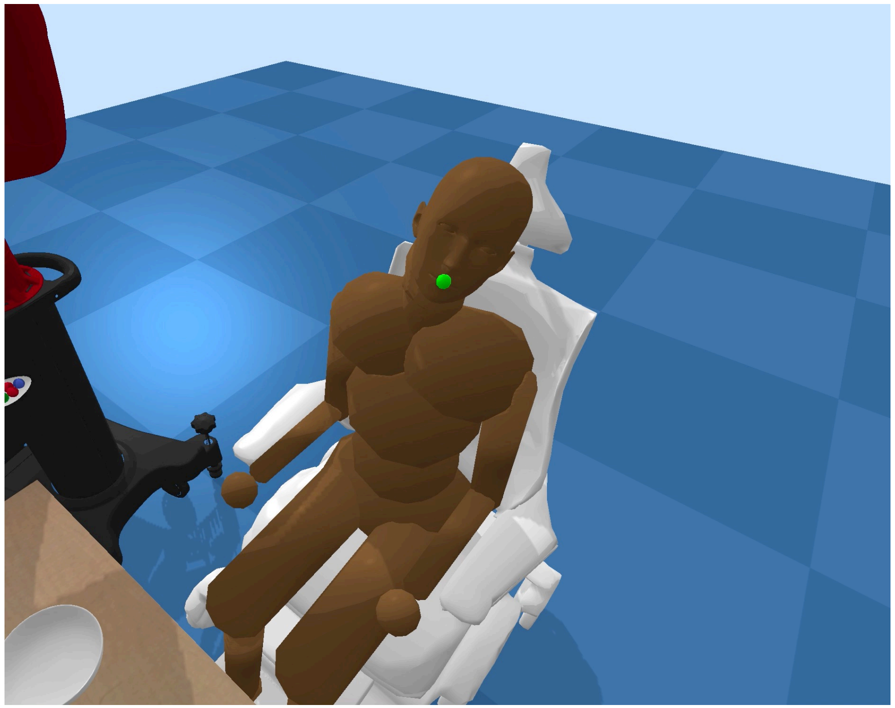
# Assistive-Gym Experiments

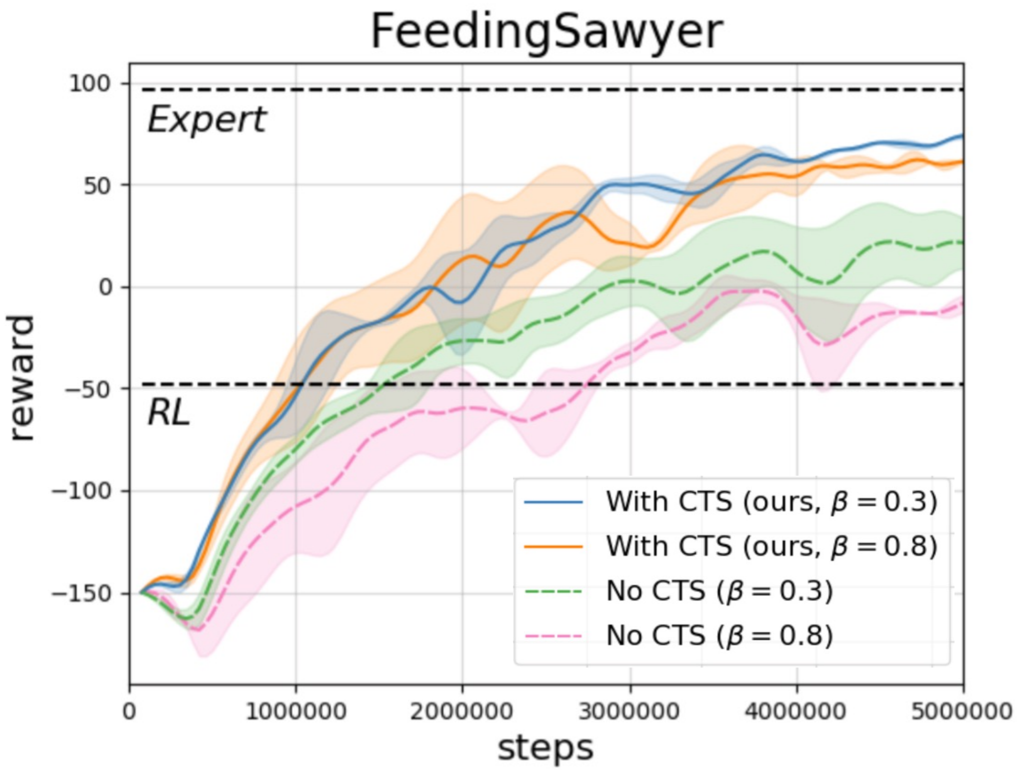
- Assistive autonomous robots as versatile caregivers
  - Assistive-Gym environment [Erickson et al. 2020]
- Tasks include: Feeding, Dressing, Bathing, Drinking, etc.
- Context: weight, height, gender, disability (mobility, shaking), preferences
- State: Robot state
- Action space: Joint forces
- Reward: Success in task + specific user preferences
- Expert: trained on dense reward
- Online: sparse reward
- Shifted context distribution sampled w.p. $\beta$



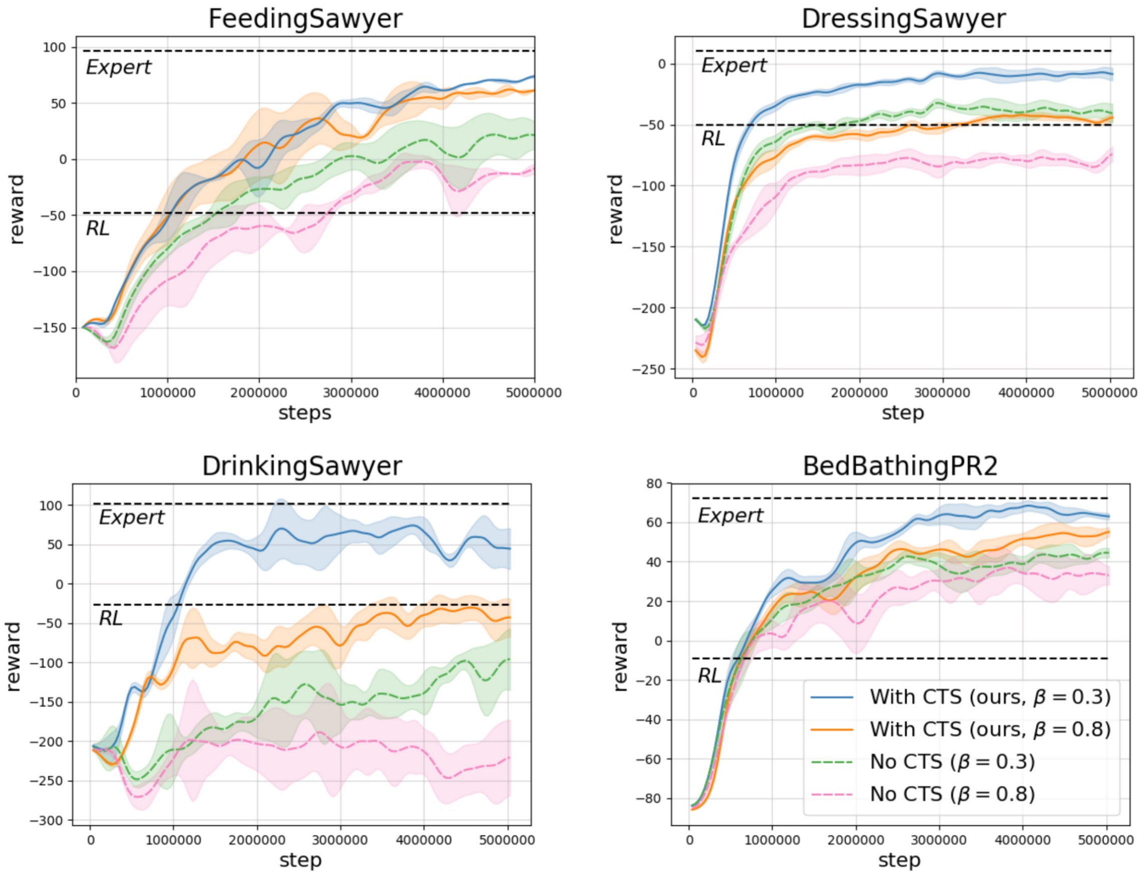
# Experiments (Feeding)

β ∈ [0,1] indicates strength of shift

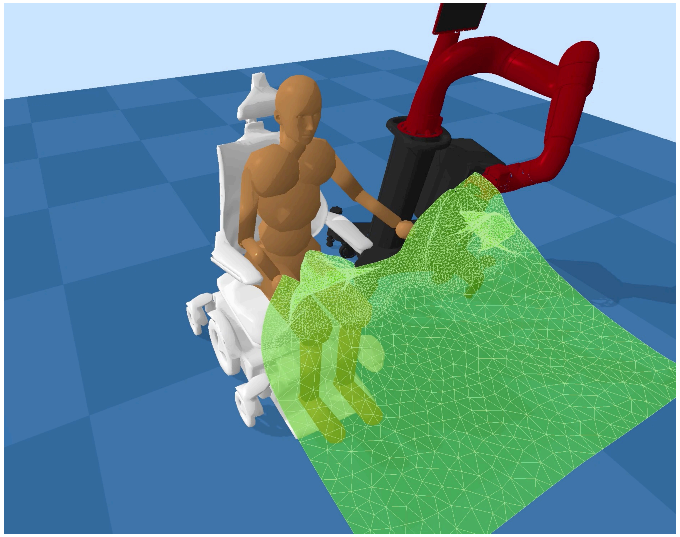




# Experiments



# Experiments (Dressing)

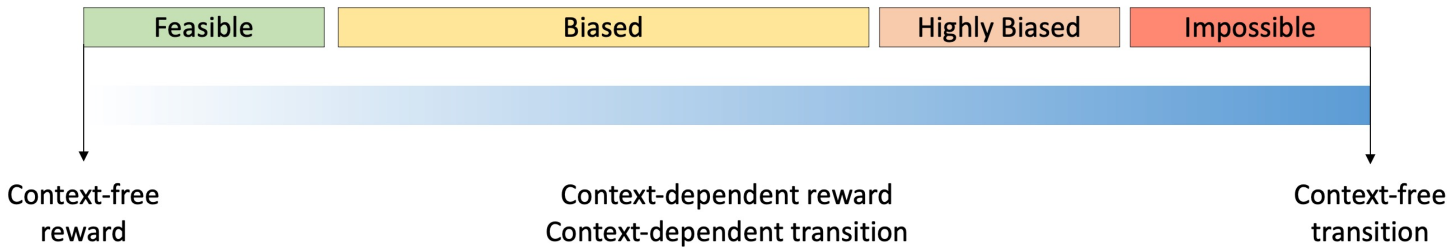


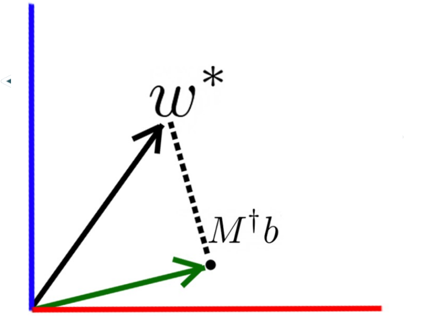
No Sample Correction

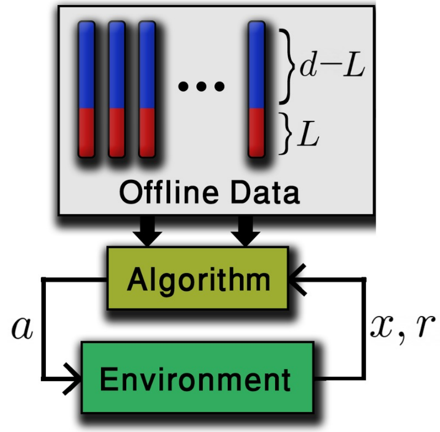
With Sample Correction

# Summary

- Offline data can make online learning more efficient
- Yet offline data often does not match online data
- Map failure modes and necessary conditions for success
- We examined **partial observability and distribution shifts**
- In linear bandits: offline data + sampling from offline policy sometimes allows us to accelare online learning
- In imitation-learning on contextual MDPs: ”it depends"
- In RL with expert data, can empirically acceleareet convergence under distribution shifts




# Thank you

- **Guy Tennenholtz** (Technion)
- Shie Mannor (Technion, NVIDIA)
- Yonatan Efroni (Technion)
- Assaf Hallak (NVIDIA)
- Gal Dalal (NVIDIA)
- Gal Chechik (NVIDIA, Bar-Ilan University)