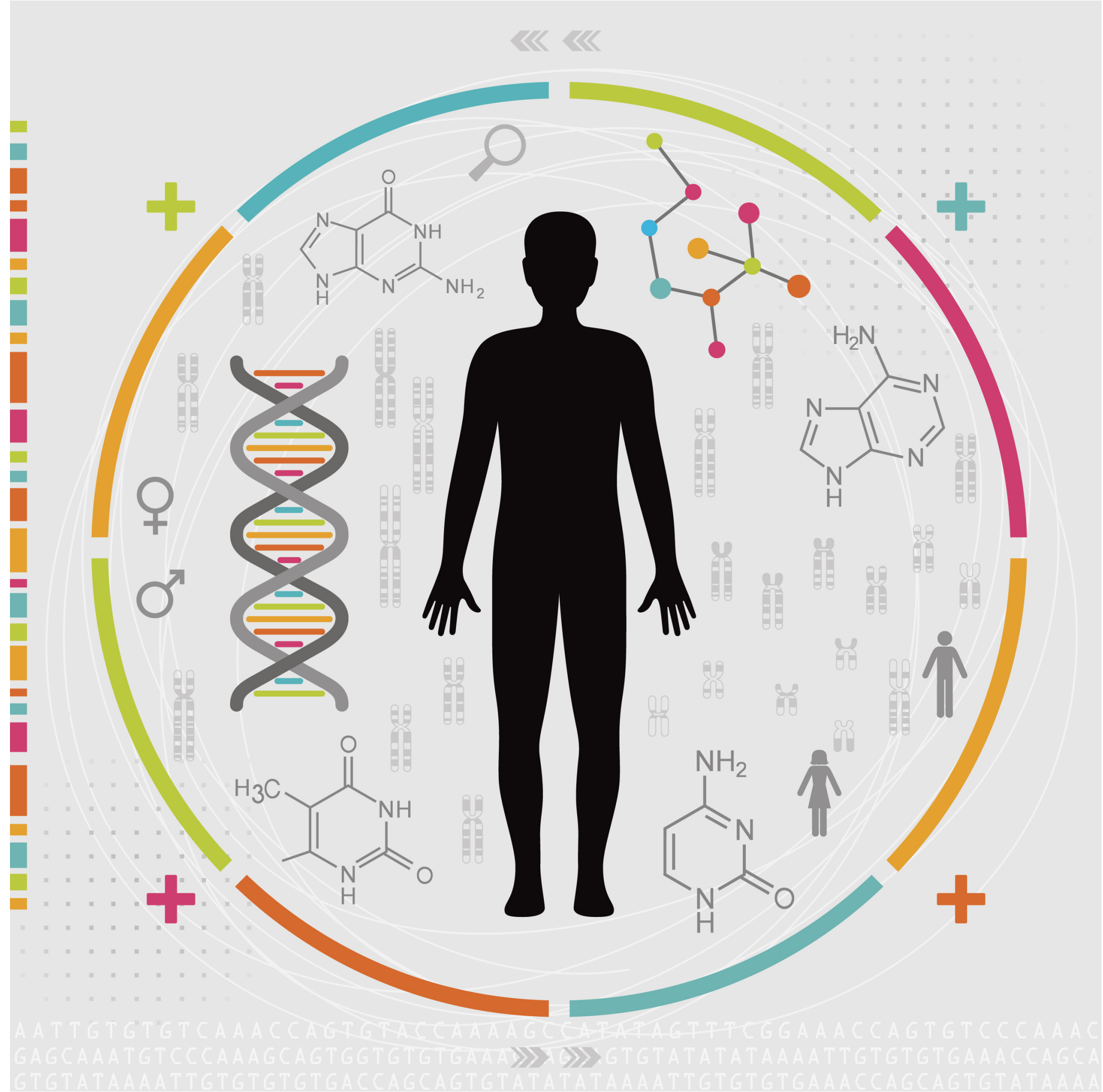
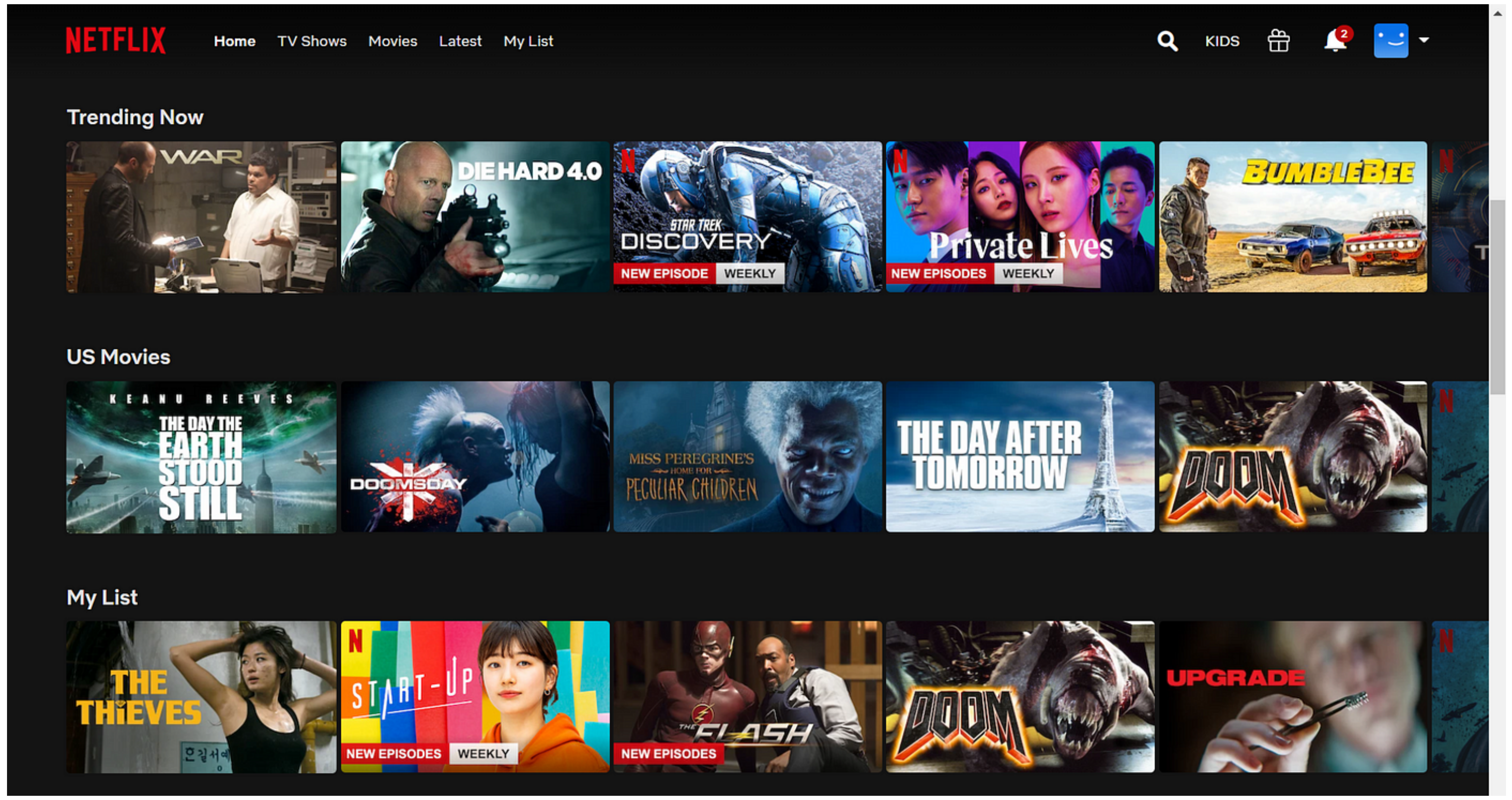
Towards Optimal Differentially Private Regret Bounds in Linear MDPs

Sharan Sahu, Statistics and Data Science, Cornell University

Recent Successes In Reinforcement Learning (RL)



Precision Medicine



Recommender Systems

Autonomous Driving

Contextual Bandits

User

Decision-Making Agent

π

Actions

a_t^1

a_t^2

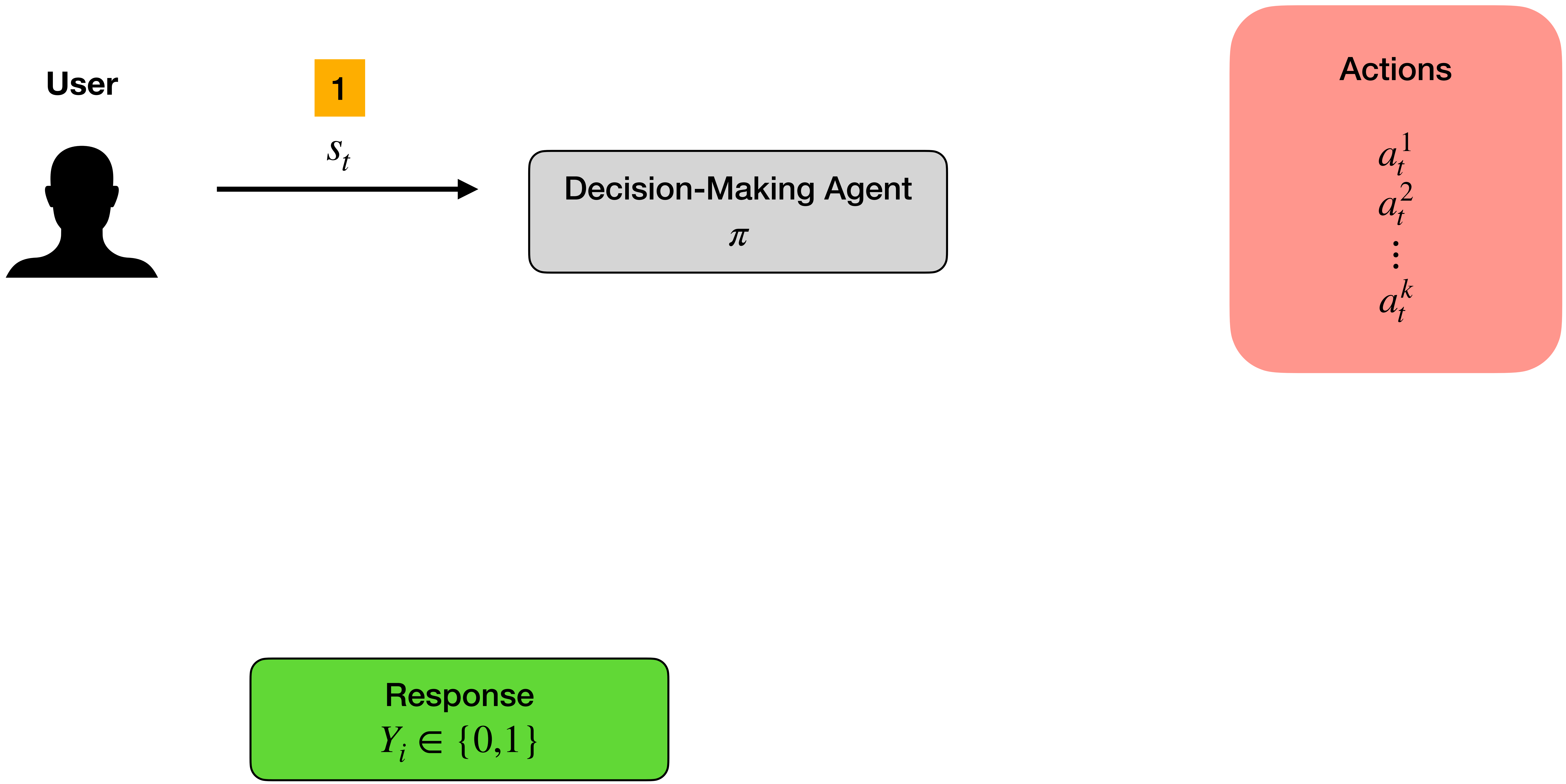
$\vdots$

a_t^k

Response

$Y_i \in \{0,1\}$

Contextual Bandits



User

1

s_t

Decision-Making Agent

π

Actions

a_t^1

a_t^2

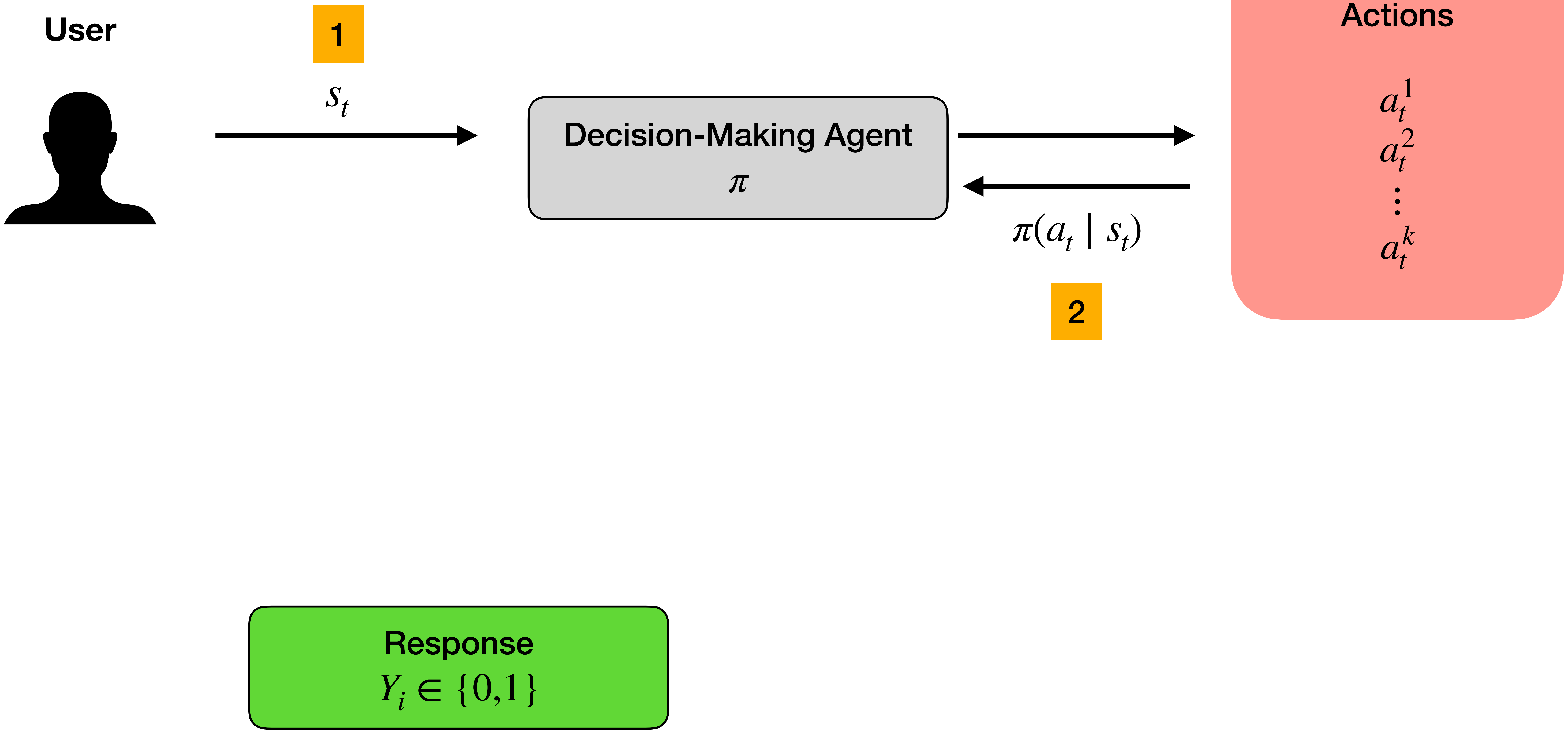
$\vdots$

a_t^k

Response

$Y_i \in \{0,1\}$

Contextual Bandits



User

1

s_t

Decision-Making Agent

π

$\pi(a_t \mid s_t)$

2

Actions

a_t^1

a_t^2

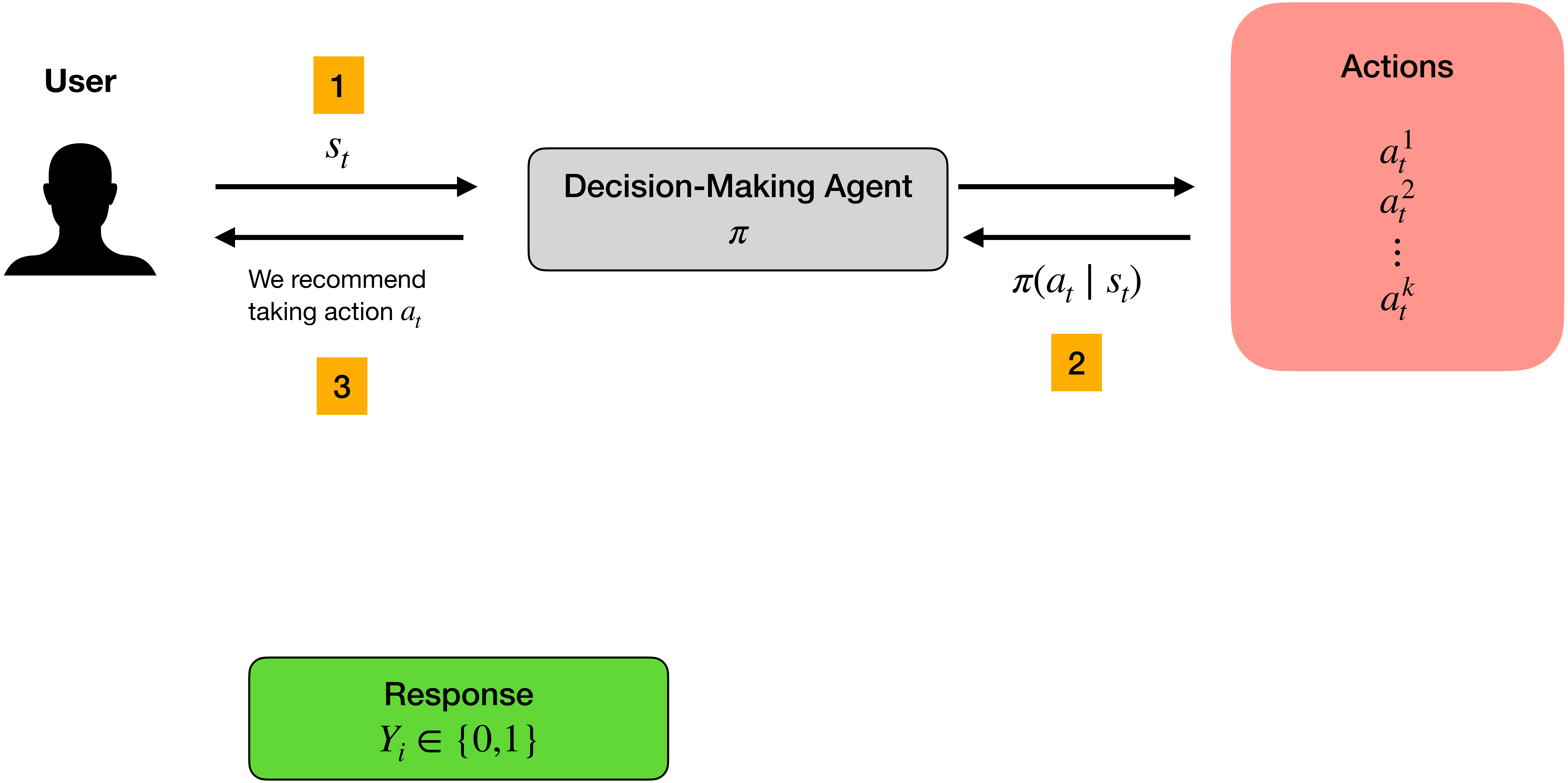
$\vdots$

a_t^k

Response

$Y_i \in \{0,1\}$

Contextual Bandits



User

1

s_t

Decision-Making Agent

π

Actions

a_t^1

a_t^2

$\vdots$

a_t^k

$\pi(a_t \mid s_t)$

2

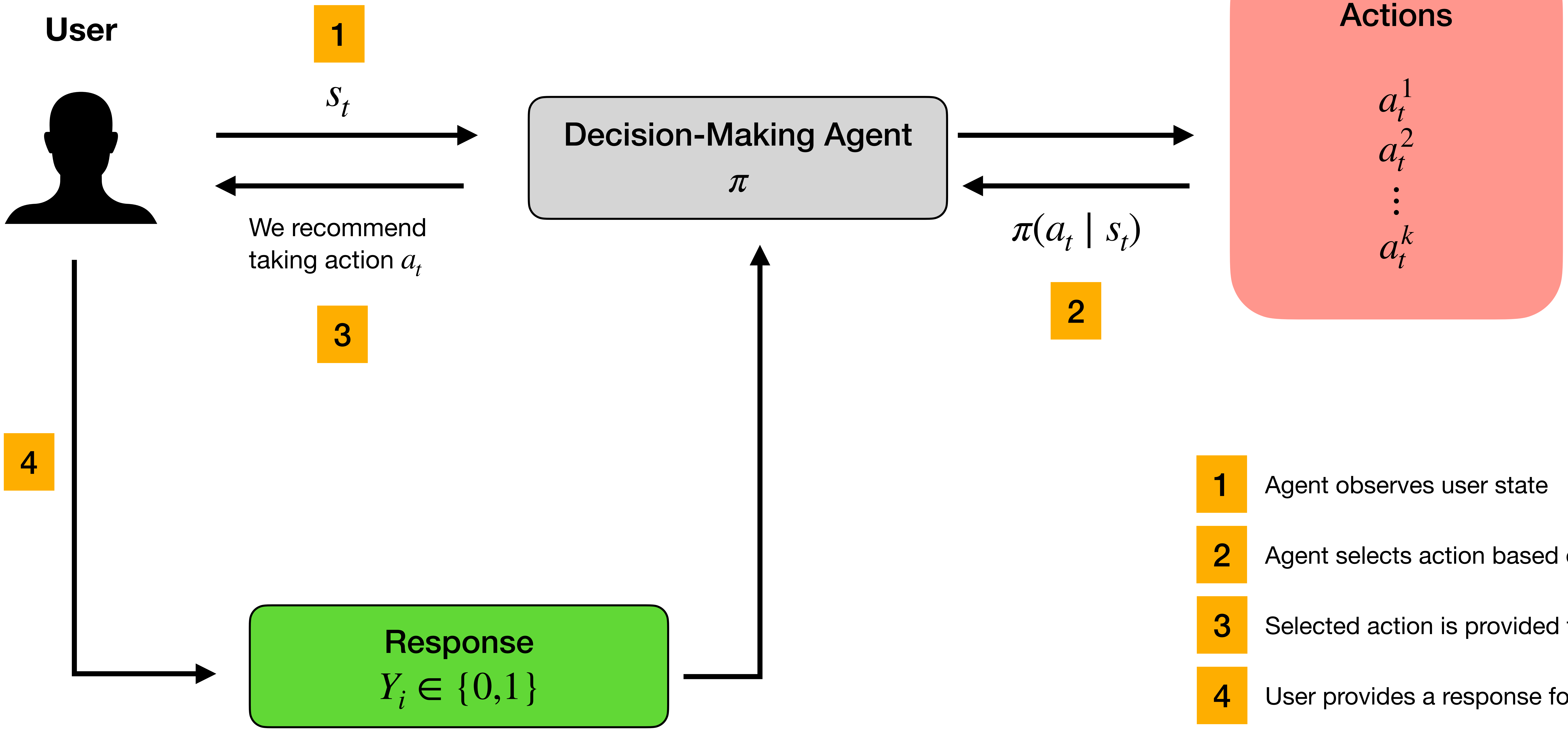
We recommend taking action a_t

3

Response

$Y_i \in \{0,1\}$

Contextual Bandits



User

1

s_t

Decision-Making Agent

π

Actions

a_t^1

a_t^2

$\vdots$

a_t^k

We recommend taking action a_t

3

$\pi(a_t \mid s_t)$

2

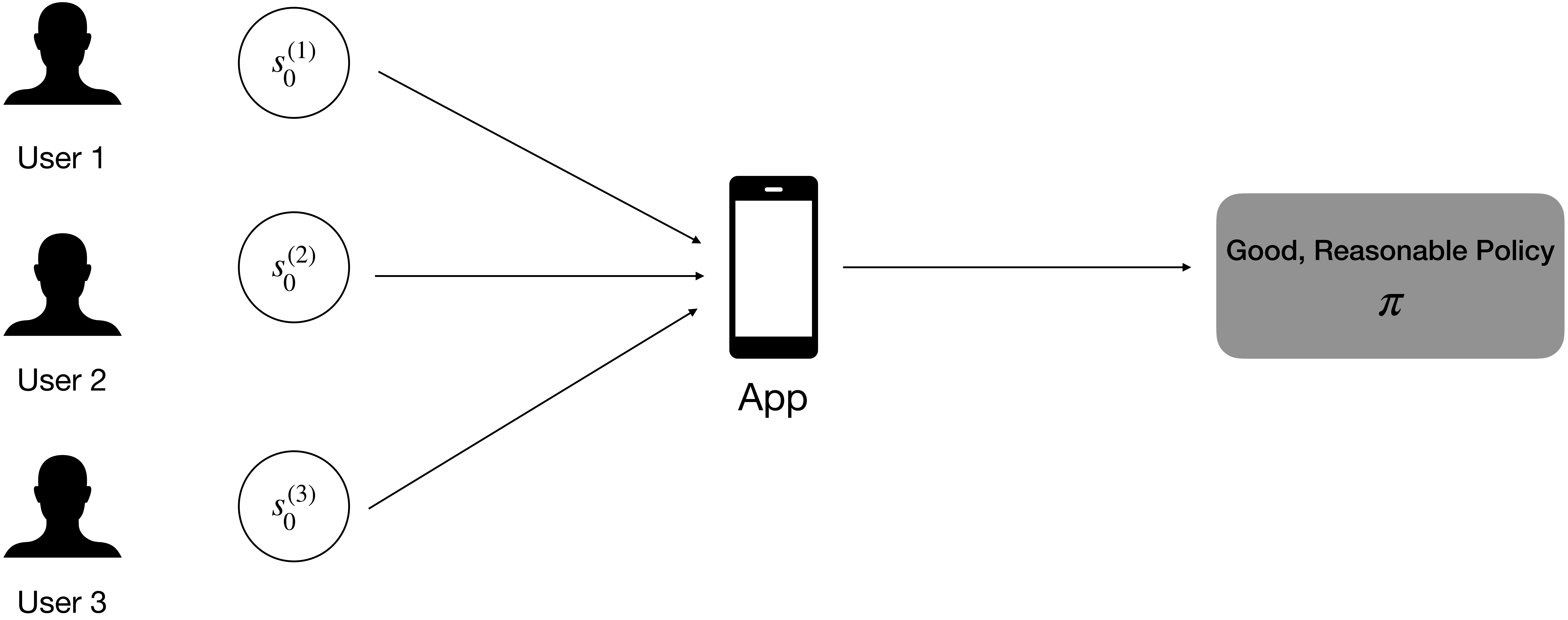
4

Response

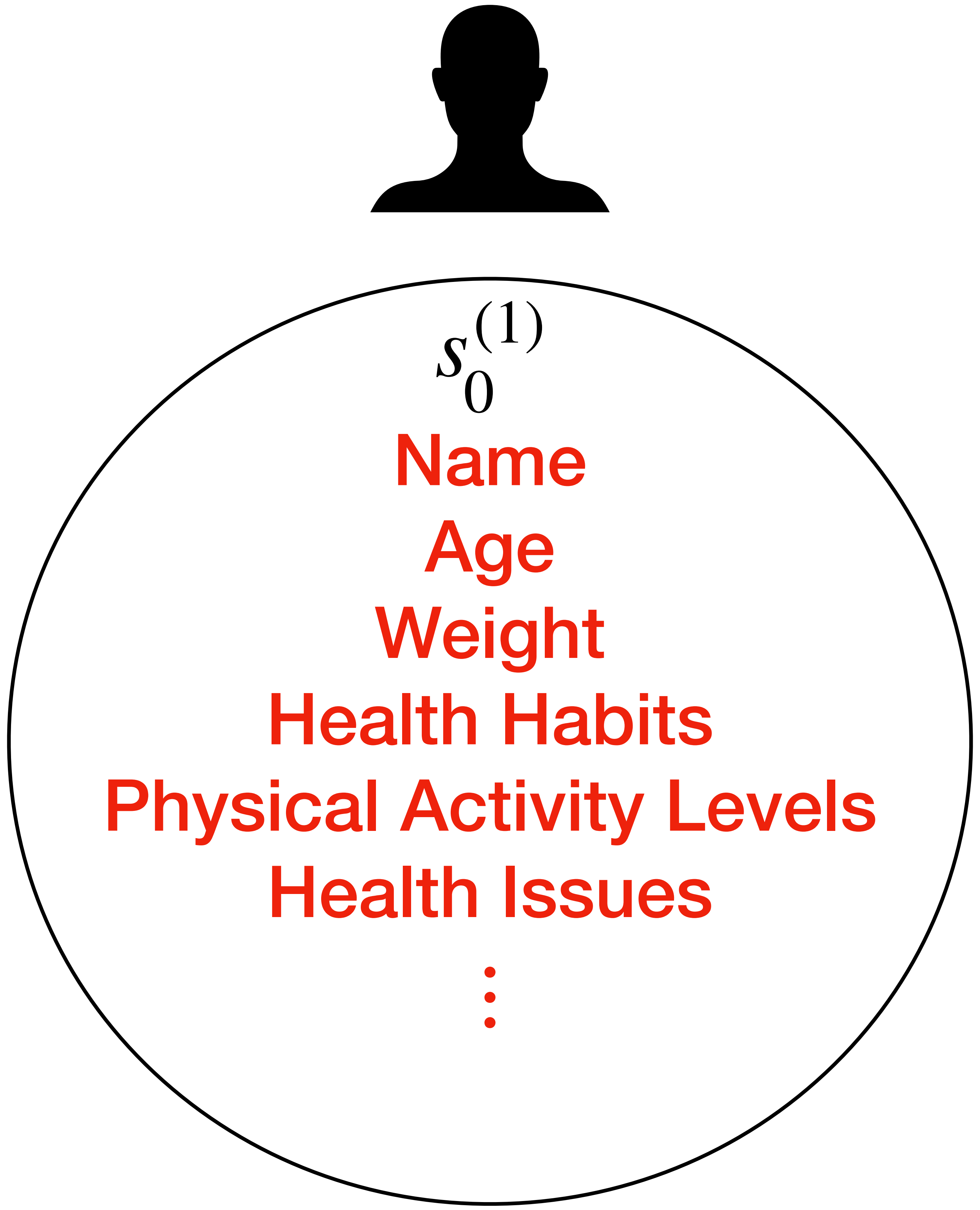
$Y_i \in \{0,1\}$

1	Agent observes user state
2	Agent selects action based on user state
3	Selected action is provided to user
4	User provides a response for agent to learn

Reasonable Policies May Use Sensitive Data



Reasonable Policies May Use Sensitive Data



$s_0^{(1)}$

Name

Age

Weight

Health Habits

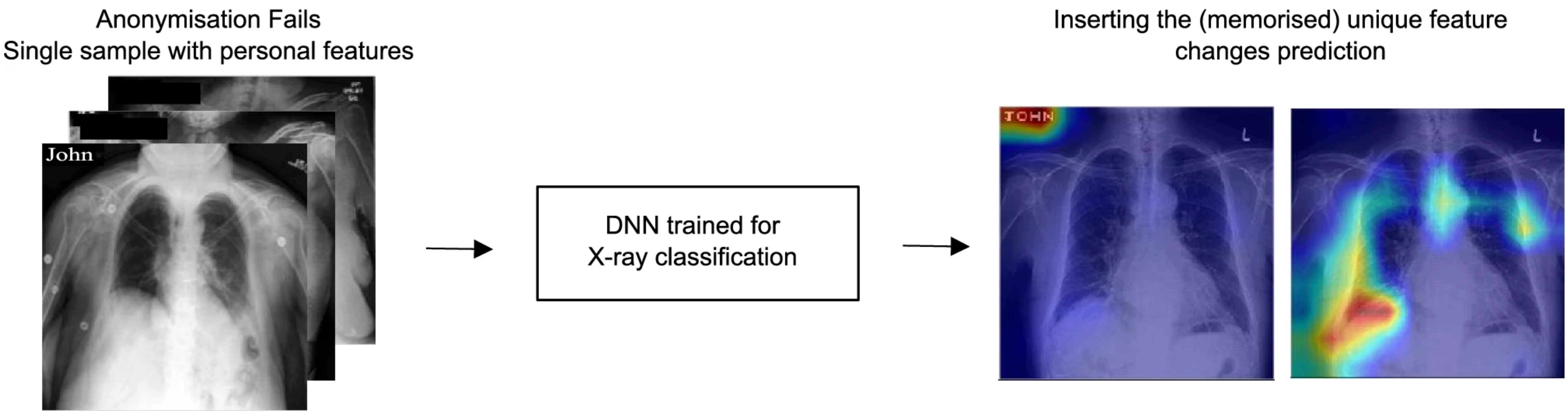
Physical Activity Levels

Health Issues

⋮

The policy has **access to information** that users may consider **sensitive** or **private**

Neural Networks Can Memorize Personal Information From One Example



Anonymisation Fails
Single sample with personal features

[redacted]

[redacted]

DNN trained for
X-ray classification

Inserting the (memorised) unique feature
changes prediction

Hartley et al. 2023

We Must Incorporate Privacy-Preserving Mechanisms Into RL

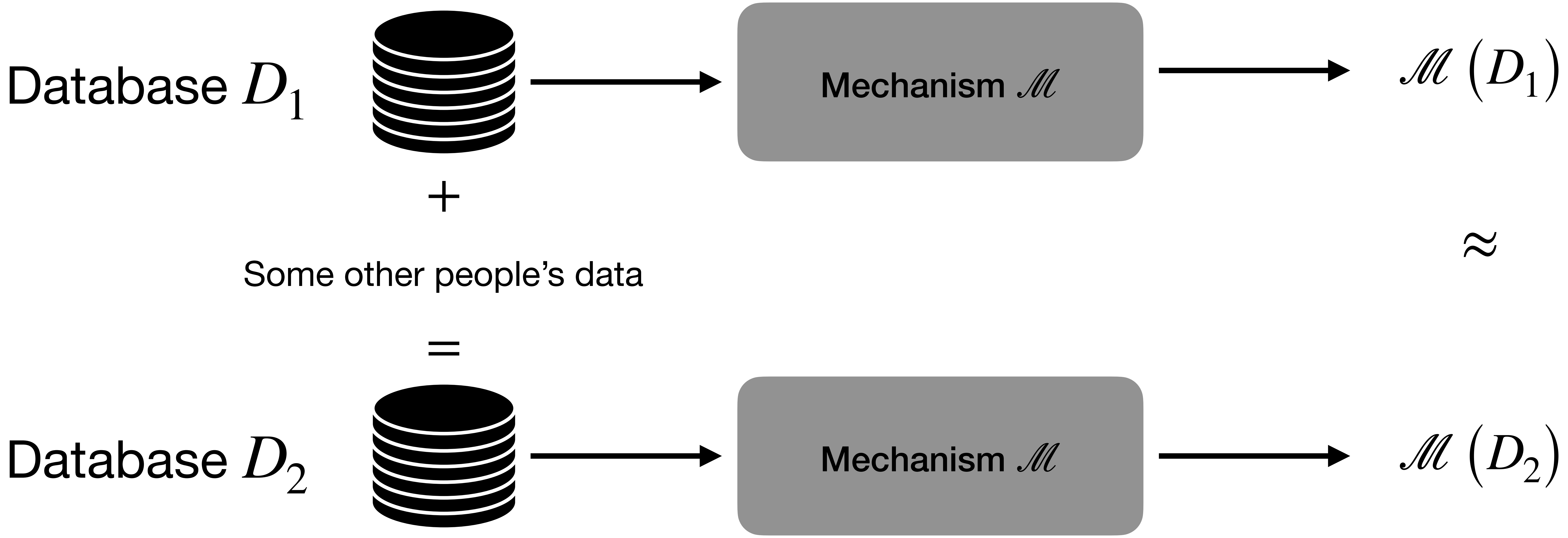
We require a mathematically rigorous framework that provides statistical guarantees for our (possibly randomized) mechanism:

Definition (Approximate Differential Privacy). A mechanism $\mathscr{M}$ is (ε, δ)-DP if for all neighboring datasets $\mathscr{U}, \mathscr{U}'$ that differ by one record and for all event E in the output range

$$\mathbb{P}\left(\mathscr{M}(\mathscr{U}) \in E\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathscr{M}(\mathscr{U}') \in E\right) + \delta$$

Remark: This is a relaxation of ε-DP as in many settings, achieving ε-DP is nearly impossible or comes at high utility cost

Differential Privacy (DP)



Database D_1 → Mechanism $\mathscr{M}$ → $\mathscr{M}\left(D_1\right)$

$+$

Some other people's data

$=$

Database D_2 → Mechanism $\mathscr{M}$ → $\mathscr{M}\left(D_2\right)$

$\approx$

Mechanism $\mathscr{M}$ is differentially private if …

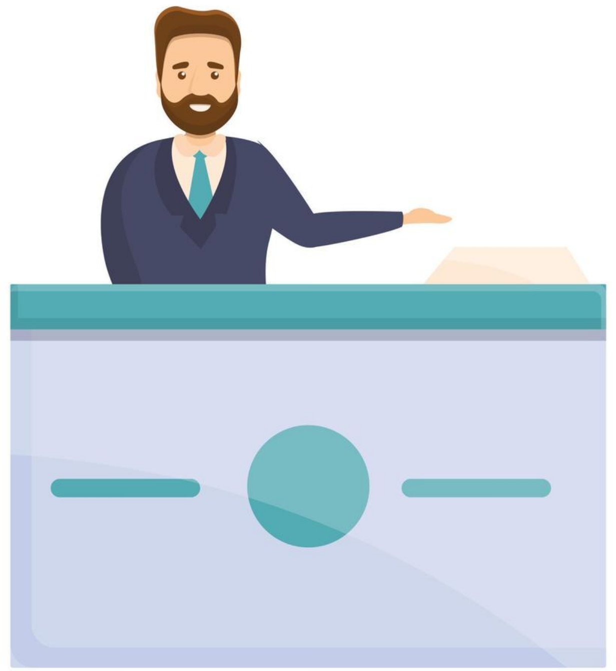
$\forall D_1, D_2$ that differ by at-most one record

$\mathscr{M}\left(D_1\right), \mathscr{M}\left(D_2\right)$ are indistinguishable

There are a few issues with (ε, δ)-DP

Trusted Individual of the Central Agency

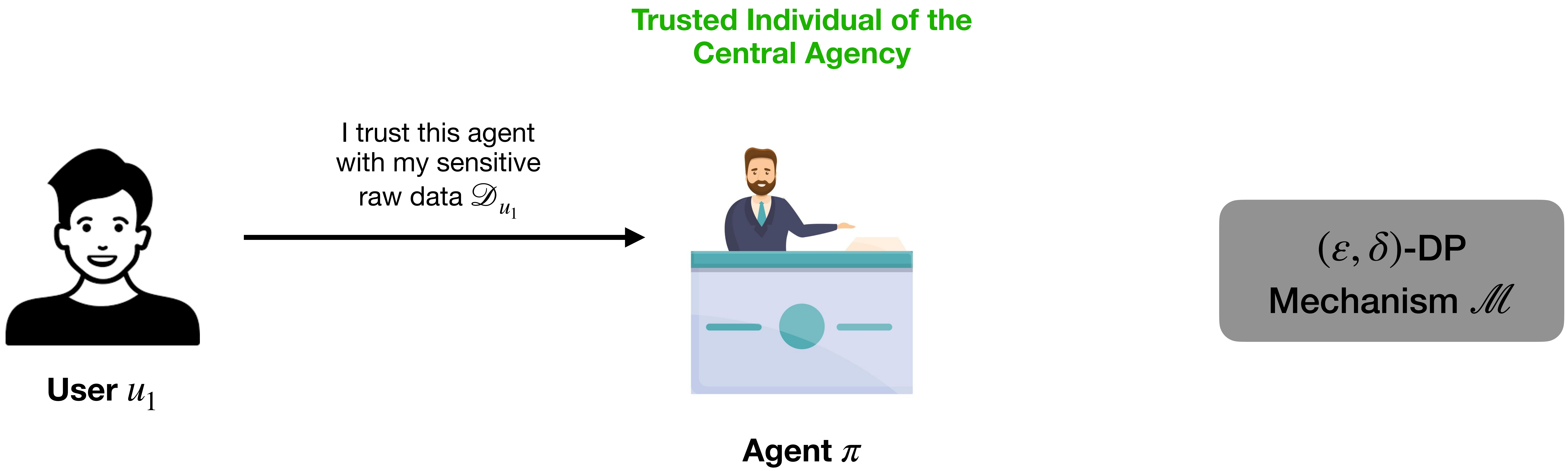
User u_1



Agent π

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

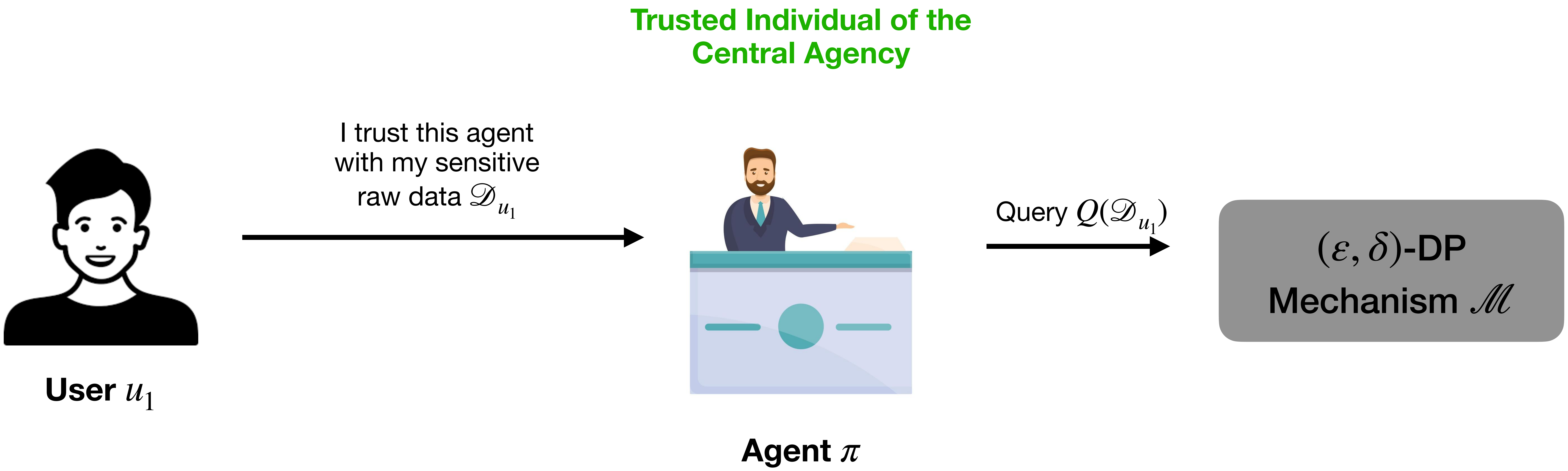
I trust this agent with my sensitive raw data $\mathscr{D}_{u_1}$

User u_1

Agent π

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

I trust this agent with my sensitive raw data $\mathscr{D}_{u_1}$

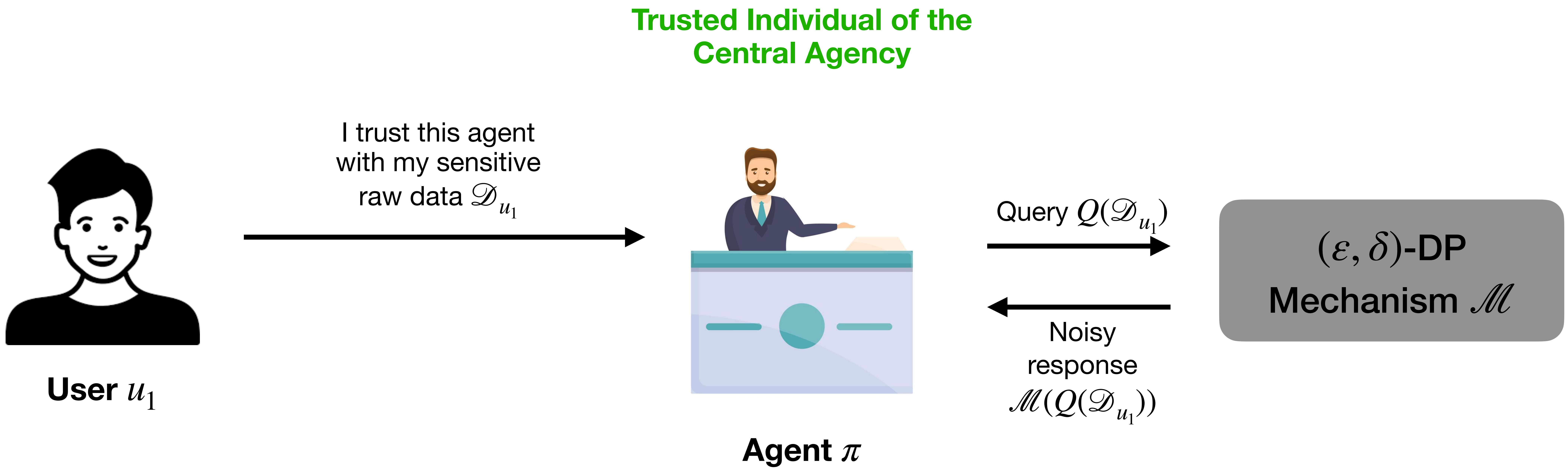
User u_1

Agent π

Query $Q(\mathscr{D}_{u_1})$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

I trust this agent with my sensitive raw data $\mathscr{D}_{u_1}$

User u_1

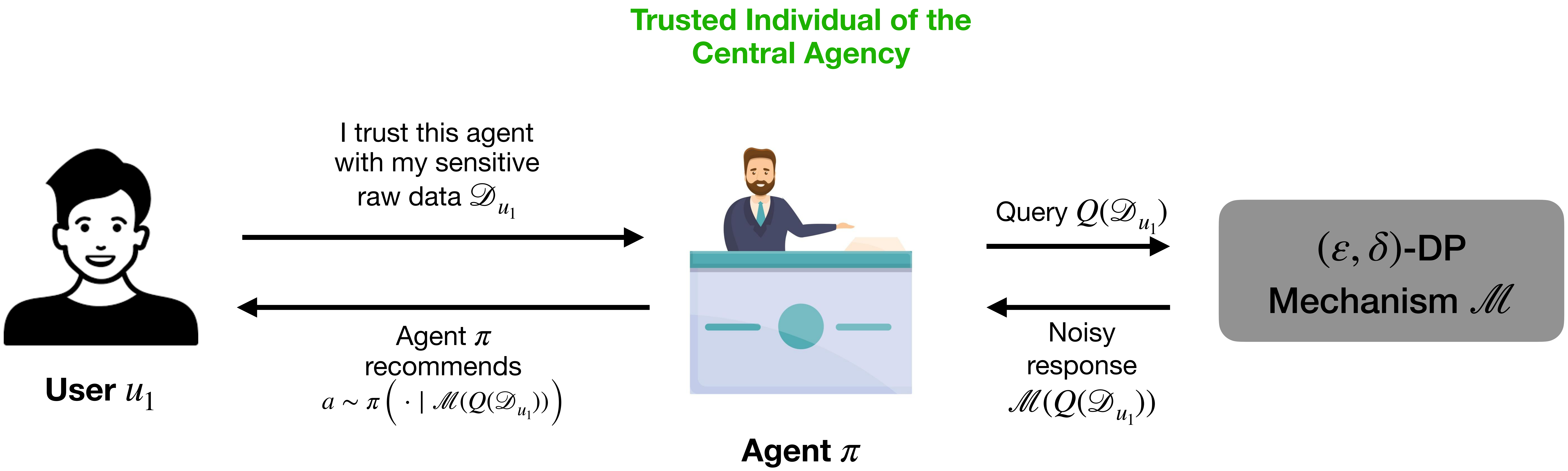
Agent π

Query $Q(\mathscr{D}_{u_1})$

Noisy response $\mathscr{M}(Q(\mathscr{D}_{u_1}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

I trust this agent with my sensitive raw data $\mathscr{D}_{u_1}$

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}_{u_1}))\right)$

User u_1

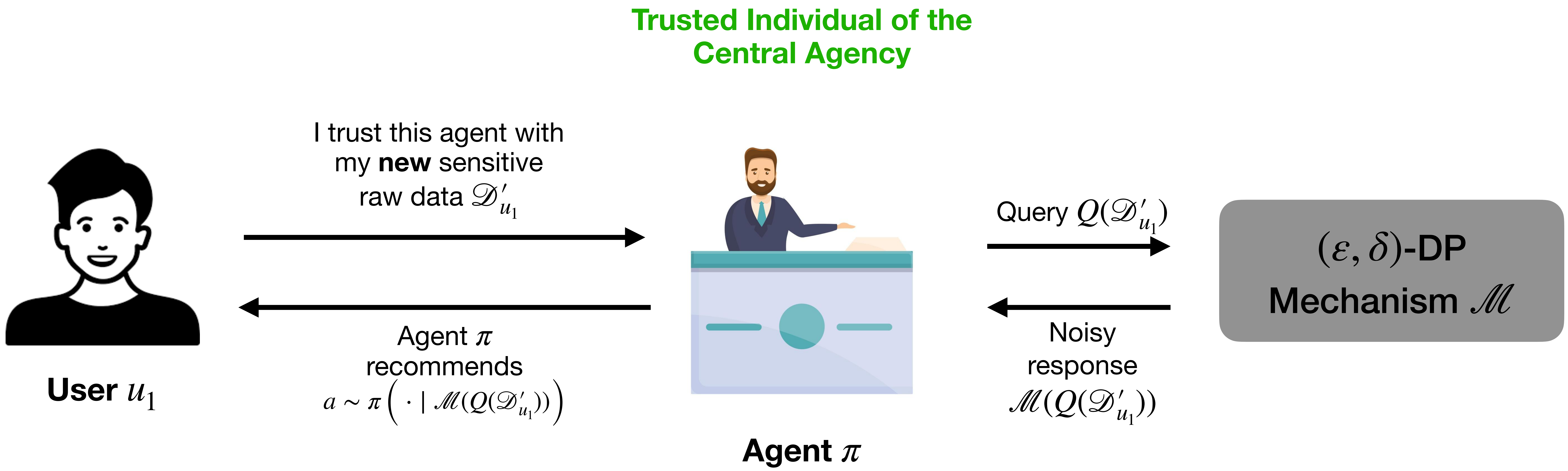
Agent π

Query $Q(\mathscr{D}_{u_1})$

Noisy response $\mathscr{M}(Q(\mathscr{D}_{u_1}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

I trust this agent with my **new** sensitive raw data $\mathscr{D}'_{u_1}$

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}'_{u_1}))\right)$

User u_1

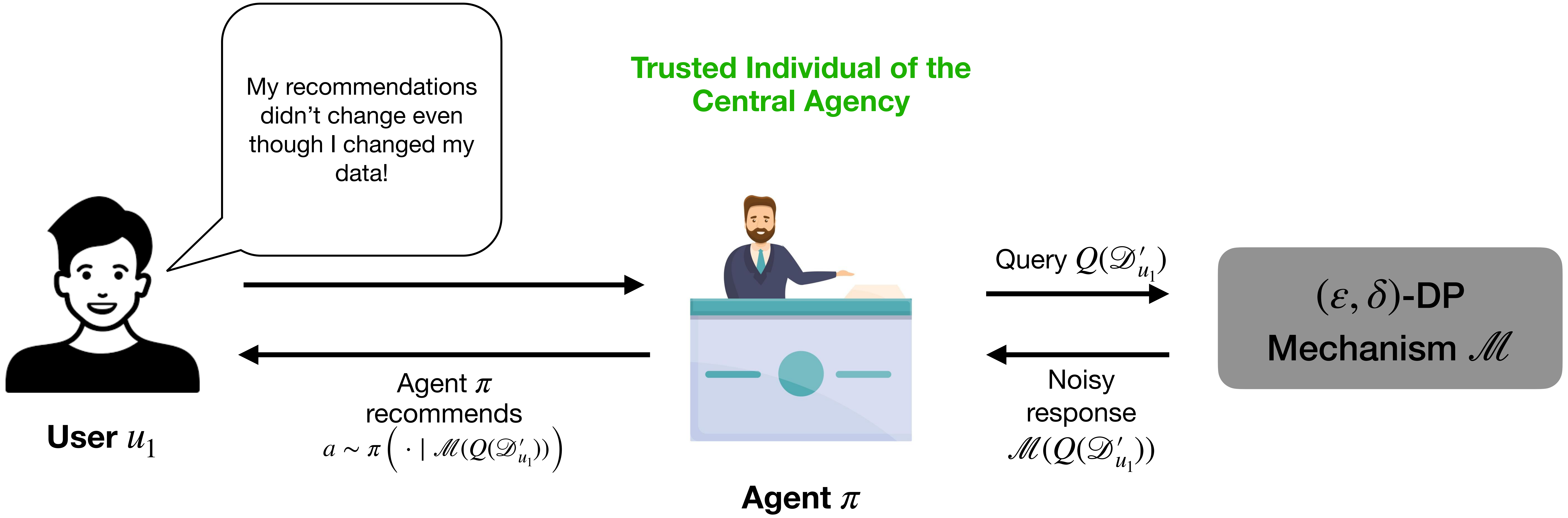
Agent π

Query $Q(\mathscr{D}'_{u_1})$

Noisy response $\mathscr{M}(Q(\mathscr{D}'_{u_1}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



My recommendations didn't change even though I changed my data!

Trusted Individual of the Central Agency

Query $Q(\mathscr{D}'_{u_1})$

(ε, δ)-DP Mechanism $\mathscr{M}$

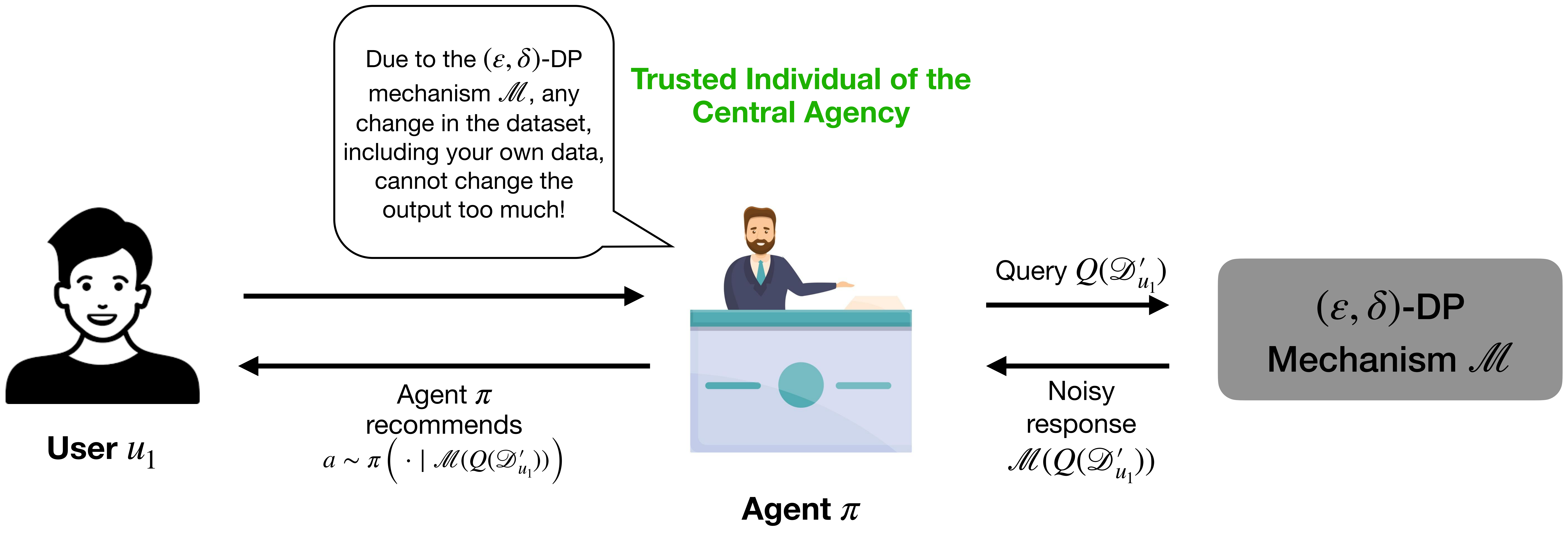
User u_1

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}'_{u_1}))\right)$

Noisy response $\mathscr{M}(Q(\mathscr{D}'_{u_1}))$

Agent π

There are a few issues with (ε, δ)-DP



Due to the (ε, δ)-DP mechanism $\mathscr{M}$, any change in the dataset, including your own data, cannot change the output too much!

Trusted Individual of the Central Agency

User u_1

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}'_{u_1}))\right)$

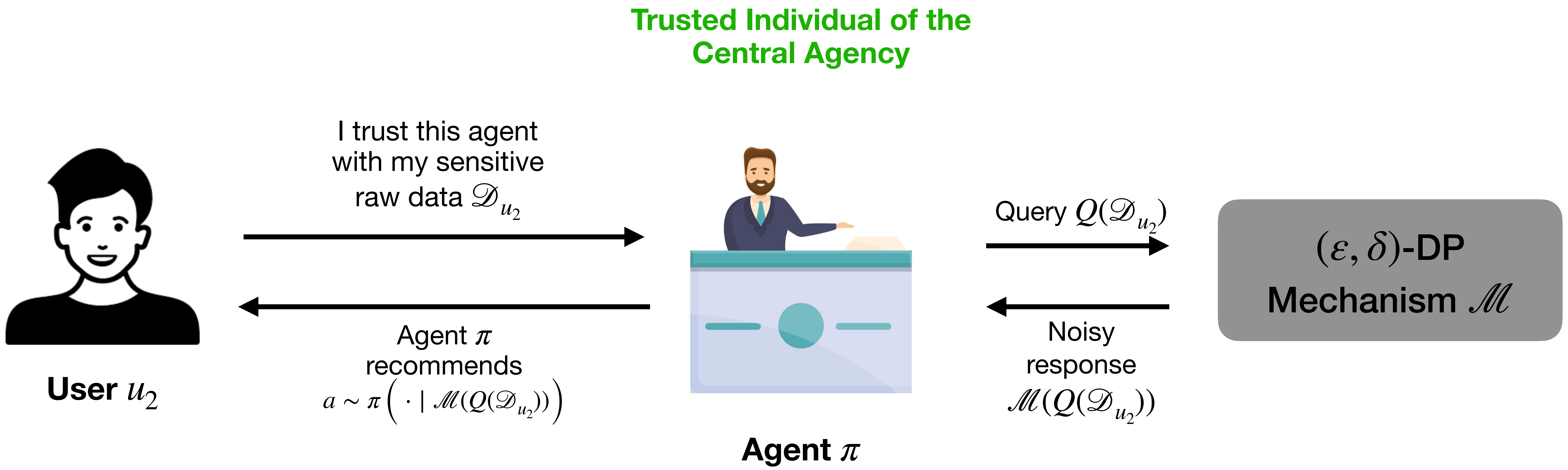
Agent π

Query $Q(\mathscr{D}'_{u_1})$

Noisy response $\mathscr{M}(Q(\mathscr{D}'_{u_1}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Trusted Individual of the Central Agency

I trust this agent with my sensitive raw data $\mathscr{D}_{u_2}$

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}_{u_2}))\right)$

Query $Q(\mathscr{D}_{u_2})$

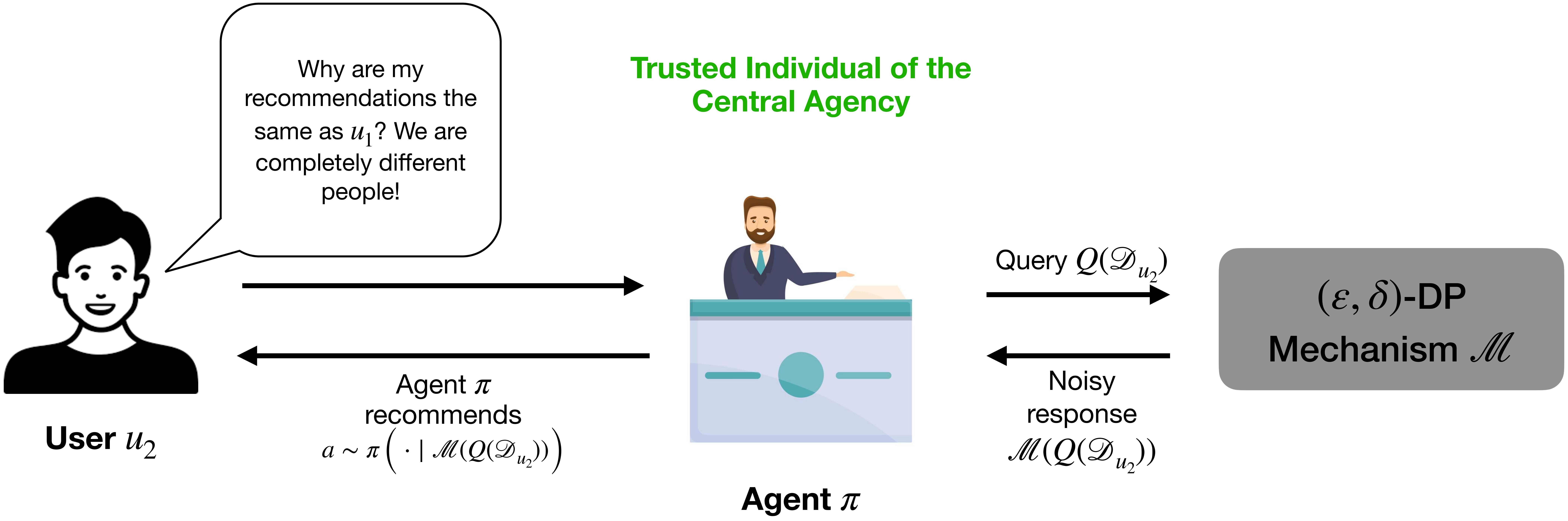
Noisy response $\mathscr{M}(Q(\mathscr{D}_{u_2}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

User u_2

Agent π

There are a few issues with (ε, δ)-DP



Why are my recommendations the same as u_1? We are completely different people!

User u_2

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}_{u_2}))\right)$

Trusted Individual of the Central Agency

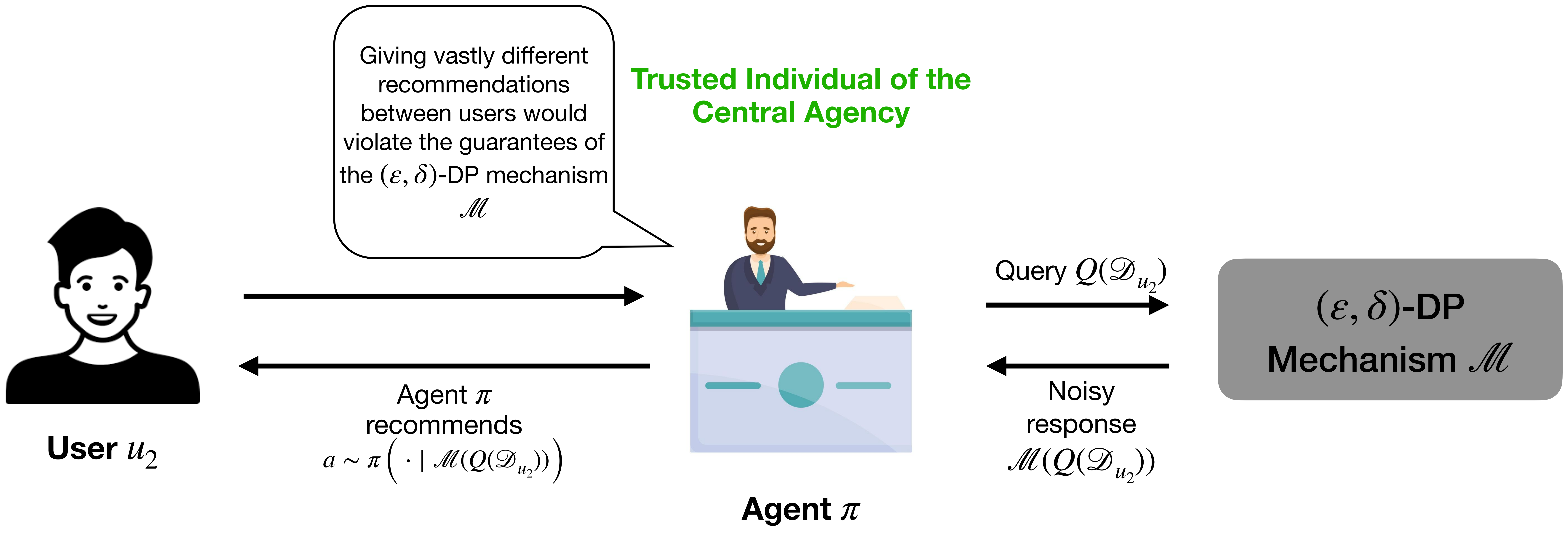
Agent π

Query $Q(\mathscr{D}_{u_2})$

Noisy response $\mathscr{M}(Q(\mathscr{D}_{u_2}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

There are a few issues with (ε, δ)-DP



Giving vastly different recommendations between users would violate the guarantees of the (ε, δ)-DP mechanism $\mathscr{M}$

Trusted Individual of the Central Agency

User u_2

Agent π recommends $a \sim \pi\left(\cdot \mid \mathscr{M}(Q(\mathscr{D}_{u_2}))\right)$

Agent π

Query $Q(\mathscr{D}_{u_2})$

Noisy response $\mathscr{M}(Q(\mathscr{D}_{u_2}))$

(ε, δ)-DP Mechanism $\mathscr{M}$

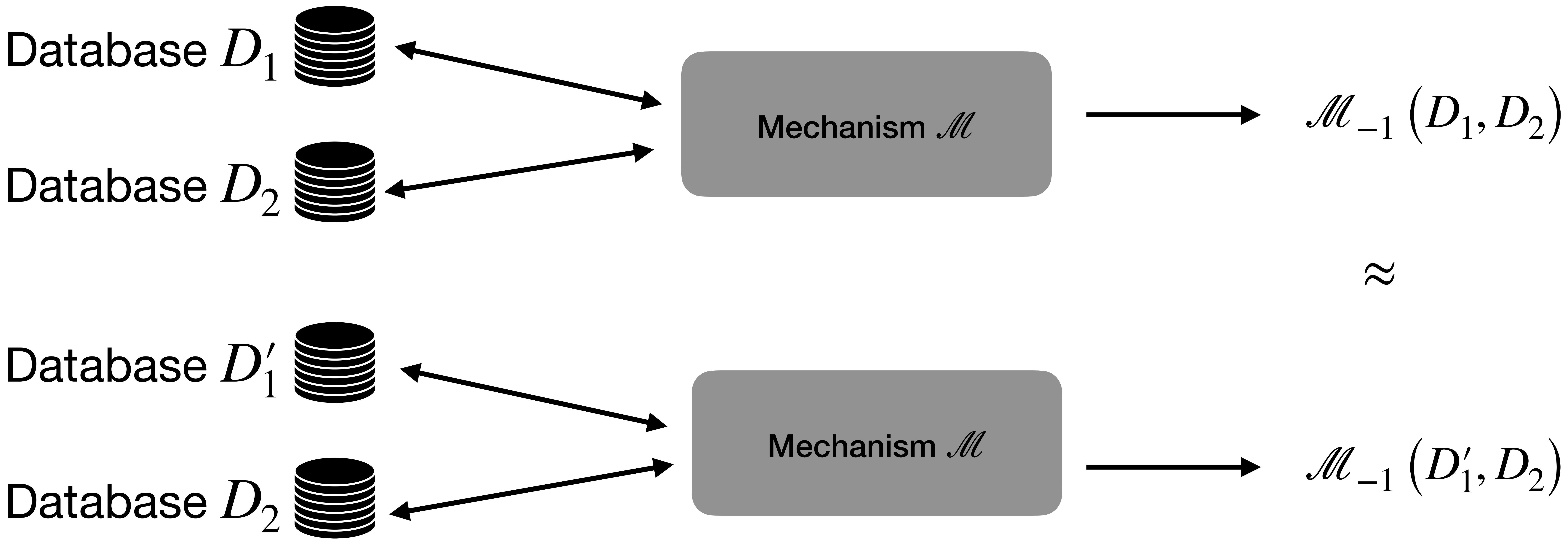
We need a further relaxation of DP …

One which works nicely with contextual bandit problems on a per-user level but does not sacrifice privacy on a per-decision or per-context level, ensuring that individual contexts do not overly influence the learned policy:

Definition (Approximate Joint Differential Privacy). A mechanism $\mathscr{M}$ is (ε, δ)-JDP if for any $k \in [K]$, any user sequences $\mathscr{U}, \mathscr{U}'$ differing on the k-user and any $E \subset \mathscr{A}^{(K-1)H}$

$$\mathbb{P}\left(\mathscr{M}_{-k}(\mathscr{U}) \in E\right) \leq e^{\varepsilon}\mathbb{P}\left(\mathscr{M}_{-k}(\mathscr{U}') \in E\right)$$

Joint Differential Privacy (JDP)



Database D_1

Database D_2

Mechanism $\mathscr{M}$

$\mathscr{M}_{-1}\left(D_1, D_2\right)$

$\approx$

Database D_1'

Database D_2

Mechanism $\mathscr{M}$

$\mathscr{M}_{-1}\left(D_1', D_2\right)$

Mechanism $\mathscr{M}$ is joint differentially private if …

$\forall\left(D_1, D_2, \ldots, D_k\right), \left(D_1', D_2, \ldots, D_k\right)$ where only one party's data differs by at most one record

$\mathscr{M}_{-1}\left(D_1, D_2\right), \mathscr{M}_{-1}\left(D_1', D_2\right)$ are indistinguishable

In this talk:

Can we develop an efficient (ε, δ)-JDP algorithm for sequential decision-making problems with **linear parametric representations**, and provide a novel algorithm with provably efficient guarantees for **privacy-preserving exploration**?

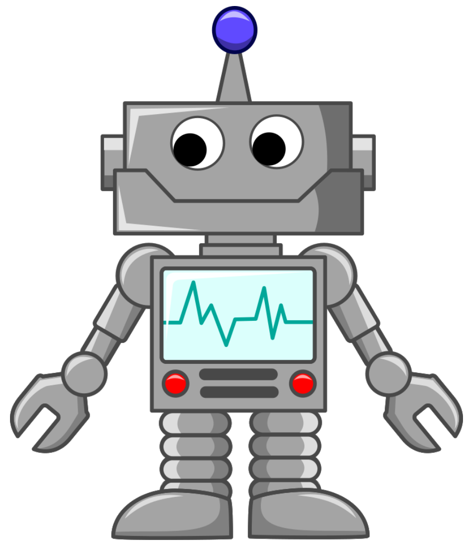
Outline

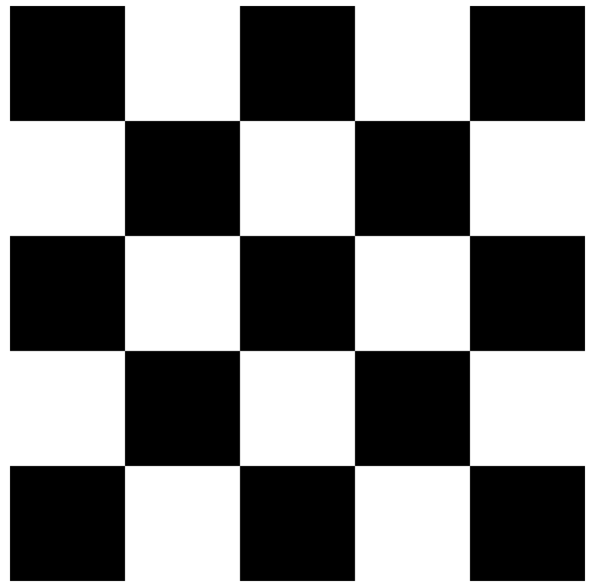
1. Problem Setup + Previous Work and Motivation

2. Can we do better?

3. Our regret bound

Episodic Time-Inhomogeneous Finite-Horizon MDPs

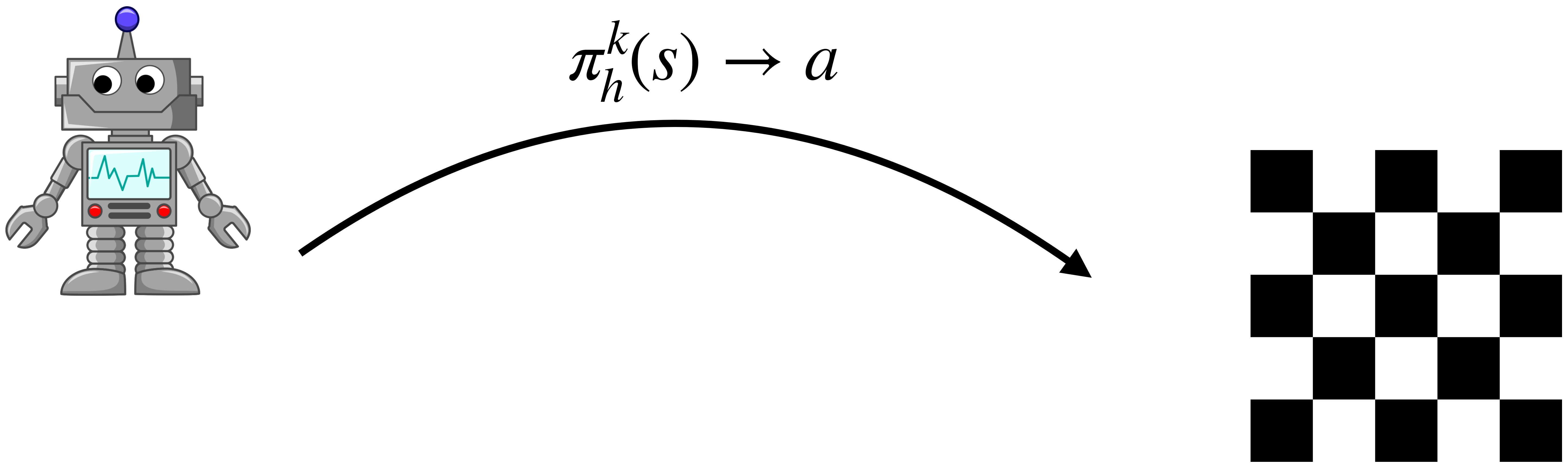




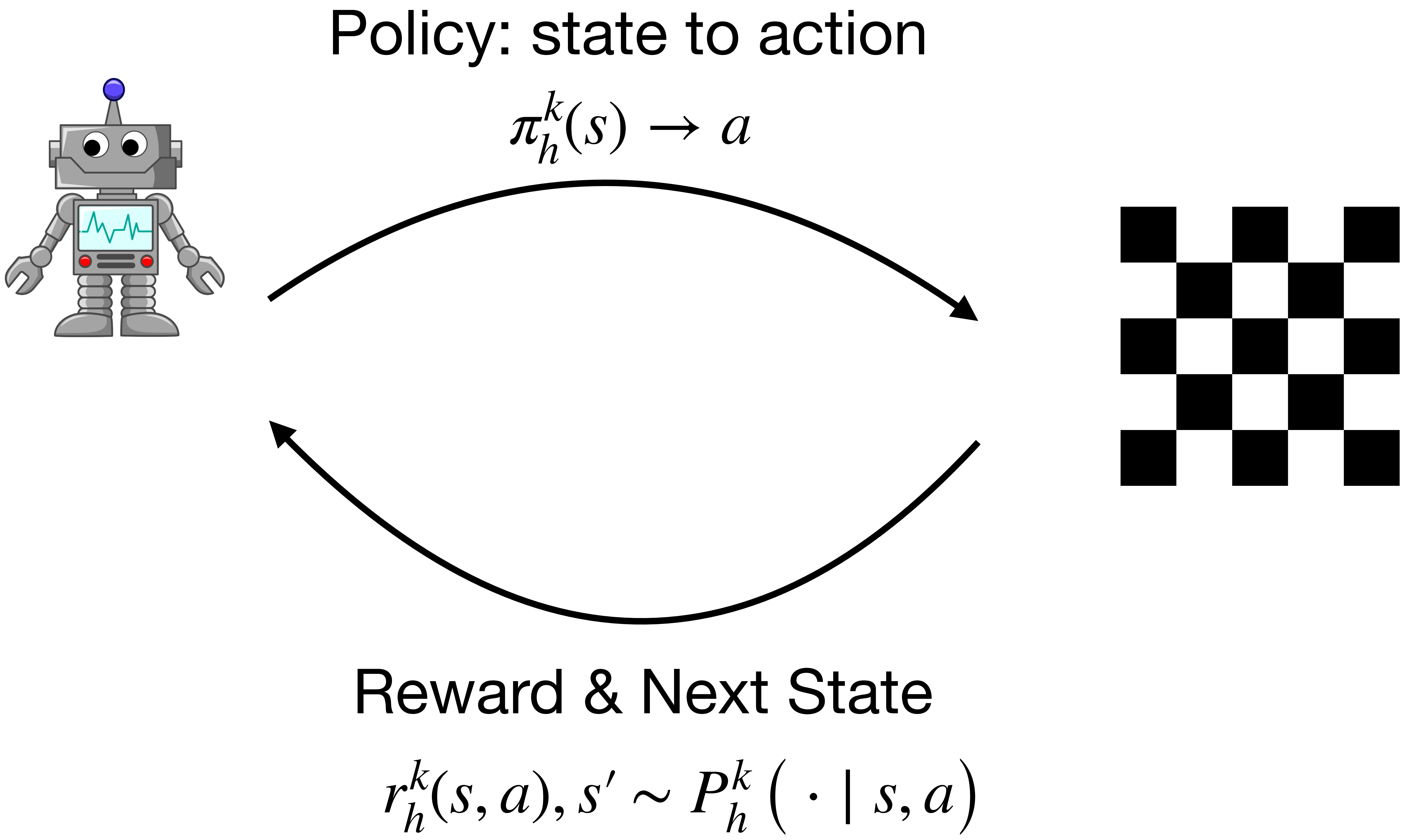
Episodic Time-Inhomogeneous Finite-Horizon MDPs

Policy: state to action

$$\pi_h^k(s) \to a$$



Episodic Time-Inhomogeneous Finite-Horizon MDPs



Policy: state to action

$\pi_h^k(s) \to a$

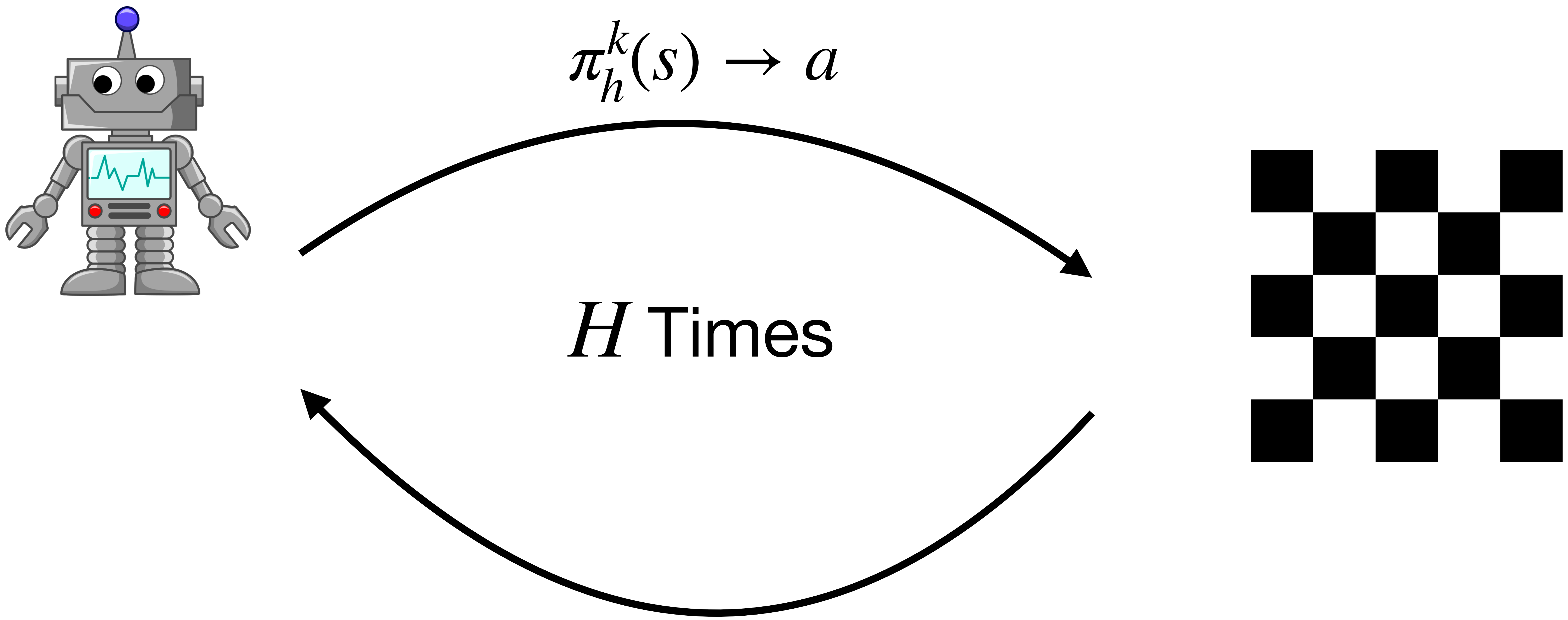
Reward & Next State

$r_h^k(s, a), s' \sim P_h^k\left(\,\cdot \mid s, a\right)$

Episodic Time-Inhomogeneous Finite-Horizon MDPs

Policy: state to action

$\pi_h^k(s) \to a$



H Times

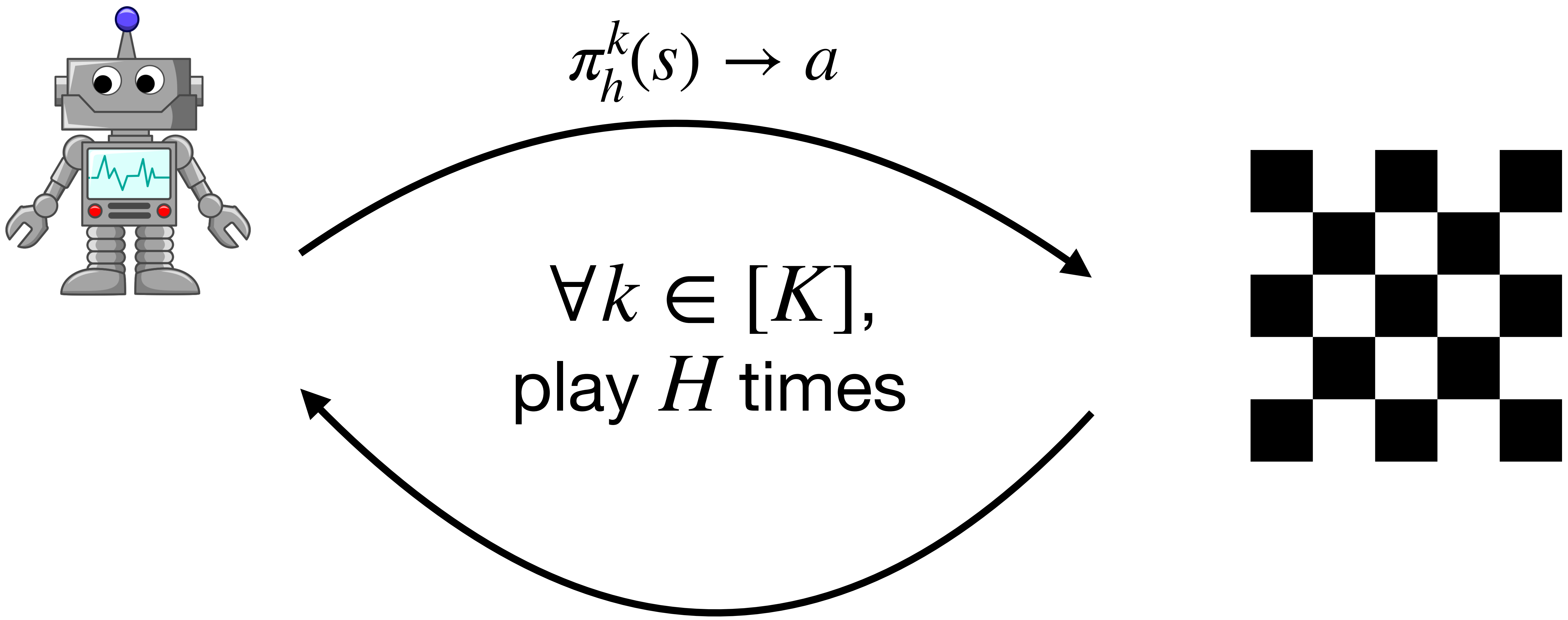
Reward & Next State

$r_h^k(s, a), s' \sim P_h^k\left(\cdot \mid s, a\right)$

Episodic Time-Inhomogeneous Finite-Horizon MDPs

Policy: state to action

$\pi_h^k(s) \to a$



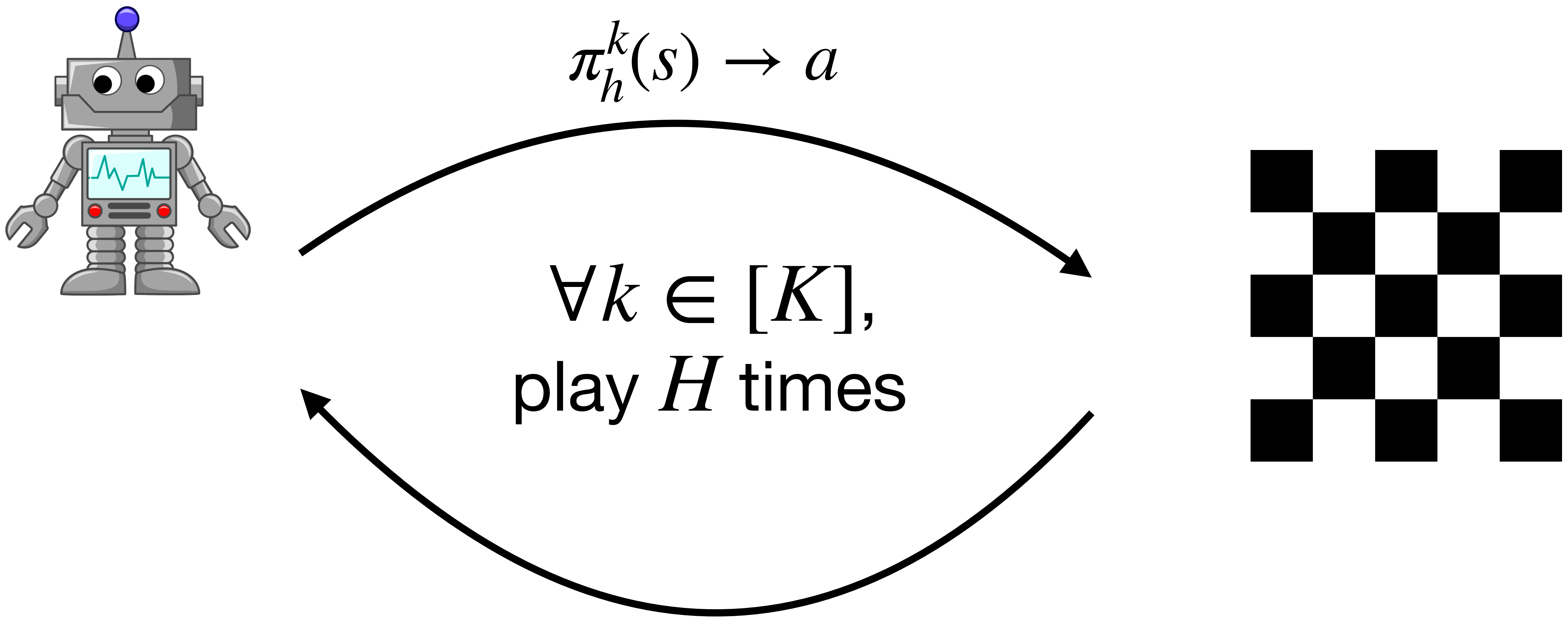
$\forall k \in [K]$,
play H times

Reward & Next State

$r_h^k(s,a), s' \sim P_h^k\left(\cdot \mid s,a\right)$

Episodic Time-Inhomogeneous Finite-Horizon MDPs

$$\tau^k = \left\{s_h^k, a_h^k\right\}_{h=1}^{H}$$



Policy: state to action

$$\pi_h^k(s) \to a$$

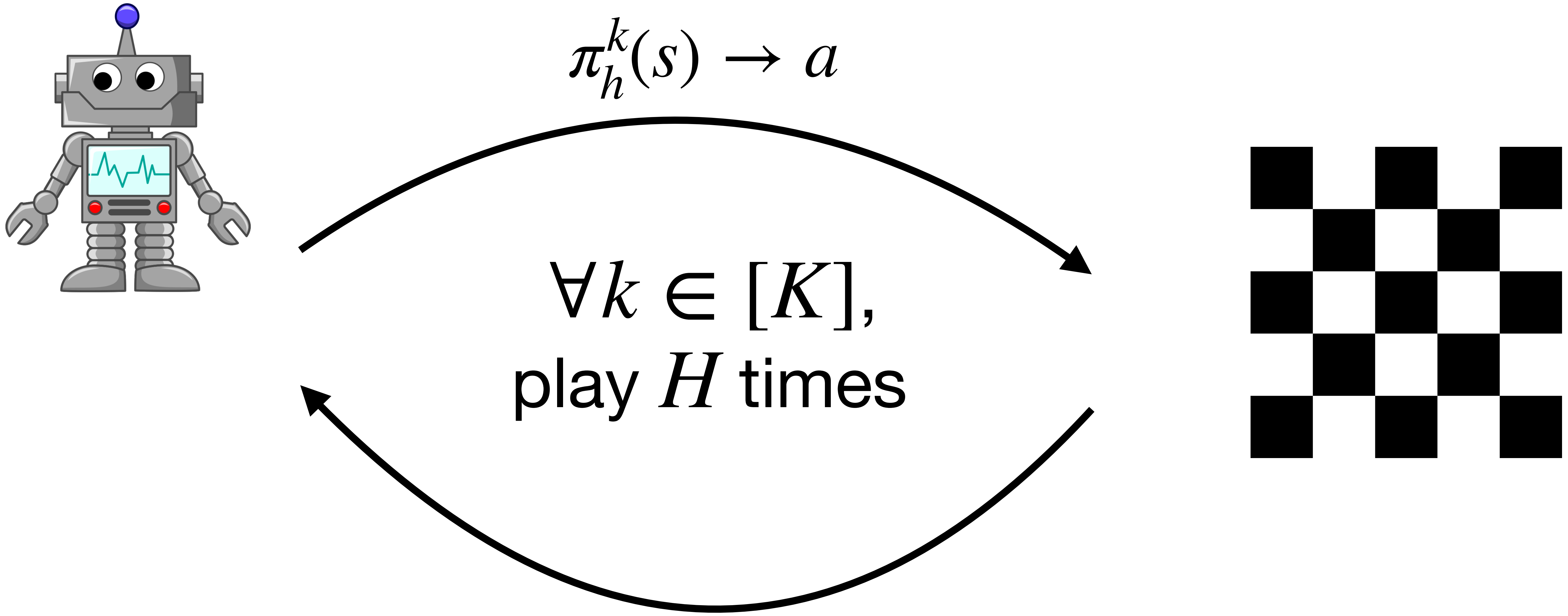
$\forall k \in [K]$,
play H times

Reward & Next State

$$r_h^k(s,a), s' \sim P_h^k\left(\,\cdot \mid s, a\right)$$

Episodic Time-Inhomogeneous Finite-Horizon MDPs

$$\tau^k = \left\{s_h^k, a_h^k\right\}_{h=1}^H$$



Policy: state to action

$$\pi_h^k(s) \to a$$

$\forall k \in [K]$,
play H times

Reward & Next State

$$r_h^k(s,a), s' \sim P_h^k\left(\cdot \mid s,a\right)$$

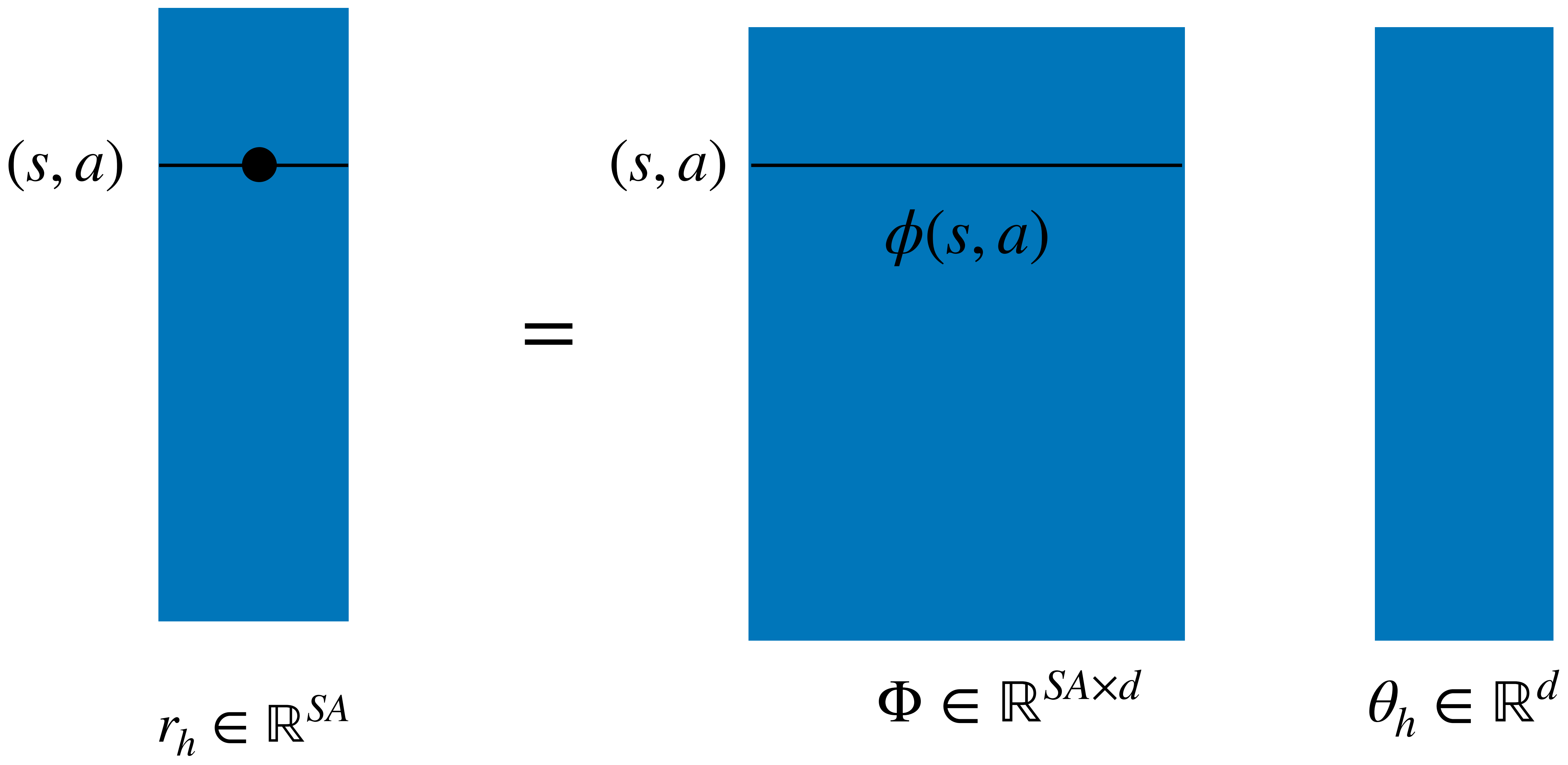
Finite-Horizon MDP: $\mathscr{M} = \left\{\mathcal{S}, \mathcal{A}, \left\{r_h\right\}_{h=1}^H, \left\{P_h\right\}_{h=1}^H, H\right\}$ $H < \infty$

Formal RL Problem Setting

Setting: Episodic inhomogeneous finite horizon MDP $\mathscr{M} = \{\mathscr{S}, \mathscr{A}, \{\mathbb{P}_h\}_h, \{r_h\}_h, H\}$ where $\mathscr{S}, \mathscr{A}$ are the states and actions, respectively, $H \in \mathbb{Z}$ is the length of each episode, $\mathbb{P}_h : \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})$, $r_h : \mathscr{S} \times \mathscr{A} \to [0,1]$ are the time-dependent transition probability and deterministic reward function. $\mathscr{S}$ is measurable and possibly uncountable, and $\mathscr{A}$ is finite. In this setting, the policy is time-dependent and we denote this $\pi = \{\pi_1, \cdots, \pi_H\}$

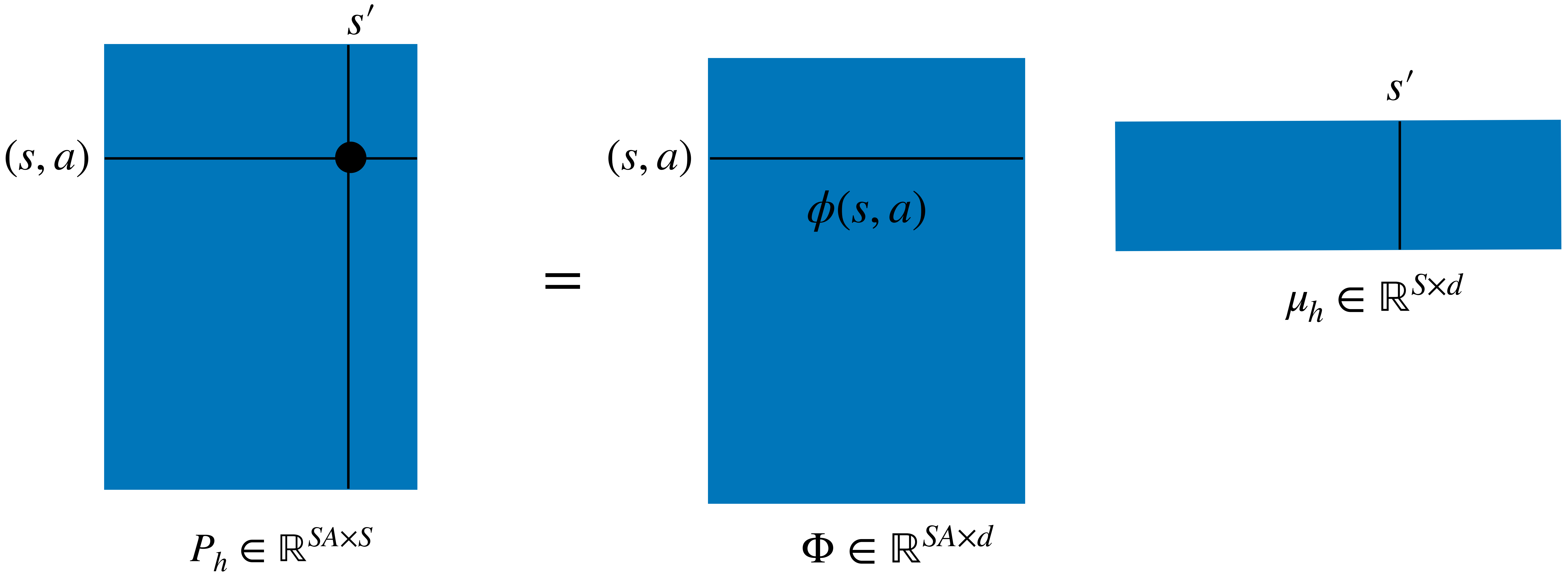
$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^* \left(s_1^k\right) - V_1^{\pi_k} \left(s_1^k\right) \right]$$

Linear MDP


(s,a)
$=$
(s,a)
$\phi(s,a)$
$r_h \in \mathbb{R}^{SA}$
$\Phi \in \mathbb{R}^{SA\times d}$
$\theta_h \in \mathbb{R}^d$

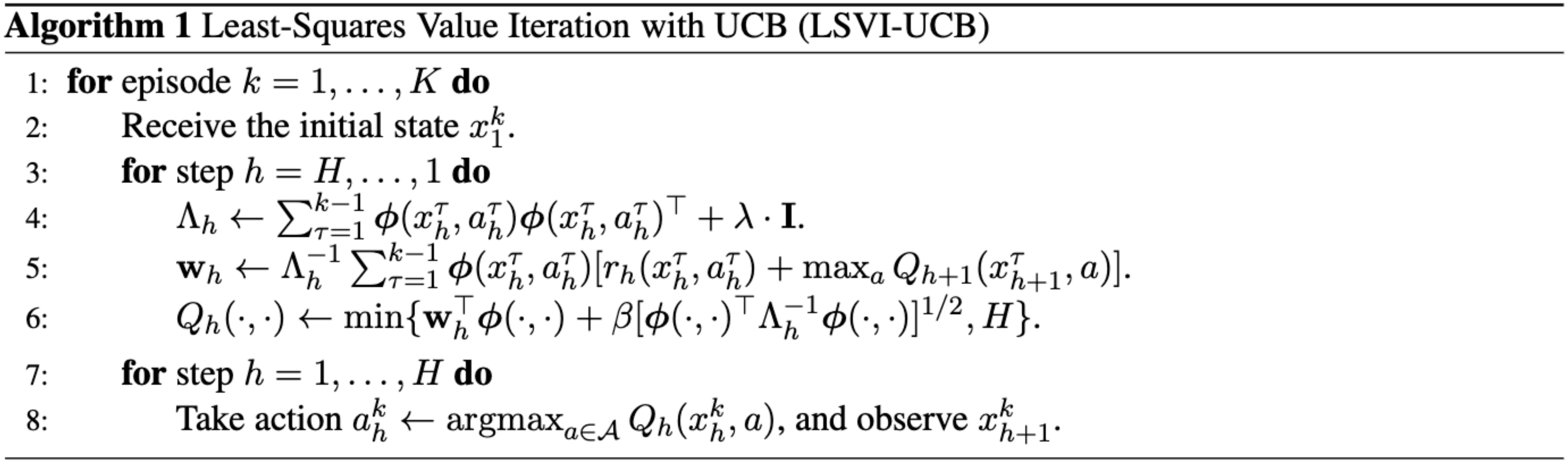
$$\exists \theta_h, \phi^\star : \forall s, a, h,\ r_h(s,a) = \phi^\star(s,a)^\top \theta_h$$

Linear MDP


s'
(s, a)
$P_h \in \mathbb{R}^{SA \times S}$
$=$
(s, a)
$\phi(s, a)$
$\Phi \in \mathbb{R}^{SA \times d}$
s'
$\mu_h \in \mathbb{R}^{S \times d}$

$$\exists \mu_h, \phi^\star : \forall s, a, h, s', P_h(s' \mid s, a) = \phi^\star(s, a)^\top \mu_h(s')$$

LSVI-UCB

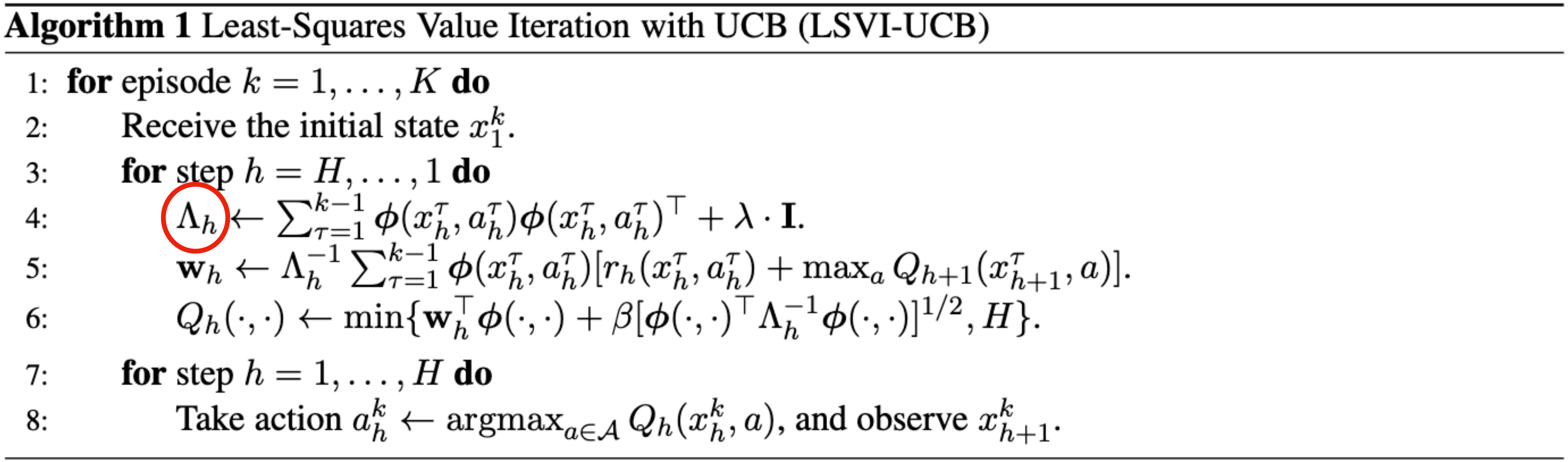


Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \ldots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \ldots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)\boldsymbol{\phi}(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot, \cdot) \leftarrow \min\{\mathbf{w}_h^\top \boldsymbol{\phi}(\cdot, \cdot) + \beta[\boldsymbol{\phi}(\cdot, \cdot)^\top \Lambda_h^{-1} \boldsymbol{\phi}(\cdot, \cdot)]^{1/2}, H\}$.
    for step $h = 1, \ldots, H$ do
        Take action $a_h^k \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

LSVI-UCB Algorithm [Jin et al. 2020]

LSVI-UCB

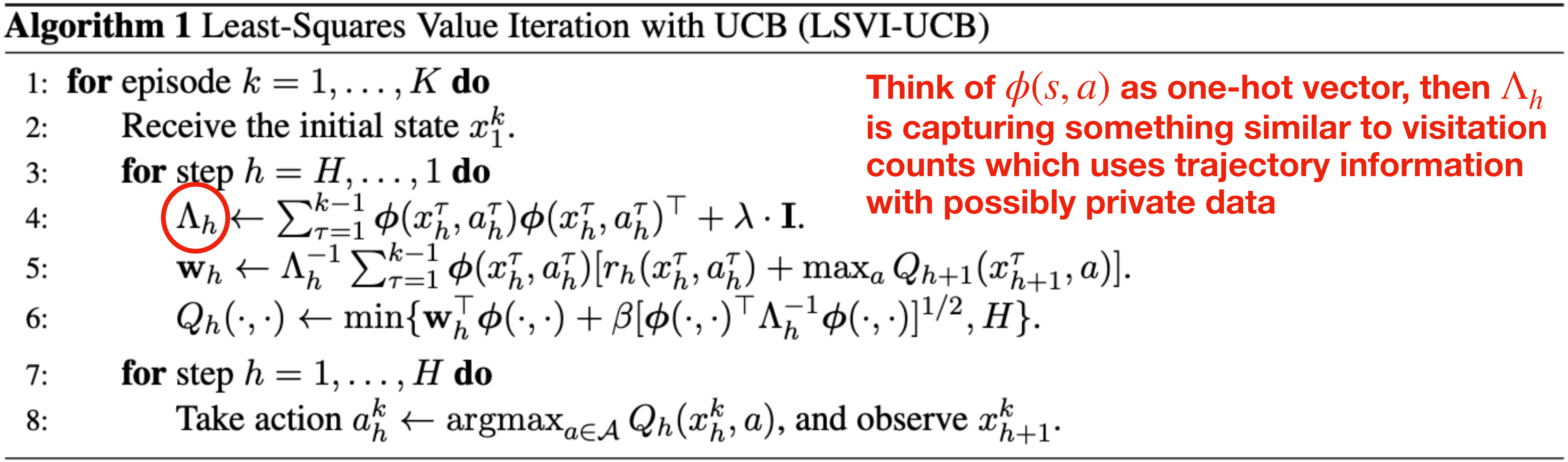


Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \ldots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \ldots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot,\cdot) \leftarrow \min\{\mathbf{w}_h^\top \phi(\cdot,\cdot) + \beta[\phi(\cdot,\cdot)^\top \Lambda_h^{-1} \phi(\cdot,\cdot)]^{1/2}, H\}$.
    for step $h = 1, \ldots, H$ do
        Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

LSVI-UCB Algorithm [Jin et al. 2020]

LSVI-UCB



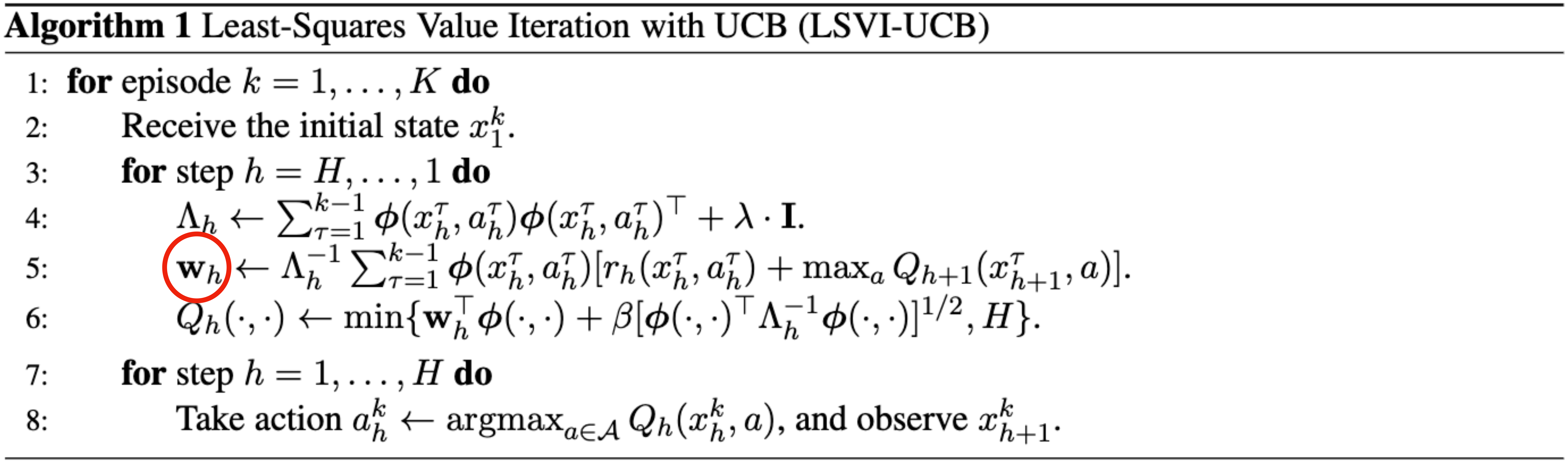
Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \dots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \dots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)\boldsymbol{\phi}(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot,\cdot) \leftarrow \min\{\mathbf{w}_h^\top \boldsymbol{\phi}(\cdot,\cdot) + \beta[\boldsymbol{\phi}(\cdot,\cdot)^\top \Lambda_h^{-1} \boldsymbol{\phi}(\cdot,\cdot)]^{1/2}, H\}$.
    for step $h = 1, \dots, H$ do
        Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

Think of $\phi(s, a)$ as one-hot vector, then Λ_h is capturing something similar to visitation counts which uses trajectory information with possibly private data

LSVI-UCB Algorithm [Jin et al. 2020]

LSVI-UCB

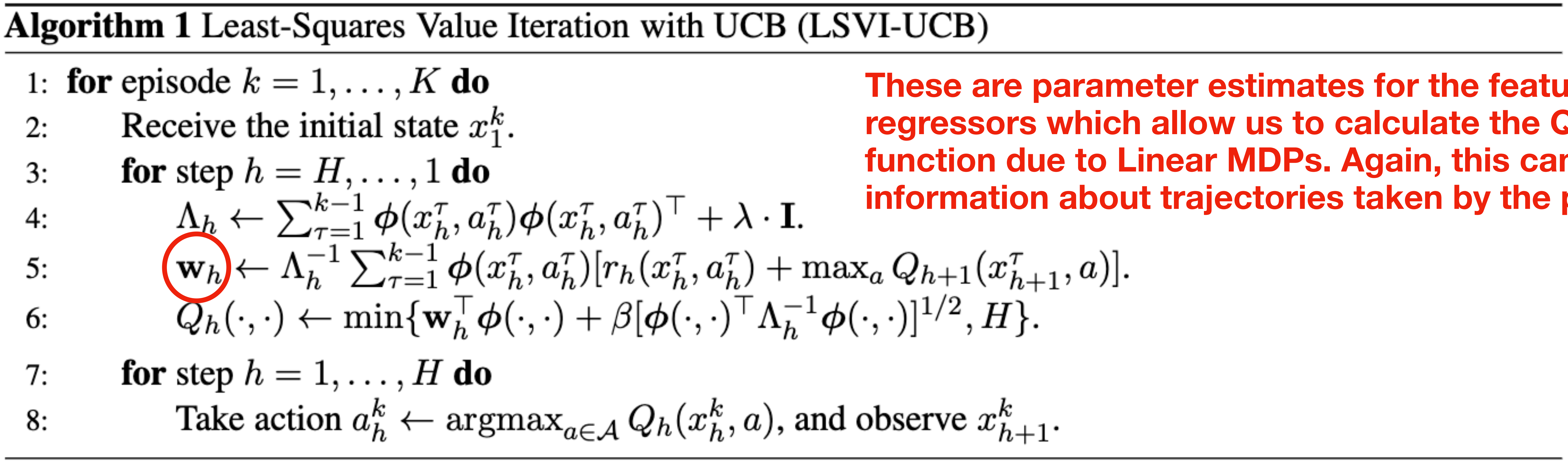


Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \ldots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \ldots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)\boldsymbol{\phi}(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot, \cdot) \leftarrow \min\{\mathbf{w}_h^\top \boldsymbol{\phi}(\cdot, \cdot) + \beta[\boldsymbol{\phi}(\cdot, \cdot)^\top \Lambda_h^{-1} \boldsymbol{\phi}(\cdot, \cdot)]^{1/2}, H\}$.
    for step $h = 1, \ldots, H$ do
        Take action $a_h^k \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

LSVI-UCB Algorithm [Jin et al. 2020]

LSVI-UCB



Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \ldots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \ldots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)\boldsymbol{\phi}(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot,\cdot) \leftarrow \min\{\mathbf{w}_h^\top \boldsymbol{\phi}(\cdot,\cdot) + \beta[\boldsymbol{\phi}(\cdot,\cdot)^\top \Lambda_h^{-1} \boldsymbol{\phi}(\cdot,\cdot)]^{1/2}, H\}$.
    for step $h = 1, \ldots, H$ do
        Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

These are parameter estimates for the feature regressors which allow us to calculate the Q-function due to Linear MDPs. Again, this can leak information about trajectories taken by the policy

LSVI-UCB Algorithm [Jin et al. 2020]

Differential Privacy Techniques

Theorem (Gaussian Mechanism). $M_G(x, f(\cdot), \varepsilon, \delta) = f(x) + (Y_1, \ldots, Y_k)$ where $Y_i \sim \mathcal{N}\left(0, \sigma^2\right)$ with $\sigma^2 = \dfrac{\Delta_2(f)\sqrt{2\log(2/\delta)}}{\varepsilon}$ is (ε, δ)-DP

Theorem (Billboard Lemma). If you have a mechanism $\mathscr{M}_{DP}$ that is (ε, δ)-DP, then any function $f_i : \mathscr{U}_i \times \mathscr{R} \to \mathscr{R}_i$ that depends on user i's data and the output of the mechanism satisfies (ε, δ)-JDP

Previous Work

[Luyo et al. 2021]. Fix any privacy level $\varepsilon, \delta \in (0,1)$. For any $p \in (0,1)$, their algorithm is (ε, δ)-JDP and, with probability at least $1 - p$, its regret is bounded as follows:

$$R(K) = \tilde{O}(\sqrt{d^3H^4K} + H^{11/5}d^{8/5}K^{3/5}/\varepsilon^{2/5})$$

Techniques Used

$$\tilde{\Lambda}_h = \Lambda_h + \mathcal{N}\left(0,\ \mathcal{O}\left(\frac{1}{\varepsilon}\sqrt{BH}\log\left(1/\delta\right)\right)\right)$$

$$\tilde{u}_h = u_h + \mathcal{N}\left(0,\ \mathcal{O}\left(\frac{1}{\varepsilon}\sqrt{H^2B}\log\left(1/\delta\right)\right)\right)$$

Static Batching to reduce the number of policy switches to $\mathcal{O}(\text{poly}(K))$

Previous Work

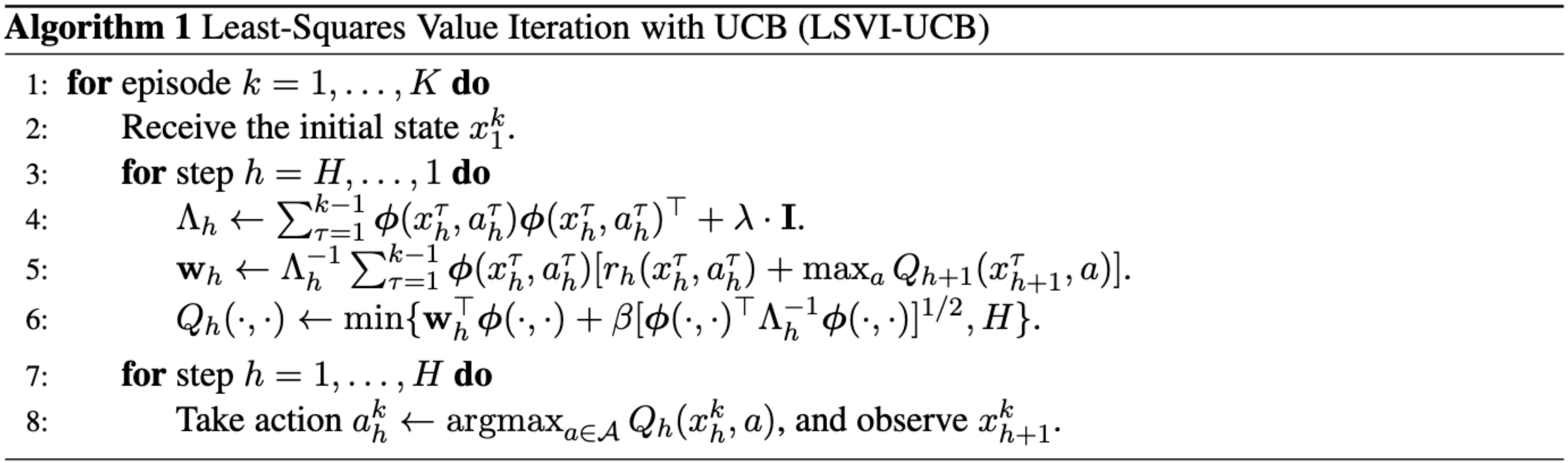
[Ngo et al. 2022]. Fix any privacy level $\varepsilon, \delta \in (0,1)$. For any $p \in (0,1)$, their algorithm is (ε, δ)-JDP and, with probability at least $1 - p$, its regret is bounded as follows:

$$R(K) = \tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\varepsilon^{1/2})$$

Techniques Used

Same techniques as previous work but instead of a static batching schedule, they use **Adaptive Batching** to reduce the number of policy switches to $\mathcal{O}(\log(K))$

LSVI-UCB



Algorithm 1 Least-Squares Value Iteration with UCB (LSVI-UCB)

```
for episode $k = 1, \dots, K$ do
    Receive the initial state $x_1^k$.
    for step $h = H, \dots, 1$ do
        $\Lambda_h \leftarrow \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)\boldsymbol{\phi}(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I}$.
        $\mathbf{w}_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(x_h^\tau, a_h^\tau)[r_h(x_h^\tau, a_h^\tau) + \max_a Q_{h+1}(x_{h+1}^\tau, a)]$.
        $Q_h(\cdot, \cdot) \leftarrow \min\{\mathbf{w}_h^\top \boldsymbol{\phi}(\cdot, \cdot) + \beta[\boldsymbol{\phi}(\cdot, \cdot)^\top \Lambda_h^{-1} \boldsymbol{\phi}(\cdot, \cdot)]^{1/2}, H\}$.
    for step $h = 1, \dots, H$ do
        Take action $a_h^k \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h(x_h^k, a)$, and observe $x_{h+1}^k$.
```

LSVI-UCB Algorithm [Jin et al. 2020]

Achieves regret $R(K) = \tilde{\mathcal{O}}\left(H^2\sqrt{d^3K}\right)$

Motivating Work: LSVI-UCB+

[Hu et al. 2022]. Set $\lambda = 1/\left(H^2\sqrt{d}\right)$. Then, with probability at least $1 - 10\delta$, the regret of LSVI-UCB+ is upper bounded by

$$R(K) = \tilde{\mathcal{O}}\left(d\sqrt{H^3K}\right)$$

Techniques Used

Instead of solving a **ridge regression** problem, we solve a **weighted ridge regression** problem using estimated weights from data. This allows us to use a self-normalized martingale argument using **Azuma-Bernstein** rather than **Azuma-Hoeffding** to get a bonus that improves our regret

Motivating Work: JDP In Tabular MDPs

[Qiao and Wang. 2023]. For any privacy budget $\varepsilon > 0$, failure probability $0 < \beta < 1$, and any privatizer where the private counts are close to the true counts with high probability, with probability at least $1 - \beta$, their algorithm is (ε, δ)-JDP and achieves regret upper bounded by:

$$R(K) = \tilde{\mathcal{O}}\left(\sqrt{H^3SAK} + S^2AH^3/\varepsilon\right)$$

Techniques Used

In previous work, since we would use a **Hoeffding-bound** that only depends on the counts, it is sufficient to **privatize the counts** loosely using Gaussian noise with sufficient variance component. However, to use a **Bernstein-bound**, we need to **carefully privatize the counts** to ensure that we can upper bound the variance term in a Bernstein-bound

Can we design a (ε, δ)-JDP algorithm that is near minimax optimal for non-private learning and improves the cost of privacy using more refined privatization and concentration techniques?

$$R(K) = \tilde{O}(\sqrt{d^3 H^4 K} + H^3 d^{5/4} K^{1/2} / \varepsilon^{1/2})$$

Can we design a (ε, δ)-JDP algorithm that is near minimax optimal for non-private learning and improves the cost of privacy using more refined privatization and concentration techniques?

$$R(K) = \tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\varepsilon^{1/2})$$

Non-private learning regret: We can do better using LSVI-UCB+

Can we design a (ε, δ)-JDP algorithm that is near minimax optimal for non-private learning and improves the cost of privacy using more refined privatization and concentration techniques?

$$R(K) = \tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\varepsilon^{1/2})$$

Non-private learning regret: We can do better using LSVI-UCB+

Cost of privacy: can we improve this $\mathcal{O}\left(\text{poly}\left(HdK\right)/\varepsilon\right)$

Our Work

Fix any privacy level $\varepsilon, \delta \in (0,1)$. For any $p \in (0,1)$, their algorithm is (ε, δ)-JDP and, with probability at least $1 - p$, its regret is bounded as follows:

$$R(K) = \tilde{O}\left(d\sqrt{H^3K} + \frac{H^{19/8}d^{15/8}K^{3/4}}{\varepsilon} \right)$$

Compared to Luyo et al. (2021) and Ngo et al. (2022), this regret bound achieves tighter dependence on H, d for the non-private terms and tighter dependence on H, ε for the private terms

Proof Sketch

1. Identify terms in the non-private algorithm that are used for estimating
2. Privatize them by (cleverly) adding noise to the terms
3. Prove the utility of the privatized terms (how close are they to the non-private terms)
4. Use the private terms in place of the non-private terms and use your standard LSVI-UCB techniques (i.e. self-normalized martingale concentrations, uniform covering arguments, elliptical potentials, and utility of the privatized terms)

Questions?