REINFORCEMENT LEARNING IN THE CONTINUING SETTING

Abhishek Naik, Zaheer Abbas, Adam White, Richard Sutton

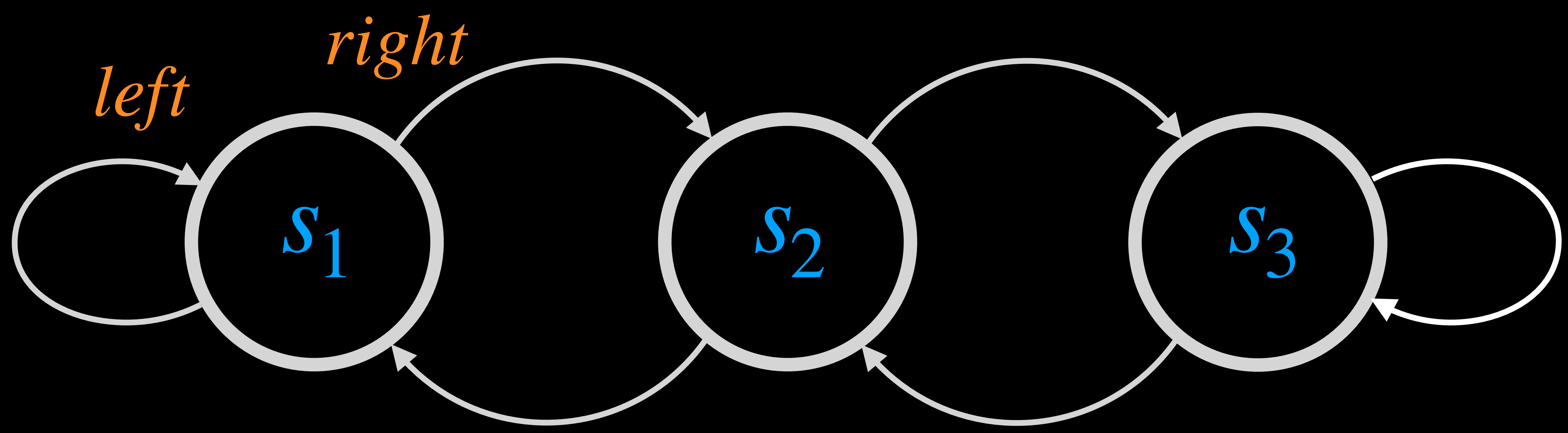
A Roadmap to Never-Ending RL @ ICLR 2021

OUTLINE

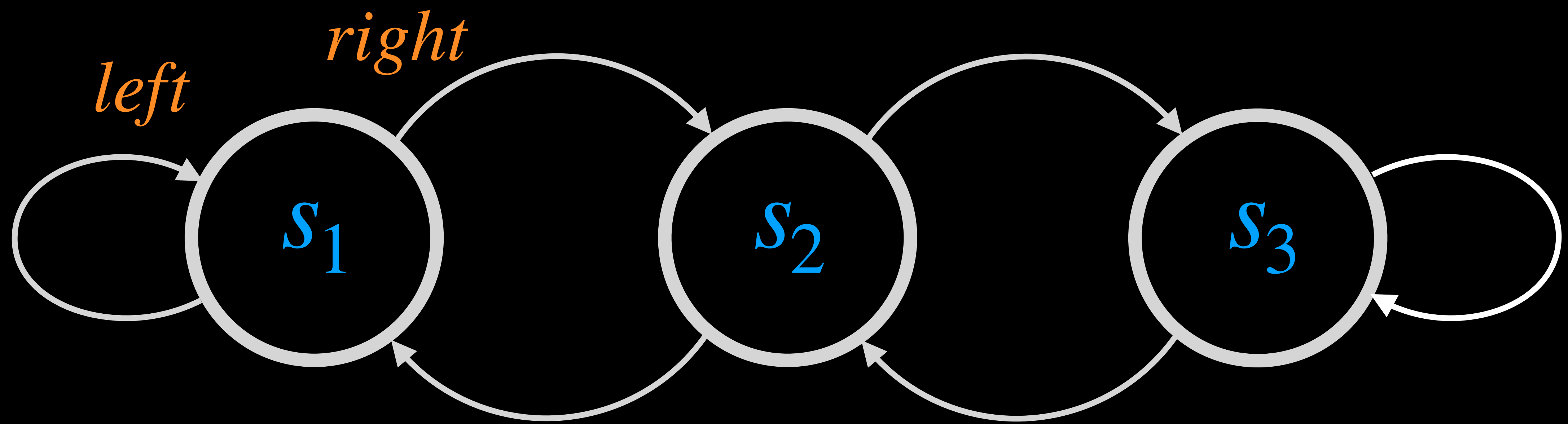
- The Problem Setting
  - continuing, compared to episodic or single-life
  - difference with continual/lifelong/never-ending and continuous
- The State of Research in the Continuing Setting
  - discounted and average-reward formulations
  - what is missing in existing problem-suites
- C-suite
  - the two broad categories of problems in C-suite
  - where do we go from here, with C-suite, and the continuing setting in general

THE PROBLEM SETTING

TYPES OF PROBLEMS



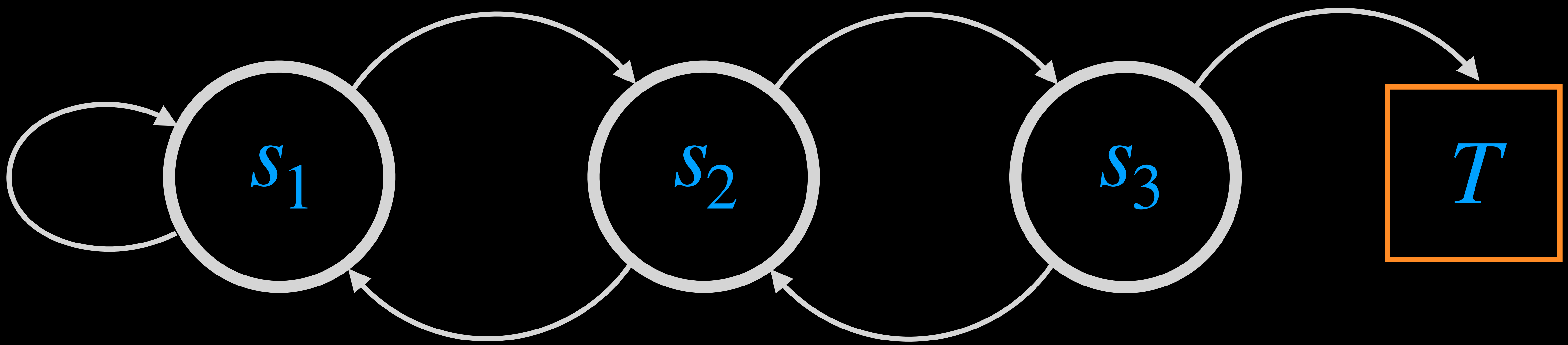
TYPES OF PROBLEMS



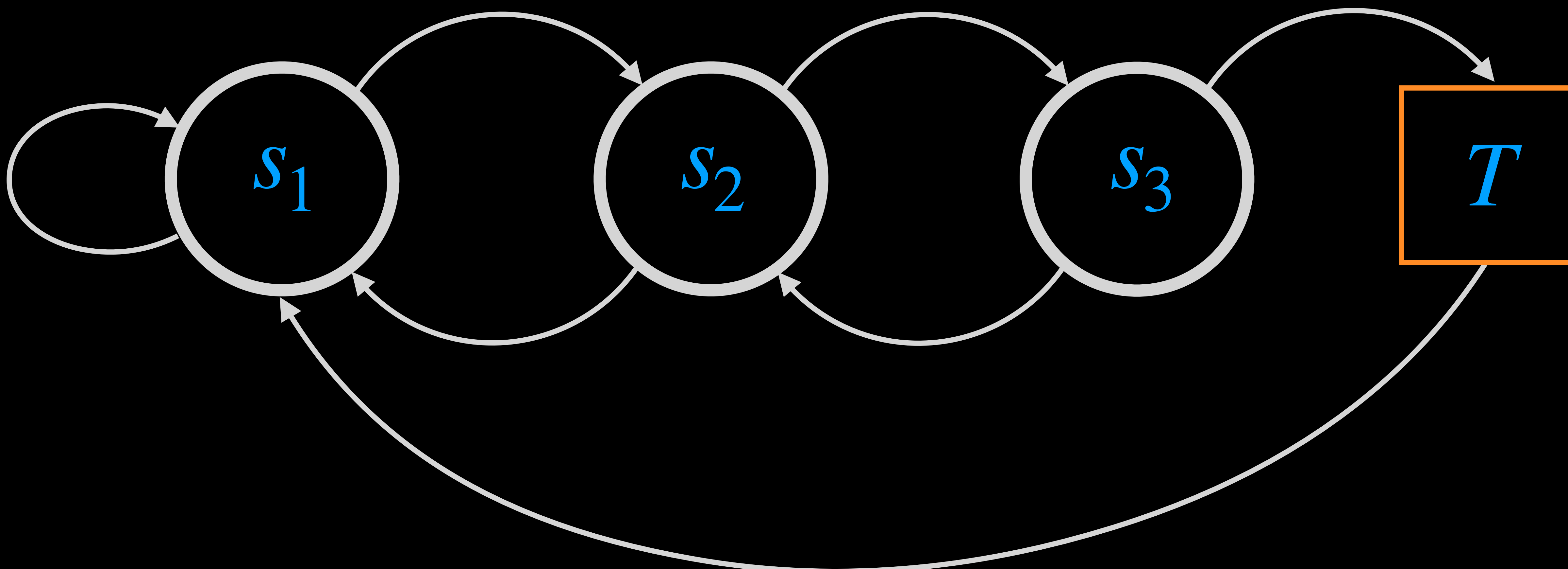
a continuing problem

TYPES OF PROBLEMS

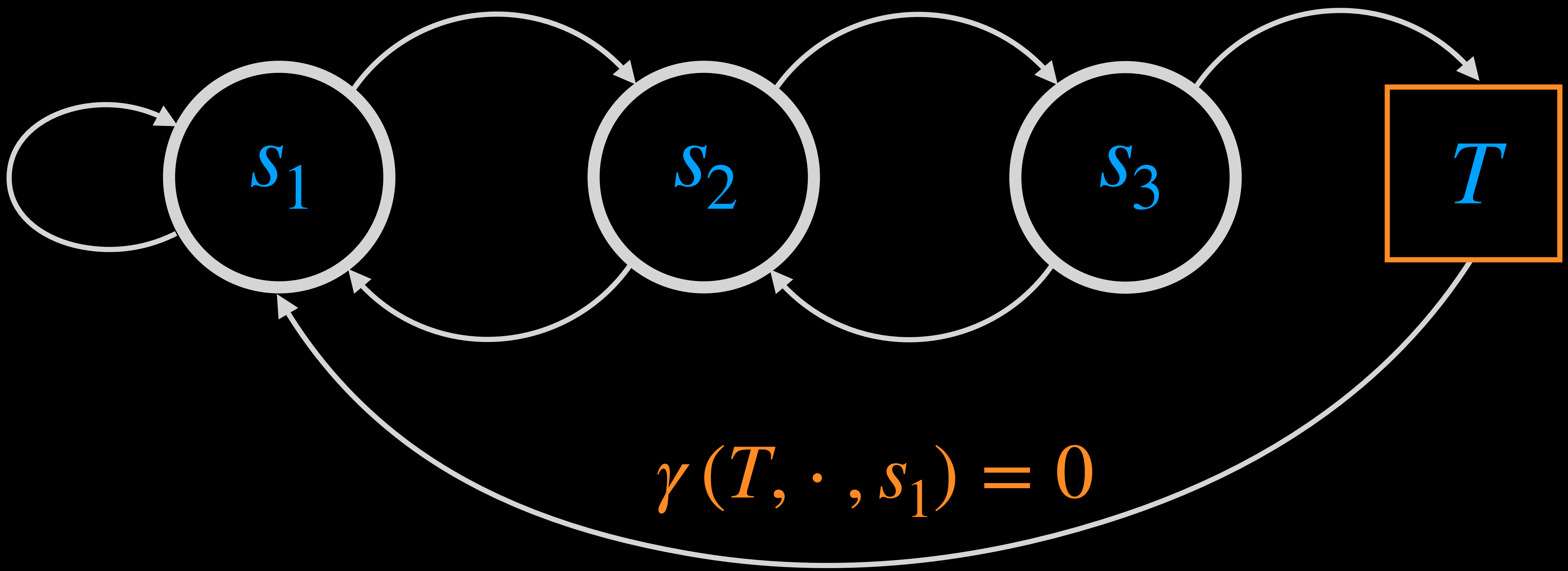
TYPES OF PROBLEMS



TYPES OF PROBLEMS

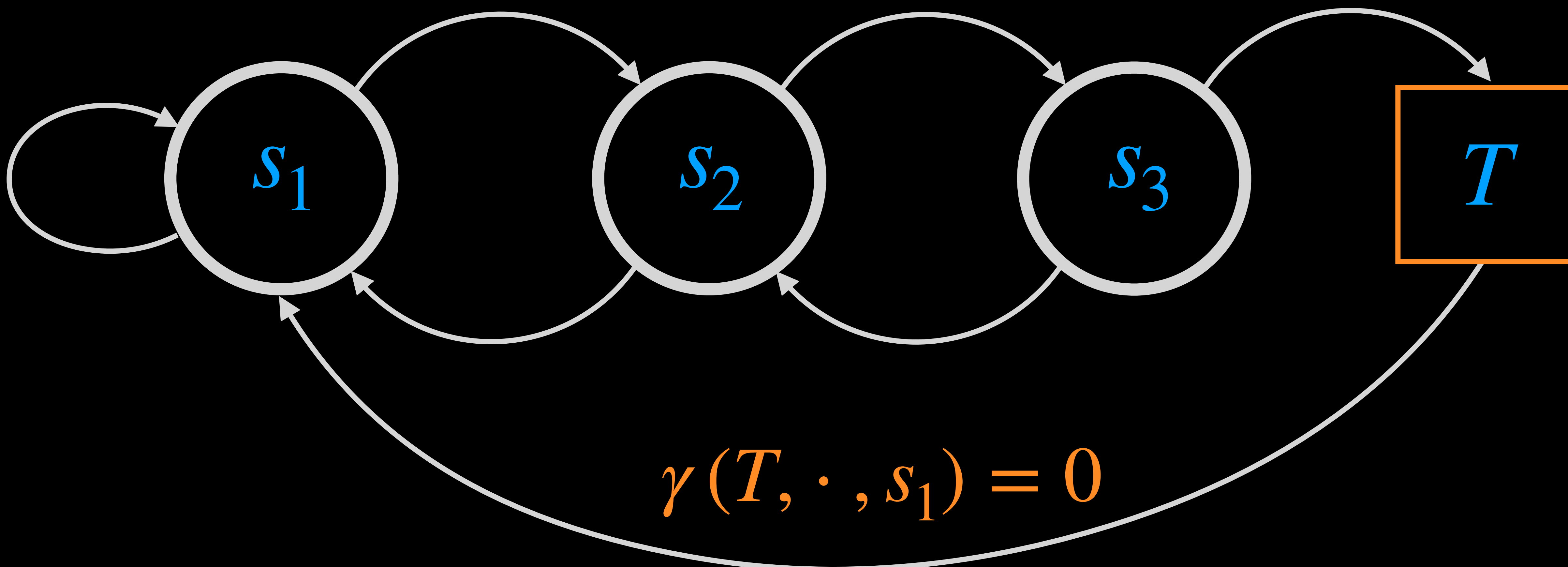


TYPES OF PROBLEMS



White, M. (2017). Unifying task specification in reinforcement learning.

TYPES OF PROBLEMS

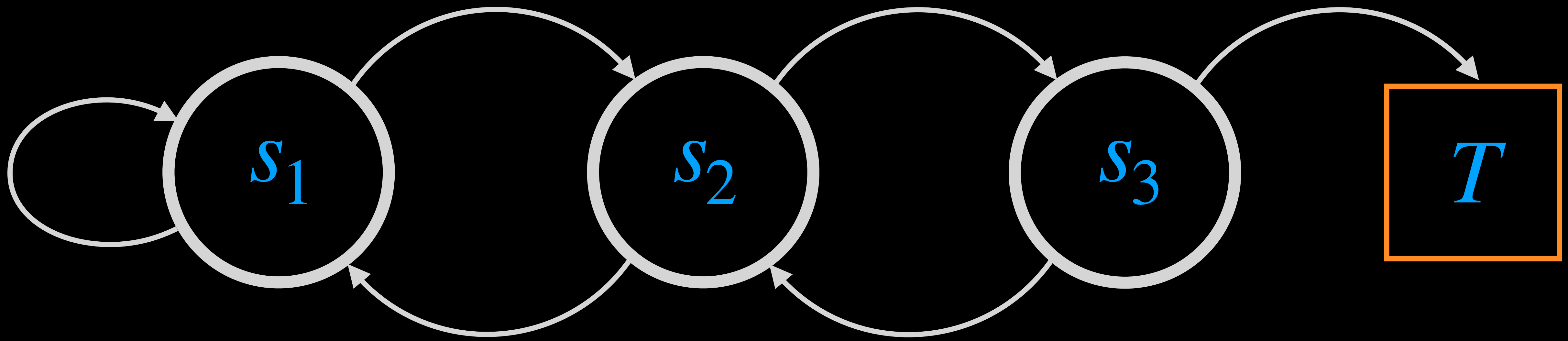


an episodic problem

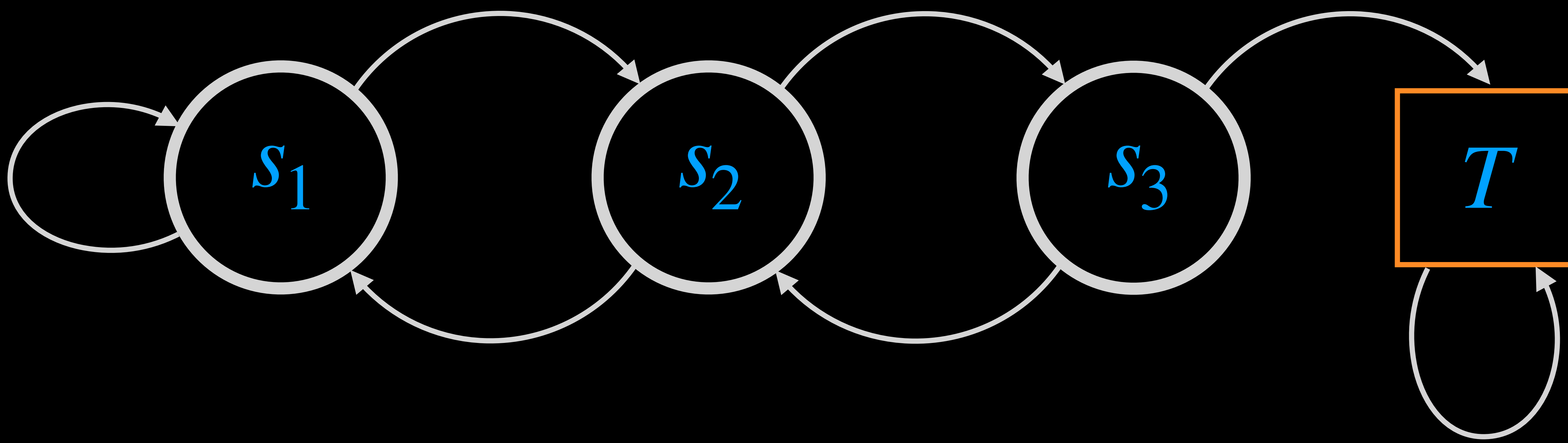
White, M. (2017). Unifying task specification in reinforcement learning.

TYPES OF PROBLEMS

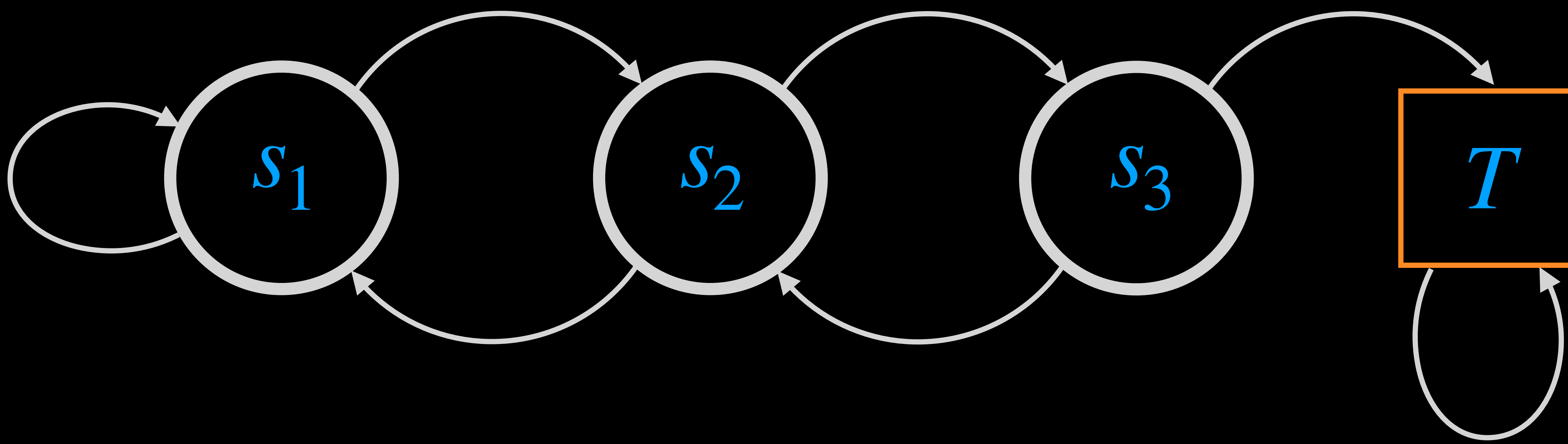
TYPES OF PROBLEMS



TYPES OF PROBLEMS

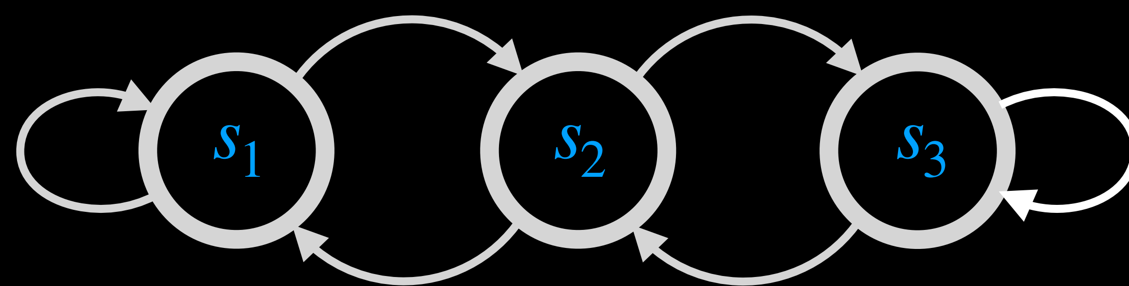


TYPES OF PROBLEMS

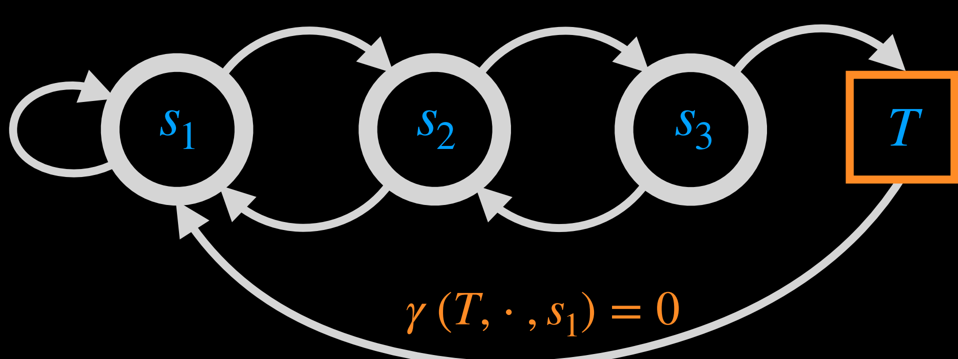


a 'single-life' problem

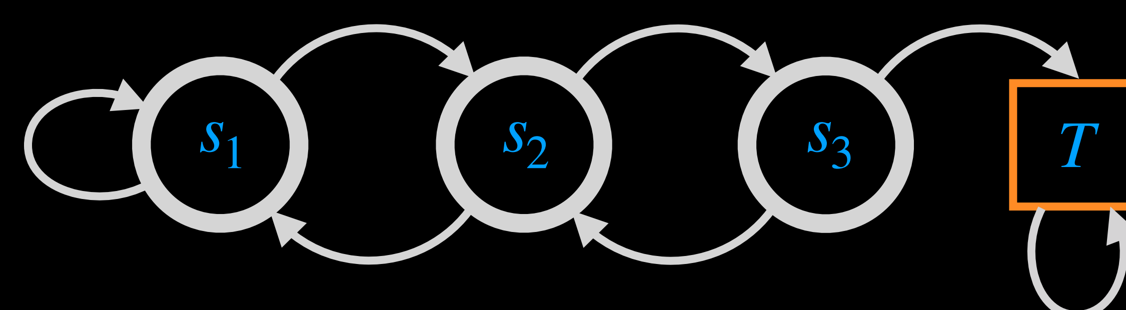
TYPES OF PROBLEMS



continuing

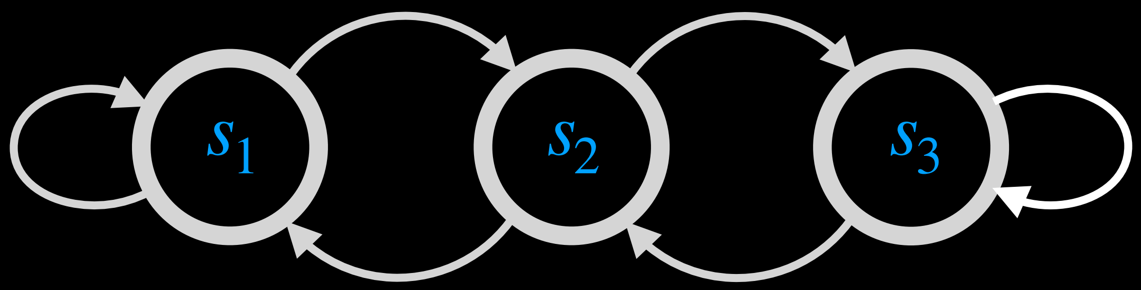


episodic

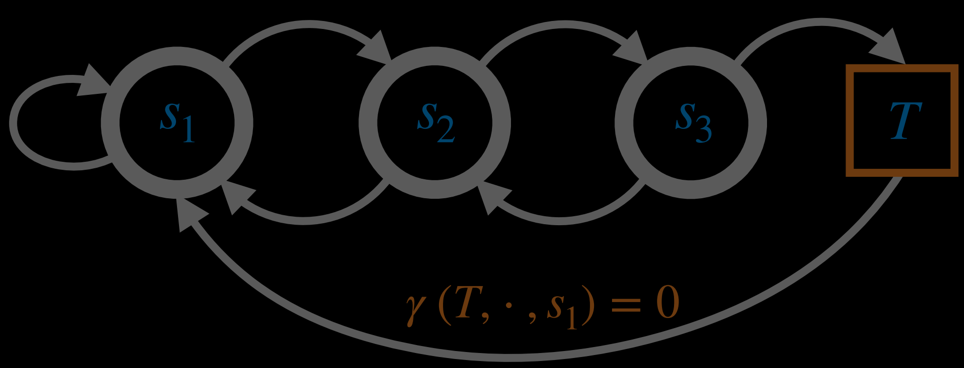


single-life

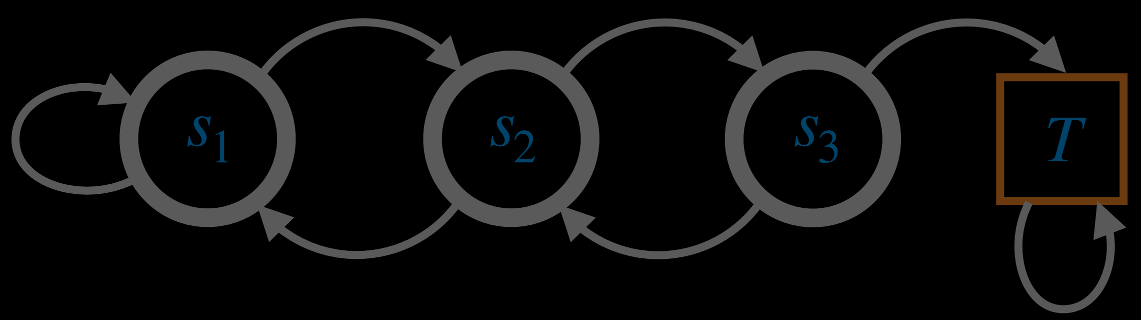
WE FOCUS ON THE CONTINUING SETTING



continuing



episodic



single-life

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:
  emphasizes a learning agent's *continual* need to adapt to a non-stationary world.

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:
  emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
    - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning: emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
  - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.
  - Continuing problems can have non-stationary aspects.

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning: emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
  - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.
  - Continuing problems can have non-stationary aspects.
- Continuous problems:

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:
emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
    - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.
    - Continuing problems can have non-stationary aspects.
- Continuous problems:
have *continuous* state and/or action spaces

IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:
  emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
  - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.
  - Continuing problems can have non-stationary aspects.
- Continuous problems:
  have *continuous* state and/or action spaces
  - Continuing problems can have continuous state/action spaces.

THE STATE OF RESEARCH IN THE CONTINUING SETTING

THE DISCOUNTED FORMULATION

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

  - Formulation widely studied first in the DP literature then RL.

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

  - Formulation widely studied first in the DP literature then RL.
  - Solution methods: SARSA, Q-learning, etc.

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

  - Formulation widely studied first in the DP literature then RL.
  - Solution methods: SARSA, Q-learning, etc.
  - Several TD-based methods exist with theoretical guarantees in the linear function approximation setting (e.g., ETD, GQ)

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

  - Formulation widely studied first in the DP literature then RL.
  - Solution methods: SARSA, Q-learning, etc.
  - Several TD-based methods exist with theoretical guarantees in the linear function approximation setting (e.g., ETD, GQ)
  - Non-linear versions successfully applied in several episodic applications (e.g., DQN on Atari).

THE DISCOUNTED FORMULATION

- Objective: maximize the discounted sum of rewards across states

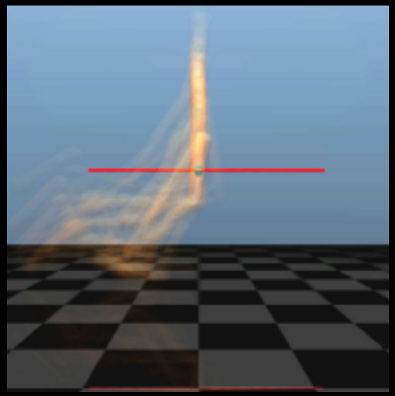
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

  - Formulation widely studied first in the DP literature then RL.
  - Solution methods: SARSA, Q-learning, etc.
  - Several TD-based methods exist with theoretical guarantees in the linear function approximation setting (e.g., ETD, GQ)
  - Non-linear versions successfully applied in several episodic applications (e.g., DQN on Atari).
  - Applications to continuing problems scarce, despite the notion of discounting originally introduced for the continuing setting.

DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS

DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS

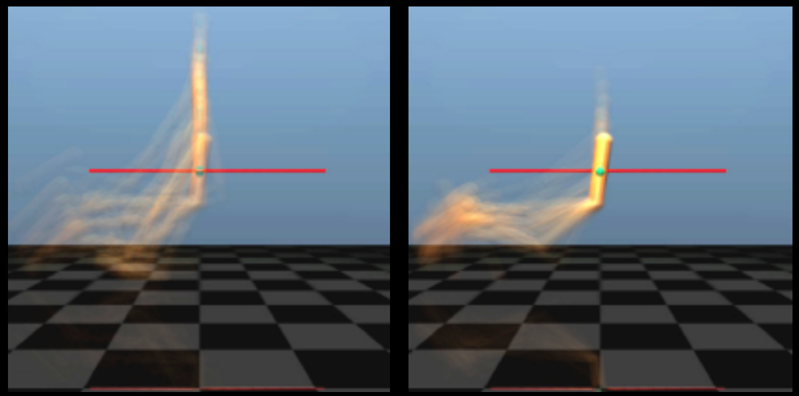
- Pardo et al. (2018) highlighted issues with artificial time limits.



Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.

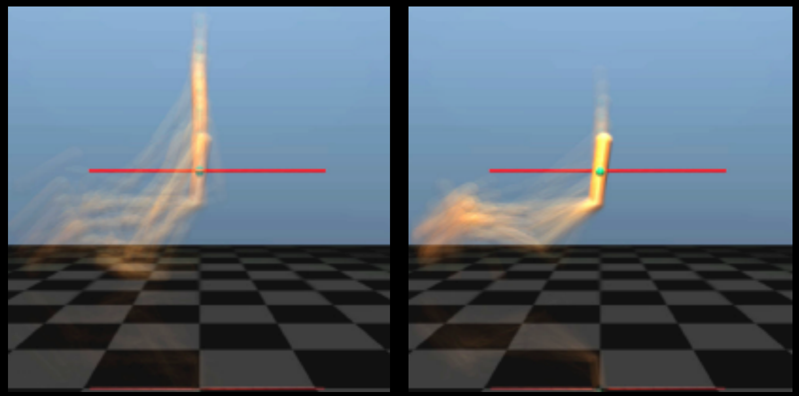
DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS

- Pardo et al. (2018) highlighted issues with artificial time limits.



Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.

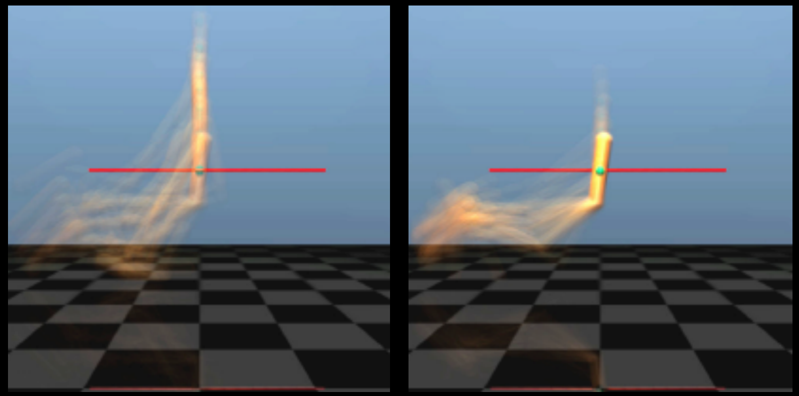
DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS



- Pardo et al. (2018) highlighted issues with artificial time limits.
- Machado et al.'s (2020) results showed resets might be sweeping challenges of exploration under the rug.

Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.
Machado, M. C., Bellemare, M. G., Bowling, M. (2020). Count-based exploration with the successor representation.

DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS



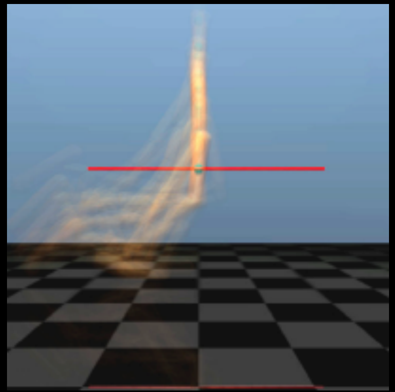
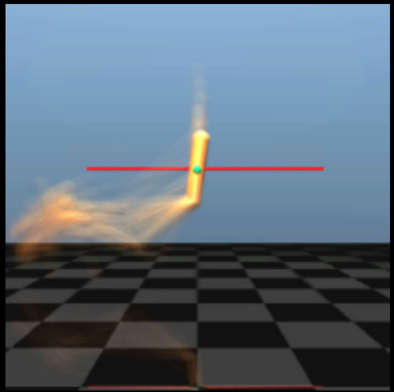
- Pardo et al. (2018) highlighted issues with artificial time limits.

  "… desirable to use time limits in order to frequently reset the environment and increase the diversity of the agent's experiences"

- Machado et al.'s (2020) results showed resets might be sweeping challenges of exploration under the rug.

Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.
Machado, M. C., Bellemare, M. G., Bowling, M. (2020). Count-based exploration with the successor representation.

DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS

- Pardo et al. (2018) highlighted issues with artificial time limits.

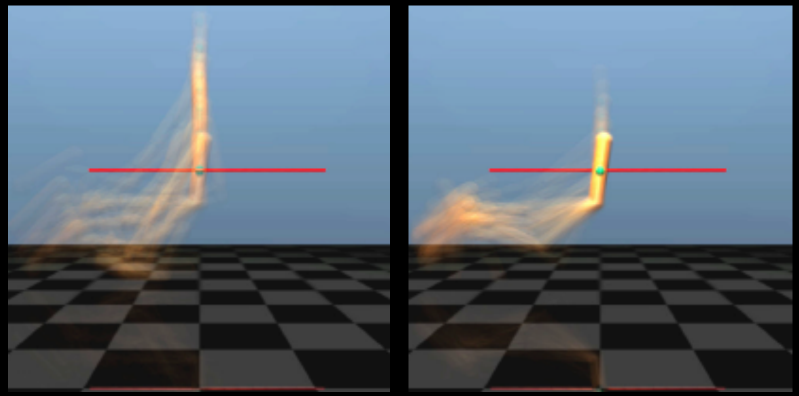
  "… desirable to use time limits in order to frequently reset the environment and increase the diversity of the agent's experiences"

- Machado et al.'s (2020) results showed resets might be sweeping challenges of exploration under the rug.

- Platanios et al. (2020, Case Study #1) found common solution methods like DQN and PPO failed in the Jelly Bean World, a continuing domain.

Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.
Machado, M. C., Bellemare, M. G., Bowling, M. (2020). Count-based exploration with the successor representation.
Platanios, E. A., Saparov, A., Mitchell, T. (2020). Jelly Bean World: A Testbed for Never-Ending Learning.

DISCOUNTED METHODS FOR EPISODIC PROBLEMS DON'T REALLY APPLY AS IS IN CONTINUING PROBLEMS



- Pardo et al. (2018) highlighted issues with artificial time limits.

  "… desirable to use time limits in order to frequently reset the environment and increase the diversity of the agent's experiences"

- Machado et al.'s (2020) results showed resets might be sweeping challenges of exploration under the rug.

- Platanios et al. (2020, Case Study #1) found common solution methods like DQN and PPO failed in the Jelly Bean World, a continuing domain.

- Sutton and Barto (2018, Ch 10) and Naik et al. (2019) claimed that discounting is incompatible with continuing control with function approximation.

Pardo, F., Tavakoli, A., Levdik, V., Kormushev, P. (2018). Time limits in reinforcement learning.
Machado, M. C., Bellemare, M. G., Bowling, M. (2020). Count-based exploration with the successor representation.
Platanios, E. A., Saparov, A., Mitchell, T. (2020). Jelly Bean World: A Testbed for Never-Ending Learning.
Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction.
Naik, A., Shariff, R., Yasui, N., Sutton, R. S. (2019). Discounted Reinforcement Learning Is Not an Optimization Problem.

THE AVERAGE-REWARD FORMULATION

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

$$r(\pi) \doteq \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[R_t \,|\, S_0, A_{0:t-1} \sim \pi\right]$$

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[R_t \,|\, S_0, A_{0:t-1} \sim \pi\right]$$

- Formulation widely studied in the DP literature, not much in RL.

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[R_t \mid S_0, A_{0:t-1} \sim \pi\right]$$

  - Formulation widely studied in the DP literature, not much in RL.
  - Solution methods: RVI Q-learning, Differential Q-learning, etc.

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

$$r(\pi) \doteq \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[R_t \,|\, S_0, A_{0:t-1} \sim \pi\right]$$

  - Formulation widely studied in the DP literature, not much in RL.
  - Solution methods: RVI Q-learning, Differential Q-learning, etc.
  - Most theoretical results in the function approximation setting are restricted to the *prediction* problem.

THE AVERAGE-REWARD FORMULATION

- Objective: maximize the average reward over time

$$r(\pi) \doteq \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\left[R_t \mid S_0, A_{0:t-1} \sim \pi\right]$$

- Formulation widely studied in the DP literature, not much in RL.
- Solution methods: RVI Q-learning, Differential Q-learning, etc.
- Most theoretical results in the function approximation setting are restricted to the *prediction* problem.
- Not much empirical experience, especially with large-scale problems.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

CONTINUING PROBLEMS ARE A STEPPING STONE TO AI

- We want to build general-purpose intelligent systems that learn to perform various tasks throughout their lifetimes via a single stream of experience.

THE CONTINUING SETTING IS UNDER-STUDIED

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
    - OpenAI Gym

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
    - OpenAI Gym
    - MuJoCo

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.

- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
  - OpenAI Gym
  - MuJoCo
  - DM control suite

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
    - OpenAI Gym
    - MuJoCo
    - DM control suite
    - Behavioural suite (bsuite)

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
    - OpenAI Gym
    - MuJoCo
    - DM control suite
    - Behavioural suite (bsuite)
    - Arcade Learning Environment (ALE)

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
  - OpenAI Gym
  - MuJoCo
  - DM control suite
  - Behavioural suite (bsuite)
  - Arcade Learning Environment (ALE)
  - PyGame Learning Environment (PLE)

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
    - OpenAI Gym
    - MuJoCo
    - DM control suite
    - Behavioural suite (bsuite)
    - Arcade Learning Environment (ALE)
    - PyGame Learning Environment (PLE)

no continuing problems

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
  - OpenAI Gym
  - MuJoCo
  - DM control suite
  - Behavioural suite (bsuite)
  - Arcade Learning Environment (ALE)
  - PyGame Learning Environment (PLE)

no continuing problems

no continuing research

THE CONTINUING SETTING IS UNDER-STUDIED

- …mainly due to lack of continuing problems in the literature.
- Most problem suites have episodic problems, or naturally-continuing problems that are made episodic via timeouts.
  - OpenAI Gym
  - MuJoCo
  - DM control suite
  - Behavioural suite (bsuite)
  - Arcade Learning Environment (ALE)
  - PyGame Learning Environment (PLE)

no continuing problems

no continuing research

C-SUITE

TYPES OF PROBLEMS IN C-SUITE

TYPES OF PROBLEMS IN C-SUITE

- Two broad categories:

TYPES OF PROBLEMS IN C-SUITE

- Two broad categories:

    1. Continuing problems from the literature
    2. New problems (inspired from the real world)

TYPES OF PROBLEMS IN C-SUITE

- Two broad categories:

  1. Continuing problems from the literature
  2. New problems (inspired from the real world)

- Can also classify the problems as:

TYPES OF PROBLEMS IN C-SUITE

- Two broad categories:

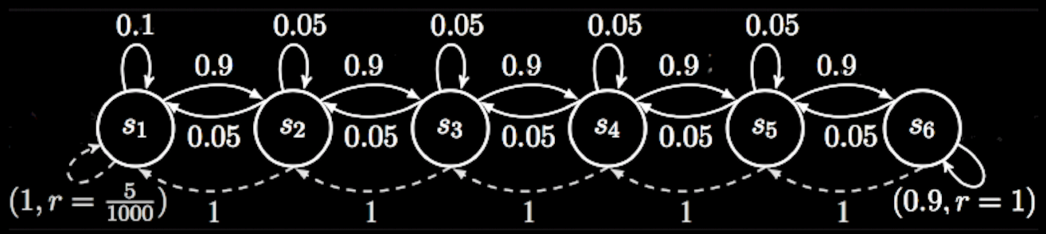
  1. Continuing problems from the literature
  2. New problems (inspired from the real world)

- Can also classify the problems as:
  1. Small-scale pedagogical problems
  2. Large-scale challenge problems

1. CONTINUING PROBLEMS FROM THE LITERATURE

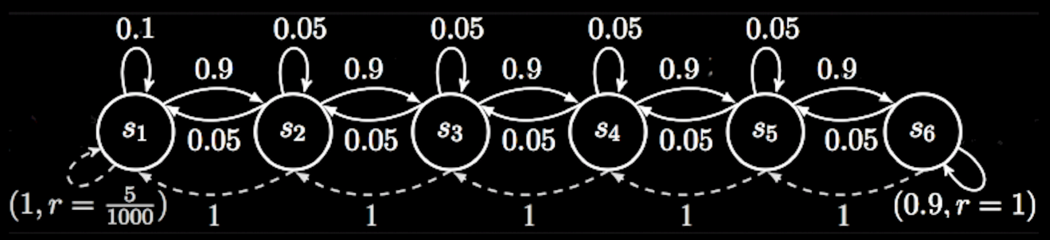
1. CONTINUING PROBLEMS FROM THE LITERATURE

- MDPs such as RiverSwim

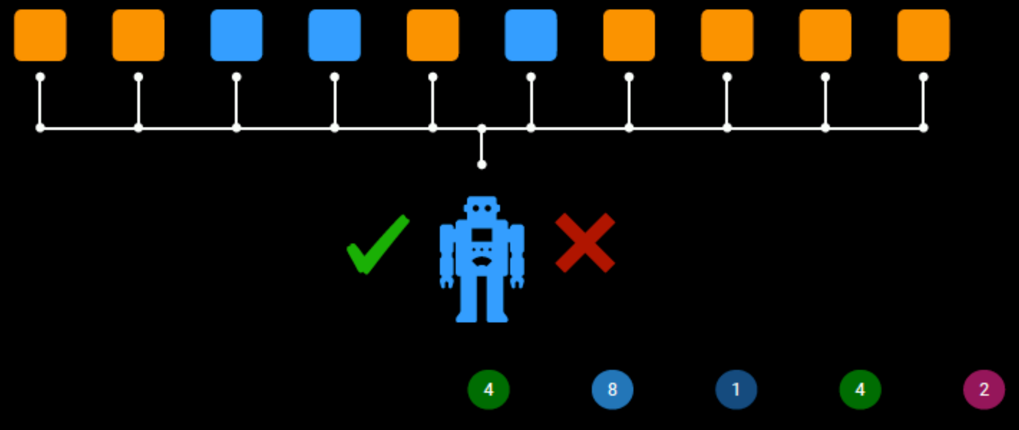


1. CONTINUING PROBLEMS FROM THE LITERATURE

- MDPs such as RiverSwim

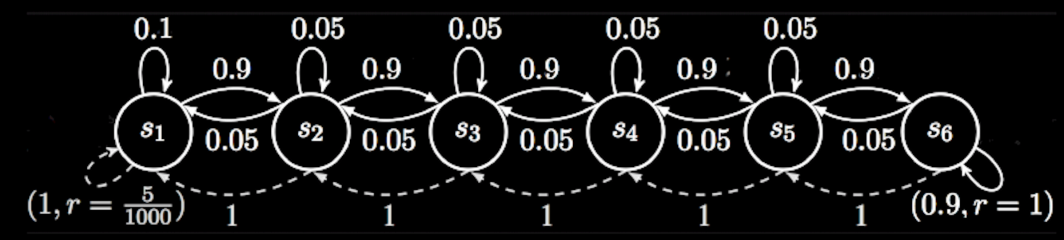


- Other tabular problems such as the Access-Control Queueing Task

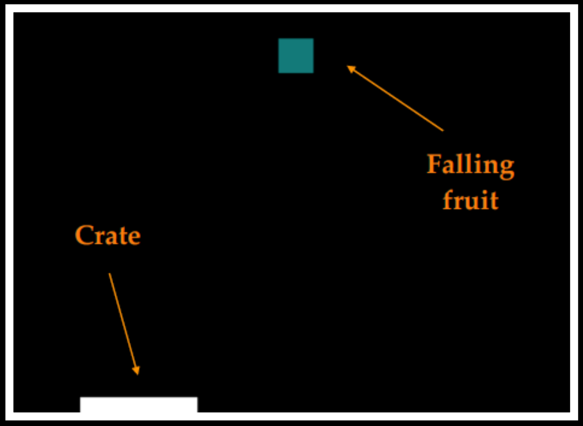


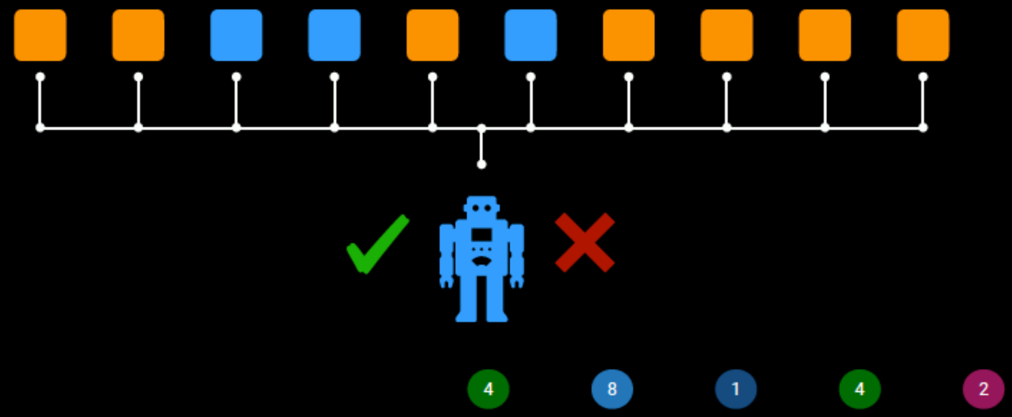
1. CONTINUING PROBLEMS FROM THE LITERATURE

- MDPs such as RiverSwim



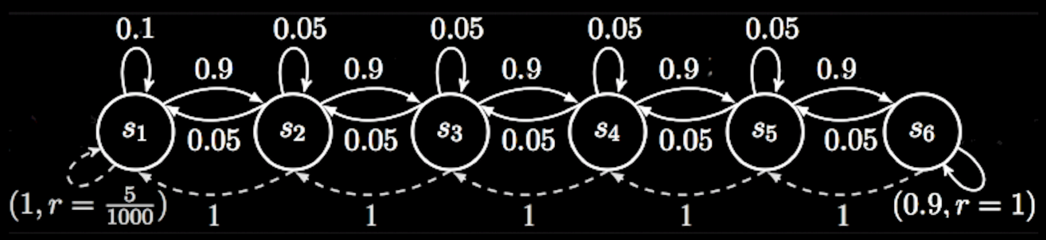
- Other tabular problems such as the Access-Control Queueing Task
- Games such as PuckWorld, Catcher



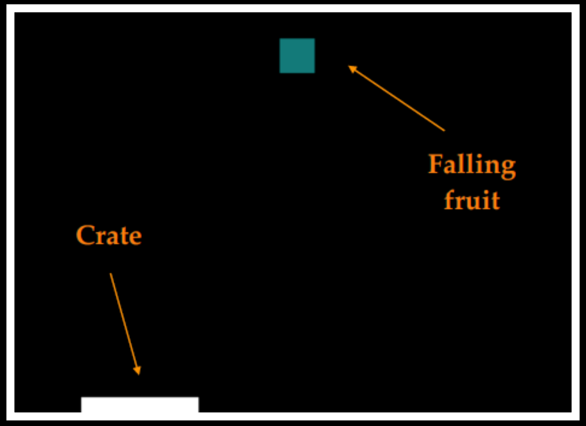


1. CONTINUING PROBLEMS FROM THE LITERATURE

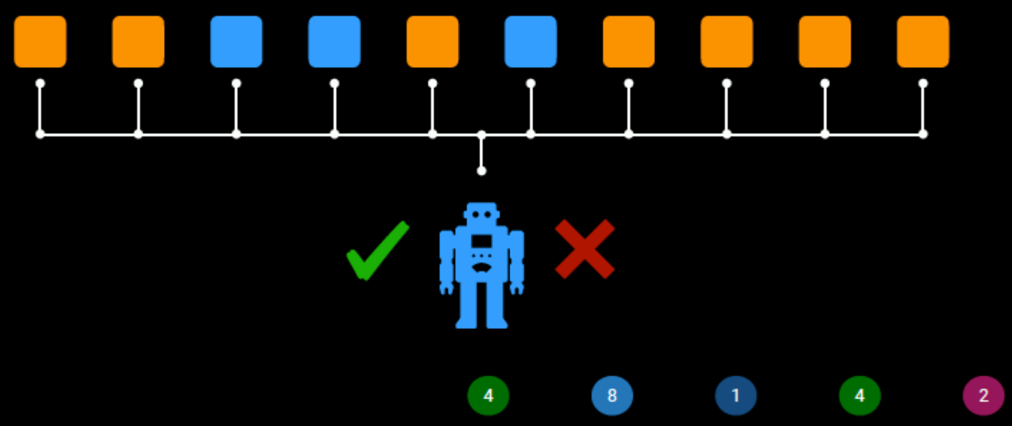
- MDPs such as RiverSwim

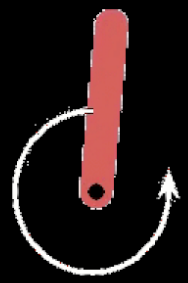


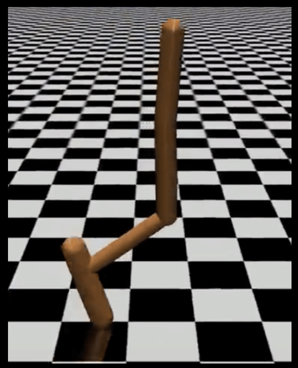
- Other tabular problems such as the Access-Control Queueing Task
- Games such as PuckWorld, Catcher

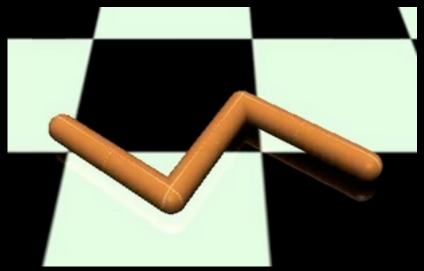


- Classic control tasks such as Pendulum, Acrobot, Walker, Hopper, Reacher, Swimmer





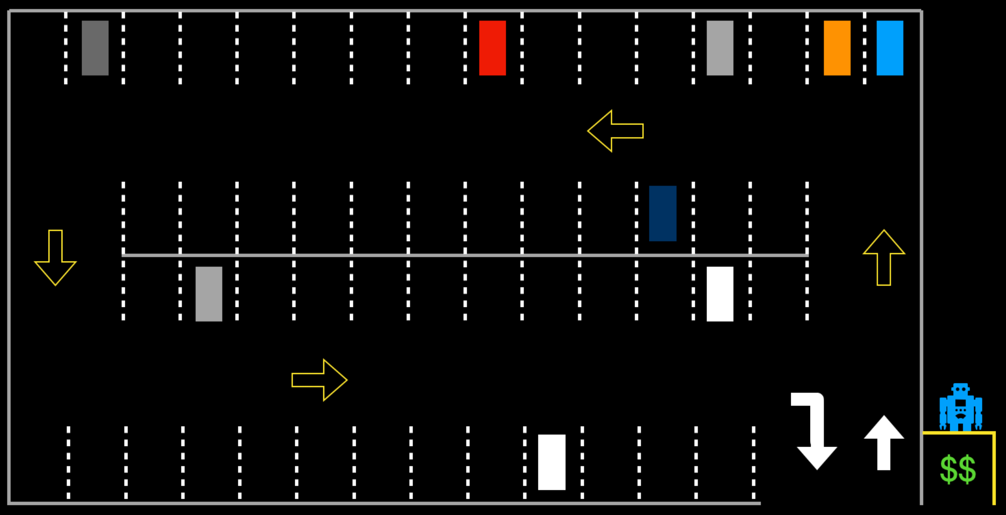




2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

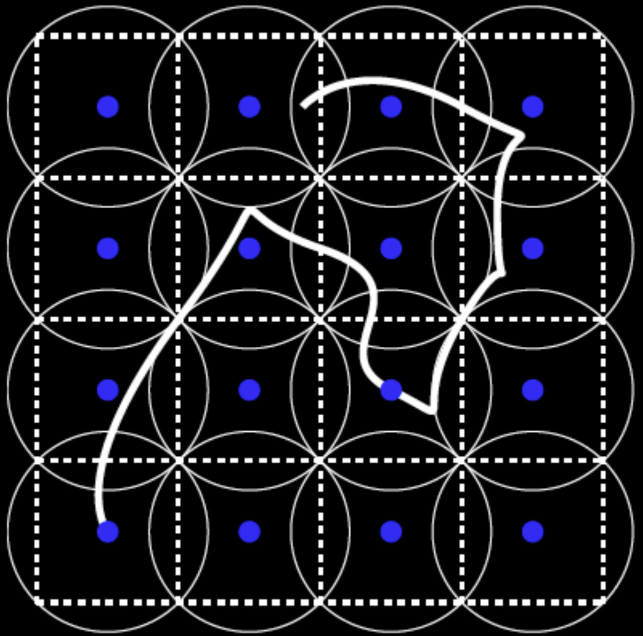
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

- Adaptive parking pricing

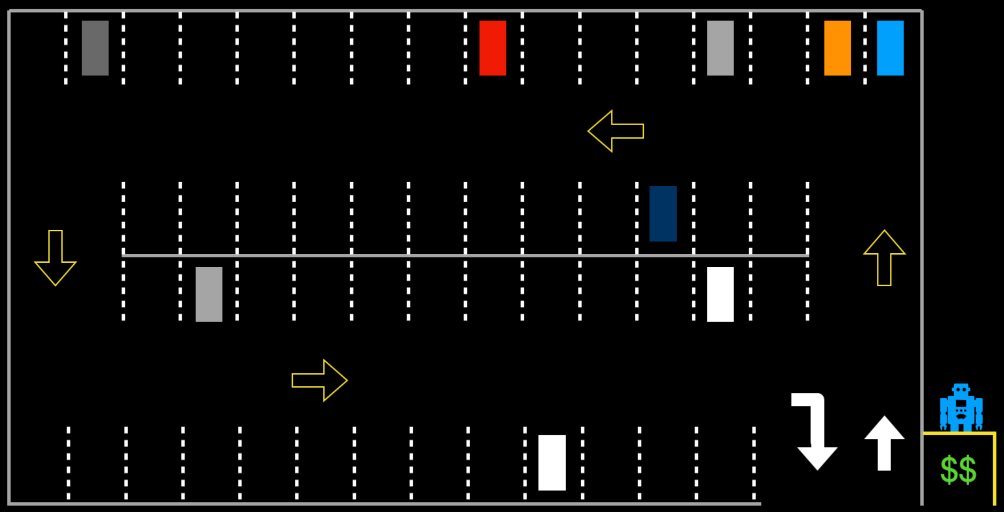


2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

- Adaptive parking pricing
- Intruder detection

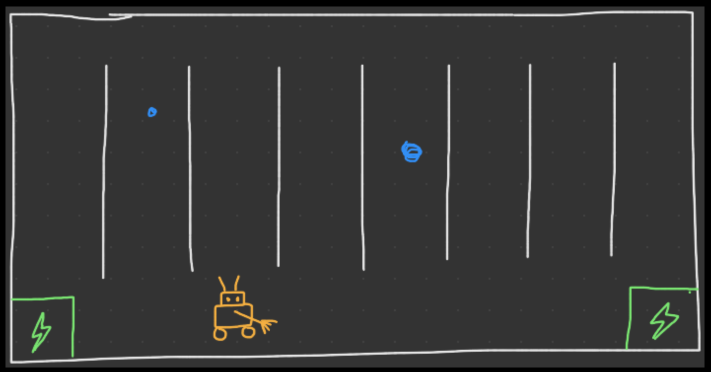


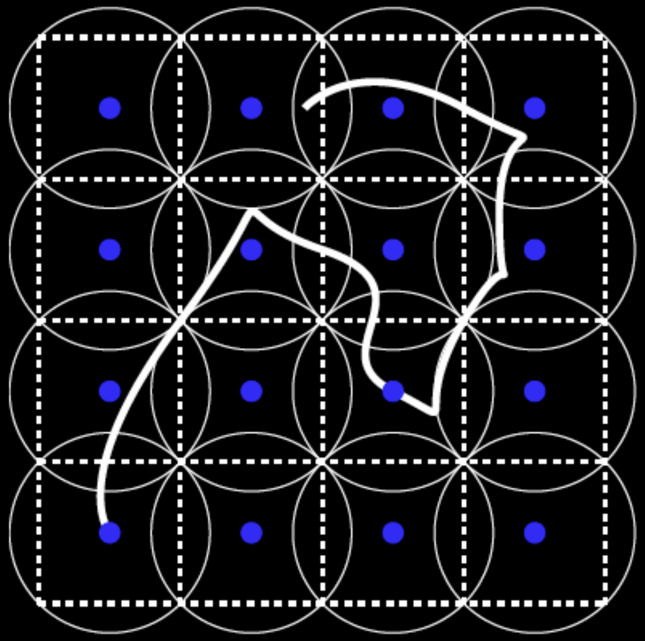
Prashanth et al. (2014)



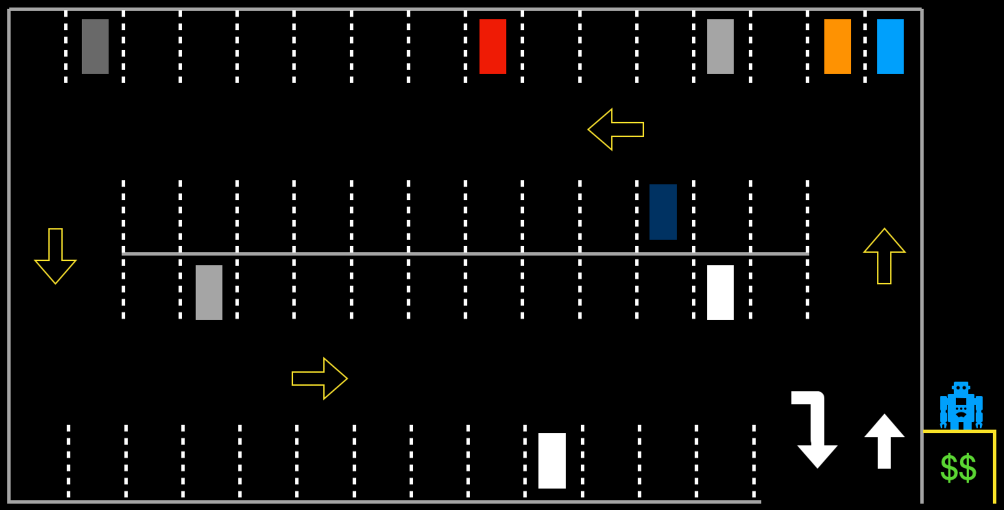
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

- Adaptive parking pricing
- Intruder detection
- Aisle clean-up



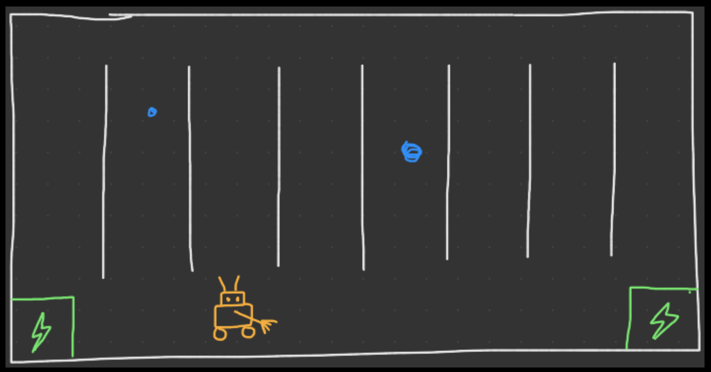


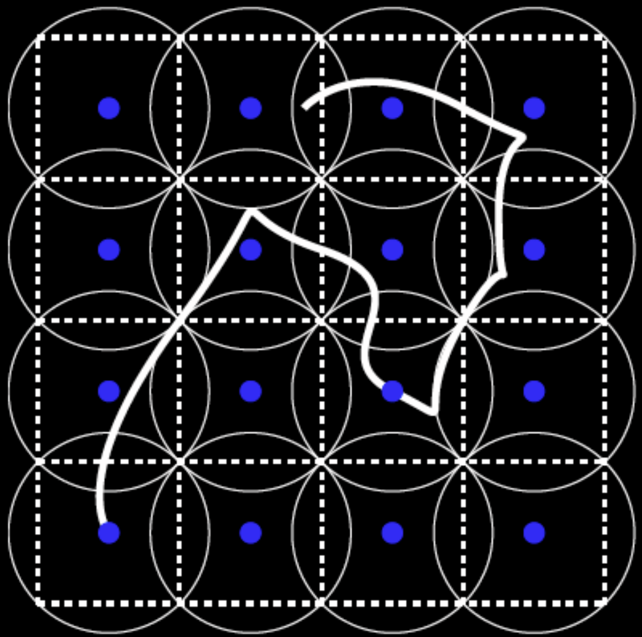
Prashanth et al. (2014)



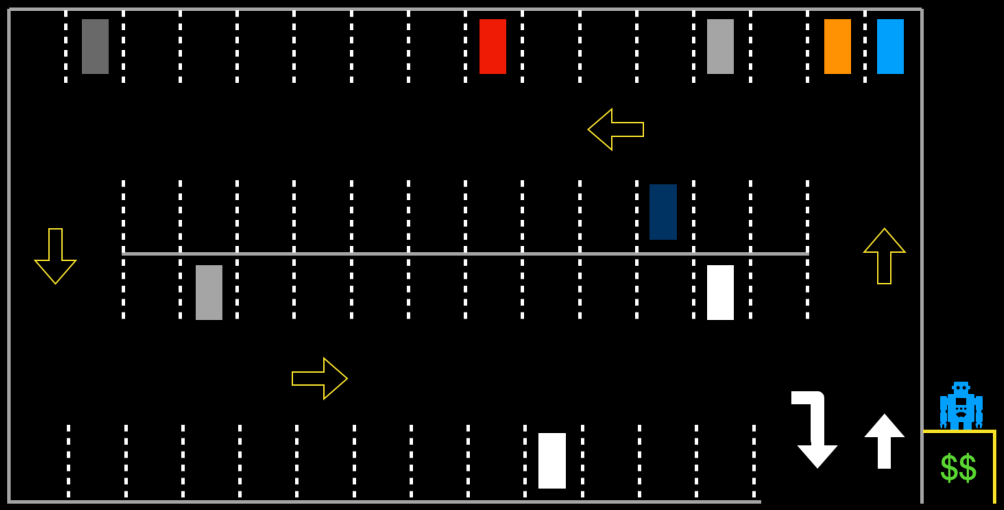
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

- Adaptive parking pricing
- Intruder detection
- Aisle clean-up
- Cache management / Inventory control



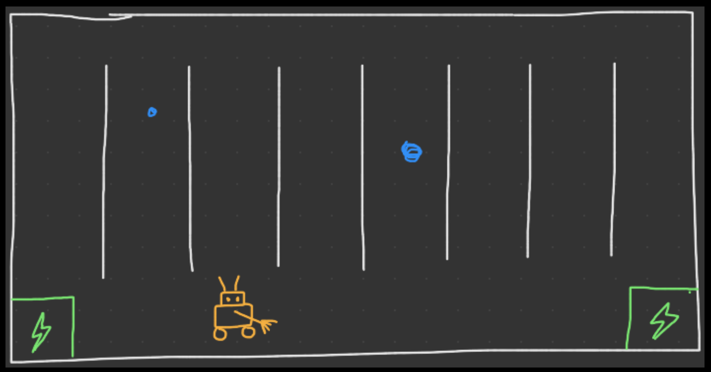


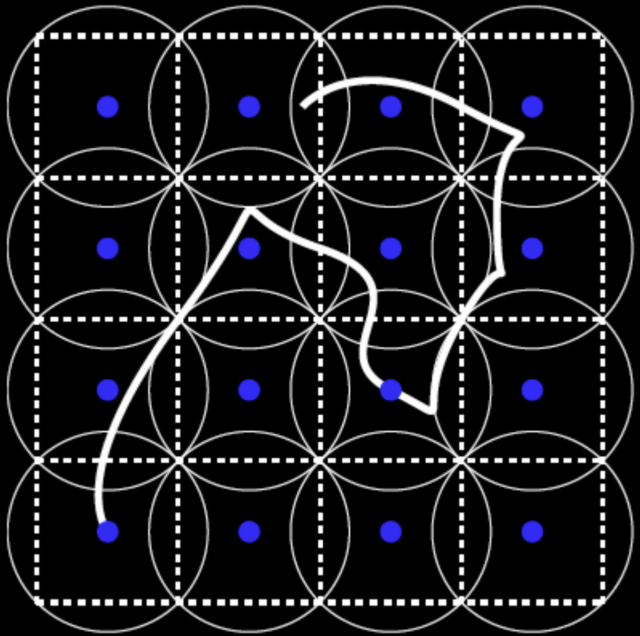
Prashanth et al. (2014)



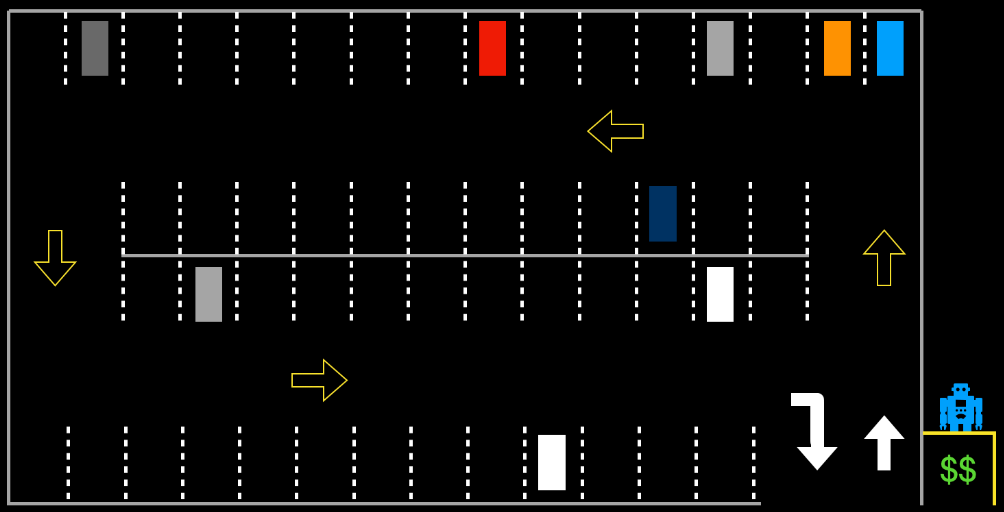
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

- Adaptive parking pricing
- Intruder detection
- Aisle clean-up
- Cache management / Inventory control
- Job/packet scheduling

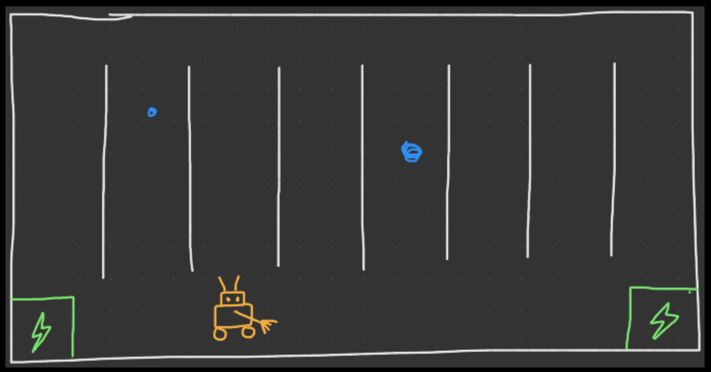


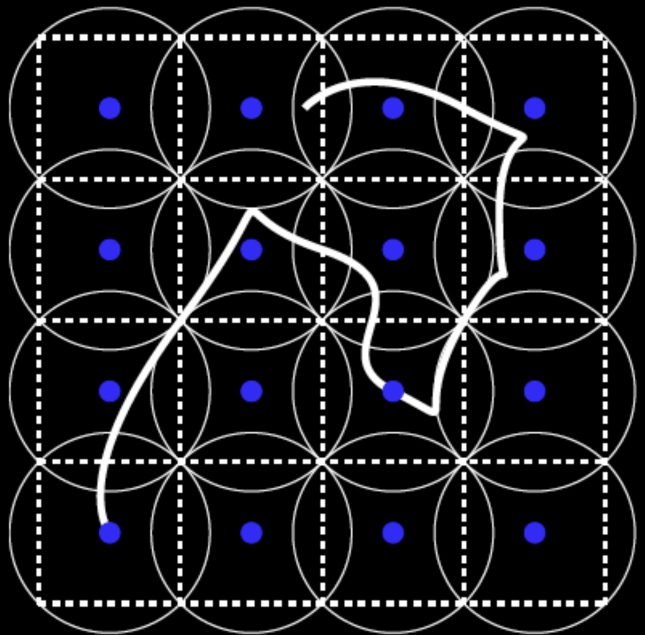


Prashanth et al. (2014)



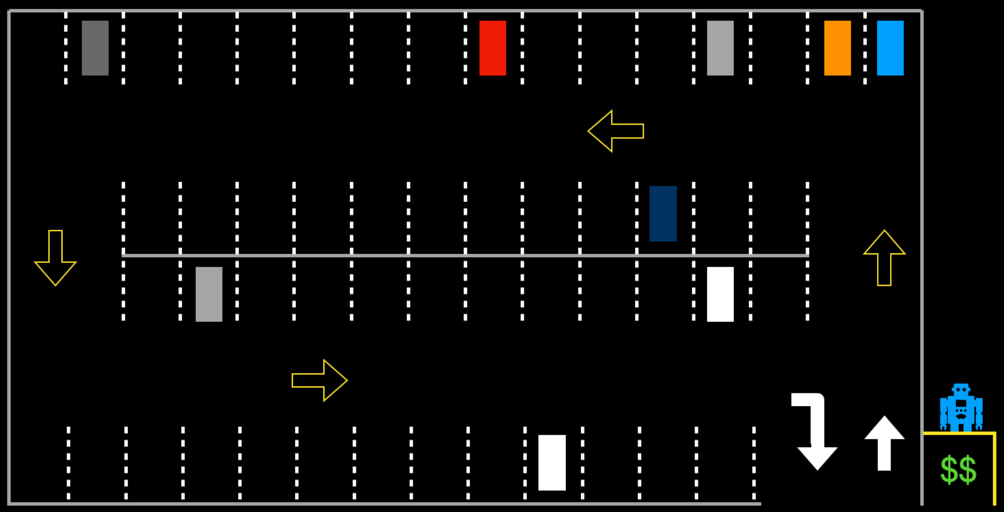
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)



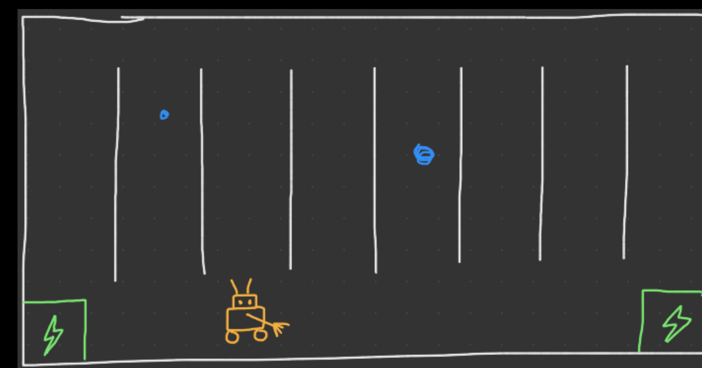


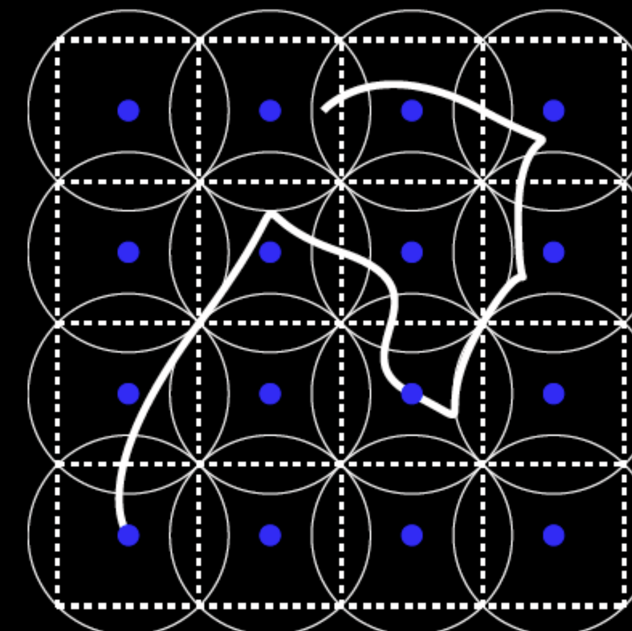
Prashanth et al. (2014)

- Adaptive parking pricing
- Intruder detection
- Aisle clean-up
- Cache management / Inventory control
- Job/packet scheduling
- Local temp/humidity control



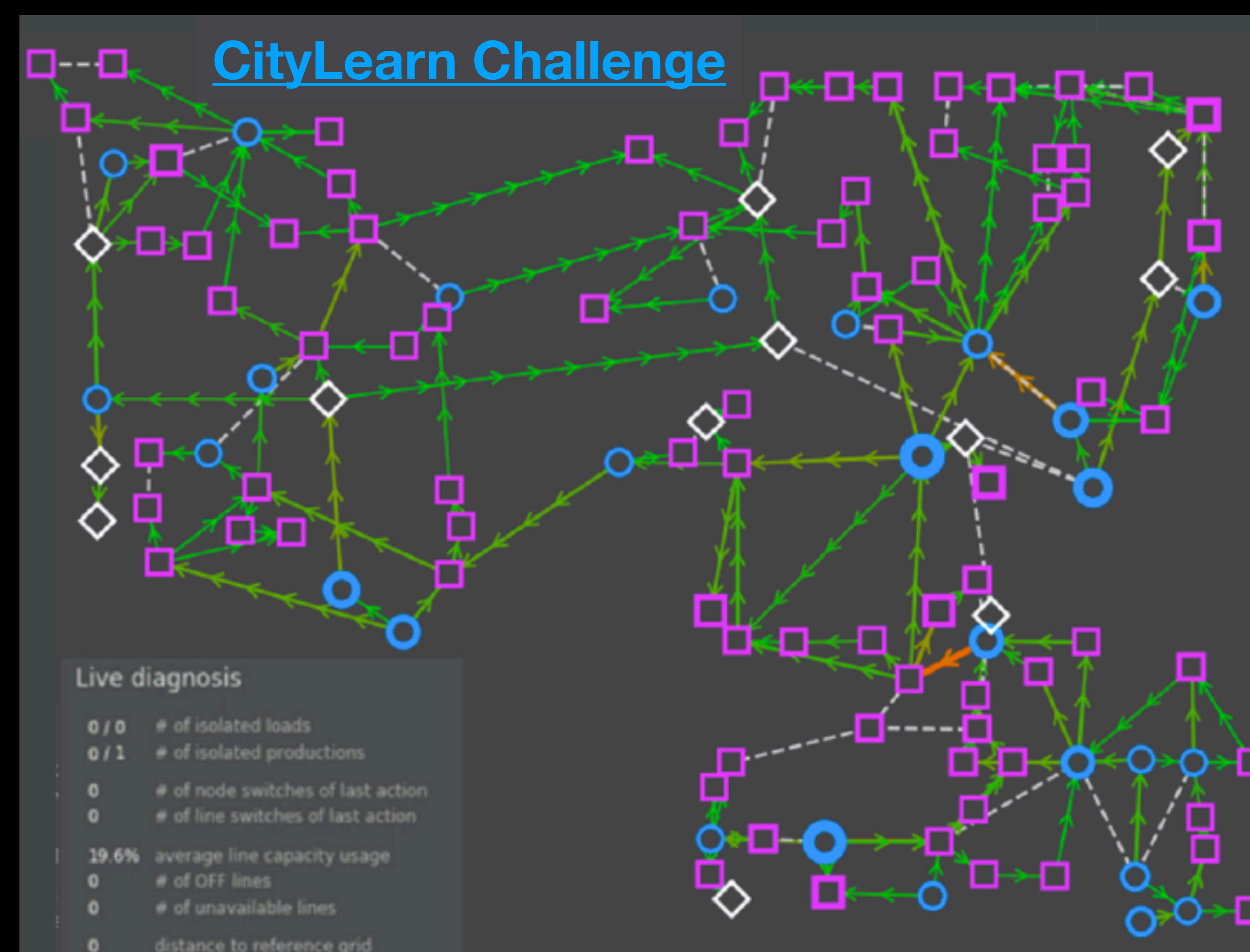
2. NEW PROBLEMS (INSPIRED FROM THE REAL WORLD)

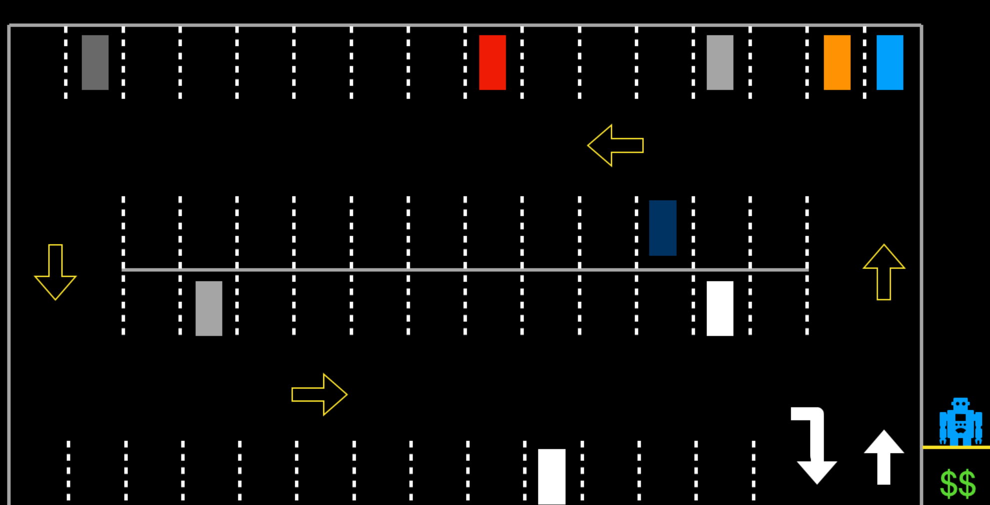




Prashanth et al. (2014)

- Adaptive parking pricing
- Intruder detection
- Aisle clean-up
- Cache management / Inventory control
- Job/packet scheduling
- Local temp/humidity control
- Powergrid management





C-SUITE IS A MEANS TO AN END

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

- Compare the discounted and average-reward formulations empirically

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

- Compare the discounted and average-reward formulations empirically
- Get more experience with average-reward methods, especially in the function approximation setting

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

- Compare the discounted and average-reward formulations empirically
- Get more experience with average-reward methods, especially in the function approximation setting
- Come with up new/better solution methods for continuing problems

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

- Compare the discounted and average-reward formulations empirically
- Get more experience with average-reward methods, especially in the function approximation setting
- Come with up new/better solution methods for continuing problems
- More generally, explore connections between the two formulations and perhaps leverage the best of both worlds

C-SUITE IS A MEANS TO AN END

The end: to study problem formulations and solution methods for continuing problems, enroute to AI.

- Compare the discounted and average-reward formulations empirically
- Get more experience with average-reward methods, especially in the function approximation setting
- Come with up new/better solution methods for continuing problems
- More generally, explore connections between the two formulations and perhaps leverage the best of both worlds
- Add more pedagogical as well as challenge problems to C-suite

THANK YOU

Join the effort! Contact:

abhishek.naik@ualberta.ca
zaheersm@google.com
adamwhite@google.com
rsutton@ualberta.ca