Reinforcement Learning

2.3 Introducción al Aprendizaje Reforzado

Jorge Vasquez

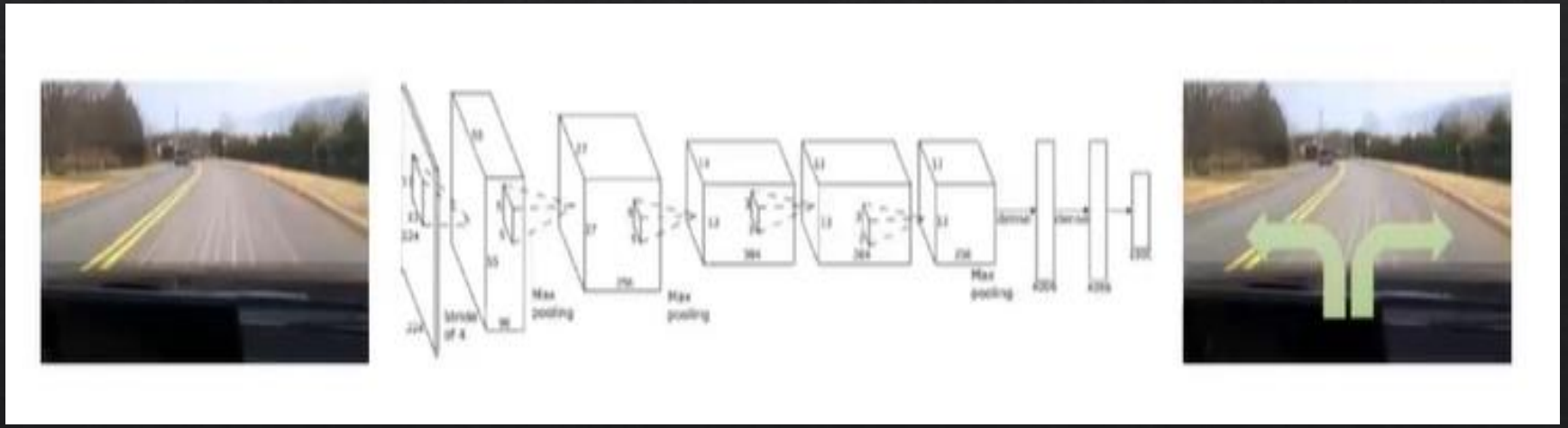
*La mayoría de las slides las tome de la clase de DRL en Berkeley, Abbel

Esta Clase

- Deep Reinforcement Learning
- Procesos de Decisión de Markov (MDP)
- Métodos de Soluciones Exactas (Iteración de valor)

Deep Reinforcement Learning

- Cuando la política, el modelo o las funciones de valor (rewards esperados) son representados como redes neuronales profundas (DL).



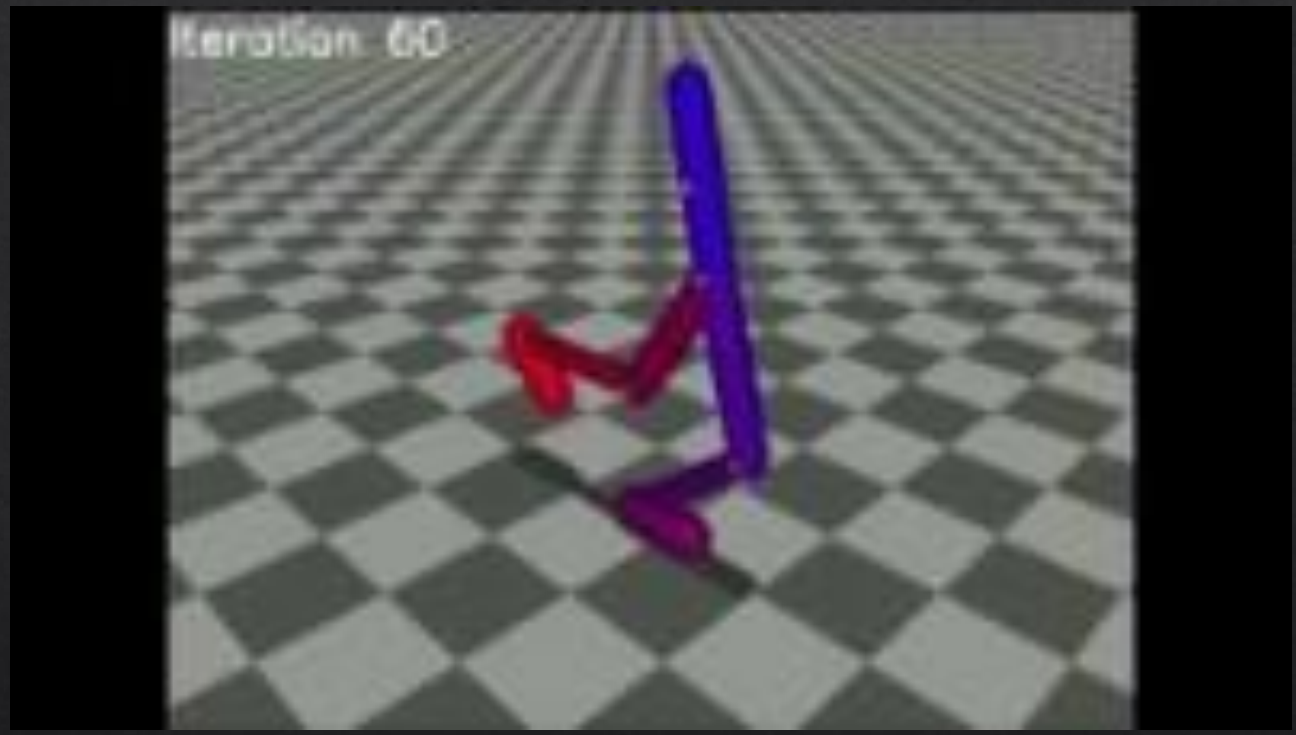
Logros de Deep Reinforcement Learning

2013 – Atari (DeepMind)

Logros de Deep Reinforcement Learning

2013 – Atari (DeepMind)

2014 – 2D Locomotion



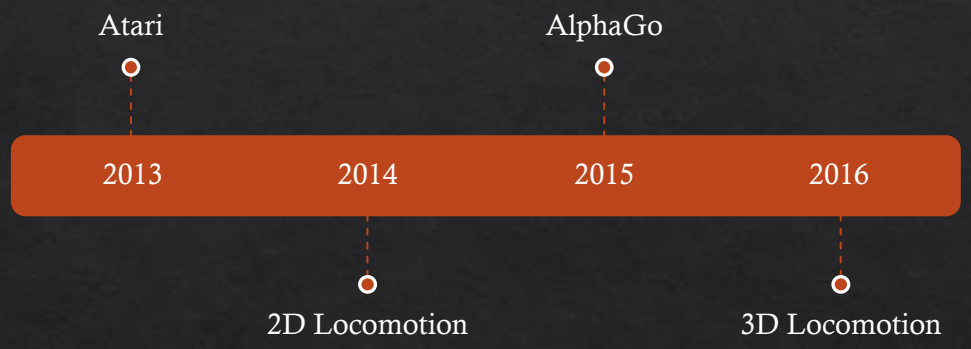
Logros de Deep Reinforcement Learning

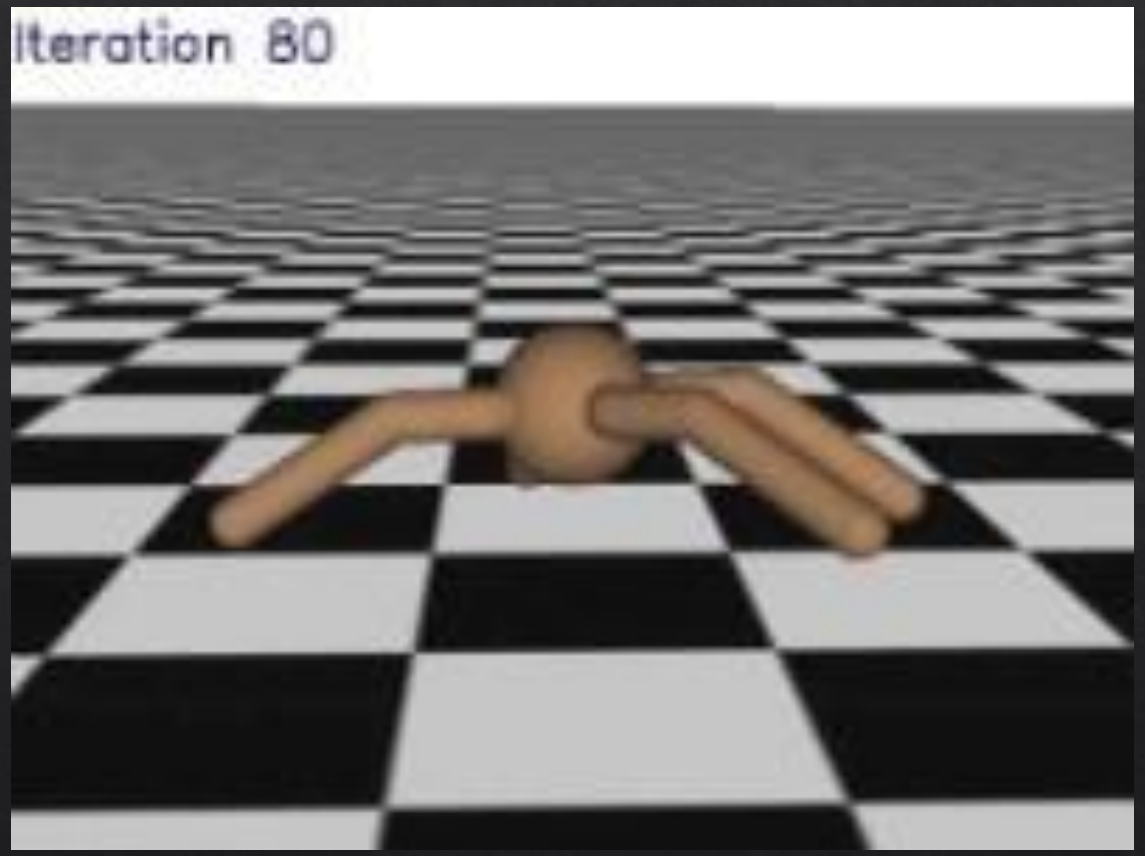
2013 – Atari (DeepMind)

2014 – 2D Locomotion

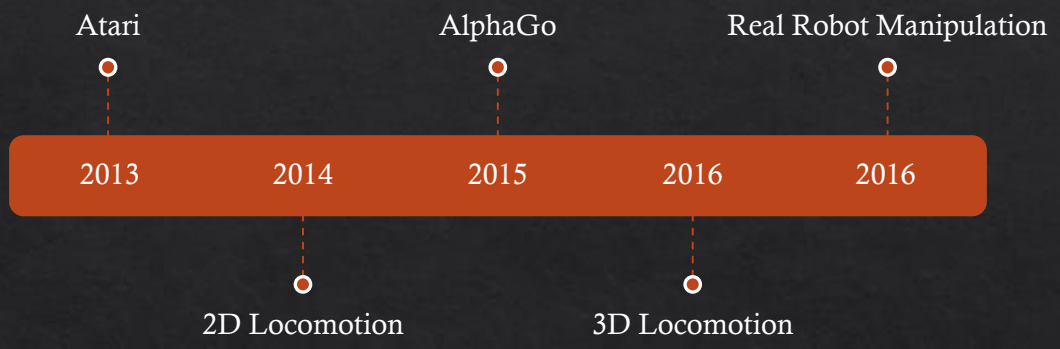
2015 – AlphaGo (DeepMind)

Logros de Deep Reinforcement Learning

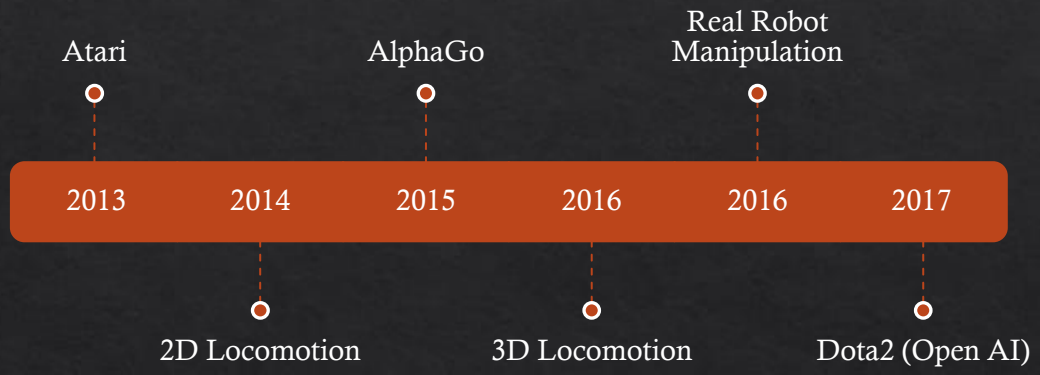




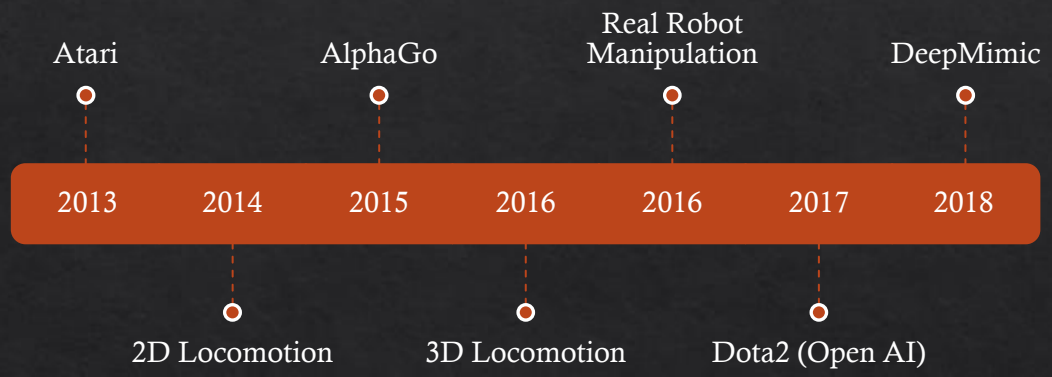
Logros de Deep Reinforcement Learning



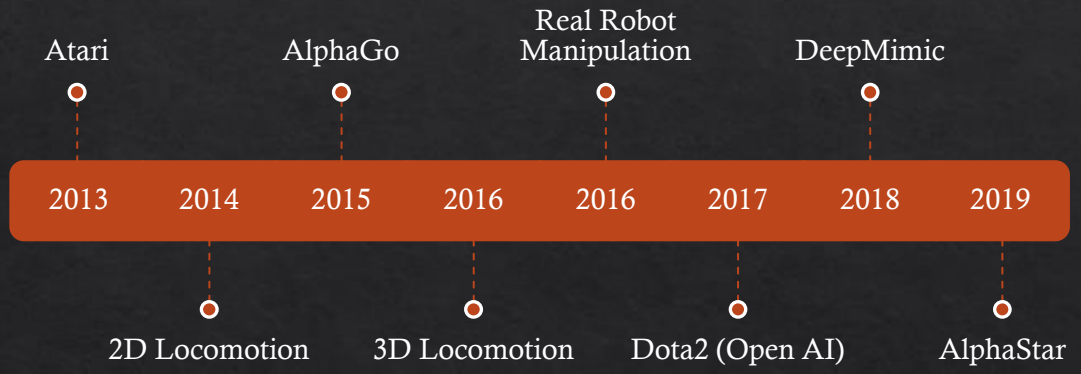
Logros de Deep Reinforcement Learning



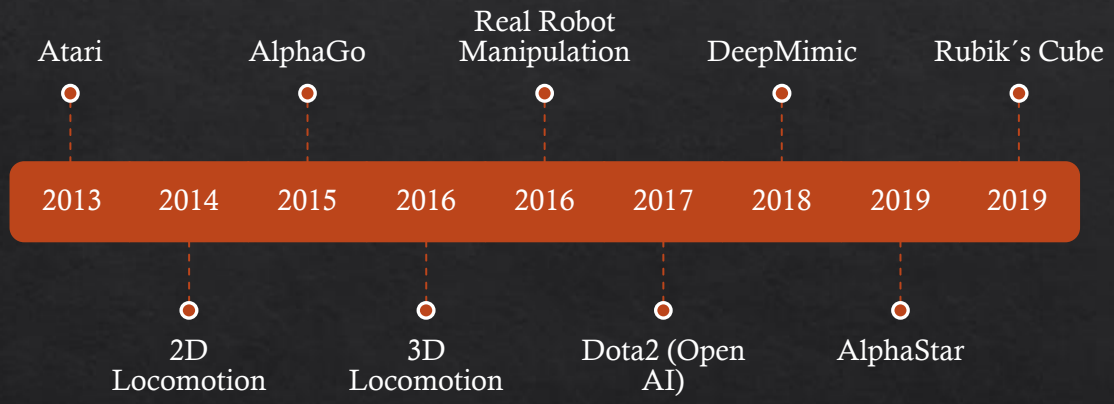
Logros de Deep Reinforcement Learning



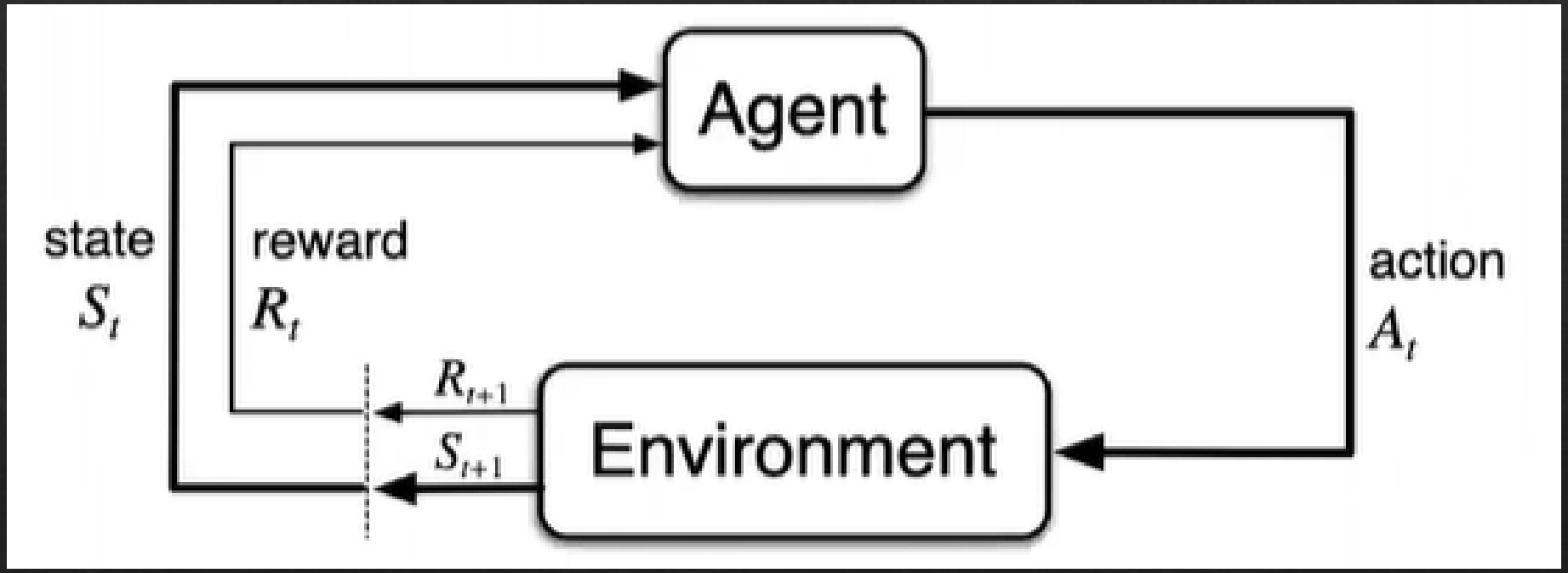
Logros de Deep Reinforcement Learning



Logros de Deep Reinforcement Learning



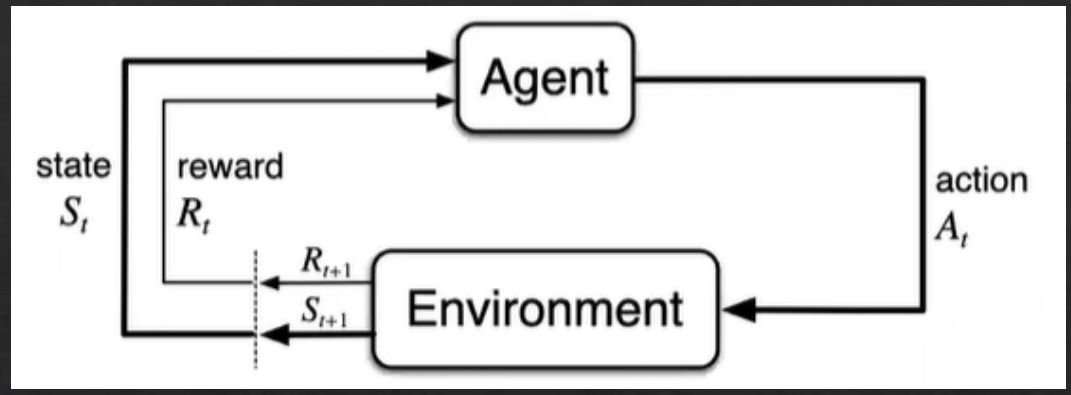
Marco de Trabajo: Proceso de Decisiones de Markov



Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

Un MDP está definido por:

- Set de Estados S
- Set de acciones A
- Función de Transición P(s´ | s, a)
- Función de Refuerzo R (s, a, s')
- Estado inicial So
- Factor descuento γ
- Horizonte H



Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

Estados Markovianos

- Un estado captura cualquier tipo de información disponible para el agente en el periodo de tiempo t sobre su entorno
- El estado puede incluir sensaciones inmediatas, sensaciones muy procesadas, estructuras construidas a través del tiempo, memorias, etc..
- Un estado debe resumir sensaciones pasadas y retener las esenciales

Recompensas refeljan objetivos

- Recompensas son valores escalares entregados por el entorno al agente, que le indican si el obj ha sido logrado.
- Los objetivos especifican lo que el agente debe lograr, no el como.
- Maximización de una suma acumulativa de recompensas

$$r(s,a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

El agente aprende una política

- La política es una distribución de probabilidades sobre acciones dado estados.

$$\pi(a \mid s) = \mathbf{Pr}(A_t = a \mid S_t = s), \forall t$$

- Una política define totalmente el comportamiento de un agente
- La política es estacionaria (no depende del tiempo)
- Durante el entrenamiento, el agente cambia su política como resultado de la experiencia

Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Robot Aspiradora

Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

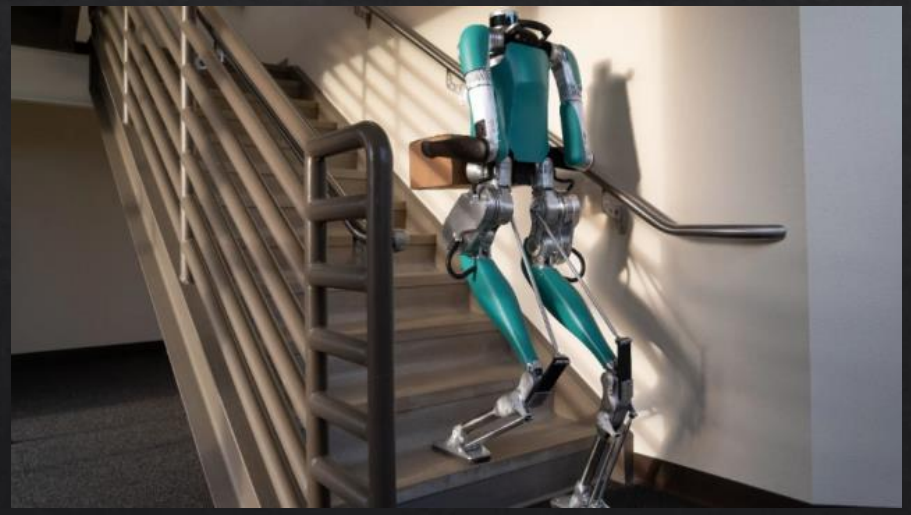
Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Robot que camina

Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$



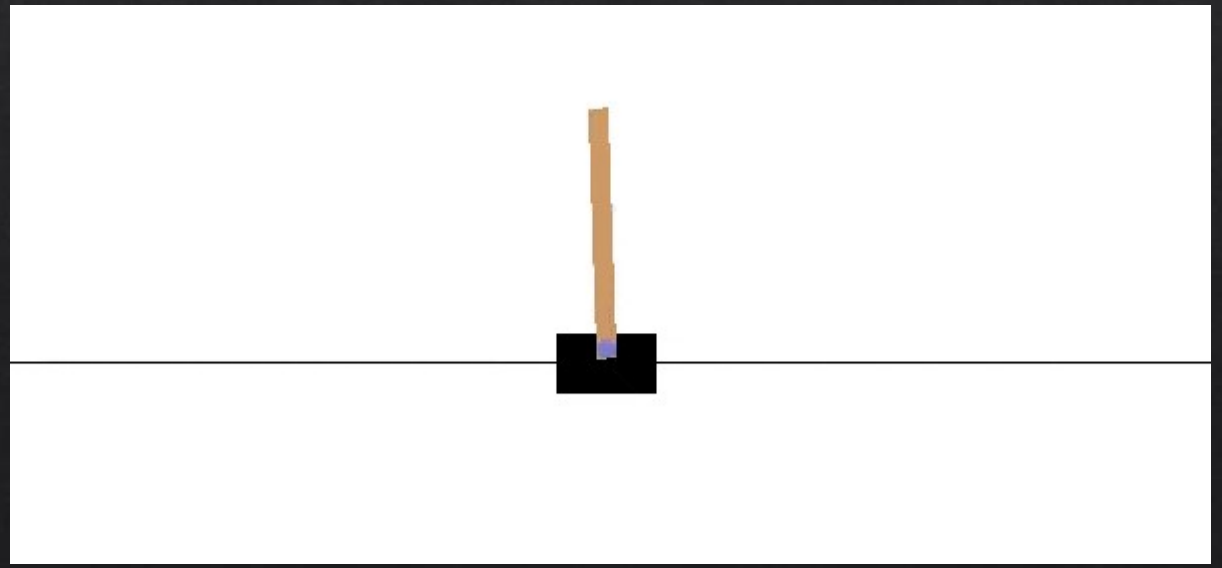
Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Estabilizar un pole

Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$



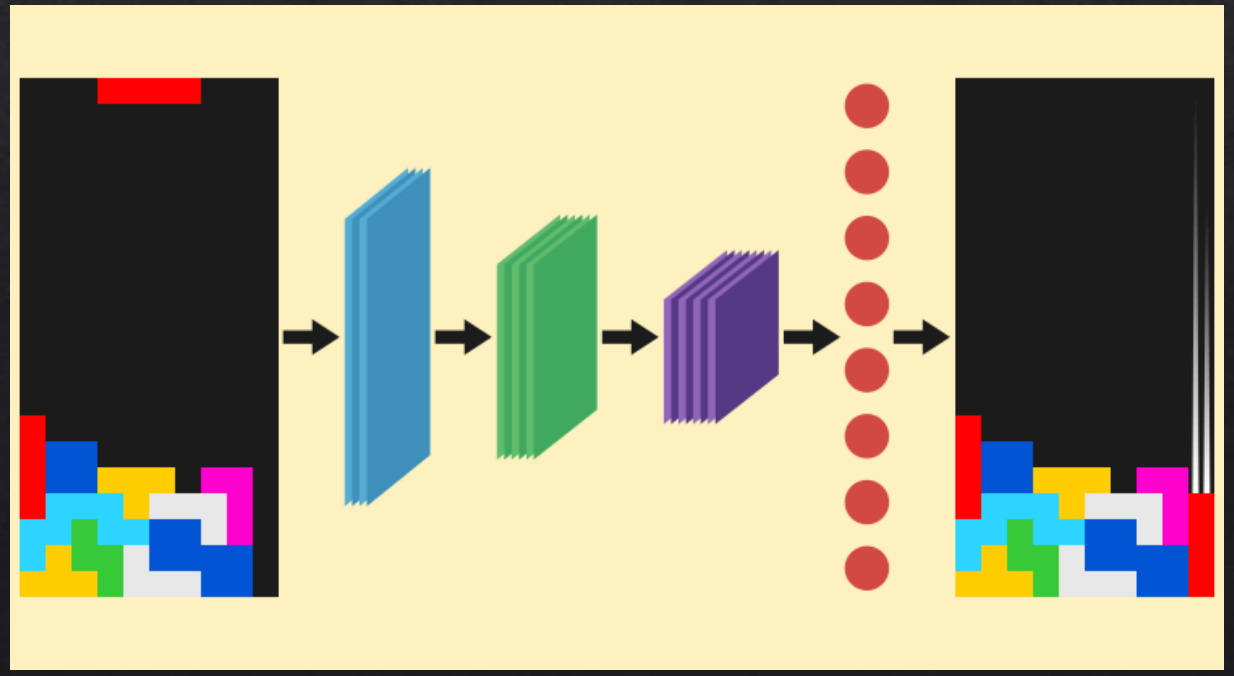
Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Juegos

Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$



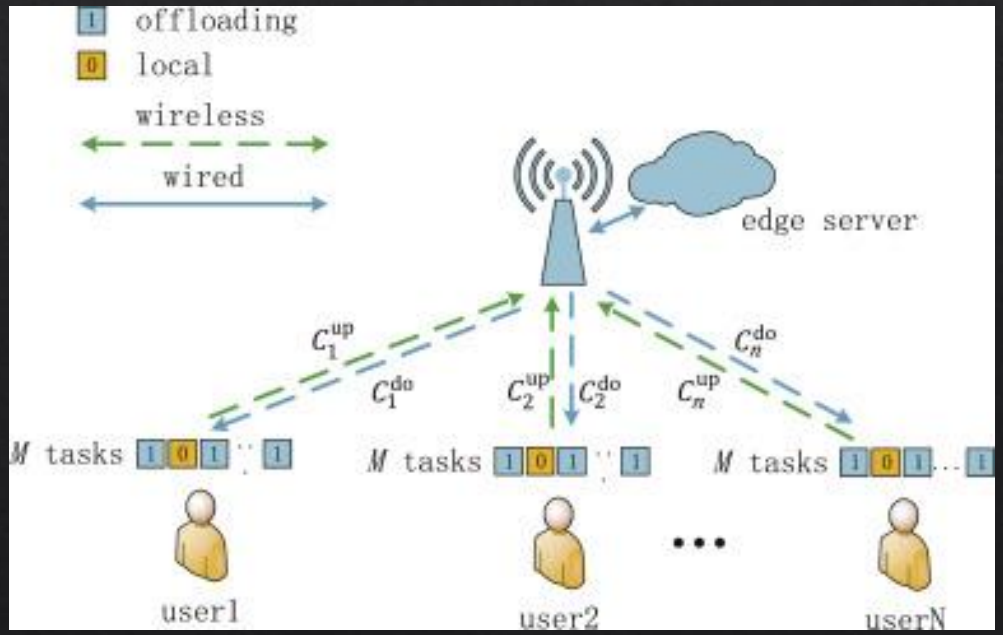
Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Administracion de Servidor

Objetivo:

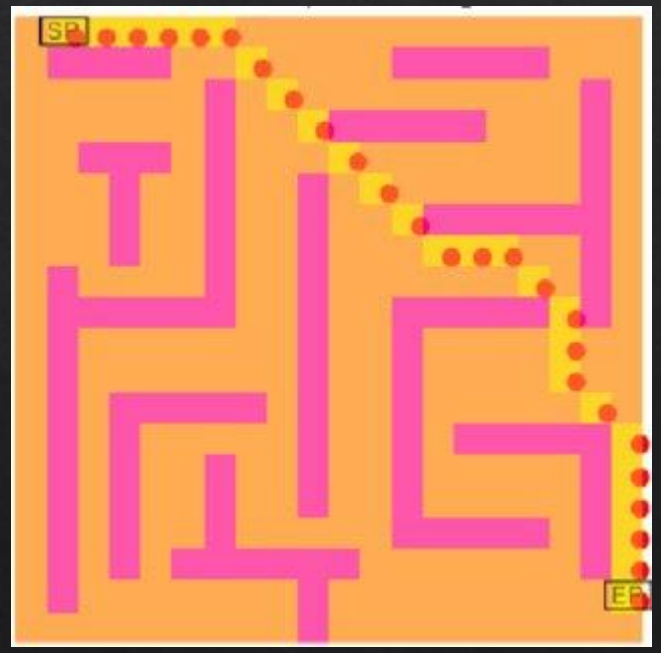
$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$



Marco de Trabajo: Proceso de Decisiones de Markov (MDP)

MDP (S, A, T, R, γ, H)

- Problemas de optimización de planificación



Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

Proceso de Decisiones de Markov (MDP)

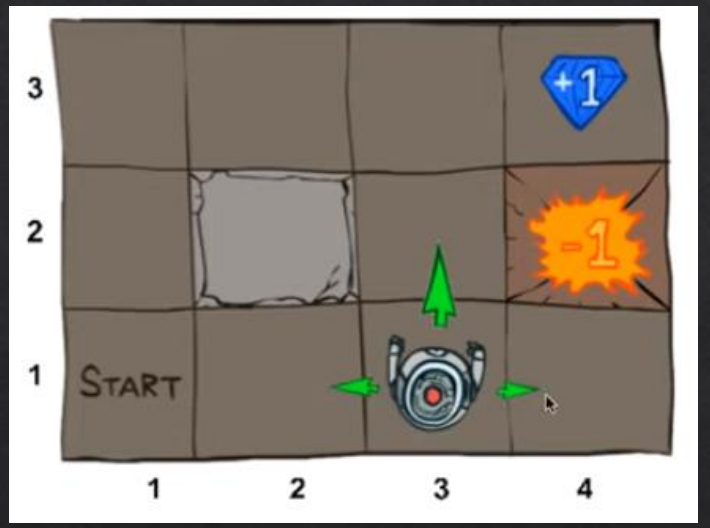
Un MDP está definido por:

- Set de Estados S
- Set de acciones A
- Función de Transición P(s´ | s, a)
- Función de Refuerzo R (s, a, s')
- Estado inicial So
- Factor descuento γ
- Horizonte H

Ejercicio de Grilla

Un MDP está definido por:

- Set de Estados S
- Set de acciones A
- Función de Transición P(s´ | s, a)
- Función de Refuerzo R (s, a, s')
- Estado inicial So
- Factor descuento γ
- Horizonte H



Objetivo:

$$max_{\pi} \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

Política: