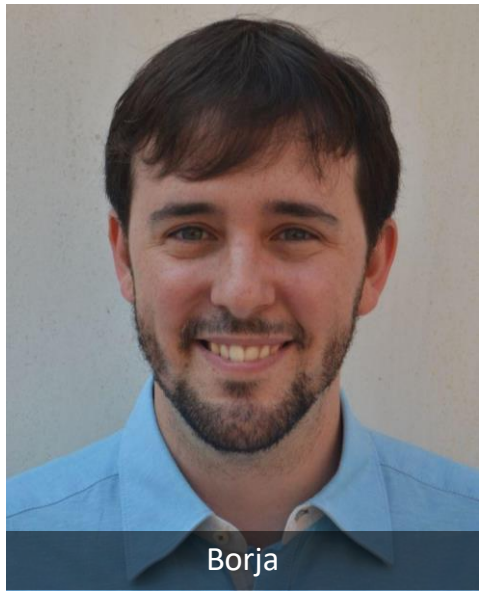
Aprendizaje por Refuerzo

Basado en Modelos



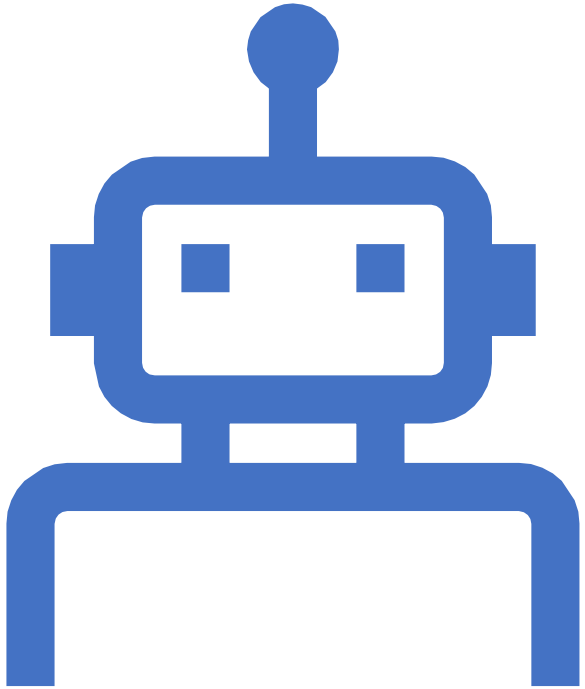
Imperial College London

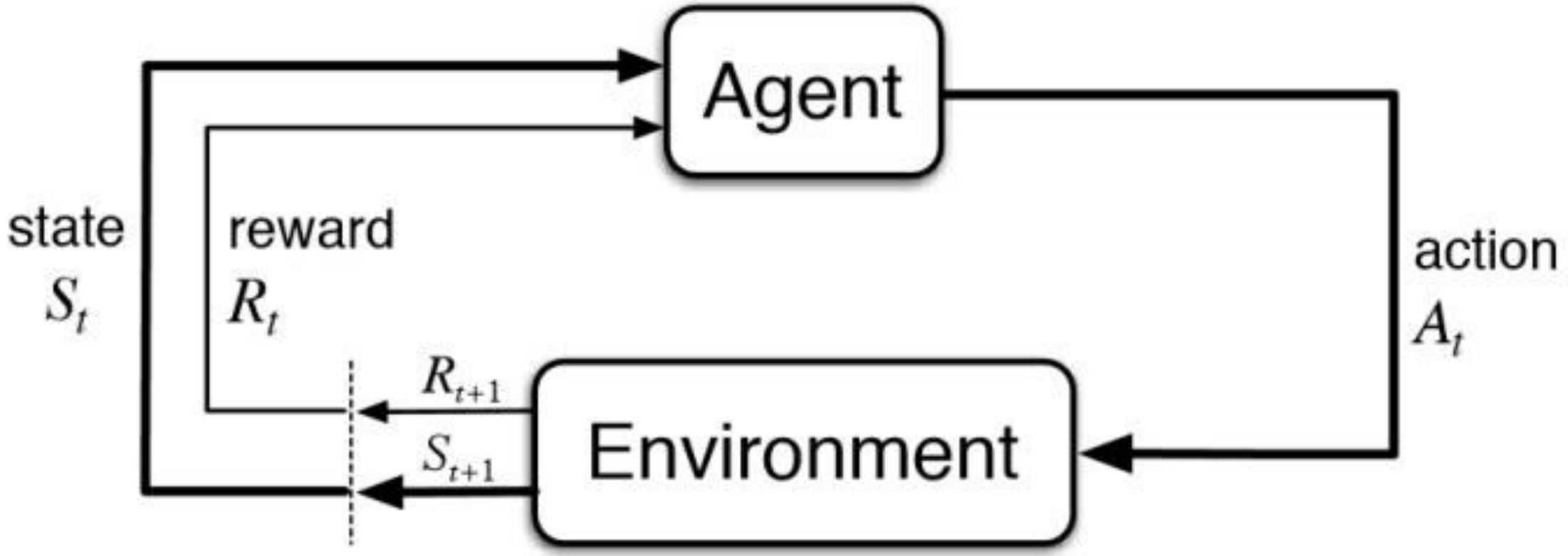
Sobre mi

- Investigador Doctoral en el **Imperial College de Londres**

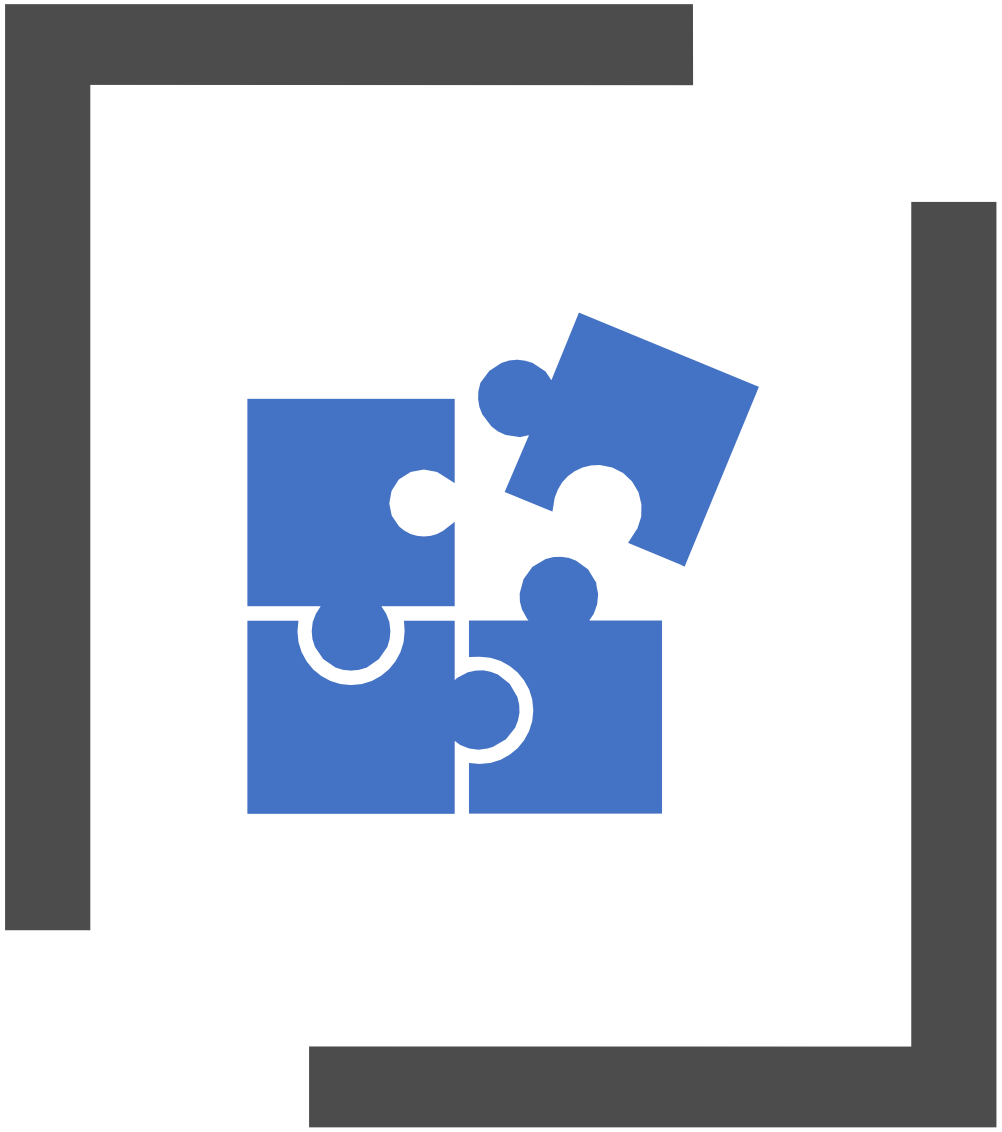
- Mi trabajo se centra en mejorar las capacidades de generalización del **aprendizaje por refuerzo profundo** mediante el uso de **IA simbólica**. Con especial atención a las aplicaciones en **sistemas multi-agente**.

¿Qué es el Aprendizaje por Refuerzo?





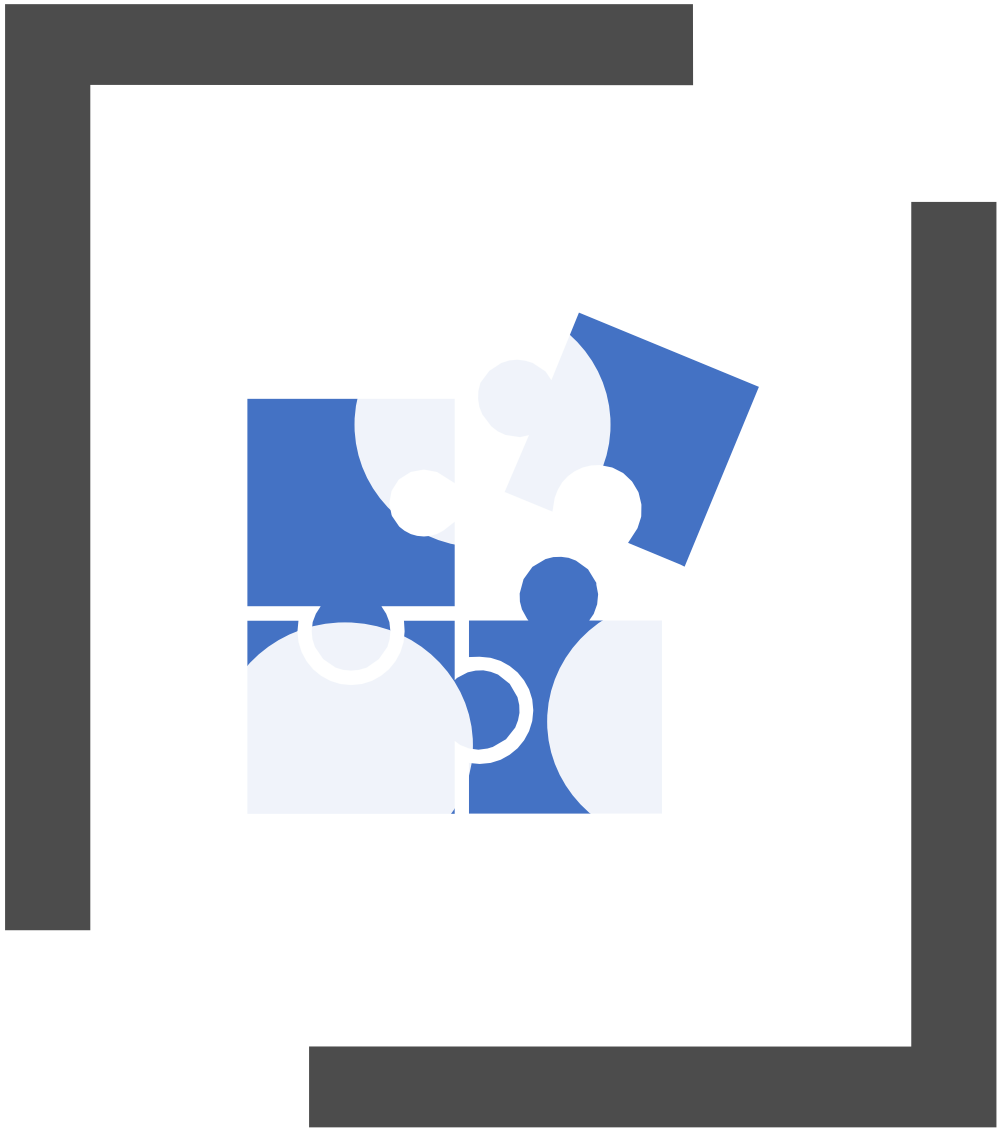
Sutton, Richard S., and Andrew G. Barto. Introduction to reinforcement learning.
Vol. 135. Cambridge: MIT press, 1998.



Modelo de Decision de Markov (MDP)

Un Modelo de Decision de Markov es una tupla $< S,\ A, P_{s,a}, \gamma, R >$ donde:

- S es un conjunto de estados.
- $A = \{a_1, a_2, \ldots a_k\}$ es un conjunto de acciones.
- $P_{s,a}(\cdot)$ son las probabilidades de transición desde un estado s tomando la acción a.
- $\gamma \in [0,1)$ es el factor de descuento.
- $R: S \rightarrow \mathbb{R}$ es la función de recompensas



MDP Parcialmente Observable (POMDP)

Un Modelo de Decision de Markov parcialmente observable es una tupla $< S, A, P_{s,a}, O, Z, \gamma, R >$ donde:

- S, A, P, γ y R son iguales que en el modelo anterior
- Z es el conjunto de posibles observaciones.
- O(s,a): $S \text{x} A \rightarrow Z$ es la función de observación

Ecuaciones de Bellman y Optimalidad

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s\pi(s)}(s') V^{\pi}(s')$$

$$Q^{\pi}(s,a) = R(s) + \gamma \sum_{s'} P_{sa}(s') V^{\pi}(s')$$

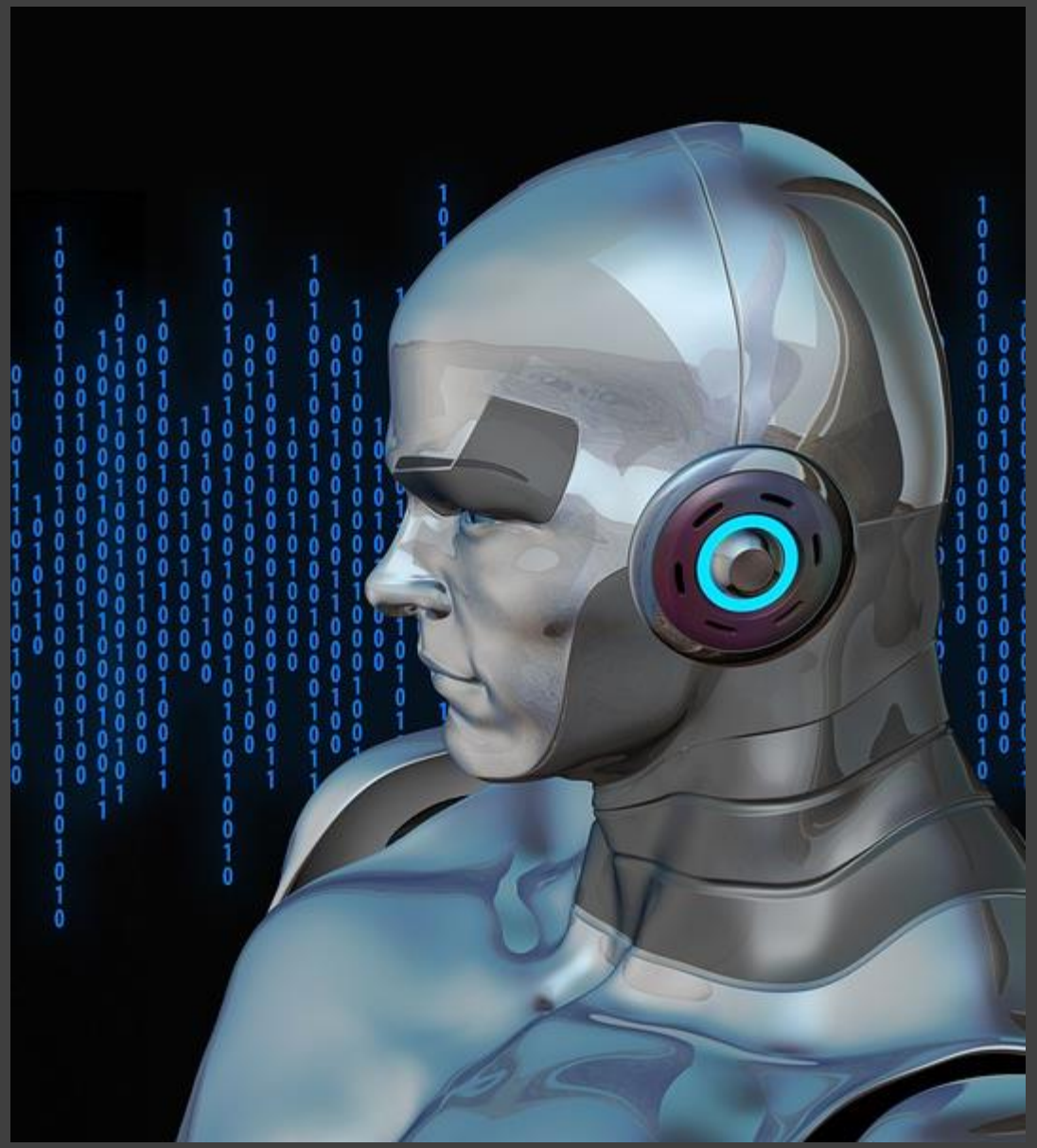
$$\pi^{*}(s) \in \arg\max_{a \in A} Q^{\pi}(s,a)$$

Model-Based vs Model-Free

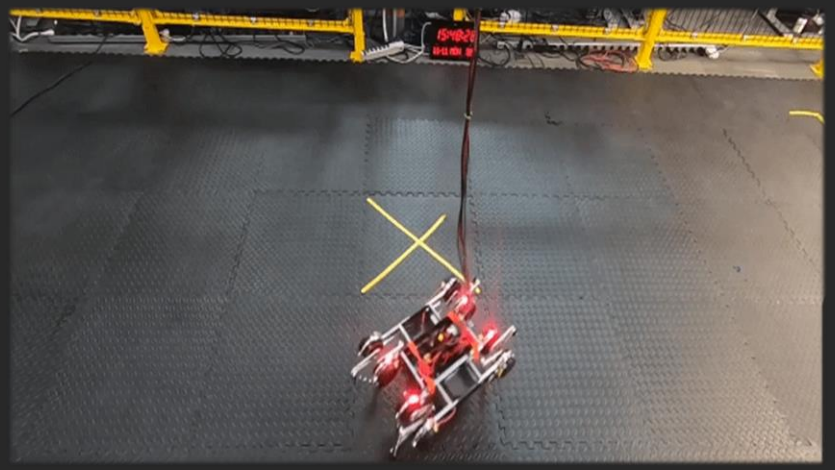
¿Debemos pensar antes de actuar?

- Un agente sin modelo (Model-free) puede ser visto como uno que actua por intuicion

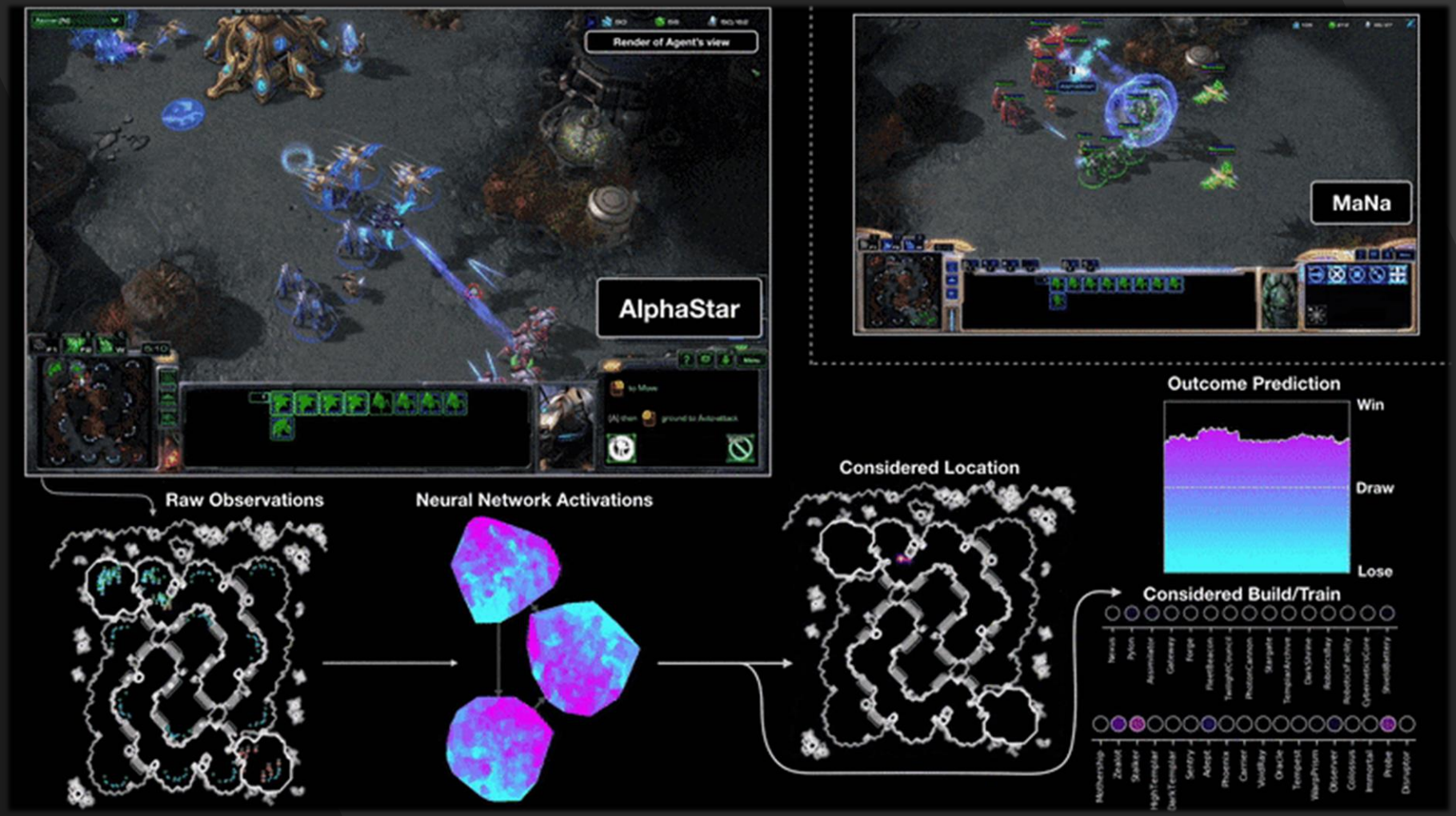
- Un agente que usa modelos (Model-based) “imagina” las posibilidades que puede explotar con sus acciones antes de actuar



Casos de éxito de Model-free RL

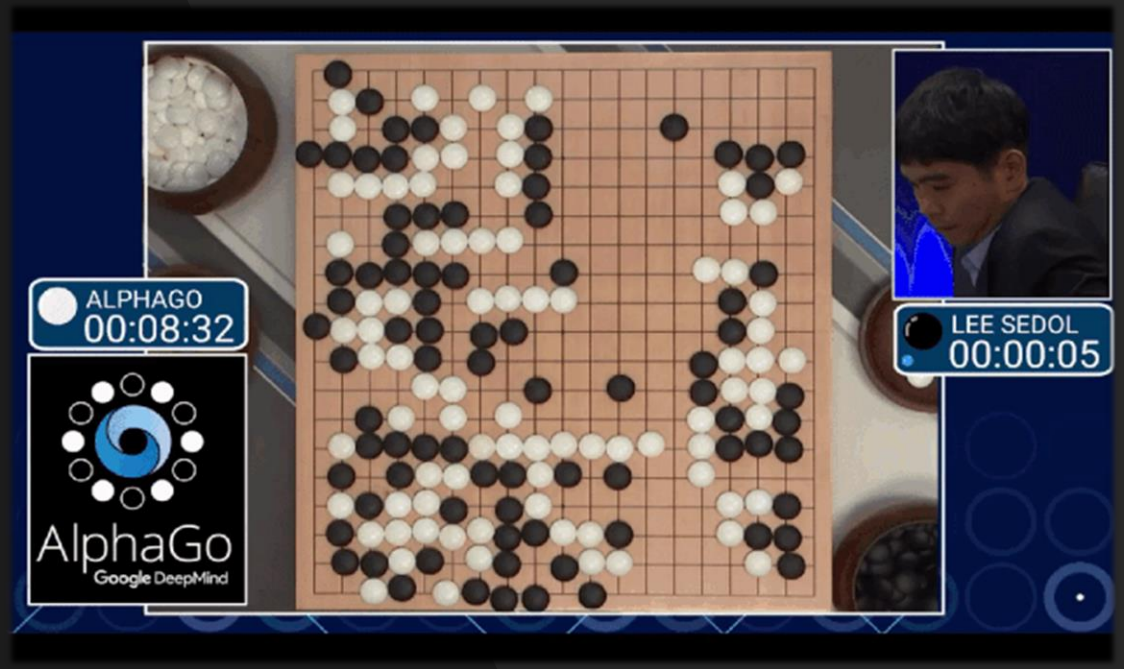


Casos de éxito de Model-free RL

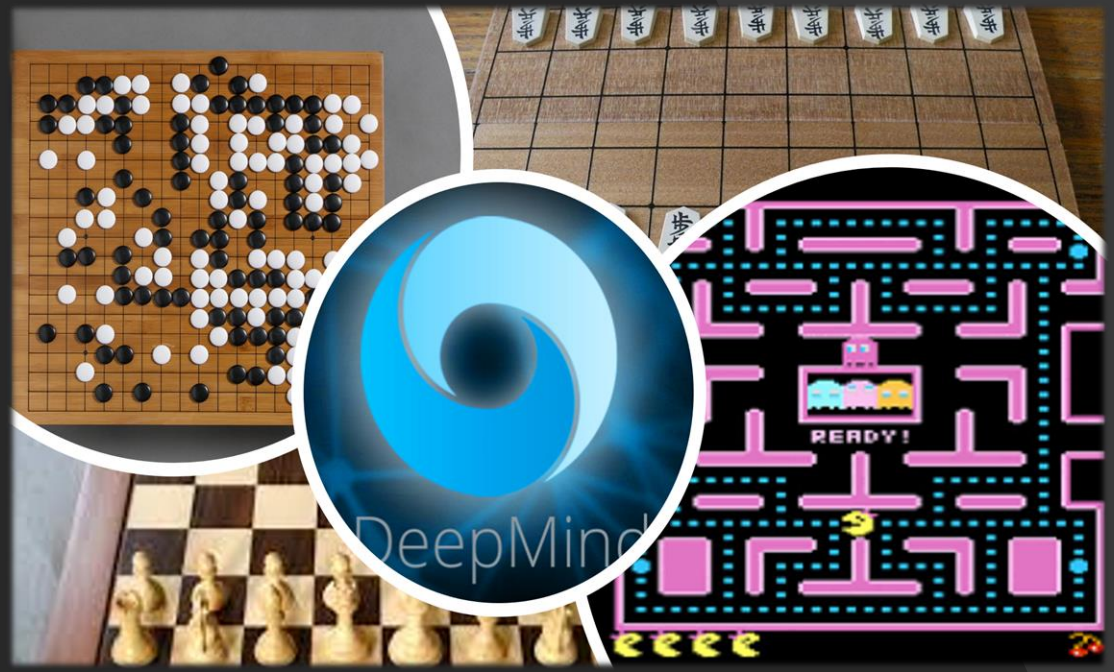


Casos de éxito de Model-Based RL

AlphaGo & AlphaZero

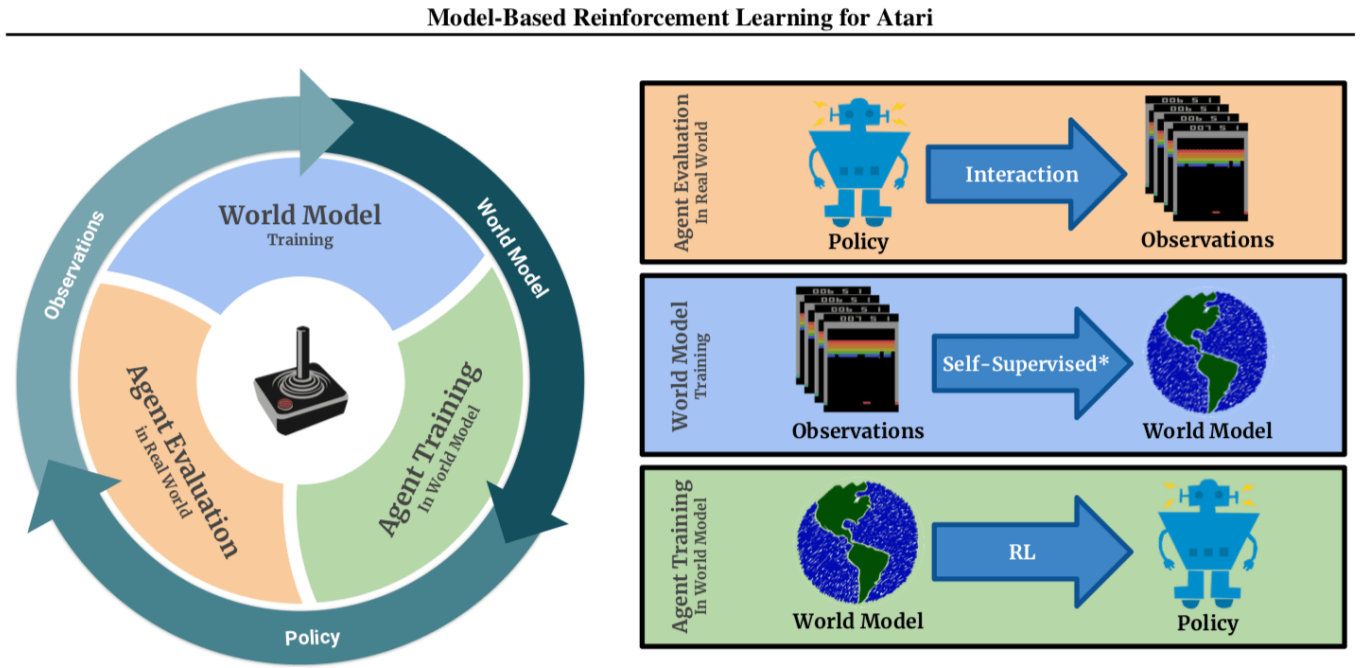


MuZero



Aprendizaje por Refuerzo

Basado en Modelos

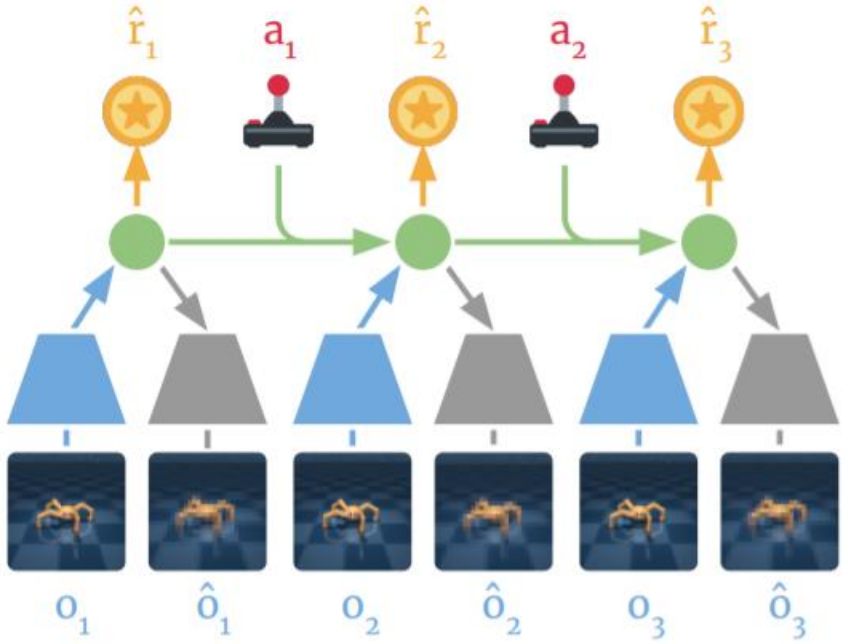


¿Cómo va esto?

- Muy similar al aprendizaje por refuerzo que no usa modelos: MDP, equaciones de Bellman…
- La mayor diferencia está en la construcción y utilización de un modelo para “imaginar” o simular lo que puede pasar.
- Una vez construido, si el modelo es bueno, podemos usar algoritmos de aprendizaje por refuerzo o de planificación
- Reduce enormemente el número de iteracciones necesarias con el entorno real.

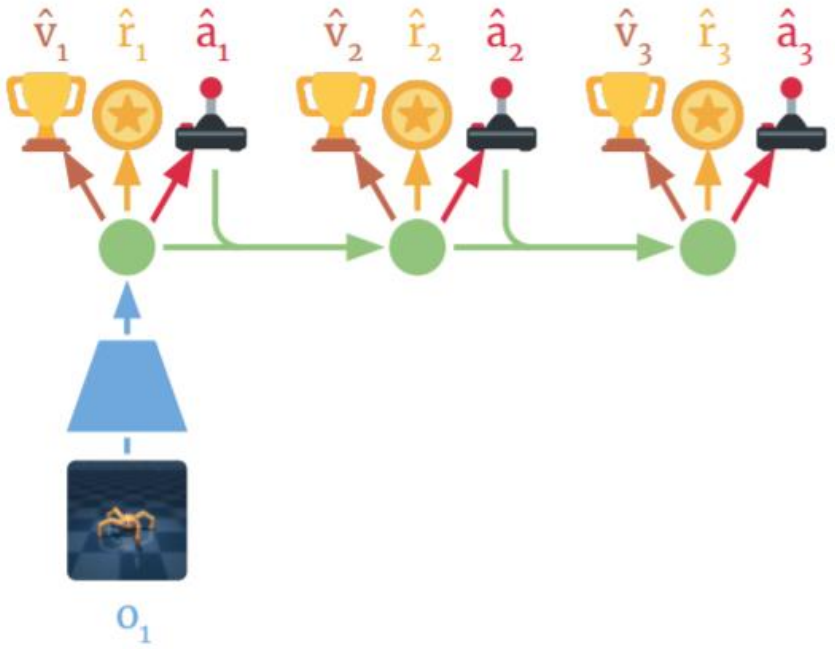
Iteraciones típicas

- Aprender las dinamicas del mundo en el que nos movemos



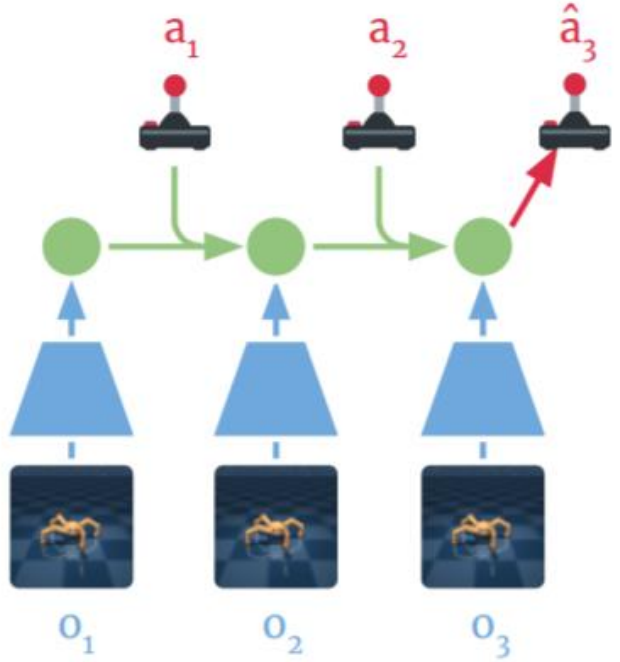
Iteraciones típicas

- Aprender las dinamicas del mundo en el que nos movemos
- Aprender a predecir las recompensas que se pueden obtener y las secuencias de estados y acciones que llevan a ellas

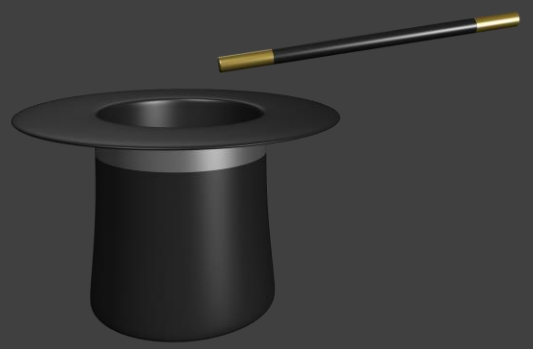


Iteraciones típicas

- Aprender las dinamicas del mundo en el que nos movemos
- Aprender a predecir las recompensas que se pueden obtener y las secuencias de estados y acciones que llevan a ellas
- Actuar en el entorno real



¿Dónde está el truco?

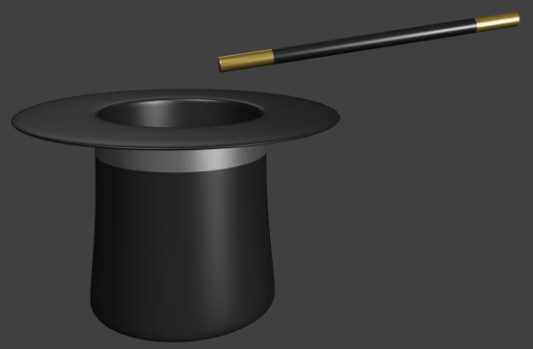


AQUÍ GRAN PARTE DEL TRABAJO SE CENTRA EN CONSTRUIR UN MODELO QUE SE AJUSTE A LA REALIDAD

EL RETO ESTÁ CUANDO NO TIENES NI IDEA DE A QUE TE ENFRENTAS DE ANTEMANO (LO QUE SUELE PASAR EN LA VIDA REAL)

TRADICIONALMENTE SE HA PRESTADO MENOS ATENCIÓN A ESTA FAMILIA DE ALGORITMOS YA QUE HA DADO PEORES RESULTADOS QUE LOS MODELOS “INTUITIVOS”

¿Dónde está el truco?



AUNQUE REQUIERA MENOS ITERACIONES CON EL ENTORNO SUELE DEMANDER MÁS “CEREBRO” (MÁS PODER DE CALCULO)

CUANDO LOS MODELOS HAY QUE APRENDERLOS SUELEN TENER SESGOS…..

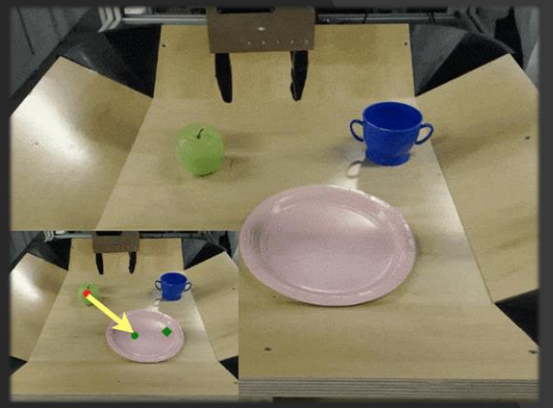
…..Y LOS AGENTES, TANTO DE RL COMO DE PLANNING, LES ENCANTA EXPLOTAR ESOS SESGOS EN LOS MODELOS….

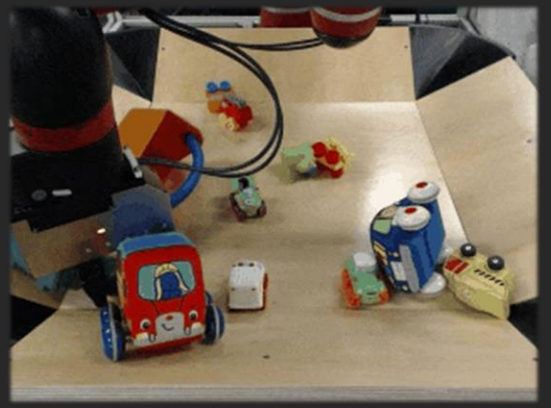
Estado del Arte

Cuando no conocemos el modelo y trabajamos con imagenes

Visual Model-Based Reinforcement Learning as a Path towards Generalist Robots



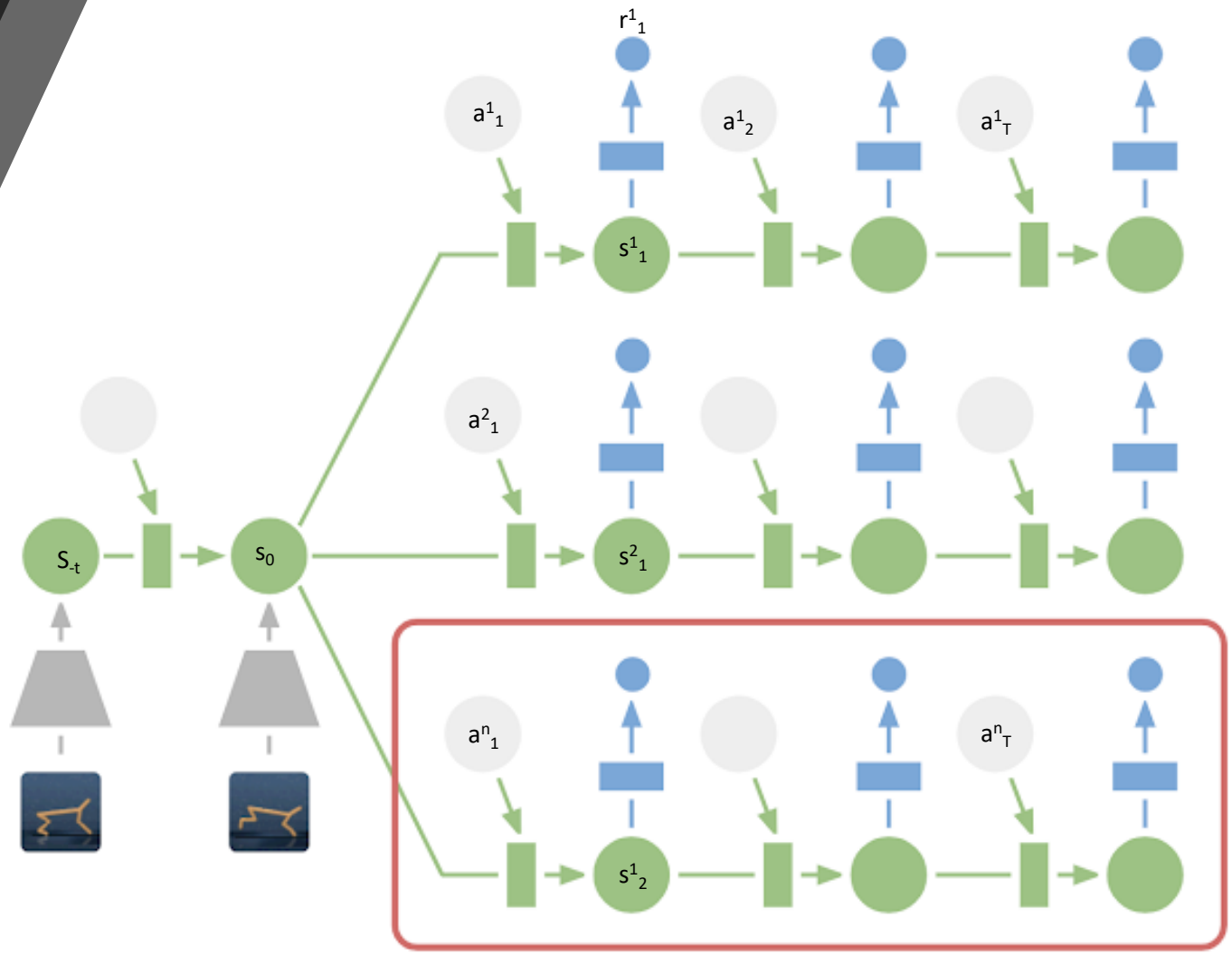





BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

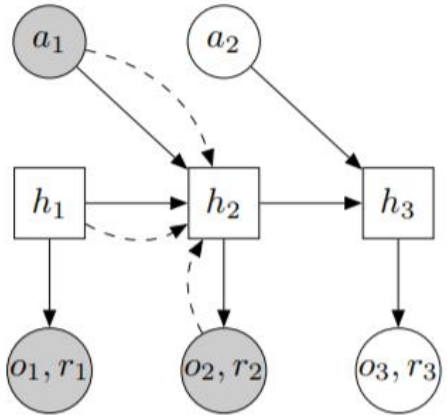
PlaNet: A Deep Planning Network for Reinforcement Learning

PlaNet: A Deep Planning Network for Reinforcement Learning

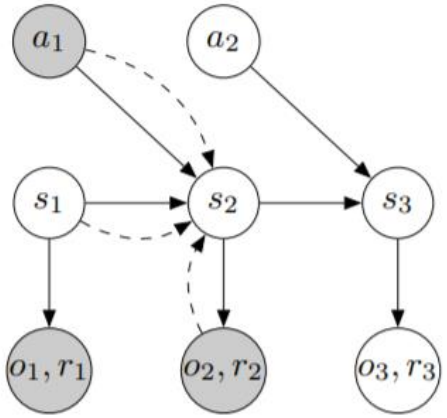


PlaNet: A Deep Planning Network for Reinforcement Learning

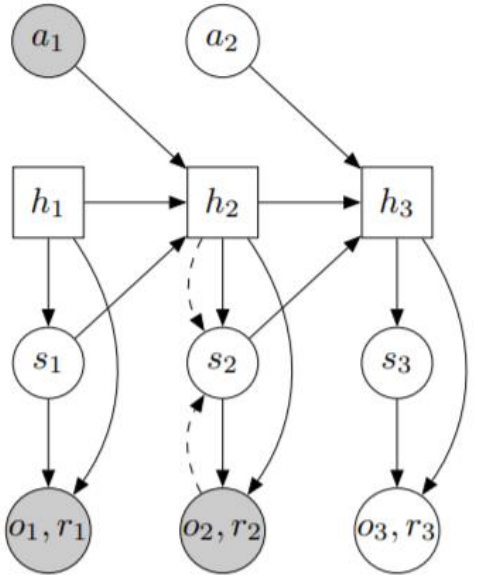
Learning Latent Dynamics for Planning from Pixels



(a) Deterministic model (RNN)

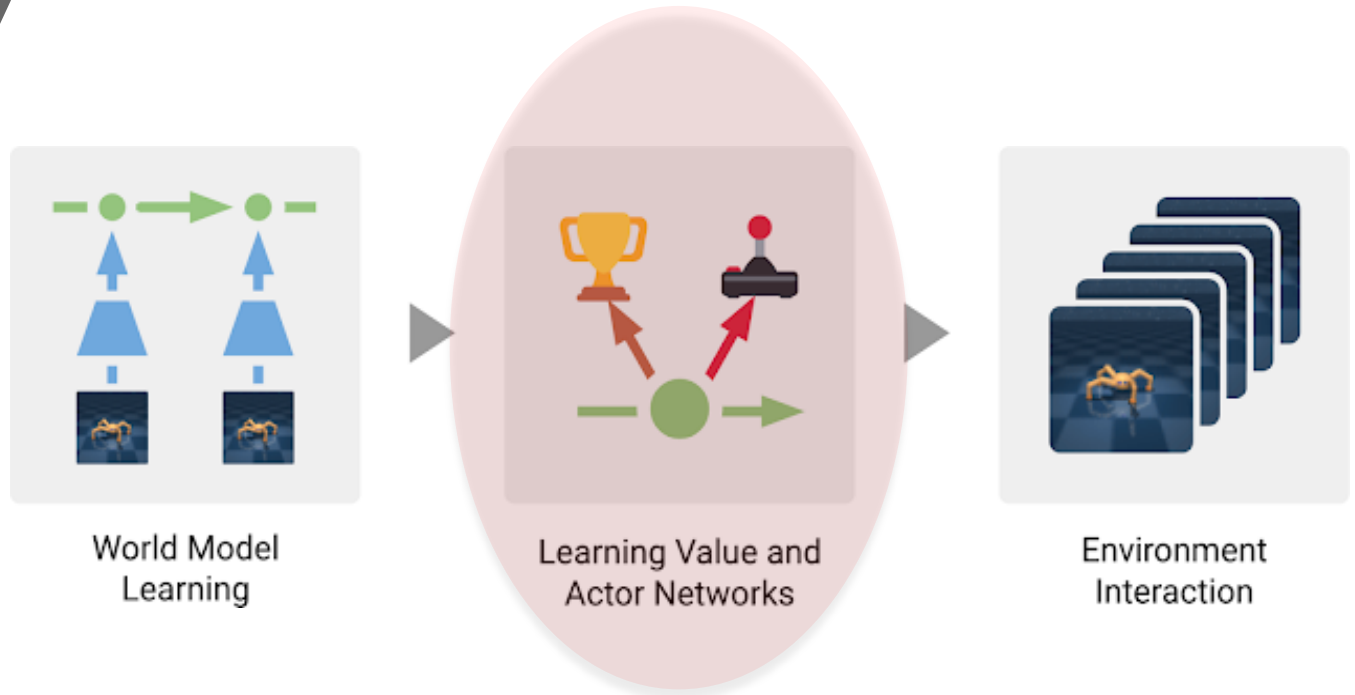


(b) Stochastic model (SSM)



(c) Recurrent state-space model (RSSM)

Dreamer: Scalable Reinforcement Learning Using World Models



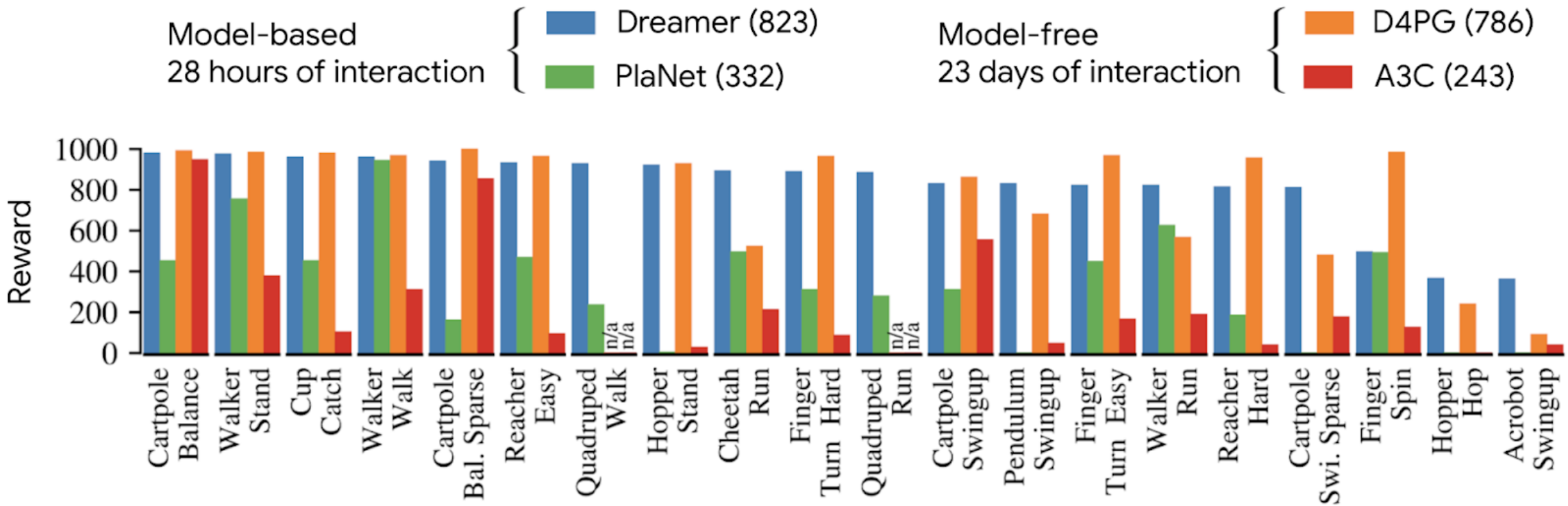
Dreamer: Scalable Reinforcement Learning Using World Models

encode images

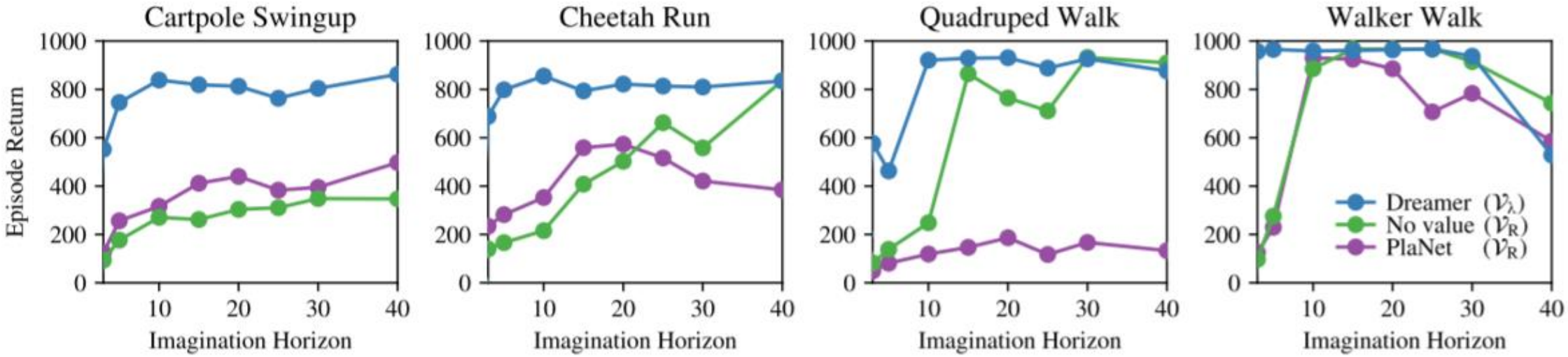
o_1

Algunos resultados

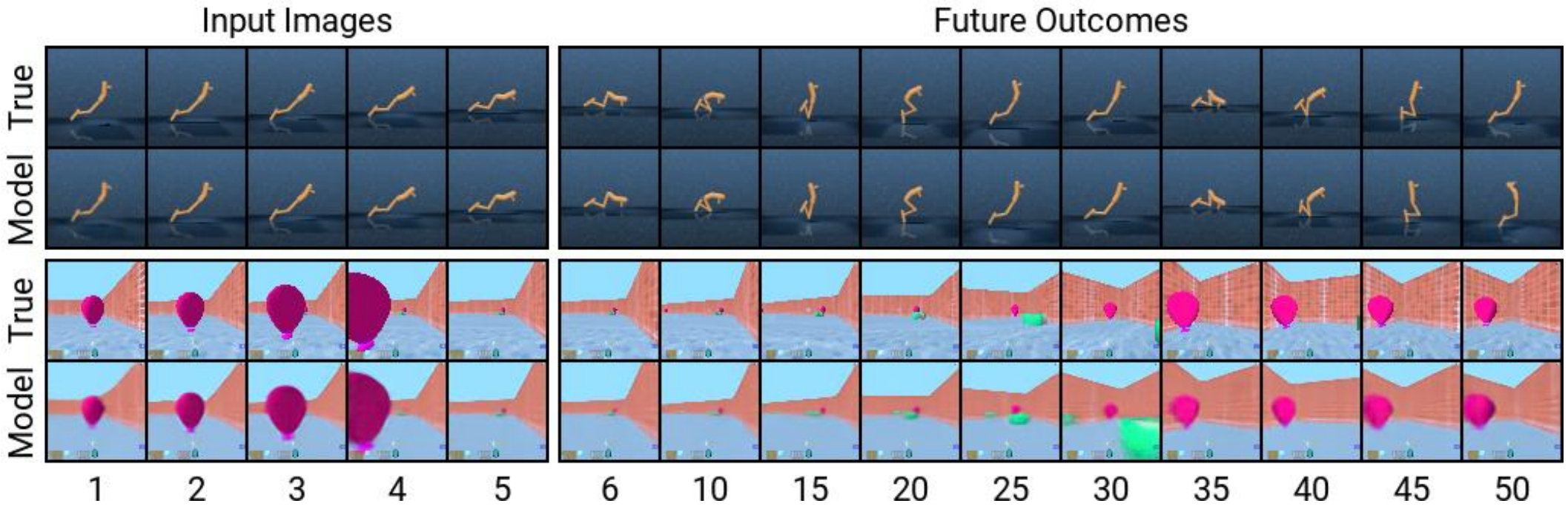
Algunos resultados



Algunos resultados



Siempre podemos ver que “imaginaba”



Para más detalle

- **PlaNet:**Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2018). Learning latent dynamics for planning from pixels. *arXiv preprint arXiv:1811.04551*.

- **Dreamer:** Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

DeepMind

Google AI

UNIVERSITY OF TORONTO

¿Preguntas?

@borruell