Aprendizaje por refuerzo y su impacto en el desarrollo de la humanidad

Ricardo Corral-Corral
@doctorcorral

VP of Engineering
Chief Data Scientist at Suggestic
https://www.suggestic.com/

AI Learners

ABRIL 10, 2019

Reinforcement Learning

|> Aprendizaje de **Refuerzo**

Conceptos particulares

vocabulario

|> Ejemplos de uso

Reinforcement

Reinforcement learning is a computational approach to understanding goal-directed learing and decision making. It is distinguished from other computational approaches by its emphasis on learning by an **agent** from direct interaction with the **environment**.

Reinforcement Learning
Richard S. Sutton, Andrew G. Barto

Learning

Reinforcement

Crédito: **RISELab - UC Berkeley**
https://www.youtube.com/watch?v=jymFj7bNsKg

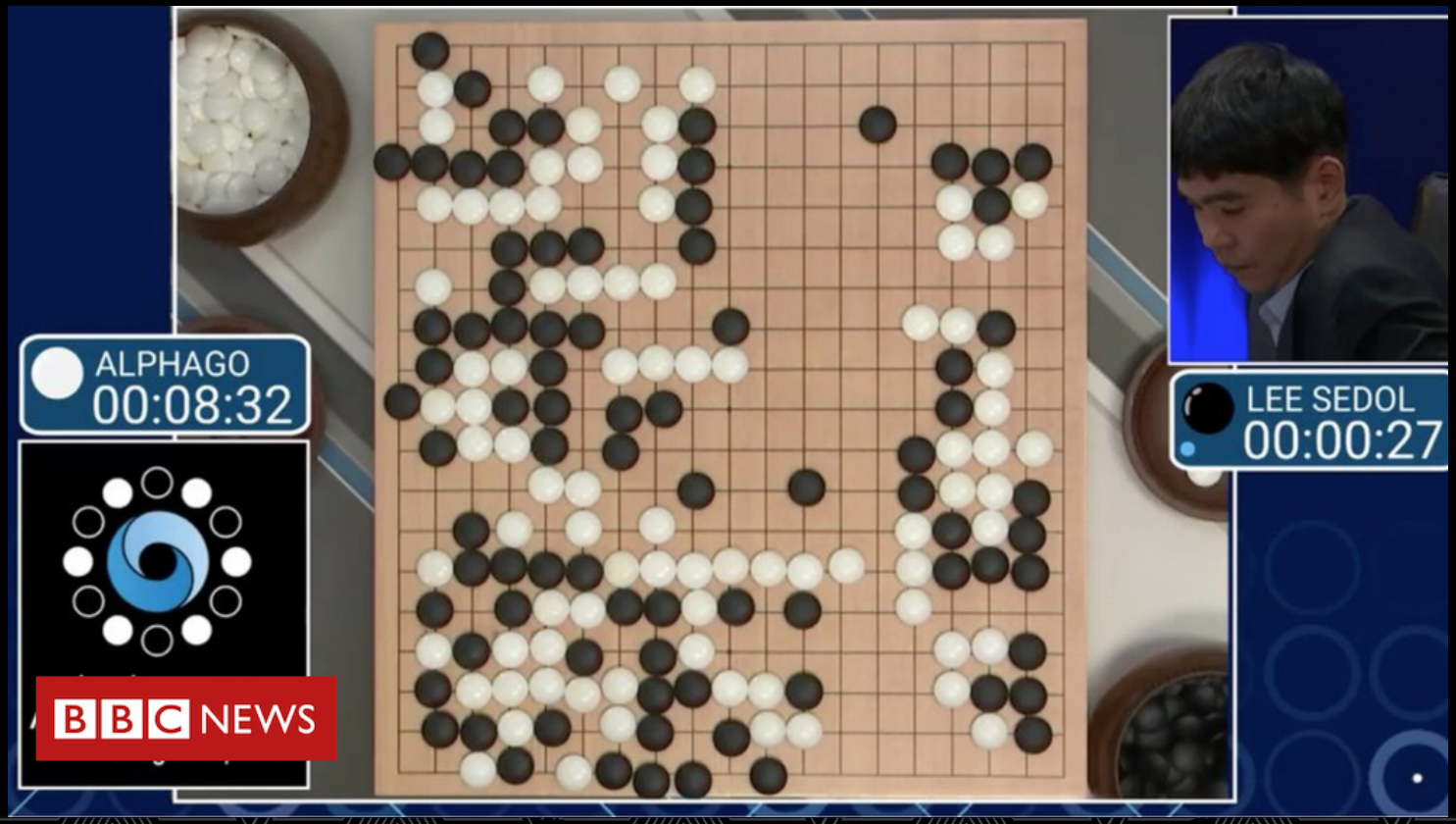
Learning

Reinforcement

Learning

Crédito: **@simoninithomas**

Reinforcement

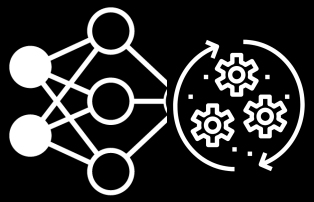


Learning

Google DeepMind Challenge Match
AlphaGo versus Lee Sedol
9 -15 March 2016

Reinforcement Learning

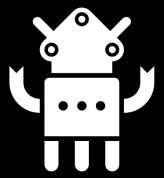
environment



action

reward

observation

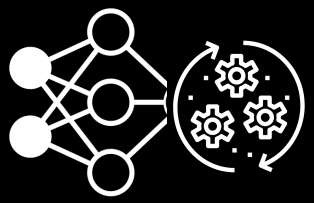


agent

- El agente es solo la parte capaz de tomar decisiones
- El ambiente recompensa al agente por sus acciones
- El agente observa los cambios en el ambiente
- El agente aprende a partir de la experiencia

Proceso de Decisión Markoviano

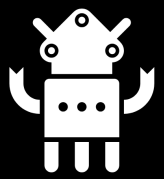
La decisión de cada acción solo considera el estado actual, ignorando estados y acciones previas



action

$$q_\pi(s,a)$$

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t=s,A_t=a]$$



agent

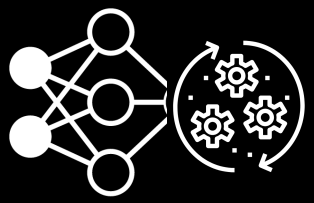
El conocimiento adquirido es utilizado para estimar la recompensa esperada y derivar la *mejor* política

$$q_*(s,a) = \sum_{s'\in\mathcal{S},r\in\mathcal{R}} p(s',r|s,a)(r+\gamma \max_{a'\in\mathcal{A}(s')} q_*(s',a'))$$

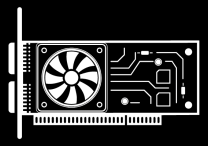
$$q_\pi(s,a) = \sum_{s'\in\mathcal{S},r\in\mathcal{R}} p(s',r|s,a)(r+\gamma \sum_{a'\in\mathcal{A}(s')} \pi(a'|s')q_\pi(s',a'))$$

Principio de optimalidad de Bellman

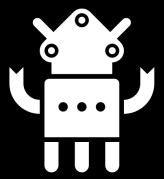
Redes Neuronales Artificiales



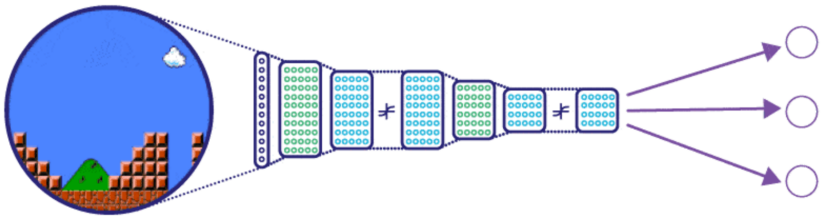
action



GPU intensive



agent



Deep Q-Learning

Estado | capas densas y convolutivas | valores Q / política

Value based

Vs

Policy based

methods

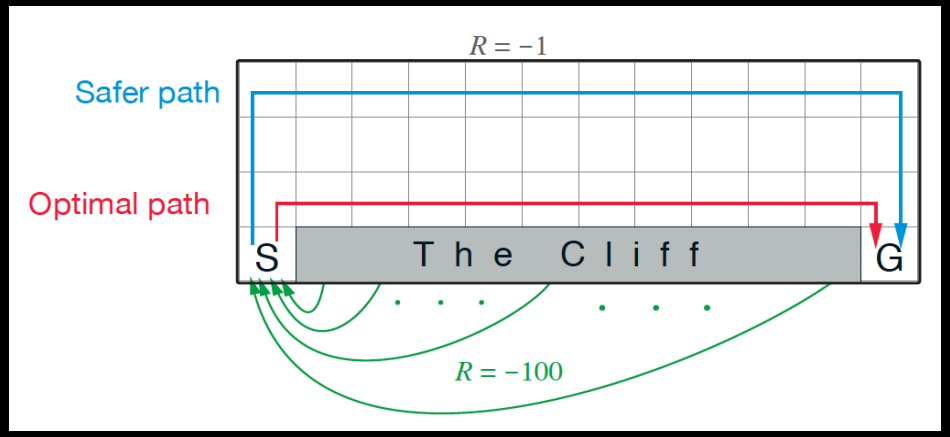
Actor Critic Methods

En los métodos basados en valor, la política se obtiene indirectamente a partir de la estimación de recompensa (descontada futura).

On policy

Vs

Off policy



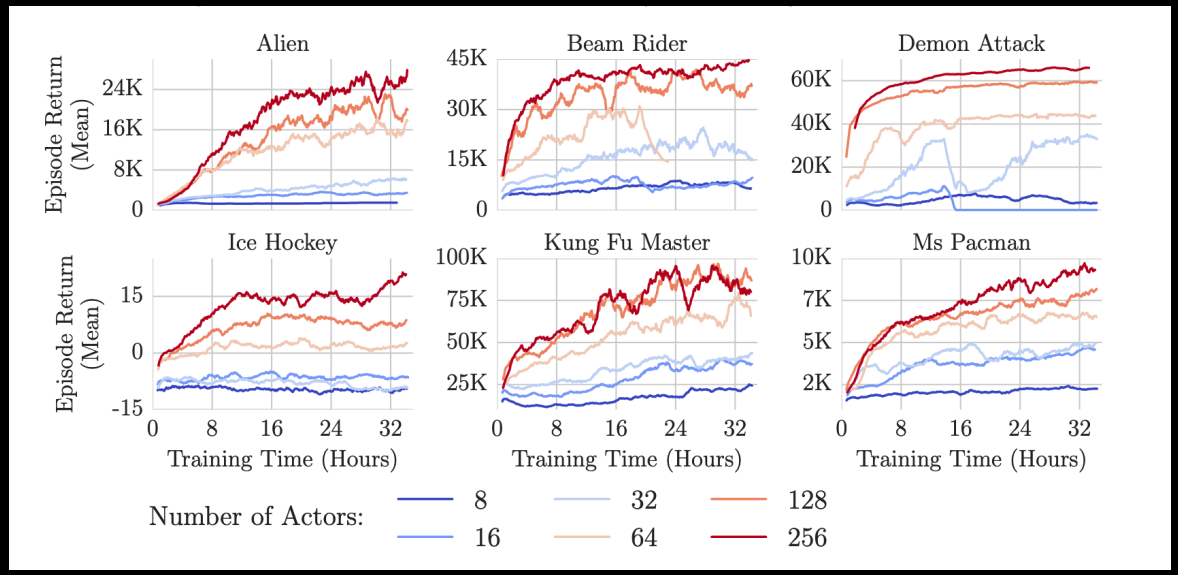
Sarsa

Q-Learning

On policy: el proceso de aprendizaje incorpora directamente las acciones producidas por la política actual.

Example 6.6: Cliff Walking,
Richard S. Suttonand, Andrew G. Barto.
Reinforcement Learning: An Introduction

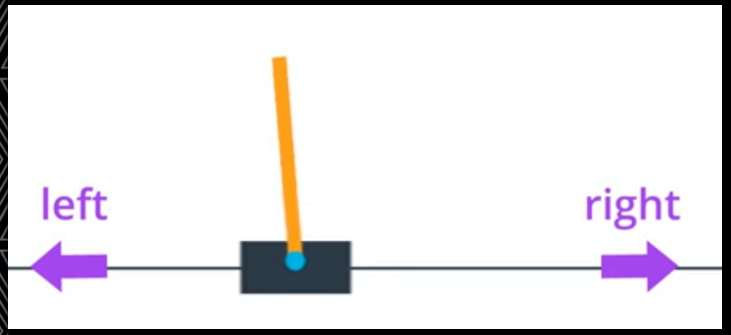
Buffer Replay

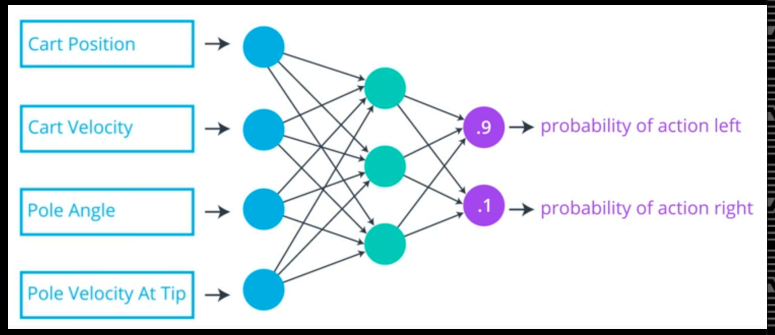


Métodos basados en gradiente se benefician de muestreo de datos independientes e idénticamente distribuidos.

DISTRIBUTED PRIORITIZED EXPERIENCE REPLAY
https://arxiv.org/pdf/1803.00933.pdf

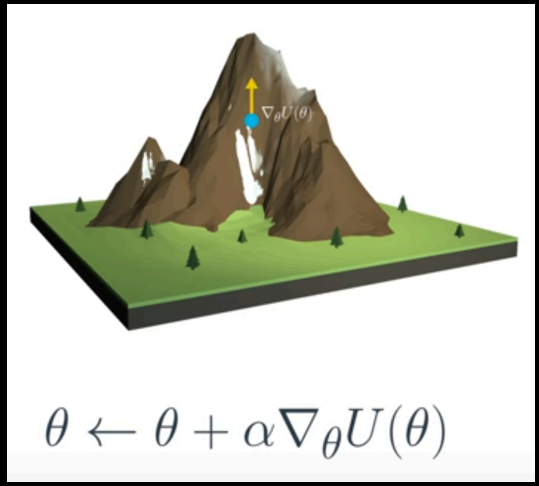
Gradient policy based





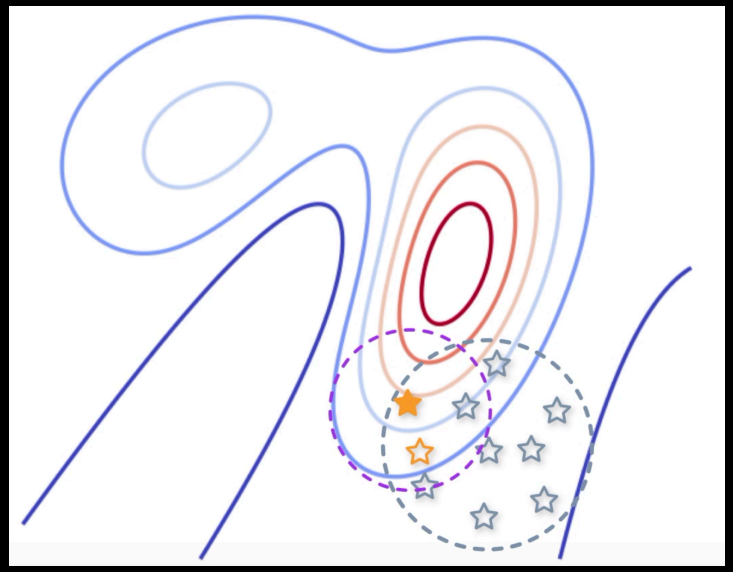
Una trayectoria es una secuencia de acciones ejecutadas por el agente y los estados visitados

Gradient policy based



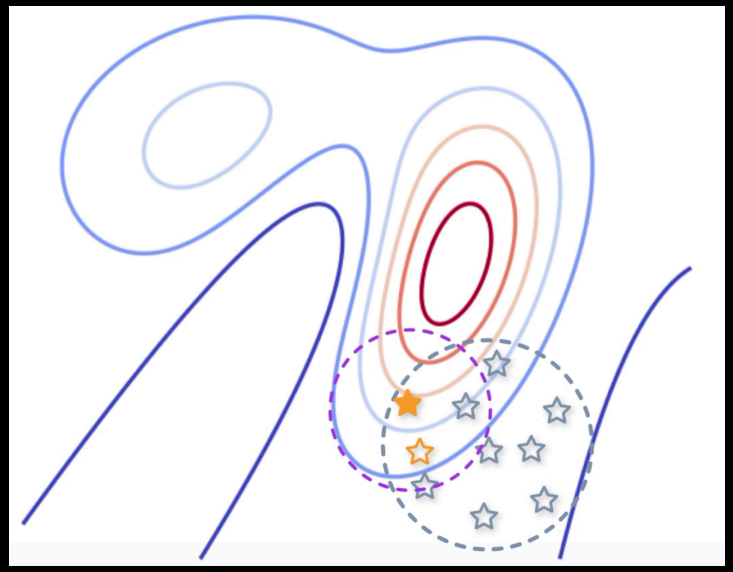
$$U(\theta) = \sum_{\tau} P(\tau;\theta) R(\tau)$$

Black Box Optimization



Proximal Policy Optimization (PPO)

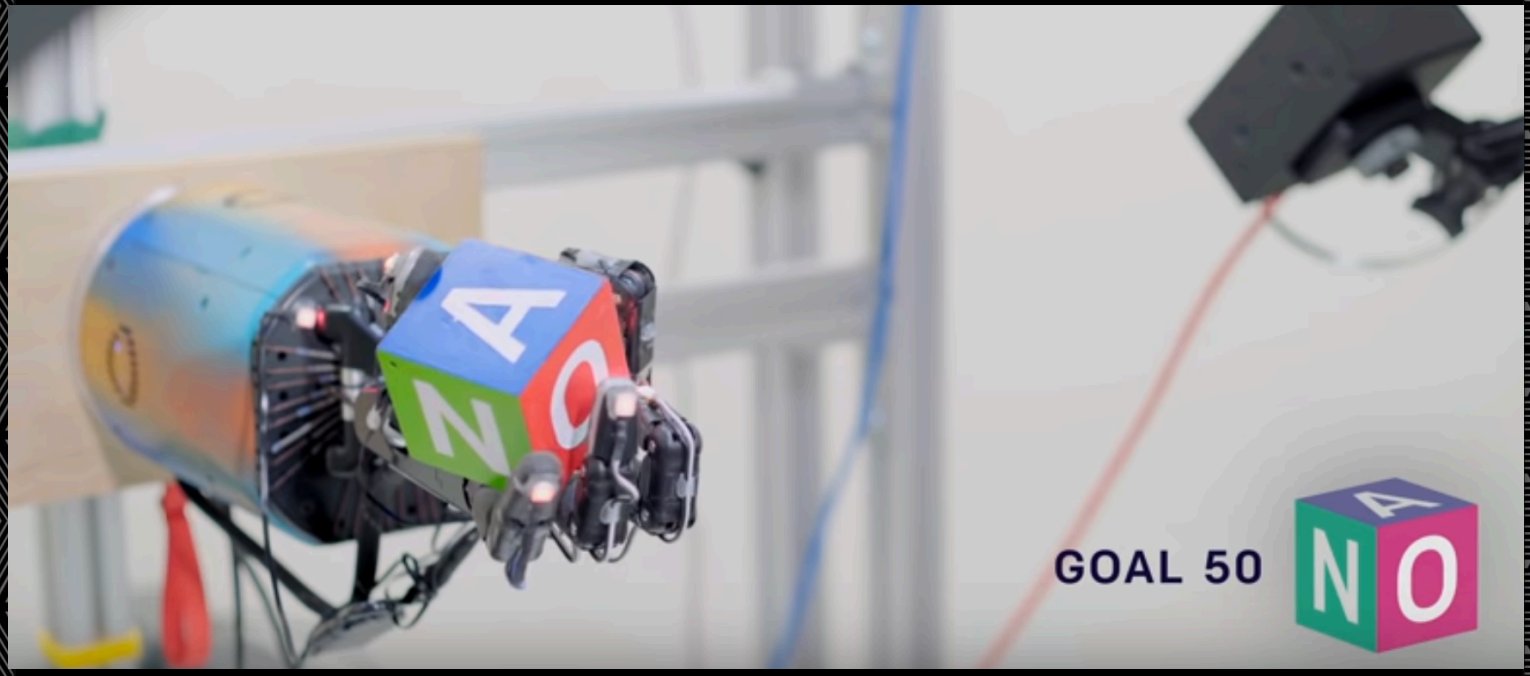
Black Box Optimization



Proximal Policy Optimization (PPO)

Casos de uso

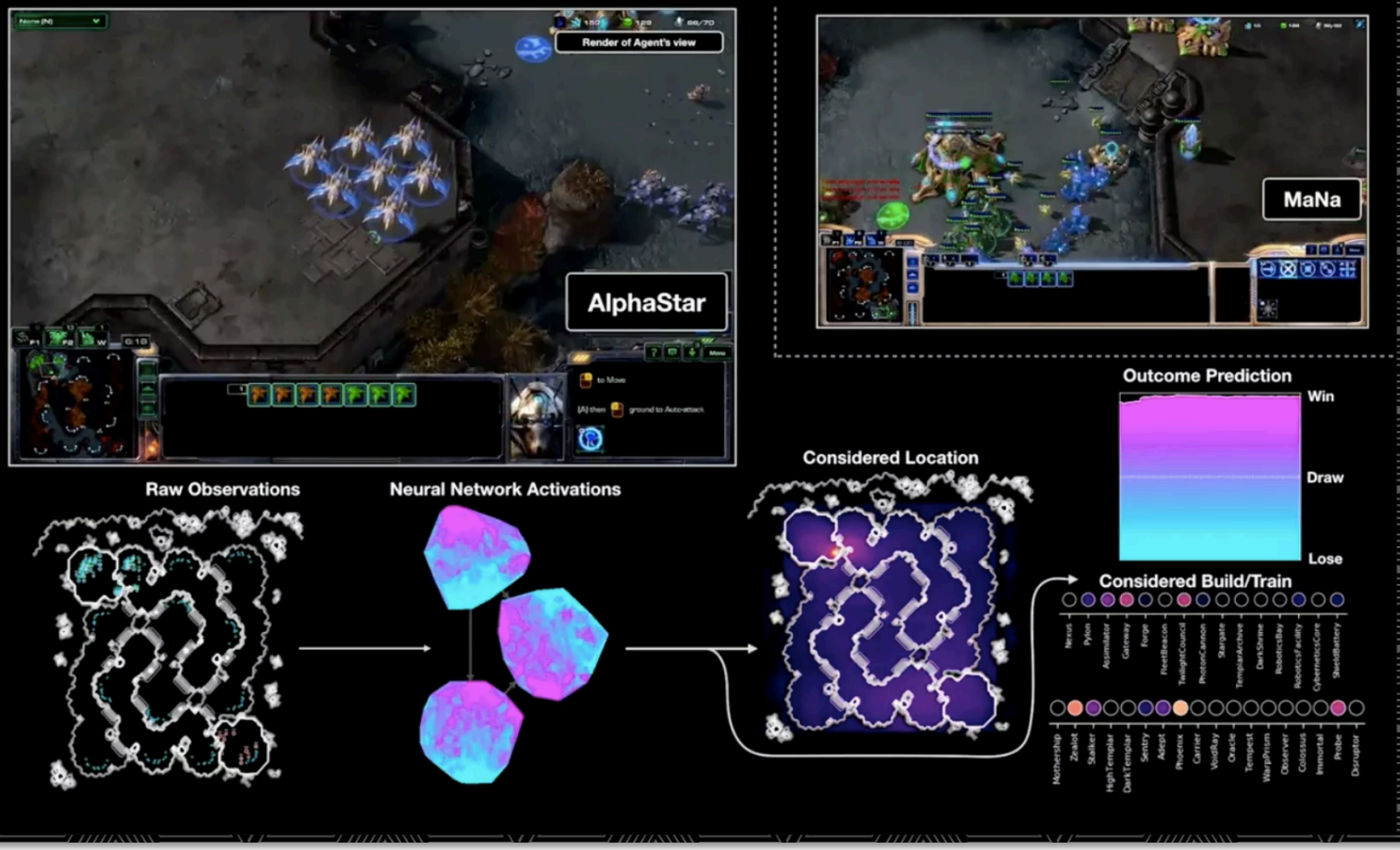
Robótica



https://openai.com/blog/learning-dexterity/

Transfer learning

AlphaStar

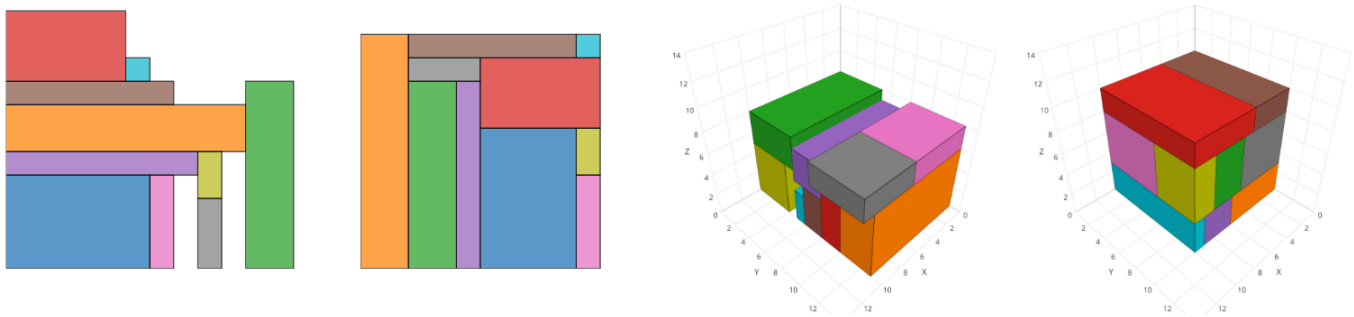


https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

BostonDyamics

https://www.bostondynamics.com/spot-mini

InstaDeep



a Lego **b** Rank-75% **c** Lego **d** Rank-75%

Figure 3: Visualization of the solution by Lego and Rank-75% in 2D and 3D.

https://www.instadeep.com/
https://arxiv.org/pdf/1807.01672.pdf

tooling

Gym https://github.com/openai/gym

ML-Agents Toolkit https://github.com/Unity-Technologies/ml-agents

https://github.com/google/dopamine

https://github.com/facebookresearch/Horizon

https://github.com/doctorcorral/gyx

Recursos

http://incompleteideas.net/book/RLbook2018.pdf

https://spinningup.openai.com/en/latest/

http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html

https://github.com/doctorcorral/gyx

https://codesync.global/speaker/ricardo-corral-corral302/

¡GRACIAS!