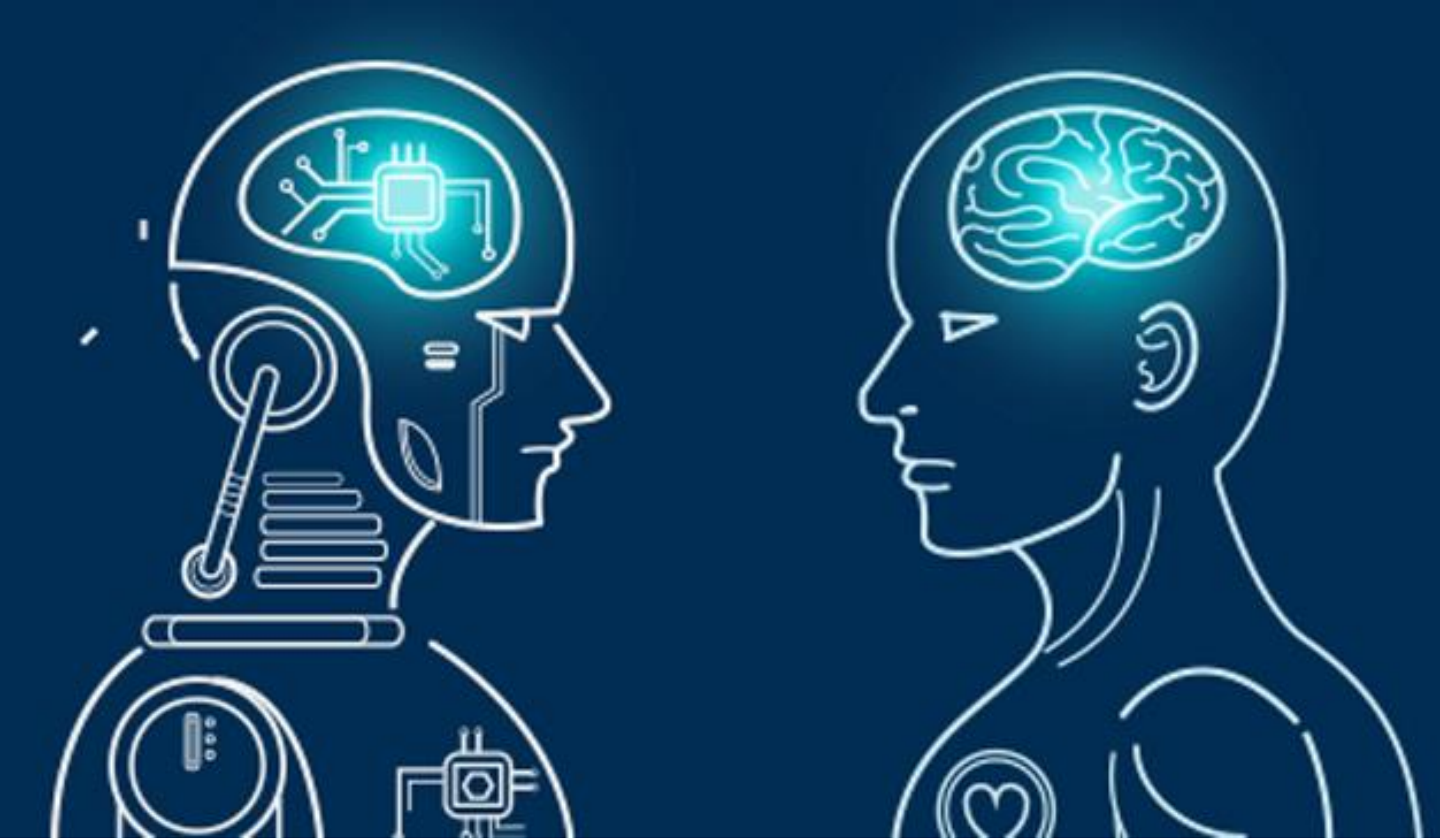
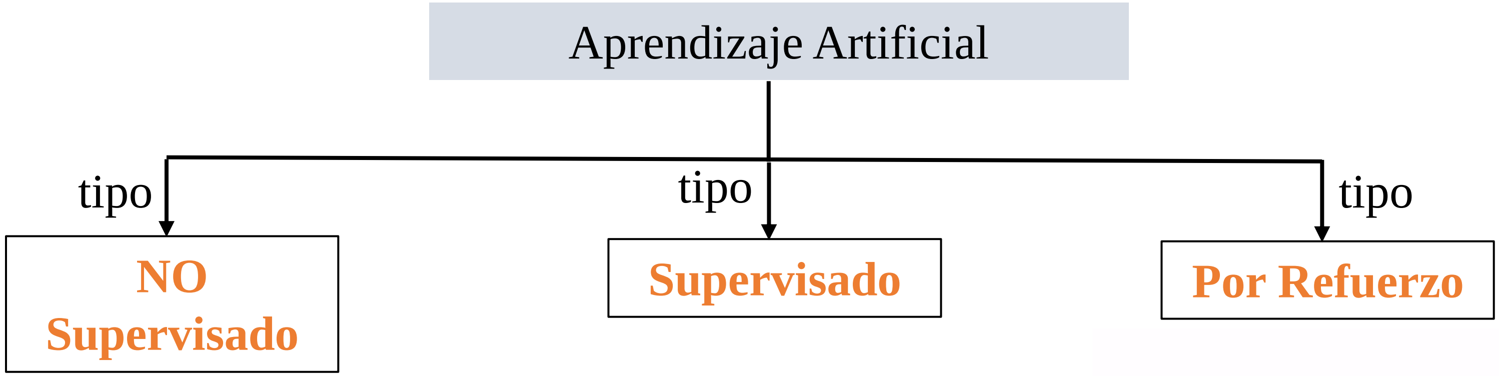
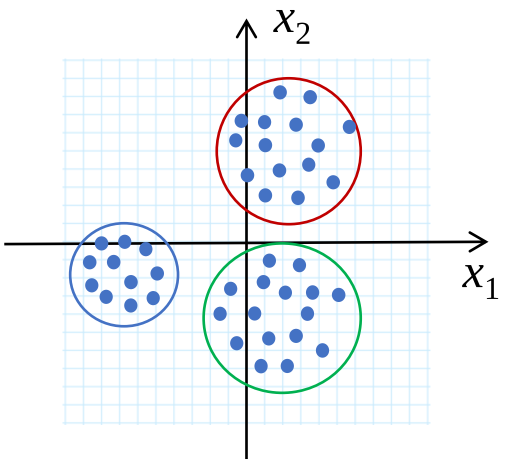
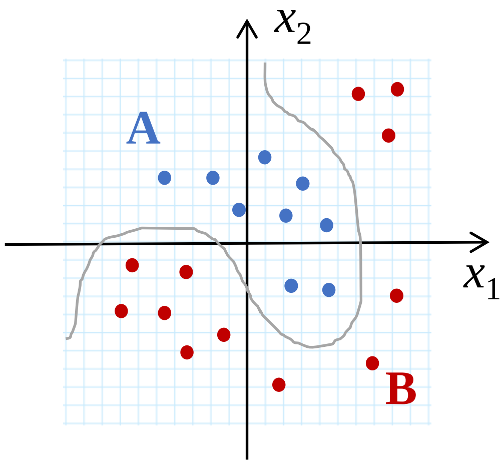
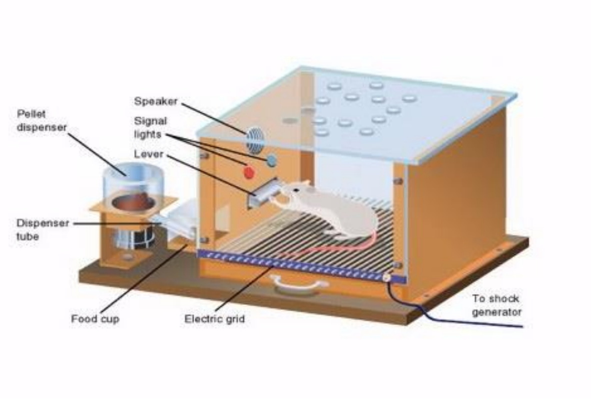


Aprendizaje Por Refuerzo





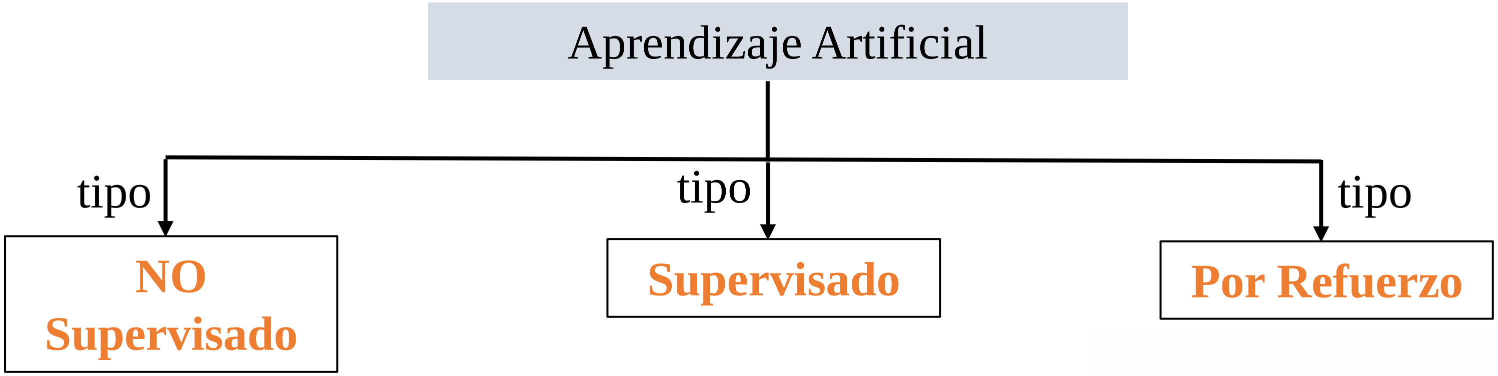




Datos	NO Etiquetados NO Estructurados	Retroalimentación	X
Función	Agrupación ↓Dimensionalidad		
Objetivo	Encontrar estructura en los Datos		

Datos	Etiquetados Estructurados	Retroalimentación	✓
Función	Clasificación Regresión		
Objetivo	Predicción		

Datos	Estados Recompensa	Retroalimentación	✓
Función	Asociar acciones a estados		
Objetivo	Encontrar las mejores Acciones		



Aprendizaje Artificial

tipo

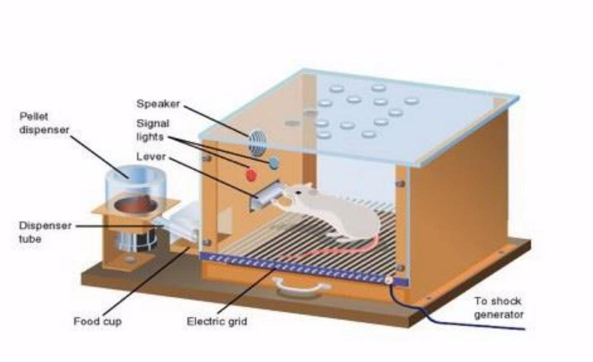
NO Supervisado

tipo

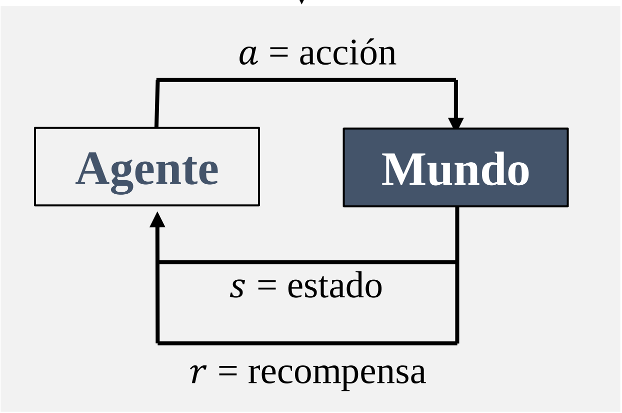
Supervisado

tipo

Por Refuerzo



estructura



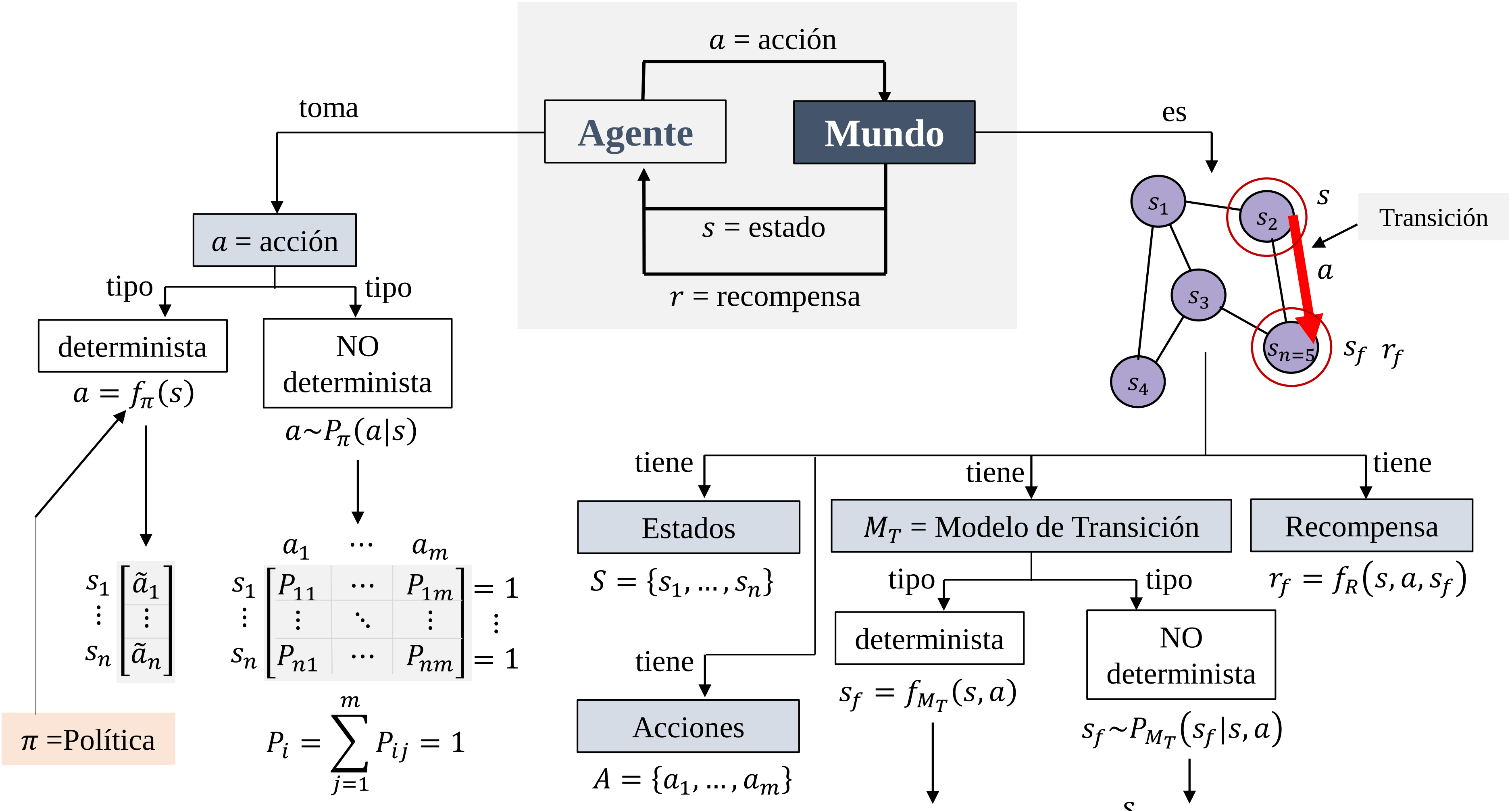
a = acción

Agente

Mundo

s = estado

r = recompensa



a = acción

Agente

Mundo

s = estado

r = recompensa

toma

a = acción

tipo

tipo

determinista

$a = f_\pi(s)$

NO determinista

$a \sim P_\pi(a|s)$

$$\begin{array}{c} s_1 \\ \vdots \\ s_n \end{array} \begin{bmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_n \end{bmatrix}$$

$$\begin{array}{c|ccc|c} & a_1 & \cdots & a_m & \\ s_1 & P_{11} & \cdots & P_{1m} & = 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ s_n & P_{n1} & \cdots & P_{nm} & = 1 \end{array}$$

$$P_i = \sum_{j=1}^{m} P_{ij} = 1$$

π =Política

es

tiene

Estados

$S = \{s_1, \dots, s_n\}$

tiene

Acciones

$A = \{a_1, \dots, a_m\}$

tiene

M_T = Modelo de Transición

tipo

tipo

determinista

$s_f = f_{M_T}(s, a)$

NO determinista

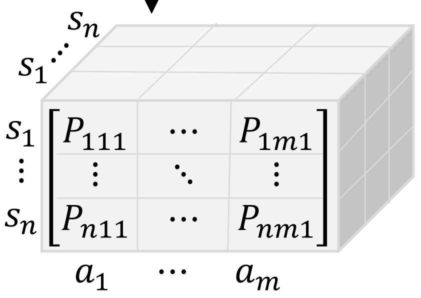
$s_f \sim P_{M_T}(s_f|s, a)$

tiene

Recompensa

$r_f = f_R(s, a, s_f)$

$$\begin{array}{c|ccc} & a_1 & \cdots & a_m \\ s_1 & s_{11} & \cdots & s_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & s_{n1} & \cdots & s_{nm} \end{array}$$



$$\begin{array}{c|ccc} s_1 & P_{111} & \cdots & P_{1m1} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & P_{n11} & \cdots & P_{nm1} \\ & a_1 & \cdots & a_m \end{array}$$

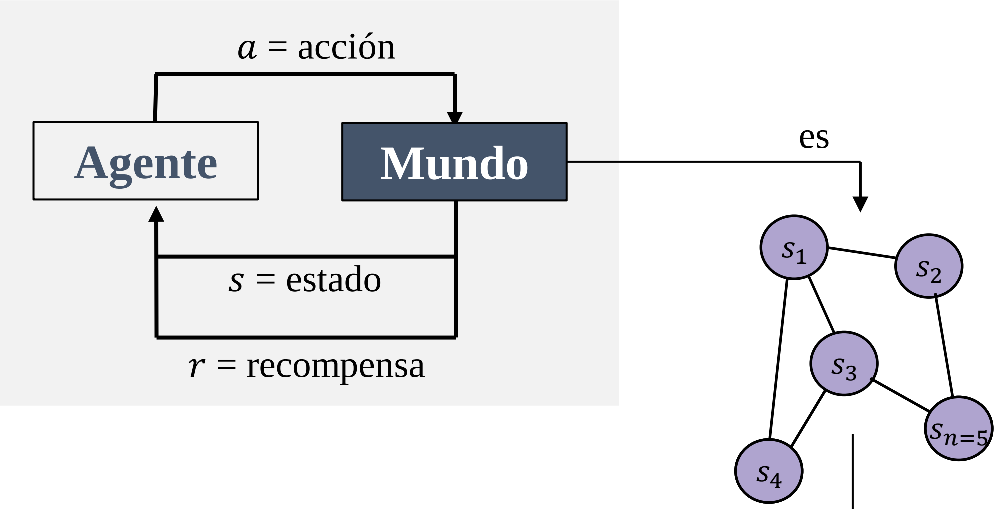
Variables:

$s, \tilde{s}, s_f, s_{ij} \in \{s_1, \dots, s_n\}$

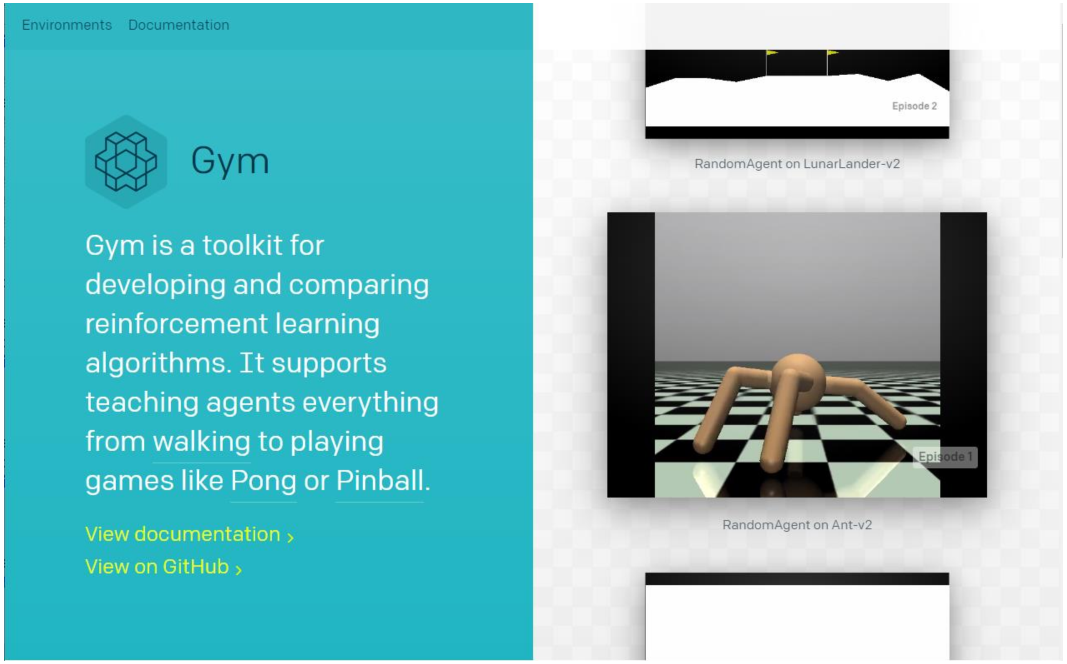
$a, \tilde{a}_i \in \{a_1, \dots, a_m\}$

$i \in \{1, \dots, n\}$

$j \in \{1, \dots, m\}$



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

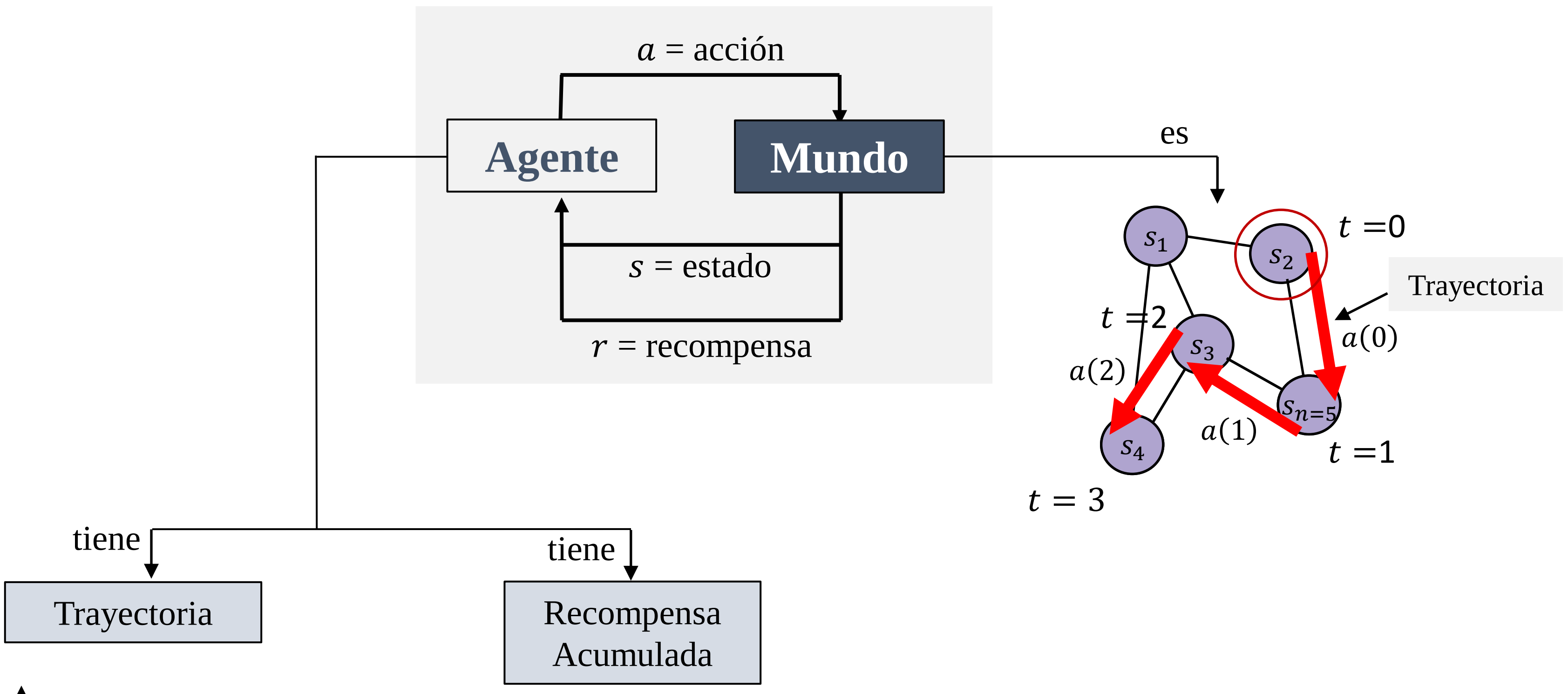


Gym

Gym is a toolkit for developing and comparing reinforcement learning algorithms. It supports teaching agents everything from walking to playing games like Pong or Pinball.

RandomAgent on LunarLander-v2

RandomAgent on Ant-v2



a = acción

Agente

Mundo

s = estado

r = recompensa

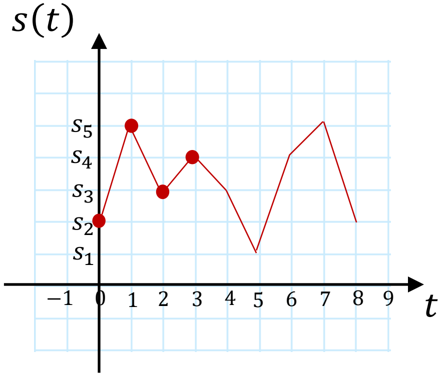
es

tiene

tiene

Trayectoria

Recompensa Acumulada



$$\tau = s(0), s(1), s(2), s(3), \ldots, s(T)$$

$$= s_2, s_5, s_3, s_4, \ldots, s_T$$

con $t \in \{0, \ldots, T\}$

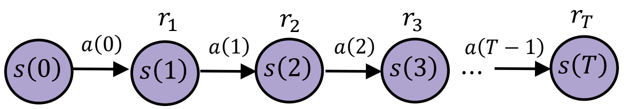
Trayectoria

$$f_{RA}(\tau) = \sum_{t=0}^{(T-1)} \gamma^t f_R(s(t), a(t), s(t+1))$$

con $r_{(t+1)} = f_R(s(t), a(t), s(t+1))$

$\gamma \in [0,1]$

$$= r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{(T-1)} r_T$$



NOTA: Convergencia

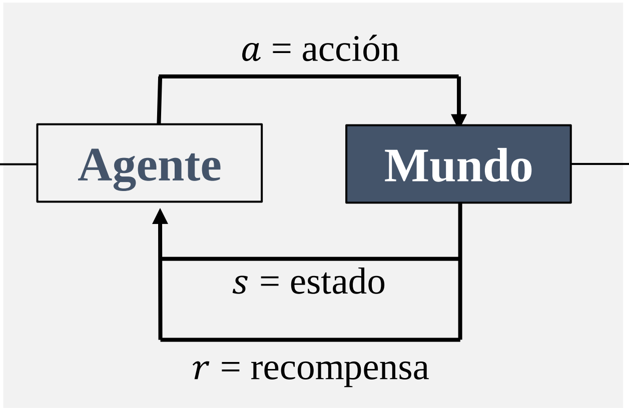
$$\sum_{n=0}^{N} a x^n = a\left[\frac{1 - x^{(N+1)}}{1 - x}\right], x \neq 1$$

$$= r_1 + \gamma[r_2 + \gamma r_3 + \cdots + \gamma^{(T-2)} r_T]$$

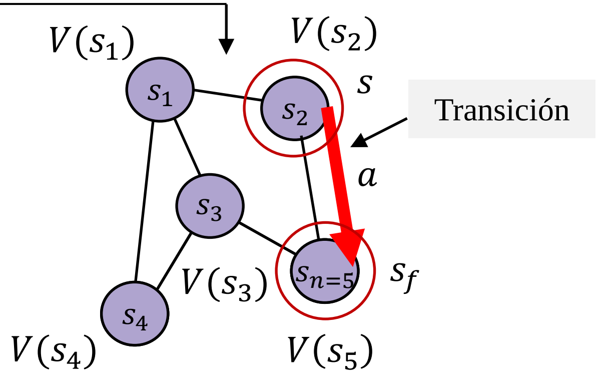
$$= r_1 + \gamma\left[\sum_{t=1}^{(T-1)} \gamma^{(t-1)} f_R(s(t), a(t), s(t+1))\right]$$

$$= r_1 + \gamma[f_{RA}(\tilde{\tau})]$$

con $\tilde{\tau} = s(1), s(2), s(3), \ldots, s(T)$



es



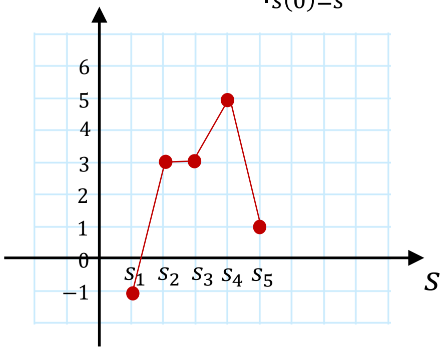
tiene

Recompensa Acumulada Promedio

tiene

Ecuación de Bellman

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{\tilde{s}(0)=s}$$



$$V(s) = \overline{f_{RA}(\tau)}\Big|_{\tilde{s}(0)=s}$$

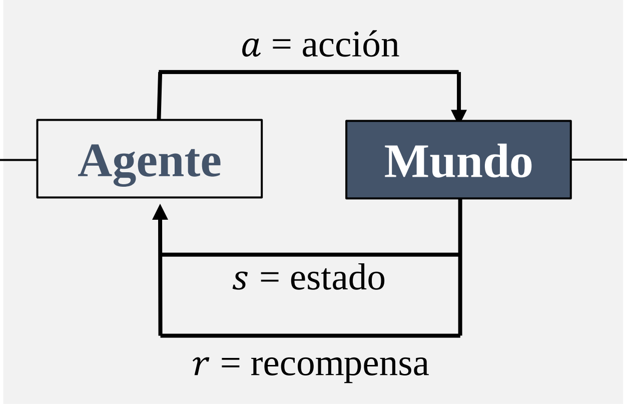
$$= \overline{f_R(s,a,s_f) + \gamma V(s_f)}$$

$$V(s) = \sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)]$$

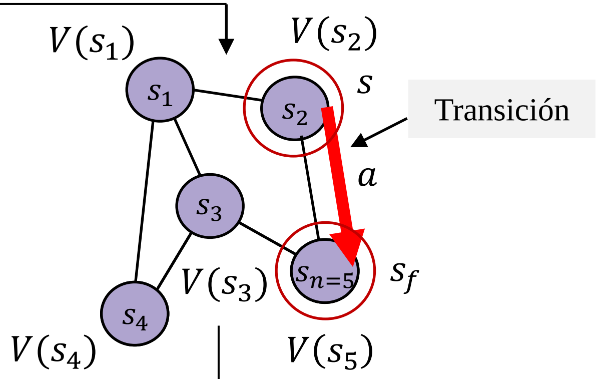
con: $a = f_\pi(s)$

$$= \sum_{s_f \in S} P_{M_T}\left(s_f|s,f_\pi(s)\right)[f_R(s,f_\pi(s),s_f) + \gamma V(s_f)]$$

EJEMPLO:



es



toma

a = acción

tipo

determinista

$a = f_\pi(s)$

3	→	→	→	+1
2	↑		↑	−1
1	↑	←	←	←
	1	2	3	4

da

Recompensa Acumulada Promedio

$$V(s) = \sum_{s_f \in S} P_{M_T}(s_f|s,a)\left[f_R(s,a,s_f) + \gamma V(s_f)\right]$$

$\gamma = 1$

3	0.8516	0.9078	0.9578	+1
2	0.8016		0.7003	−1
1	0.7453	0.6953	0.6514	0.4279
	1	2	3	4

tiene

Estados

$S = \{s_{\mathrm{I}}, s_{21}, \dots, s_{\mathrm{F1}}, s_{\mathrm{F2}}\}$

3	s_{13}	s_{23}	s_{33}	s_{F1} +1
2	s_{12}		s_{32}	s_{F2} −1
1	s_{I} START	s_{21}	s_{31}	s_{41}
	1	2	3	4

tiene

Acciones

$A = \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$

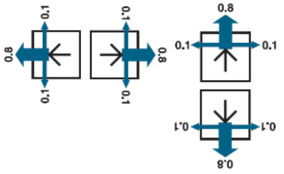
tiene

M_T = Modelo de Transición

tipo

NO determinista

$s_f \sim P_{M_T}(s_f|s,a)$

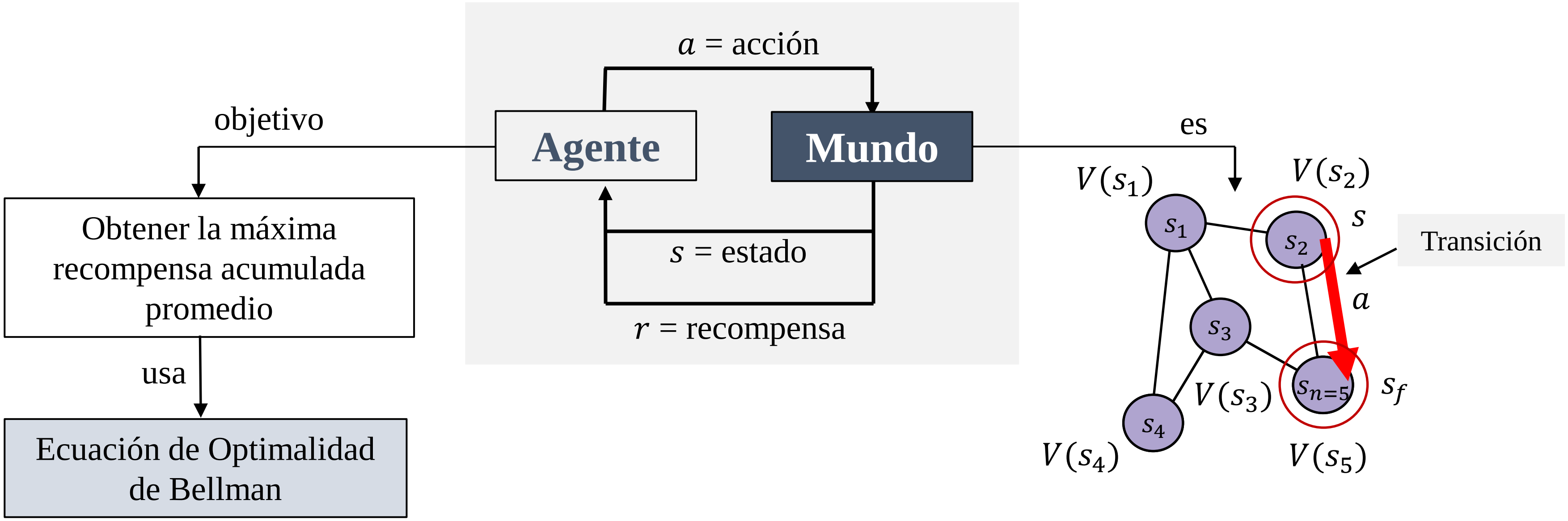


tiene

Recompensa

$r_f = f_R(s_f)$

3	-0.04	-0.04	-0.04	+1
2	-0.04		-0.04	−1
1	START	-0.04	-0.04	-0.04
	1	2	3	4



objetivo

Agente

Mundo

a = acción

s = estado

r = recompensa

es

Obtener la máxima recompensa acumulada promedio

usa

Ecuación de Optimalidad de Bellman

$$V(s) = \max_a \left[\overline{f_R(s,a,s_f) + \gamma V(s_f)}\right] = \max_a \left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)]\right]$$

definir

$$Q(s,a) = \left[\overline{f_R(s,a,s_f) + \gamma V(s_f)}\right]$$

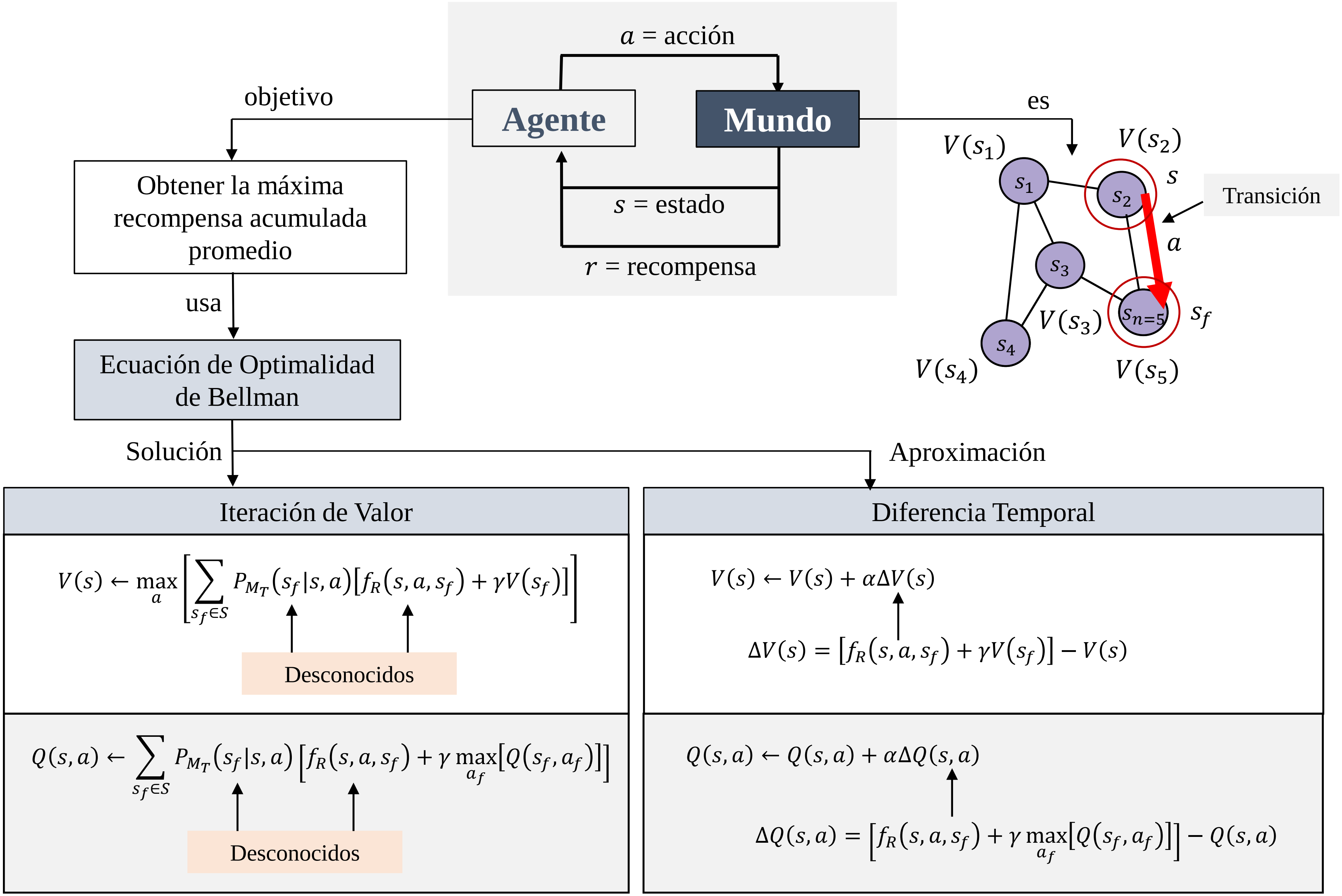
$$= \left[\overline{f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)]}\right] = \sum_{s_f \in S} P_{M_T}(s_f|s,a) \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)]\right]$$

$$V(s) = \max_a [Q(s,a)]$$

	V(s)
s1	V(s1)
s2	V(s2)
s3	V(s3)
s4	V(s4)
s(n=5)	V(s5)

⬌

		Q(s,a)	
	a1	a2	a(m=3)
s1	Q(s1,a1)	Q(s1,a2)	Q(s1,a3)
s2	Q(s2,a1)	Q(s2,a2)	Q(s2,a3)
s3	Q(s3,a1)	Q(s3,a2)	Q(s3,a3)
s4	Q(s4,a1)	Q(s4,a2)	Q(s4,a3)
s(n=5)	Q(s5,a1)	Q(s5,a2)	Q(s5,a3)



Obtener la máxima recompensa acumulada promedio

Ecuación de Optimalidad de Bellman

Iteración de Valor

$$V(s) \leftarrow \max_{a}\left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)\left[f_R(s,a,s_f) + \gamma V(s_f)\right]\right]$$

Desconocidos

$$Q(s,a) \leftarrow \sum_{s_f \in S} P_{M_T}(s_f|s,a)\left[f_R(s,a,s_f) + \gamma \max_{a_f}\left[Q(s_f,a_f)\right]\right]$$

Desconocidos

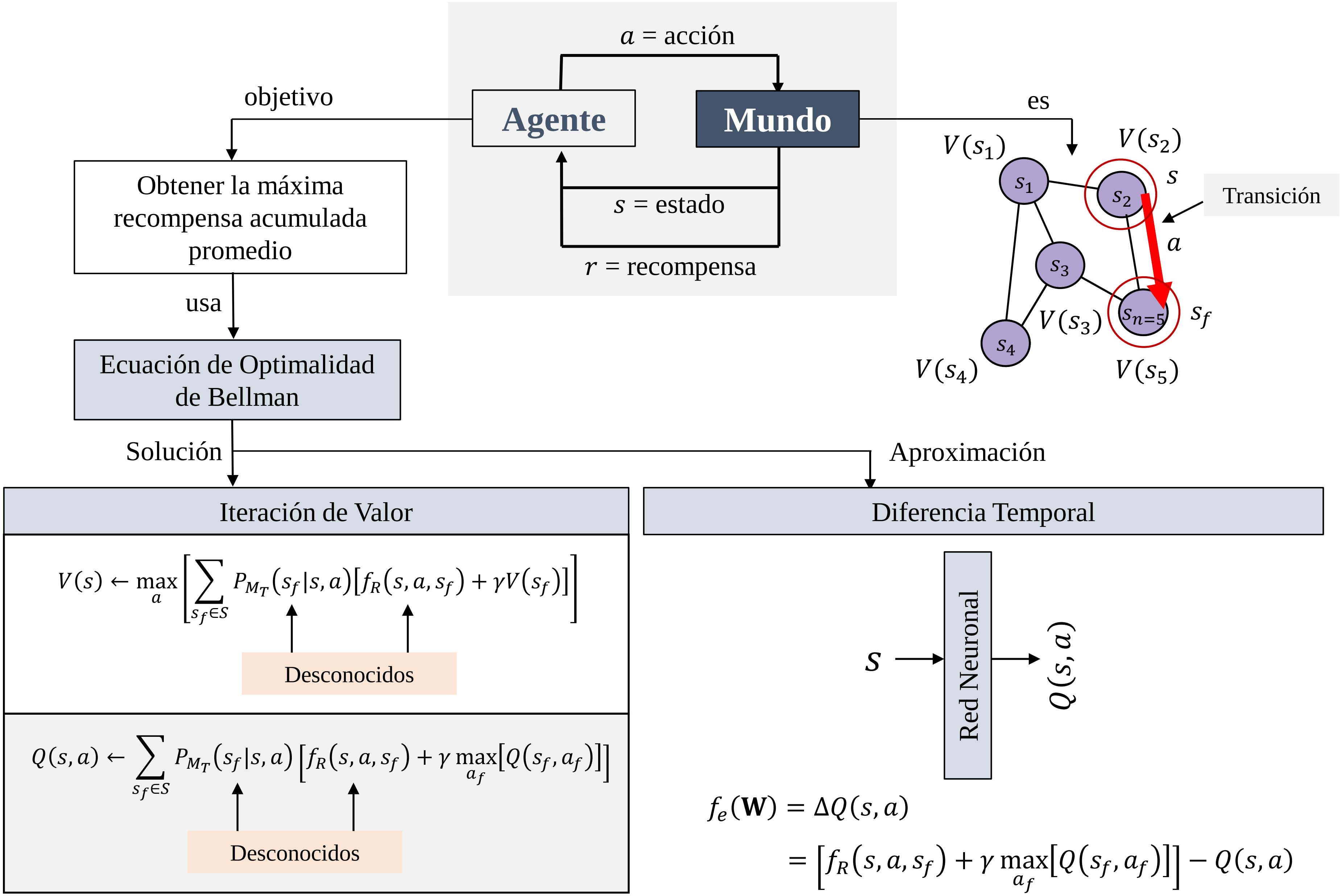
Diferencia Temporal

$$V(s) \leftarrow V(s) + \alpha \Delta V(s)$$

$$\Delta V(s) = \left[f_R(s,a,s_f) + \gamma V(s_f)\right] - V(s)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha \Delta Q(s,a)$$

$$\Delta Q(s,a) = \left[f_R(s,a,s_f) + \gamma \max_{a_f}\left[Q(s_f,a_f)\right]\right] - Q(s,a)$$



Obtener la máxima recompensa acumulada promedio

Ecuación de Optimalidad de Bellman

Iteración de Valor

$$V(s) \leftarrow \max_{a} \left[\sum_{s_f \in S} P_{M_T}(s_f|s,a)[f_R(s,a,s_f) + \gamma V(s_f)] \right]$$

Desconocidos

$$Q(s,a) \leftarrow \sum_{s_f \in S} P_{M_T}(s_f|s,a) \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right]$$

Desconocidos

Diferencia Temporal

s → Red Neuronal → $Q(s,a)$

$$f_e(\mathbf{W}) = \Delta Q(s,a)$$
$$= \left[f_R(s,a,s_f) + \gamma \max_{a_f}[Q(s_f,a_f)] \right] - Q(s,a)$$

APENDICE: Ecuaciones

Recompensa Acumulada Promedio

$$V(s) = \overline{f_{RA}(\tau)}\Big|_{\tilde{s}(0)=s}$$

$$= \frac{1}{N}\sum_{k=1}^{N} f_{RA}(\tau_k)\Big|_{\widetilde{s_k}(0)=s}$$

$$= \frac{1}{N}\left[f_{RA}(\tau_1)\Big|_{\widetilde{s_1}(0)=s} + \cdots + f_{RA}(\tau_N)\Big|_{\widetilde{s_N}(0)=s} \right]$$

con $f_{RA}(\tau_k)|_{\widetilde{s_k}(0)=s} = r_0^k + \gamma r_1^k + \gamma^2 r_2^k + \cdots + \gamma^T\ r_{T_k}^k$

$$\begin{aligned}
= \; & \frac{1}{N}\left[r_0^1 + \gamma r_1^1 + \gamma^2 r_2^1 + \gamma^3 r_3^1 + \cdots + \gamma^{T_1}\ r_{T_1}^1\right] + \\
& \frac{1}{N}\left[r_0^2 + \gamma r_1^2 + \gamma^2 r_2^2 + \gamma^3 r_3^2 + \cdots + \gamma^{T_2}\ r_{T_2}^2\right] + \\
& \vdots \\
& \frac{1}{N}\left[r_0^N + \gamma r_1^N + \gamma^2 r_2^N + \gamma^3 r_3^N + \cdots + \gamma^{T_N}\ r_{T_N}^N\right]
\end{aligned}$$

$$\begin{aligned}
= \; & \frac{1}{N}[r_0^1] + \frac{1}{N}\gamma\left[r_1^1 + \gamma^1 r_2^1 + \gamma^2 r_3^1 + \cdots + \gamma^{(T_1-1)}\ r_{T_1}^1\right] + \\
& \frac{1}{N}[r_0^2] + \frac{1}{N}\gamma\left[r_1^2 + \gamma^1 r_2^2 + \gamma^2 r_3^2 + \cdots + \gamma^{(T_2-1)}\ r_{T_2}^2\right] + \\
& \vdots \\
& \frac{1}{N}[r_0^N] + \frac{1}{N}\gamma\left[r_1^N + \gamma^1 r_2^N + \gamma^2 r_3^N + \cdots + \gamma^{(T_N-1)}\ r_{T_N}^N\right]
\end{aligned}$$

$$\begin{aligned}
\tau_1 &= \widetilde{s_1}(0), \widetilde{s_1}(1), \widetilde{s_1}(2), \widetilde{s_1}(3), \ldots, \widetilde{s_1}(T_1) \\
\tau_2 &= \widetilde{s_2}(0), \widetilde{s_2}(1), \widetilde{s_2}(2), \widetilde{s_2}(3), \ldots, \widetilde{s_2}(T_2) \\
& \vdots \\
\tau_N &= \widetilde{s_N}(0), \widetilde{s_N}(1), \widetilde{s_N}(2), \widetilde{s_N}(3), \ldots, \widetilde{s_N}(T_N)
\end{aligned}$$

$$\begin{aligned}
\tau_1 &= \widetilde{s_1}(1), \widetilde{s_1}(2), \widetilde{s_1}(3), \ldots, \widetilde{s_1}(T_1) \\
\tau_2 &= \widetilde{s_2}(1), \widetilde{s_2}(2), \widetilde{s_2}(3), \ldots, \widetilde{s_2}(T_2) \\
& \vdots \\
\tau_N &= \widetilde{s_N}(1), \widetilde{s_N}(2), \widetilde{s_N}(3), \ldots, \widetilde{s_N}(T_N)
\end{aligned}$$

$$\begin{aligned}
= \; & \frac{1}{N}[r_0^1] + \frac{1}{N}\gamma\left[f_{RA}(\tau_1)\Big|_{\widetilde{s_1}(0)=\widetilde{s_1}(1)} = r_1^1 + \gamma^1 r_2^1 + \gamma^2 r_3^1 + \cdots + \gamma^{(T_1-1)}\ r_{T_1}^1\right] + \\
& \frac{1}{N}[r_0^2] + \frac{1}{N}\gamma\left[r_1^2 + \gamma^1 r_2^2 + \gamma^2 r_3^2 + \cdots + \gamma^{(T_2-1)}\ r_{T_2}^2\right] + \\
& \vdots \\
& \frac{1}{N}[r_0^N] + \frac{1}{N}\gamma\left[r_1^N + \gamma^1 r_2^N + \gamma^2 r_3^N + \cdots + \gamma^{(T_N-1)}\ r_{T_N}^N\right]
\end{aligned}$$