

3

---

# FUNDAMENTOS DEL APRENDIZAJE POR REFUERZO (2)

## VALOR DE UN ESTADO

- ▶ Si lanzamos un dado de seis caras, el **valor esperado** del resultado es:

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$



- ▶ Esto significa que, en promedio, si lanzamos el dado muchas veces, el **valor esperado** del resultado será 3.5.
- ▶ **Esperanza matemática:** es un valor que indica el resultado promedio que se espera obtener de una variable aleatoria.

$$\mathbb{E}[X] = \sum_i x_i P(x_i)$$

## VALOR DE UN ESTADO



Fuente imagen: Cast away, 2000

## VALOR DE UN ESTADO

- ▶ Ir a la derecha (30% de probabilidad) → Destino: una ciudad con una recompensa de 4 puntos.
- ▶ Ir a la izquierda (70% de probabilidad) → Destino: un bosque con una recompensa de 10 puntos.
- ▶ ¿Cuánto vale este cruce de caminos? No hay una única respuesta, porque depende de a dónde se elija ir.
- ▶ Una forma de responder es calcular el **valor esperado**, ponderando cada recompensa por su probabilidad:

$$V(\text{Cruce}) = 0.3(4) + 0.7(10) = 1.2 + 7 = 8.2$$

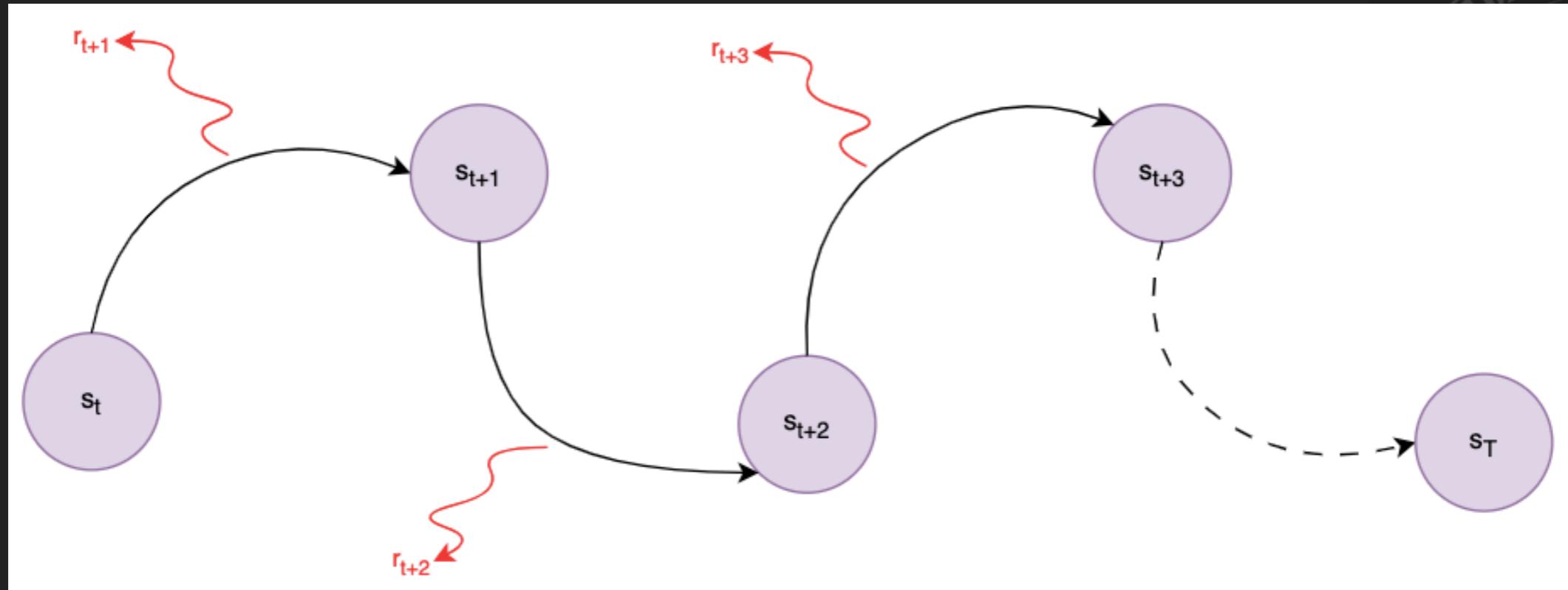
- ▶ ∴ el valor de este estado es un promedio ponderado de los posibles futuros.



Fuente imagen: Cast away, 2000

## VALOR DE UN ESTADO

- ▶ ¿Cómo sé “cuán bueno” es un estado?



## VALOR DE UN ESTADO

- La manera más simple de medir el valor de un estado de un MDP es obtener la recompensa inmediata que recibe el agente por moverse al estado siguiente  $s_{t+1}$ .

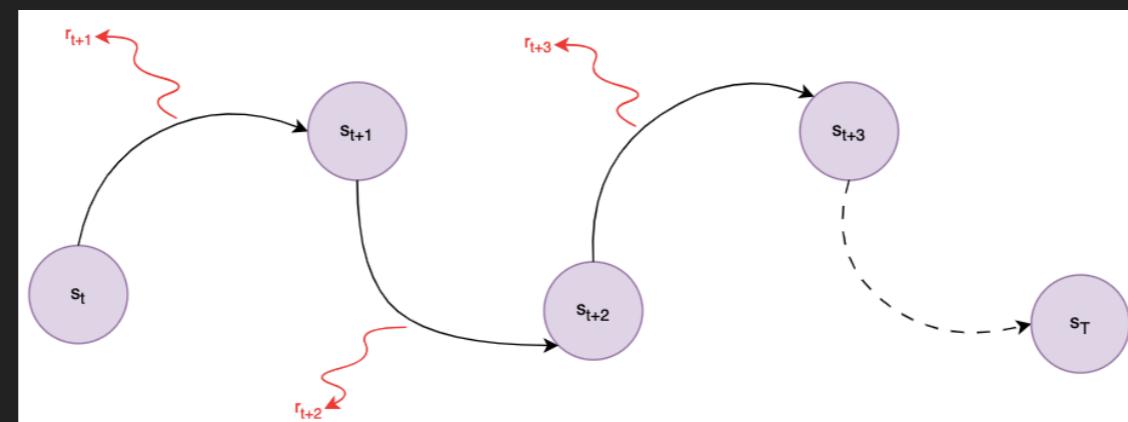
$$V(s_t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots + r_T$$

$$V(s_{t+1}) = r_{t+2} + \gamma r_{t+3} + \gamma^2 r_{t+4} + \dots + r_T$$

- Identificamos el patrón recursivo. El término  $\gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$  es muy similar a la función de valor del siguiente estado  $s_{t+1}$
- Podemos sustituir  $V(s_{t+1})$  en la expresión  $V(s_t)$

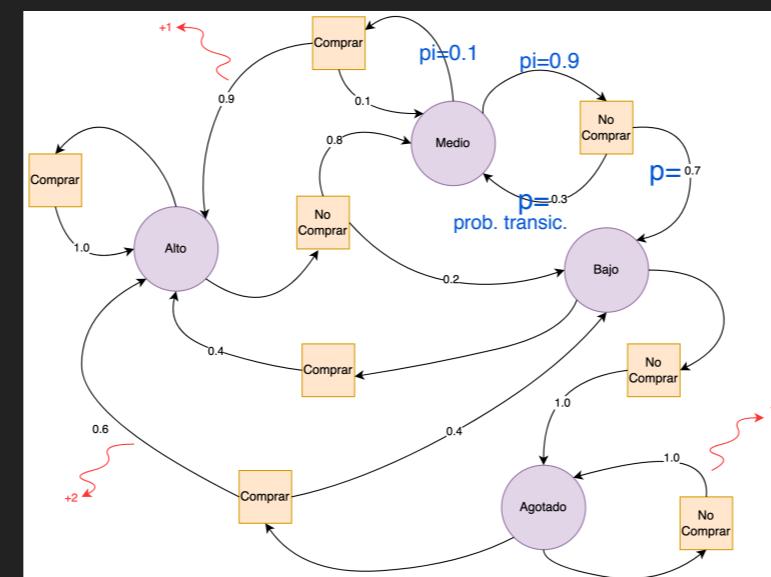
$$V(s_t) = r_{t+1} + \gamma * [r_{t+2} + \gamma r_{t+3} + \dots + r_T]$$

$$V(s_t) = r_{t+1} + \gamma * V(s_{t+1})$$

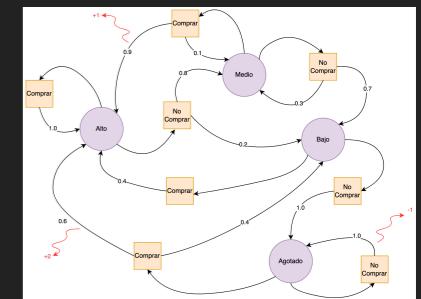


## VALOR DE UN ESTADO

- ▶ Si el agente se encuentra en un **estado  $s$** , la **política  $\pi$**  es la estrategia o regla de cual **acción** decide tomar el agente desde ese estado  $s$ .
- ▶ Una política podría ser elegir la mejor acción posible desde un estado  $s$ .



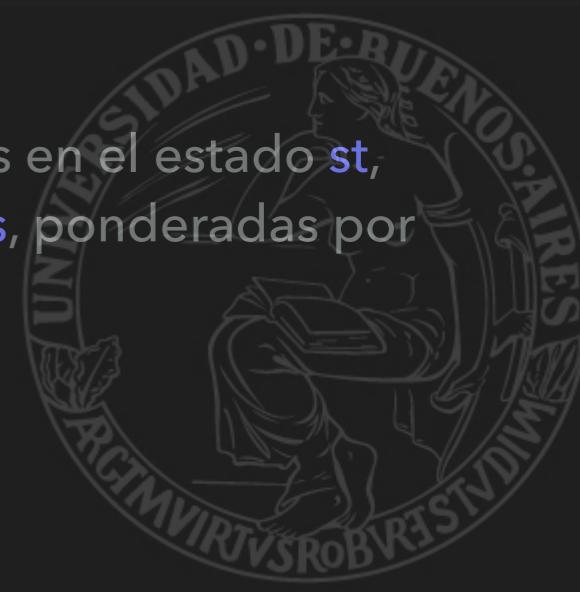
# VALOR DE UN ESTADO



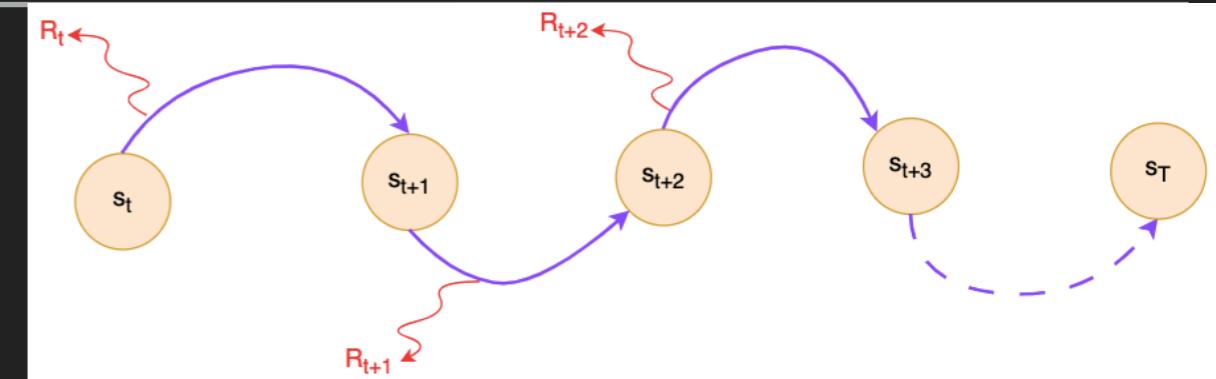
- ▶ Dado que estamos en un entorno incierto o de incertidumbre (porque no podemos predecir el estado siguiente debido a la existencia de probabilidades) el **valor de un estado** se puede calcular a partir del **valor esperado (esperanza)** por emplear una **política  $\pi$** .
- ▶ La **esperanza** representa el valor promedio que podemos esperar obtener si estamos en el estado  $s_t$ , seguimos la **política  $\pi$** , y consideramos todas las **posibles transiciones y recompensas**, ponderadas por sus probabilidades.

$$V_{\pi}(s_t) = \mathbb{E}_{\pi}[r_t + \gamma V_{\pi}(s_{t+1})]$$

- ▶ El **valor de estar en el estado  $s_t$  siguiendo la **política  $\pi$****  es igual al **valor esperado** de la **recompensa inmediata  $r_t$**  mas el **valor descontado** del siguiente estado  $s_{t+1}$ , también evaluado siguiendo la **política  $\pi$** .
- ▶ El **valor esperado** bajo la **política  $\pi$**  implica promediar sobre todas las posibles acciones que la **política  $\pi$**  podría recomendar en ese **estado**, y sobre todos los posibles **estados siguientes** que se podría alcanzar al tomar esas **acciones**.
- ▶ Esto se conoce como **Ecuación de Bellman para la evaluación de una política en un estado  $s_t$** .



## VALOR DE UN ESTADO



- ▶ Demostración detallada de la **ecuación de Bellman** para determinar el valor de un estado  $s$ :
- ▶ 1. El retorno total ( $G_t$ ) es la suma de recompensas futuras:

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

- ▶ 2. Si consideramos la **tasa de descuento** obtenemos la suma de recompensas futuras descontadas:

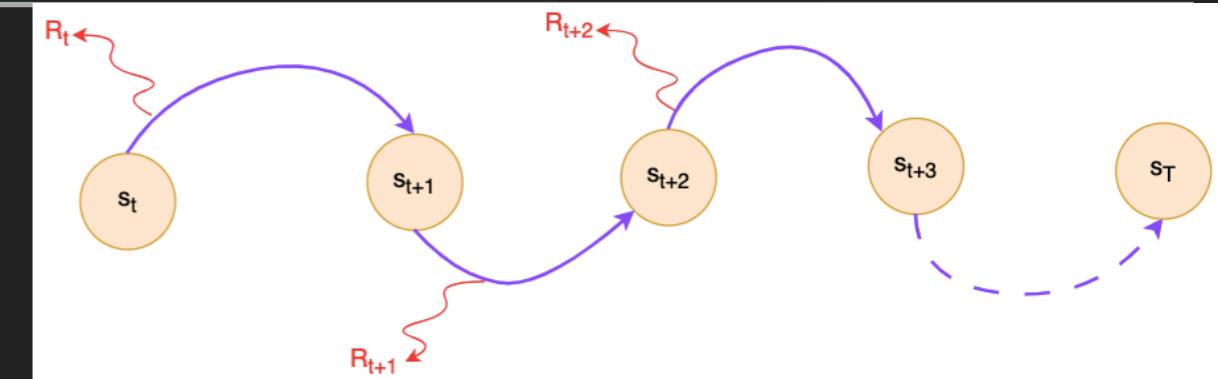
$$G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \dots$$

$$G_{t+1} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \dots$$

$$G_t = R_t + \gamma G_{t+1}$$



## VALOR DE UN ESTADO



- 3. Se aplica esperanza condicional a ambos miembros de la ecuación respecto a  $s_t = s$ :

$$G_t = R_t + \gamma G_{t+1} \quad \mathbb{E}[G_t | s_t = s] = \mathbb{E}[R_t + \gamma G_{t+1} | s_t = s]$$

- 4. Por propiedad de linealidad de la esperanza:

$$\mathbb{E}[G_t | s_t = s] = \mathbb{E}[R_t | s_t = s] + \gamma \mathbb{E}[G_{t+1} | s_t = s]$$

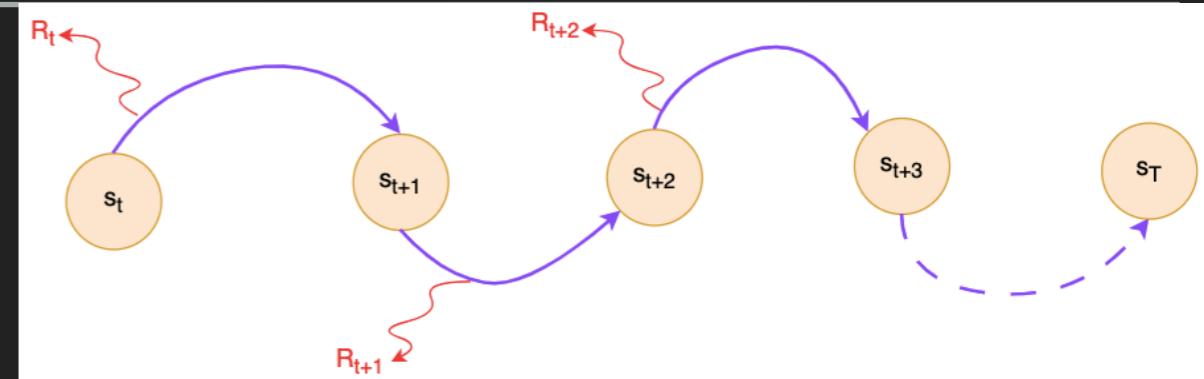
- 5. Por definición de valor de estado:

$$V(s) = \mathbb{E}[G_t | s_t = s]$$

- 6. Sustituyendo el segundo miembro en la ecuación anterior queda  $V(s)$ :

$$V(s) = \mathbb{E}[R_t | s_t = s] + \gamma \mathbb{E}[G_{t+1} | s_t = s]$$

## VALOR DE UN ESTADO



- 7. En este punto la esperanza  $E[G_{t+1}]$  debería estar expresada en términos de  $V(s)$ , por lo cual sabiendo que:

$$V(s_{t+1}) = \mathbb{E}[G_{t+1} | s_{t+1}]$$

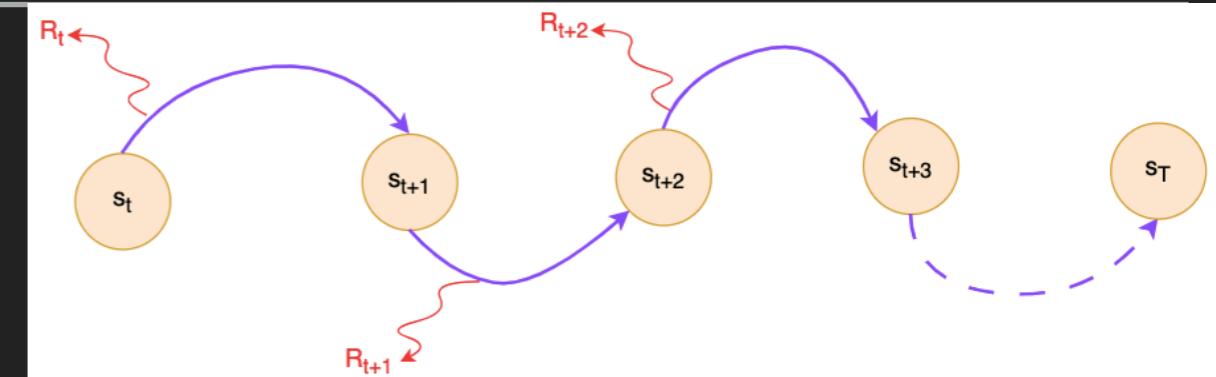
- 8. Se aplica al segundo término  $E[G_{t+1}]$  la regla de la esperanza iterada que establece que dadas 2 variables aleatorias  $X$  e  $Y$ , y a su vez dada una función  $h(X, Y)$ , la esperanza de  $h$  es igual a la esperanza condicional (anidada o iterada) respecto a ambas variables aleatorias  $X$  e  $Y$ :

$$\mathbb{E}[h(X, Y)] = \mathbb{E}\{\mathbb{E}[h(X, Y)|X]\}$$

- En nuestro caso, la esperanza iterada de  $E[G_{t+1}]$  es:

$$\mathbb{E}[G_{t+1} | s_t = s] = \mathbb{E}[\mathbb{E}[G_{t+1} | s_{t+1}] | s_t = s]$$

# VALOR DE UN ESTADO



► 9. Sustituyendo la esperanza  $\mathbb{E}[G_{t+1}]$  queda:

$$V(s_{t+1}) = \mathbb{E}[G_{t+1} | s_{t+1}]$$

$$\mathbb{E}[G_{t+1} | s_t = s] = \mathbb{E}[\mathbb{E}[G_{t+1} | s_{t+1}] | s_t = s]$$

$$\mathbb{E}[G_{t+1} | s_t = s] = \mathbb{E}[V(s_{t+1}) | s_t = s]$$

► Teníamos que  $V(s)$  era:

$$V(s) = \mathbb{E}[R_t | s_t = s] + \gamma \mathbb{E}[G_{t+1} | s_t = s]$$

$$V(s) = \mathbb{E}[R_t | s_t = s] + \gamma \mathbb{E}[V(s_{t+1}) | s_t = s]$$

$$V_\pi(s_t) = \mathbb{E}_\pi[r_t + \gamma V_\pi(s_{t+1})]$$

► 10. Expandiendo la esperanza la **ecuación de Bellman** queda expresada como:

$$V(s) = \sum_{s'} P(s' | s, a) R(s, a, s') + \gamma \sum_{s'} P(s' | s, a) V(s')$$

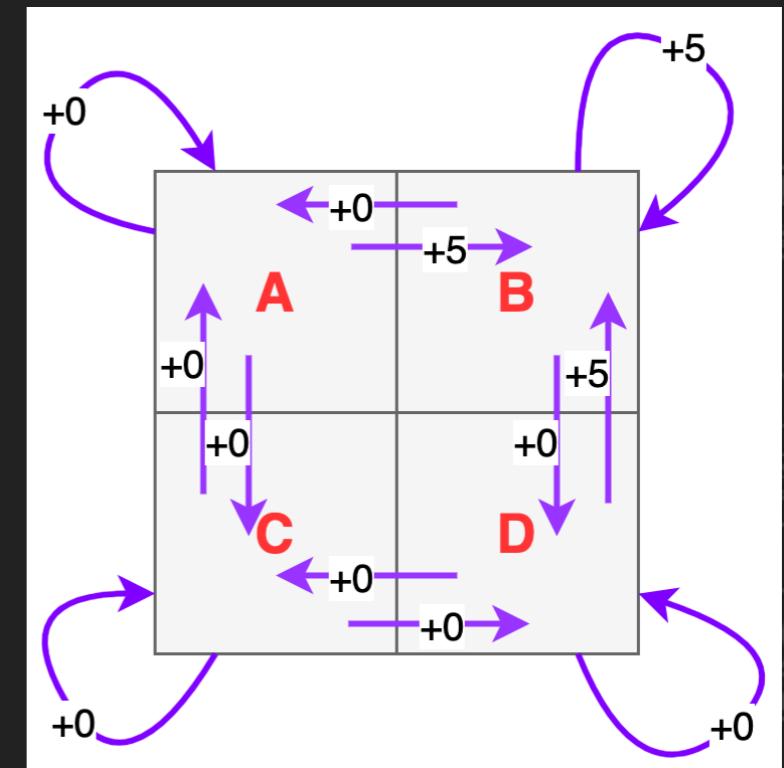
$$V(s) = \sum_{s'} P(s' | s, a) [R(s, a, s') + \gamma V(s')]$$

$$v_\pi(s) = \sum_a \pi(a | s) \sum_{s', r} p(s', r | s, a) [r + \gamma v_\pi(s')],$$

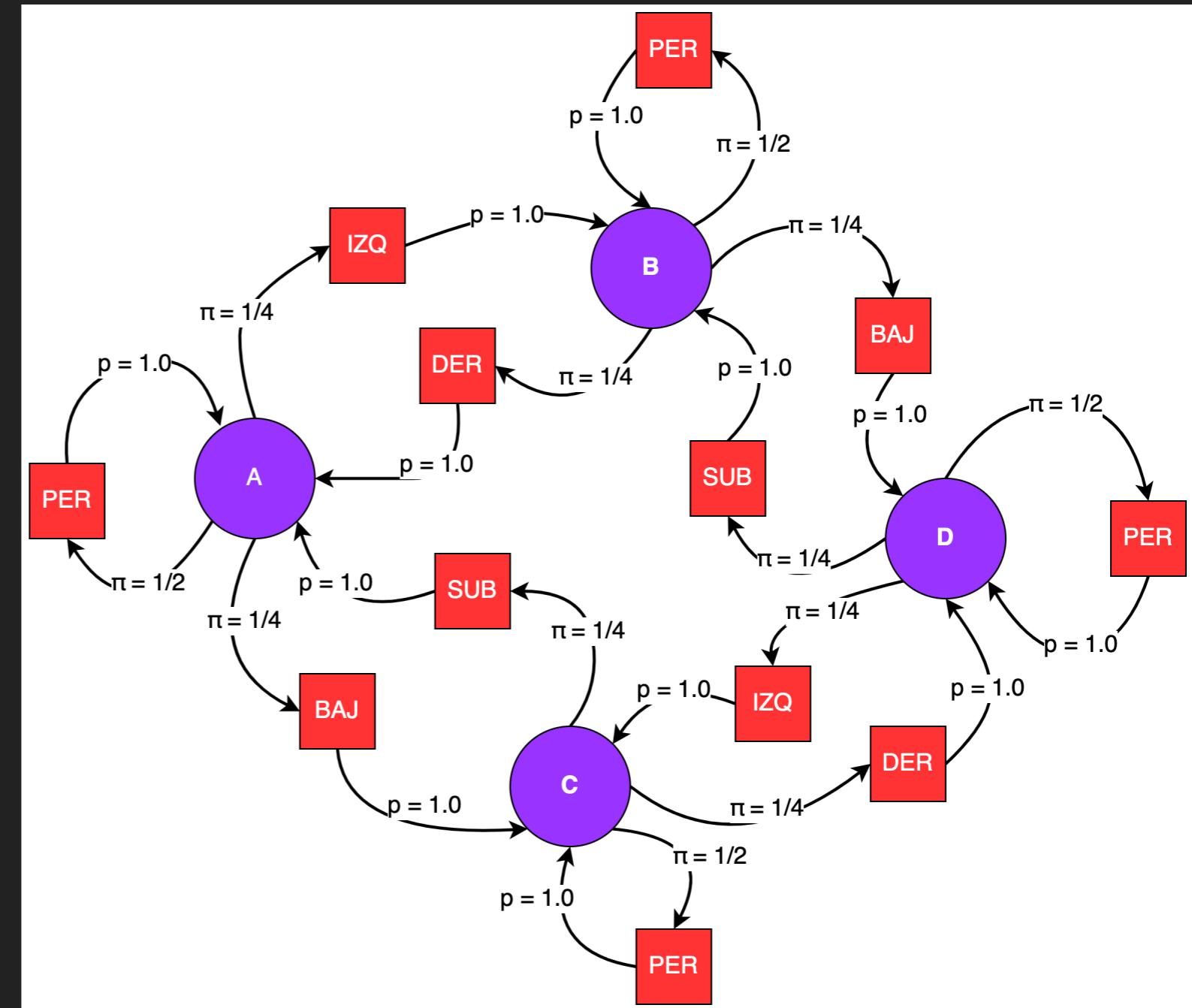
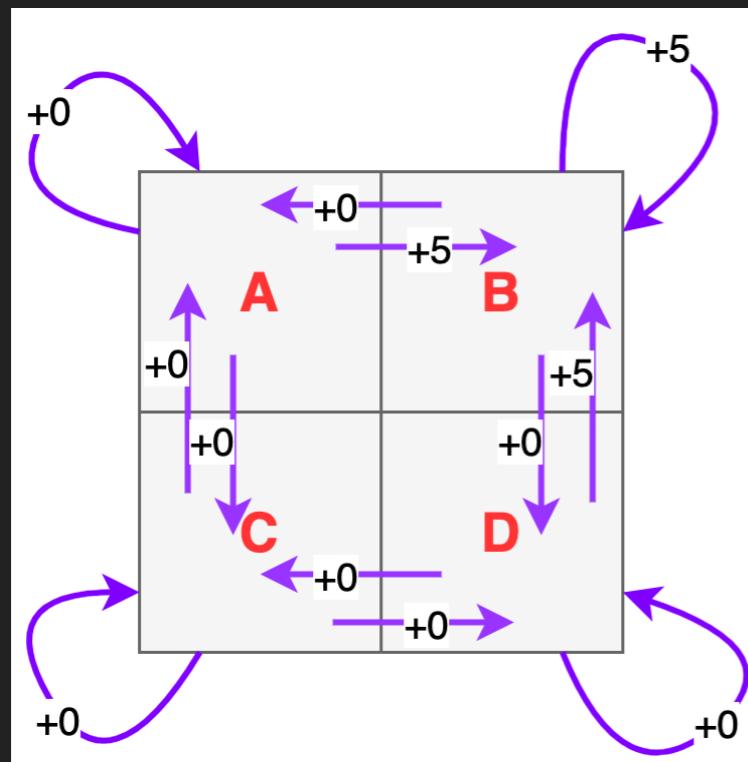
# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

- ▶ Ejemplo 1: Gridworld de 2x2
- ▶ Factor de descuento = 0.7

$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_{\pi}(s')],$$



# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN



# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

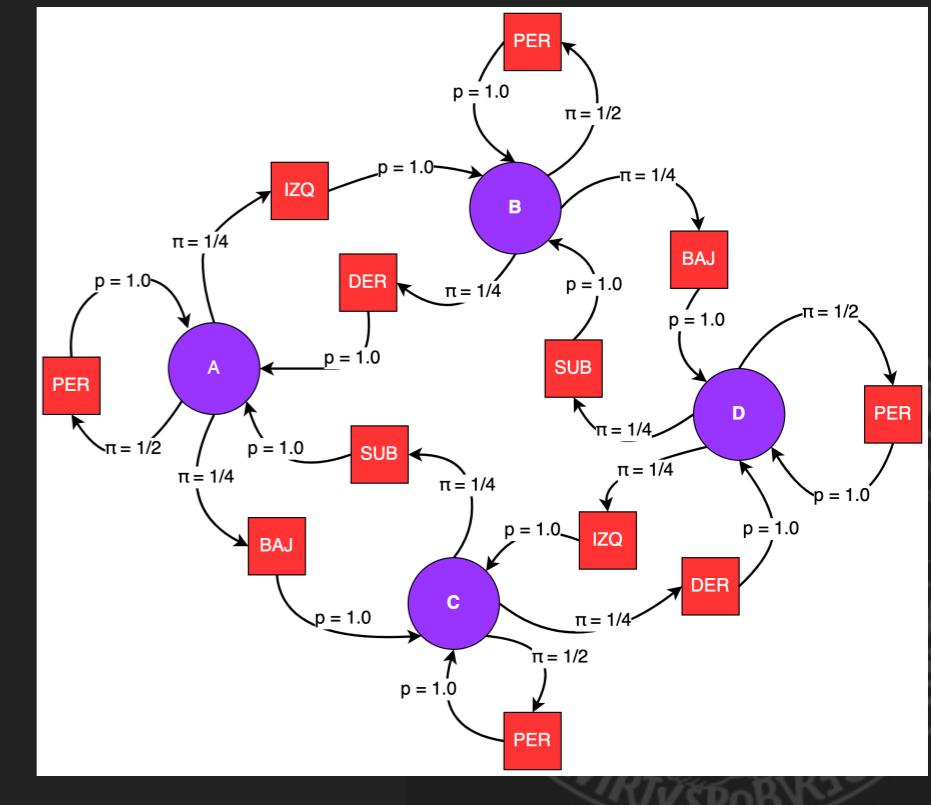
$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_{\pi}(s')],$$

$$v_{\pi}(A) = \pi(a|A)[r + 0.7v_{\pi}(s')]$$

$$v_{\pi}(B) = \pi(a|B)[r + 0.7v_{\pi}(s')]$$

$$v_{\pi}(A) = \frac{1}{4} [5 + 0.7v_{\pi}(B)] + \frac{1}{4} [0 + 0.7v_{\pi}(C)] + \frac{1}{2^4} [0 + 0.7v_{\pi}(A)]$$

$$v_{\pi}(B) = \frac{1}{4} [0 + 0.7v_{\pi}(A)] + \frac{1}{4} [0 + 0.7v_{\pi}(D)] + \frac{1}{4} [5 + 0.7v_{\pi}(B)]$$



# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

$$v_{\pi}(A) = \frac{1}{4} [5 + 0.7v_{\pi}(B)] + \frac{1}{4} [0 + 0.7v_{\pi}(C)] + \frac{1}{4} [0 + 0.7v_{\pi}(A)]$$

- ▶ De forma similar podemos calcular  $V(C)$  y  $V(D)$ .
- ▶ Tenemos por lo tanto 4 ecuaciones con 4 incógnitas.

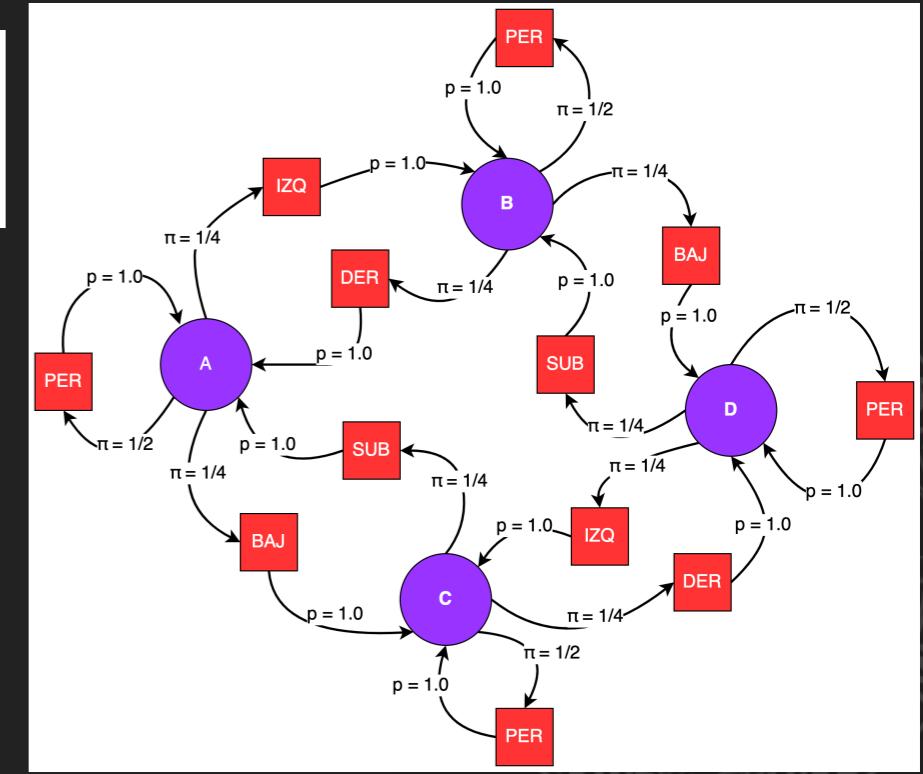
✓ A: 4.1666

✓ B: 6.0897

✓ C: 2.2436

✓ D: 4.1666

- ▶ Estado B: 6.0897: El hecho de que el valor  $V(B)$  sea más alto que el resto indica que estar en B (y seguir la política prefijada) es más beneficioso a largo plazo.

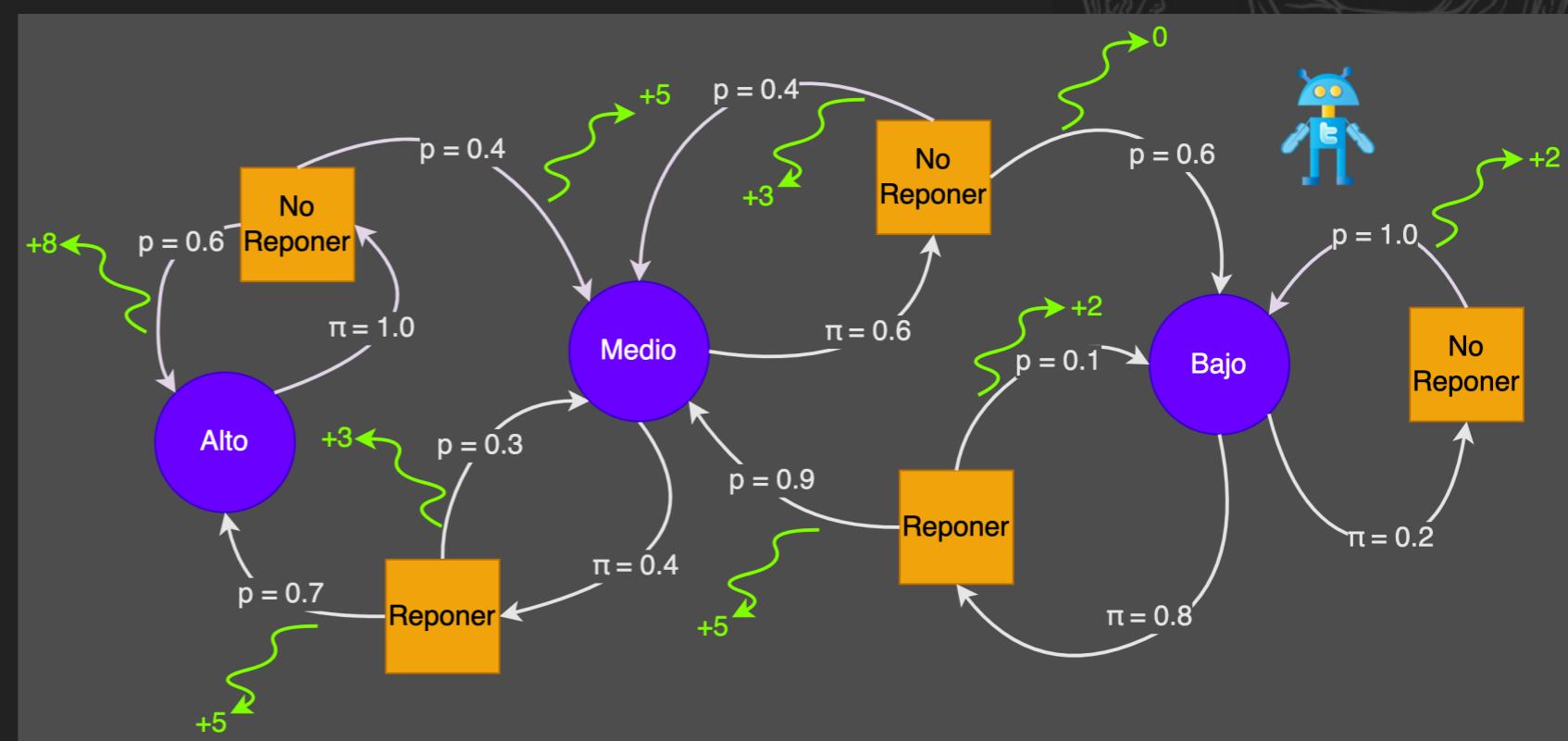
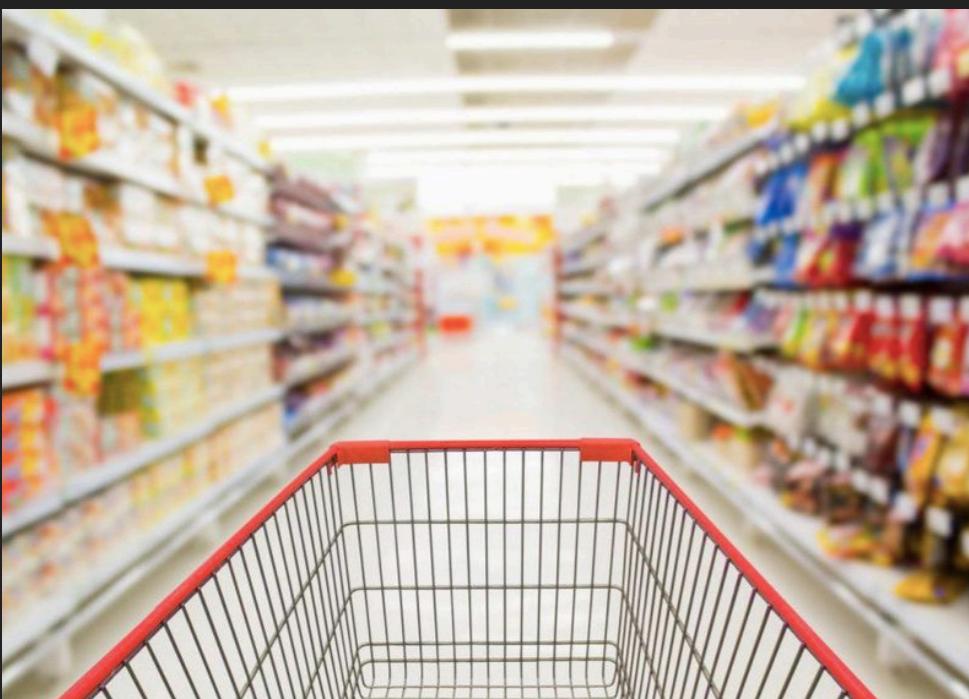


# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

- ▶ (Sutton et al., 2018)

$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_{\pi}(s')],$$

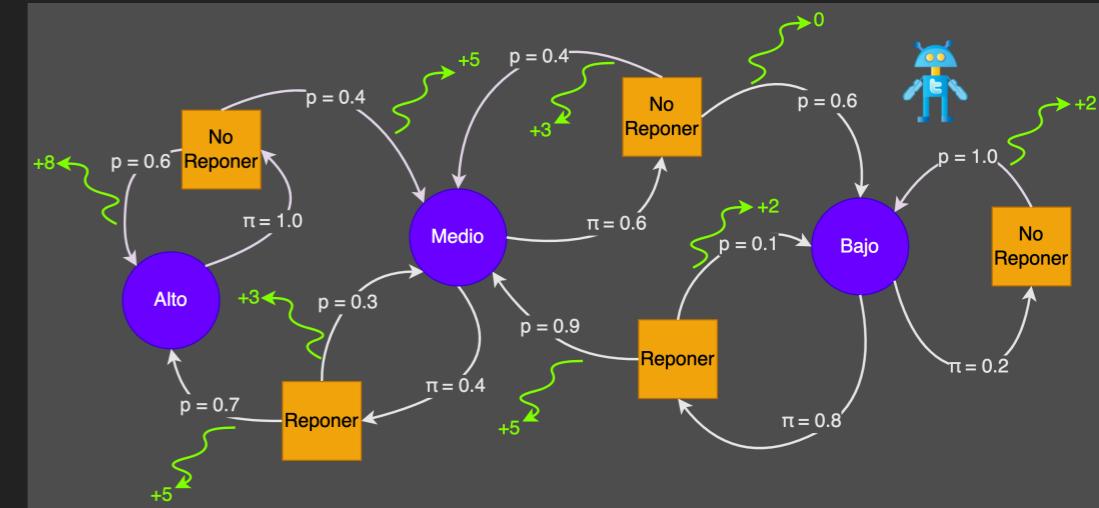
- ▶ Ejemplo: Robot repositor.



# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

- ▶ (Sutton et al., 2018)

$$v_{\pi}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_{\pi}(s')],$$

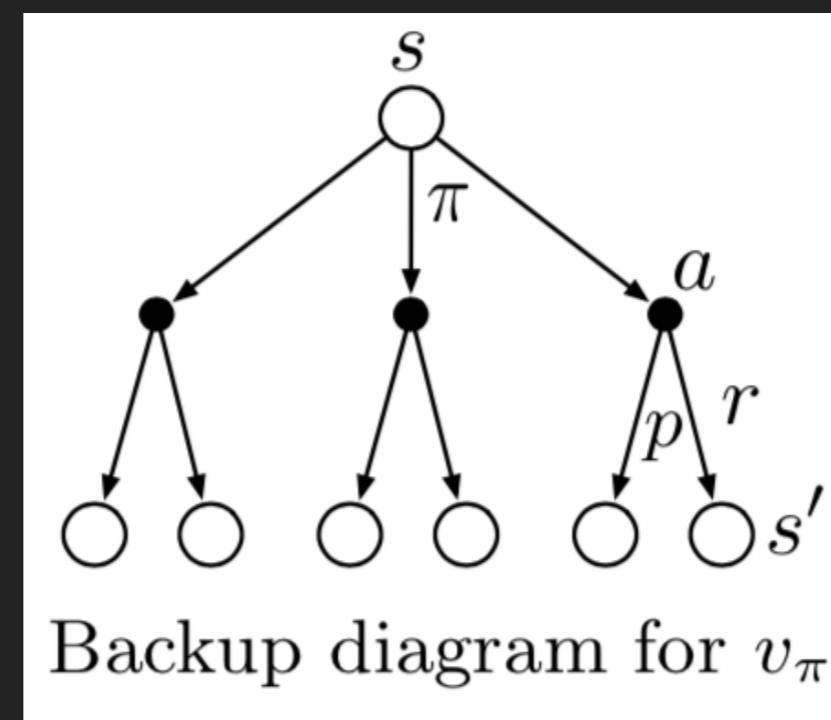
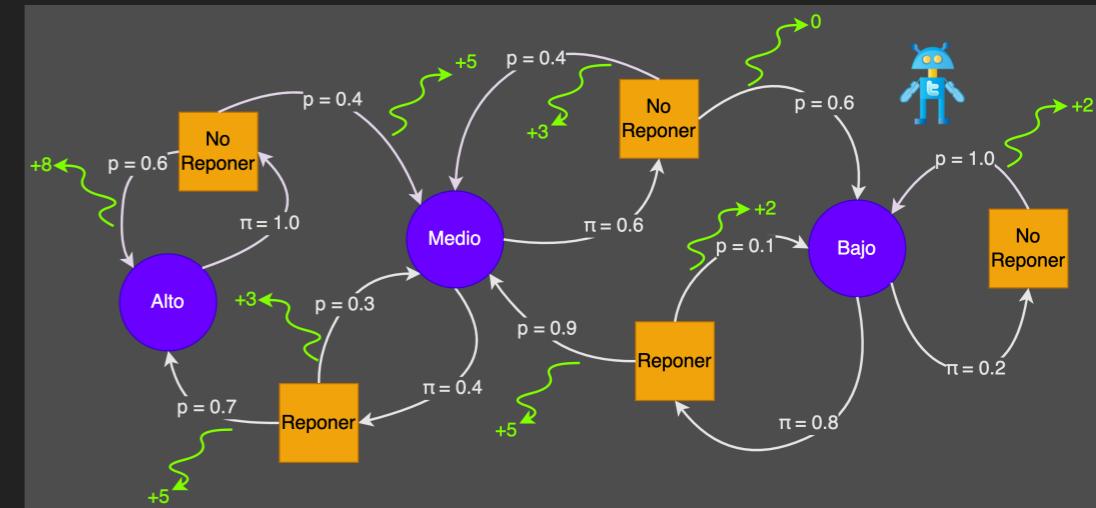


- ▶ La política ( $\pi$ ) que define la probabilidad de tomar cada acción en cada estado.
- ▶ La dinámica del entorno ( $p(s', r | s, a)$ ) define las probabilidades de transición de un estado a otro ( $s \rightarrow s'$ ) y las recompensas.
- ▶ Un MDP se define por una tupla  $M = \langle S, A, \Phi, R \rangle$ , o bien  $M = \langle S, A, T, R \rangle$ .

# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

► (Sutton et al., 2018)

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_\pi(s')] ,$$



# INTERPRETACIÓN DE LA ECUACIÓN DE BELLMAN

## ► Resultados:

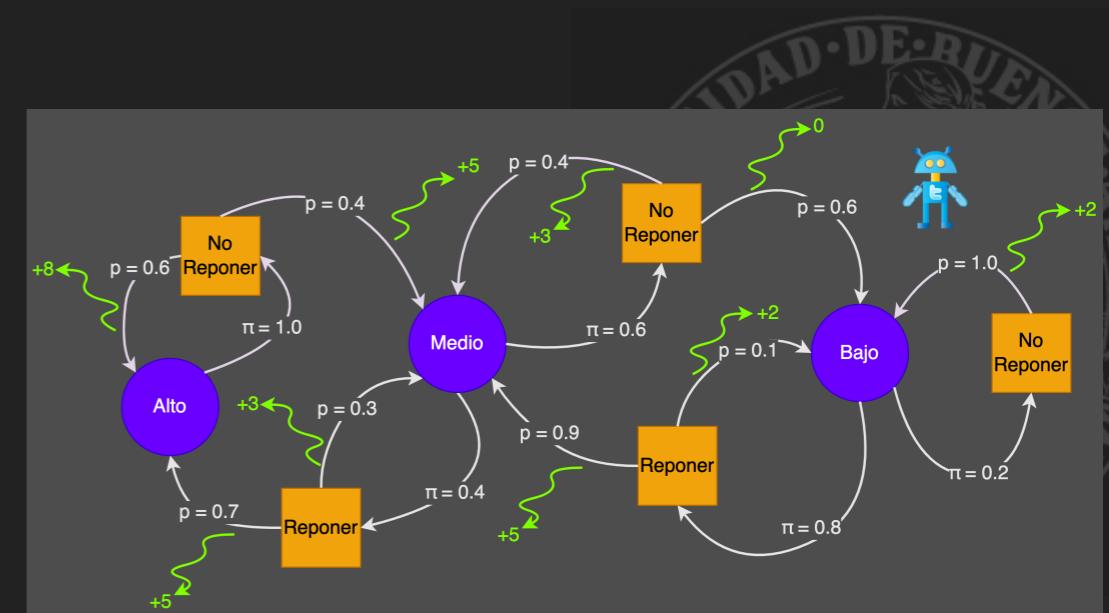
✓ alto: 43.0482

✓ medio: 36.1172

✓ bajo: 34.2194

► "alto" es el mejor estado. Esto tiene sentido, ya que indica que es más valioso a largo plazo mantener la góndola llena, incluso si no hay recompensas inmediatas por ello.

► "medio" es mejor que "bajo". También es lógico, ya que un estante a medio llenar es más valioso que un estante casi vacío.



## REFERENCIAS BIBLIOGRÁFICAS Y WEB (I)

- ▶ R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

