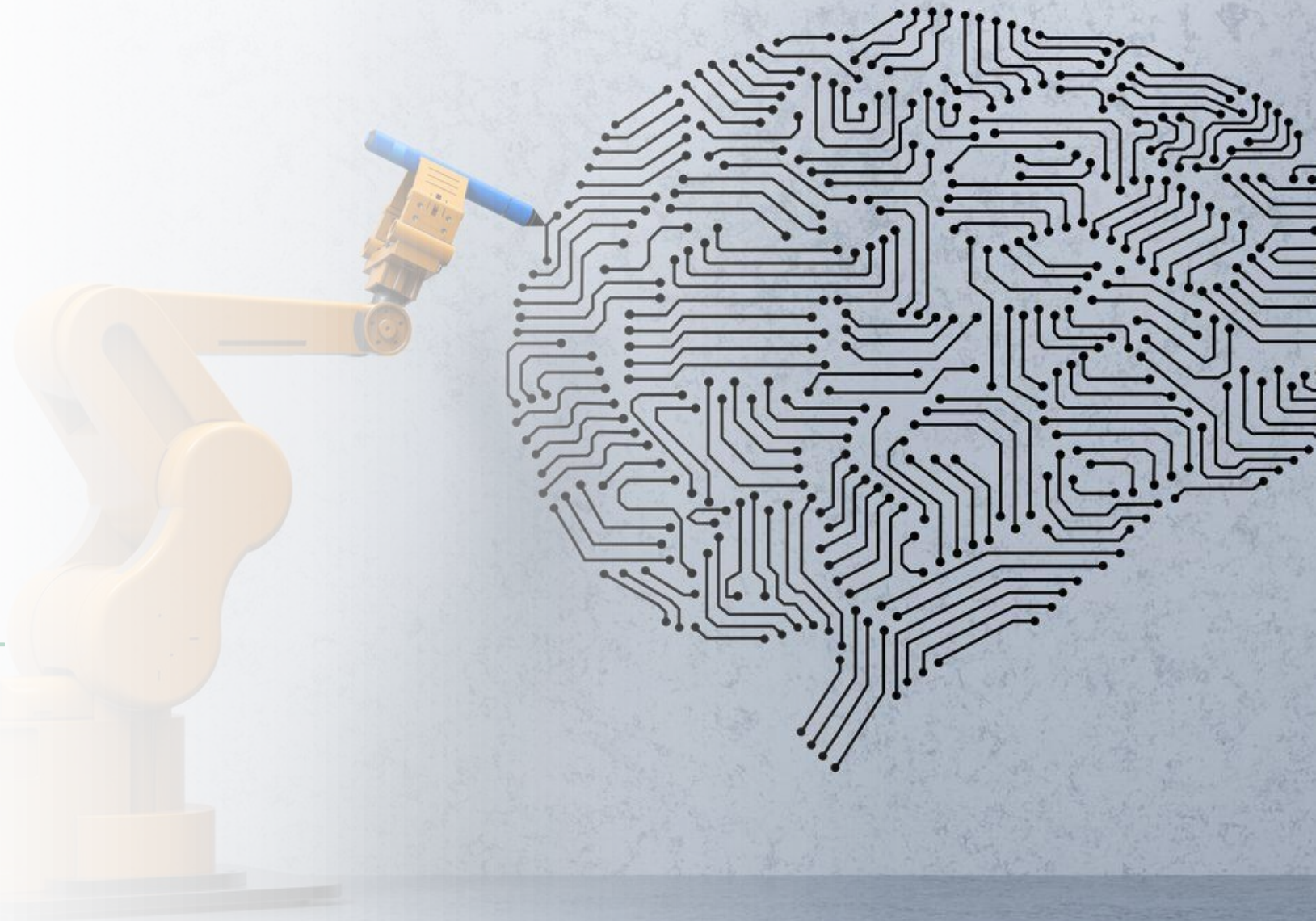
# Aprendizaje por refuerzo

Clase 2: Conceptos básicos

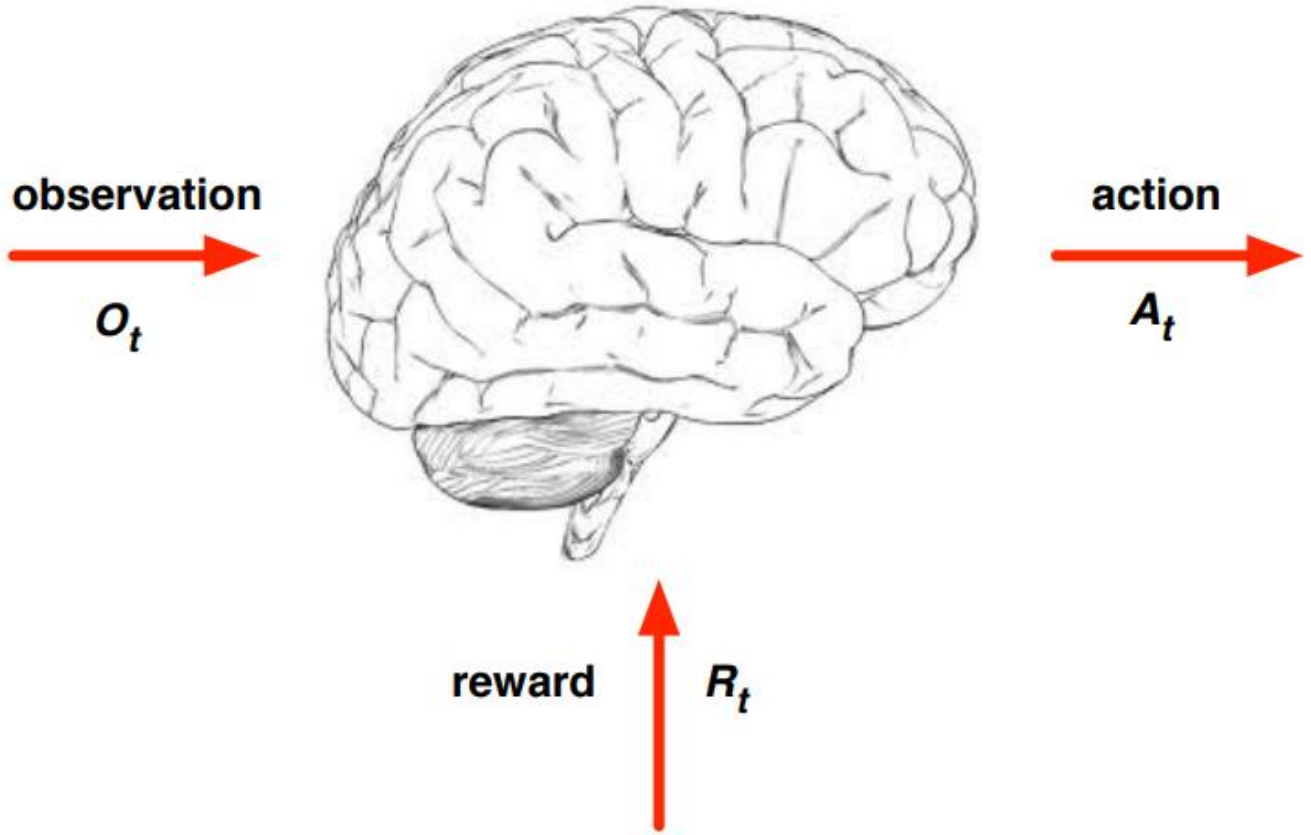


# Para el día de hoy…

- El problema
- Dentro de un agente de aprendizaje por refuerzo
- Los problemas dentro de aprendizaje por refuerzo

- Classroom: n4zyeed

# El modelo



# Características de aprendizaje por refuerzo

- No existe un supervisor, solo una señal de recompensa
- La retroalimentación tiene retraso
- El tiempo importa
- Las acciones del agente afecta los datos que recibe

# La recompensa

- Una recompensa $R_t \in \mathbb{R}$ es una señal de retroalimentación
- Indica que tan bien le va a un agente en el paso $t$
- El trabajo del agente es maximizar la recompensa acumulada
- Aprendizaje por refuerzo se basa en la hipótesis de la recompensa “Toda meta puede ser descrita como la maximización de recompensas esperadas”

# Ejemplos de recompensas

## Volar un helicóptero

- + seguir una trayectoria deseada
- - estrellarse

## Ajedrez

- + ganar el juego
- - perder el juego

## Portafolios de inversión

- +/- valor del portafolio

## Un robot caminando

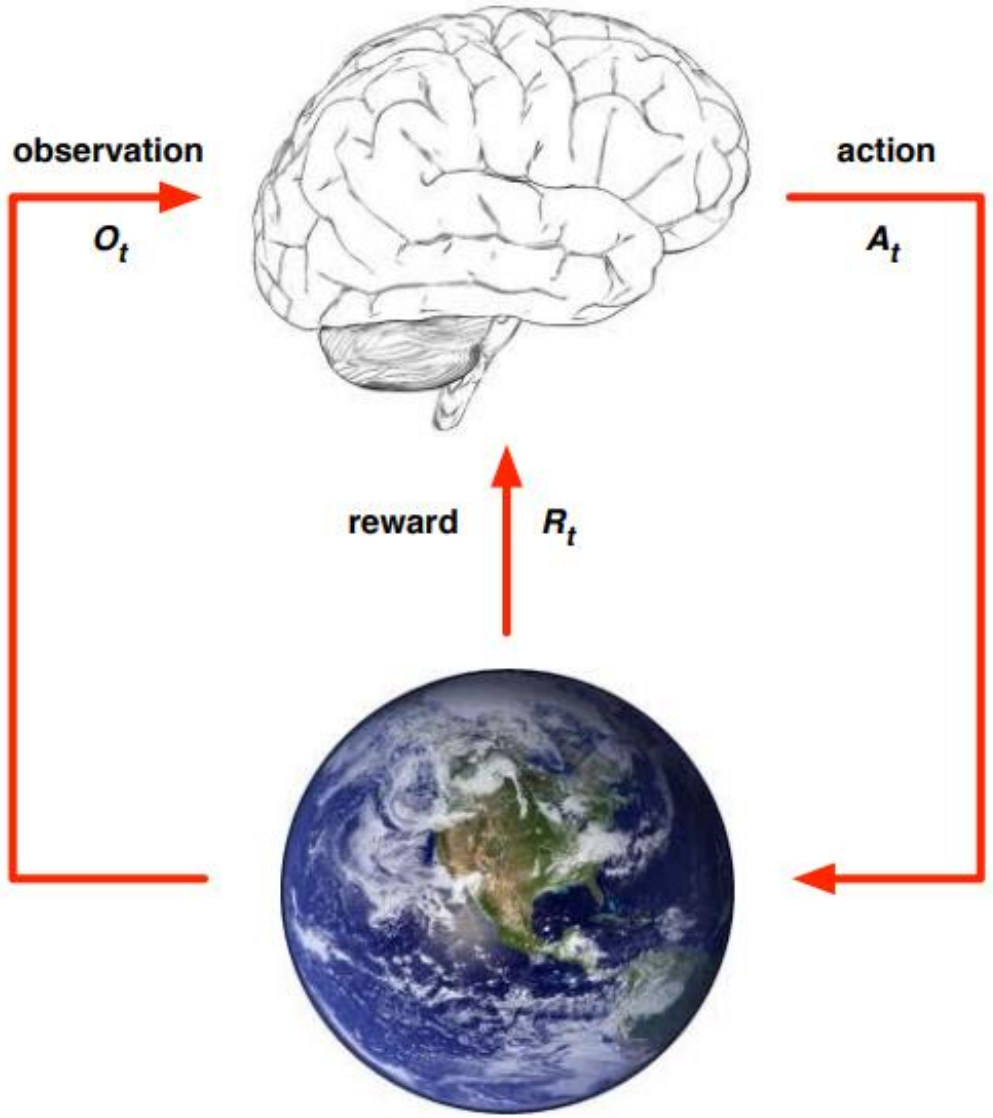
- + moverse hacia adelante
- - Caerse

# Toma de decisiones secuencial

- Meta: seleccionar las acciones que maximicen la recompensa futura total
- Las acciones pueden tener consecuencias a largo plazo
- La recompensa puede tener retraso
- Puede ser mejor sacrificar recompensa inmediata para ganar más a largo plazo

# El agente y su ambiente

- En cada paso $t$ el agente
  - Ejecuta una acción $A_t$
  - Recibe una observación $O_t$
  - Recibe una recompensa $R_t$
- El ambiente
  - Recibe una acción $A_t$
  - Emite una observación $O_{t+1}$
  - Emite una recompensa $R_{t+1}$
- $t$ se incrementa



# La historia y el estado

- La historia es la secuencia de observaciones, acciones y recompensas

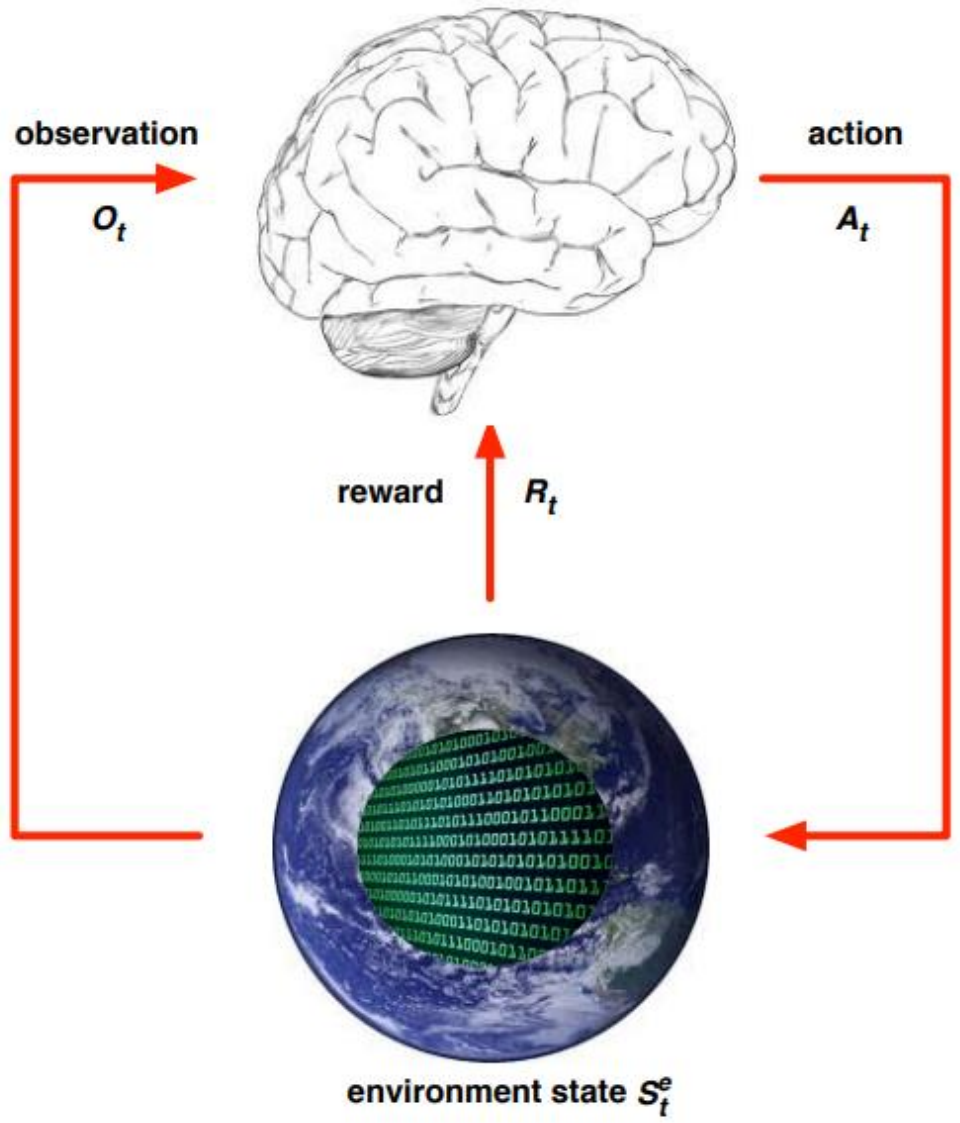
$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

- Lo que sucede después depende de la historia
  - El agente selecciona sus acciones
  - El ambiente selecciona observaciones y recompensas
- El estado es la información utilizada para determinar que sucede después

$$S_t = f(H_t)$$

# El estado del ambiente

- El estado del ambiente $S_t^e$ es la representación privada del ambiente
- Cualquier información utilizada por el ambiente para elegir la siguiente observación y recompense
- El estado del ambiente no es visible para el agente
- Aún si $S_t^e$ es visible, contiene información irrelevante



# Estado del agente

- El estado del agente $S_t^a$ es la representación interna del agente
- Cualquier información que el agente use para elegir su siguiente acción
- Puede ser cualquier función de la historia

$$S_t^a = f(H_t)$$

# La información del estado

- La información del estado contiene la información útil de la historia
- Un estado $S_t$ es Markoviano si y solo si

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \dots, S_t]$$

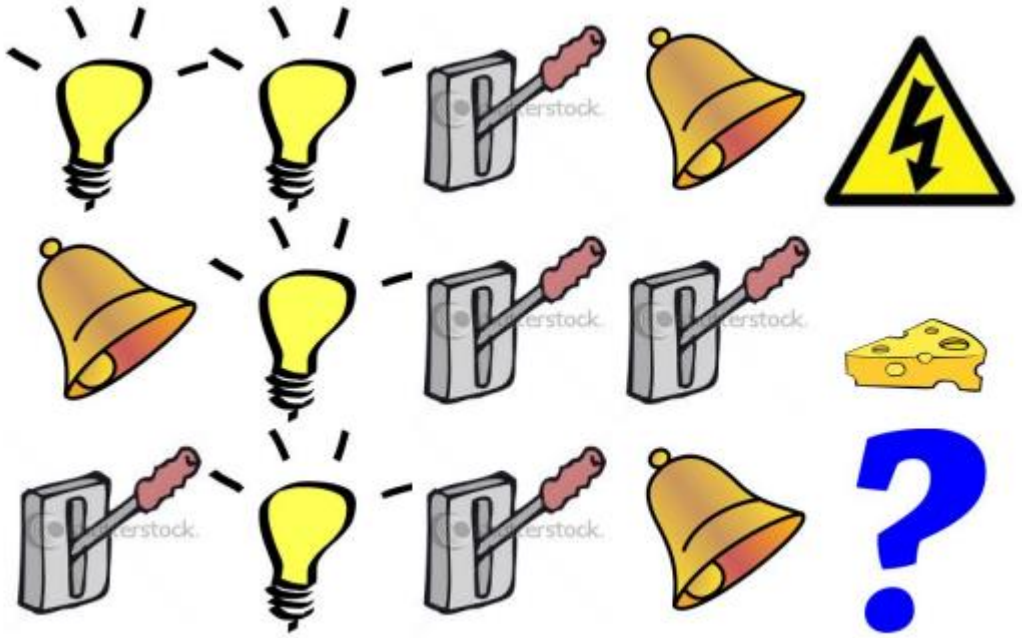
# La información del estado

- La información del estado contiene la información útil de la historia
- Un estado $S_t$ es Markoviano si y solo si

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \dots, S_t]$$

- El futuro es independiente del pasado dado el presente
- Si el estado es conocido, la historia se puede olvidar
- El estado del ambiente $S_t^e$ es Markov
- La historia $H_t$ es Markov

# Un ejemplo

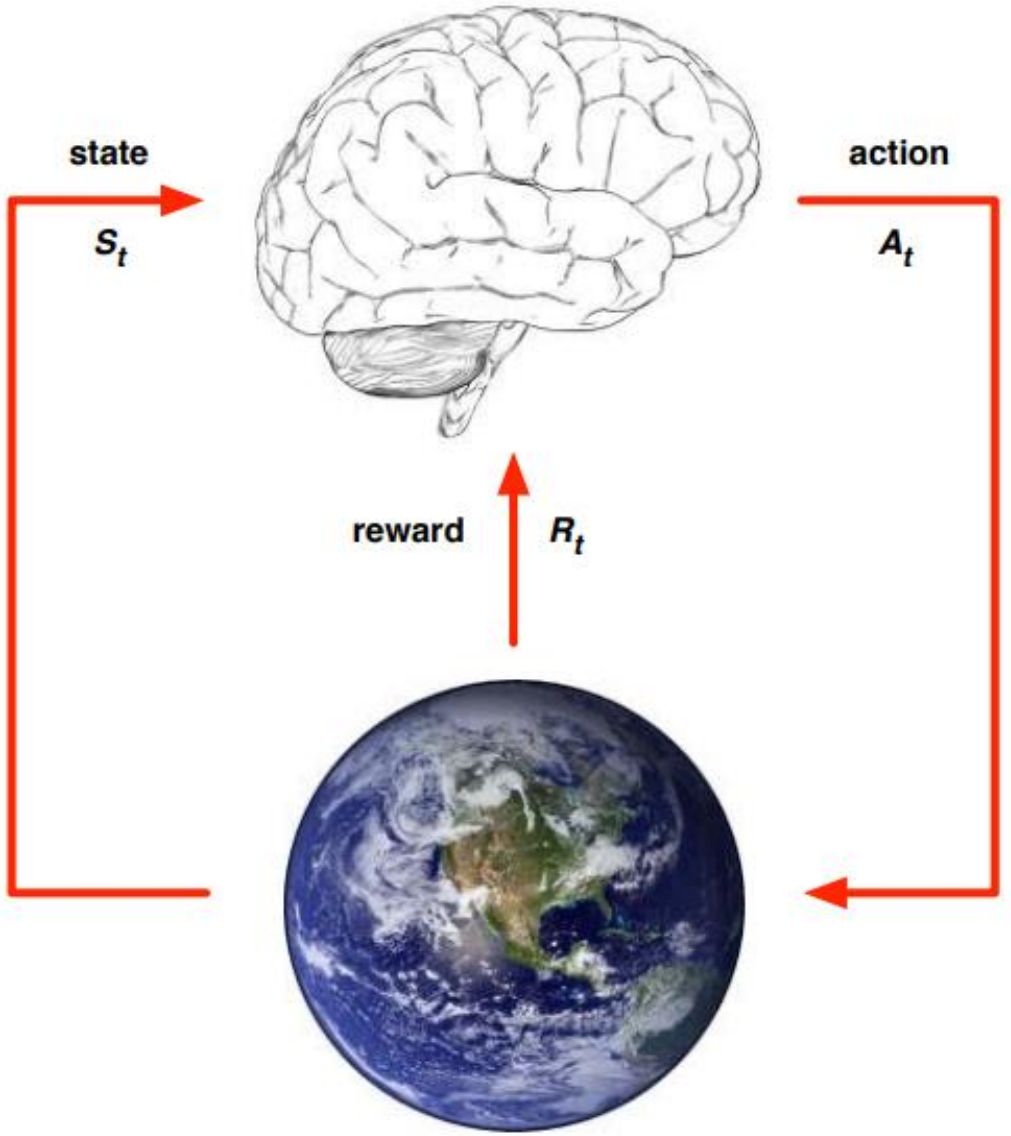



# Ambientes completamente observables

- El agente directamente observa el estado del ambiente

$$O_t = S_t^a = S_t^e$$

- Formalmente, esto es un proceso de decisión de Markov (MDP)

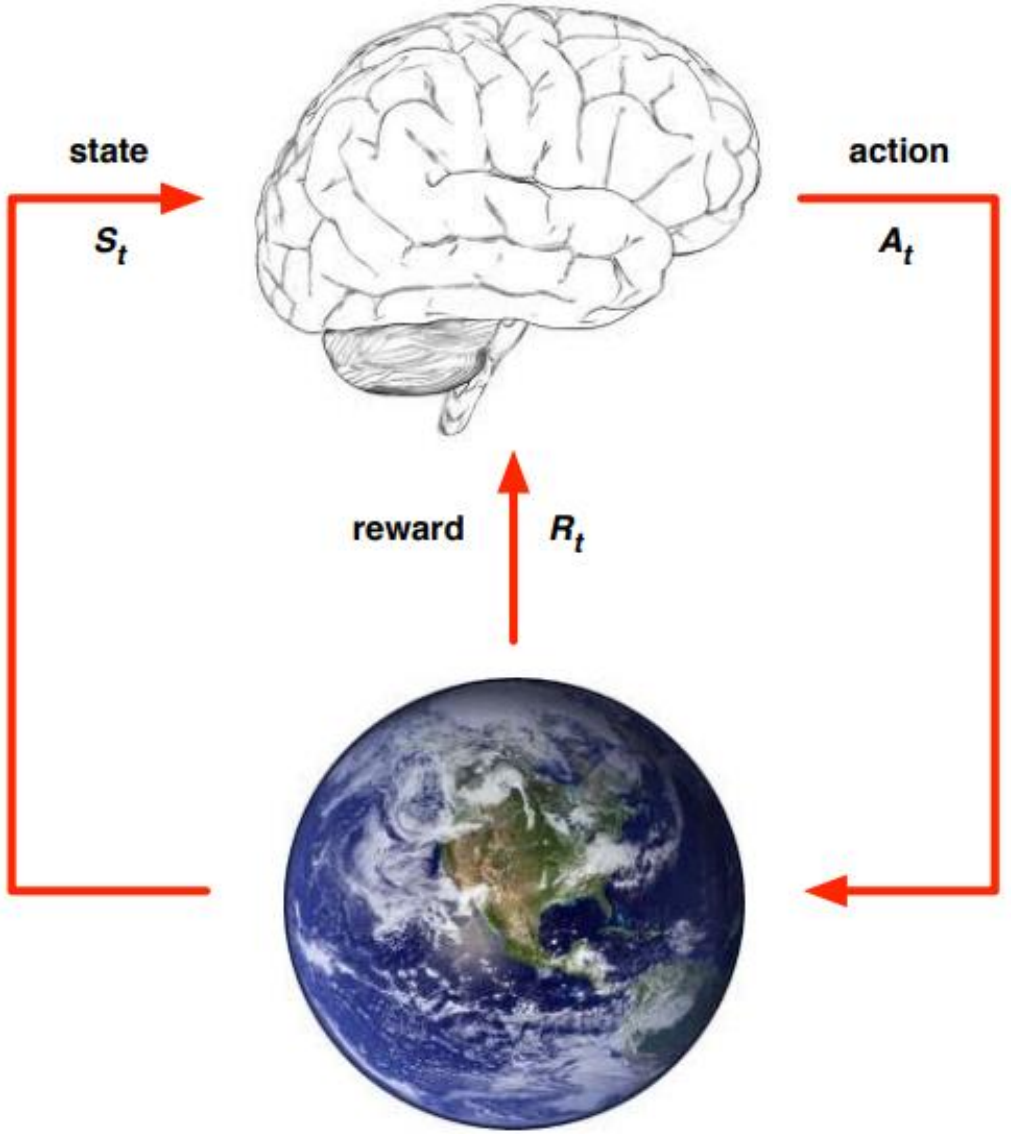


# Ambientes parcialmente observados

- El agente observa de forma indirecta el ambiente

$$S_t^a \neq S_t^e$$

- Formalmente, esto es un proceso de decisión de Markov parcialmente observable (POMDP)
- El agente debe construir su propia representación $S_t^a$
  - La historia completa: $S_t^a = H_t$
  - Creencias del estado del ambiente
  - Alguna red neuronal



# Componentes de un agente de aprendizaje por refuerzo

- Política
  - el comportamiento del agente
- Función de valor
  - que tan bueno es cada par estado/acción
- Modelo
  - la representación del ambiente usada por el agente

# Política $\pi$

- Es un mapeo de estados a acciones
- Tipos
  - Determinista: $a = \pi(s)$
  - Estocástica: $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

# Función de valor

- Es una predicción de la recompensa futura
- Evalúa que tan bueno o malo es un estado
- Ayuda a seleccionar entre estados
- Un ejemplo:

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s]$$

# Modelo

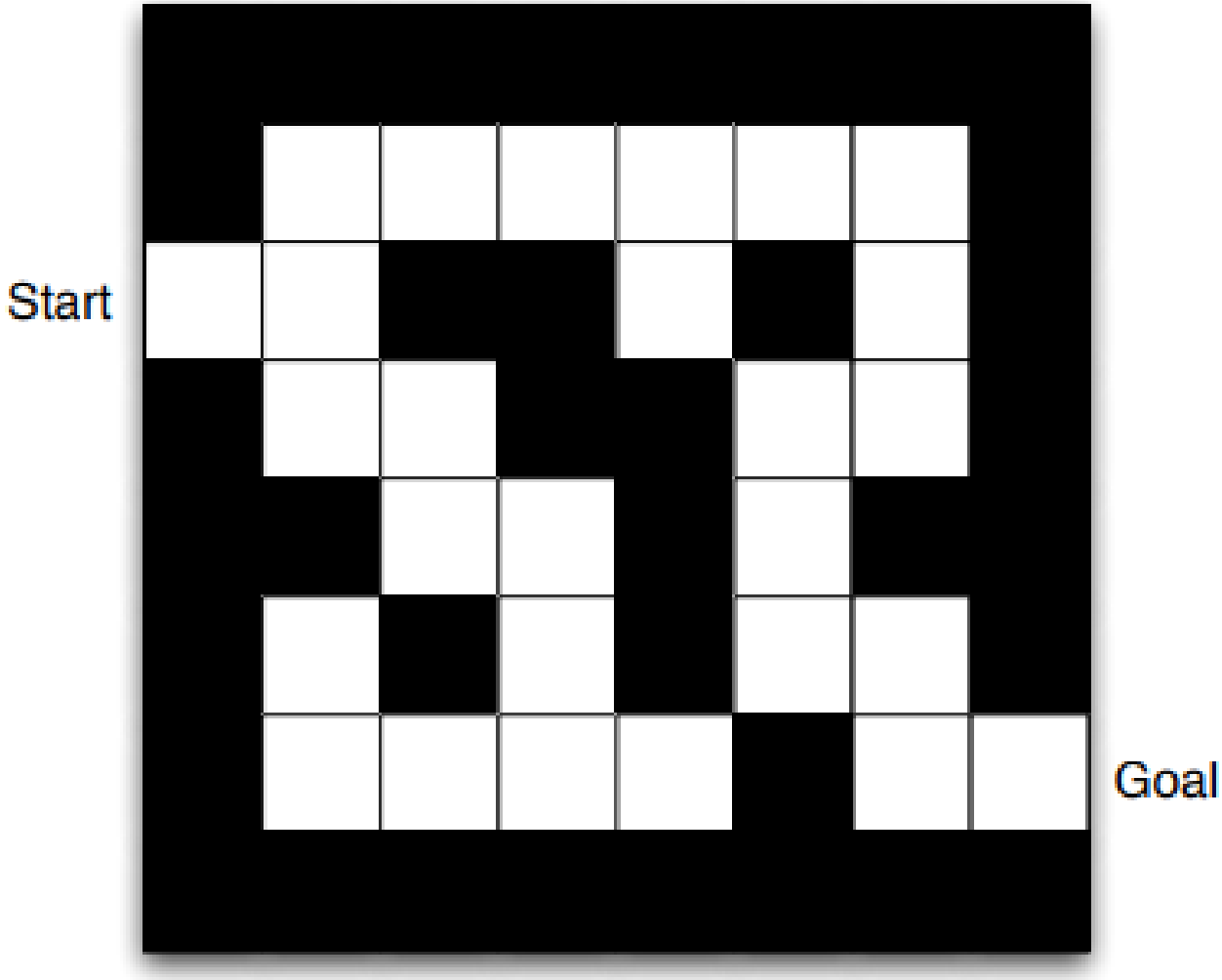
- Predice que hará el ambiente
- $\mathcal{P}$ predice el siguiente estado
- $\mathcal{R}$ predice la siguiente recompensa

$$\mathcal{P}_{ss'}^{a} = p(s, a, s') = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$

$$\mathcal{R}_{s}^{a} = r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

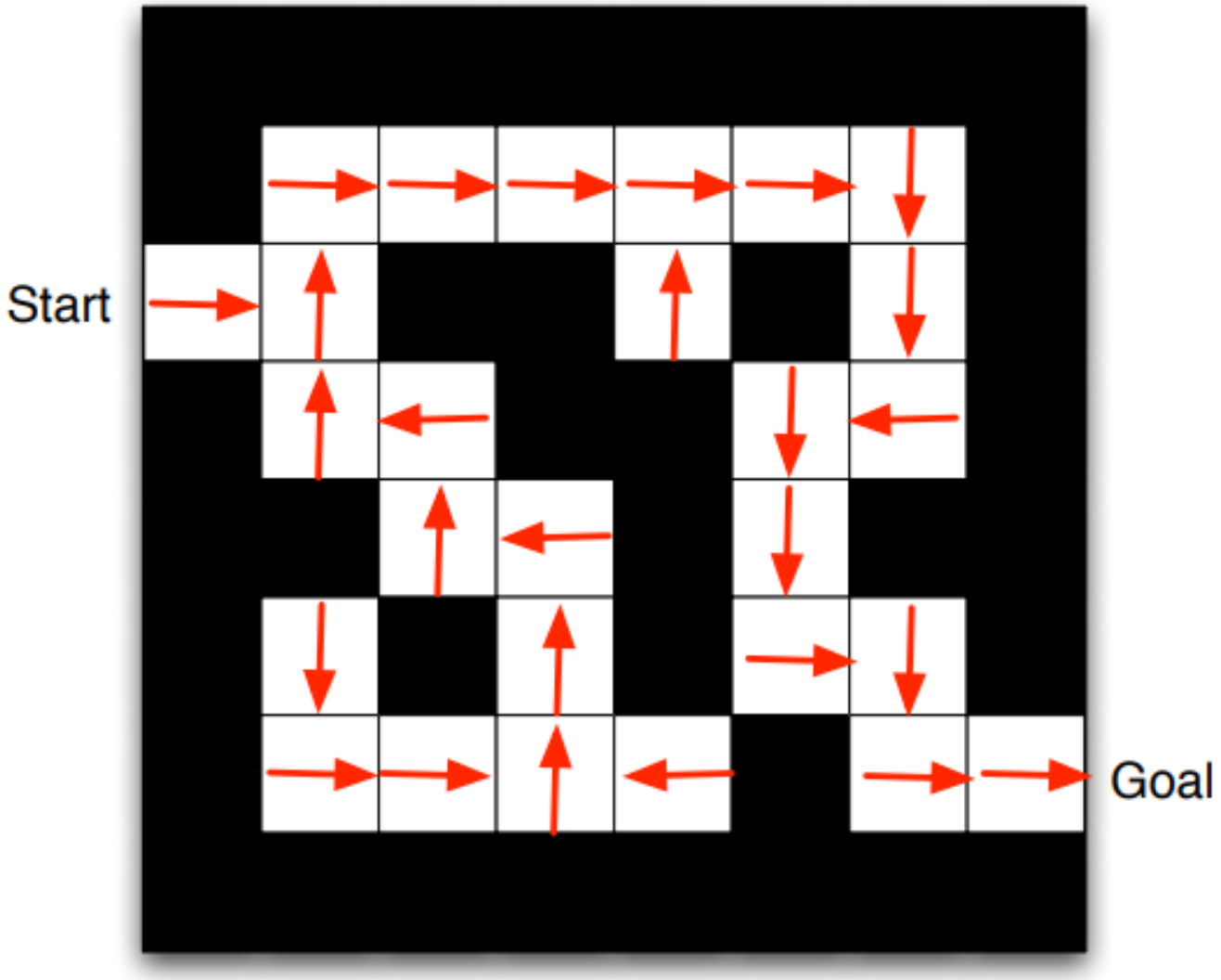
# Un ejemplo

- Recompensas: -1 por paso
- Acciones: N, S, E, O
- Estados: ubicación del agente



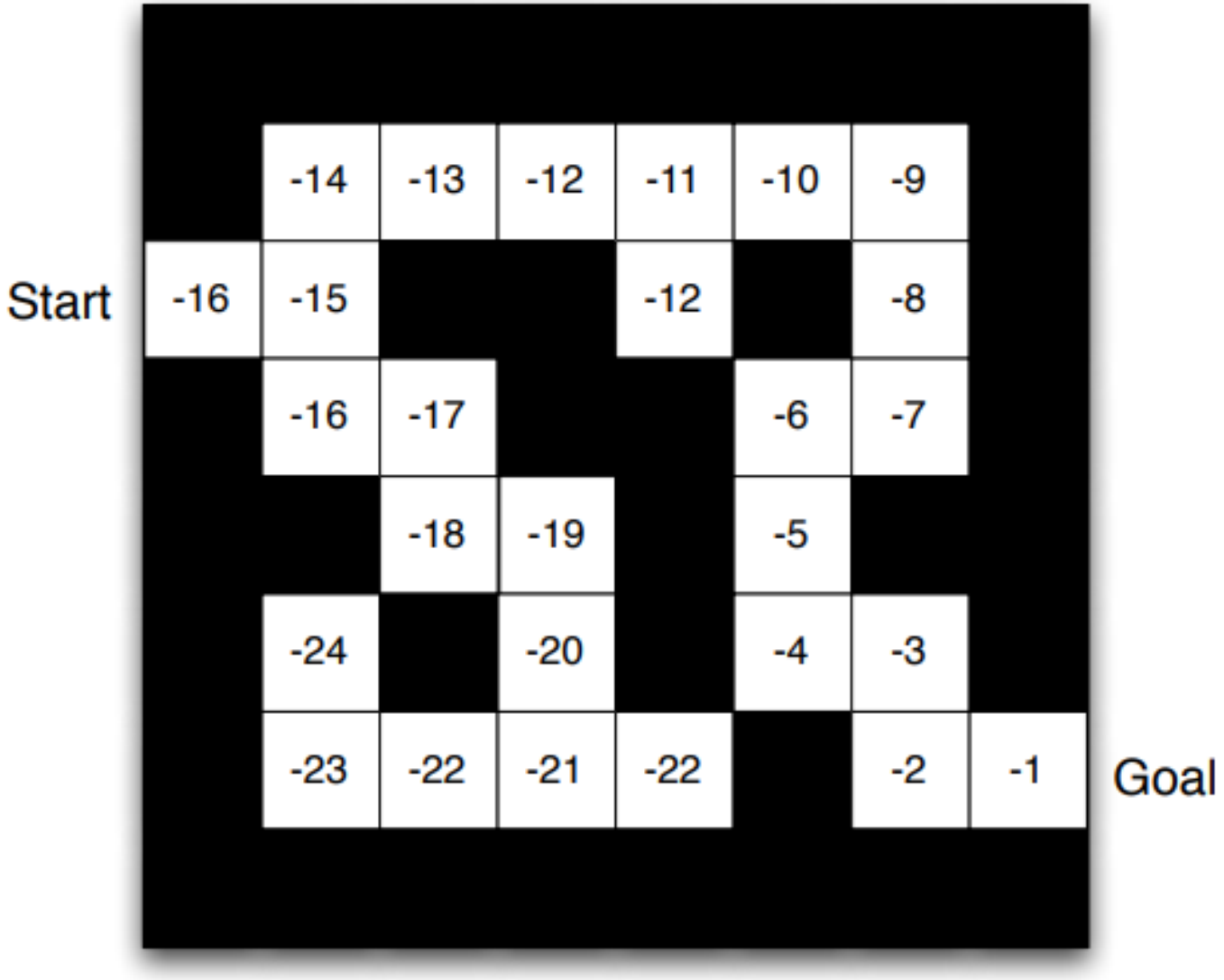
# Política

- Recompensas: -1 por paso
- Acciones: N, S, E, O
- Estados: ubicación del agente



# Función de valor

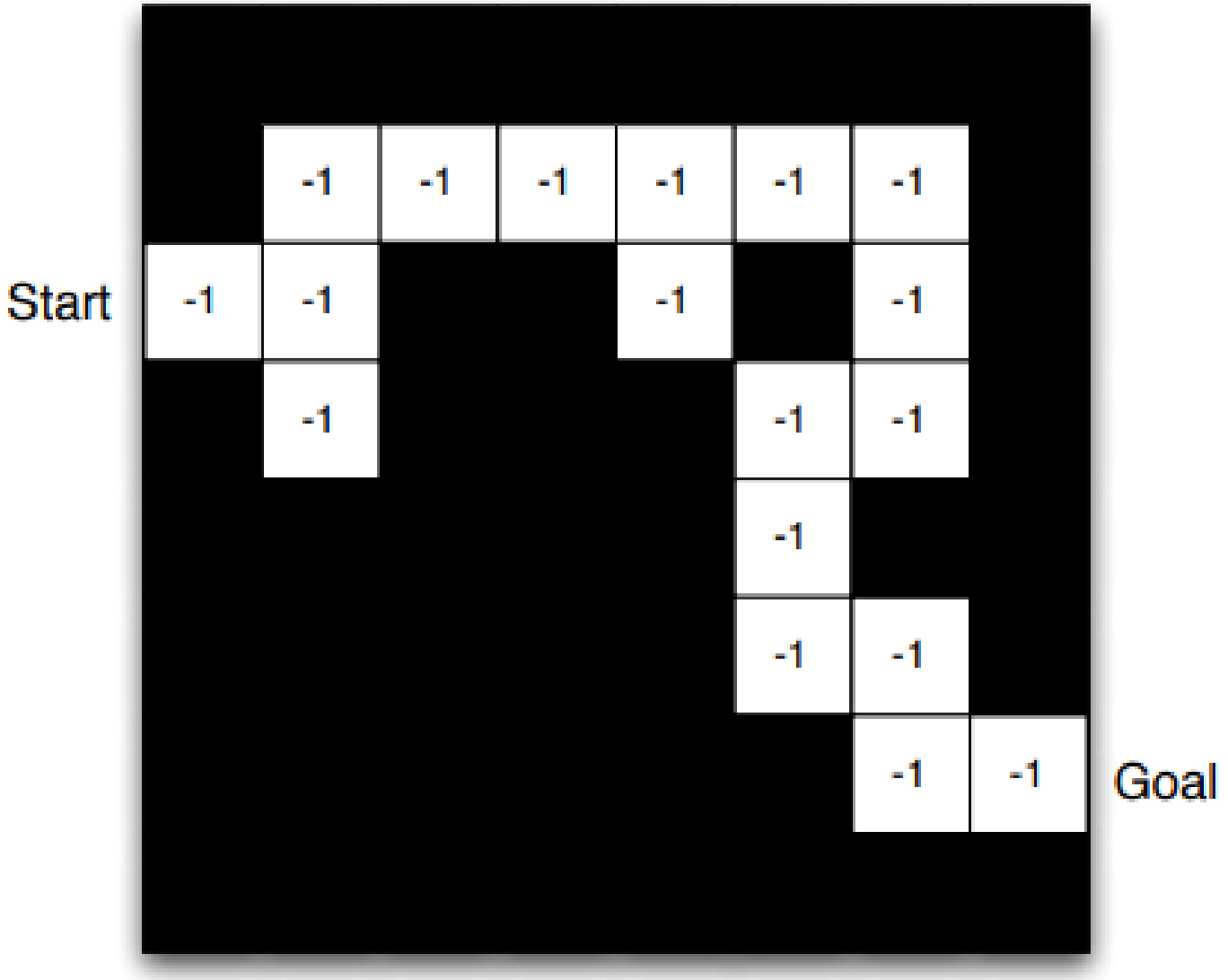
- Recompensas: -1 por paso
- Acciones: N, S, E, O
- Estados: ubicación del agente



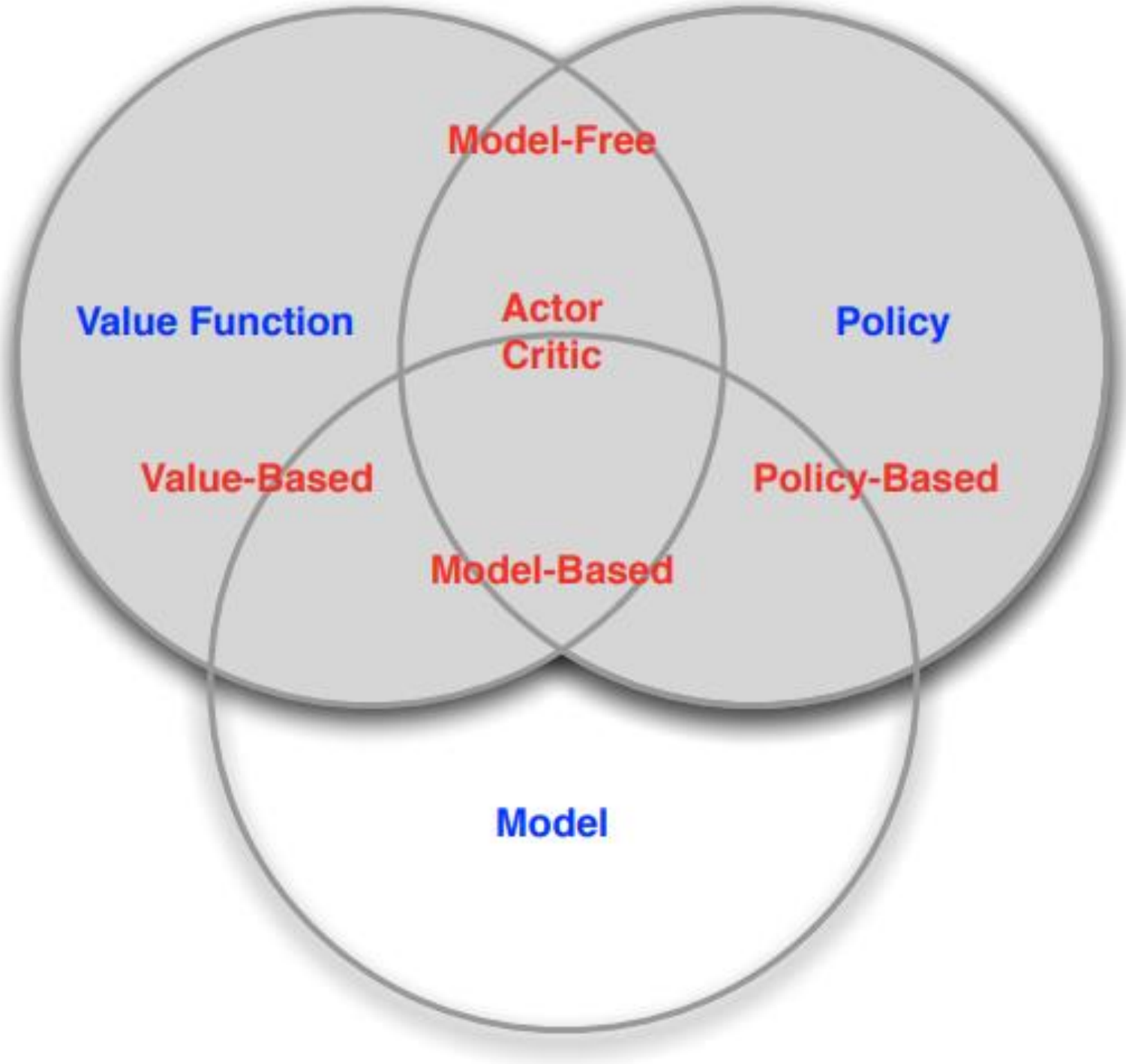
# Modelo

- Recompensas: -1 por paso
- Acciones: N, S, E, O
- Estados: ubicación del agente

- Puede ser imperfecto
- $\mathcal{P}_{ss'}^{a}$, es representado por la rejilla
- $\mathcal{R}_{s}^{a}$ representa la recompensa inmediata



# Taxonomía de agentes de aprendizaje por refuerzo



# Exploración y explotación

Exploración: encuentra más información acerca del ambiente

Explotación: utiliza la información conocida para maximizar la recompensa

Usualmente queremos hacer las dos

# Predicción y control

Predicción: evaluar el futuro dada una política

Control: optimizar el futuro. Es decir, encontrar la mejor política

# Para la otra vez…

- Procesos de decisión de Markov