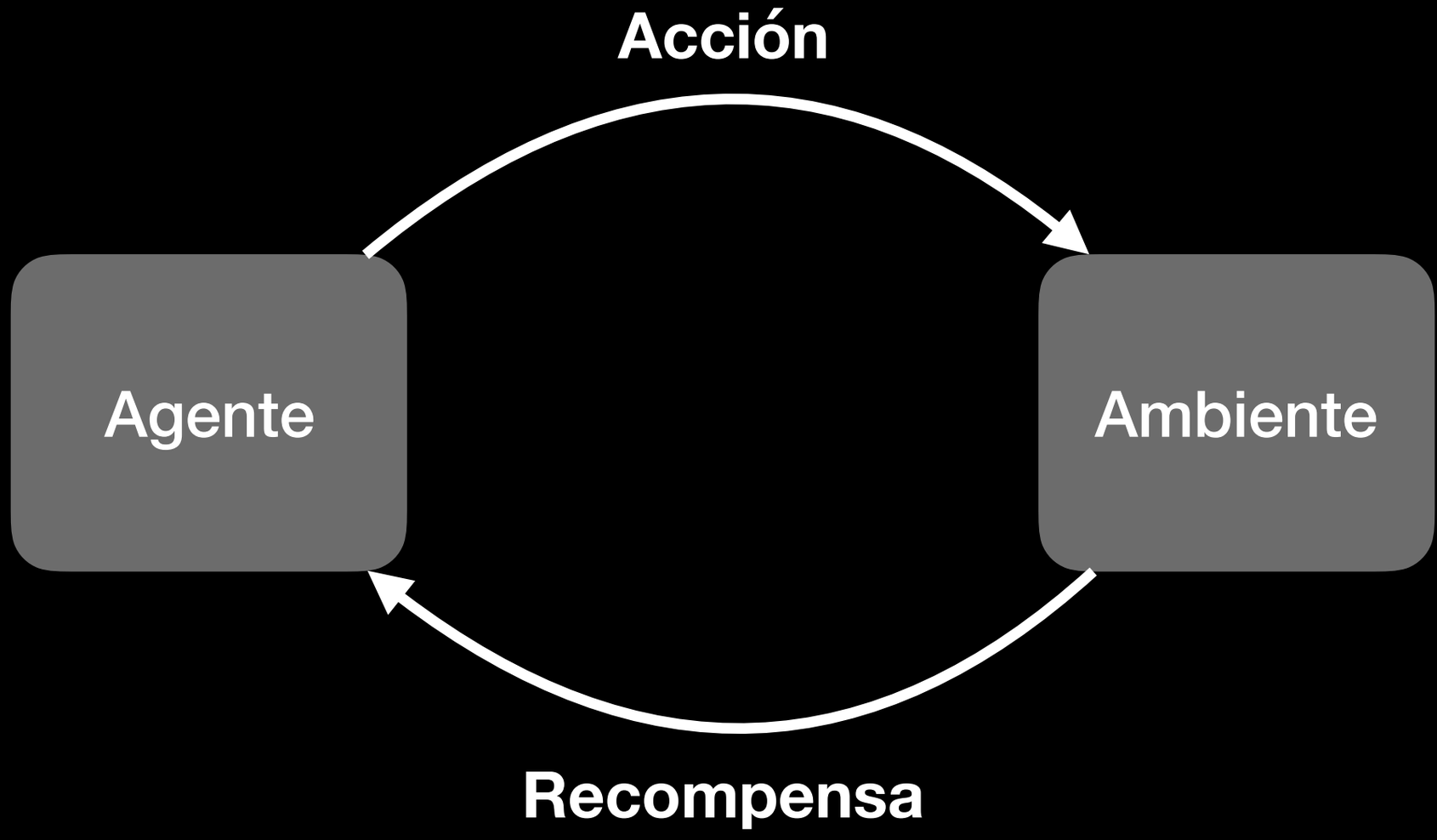
# Interactive Recommendation via Deep Neural Memory Augmented Contextual Bandits

Yilin Shen, Yue Deng, Avik Ray, Hongxia jin

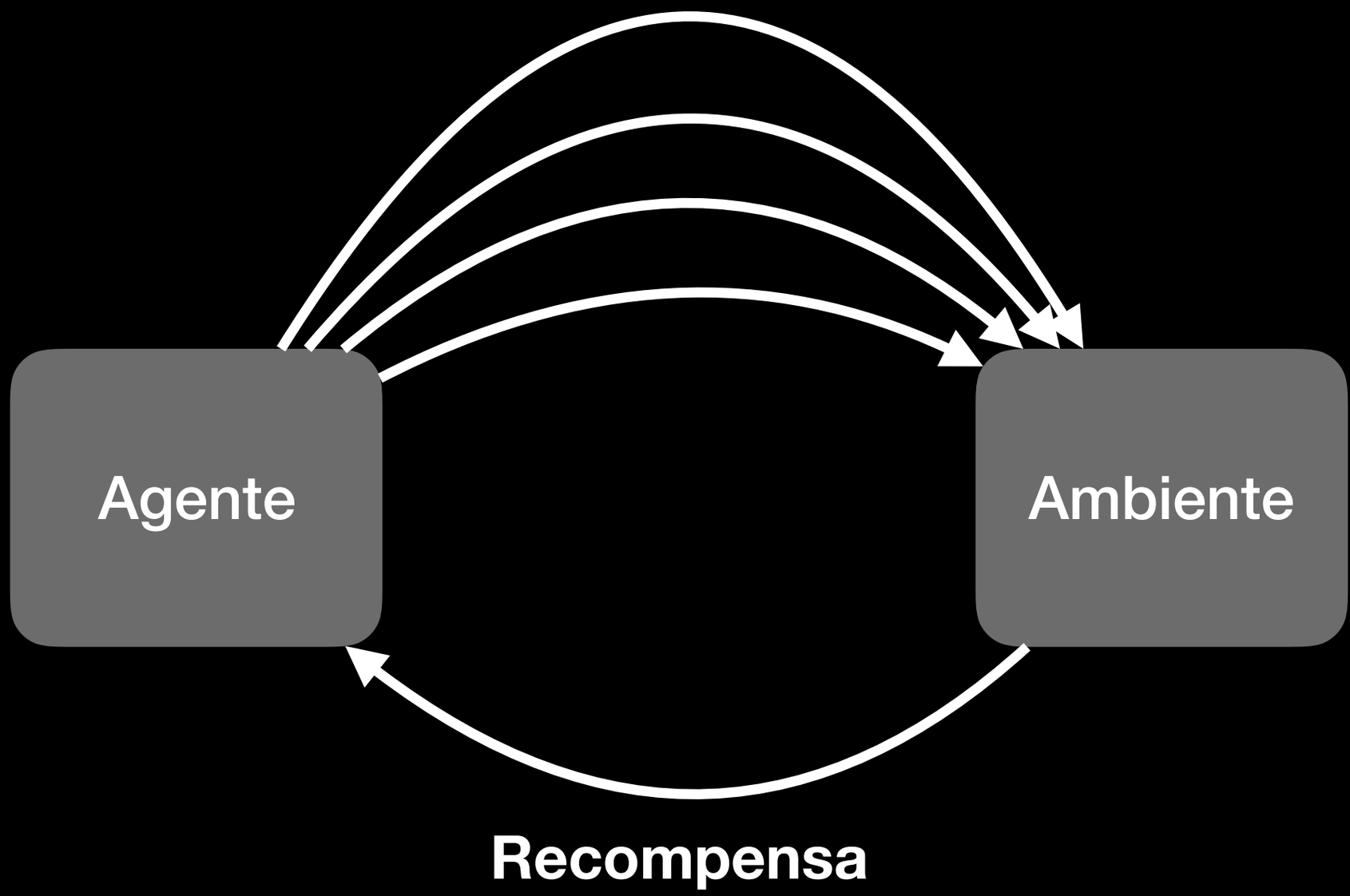
# Introducción

- Recomendación Interactiva
- Ej: Chat Bot
- Problema Modelado tradicionalmente como un Contextual Bandit (por usuario)

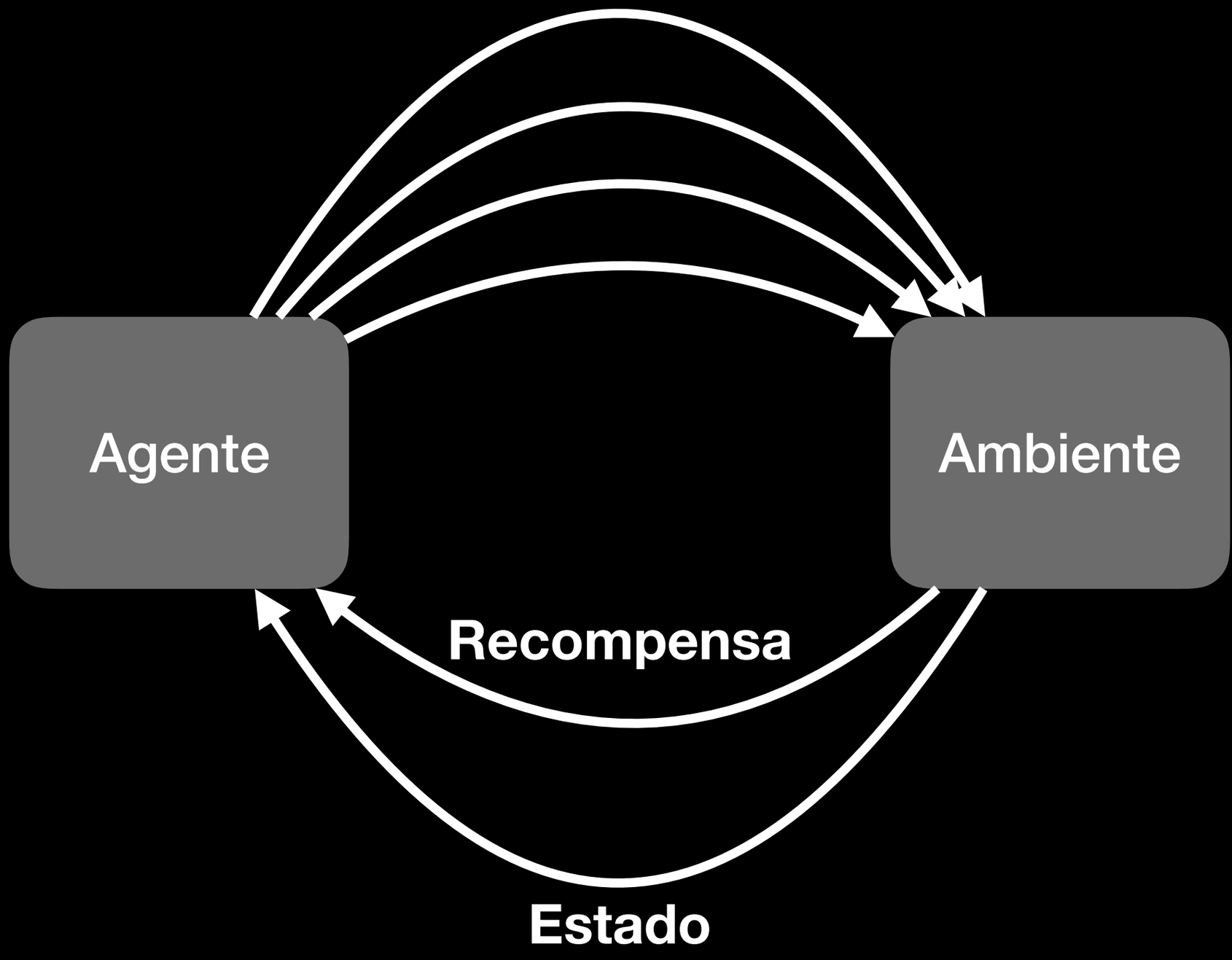
# One-Armed Bandit



# Multi-Armed Bandit

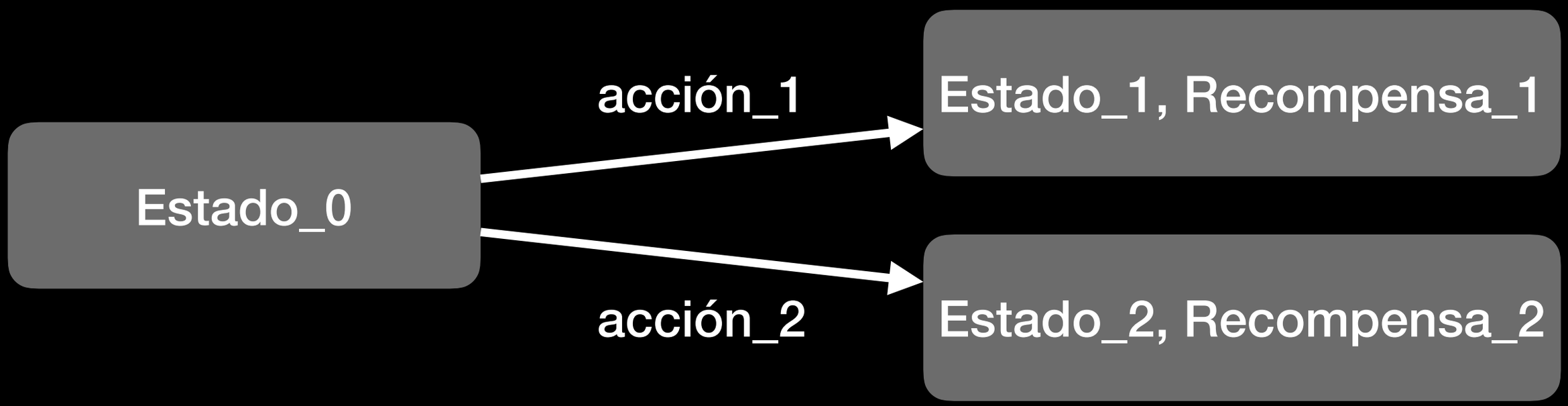


# Contextual Bandit



# Contextual Bandit

- Objetivo: Maximizar recompensa
- Se intenta aprender las transiciones entre estados y recompensas asociadas a (estado, acción)
- A diferencia de RL, no modifica el ambiente



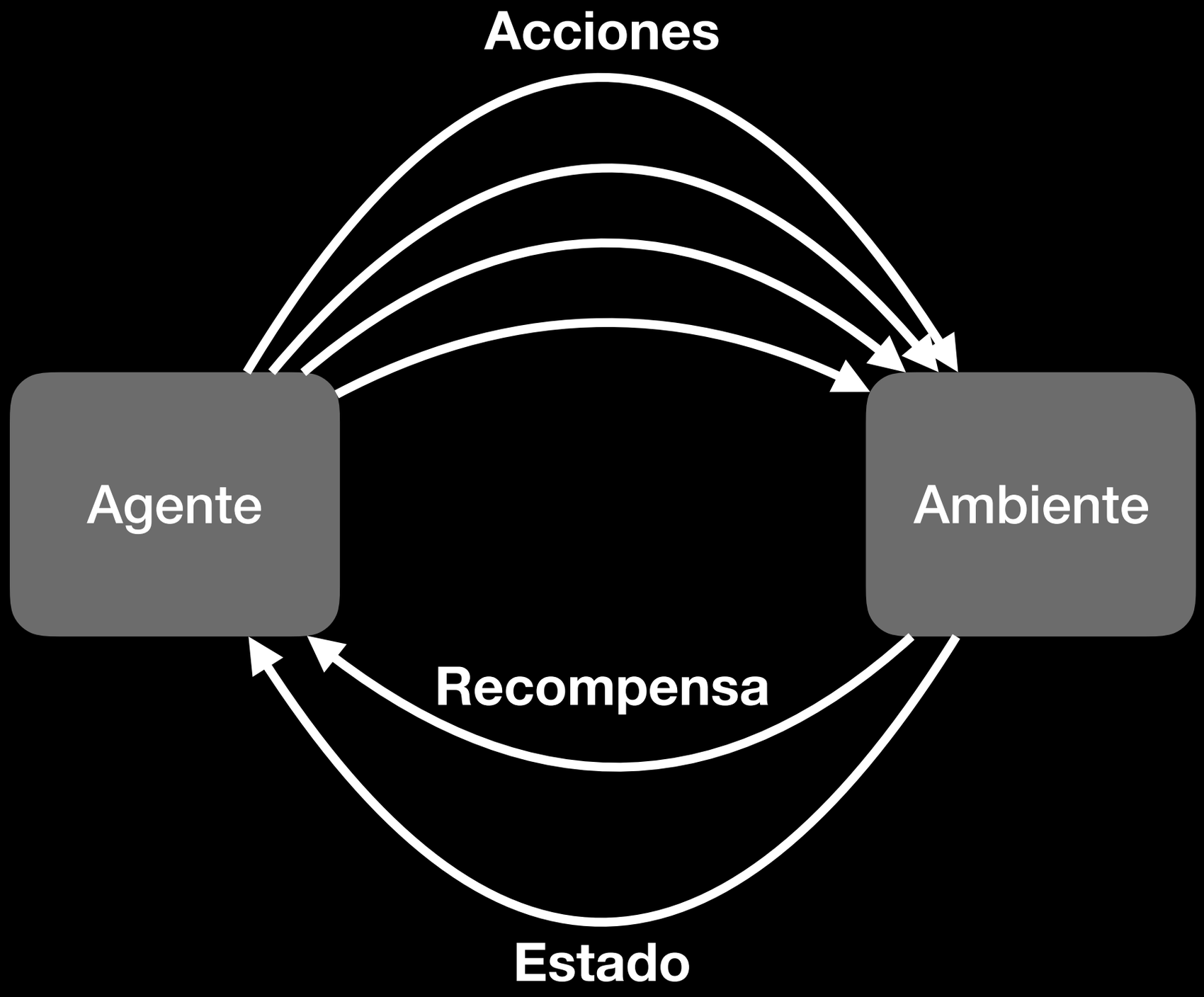
# Contextual Bandit

- Problemas:
  - Explore / Exploit Dilemma
  - Muchas iteraciones para converger (Propagar recompensas diferidas por muchas transiciones)
  - Dificultad para adaptarse a ambientes cambiantes (Online)
  - Dificultad para recomendar items novedosos como noticias recientes (cold start)

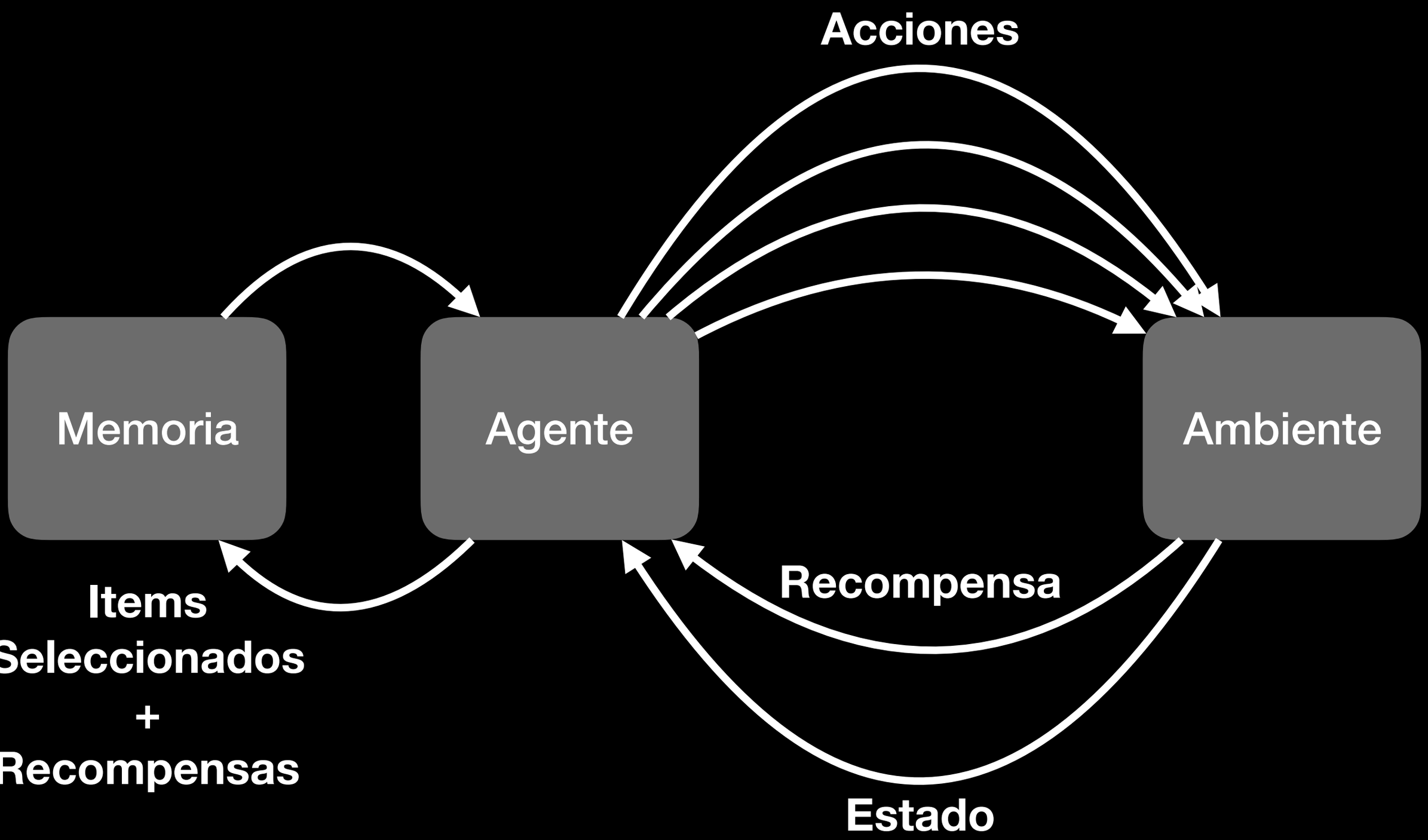
# Contribución Paper

- History state tracking enabled contextual bandit algorithm
- “Recuerda” las features de los estados de los items (acciones) seleccionados
- Rápido aprendizaje de preferencias de usuarios
- Aprende a partir de una pequeña cantidad de interacciones del usuario, pero en diversos items.

# Contribución Paper



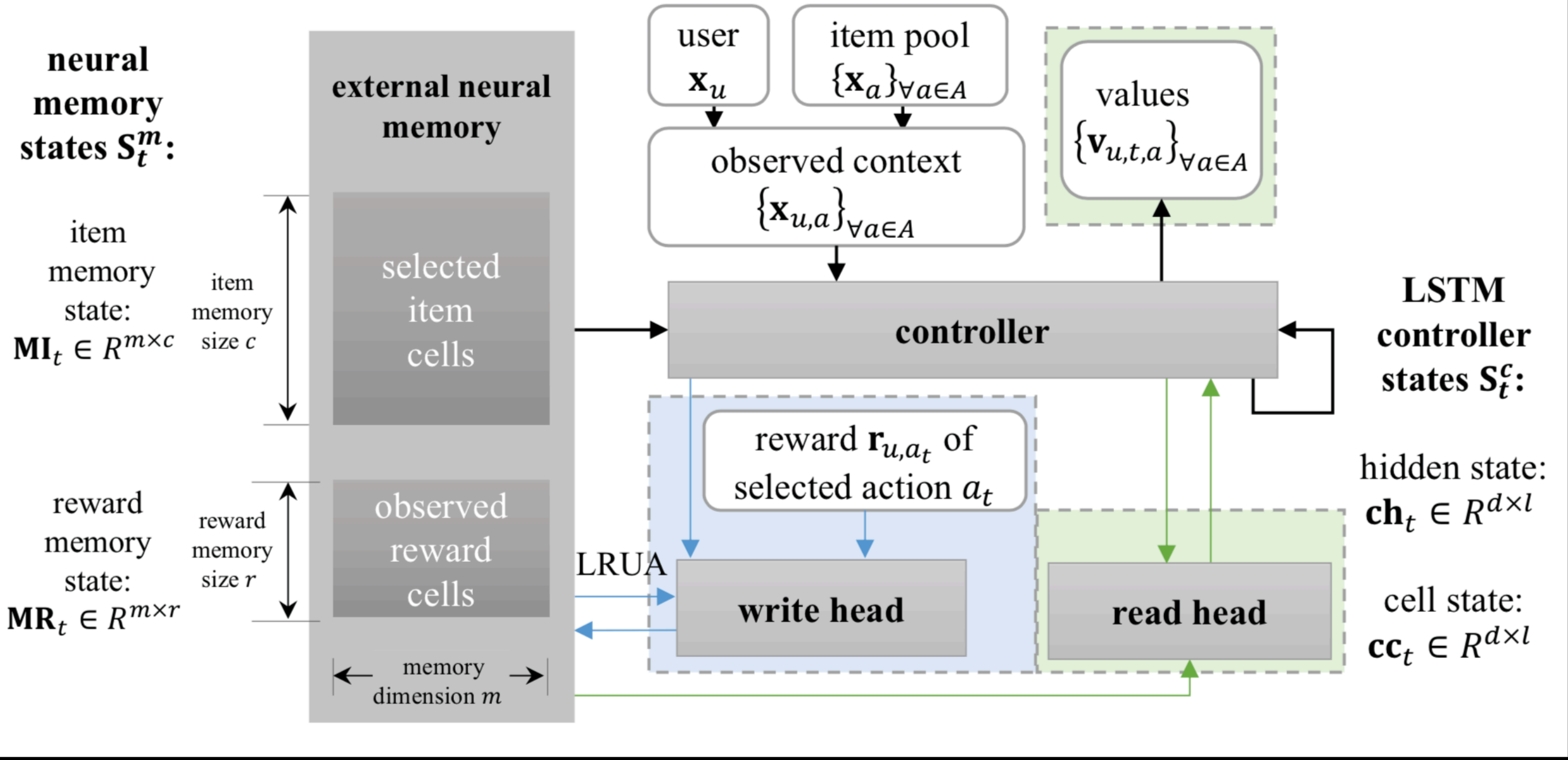
# Contribución Paper



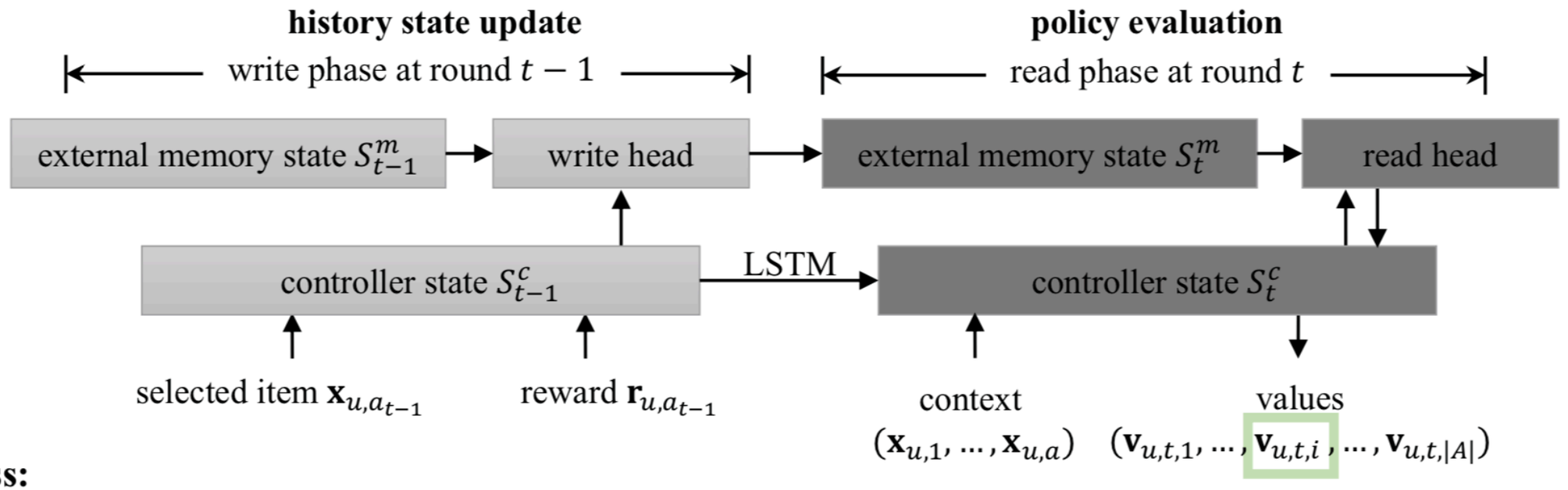
# Solución del Paper

- Se utiliza Experience Replay
- (S_t, A_t, R_t, S_t+1)
- Para acelerar el aprendizaje
- Por costo de la exploración.

# Solución del Paper



# Controller




**$L_2$ loss:**

$$\mathcal{L}_{\text{online}} = \left\|\mathbf{r}_{u,a_t} - \mathbf{v}_{u,t,a_t}\right\|_2^2 + \mathbf{E}_{\mathbf{Tr}_{t'-1}^{t'} \sim U(\mathbf{D})} \left\|\mathbf{r}_{u,a_{t'}} - \mathbf{v}_{u,t',a_{t'}}\right\|_2^2$$

**experience replay**

**select item** $a_t = i$ using decayed-$\epsilon$-greedy
**observe reward** $\mathbf{r}_{u,a_t}$

**Transition Instance** $\mathbf{Tr}_{t-1}^{t} = \langle S_{t-1}, (\mathbf{x}_{u,a_{t-1}}, r_{u,a_{t-1}}), S_t, \mathbf{x}_{u,a_t} \rangle$ **from round** t-1 **to round** t

# Recomendación de item

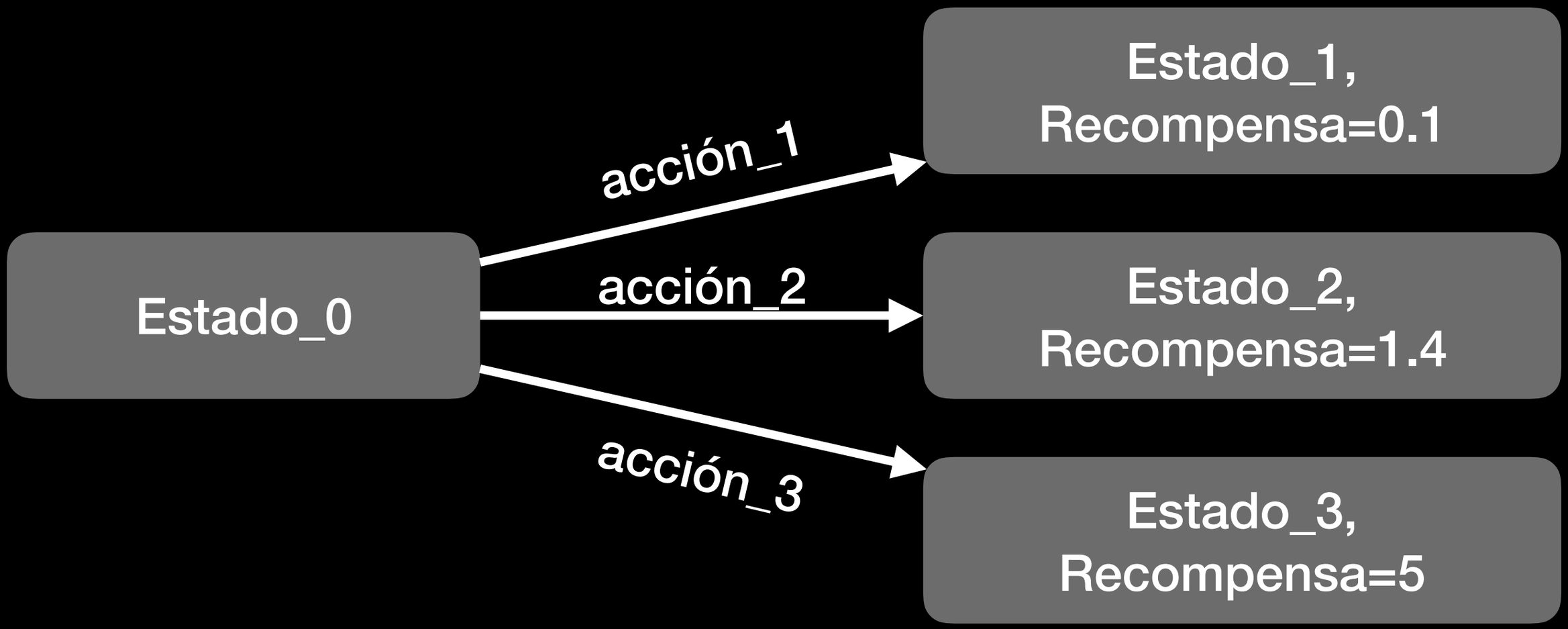
- ε - Greedy bandit strategy
  - Acción seleccionada con probabilidad (1 - ε)

$$a_t = argmax_{a \in A}(v_{u,t,a})$$

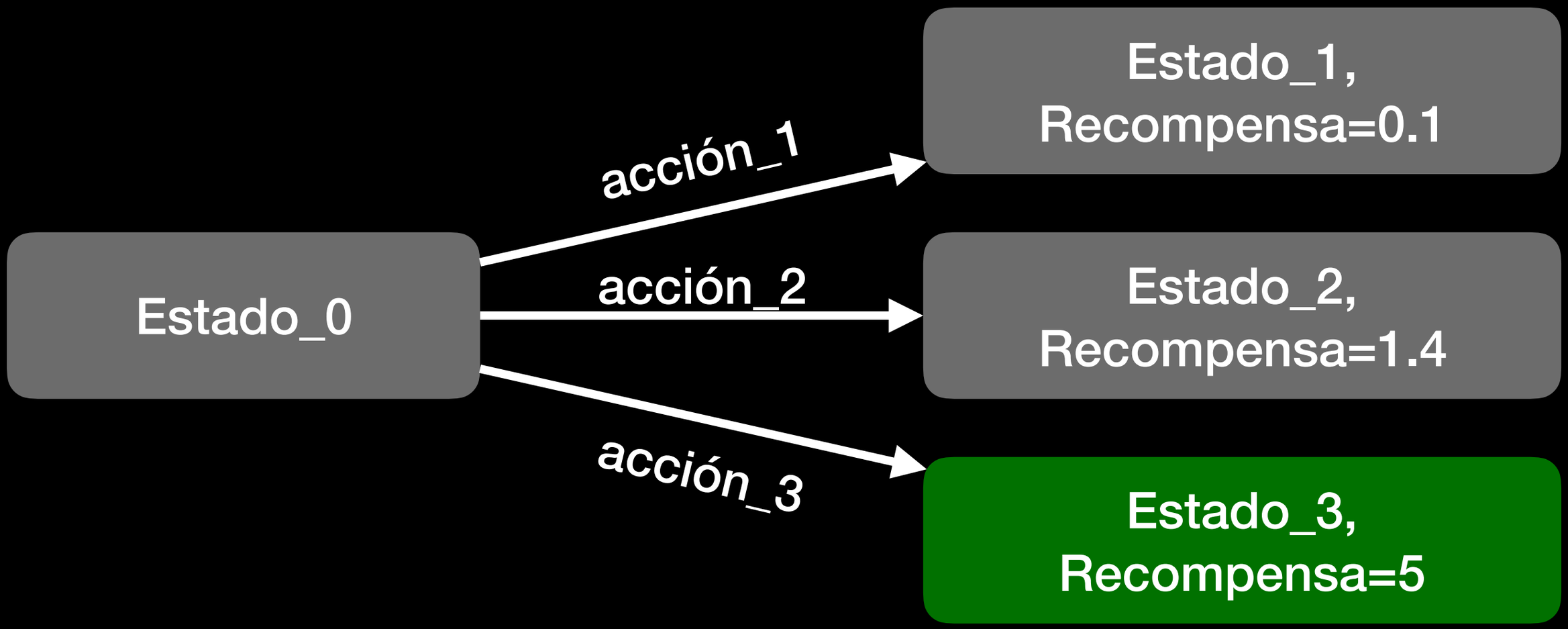
  - Random con probabilidad ε

  - ε decae exponencialmente luego de cada iteración.
    (Ej $\varepsilon_t = \varepsilon_{t-1} * 0{,}9$)

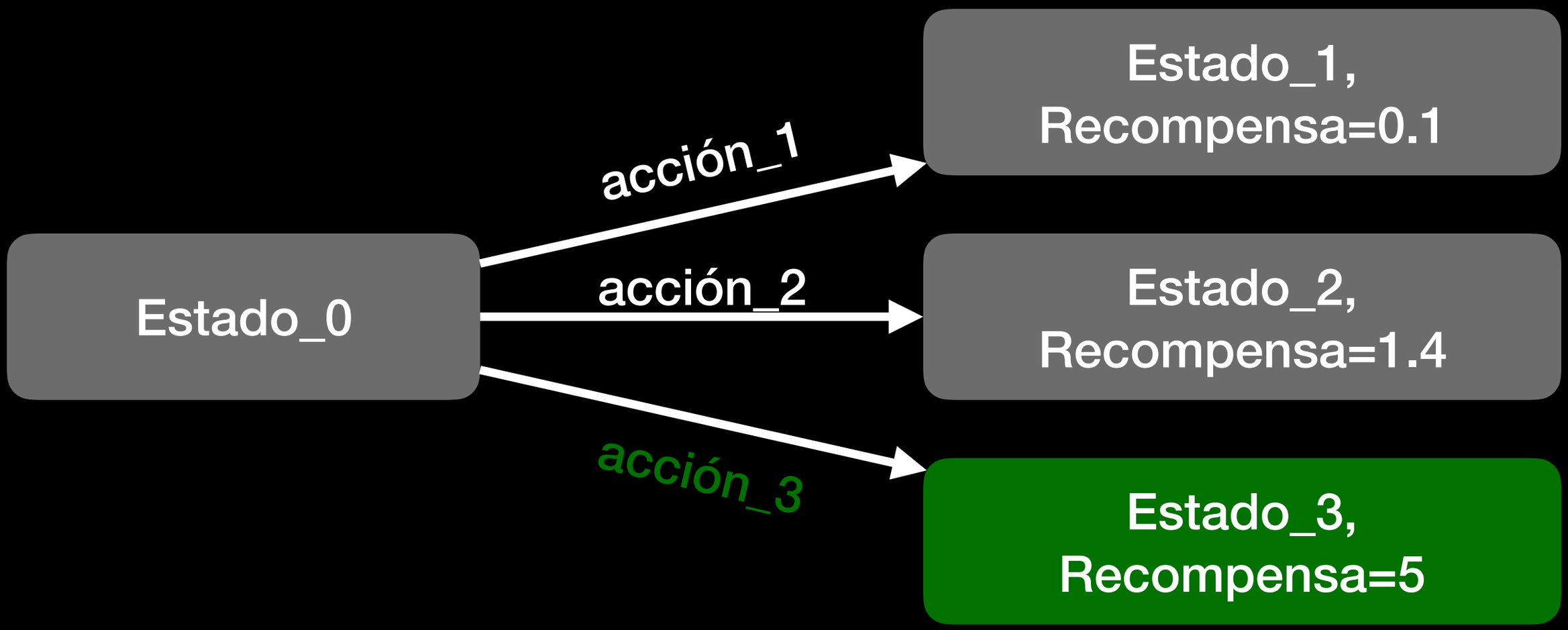
# ε - Greedy bandit strategy



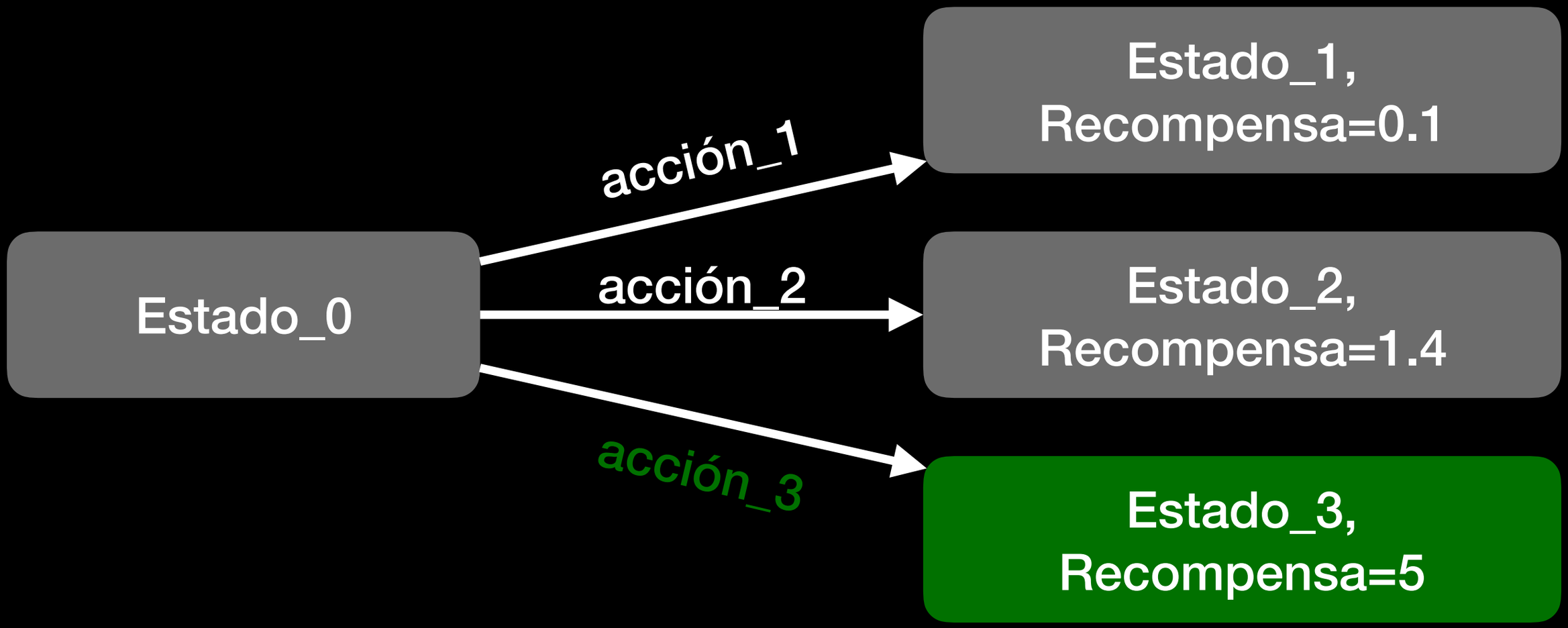
# ε - Greedy bandit strategy



# ε - Greedy bandit strategy



# ε - Greedy bandit strategy




Se toma con probabilidad 1 - ε: acción_3

probabilidad ε: random entre 1 y 3

# Ventajas DMCB

- Entiende rápidamente las preferencias de los usuarios.
- DMCB permite compartir el modelo entre usuarios pero manteniendo el history state particular para cada usuario en memoria externa.

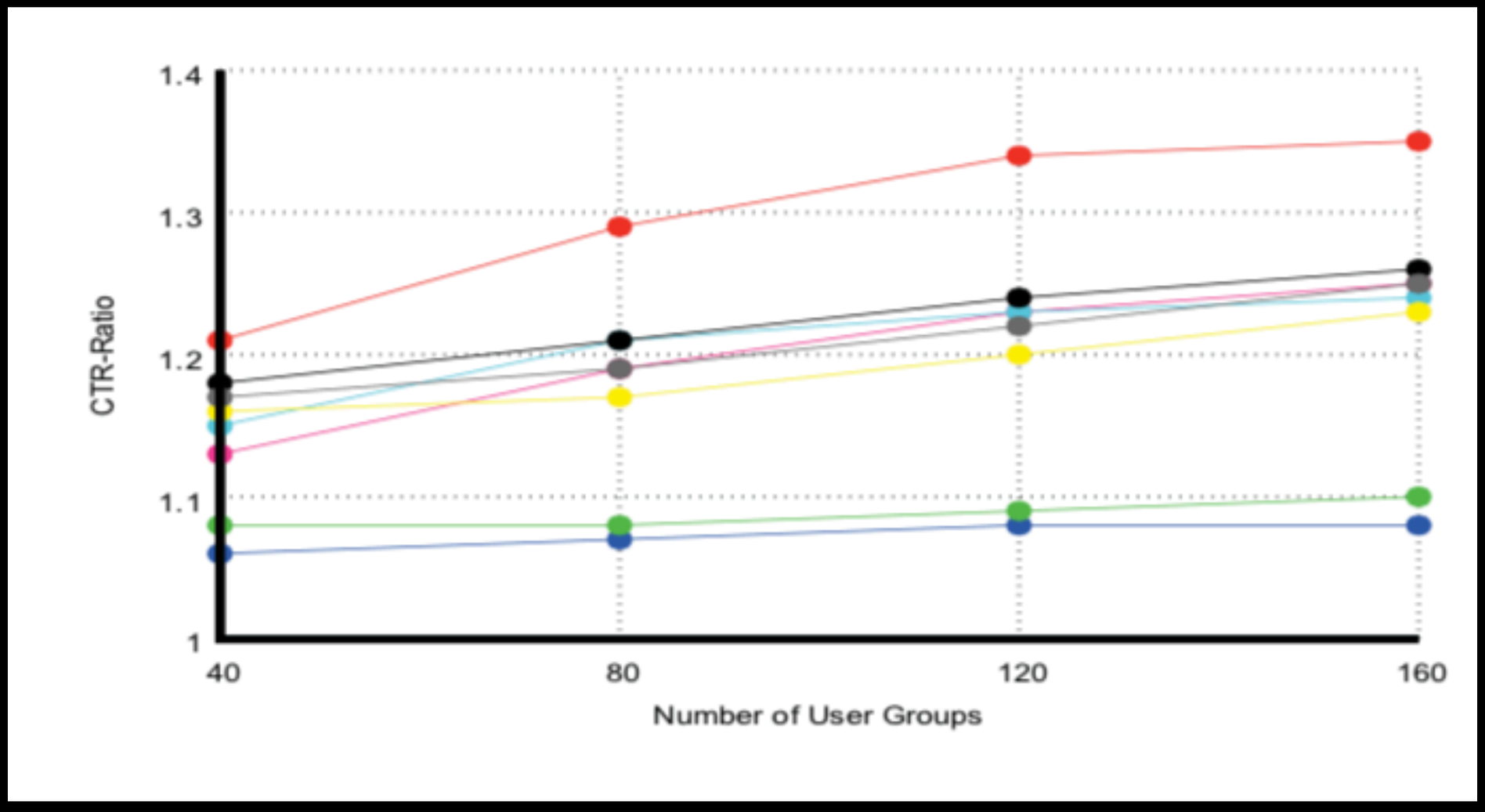
# Experimentos

- 3 Datasets reales (y otros sintéticos)
  - Yahoo Dataset (Métrica: Relative Click Through Rate)
  - Last.FM Dataset (Métrica: Cumulative Reward)
  - Delicious Dataset (Métrica: Cumulative Reward)

# Resultados

- Yahoo Dataset
  - 5-8% Mejora
- Last.FM Dataset y Delicious Dataset
  - Rendimiento se dispara desde iteración 12
- Entre iteración 15 y 25 items nuevos:
  - DMCB: 55%-65%
  - Otros: Sobre 75%

# Resultados

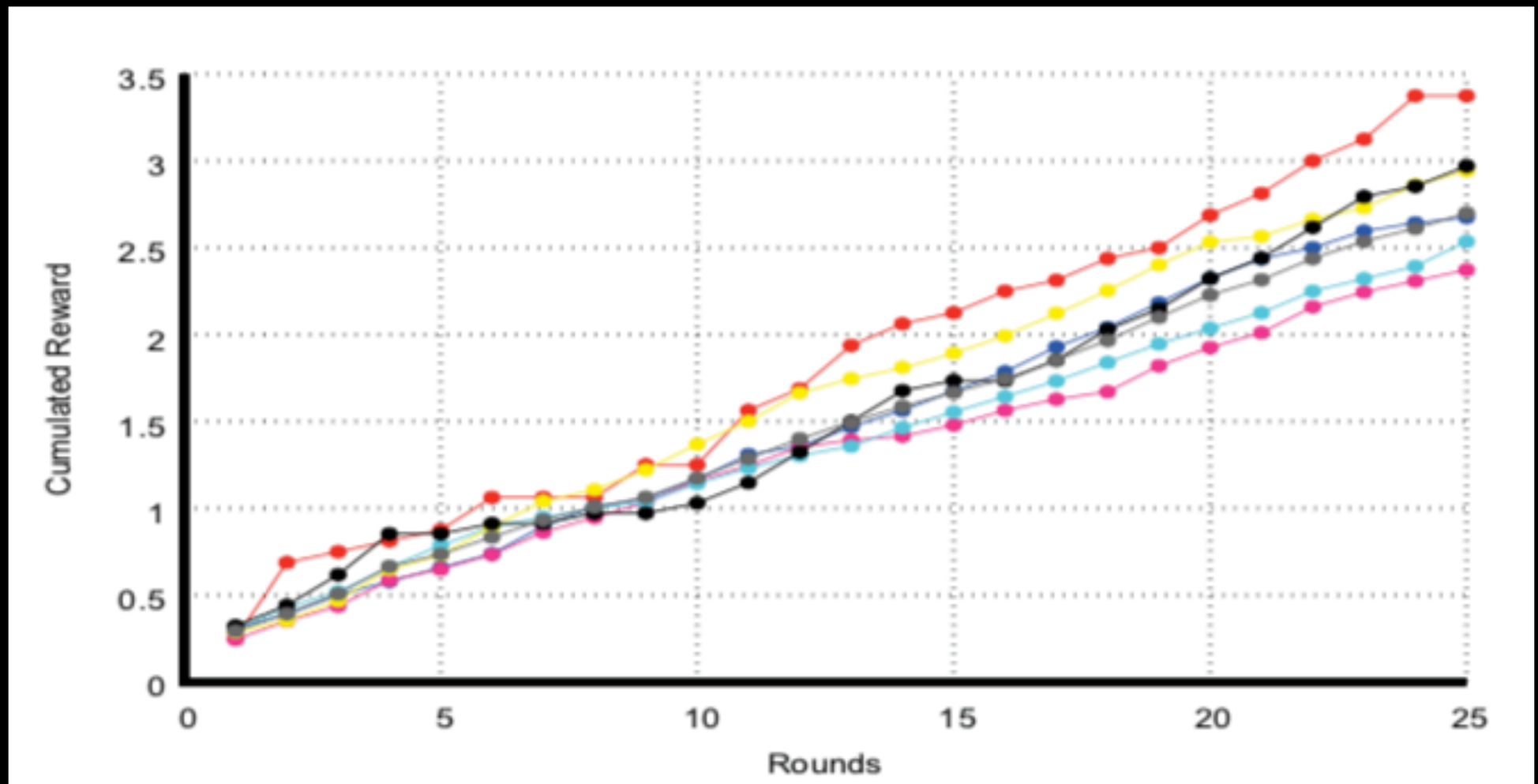



**(b) Policy Evaluation via Online DMCB Training**

DMCB
LinUCB
PTS
UCB-PMF
Hybrid-LinUCB
Hidden-LinUCB
CLUB
CoLin

# Resultados

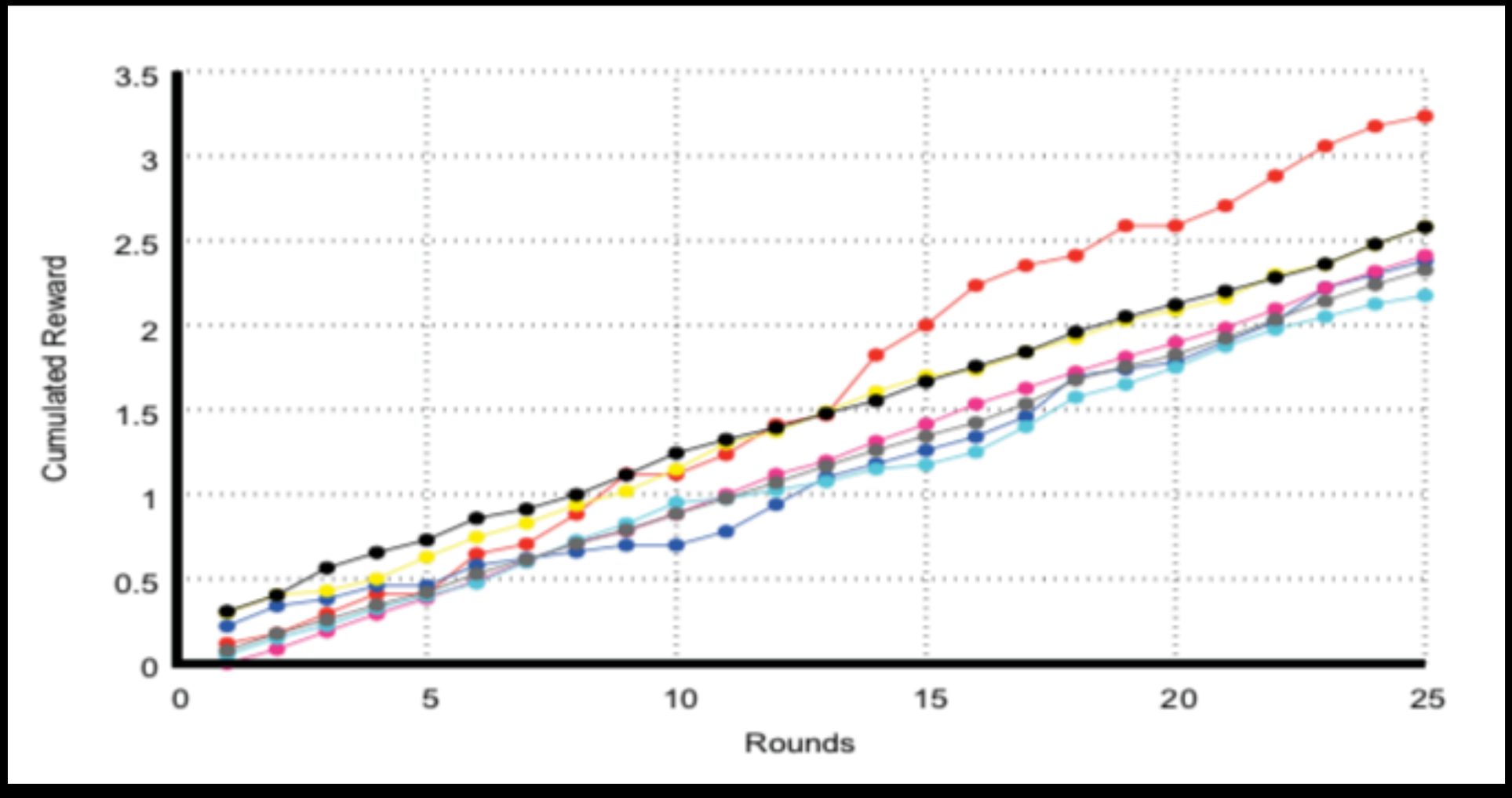
**Last.FM**




**(b) Policy Evaluation via Online DMCB Training**

DMCB
LinUCB
PTS
UCB-PMF
Hybrid-LinUCB
Hidden-LinUCB
CLUB
CoLin

# Resultados




**(b) Policy Evaluation via Online DMCB Training**

DMCB
LinUCB
PTS
UCB-PMF
Hybrid-LinUCB
Hidden-LinUCB
CLUB
CoLin

# Trabajo Futuro

- Modelo actual esta pensado para datasets pequeños, modificarlo para que funcione bien en datasets con muchos items.

- Se quiere diseñar un End-to-End bandit policy learning with latent policy evaluation.