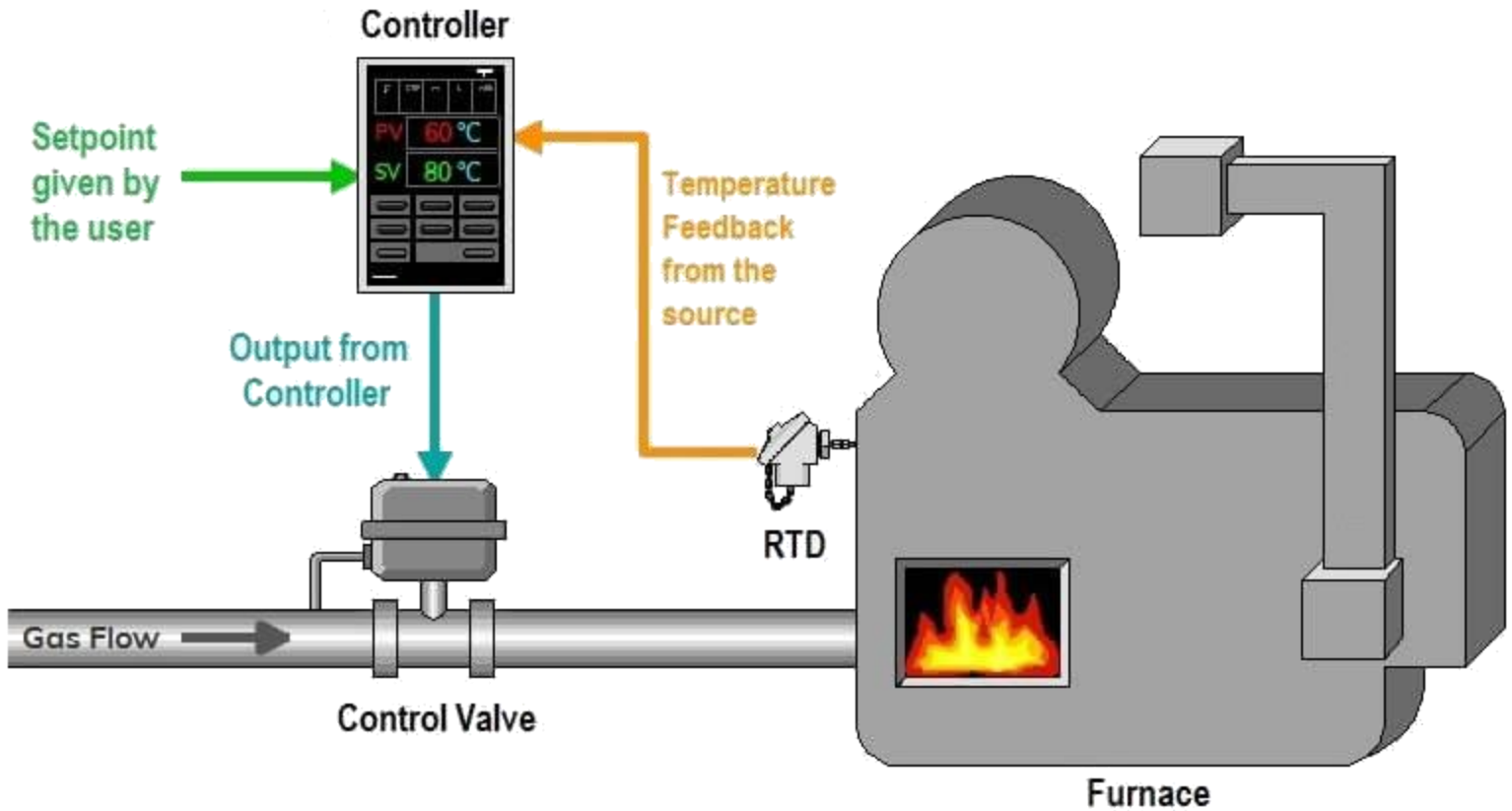
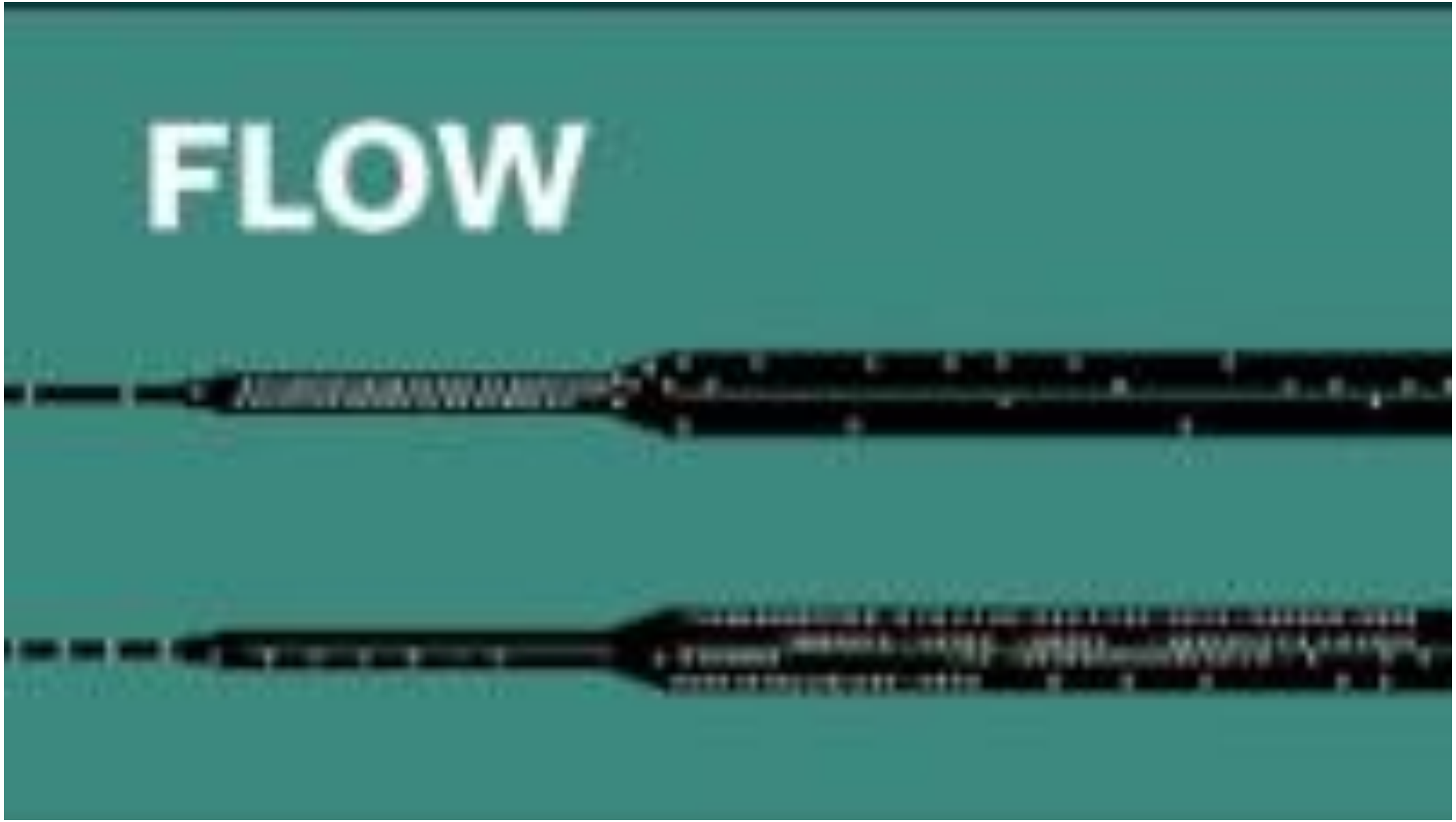
Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 - Inteligencia Artificial

Control de agentes basado en aprendizaje

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación


Controller
PV 60 °C
SV 80 °C
Setpoint given by the user
Temperature Feedback from the source
Output from Controller
RTD
Gas Flow
Control Valve
Furnace

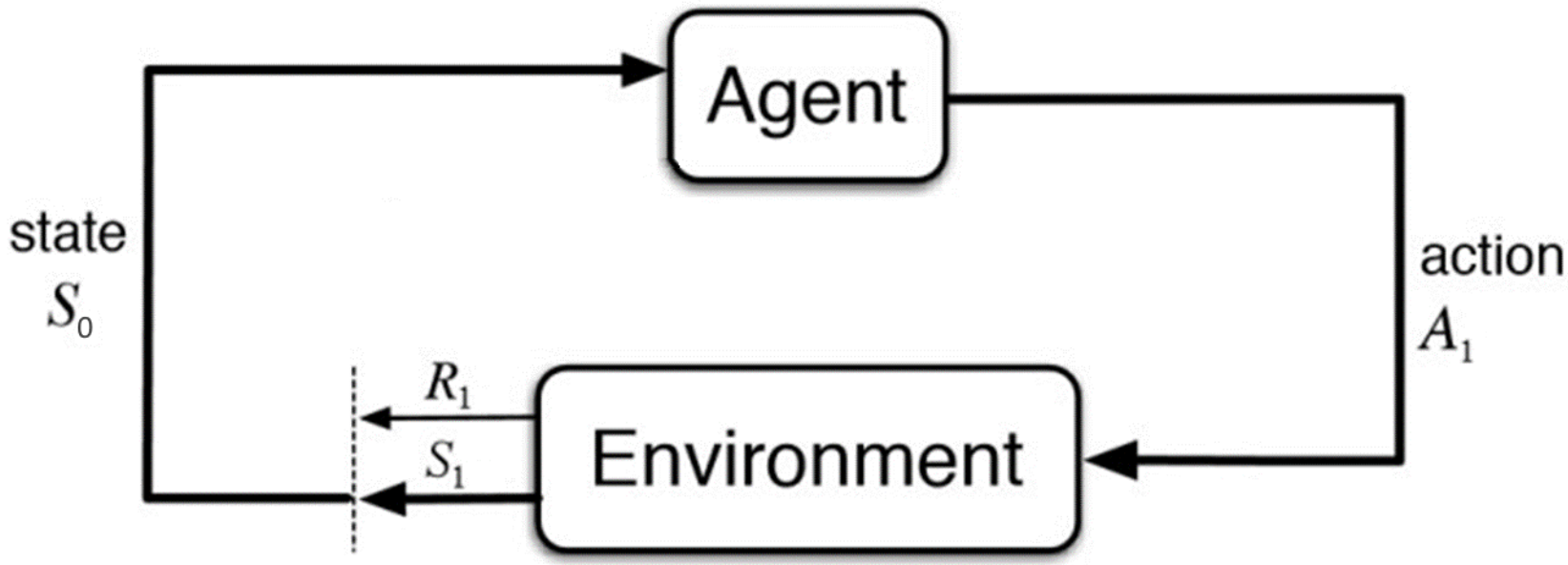


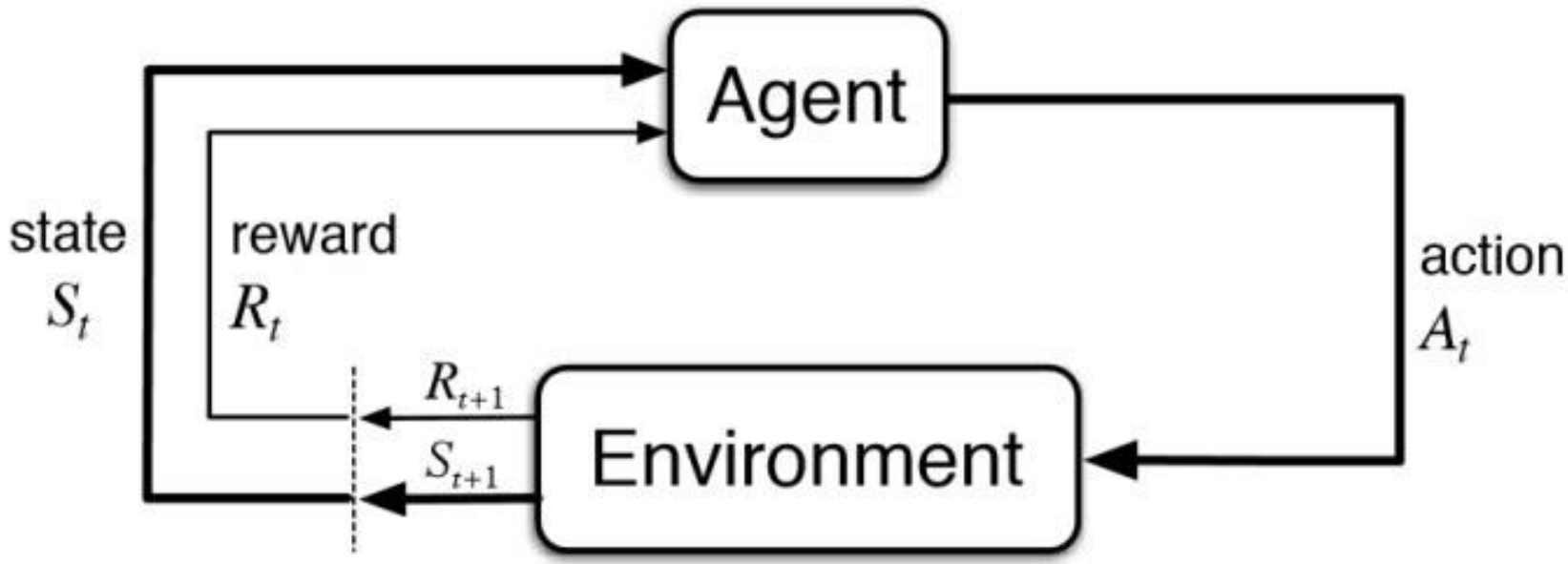
https://www.youtube.com/watch?v=P7xx9uH2i7w

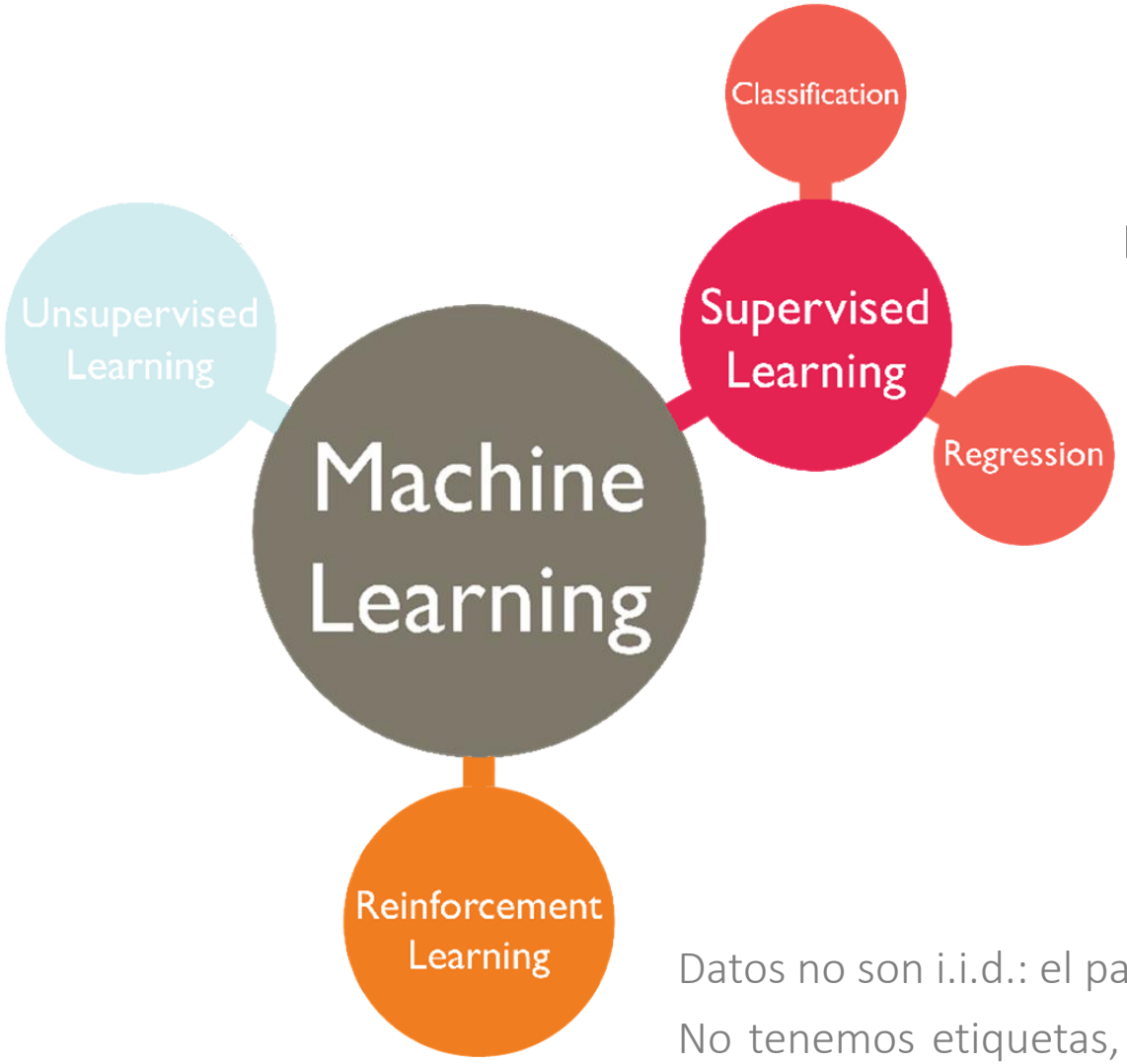
Para esto, utilizaremos aprendizaje reforzado

Aprendizaje reforzado es:

- Formalismo matemático para la toma de decisiones basada en aprendizaje
- Enforque para aprender a tomar decisiones y controlar agentes basado en la experiencia


Agent
state
S_0
action
A_1
R_1
S_1
Environment


Agent
state
S_t
reward
R_t
action
A_t
R_{t+1}
S_{t+1}
Environment



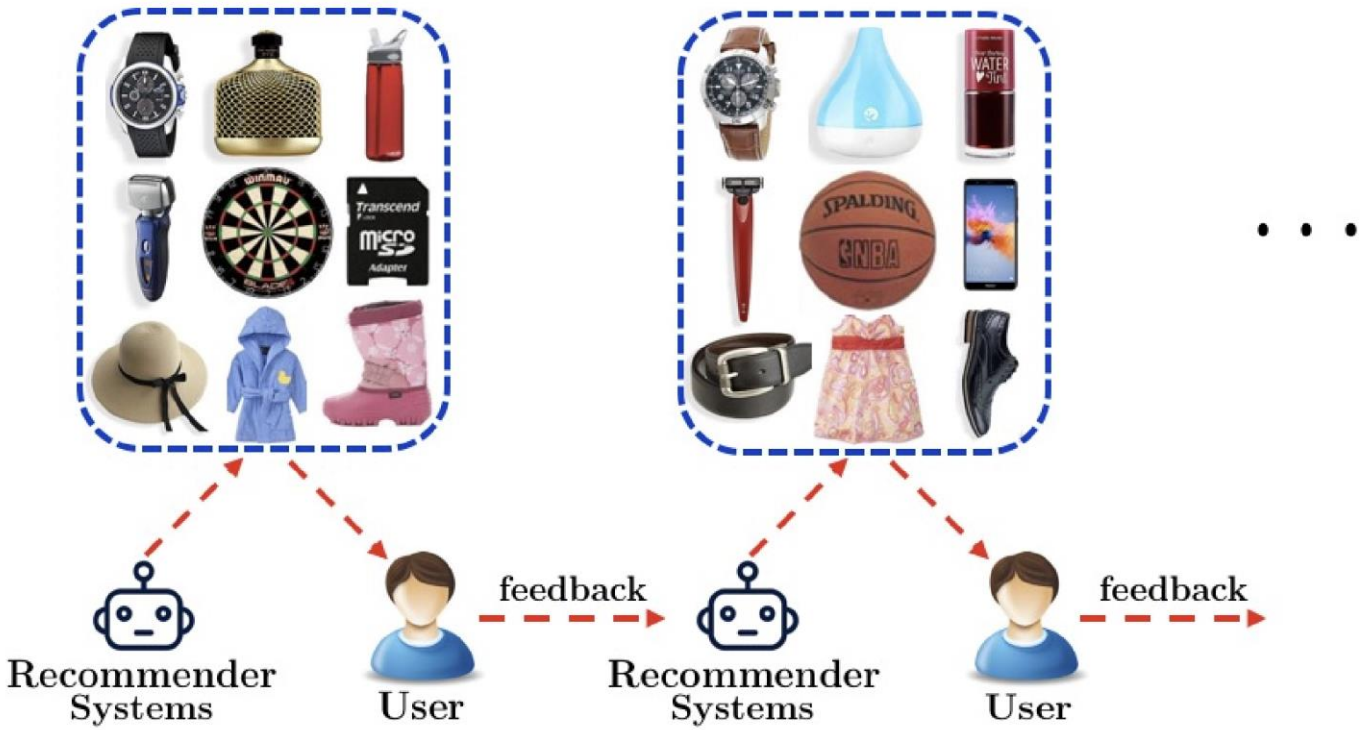
Datos etiquetados e i.i.d.

Datos no son i.i.d.: el pasado influencia el futuro

No tenemos etiquetas, solo sabemos si tuvimos éxito o no (o si recibimos una recompensa)

- Acciones: ?
- Observaciones (estado): ?
- Recompensa: ?

- Acciones: movimientos musculares
- Observaciones (estado): vista, olfato, tacto, oído, gusto
- Recompensa: comida



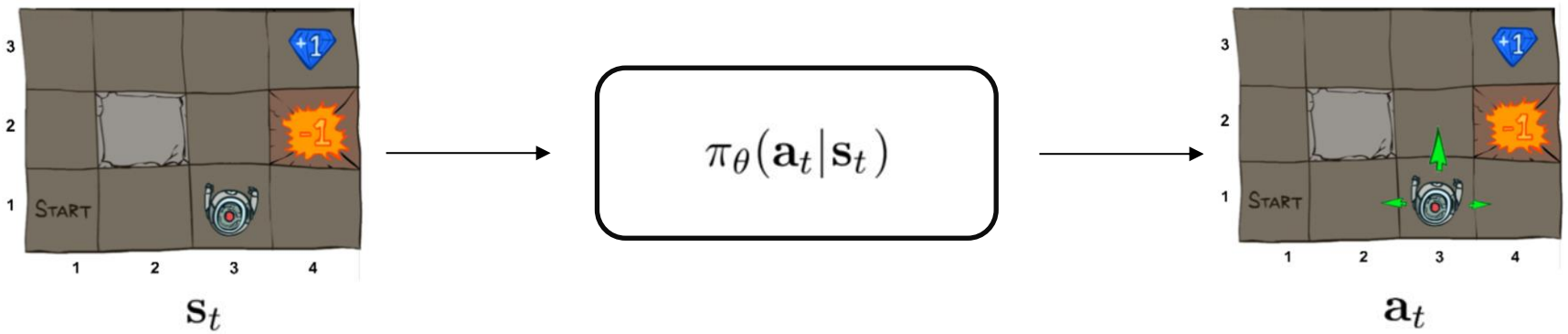
- Acciones: ?
- Observaciones (estado): ?
- Recompensa: ?

- Acciones: qué recomendar al usuario
- Observaciones (estado): elecciones del usuario (historial)
- Recompensa: feedback o elección del usuario (o nada!!)

- Acciones: ?
- Observaciones (estado): ?
- Recompensa: ?

- Acciones: qué y cuánto comprar
- Observaciones (estado): niveles de inventario
- Recompensa: ganancia

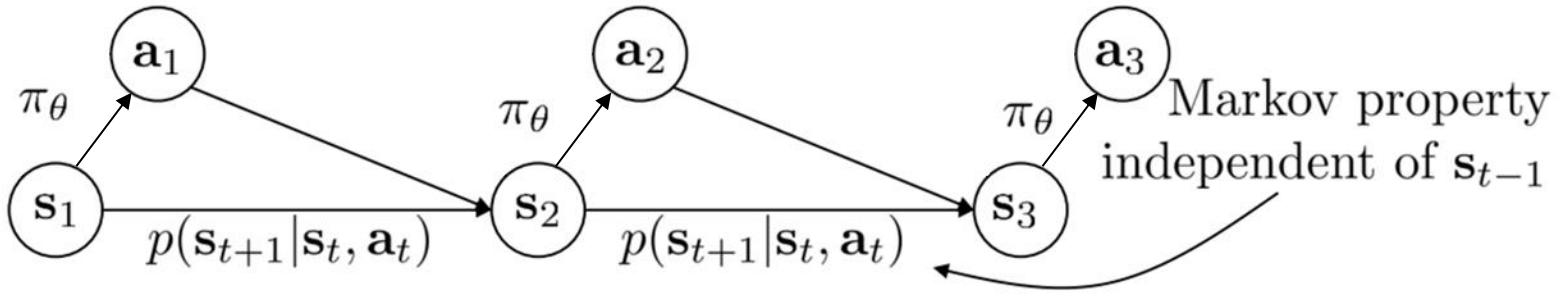
Antes de empezar con las técnicas, un poco de notación



$\mathbf{s}_t$ – state

$\mathbf{a}_t$ – action

$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy



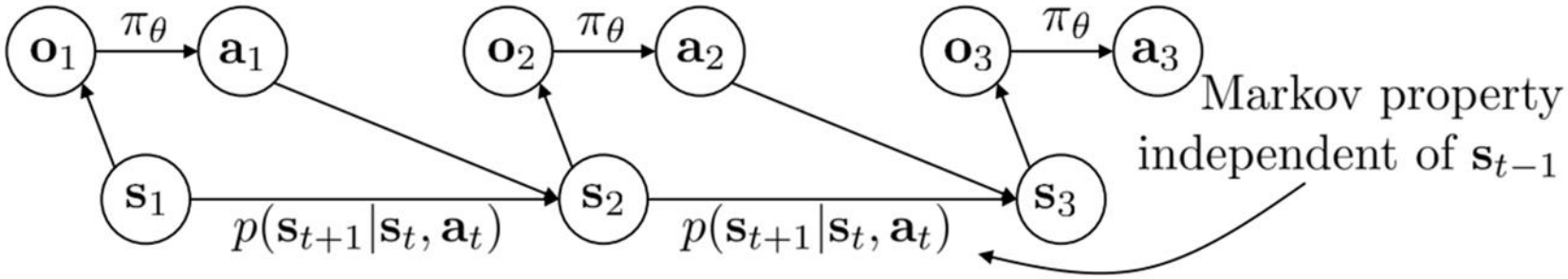
Antes de empezar con las técnicas, un poco de notación

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$

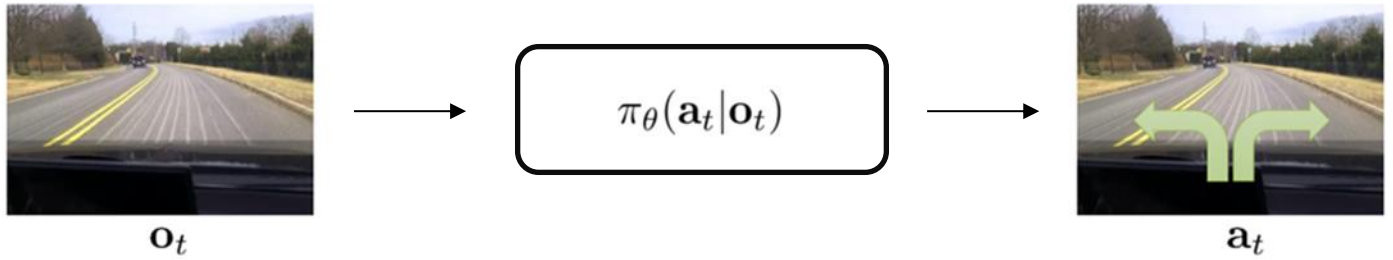
$\mathbf{o}_t$ $\mathbf{a}_t$

$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)



La recompensa actúa como una especie de supervisión



which action is better or worse?

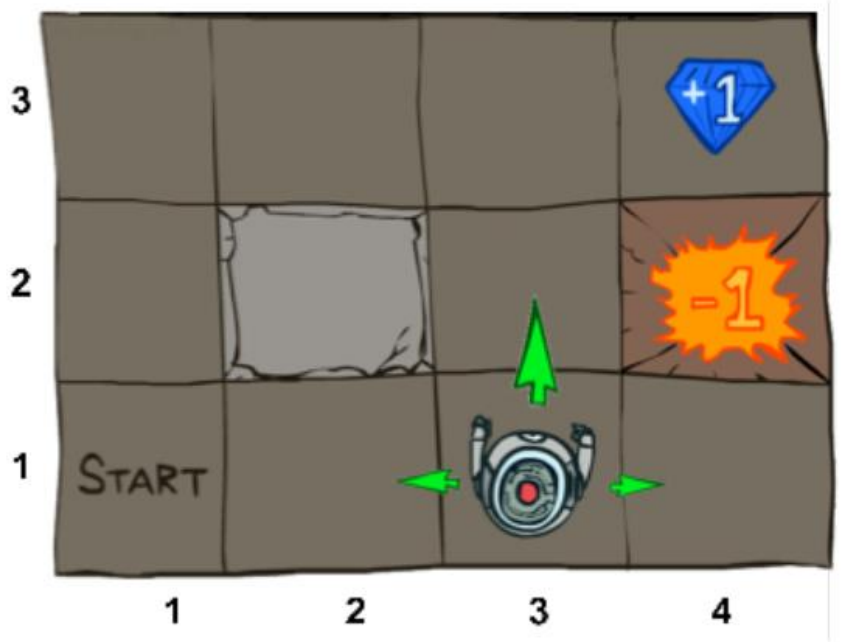
$r(\mathbf{s},\mathbf{a},\mathbf{s}')$: reward function ⟶ tells us which states and actions are better

high reward 

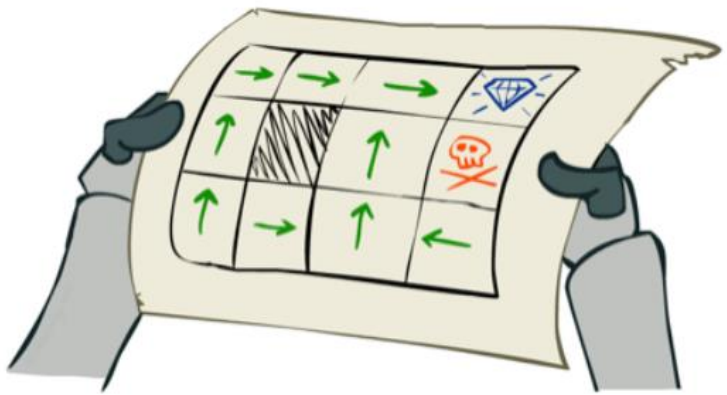
 low reward

$S, A, r(s, a, s')$ y $p(s'|s, a)$ definen un proceso de decisión markoviano (MDP)

Comencemos enfrentando esto de manera intuitiva



π:



$$max_\pi \mathrm{E}[\sum_{t=0}^{H} \gamma^t R(S_t, A_t, S_{t+1}) | \pi]$$

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

INF3813 – Deep Learning Avanzado

Control de agentes basado en aprendizaje

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación