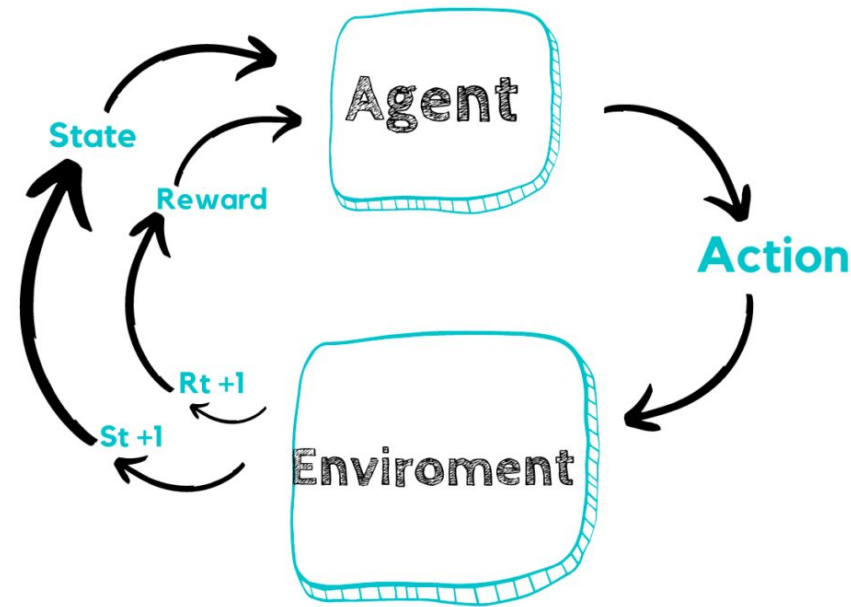


# *Aprendizado por Reforço*



Artur Hugo (18/0030400)

Felipe Neves (18/0016296)

# O que é aprendizado por reforço?

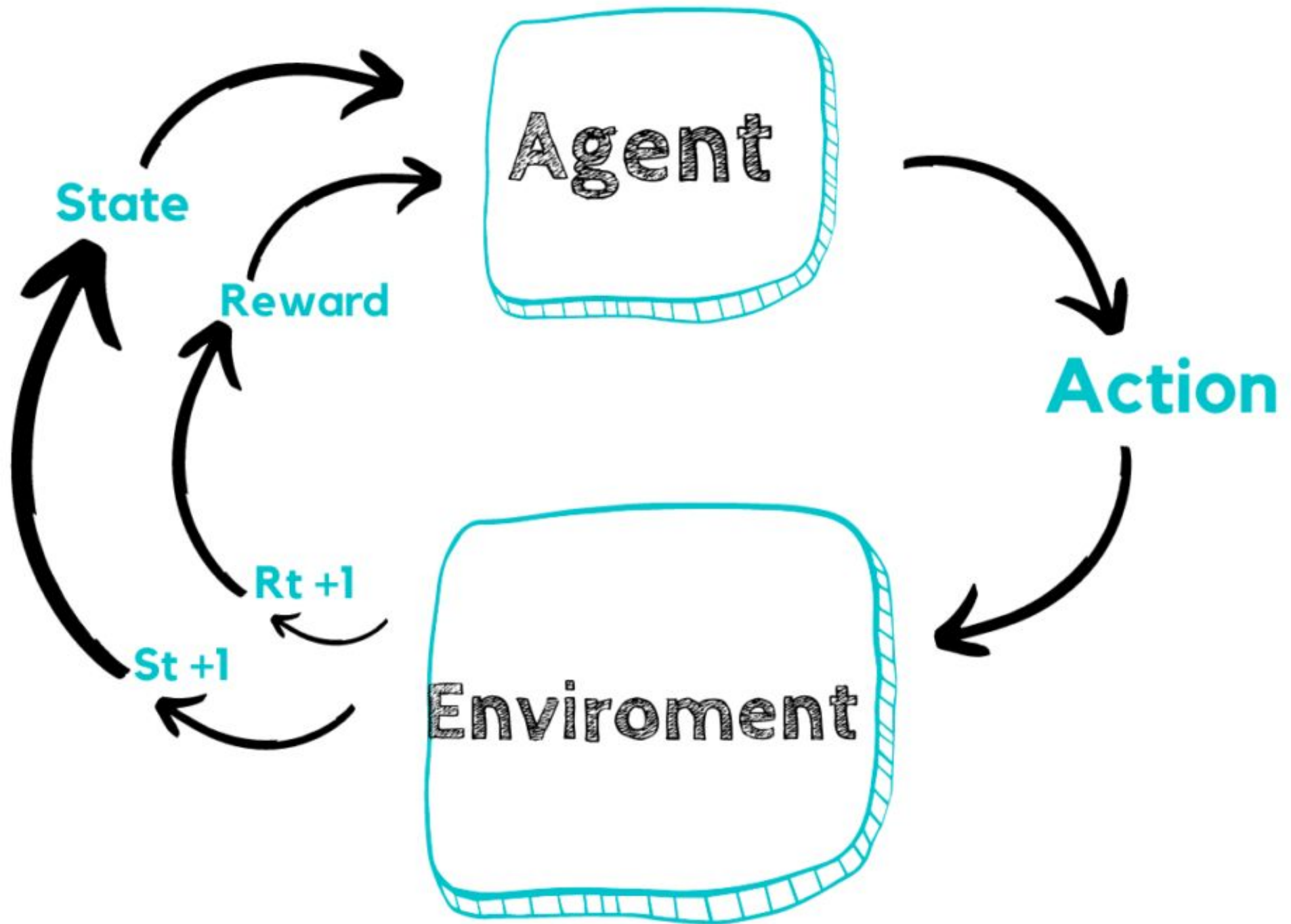
*Aprendizado por reforço é a área de aprendizado de máquina que estuda como agentes inteligentes desempenham ações em um ambiente de forma a maximizar a noção de recompensa cumulativa*

# O que é aprendizado por reforço?

- Aprendizado a partir de recompensas e penalidades
- Sem uso de dados pré-classificados (labels)
- Exploração vs Otimização (não há conhecimento prévio)
- Decisões sequenciais (consequência das características acima)

# Conceitos Básicos:

- Agente: Realiza ações + Ganha recompensas
- Ambiente: cenário no qual agente se encontra
- Estado: configuração atual do ambiente
- Recompensa: consequência imediata recebida ao realizar ação
- Política: estratégia usada pelo agente para escolher próxima ação
- Valor: retorno esperado a longo-prazo



# Aplicações

- Jogos

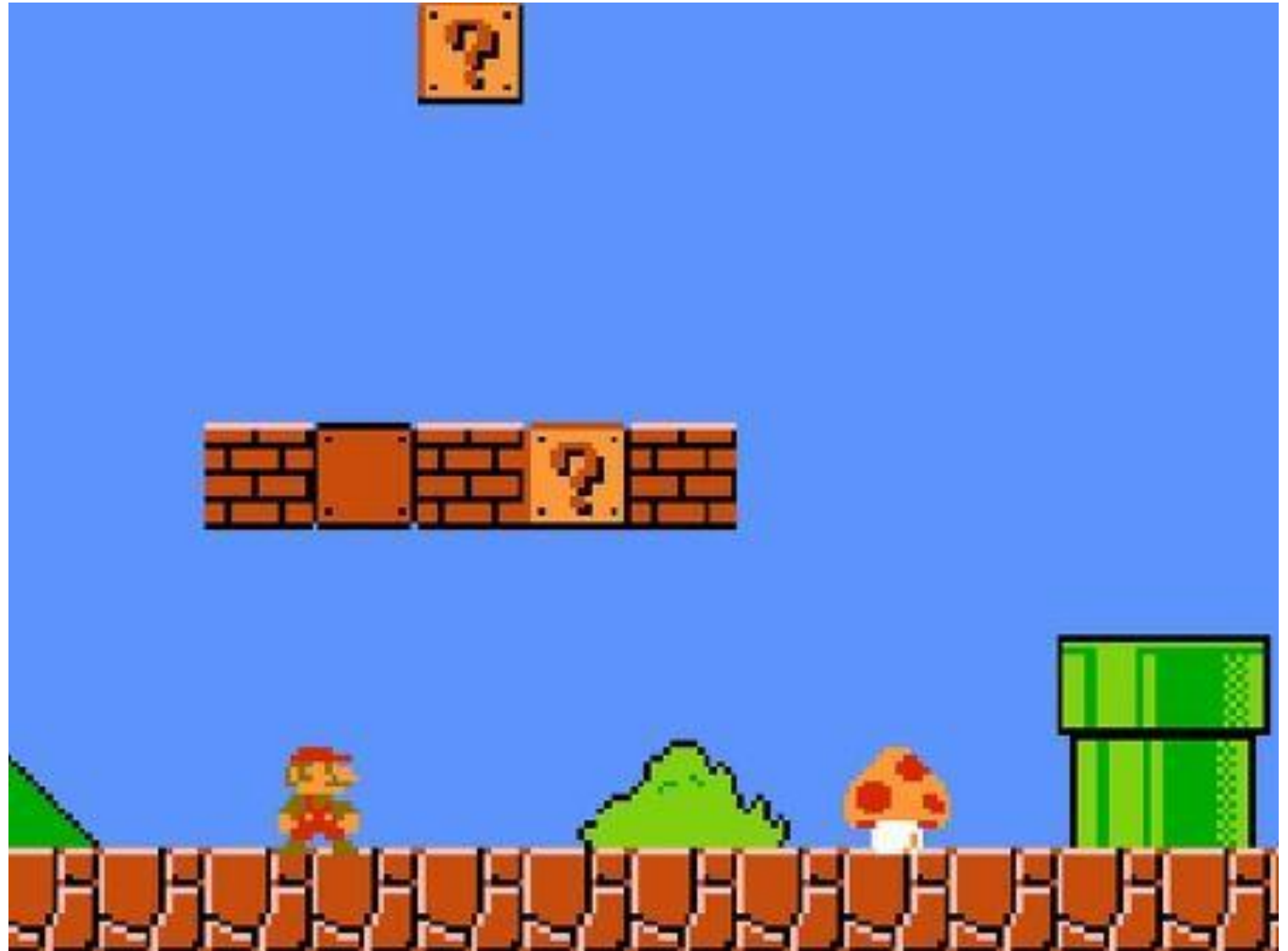


Imagem retirada  
de <https://www.uol.com.br/start/ultimas-noticias/2021/07/21/cientistas-criam-robo-capaz-de-jogar-super-mario-bros-veja-video.htm>

# Aplicações

- Jogos
- Redes

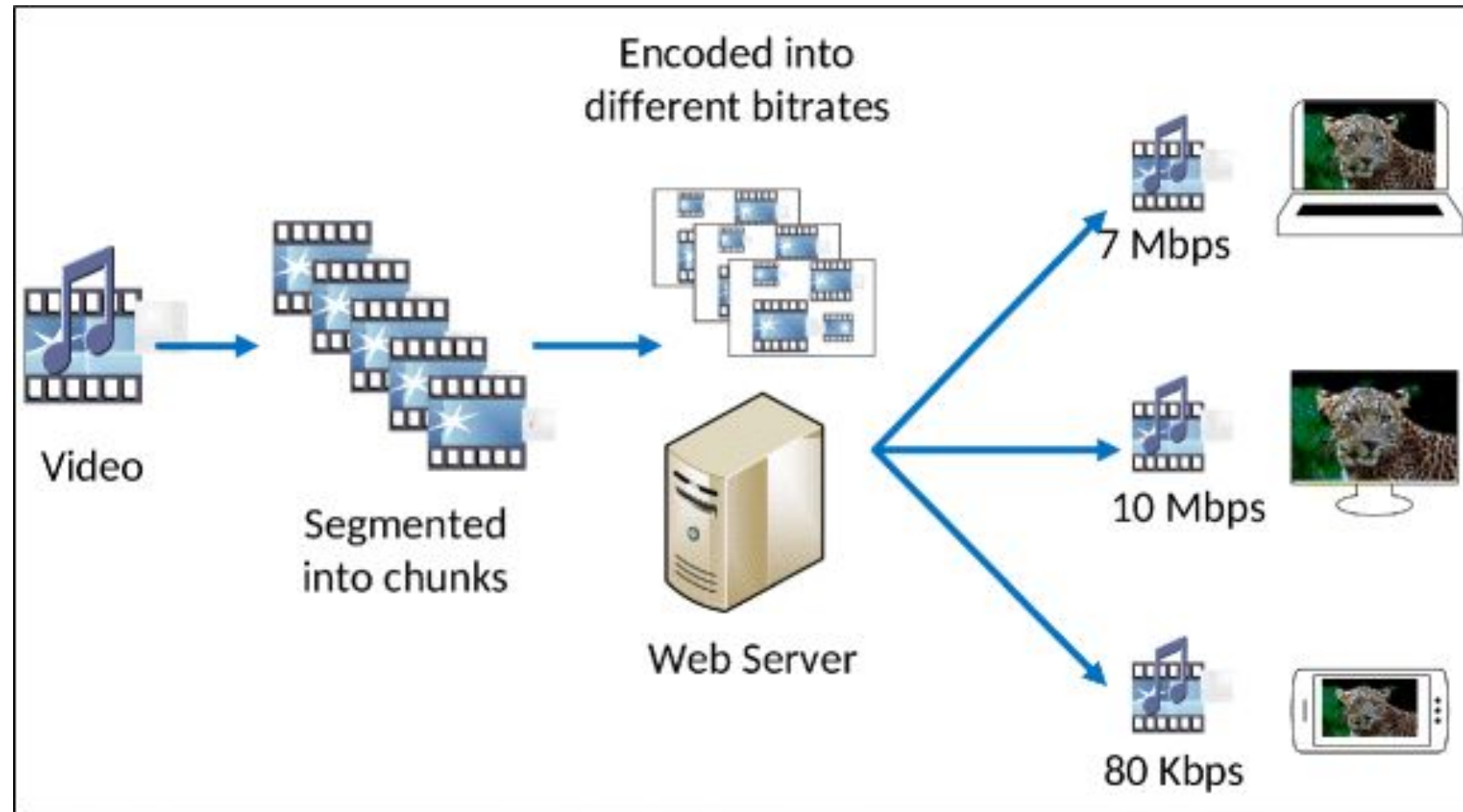


Imagem retirada de <https://www.gta.ufrj.br/ensino/eel879/vf/mpeg-dash/>

# Aplicações

- Jogos
- Redes
- Robótica



Imagem retirada de <https://www.gta.ufrj.br/ensino/eel879/vf/mpeg-dash/>



# Aplicações

- Jogos
- Redes
- Robótica
- NLP
- E muito mais

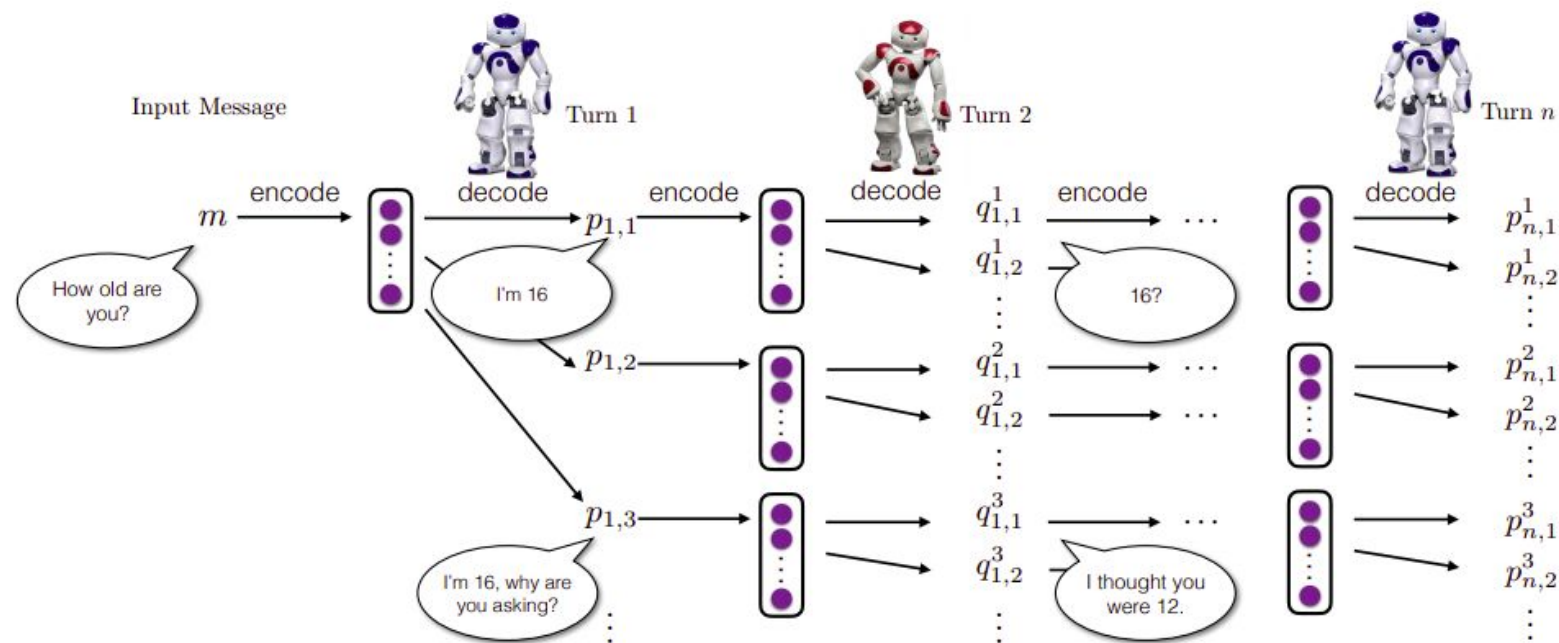



Figure 1: Dialogue simulation between the two agents.

Imagem retirada de <https://neptune.ai/blog/reinforcement-learning-applications>

# Algoritmos de Aprendizado por Reforço:



Algorithm	Description	Policy	Action Space	State Space	Operator
Monte Carlo	Every visit to Monte Carlo	Either	Discrete	Discrete	Sample-means
Q-learning	State-action-reward-state	Off-policy	Discrete	Discrete	Q-value
SARSA	State-action-reward-state-action	On-policy	Discrete	Discrete	Q-value
Q-learning - Lambda	State-action-reward-state with eligibility traces	Off-policy	Discrete	Discrete	Q-value
SARSA - Lambda	State-action-reward-state-action with eligibility traces	On-policy	Discrete	Discrete	Q-value
DQN	Deep Q Network	Off-policy	Discrete	Continuous	Q-value
DDPG	Deep Deterministic Policy Gradient	Off-policy	Continuous	Continuous	Q-value
A3C	Asynchronous Advantage Actor-Critic Algorithm	On-policy	Continuous	Continuous	Advantage
NAF	Q-Learning with Normalized Advantage Functions	Off-policy	Continuous	Continuous	Advantage
TRPO	Trust Region Policy Optimization	On-policy	Continuous	Continuous	Advantage
PPO	Proximal Policy Optimization	On-policy	Continuous	Continuous	Advantage
TD3	Twin Delayed Deep Deterministic Policy Gradient	Off-policy	Continuous	Continuous	Q-value
SAC	Soft Actor-Critic	Off-policy	Continuous	Continuous	Advantage

# O que é Q-Learning?

- Espaço de estados **S**
- Espaço de ações **A**
- Q-Tabela: todos os pares **(s,a)**
- Taxa de aprendizado  **$\alpha$**  e fator de desconto  **$\gamma$**

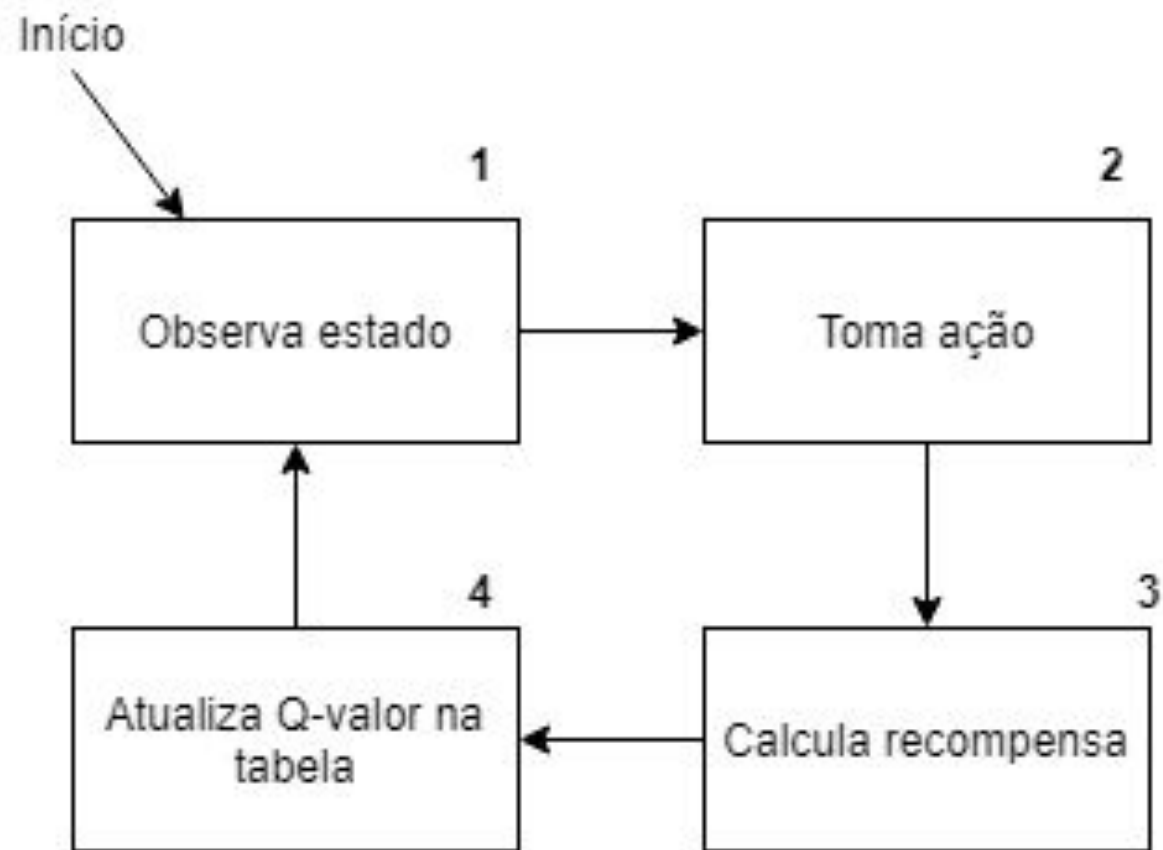
$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(R_t + \gamma \max_a Q(S_{t+1}, a))$$

# Como o agente escolhe uma ação?

- Exploration x Exploitation
- Política de exploração  $\pi(a|s)$
- Política  **$\epsilon$ -greedy**:
  - Escolhe ação ótima com probabilidade  $1 - \epsilon$
  - Escolhe ação aleatória com probabilidade  $\epsilon$
- Decaimento do fator de exploração

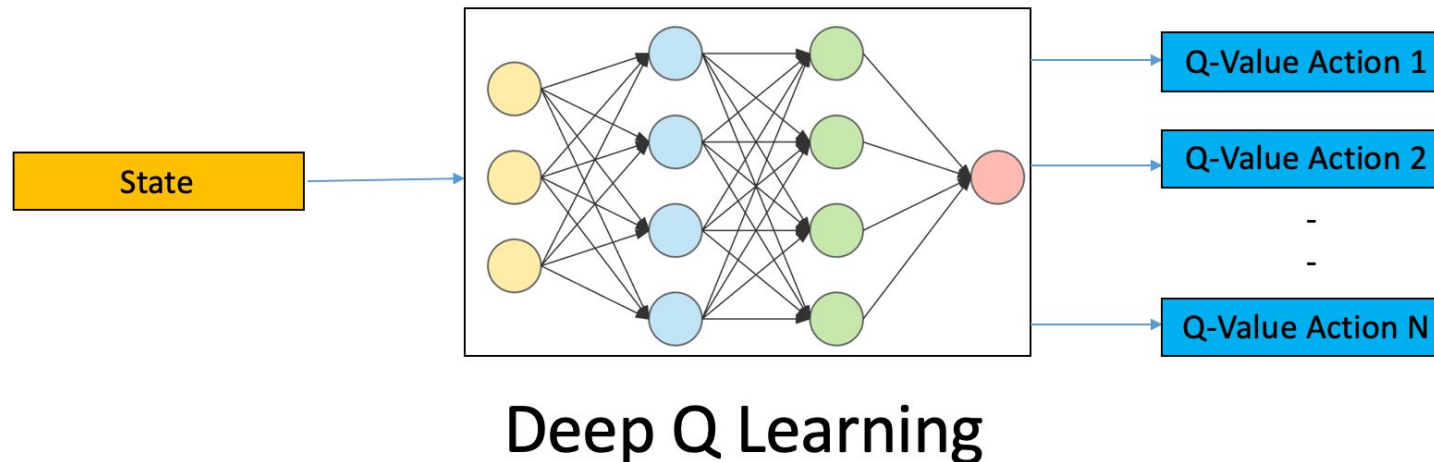
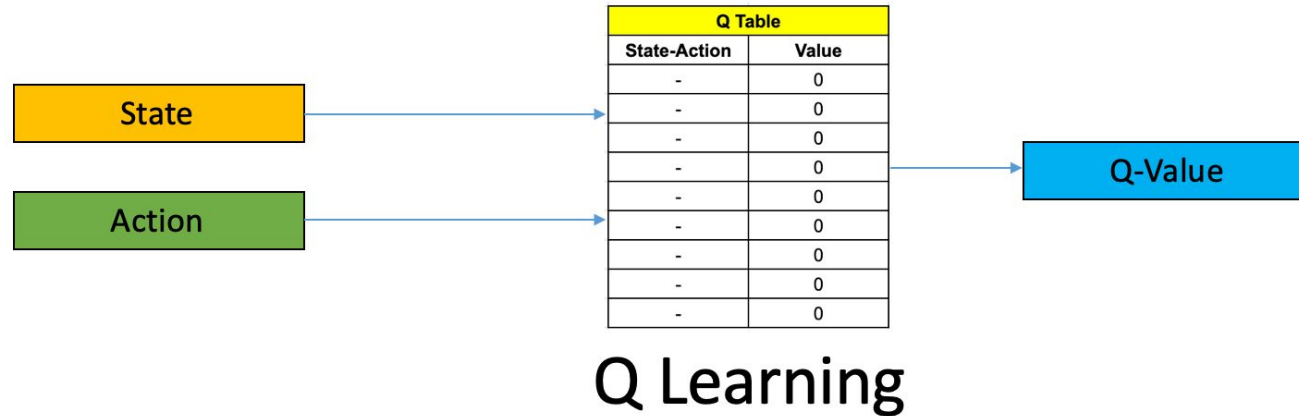
# Processo iterativo

- Agente começa um episódio
- Cada passo do episódio:
- Verifica se acabou o episódio (e.g. alcançou objetivo ou excedeu limite de passos)



# Demonstração de Q-Learning

# Abordagem clássica vs Aprendizado profundo



# Supervisado vs Não supervisionado vs Reforço

	Aprendizado Supervisionado	Aprendizado não Supervisionado	Aprendizado por reforço
Tipos de Problemas	Regressão Classificação	Associação Clustering	Baseados em recompensa
Organização de dados	Dados rotulados	Dados não rotulados	Dados não predefinidos*
Treinamento	Supervisão externa	Sem supervisão	Sem supervisão
Pontos Fracos	<ul style="list-style-type: none"><li>- Trabalho de rotulação</li><li>- Tempo de computação de treinamento</li></ul>	<ul style="list-style-type: none"><li>- Menor acurácia de resultados</li><li>- necessidade de interpretação dos resultados</li></ul>	<ul style="list-style-type: none"><li>- Aprender do zero, sem usar instruções</li><li>- Problemas com espaços de ação grandes</li><li>- extremamente caro em computação e tempo</li></ul>



# Referências

- <https://deeplizard.com/learn/video/nyjbcRQ-uQ8>
- <https://www.youtube.com/watch?v=OYhFoMySoVs>
- [https://towardsdatascience.com/the-complete-reinforcement-learning-dictionary-e16230b7d24e#:~:text=Episode%3A%20All%20states%20that%20come,it%20receives%20during%20an%20episode.&text=Episodic%20Tasks%3A%20Reinforcement%20Learning%20tasks,episode%20has%20a%20terminal%20state\).](https://towardsdatascience.com/the-complete-reinforcement-learning-dictionary-e16230b7d24e#:~:text=Episode%3A%20All%20states%20that%20come,it%20receives%20during%20an%20episode.&text=Episodic%20Tasks%3A%20Reinforcement%20Learning%20tasks,episode%20has%20a%20terminal%20state).)
- [https://en.wikipedia.org/wiki/Reinforcement\\_learning#Associative\\_reinforcement\\_learning](https://en.wikipedia.org/wiki/Reinforcement_learning#Associative_reinforcement_learning)
- [https://en.wikipedia.org/wiki/Q-learning#Deep\\_Q-learning](https://en.wikipedia.org/wiki/Q-learning#Deep_Q-learning)
- <https://www.guru99.com/reinforcement-learning-tutorial.html>
- [https://www.youtube.com/watch?v=iKdIKYG78j4&ab\\_channel=Dr.DanielSoper](https://www.youtube.com/watch?v=iKdIKYG78j4&ab_channel=Dr.DanielSoper)

Obrigado, :)