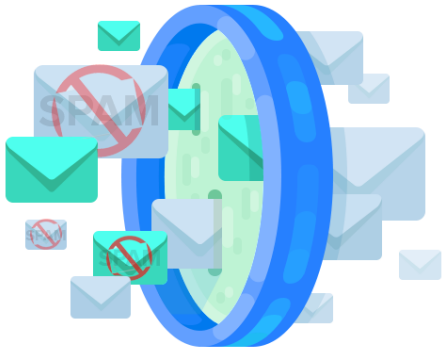
Introduction to Reinforcement Learning

Mohammad **Dehghani**

Mechanical and Industrial Engineering Department

Jan-20

Introduction



Email Filters in Gmail

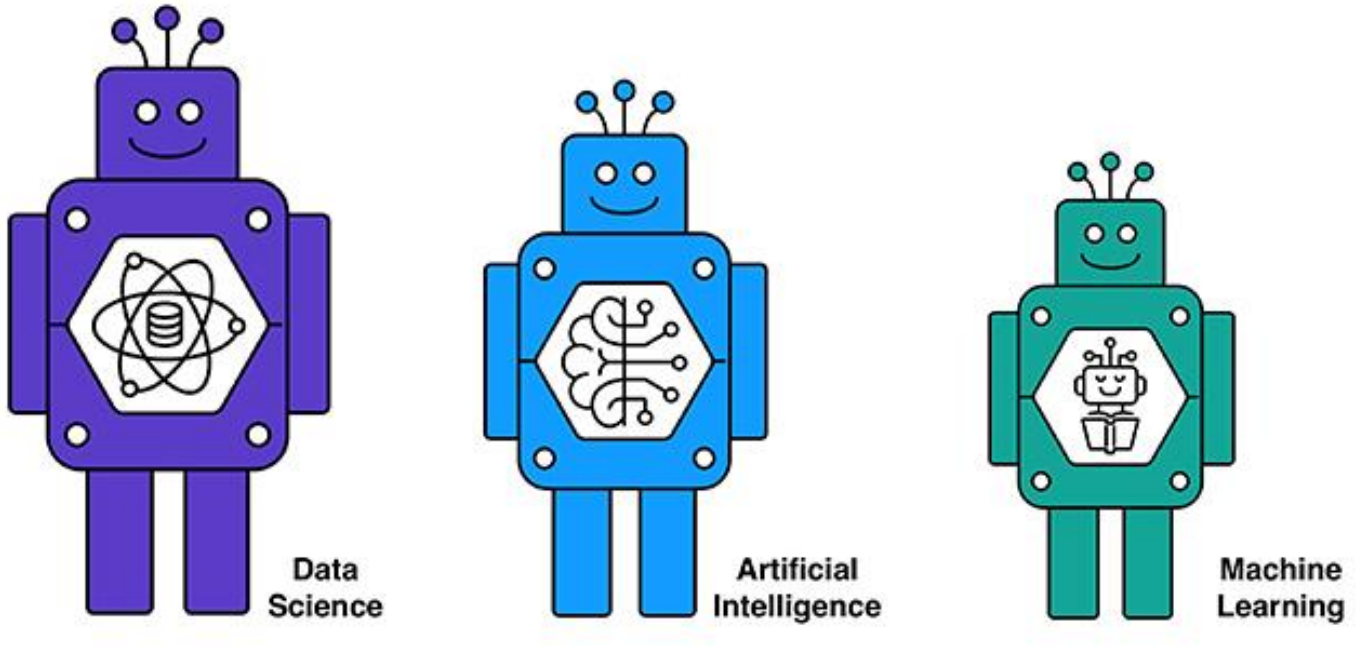
Product Recommendations

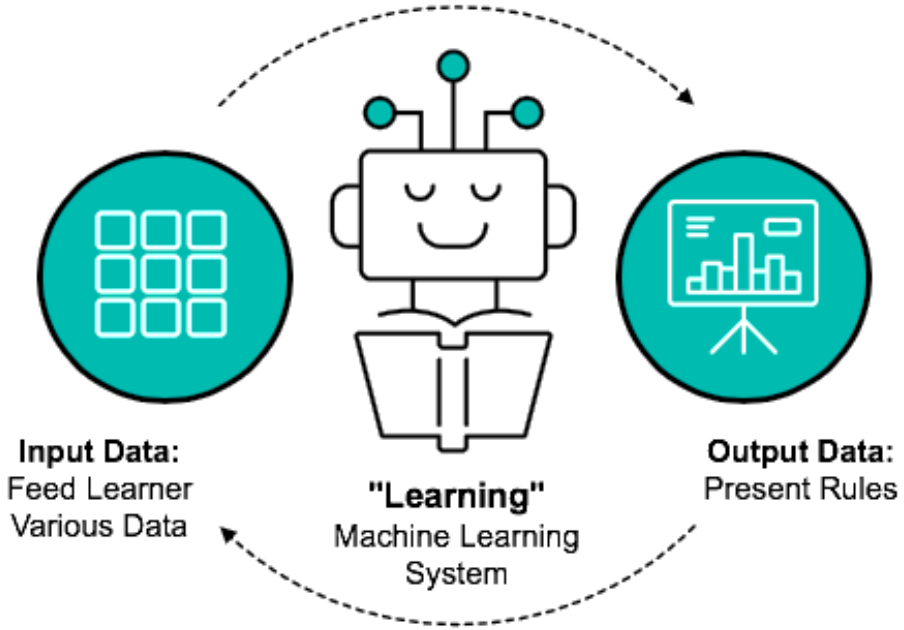
Airline Flights

Manually flying? 7 min

Ride-sharing Apps

Machine Learning





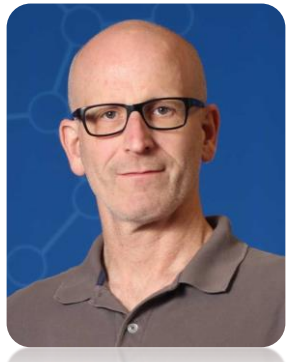
A critique of AI

Herbert A. Simon

- Pioneer in the foundations of AI
- 1978 Nobel Prize in economics

- **Simon predicted in 1965**
  - "machines will be capable, within twenty years, of doing any work a man can do"

A critique of AI



Anthony Zador

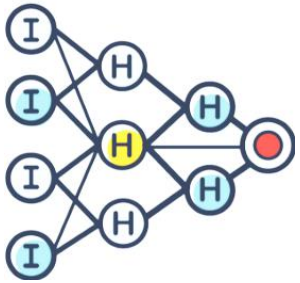
- Professor of Biology
- Cold Spring Harbor Laboratory

nature communications

A critique of pure learning and what artificial neural networks can learn from animal brains," (2019)

"ANNs remain far from approaching human intelligence"

- ANNs can crush human opponents in games, they cannot approach the cognitive capabilities of a four-year old.



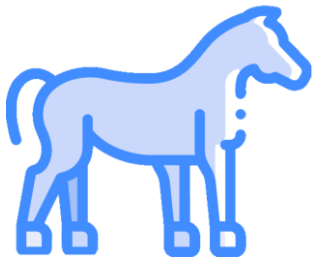
ANN

- 10^7 "labeled" examples

Newborn Baby

- 10^7 observations/sec ~ 1 Year

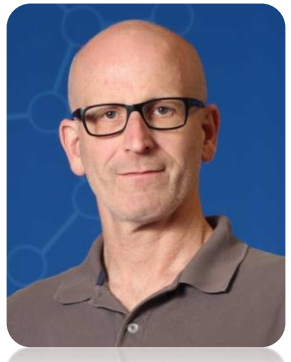
Squirrel



Colt

Spider

A critique of AI



Anthony Zador

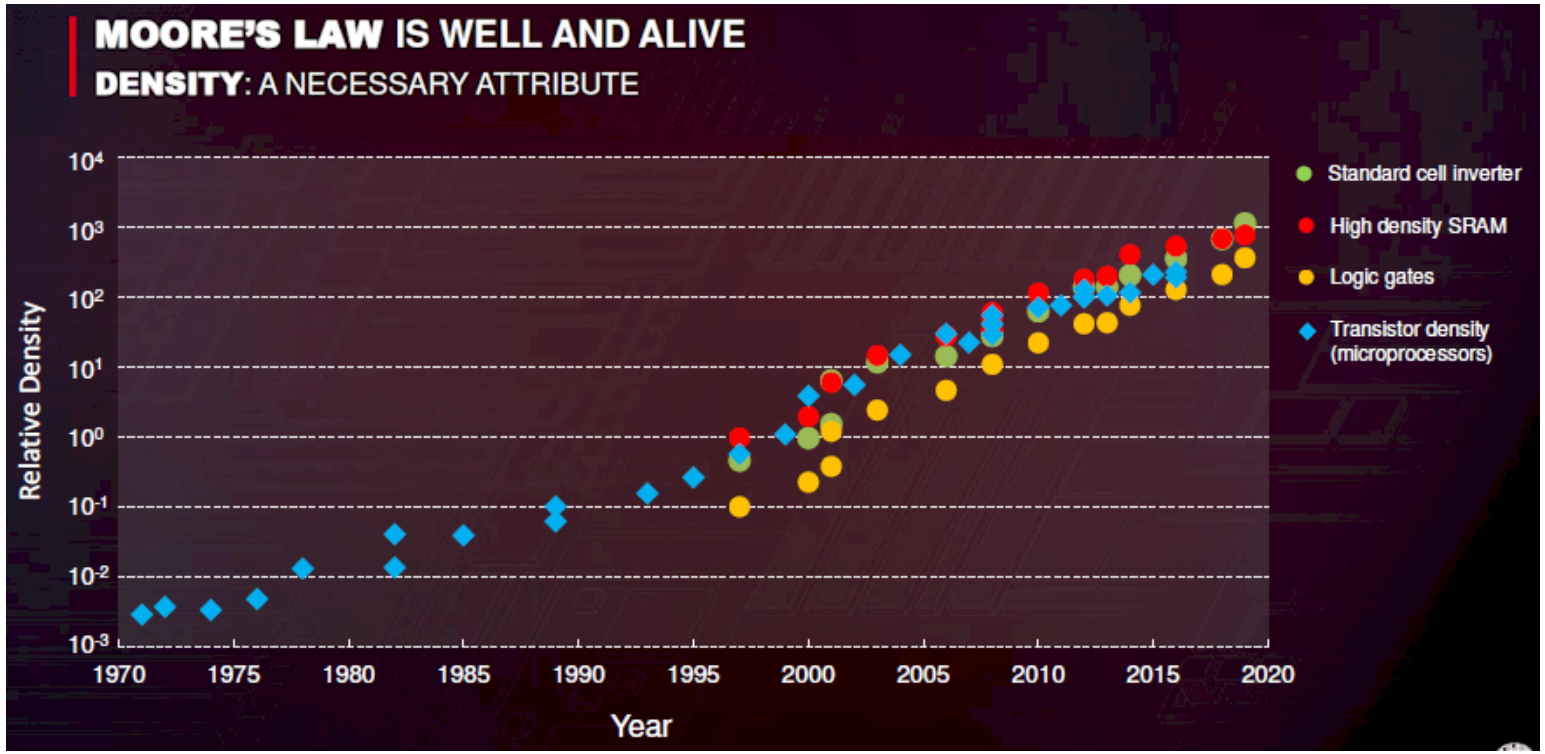
- Professor of Biology
- Cold Spring Harbor Laboratory

nature communications

A critique of pure learning and what artificial neural networks can learn from animal brains," (2019)

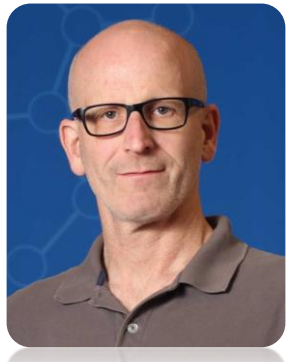
"ANNs remain far from approaching human intelligence"

- ANNs can crush human opponents in games, they cannot approach the cognitive capabilities of a four-year old.



Source: https://curiosity.com/ : number of transistors on a microchip doubles every two years

A critique of AI



Anthony Zador

- Professor of Biology
- Cold Spring Harbor Laboratory

nature communications

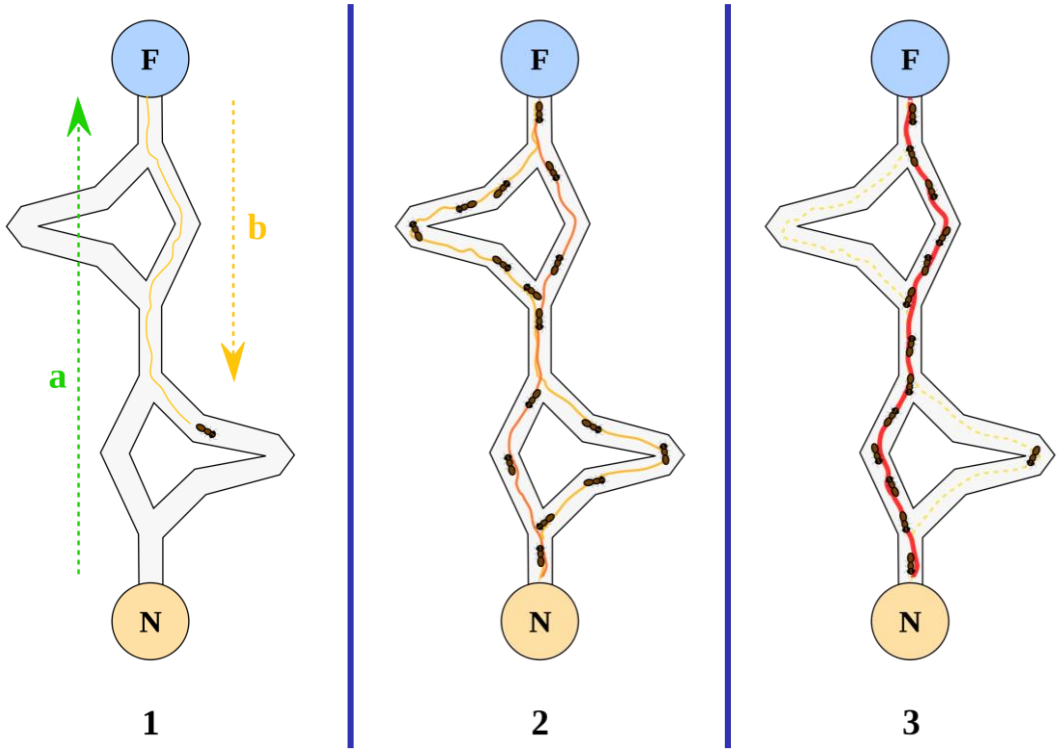
A critique of pure learning and what artificial neural networks can learn from animal brains," (2019)

- **Much of an animal's behavior is innate.**
  - These innate behaviors and representations arise through evolution by natural selection.
- **Evolution can be thought of as a kind of *reinforcement algorithm*, operating on the timescale of generations.**
- **Most of the data that contribute an animal's fitness are encoded by evolution into the genome**
  - Evolution, like learning, can also be viewed as a mechanism for extracting statistical regularities, albeit on a much longer time scale than learning

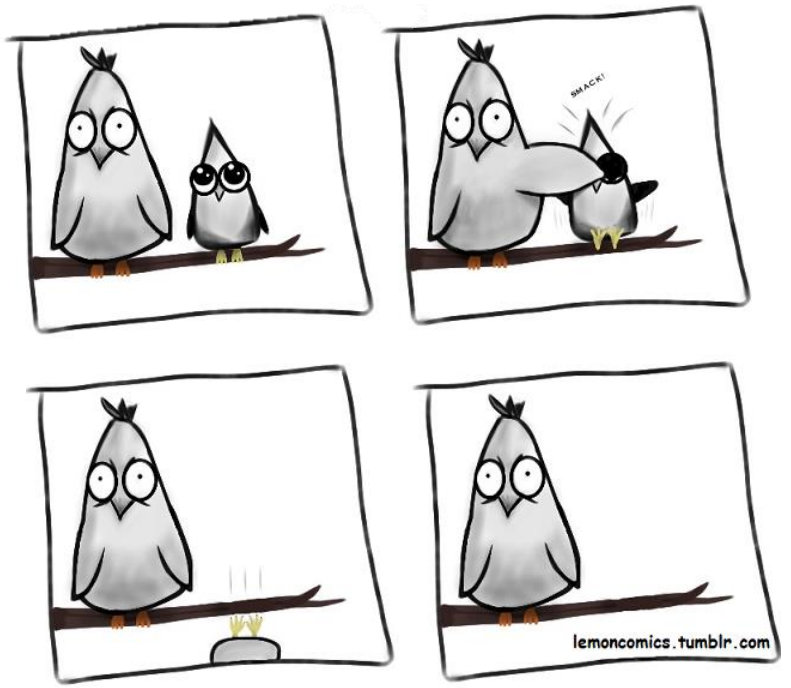
Learning

- **Learning is a change in behavior that occurs as a result of experience.**
  - ***Ability to learn*** is the most striking behavioral characteristic of spices (particularly mammals)

Ants
(What is the shortest path?)



Birds
(how to fly?)



Learning

- **Learning is a change in behavior that occurs as a result of experience.**
  - ***Ability to learn*** is the most striking behavioral characteristic of spices (particularly mammals)

Evolution Theory

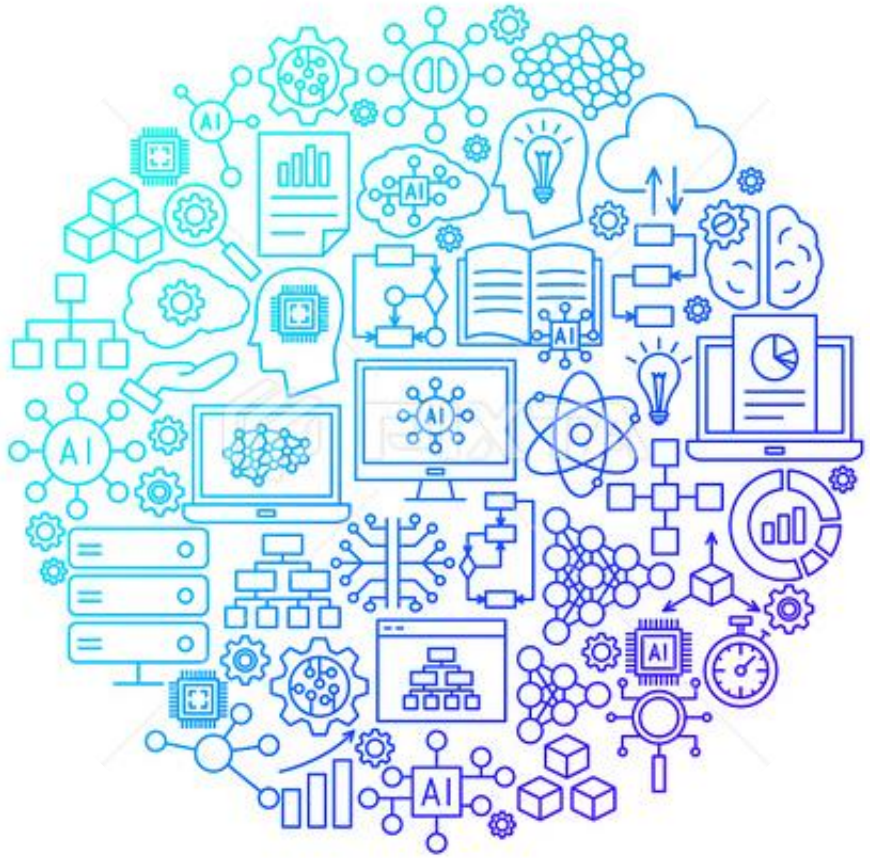
  - Human infants can discriminate faces soon after birth
  - Monkeys raised with no exposure to faces show a preference for faces upon first exposure

Learning

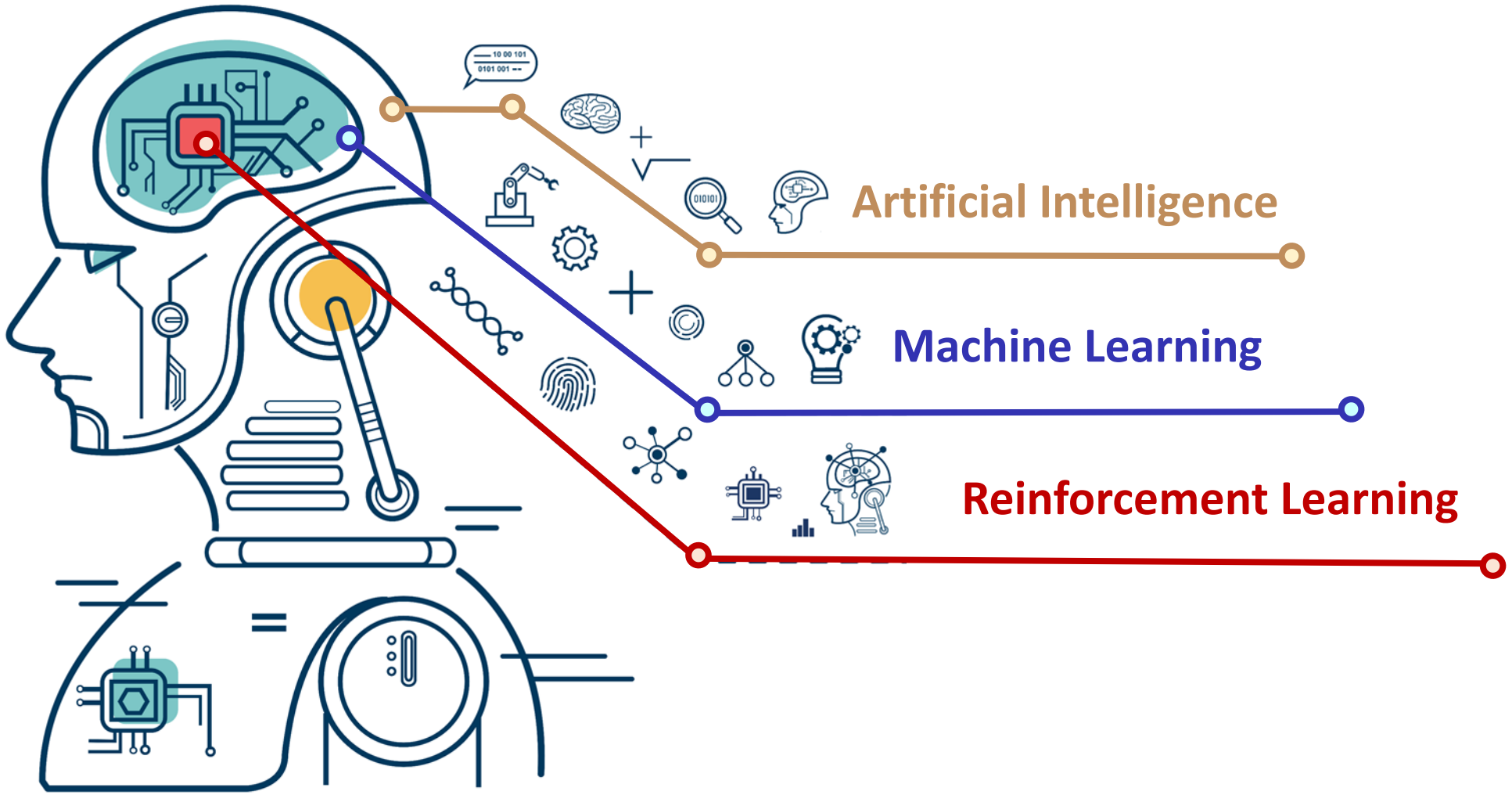
- **Learning is a change in behavior that occurs as a result of experience.**
  - ***Ability to learn*** is the most striking behavioral characteristic of spices (particularly mammals)

» **Intelligence = Learning Capability** «

The Big Picture

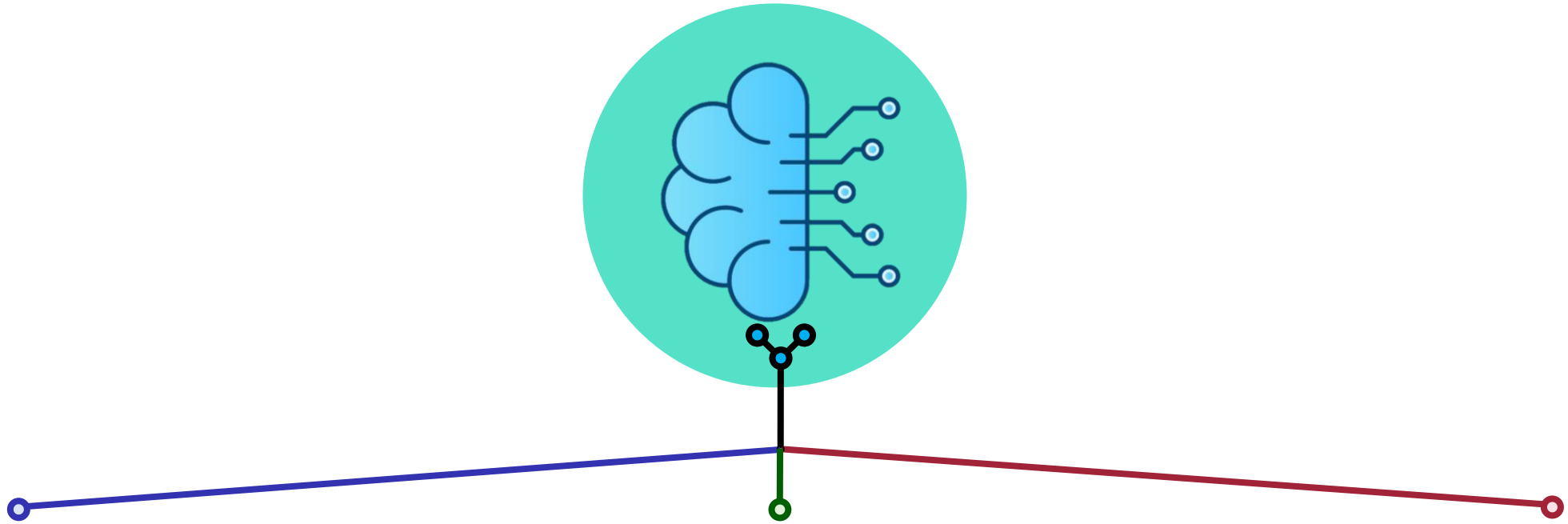


Big Picture


Artificial Intelligence
Machine Learning
Reinforcement Learning

Big Picture

Machine Learning



Supervised Learning

Unsupervised Learning

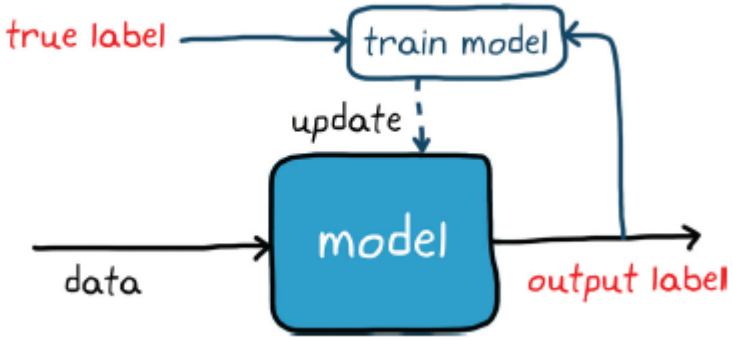
Reinforcement Learning

Machine Learning - Supervised Learning

- Using ***Supervised Learning***, you train the computer to apply a label to a given input.
  - For example, if one of the columns of your dataset of animal features is the species, you can treat species as the label and the rest of the data as inputs into a mathematical model.

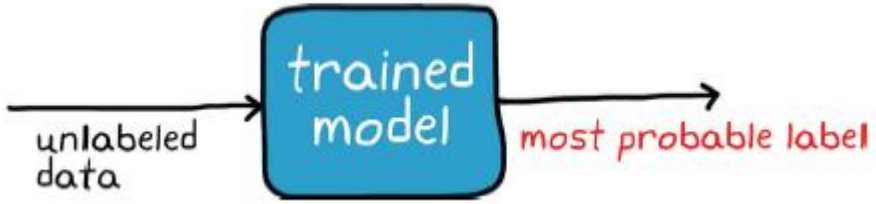
animal dataset (labeled)

species	weight	height	num. of legs	communal living	domesticatable	..
rat	1.3	1.1	4	yes	yes	..
robin	1.2	0.8	4	no	no	..
elephant	48.5	12.2	4	yes	no	..
rabbit	2.5	2.1	4	yes	yes	..
spider	0.1	0.2	8	no	no	..
..	..	..	..	..	..	..



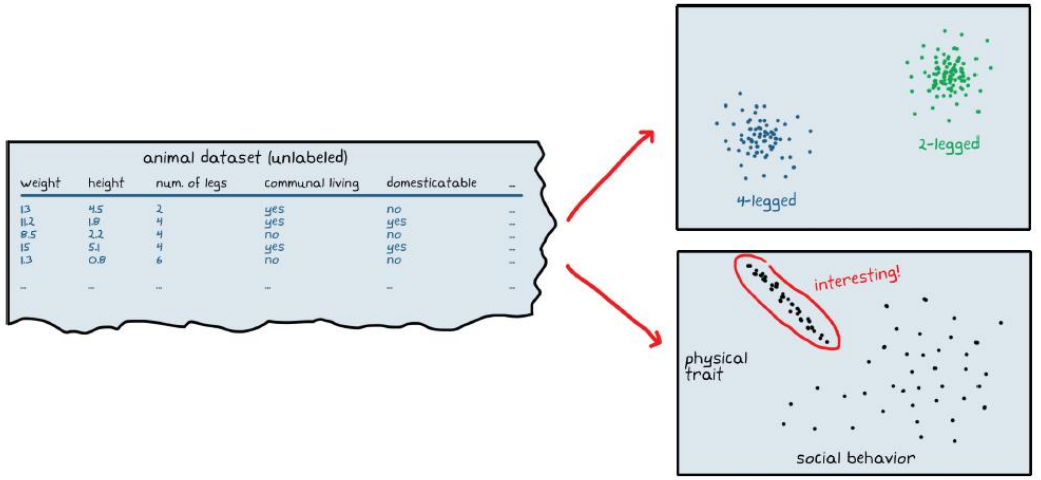
- The model guesses the species, and then the machine learning algorithm systematically tweaks the model

- With enough training data to get a reliable model, you could then input the features for a new, unlabeled animal, and the trained model would apply the most probable species label to it.



Machine Learning - Unsupervised Learning

- Unsupervised learning is used to find patterns or hidden structures in datasets that have not been categorized or labeled.
  - For example, say you have information on the physical attributes and social tendencies of 100,000 animals.
  - You could use unsupervised learning to group the animals or cluster them into similar features.
  - These groups could be based on number of legs, or based on patterns that might not be as obvious, such as correlations between physical traits and social behavior that you didn't know about ahead of time.



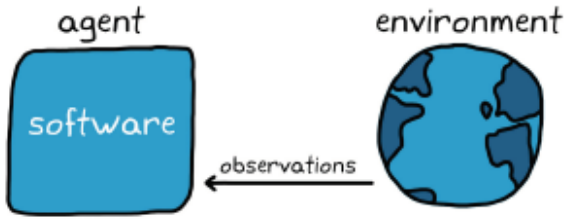
Machine Learning - Reinforcement Learning

- Reinforcement learning is a different beast altogether.
  - Unlike the other two learning frameworks, which operate using a static dataset, RL works with data from a dynamic environment.
  - And the goal is not to cluster data or label data, but to find the best sequence of actions that will generate the optimal outcome.
  - The way reinforcement learning solves this problem is by allowing a piece of software called an ***agent to explore***, interact with, and learn from the environment.

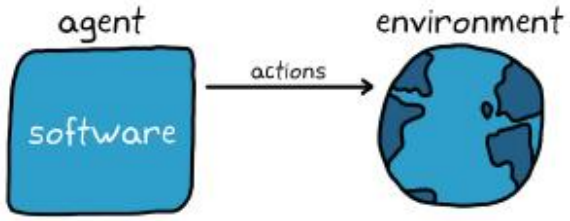
Machine Learning - Reinforcement Learning

- Reinforcement learning is a different beast altogether.

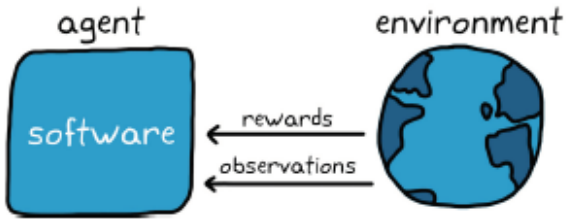
1 The agent is able to observe the current state of the environment.



2 From the observed state, it decides which action to take.

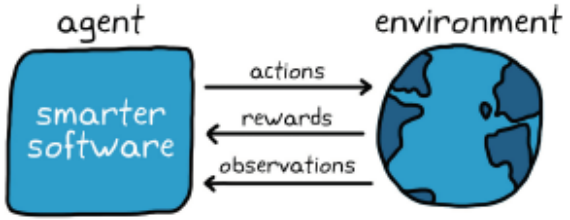


3 The environment changes state and produces a reward for that action. Both of which are received by the agent.



4 Using this new information, the agent can determine whether that action was good and should be repeated, or if it was bad and should be avoided.

The observation-action-reward cycle continues until learning is complete.



Learning Methods

Supervised Learning

- Learn a mapping between inputs and outputs;
- An oracle provides labelled examples of this mapping;

$$y = f(x)$$

Unsupervised Learning

- Learn a structure in a data set (capture the distribution);
- No oracle;

$$f(x)$$

Reinforcement Learning

- Learn to Behave!
- Online Learning.
- Sequential decision making, control.

$$y = f(x)$$

$$z$$

Yann LeCun

- Chief AI Scientist, Facebook

NIPS 2016 Keynote:

"If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning."

Pieter Abbeel

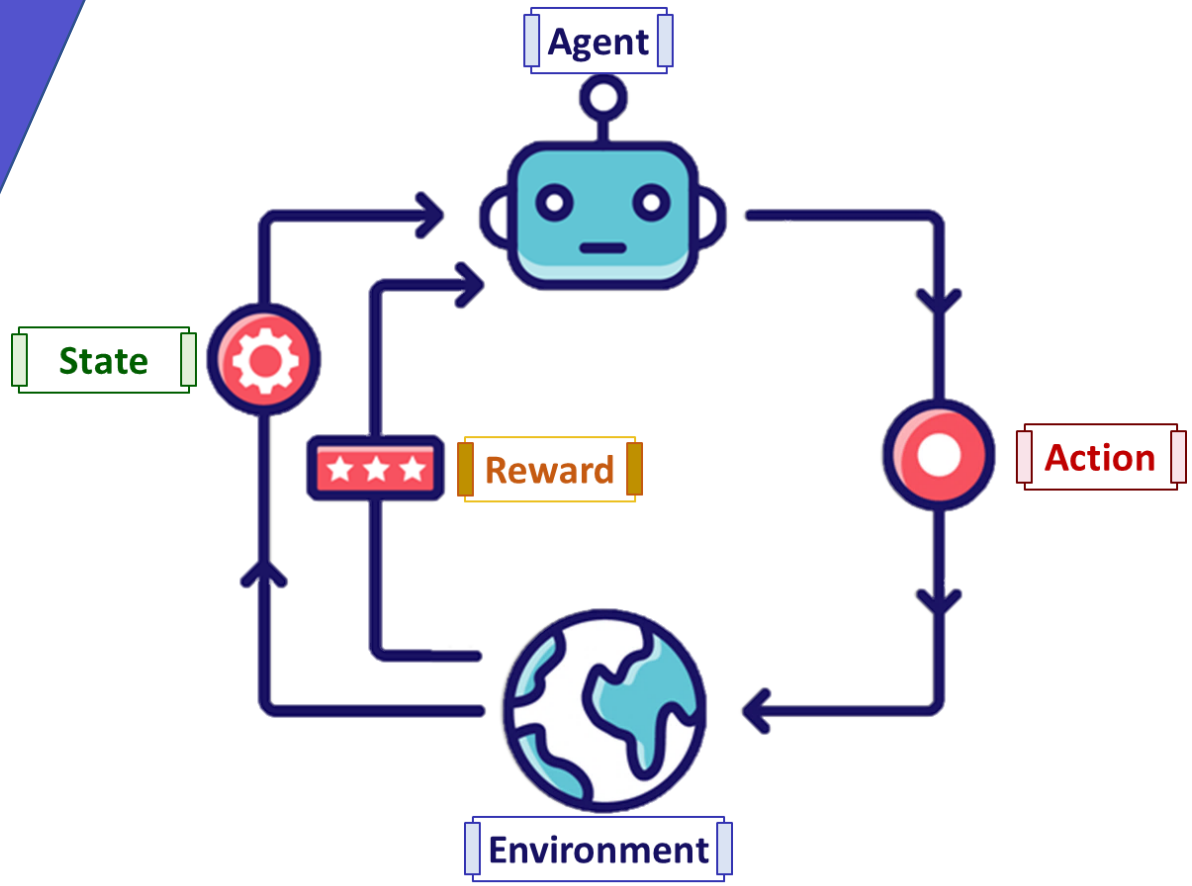
- Prof. At UC Berkeley

NIPS 2017 Keynote:

"I prefer to eat a cake with a lot of cherries because I like reinforcement learning."

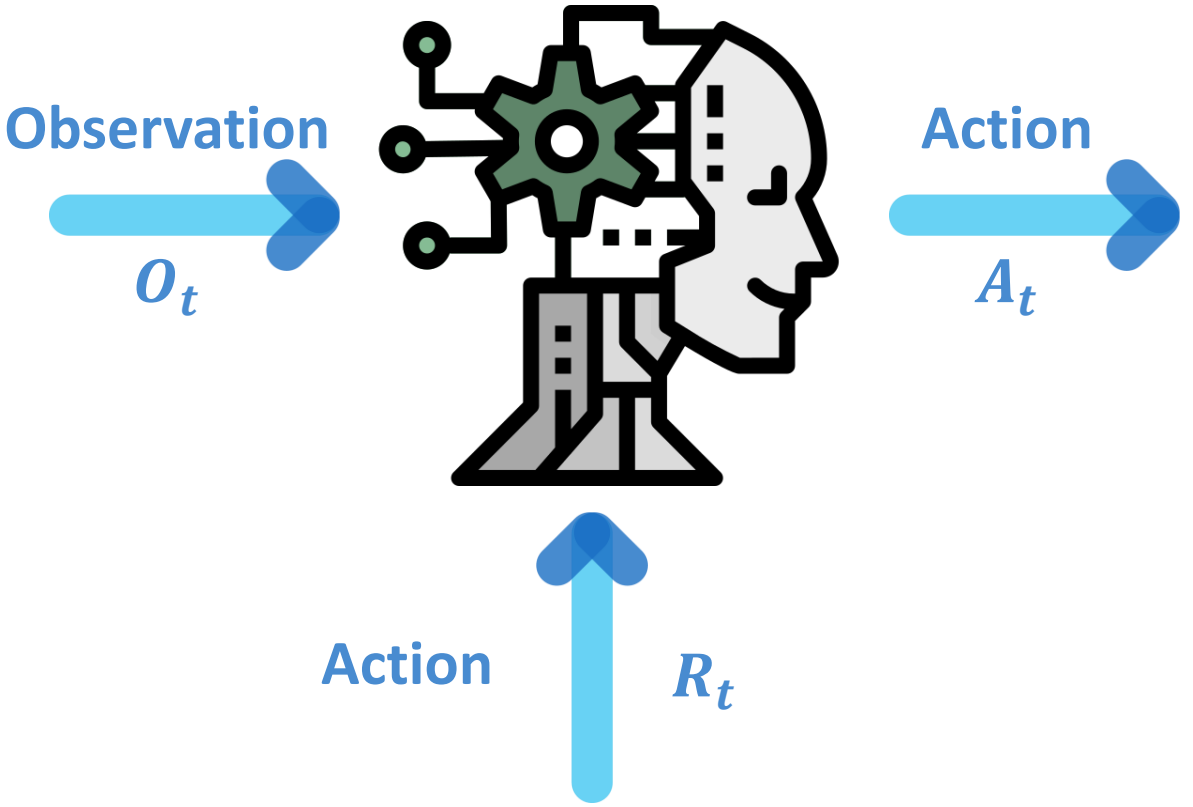
- Although the classic reinforcement learning method regards failure as zero reward, Abbeel proposed that even failures can be used to train machines.
- Intelligence is a cake with many cherries

RL Intro

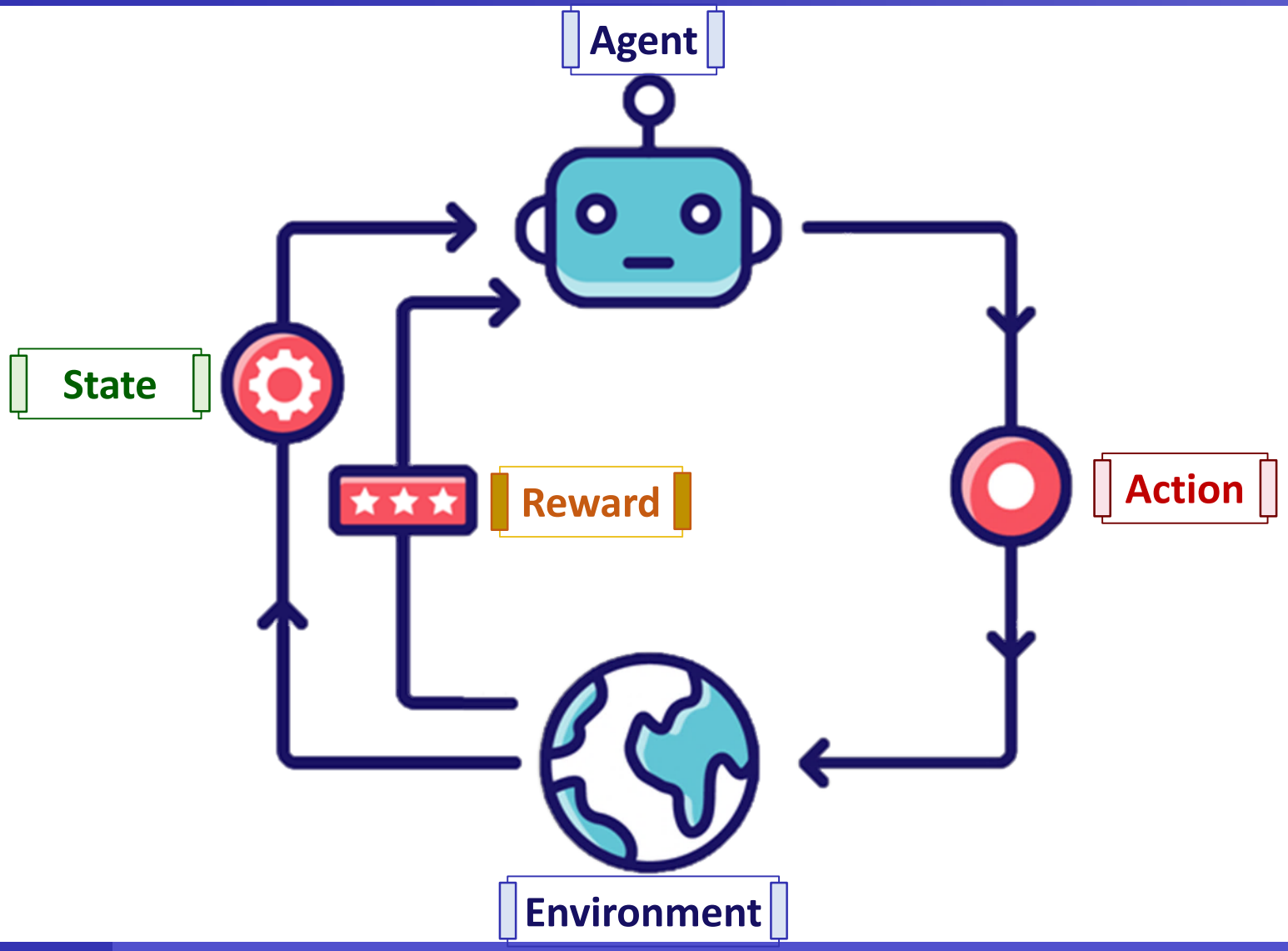


Reinforcement Learning

- An agent interacts with its environment
- Agent takes actions that affect the state of the environment
- Feedback is limited to a reward signal that indicates how well the agent is performing
- Goal: improve the behavior given only this limited feedback



Reinforcement Learning (RL)



- At each time t, the agent:
  - Executes and action A_t
  - Receives observation S_t
  - Receives scalar reward R_t
- The environment
  - Changes upon action A_t
  - Emits observation S_{t+1}
  - Emits scalar reward R_{t+1}

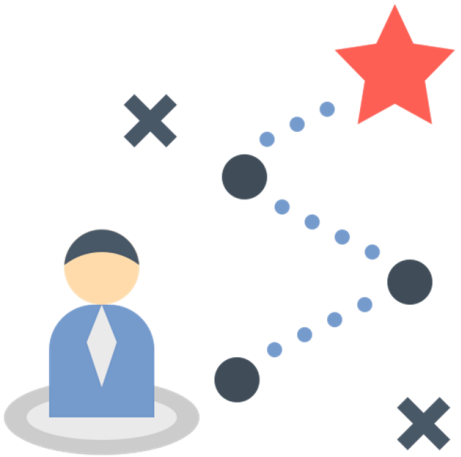
Time step t is incremented after each iteration

RL Characteristics

Main Characteristics of *Reinforcement Learning*

- Reinforcement learning is learning what to do—how to map *situations* to *actions*
- Goal is to maximize a numerical reward signal.
  - There is no supervisor, only a *reward signal*
- The learner is not told which actions to take, but instead must *discover* which actions yield the most reward by trying them.
- Feedback is *delayed*, not *instantaneous*
  - actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards.

RL Characteristics



trial-and-error

Delayed Reward

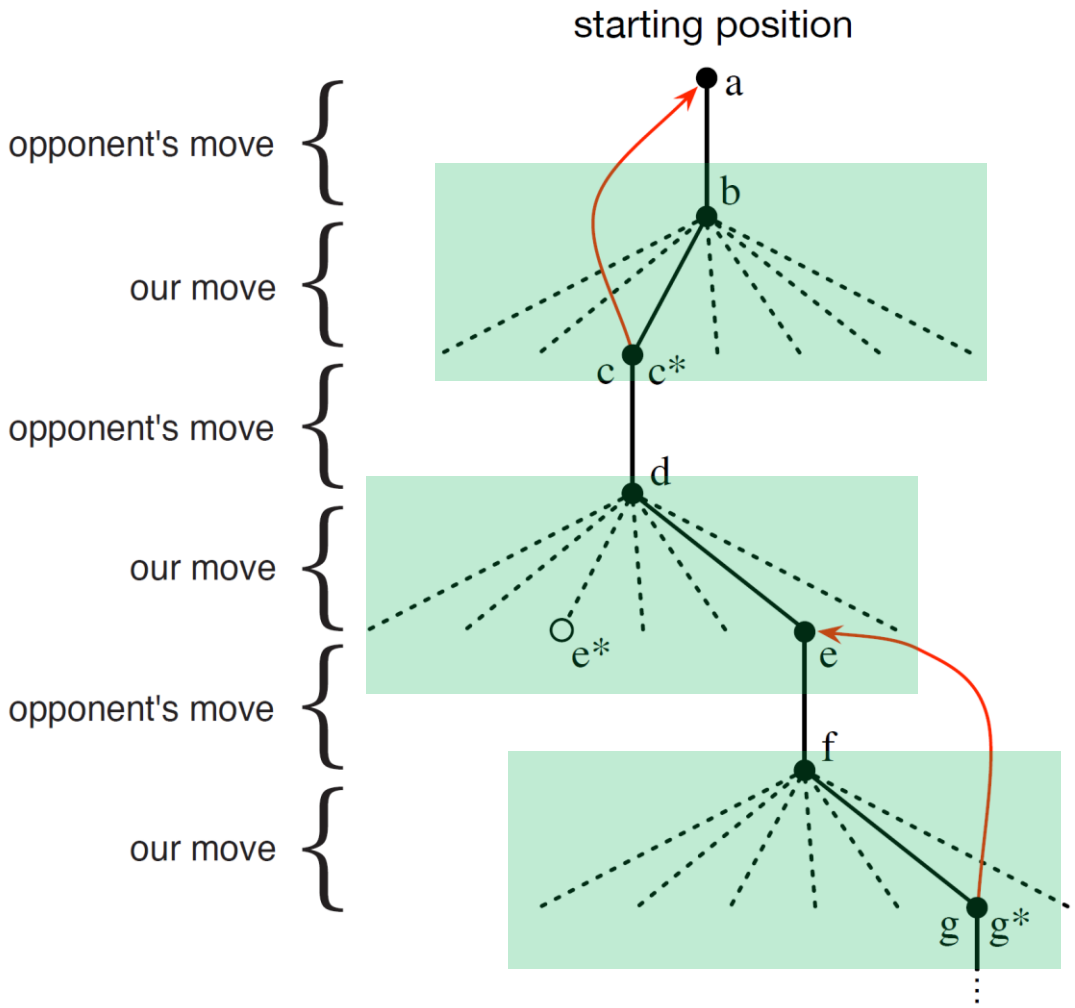
RL Distinctions

- Time really matters (sequential, non i.i.d data)
- Uncertain environment
  - RL explicitly considers the ***whole problem*** of a goal-directed agent interacting with an uncertain environment
- Agent's action affect the subsequent data it receives
- Capabilities to be applied in larger behaving system
  - an agent that monitors the charge level of robot's battery and sends commands to the robot's control architecture
- Curse of dimensionality

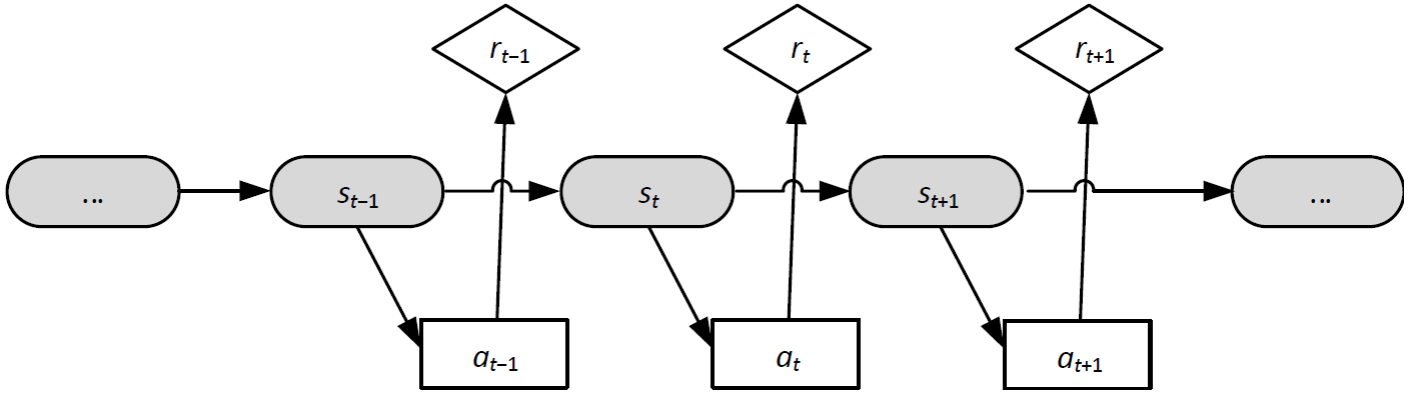
RL Example (Tic-Tac-Toe)

X	O	O
O	X	X
		X

(Tic-Tac-Toe)

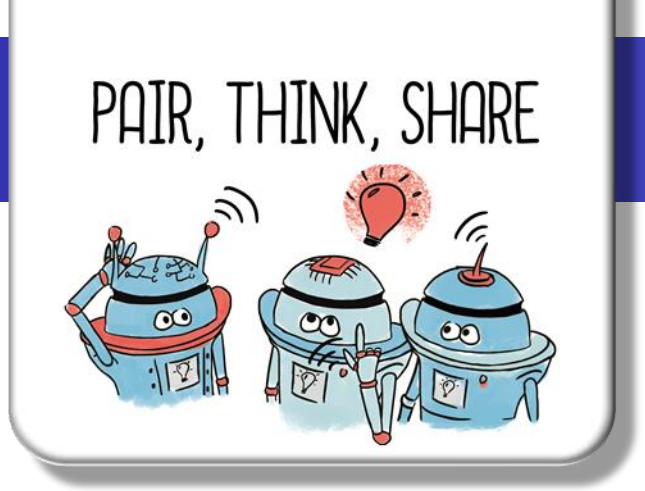


History and State



- The history is the sequence of observations, actions, and rewards

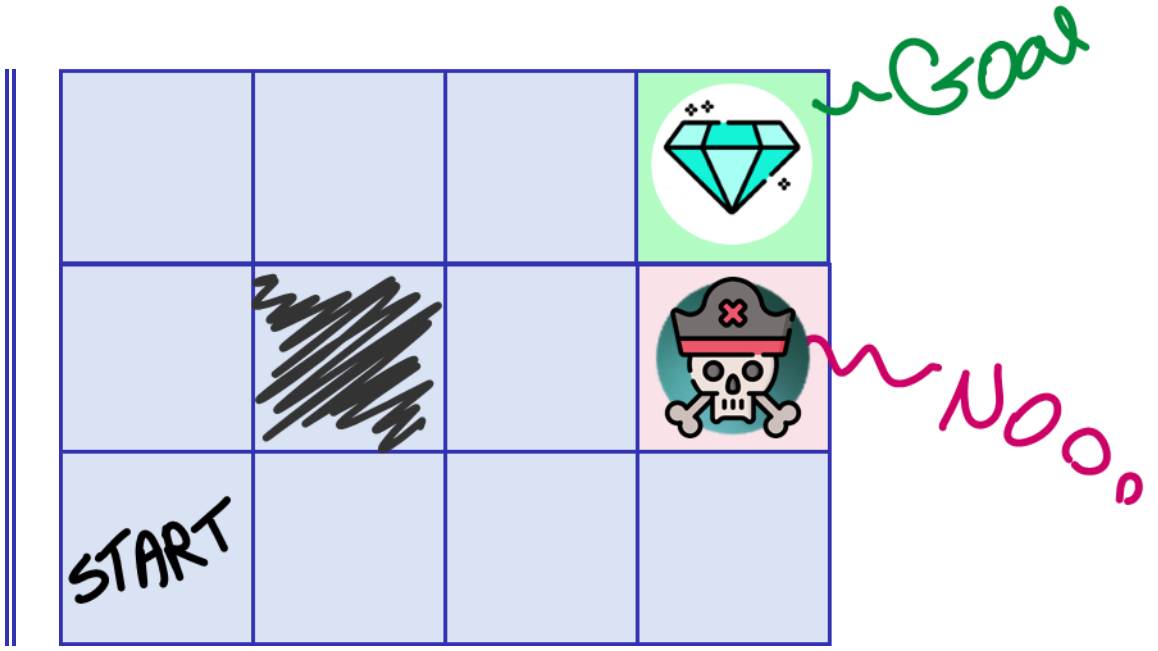
$$H_t = A_1, O_1, R_1, \dots, A_t, O_t, R_t$$

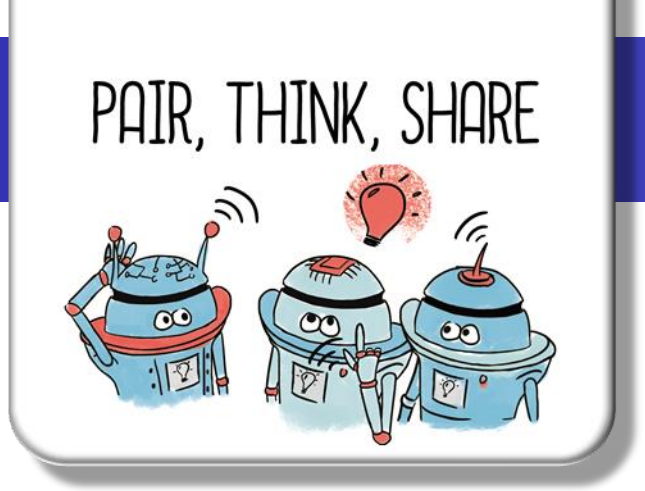


The World – Part (a)

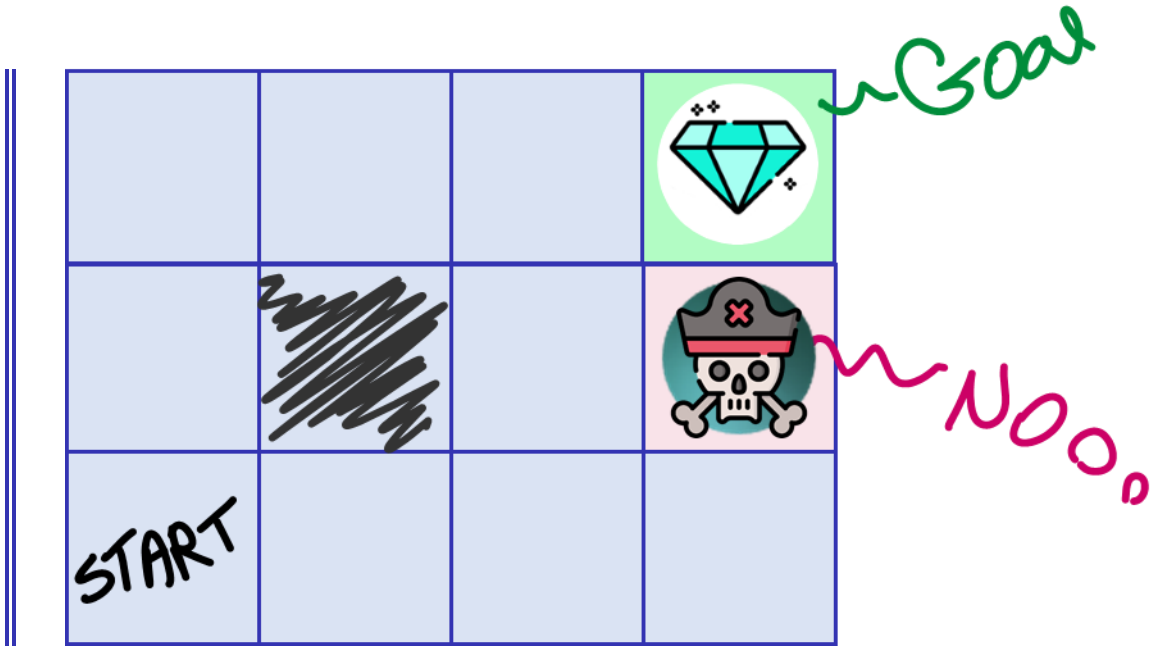
What is the shortest sequence getting from Start to Goal?

Possible Actions: **Up, Down, Left, Right**



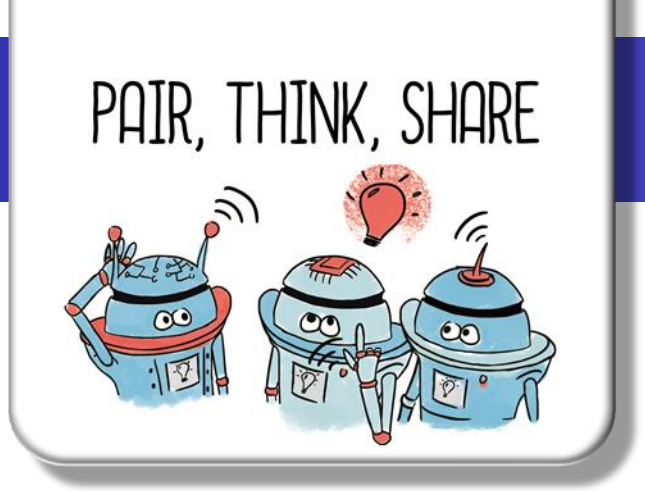


The World – Part (b)



- Possible Actions: **Up, Down, Left, Right**
  - Action executes correctly: 0.8
  - Moves at Right Angle: 0.2
    - Right 0.1
    - Left 0.1
- What is the reliability of your sequence:

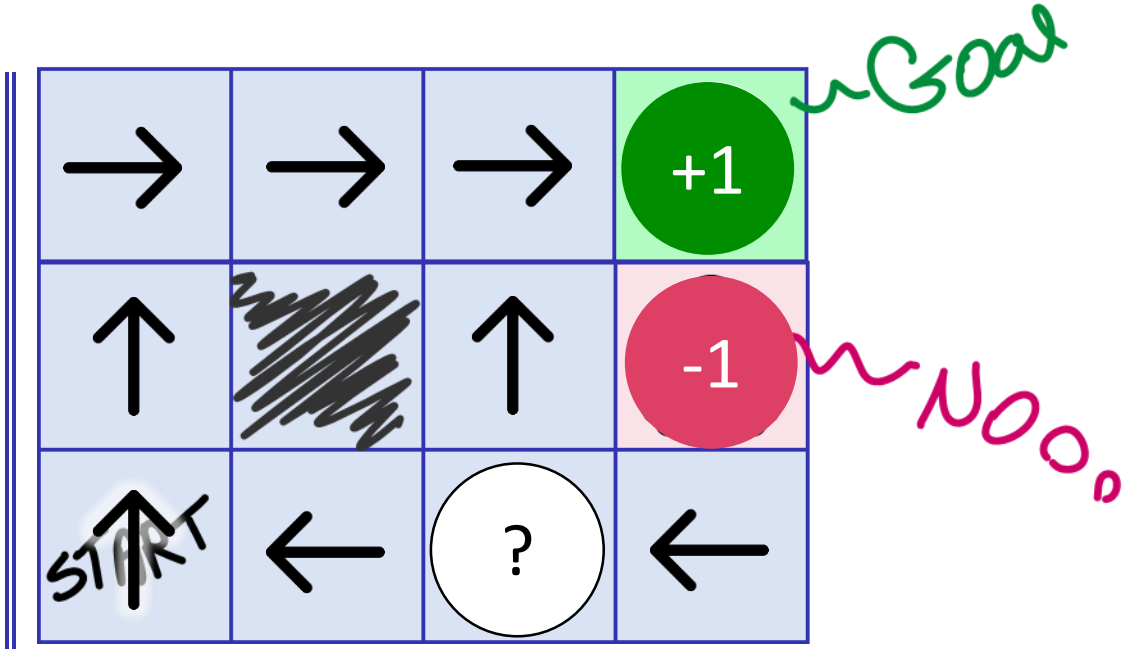
  Up, Up , Right, Right, Right

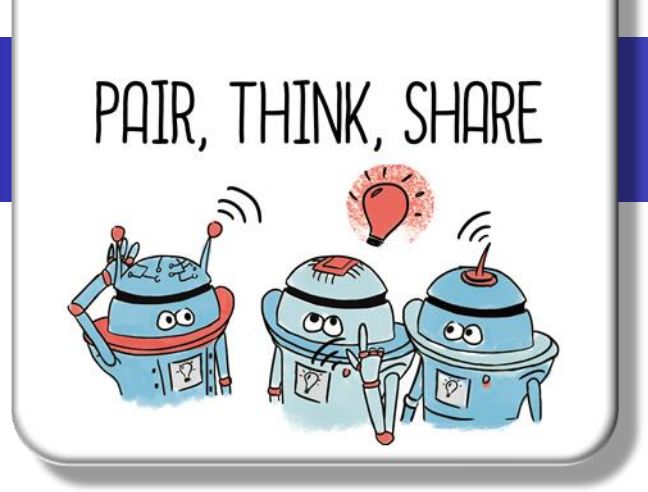


The World – Part (c)

- Assume, for each state, $R(s) = -0.04$, except the termination states
- What is the best action to take in the highlighted state?

  Possible Actions: Up, Down, Left, Right
  - Action executes correctly: 0.8
  - Moves at Right Angle: 0.2

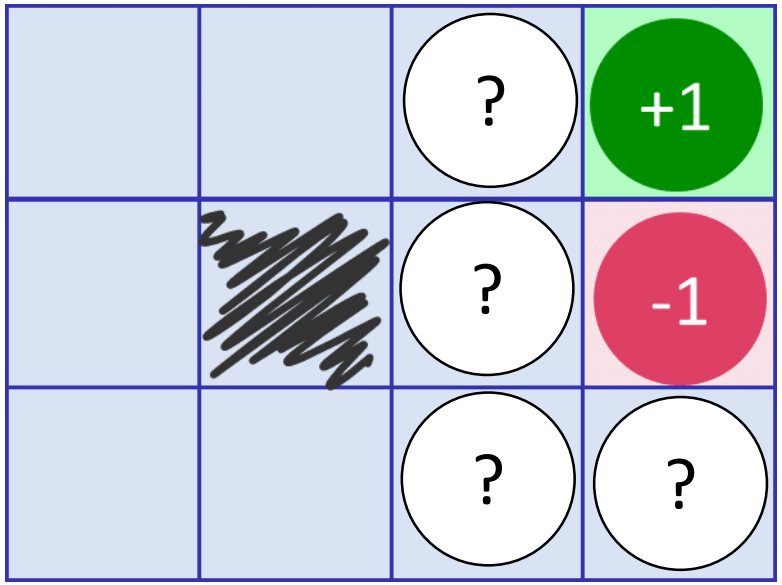




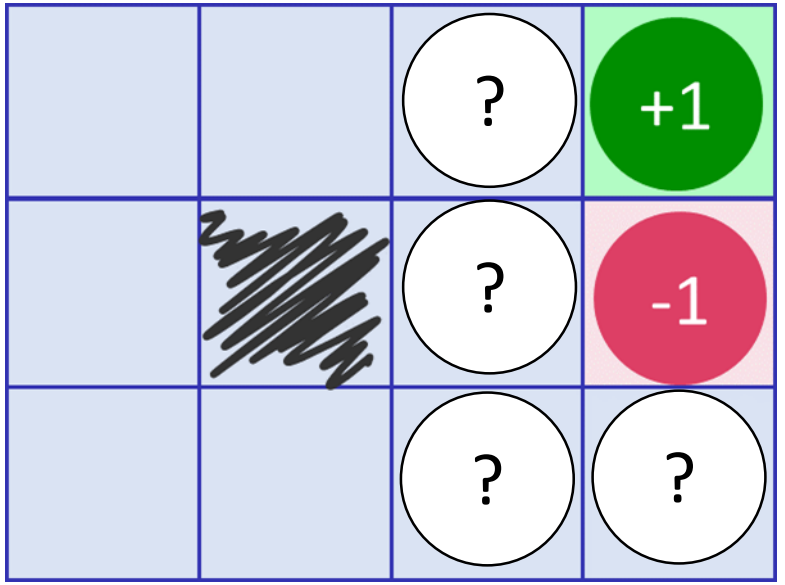
The World – Part (d)

What is the best action to take in the highlighted state?

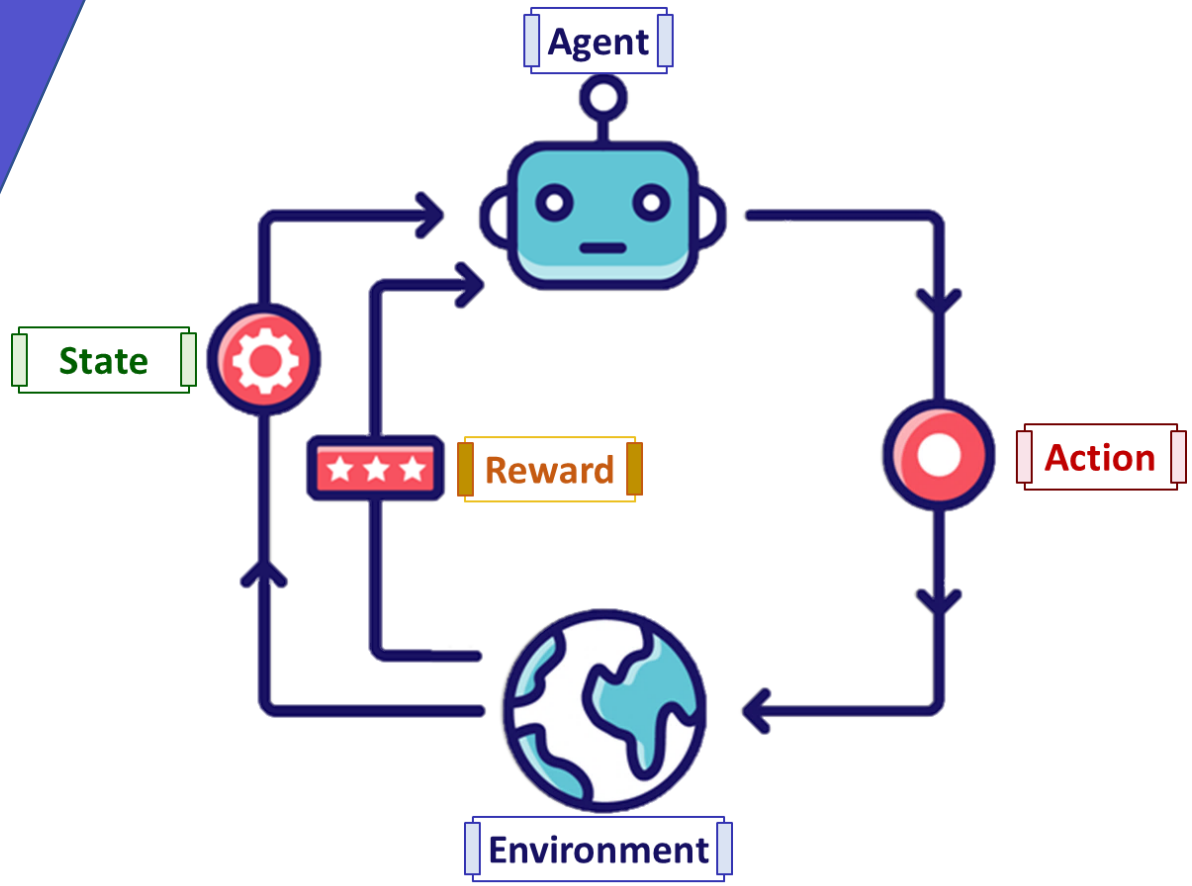
$R(s) = +2$



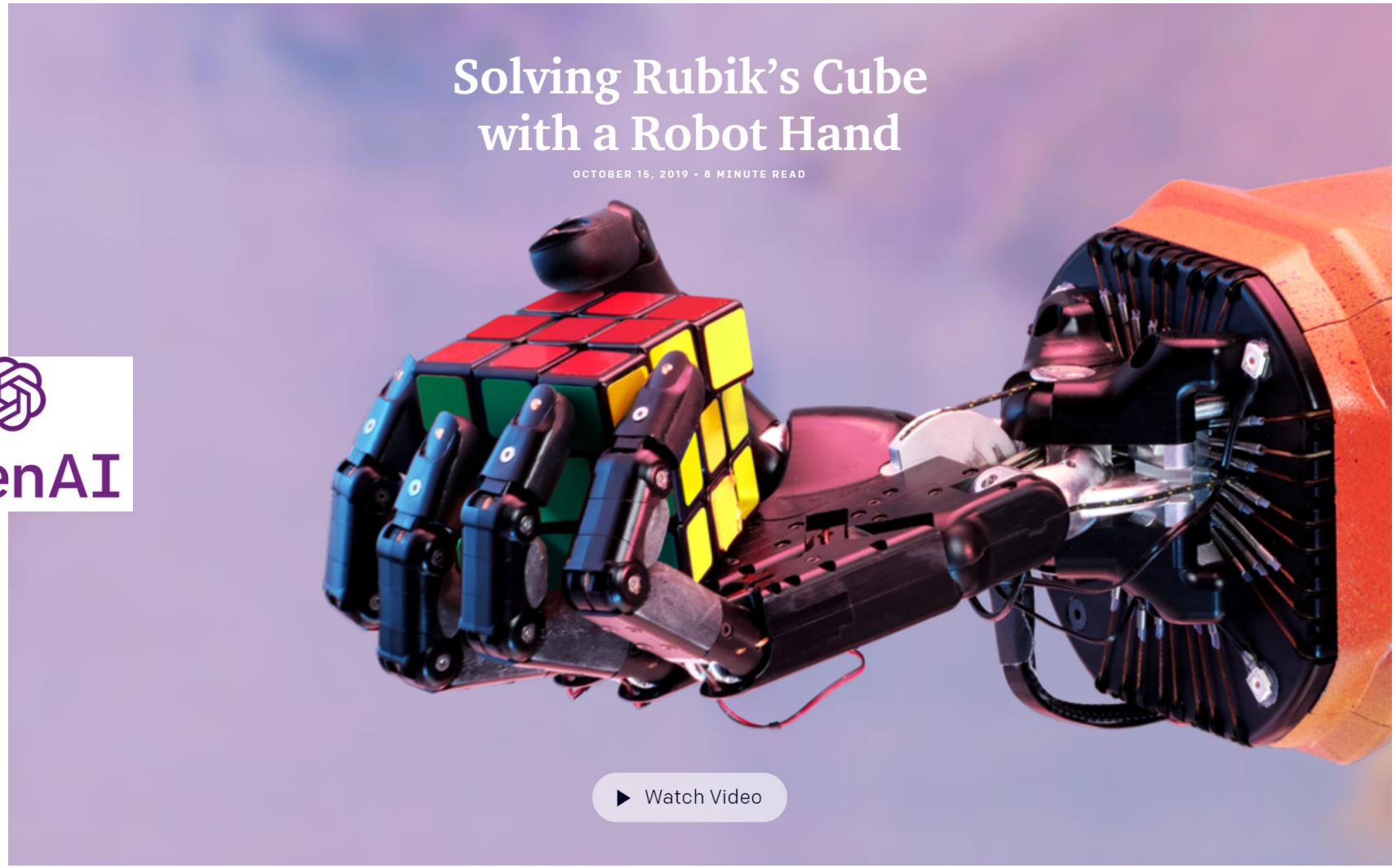
$R(s) = -2$



RL Examples

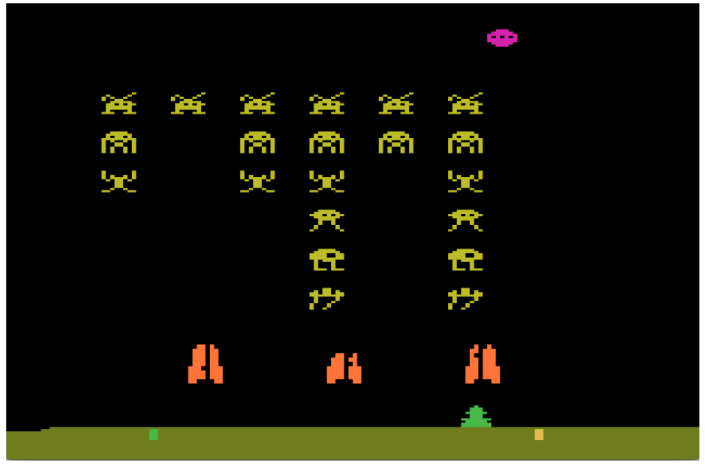


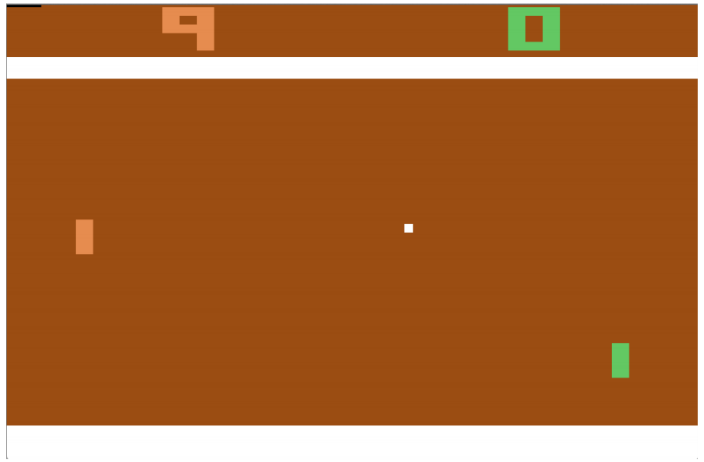
RL Examples (Rubik's Cube by OpenAI)



OpenAI

RL Examples (Atari Games – Google DeepMind)





Google DeepMind

RL Example

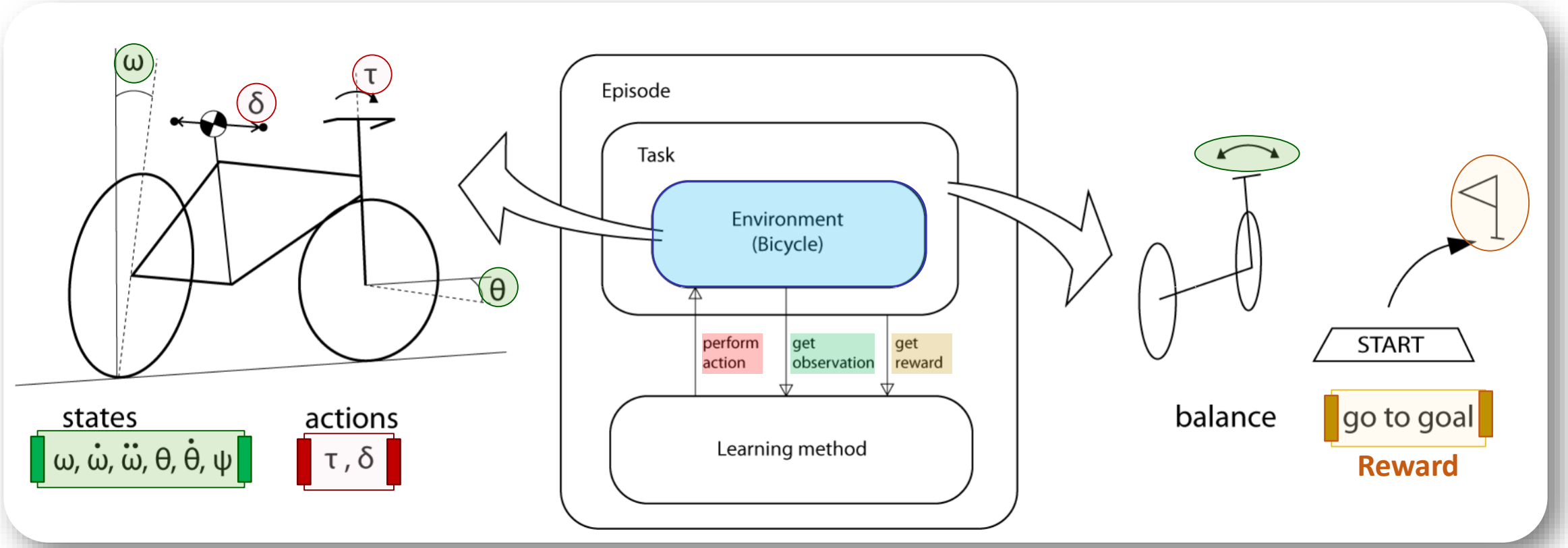
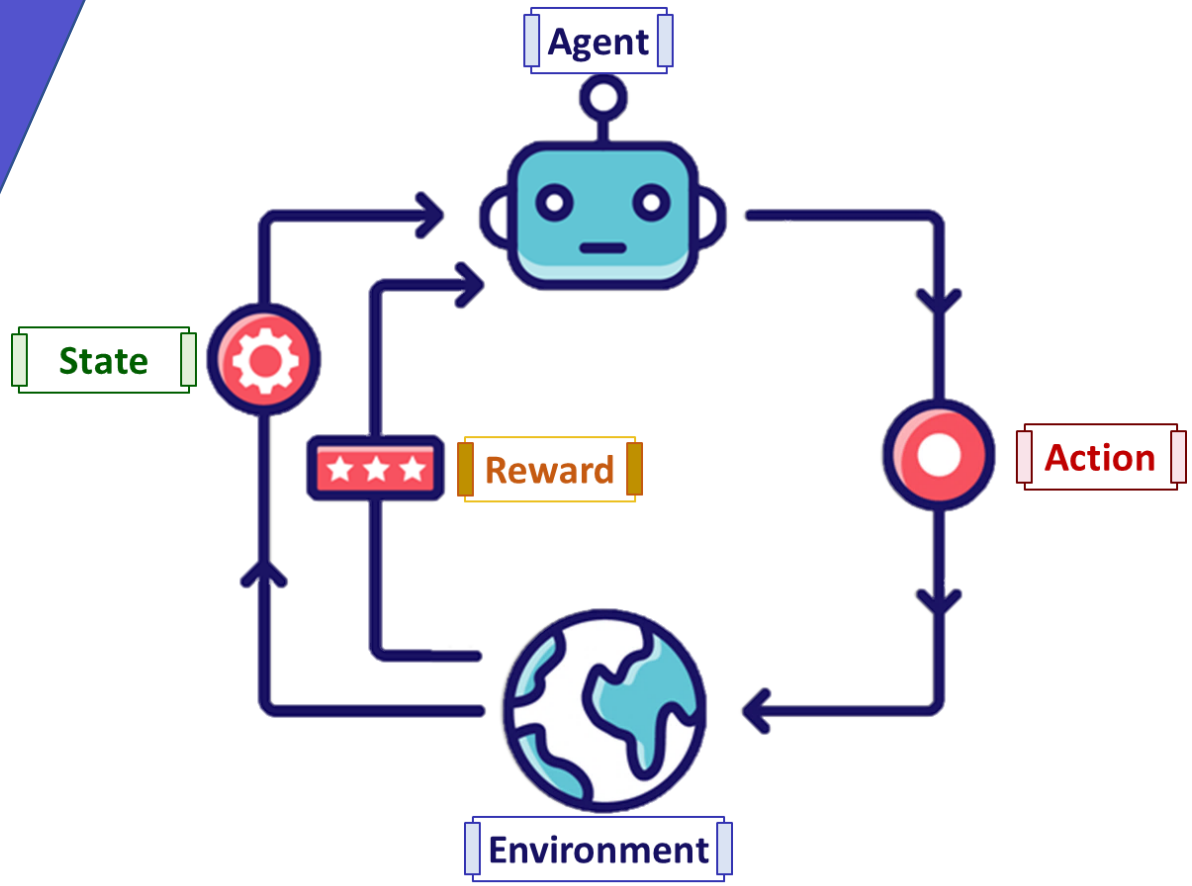


Image: borrowed from *B. Cam, C. Dembia, and J. Israeli, "Reinforcement learning for bicycle control," 2013*

RL Elements



Reinforcement Learning Elements

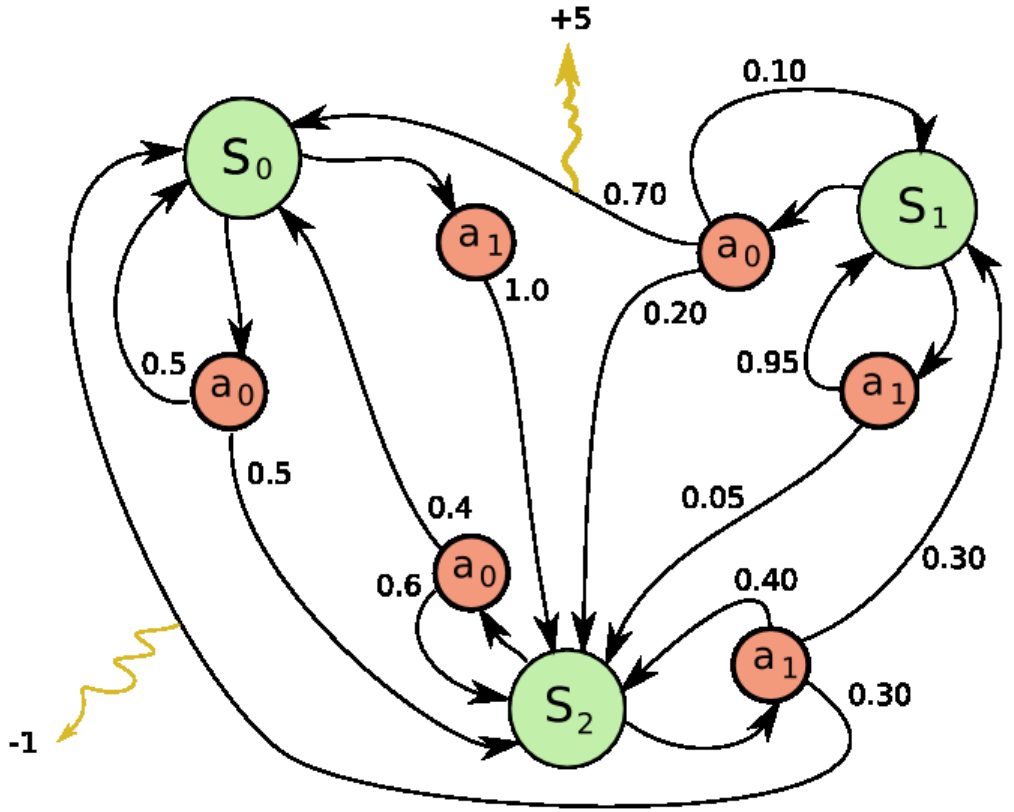
- **Environment:** Physical world in which the agent operates
- **State:** Current situation of the agent
- **Reward:** Feedback from the environment
- **Policy:** Method to map agent's state to actions
- **Value**: Future reward that an agent would receive by taking an action in a particular state

Markov Decision Process

- A Markov decision process (MDP) specifies a setup for reinforcement learning
- MDPs allow to model decision making in situations where outcomes are partly random and partly under the control of a decision maker
- **Definition**
  - A Markov Decision Process is a 4-tuple (S;A;R;T) with
    - A set of possible world states S
    - A set of possible actions A
    - A real-valued reward function R
    - Transition probabilities T
  - A MDP must fulfill the so-called Markov property
    - The effects of an action taken in a state depend only on that state and NOT on the prior history

Markov Decision Process

- **State:**
  - A state S_t is a representation of the environment at time step t
  - Can be directly observable to the agent or hidden
- **Actions**
  - At each state, the agent is able to perform a action at that affects the subsequent state the environment st+1
  - Actions can be any decisions which one wants to learn

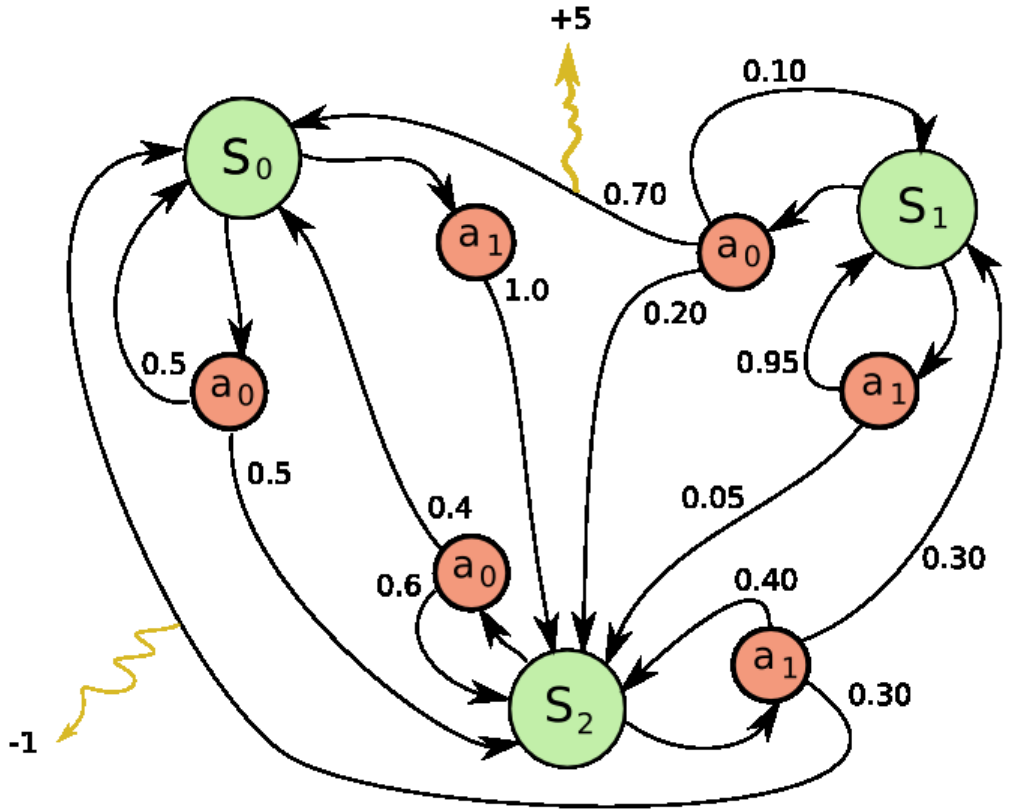


Markov Decision Process

- Transition probabilities
  - Given a current state s, a possible subsequent state s' and an action a
  - The transition probability is defined as:

$$T_{ss'}^{a} = P[s_t = s, a_t = a]$$

$$T_{ss'}^{a} \sim \quad Pr(s'|s, a)$$



Markov Decision Process

$\mathbf{R_{t+1}}$: ***reward signal*** is a scalar feedback signal emitted by the environment

- Indicates how well agent is performing when reaching step $t + 1$
- The agent's sole objective is to maximize the ***total reward*** it receives over the ***long run***.

Definition: Reward Hypothesis

All goals can be described by the maximization of the expected cumulative reward.

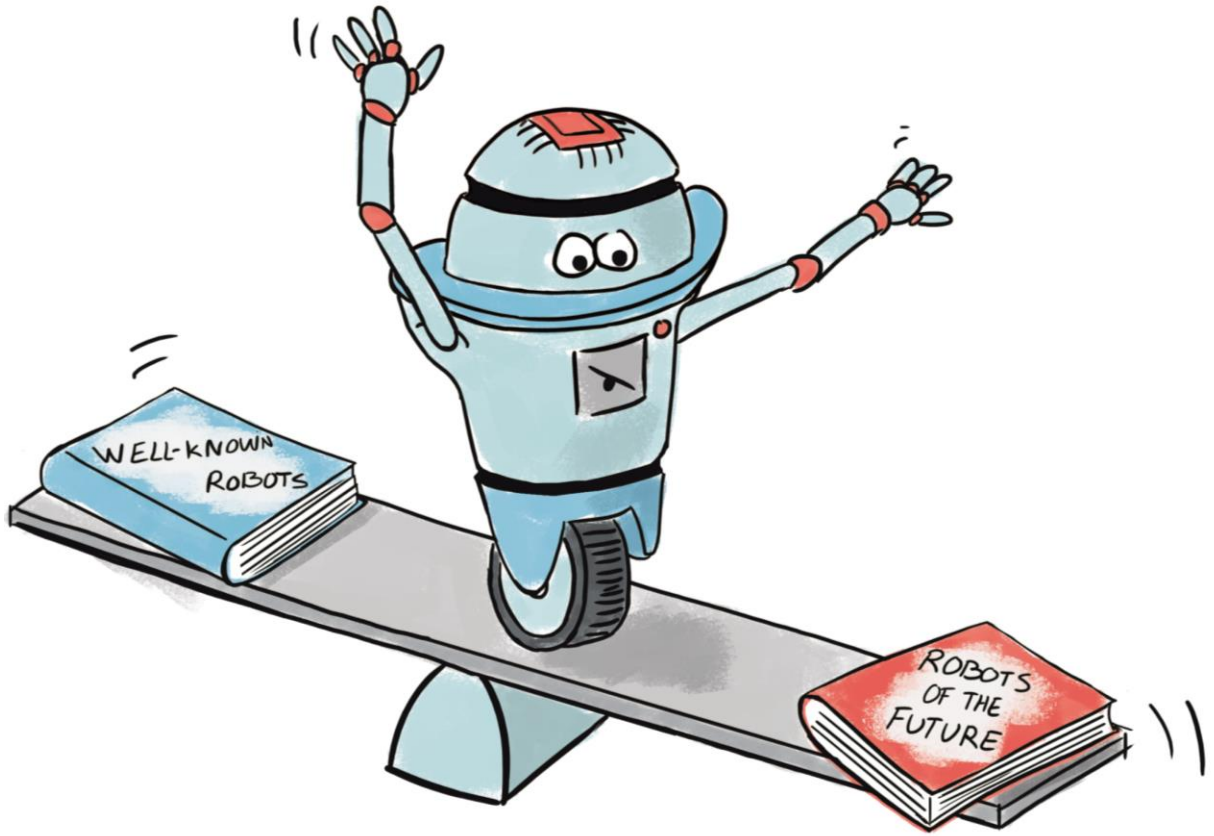
Examples of Reward

- **Bicycle:**
  - The distance that the bike travels
- **Atari Games**
  - Achieving more points
- **Chess**
  - Winning the game

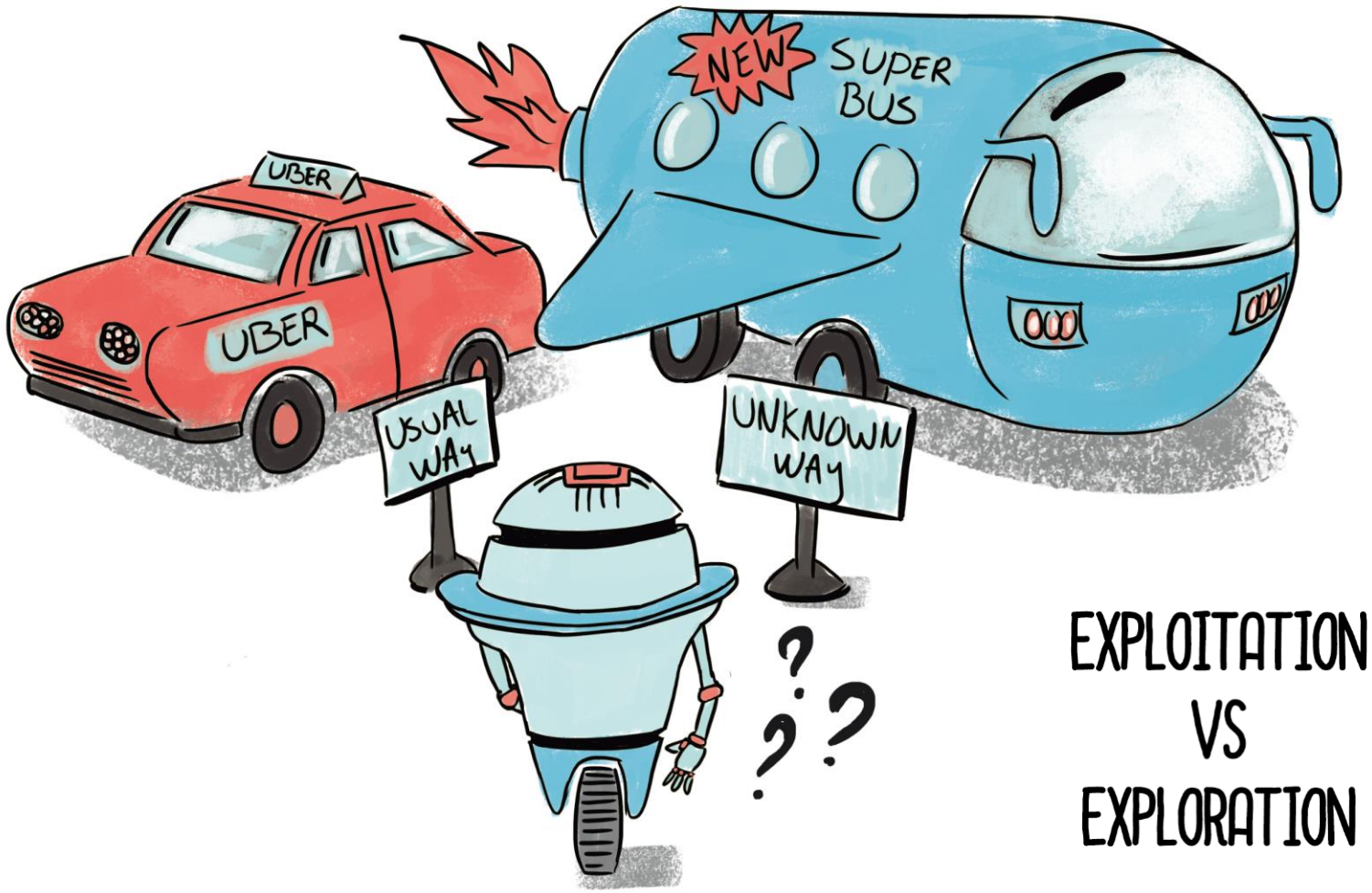
Markov Decision Process (MDP)

- **State:** S
- **Model**: $T(s, a, s') \sim Pr(s'|s, a)$
- ***Actions:*** *A(s), A*
- ***Rewards:*** $R(S), R(s, a), R(s, a, s')$
- ***Policy:*** $\pi(s) \rightarrow a, \pi^*$

Exploration vs. Exploitation

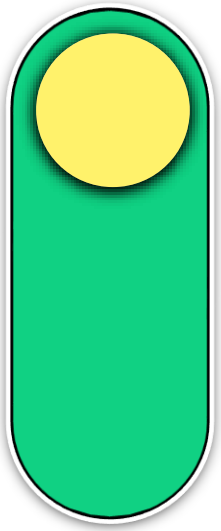


Exploration vs. Exploitation

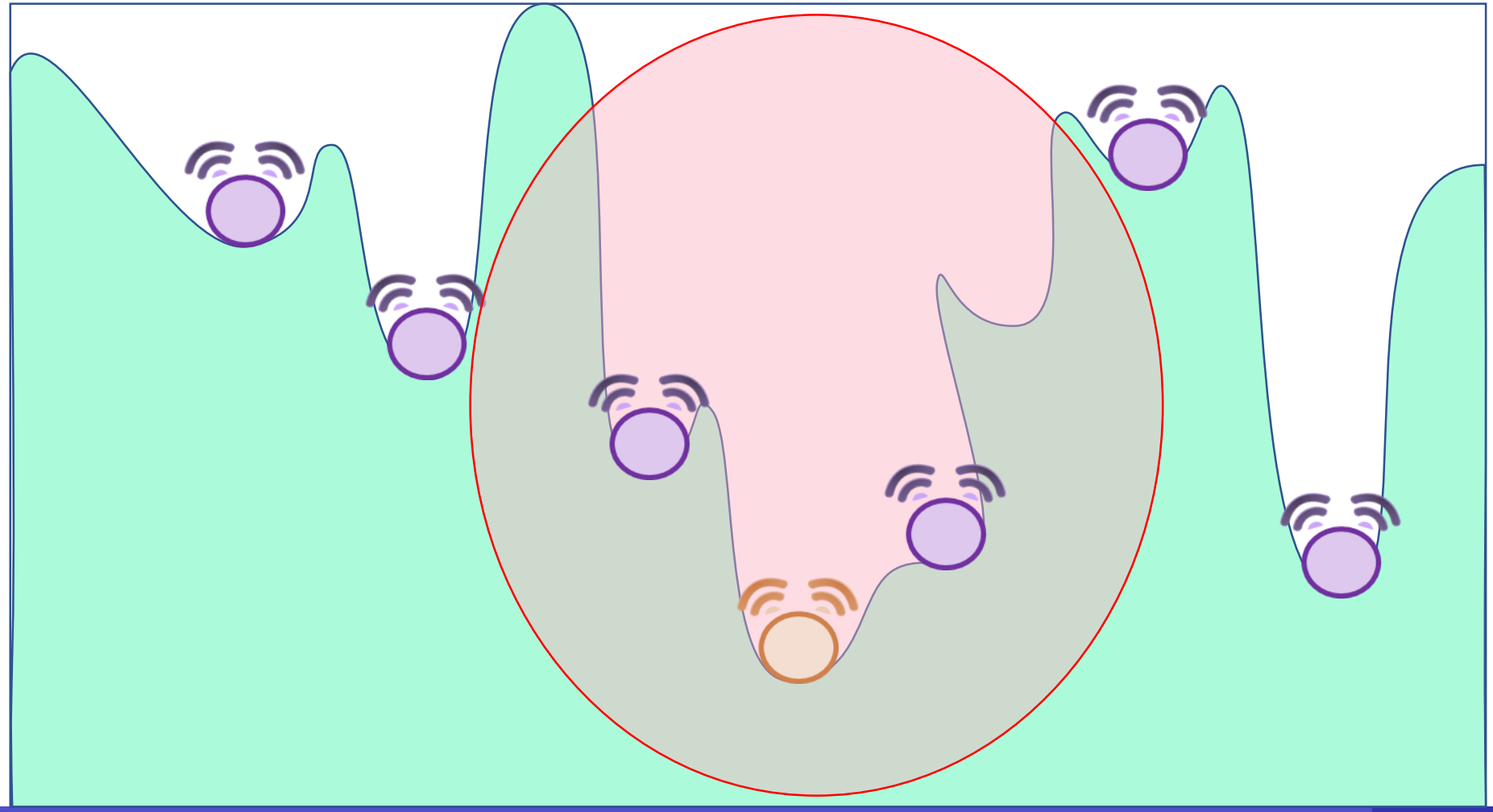


Exploration vs. Exploitation

Exploration



Exploitation



Exploration vs. Exploitation

Exploration

Exploitation

Diversify the Solution

Finding Unexplored Solutions

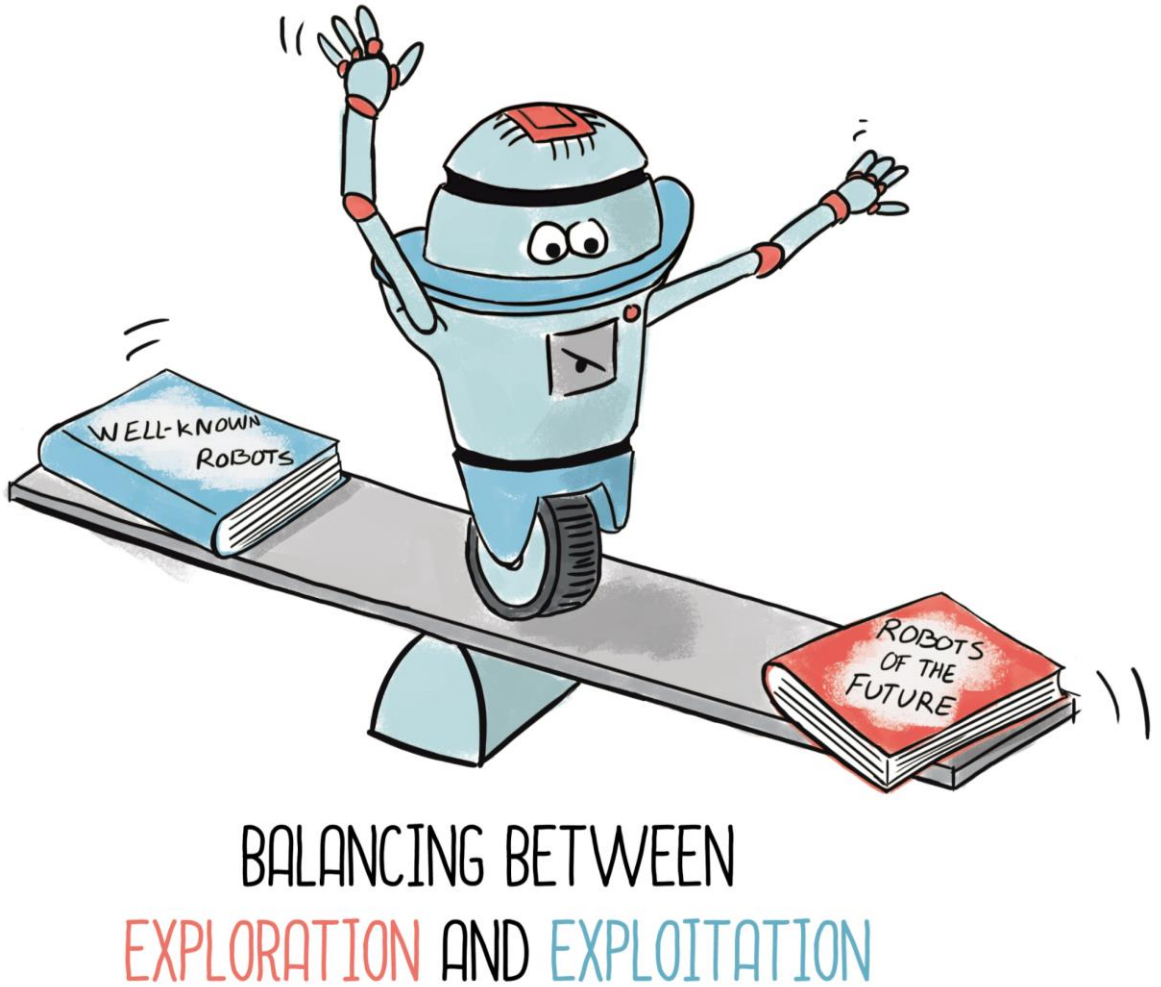
Random Search

Short-term Goals

Local Optimum Trap

Greedy Search

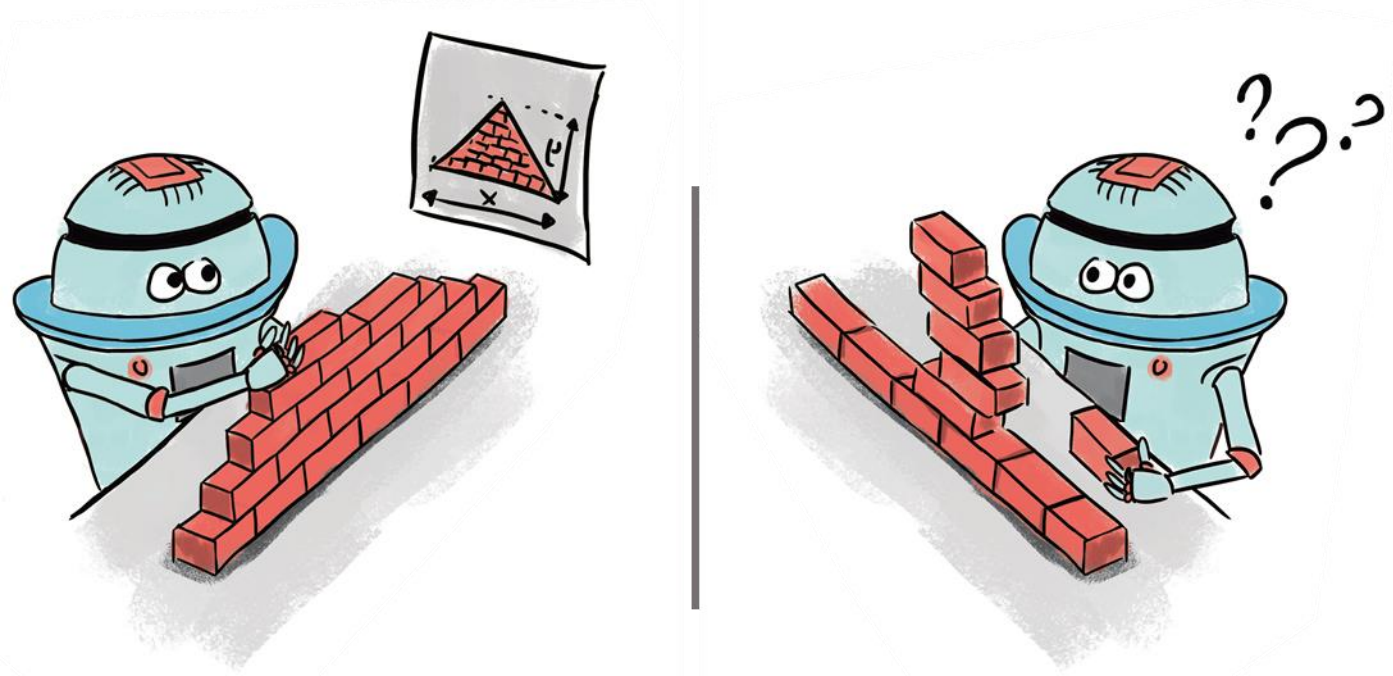
Balance between Exploration and Exploitation



Exploration vs. Exploitation Examples

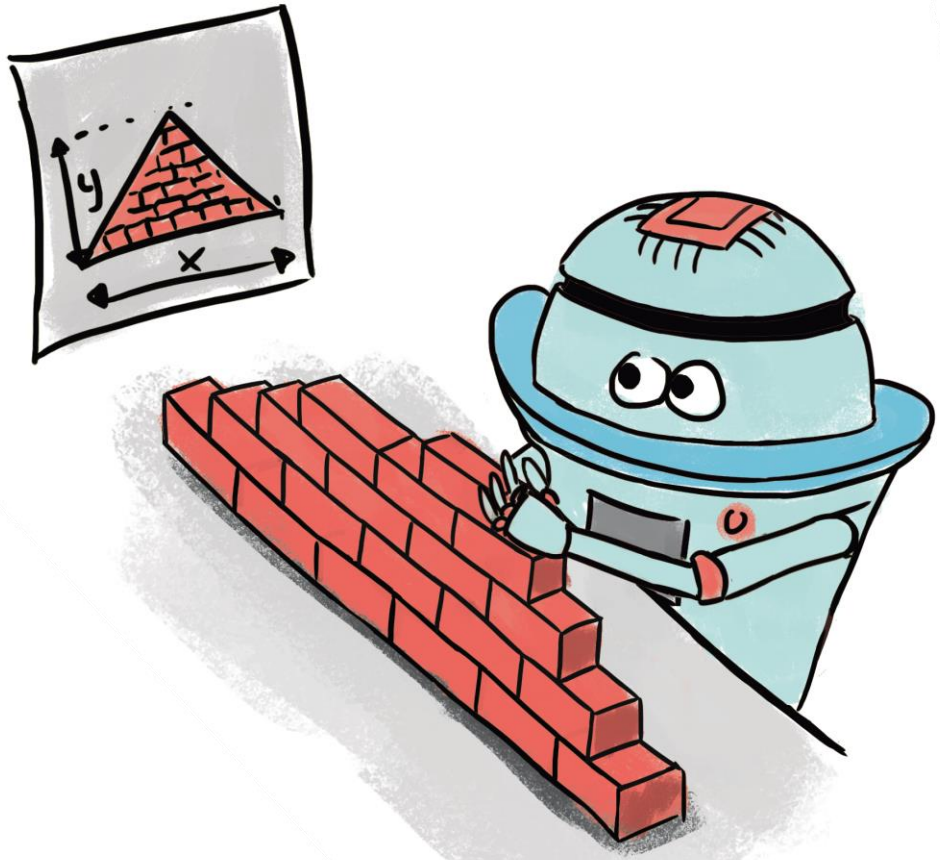
- **Overcome afternoon fatigue**
  - Getting a coffee
  - Taking a nap, trying
  - Power through
- **Dating**
  - How long do you date around?
  - You can't possibly date everyone, so what confidence level do you need that you've found the right person?
- **Our life journeys**
  - We explore while we're young and exploit as we get older
  - In our youth, we develop many friends, we try new experiences, we travel.
  - When we're old, we find the things we truly care about and dedicate our lives to them
    - our marriages, our families, a tightly knit group of friends.

Model-based vs. Model-free



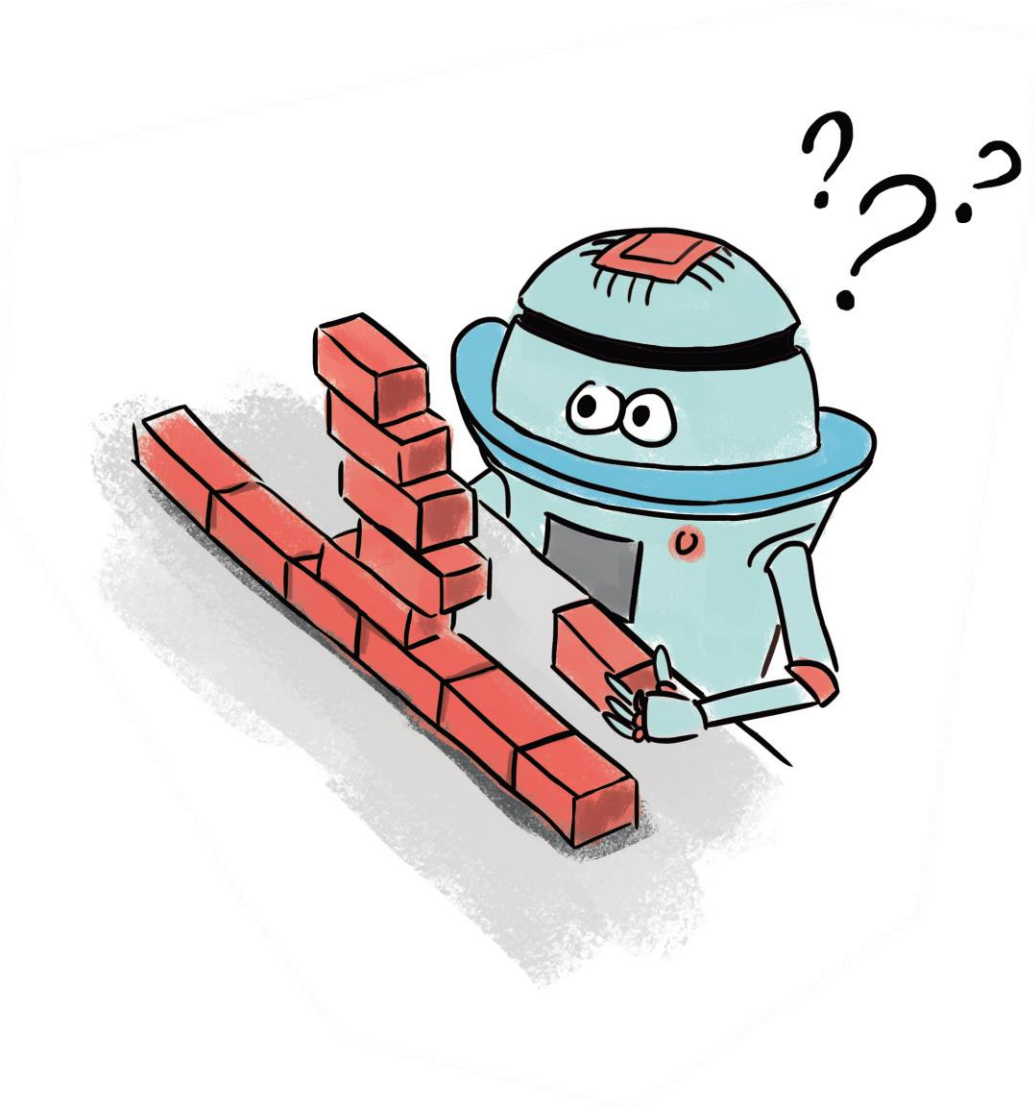
Model-based vs Model-free RL

- **Model-based learning**
  - **Aim:** find optimal policy and value functions
  - Model of the environment is as MDP with transition probabilities
  - **Approach:** learn the MDP model or an approximation of it



Model-based vs Model-free RL

- **Model-free learning**
  - Explicit model of the environment model is not available
    - i.e. transition probabilities are unknown
  - **Approach:** derive the optimal policy without explicitly formalizing the model



References

- [1] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
- [2] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, 2016.
- [3] R. Bradley, “16 Examples of Artificial Intelligence (AI) in Your Everyday Life | The Manifest.” [Online]. Available: https://themanifest.com/development/16-examples-artificial-intelligence-ai-your-everyday-life. [Accessed: 28-Dec-2019].
- [4] C. Osborne, “Uber uses artificial intelligence to figure out your personal price hike,” ZDNet. [Online]. Available: https://www.zdnet.com/article/uber-uses-artificial-intelligence-to-figure-out-your-personal-price-hike/. [Accessed: 28-Dec-2019].
- [5] S. Kennedy, “Artificial Intelligence and Machine Learning: What Are They and Why Are They Important? | MapR.” [Online]. Available: https://mapr.com/blog/artificial-intelligence-and-machine-learning-what-are-they-and-why-are-they-important/. [Accessed: 28-Dec-2019].
- [6] A. M. Zador, “A critique of pure learning and what artificial neural networks can learn from animal brains,” Nature communications, vol. 10, no. 1, pp. 1–7, 2019.
- [7] T. Peng, “Yann LeCun Cake Analogy 2.0,” Medium, 22-Feb-2019. [Online]. Available: https://medium.com/syncedreview/yann-lecun-cake-analogy-2-0-a361da560dae. [Accessed: 29-Dec-2019].
- [8] J. Randløv and P. Alstrøm, “Learning to Drive a Bicycle Using Reinforcement Learning and Shaping.,” in ICML, 1998, vol. 98, pp. 463–471.
- [9] B. Cam, C. Dembia, and J. Israeli, “Reinforcement learning for bicycle control,” 2013.
- [10] D. Finzer, “Exploration vs. exploitation,” Medium, 04-May-2016. [Online]. Available: https://medium.com/@devinfinzer/exploration-vs-exploitation-b17612698511. [Accessed: 04-Jan-2020].
- [11] V. Mnih et al., “Playing atari with deep reinforcement learning,” arXiv preprint arXiv:1312.5602, 2013.