Lecture 4: Introduction to Reinforcement Learning

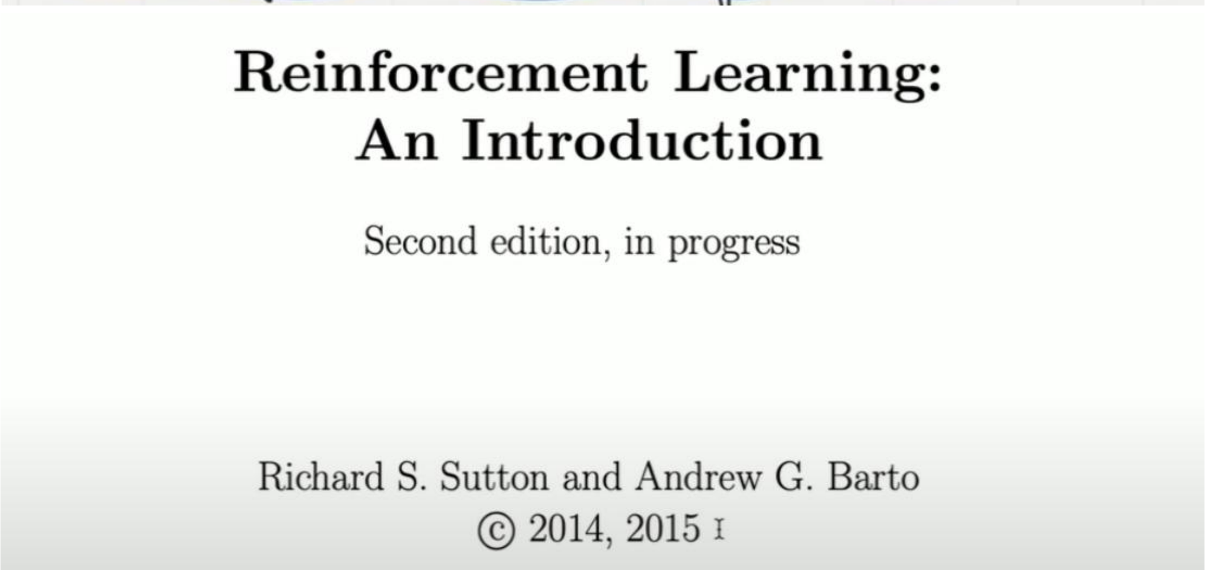
So far, we have covered the following methods for inducing reasoning in LLMs:-

(1) Inference- Time Compute Scaling ✓

Today, we will start to understand the second method for inducing reasoning in LLMs - Pure Reinforcement Learning

We will begin our journey by understanding first about classical reinforcement learning:-

The Reinforcement Learning Problem:-



Reinforcement Learning: An Introduction

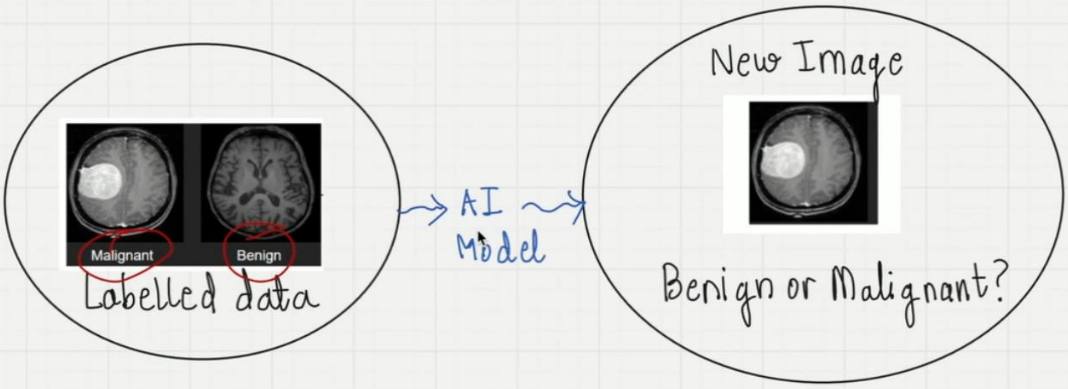
Second edition, in progress

Richard S. Sutton and Andrew G. Barto

© 2014, 2015

To understand about the RL problem, it would help us to understand how different is it from supervised learning and unsupervised learning.

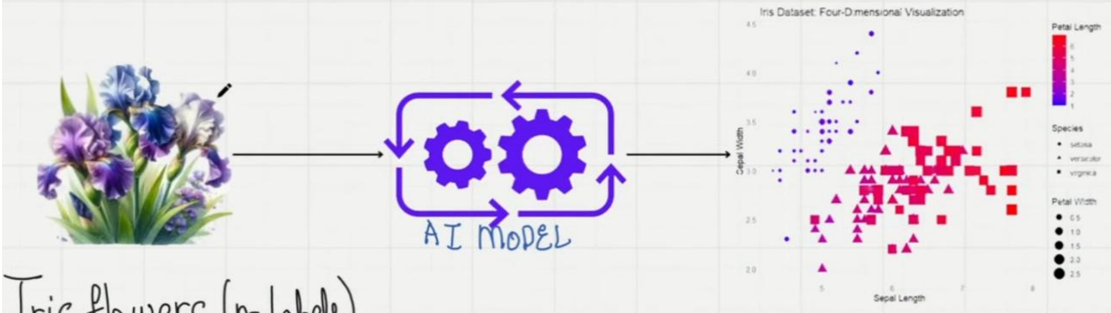
Supervised Learning:



Prediction on Unseen data

Objective: Be able to Generalize or extrapolate in situations not present in the training dataset.

Unsupervised Learning:-



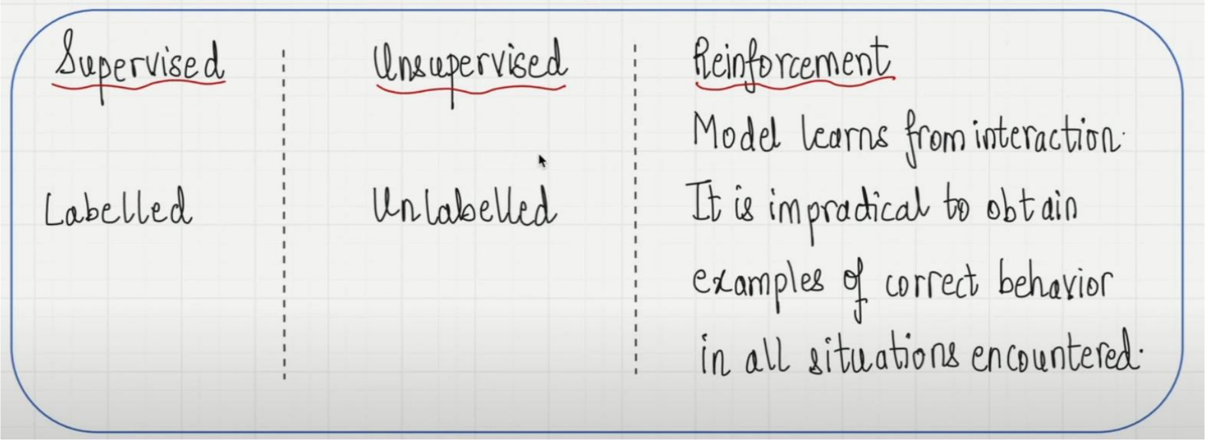
Iris flowers (no labels)

Clustering of data

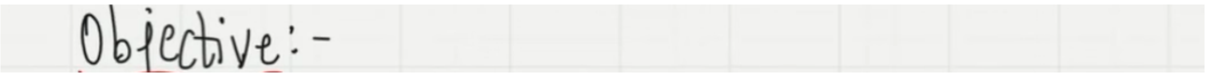
Objective: Finding structures hidden in collections of unlabelled data

Now, let us understand how reinforcement learning compares to both these categories

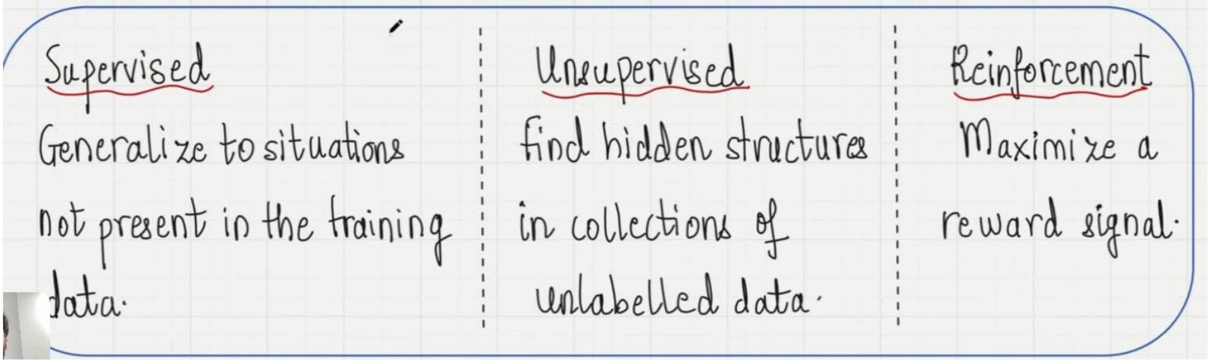
Labelling of data:-



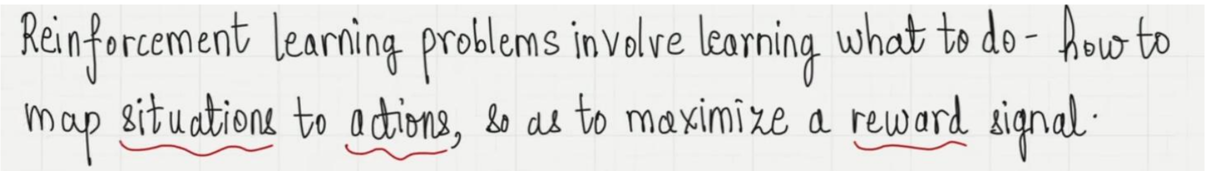
Supervised	Unsupervised	Reinforcement
Labelled	Unlabelled	Model learns from interaction. It is impractical to obtain examples of correct behavior in all situations encountered.



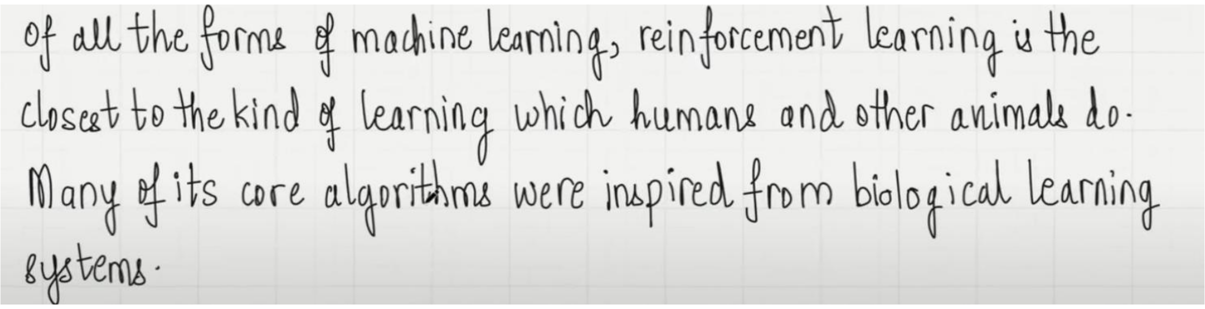
Objective:-



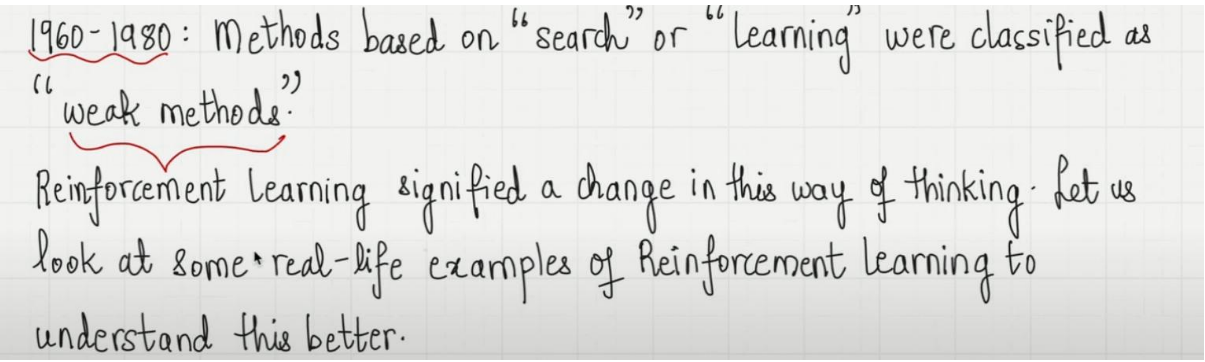
Supervised	Unsupervised	Reinforcement
Generalize to situations not present in the training data.	Find hidden structures in collections of unlabelled data.	Maximize a reward signal.



Reinforcement learning problems involve learning what to do - how to map situations to actions, so as to maximize a reward signal.



Of all the forms of machine learning, reinforcement learning is the closest to the kind of learning which humans and other animals do. Many of its core algorithms were inspired from biological learning systems.



1960-1980: Methods based on "search" or "learning" were classified as "weak methods".

Reinforcement Learning signified a change in this way of thinking. Let us look at some real-life examples of Reinforcement learning to understand this better.

The following examples have guided the development of Reinforcement Learning as a field:-

(1) A master chess player makes a move:-

The choice of the move is based on 2 things:

(1) Planning by anticipating replies and counterreplies.

(2) Intuitive judgements about the desirability of moves.

(2) How does a gazelle (animal) calf learn to run:

After birth:

2-3 minutes later:- Struggles to get up.

30 minutes later:- Runs at 36 km/hr.

(3) How does a mobile robot make decisions

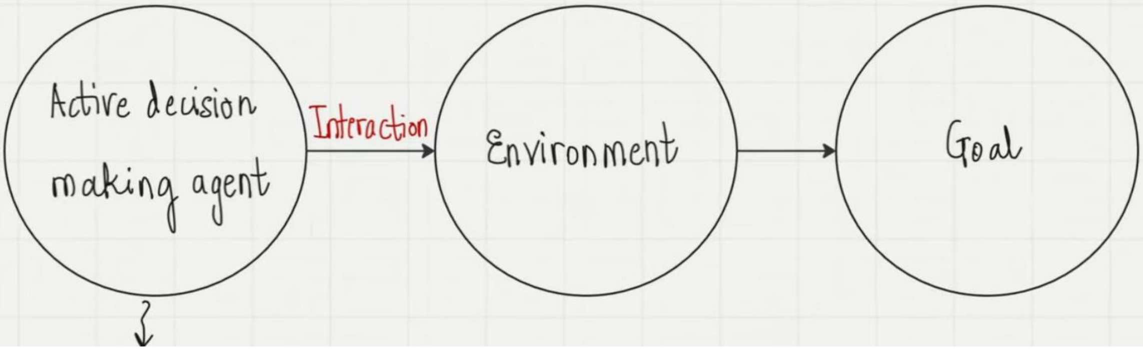
Decision Questions	Decision Parameters
Should I go to a new room in search of trash or go back to charging station?	Current battery charge How much time it has taken to find recharger in the past?

(4) Phil prepares his breakfast

Actions:- Walking to the cupboard, openingit, selecting a cereal box, reaching for, grasping and retrieving the box, obtain a plate, spoon.

Each step involves a series of eye-movements to obtain information and to guide reaching and locomotion. Each step is guided by goals in service of other goals with the final goal of obtaining nourishment.

What is common in all these examples?



Chess player, gazelle, mobile robot, Phil	Chess Board, Nature + Internal* for gazelle, Room + Battery, Kitchen + Internal* for Phil	Win game, run fast, max. trash collection, obtain nourishment

*refers to the agent's internal memories, preferences which also form a part of the environment.

In all these examples, the agent uses it experience to improve its performance overtime by interacting with the environment. Think how?

Elements of Reinforcement Learning:-

(1) Policy:-

Informally, policy defines the agent's way of behaving at a given time.

(2) Reward Signal:- Reward signal defines the goal in a reinforcement learning problem. At each time step, the environment sends to the reinforcement learning agent, a single number - a reward. The sole objective of the agent is to maximize the total rewards received over time.

Reward is analogous to pleasure or pain in a biological system. If you touch a boiling vessel, your body gives you a negative reward signal.

(3) Value function:- While reward signals indicate what is good in a immediate sense, value function specifies what is good in the long run.

Value of a state is the total amount of reward the agent can expect to accumulate over the future, starting from that state.

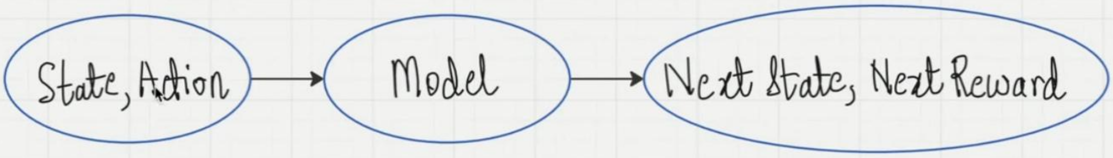
As an example, consider a sports tournament in which there are 14 matches. Let us say, a player X is selected by a team. Now, in the first 2 matches, this player does not play well. The reward signal is low. But the captain has faith in the player and believes that he will contribute in later matches. Hence, even if the reward signal is low, the value is high, since the long-term desirability of the player is high.

Reward Signal vs Value Function

In the field of reinforcement learning, the central role of value estimation is the most important thing which researchers learnt from 1960-1990.

(4) Model of the environment:-

Model is something which mimics the behavior of the environment.



There are model-based (used for planning) as well as model-free methods in reinforcement learning.

A Practical Example:- Game of Tic-Tac-Toe

We will understand the general idea of reinforcement learning using this example.

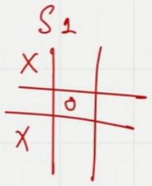
X	O	O
X	O	O
X		X

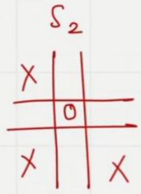
X: Player 1 (Us)

O: Player 2 (Opponent)

Our goal: To construct a player which can find flaws in the opponent's play and maximize the chances of winning.

State: Configuration of X's and O's on a 3x3 board.

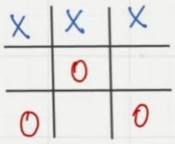
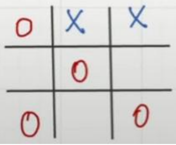
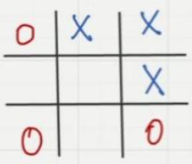




Policy: Rule which tells the player what move to make for every state of the game.

Value Function:-

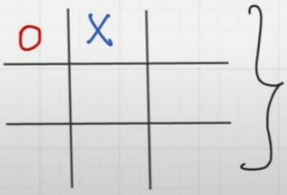
For every state, we assign a number which gives the latest estimate of the probability of winning from that state. Then we draw up a table.

State	Value
	1
	0
⋮	
	$\boxed{0.5}$

(Initial value of all other states - 0.5)

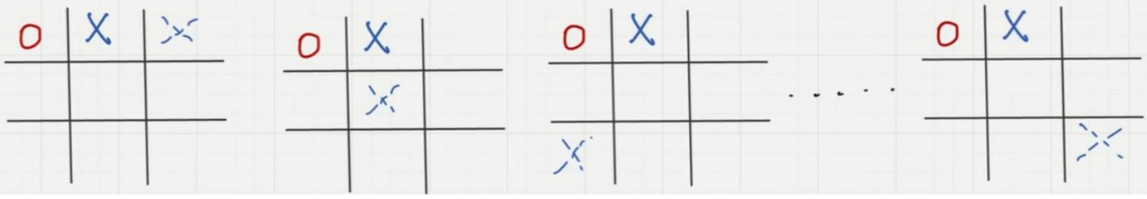
This table is the value function. The values corresponding to the states will change as our agent plays the game more number of times.

Now, let us see how we modify the value function estimates as we play the game more number of times such that they reflect the true probabilities.



Let us say, we are here. Now, to play our next move, we look at all possible next states and select the one with highest probability.

All Possible Next States:-



Value function: 0.5 0.6 0.55 0.4

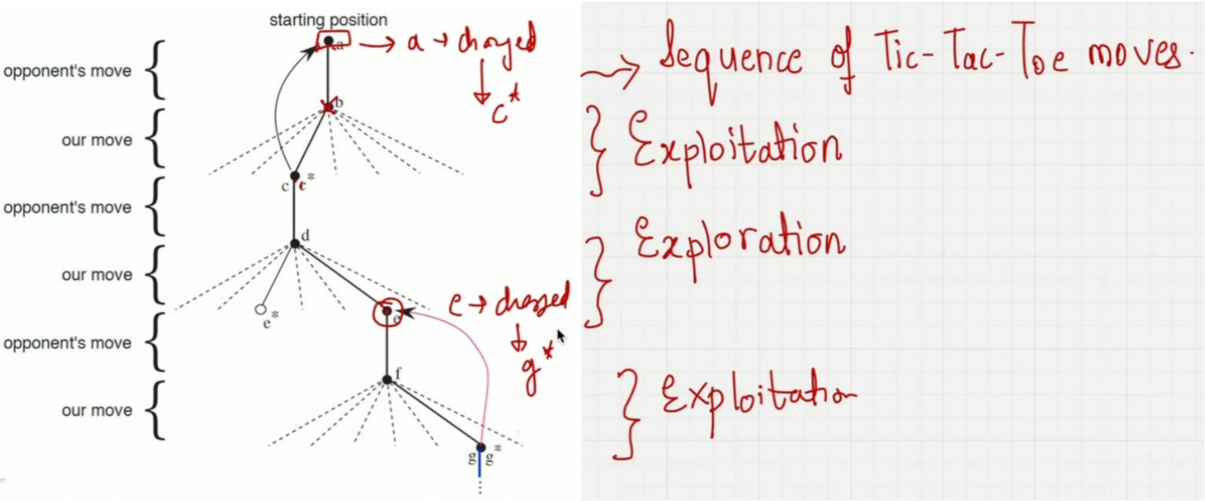
This is called exploitation: Selecting the state with the greatest value.

While we are playing, we change the value of the states we find ourselves. We make them more accurate estimates of the probability of winning.

IMPORTANT INTUITION:-

After each greedy move, the current value of the earlier state is adjusted to be closer to the value of the latter state.

This is done by moving the earlier state value a fraction of a way towards the latter state. This is called "backing-up"

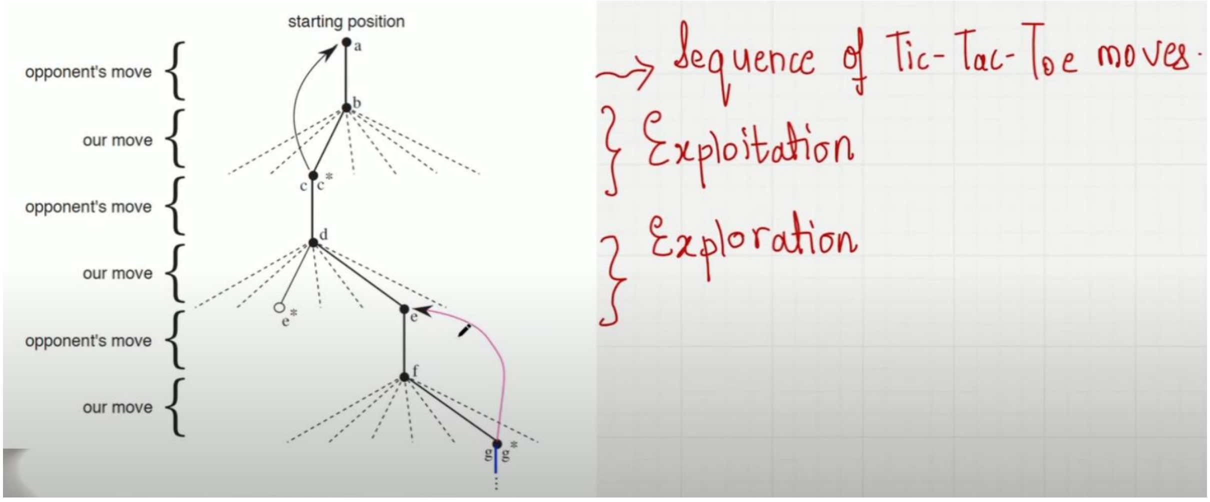


⤳ Sequence of Tic-Tac-Toe moves.

} Exploitation

} Exploration

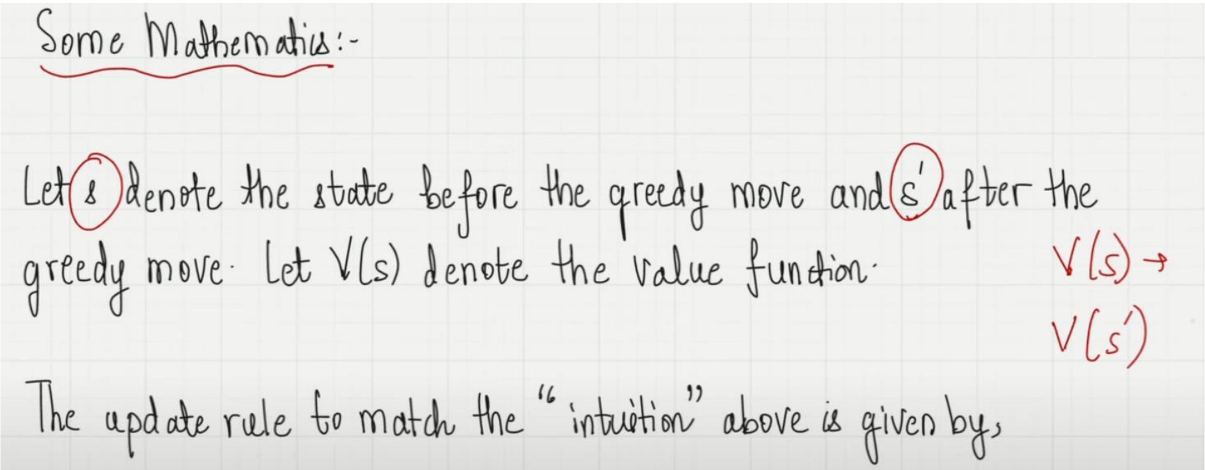
} Exploitation



⤳ Sequence of Tic-Tac-Toe moves.

} Exploitation

} Exploration

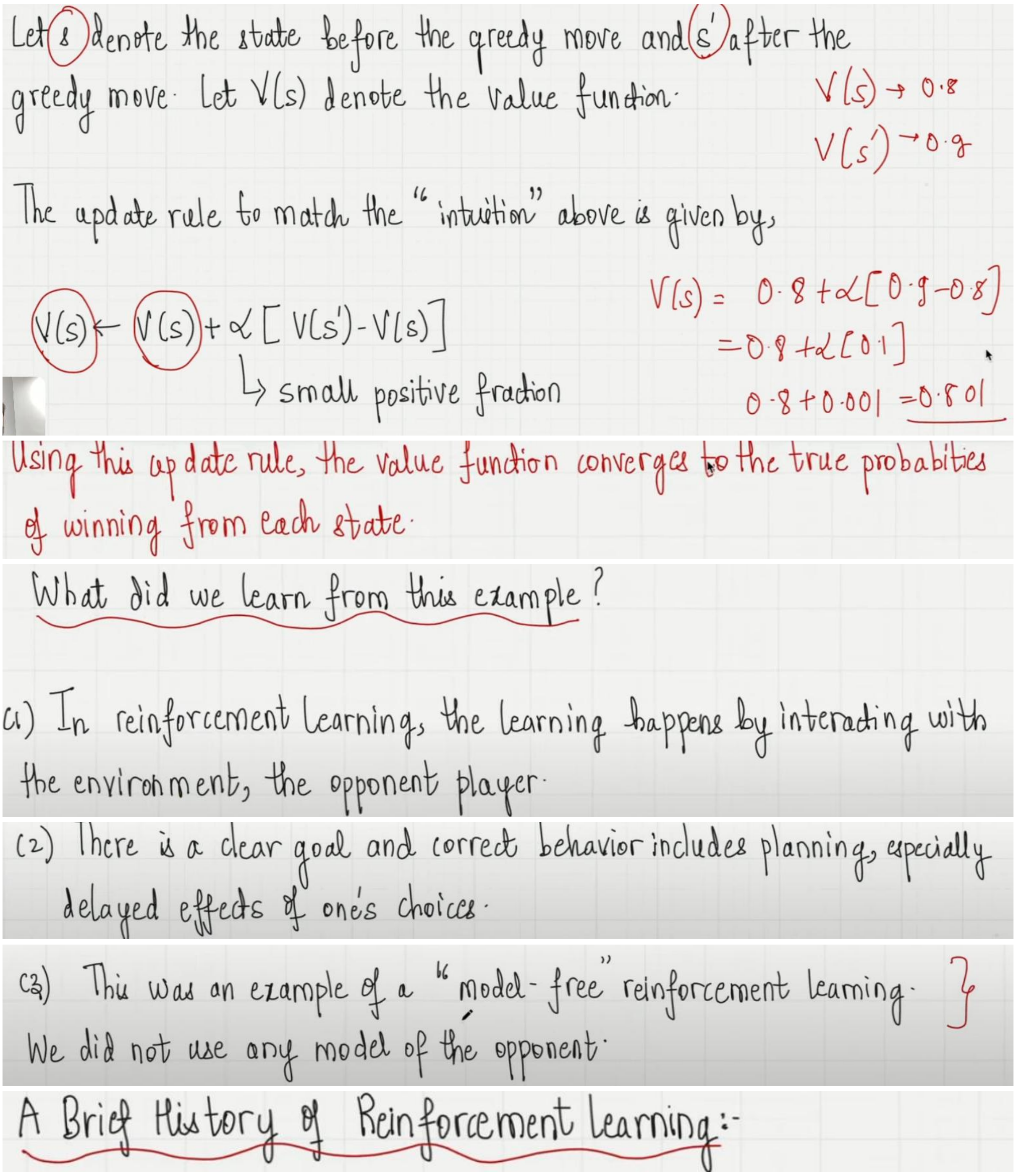


Some Mathematics:-

Let s denote the state before the greedy move and s' after the greedy move. Let $V(s)$ denote the value function.

$V(s) \rightarrow$
$V(s')$

The update rule to match the "intuition" above is given by,



Let s denote the state before the greedy move and s' after the greedy move. Let V(s) denote the value function.

$V(s) \rightarrow 0.8$

$V(s') \rightarrow 0.9$

The update rule to match the "intuition" above is given by,

$$V(s) \leftarrow V(s) + \underbrace{\alpha}_{\text{small positive fraction}} [V(s') - V(s)]$$

$$\begin{aligned} V(s) &= 0.8 + \alpha[0.9 - 0.8] \\ &= 0.8 + \alpha[0.1] \\ 0.8 + 0.001 &= \underline{0.801} \end{aligned}$$

Using this update rule, the value function converges to the true probabilties of winning from each state.

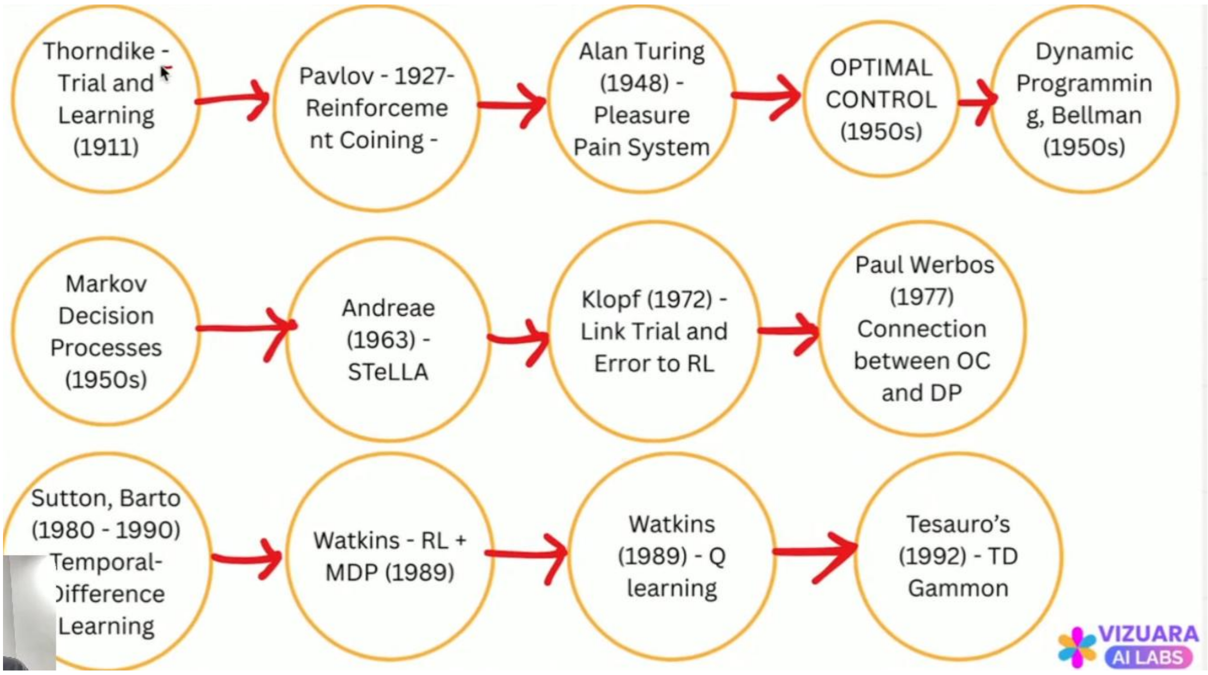
What did we learn from this example?

(1) In reinforcement Learning, the learning happens by interacting with the environment, the opponent player.

(2) There is a clear goal and correct behavior includes planning, especially delayed effects of one's choices.

(3) This was an example of a "model-free" reinforcement learning. We did not use any model of the opponent. }

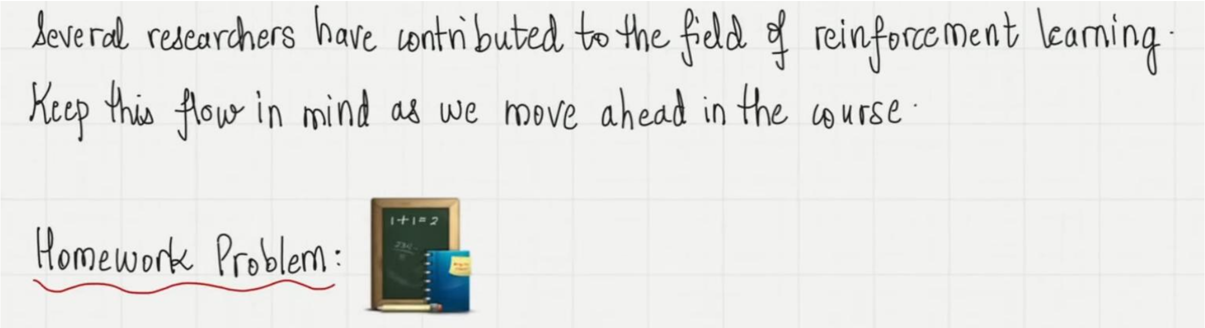
A Brief History of Reinforcement Learning:-



Thorndike - Trial and Learning (1911) → Pavlov - 1927- Reinforcement Coining - → Alan Turing (1948) - Pleasure Pain System → OPTIMAL CONTROL (1950s) → Dynamic Programming, Bellman (1950s)

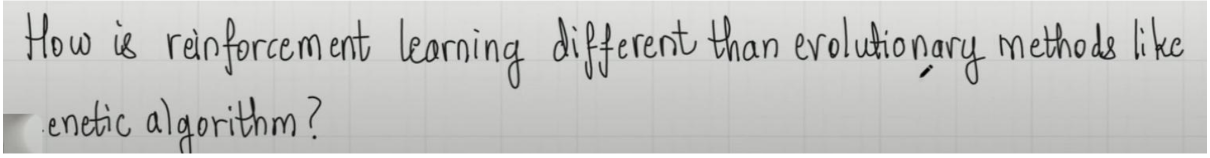
Markov Decision Processes (1950s) → Andreae (1963) - STeLLA → Klopf (1972) - Link Trial and Error to RL → Paul Werbos (1977) Connection between OC and DP

Sutton, Barto (1980 - 1990) Temporal-)ifference Learning → Watkins - RL + MDP (1989) → Watkins (1989) - Q learning → Tesauro's (1992) - TD Gammon



Several researchers have contributed to the field of reinforcement learning. Keep this flow in mind as we move ahead in the course.

Homework Problem:



How is reinforcement learning different than evolutionary methods like .enetic algorithm?