Safe Landings In Deep Space Using RL Techniques

RL 3547 – University Of Toronto – December 2019

The Team

Adnan Lanewala

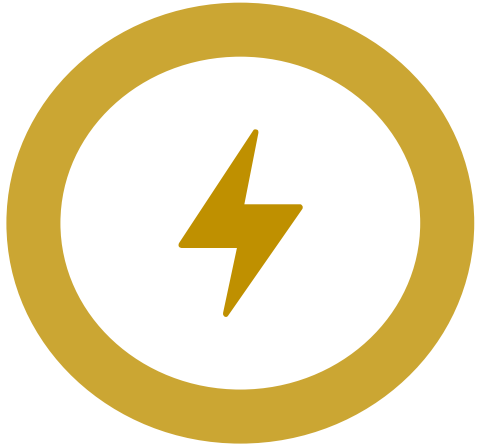
Naresh Patel

Nisarg Patel

AGENDA

Problem

Key Findings



Solution

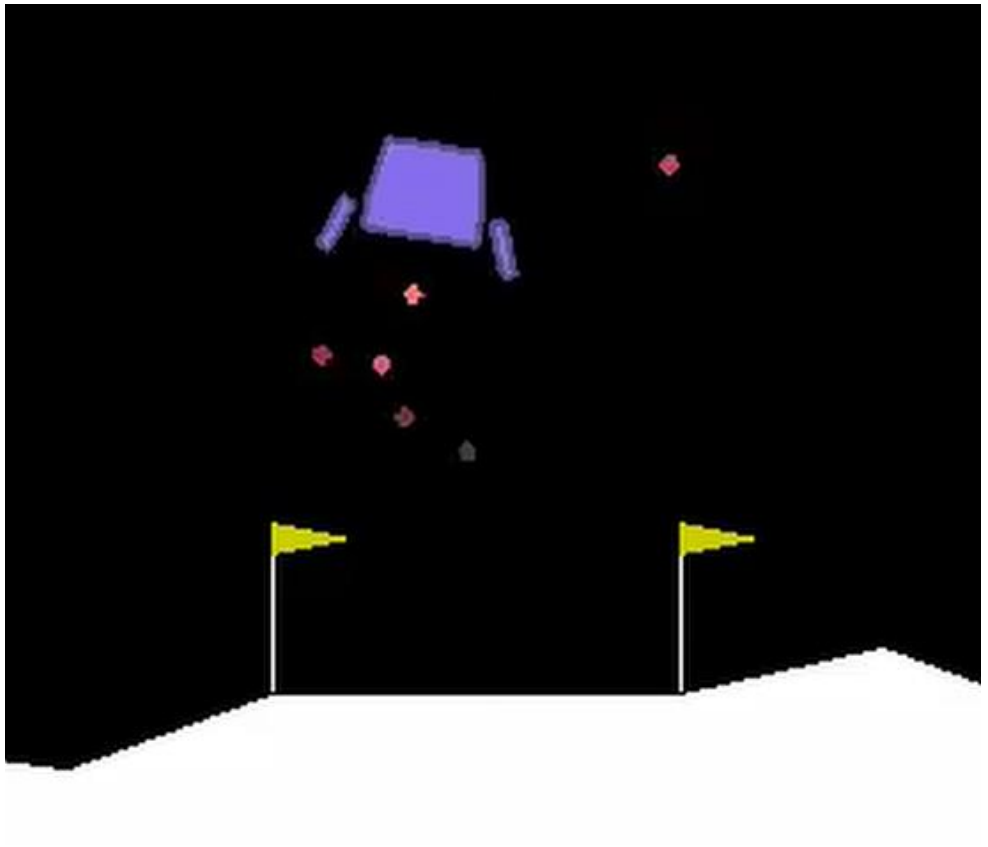
Problem Statement

Satellite Cost

USD $10 – 400* million

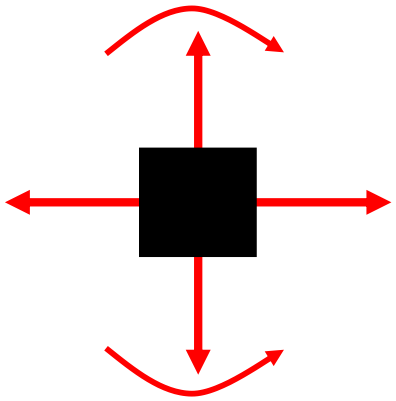
*According to https://science.howstuffworks.com/satellite10.htm

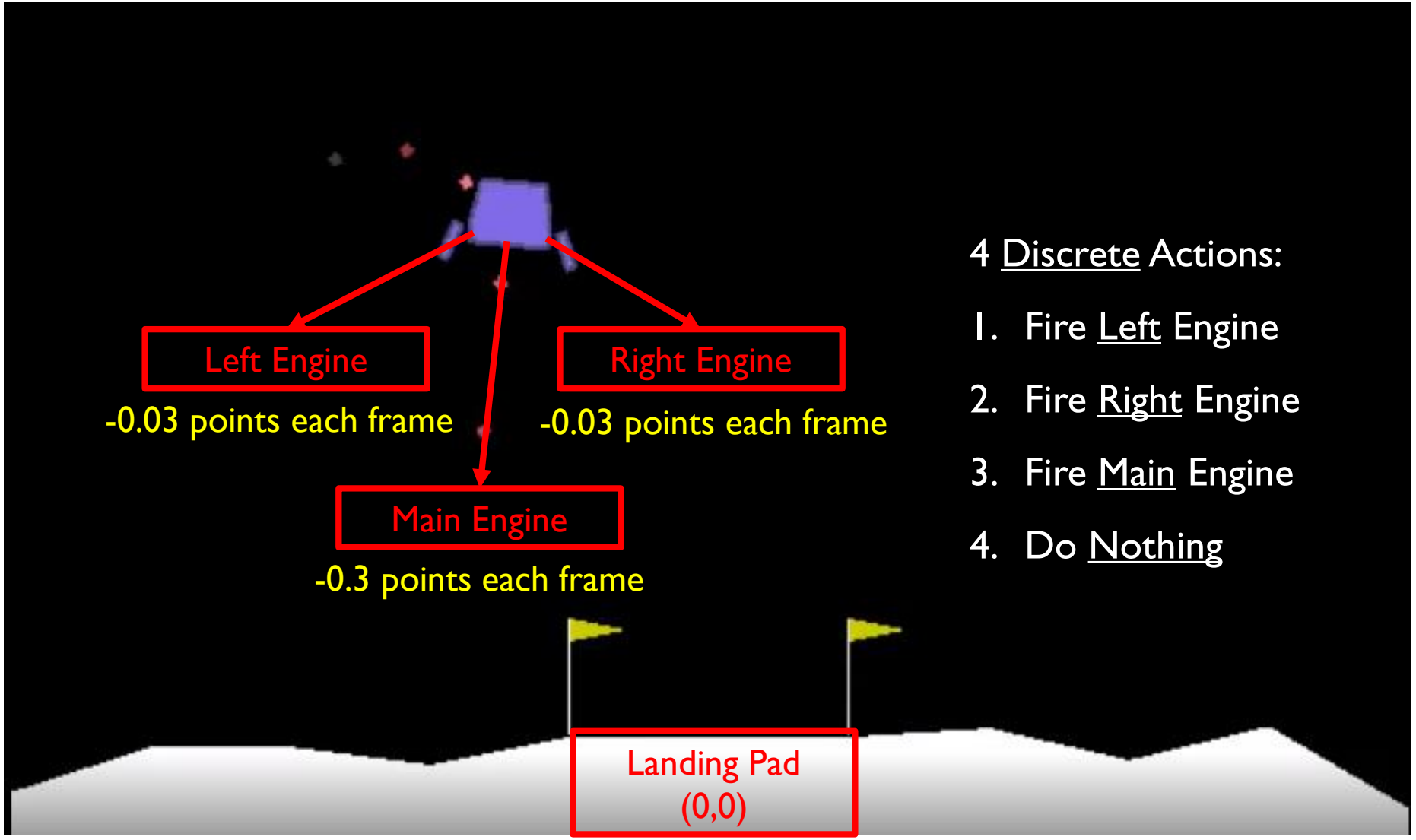
Problem Statement



Open AI Gym – Lunar Lander (v2) Environment

Action Space

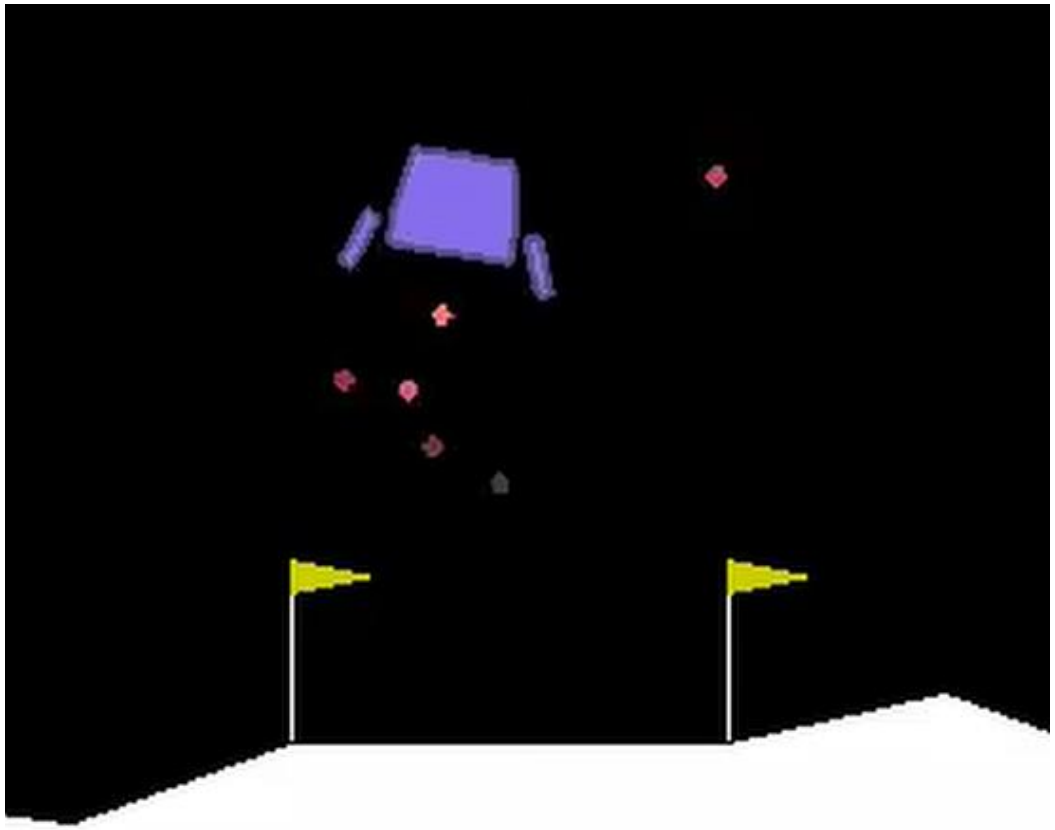




4 Discrete Actions:

1. Fire Left Engine
2. Fire Right Engine
3. Fire Main Engine
4. Do Nothing

Reward



Crash	– 100 points
Come to Rest	+ 100 points
Leg Contact	+ 10 points
Zero Speed Touchdown	+ (100 – 140) points
Solved	+ 200 points
Firing Main Engine	– 0.3 points per frame
Firing Side Engine	– 0.03 points per frame
Fuel Supply	Unlimited

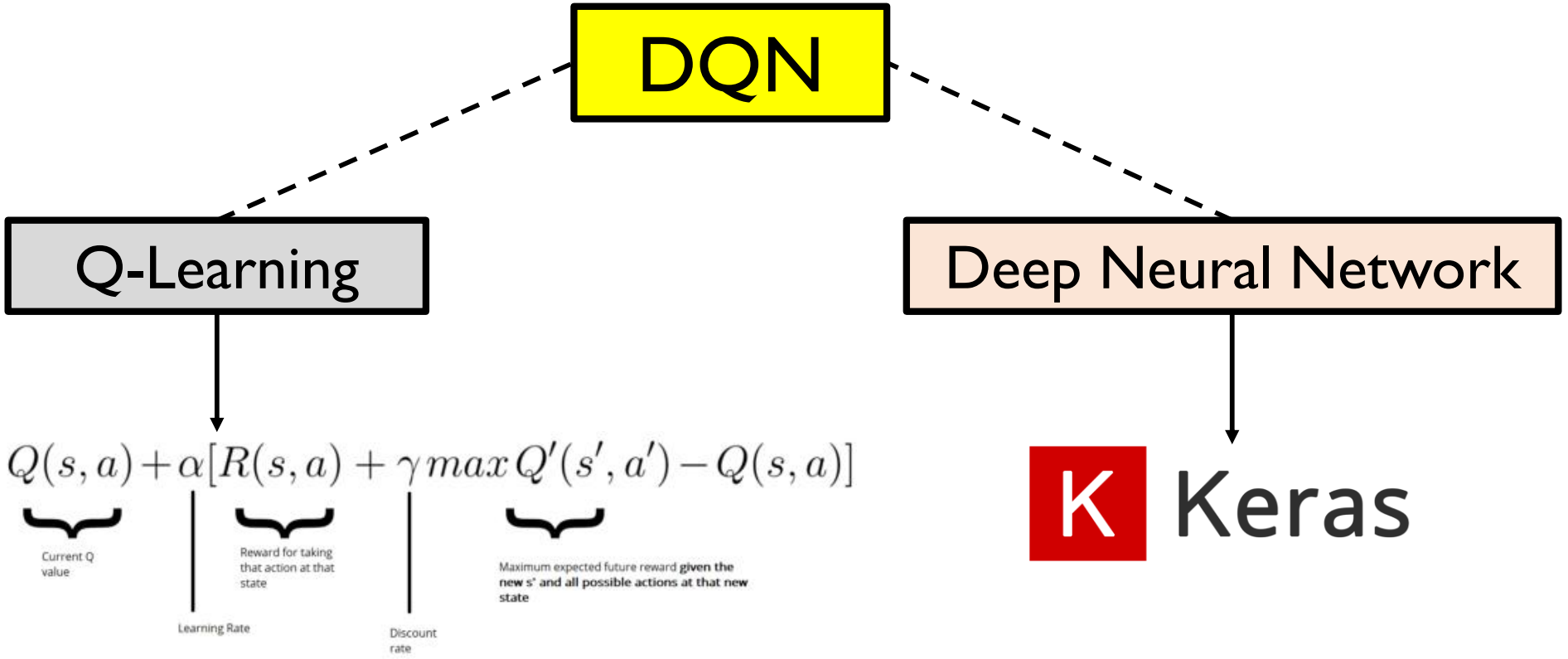
Solution – Algorithm

Comparison of reinforcement learning algorithms

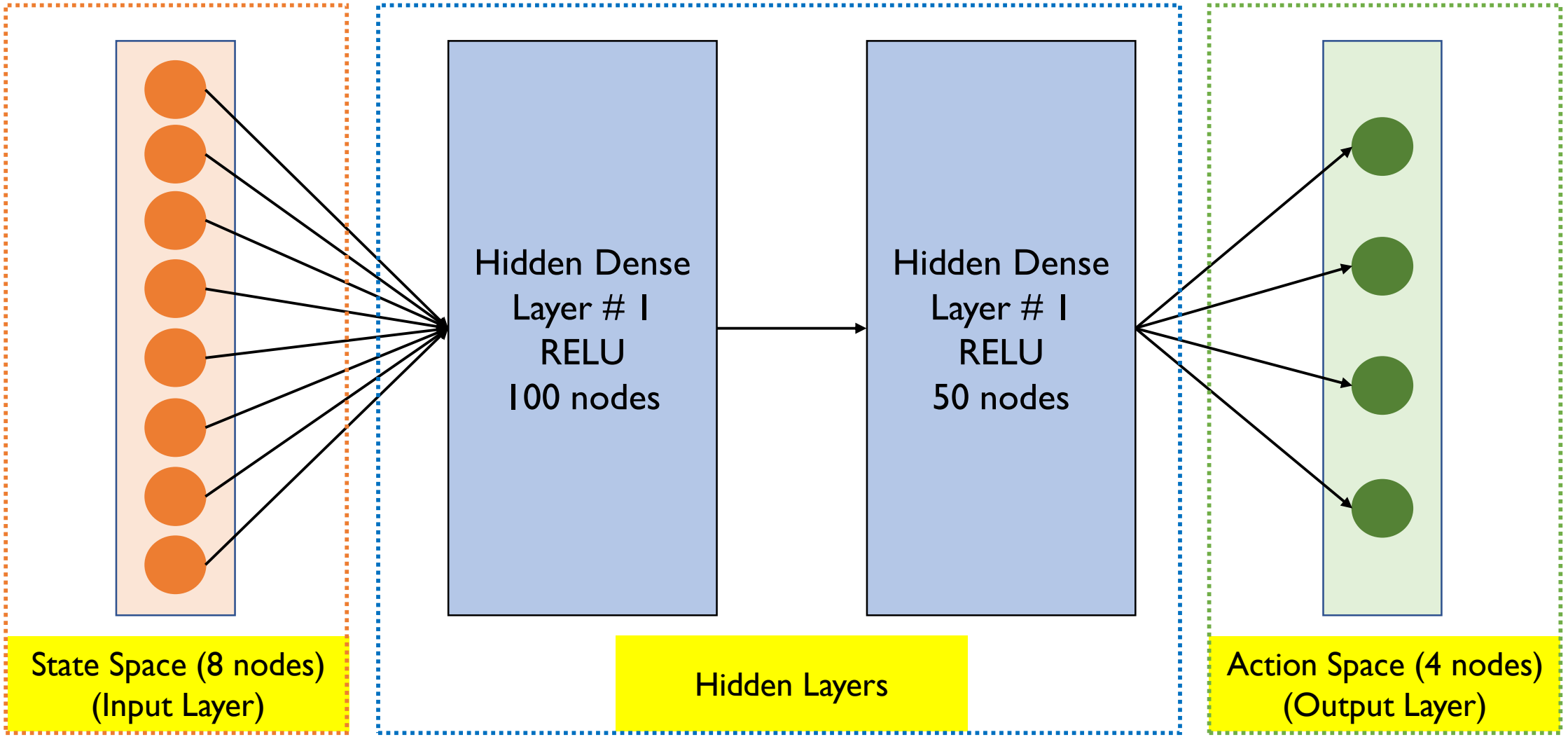
Algorithm	Description	Model	Policy	Action Space	State Space	Operator
Monte Carlo	Every visit to Monte Carlo	Model-Free	Off-policy	Discrete	Discrete	Sample-means
Q-learning	State–action–reward–state	Model-Free	Off-policy	Discrete	Discrete	Q-value
SARSA	State–action–reward–state–action	Model-Free	On-policy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

Information retrieved from https://en.wikipedia.org/wiki/Reinforcement_learning

Solution – DQN



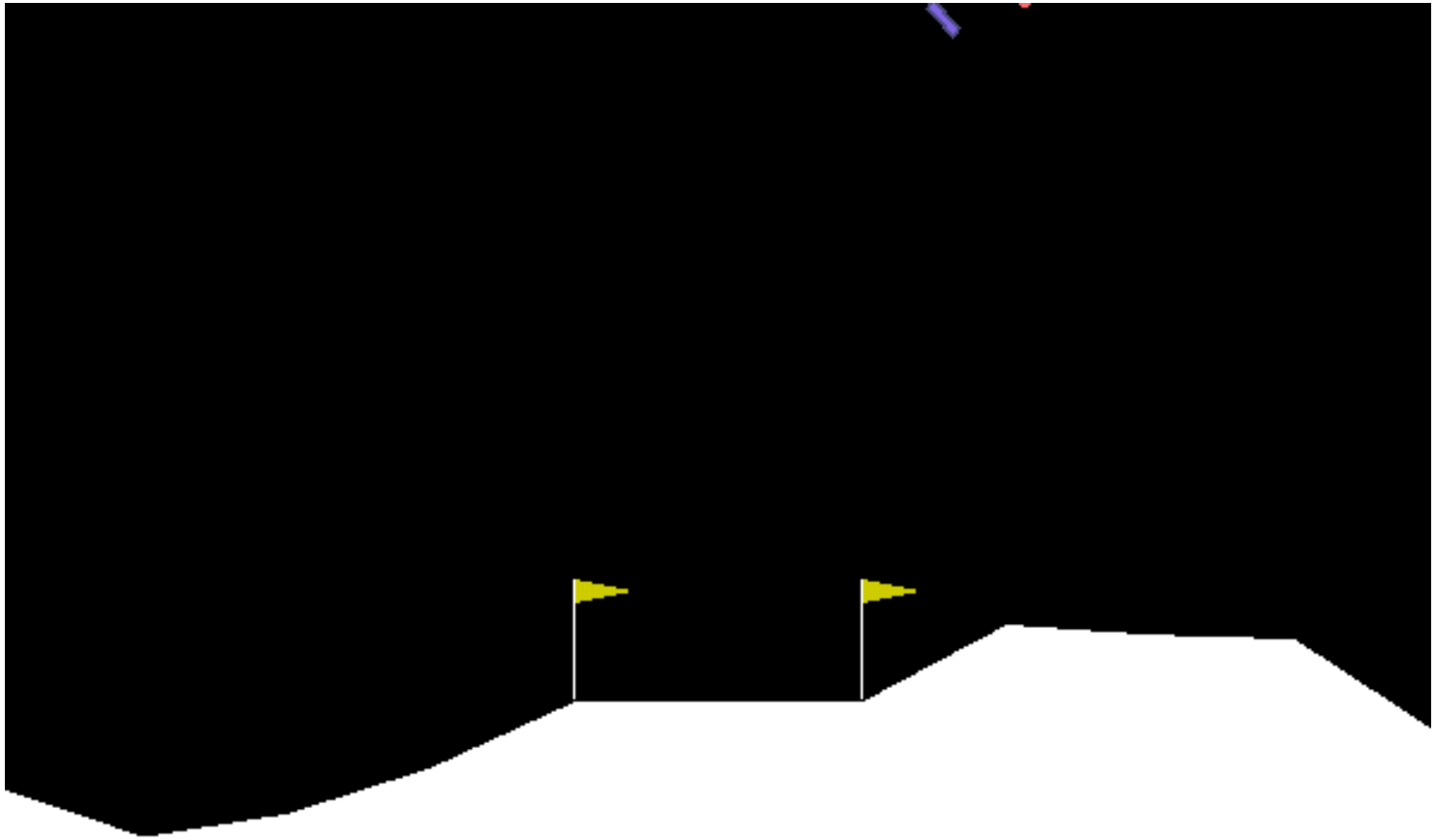
Solution – Network



Solution – Parameters

Discount Rate – long term reward	γ = 0.99
Exploration Rate at Start	ϵ_{max} = 1.0
Exploration Rate min. value	ϵ_{min} = 0.01
Exploration Rate decay rate	ϵ_{decay} = 0.997
Learning Rate	α = 0.001
Score Threshold	+ 50 points
# of Episodes for Avg Score	100 episodes
Early Stopping	Yes
Batch Size	64
Max Steps	2,000
Max Episodes	500

Solution – Training



Episode = 0

Solution – Training

Episode = 10

Episode = 50

Solution – Training

Episode = 100

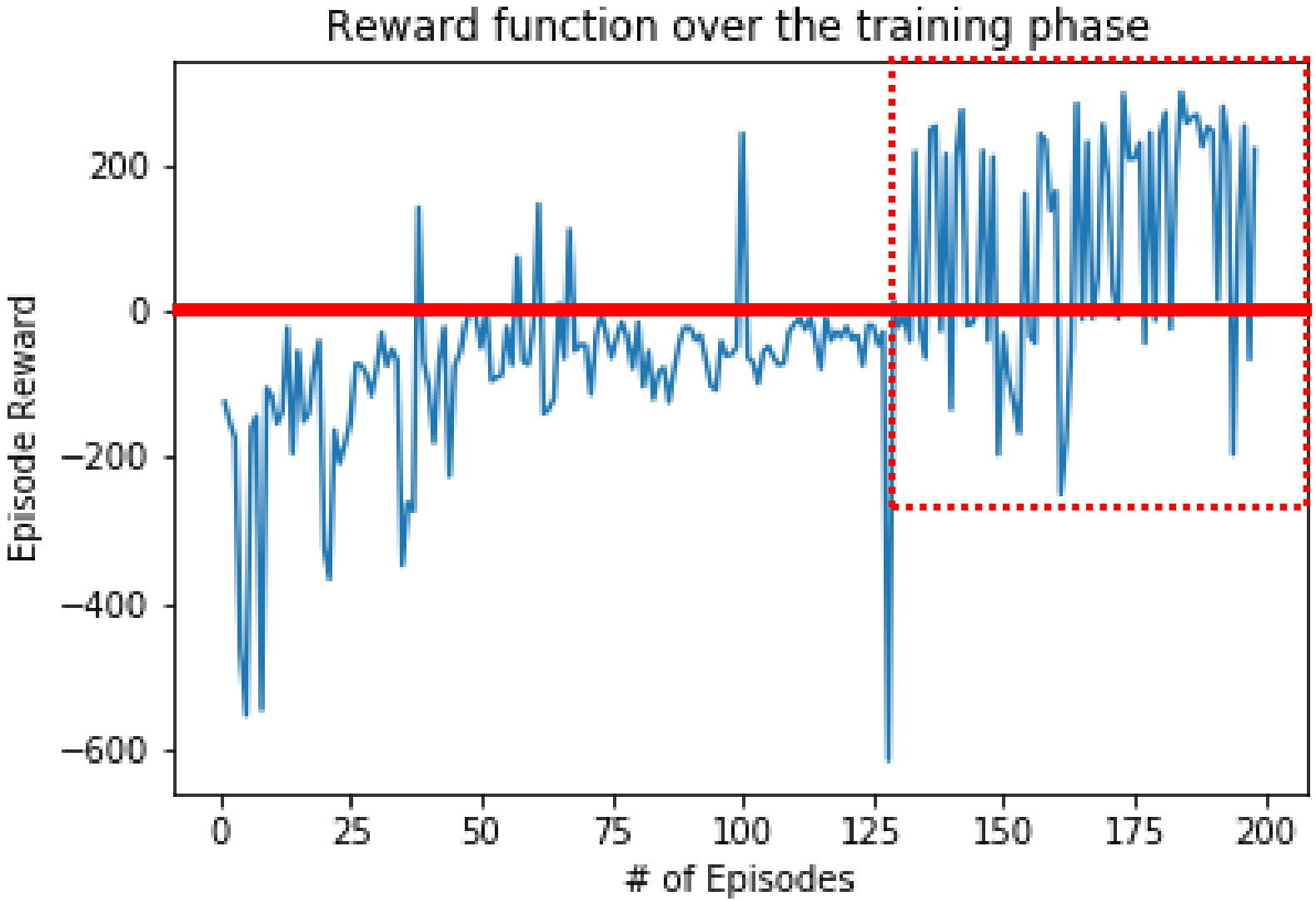
Episode = 150

Solution – Training

Episode = 180

Episode = 198

Solution – Outcome



Key Outcomes

Produced a generalized model that safely lands the satellite with good landing principles (approach)

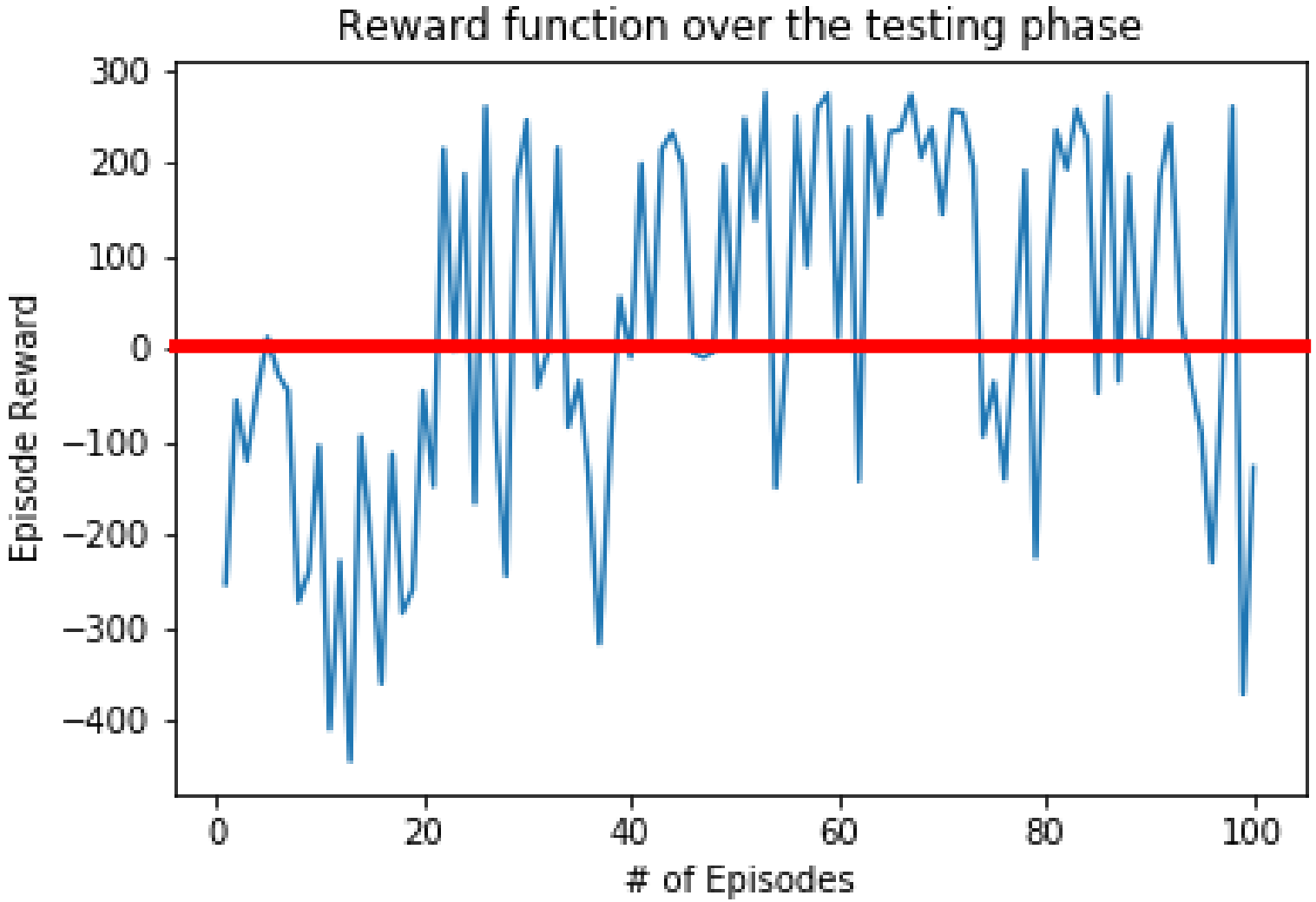
+50 points

Average Score

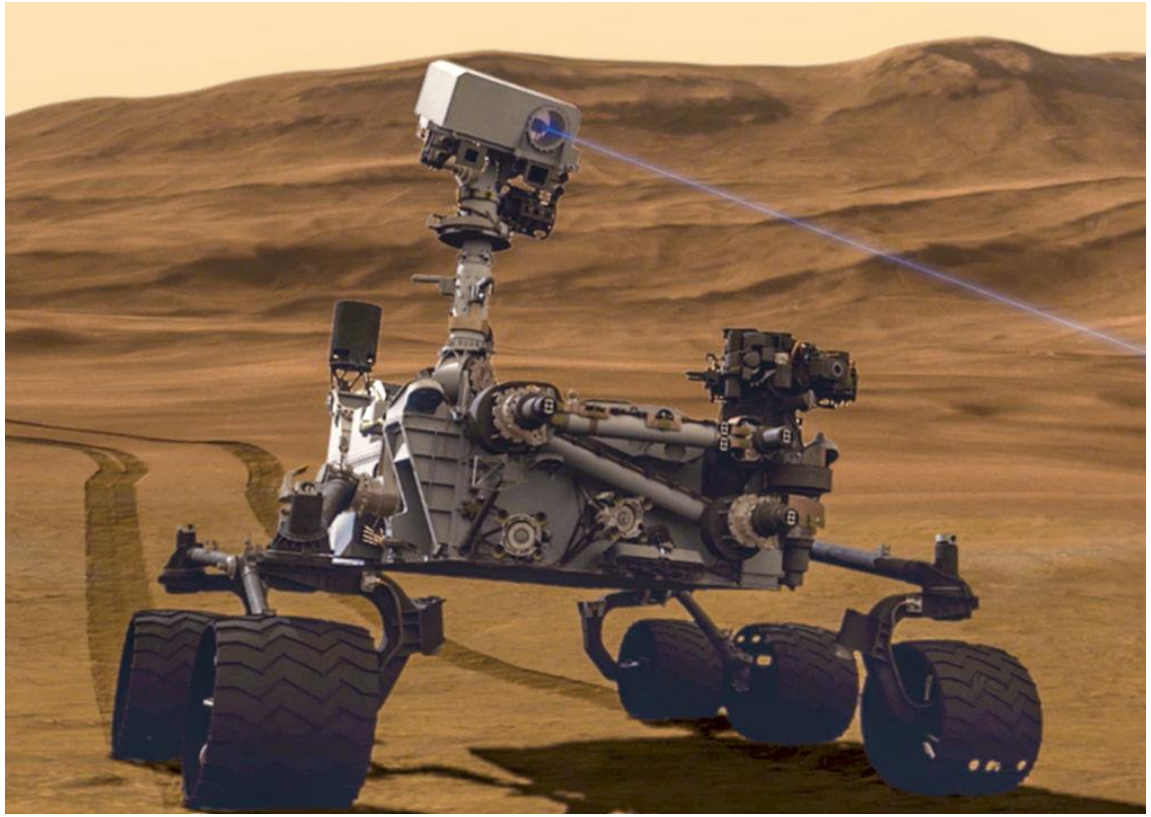
100 Episodes

# of Episodes

Test Results



Thank You



SCAN ME

https://github.com/nishp763/SCS-RL-3547-Final-Project

Appendix – Next Steps

- Multi-agent & Collaboration
- Different Algorithms (SARSA, Policy Iteration, RL techniques, Genetic mutations)
- More Training with Higher Score threshold
- Failure Simulation & Auto-correction (Engine Failure, Crosswind, Tumbling, Loss of Control)
- Work with Limited Fuel Supply

Appendix – Keras