# Improving Earth Science Missions with Deep Reinforcement Learning

Alberto Candela, Jason Swope, and Steve Chien

Jet Propulsion Laboratory, California Institute of Technology, CA, USA

## INTRODUCTION

Fundamental physics of remote sensing dictates that high spatial resolution at reduced size (and therefore power, cost) forces reduced swath. This places a premium on measurement on acquiring the highest science value data enabled by pointable instruments in Earth science missions. Dynamic targeting (DT) can improve the efficiency of conventional expensive narrow swath instruments. DT is a decision-making approach that leverages information from a lookahead sensor to identify targets for the primary instrument, which can then be pointed to improve science yield (Fig. 1).

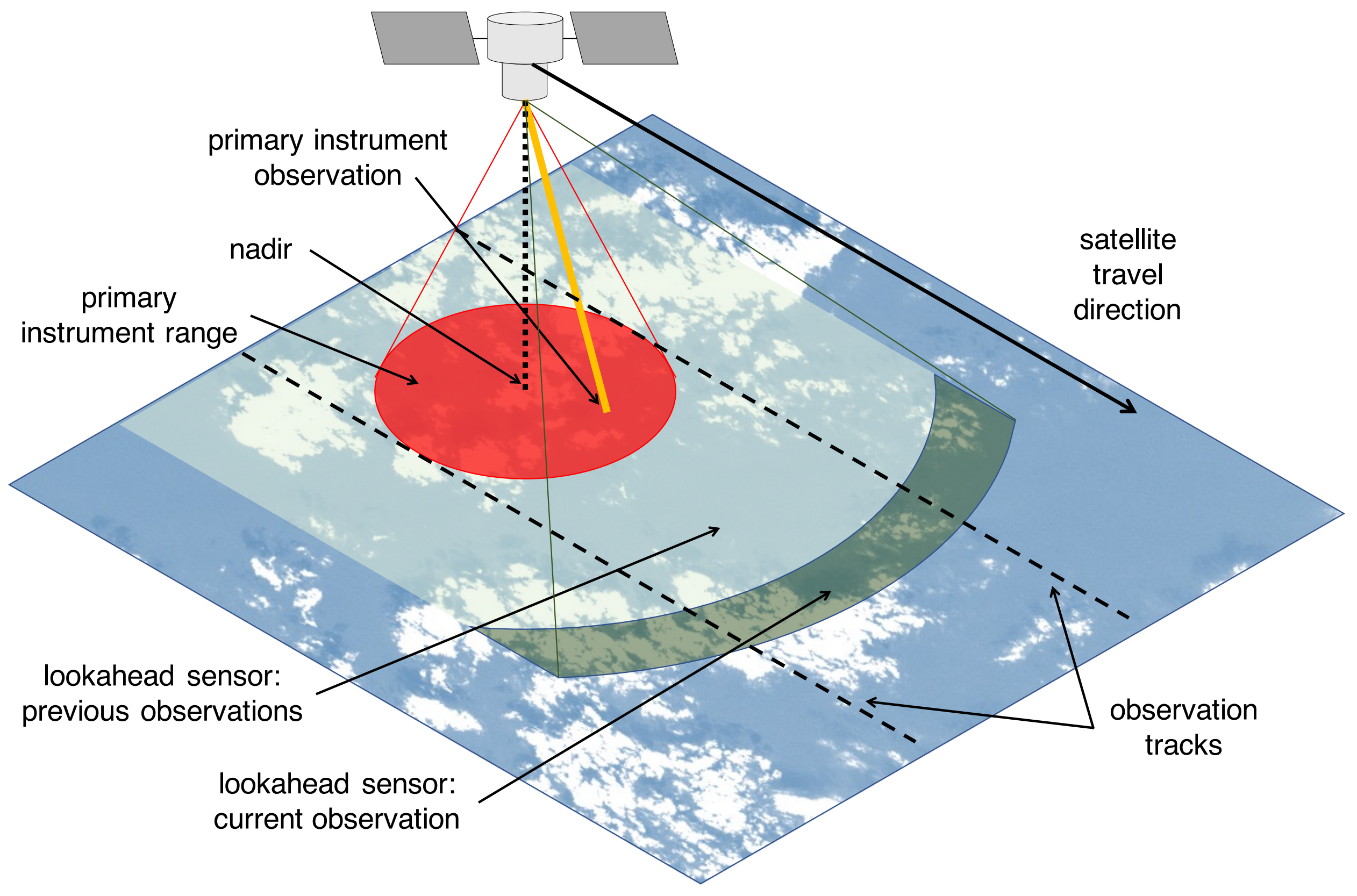
primary instrument observation
nadir
primary instrument range
satellite travel direction
lookahead sensor: previous observations
observation tracks
lookahead sensor: current observation

*Fig 1. Dynamic targeting leverages information from a lookahead sensor to identify targets for the primary instrument to improve science yield given energy constraints.*

## RELATED WORK

Most work has focused on screening cloud cover and other poor observing conditions from airborne and spaceborne missions. This work is an extension of a NASA study for the Smart Ice Cloud Sensing (SMICES) satellite concept, whose objective is to employ Artificial Intelligence (AI) to make better decisions while collecting dynamic measurements of ice clouds and storms.

## SIMULATION STUDY

The approach is evaluated in a simulation study that consists of an Earth-observing satellite with two onboard instruments: a primary radar with a narrow swath of 217 km, and a secondary radiometer with a lookahead of 420 km. General Mission Analysis Tool (GMAT) was used to simulate and generate realistic satellite trajectories. The simulation study consists of the following mission scenarios and datasets:

1) **Storm Hunting**: the goal is to observe storm clouds; global data comes from the Global Precipitation Measurement (GPM) mission (Fig. 2)

2) **Cloud Avoidance**: the goal is to acquire clear-sky measurements; data comes from global cloud fraction products from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Fig. 3).

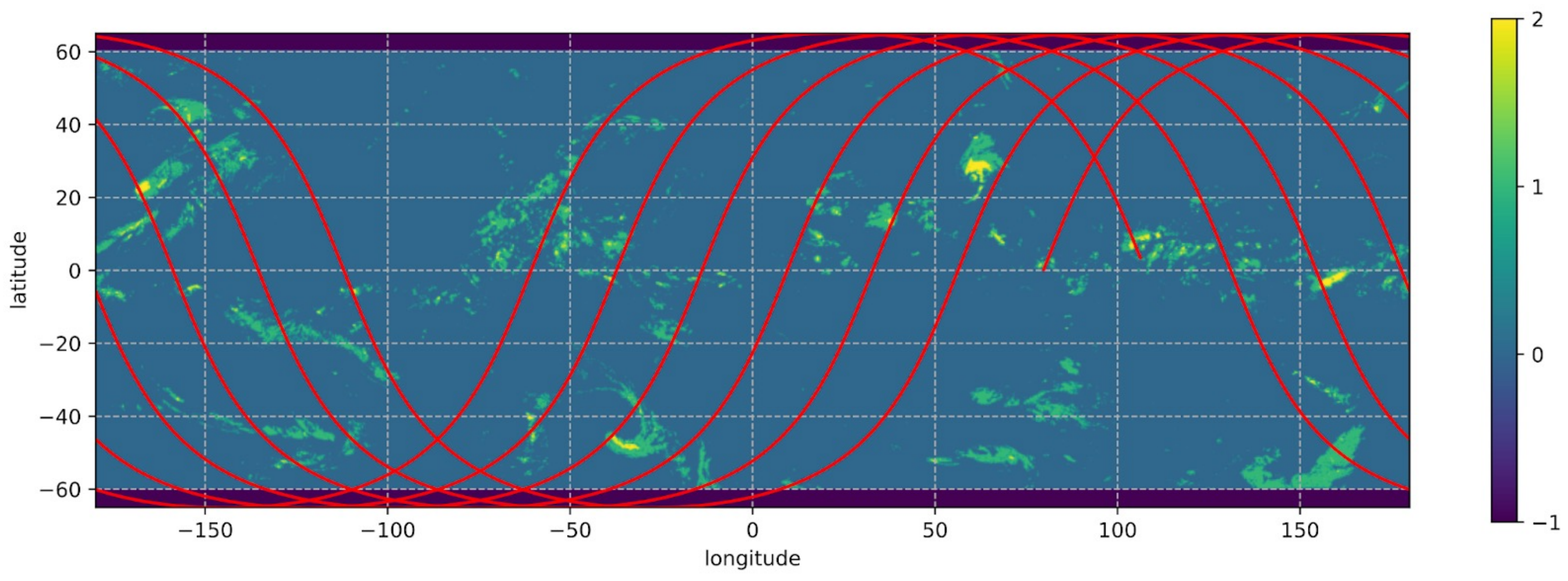

*Fig 2. GPM global storm data set and simulated satellite orbit with a 65 degree inclination.*

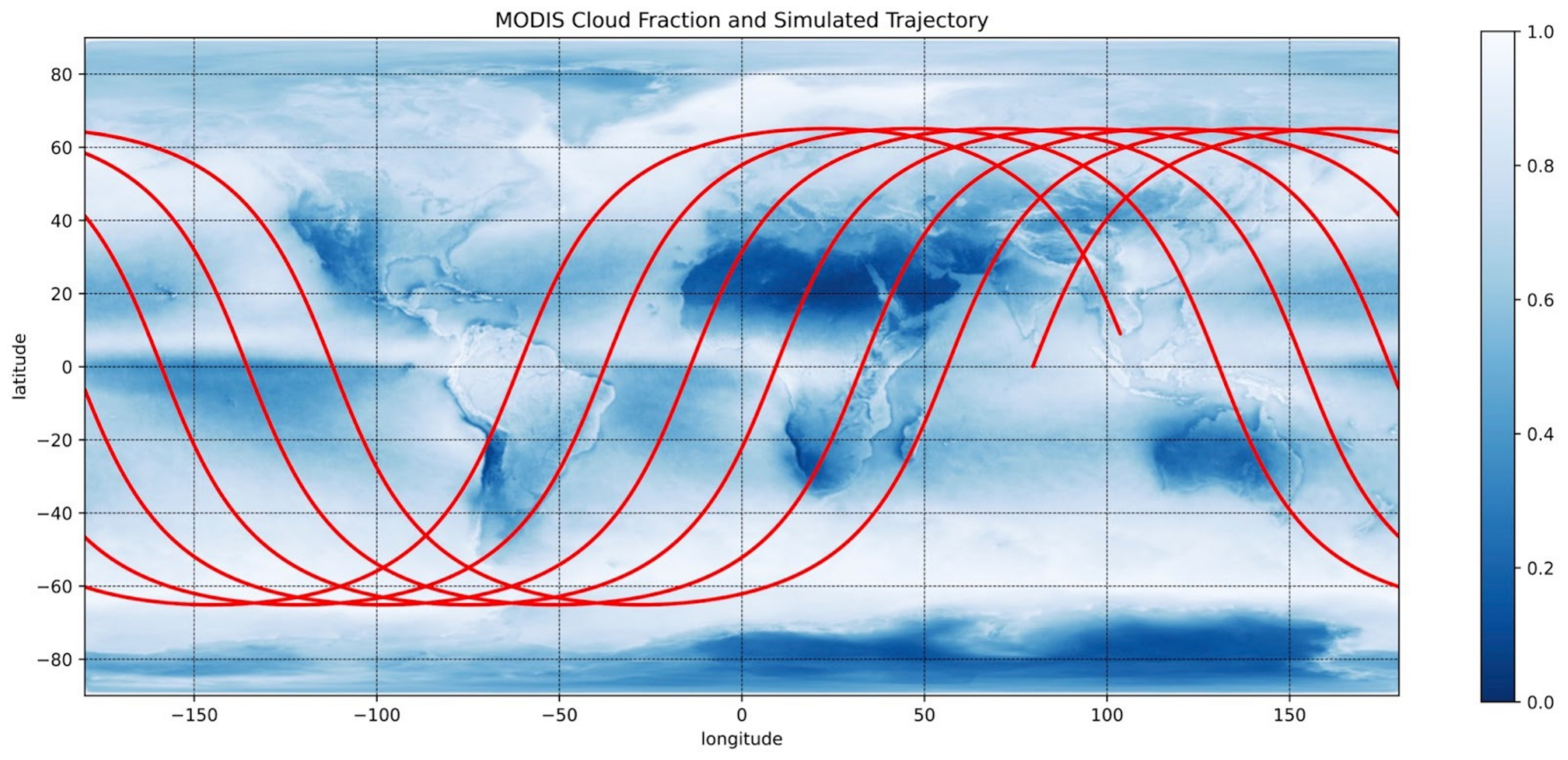

*Fig 3. MODIS global cloud fraction dataset and simulated satellite orbit with a 65 degree inclination.*

**National Aeronautics and Space Administration**

**Jet Propulsion Laboratory**
California Institute of Technology
Pasadena, California

**www.nasa.gov**

## APPROACH

We formulate DT as a pointing planning problem. We want to observe scientific phenomena of interest more often while screening poor observing conditions and respecting the energy constraints. In previous work, we developed several DT algorithms that draw from a rich heritage of decision-making methods involving AI, operations research, and heuristic search.

We use deep reinforcement learning to build upon this previous work as follows:

1) RL Simulator: we adapted our spacecraft mission simulator for the storm hunting and cloud avoidance scenarios so it conforms to this Markov Decision Process (MDP) formulation (Fig. 4):

   - state space (continuous): [image derived from lookahead instrument, state of charge]
   - action space (continuous): [sample flag variable, radius, angle]
   - time step: 1 second

2) Optimal actions for a few orbits: to this end we used a dynamic programming (DP) algorithm (Fig. 5); it is not deployable on missions as it requires unrealistic instrument and compute resources

3) Imitation learning on a few orbits: we conduct behavioral cloning by training a convolutional neural network (CNN) to predict optimal actions from states (trained with ~1 million states and actions)

4) Reinforcement learning on more orbits (*ongoing work*): we perform transfer learning by using the pretrained CNN, we continue its training on new orbits using the Proximal Policy Optimization (PPO) algorithm, which is an actor-critic method that supports continuous action spaces

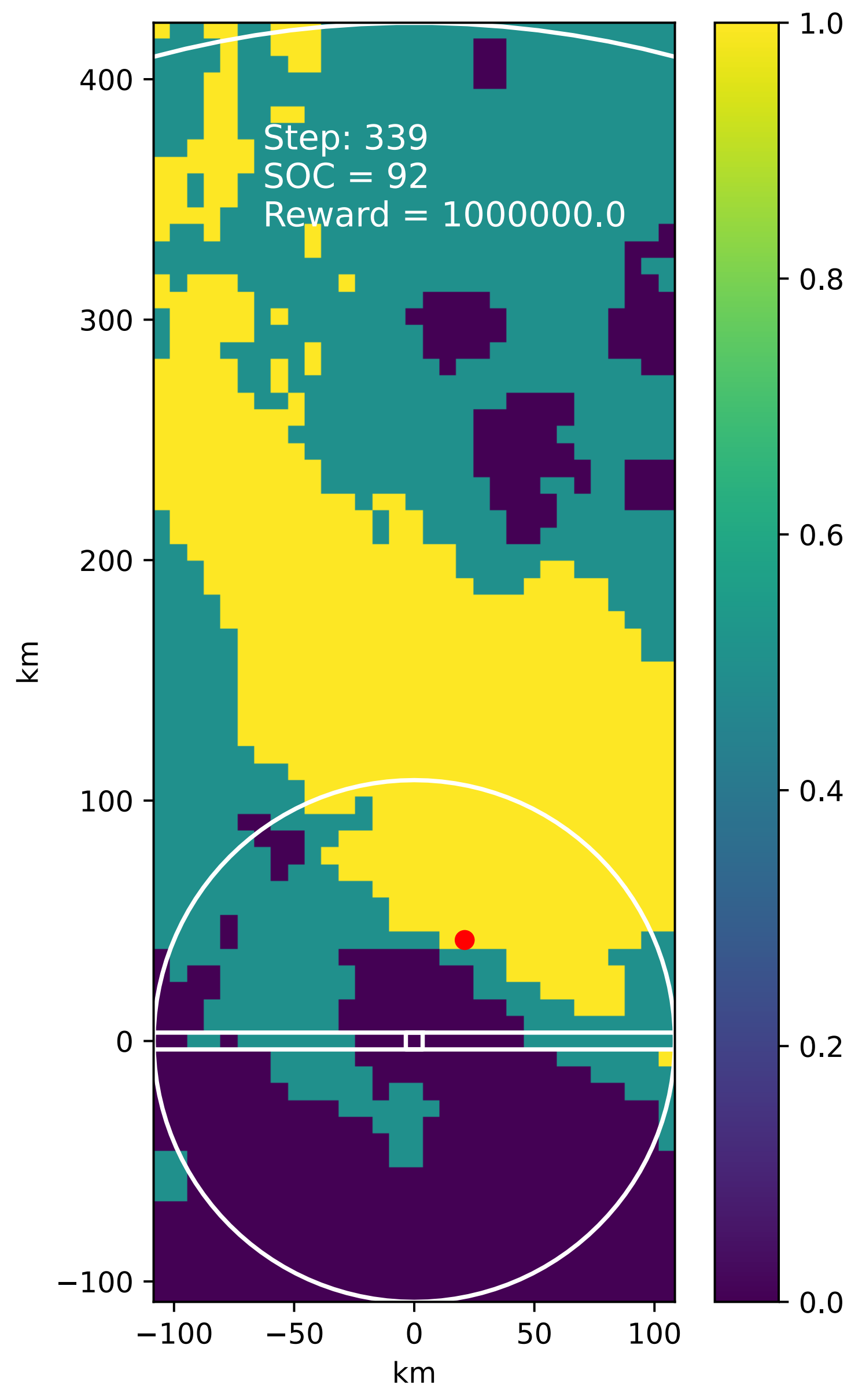

*Fig. 4. The Earth science mission simulator provides primary and lookahead instrument observations while conforming to an MDP formulation for reinforcement learning.*

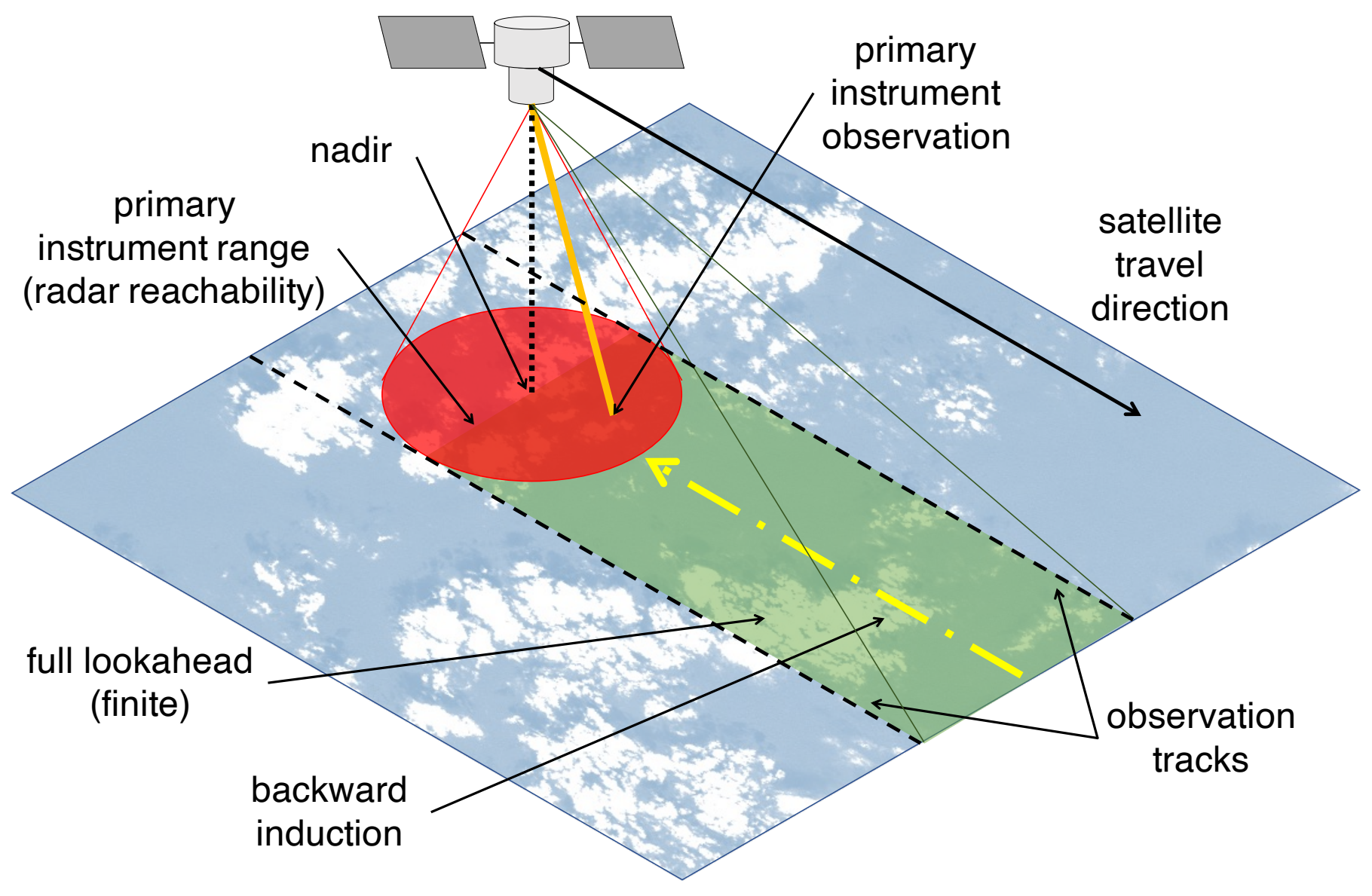
nadir
primary instrument observation
primary instrument range (radar reachability)
satellite travel direction
full lookahead (finite)
observation tracks
backward induction

*Fig. 5. The DP algorithm has a full lookahead (assuming the path is finite) and achieves optimality, but in general it cannot be deployed on missions.*

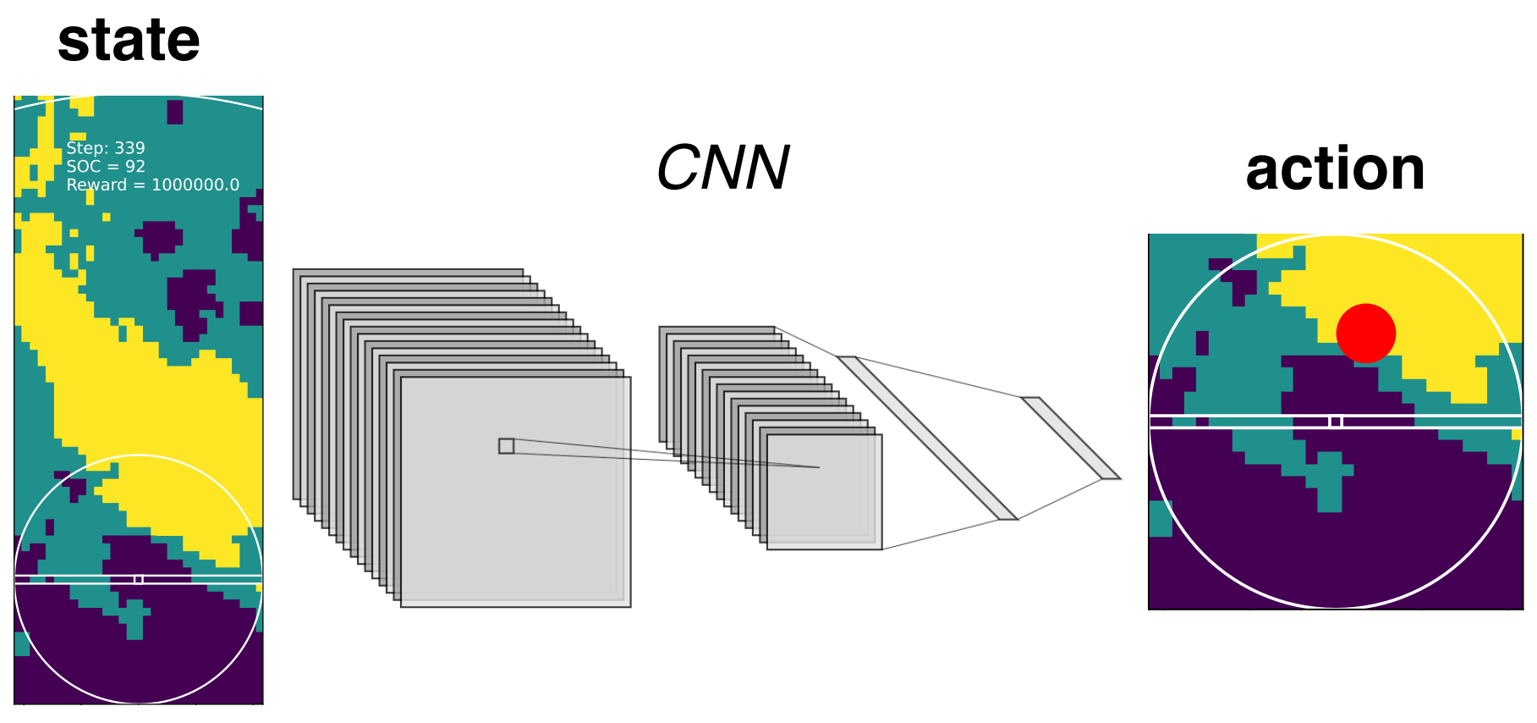

*Fig. 6. Imitation learning: behavioral cloning using deep learning to predict optimal actions from states*

## RESULTS

| Scenario | Random | Greedy | Imitation Learning | DP |
|---|---|---|---|---|
| storm hunting | 0.01 % | 0.97 % | 1.28 % | 1.46 % |
| cloud avoidance | 0.99 % | 2.61 % | 3.01 % | 3.28 % |

*Table 1. Algorithms' performance in terms of observed targets of interest*

|  | Random | Greedy | Imitation Learning | DP |
|---|---|---|---|---|
| average time (ms) | *3.9* | 20.8 | 31.52 | 1,322.75 |

*Table 2. Average computation times per timestep for each algorithm*

## CONCLUSIONS AND FUTURE WORK

Experimental results indicate that DT together with deep imitation learning is a promising approach. When comparing it against the baseline algorithms, significantly more targets of interest are observed while respecting energy constraints. Also, its performance is relatively close to optimal (~90%) while being much faster.

Future work will wrap up ongoing work using PPO. Furthermore, it will keep improving our simulation studies so they reflect each unique use case more realistically. For each different mission scenario, we plan to capture its physical costs and constraints such as instrument warm-up times, variable power consumption, slew times, on-board reaction times, and quality degradation for off-nadir measurements collected by the primary instrument. Additionally, we want to explore more sophisticated reward functions that are nonlinear and mutually dependent, especially those that model diminishing (or increasing) returns in repeated measurements from the same point or cloud.