The Dance of Data and Doers:

A Trajectory-Oriented Perspective on RL

Ben Eysenbach

10-403 Guest Lecture
Apr 23, 2020

Before getting started

- Who am I?
- You must ask questions :)

RL is the search for good experience

Typical "black box" perspective: find policy with large reward

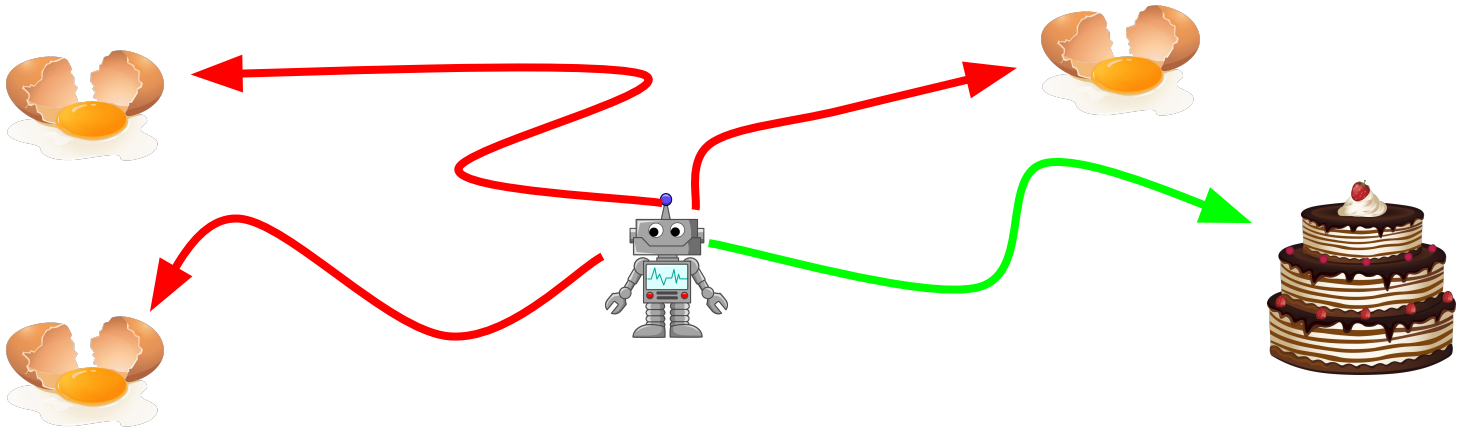
$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t} \gamma^t r(s_t, a_t) \right]$$

RL is the search for good experience

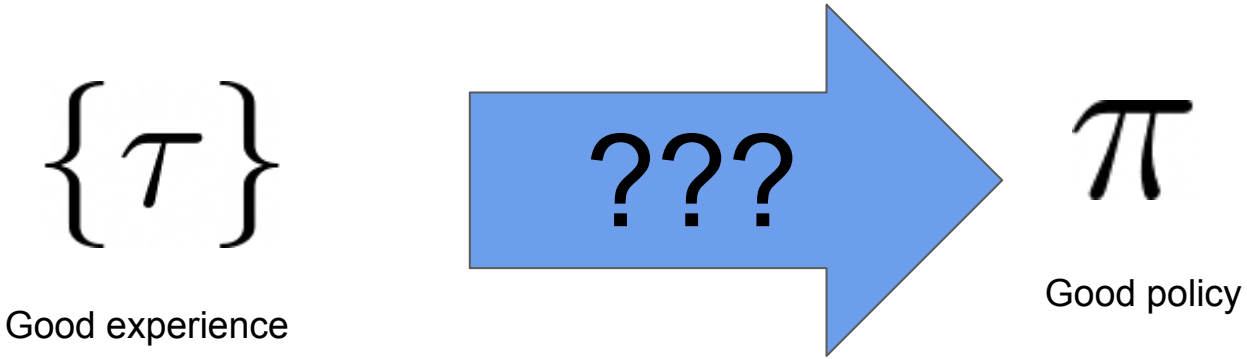
Typical "black box" perspective: find policy with large reward

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t} \gamma^t r(s_t, a_t) \right]$$

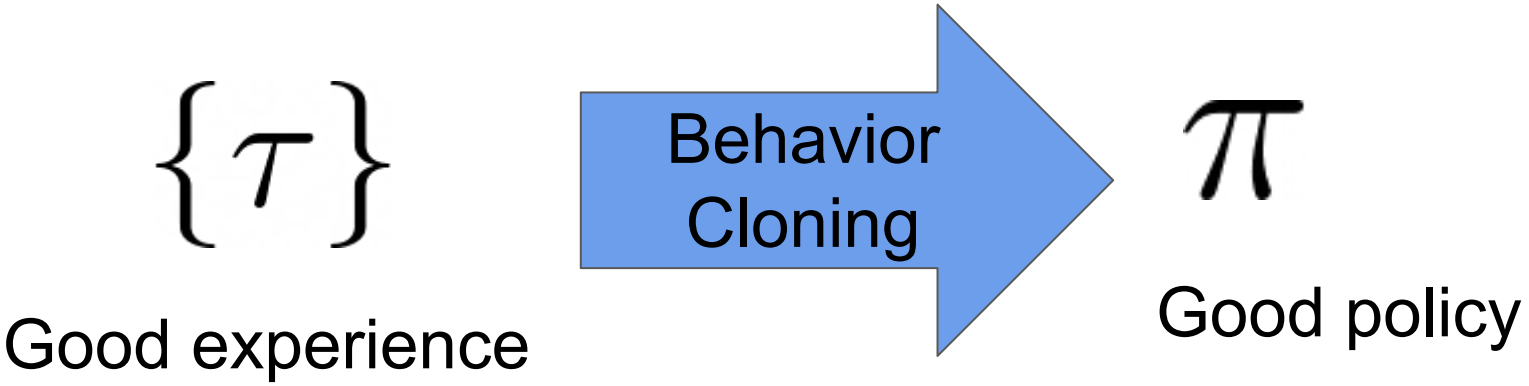
Trajectory-optimization perspective: find good experience



How to extract a policy from good experience?

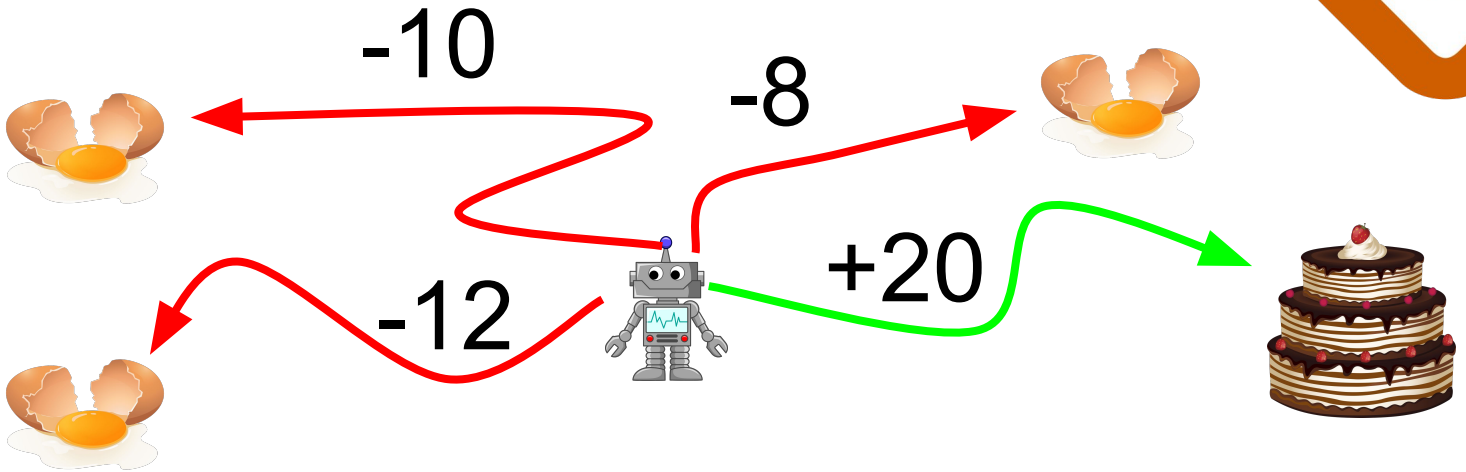


How to extract a policy from good experience?



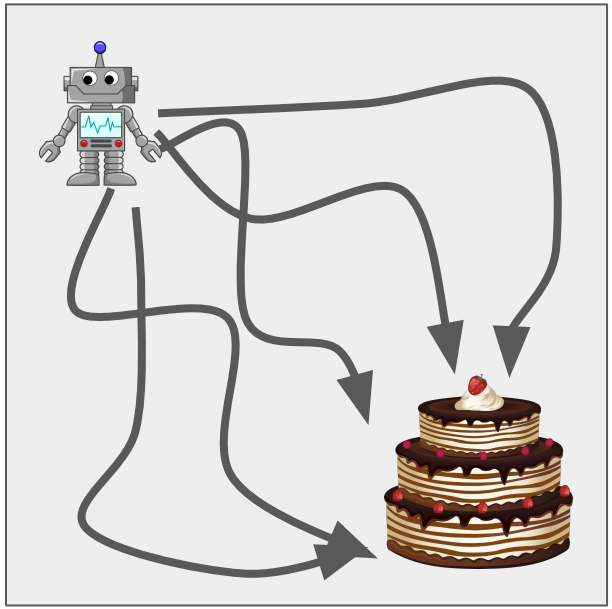
$$\max_{\theta} \mathbb{E}_{s,a\sim\{\tau\}} \left[\log \pi_\theta(a \mid s)\right]$$

Use the reward function to identify good exp

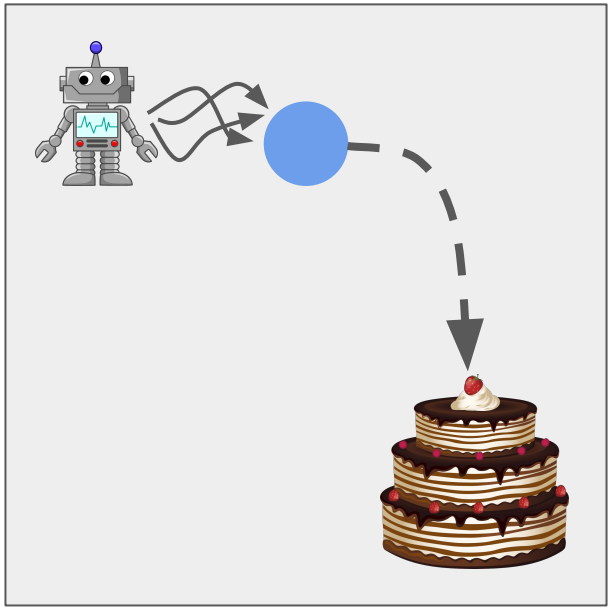


$$\max_{\tau=s_1,a_1,\cdots} R(\tau) \triangleq \sum_t \gamma^t r(s_t, a_t)$$

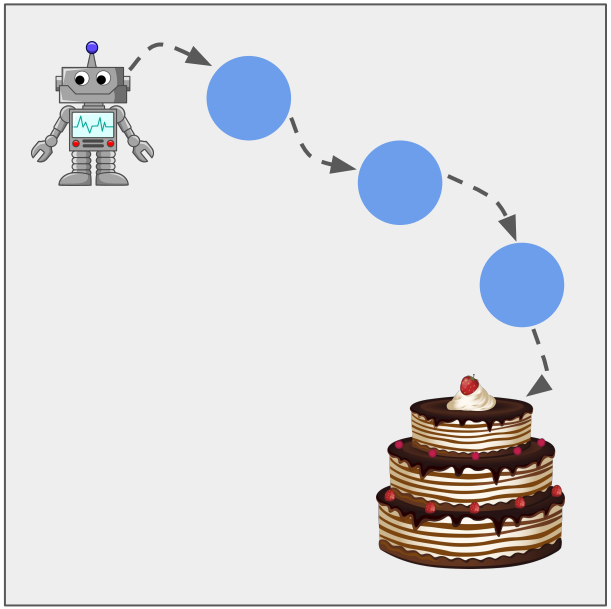
How do you find good experience?



1

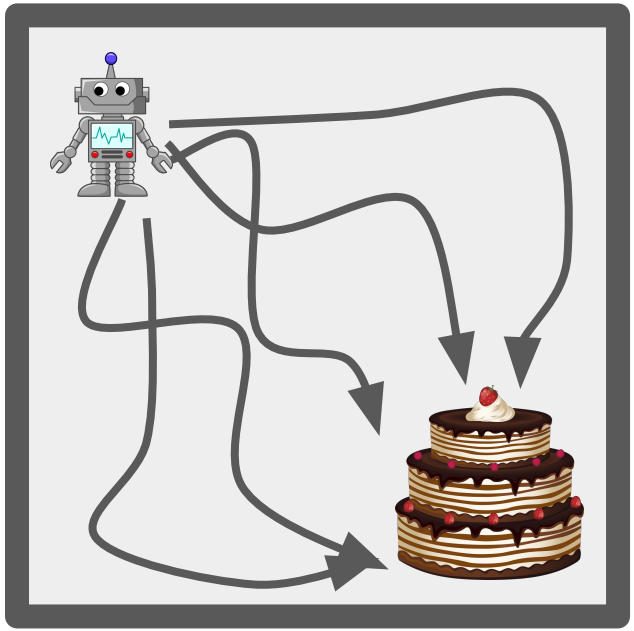


2

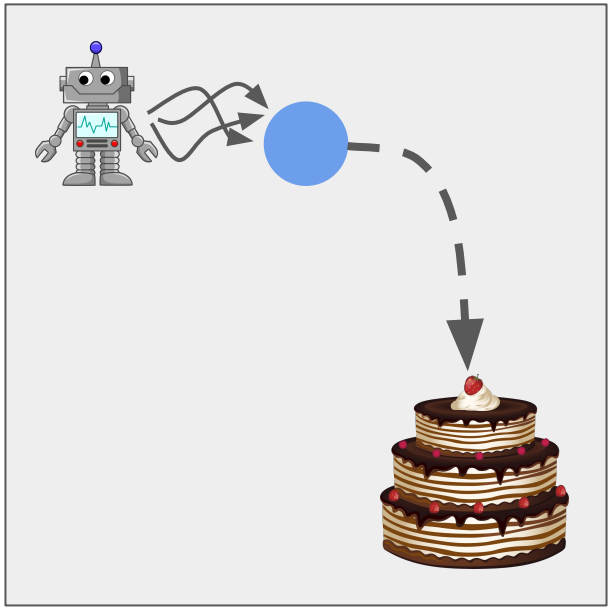


3

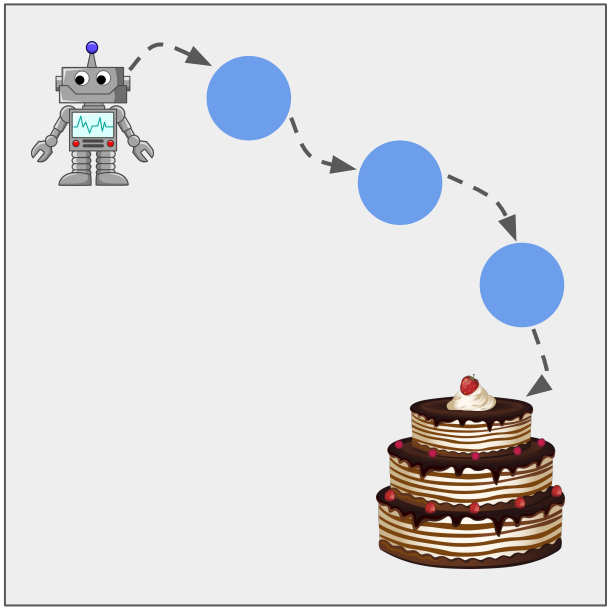
How do you find good experience?



1

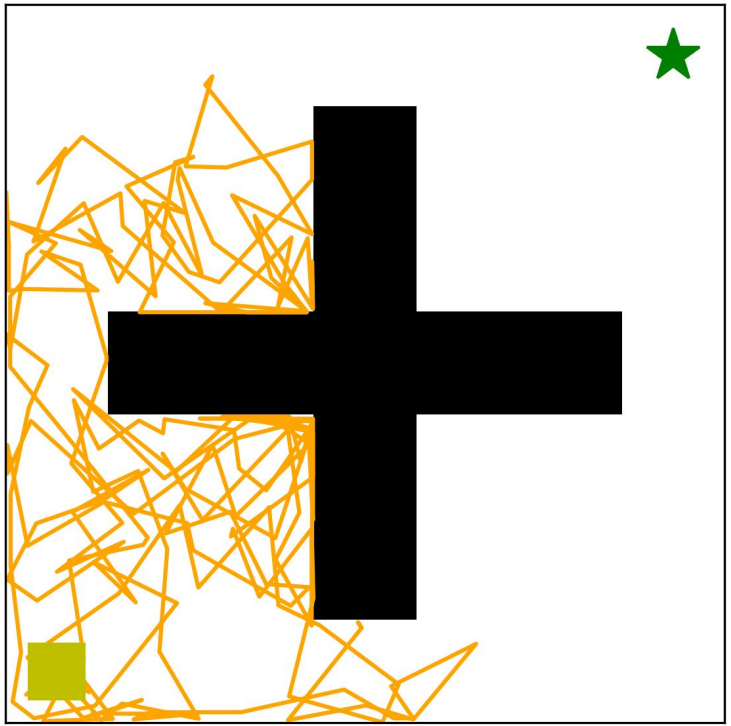


2

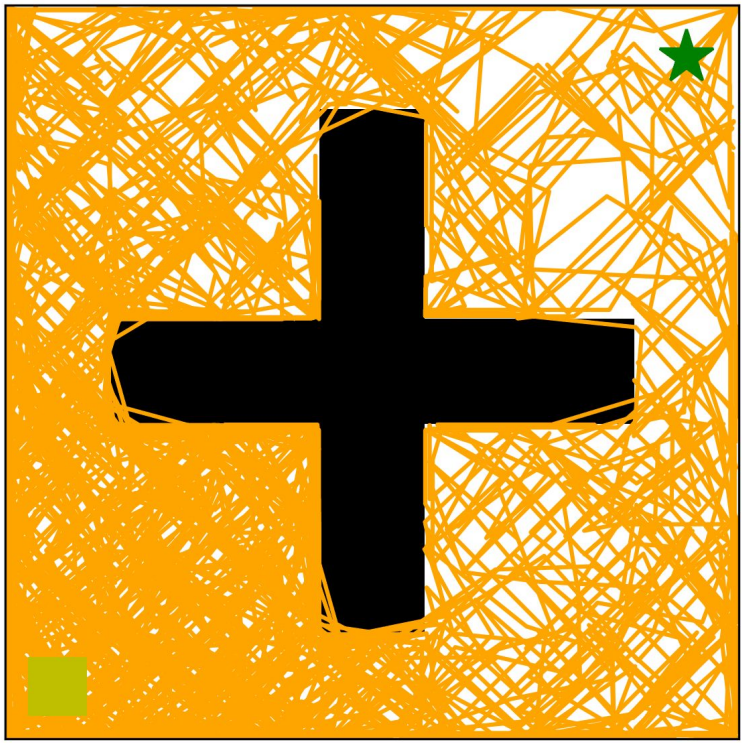


3

Optimize the entire trajectory: random search

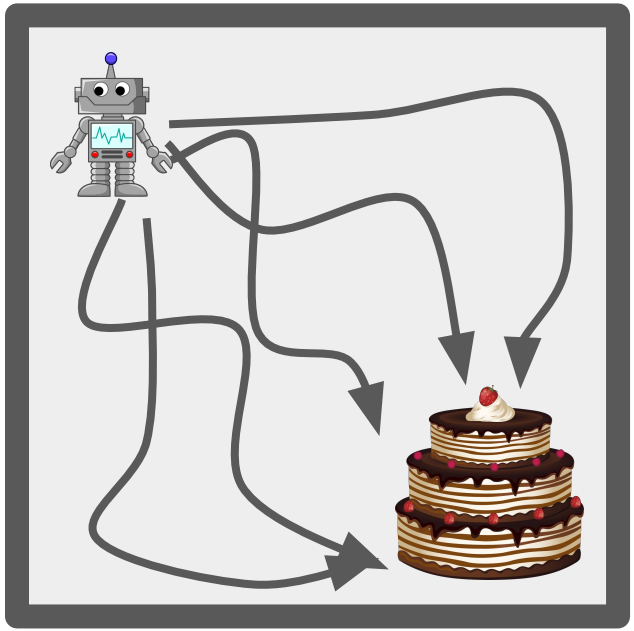


Optimize the entire trajectory: evolutionary strategies

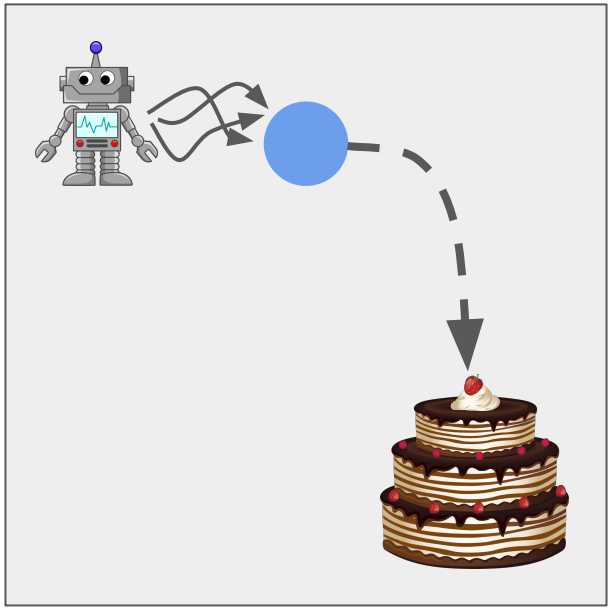


CMA-ES [Hansen 16]

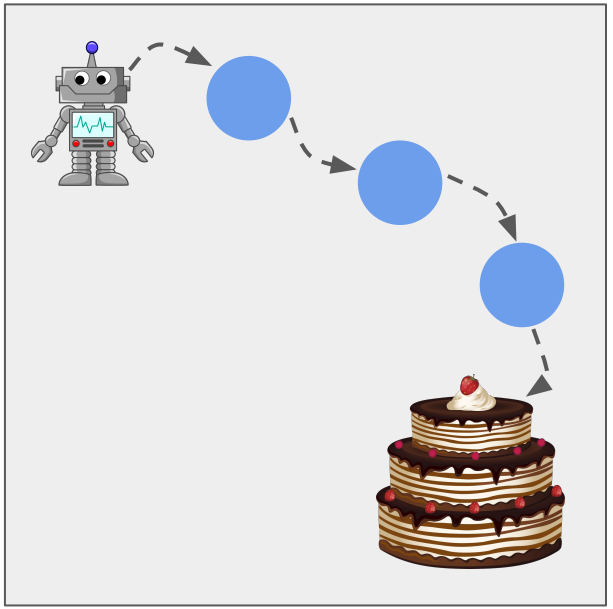
How do you find good experience?



1

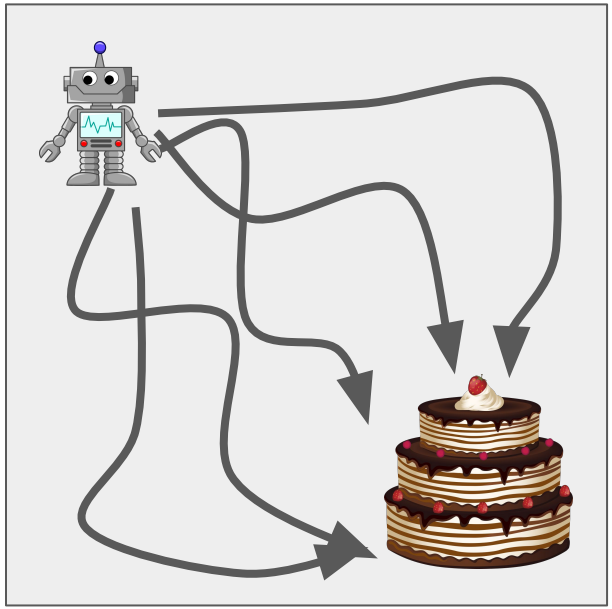


2

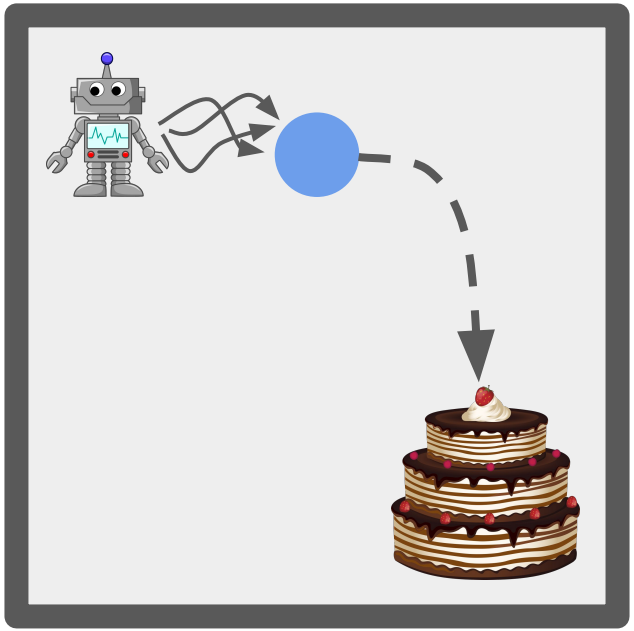


3

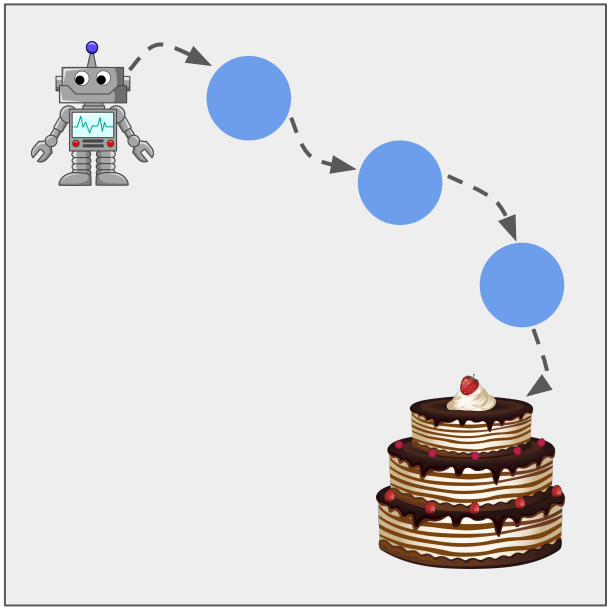
How do you find good experience?



1



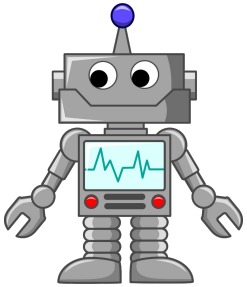
2



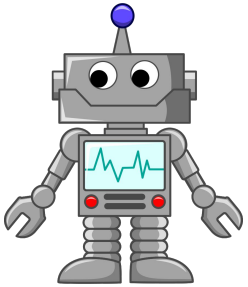
3

Optimize partial trajectory + end point

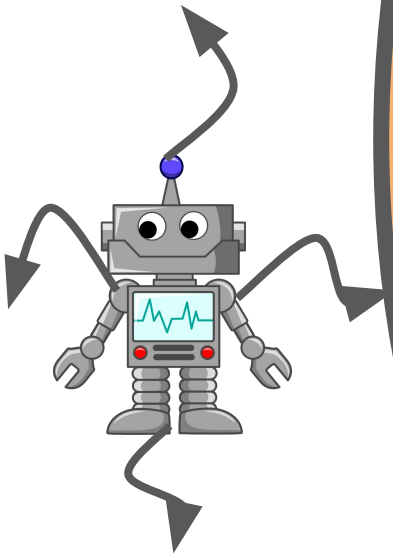
Main idea: get off to a good start, and end at a promising state



$$V(s_k)$$



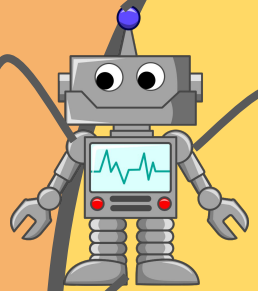
$V(s_k)$



+5

+10

+20



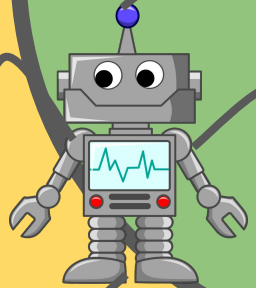
+5

+10

+20

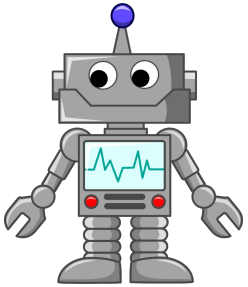
+5

+10



+20

Q-learning estimates value of states



+???

+???

+???

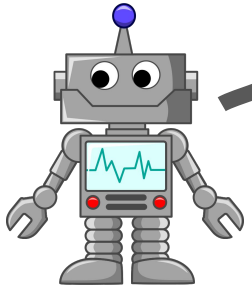
Q-learning estimates value of states

$$Q^* = r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$

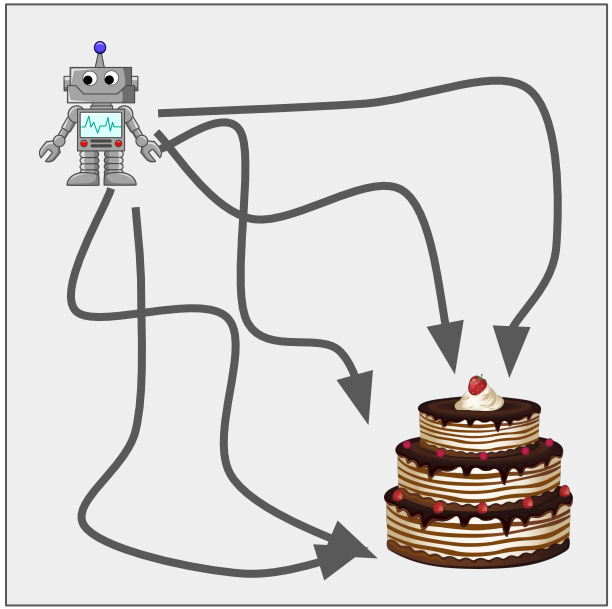
$$\min_{\theta} \left(Q_\theta(s_t, a_t) - Q^*\right)^2$$

First part of the plan

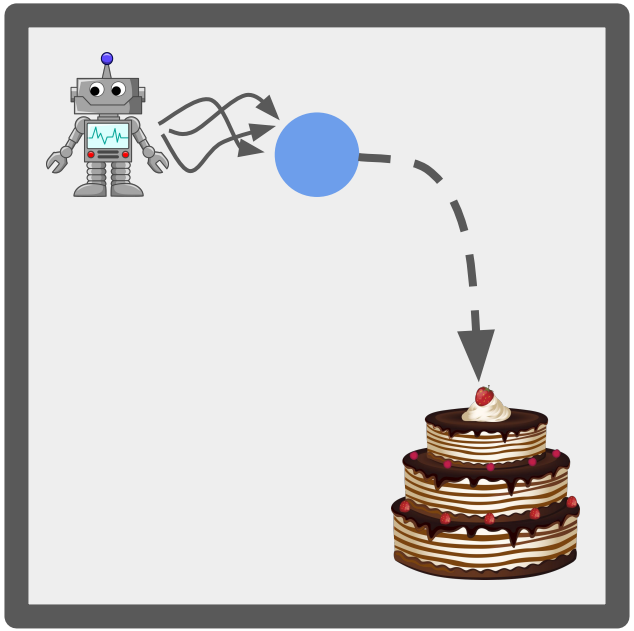
Rest of the plan



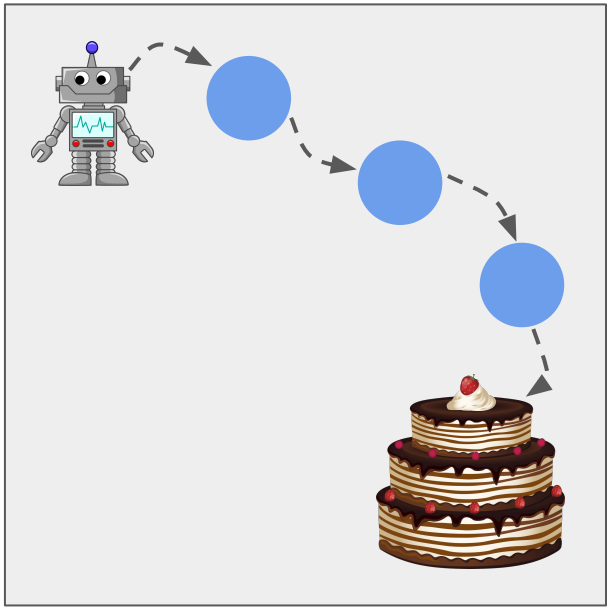
How do you find good experience?



1

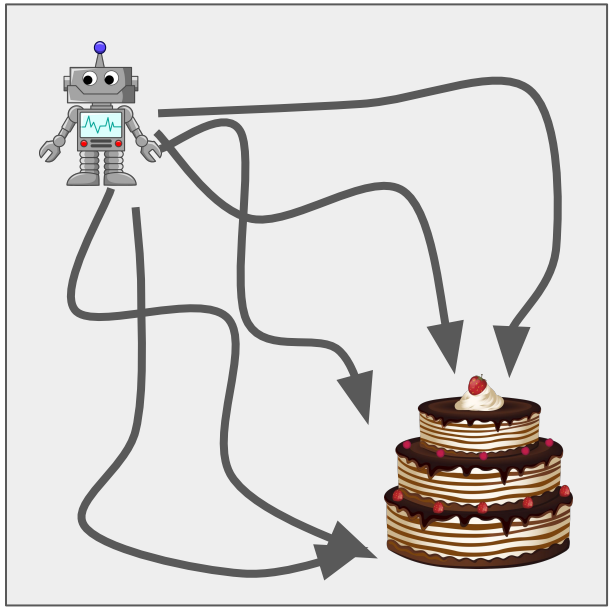


2

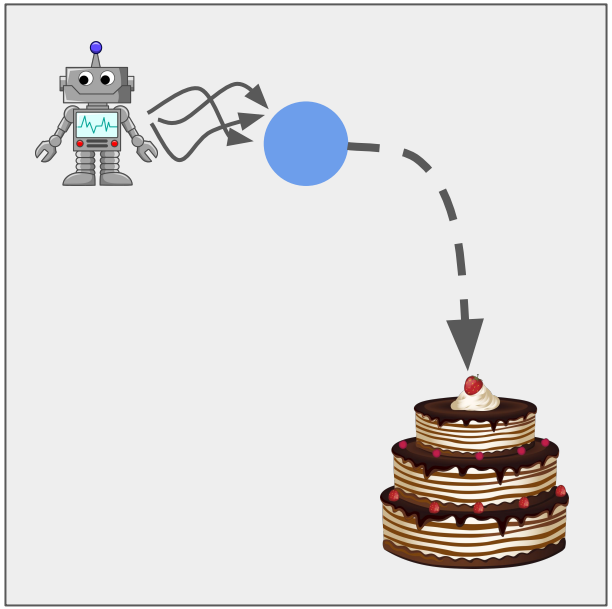


3

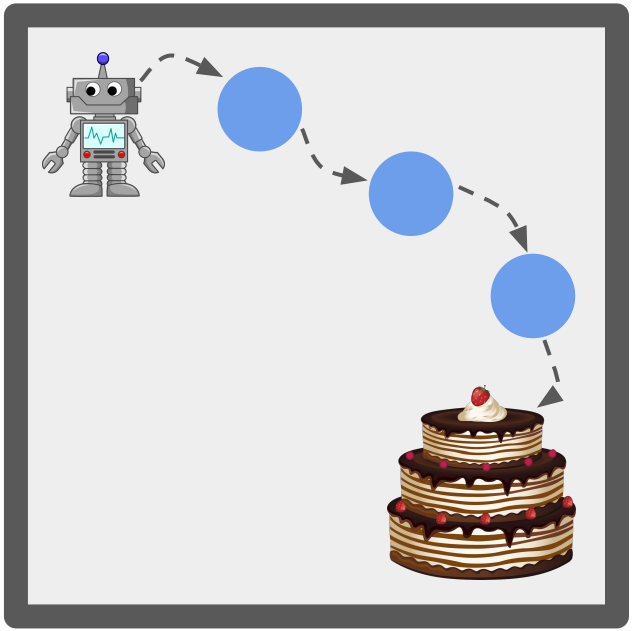
How do you find good experience?



1



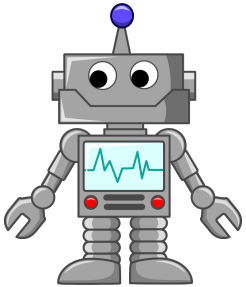
2

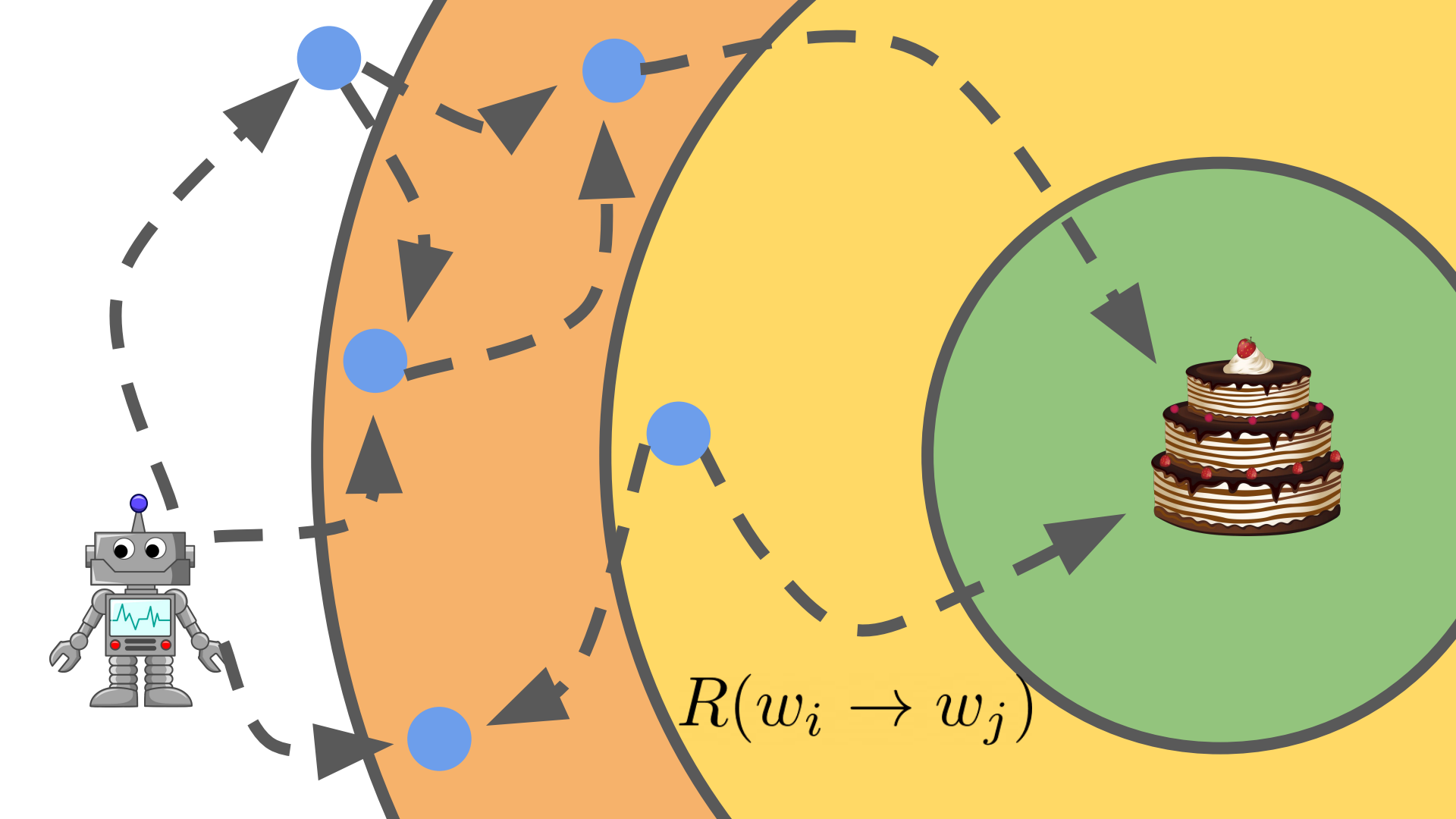


3

Optimize a coarse plan

Idea: Find a sequence of waypoints to our destination state.



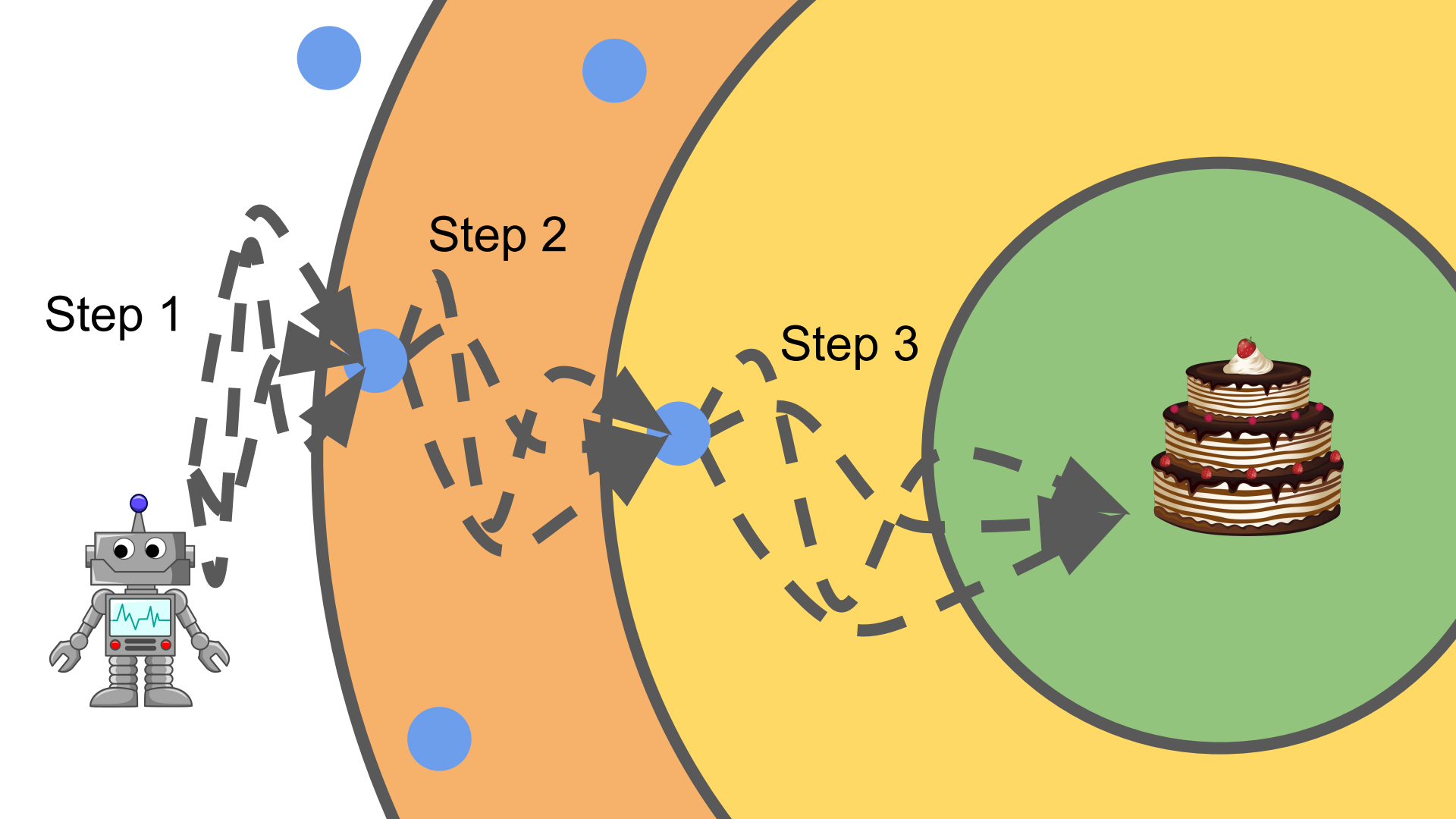

$R(w_i \rightarrow w_j)$

Optimize a coarse plan

Idea: Find a sequence of waypoints enroute to our destination state.

$$\max_{w_1,\cdots,w_k} \sum_{i=1}^{k} R(w_i \rightarrow w_{i+1})$$

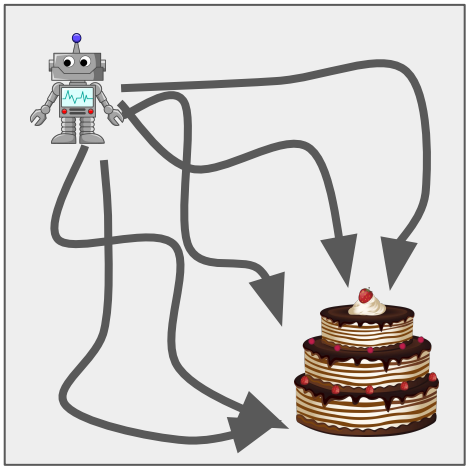
Just a shortest path problem! Solve with Dijkstra's Algorithm.

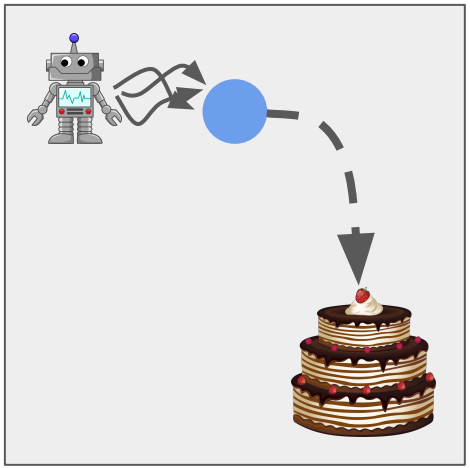

Step 1
Step 2
Step 3

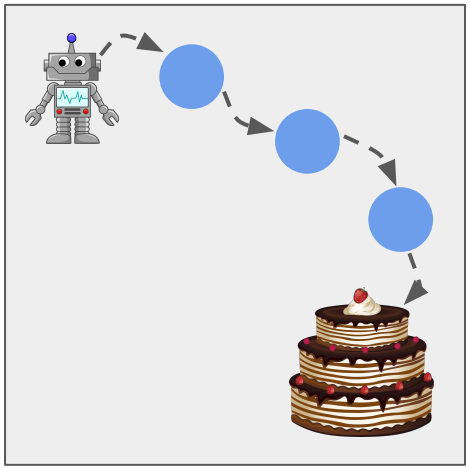
Summary

Main idea: find high-reward trajectories, learn a policy via behavior clo

Three ways to find high-reward trajectories:







Example papers: [Tassa 12, Chua 18] [Lowrey 18, Silver 17] [Savinov 18, E. 19, Nasiriany 19]