Path following with reinforcement learning for autonomous cars

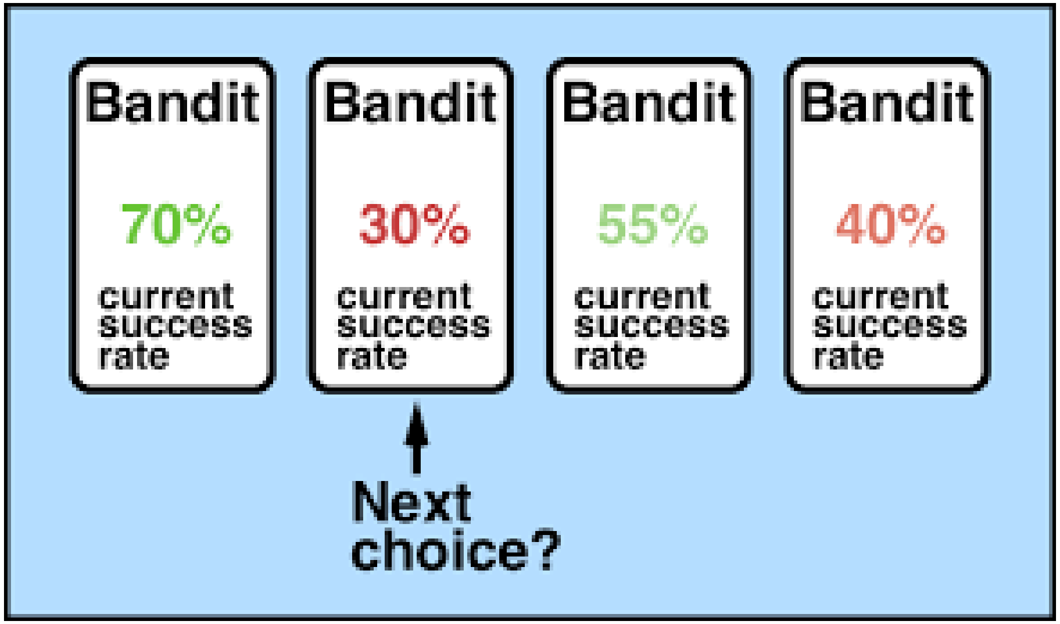
- Mozzam Motiwala (IAS)

Index

- Basics of Reinforcement Learning
- Model Based vs Model Free Reinforcement Learning
- Autonomous Car collision avoidance

What is Reinforcement Learning?

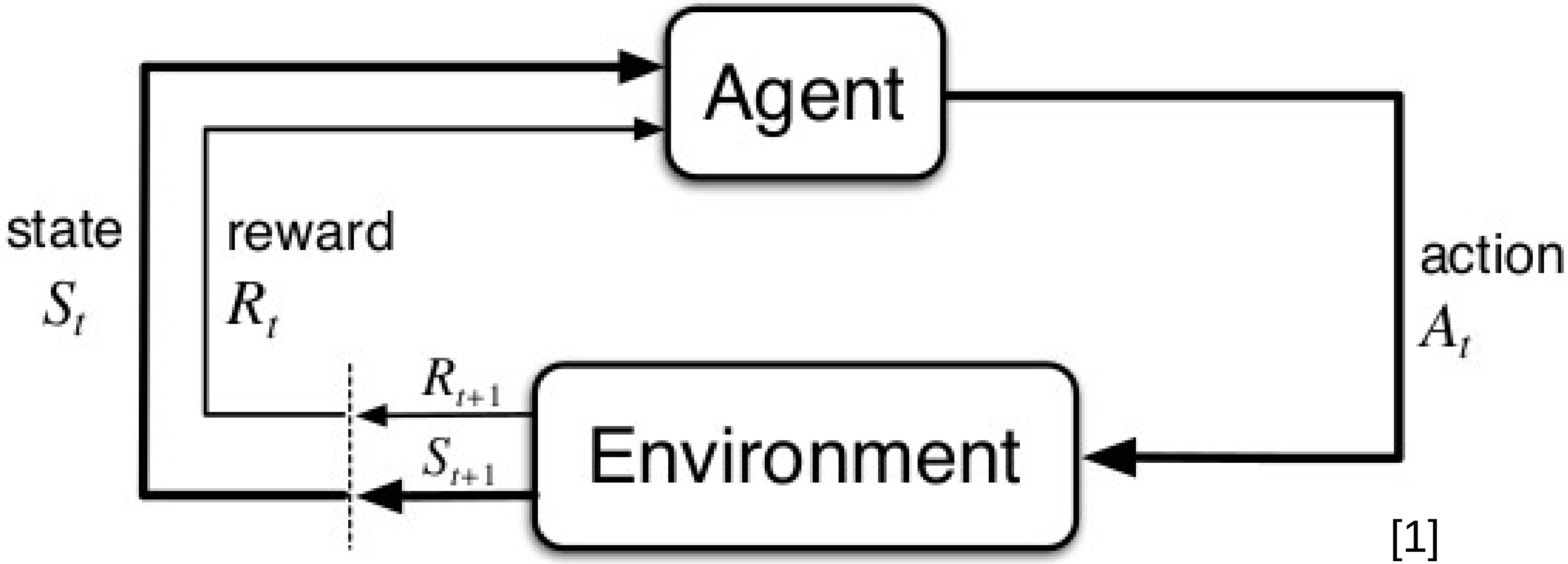
- Learning by trial and error only based on a reward signal[1]



https://towardsdatascience.com/solving-the-multi-armed-bandit-problem-b72de40db97c

Exploration vs Exploitation?

Markov-Desicion Process



[1]

Reward Function?

Transition Function?

Policy?

Optimal Policy?

Some terminalogy

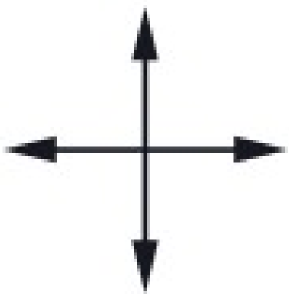
- Value Function: $v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma v_\pi(s')\Big], \quad \text{for all } s \in \mathcal{S},$$

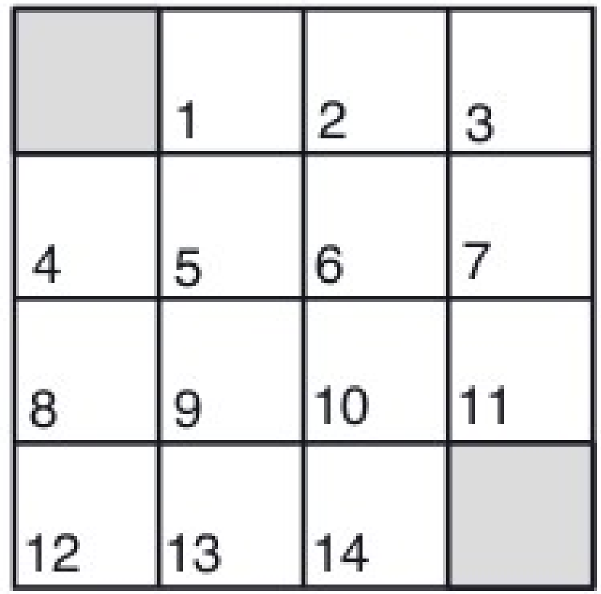
- Action Value Function: $q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$

Why Discounting Factor?

Gridworld



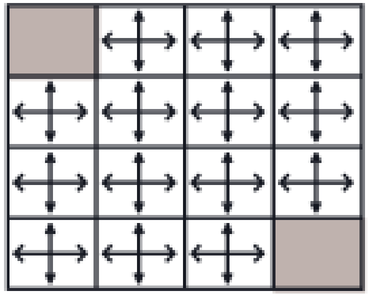
actions



$$R_t = -1$$

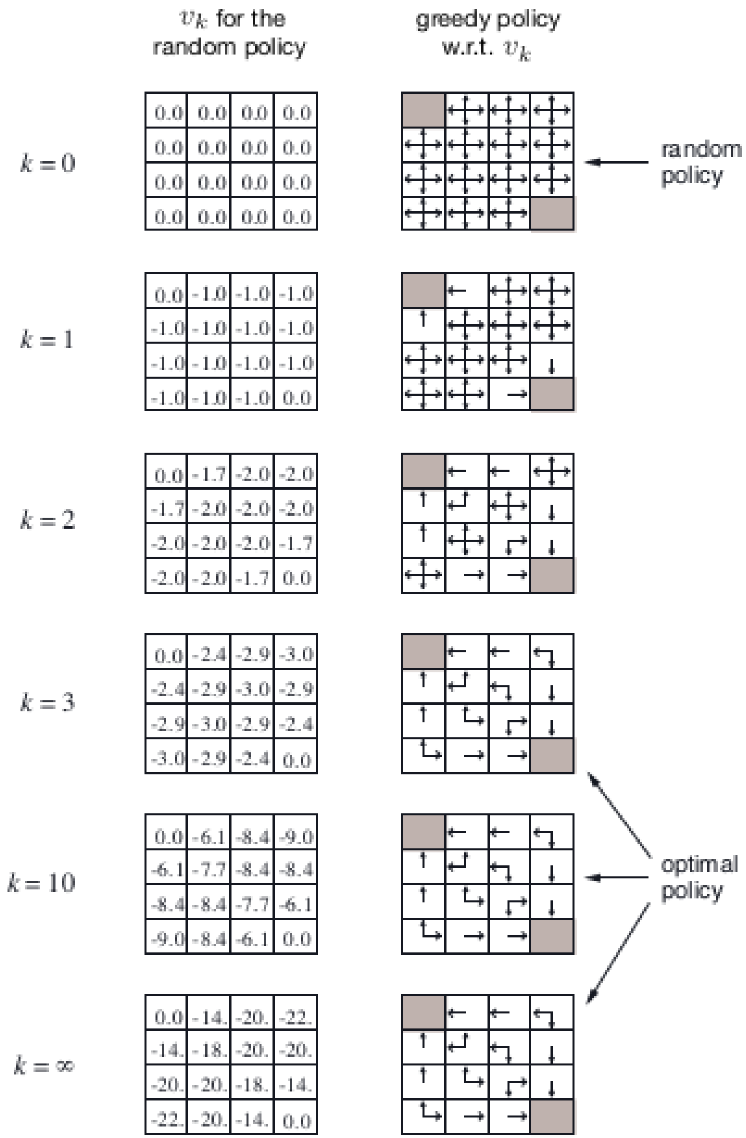
on all transitions

[1]

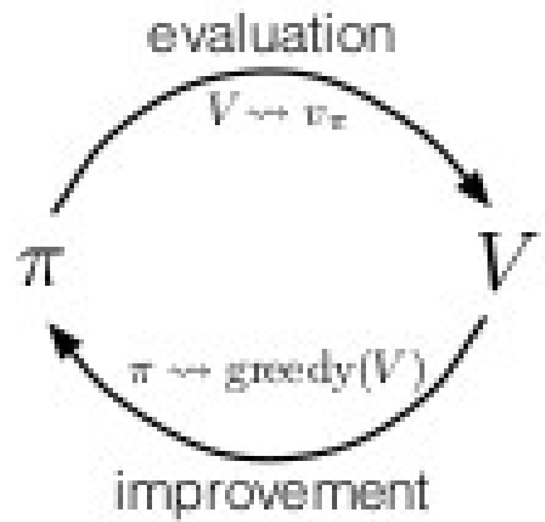


random policy

Finding Optimal Policy



[1]



Cart Pole Balancing Problem

```
Observation:
    Type: Box(4)
    Num     Observation               Min          Max
    0       Cart Position             -4.8         4.8
    1       Cart Velocity             -Inf         Inf
    2       Pole Angle                 -24°        24°
    3       Pole Velocity At Tip

Action:
    Type: Discrete(2)
    Num     Action
    0       Push cart to the left
    1       Push cart to the right
```

https://towardsdatascience.com/cartpole-introduction-to-reinforcement-learning-ed0eb5b58288

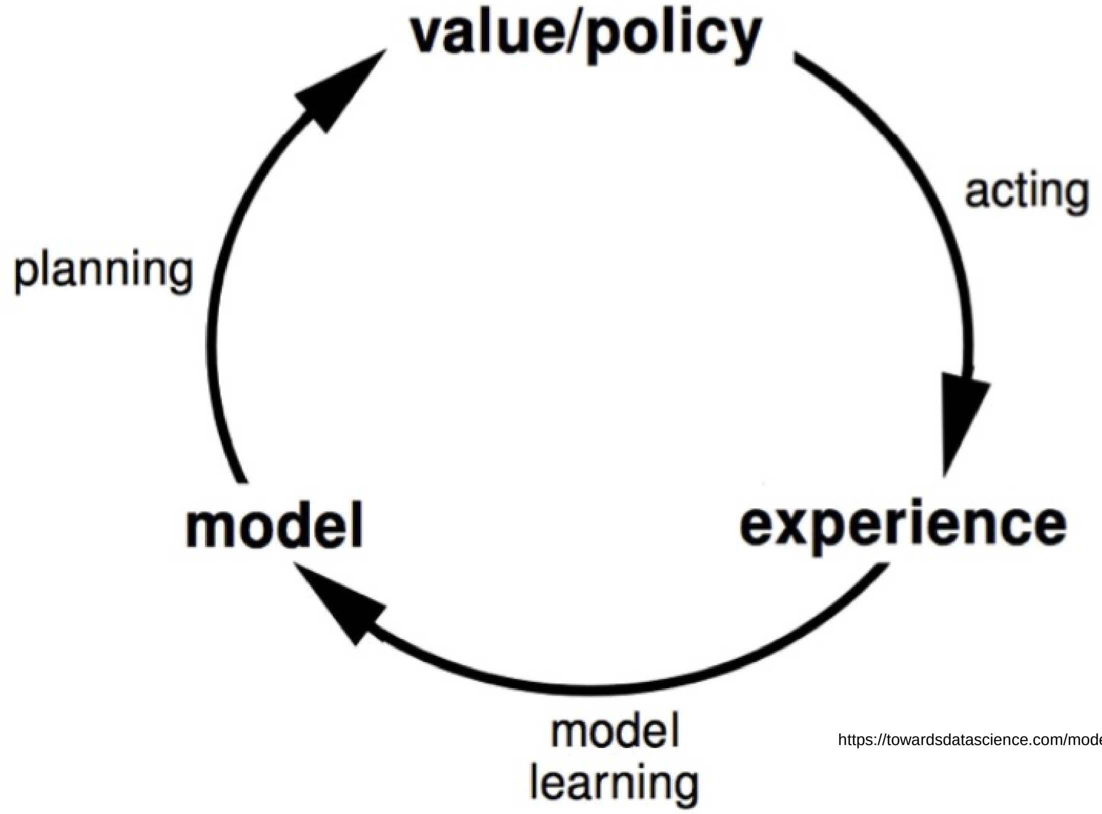
https://www.youtube.com/watch?v=Lt-KLtkDlh8

Index

- Basics of Reinforcement learning
- Model Based vs Model Free Reinforcement Learning
- Autonomous Car collision avoidance

Model-based

By a model of the environment we mean anything that an agent can use to predict how the environment will respond to its actions[2].



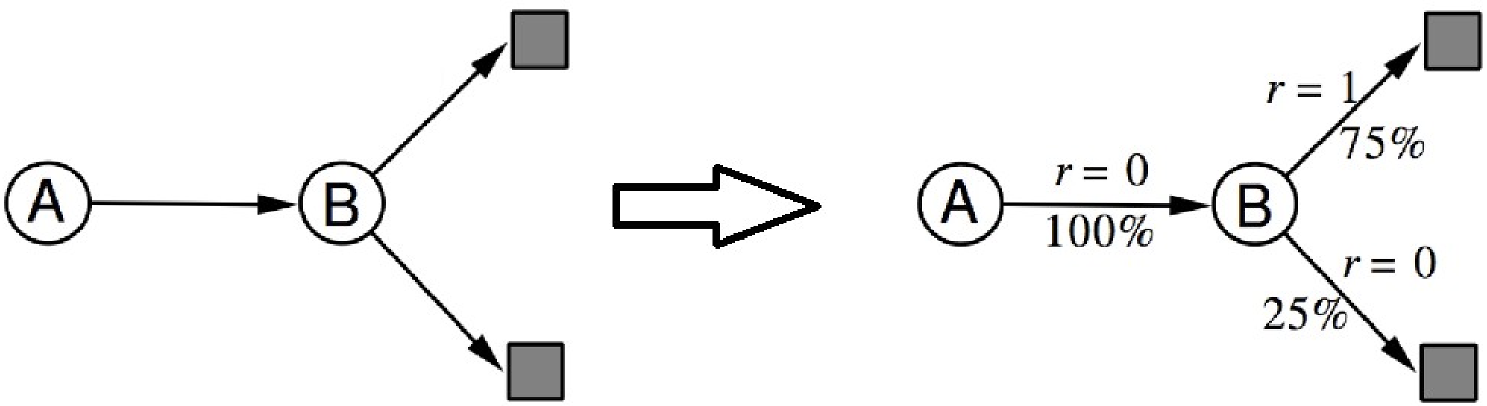
https://towardsdatascience.com/model-based-reinforcement-learning-cb9e41ff1f0d

Example

https://towardsdatascience.com/model-based-reinforcement-learning-cb9e41ff1f0d

Two states A, B; no discounting; 8 episodes of experience

A, 0, B, 0
B, 1
B, 1
B, 1
B, 1
B, 1
B, 1
B, 0



Whats Next? :: Now lets sample from it to adjust the policy..

Why model-based RL?

Reduced number of interaction with the real environment while learning.

Types: Neural Network Model, Guassian Process Model.. etc

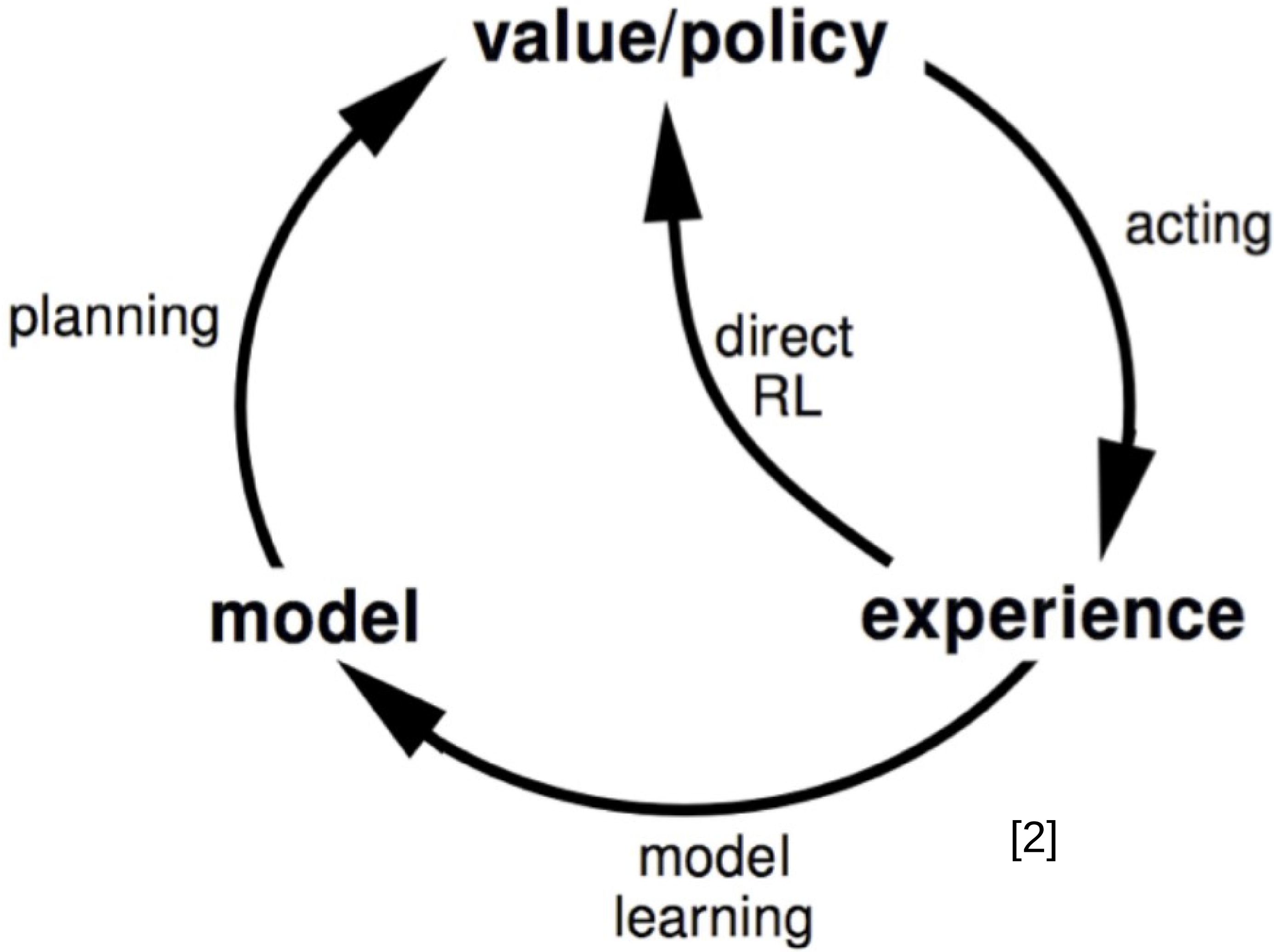
Advantages?

- Fast
- Need less data

Problems?

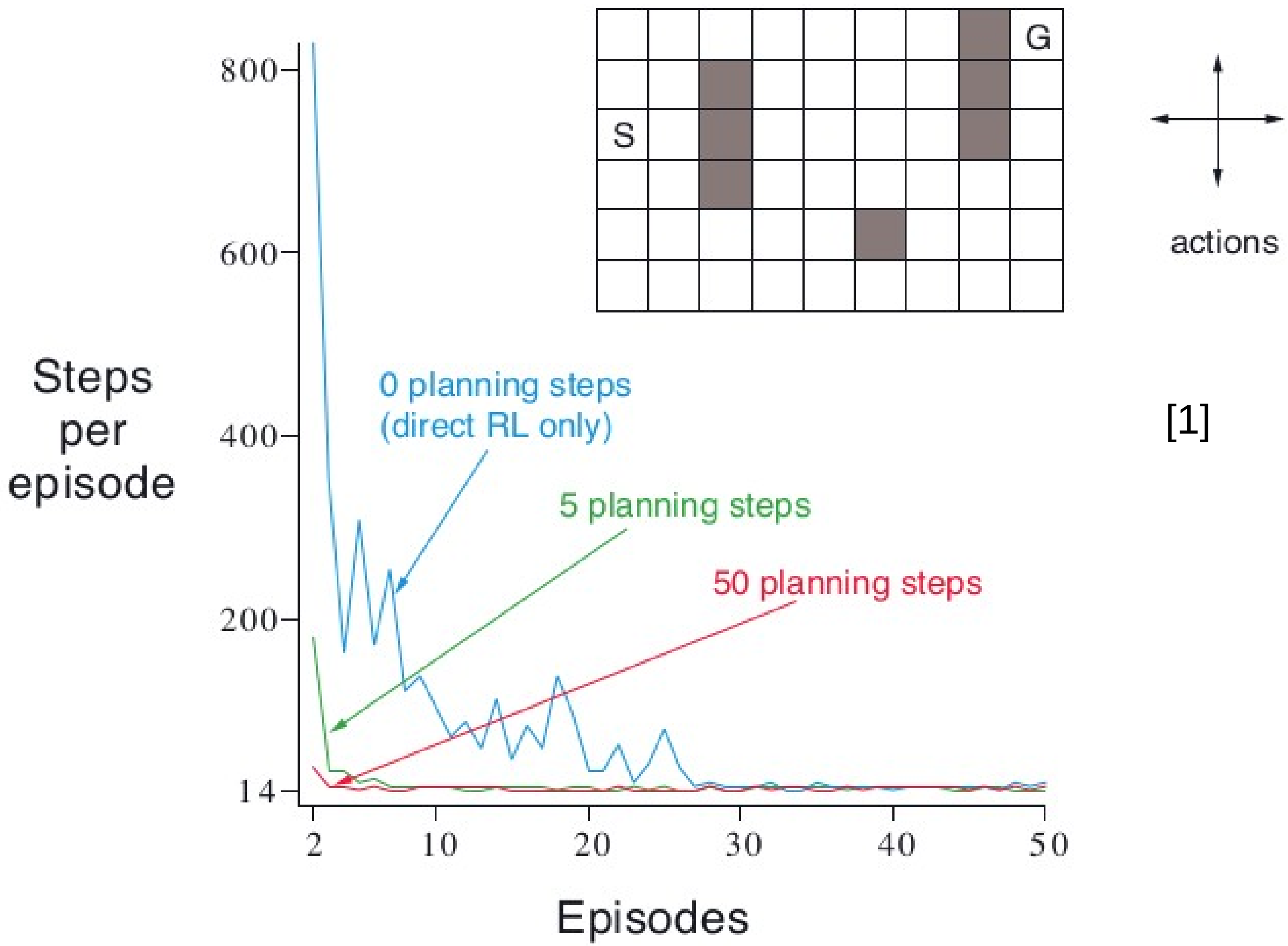
- What if the model is wrong?

Model Based+ Model Free



[2]

Results


S
G
actions
Steps
per
episode
800
600
400
200
14
2
10
20
30
40
50
Episodes
0 planning steps
(direct RL only)
5 planning steps
50 planning steps

[1]

Why better result?

Tabular Dyna-Q

Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$

Loop forever:

(a) $S \leftarrow$ current (nonterminal) state

(b) $A \leftarrow \varepsilon$-greedy(S, Q)

(c) Take action A; observe resultant reward, R, and state, S'

(d) $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

(e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment)

(f) Loop repeat n times:

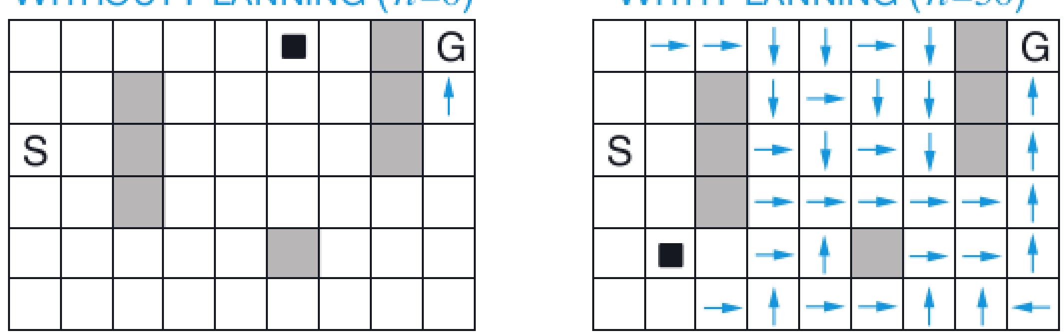
  $S \leftarrow$ random previously observed state

  $A \leftarrow$ random action previously taken in S

  $R, S' \leftarrow Model(S,A)$

  $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

WITHOUT PLANNING (n=0)   WITH PLANNING (n=50)



[1]

Index

- Basics of Reinforcement learning
- Model Based vs Model Free Reinforcement Learning
- Autonomous Car Collision Avoidance

Application: Autonomous Car

Why Reinforcement Learning?

Problem with traditional methods

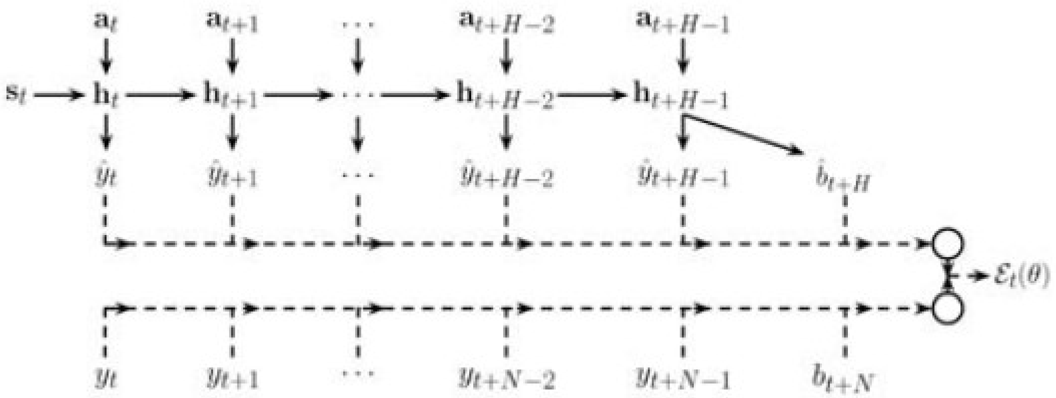
- Slow
- Assumptions

Learning in RL

- Adapting to environment
- Learning from mistakes

Generalized Computation Graph

Self-supervised Deep Reinforcement Learning with Generalized Computation Graphs(GCG) for Robot Navigation[3]



[3]

- H=1 : Model-Free
- H= N (Length of Episode): Model-Based

Model Details

- Deep RNN as Model
- Model output 1= Current Reward ŷ: Robots speed
- Model output 2= Future Value to go(value of the state) ^b: Distance travelled before collision
- Policy Evaluation Function :

$$J(\mathbf{s}_t, \mathbf{A}_t^H) = \sum_{h=0}^{H-1} -\hat{y}_{t+h} - \hat{b}_{t+H}.$$

- Policy Evaluation by sampling k random action sequence and selecting the one with max reward.

GCG : Algorithm

Algorithm 1 Reinforcement learning with generalized computation graphs

input: computation graph $G_\theta(\mathbf{s}_t, \mathbf{A}_t^H)$, error function $\mathcal{E}_t(\theta)$, and policy evaluation function $J(\mathbf{s}_t, \mathbf{A}_t^H)$
initialize dataset $\mathcal{D} \leftarrow \emptyset$
for $t = 1$ to T **do**
    get current state $\mathbf{s}_t$
    $\mathbf{A}_t^H \leftarrow \arg\max_{\mathbf{A}} J(\mathbf{s}_t, \mathbf{A})$
    execute first action $\mathbf{a}_t$
    receive labels y_t and b_t
    add $(\mathbf{s}_t, \mathbf{a}_t, y_t, b_t)$ to dataset $\mathcal{D}$
    update G_θ by $\theta \leftarrow \arg\min_\theta \mathcal{E}_{t'}(\theta)$ using $\mathcal{D}$
end for

[3]

Evaluation and Results

https://www.youtube.com/watch?v=NlFbLVG6LpA

Distance until crash (m)	Random policy	Double Q-learning with off-policy data	Our approach
Mean	3.4	7.2	52.4
Median	2.8	6.1	29.3
Max	8.0	21.5	197.0

TABLE I: Evaluation of our learned policy navigating at 1.2m/s using only monocular images in a real-world indoor environment after 4 hours of self-supervised training, compared to a random policy and double Q-learning trained with the same data gathered by our approach.

[3]

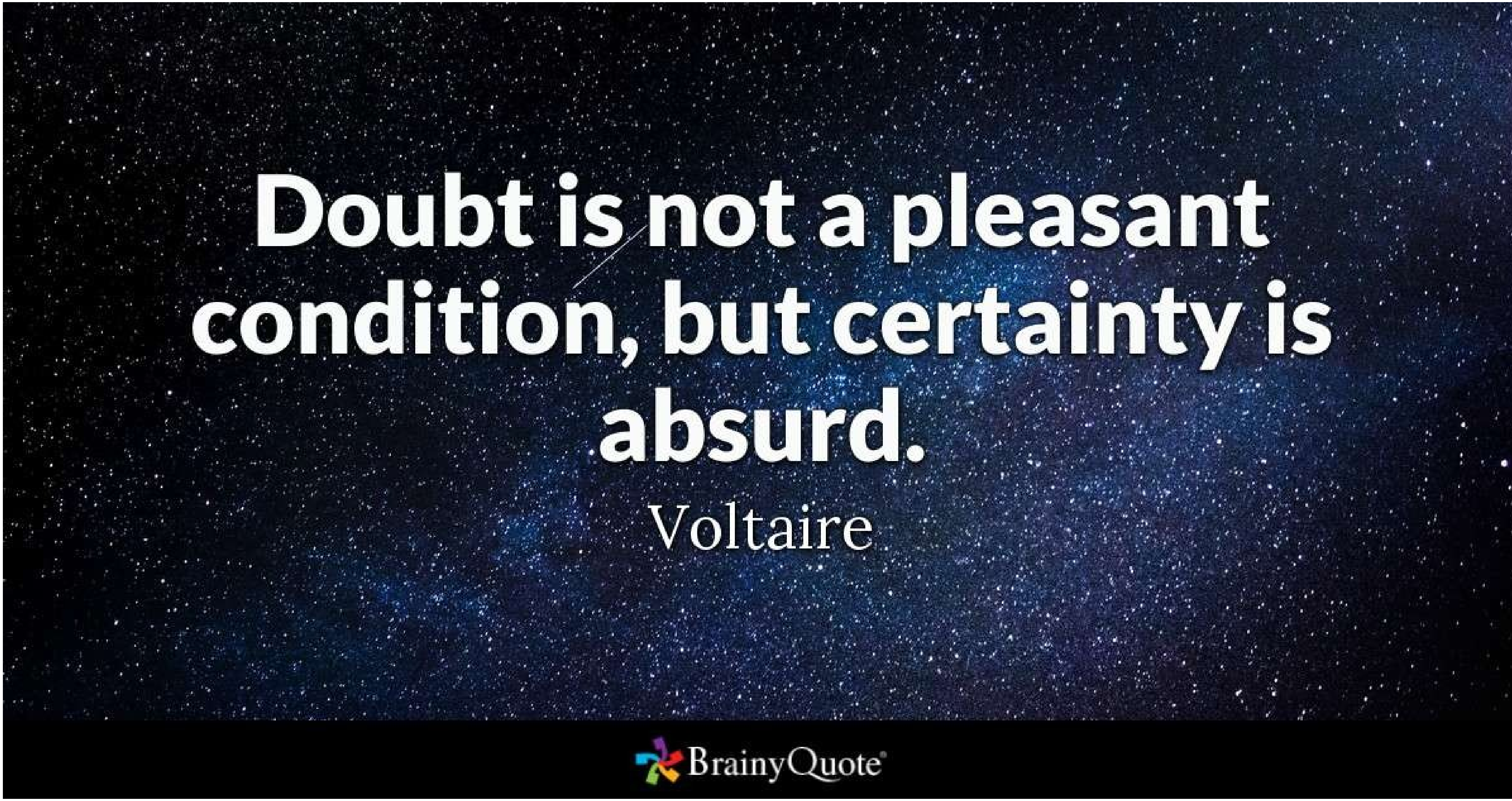
Summary

- Benefits of Reinforcement Learning
- Model-Free vs Model-Based
- Combined approach that subsumes Model-free and Model-based

References

1. R. Sutton and A. Barto, Reinforcement Learning: An Introduction
2. R. Sutton, “Dyna, an Integrated Architecture for Learning, Planning,and Reacting,” in AAAI, 1991.
3. G. Kahn, A. Villaflor, B. Ding, P. Abbeel, and S. Levine. Self-Supervised Deep ReinforcementLearning with Generalized Computation Graphs for Robot Navigation. InIEEE InternationalConference on Robotics and Automation, 2018.

Question?


Doubt is not a pleasant
condition, but certainty is
absurd.
Voltaire
BrainyQuote