

# Reinforcement Learning

---

## Course Wrap-Up

Christopher Mutschler

# Course Wrap-Up

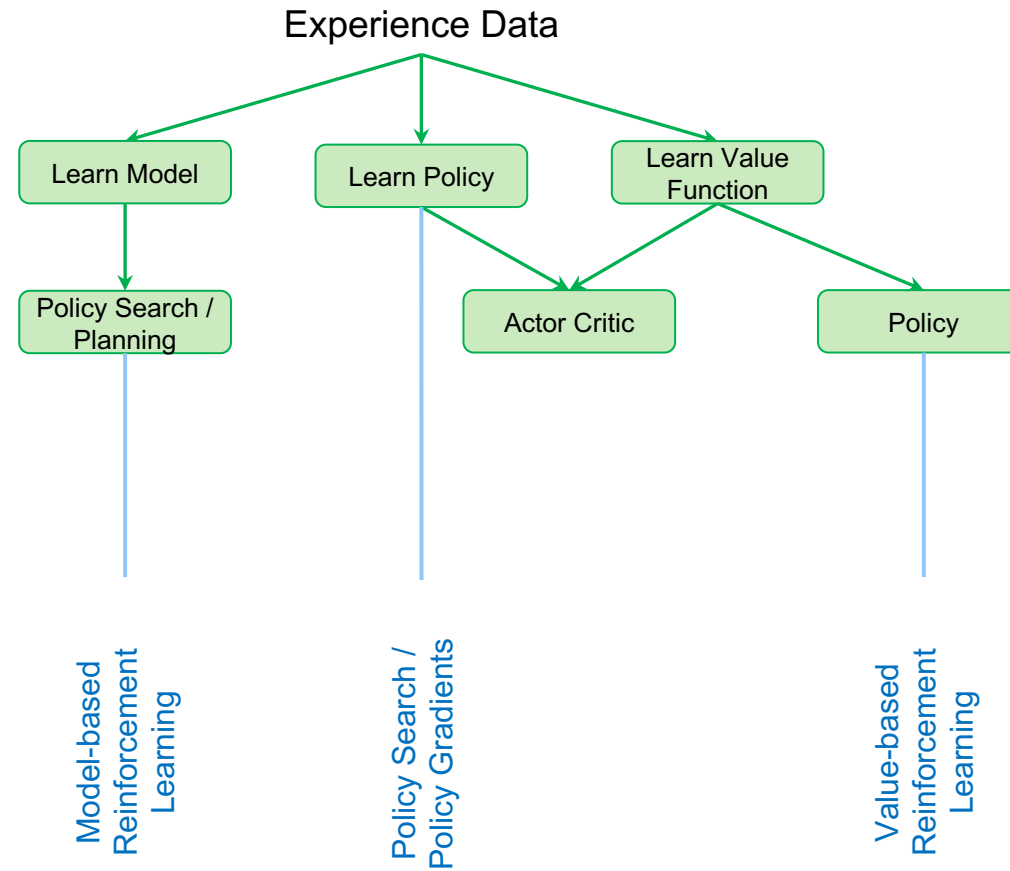
## Agenda

---

- Discussion of Exercise #12: BCQ
- Course WrapUp
- Course Evaluation
- Q&A on exam
- Guest Lecture: ChatGPT (Georgios Kontes)

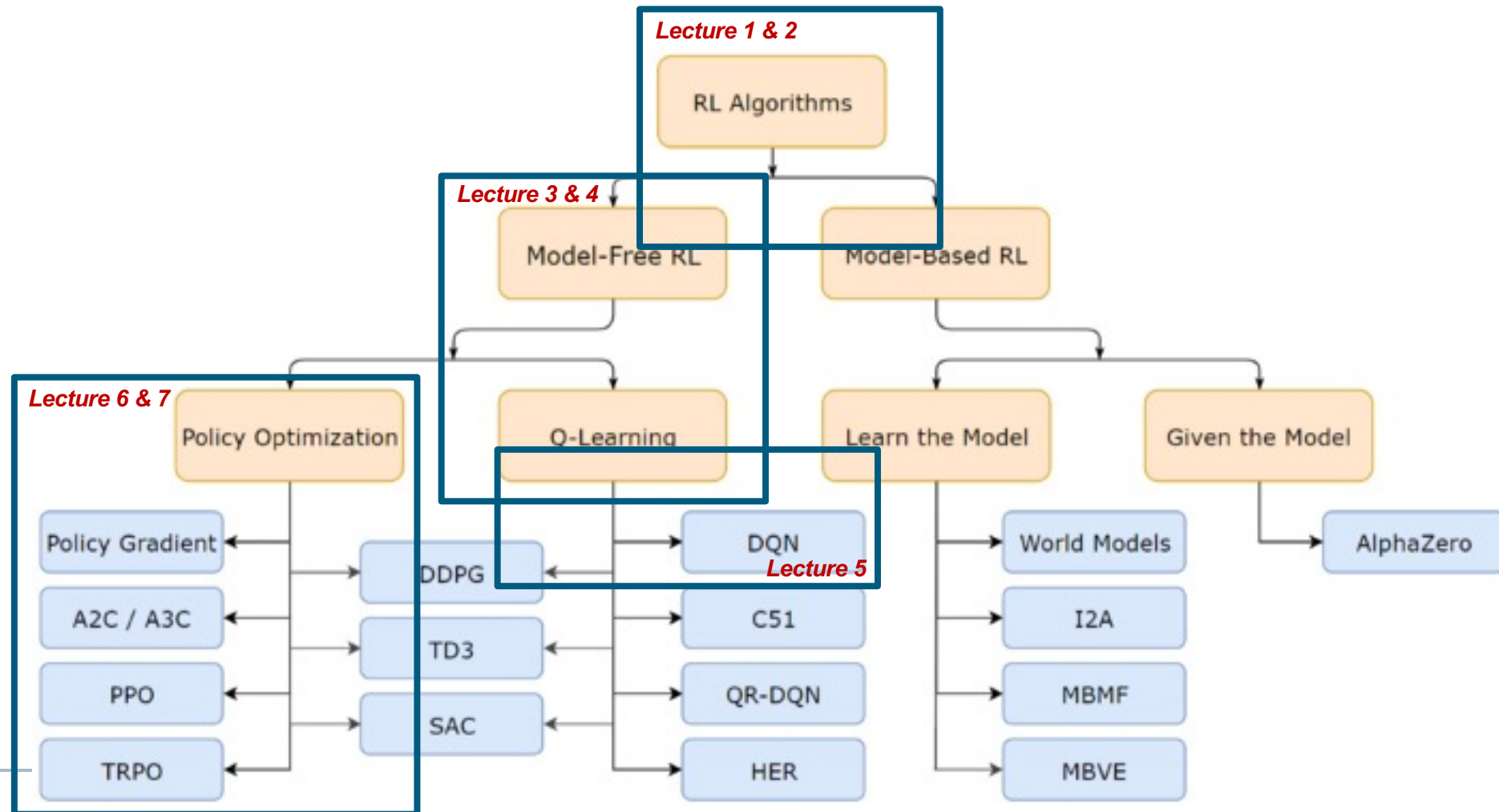
# Course Wrap-Up

## Summary of Content



# Course Wrap-Up

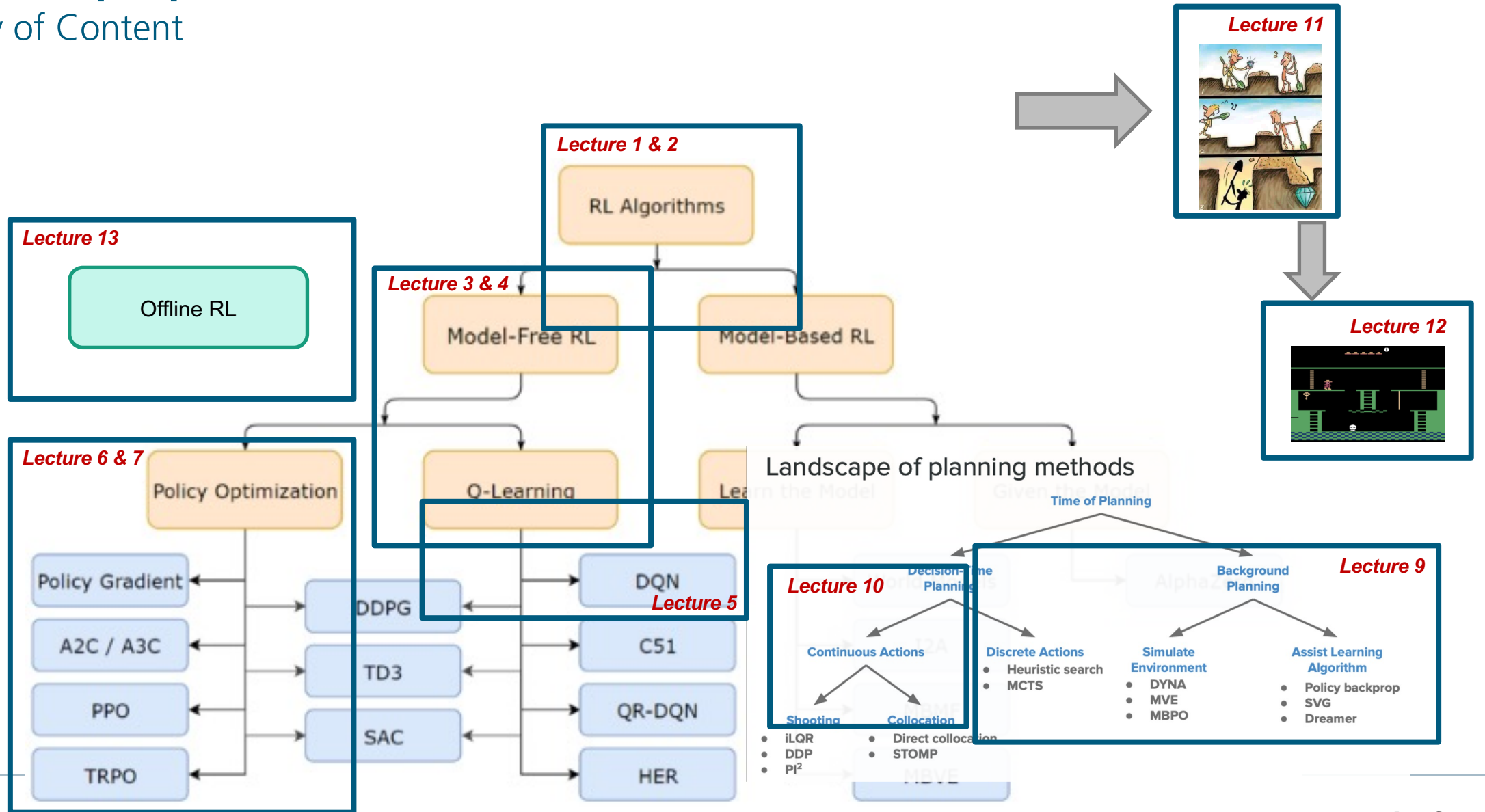
## Summary of Content



<https://smartlabai.medium.com/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc>

# Course Wrap-Up

## Summary of Content



<https://smartlabai.medium.com/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc>

# Course Wrap-Up

## Agenda

---

- Discussion of Exercise #12: BCQ
- Course WrapUp
- **Course Evaluation**
- Q&A on exam
- Guest Lecture: ChatGPT (Georgios Kontes)

# Course Wrap-Up

## Course Evaluation: Lecture

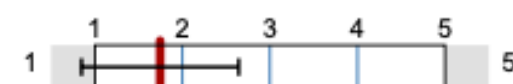
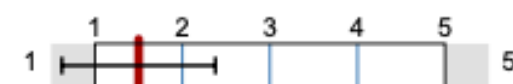
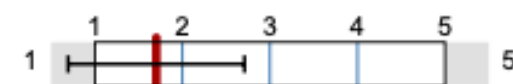
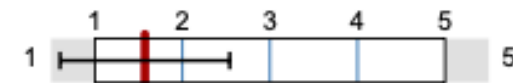
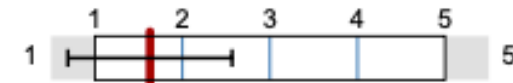
### Globalindikator

Indikator • Organisation, Inhalte und Kompetenzen der LV (Kap. 3)

Indikator • Struktur der LV (Kap. 4)

Indikator • Durchführung der LV (Kap. 5)

Indikator • Zufriedenheit und Kompetenzerwerb (Kap. 6)



→ [Open PDF](#)

# Course Wrap-Up

## Agenda

---

- Discussion of Exercise #12: BCQ
- Course WrapUp
- Course Evaluation
- **Q&A on exam**
- Guest Lecture: ChatGPT (Georgios Kontes)



# Course Wrap-Up

Q&A on the exam

## 2. Bellman Equations & Value Functions (15 Points)

(a) **Bellman Expectation Equation:** Define the recursive definition of value function  $V^\pi(s)$ , given a policy  $\pi$ .

$$V^\pi(s) =$$

(c) Define the three different variants of states we might deal with in MDPs.

- i. ....
- ii. ....
- iii. ....

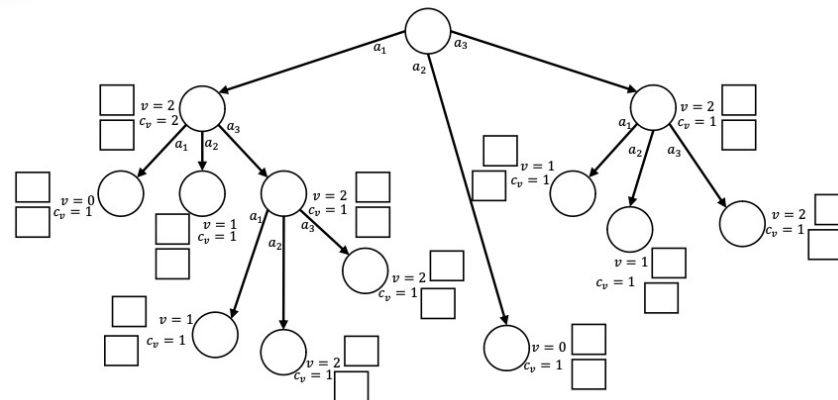
- 25% of total points are multiple choice
- 2 Programming tasks
- Basic formulas should be known & need to be applied
- Central objective functions should be known

(b) Consider the current state of the tree below.

- at each node there are three different actions available, i.e.,  $a_1$ ,  $a_2$  and  $a_3$ .
- $c_v$  denotes the current visitation count at the node
- $v$  denotes the current value estimate of the node
- select actions according to UCB with an exploration parameter of  $\sqrt{2}$

Execute one iteration of MCTS. The tree should be expanded accordingly. Please update the value estimates  $v$  and the visitation counts  $c_v$ .

You will have access to a simulator. It will return an average state value of  $v = 1$  for the state reached from taking  $a_1$ , a value  $v = 2$  for the state reached from taking  $a_2$ , and a value  $v = 3$  for the state reached from taking  $a_3$ . There are various ways to deal with the expansion step. Feel free to choose any expansion strategy that makes MCTS converge.



## 3. Policy Iteration (8 Points)

(a) Consider the simple gridworld below. The possible actions are **up**, **down**, **left** and **right**, and transitions are deterministic. Actions leading out of the grid leave the state unchanged. The top-left and bottom-right corners are terminal states, any of the other cells are nonterminal states. The agent receives a reward of -1 on every transition. We consider a discounted MDP with  $\gamma = 0.5$ . Please apply policy iteration. The left column shows the current state-value function estimates (initially initialized with 0). The right column shows the greedy policy with respect to the current value function (initially initialized with up in every state. In every iteration (i.e., line) please do a single sweep over the state space (instead of running until convergence) to evaluate the policy. You are only required to fill out the gray boxes.

0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0

█	↑	↑	↑	↑
↑	↑	↑	↑	↑
↑	↑	↑	↑	↑
↑	↑	↑	↑	↑
↑	↑	↑	↑	█

█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█

█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█

█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█

█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█
█	█	█	█	█