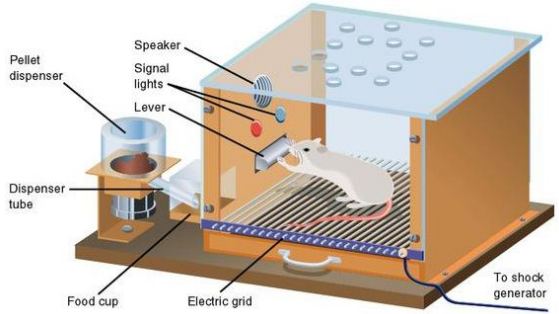
High-level intuition & course overview

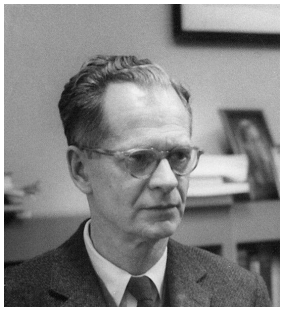
Introduction to Reinforcement Learning

Leiden University, The Netherlands

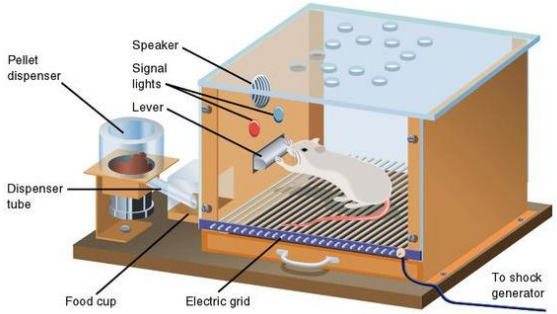
Biology

Biology

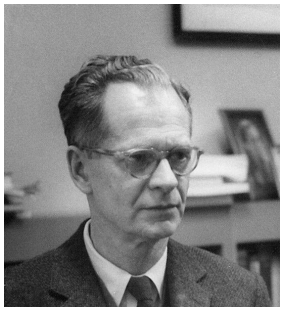

Pellet dispenser
Speaker
Signal lights
Lever
Dispenser tube
Food cup
Electric grid
To shock generator



Biology

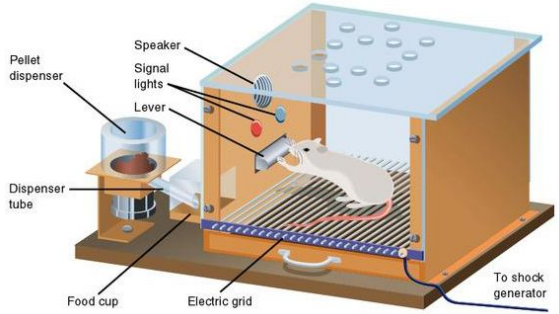


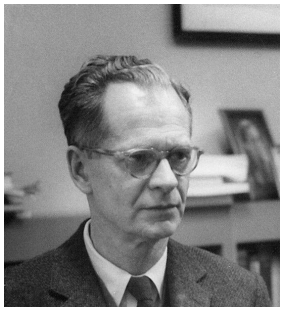
Skinner box



B.F. Skinner (1904 – 1990)

Biology



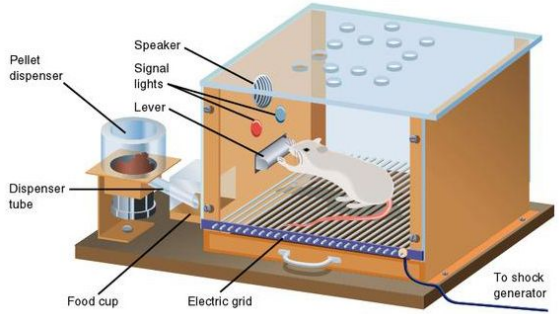


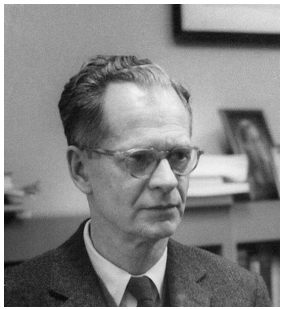
Skinner box

B.F. Skinner (1904 – 1990)

Instrumental conditioning:
Learning behaviour based on reward and punishment (trial and error)

Biology





Skinner box

B.F. Skinner (1904 – 1990)

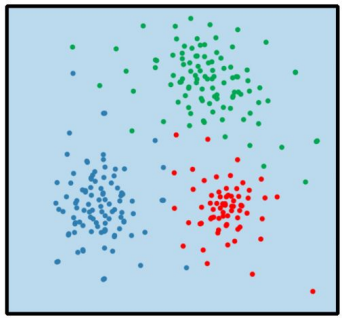
Instrumental conditioning:
Learning behaviour based on reward and punishment (trial and error)

RL is the computational specification of this idea

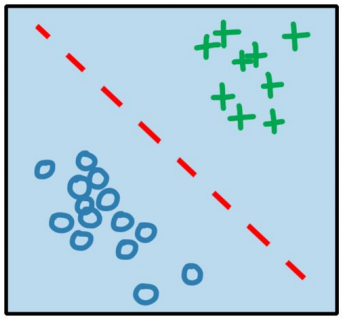

machine learning

machine learning

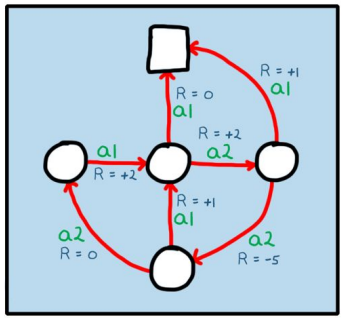
unsupervised learning



supervised learning



reinforcement learning



Supervised vs reinforcement learning

Supervised vs reinforcement learning

	Supervised learning	Reinforcement learning
Dataset		
Feedback		

Supervised vs reinforcement learning

	Supervised learning	Reinforcement learning
Dataset	Given	
Feedback		

Supervised vs reinforcement learning

	Supervised learning	**Reinforcement learning**
Dataset	Given	Active collection
Feedback		

Supervised vs reinforcement learning

	Supervised learning	**Reinforcement learning**
Dataset	Given	Active collection
Feedback	Full (x with correct y)	

Supervised vs reinforcement learning

	Supervised learning	**Reinforcement learning**
Dataset	Given	Active collection
Feedback	Full (x with correct y)	Partial (~~state with correct action~~) (feedback on some outcomes)

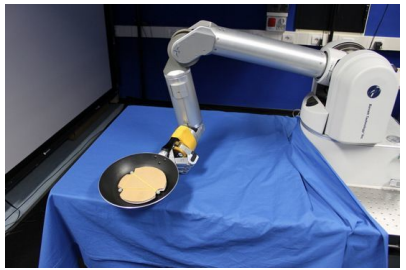
Benefits of Reinforcement Learning

Benefits of Reinforcement Learning

Autonomous behaviour/learning
(only specify goals)

Benefits of Reinforcement Learning

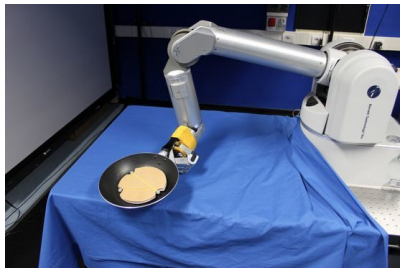
Autonomous behaviour/learning
(only specify goals)



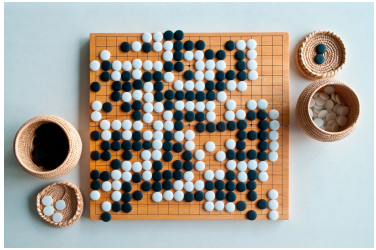
Solve tasks that you can't label
(only need to label the outcome)

Benefits of Reinforcement Learning

Autonomous behaviour/learning
(only specify goals)



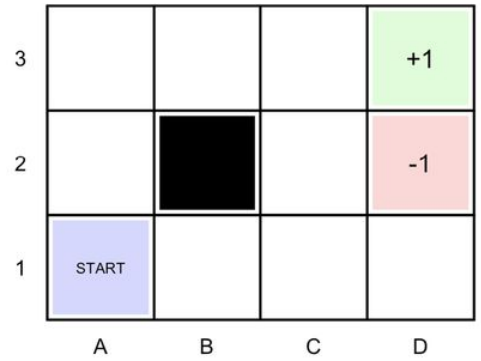
Solve tasks that you can't label
(only need to label the outcome)



Outperform human solution
(only need to label the outcome)

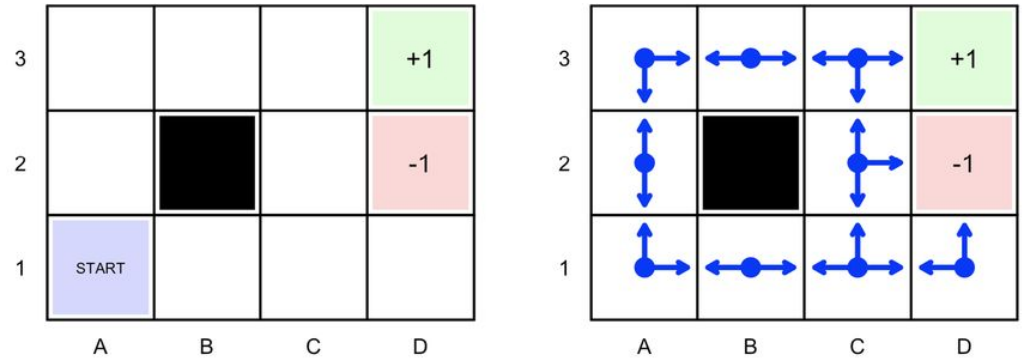
Sequential decision-making

Sequential decision-making



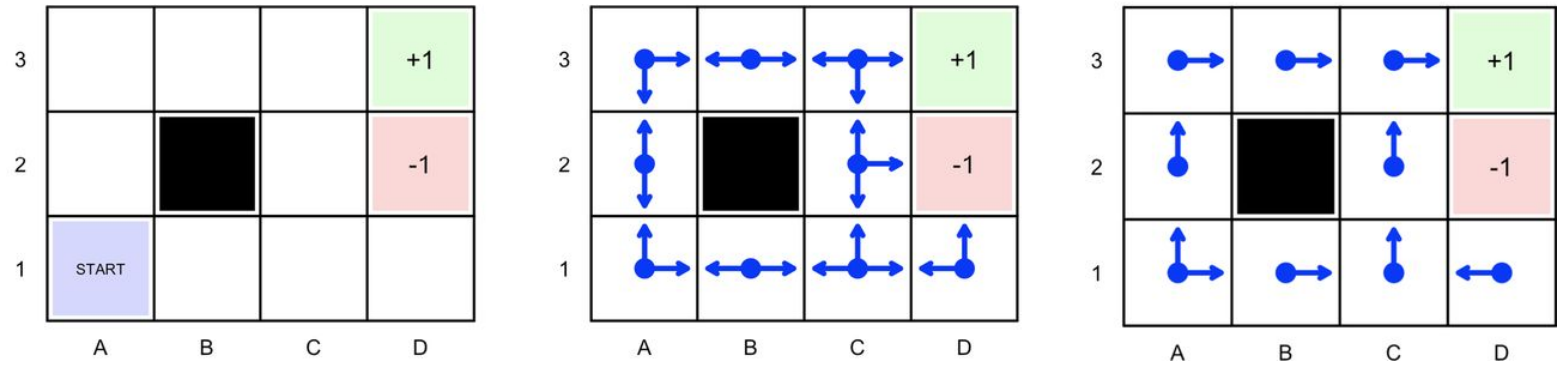
We start in a certain state, we can take actions that may bring us to other states, and there are some unknown rewards (goals) and punishments

Sequential decision-making



In advance we have no clue how to act in every state (*random* policy)

Sequential decision-making



We eventually want to find the action selection strategy in each state that in the long run gets us as much total reward as possible (the *optimal policy*)

Course structure

Block I

Block II

Block III

Block I:

Exploration & Dynamic Programming

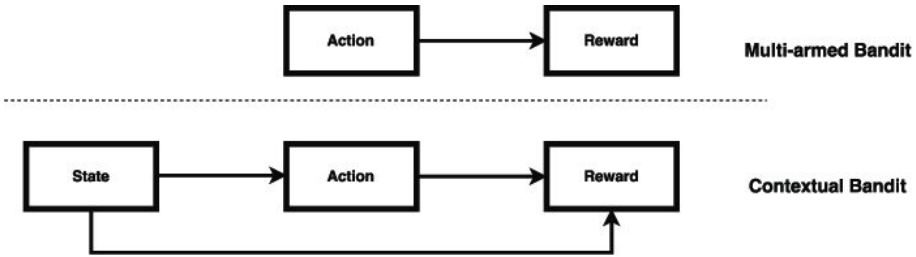
Week 2: Exploration

Week 2: Exploration

Start with the simplest one-step decision making problem: **Multi-armed Bandit**

Week 2: Exploration

Start with the simplest one-step decision making problem: **Multi-armed Bandit**



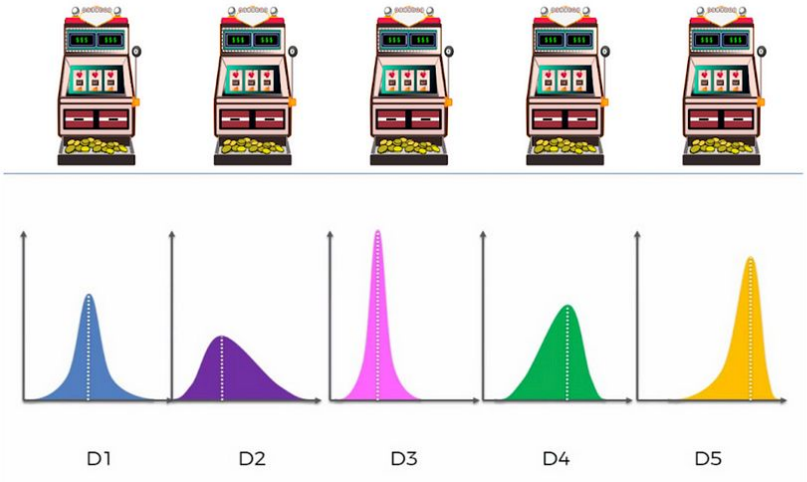
Week 2: Exploration

Week 2: Exploration

Key challenge: **exploration-exploitation trade-off**

Week 2: Exploration

Key challenge: **exploration-exploitation trade-off**



Multiple available actions with unknown reward distributions

Learn ways to trade-off exploration (something novel) & exploitation (known to work well)

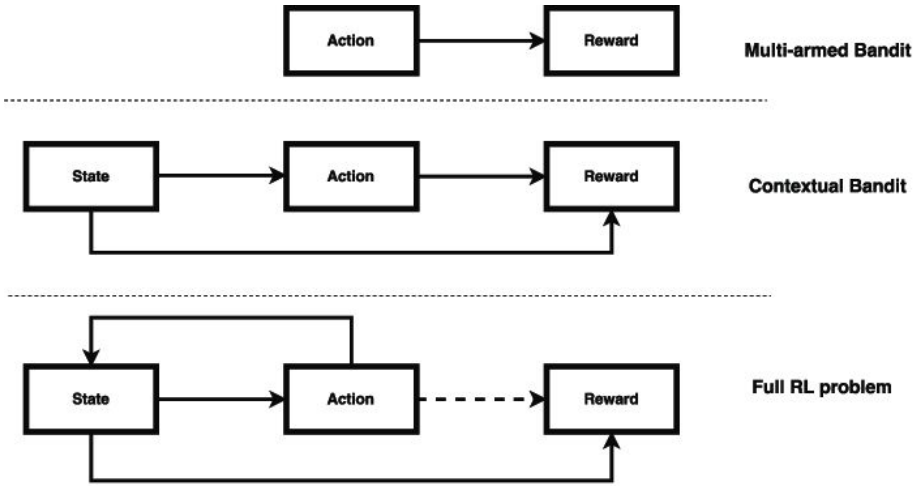
Week 3: Markov Decision Process

Week 3: Markov Decision Process

Make the decision-making problem *sequential*

Week 3: Markov Decision Process

Make the decision-making problem *sequential*



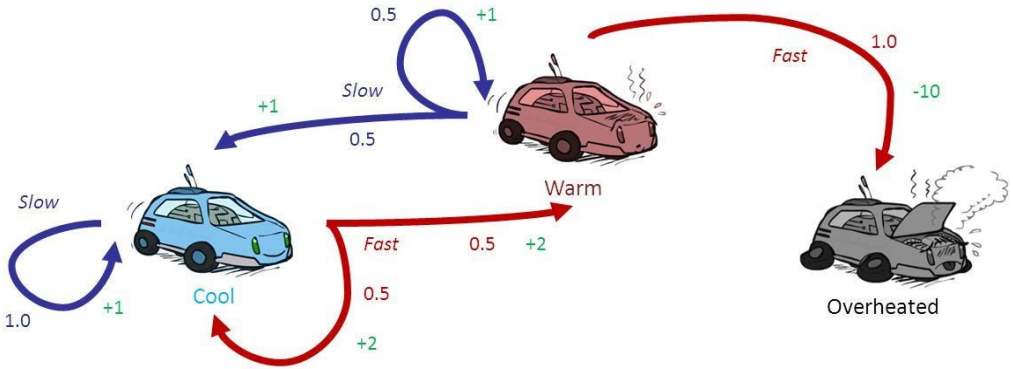
Week 3: Markov Decision Process

Week 3: Markov Decision Process

Learn to define sequential decision-making problems in a generic way

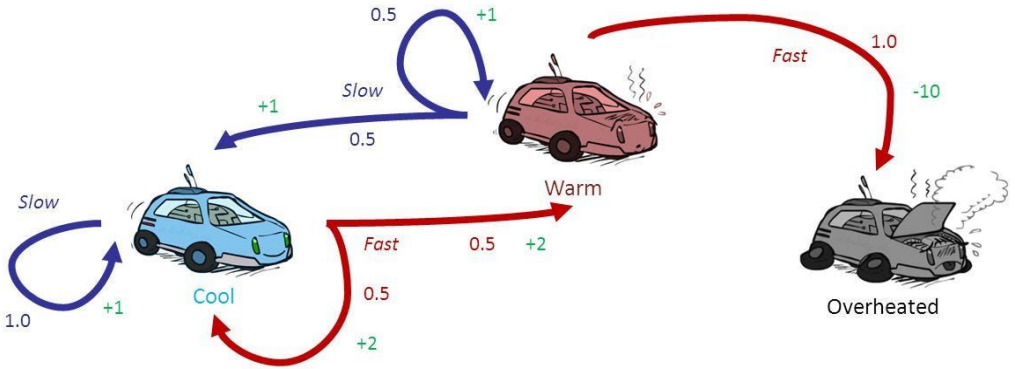
Week 3: Markov Decision Process

Learn to define sequential decision-making problems in a generic way



Week 3: Markov Decision Process

Learn to define sequential decision-making problems in a generic way



- States
- Actions
- Transition probabilities
- Reward function

- Policy
- Value function

Week 4: Dynamic Programming

Week 4: Dynamic Programming

If we have a *full transition model* we can *exactly* solve for the optimal policy

Week 4: Dynamic Programming

If we have a *full transition model* we can *exactly* solve for the optimal policy

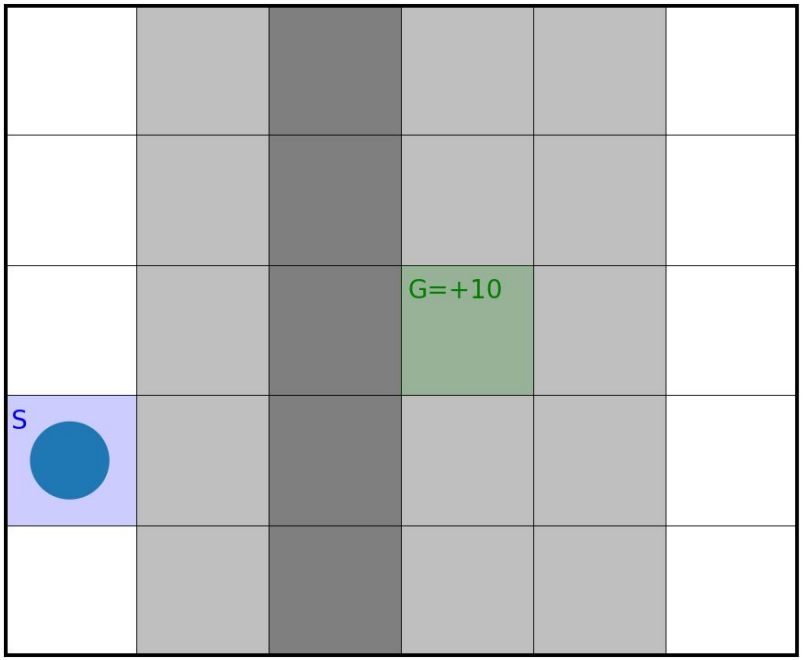
By:

1) Sweeping through all states, each time propagating rewards 1 step further back

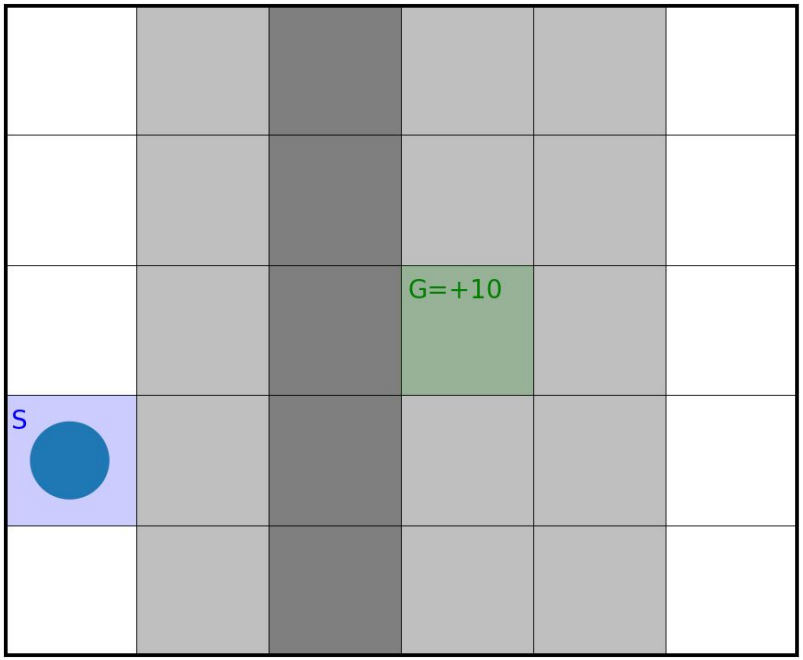
2) Until we converge on the 'optimal value' (highest total reward obtainable) of each state-action

Week 4: Dynamic Programming - demonstration

Week 4: Dynamic Programming - demonstration

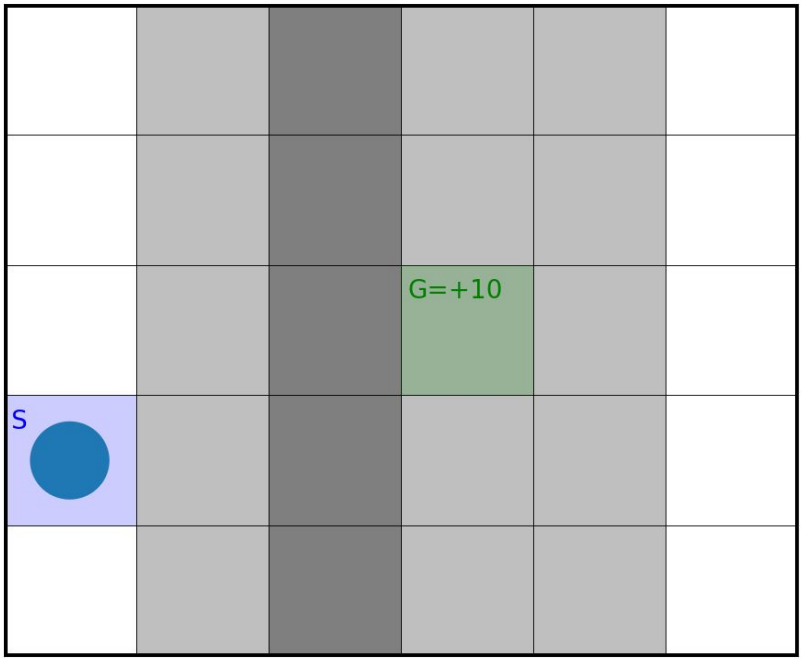


Week 4: Dynamic Programming - demonstration



- **States:** all locations
- **Actions**: Can move up/down/left/right
- **Transitions**: In the dark columns there blows an upward wind (up) with strength 1 (light grey) or 2 (dark grey)
- **Rewards**: The goal has a reward of +10, all other steps have a reward of -1

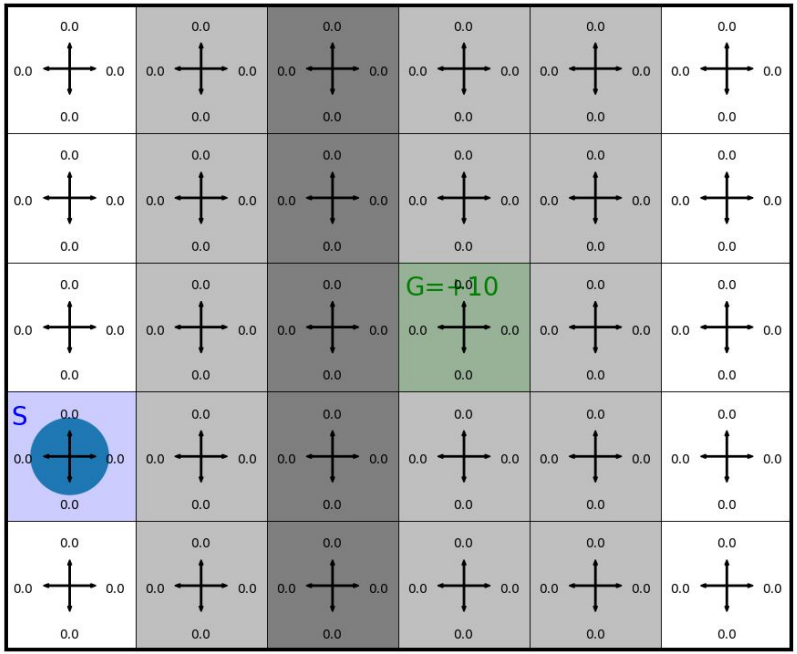
Week 4: Dynamic Programming - demonstration



- **States:** all locations
- **Actions**: Can move up/down/left/right
- **Transitions**: In the dark columns there blows an upward wind (up) with strength 1 (light grey) or 2 (dark grey)
- **Rewards**: The goal has a reward of +10, all other steps have a reward of -1

Question: What path do you think an optimal agent take?

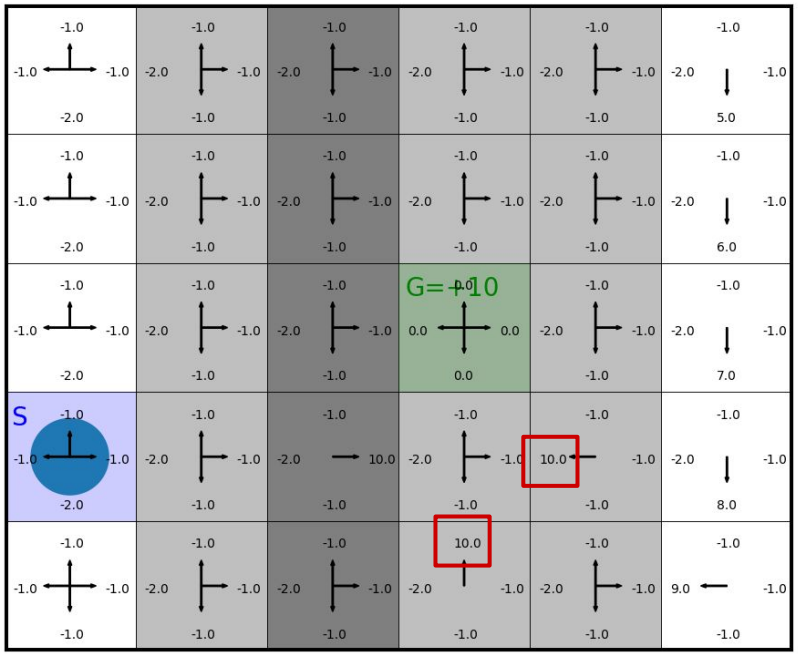
Week 4: Dynamic Programming - demonstration



Initialize all state-action values

(e.g. to 0.0)

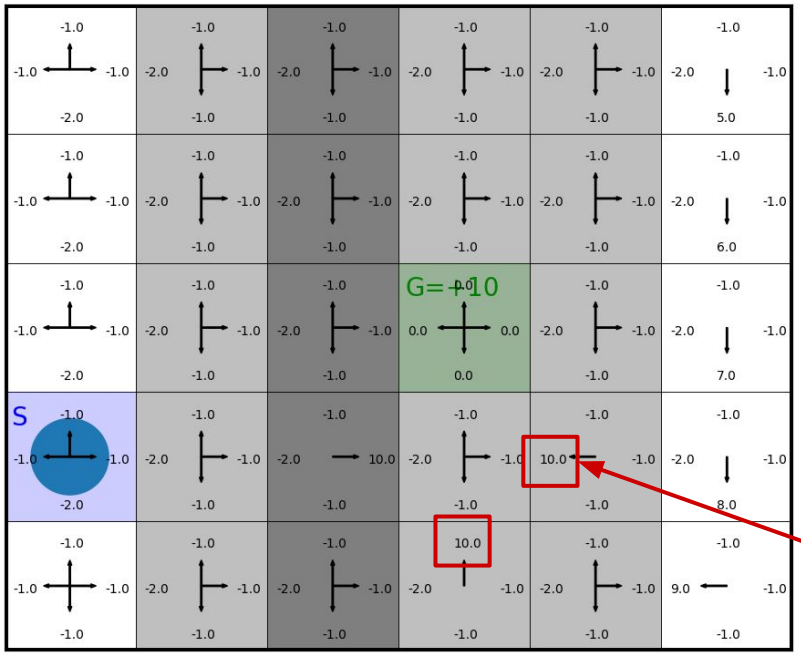
Week 4: Dynamic Programming - demonstration



Iterate (sweep through all states)

→ Rewards start propagating in value estimates

Week 4: Dynamic Programming - demonstration

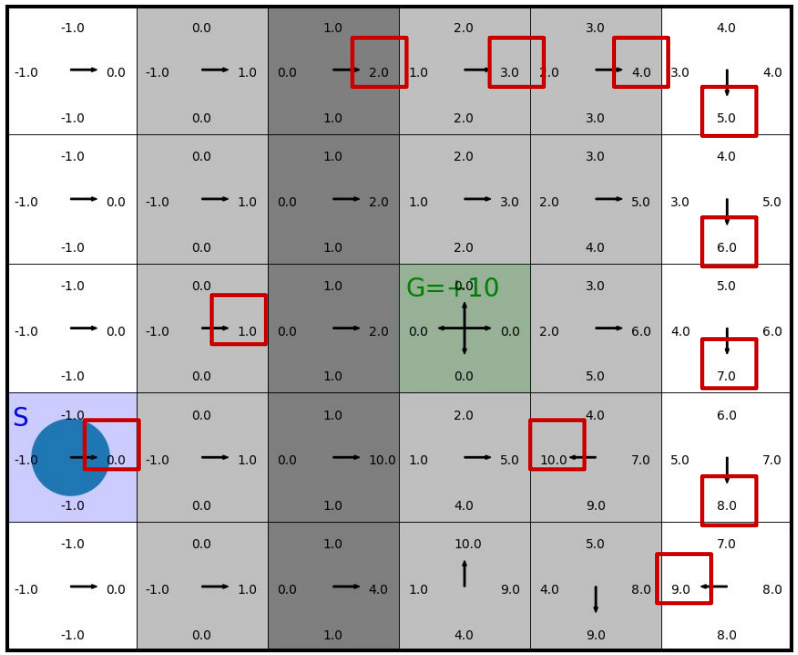


Iterate (sweep through all states)

→ Rewards start propagating in value estimates

'From this action, you can at best get a total reward of 10'

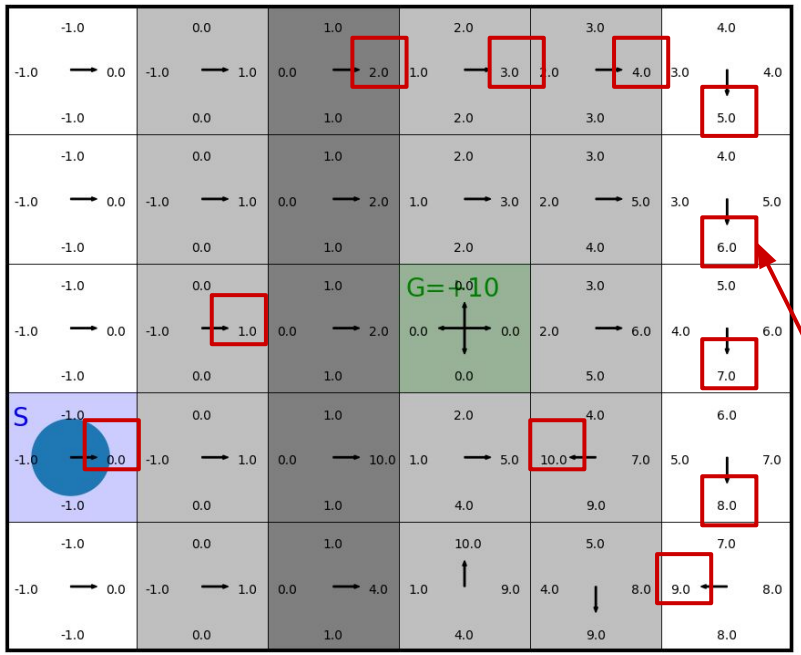
Week 4: Dynamic Programming - demonstration



Until convergence

→ All reward information has propagated

Week 4: Dynamic Programming - demonstration

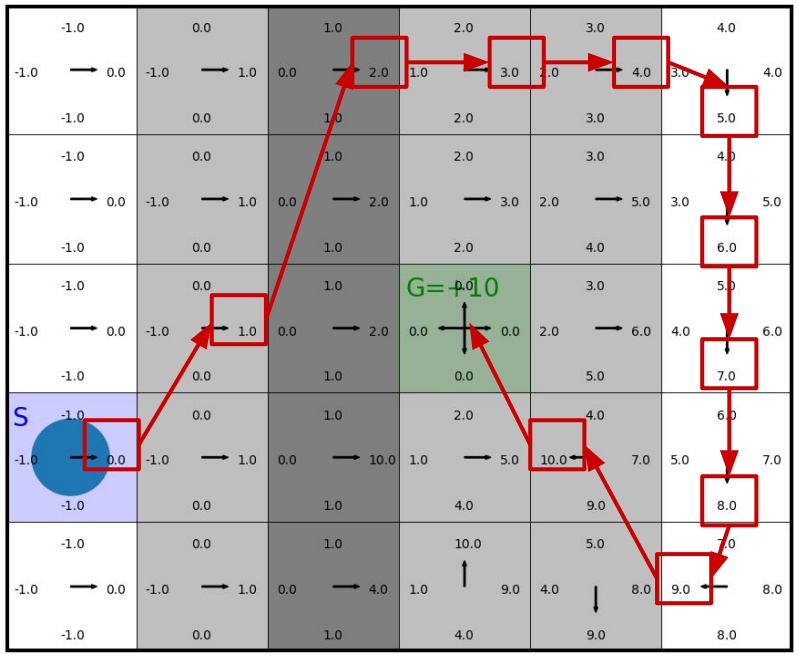


Until convergence

→ All reward information has propagated

'From this action, you can at best get a total reward of 6'
(4 steps of -1, and one final reward of +10)

Week 4: Dynamic Programming - demonstration



Found the optimal policy:

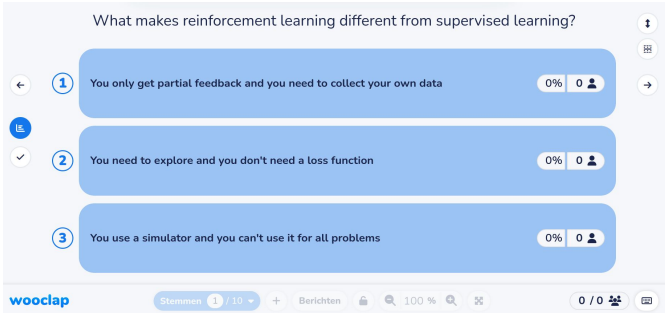
by repeatedly taking the available action with the highest value estimate

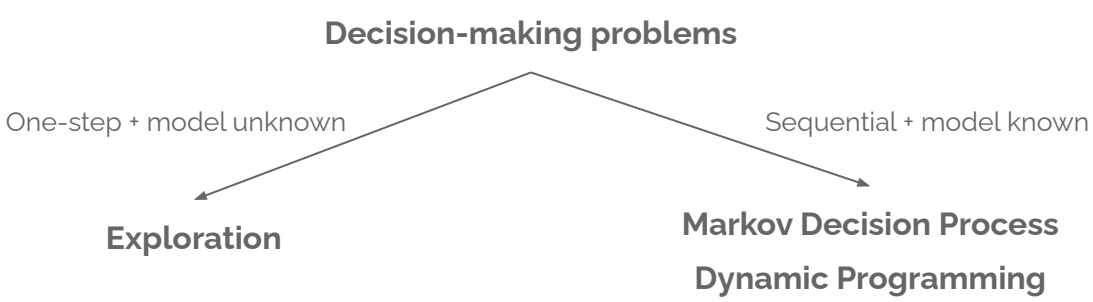
Week 5: Q&A + Quiz + Assignment 1 deadline

In the last week of each block you get time to reflect & complete the assignment

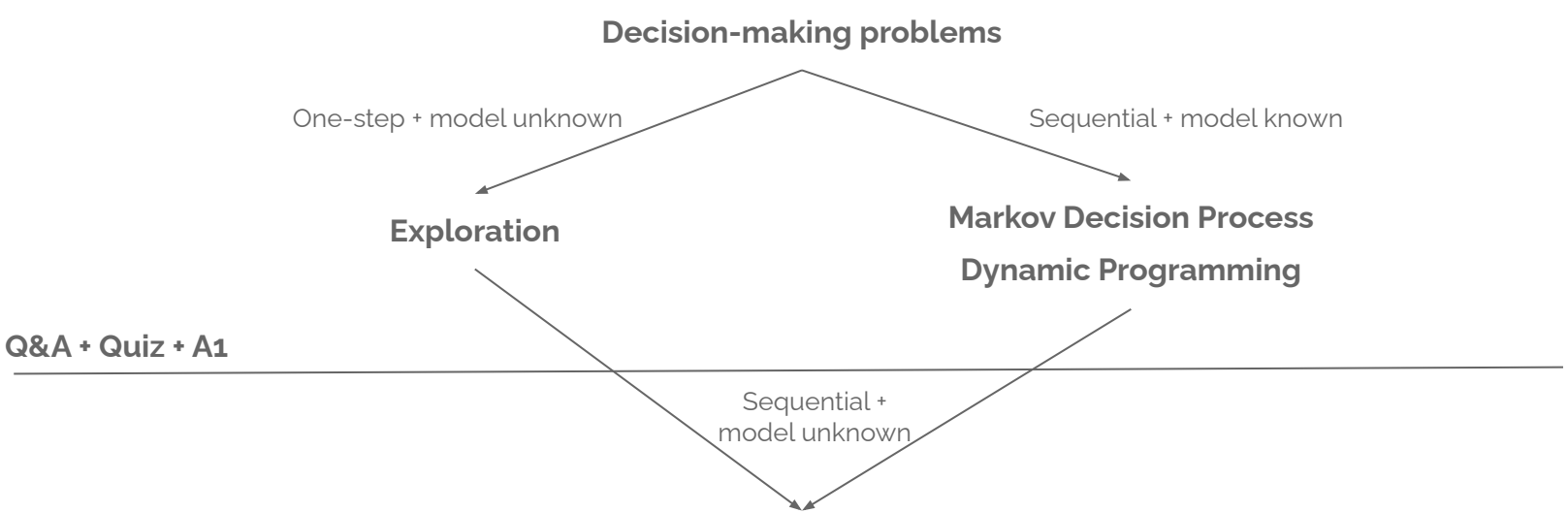
Week 5: Q&A + Quiz + Assignment 1 deadline

In the last week of each block you get time to reflect & complete the assignment




Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming

Q&A + Quiz + A1


Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming
Q&A + Quiz + A1
Sequential +
model unknown

Block II:

Model-free Reinforcement Learning

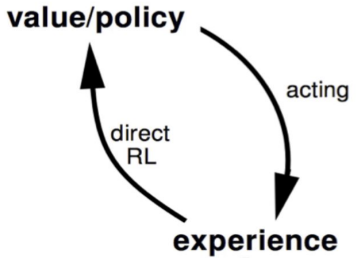
Week 6+7: Model-free reinforcement learning

Week 6+7: Model-free reinforcement learning

Typically we don't have a model of the MDP, but only a *simulator* (or the *real-world*)

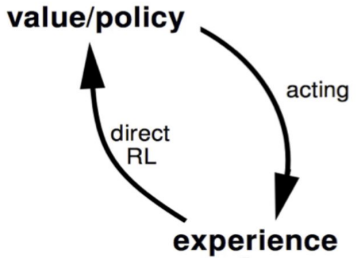
Week 6+7: Model-free reinforcement learning

Typically we don't have a model of the MDP, but only a *simulator* (or the *real-world*)



Week 6+7: Model-free reinforcement learning

Typically we don't have a model of the MDP, but only a *simulator* (or the *real-world*)



We need to iterate:

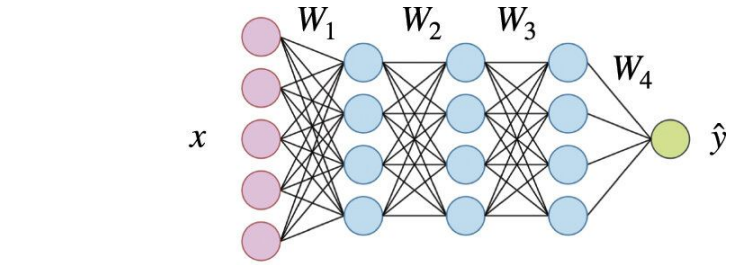
1. **Act**: sample traces (with **exploration/exploitation trade-off** – Week 2)
2. **Update**: use the obtained data to improve our solution (**credit assignment**)

Week 6+7: Model-free reinforcement learning

Key challenge = **credit assignment**

Week 6+7: Model-free reinforcement learning

Key challenge = **credit assignment**



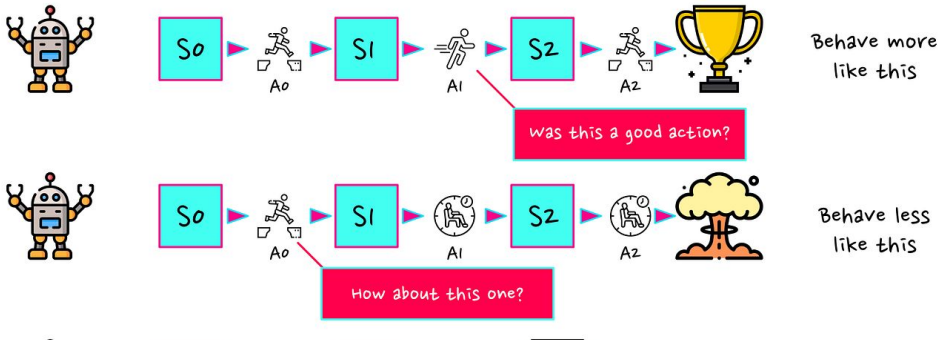
(**not** the credit assignment within a neural network – typically w. backpropagation)

Week 6+7: Model-free reinforcement learning

Key challenge = (temporal) **credit assignment**

Week 6+7: Model-free reinforcement learning

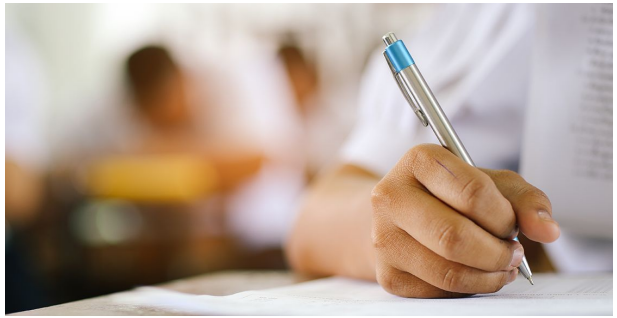
Key challenge = (temporal) **credit assignment**



(which actions in a sequence contributed to a good or bad outcome)

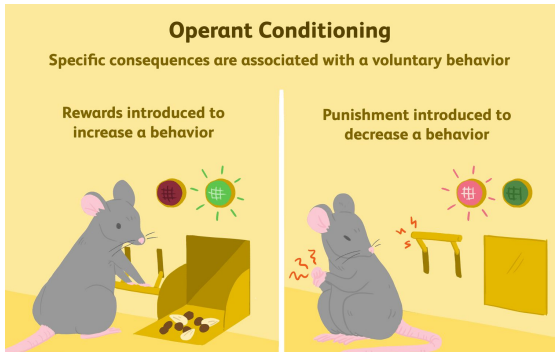
Week 8: Retake week / Free

Week 8: Retake week / Free



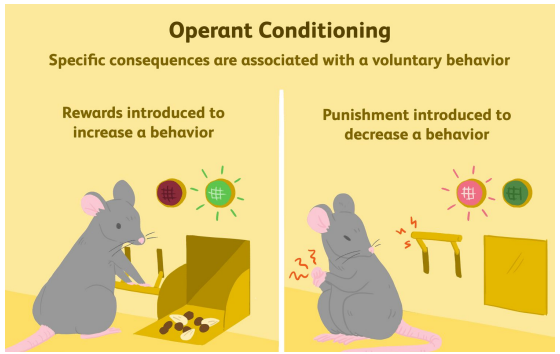
Week 9: Psychology & Neuroscience

Week 9: Psychology & Neuroscience

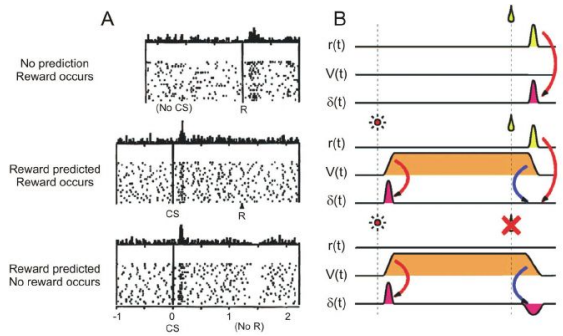


Psychology: inspiration in instrumental conditioning

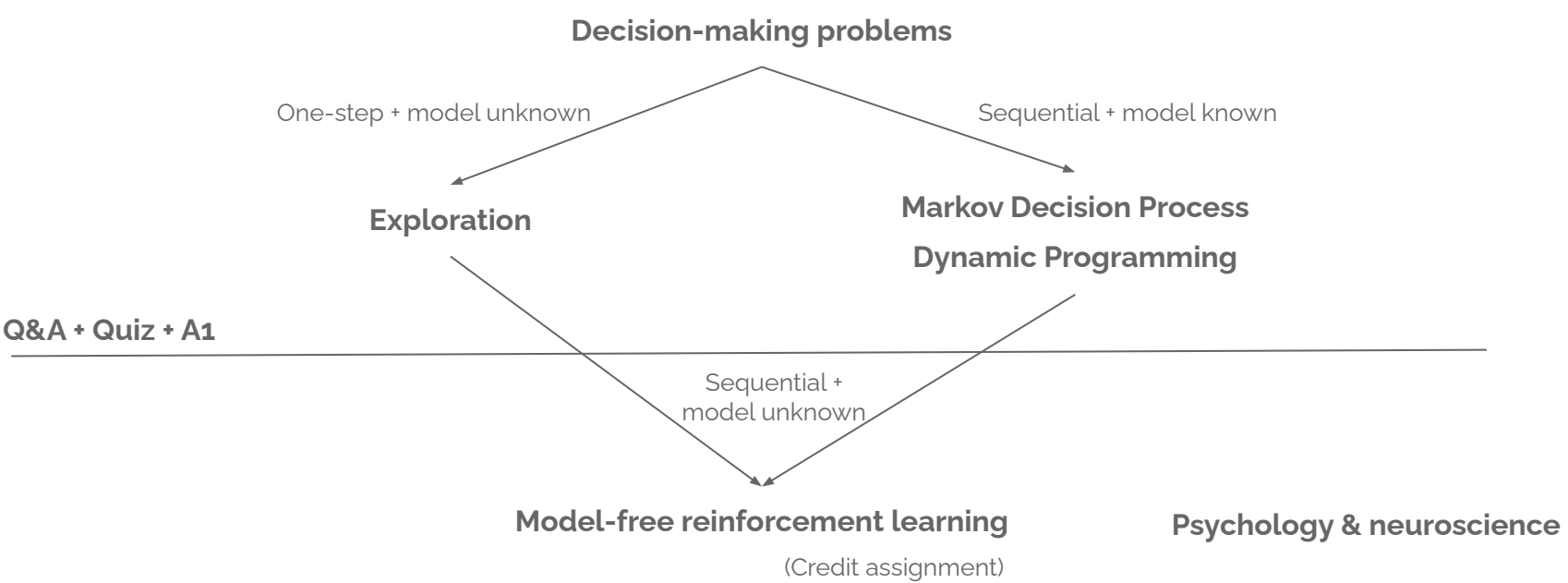
Week 9: Psychology & Neuroscience

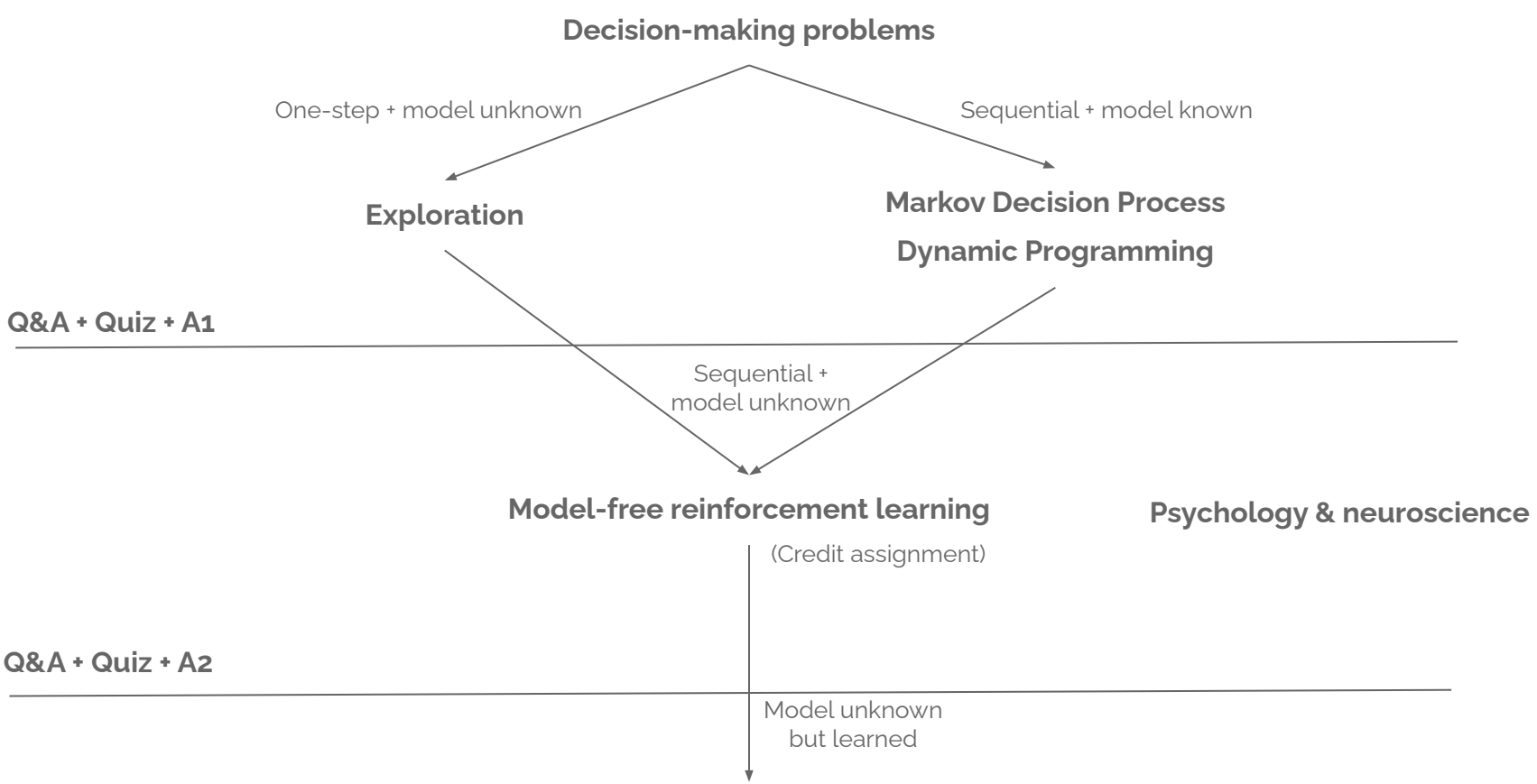


Psychology: inspiration in instrumental conditioning



Neuroscience: algorithmic RL concepts empirically match dopamine signalling


Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming
Q&A + Quiz + A1
Sequential +
model unknown
Model-free reinforcement learning
(Credit assignment)
Psychology & neuroscience
Q&A + Quiz + A2


Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming
Q&A + Quiz + A1
Sequential +
model unknown
Model-free reinforcement learning
Psychology & neuroscience
(Credit assignment)
Q&A + Quiz + A2
Model unknown
but learned

Block II:

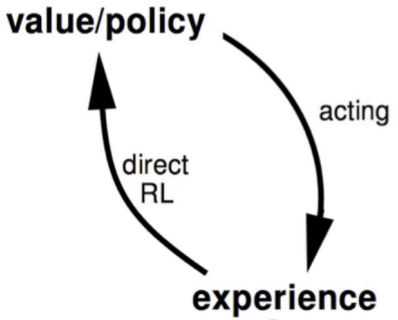
Model-based Reinforcement Learning

Week 11: Model-based reinforcement learning

We typically don't have a model at the start, but can we learn one from data?

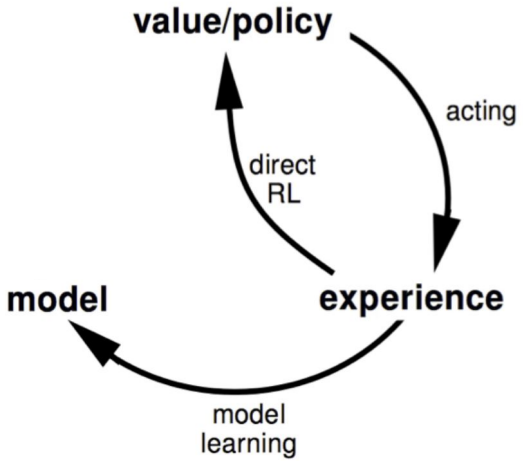
Week 11: Model-based reinforcement learning

We typically don't have a model at the start, but can we learn one from data?



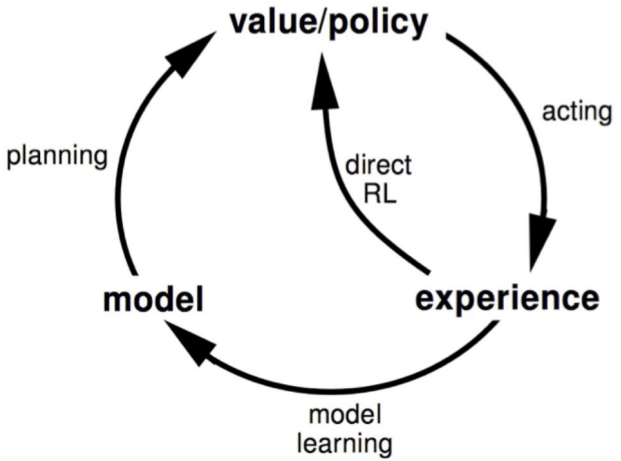
Week 11: Model-based reinforcement learning

We typically don't have a model at the start, but can we learn one from data?



Week 11: Model-based reinforcement learning

We typically don't have a model at the start, but can we learn one from data?



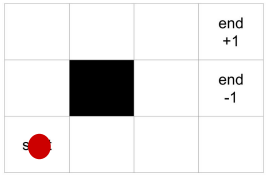
Week 12: Decision-time planning

Week 12: Decision-time planning

If we have a model, we can also try to solve the problem through repeated pure planning from a given state ('decision-time planning')

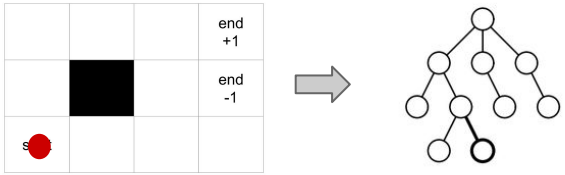
Week 12: Decision-time planning

If we have a model, we can also try to solve the problem through repeated pure planning from a given state ('decision-time planning')



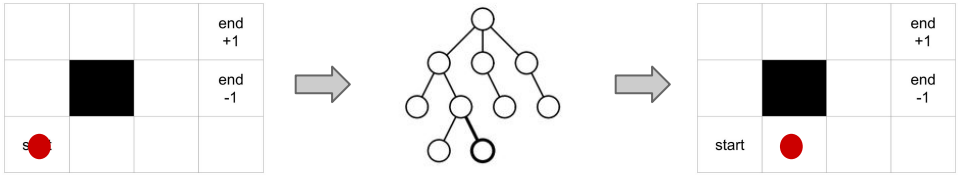
Week 12: Decision-time planning

If we have a model, we can also try to solve the problem through repeated pure planning from a given state ('decision-time planning')



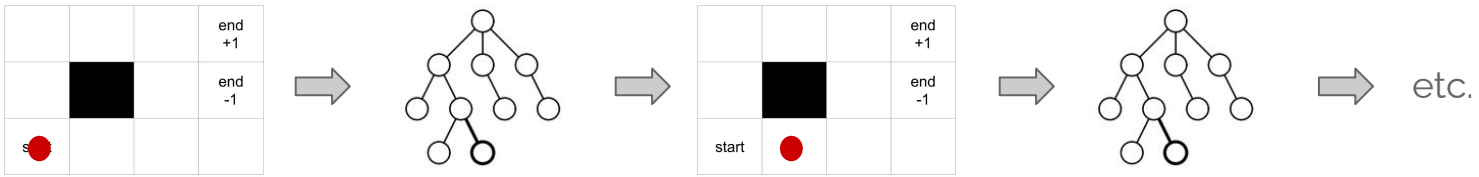
Week 12: Decision-time planning

If we have a model, we can also try to solve the problem through repeated pure planning from a given state ('decision-time planning')



Week 12: Decision-time planning

If we have a model, we can also try to solve the problem through repeated pure planning from a given state ('decision-time planning')



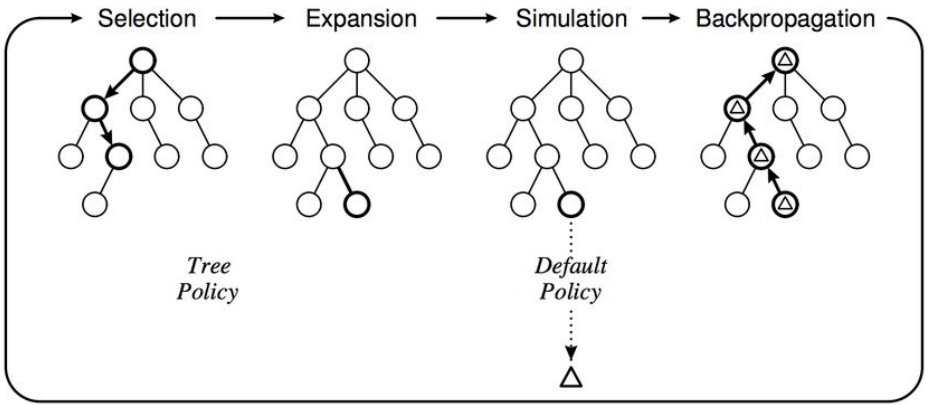
Week 12: Sample-based planning

Week 12: Sample-based planning

Instead of systematic enumeration, use *statistical sampling (with uncertainty)* to expand

Week 12: Sample-based planning

Instead of systematic enumeration, use *statistical sampling (with uncertainty)* to expand



Most popular class:
Monte Carlo Tree Search (MCTS)

Week 13: Advanced topics

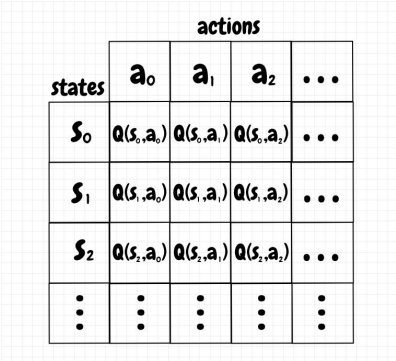
Week 13: Approximation

Week 13: Approximation

Tabular solutions don't scale to larger problems, where we need approximate solutions

Week 13: Approximation

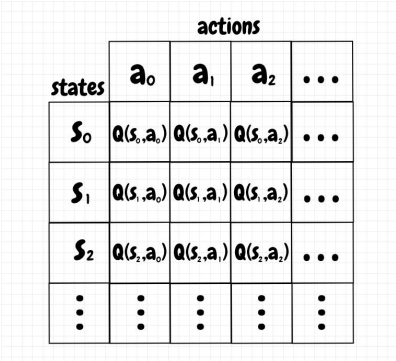
Tabular solutions don't scale to larger problems, where we need approximate solutions



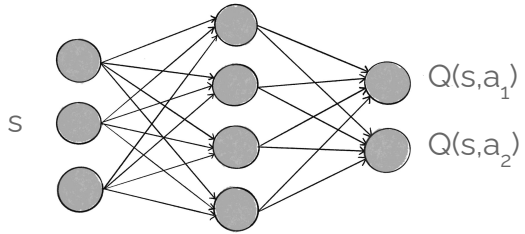
Tabular:
First part of course, focus on RL concepts

Week 13: Approximation

Tabular solutions don't scale to larger problems, where we need approximate solutions



Tabular:
First part of course, focus on RL concepts



Approximate:
Deep reinforcement learning - replace tabular solution wit neural network that generalizes

Week 13: Policy-based RL

Week 13: Policy-based RL

Instead of learning the value of actions, we may *directly learn action selection probabilities*

Week 13: Policy-based RL

Instead of learning the value of actions, we may *directly learn action selection probabilities*

Value-based:

Predict the expected sum of rewards for each action

Week 13: Policy-based RL

Instead of learning the value of actions, we may *directly learn action selection probabilities*

Value-based:

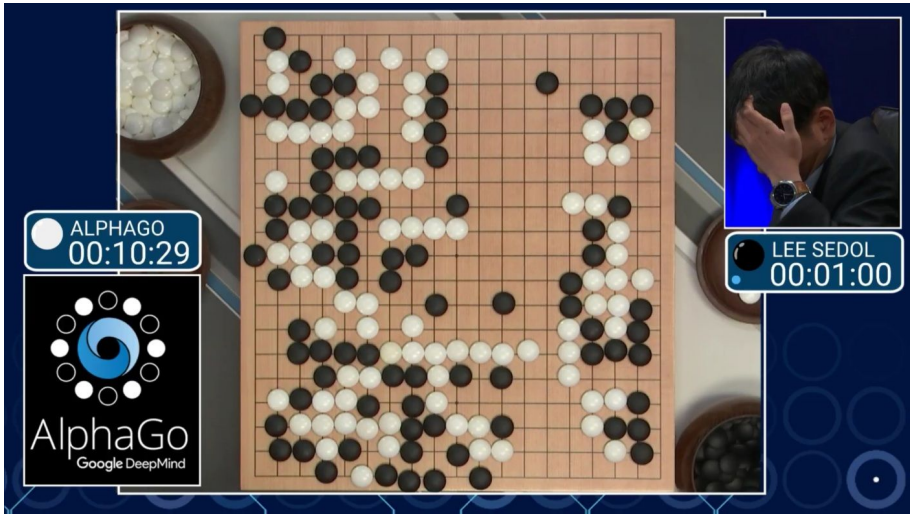
Predict the expected sum of rewards for each action

Policy-based:

Predict the desired action selection probability of each action

Week 14: Applications

Week 14: Applications



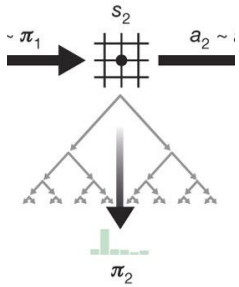
AlphaGo: One of the key successes of reinforcement learning

Week 14: AlphaGo

Brings together the main concepts of the whole course

Week 14: AlphaGo

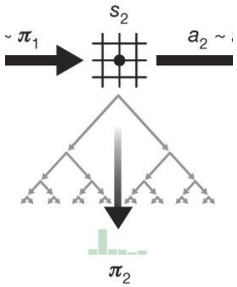
Brings together the main concepts of the whole course



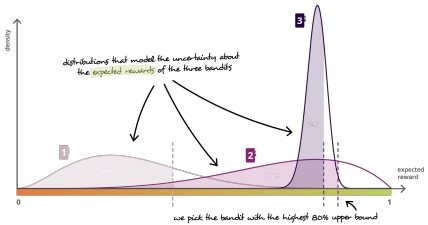
Sample-based planning
(Monte Carlo Tree Search)

Week 14: AlphaGo

Brings together the main concepts of the whole course



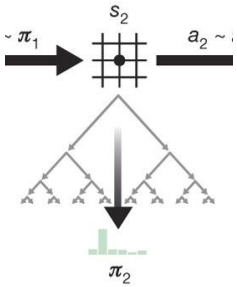
Sample-based planning
(Monte Carlo Tree Search)



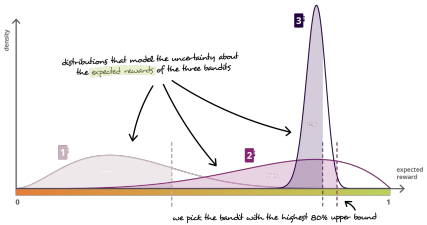
Exploration +
Credit assignment

Week 14: AlphaGo

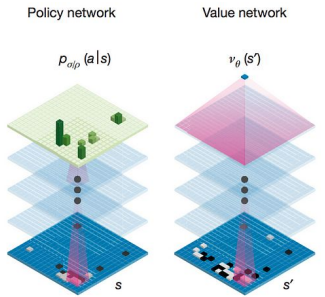
Brings together the main concepts of the whole course



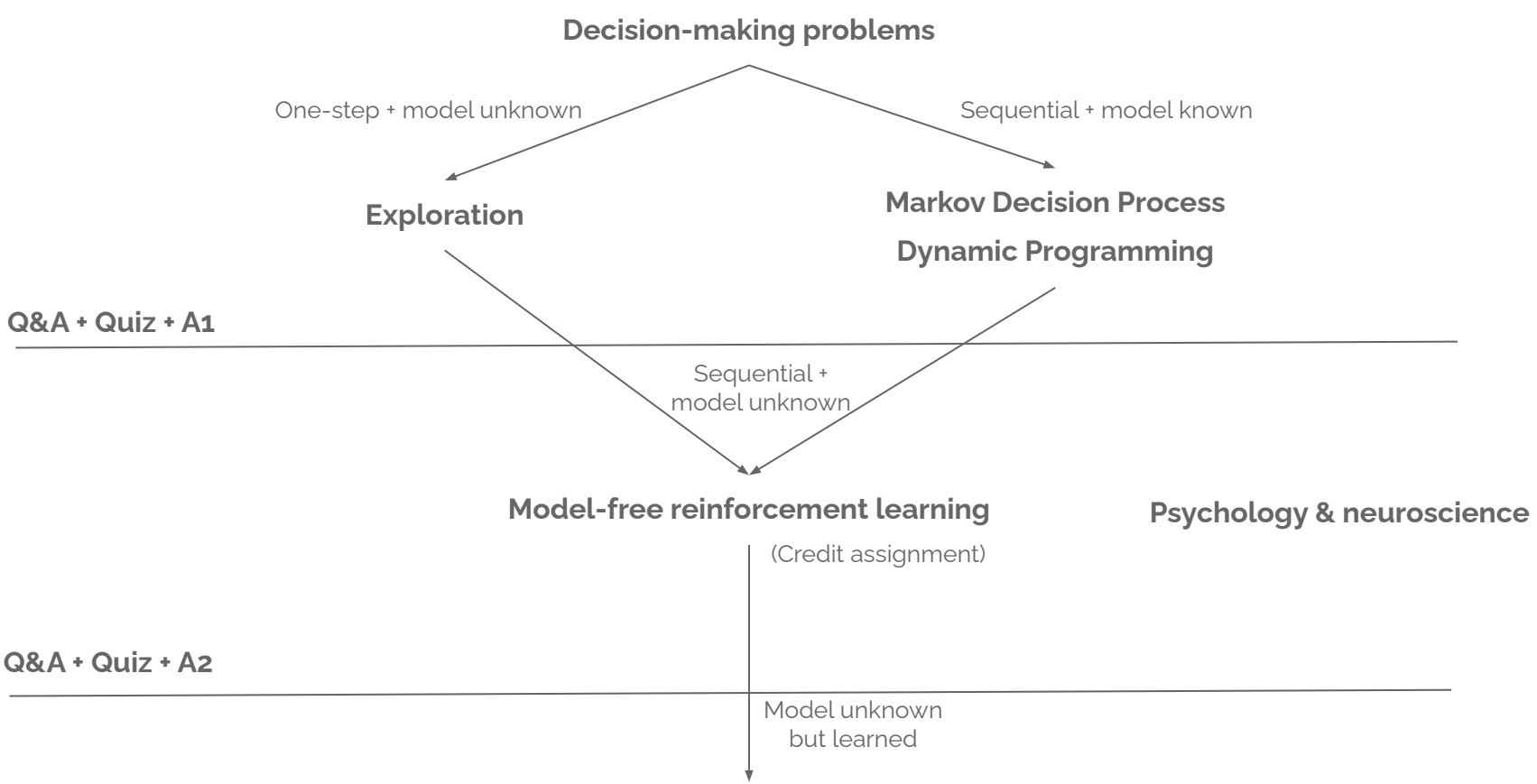
Sample-based planning
(Monte Carlo Tree Search)

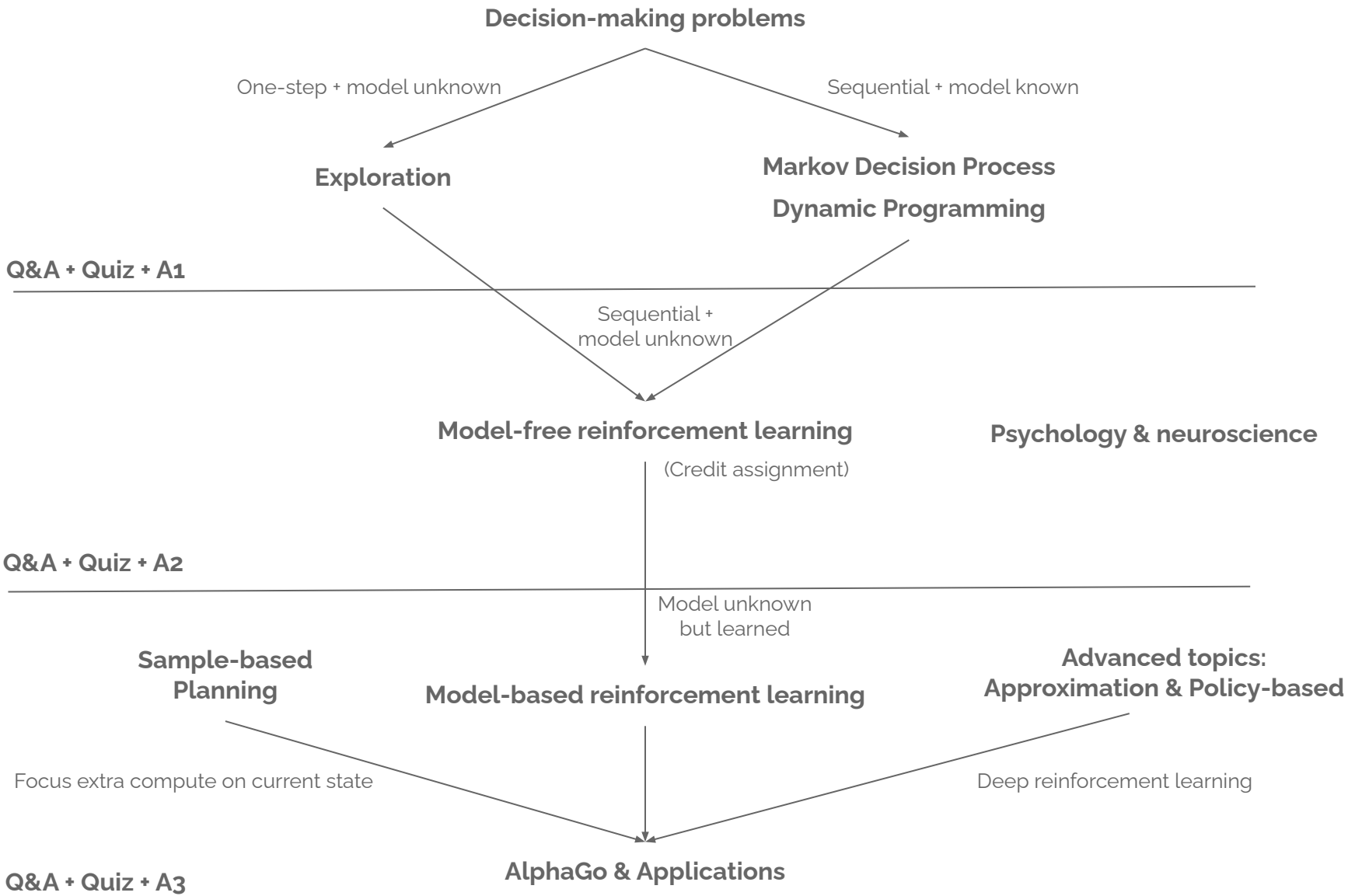


Exploration +
Credit assignment



Neural network approximation +
policy & value (actor-critic)


Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming
Q&A + Quiz + A1
Sequential +
model unknown
Model-free reinforcement learning
Psychology & neuroscience
(Credit assignment)
Q&A + Quiz + A2
Model unknown
but learned


Decision-making problems
One-step + model unknown
Sequential + model known
Exploration
Markov Decision Process
Dynamic Programming
Q&A + Quiz + A1
Sequential +
model unknown
Model-free reinforcement learning
Psychology & neuroscience
(Credit assignment)
Q&A + Quiz + A2
Model unknown
but learned
Sample-based
Planning
Model-based reinforcement learning
Advanced topics:
Approximation & Policy-based
Focus extra compute on current state
Deep reinforcement learning
AlphaGo & Applications
Q&A + Quiz + A3

At Home

At Home

- Read Chapter 1 from Sutton & Barto

At Home

- Read Chapter 1 from Sutton & Barto
- Study the slides and course website

At Home

- Read Chapter 1 from Sutton & Barto
- Study the slides and course website
- (If applicable: repeat preliminaries on ‘Probability’ and ‘Experimentation & Report Writing’)

At Home

- Read Chapter 1 from Sutton & Barto
- Study the slides and course website
- (If applicable: repeat preliminaries on ‘Probability’ and ‘Experimentation & Report Writing’)
- Start on A1 (Colab notebooks also explain the material)

Questions?