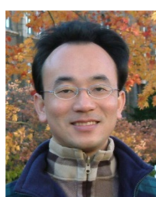
Tamkang University

Big Data Mining

Reinforcement Learning (RL)

1071BDM13
TLVXM1A (M2244) (8619) (Fall 2018)
(MBA, DBETKU) (3 Credits, Required) [Full English Course]
(Master's Program in Digital Business and Economics)
Mon, 9, 10, 11, (16:10-19:00) (B206)



Min-Yuh Day, Ph.D.
Assistant Professor
[Department of Information Management](http://mail.tku.edu.tw/myday)
[Tamkang University](http://mail.tku.edu.tw/myday)

http://mail.tku.edu.tw/myday

2018-12-17

Course Schedule (1/2)

Tamkang University

Week	Date	Subject/Topics
1	2018/09/10	Course Orientation for Big Data Mining
2	2018/09/17	ABC: AI, Big Data, Cloud Computing
3	2018/09/24	Mid-Autumn Festival (Day off)
4	2018/10/01	Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data
5	2018/10/08	Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem
6	2018/10/15	Foundations of Big Data Mining in Python
7	2018/10/22	Supervised Learning: Classification and Prediction
8	2018/10/29	Unsupervised Learning: Cluster Analysis
9	2018/11/05	Unsupervised Learning: Association Analysis

Course Schedule (2/2)

Tamkang University

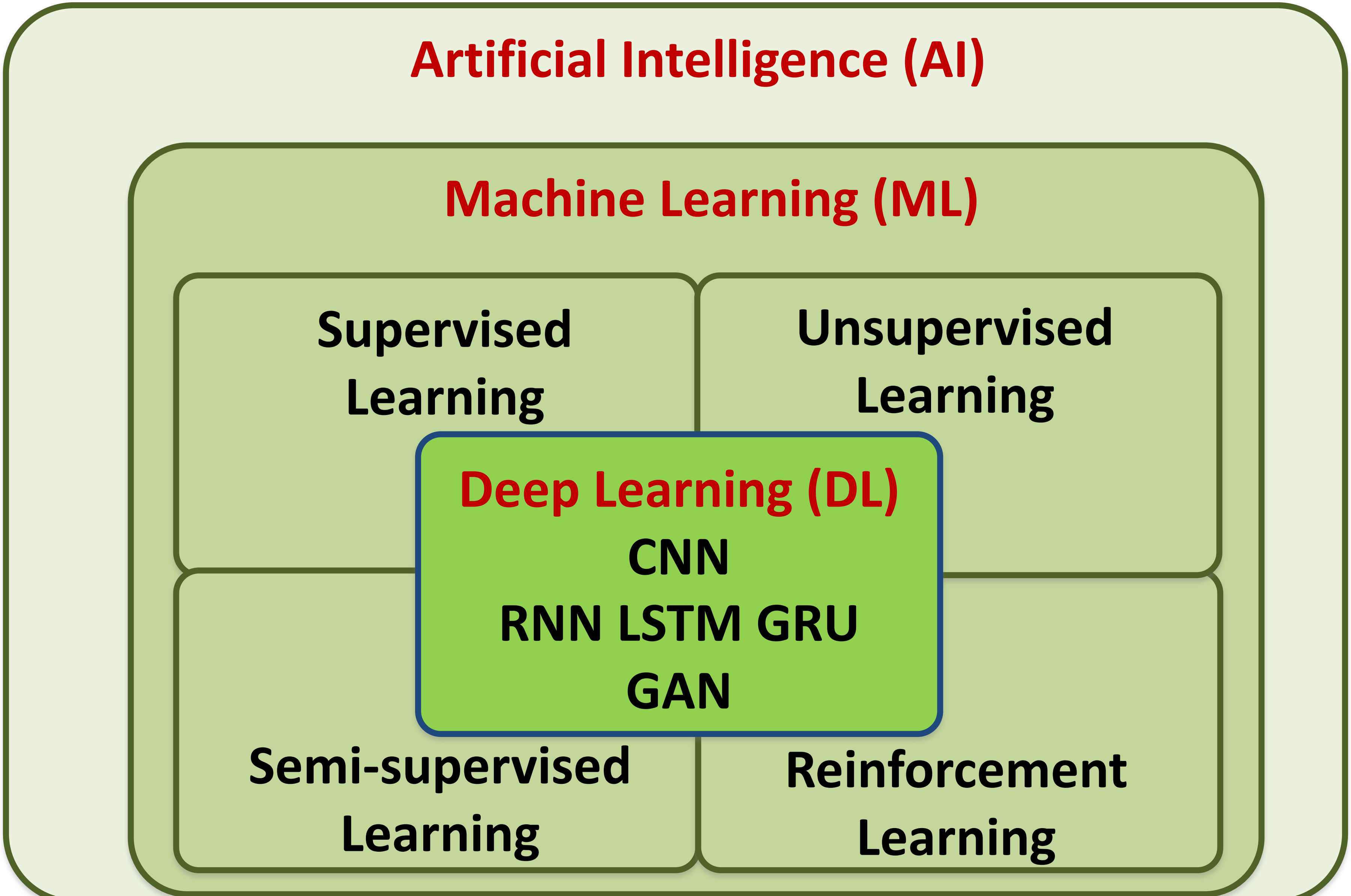
Week	Date	Subject/Topics
10	2018/11/12	Midterm Project Report
11	2018/11/19	Machine Learning with Scikit-Learn in Python
12	2018/11/26	Deep Learning for Finance Big Data with TensorFlow
13	2018/12/03	Convolutional Neural Networks (CNN)
14	2018/12/10	Recurrent Neural Networks (RNN)
15	2018/12/17	Reinforcement Learning (RL)
16	2018/12/24	Social Network Analysis (SNA)
17	2018/12/31	Bridge Holiday (Extra Day Off)
18	2019/01/07	Final Project Presentation

Reinforcement Learning (RL)

Outline

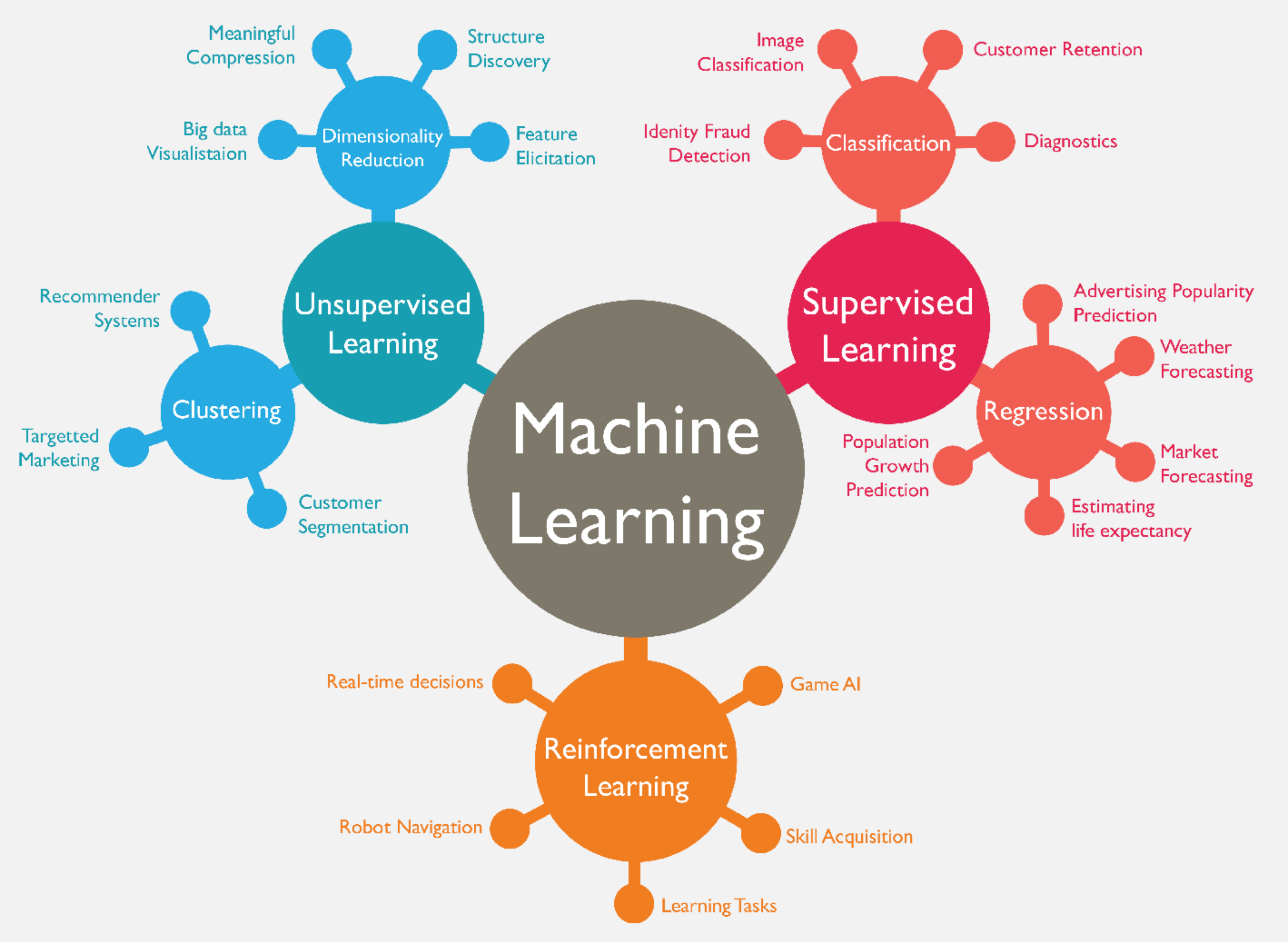
- Reinforcement Learning (RL)
  - Markov Decision Processes (MDP)
- Deep Reinforcement Learning (DRL) Algorithms
  - SARSA
  - Q-Learning
  - DQN
  - A3C
  - Rainbow
- Google Dopamine

AI, ML, DL



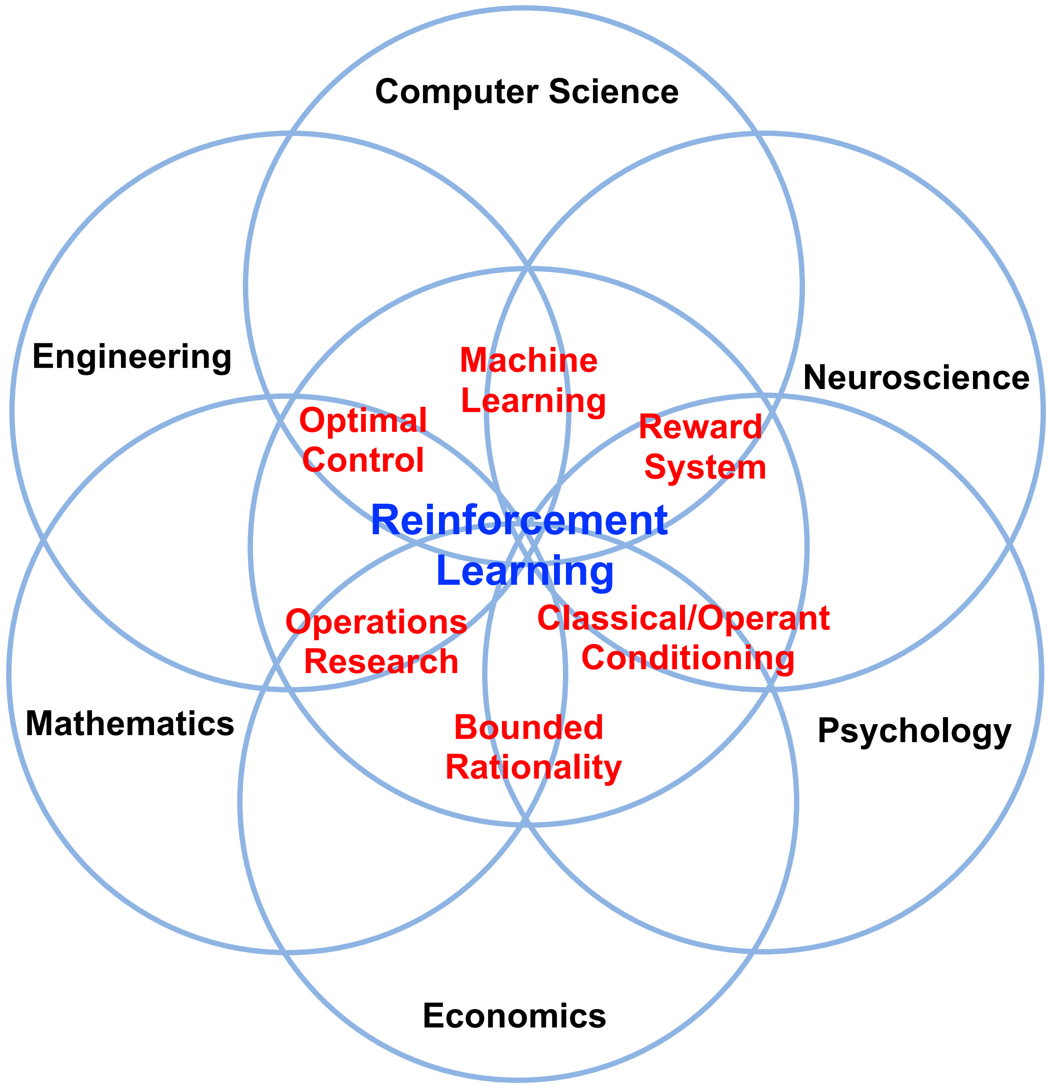
Source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/deep_learning.html

Machine Learning (ML)



Source: https://www.mactores.com/services/aws-big-data-machine-learning-cognitive-services/

Reinforcement Learning (RL)

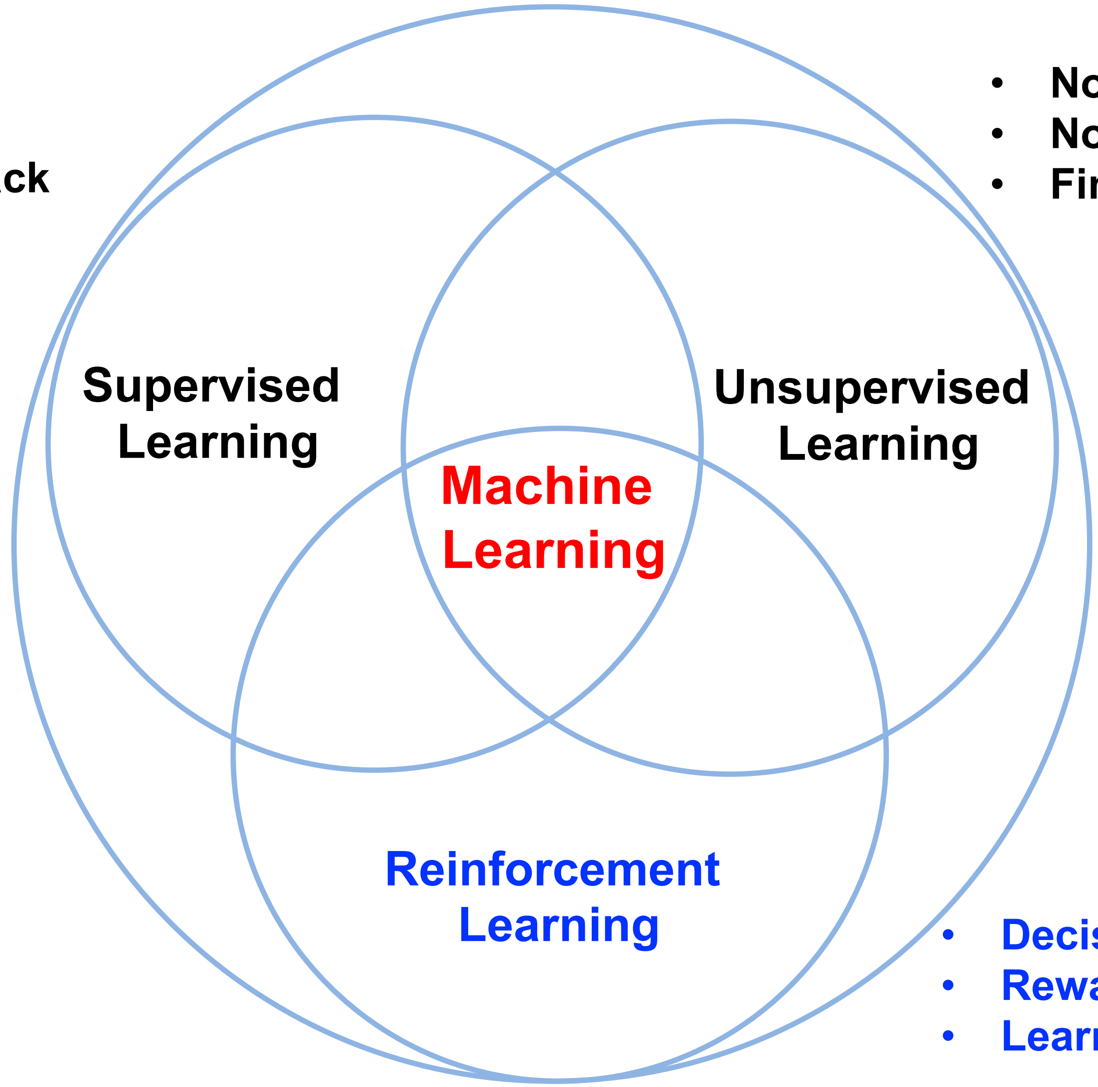


Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Branches of Machine Learning (ML)
Reinforcement Learning (RL)

- Labeled data
- Direct feedback
- Predict

- No Labels
- No feedback
- Find hidden structure



- Decision process
- Reward system
- Learn series of actions

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

David Silver (2015), Introduction to reinforcement learning

- **Elementary Reinforcement Learning**
  - 1: Introduction to Reinforcement Learning
  - 2: Markov Decision Processes
  - 3: Planning by Dynamic Programming
  - 4: Model-Free Prediction
  - 5: Model-Free Control
- **Reinforcement Learning in Practice**
  - 6: Value Function Approximation
  - 7: Policy Gradient Methods
  - 8: Integrating Learning and Planning
  - 9: Exploration and Exploitation
  - 10: Case Study: RL in Classic Games

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

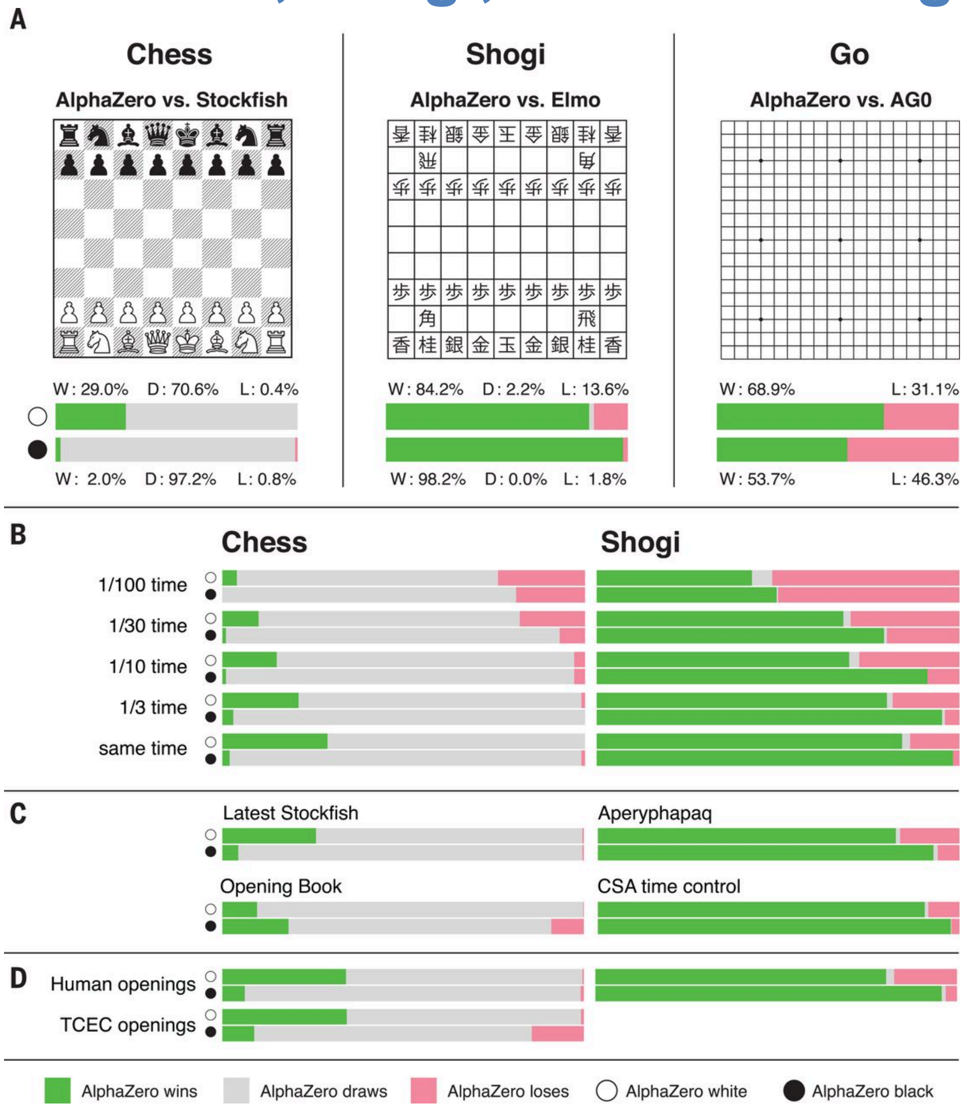
Reinforcement Learning
AlphaZero (AZ) and AlphaGo Zero (AZ0)

- AlphaZero (Silver et al., 2018)
  - A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. (Science)
- AlphaGo Zero (Silver et al., 2017)
  - Mastering the game of Go without human knowledge (Nature)

AlphaZero:
Shedding new light on the grand games of chess, shogi and Go

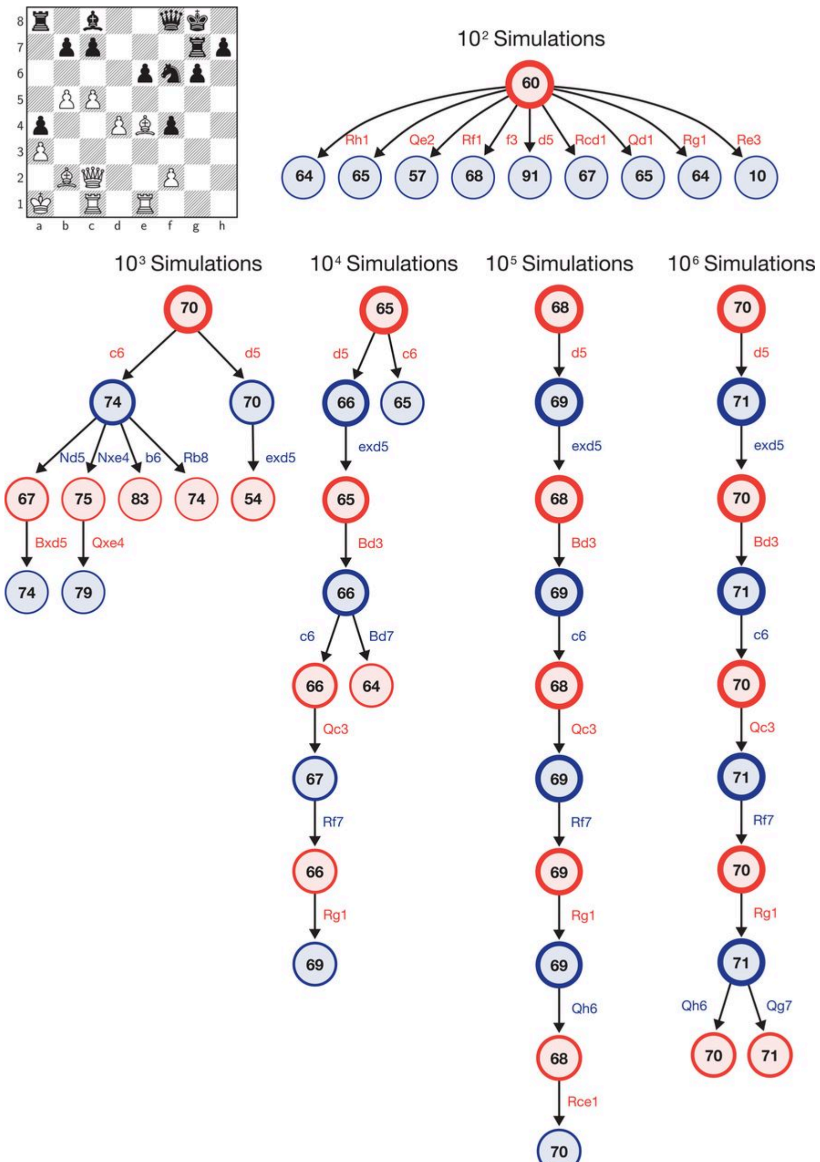
https://www.youtube.com/watch?v=7L2sUGcOgh0

AlphaZero
A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play



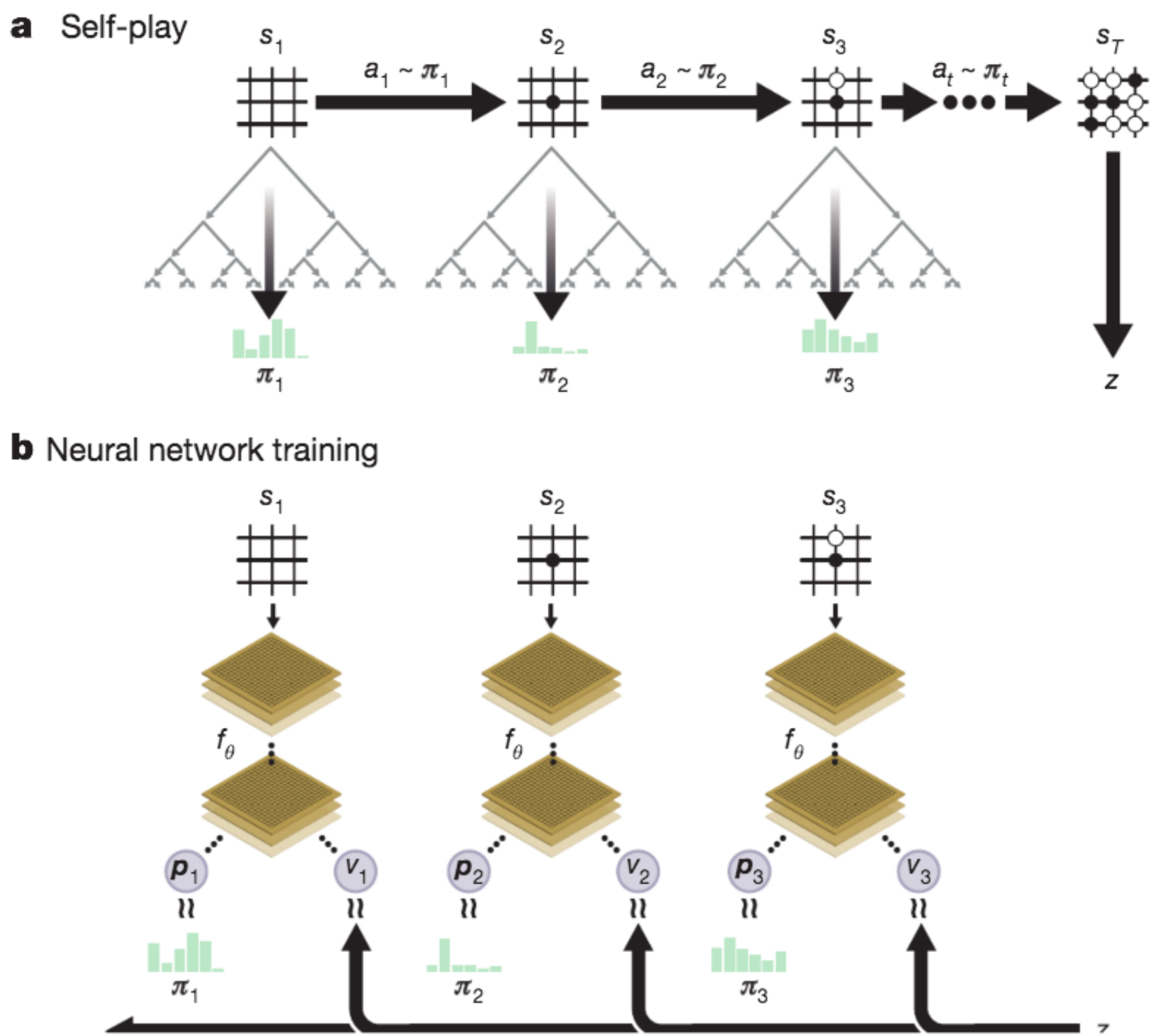
Source: David Silver et al. (2018), "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science 362, no. 6419 (2018): 1140-1144.

AlphaZero's search procedure



Source: David Silver et al. (2018), "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science 362, no. 6419 (2018): 1140-1144.

Self-play reinforcement learning in AlphaGo Zero

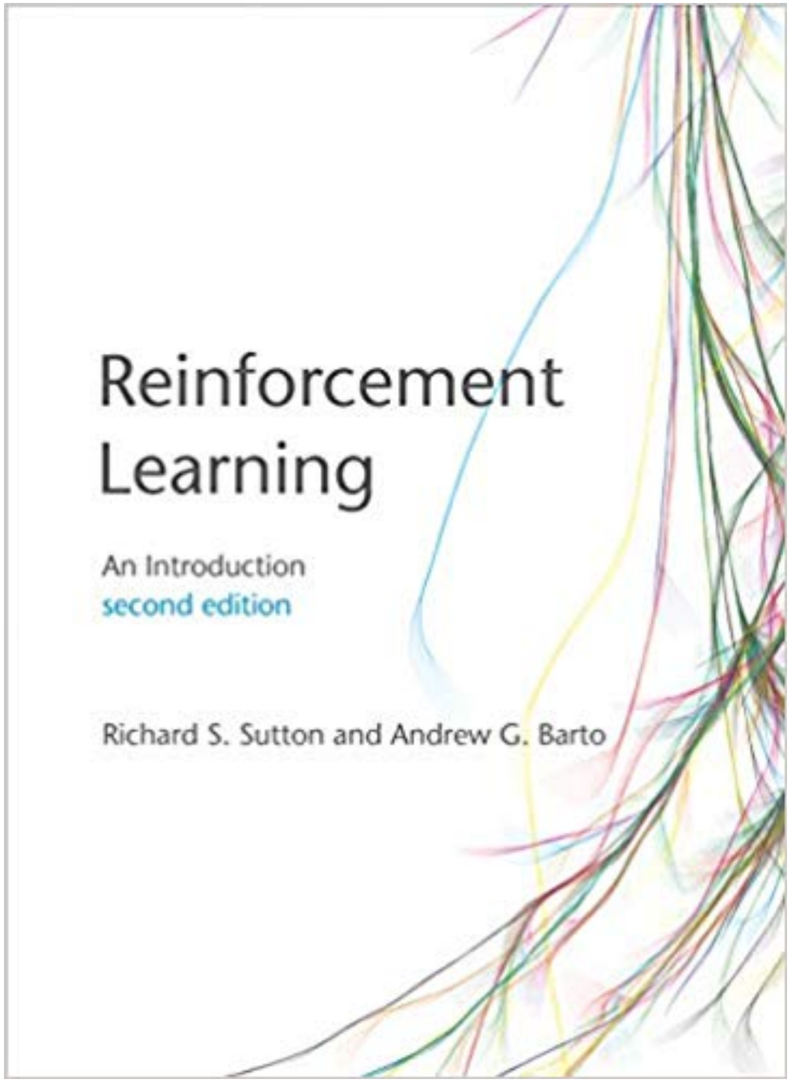


Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

Richard S. Sutton & Andrew G. Barto (2018),

Reinforcement Learning: An Introduction,

2nd Edition, A Bradford Book



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.
https://www.amazon.com/Reinforcement-Learning-Introduction-Adaptive-Computation/dp/0262039249

Reinforcement learning

- Reinforcement learning is learning what to do —how to map situations to actions —so as to maximize a numerical reward signal.

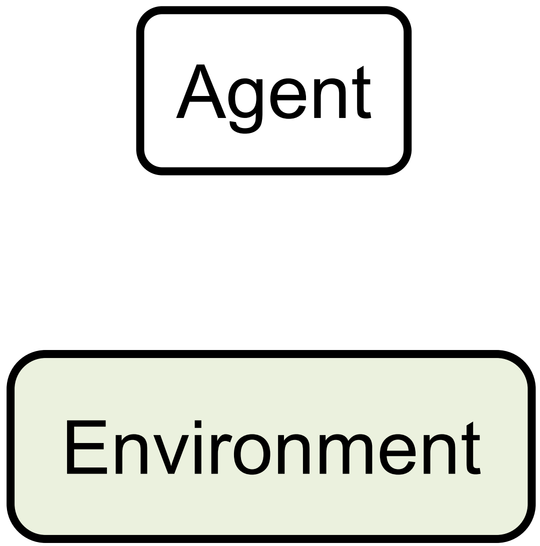
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Two most important distinguishing features of reinforcement learning

- trial-and-error search
- delayed reward

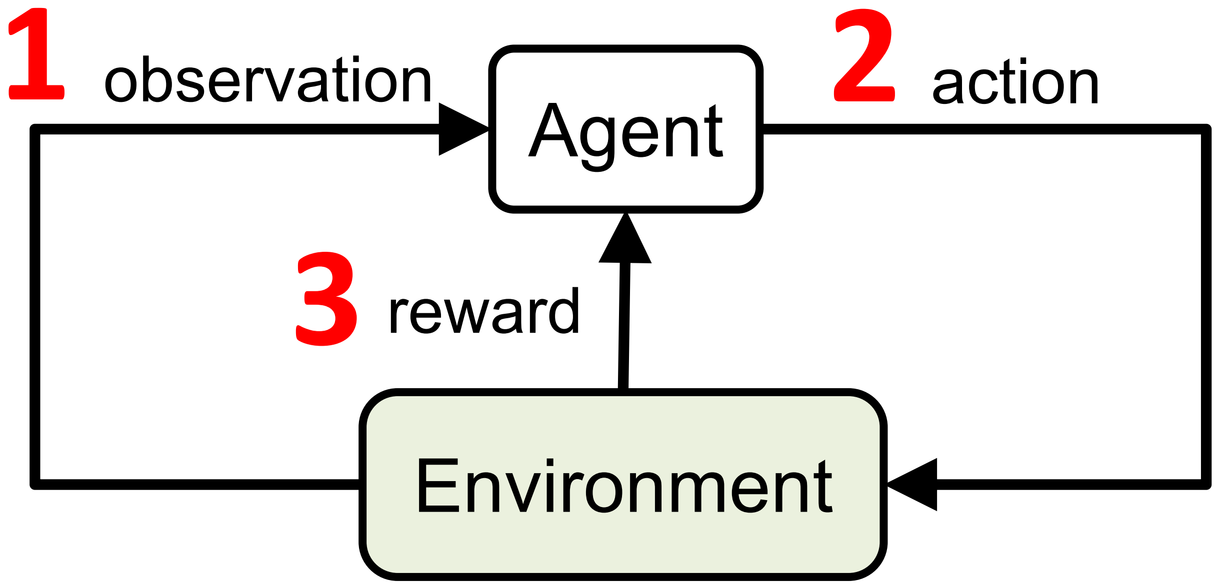
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



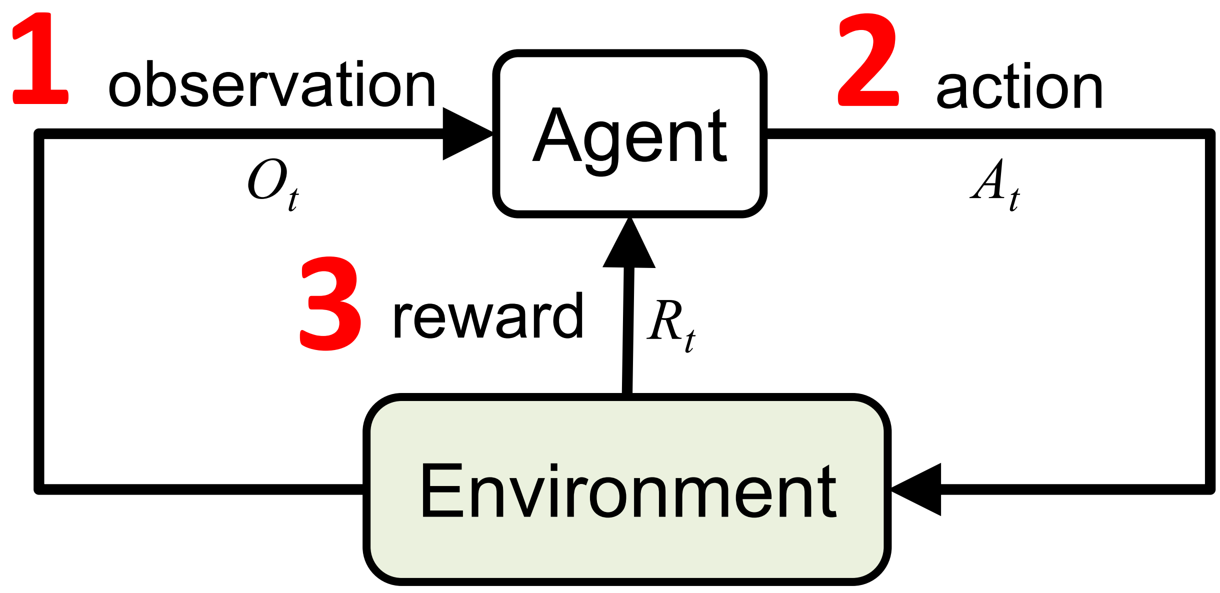
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

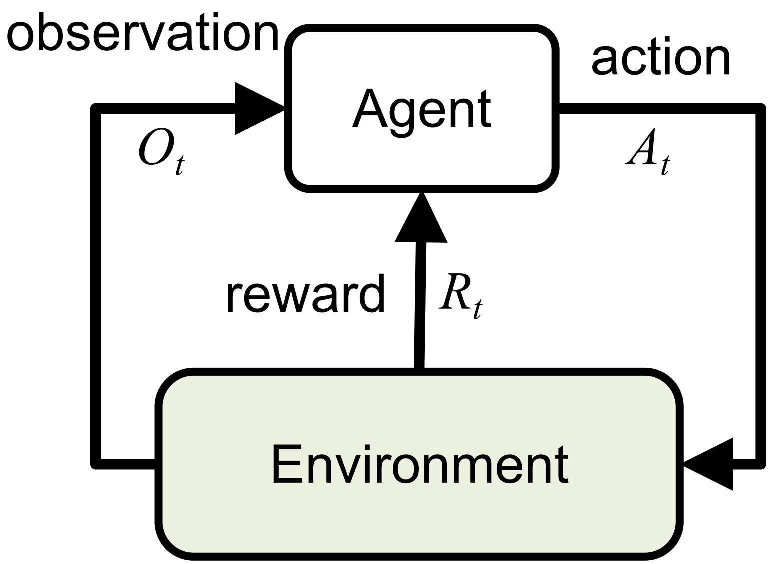
Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agent and Environment

- At each step *t* the agent:
  - Executes action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at env. step



Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

History and State

- The history is the sequence of observations, actions, rewards

$$H_t = O_1, A_1, R_1, \ldots, A_{t-1}, O_t, R_t$$

- i.e. all observable variables up to time t
- i.e. the sensorimotor stream of a robot or embodied agent
- What happens next depends on the history:
  - The agent selects actions
  - The environment selects observations/rewards
- State is the information used to determine what happens next
- Formally, state is a function of the history:

$$S_t = f(H_t)$$

Information State

- An information state (a.k.a. Markov state) contains all useful information from the history.
- Definition

  A state S_t is Markov if and only if

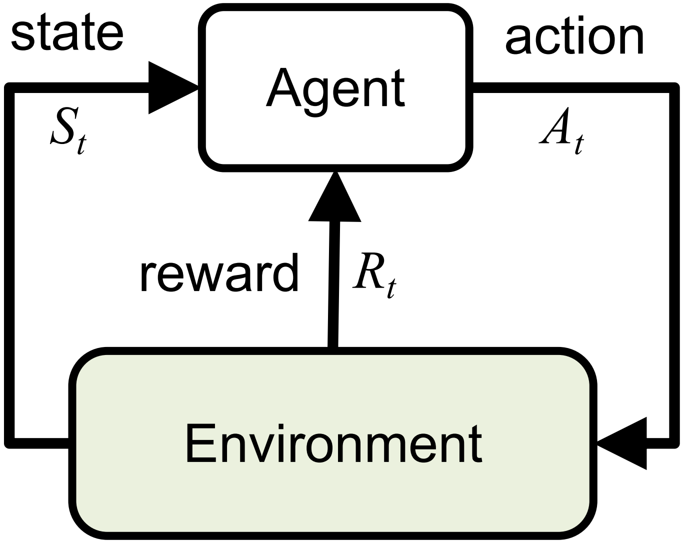
  $$P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S_1, \ldots, S_t]$$
- "The future is independent of the past given the present"

  $$H_{1:t} \rightarrow S_t \rightarrow H_{t+1:\infty}$$
- Once the state is known, the history may be thrown away i.e. The state is a sufficient statistic of the future
- The environment state S_t^e is Markov
- The history H_t is Markov

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Fully Observable Environments

- Full observability:
  - agent directly observes environment state
  - Agent state = environment state = information state
  - Formally, this is a Markov decision process (MDP)



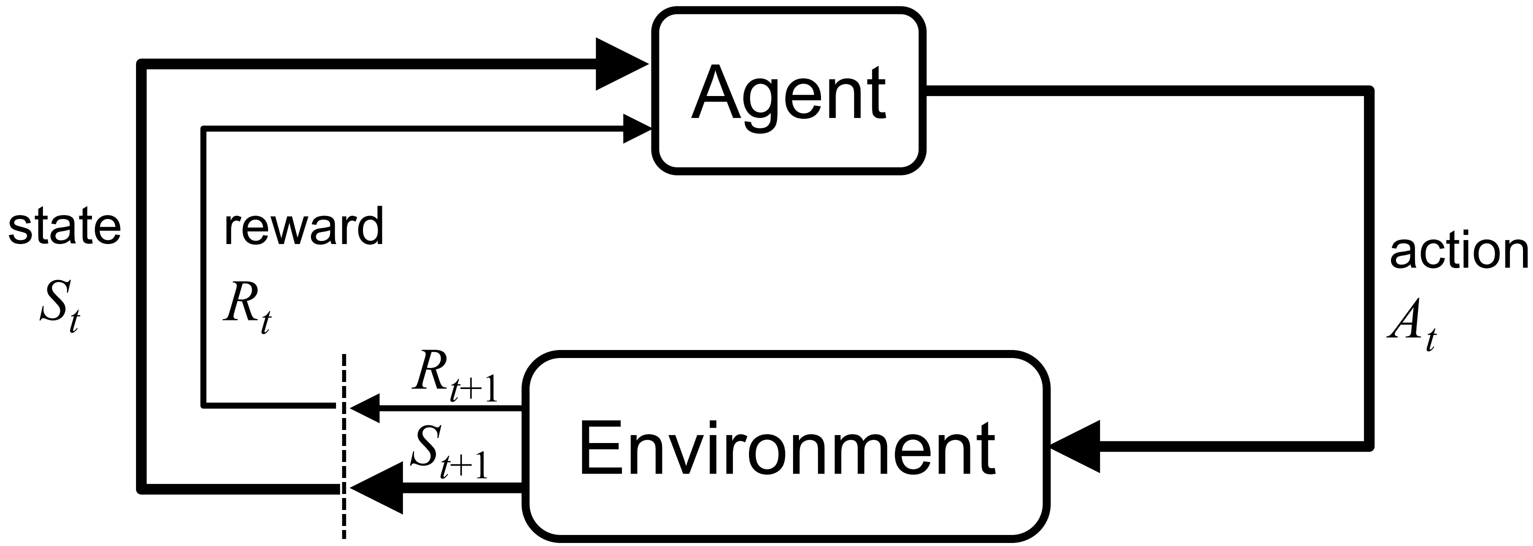
Partially Observable Environments

- Partial observability: agent indirectly observes environment
  - A robot with camera vision isn't told its absolute location
  - A trading agent only observes current prices
  - A poker playing agent only observes public cards
- Now agent state ≠ environment state
- Formally this is a partially observable Markov decision process (POMDP)
- Agent must construct its own state representation S^a_t, e.g.
  - Complete history: $S^a_t = H_t$
  - Beliefs of environment state: $S^a_t = (P[S^e_t = s_1],\ldots,P[S^e_t = s_n])$
  - Recurrent neural network: $S^a_t = \sigma(S^a_{t-1} W_s + O_t W_o)$

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Reinforcement Learning (DL)

The Agent-Environment Interaction in a Markov Decision Process (MDP)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Characteristics of Reinforcement Learning

- No supervisor, only a reward signal
- Feedback is delayed, not instantaneous
- Time really matters (sequential, non i.i.d data)
- Agent's actions affect the subsequent data it receives

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Examples of Reinforcement Learning

- Make a humanoid robot walk
- Play may different Atari games better than humans
- Manage an investment portfolio

Examples of Rewards

- Make a humanoid robot walk
  - +ve reward for forward motion
  - -ve reward for falling over
- Play may different Atari games better than humans
  - +/-ve reward for increasing/decreasing score
- Manage an investment portfolio
  - +ve reward for each $ in bank

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Sequential Decision Making

- Goal: select actions to maximize total future reward
- Actions may have long term consequence
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward
- Examples:
  - A financial investment (may take months to mature)
  - Blocking opponent moves (might help winning chances many moves from now)

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Elements of Reinforcement Learning

- Agent
- Environment
- Policy
- Reward signal
- Value function
- Model

Elements of Reinforcement Learning

- Policy
  - Agent’s behavior
  - It is a map from state to action
- Reward signal
  - The goal of a reinforcement learning problem
- Value function
  - How good is each state and/or action
  - A prediction of future reward
- Model
  - Agent’s representation of the environment

Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Major Components of an RL Agent

1. Policy: agent’s behaviour function
2. Value function: how good is each state and/or action
3. Model: agent’s representation of the environment

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Policy

- A policy is the agent's behaviour
- It is a map from state to action, e.g.
  - Deterministic policy: $a = \pi(s)$
  - Stochastic policy: $\pi(a|s) = P[A_t = a|S_t = s]$

Value Function

- Value function is a prediction of future reward
- Used to evaluate the goodness/badness of states
- And therefore to select between actions, e.g.

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s]$$

Model

- A model predicts what the environment will do next
- P predicts the next state
- R predicts the next (immediate) reward, e.g.

$$P^a_{ss'} = P[S_{t+1} = s' \mid S_{t+1} = s, A_t = a]$$

$$R^a_s = E[R_{t+1} \mid S_t = s, A_t = a]$$

Reinforcement Learning

- Value Based
  - No Policy (Implicit)
  - Value Function
- Policy Based
  - Policy
  - No Value Function
- Actor Critic
  - Policy
  - Value Function

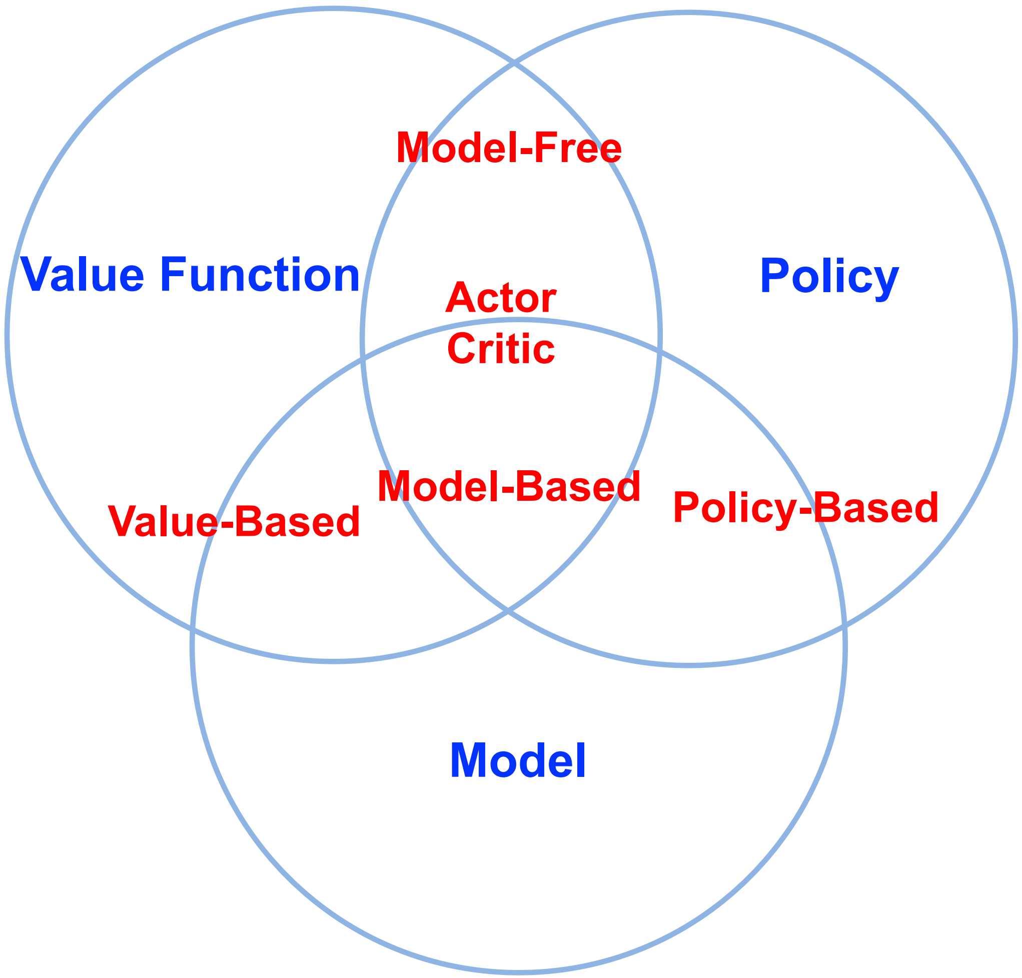
Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Reinforcement Learning

- Model Free
  - Policy and/or Value Function
  - No Model
- Model Based
  - Policy and/or Value Function
  - Model

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Reinforcement Learning (RL) Taxonomy



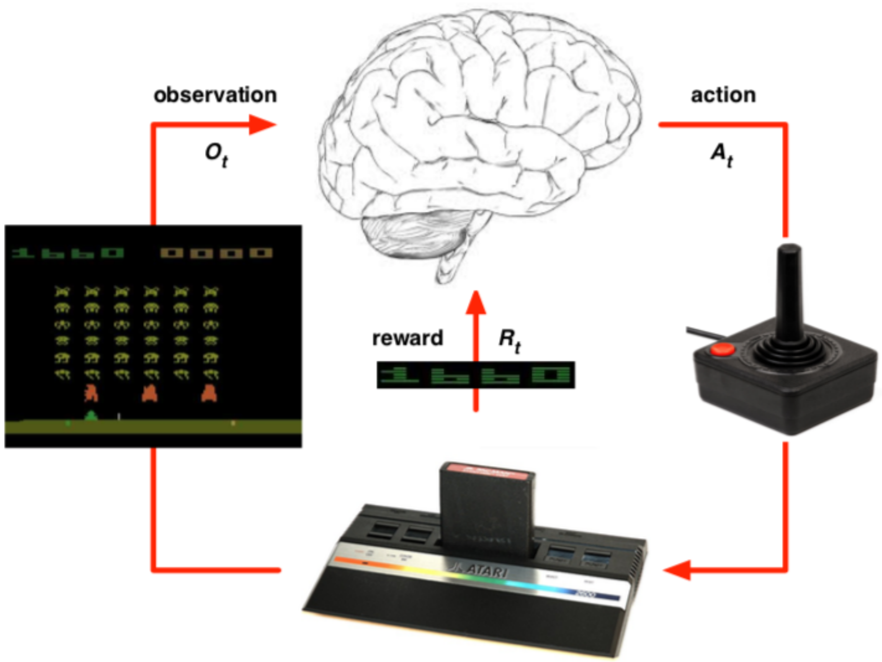
Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Learning and Planning

- Two fundamental problems in sequential decision making
  - Reinforcement Learning
    - The environment is initially unknown
    - The agent interacts with environment
    - The agent improves its policy
  - Planning
    - A model of the environment is known
    - The agent performs computations with its model (without any external interaction)
    - The agent improves its policy
    - a.k.a deliberation, reasoning, introspection, pondering, thought, search

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Atari Example: Reinforcement Learning

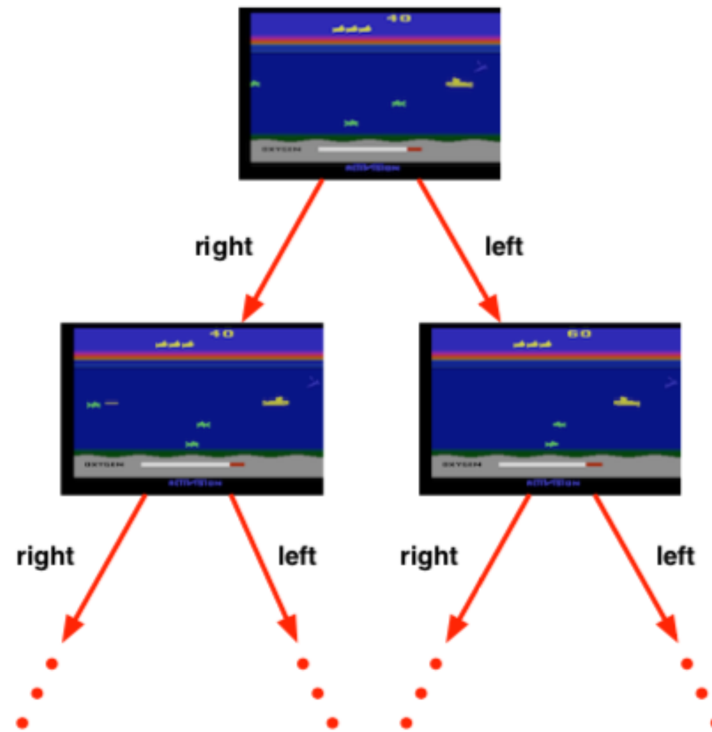


- Rules of the game are unknown
- Learn directly from interactive game-play
- Pick actions on joystick, see pixels and scores

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Atari Example: Planning

- Rules of the game are known
- Can query emulator
  – perfect model inside agent's brain
- If I take action a from state s:
  – what would the next state be?
  – what would the score be?
- Plan ahead to find optimal policy
  – e.g. tree search



Exploration and Exploitation

- Reinforcement learning is like trial-and-error learning
- The agent should discover a good policy
- From its experiences of the environment
- Without losing too much reward along the way
- **Exploration** finds more information about the environment
- **Exploitation** exploits known information to maximise reward
- It is usually important to explore as well as exploit

Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Exploration and Exploitation Examples

- Restaurant Selection
  - Exploitation: Go to your favorite restaurant
  - Exploration: Try a new restaurant Online Banner
- Advertisements
  - Exploitation: Show the most successful advert
  - Exploration: Show a different advert
- Oil Drilling
  - Exploitation: Drill at the best known location
  - Exploration: Drill at a new location
- Game Playing
  - Exploitation: Play the move you believe is best
  - Exploration: Play an experimental move

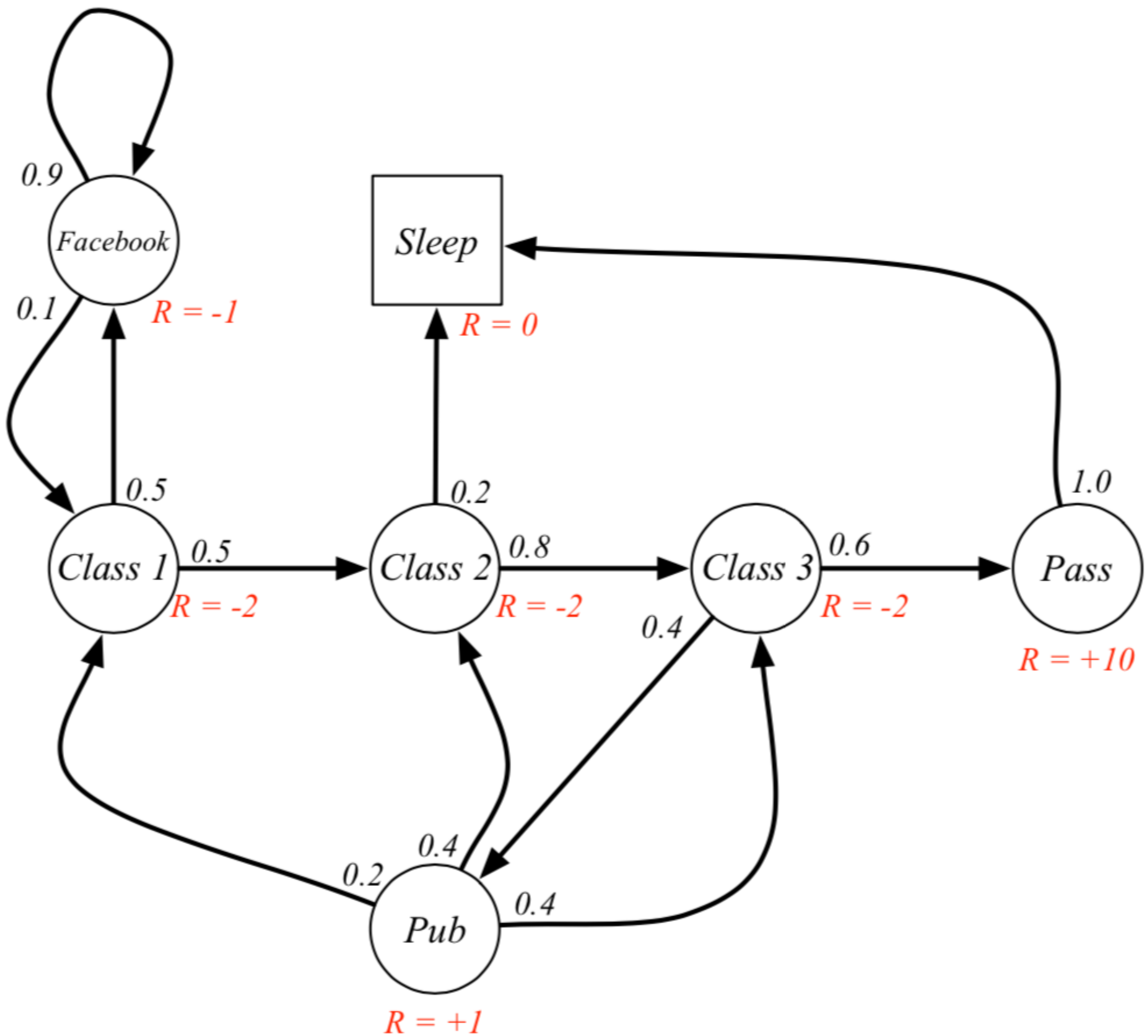
Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Prediction and Control

- Prediction: evaluate the future
  - Given a policy
- Control: optimize the future
  - Find the best policy

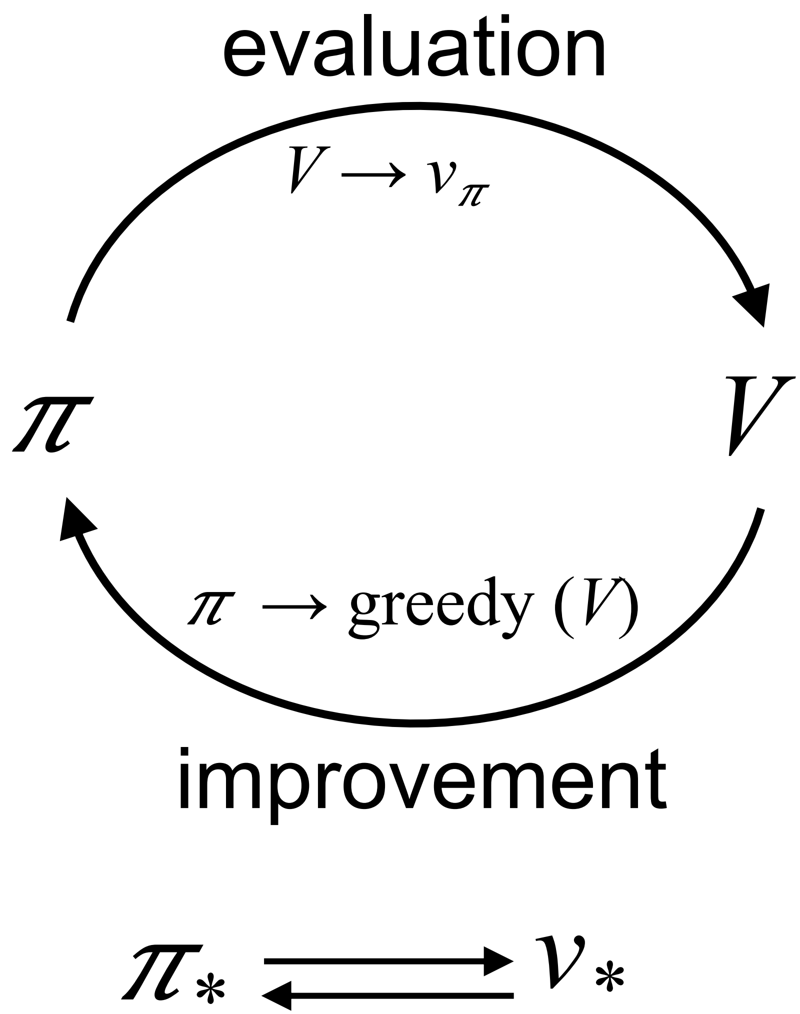
Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

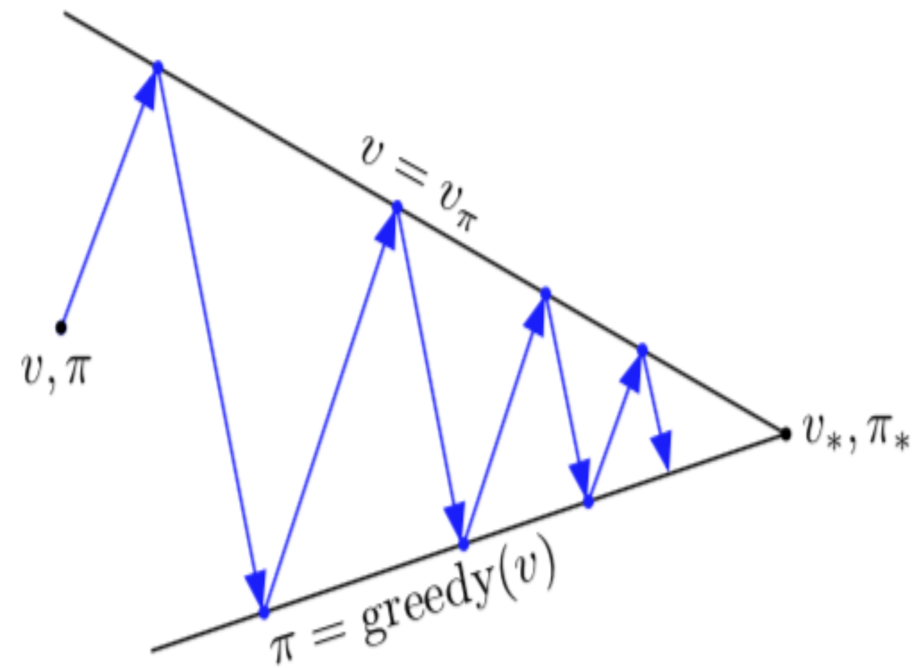
Markov Decision Processes (MDP)
Example: Student MRP



Source: David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ

Generalized Policy Iteration (GPI)

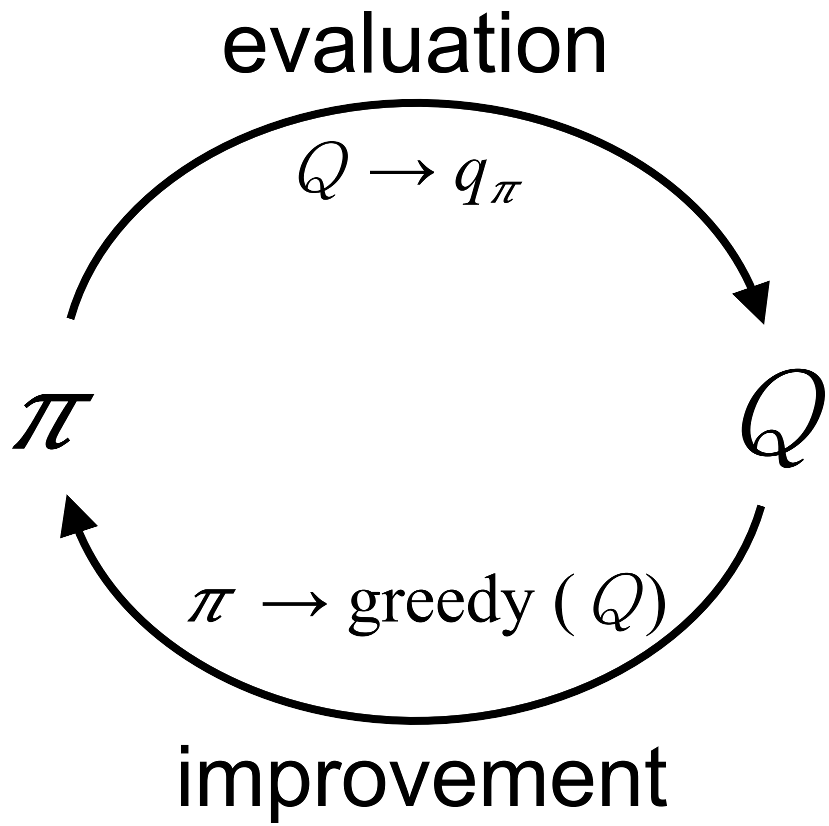




Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Generalized Policy Iteration (GPI)

Any iteration of **policy evaluation** and **policy improvement**, independent of their granularity.



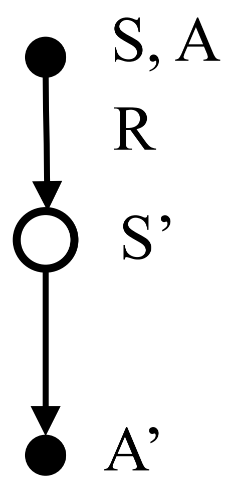
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Temporal-Difference (TD) Learning

- Sarsa: On-policy TD Control
- Q-learning: Off-policy TD Control

SARSA
(state-action-reward-state-action)
On-policy TD Control

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma\ Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\ ]$$

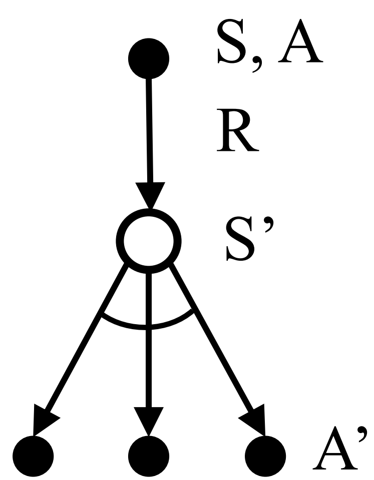


SARSA

Q-learning (Watkins, 1989)

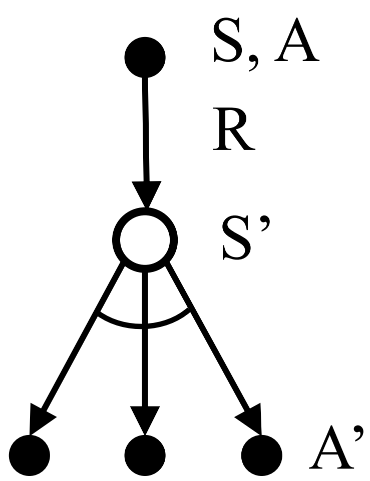
Off-policy TD Control

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max Q(S_{t+1}, a) - Q(S_t, A_t)]$$

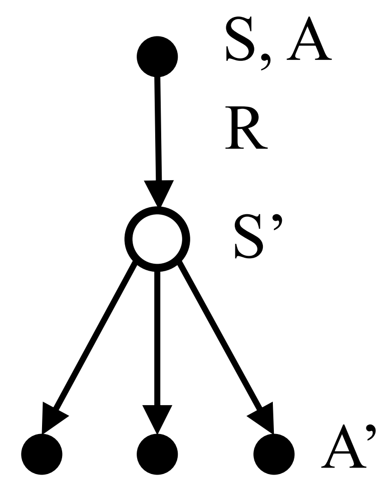


Q-learning

Q-learning and Expected SARSA



Q-learning



Expected SARSA

Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Q-learning and Double Q-learning

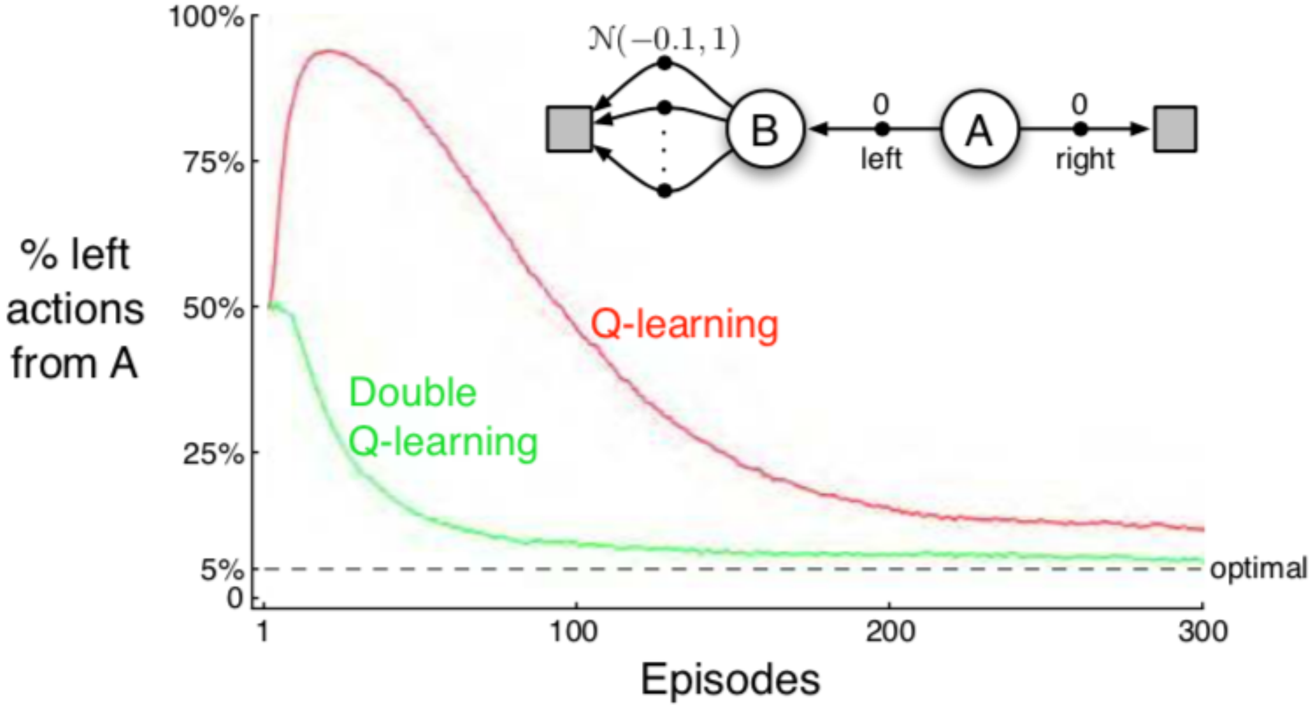


Figure 6.5: Comparison of Q-learning and Double Q-learning on a simple episodic MDP (shown inset). Q-learning initially learns to take the left action much more often than the right action, and always takes it significantly more often than the 5% minimum probability enforced by ε-greedy action selection with $\varepsilon = 0.1$. In contrast, Double Q-learning is essentially unaffected by maximization bias. These data are averaged over 10,000 runs. The initial action-value estimates were zero. Any ties in ε-greedy action selection were broken randomly.

Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

n-step methods for sate-action value

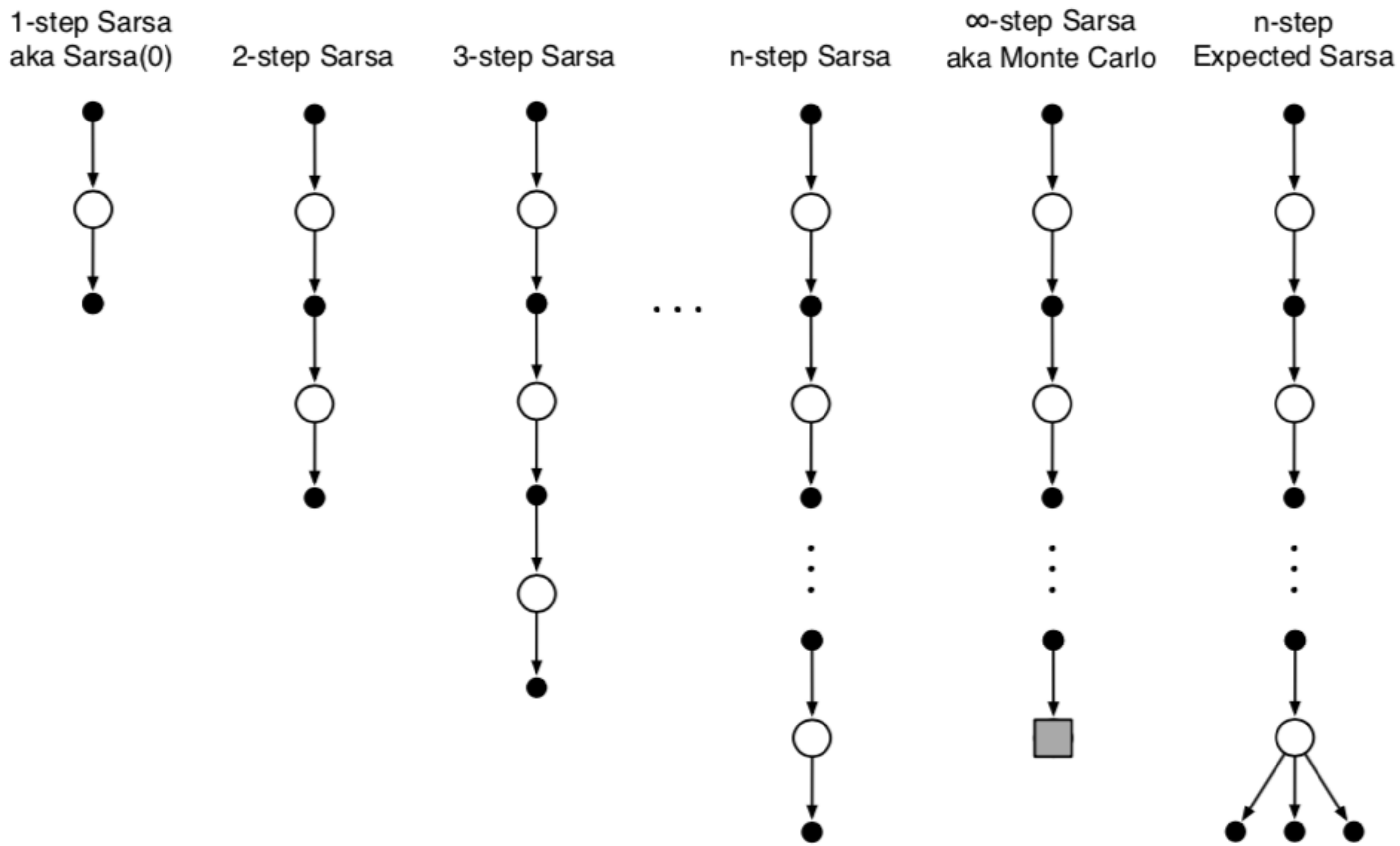
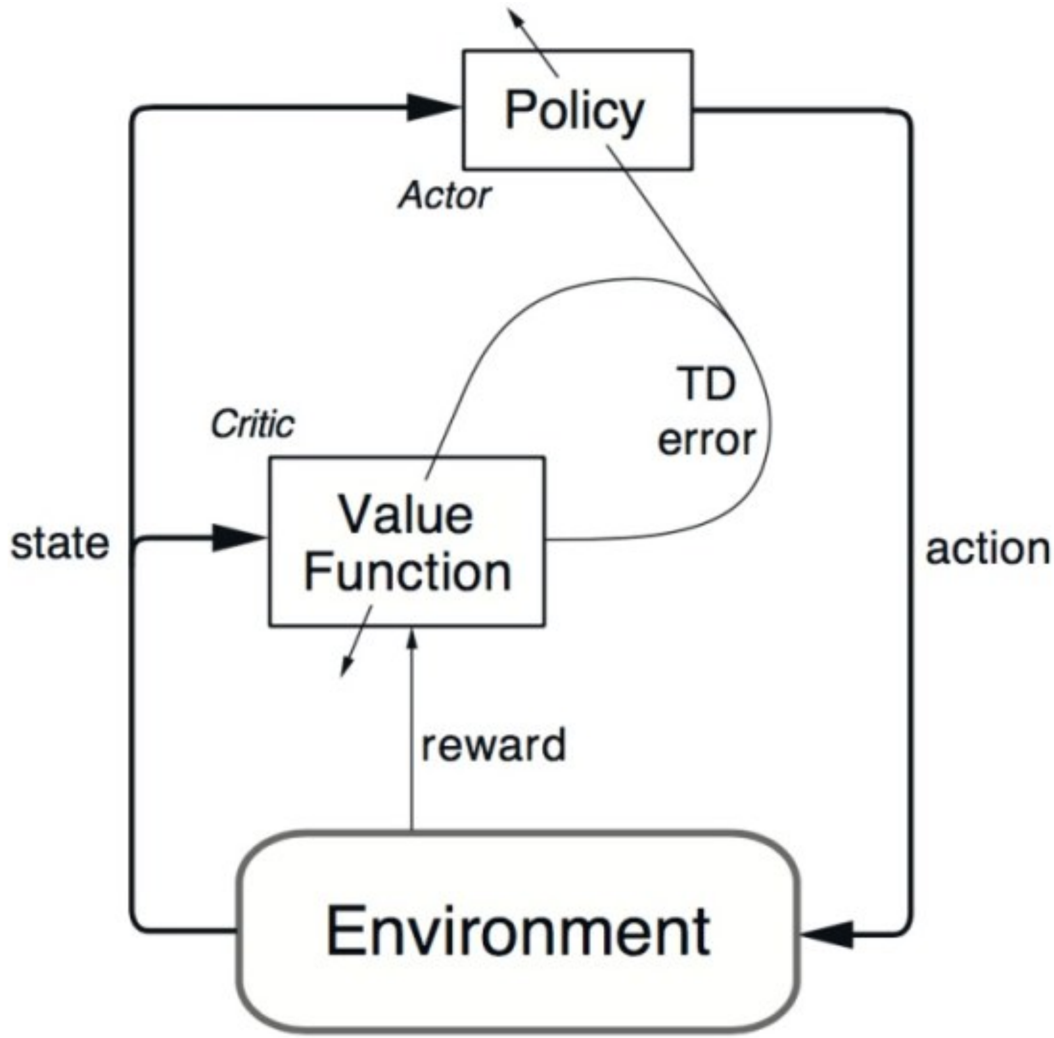


Figure 7.3: The backup diagrams for the spectrum of n-step methods for state–action values. They range from the one-step update of Sarsa(0) to the up-until-termination update of the Monte Carlo method. In between are the n-step updates, based on n steps of real rewards and the estimated value of the nth next state–action pair, all appropriately discounted. On the far right is the backup diagram for n-step Expected Sarsa.

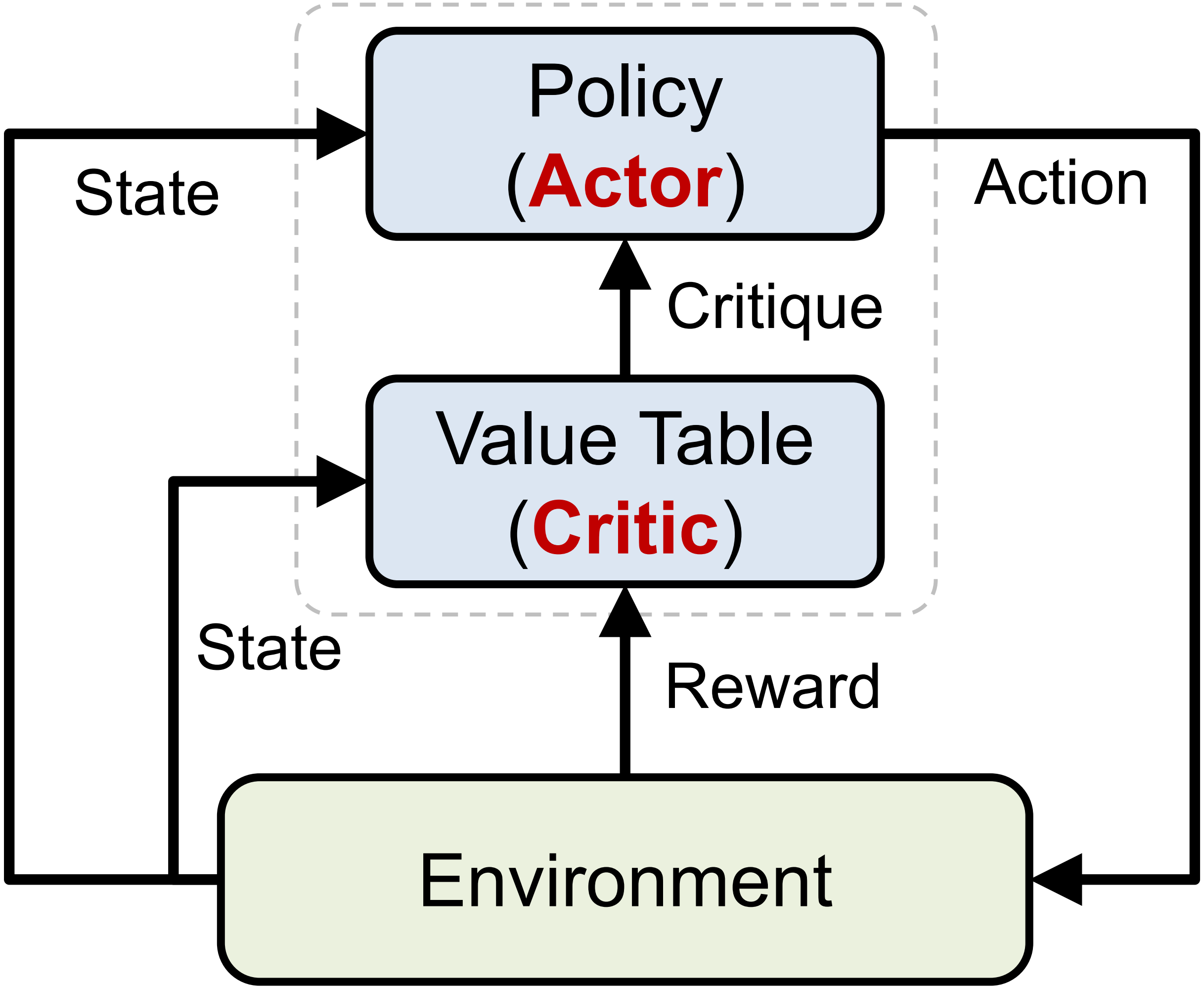
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning
Actor-Critic (AC) Architecture



Source: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

Reinforcement Learning
Actor-Critic (AC) Learning Methods



Reinforcement Learning Methods

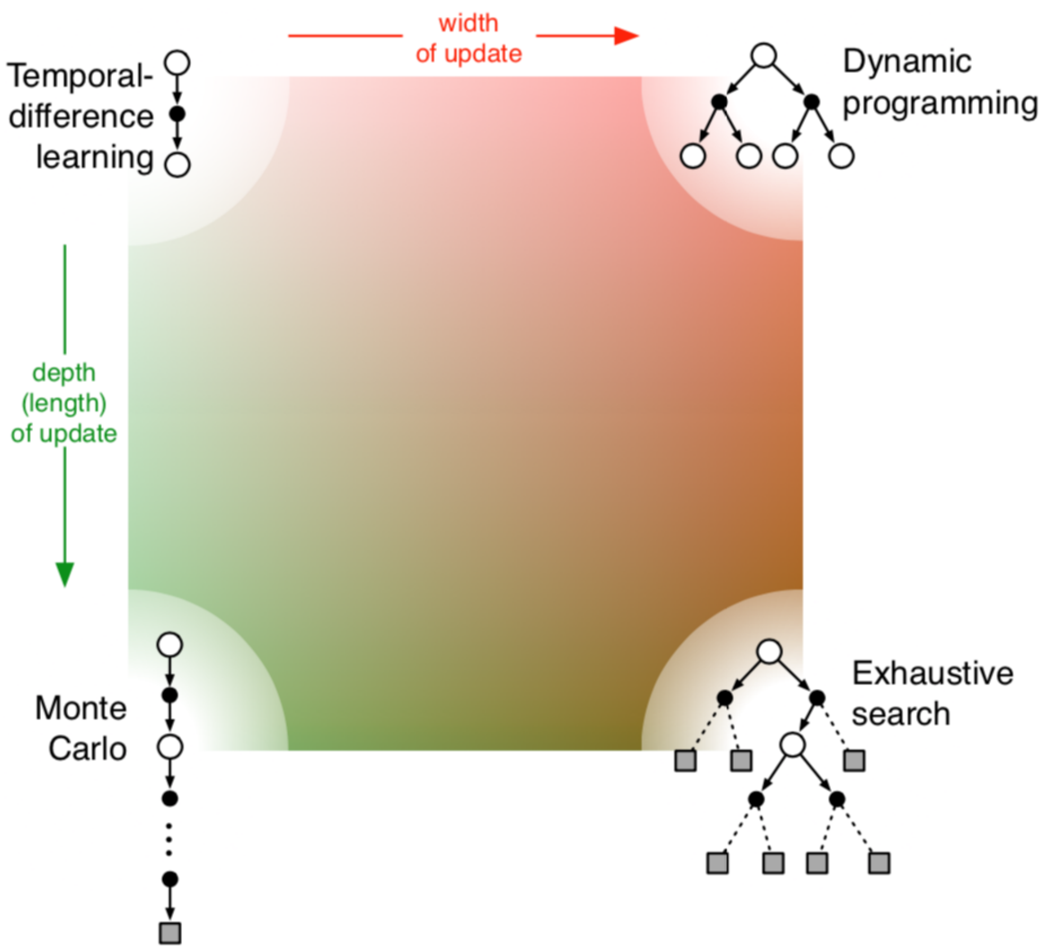


Figure 8.11: A slice through the space of reinforcement learning methods, highlighting the two of the most important dimensions explored in Part I of this book: the depth and width of the updates.

Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Monte Carlo Tree Search (MCTS)

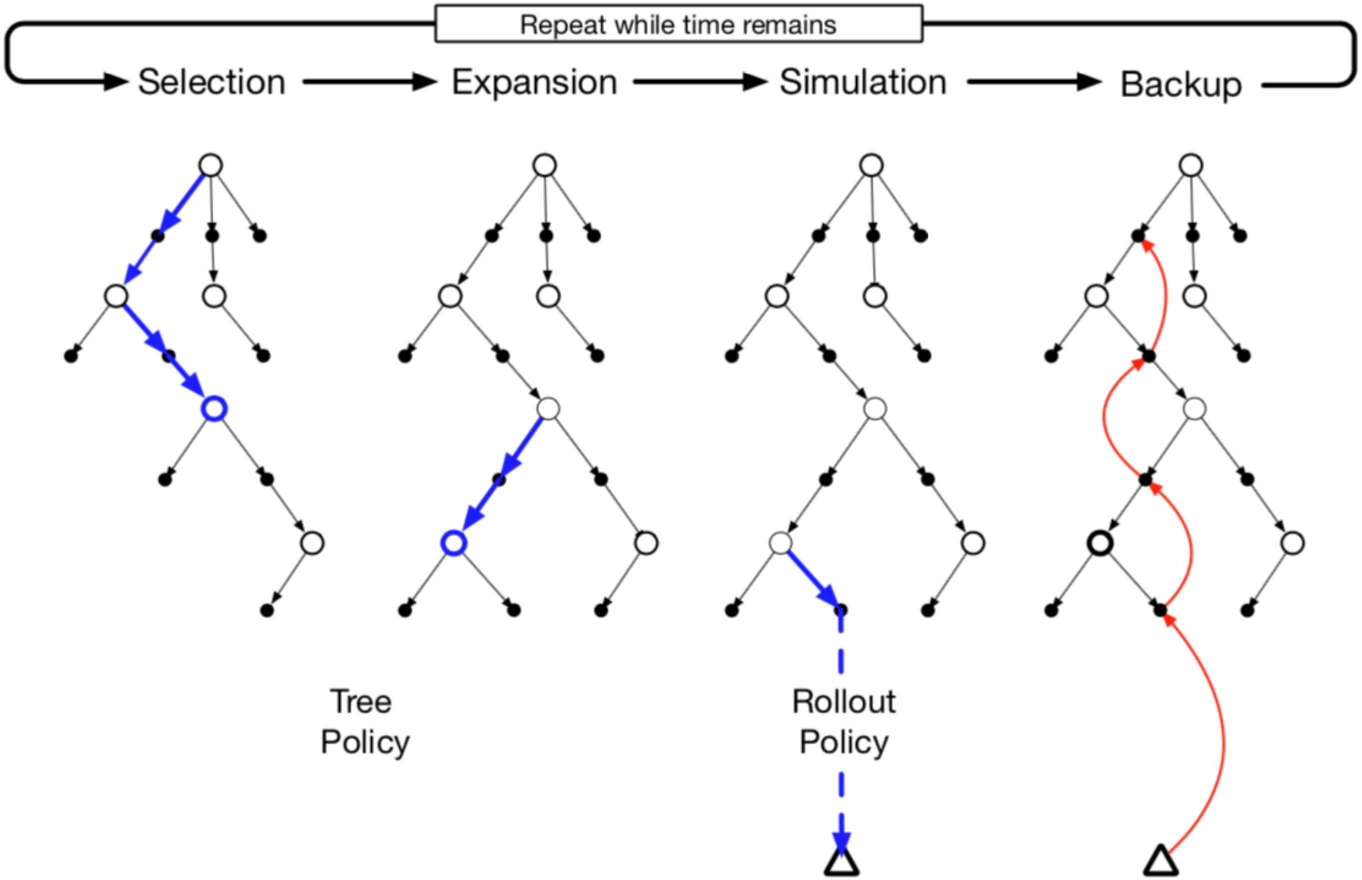
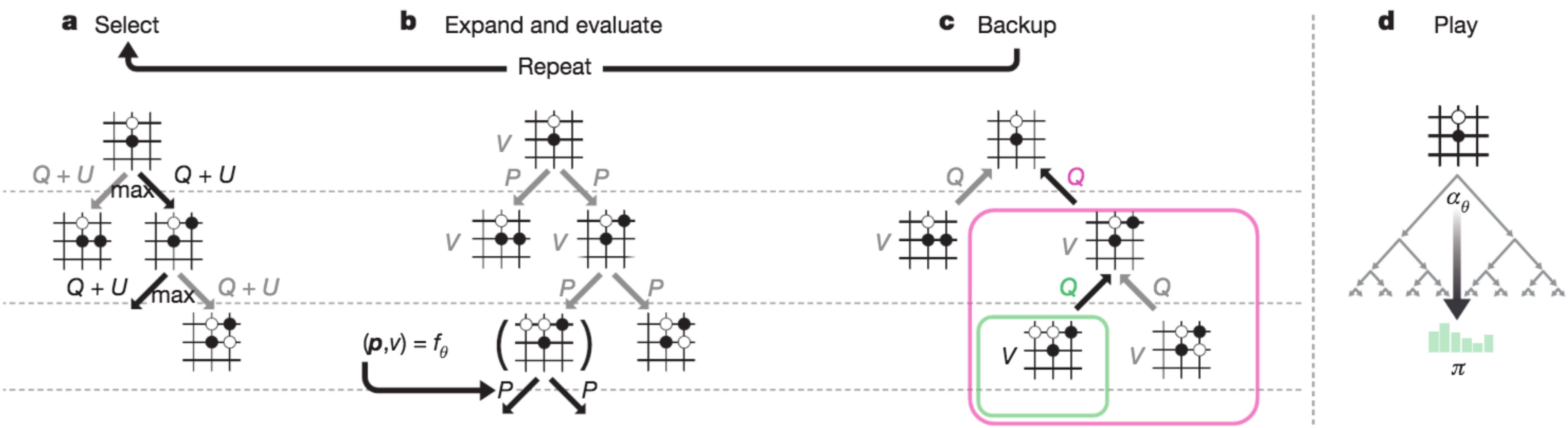


Figure 8.10: Monte Carlo Tree Search. When the environment changes to a new state, MCTS executes as many iterations as possible before an action needs to be selected, incrementally building a tree whose root node represents the current state. Each iteration consists of the four operations **Selection**, **Expansion** (though possibly skipped on some iterations), **Simulation**, and **Backup**, as explained in the text and illustrated by the bold arrows in the trees. Adapted from Chaslot, Bakkes, Szita, and Spronck (2008).

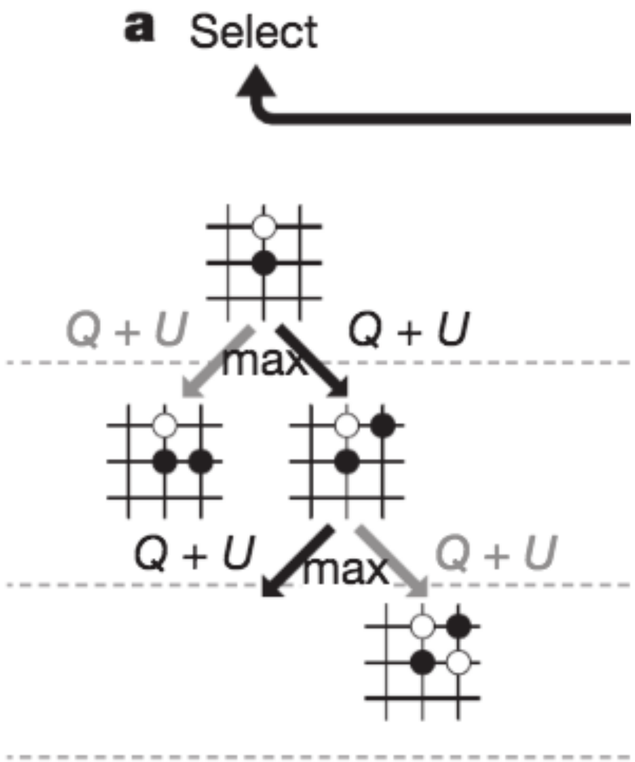
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Monte Carlo Tree Search (MCTS)
MCTS in AlphaGo Zero



Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

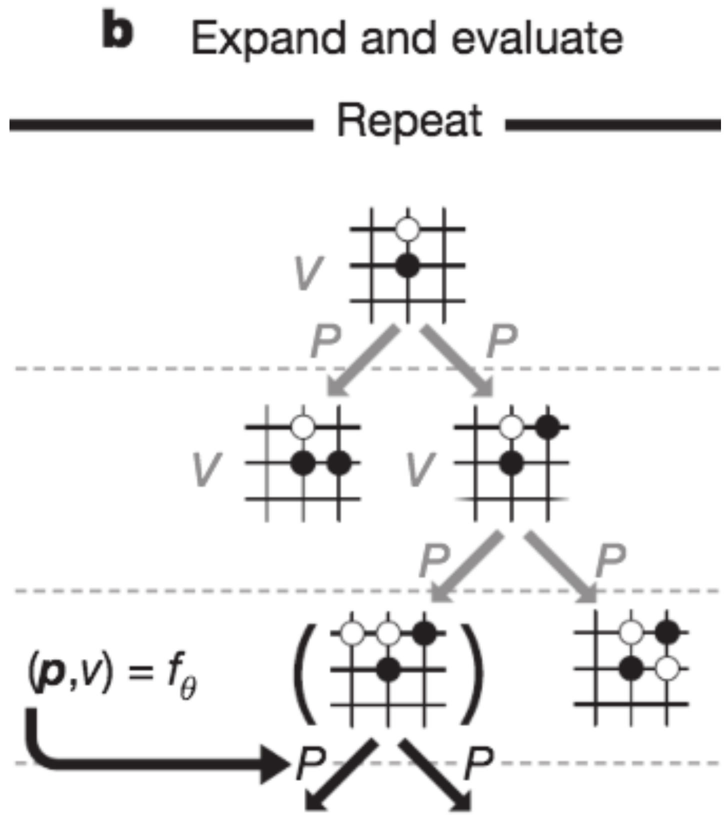
MCTS in AlphaGo Zero



a: Each simulation traverses the tree by selecting the edge with maximum action value Q, plus an upper confidence bound U that depends on a stored prior probability P and visit count N for that edge (which is incremented once traversed).

Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

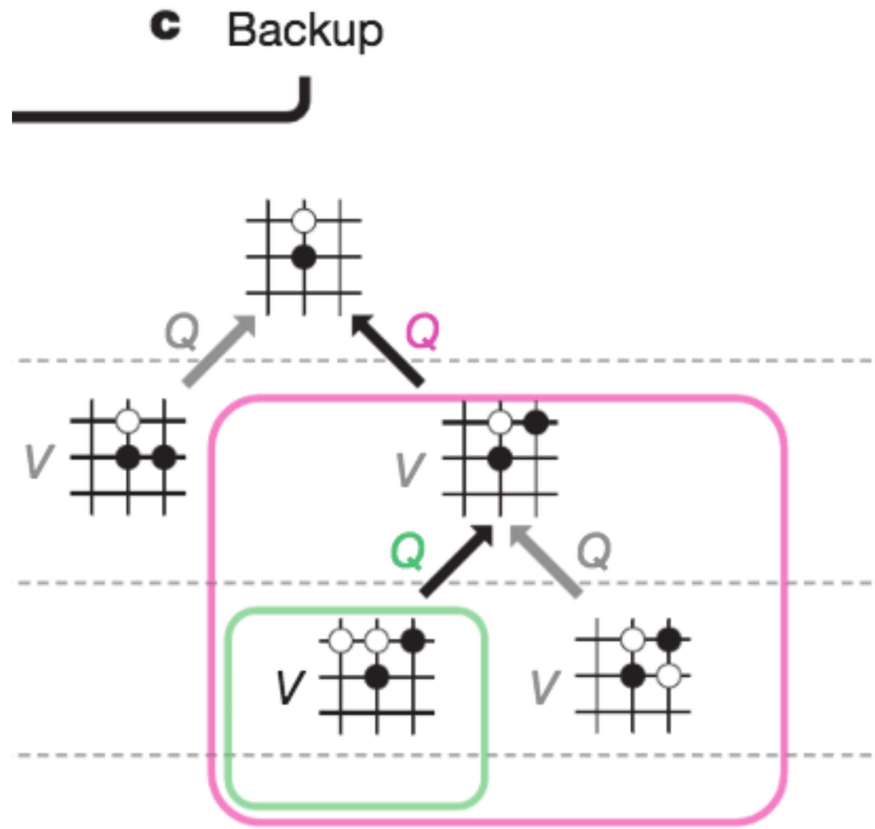
MCTS in AlphaGo Zero



b: The leaf node is expanded and the associated position s is evaluated by the neural network (P(s, ·),V(s)) = $f_\theta(s)$; the vector of P values are stored in the outgoing edges from s.

Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

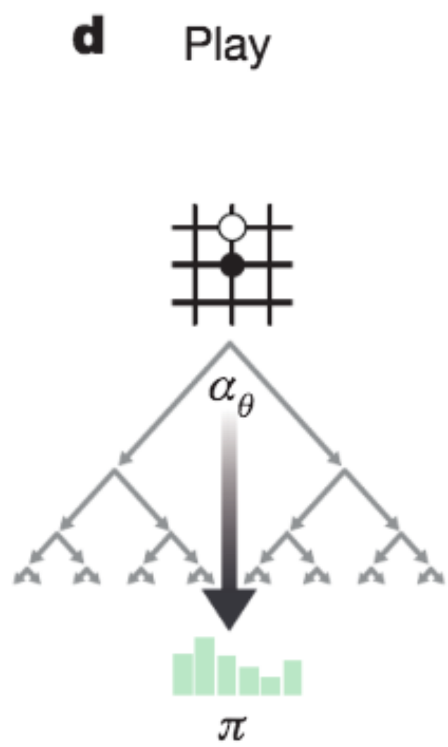
MCTS in AlphaGo Zero



c: Action value Q is updated to track the mean of all evaluations V in the subtree below that action

Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

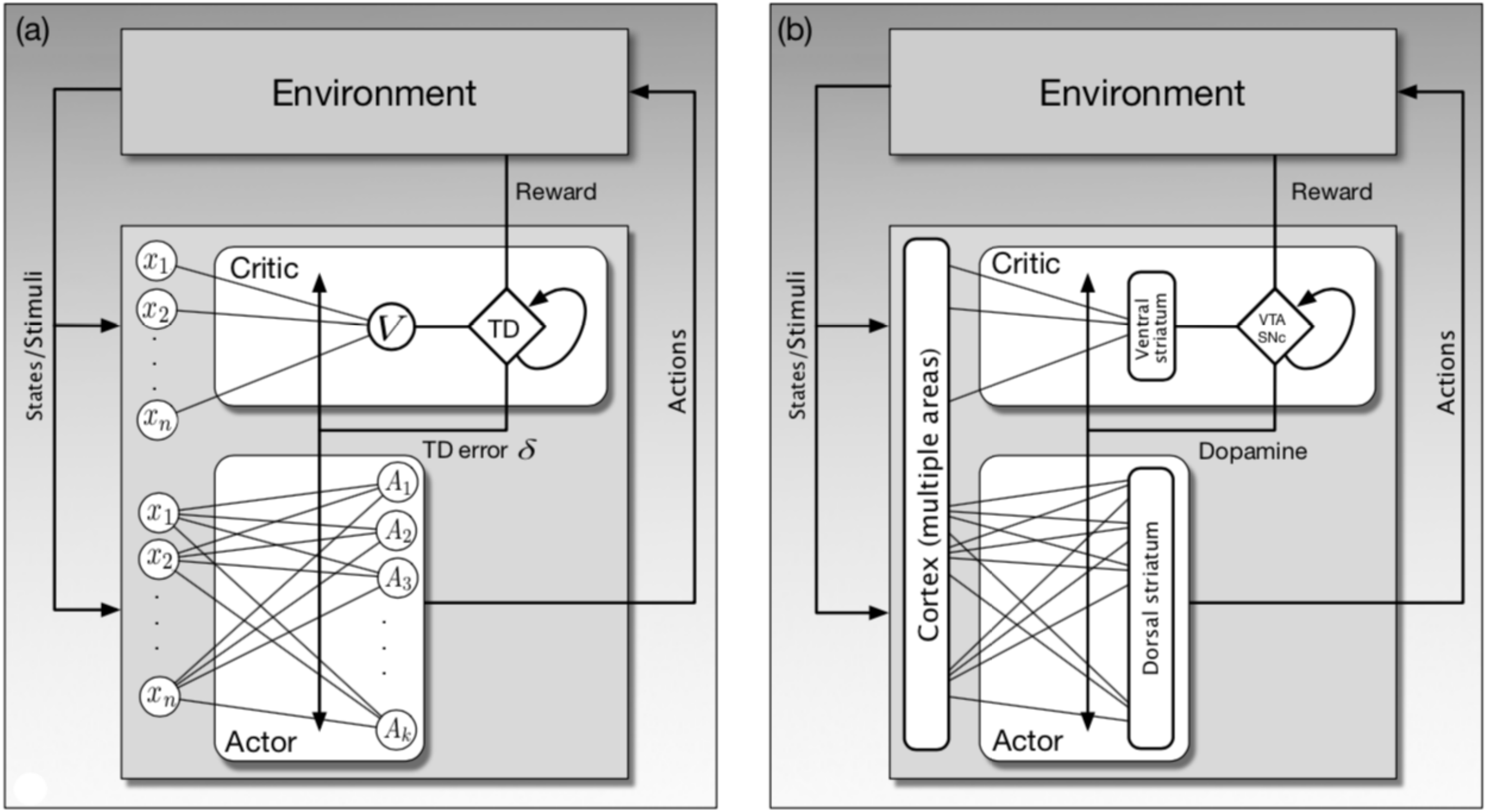
MCTS in AlphaGo Zero



d: Once the search is complete, search probabilities π are returned, proportional to $N^{1/\tau}$, where N is the visit count of each move from the root state and τ is a parameter controlling temperature.

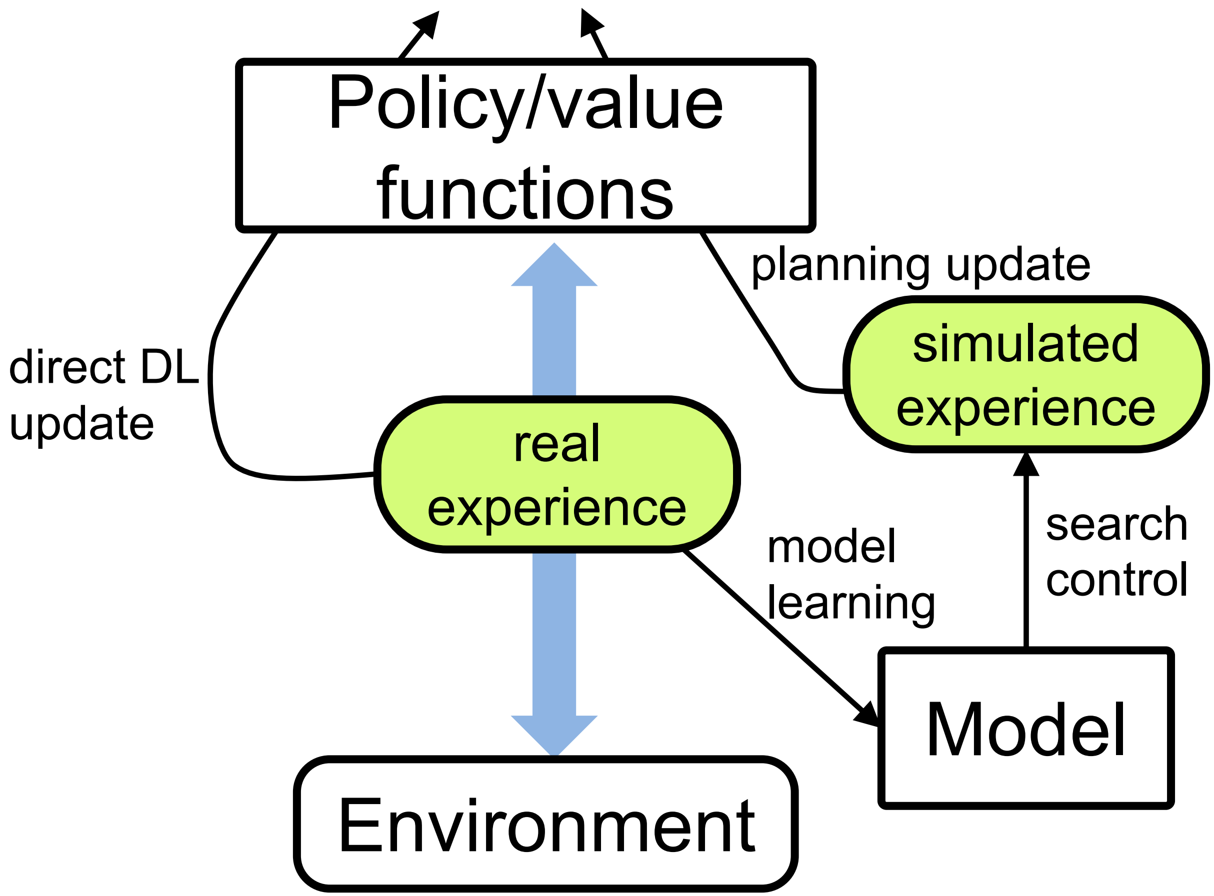
Source: David Silver et al. (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.

Reinforcement Learning
Actor Critic ANN



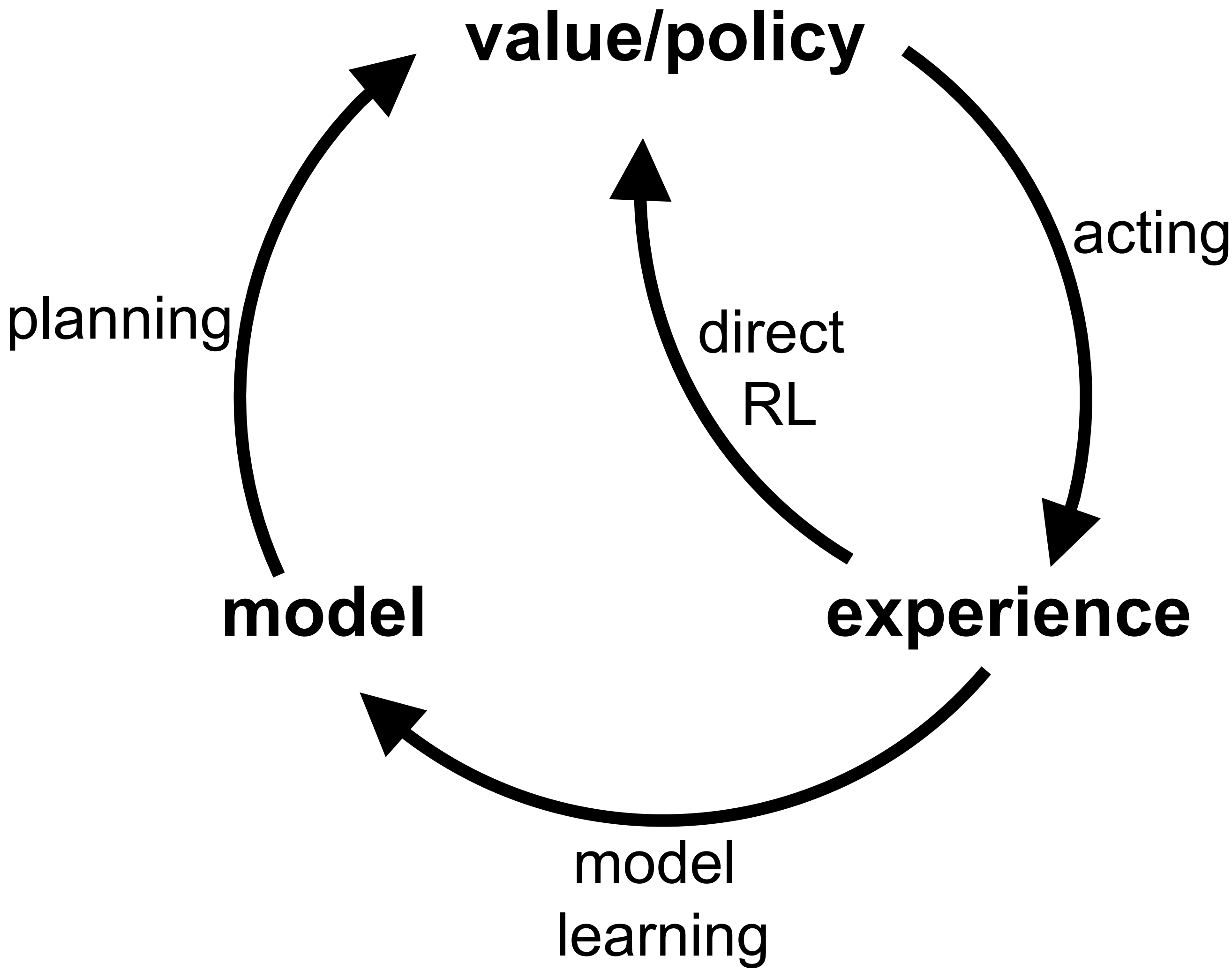
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning
General Dyna Architecture



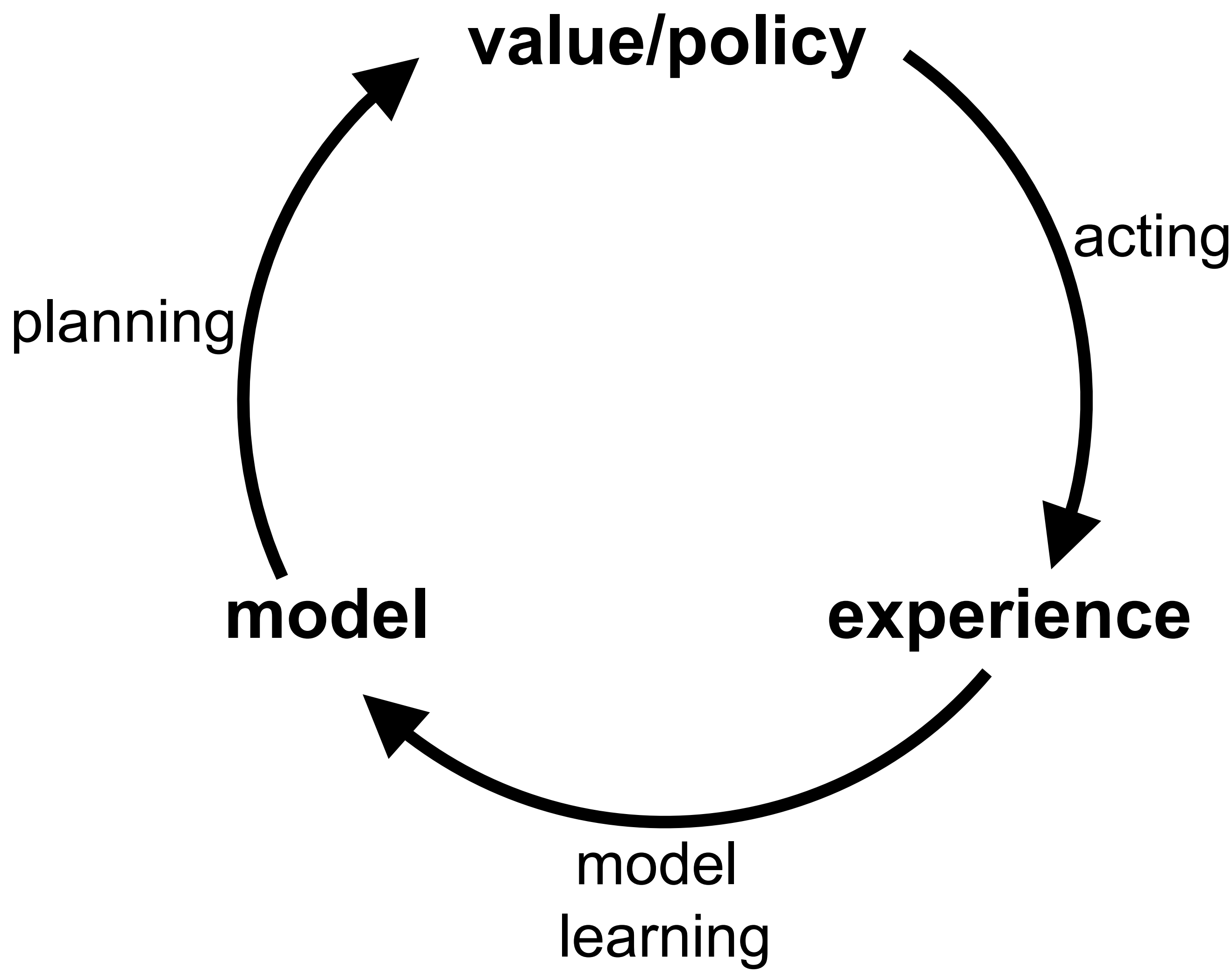
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Dyna:
Integrated Planning, Acting, and Learning



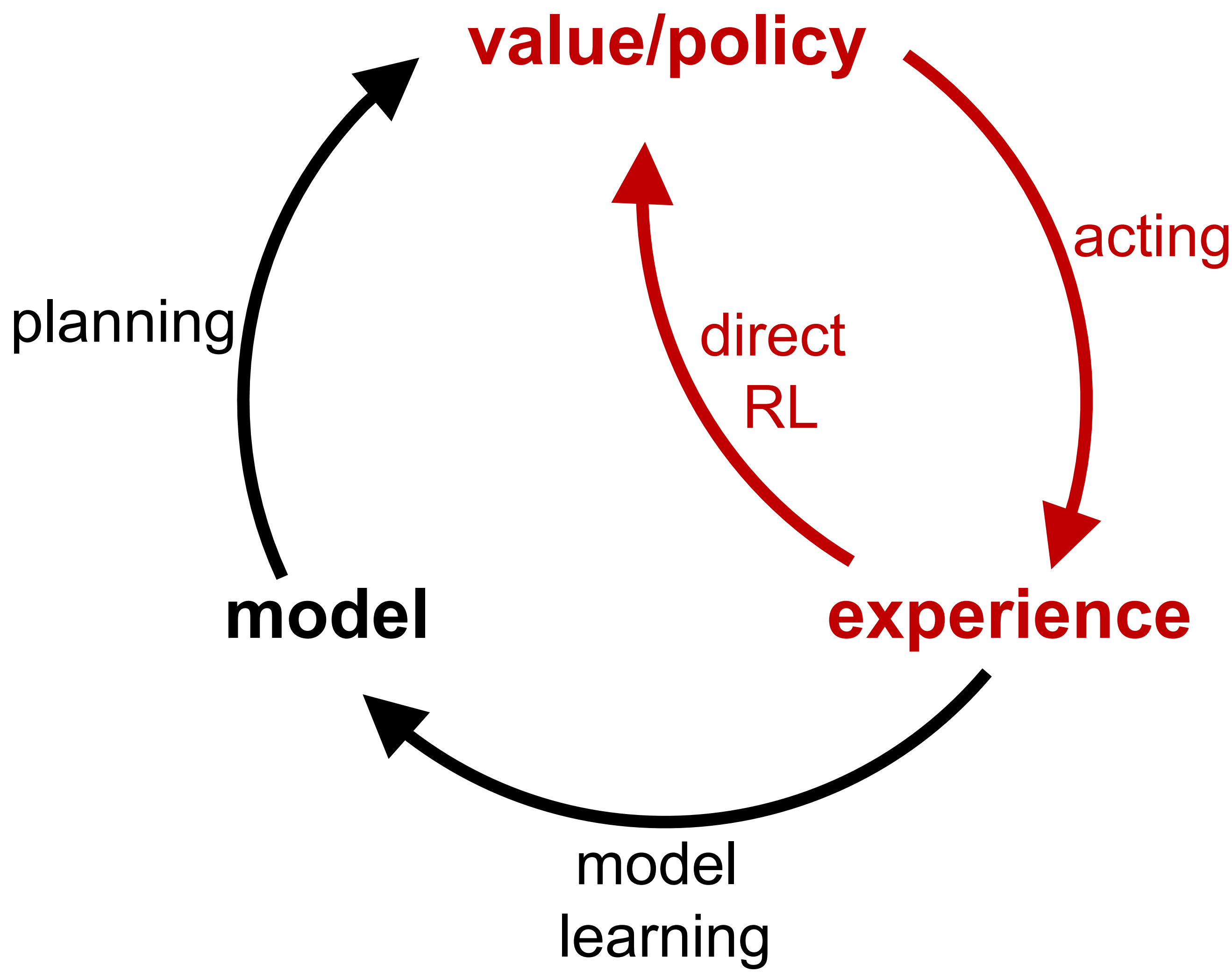
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Model-Based RL



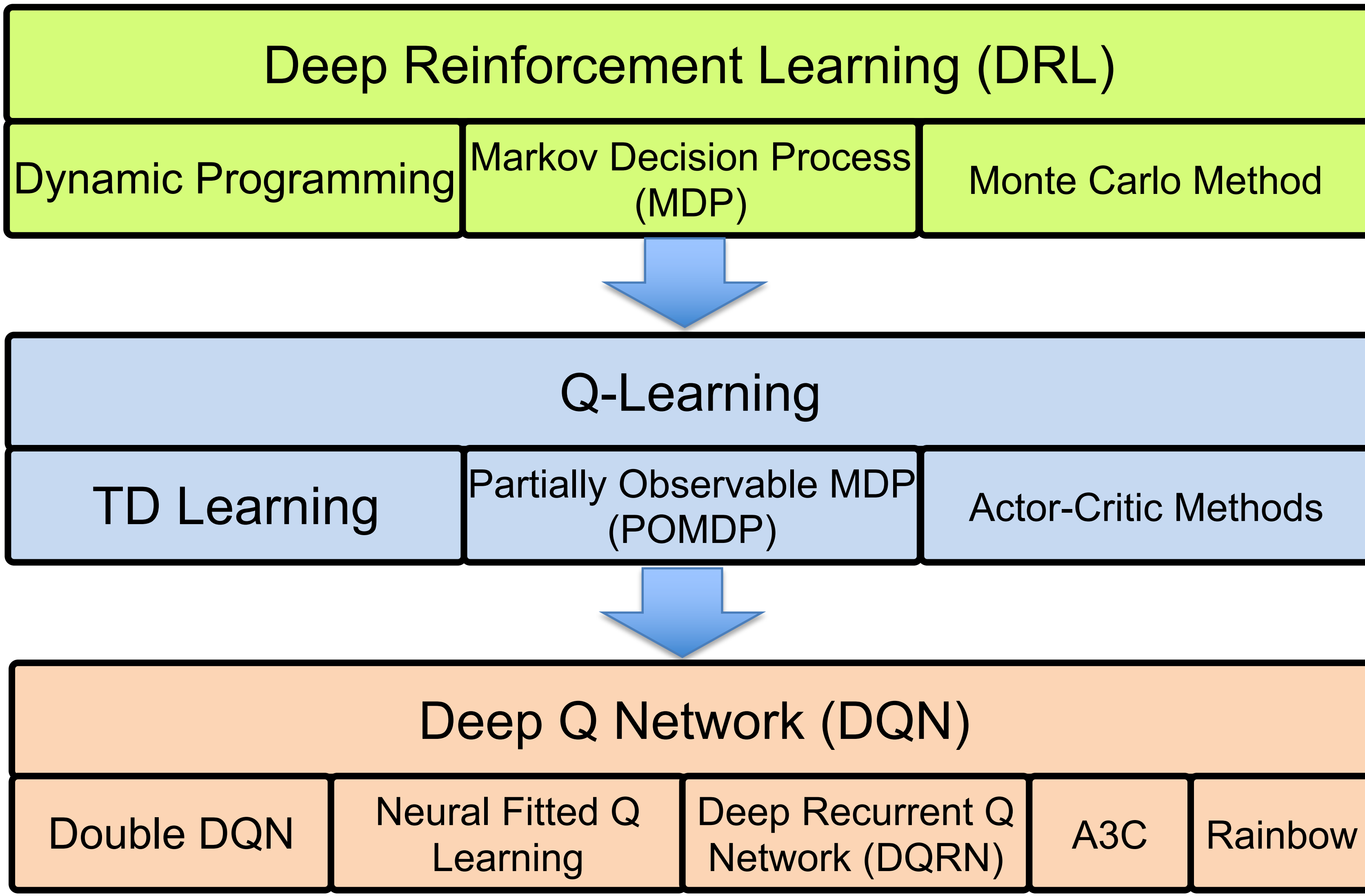
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Model-Free RL (DQN, A3C)



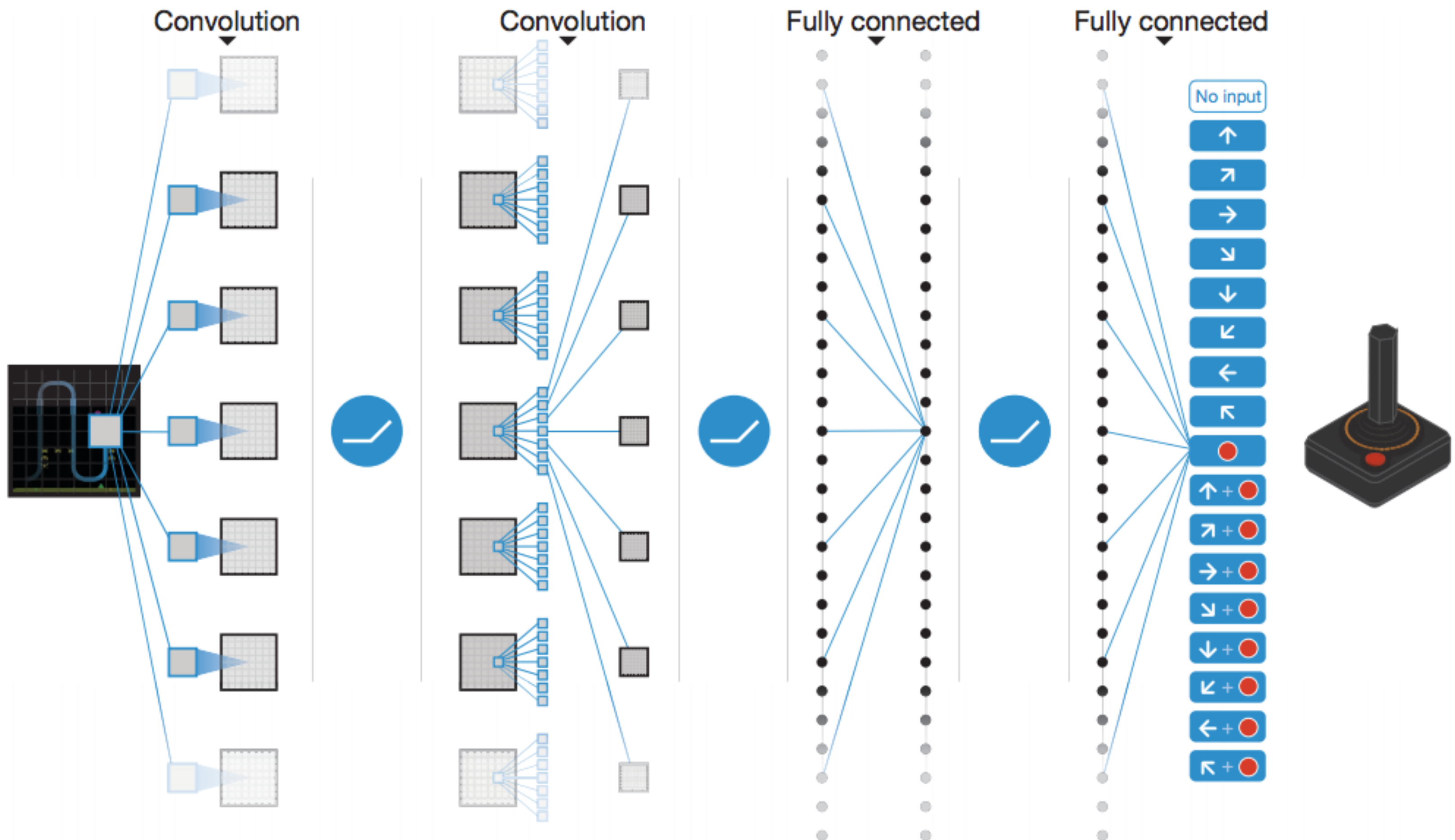
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning Algorithms



Source: https://amitray.com/deep-learning-past-present-and-future-a-review/

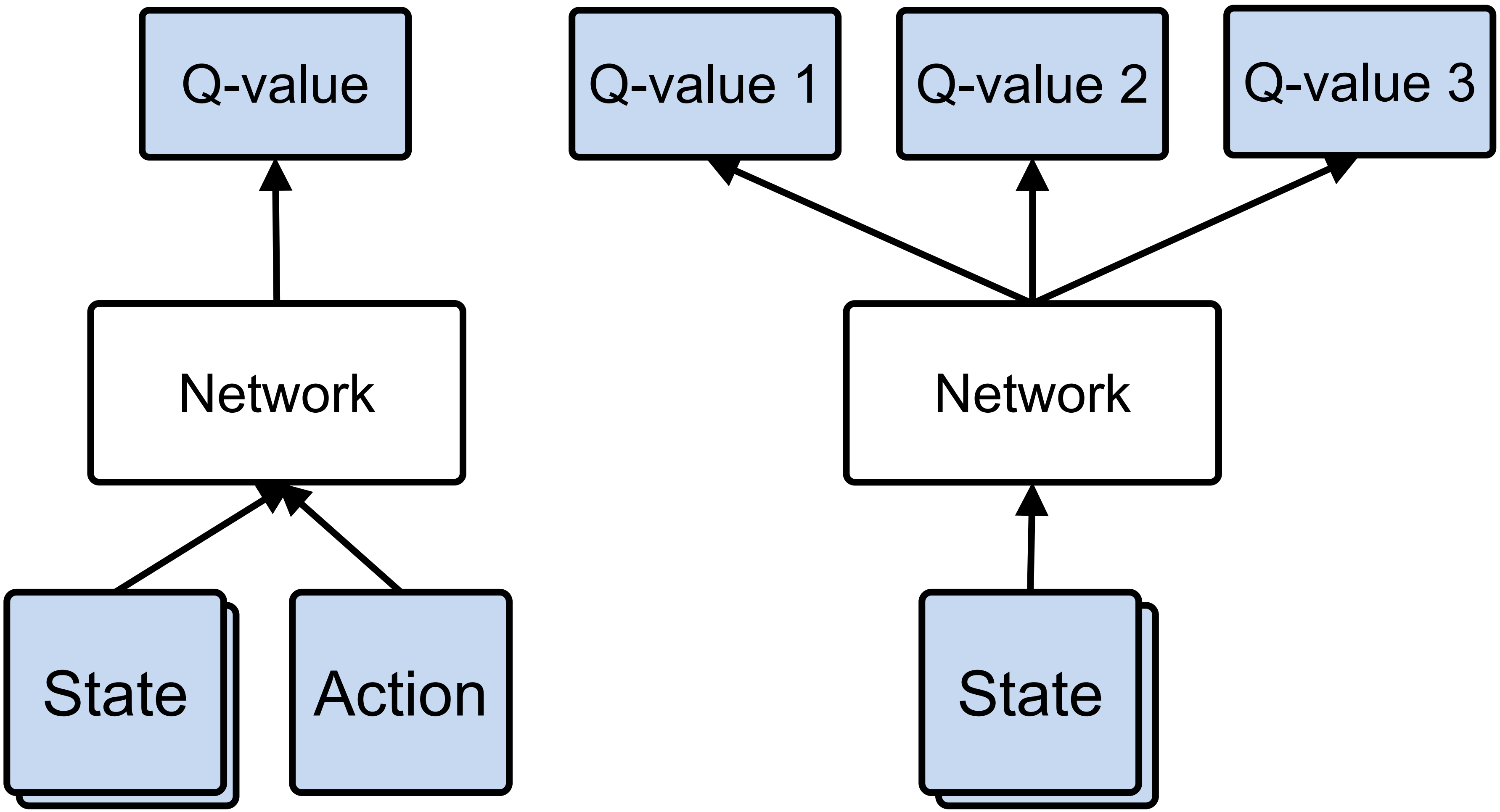
Human-level control through deep reinforcement learning (DQN)



Schematic illustration of the convolutional neural network

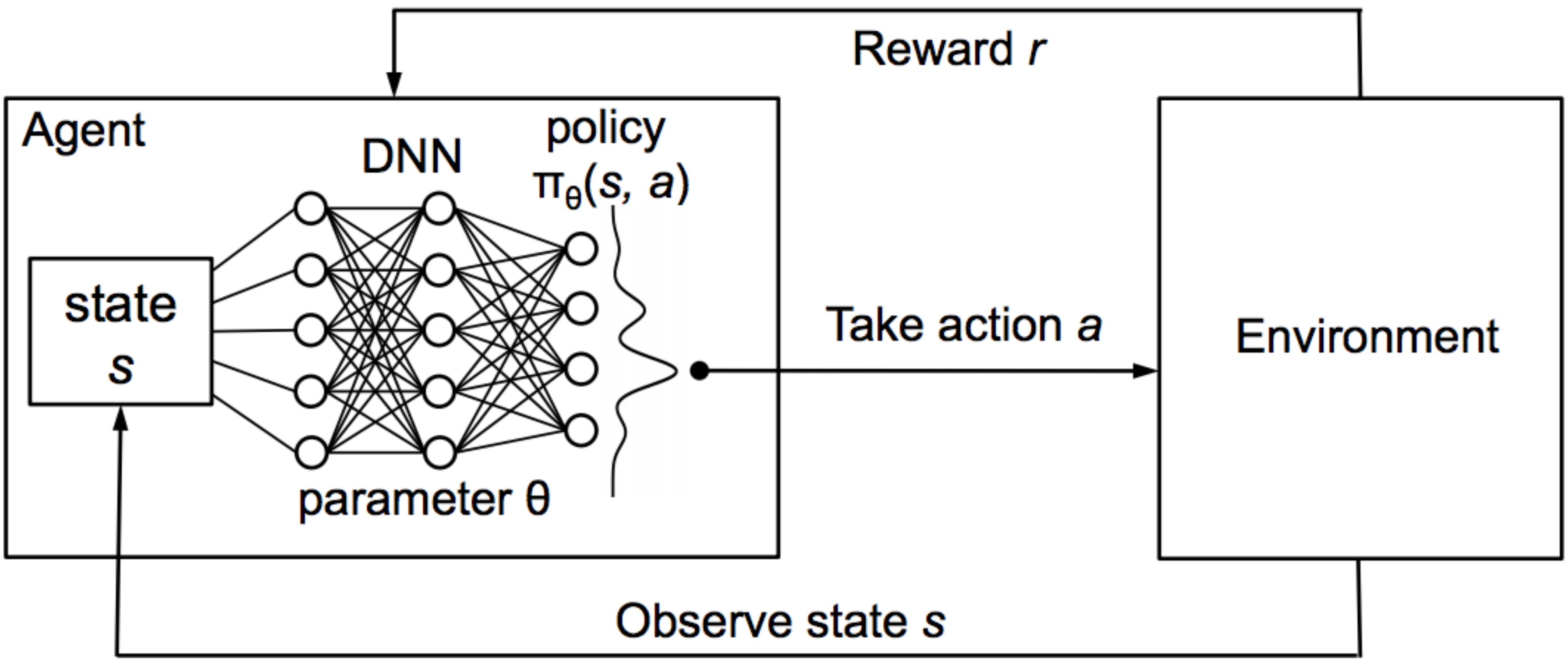
Source: Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves et al. "Human-level control through deep reinforcement learning." Nature 518, no. 7540 (2015): 529.

Deep Q-Network (DQN)



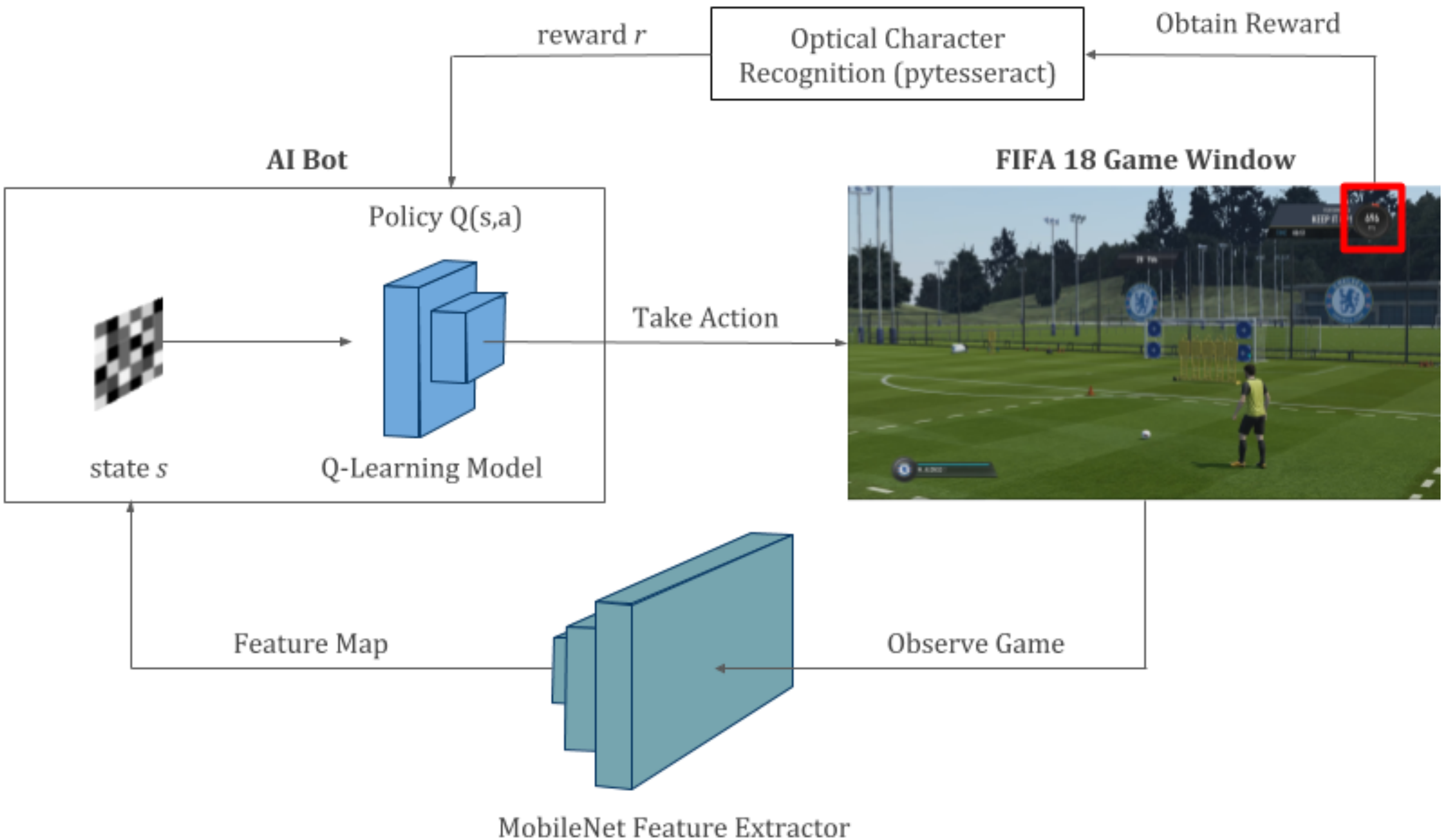
Source: https://neuro.cs.ut.ee/demystifying-deep-reinforcement-learning/

Reinforcement Learning with policy represented via DNN



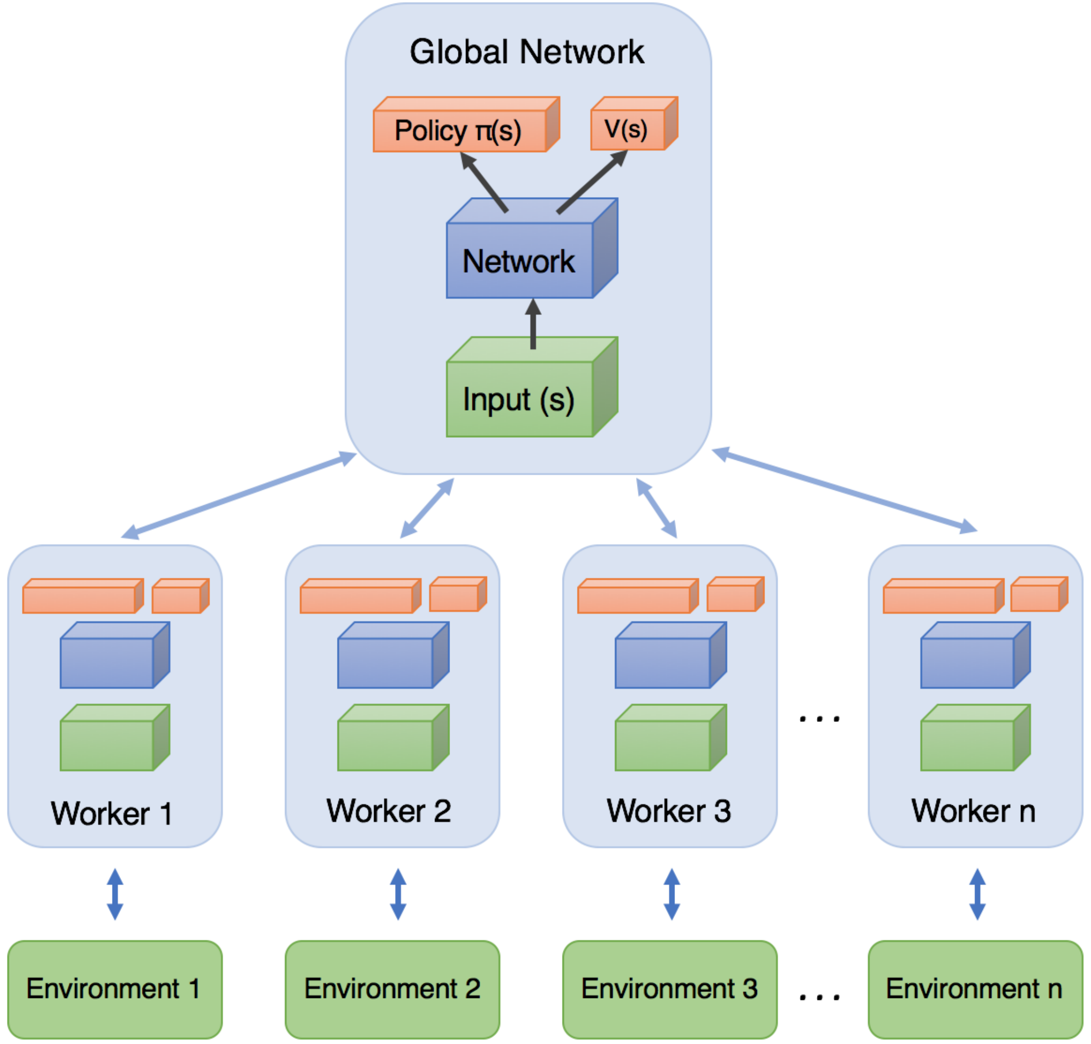
Source: Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. (2016) "Resource management with deep reinforcement learning." In Proceedings of the 15th ACM Workshop on Hot Topics in Networks, pp. 50-56. ACM, 2016.

Reinforcement Learning
Deep Q-Learning in FIFA 18



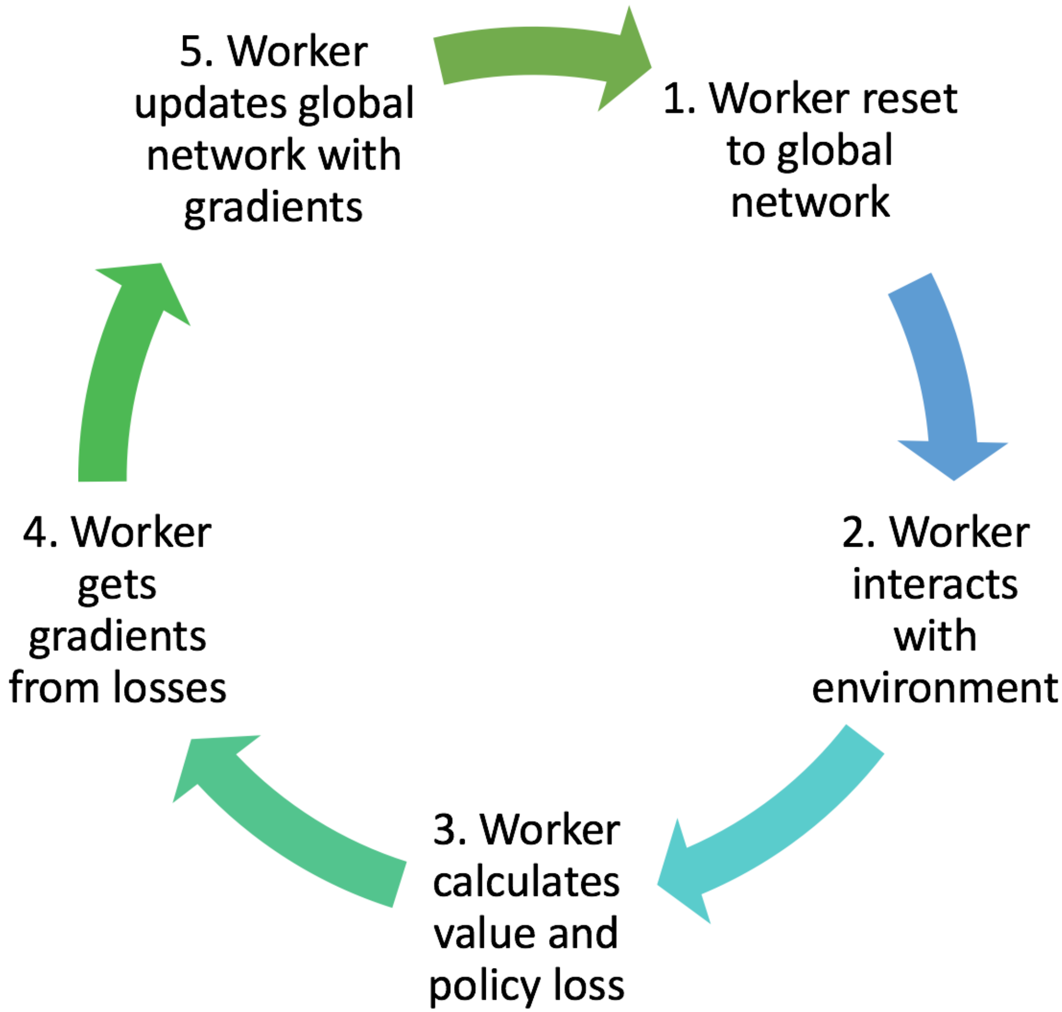
Source: https://towardsdatascience.com/using-deep-q-learning-in-fifa-18-to-perfect-the-art-of-free-kicks-f2e4e979ee66

Asynchronous Advantage Actor-Critic (A3C)



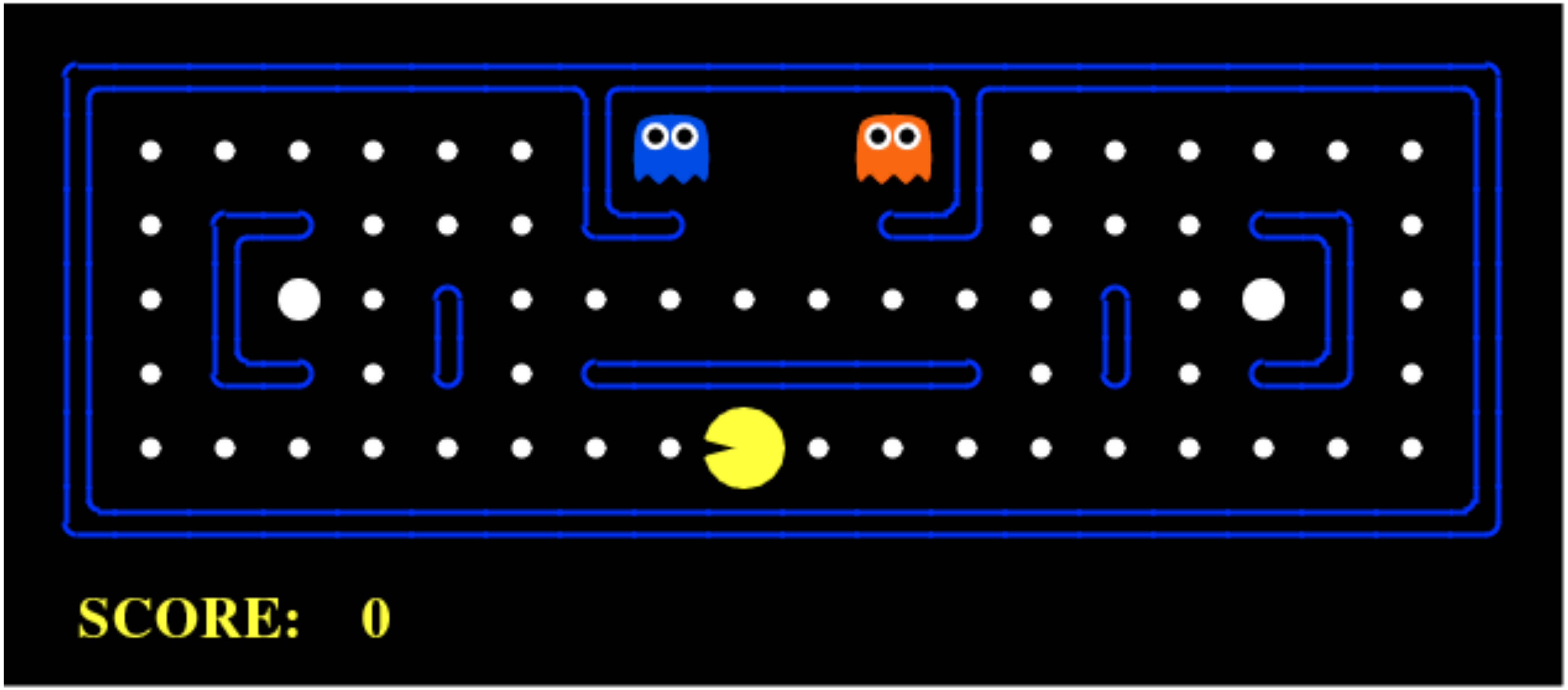
Source: https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2

Training workflow of each worker agent in A3C



Source: https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-8-asynchronous-actor-critic-agents-a3c-c88f72a5e9f2

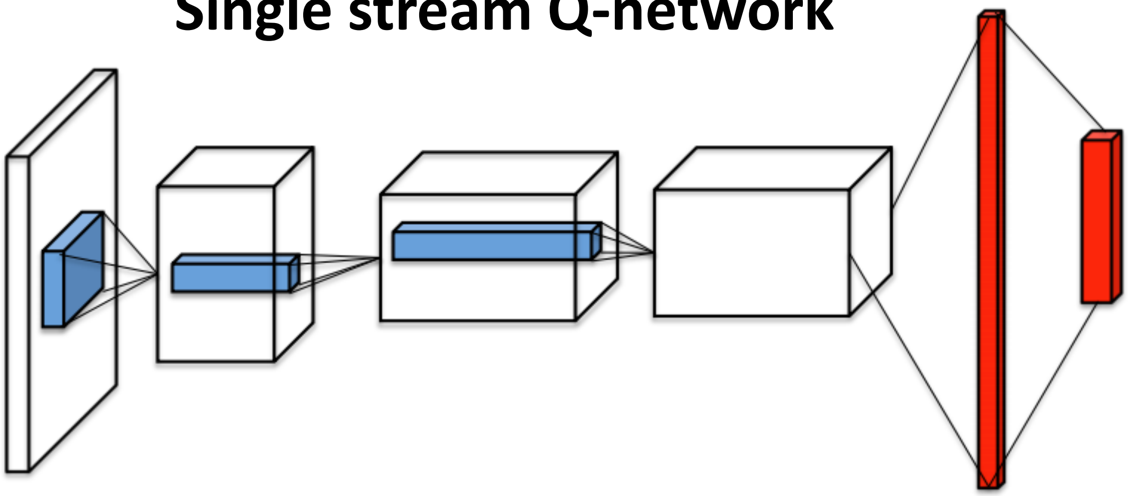
Reinforcement Learning Example: PCMAN



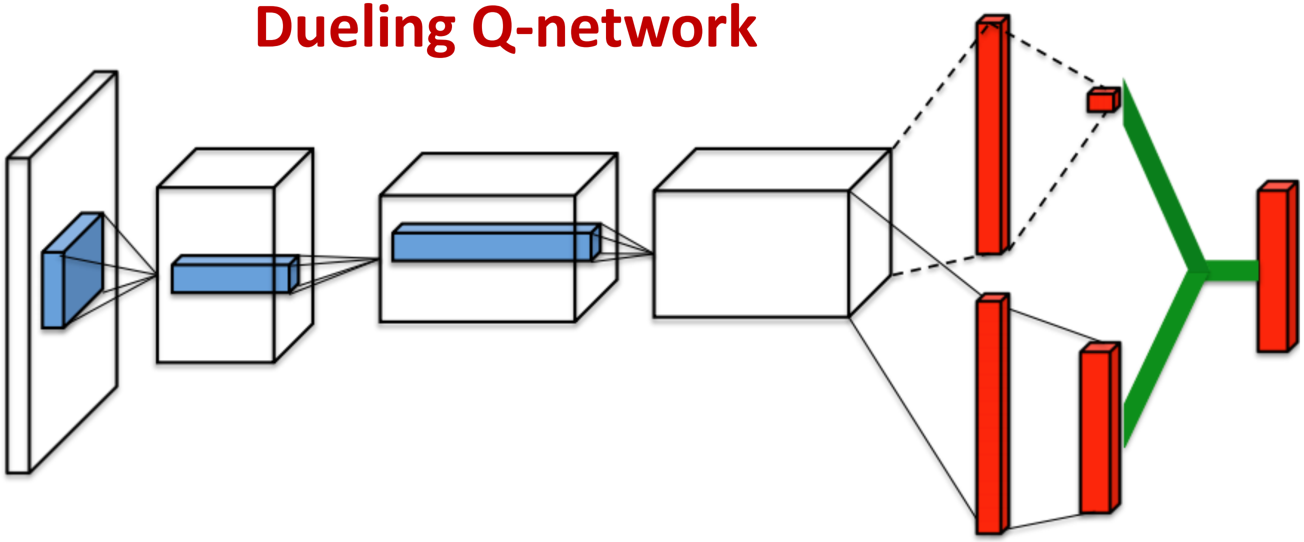
Source: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

Dueling Network Architectures for Deep Reinforcement Learning

Single stream Q-network

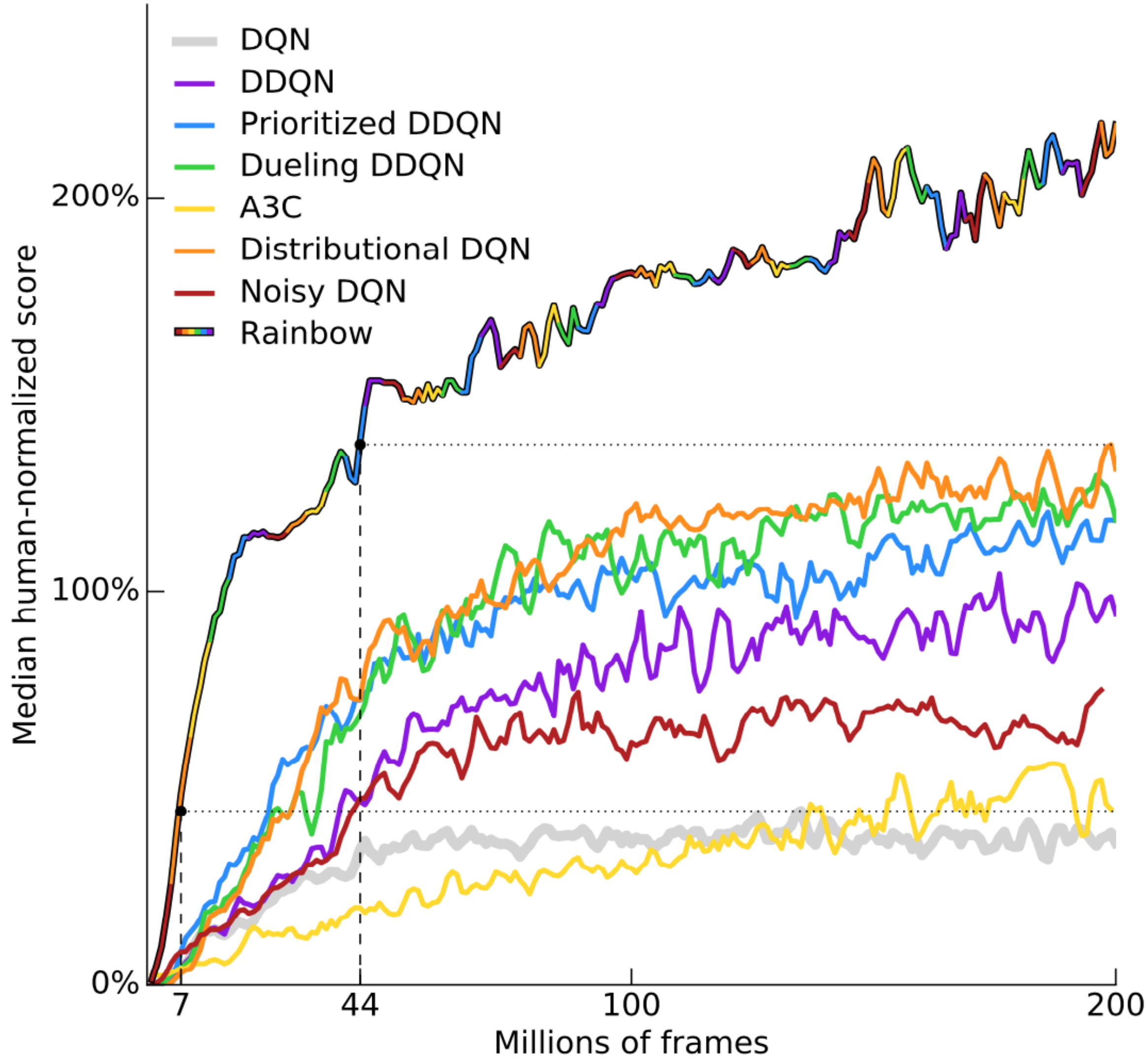


Dueling Q-network



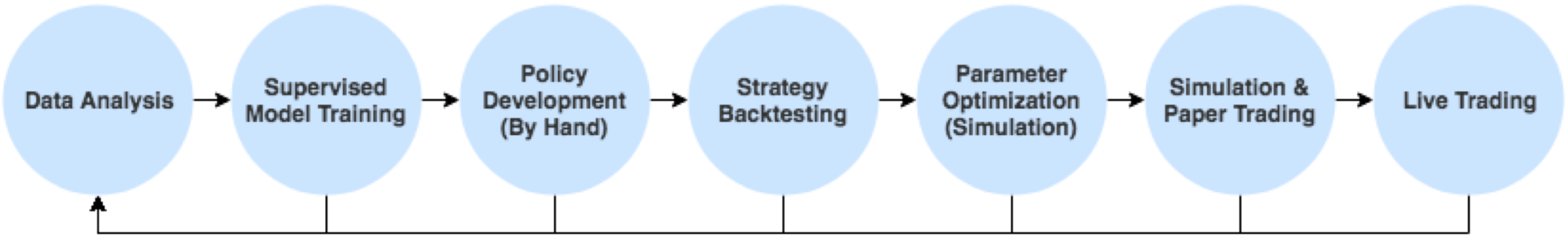
Source: Ziyu Wang et al. (2015). "Dueling network architectures for deep reinforcement learning." arXiv preprint arXiv:1511.06581 (2015)

Rainbow: Combining improvements in deep reinforcement learning



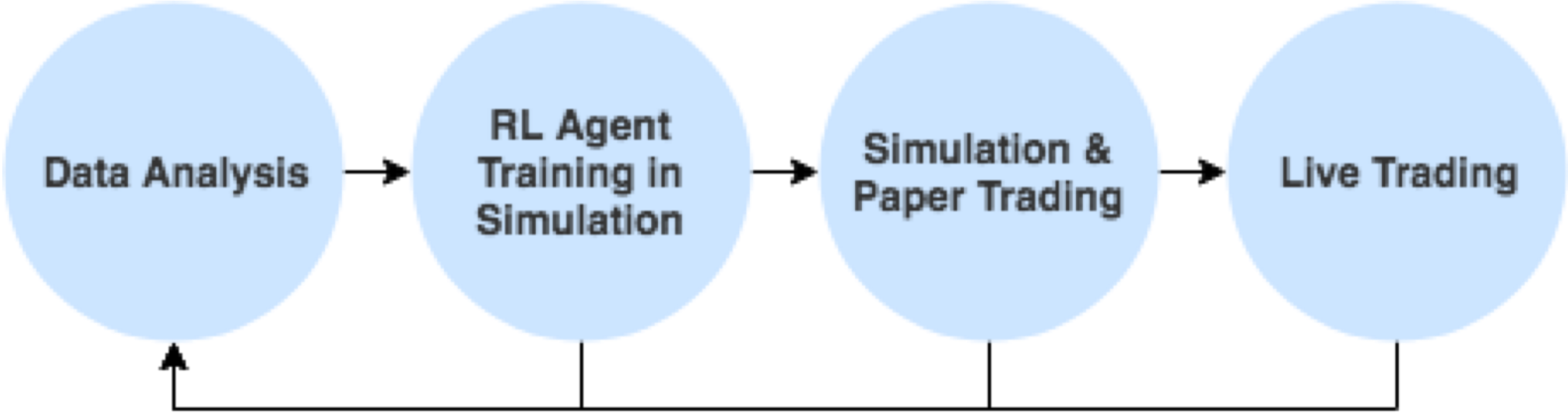
Source: Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver (2017). "Rainbow: Combining improvements in deep reinforcement learning." arXiv preprint arXiv:1710.02298 (2017).

A Typical Strategy Development Workflow



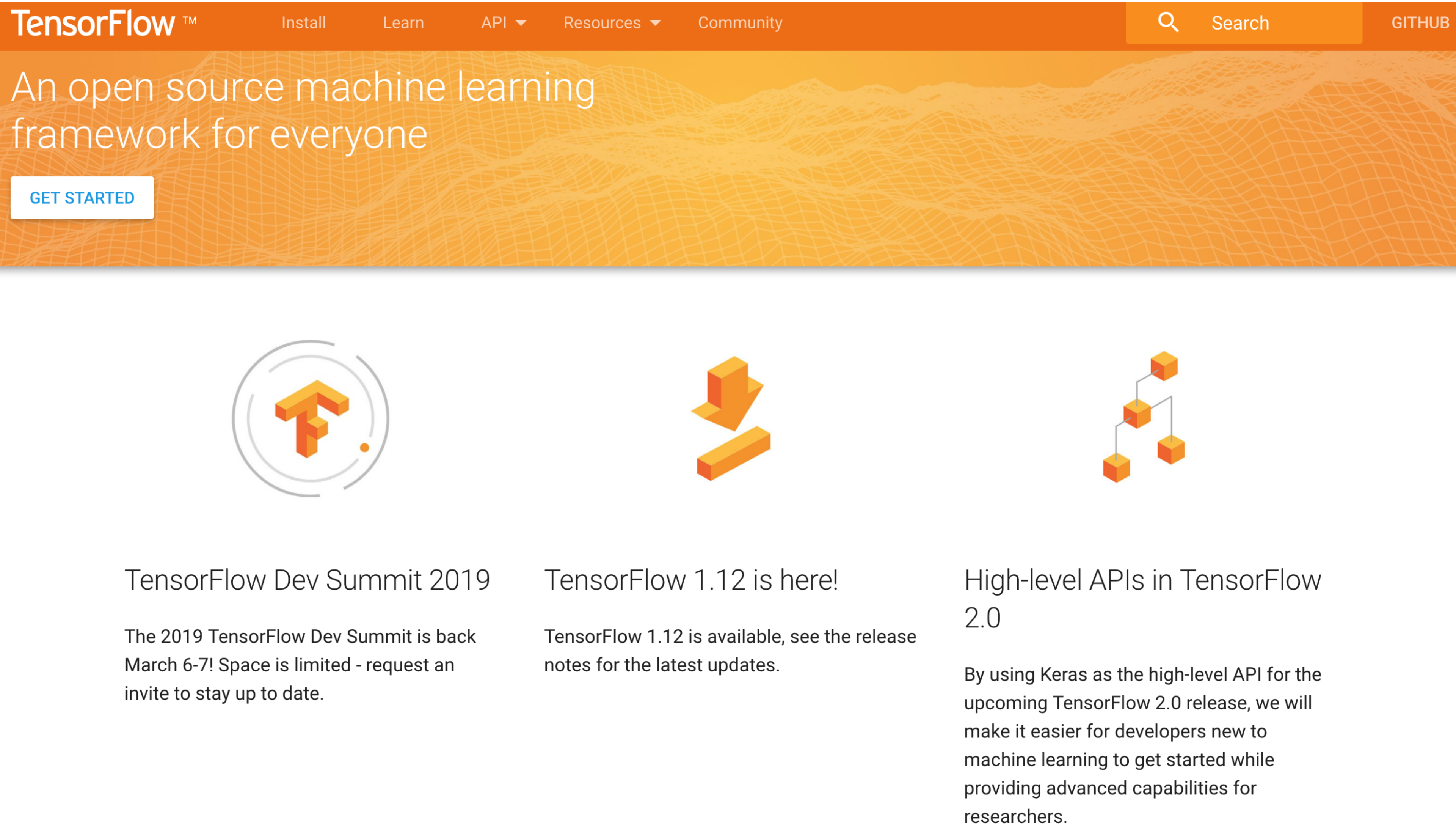
Source: http://www.wildml.com/2018/02/introduction-to-learning-to-trade-with-reinforcement-learning/

Reinforcement Learning (RL) in Trading Strategies



Source: http://www.wildml.com/2018/02/introduction-to-learning-to-trade-with-reinforcement-learning/

Google TensorFlow



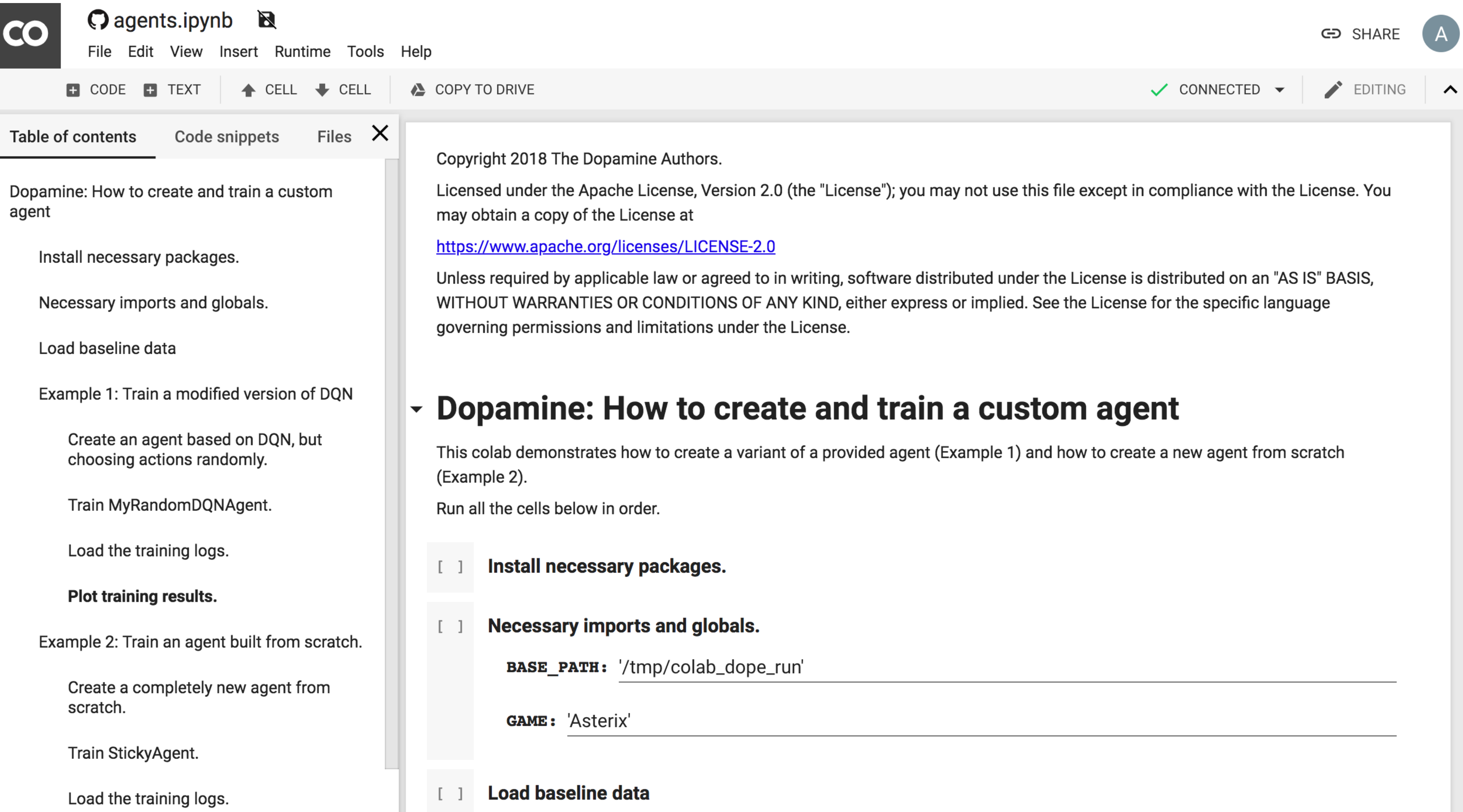
https://www.tensorflow.org/

Google Dopamine

Dopamine is a research framework for fast prototyping of reinforcement learning algorithms.

https://github.com/google/dopamine

Deep Reinforcement Learning
Dopamine Colab Examples
DQN Rainbow



https://colab.research.google.com/github/google/dopamine/blob/master/dopamine/colab/agents.ipynb

Summary

- Reinforcement Learning (RL)
  - Markov Decision Processes (MDP)
- Deep Reinforcement Learning (DRL) Algorithms
  - SARSA
  - Q-Learning
  - DQN
  - A3C
  - Rainbow
- Google Dopamine

References

- Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.
- David Silver (2015), Introduction to reinforcement learning, https://www.youtube.com/playlist?list=PLqYmG7hTraZDM-OYHWgPebj2MfCFzFObQ
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, Demis Hassabis (2018), "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play." Science 362, no. 6419 (2018): 1140-1144.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis (2017), "Mastering the game of Go without human knowledge." Nature 550 (2017): 354–359.
- Hado Van Hasselt, Arthur Guez, and David Silver (2016). "Deep Reinforcement Learning with Double Q-Learning." In AAAI, vol. 2, p. 5. 2016.
- Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver (2017). "Rainbow: Combining improvements in deep reinforcement learning." arXiv preprint arXiv:1710.02298 (2017).
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves et al. (2015) "Human-level control through deep reinforcement learning." Nature 518, no. 7540 (2015): 529.
- Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas (2015). "Dueling network architectures for deep reinforcement learning." arXiv preprint arXiv:1511.06581 (2015).