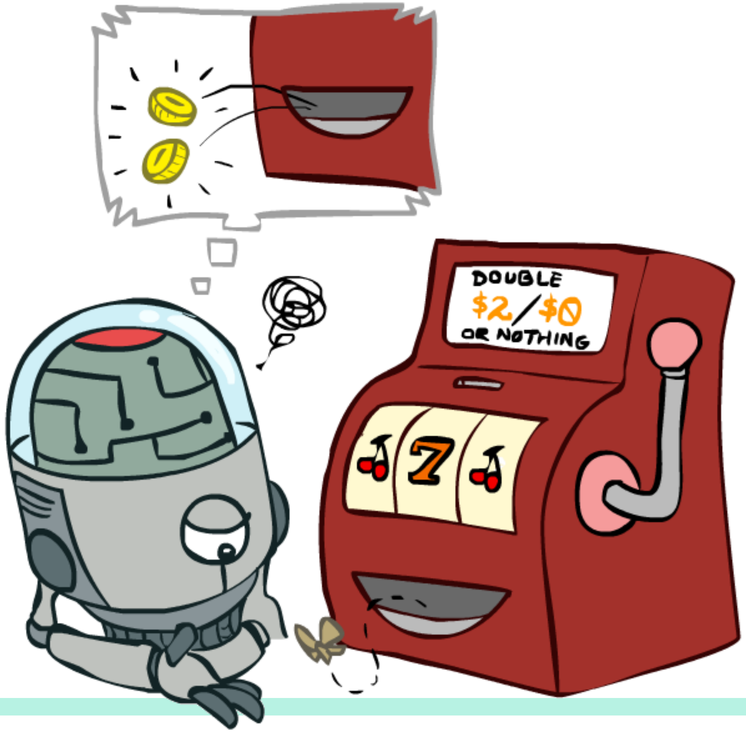
Advanced Topics in AI

Temporal Difference Value Learning



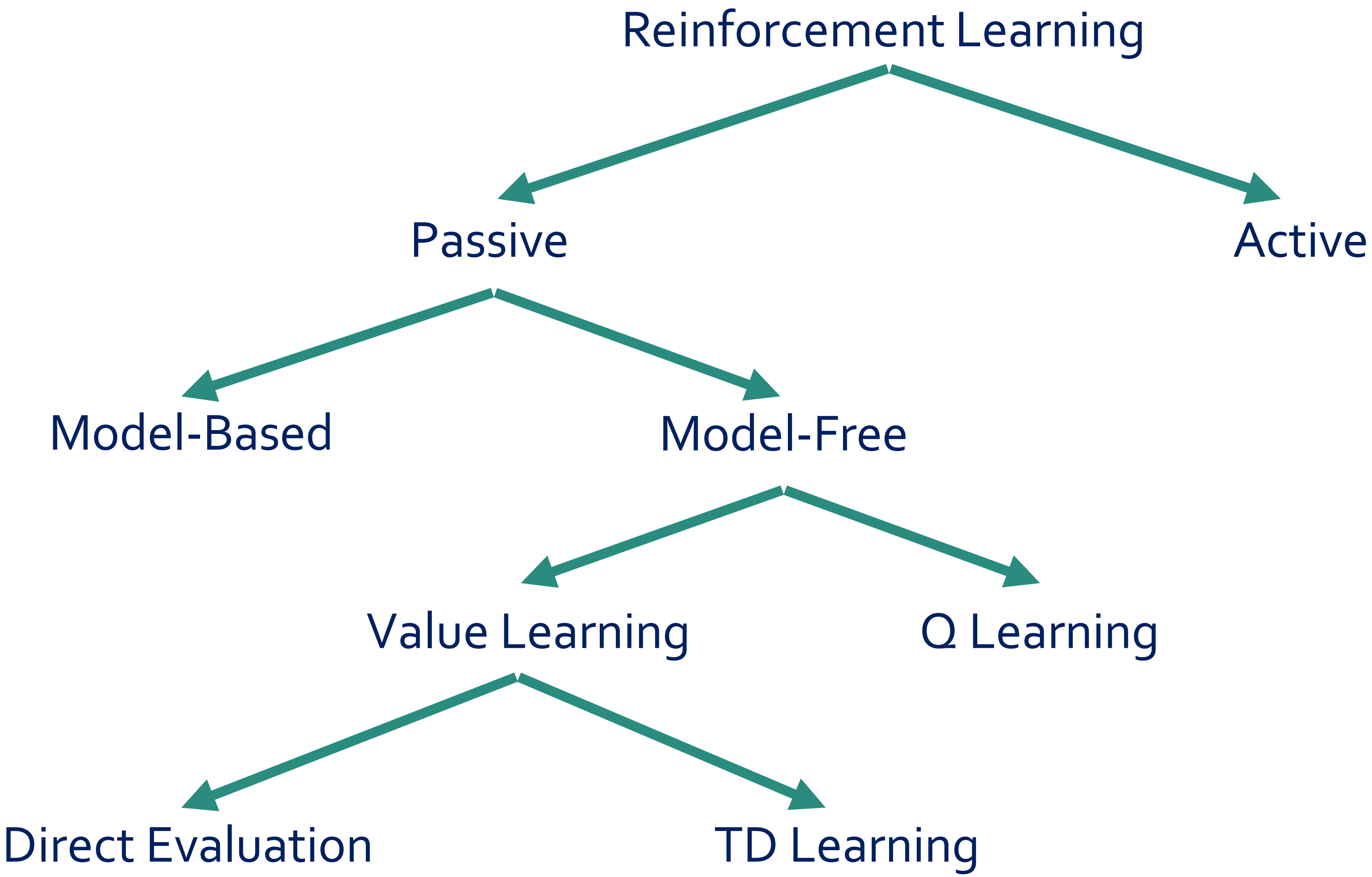
Instructor: Prof. Dr. techn. Wolfgang Nejdl

Leibniz University Hannover

Co-financed by the Connecting Europe Facility of the European Union

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All materials are available at http://ai.berkeley.edu.]

Reinforcement Learning Taxonomy



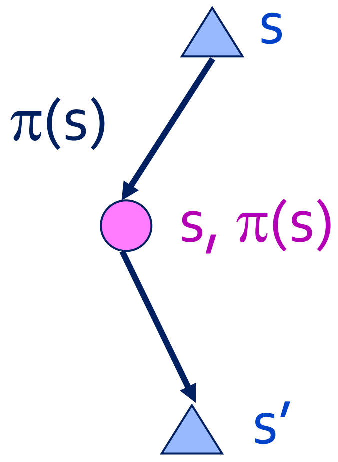
- Reinforcement Learning
  - Passive
    - Model-Based
    - Model-Free
      - Value Learning
        - Direct Evaluation
        - TD Learning
      - Q Learning
  - Active

Co-financed by the Connecting Europe Facility of the European Union

Temporal Difference Value Learning

- Big idea: learn from every experience!
  - Update V(s) each time we experience a transition (s, a, s', r)
  - Likely outcomes s' will contribute updates more often

- Temporal difference learning of values
  - Policy still fixed, still doing evaluation!
  - Move values toward value of whatever successor occurs: running average



Sample of V(s): $\text{sample} = R(s, \pi(s), s') + \gamma V^{\pi}(s')$

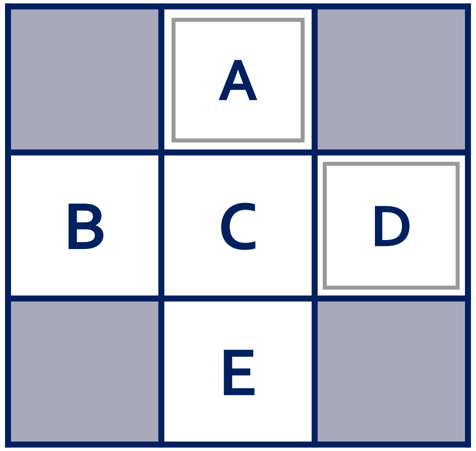
Update to V(s): $V^{\pi}(s) \leftarrow (1-\alpha) V^{\pi}(s) + \alpha \cdot \text{sample}$

Same update: $V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha\big(\text{sample} - V^{\pi}(s)\big)$

Co-financed by the Connecting Europe Facility of the European Union

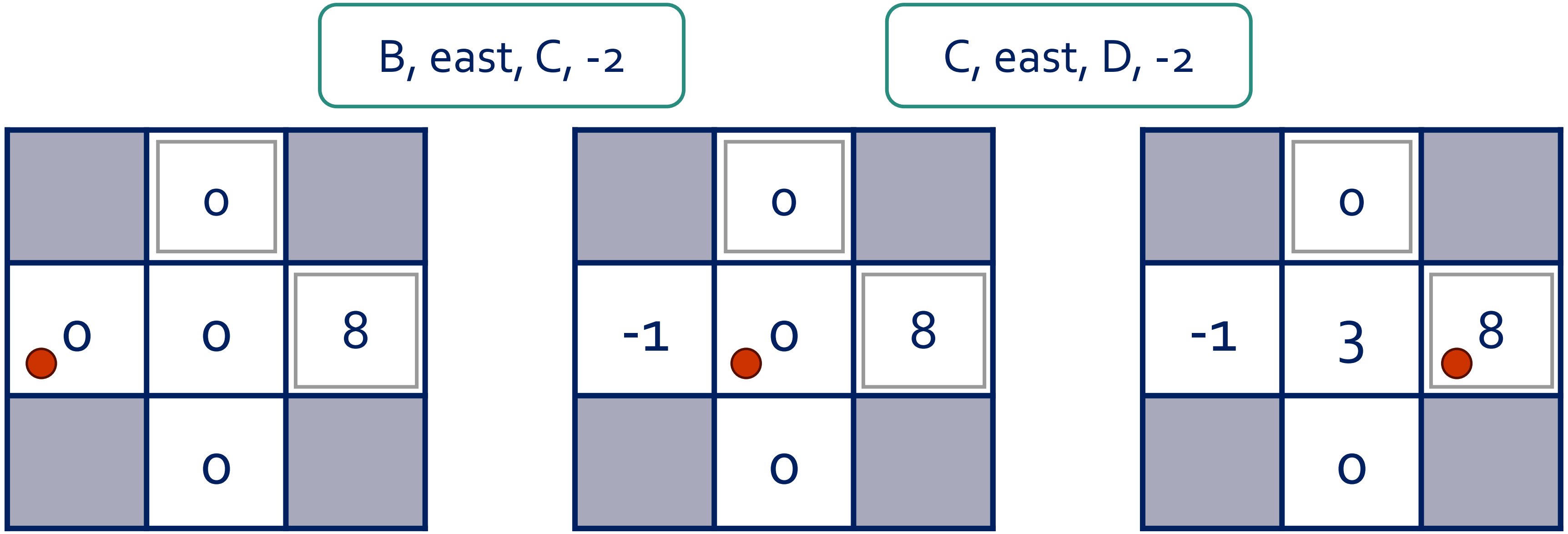
Example: TD Value Learning

States



Assume: $\gamma = 1$, $\alpha = 1/2$

Observed Transitions

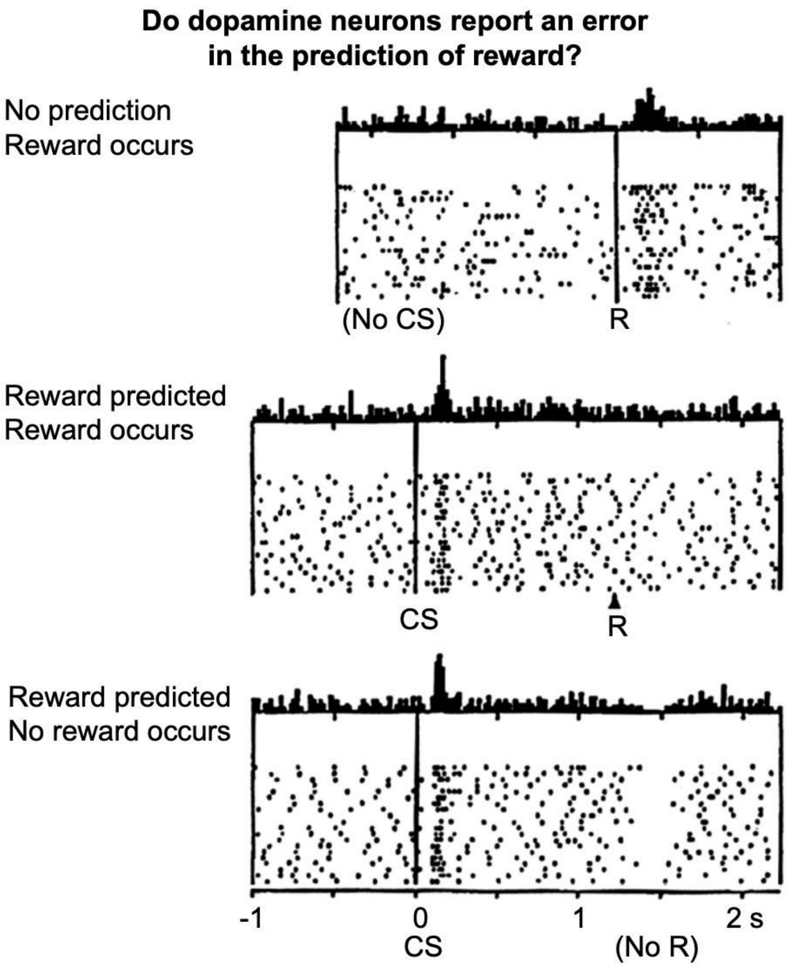


$$V^{\pi}(s) \leftarrow (1-\alpha)V^{\pi}(s) + \alpha[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

Co-financed by the Connecting Europe Facility of the European Union

TD Learning in the Brain

- Neurons transmit Dopamine to encode reward or value prediction error
  - $V^{\pi}(s) \leftarrow V^{\pi}(s) + \alpha\left(\text{sample} - V^{\pi}(s)\right)$
- Example of Neuroscience & RL informing each other
- For more examples, see [AI and Neuroscience: A virtuous circle]
  - https://www.deepmind.com/blog/ai-and-neuroscience-a-virtuous-circle



[A Neural Substrate of Prediction and Reward. Schultz, Dayan, Montague. 1997]

Co-financed by the Connecting Europe Facility of the European Union

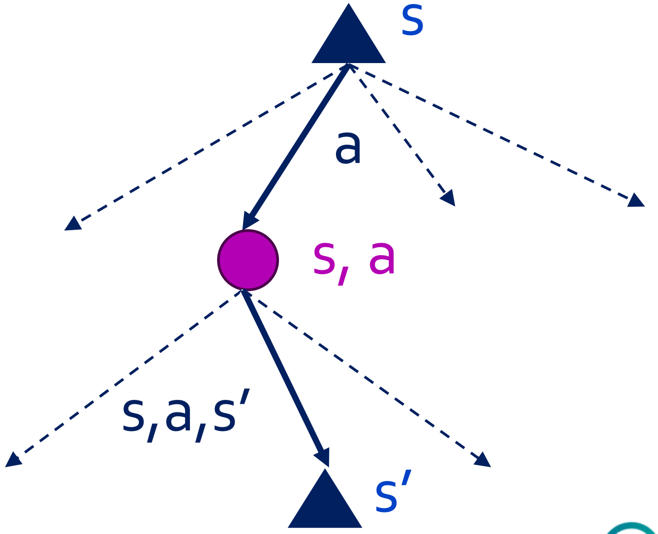
Problems with TD Value Learning

- TD value leaning is a model-free way to do policy evaluation, mimicking Bellman updates with running sample averages
- However, if we want to turn values into a (new) policy, we’re sunk:

$$\pi(s) = \arg\max_a Q(s,a)$$

$$Q(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V(s')]$$

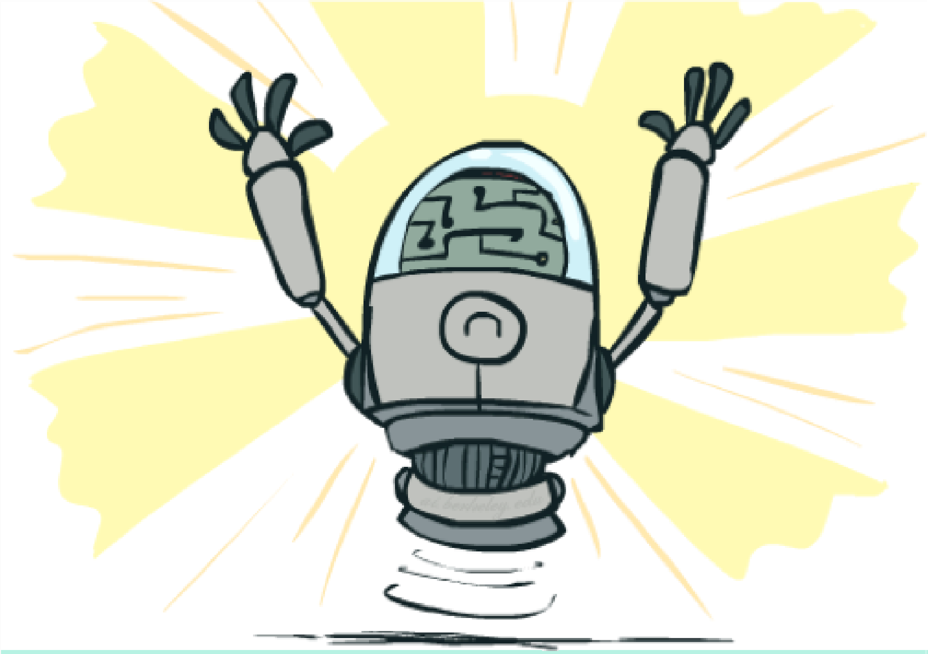
- Idea: learn Q-values, not values
- Makes action selection model-free too!



Co-financed by the Connecting Europe Facility of the European Union

Advanced Topics in AI

Next: Q-Learning



Instructor: Prof. Dr. techn. Wolfgang Nejdl

Leibniz University Hannover

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All materials are available at http://ai.berkeley.edu.]

Co-financed by the Connecting Europe Facility of the European Union