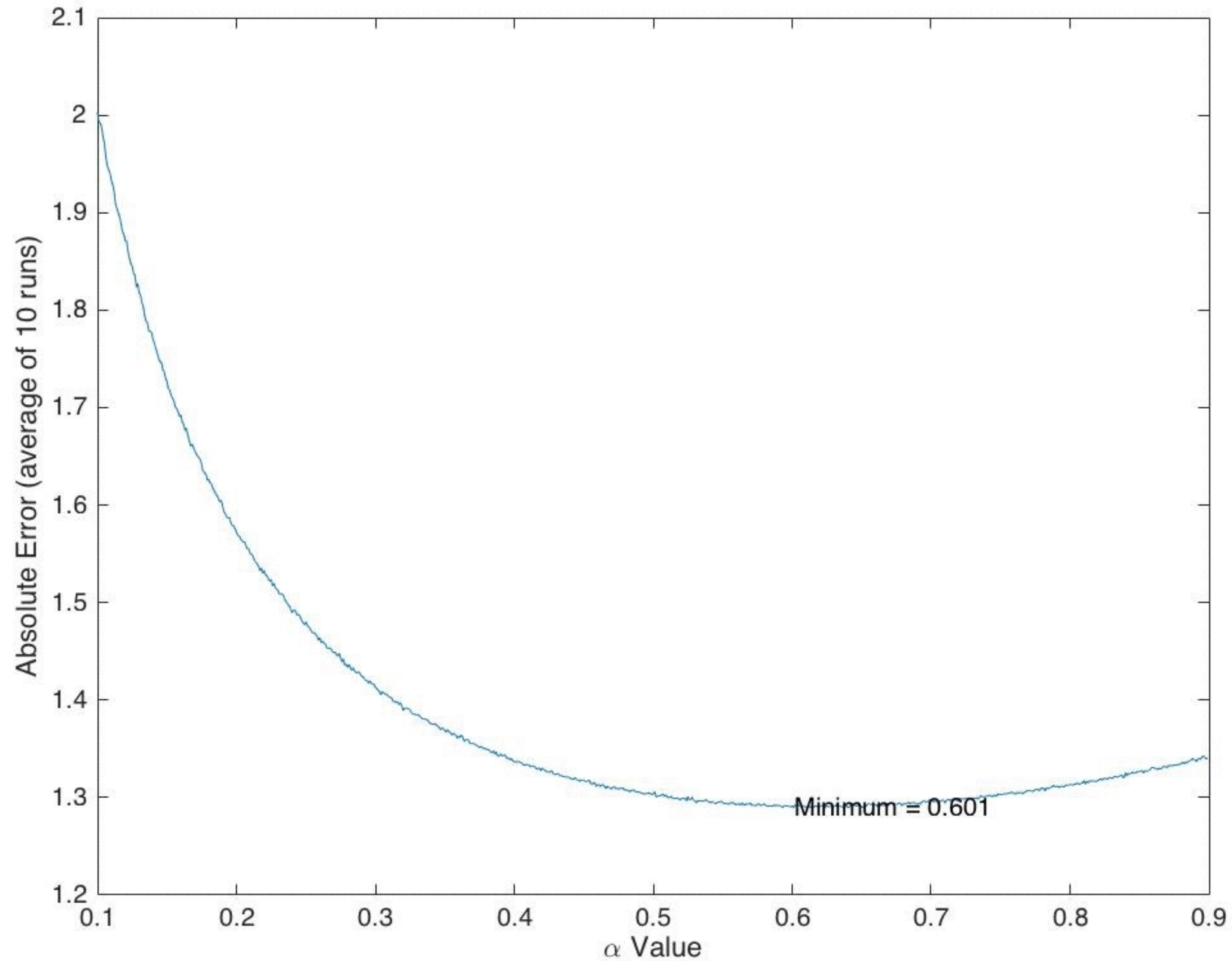


Question 1, part 5: Empirical search for best step size

by Dylan Ashley



What you have learned from bandits

- The need to tradeoff exploitation and exploration, e.g., by an ε -greedy policy
- The difference between a sample, an estimate, and a true expected value

$$R_t, \quad Q_t(a), \quad q_*(a)$$

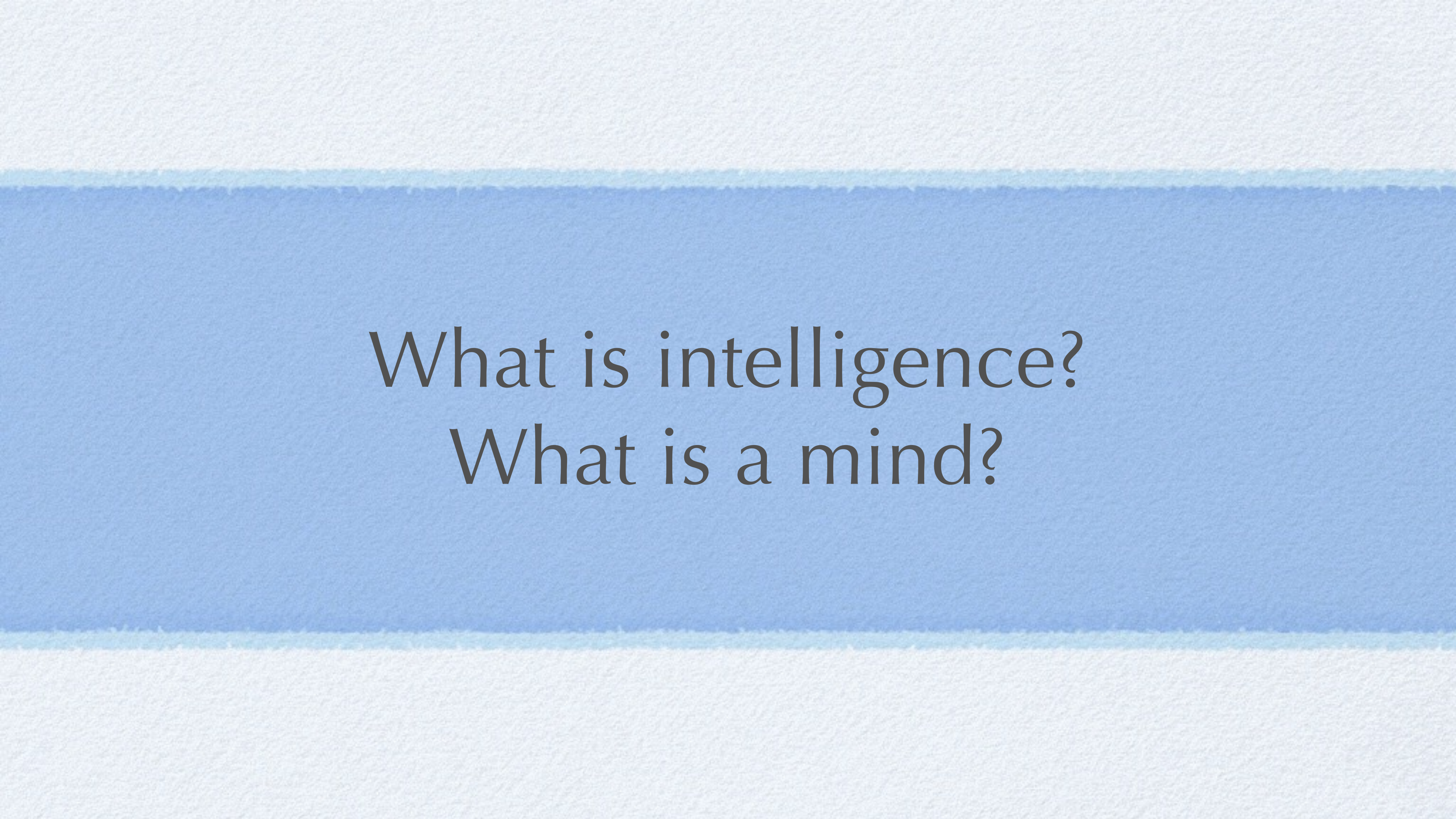
- The difference between the greedy action and the optimal action
- A learning rule. How learning can be seen as computing an average
 - The role of the step size α (how it can be too big, or too small, or “1/n”)
- A complete example of mathematically formalized goal seeking (intelligence)
—both problem and solution methods

Quiz for fun

Defining “Intelligent Systems”

Defining “System”

- A thing
 - with some recognizable identity over time (need not be physical)
 - usually with some inputs and outputs
 - may have state (not a function)
 - sometimes with a goal/purpose



The background of the slide features a wide, flat landscape under a clear sky. A prominent, solid blue horizontal band stretches across the middle of the image, creating a strong visual contrast. The text is centered within this blue band.

What is intelligence?
What is a mind?

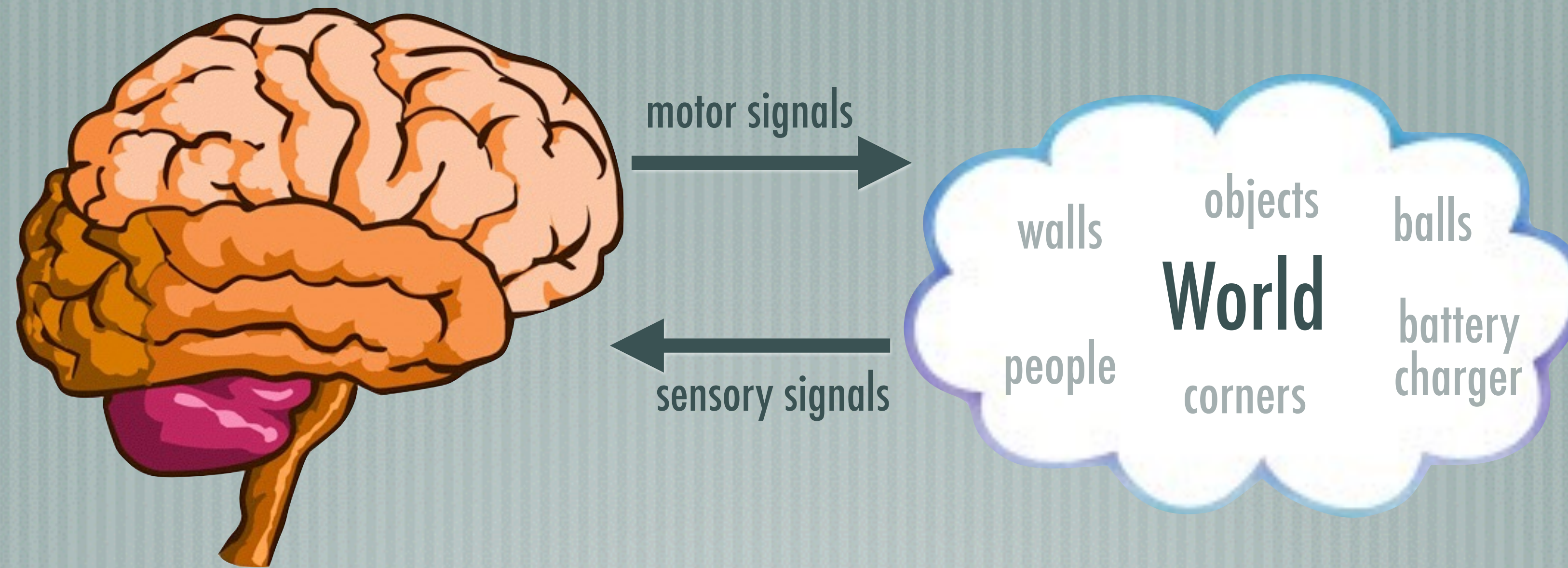
Socrative.com, Room 568225

Question: What is the right definition of intelligence?

Answer: The computational part of the ability to:

- A. improve over time with experience
- B. use language, communicate and cooperate with other intelligent agents
-  C. achieve goals
- D. predict and control your input signals
-  E. There is no such thing as a right definition

Minds are sensori-motor information processors



— [the mind's job is to predict and control its sensory signals

the mind's first responsibility is
real-time sensorimotor information processing

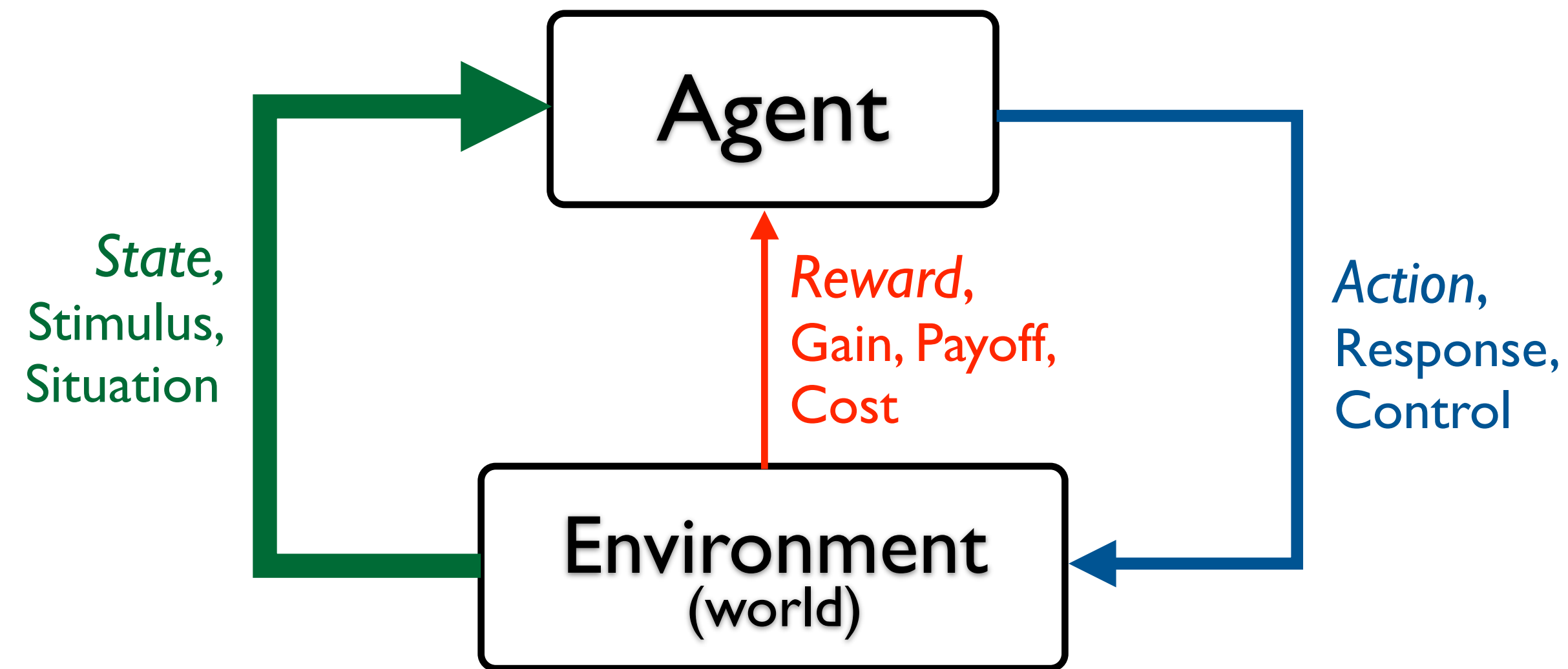
- Perception, action, & anticipation
- as fast and reactive as possible



Artificial Intelligence (definition)

- the science and technology of information processing systems with goals
 - that is, of information processing systems that observers tend to find it useful to think about in terms of goals
 - that is, in terms of outcomes rather than in terms of mechanisms

The RL Interface



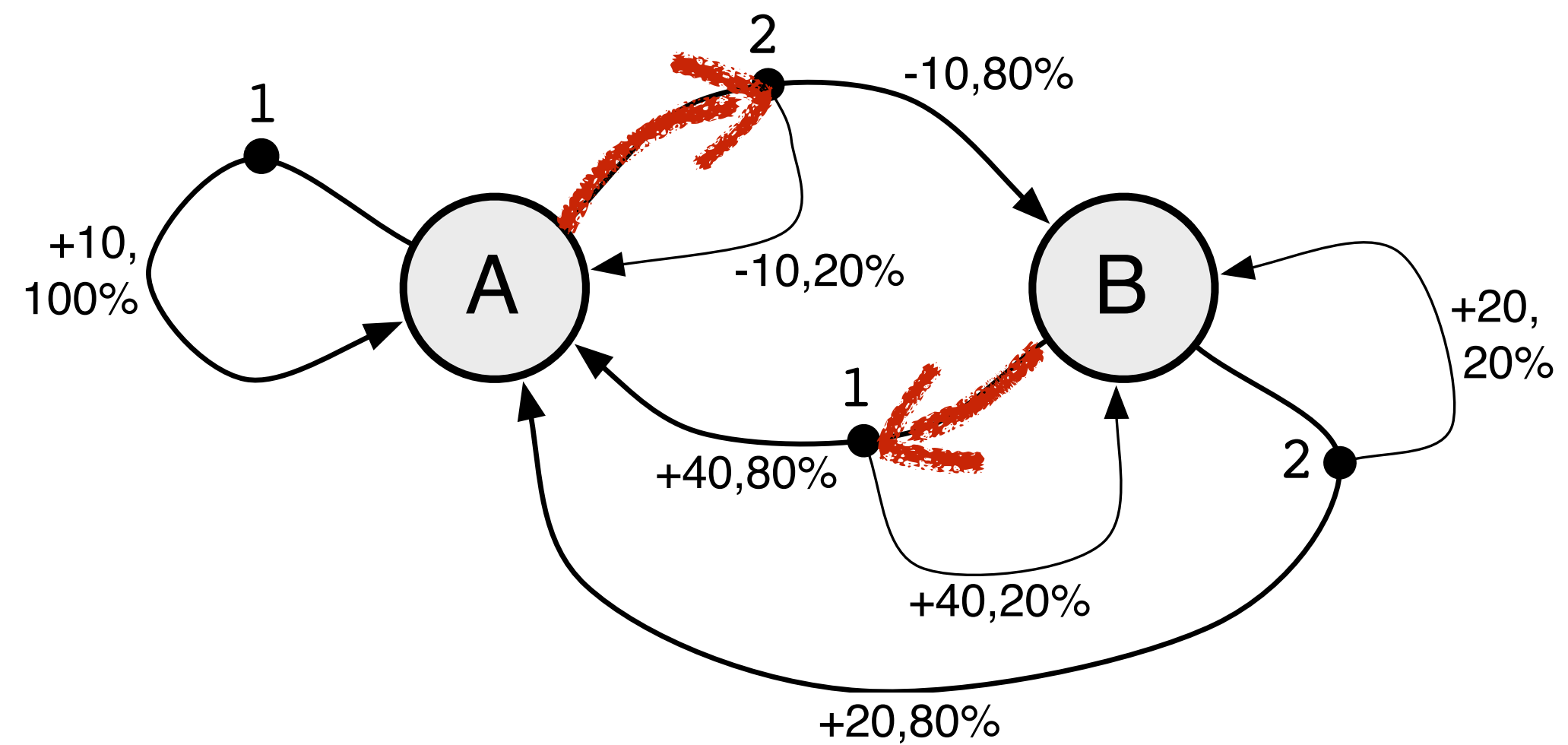
- Environment may be unknown, nonlinear, stochastic and complex
- Agent learns a policy mapping states to actions
 - Seeking to maximize its cumulative reward in the long run

You are the reinforcement learner! (interactive demo)

Optimal policy
(deterministic)

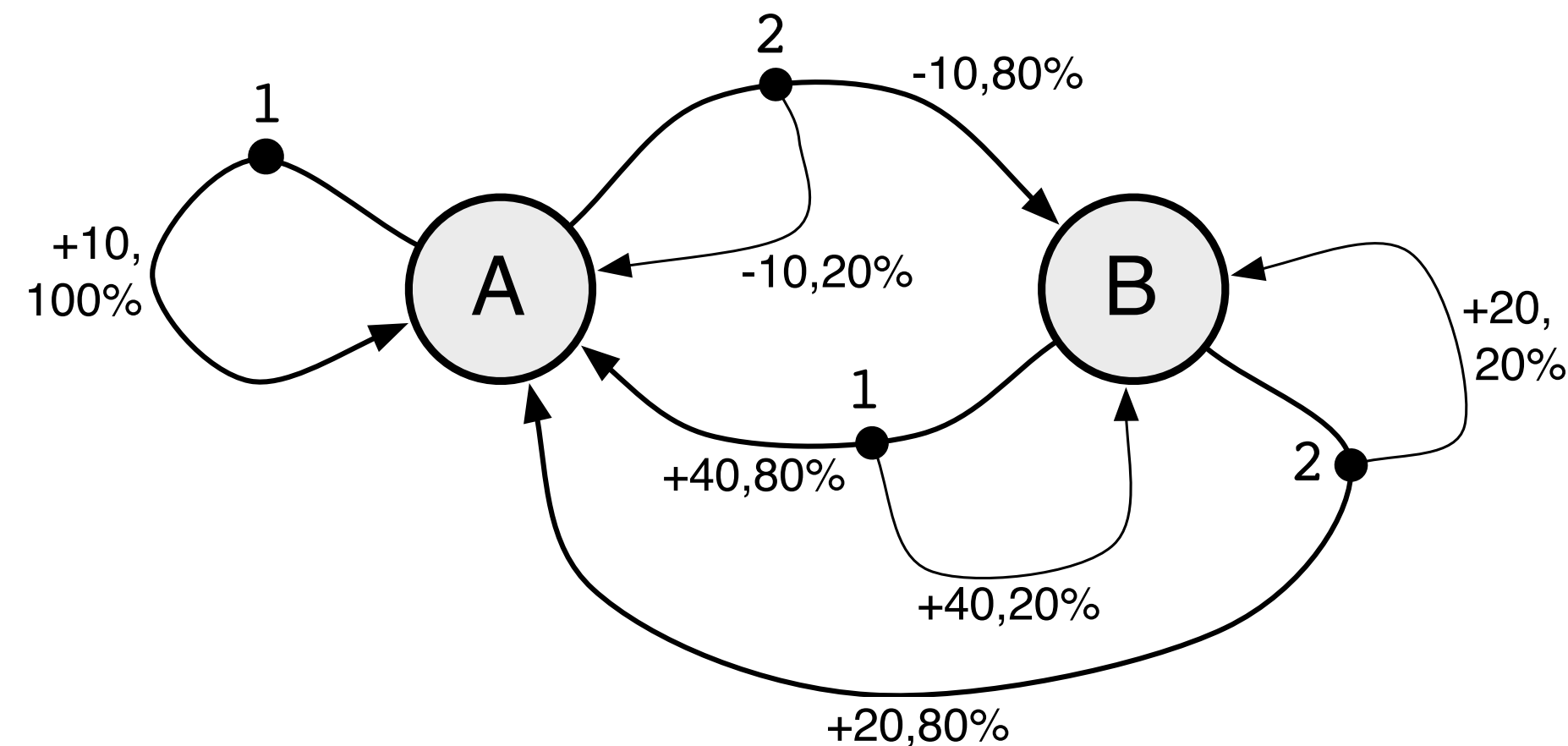
State	Action
A	2
B	1

True model of the world



The Environment: A Finite Markov Decision Process (MDP)

- Discrete time $t = 1, 2, 3, \dots$
- A finite set of **states**
- A finite set of **actions**
- A finite set of **rewards**
- Life is a trajectory:



$$\dots S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \dots$$

- With arbitrary Markov (stochastic, state-dependent) dynamics:

$$p(r, s' | s, a) = \text{Prob} \left[R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a \right]$$

Policies

- Deterministic policy

$$a = \pi(s)$$

- An agent following a policy

$$A_t = \pi(S_t)$$

- Informally the agent's goal is to choose each action so as to maximize the discounted sum of future rewards,

to choose each A_t to maximize $R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$

- We are **searching for a policy**

“gamma”, the discount rate $\in [0, 1)$

e.g.

State	Action
A	→ 2
B	→ 1

The number of deterministic policies is *exponential* in the *number of states*

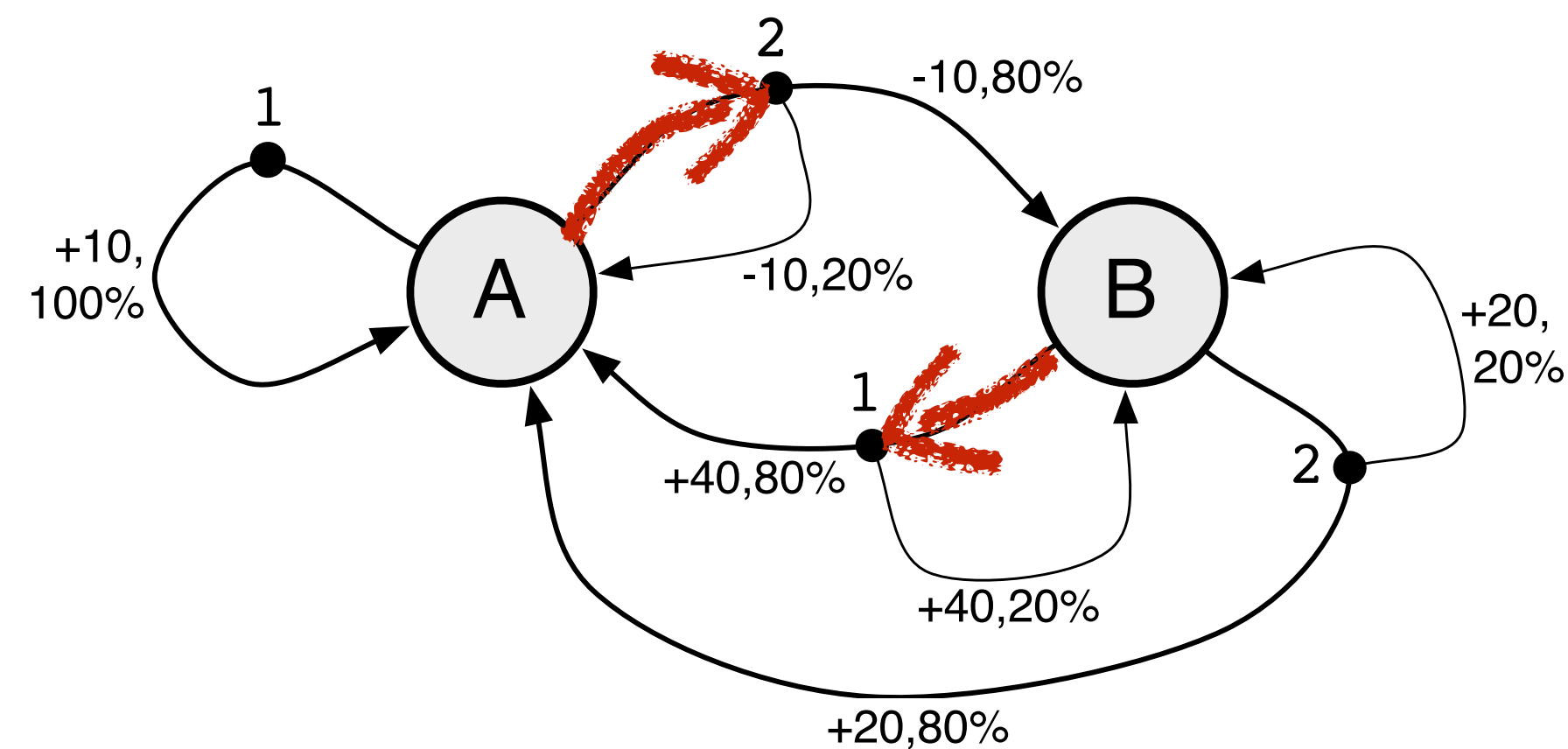
Action-value functions

- An **action-value function** says how good it is to be in a state, take an action, and thereafter follow a policy:

$$q_{\pi}(s, a) = \mathbb{E} \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, A_t = a, A_{t+1:\infty} \sim \pi \right]$$

Action-value function
for the optimal policy and $\gamma=0.9$

State	Action	Value
A	1	130.39
A	2	133.77
B	1	166.23
B	2	146.23



Optimal policies

- A policy π_* is **optimal** iff it maximizes the action-value function:

$$q_{\pi_*}(s, a) = \max_{\pi} q_{\pi}(s, a) = q_*(s, a)$$

- Thus all optimal policies share the same **optimal value function**
- Given the optimal value function, it is easy to act optimally:

$$\pi_*(s) = \arg \max_a q_*(s, a) \quad \text{“greedification”}$$

- We say that the optimal policy is **greedy** with respect to the optimal value function
- There is always at least one deterministic optimal policy

Summary from first principles

- Intelligence is all about achieving *goals*
- Goals can be formulated as maximizing *reward*
 - e.g. expected cumulative discounted reward over time
- We maximize reward by finding and following an optimal policy π_*
- To find π_* we need to first find the optimal value function q_*
- To find q_* we need to repeatedly find the value function q_π for a policy that is our current best guess at the optimal policy
- Thus, intelligence is all about estimating the value of the current policy!