

# Multi-Armed Bandits

## Machine Learning

Daniele Loiacono



**POLITECNICO**  
MILANO 1863

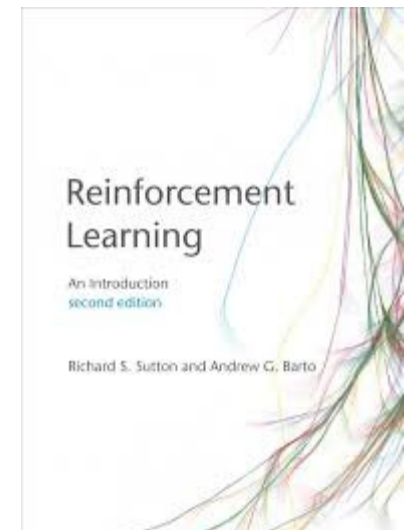
# Outline and References

## □ Outline

- ▶ K-armed problems
- ▶ Action-Values
- ▶ Incremental Update and Non-Stationary Problems
- ▶ Epsilon-Greedy Action Selection
- ▶ Optimistic Initial Values
- ▶ UCB Action Selection


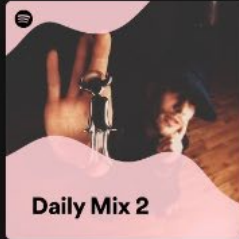




## □ References

- ▶ [Reinforcement Learning: An Introduction](#) [RL Chapter 2]
- ▶ [Fundamentals of Reinforcement Learning](#) (Coursera)

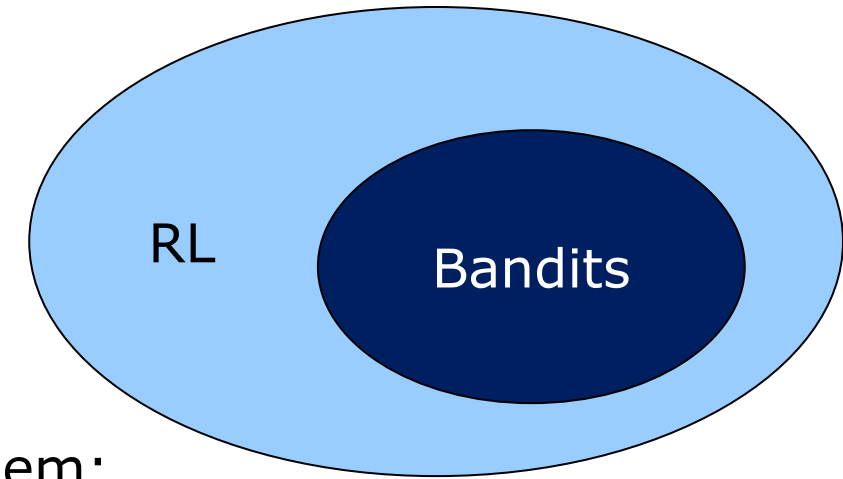


# Making decisions under uncertainty



Daily Mix 1	Daily Mix 2	Daily Mix 3	Daily Mix 4	Daily Mix 5	Daily Mix 6
					
<b>Daily Mix 1</b> New York Dolls, Talking Heads, The Clash and more	<b>Daily Mix 2</b> Jamiroquai, Maroon 5, Christina Aguilera and more	<b>Daily Mix 3</b> The Cramps, Jack White, The Gun Club and more	<b>Daily Mix 4</b> Mr. Rain, Cosmo, Nayt and more	<b>Daily Mix 5</b> Jimi Hendrix, Frank Zappa, Queen and more	<b>Daily Mix 6</b> The Undertones, Stiff Little Fingers, The Specials and...

# The k-armed Bandit Problem

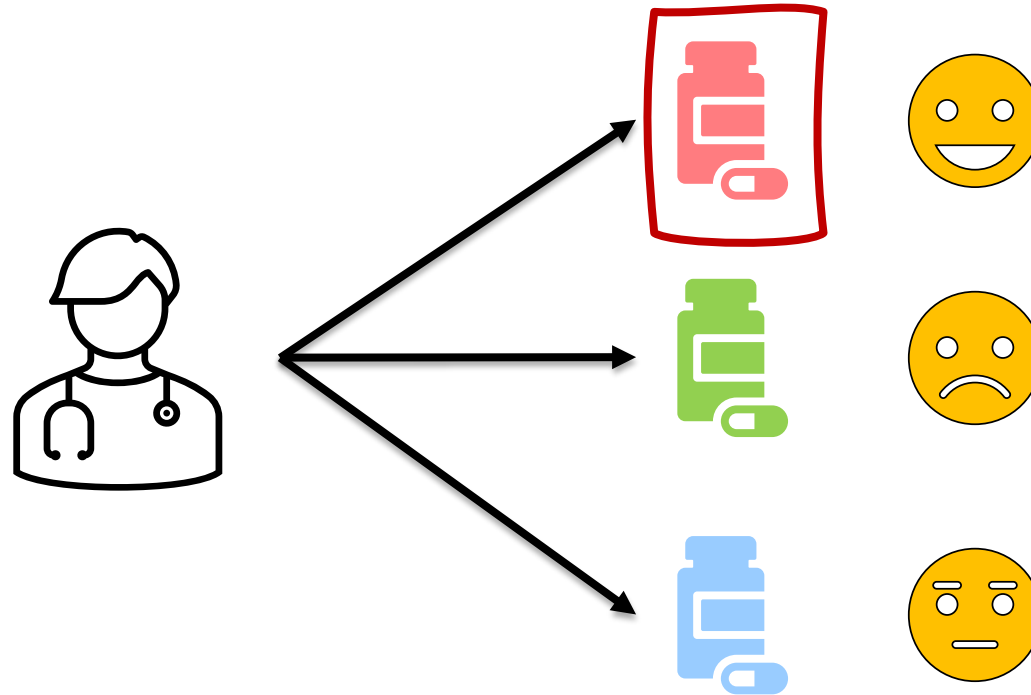


- ❑ It is the simplest form of Reinforcement Learning problem:

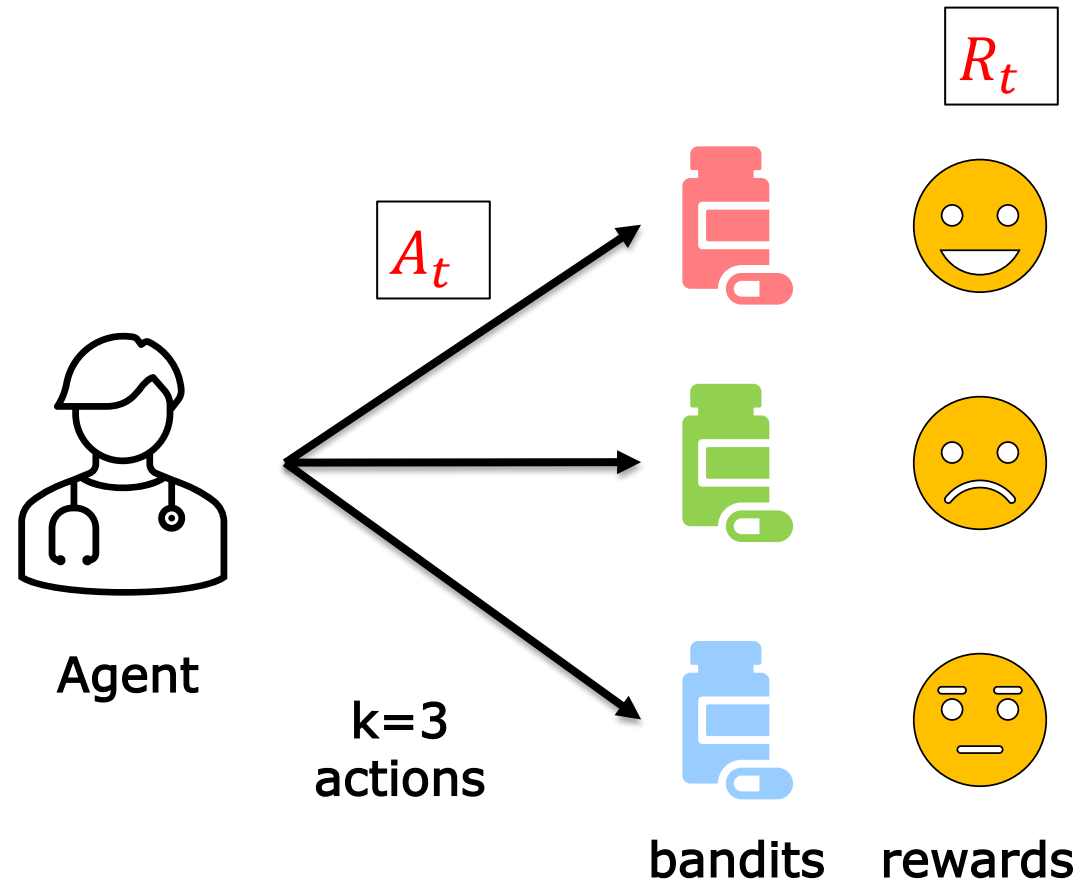
In the k-armed bandit problem, we have an **agent** who chooses between k **actions** and receives a **reward** based on action it chooses.

- ▶ Goal is to find optimal decision (**action**) among k options
- ▶ Optimal decision is not context-dependent (**no state**)
- ▶ Feedback consists of an evaluation (**reward**) of decisions under **uncertainty**
- ▶ Learning by **trial and error** and through **interaction with environment**

# The k-armed Bandit Problem



# The k-armed Bandit Problem



## Action-Values

# Action-Values

- The **value** of each action is defined as the **expected reward**:

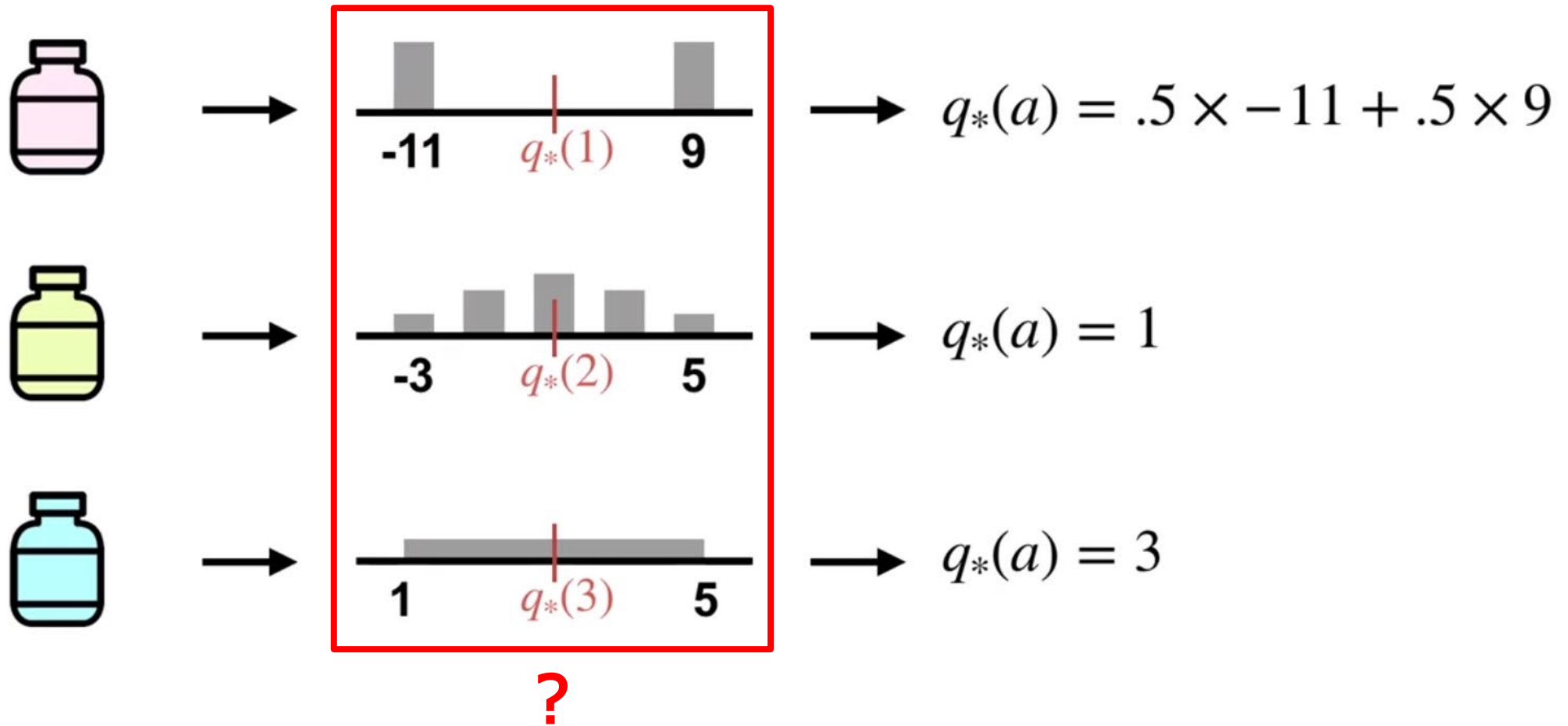
$$q^*(a) \doteq \mathbb{E}[R_t | A_t = a] = \sum p(r|a)r \quad \forall a \in \{1, \dots, k\}$$

- The goal of the agent is to **maximize** the **expected reward**:

$$\operatorname{argmax}_a q^*(a)$$



# Computing $q^*(a)$



## Estimate of $q^*(a)$

□ As  $p(r|a)$  is not known, we estimate  $q^*(a)$  from experience:

sum of rewards when  $a$  chosen before step  $t$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

times  $a$  chosen before step  $t$

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1

2

3

4

5

6

7

8

9

10

11

12

$Q_t(a) =$

0.0

0.0

0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



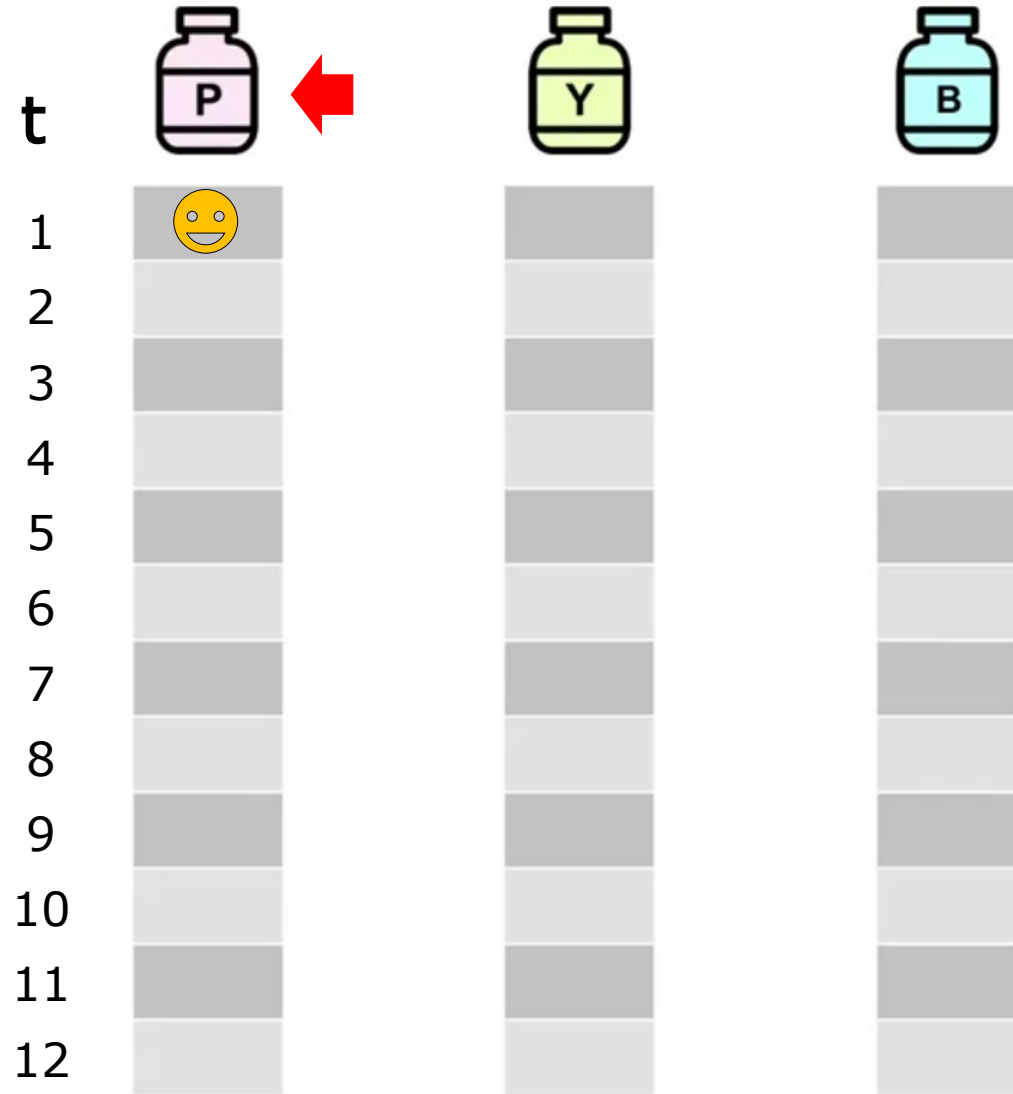
0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$



$Q_t(a) =$  1.0 0.0 0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



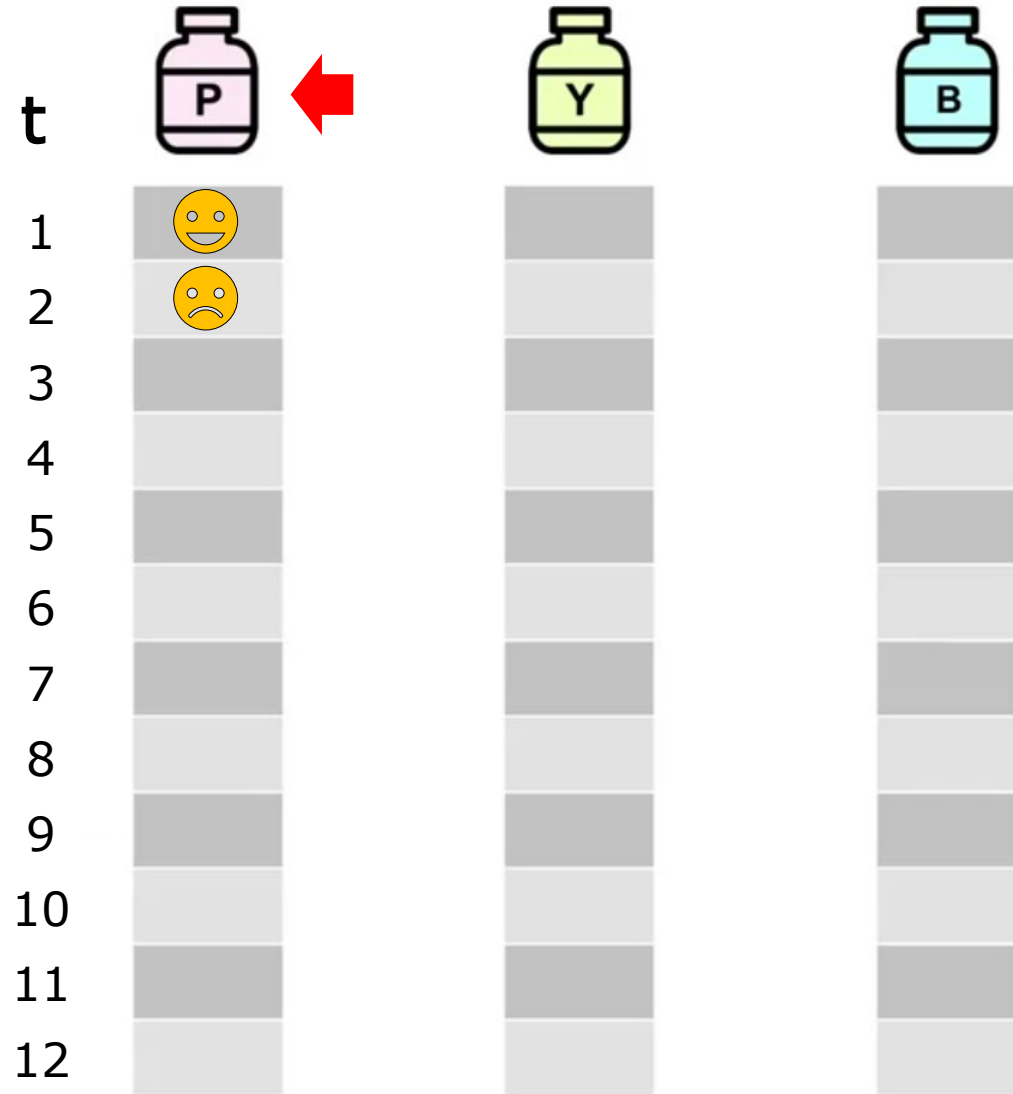
0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$



$Q_t(a) =$  0.5 0.0 0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



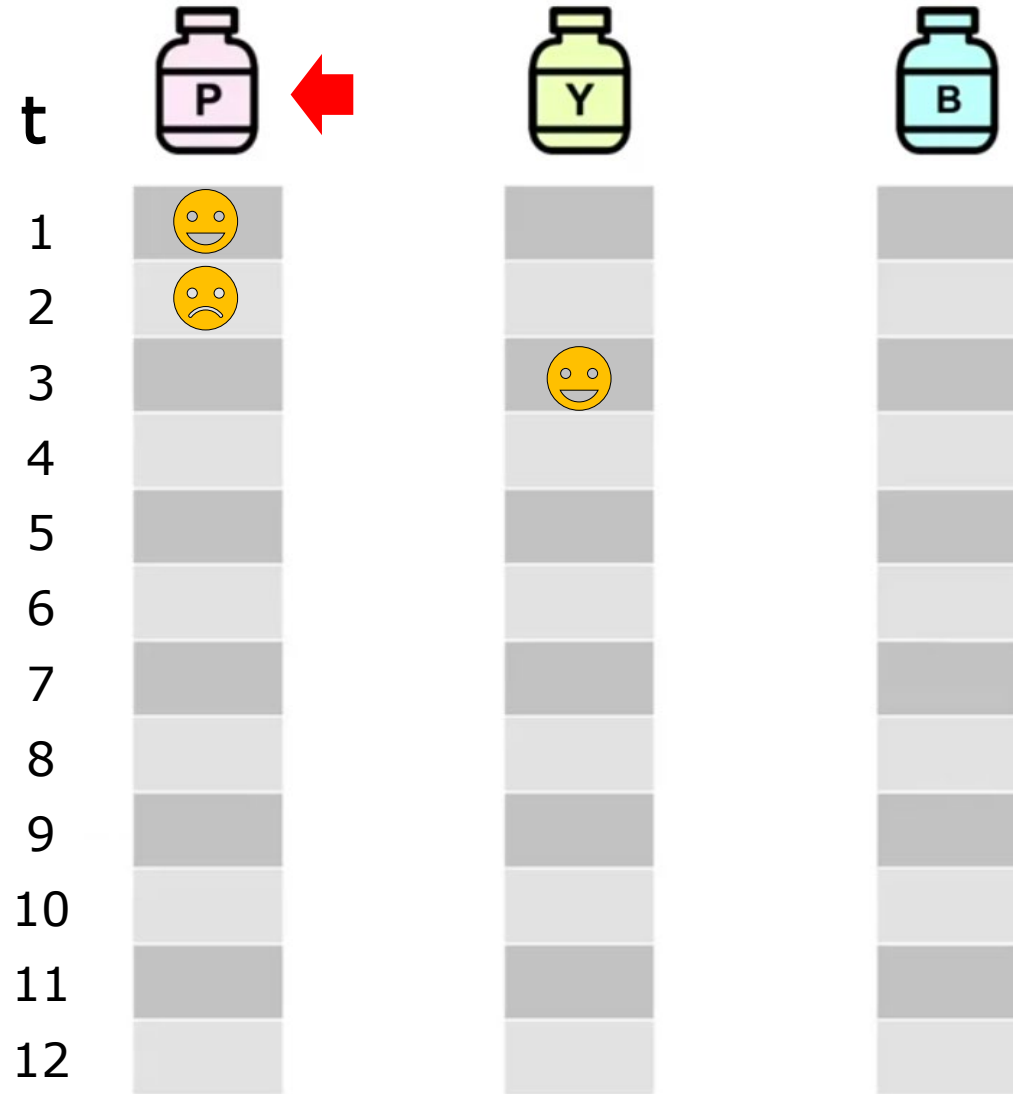
0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$



$Q_t(a) =$  0.5 1.0 0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5

6

7

8

9

10

11

12

$Q_t(a) =$

0.5

1.0

0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



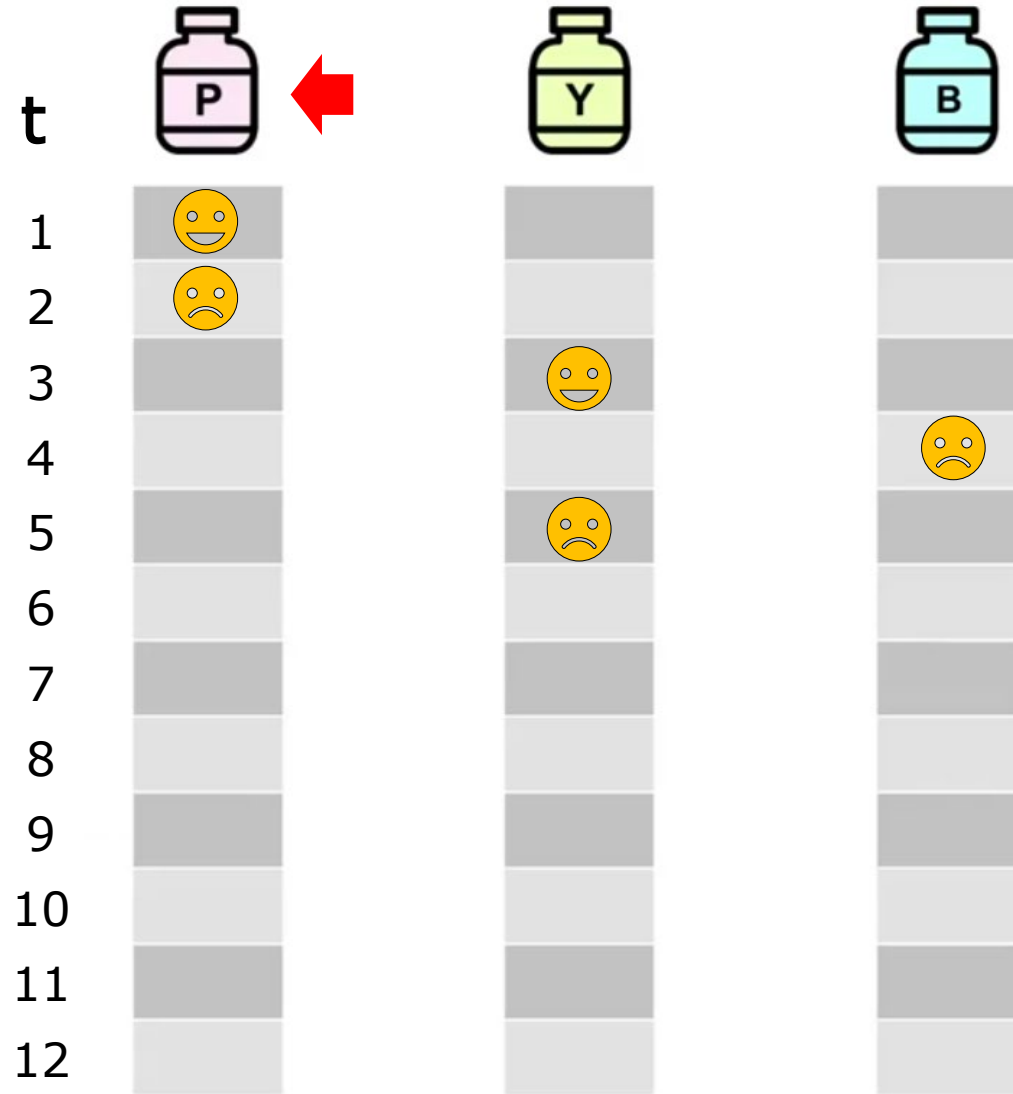
0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$



$Q_t(a) =$       0.5      0.5      0.0



# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7

8

9

10

11

12

$Q_t(a) =$

0.5

0.5

0.5

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8

9

10

11

12

$Q_t(a) = 0.33$

0.5

0.5

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8



9

10

11

12

$Q_t(a) = 0.33$

0.5

0.66

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8



9



10

11

12

$Q_t(a) = 0.33$

0.66

0.66

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8



9



10



11

12

$Q_t(a) = 0.33$

0.66

0.5

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8



9



10



11



12

$Q_t(a) = 0.25$

0.66

0.5

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_t=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_t=a}}$$

t



1



2



3



4



5



6



7



8



9



10



11



12



$Q_t(a) =$

0.25

0.75

0.5

## **Incremental update and non-stationary problems**



# Incremental update of action-values

□ Let's consider the update of a single action:

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i = \frac{1}{n} \left( R_n + (n-1) \frac{1}{n-1} \sum_{i=1}^{n-1} R_i \right) \\ &= \frac{1}{n} (R_n + (n-1)Q_n) \\ &= Q_n + \frac{1}{n} (R_n - Q_n) \end{aligned}$$

$$Q_n = \frac{1}{n-1} \sum_{i=1}^{n-1} R_i$$

# Incremental update of action-values

The diagram illustrates the incremental update of action-values formula,  $Q_{n+1} = Q_n + \frac{1}{n} (R_n - Q_n)$ . The formula is presented with color-coded boxes around its components:  $Q_{n+1}$  is in a red box,  $Q_n$  is in a blue box,  $\frac{1}{n}$  is in a green box, and  $(R_n - Q_n)$  is in a purple box. Four labels in orange boxes are connected to these components by lines: 'new estimate' points to  $Q_{n+1}$ , 'old estimate' points to  $Q_n$ , 'step size' points to  $\frac{1}{n}$ , and 'target - old estimate' points to  $(R_n - Q_n)$ .

new estimate

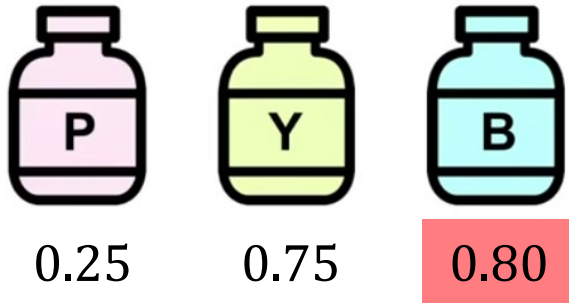
old estimate

step size

target - old estimate

$$Q_{n+1} = Q_n + \frac{1}{n} (R_n - Q_n)$$

# Non-stationary bandit problems

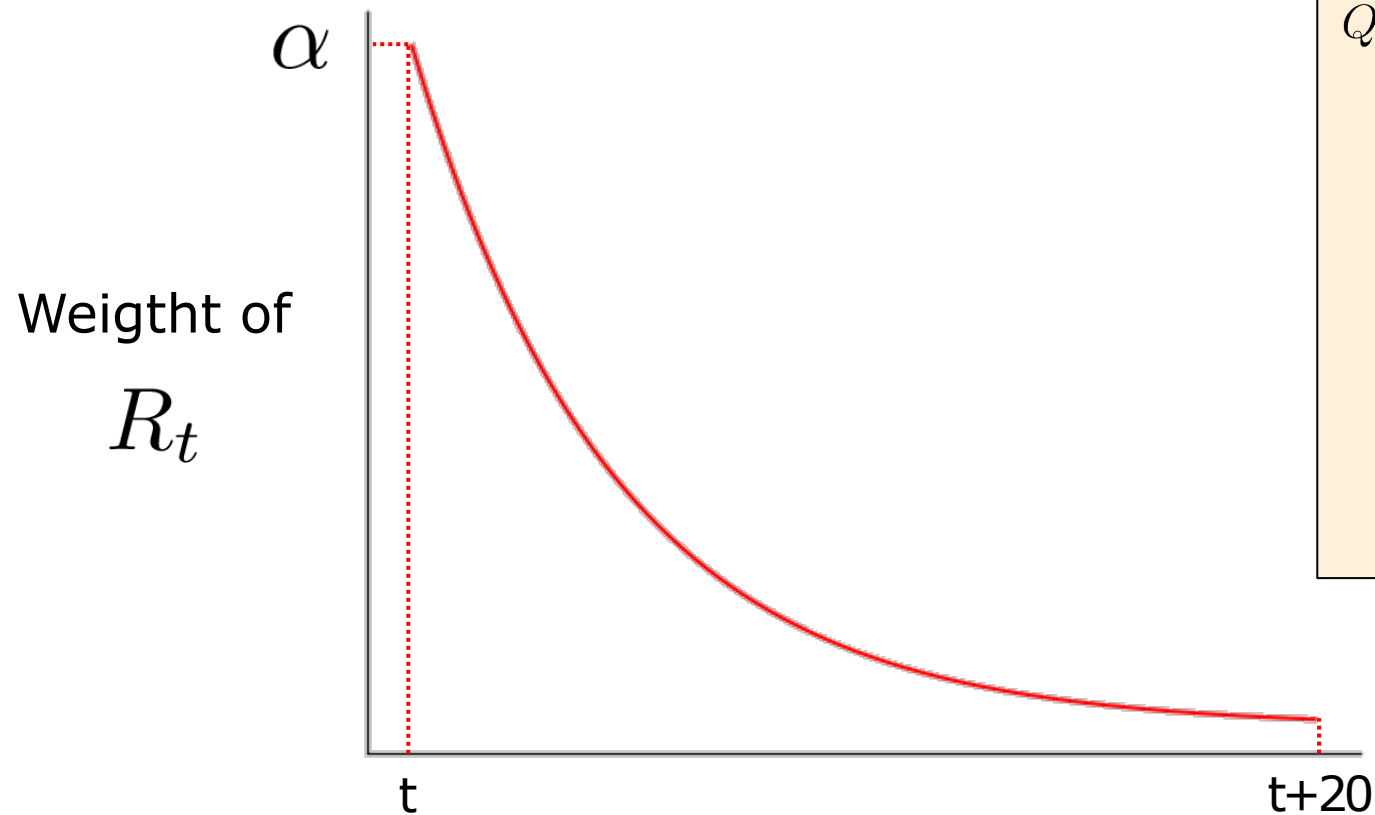


$$\alpha \in [0, 1]$$

$$Q_{n+1} = Q_n + \alpha (R_n - Q_n)$$

# Non-stationary bandit problems

$$Q_{n+1} = Q_n + \alpha (R_n - Q_n)$$



$$\begin{aligned} Q_{n+1} &= Q_n + \alpha(R_n - Q_n) \\ &= \alpha R_n + (1 - \alpha)Q_n \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\ &= \alpha R_n + (1 - \alpha)\alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \dots \\ &\quad + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n \boxed{Q_1} \end{aligned}$$

Initial action-value

## Epsilon-Greedy Action Selection

# Action Selection

$$Q_{12}(\text{P}) = 0.25$$

$$Q_{12}(\text{B}) = 0.5$$

$$Q_{12}(\text{Y}) = 0.75$$

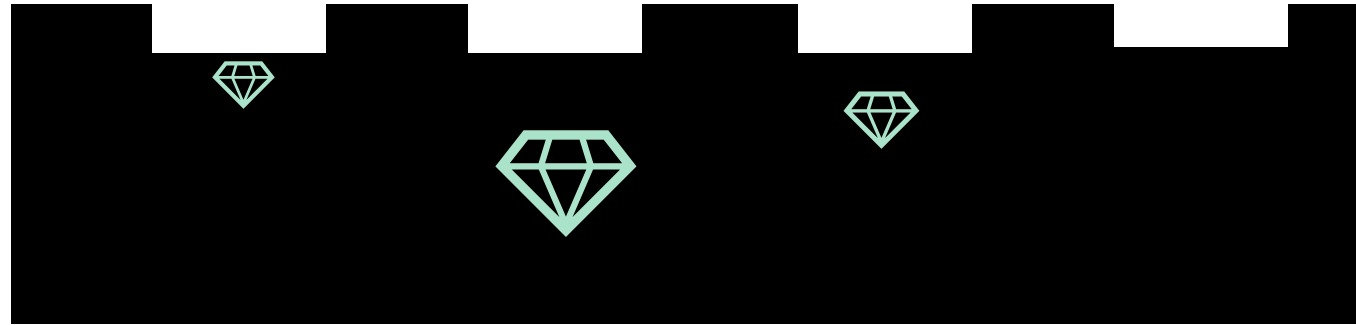
non-greedy actions

greedy action

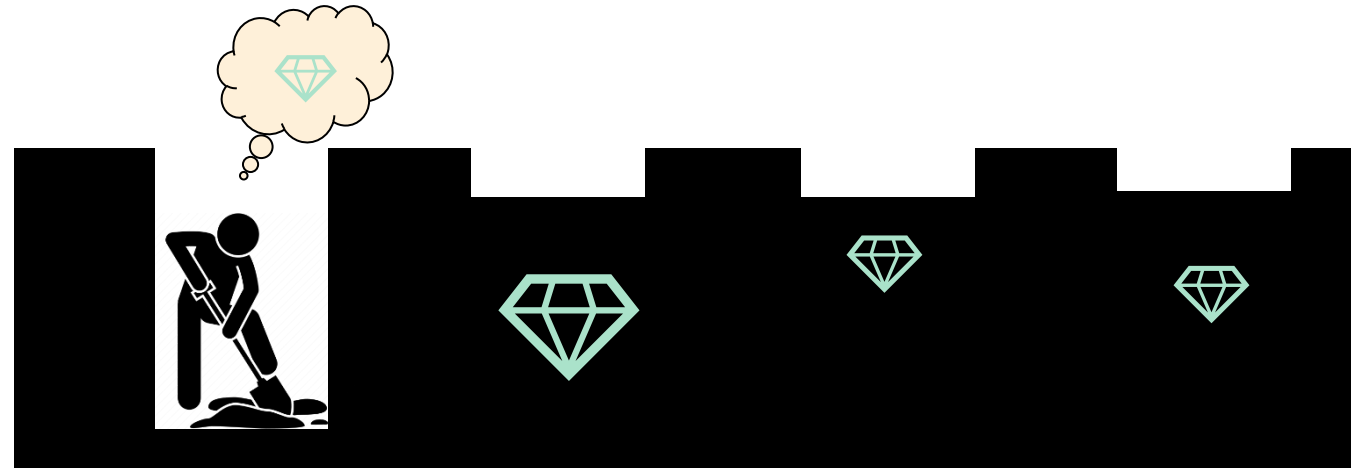
$$a_g = \underset{\text{P} \text{ B } \text{Y}}{\operatorname{argmax}} Q(a)$$

What would happen if we always select the greedy action?

# Exploration vs Exploitation



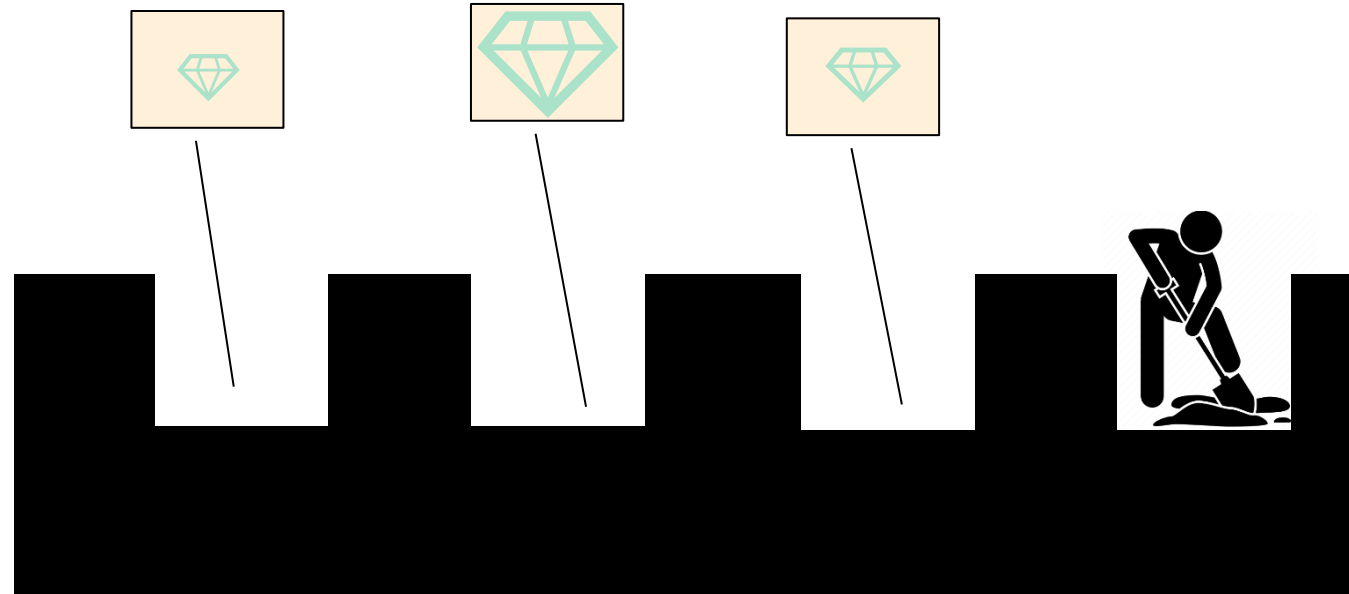
# Exploration vs Exploitation



- ❑ **Exploitation**: agent exploits knowledge for **short-term** benefit

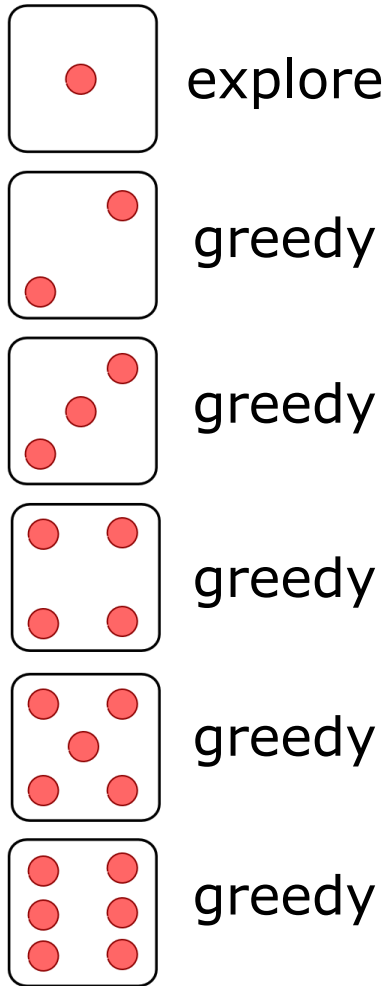


# Exploration vs Exploitation



- ❑ **Exploitation**: agent exploits its knowledge for **short-term** benefit
- ❑ **Exploration**: agent improves its knowledge for **long-term** benefit
- ❑ How to choose when to explore and when to exploit?

# Epsilon-Greedy Action Selection

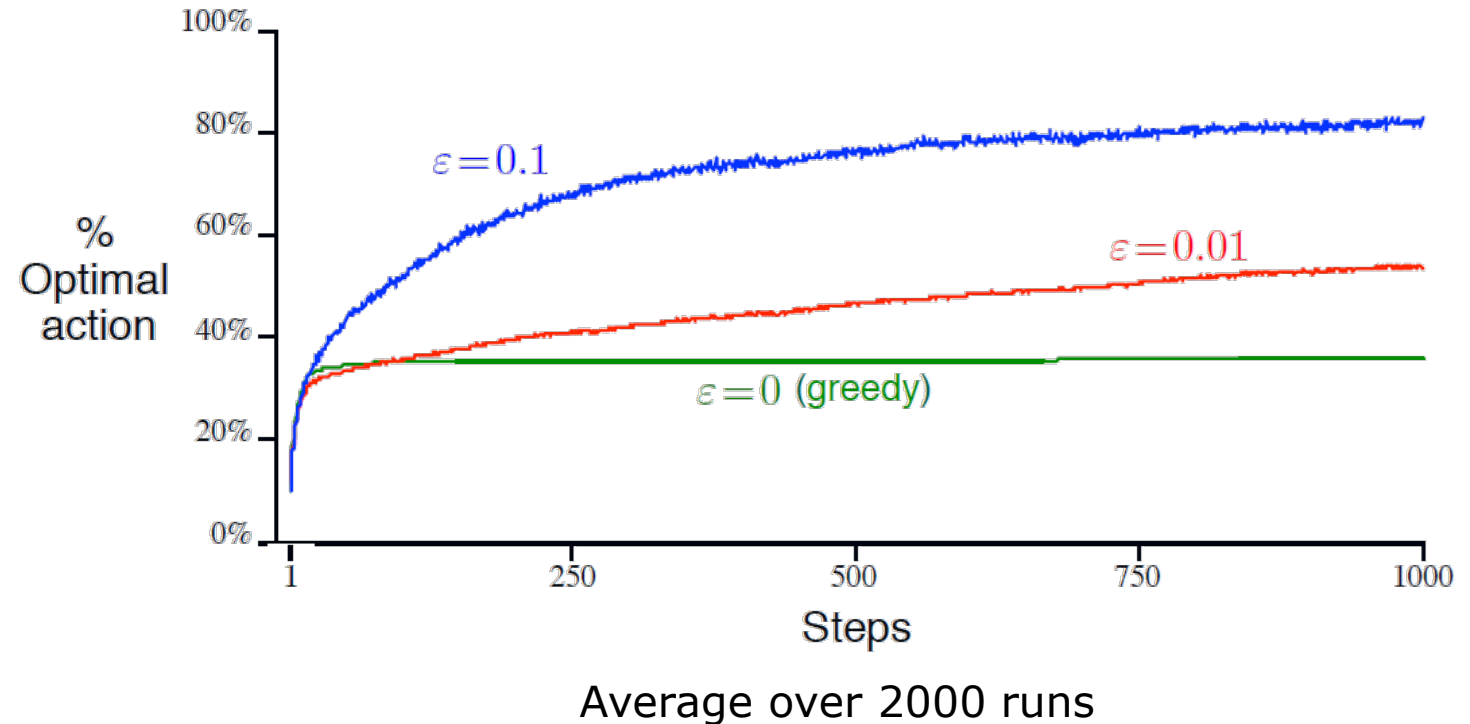
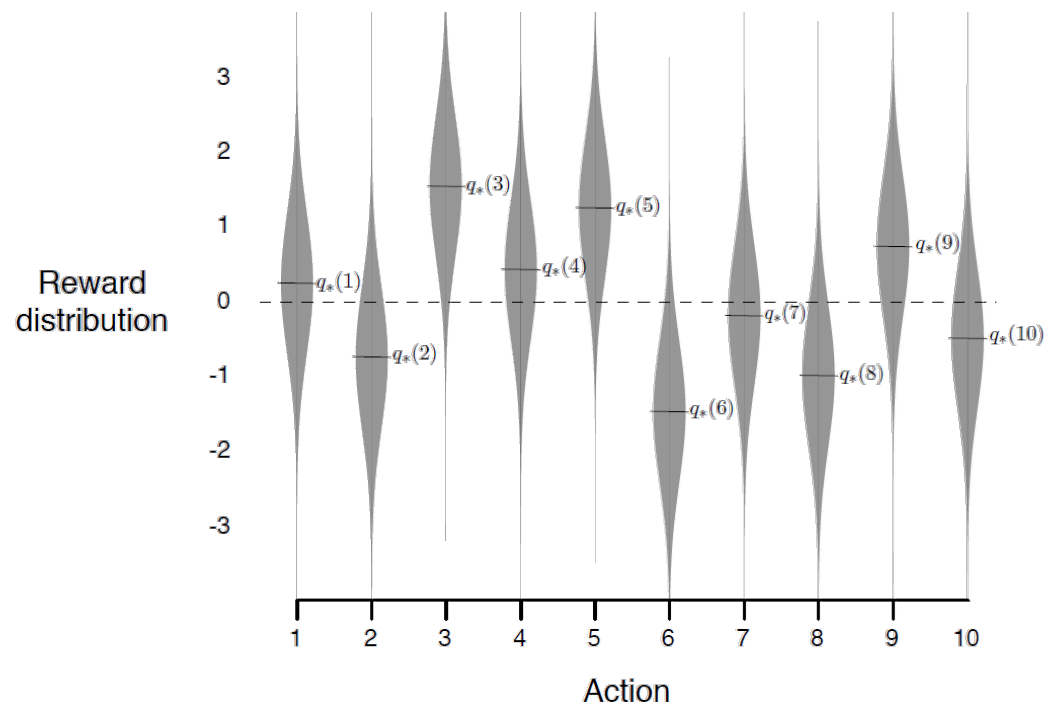


$$A_t = \begin{cases} \operatorname{argmax}_a Q_t(a) & \text{with probability } 1 - \epsilon \\ \operatorname{Uniform}(\{a_1, \dots, a_k\}) & \text{with probability } \epsilon \end{cases}$$

# Epsilon-Greedy Action Selection

$$A_t = \begin{cases} \operatorname{argmax}_a Q_t(a) & \text{with probability } 1 - \epsilon \\ \operatorname{Uniform}(\{a_1, \dots, a_k\}) & \text{with probability } \epsilon \end{cases}$$

10-armed testbed



## Optmistic Initial Values

## Optimistic Initial Values

- ❑ So far we initialized action-values to 0.0
- ❑ What happen if we initialize action-values to larger values?

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1

2

3

4

5

6

7

8

9

10

11

12

$Q_t(a) =$

0.0

0.0

0.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1

2

3

4

5

6

7

8

9

10

11

12

$Q_t(a) =$

2.0

2.0

2.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2

3

4

5

6

7

8

9

10

11

12

$Q_t(a) =$

1.5

2.0

2.0



## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3

4

5

6

7

8

9

10

11

12

$Q_t(a) =$

1.5

1.0

2.0

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4

5

6

7

8

9

10

11

12

$Q_t(a) =$

1.5

1.0

1.5

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5

6

7

8

9

10

11

12

$Q_t(a) =$

1.5

1.0

1.25

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6

7

8

9

10

11

12

$$Q_t(a) = 0.75$$

1.0

1.25

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7

8

9

10

11

12

$Q_t(a) =$

0.75

1.0

0.625

## Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8

9

10

11

12

$Q_t(a) =$

0.75

1.0

0.625

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8



9

10

11

12

$$Q_t(a) = 0.75$$

1.0

0.625

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8



9



10

11

12

$$Q_t(a) = 0.75$$

0.5

0.625



# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8



9



10



11

12

$$Q_t(a) = 0.375$$

0.5

0.625

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8



9



10



11



12

$$Q_t(a) = 0.375$$

$$0.5$$

$$0.3125$$

# Lets go back to clinical trial example...

Reward 1 if the treatment works, 0 otherwise



0.25



0.75



0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$

$$\alpha = 0.5$$

t



1



2



3



4



5



6



7



8



9



10



11



12

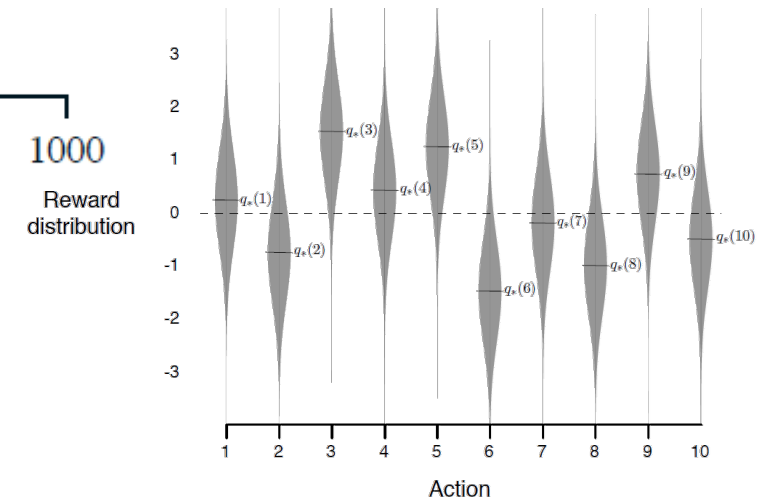
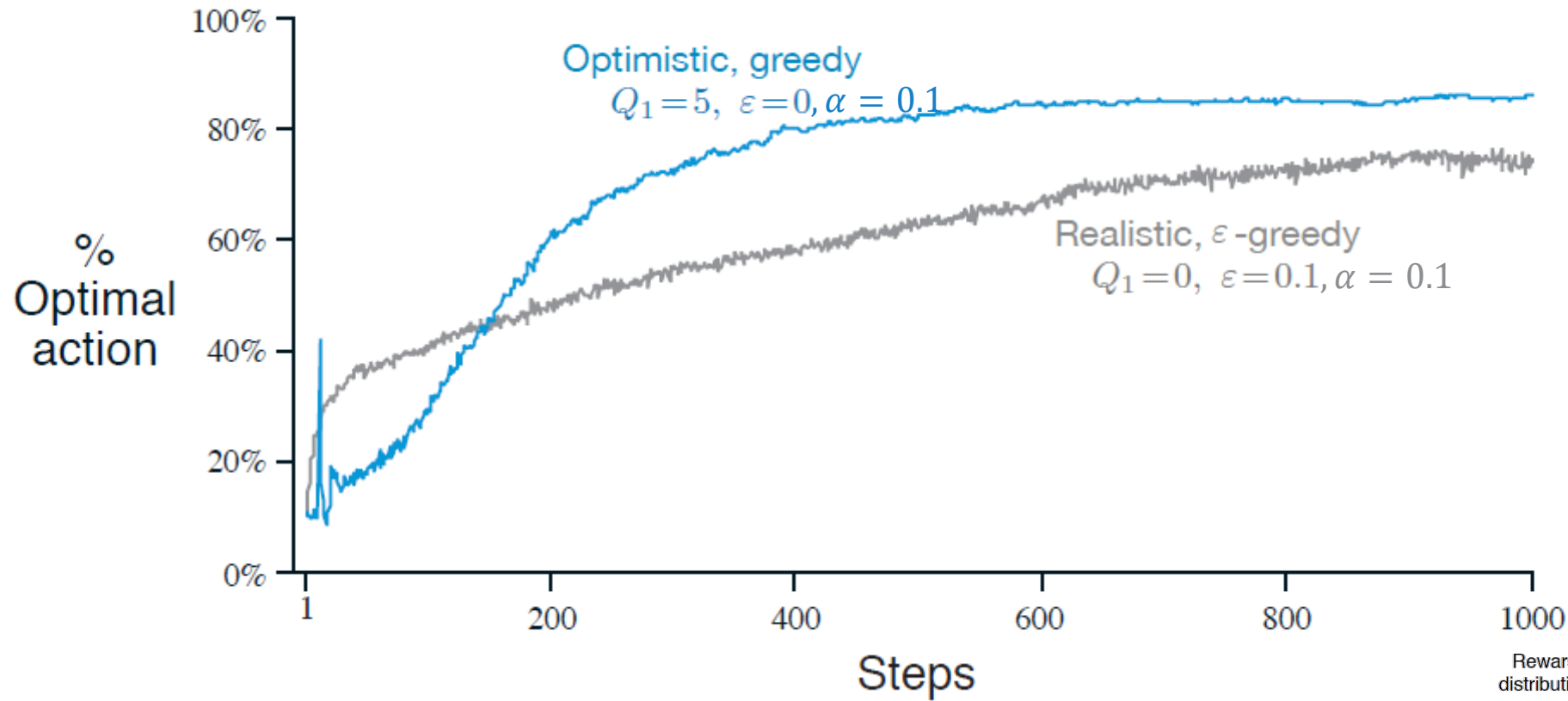


$$Q_t(a) = 0.375$$

$$0.75$$

$$0.3125$$

# Optimistic Initial Values on 10-armed testbed



## Limitations of optimistic initial values

- ❑ Optimistic initial values only drive early exploration
- ❑ They are not well-suited for non-stationary problems
- ❑ We may not know what other optimistic initial value should be

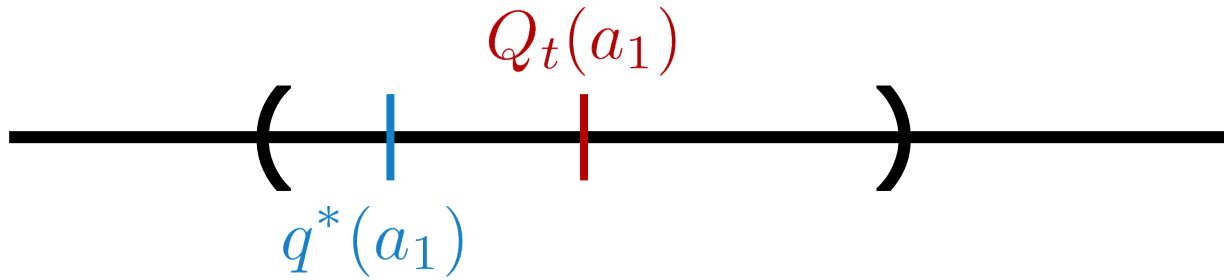
## UCB Action Selection

## Let's go back to epsilon-greedy

$$A_t = \begin{cases} \operatorname{argmax}_a Q_t(a) & \text{with probability } 1 - \epsilon \\ \boxed{\operatorname{Uniform}(\{a_1, \dots, a_k\})} & \text{with probability } \epsilon \end{cases}$$

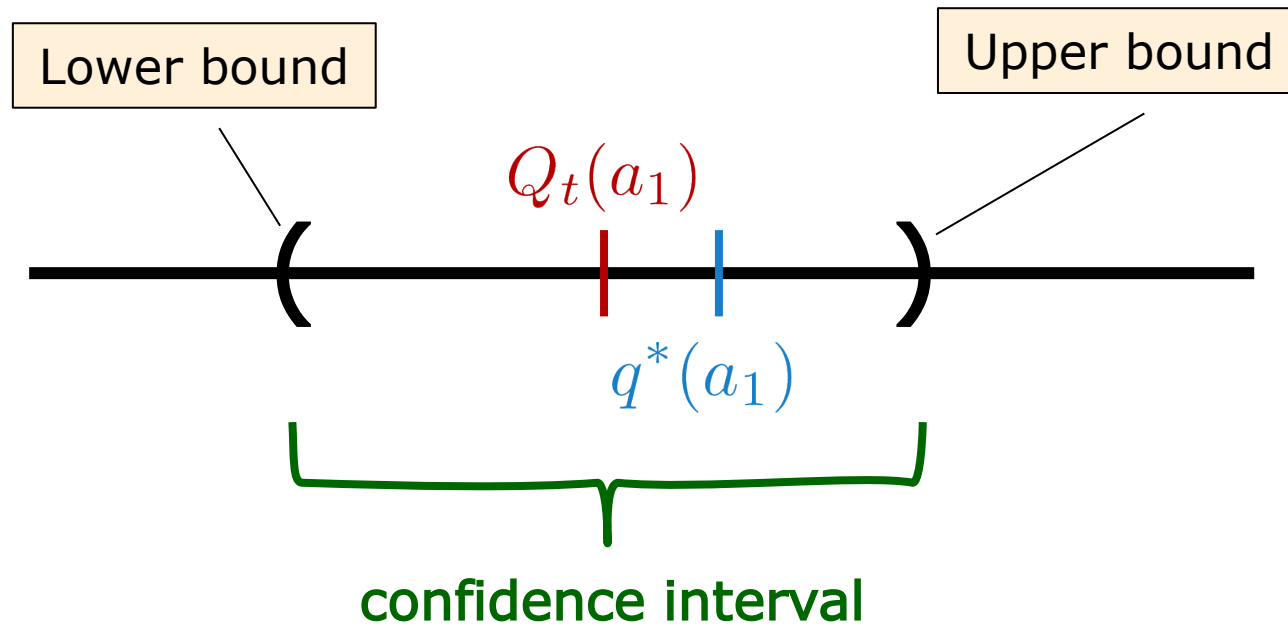
- ❑ Can we do better than uniform selection?
- ❑ Can we evaluate the uncertainty in action-value estimates?

# Uncertainty in Action-Values Estimates

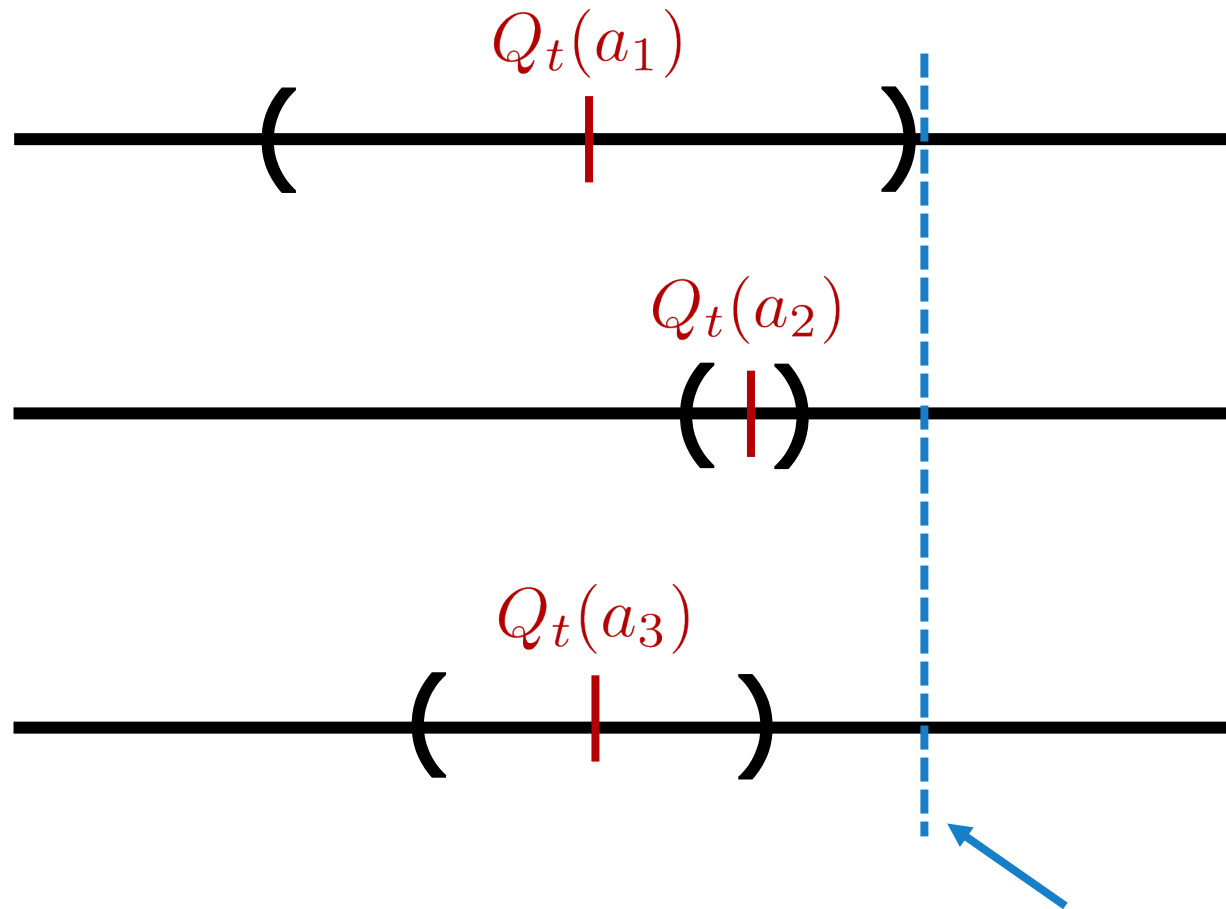




# Uncertainty in Action-Values Estimates



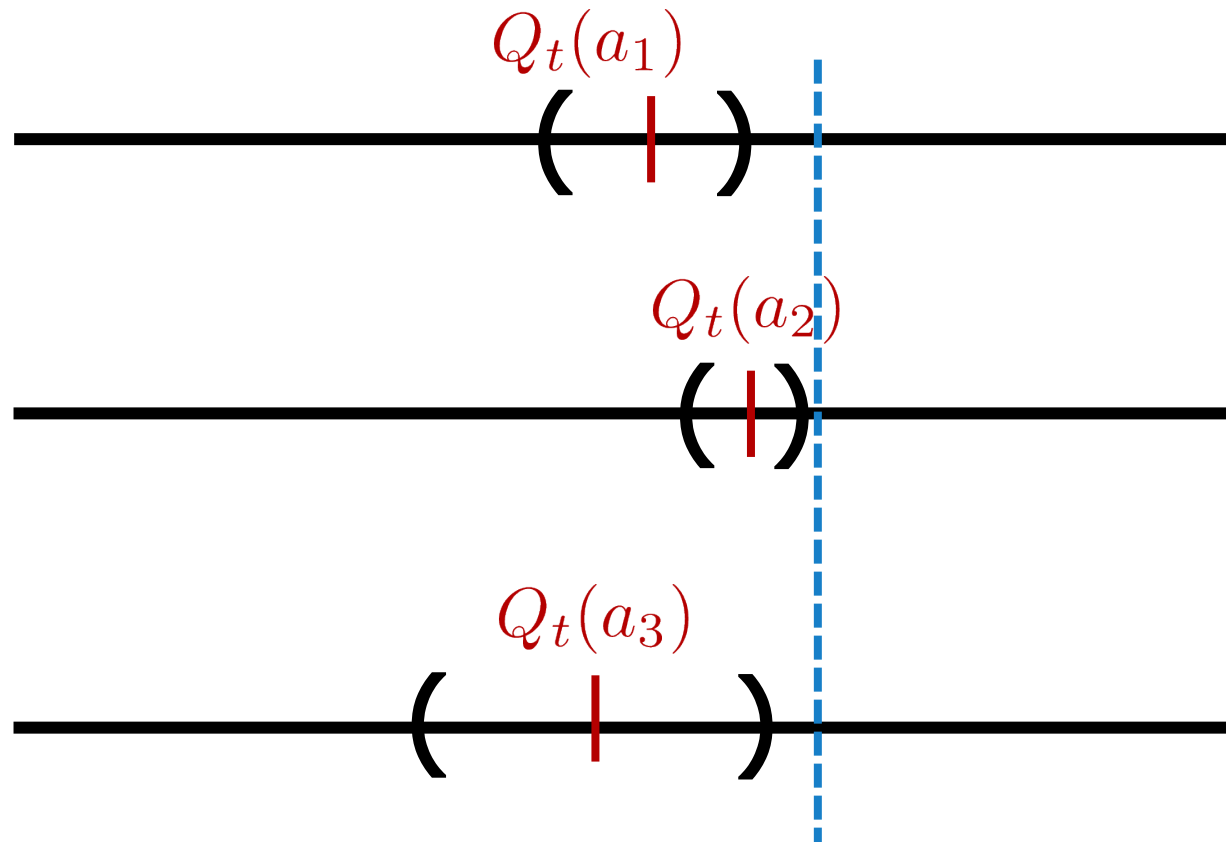
# Upper-Confidence Bound (UCB) Action Selection



Which action  
should be selected?

Optimism in the face of uncertainty

# Upper-Confidence Bound (UCB) Action Selection



How do we compute  
upper bound?

# Upper-Confidence Bound (UCB) Action Selection

The diagram illustrates the UCB action selection formula. A central equation is shown with three callout boxes pointing to its components:

$$A_t = \operatorname{argmax}_a \left[ Q_t(a) + c \sqrt{\frac{\ln(t)}{N_t(a)}} \right]$$

- The box labeled "Exploit" points to the  $Q_t(a)$  term.
- The box labeled "User-defined coefficient" points to the  $c$  term.
- The box labeled "Explore" points to the square root term  $\sqrt{\frac{\ln(t)}{N_t(a)}}$ .

# UCB on the 10-armed testbed

