Multi-Armed Bandits

Machine Learning

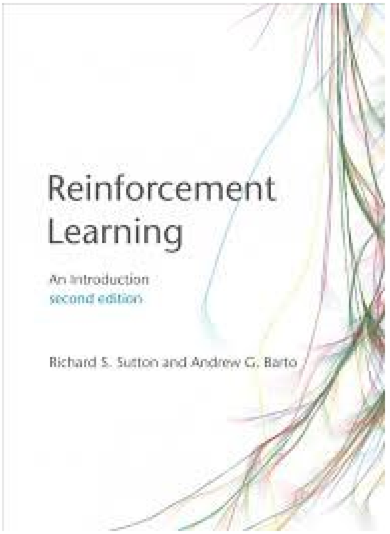
Daniele Loiacono

POLITECNICO
MILANO 1863

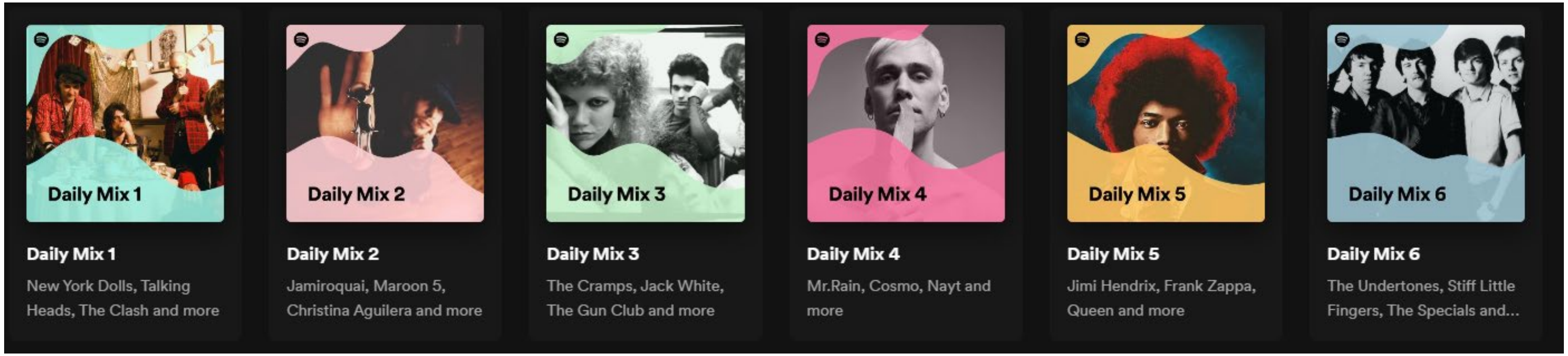
Outline and References

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

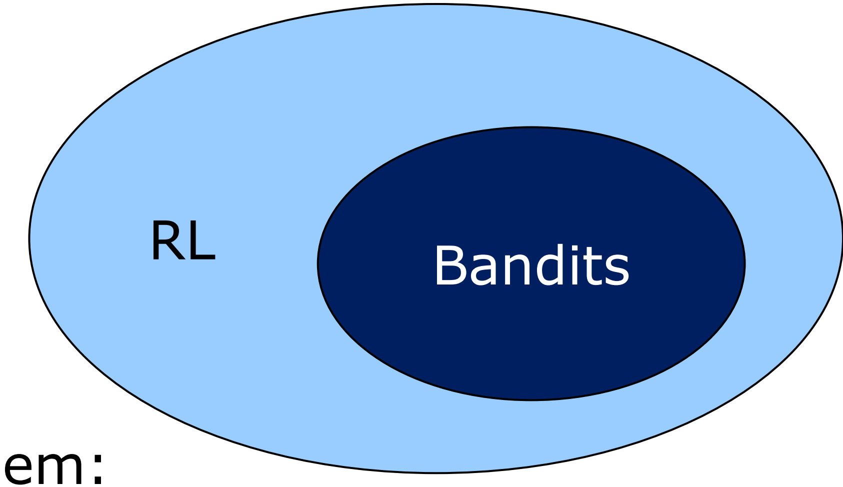
- References
  - Reinforcement Learning: An Introduction [RL Chapter 2]
  - Fundamentals of Reinforcement Learning (Coursera)



Making decisions under uncertainty



The k-armed Bandit Problem

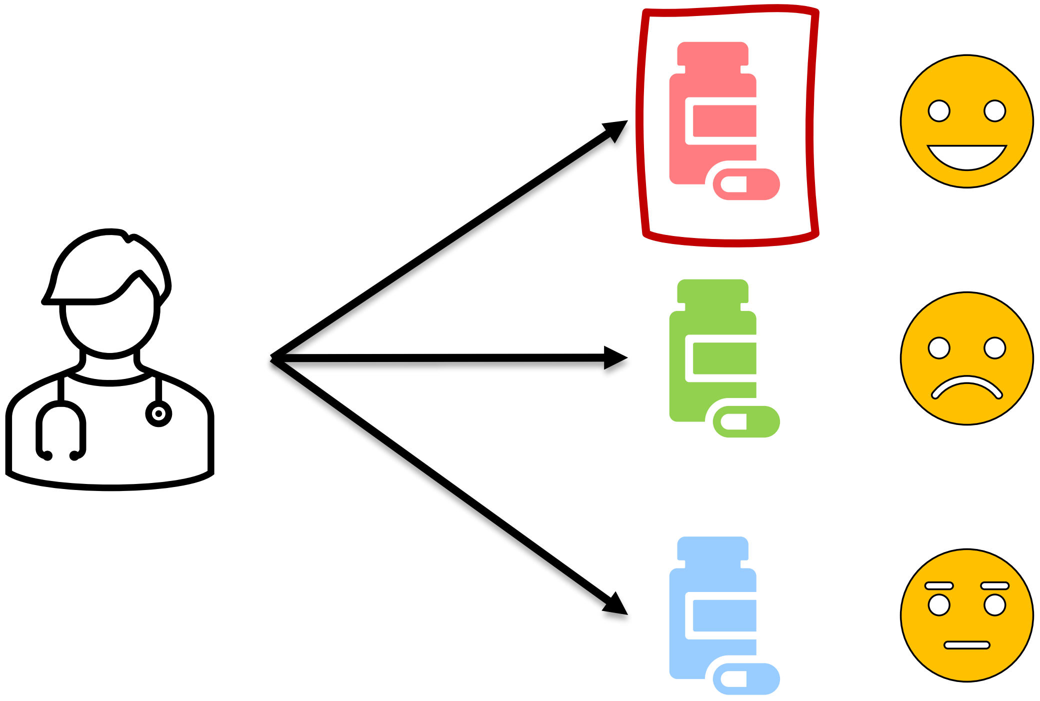


- It is the simplest form of Reinforcement Learning problem:

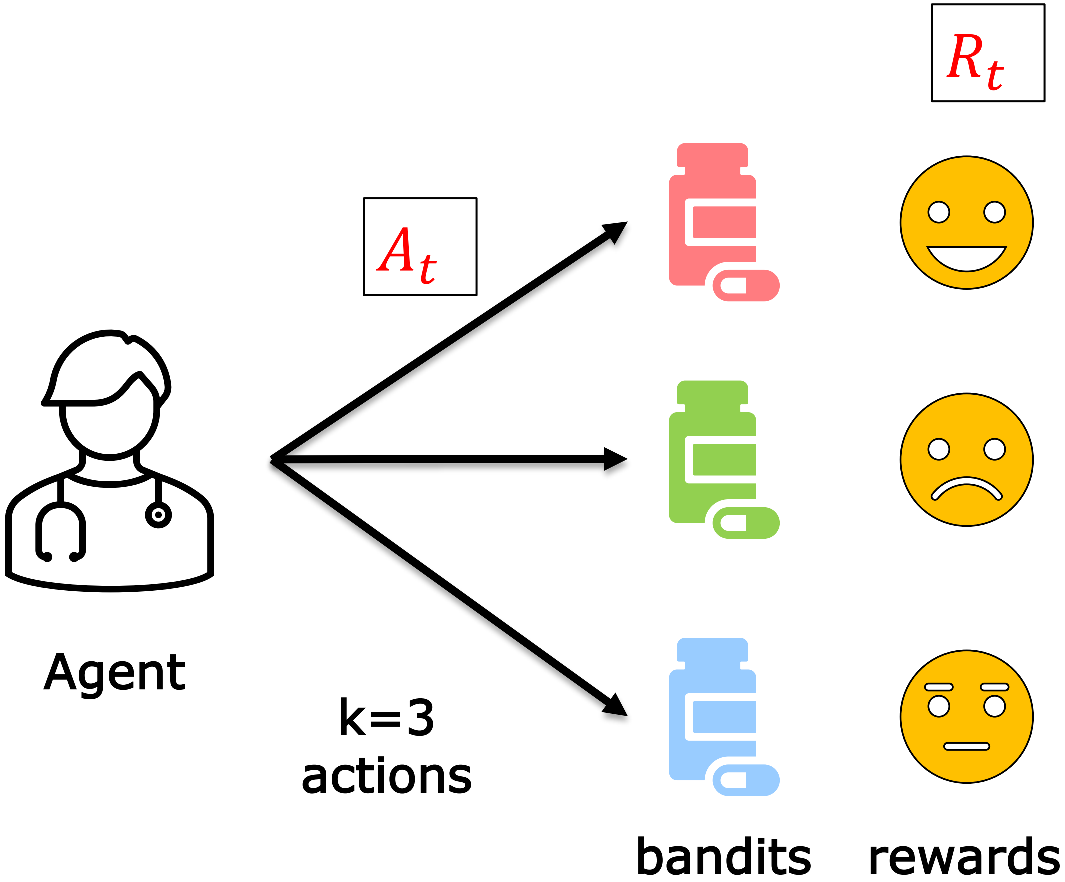
  In the k-armed bandit problem, we have an **agent** who chooses between k **actions** and receives a **reward** based on action it chooses.

  - 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

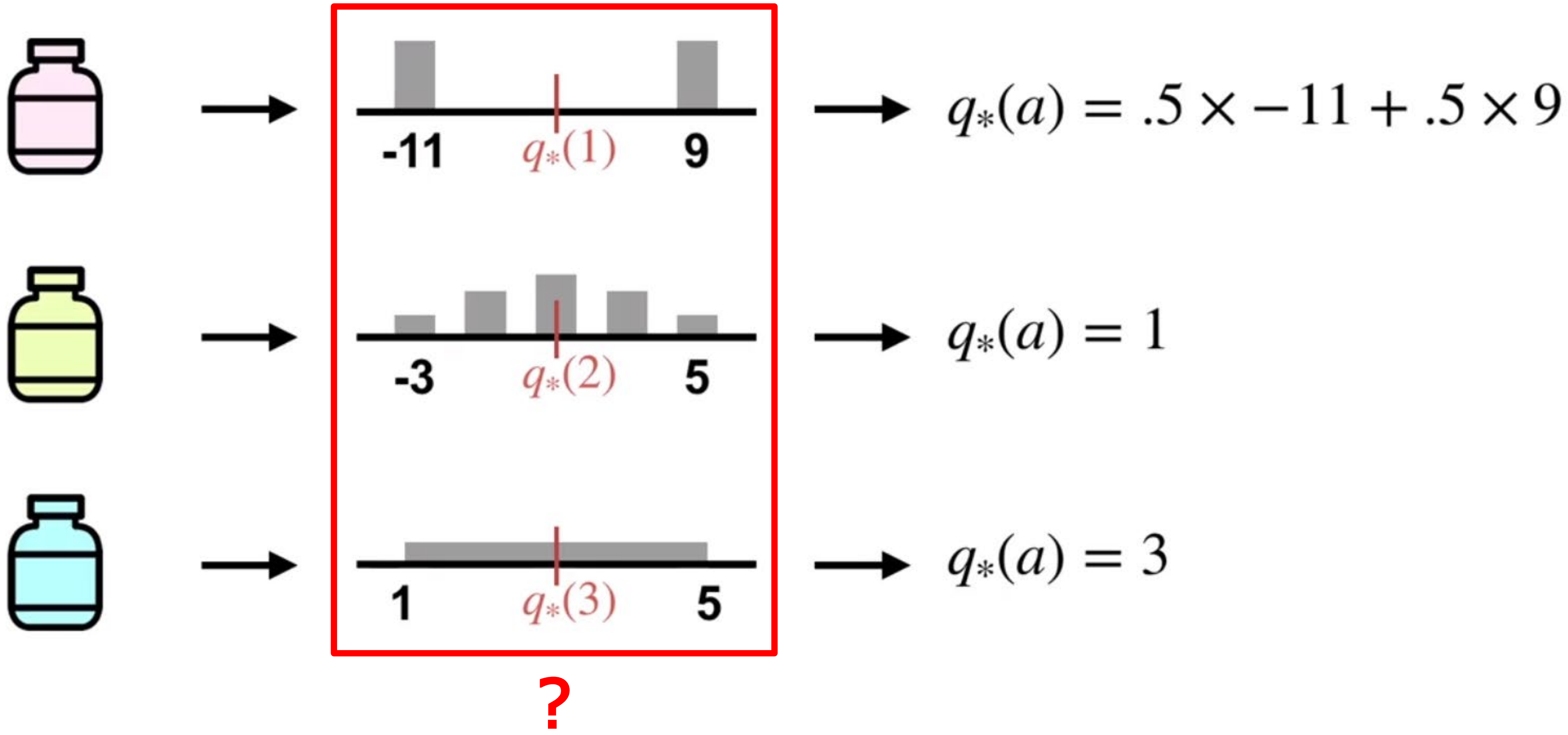
- The **value** of each action is defined as the **expected reward**:

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

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

$$argmax_a\, q^*(a)$$

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

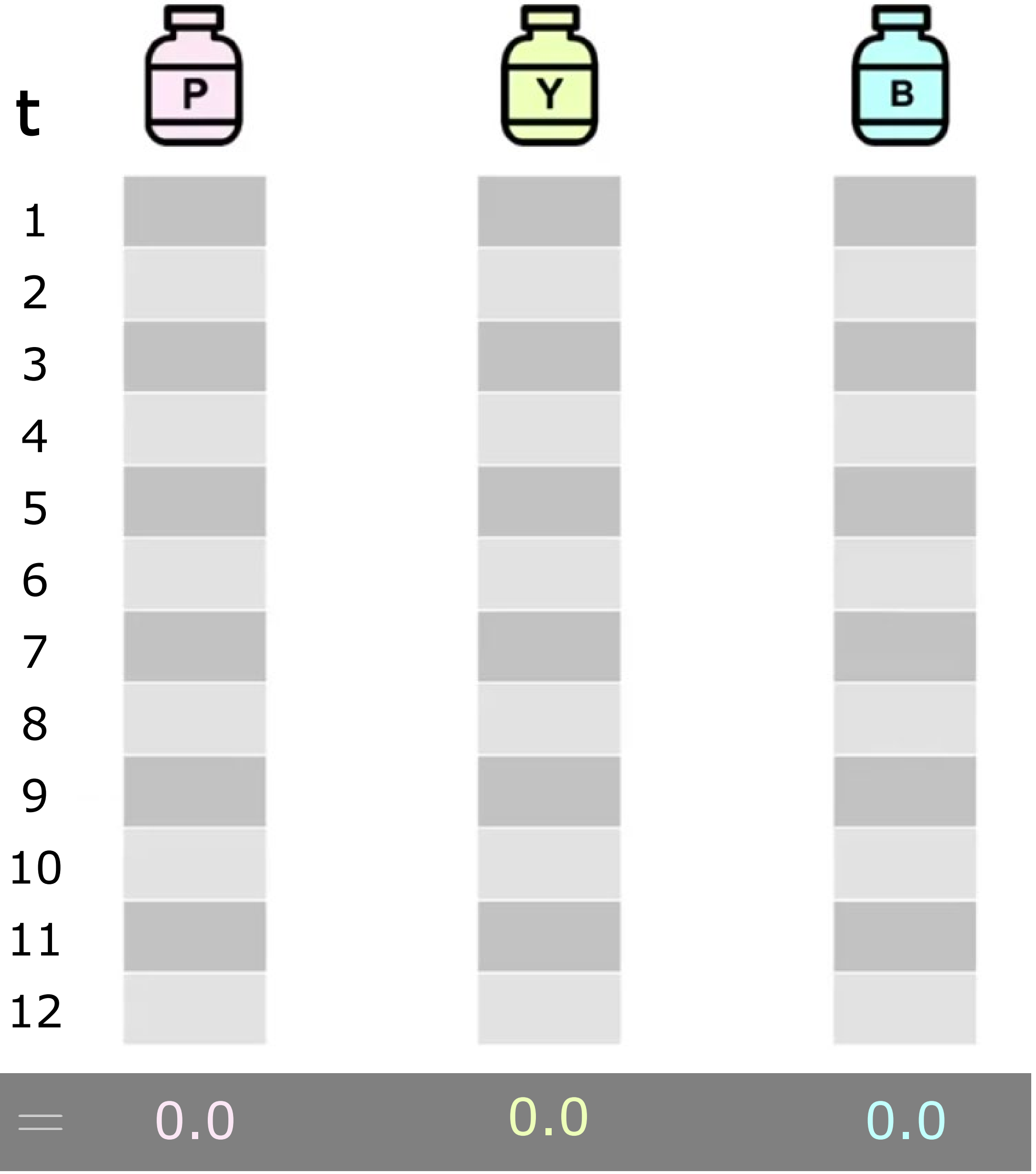
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25   0.75   0.50

$q^*(a)$

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



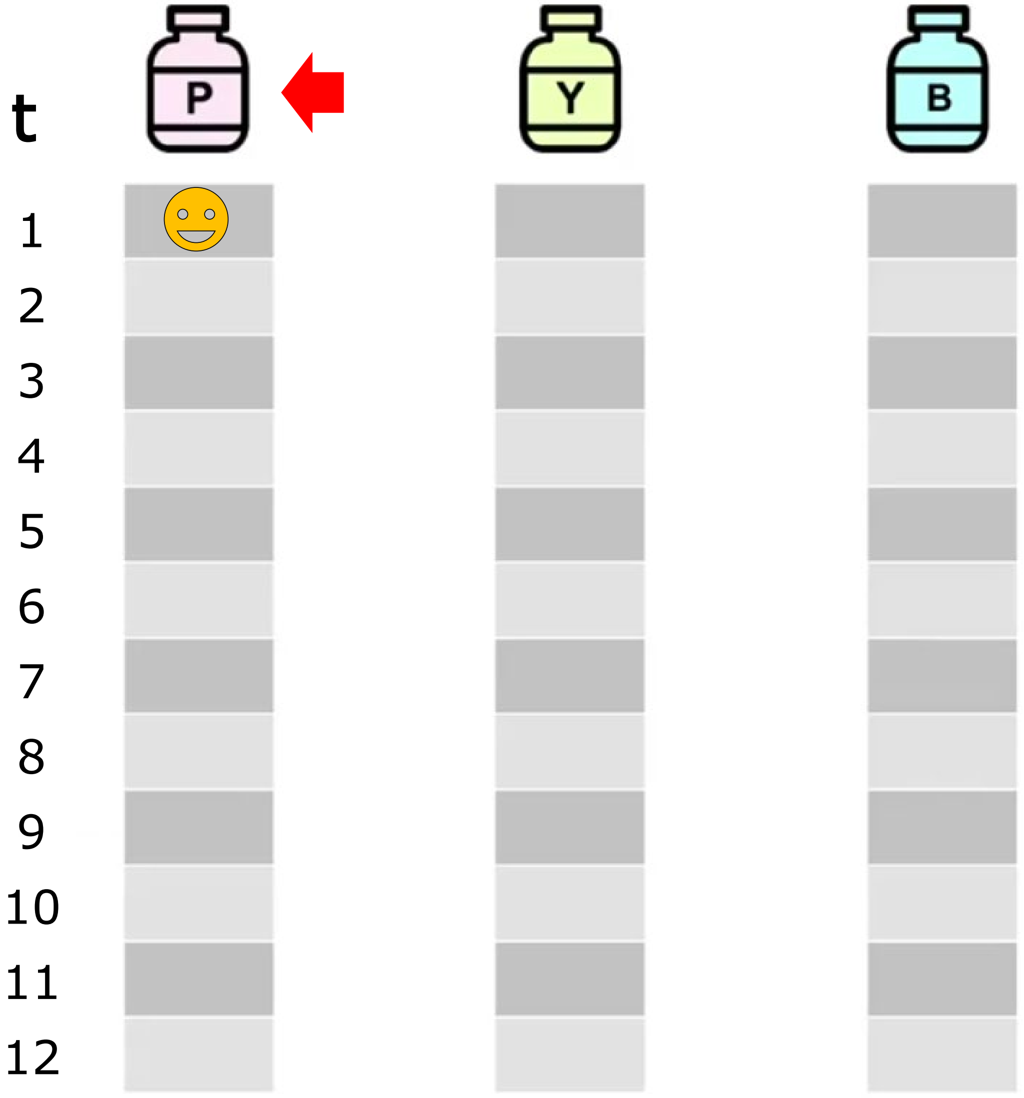
Lets go back to clinical trial example…

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

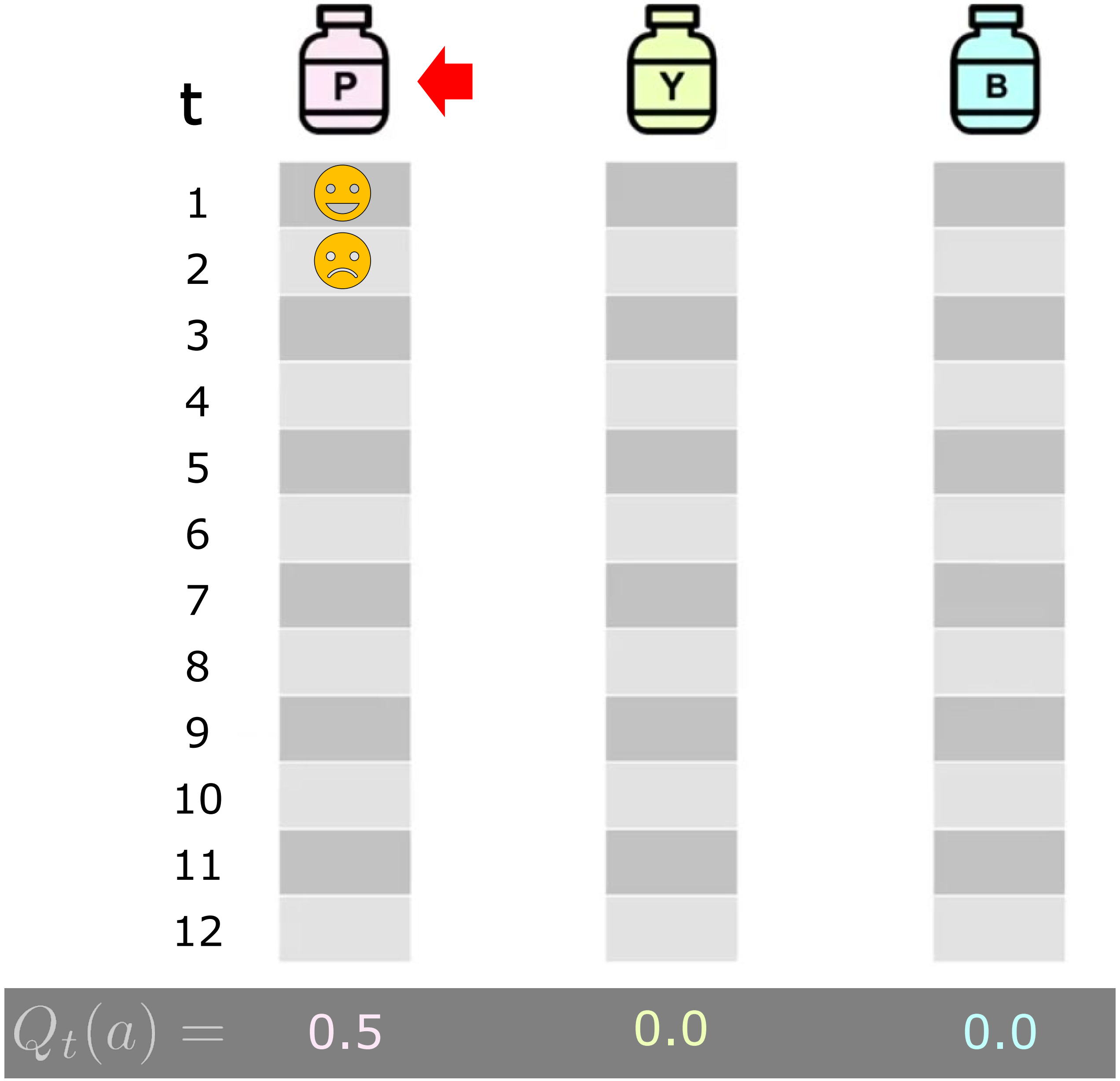
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



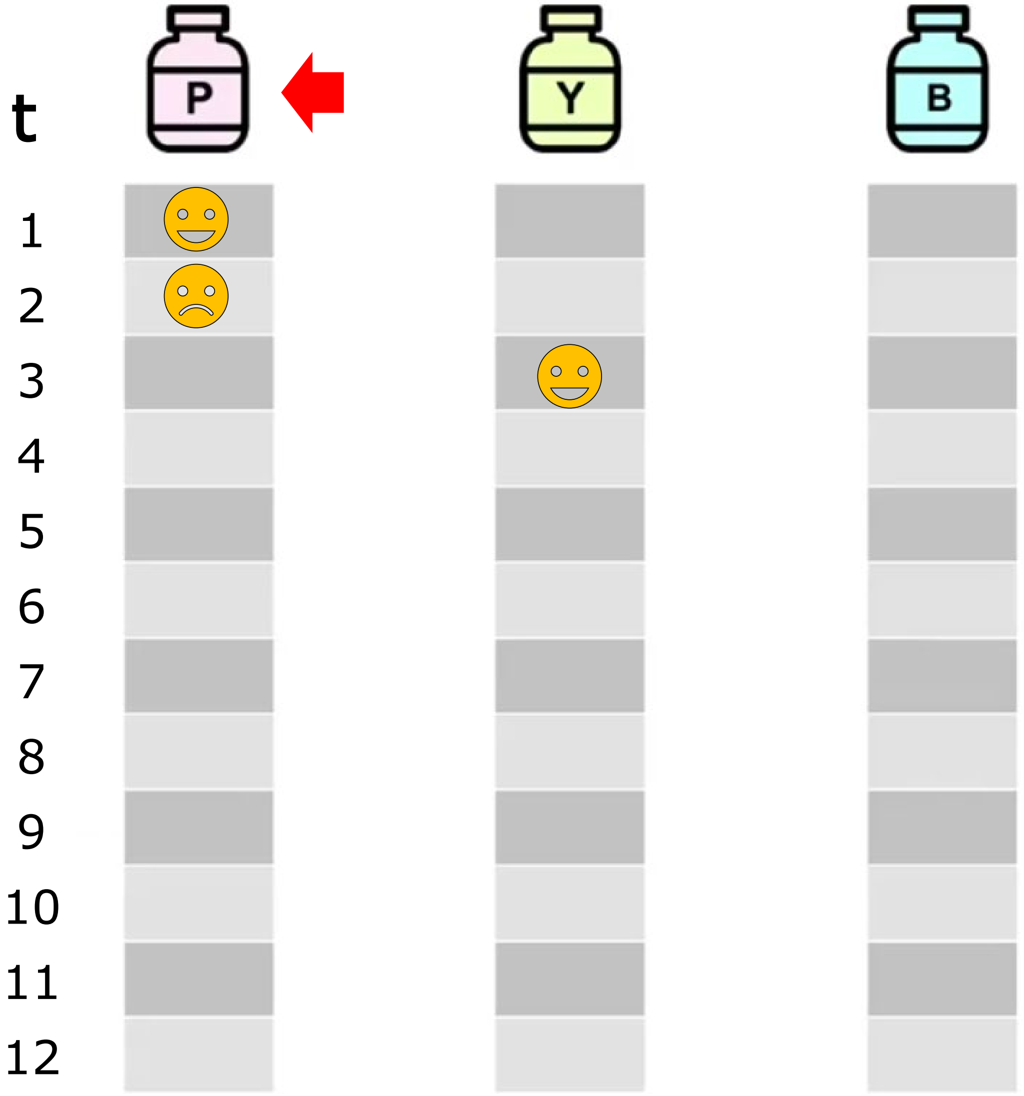
Lets go back to clinical trial example…

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

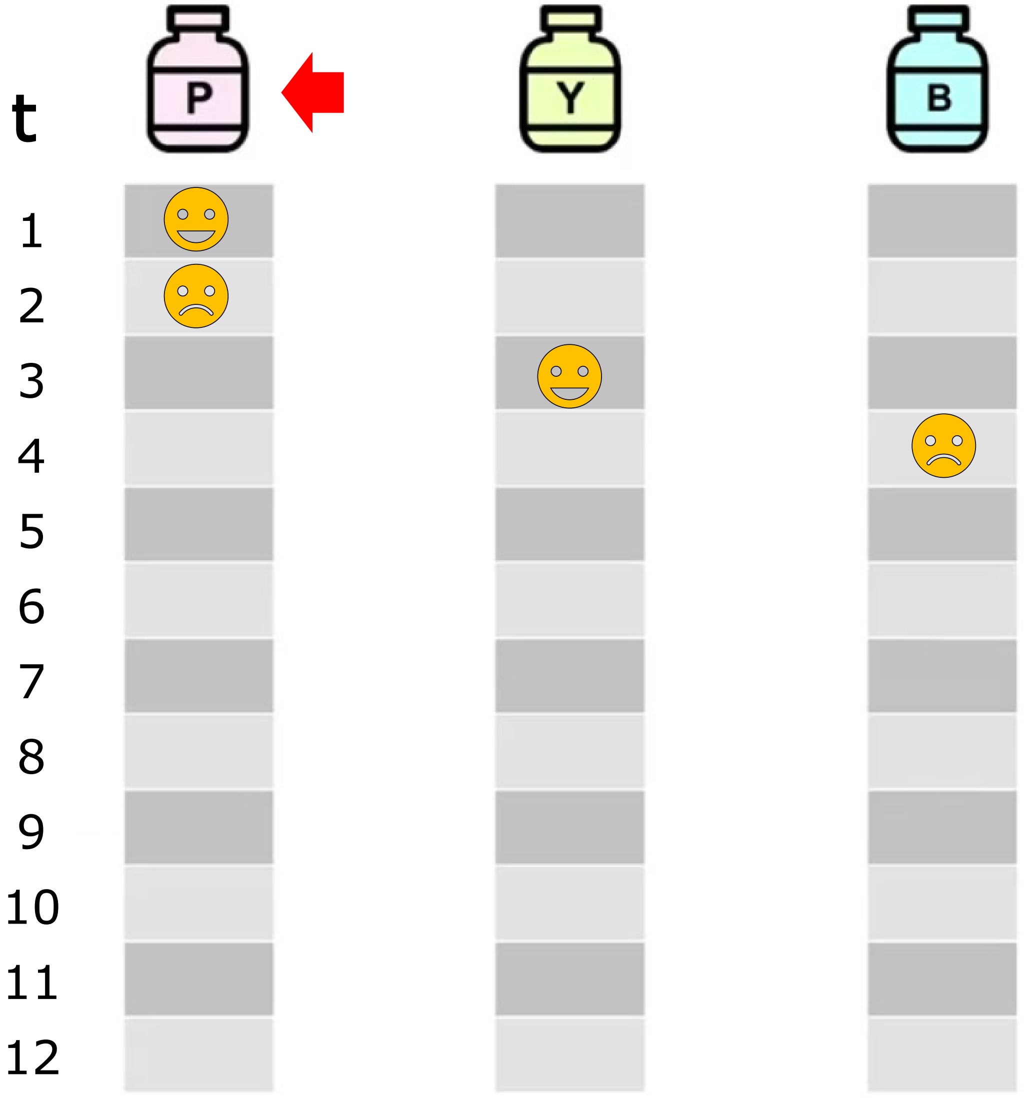
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.5 1.0 0.0

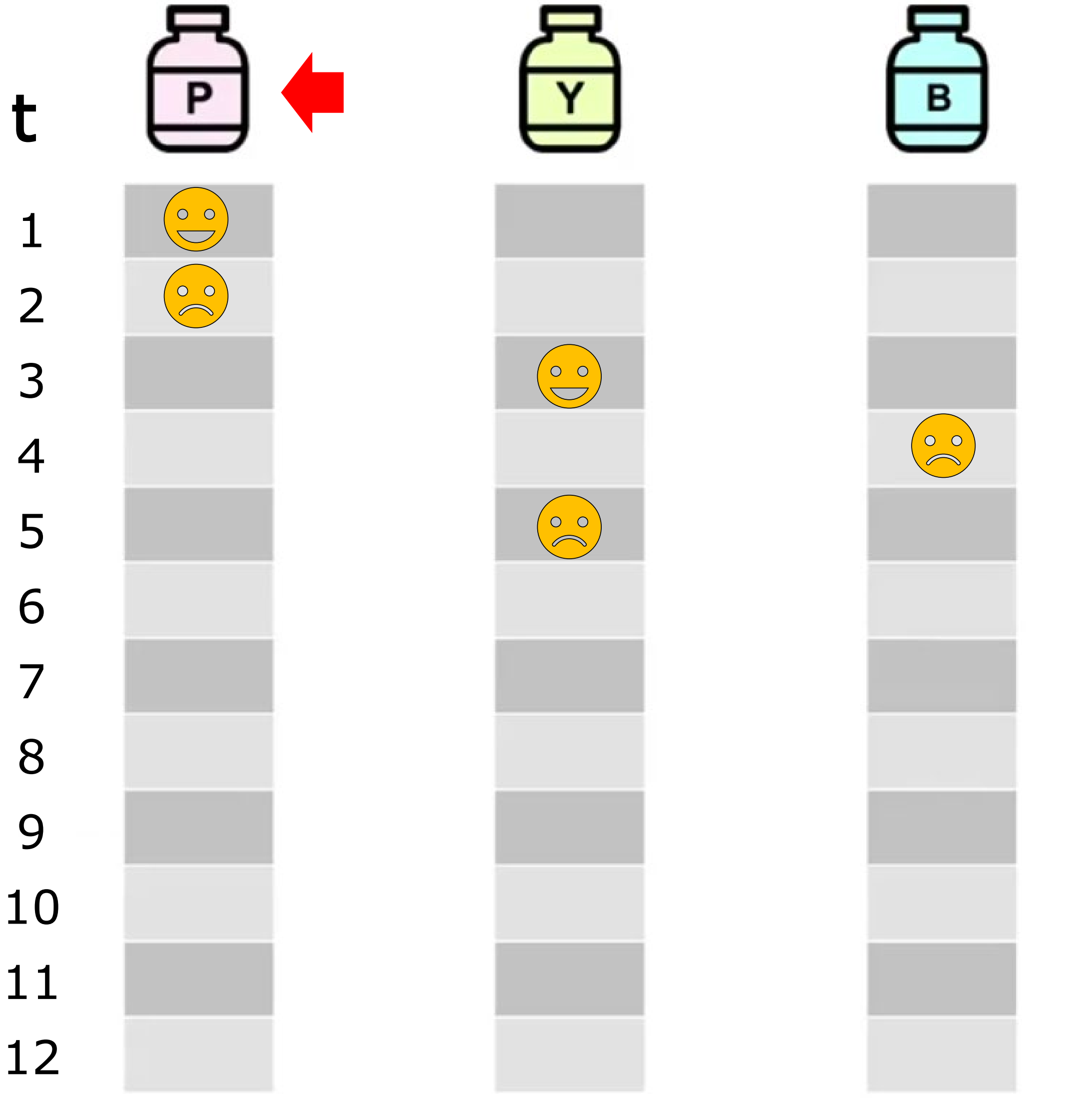
Lets go back to clinical trial example…

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

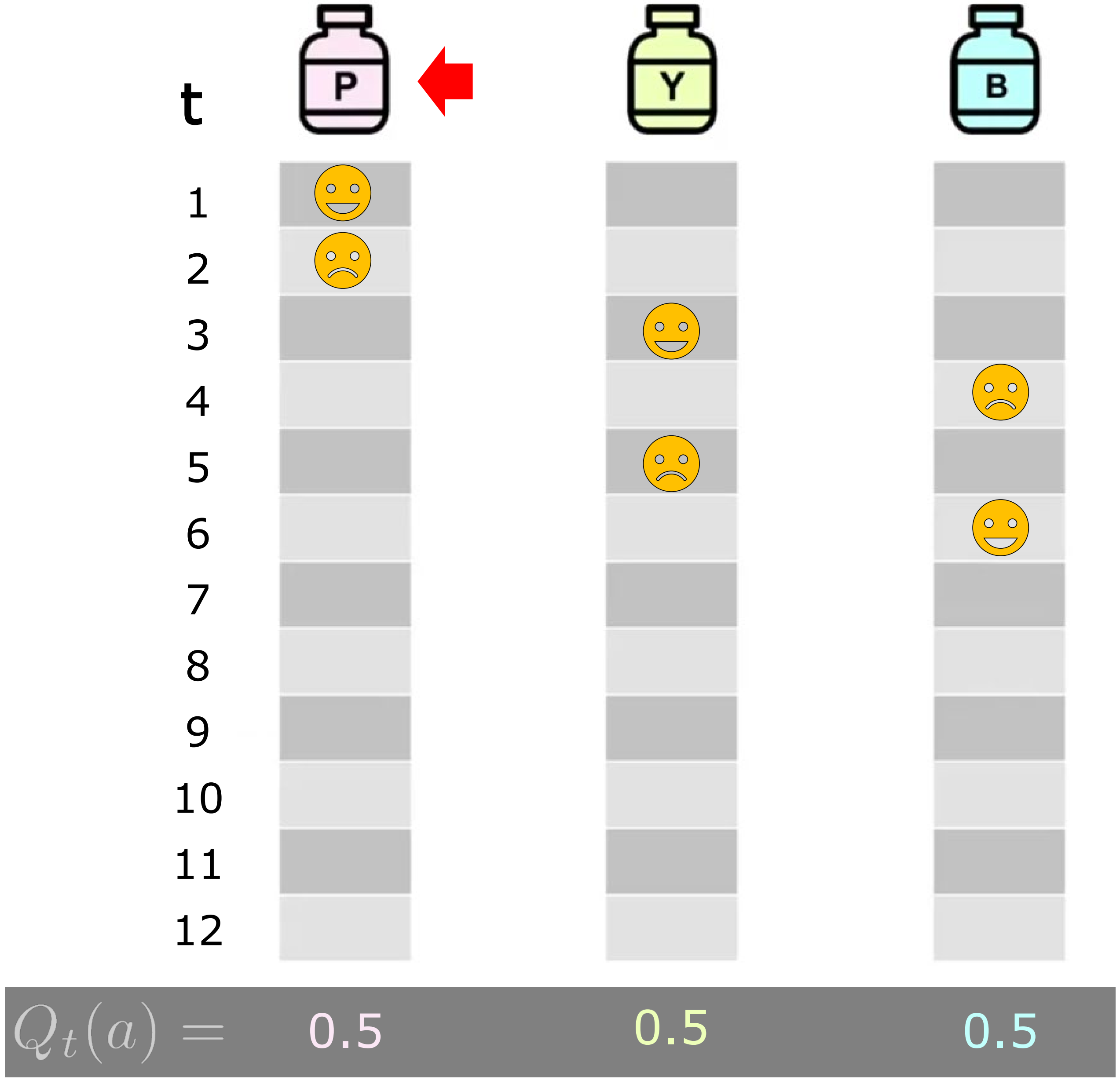
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



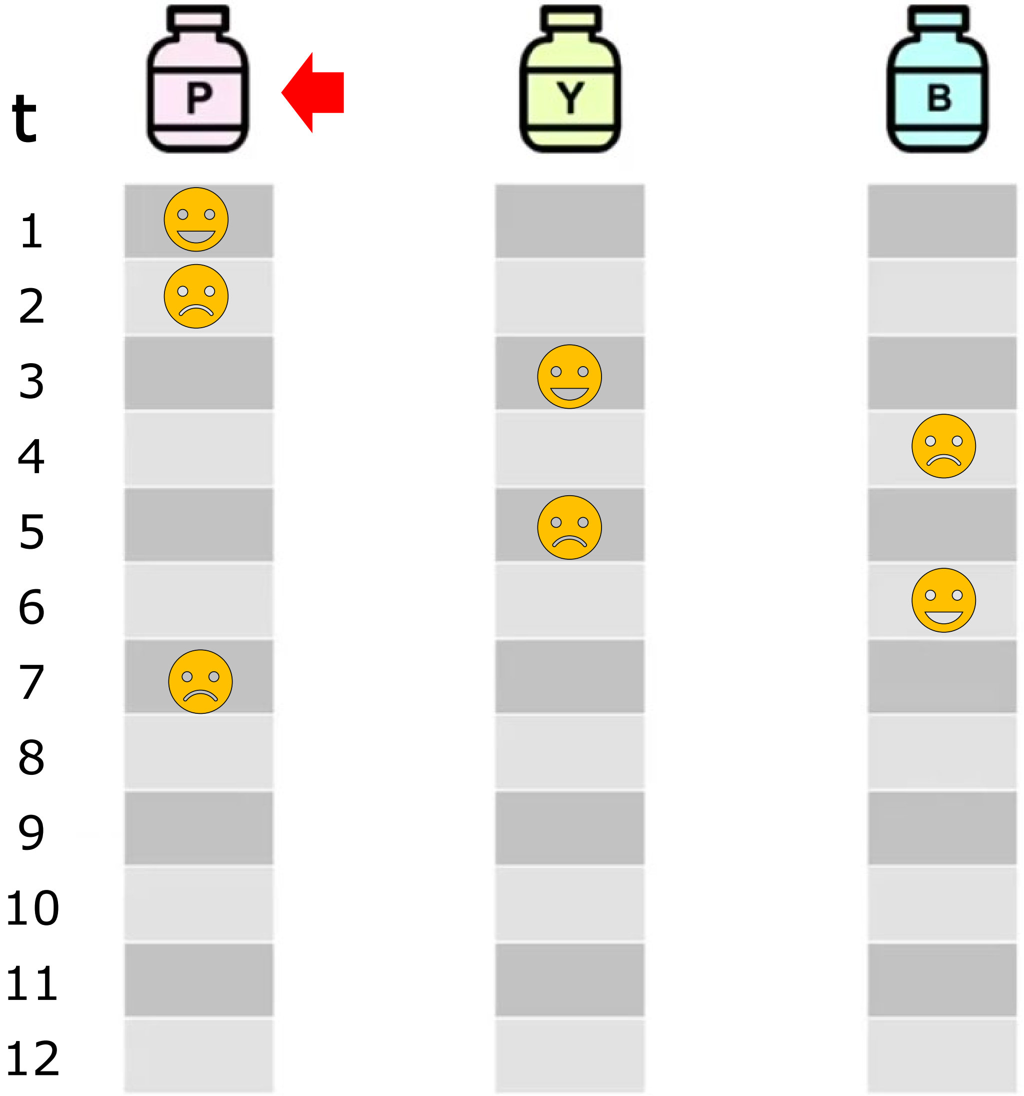
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.33 0.5 0.5

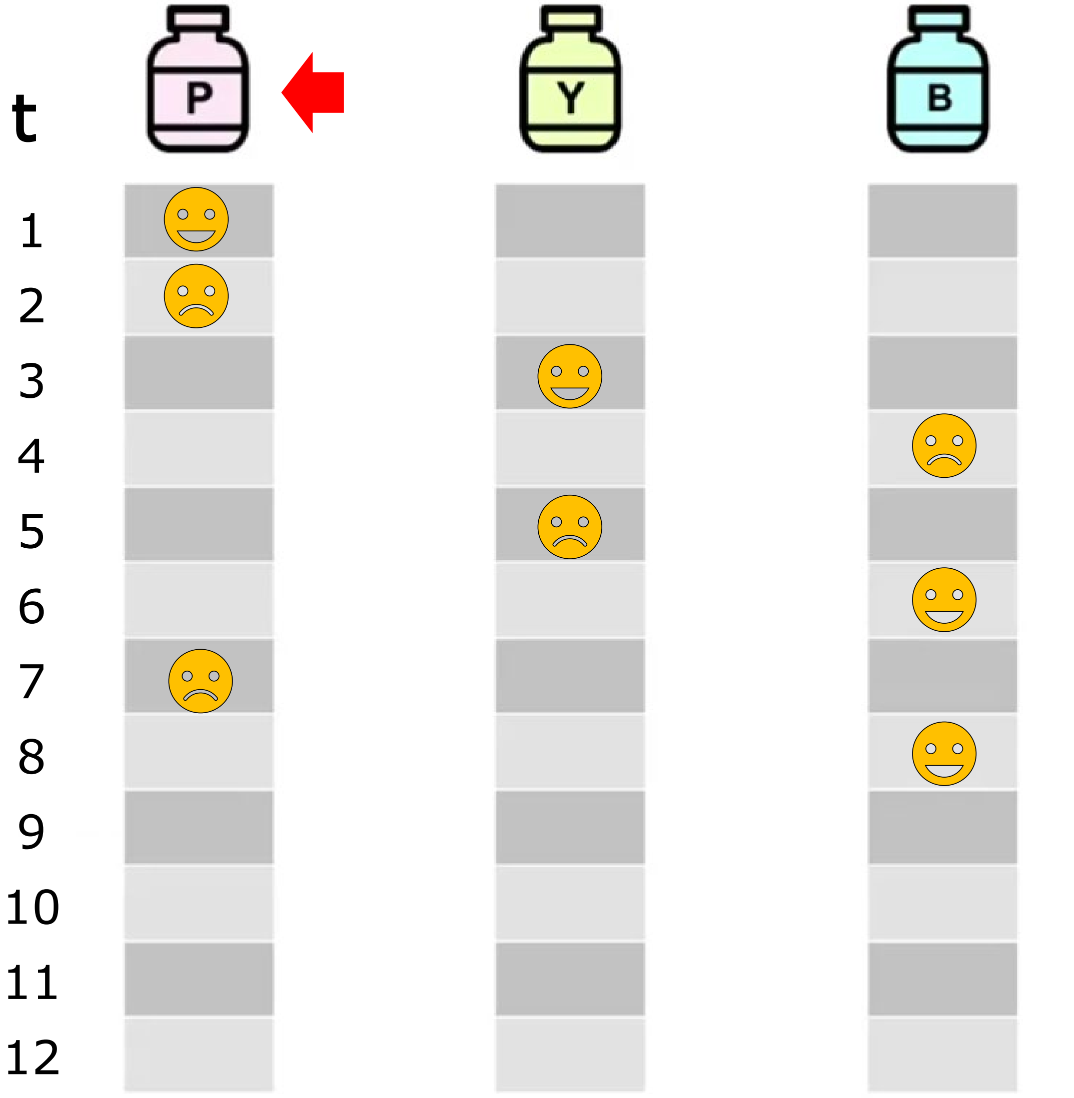
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$$q^*(a)$$

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



$Q_t(a) =$ 0.33 0.5 0.66

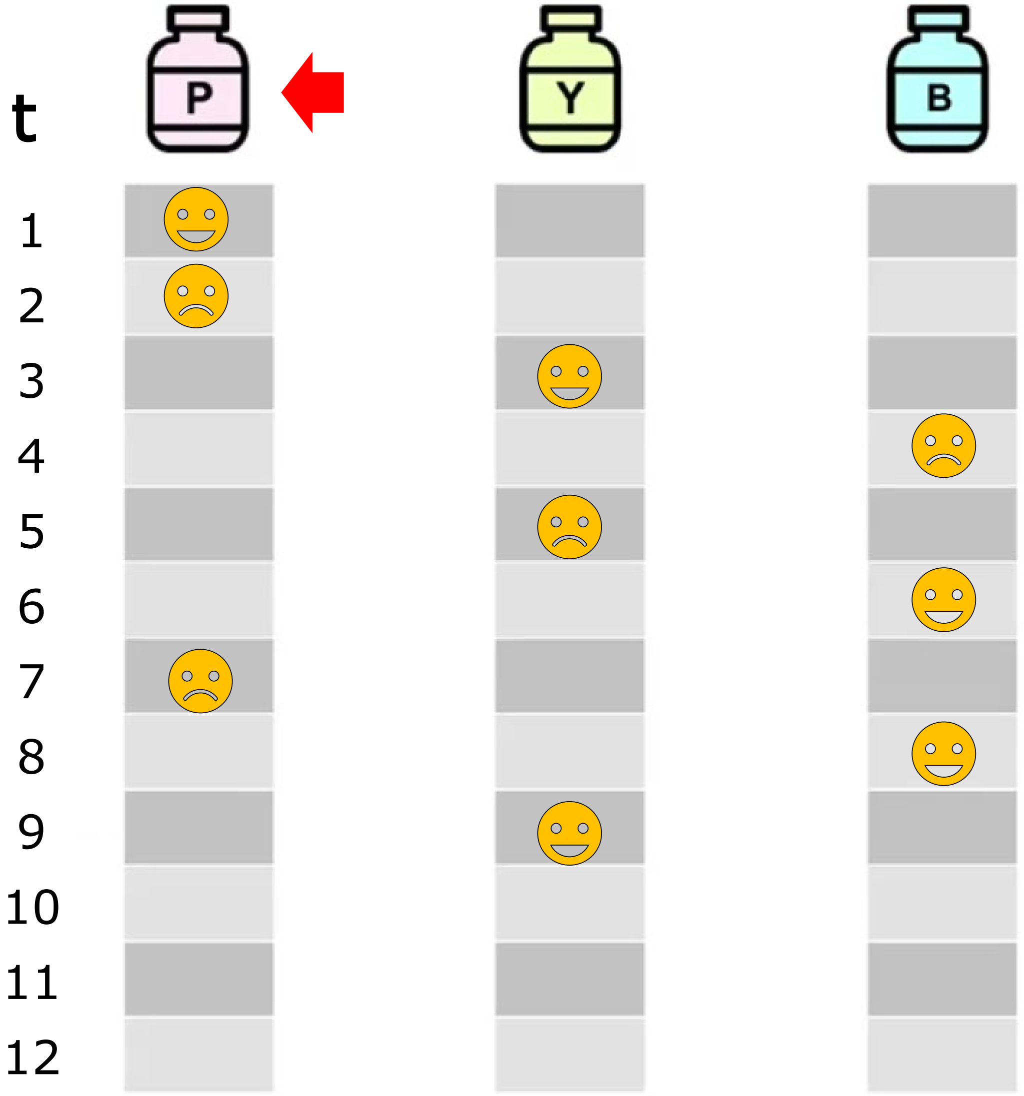
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.33 0.66 0.66

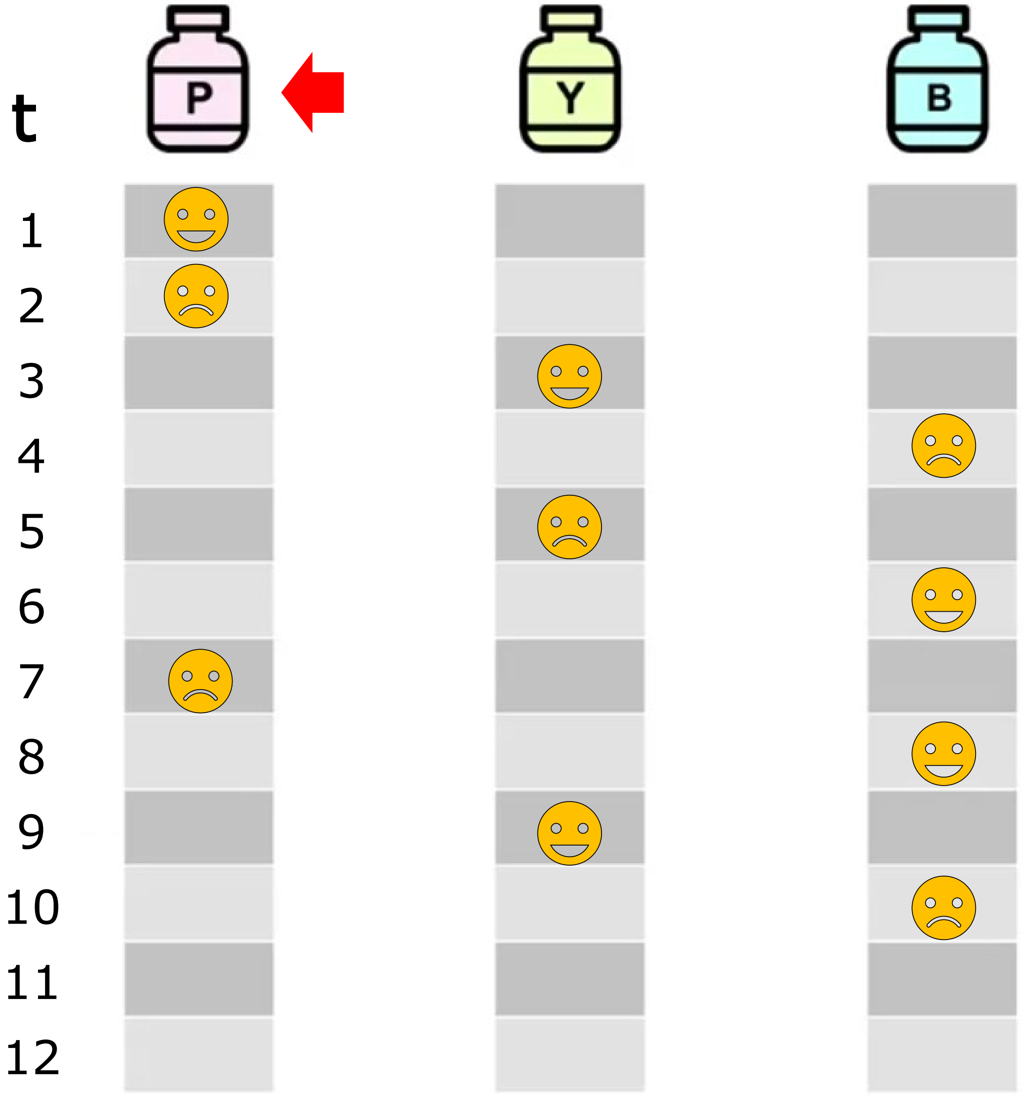
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.33 0.66 0.5

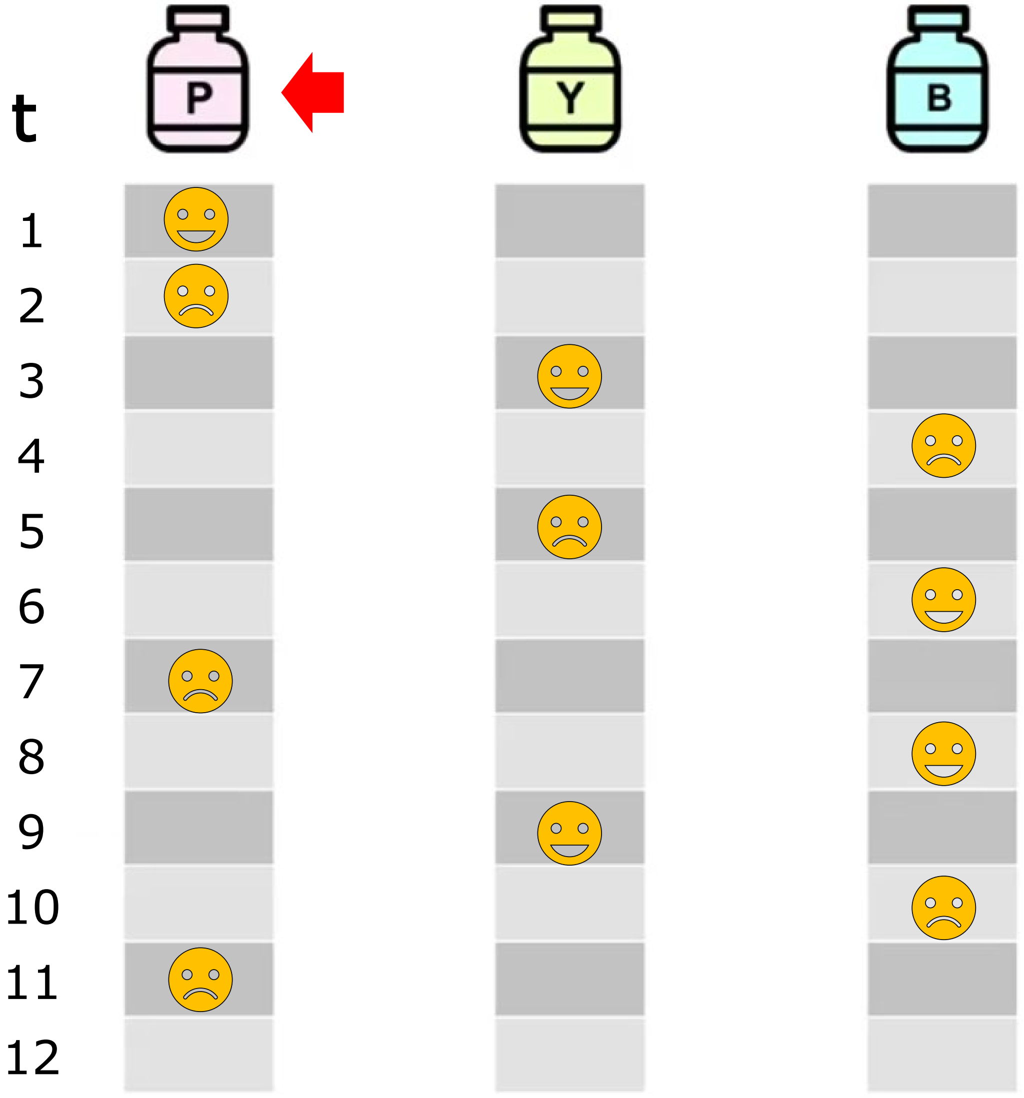
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.25 0.66 0.5

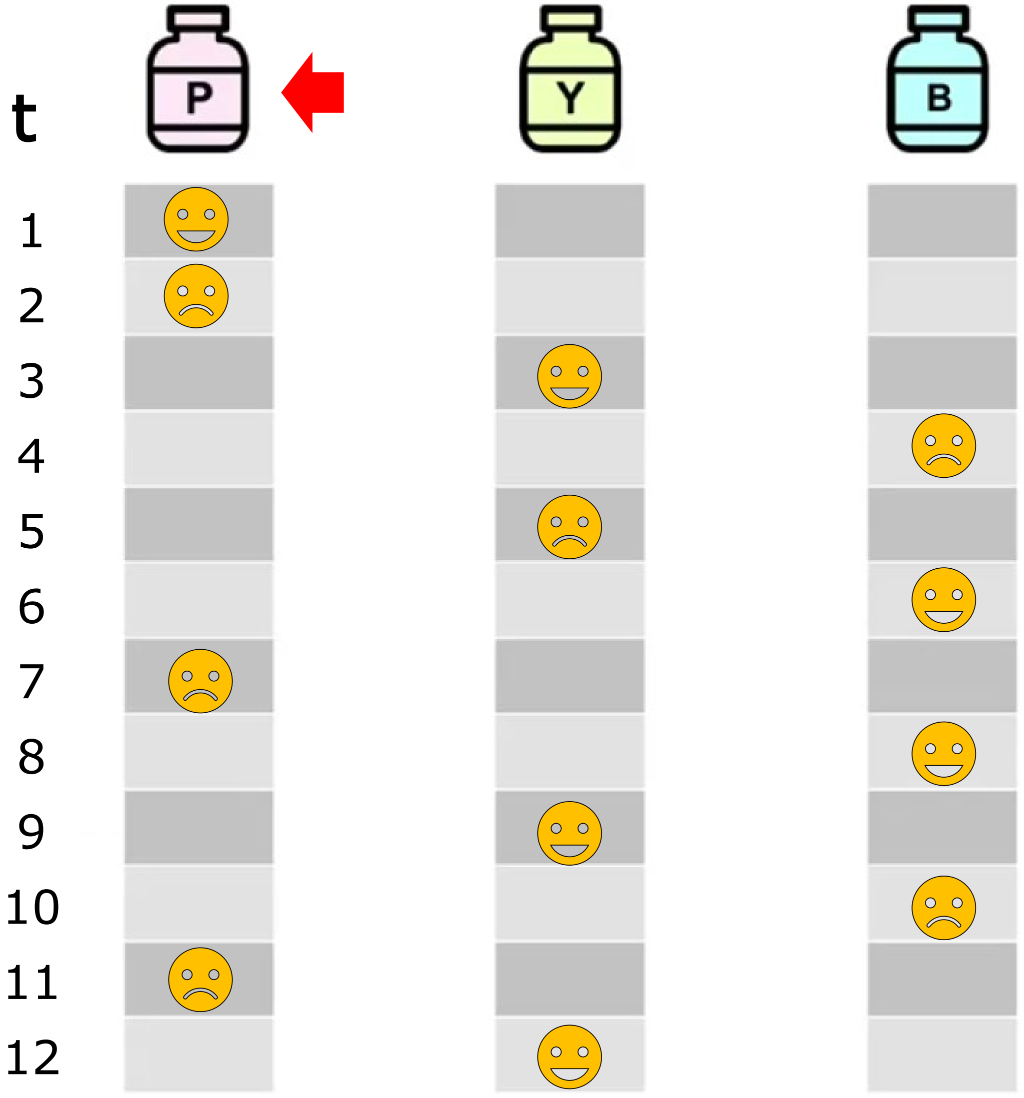
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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



$Q_t(a) =$ 0.25 0.75 0.5

Incremental update and non-stationary problems

Incremental update of action-values

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

- Let’s consider the update of a single action:

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\left(R_n + (n-1)\boxed{\frac{1}{n-1}\sum_{i=1}^{n-1} R_i}\right)$$

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

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

Incremental update of action-values

new estimate

step size

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

old estimate

target – old estimate

Non-stationary bandit problems

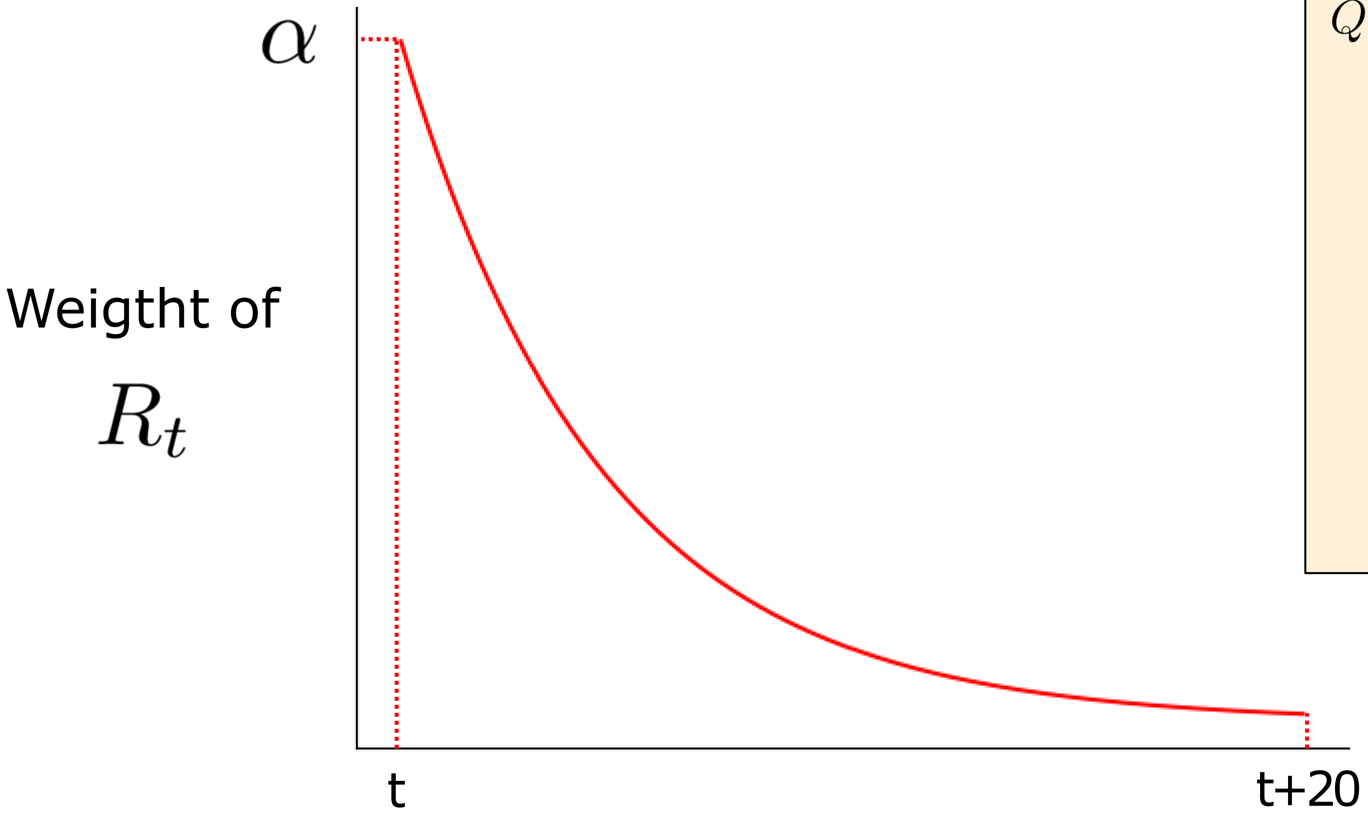
0.25 0.75 0.80

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

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

Non-stationary bandit problems

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



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

Initial action-value

Epsilon-Greedy Action Selection

Action Selection

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

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

non-greedy actions

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

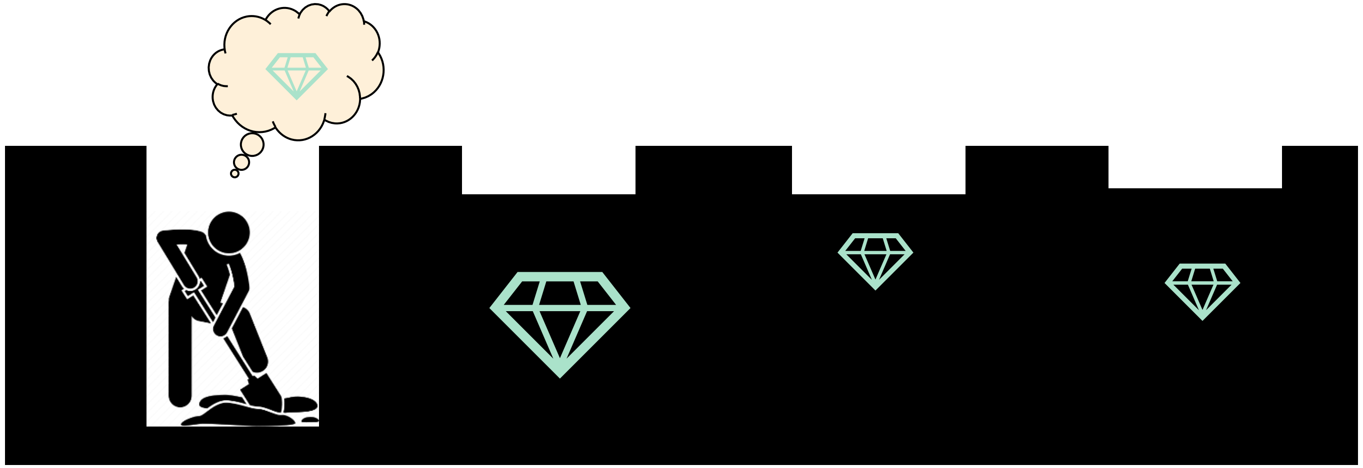
greedy action

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

What would happen if we always select the greedy action?

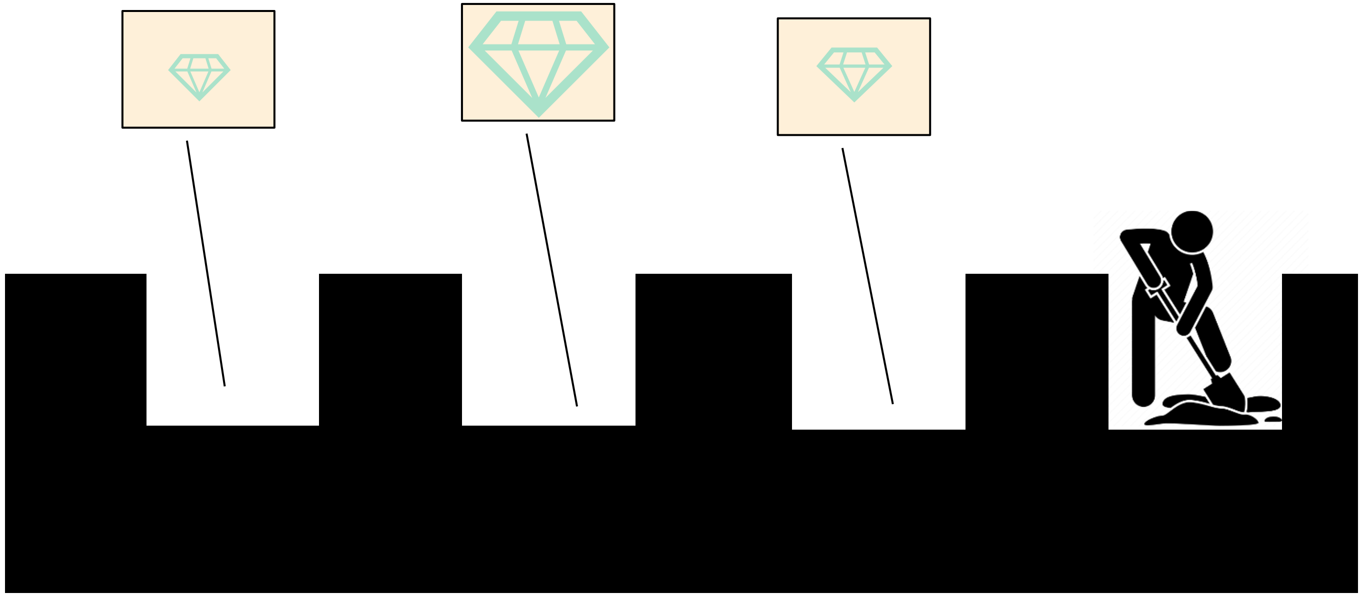
Exploration vs Exploitation

Exploration vs Exploitation



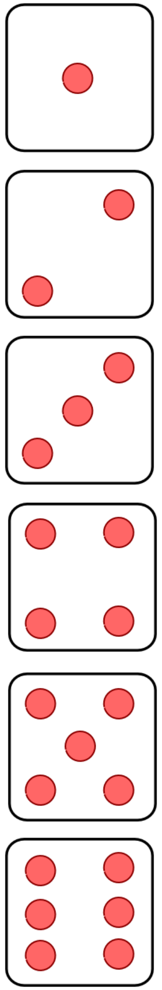
- Exploitation: agent exploits knowledge for **short-term** benefit

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



explore

greedy

greedy

greedy

greedy

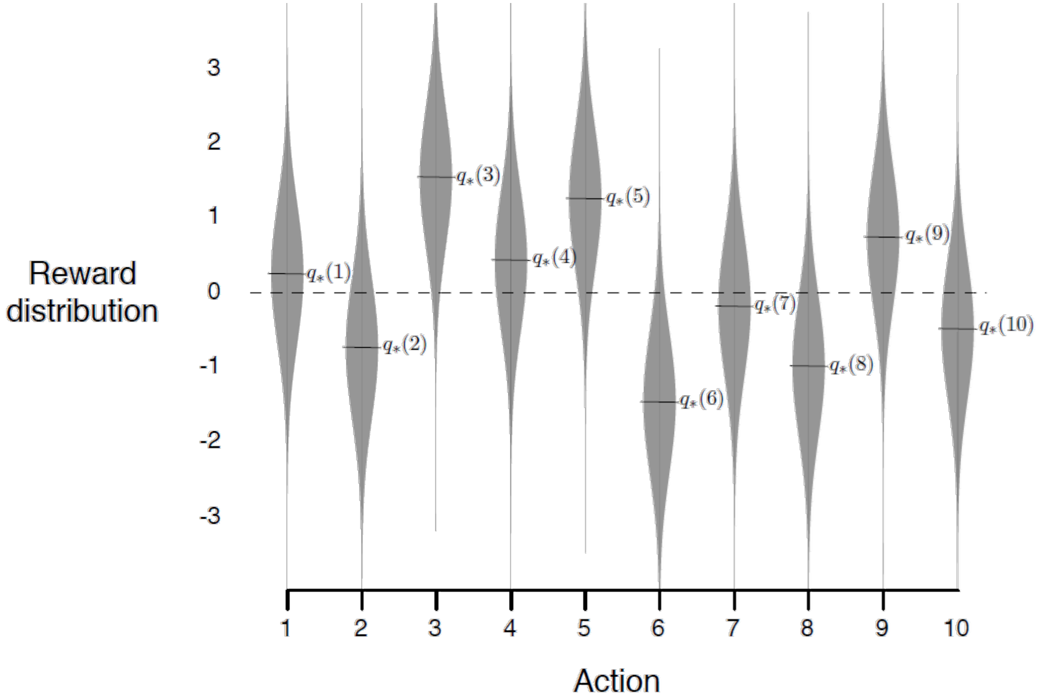
greedy

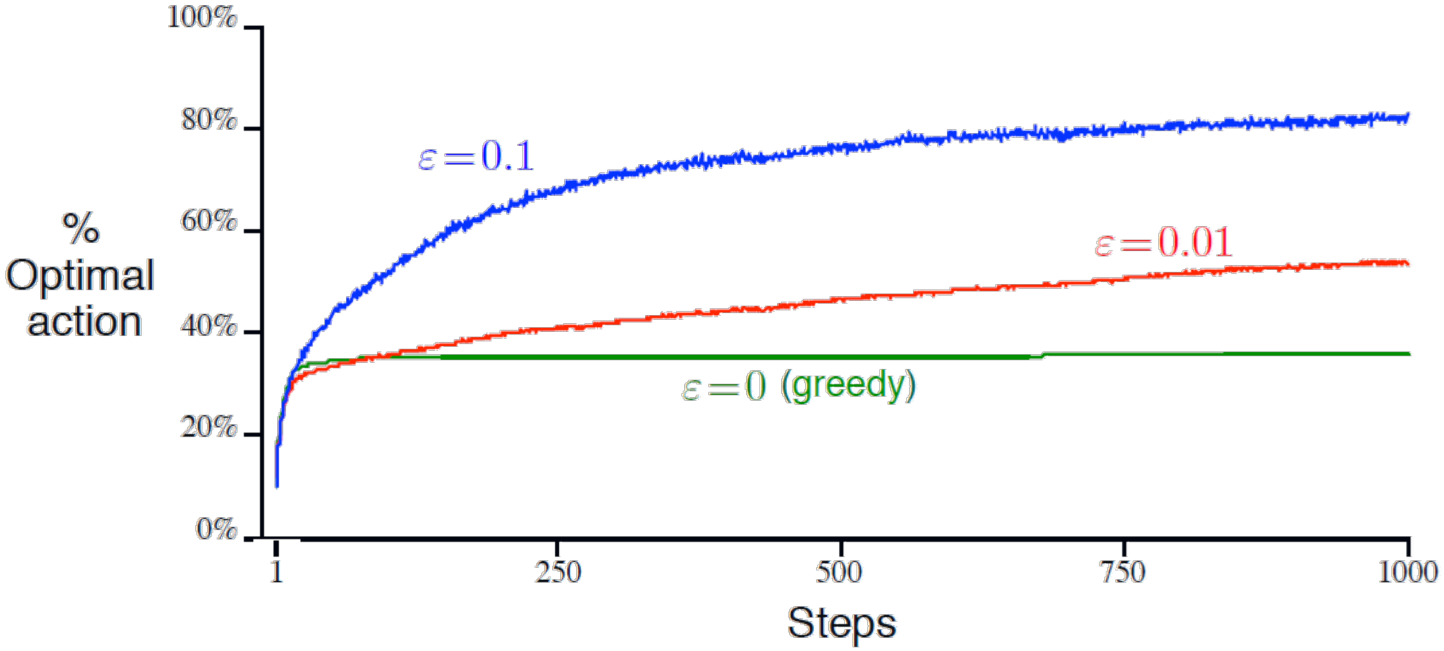
$$A_t = \begin{cases} argmax_a \, Q_t(a) & \text{with probability } 1-\epsilon \\ Uniform(\{a_1, ..., a_k\}) & \text{with probability } \epsilon \end{cases}$$

Epsilon-Greedy Action Selection

$$A_t = \begin{cases} argmax_a \; Q_t(a) & \text{with probability } 1-\epsilon \\ Uniform(\{a_1, ..., a_k\}) & \text{with probability } \epsilon \end{cases}$$

10-armed testbed





Average over 2000 runs

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	P	Y	B
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	P	Y	B
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…

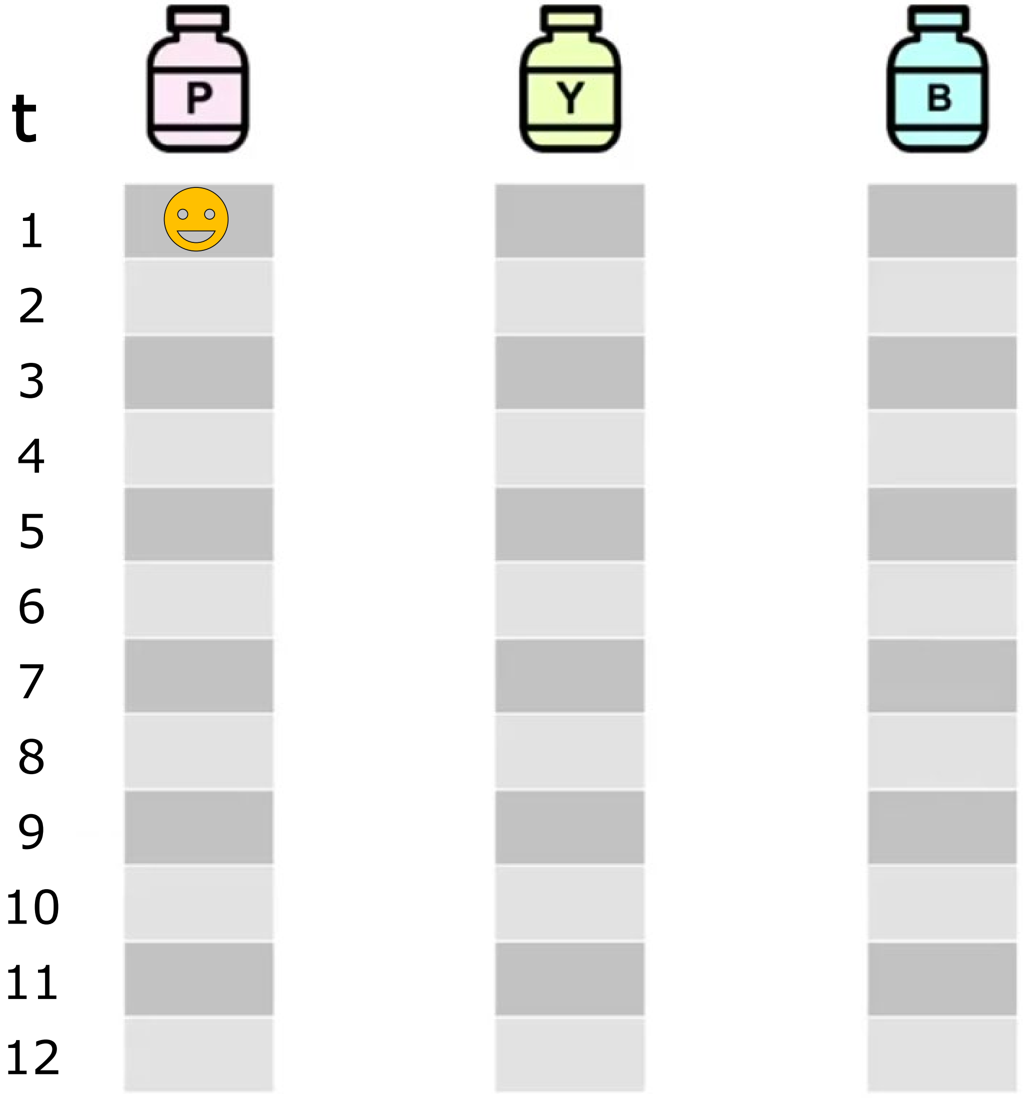
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 1.5 2.0 2.0

Lets go back to clinical trial example…

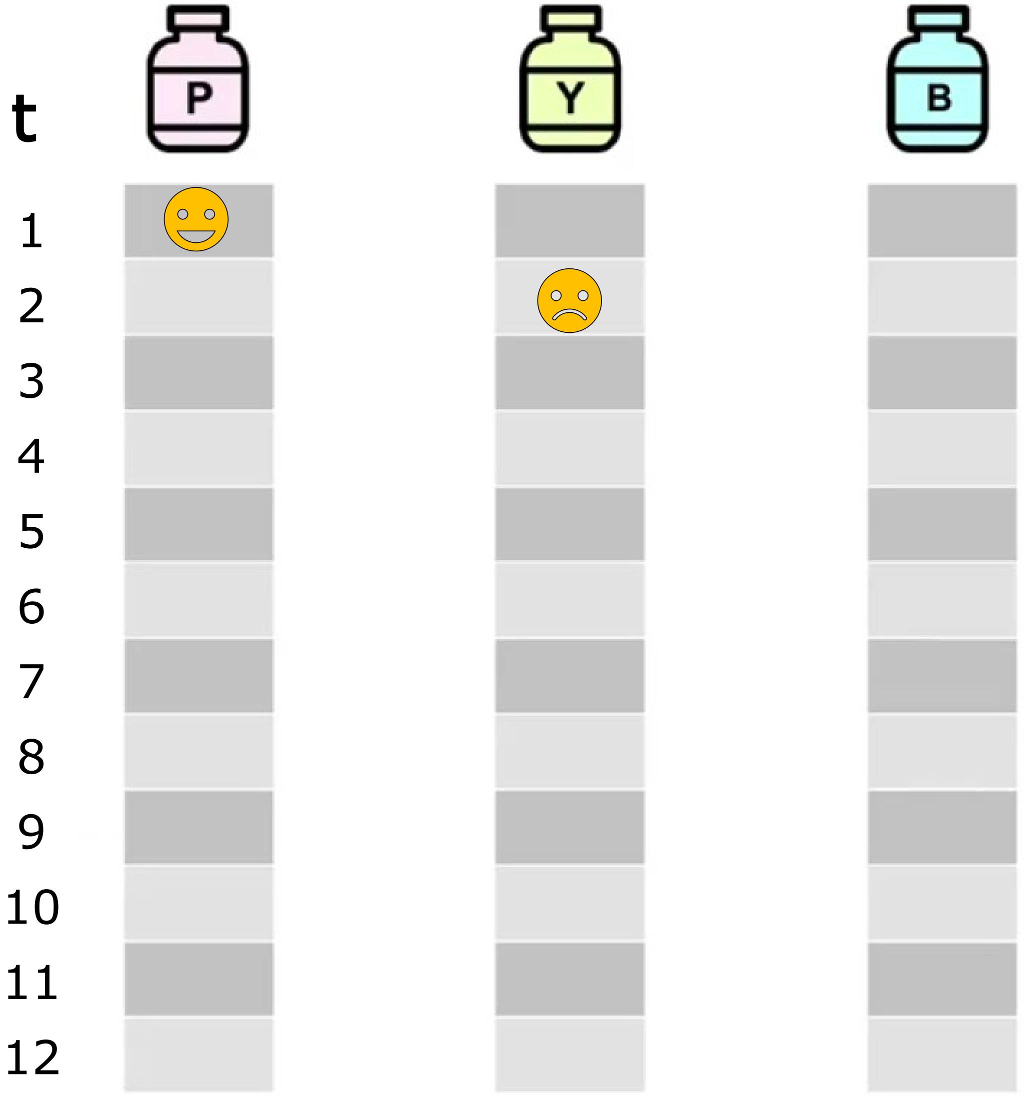
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 1.5 1.0 2.0

Lets go back to clinical trial example…

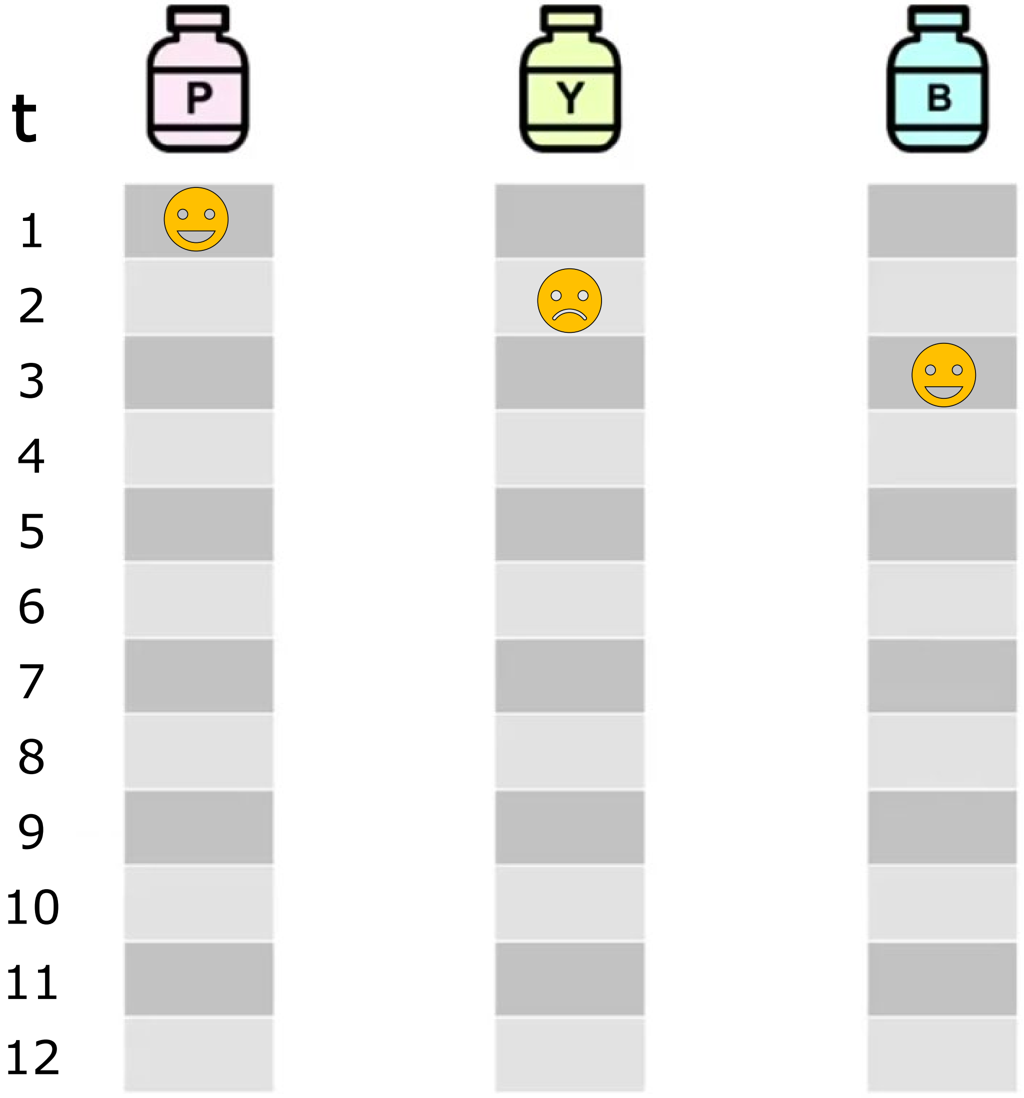
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 1.5 1.0 1.5

Lets go back to clinical trial example…

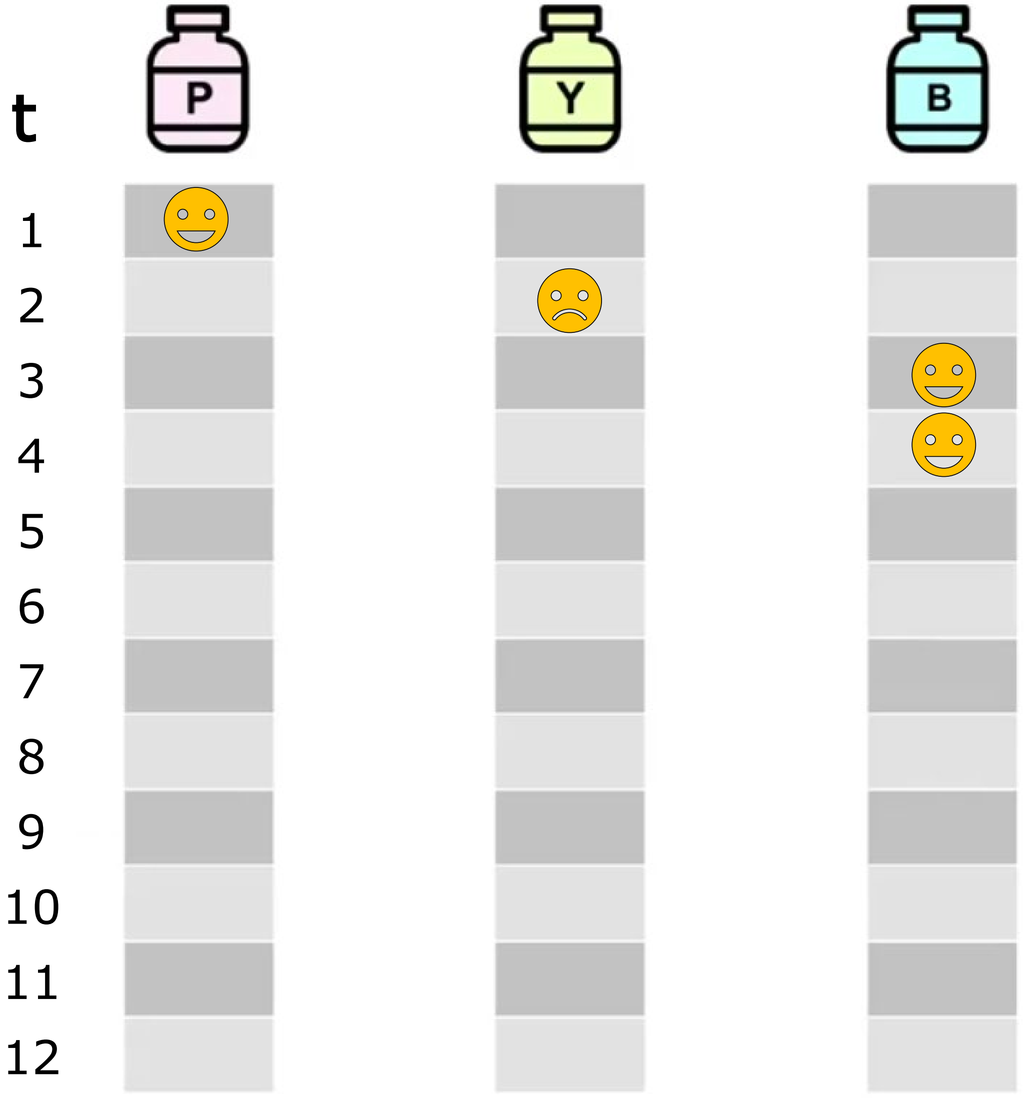
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 1.5 1.0 1.25

Lets go back to clinical trial example…

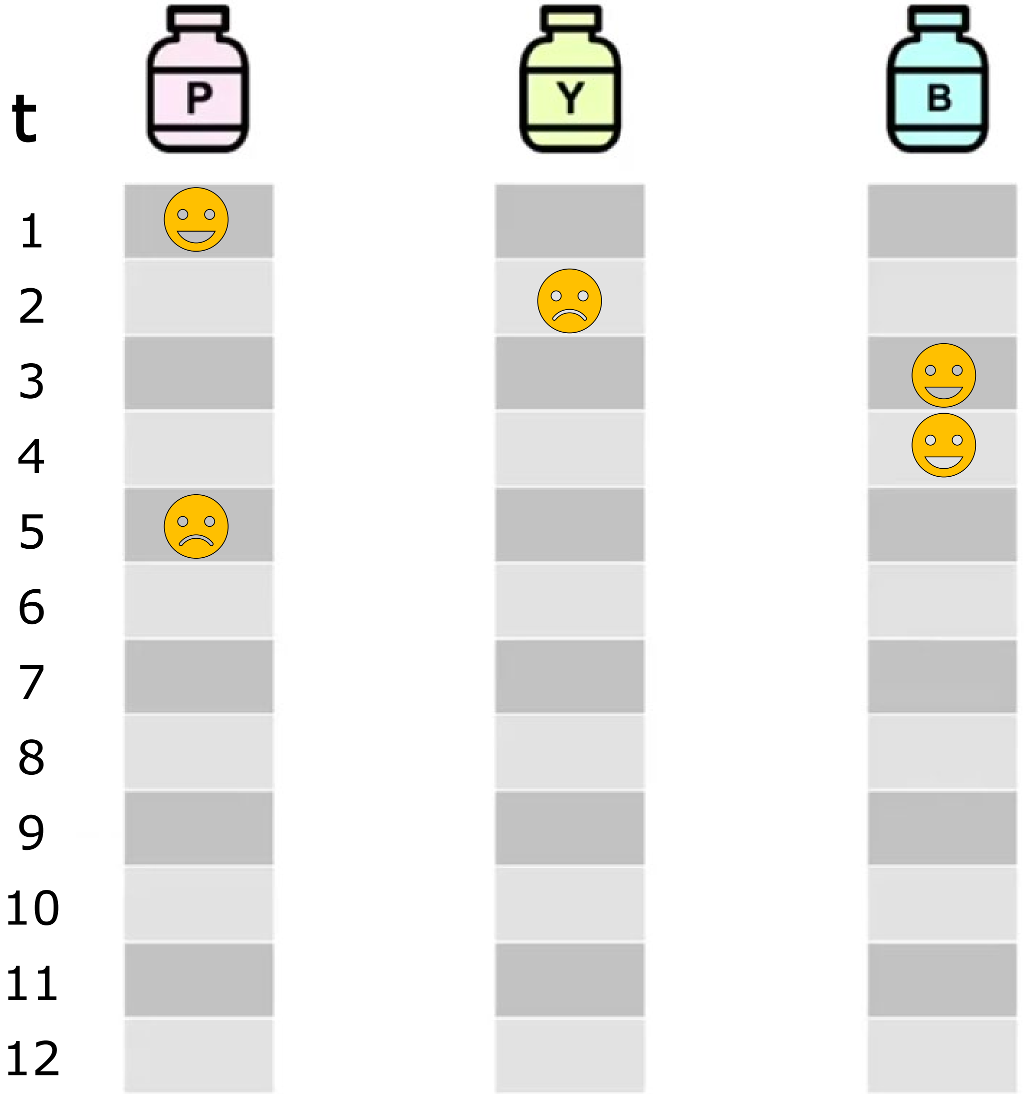
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 0.75 1.0 1.25

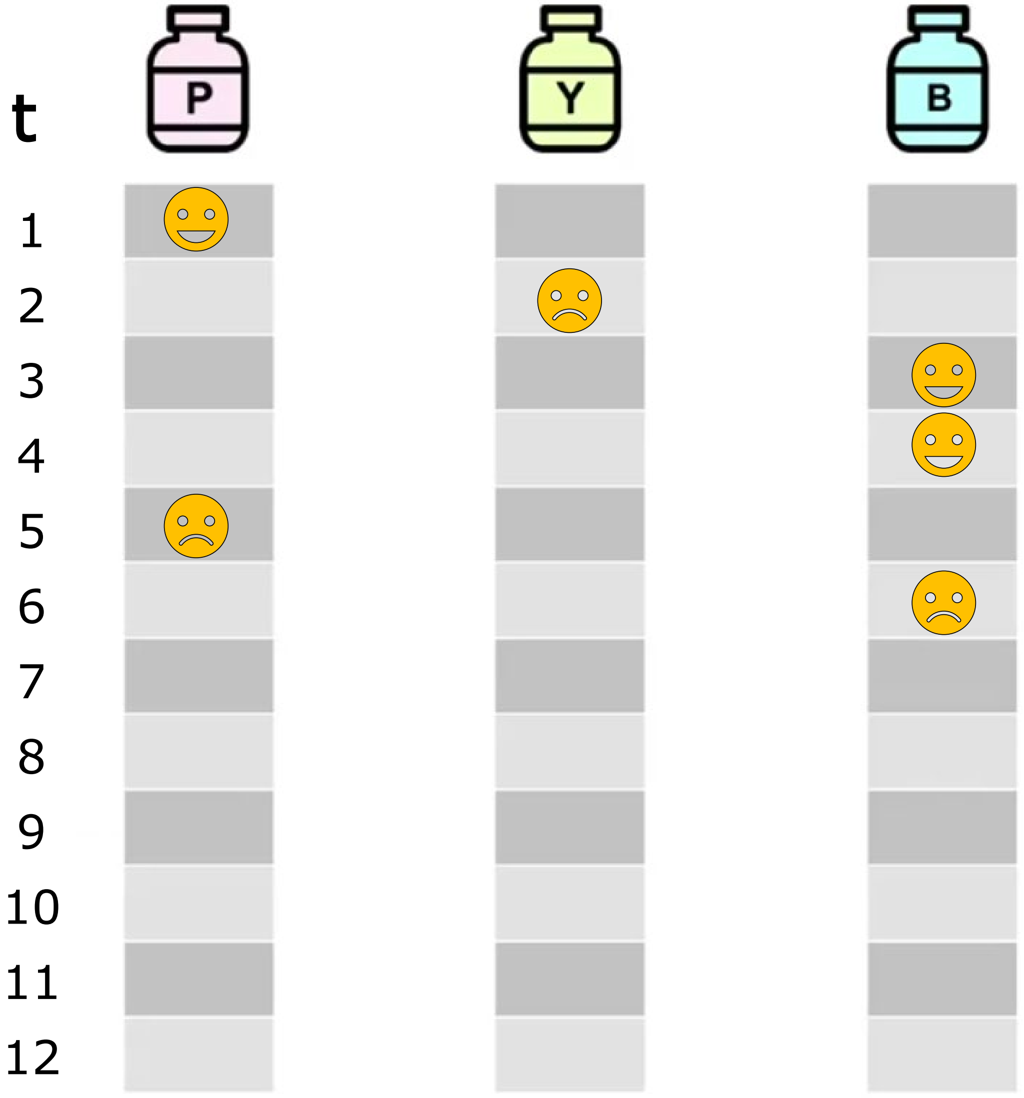
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$
$$\alpha = 0.5$$



$Q_t(a) =$ 0.75 1.0 0.625

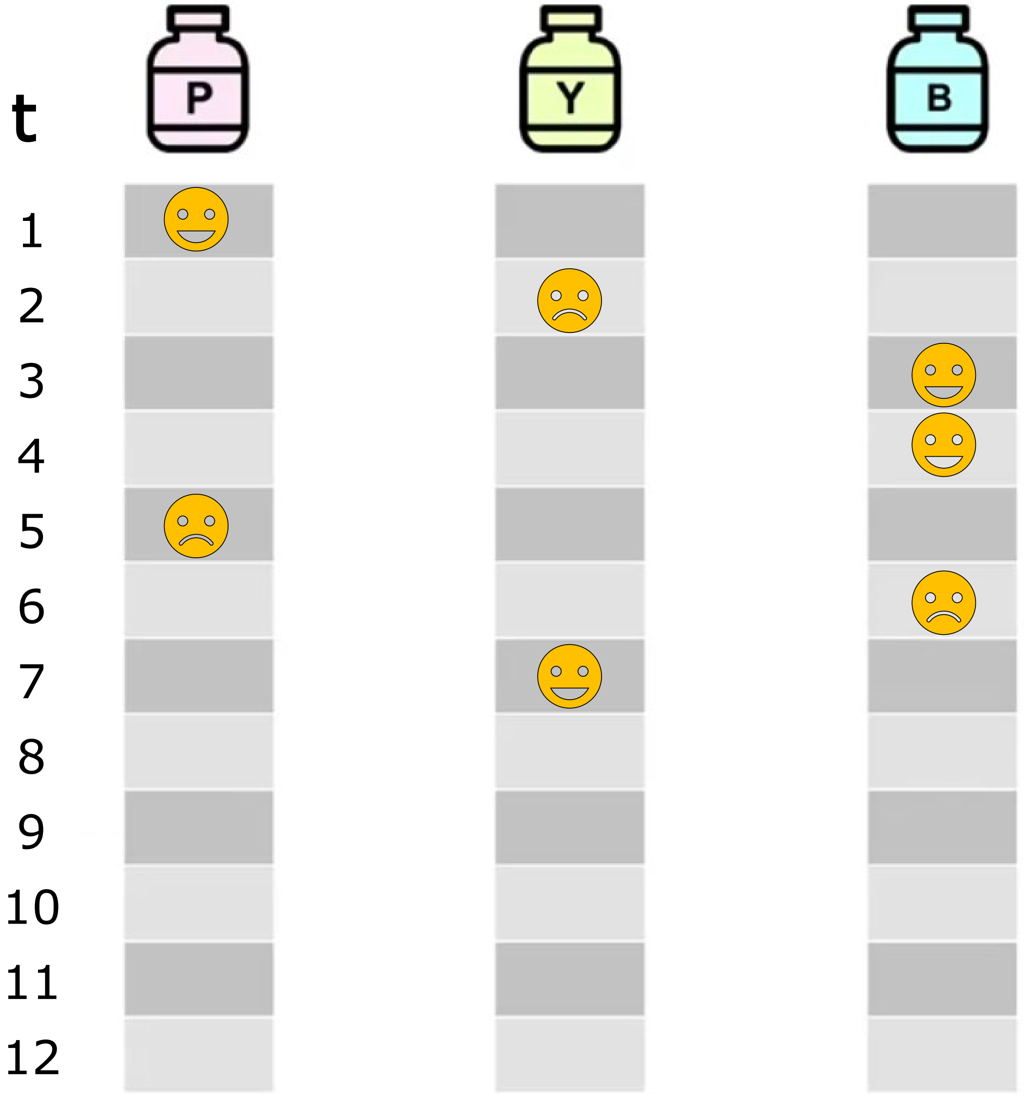
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$
$$\alpha = 0.5$$



$Q_t(a) =$ 0.75 1.0 0.625

Lets go back to clinical trial example…

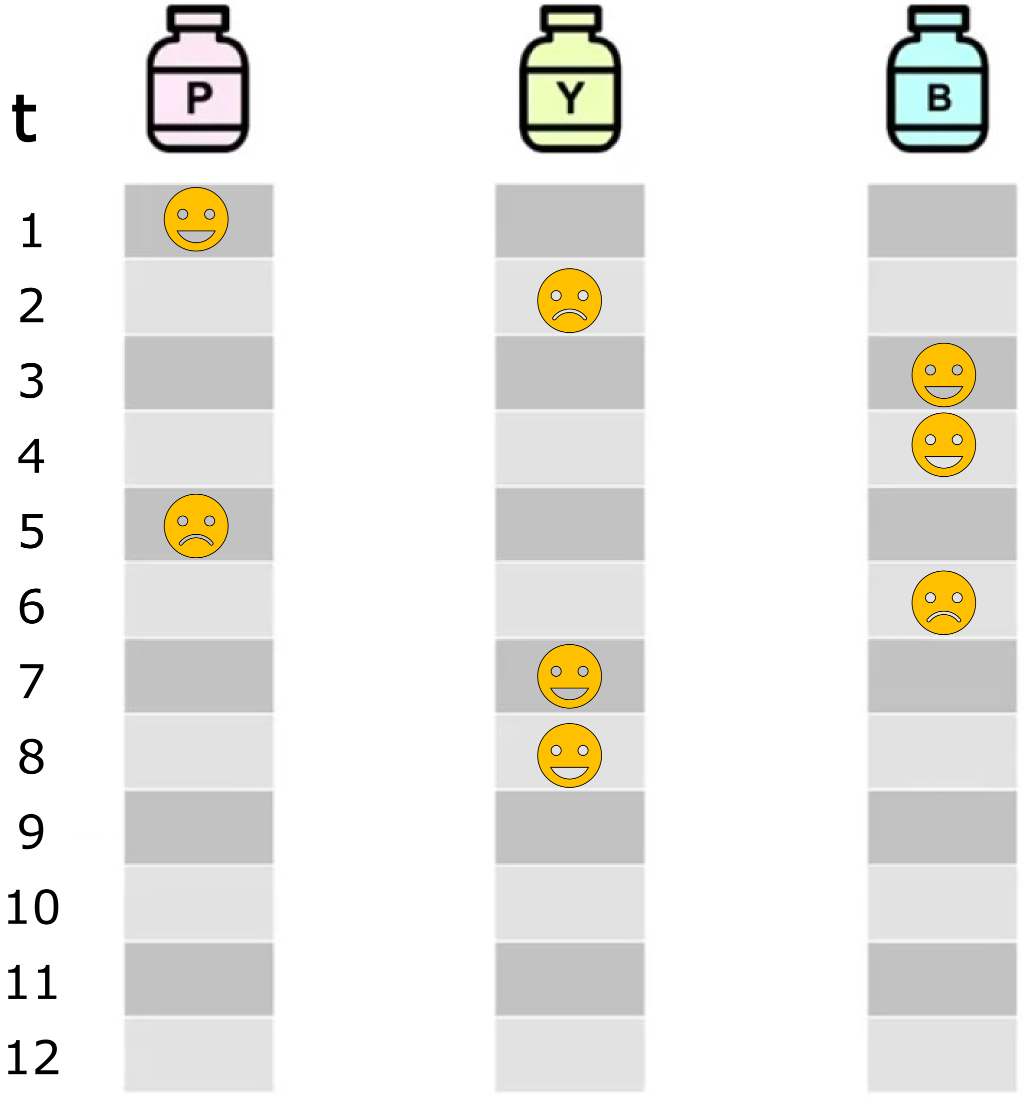
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 0.75 1.0 0.625

Lets go back to clinical trial example…

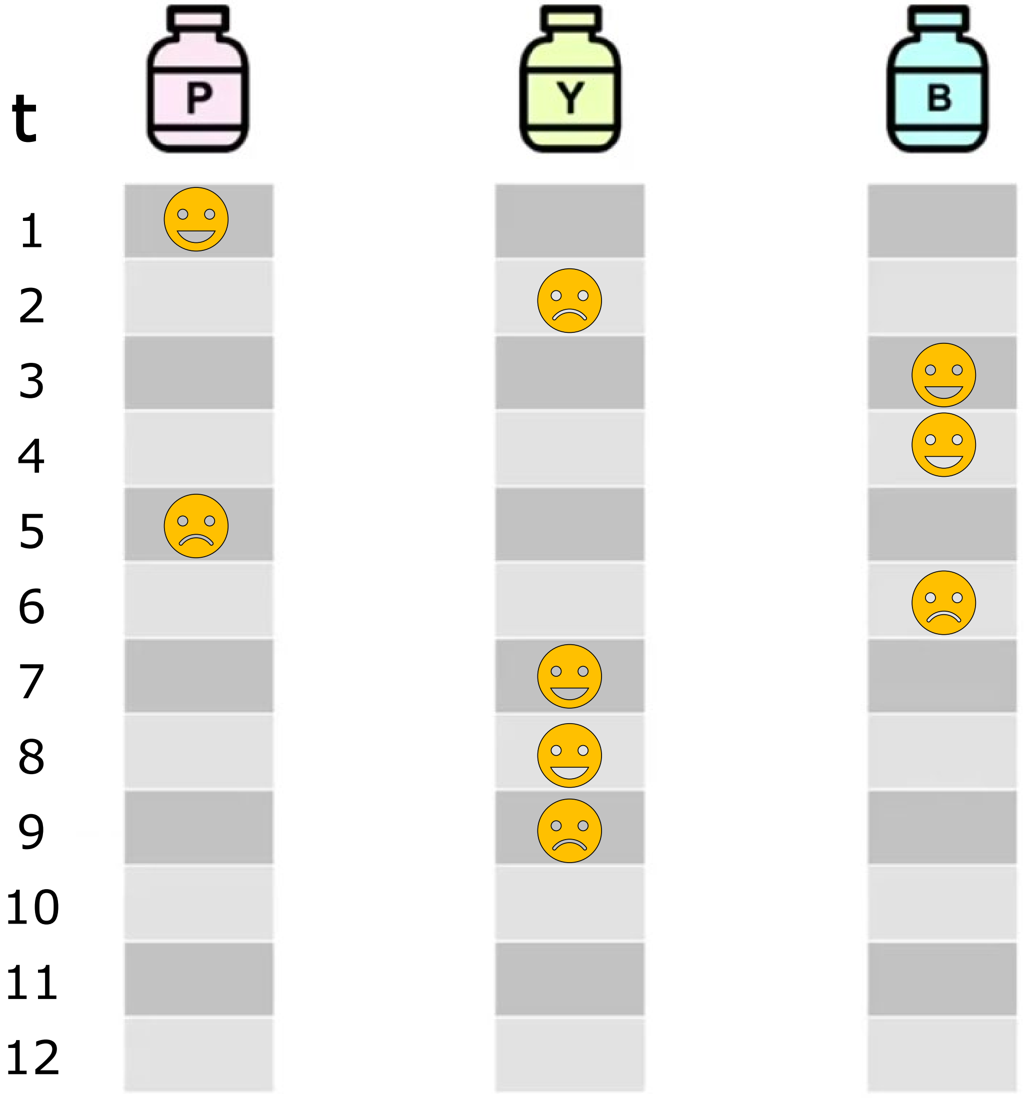
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 0.75 0.5 0.625

Lets go back to clinical trial example…

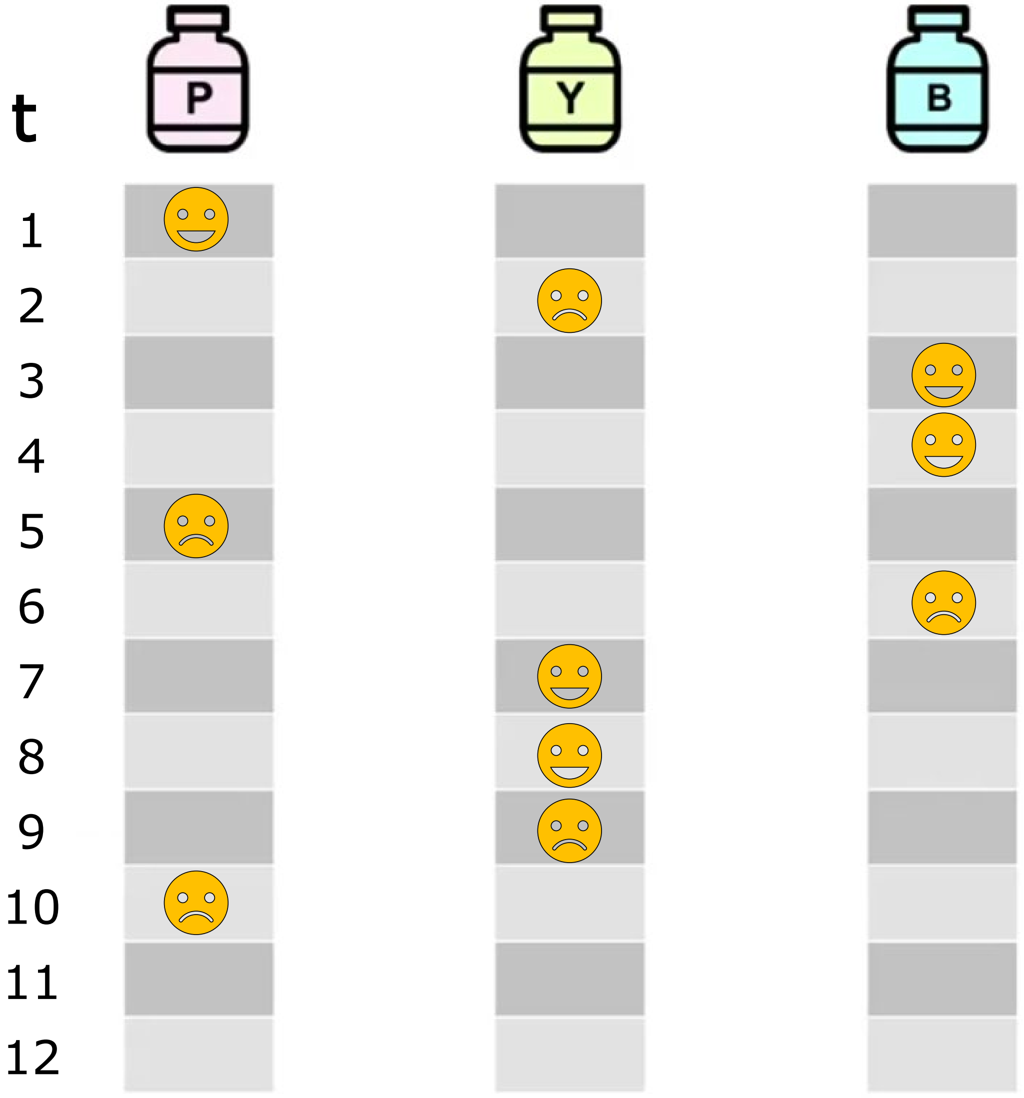
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 0.375 0.5 0.625

Lets go back to clinical trial example…

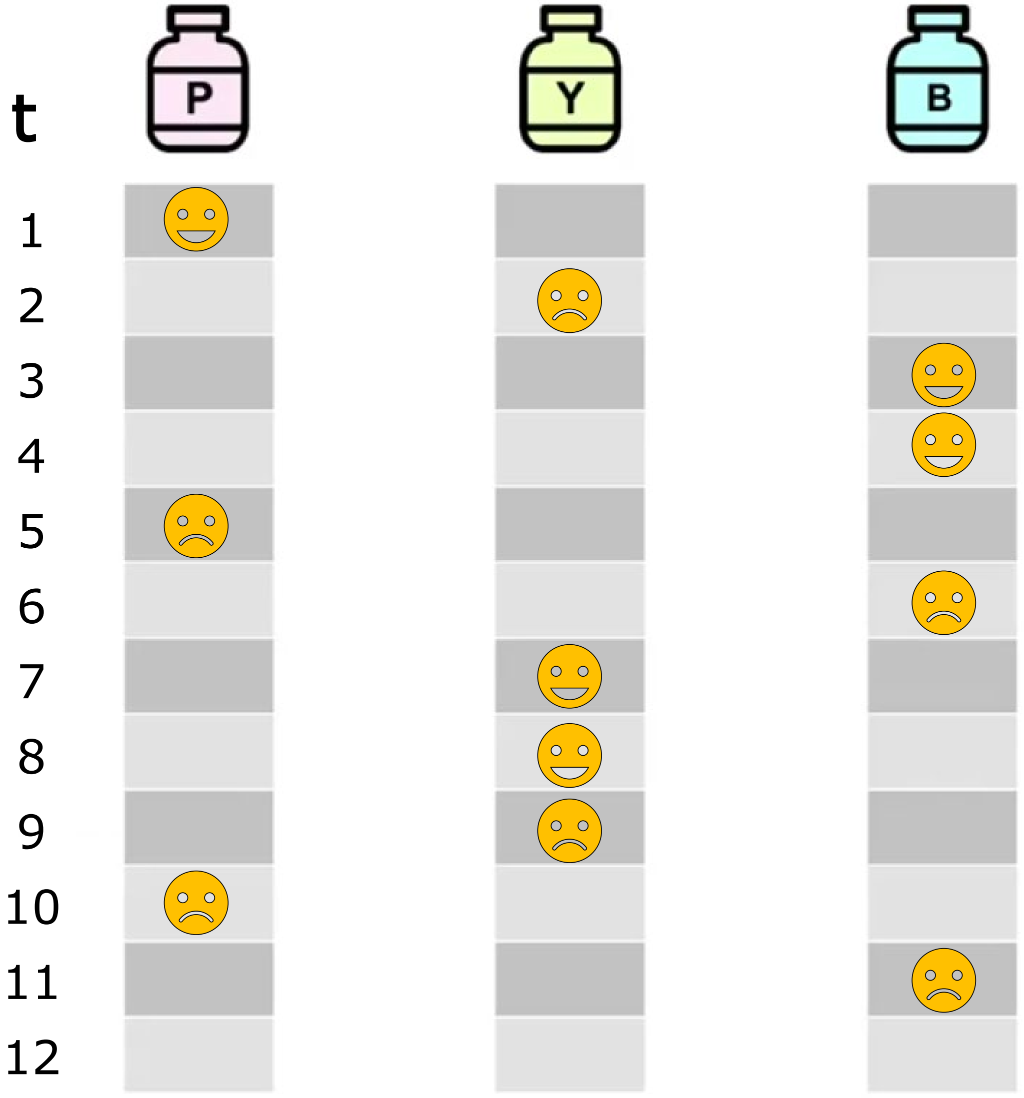
Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

$q^*(a)$

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

$$\alpha = 0.5$$



$Q_t(a) =$ 0.375 0.5 0.3125

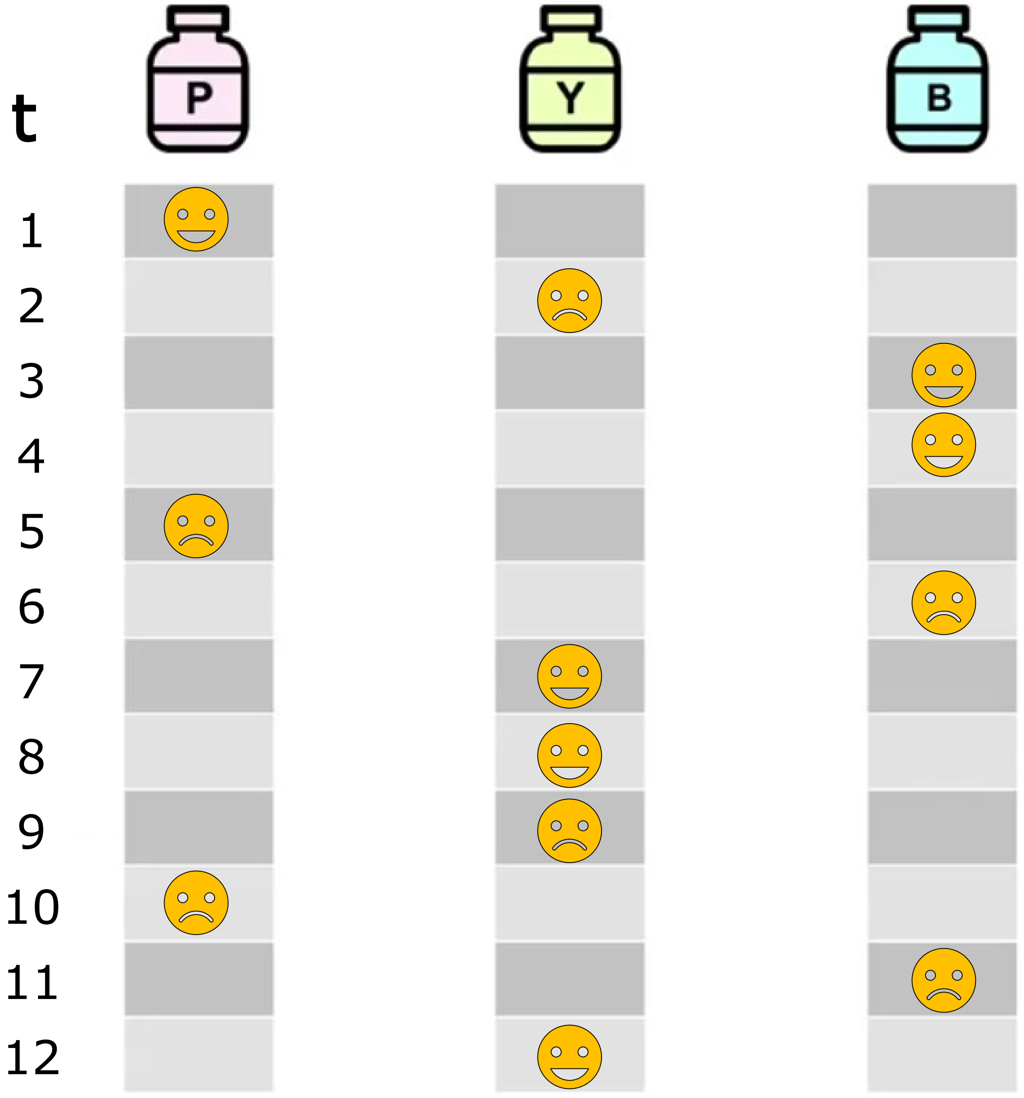
Lets go back to clinical trial example…

Reward 1 if the treatment works, 0 otherwise

0.25 0.75 0.50

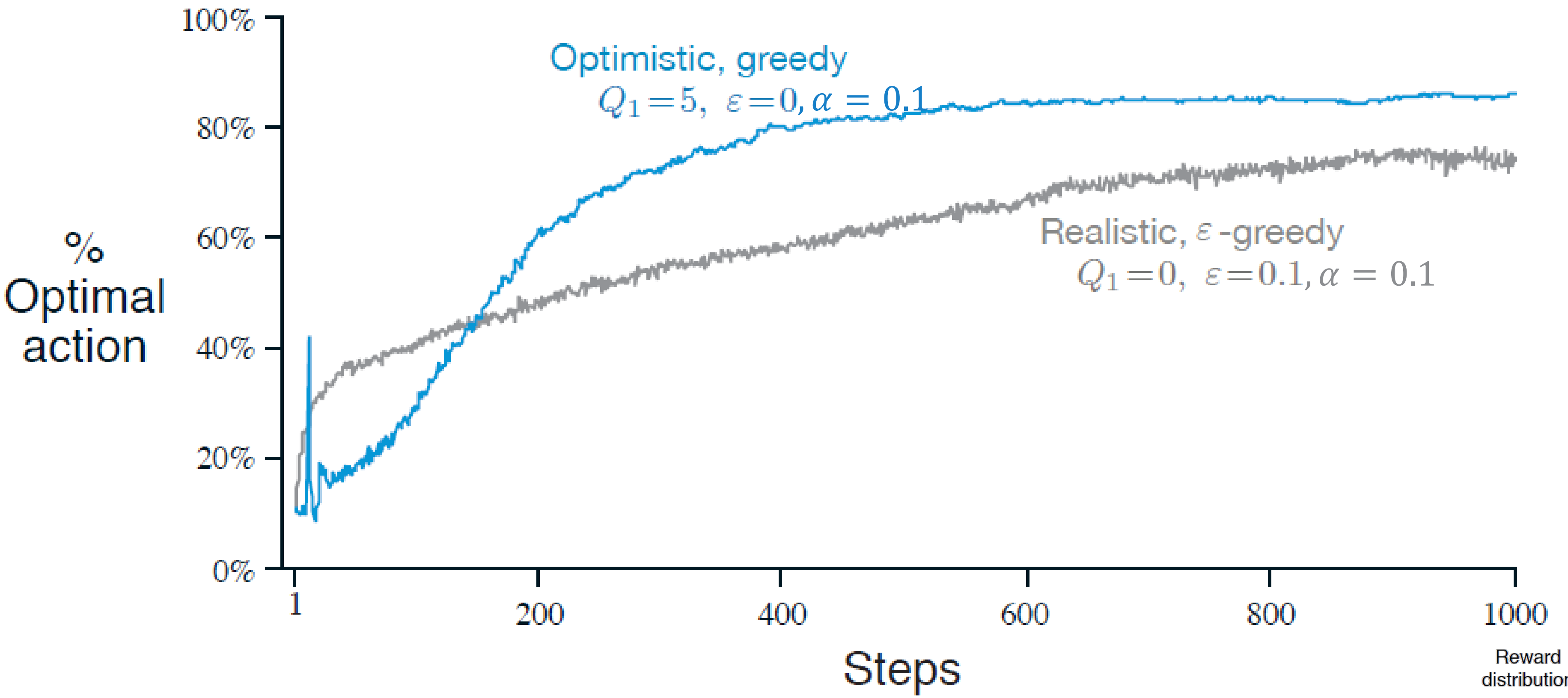
$q^*(a)$

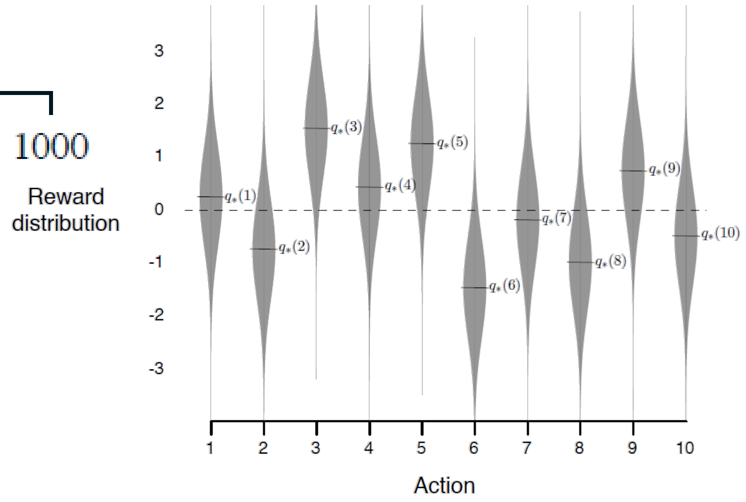
$$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$$
$$\alpha = 0.5$$



$Q_t(a) =$ 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 othe optimistic inital value should be

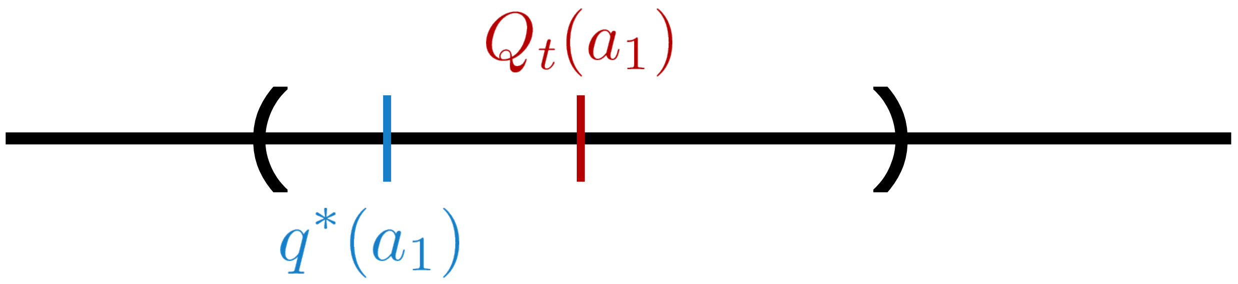
UCB Action Selection

Let’s go back to epsilon-greedy

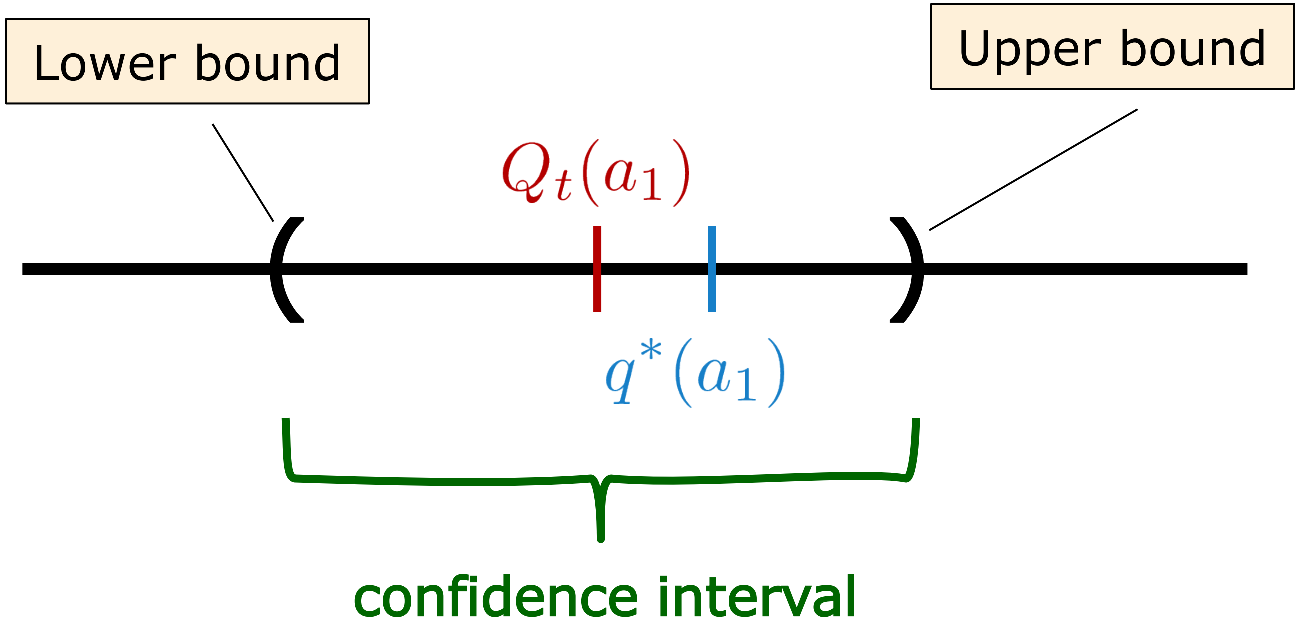
$$A_t = \begin{cases} argmax_a\, Q_t(a) & \text{with probability } 1-\epsilon \\ \boxed{Uniform(\{a_1, ..., a_k\})} & \text{with probability } \epsilon \end{cases}$$

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

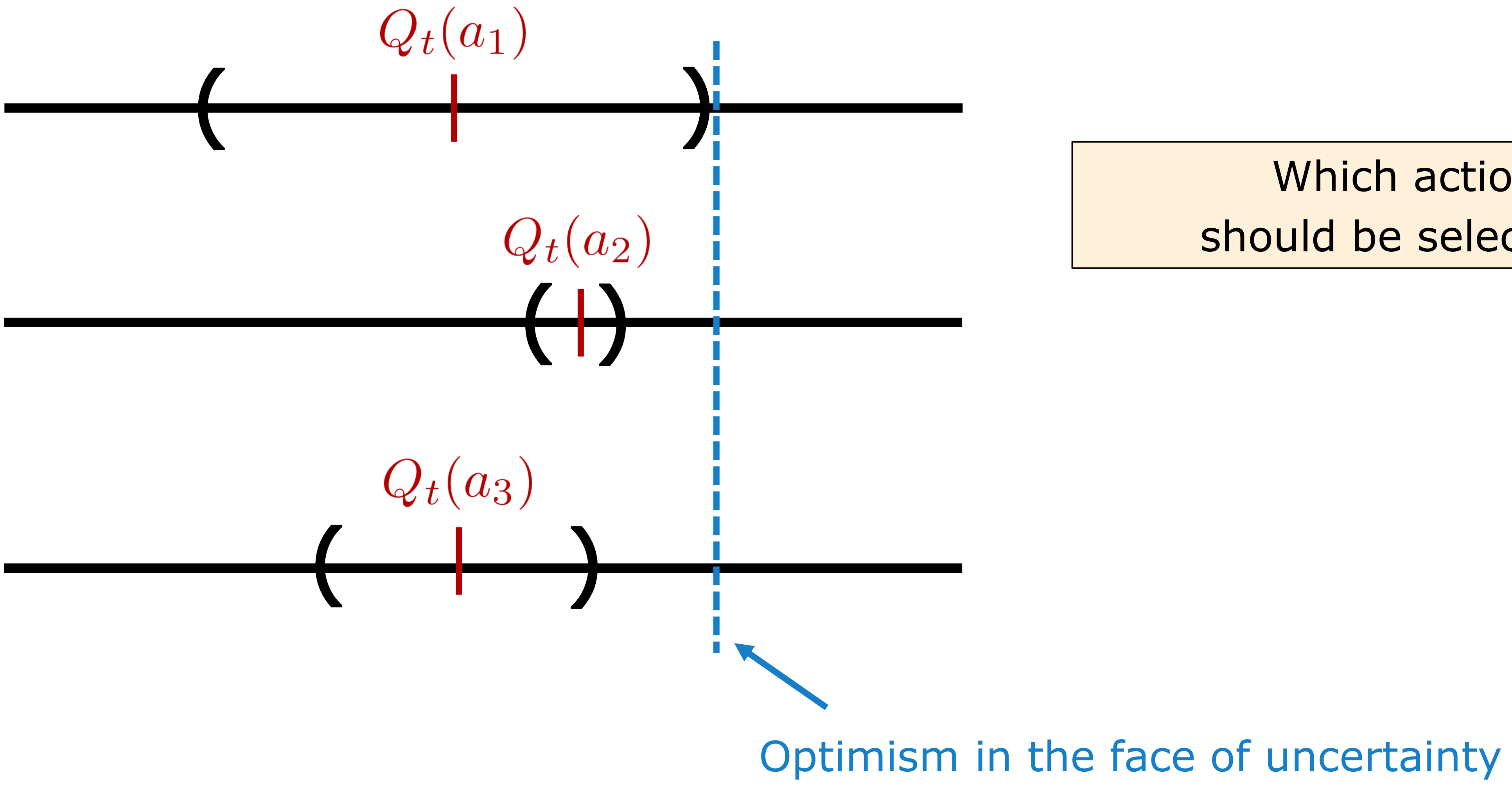
Uncertainty in Action-Values Estimates



Uncertainty in Action-Values Estimates

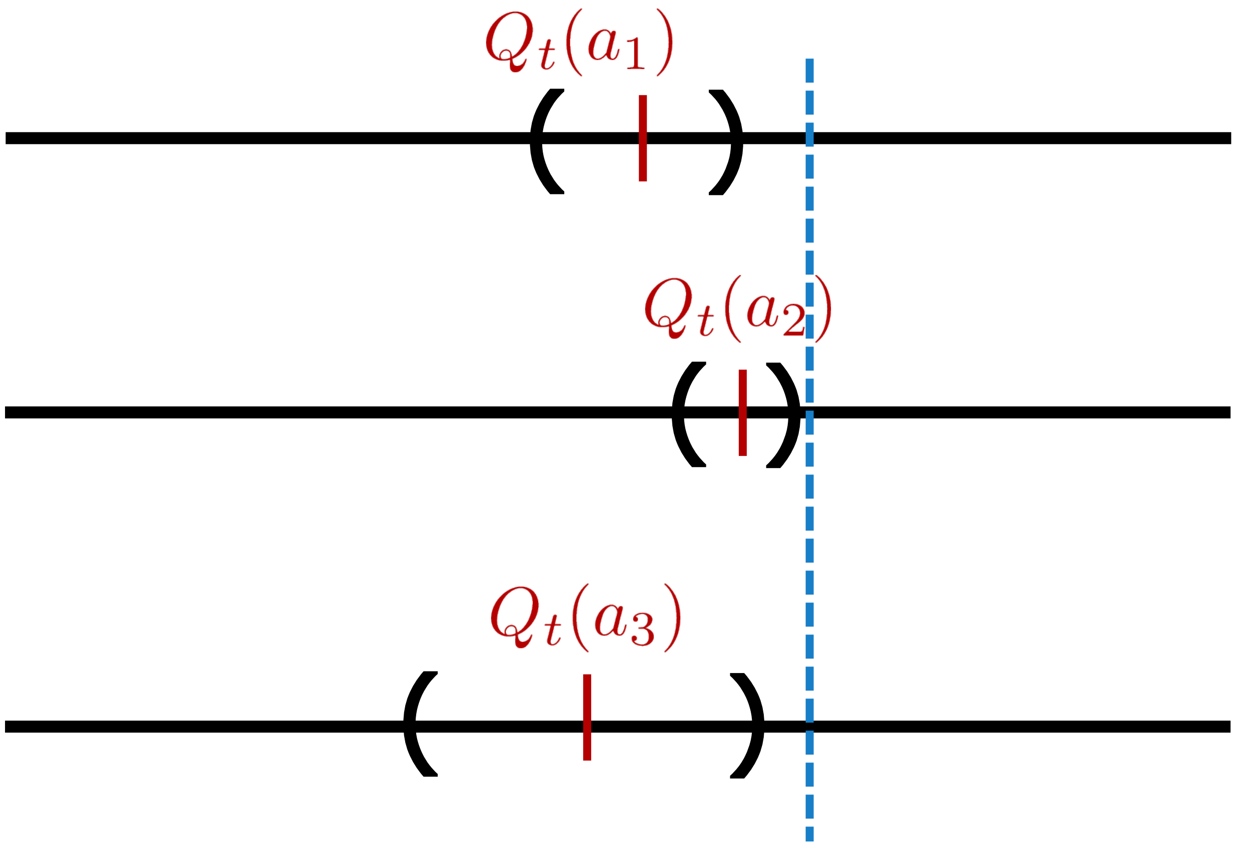


Upper-Confidence Bound (UCB) Action Selection



Which action
should be selected?

Upper-Confidence Bound (UCB) Action Selection



How do we compute upper bound?

Upper-Confidence Bound (UCB) Action Selection

Exploit

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

User-defined coefficient

Explore

UCB on the 10-armed testbed