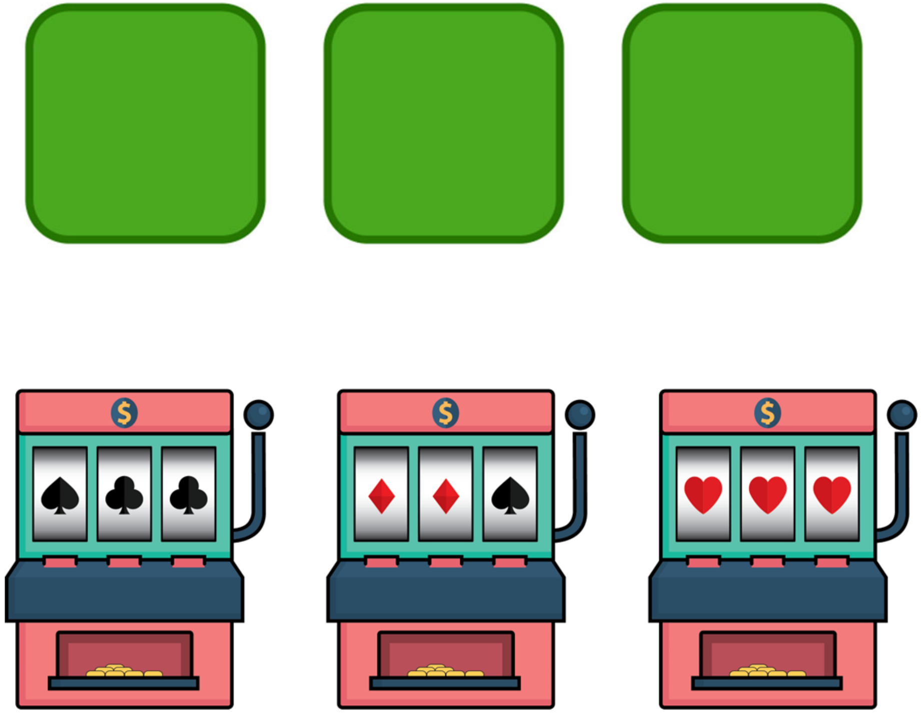
2. Exploración y Explotación

2.1 The k-armed bandit problem



2. Exploración y Explotación

2.1 The k-armed bandit problem

🟥 🟩 🟩

12

Q1(left)=12

🟩 🟥 🟩

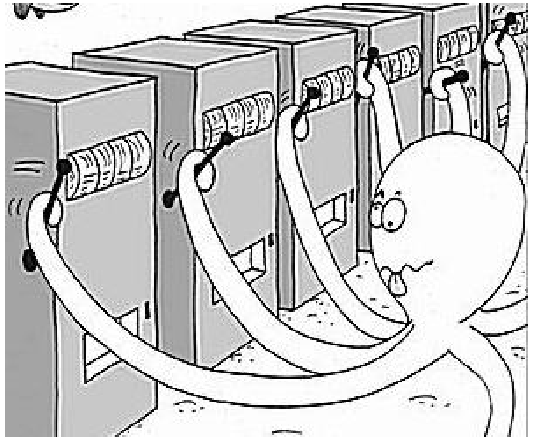
11

Q2(center)=11

🟩 🟩 🟥

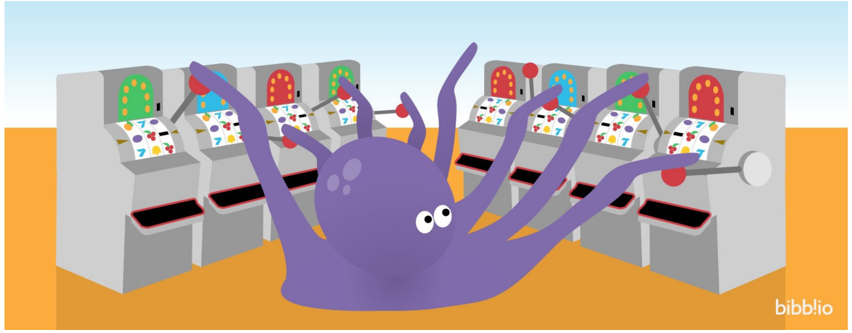
13

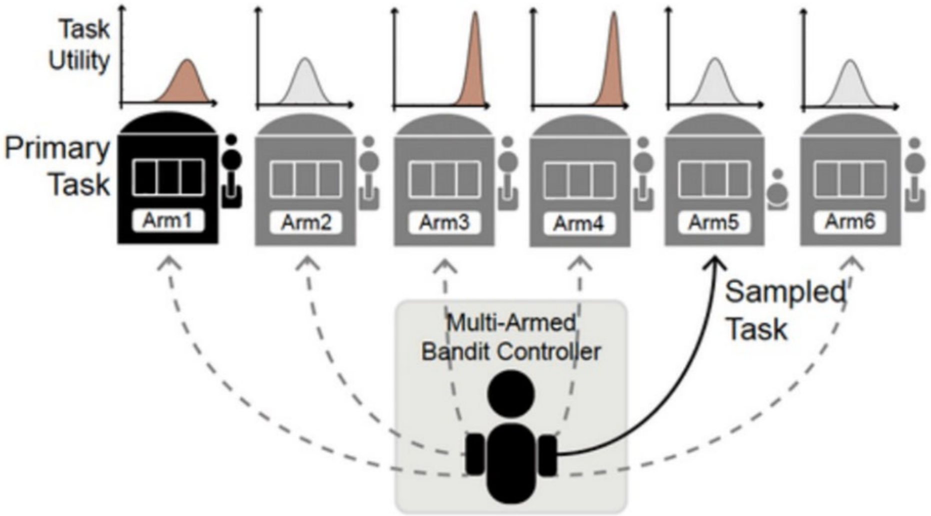
Q3(right)=13



2. Exploración y Explotación

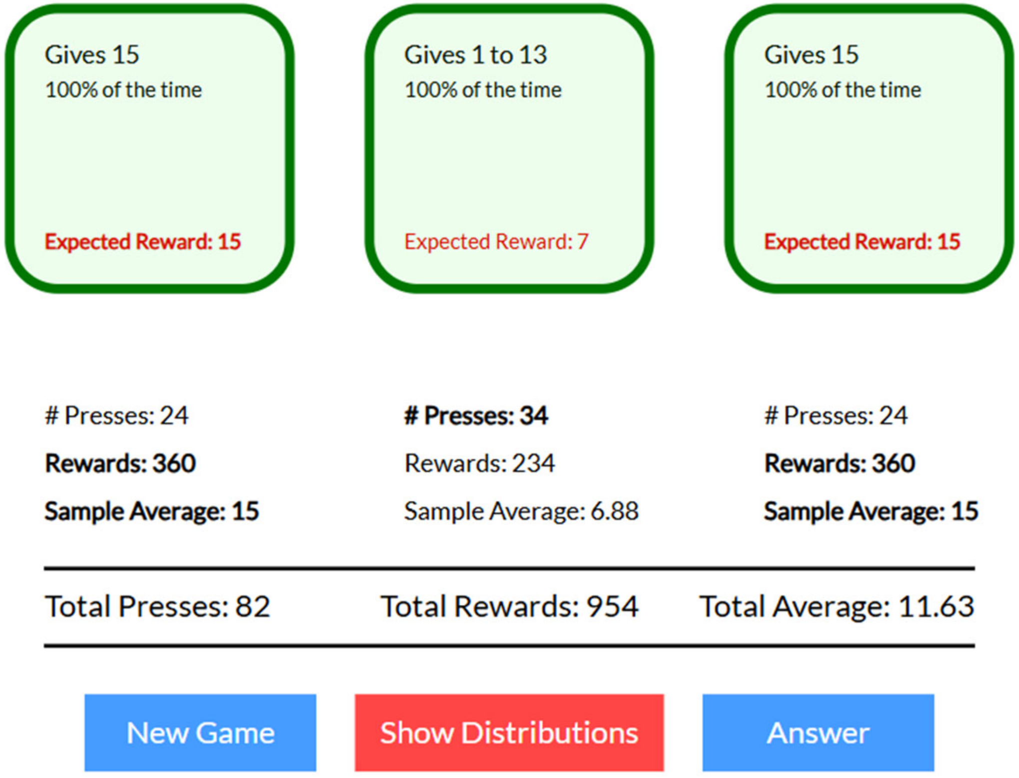
2.1 The k-armed bandit problem





2. Exploración y Explotación

2.1 The k-armed bandit problem



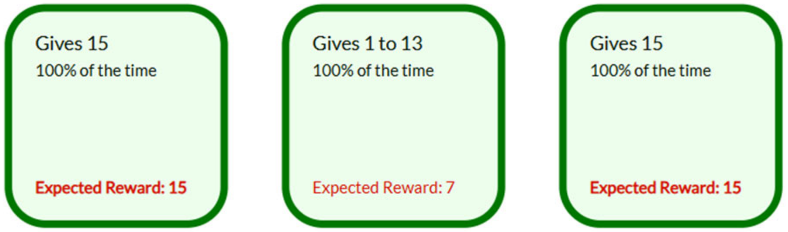
Gives 15
100% of the time
Expected Reward: 15

Gives 1 to 13
100% of the time
Expected Reward: 7

Gives 15
100% of the time
Expected Reward: 15

# Presses: 24	**# Presses: 34**	# Presses: 24
Rewards: 360	Rewards: 234	**Rewards: 360**
Sample Average: 15	Sample Average: 6.88	**Sample Average: 15**

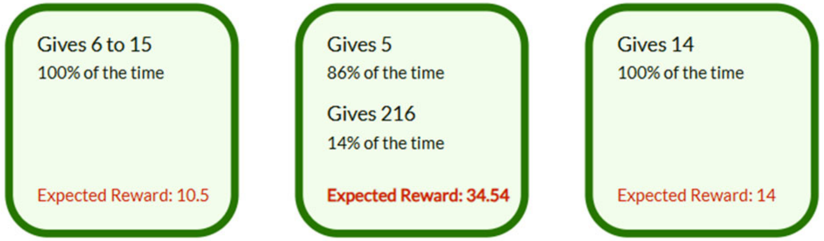
Total Presses: 82 Total Rewards: 954 Total Average: 11.63



Gives 15
100% of the time
Expected Reward: 15

Gives 1 to 13
100% of the time
Expected Reward: 7

Gives 15
100% of the time
Expected Reward: 15



Gives 6 to 15
100% of the time
Expected Reward: 10.5

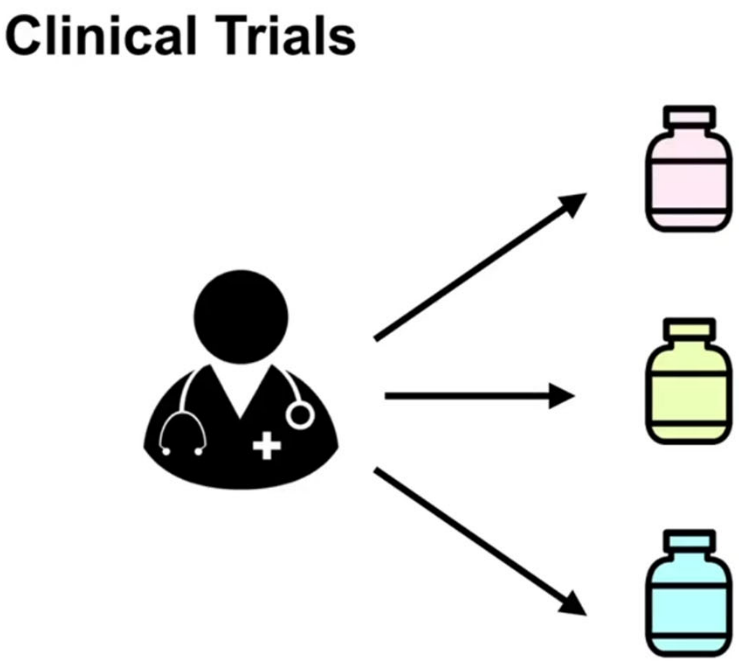
Gives 5
86% of the time
Gives 216
14% of the time
Expected Reward: 34.54

Gives 14
100% of the time
Expected Reward: 14

2. Exploración y Explotación

2.1 The k-armed bandit problem

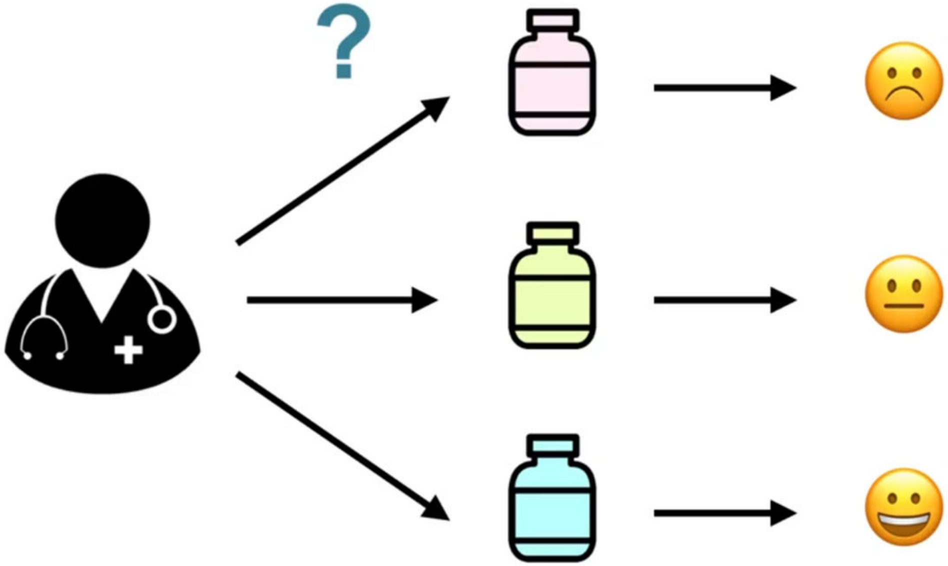
Imagine a medical trial where a doctor wants to measure the effect of three different treatments.



2. Exploración y Explotación

2.1 The k-armed bandit problem

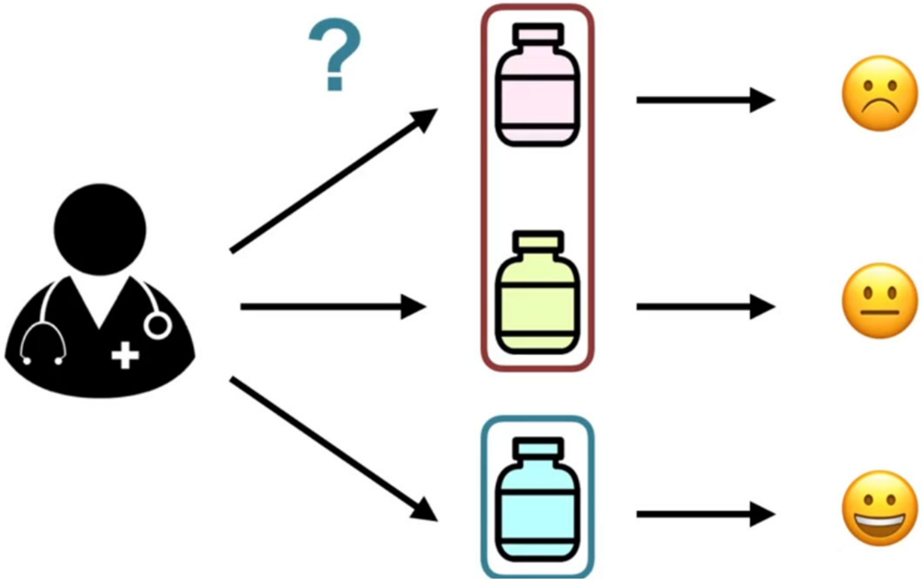
Whenever patient comes into the office, the doctor prescribes a treatment at **random.** The doctor then monitors the patient and observes any changes to their health. After a while, the doctor notices that one treatment seems to be working better than the others. The doctor must now decide between sticking with the best-performing treatment or continuing with the randomized study.



2. Exploración y Explotación

2.1 The k-armed bandit problem

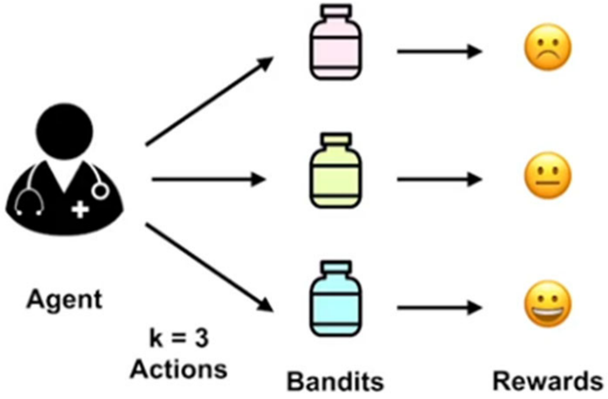
If the doctor only prescribes one treatment, then they can no longer collect data on the other two. Perhaps one of the other treatments is actually better, it only appears worse due to chance. If the other two treatments are worse, then continuing the study risk the health of the other patients. This medical trial exemplifies decision-making under **uncertainty**.



2. Exploración y Explotación

2.1 The k-armed bandit problem

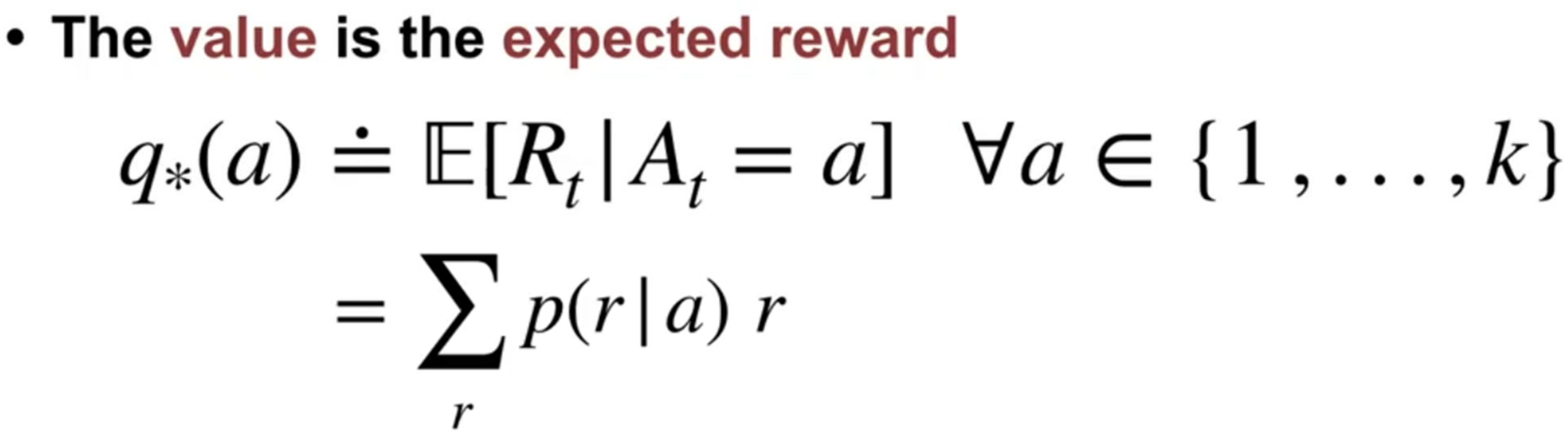
The medical trial example is a case of **the k-armed bandit problem**. In the k-armed bandit problem, we have a **decision-maker** or **agent**, who chooses between k different actions, and receives a **reward** based on the action he chooses. In the medical trial, the role of the agent is played by a doctor. The doctor has to choose between three different actions, to prescribe the pink, yellow or blue treatment. Each treatment is an action. Choosing that treatment yields some unknown reward. Finally, the welfare of the patient after the treatment is the reward that the doctor receives.



2. Exploración y Explotación

2.2 Action-Values Methods

For the doctor to decide which action is best, we must define the value of taking each action. We call these values the **action values** or the **action value function**. We define as *q star of a is defined as the expectation of R_t, given we selected action A, for each possible action one through k*. This conditional expectation is defined as a sum over all possible rewards.



- The value is the expected reward

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \quad \forall a \in \{1, \ldots, k\}$$

$$= \sum_r p(r \mid a)\, r$$

2. Exploración y Explotación

2.2 Action-Values Methods

- The value is the expected reward

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \quad \forall a \in \{1, \ldots, k\}$$
$$= \sum_r p(r \mid a)\, r$$

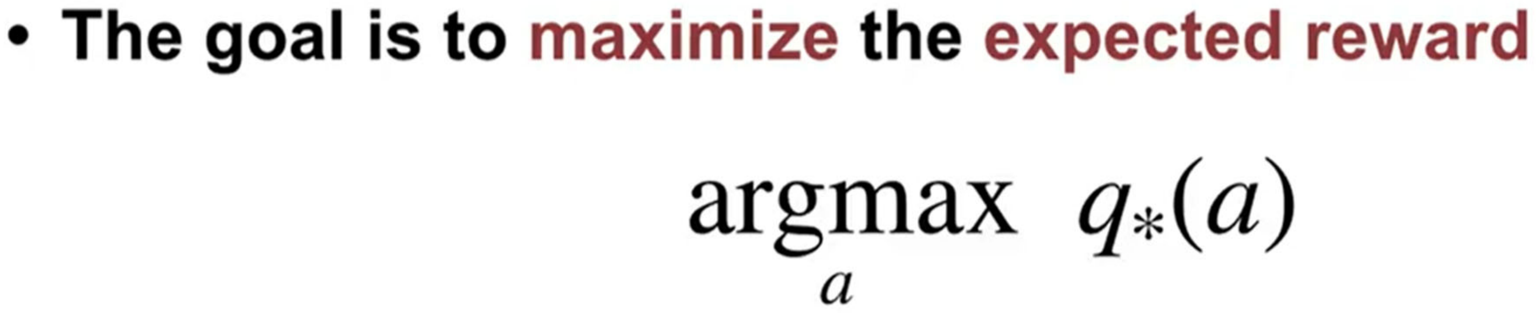
***a*:** an action (for example, choosing a bottle).
R_t: the random reward you receive.
$q_\star(a)$: the true value of the action.

The value of an action is the expected value (long-term average) of the reward when you choose that action.
Since the reward is random, it's not enough to see a single outcome because sometimes you win and sometimes you lose. You multiply each possible reward by its probability and add them up. That's the expected value.

2. Exploración y Explotación

2.2 Action-Values Methods

The goal of the agent is to maximize the expected reward. If the agent selects the action that has the highest value, it achieves that goal. We call this procedure the **argmax**, or the argument which maximizes our function q star.



- **The goal is to maximize the expected reward**

$$\underset{a}{\operatorname{argmax}} \; q_*(a)$$

2. Exploración y Explotación

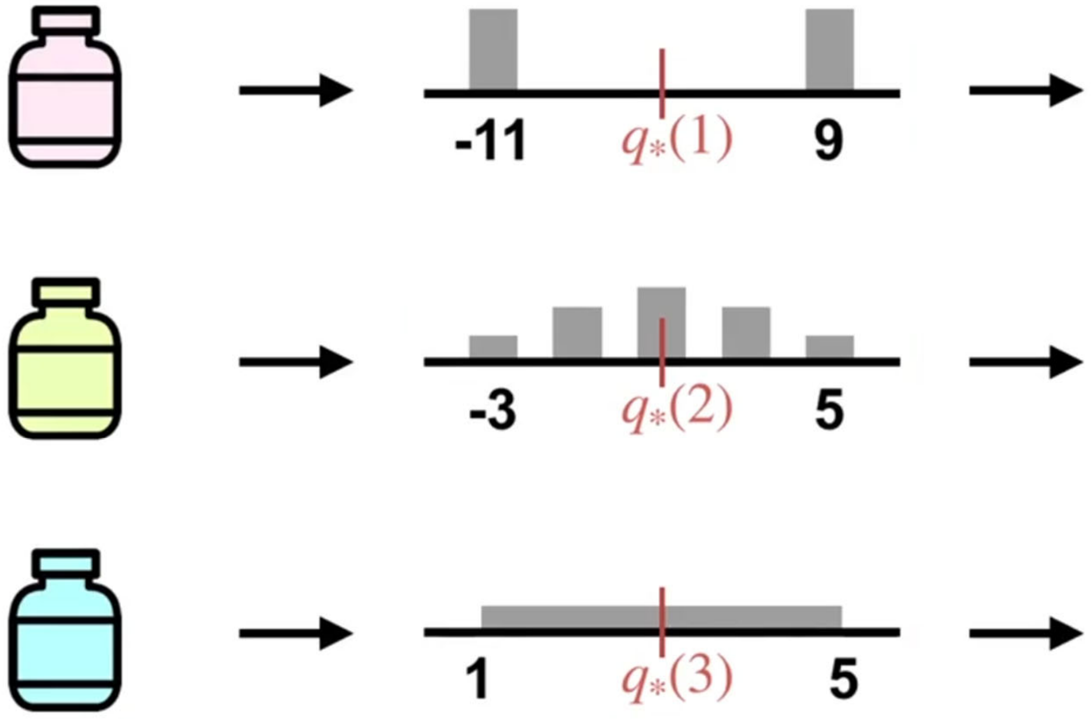
2.2 Action-Values Methods

Very simple example:
Let's use something which is easier to measure, perhaps the change in blood pressure after receiving the treatment. Each treatment may yield rewards following different probability distributions. Perhaps one is Bernoulli, one is binomial, and one is uniform. ***Q star is the mean of the distributions for each action.*** You can easily calculate the expected value of the Bernoulli distribution. Simply multiply the probability of failure by the reward when failed, plus the probability of success, times reward when succeeded.

2. Exploración y Explotación

2.2 Action-Values Methods

Calculating $q_*(a)$



$$q_*(a) = .5 \times -11 + .5 \times 9$$

$$q_*(a) = 1$$

$$q_*(a) = 3$$

2. Exploración y Explotación

2.2 Action-Values Methods

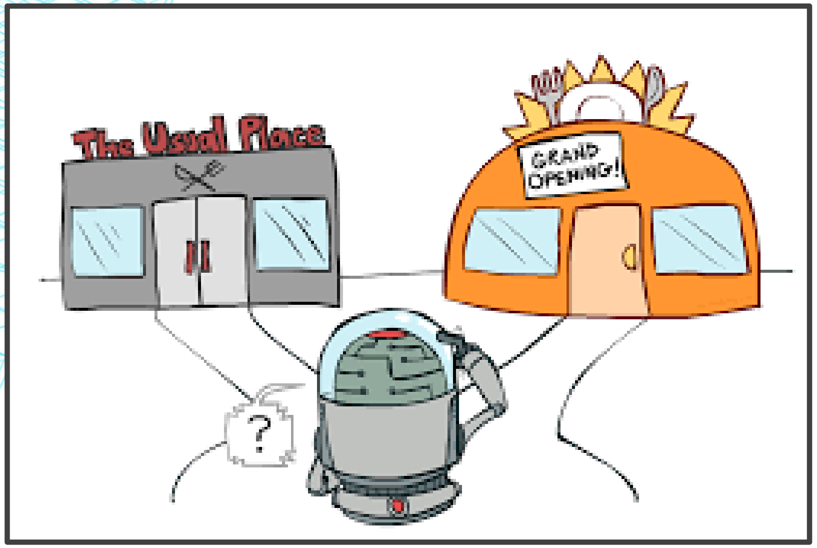
There are many examples of making decisions under uncertainty. For instance, the medical trial example, content recommendations, like what movie to watch, or what song to listen to. Even ordering food at a restaurant, you can't be certain what you like, but you make the best choice you can.

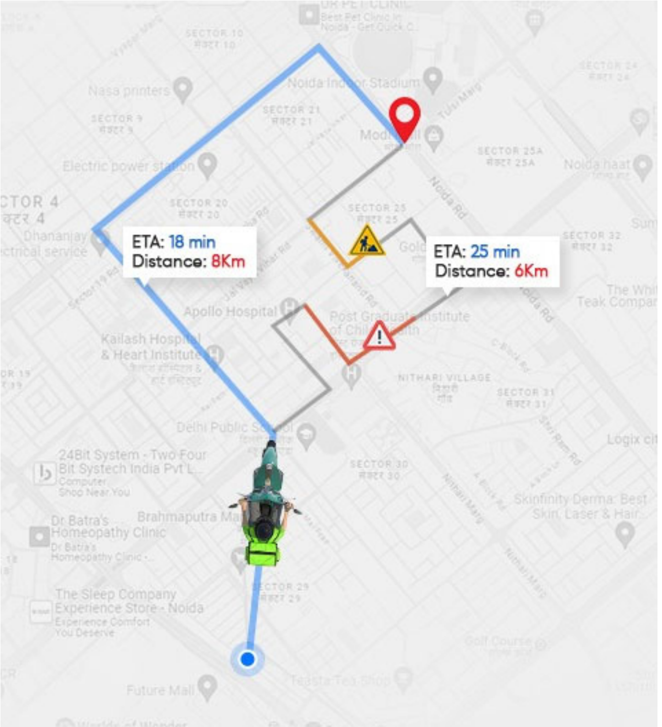
Examples of making decisions under uncertainty

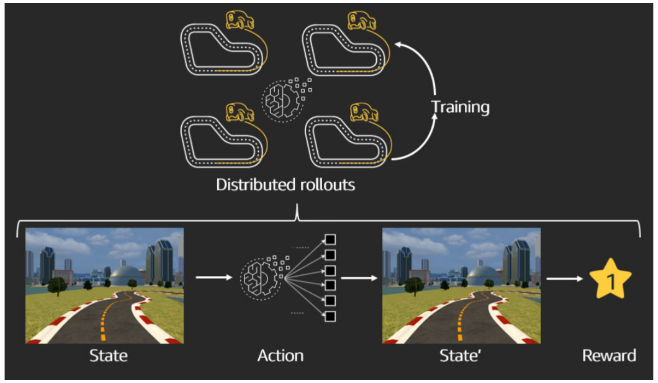
2. Exploración y Explotación

2.2 Action-Values Methods

There are many examples for decision-process





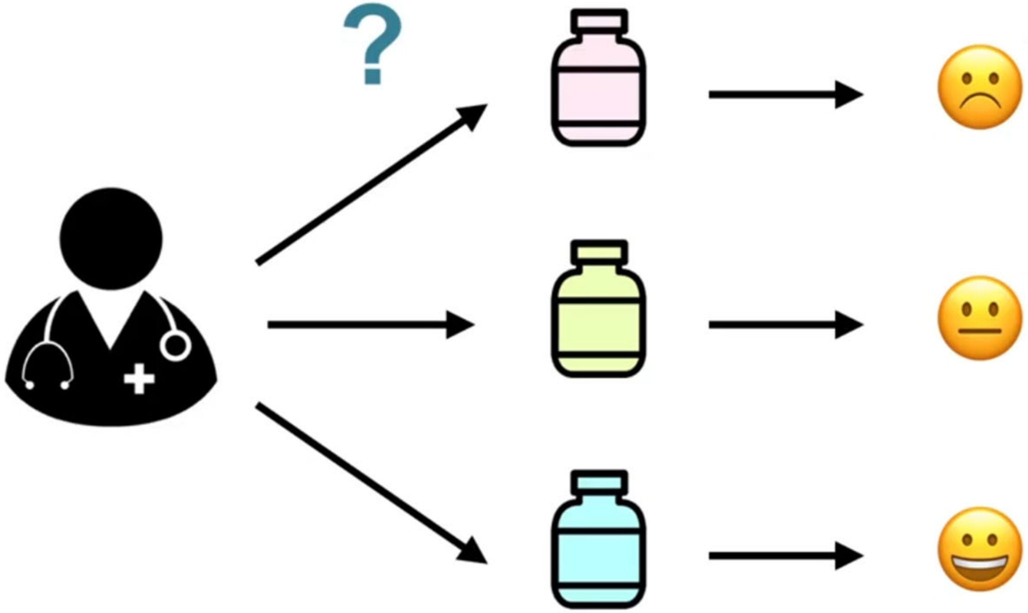


2. Exploración y Explotación

2.3 Epsilon-greedy strategies

In the doctor medical trial:

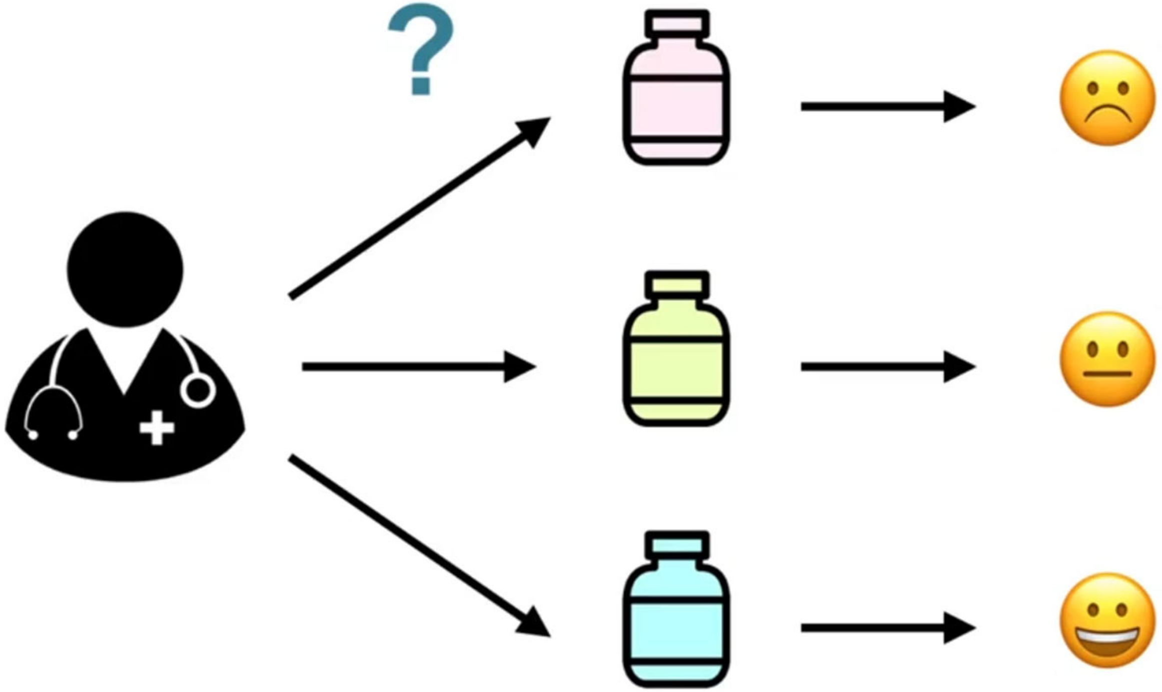
What would happen if this doctor already knew the long-term outcome of each treatment?



2. Exploración y Explotación

2.3 Epsilon-greedy strategies

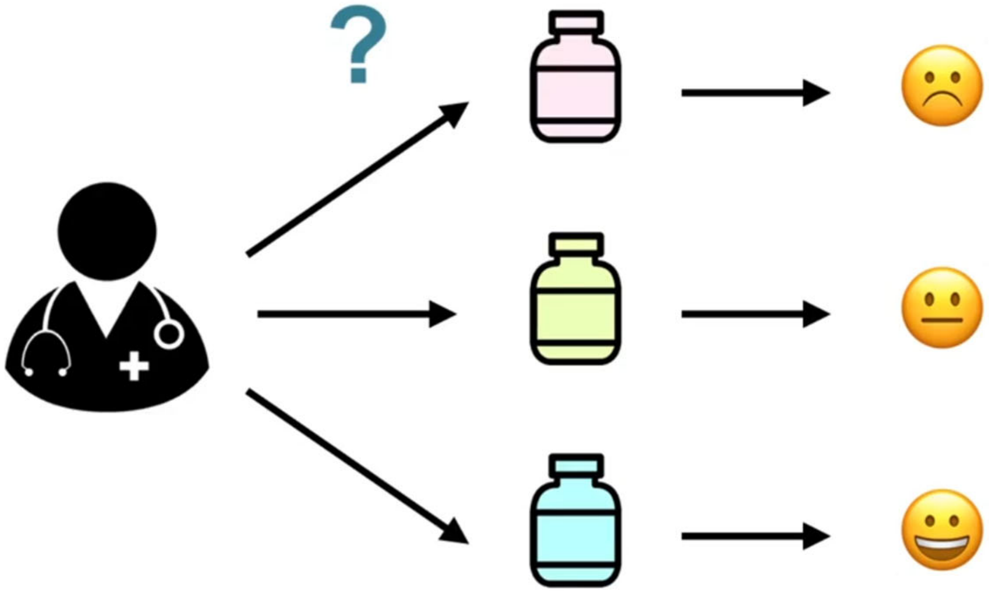
Choosing the appropriate treatment would be trivial.



2. Exploración y Explotación

2.3 Epsilon-greedy strategies

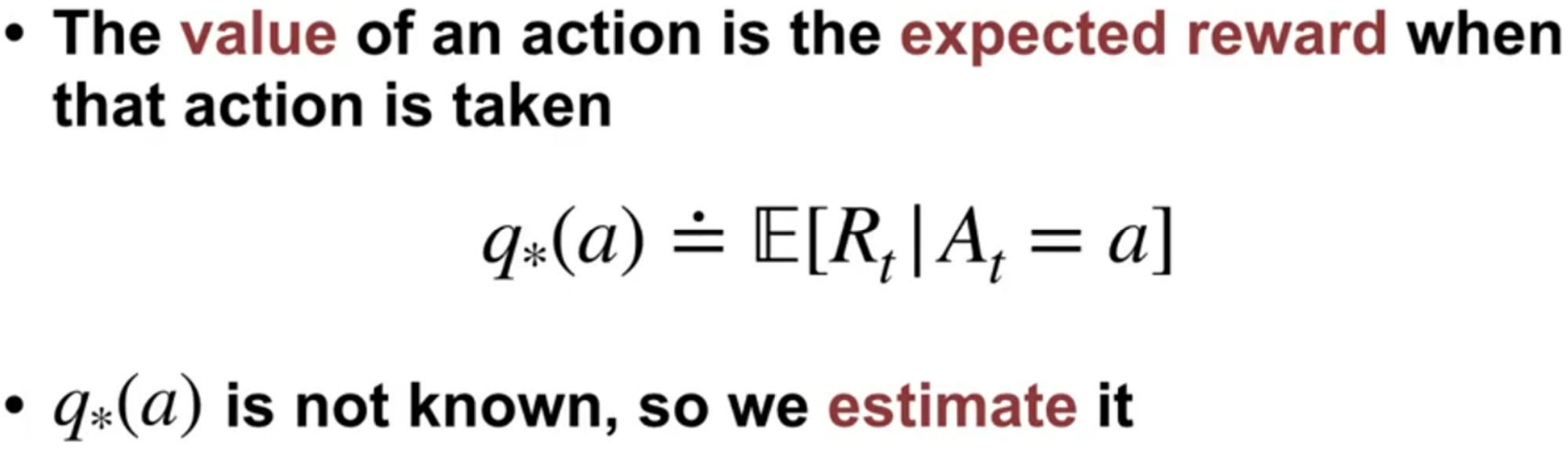
Usually, the doctor will run many trials to learn about each treatment. Each day, the doctor could use all the previously collected data to estimate which treatment they believed to be the best.



2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Remember



- The value of an action is the expected reward when that action is taken

$$q_*(a) \doteq \mathbb{E}[R_t \,|\, A_t = a]$$

- $q_*(a)$ is not known, so we estimate it

2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}$$

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

We use t-1 because the value at time t is based on actions taken prior to time t. Also, if action A has not yet been taken, we set the value to some default like zero.

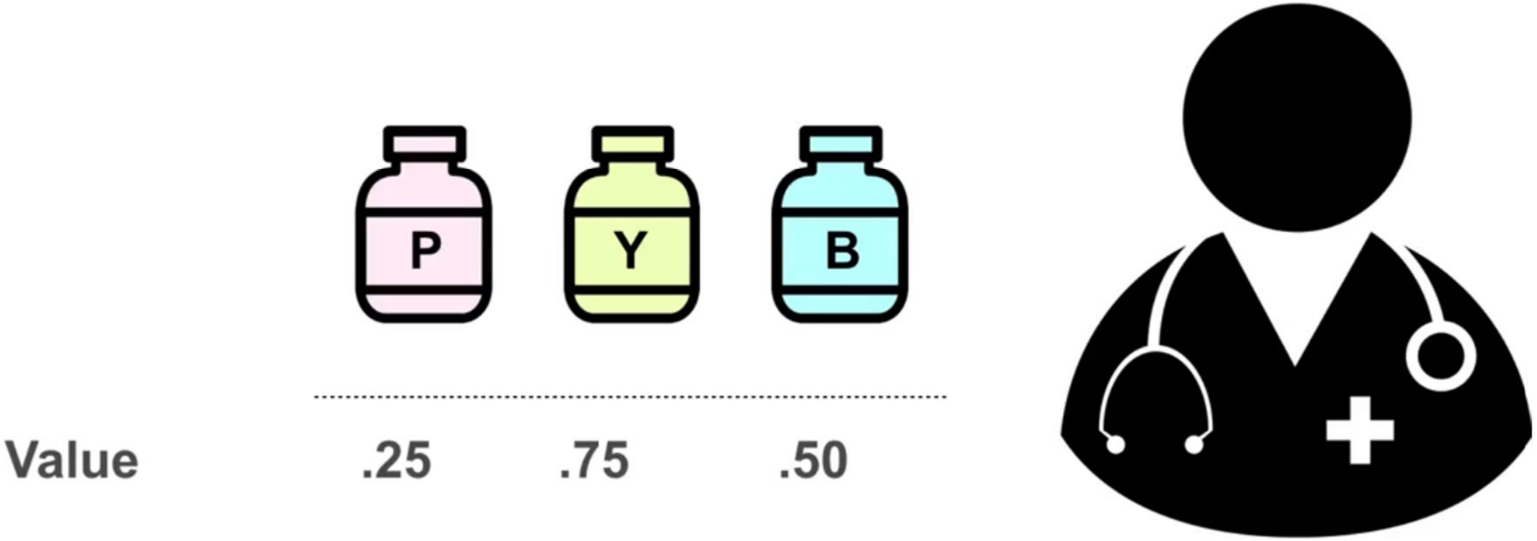
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

Example: Clinical Trials

A reward of 1 if the treatment succeeds otherwise 0



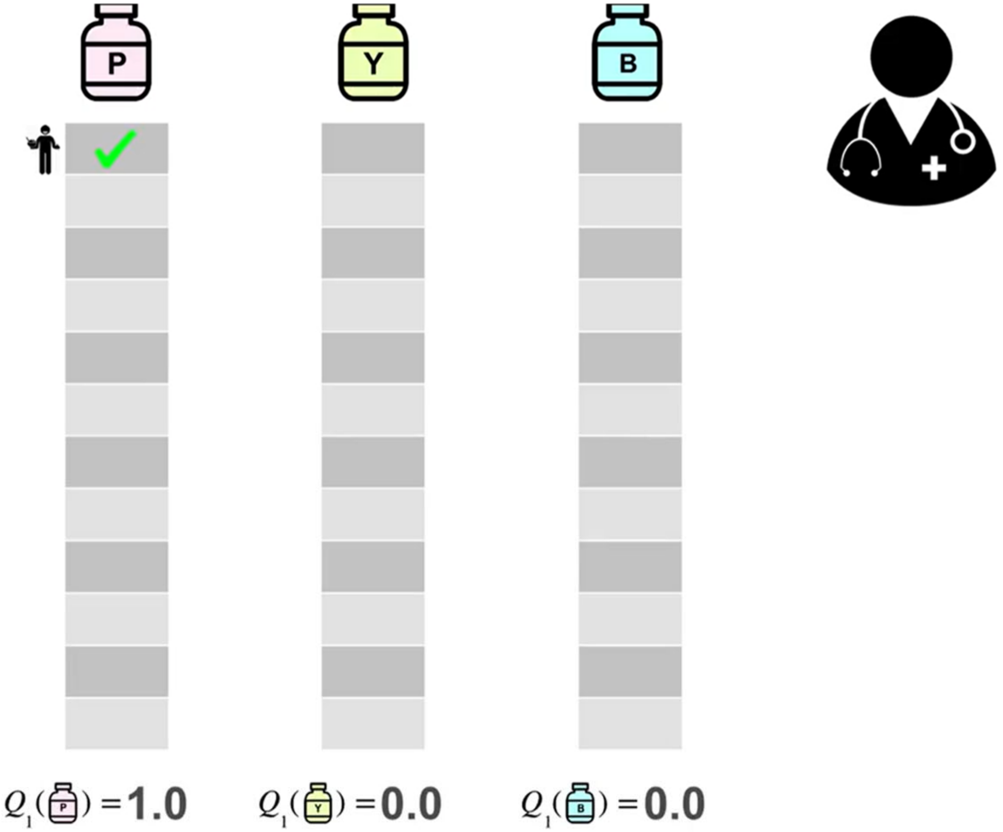
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

A reward of 1 if the treatment succeeds otherwise 0

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



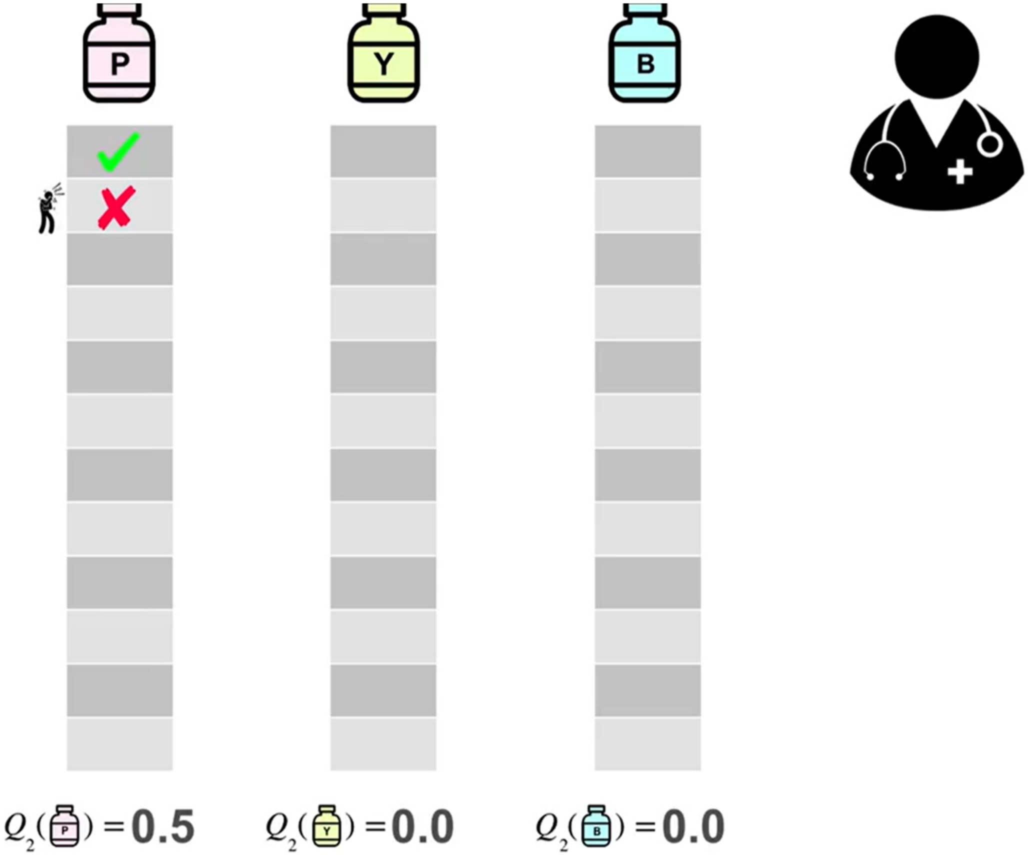
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

A reward of 1 if the treatment succeeds otherwise 0

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



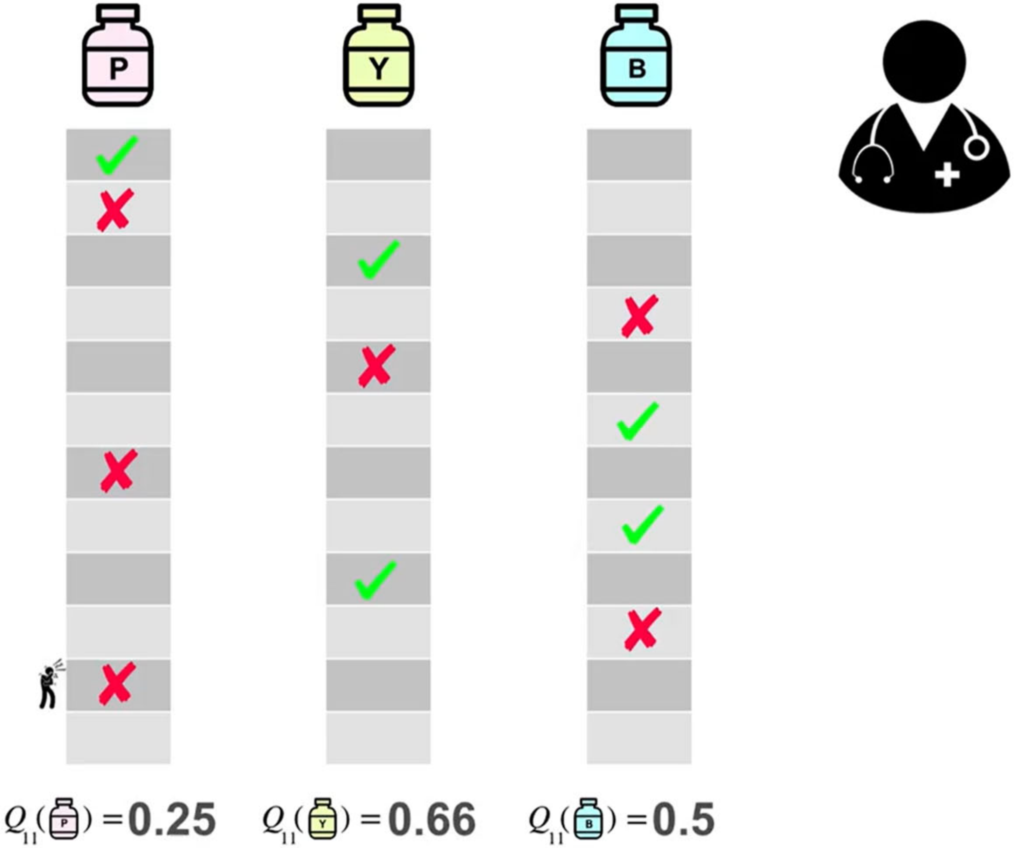
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

A reward of 1 if the treatment succeeds otherwise 0

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



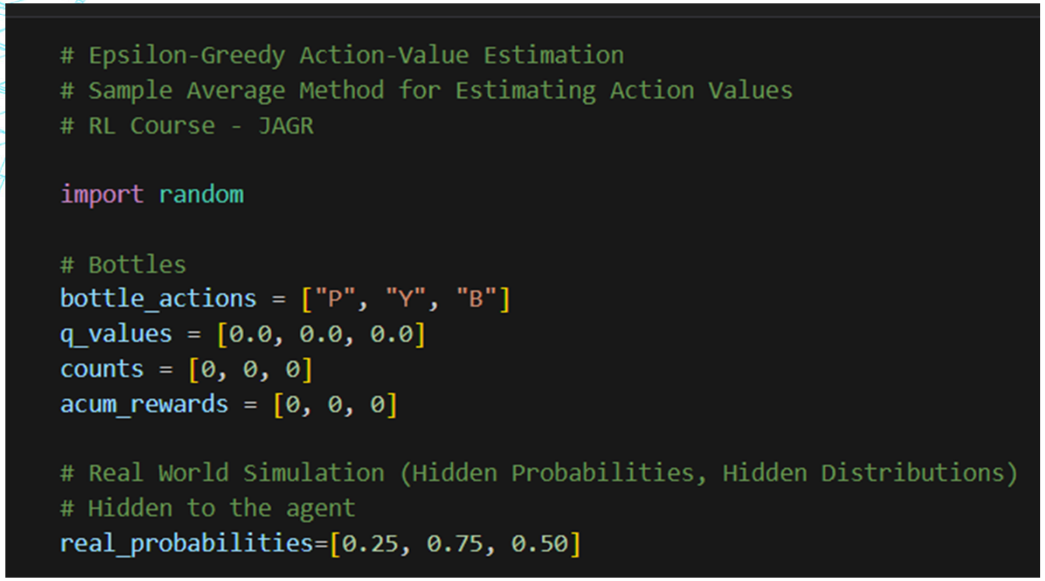
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Sample – Average Method

Python Example

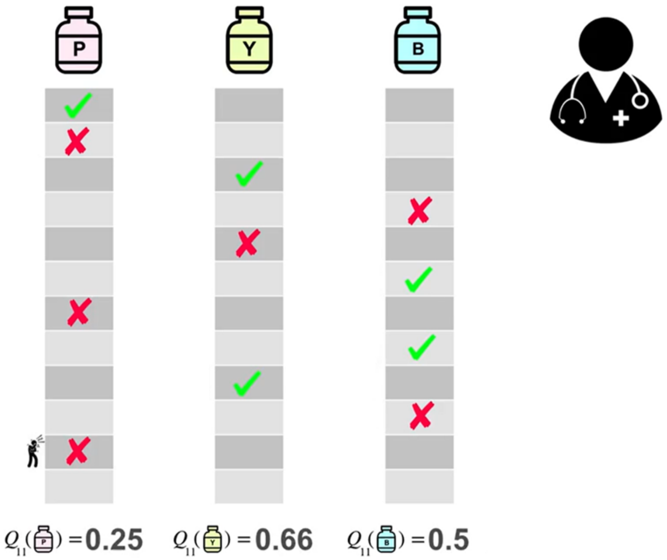
Action_values_Doctor.py



```
# Epsilon-Greedy Action-Value Estimation
# Sample Average Method for Estimating Action Values
# RL Course - JAGR

import random

# Bottles
bottle_actions = ["P", "Y", "B"]
q_values = [0.0, 0.0, 0.0]
counts = [0, 0, 0]
acum_rewards = [0, 0, 0]

# Real World Simulation (Hidden Probabilities, Hidden Distributions)
# Hidden to the agent
real_probabilities=[0.25, 0.75, 0.50]
```

A reward of 1 if the treatment succeeds otherwise 0

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

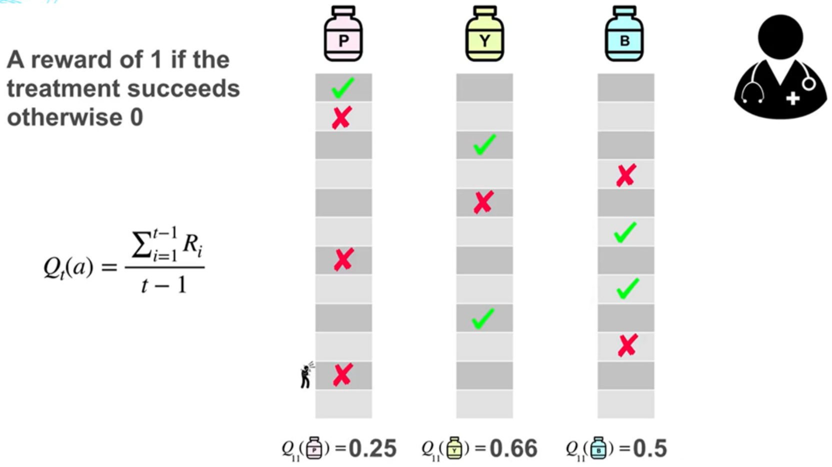


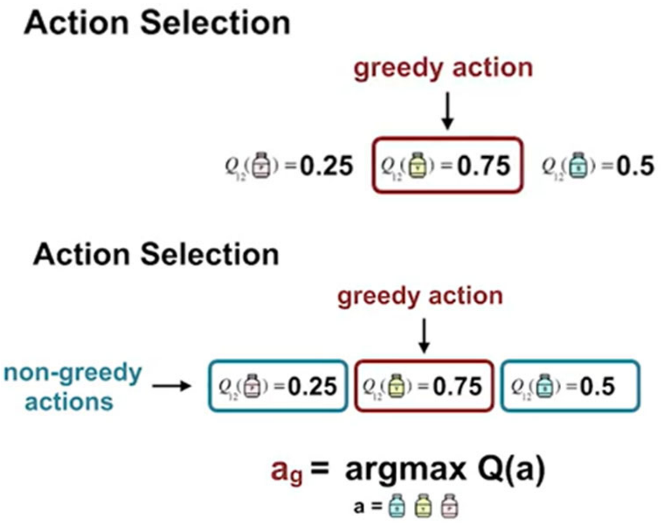
2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Greedy Action

As the doctor observes more patients, the estimates approach the true action values. In reality, our doctor would not randomly assign treatments to their patients. Instead, they would probably assign the treatment that they currently think is the best. We call this method of choosing actions greedy. The greedy action is the action that currently has the largest estimated value. Selecting the greedy action means the agent is exploiting its current knowledge. It is trying to get the most reward it can right now.


A reward of 1 if the treatment succeeds otherwise 0
$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i}{t-1}$
P
Y
B
Q_{11}(P) =0.25
Q_{11}(Y) =0.66
Q_{11}(B) =0.5

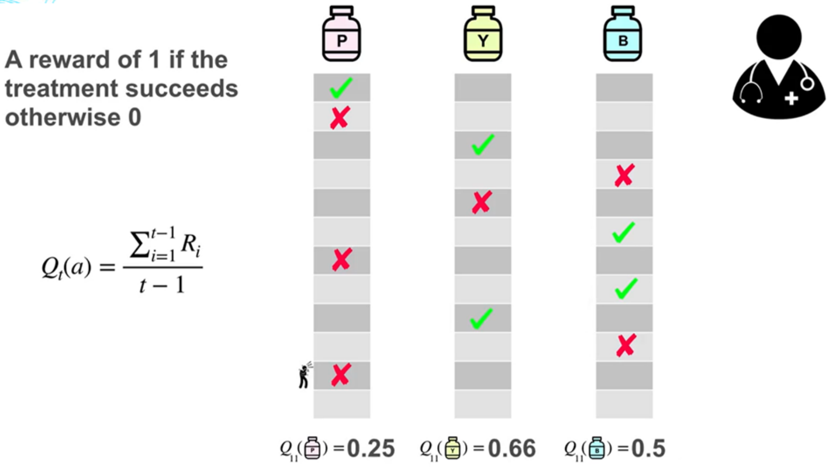


2. Exploración y Explotación

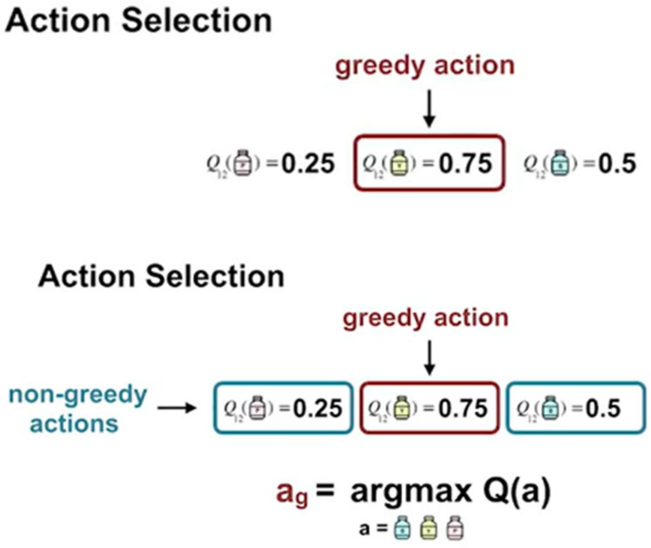
2.3 Epsilon-greedy strategies

Greedy Action

We can compute the **greedy action** by taking the **argmax** of our estimated values. Alternatively, the agent may choose to **explore** by choosing a non-greedy action. The agent would sacrifice immediate reward hoping to gain more information about the other actions. The agent can not choose to both **explore** and **exploit** at the same time. *This is one of the fundamental problems in reinforced learning:* **<u>The exploration-exploitation dilemma.</u>**



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



a_g = argmax Q(a)

2. Exploración y Explotación

2.3 Epsilon-greedy strategies

Project U2_Part 01