Welcome to Part 6 - Reinforcement Learning

Section 26, Lecture 168

Welcome to Part 6 - Reinforcement Learning!

Reinforcement Learning is a branch of Machine Learning, also called Online Learning. It is used to solve interacting problems where the data observed up to time t is considered to decide which action to take at time t + 1. It is also used for Artificial Intelligence when training machines to perform tasks such as walking. Desired outcomes provide the AI with reward, undesired with punishment. Machines learn through trial and error.

In this part, you will understand and learn how to implement the following Reinforcement Learning models:

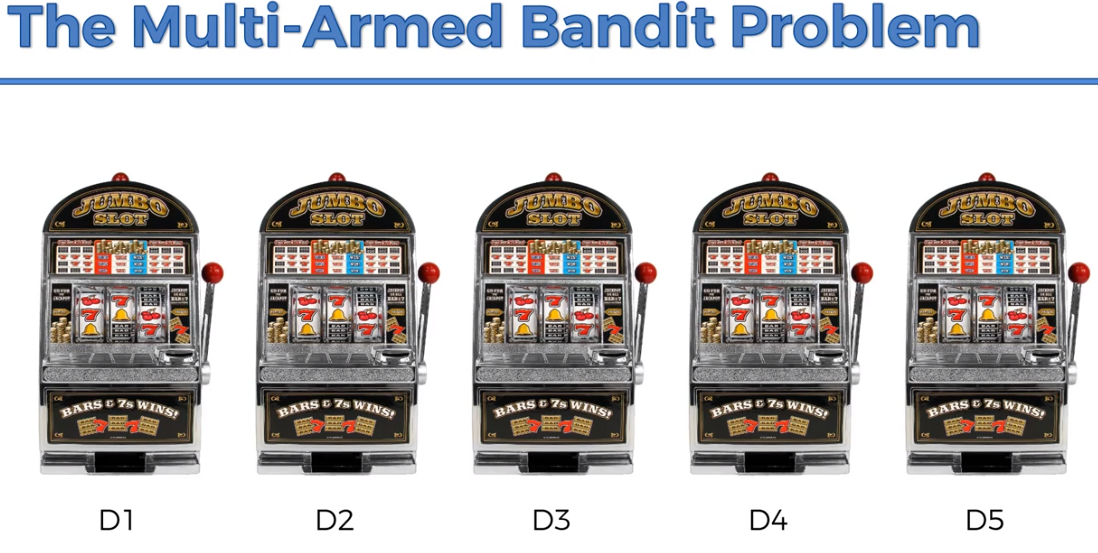
1. Upper Confidence Bound (UCB)
2. Thompson Sampling

To train robot dogs to walk,

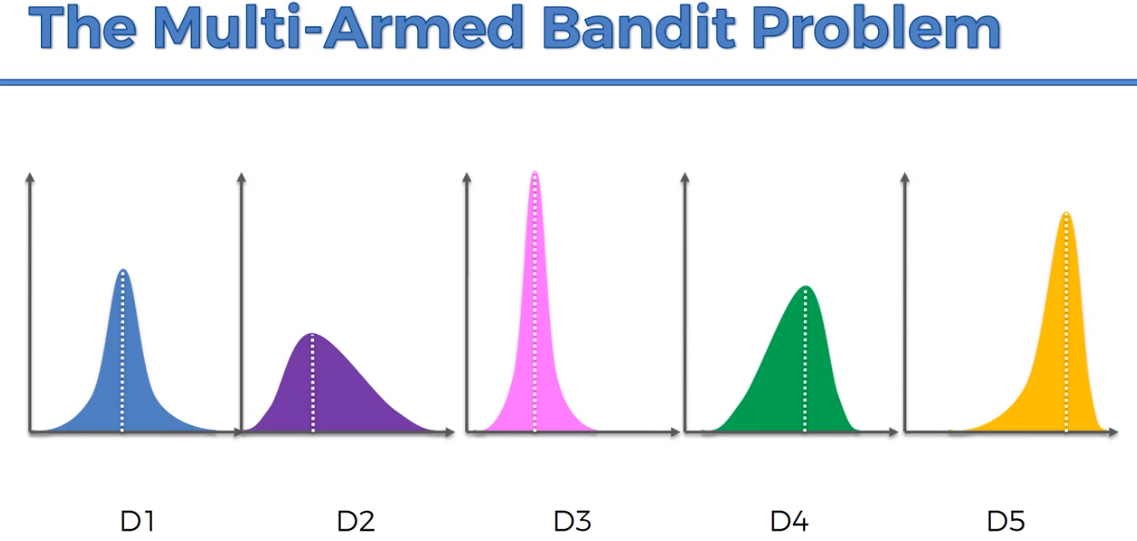


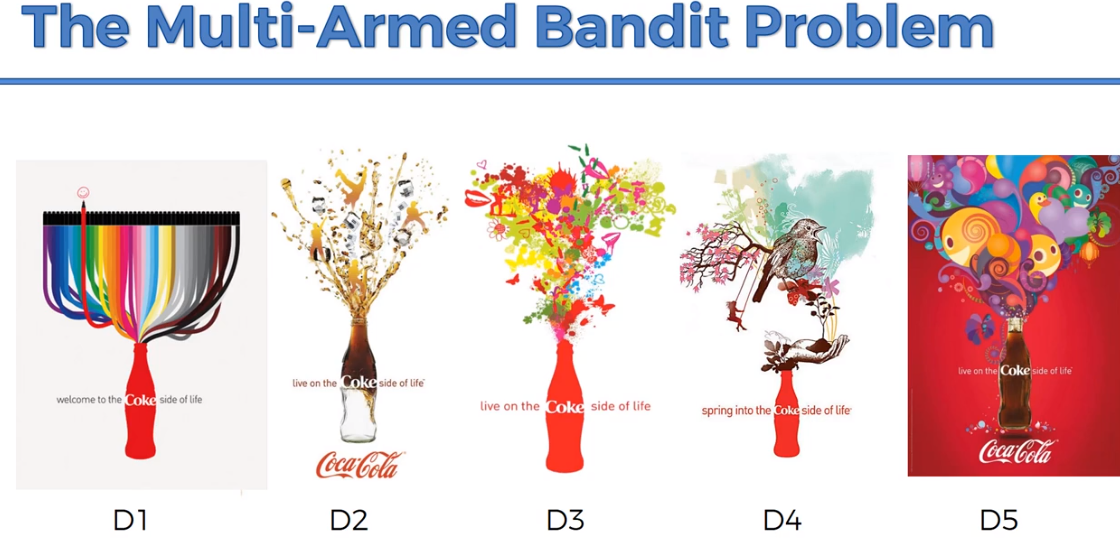
Reward🡪1

Punishment🡪0



How to bet on a machine and maximize the return.D5(Distribution 5 is thee best)

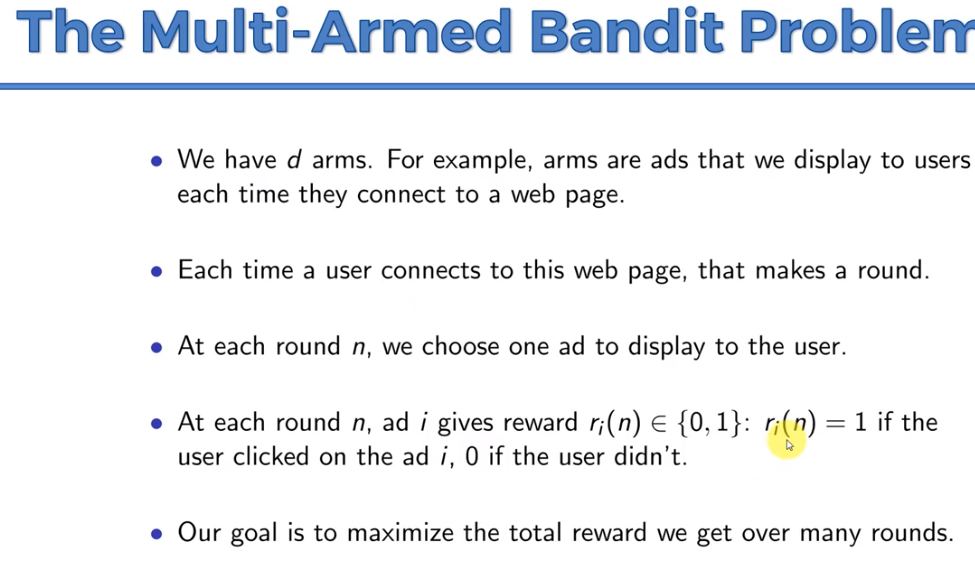


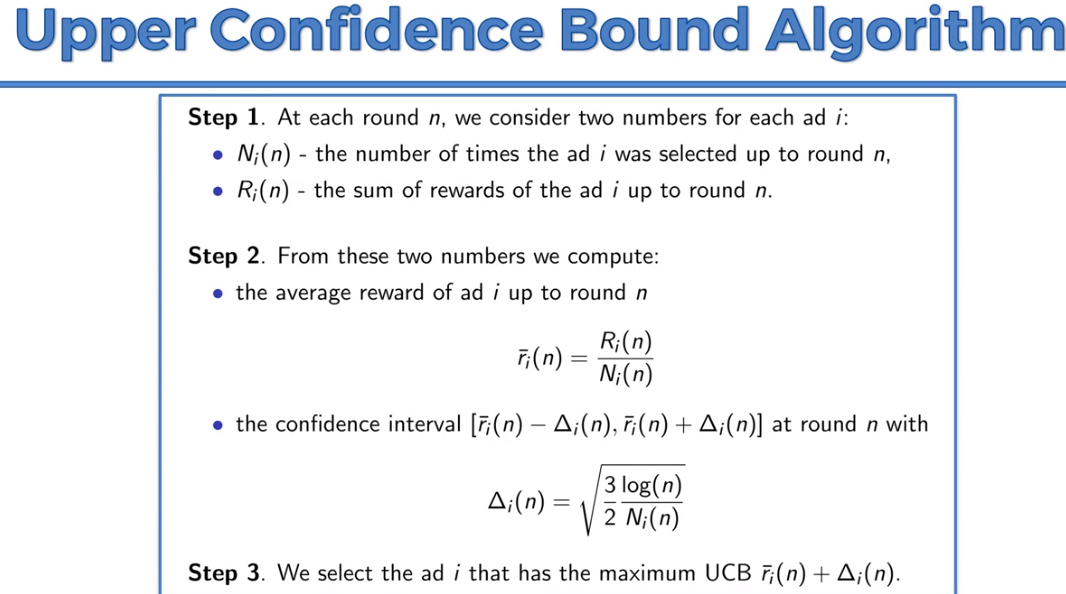


We have to conclude which is the best advertisement for coke, rather than going through all the advertisement.



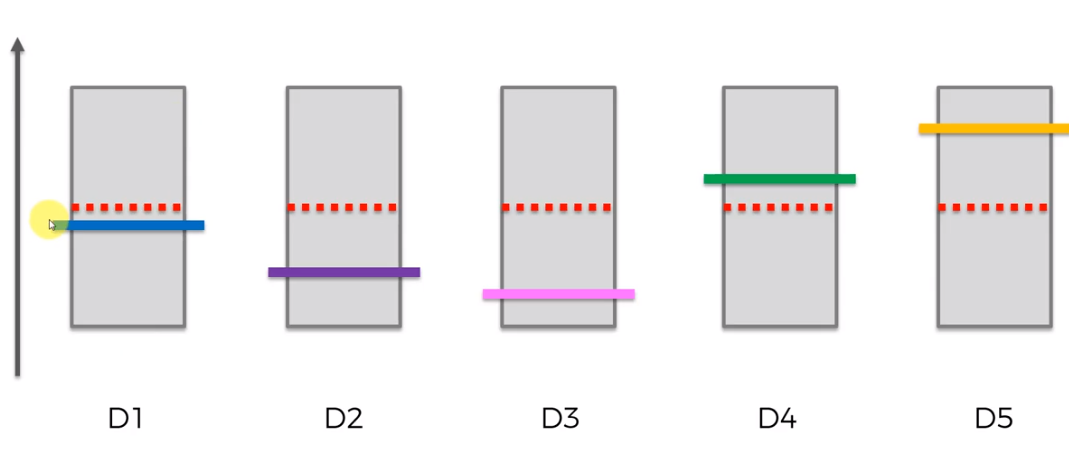
We are combining exploration and analysis to get the result in a optimal way.



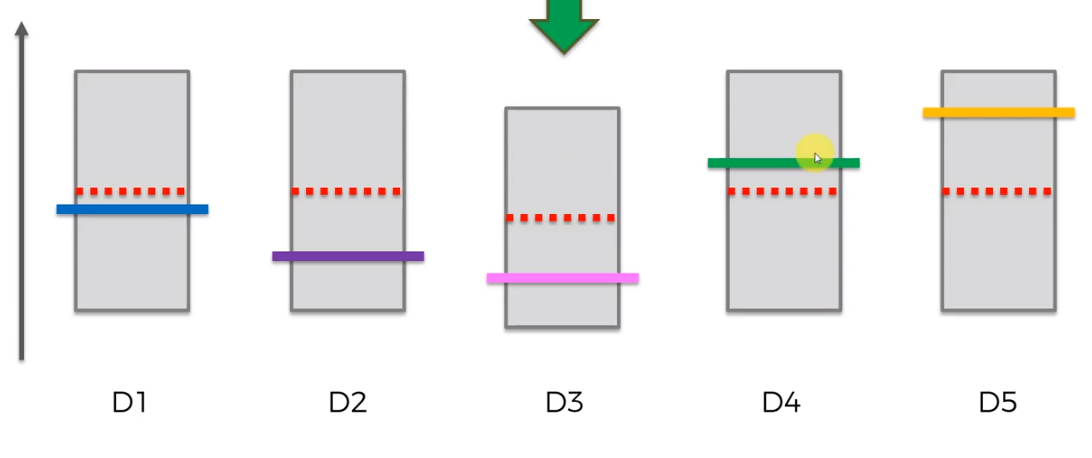


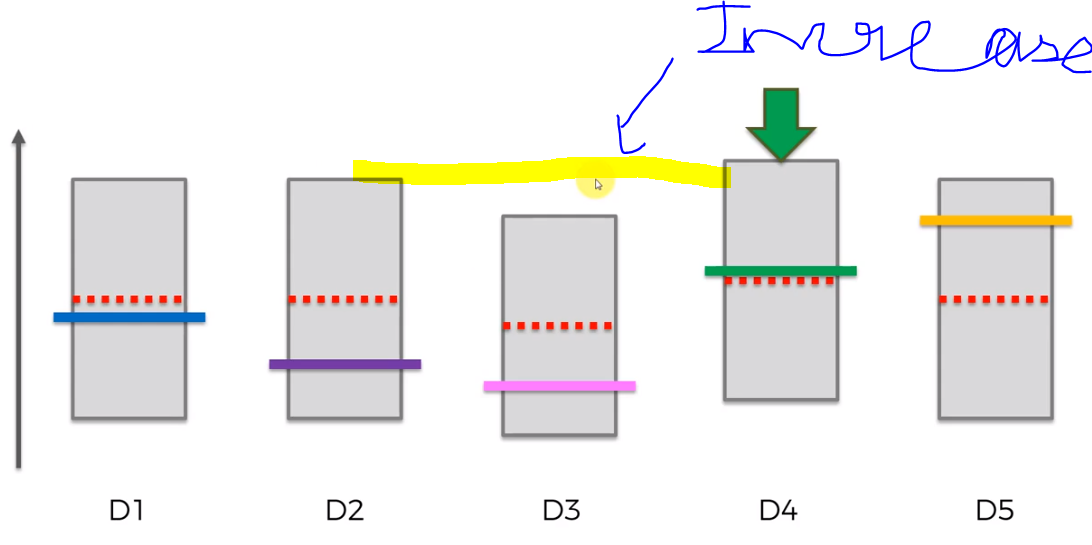


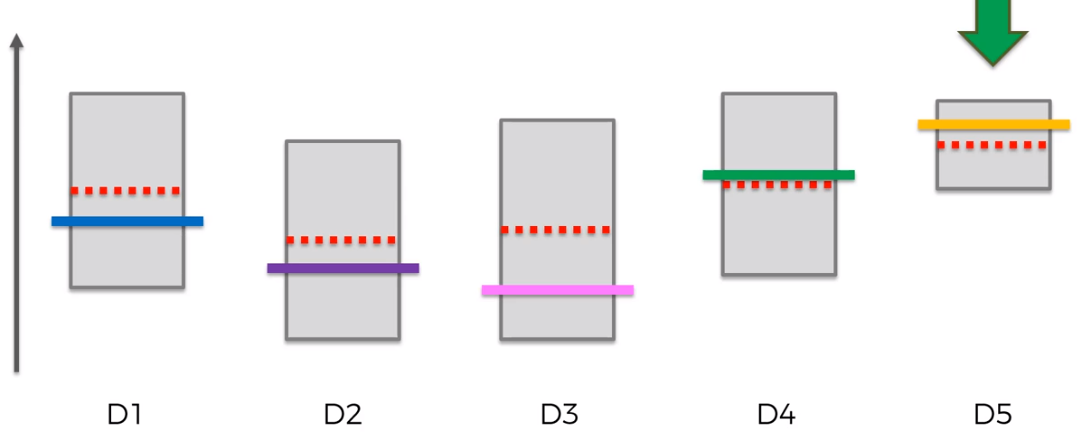


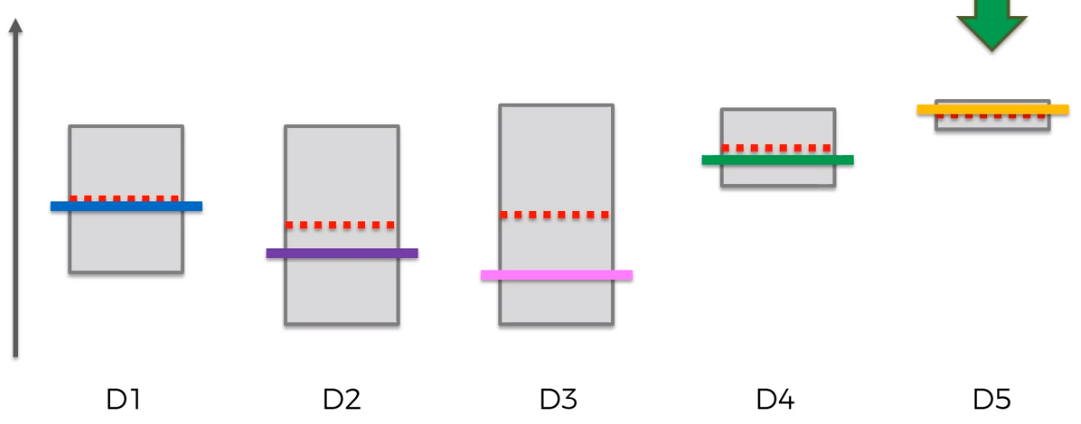


Pick one ‘D’ and display the ad. If the user doesn’t click the confidence will reduce.



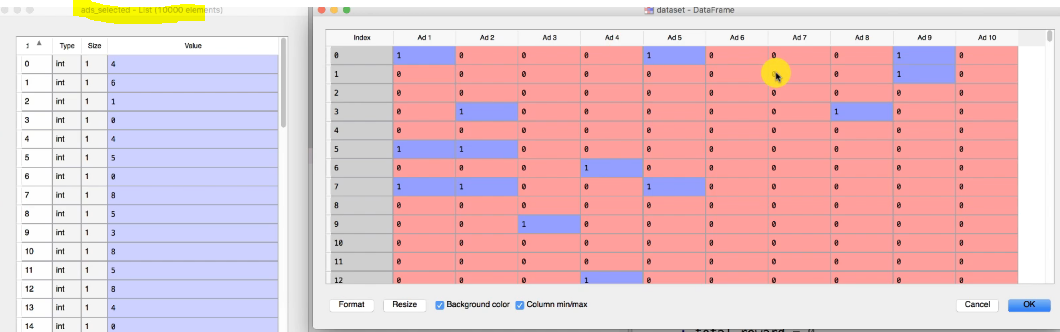


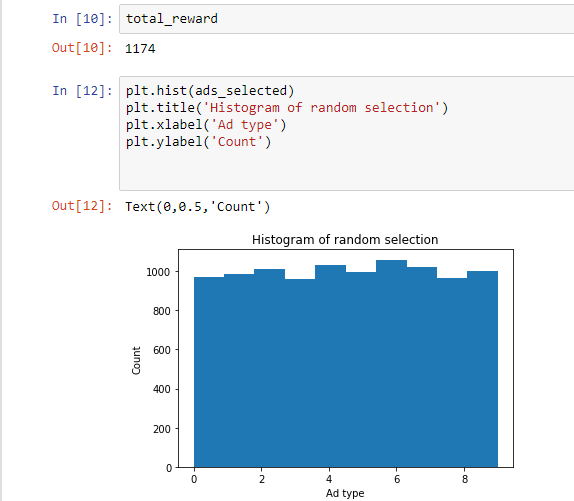






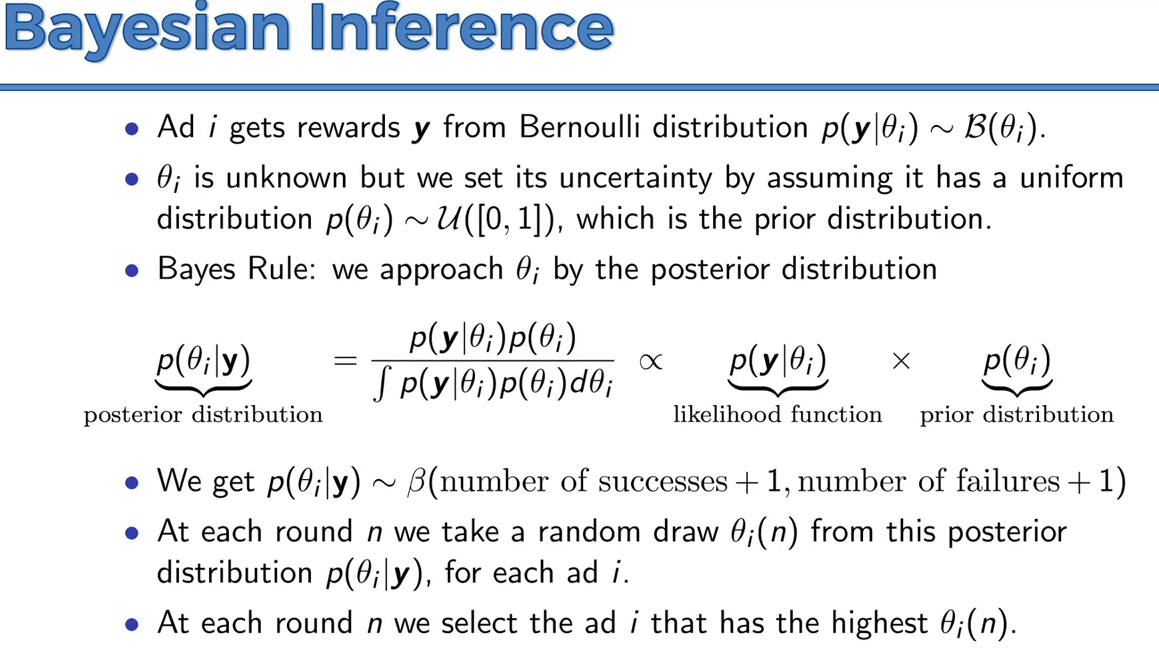
Random selection without algorithm or another strategy,

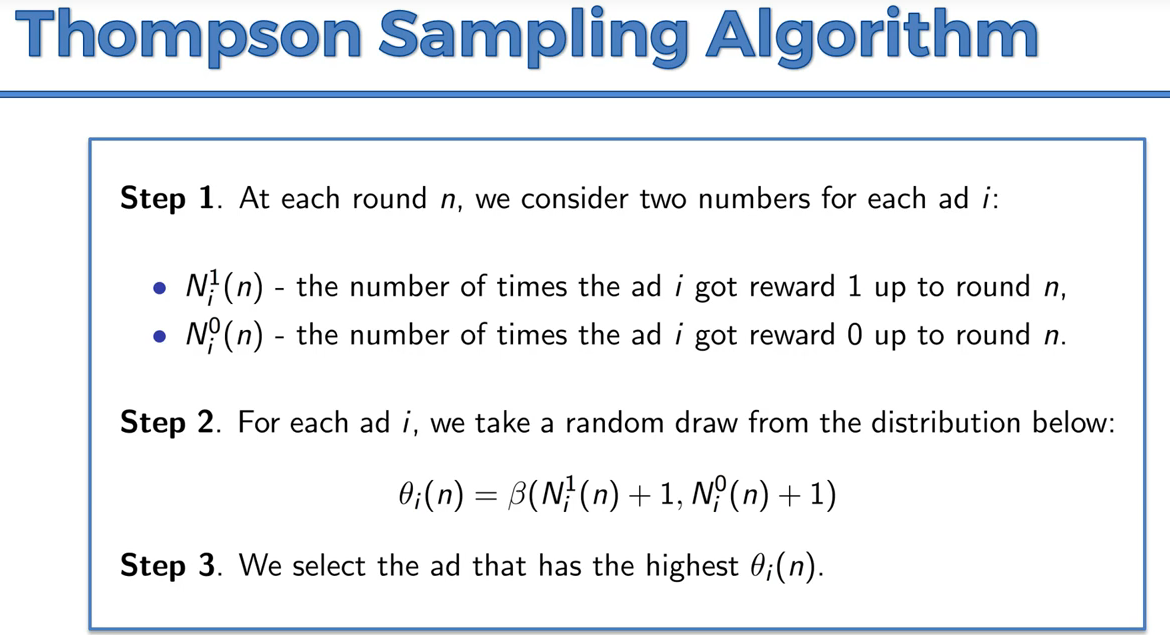


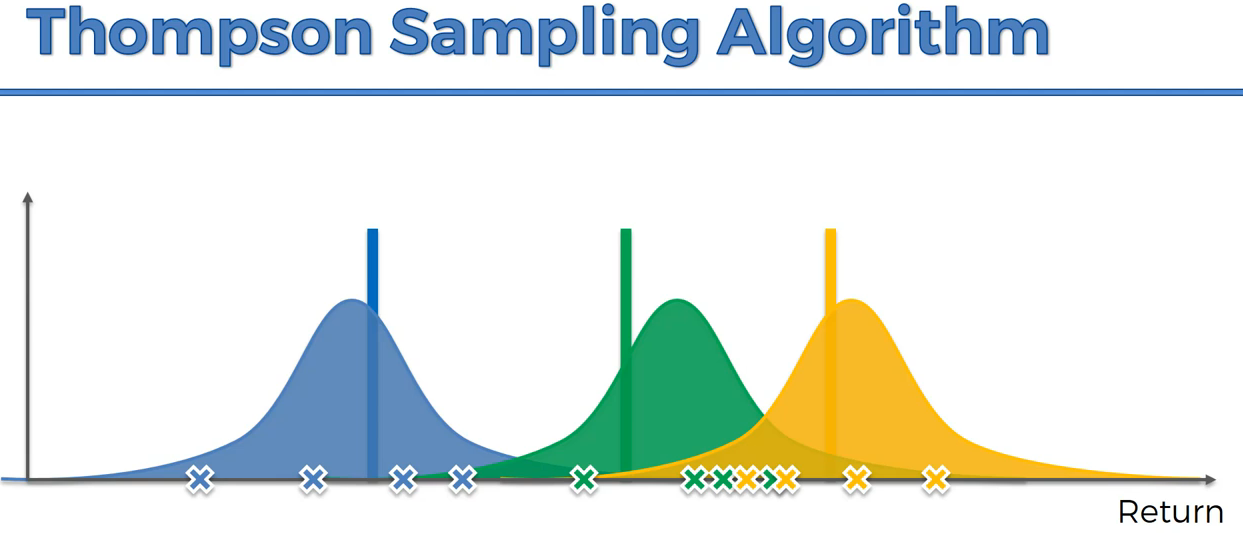


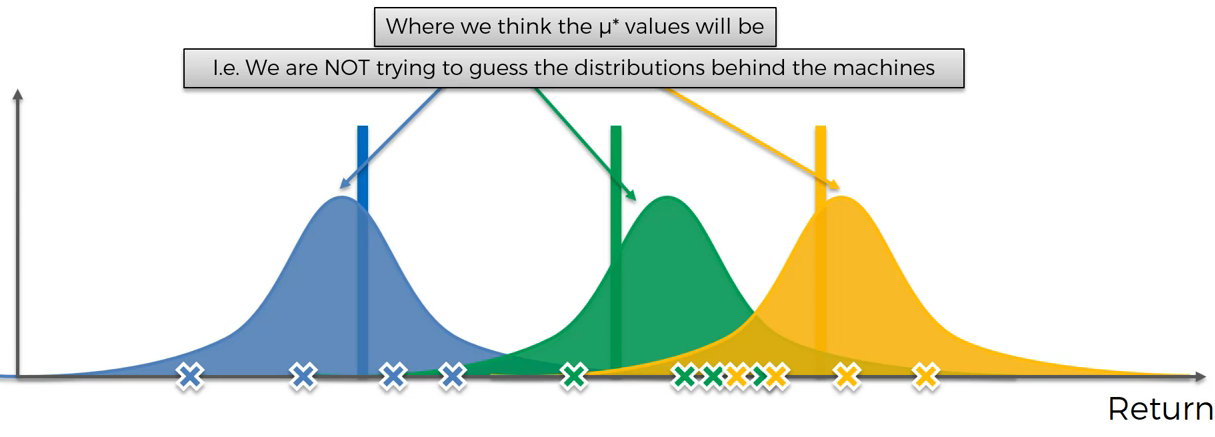
Out of 10000 people only 1174 people has clicked the ad.

Thompson Sampling









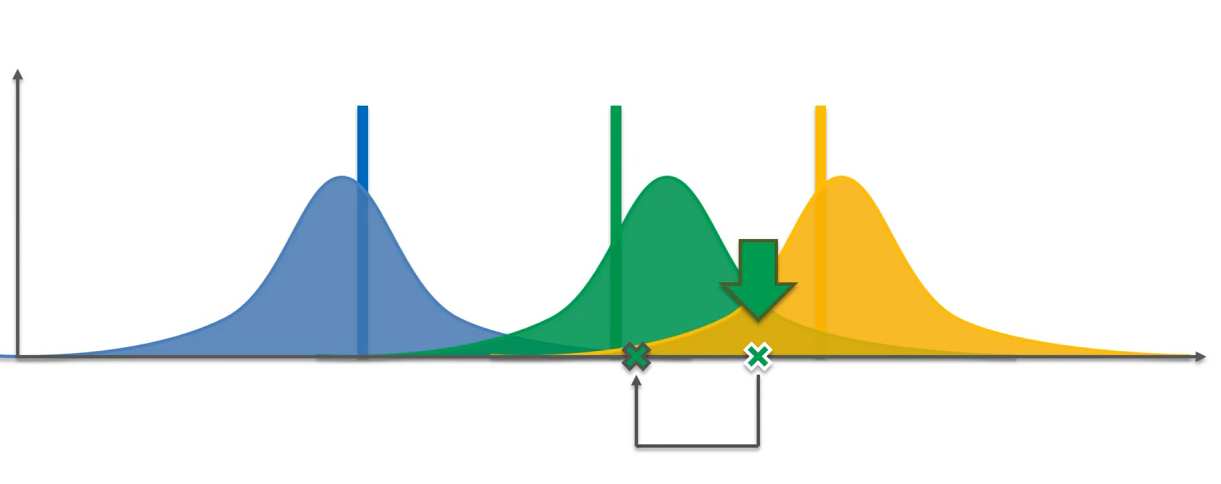
Algorithm will create the distribution based upon the points on the x-axis.



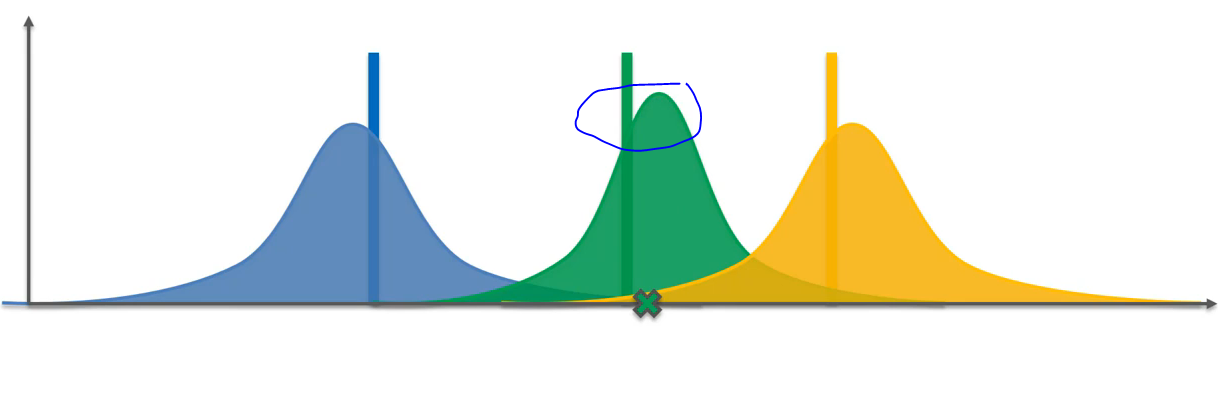
Machine will pull a value from each distribution.

Green has highest expected return based on the x-axis.

It will adjust the point.



Everytime we add a new information, our distribution will change.



Round: 2



We pick the yellow since the return is high.

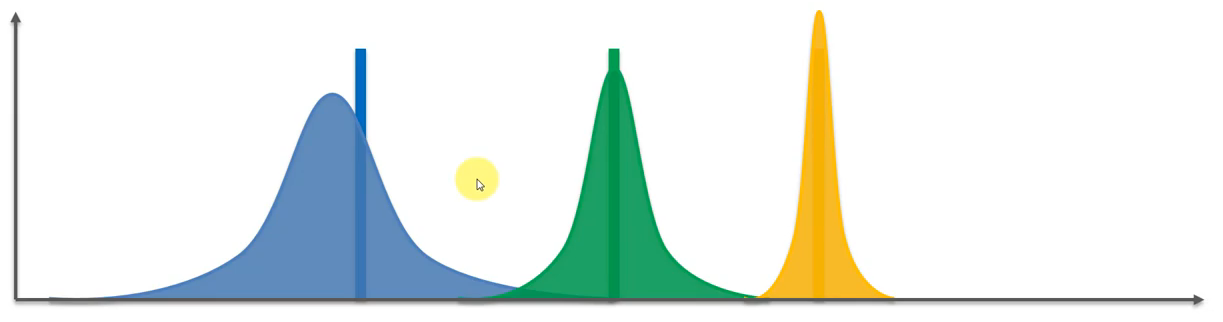


Our point will move and the distribution too.

**Round 3:**

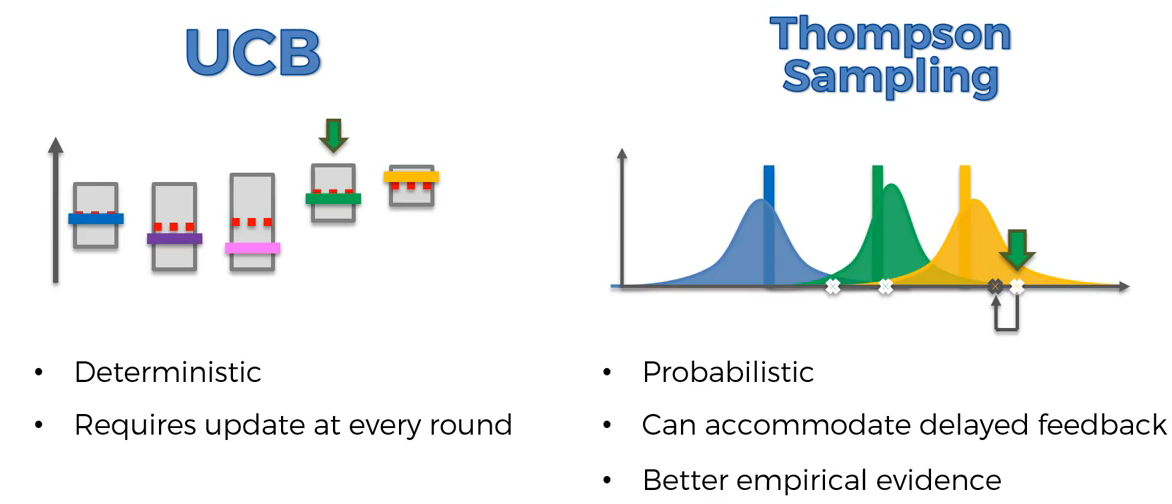


**End:**



Yellow is the best.





**UCB:**

UCB is straight forward(deterministic)

After every computation,it should be updated

**Thompson**:

It will detemine probability based on the distribution area

Delayed in feedback wont be a problem.ie the update will done based on the **BATCH**

Thompson works better and the output is better.