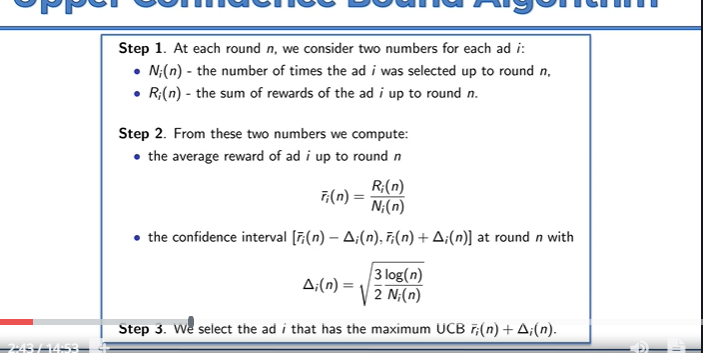
Reinforcement Learning (a.k.a online learning, interactive learning), eg train the dog to walk, if the action is correct, give it a record, and vice-versa.

* Multi-armed bandit problem: all slot machines have different win rate distributions. The goal is to figure out which machine gives you the best winning distribution.
* The purpose is to find the best choice in the shortest time

Upper Confidence Bounce



Summary:

1. We assume all machines have their distribution starts at the same level
2. We construct a confidence band, that could ensure the value will fall within this range. We did this with a few trial runs at the start.
3. Then, we pick one of them (or possibly the one with the highest bound)
4. We run that machine. If the result is not positive, then we reduce the value, or shorten the confidence bound for that machine.
5. Then, we find the machine with the highest bound again. If it is positive, increase the value, and shorten the confidence bound for that machine.