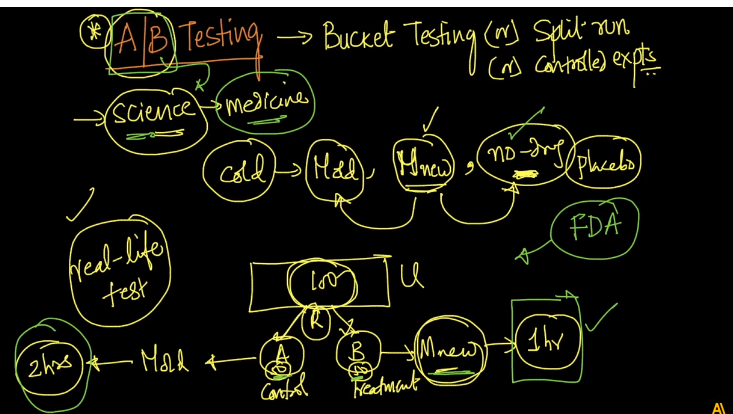
AB testing / Bucket Testing / Split run / Controlled experiments:

Let’s take an example to understand basic concept of AB testing.

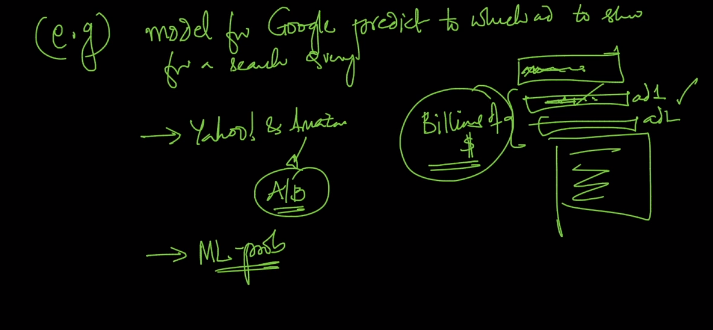
Suppose there is some existing drug for any problem and another company produced another drug and says it is better than old one.

So govt. body before approving it will test it one number of random patients i.e. suppose 100 are given old drug and 100 are given new drug.

And say drug 1 took 2hrs to recover and drug 2 took 1hr to recover than second company can claim that the new drug can help you recover in half the time than previous drug.



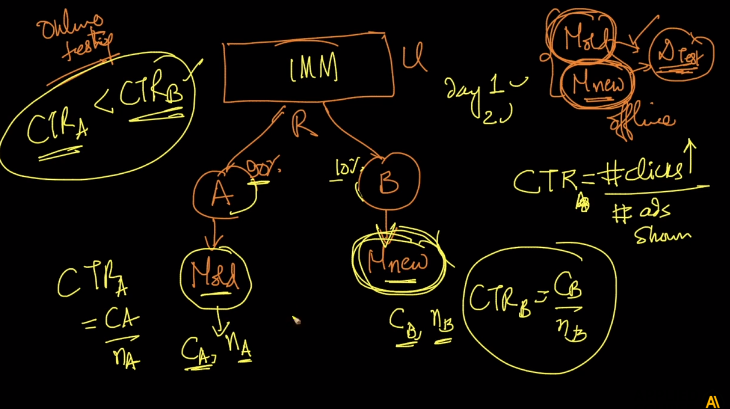
Now how it works in ML



The problem is depicted in above image.

So the solution to this is say we have a new model which in offline testing working well than existing model.

Than to test it in real world or production we split our traffic in say 90:10 or 95:5 ratio and calculate CTR i.e. #clicks/#ads shown for both models.  
if CTR(A) <CTR(B) than it means that new model is working well.



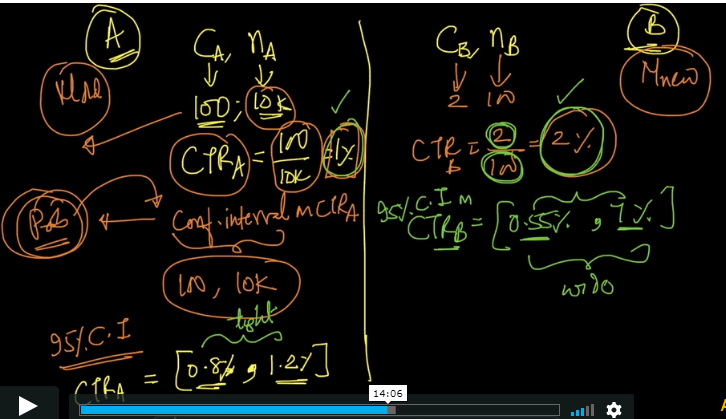
But we don’t just deploy it for 100% traffic but we increase the ration say 80:20 or 75:25.

And again calculate CTR(A) and CTR(B).

So now till the time we are extremely sure we repeat the process and we are sure we will send 100% traffic to new model.

But how we compare CTR values which is not as straightforward as it seems to be.

Because we are posting much more ads in model A and getting some clicks and posting few ads on model B and getting some clicks.



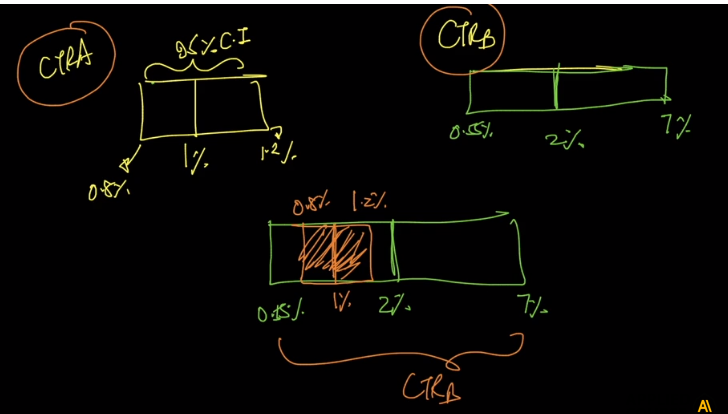
For example 10k ads posted and got 100 clicks for model A than CTR of A is 1% and 100 ads posted and 2 clicks then CTR(B) is 2%.

But comparing this values is not correct.

So now we calculate Confidence Interval(CI) and compare them. So when we compute CTR using CI we get say [0.8%,1.2%] which is quite a tough split.

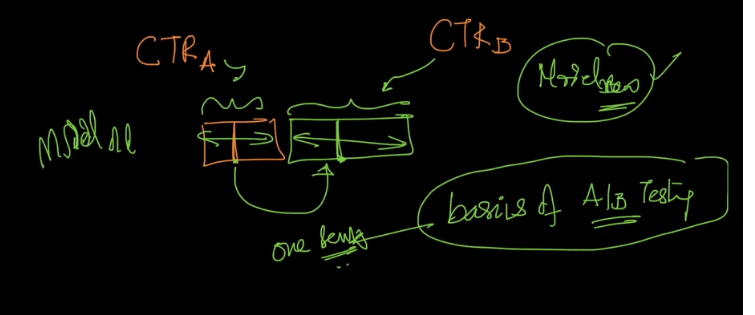
But if we calculate CTR(B) using CI we will get something like [0.55%,7%] which is very wide window.

When we draw CI in blocks it will somewhat look like below



And if your blocks are overlapping when we draw them one top of the other than you can’t be sure which model performs better.

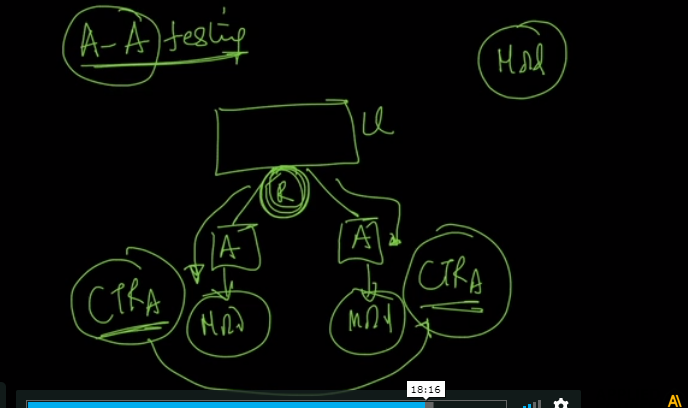
SO to make sure which models performance is better.



If 95% CI does not overlap then we can determine which model performs better.

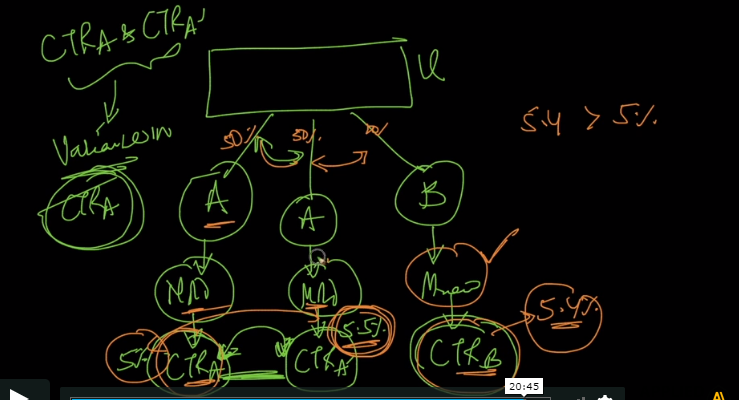
This is all just basic overview of what AB testing is and there is still whole lot of concept about AB testing.

Now let’s see another concept i.e. A-A testing.



Say you split your users in two groups and use model A for both and check change in CTR values when there is slight change in data.

Now how people do it is they split users into 3 sets A-A-B.



So know suppose you get CTR(A) is 5% and CTR(B) is 5.4% but CTR(A[dash]) is say 5.5% so this means that new model isn’t doing anything great but it is just because of variance of data.

And this happens few times and this sort of testing is called A-A-B testing.

**Comments:**

