**Deployment Patterns**

**When deployment number of use cases are there**

Deployment vs maintenance

1. New product : speech recognition (small traffic and gradually increase)
2. Automated deployment (Scratch detection)
3. Replace previous ML System

Key Ideas

1. In other words, rather than sending tons of traffic to a maybe not fully proven learning algorithm, you may send it only a small amount of traffic and monitor it and then ramp up the percentage or amount of traffic.
2. you see a few times is rollback. Meaning that if for some reason the algorithm isn't working, it's nice if you can revert back to the previous system if indeed there was an earlier system

Example

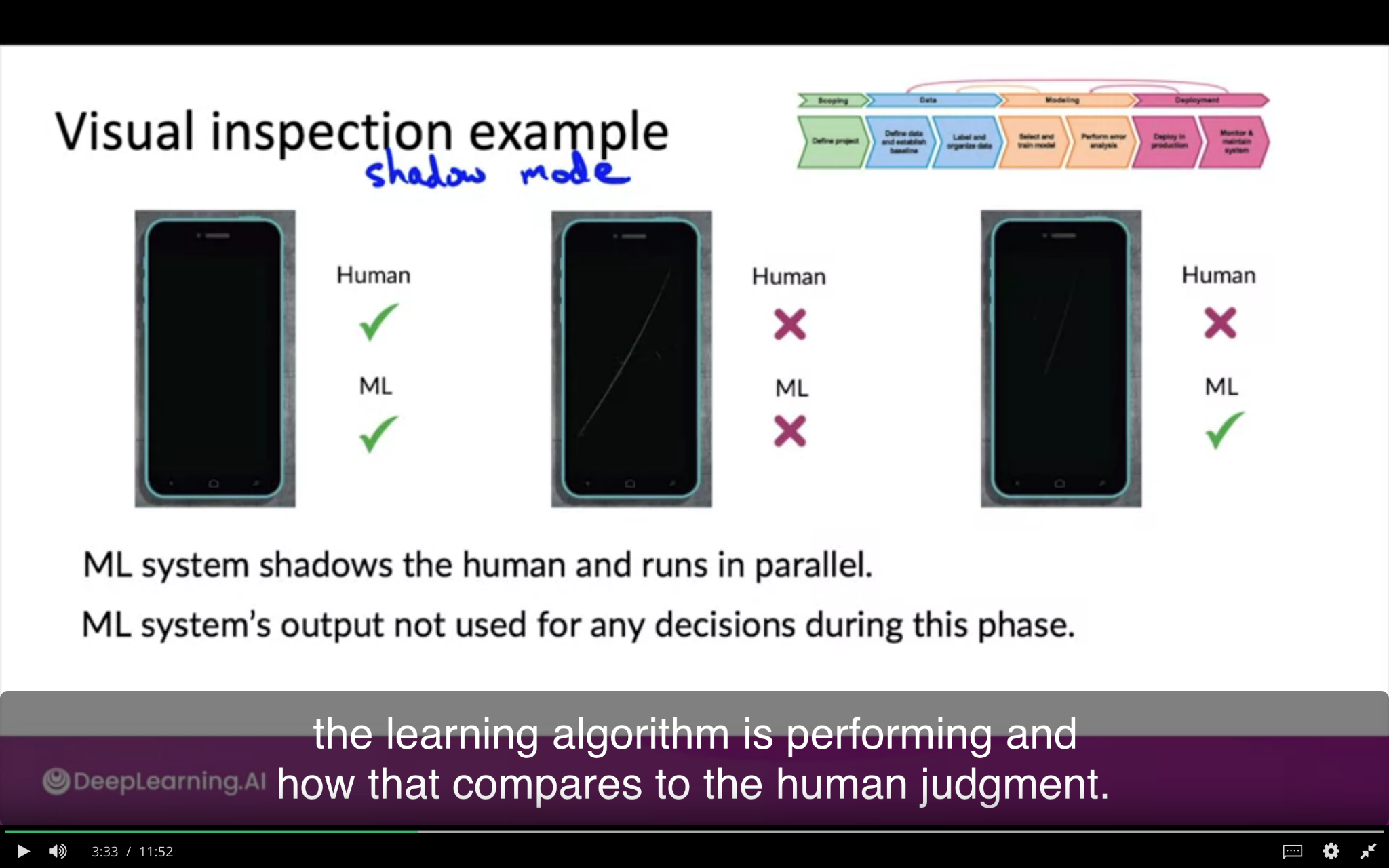
Smart phone defect detection

1. **Shadow Mode :**

ML system shadows the human and runs in parallel

MLs output not used for any decisions

The use cases are shown below.

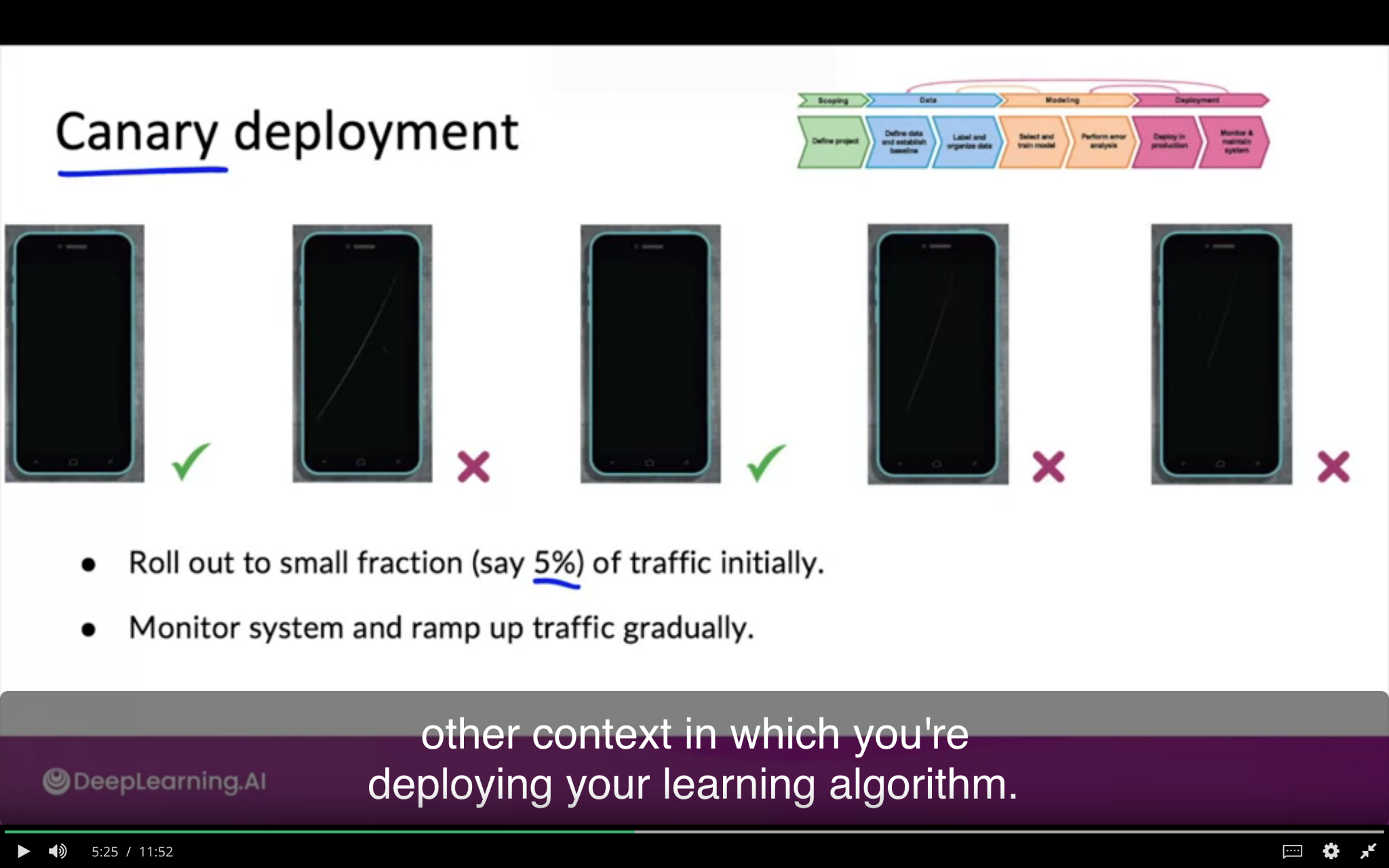
****

In this case you gradually replace machine to decide instead of human to make decisions in future.

1. **Canary Deployment:**

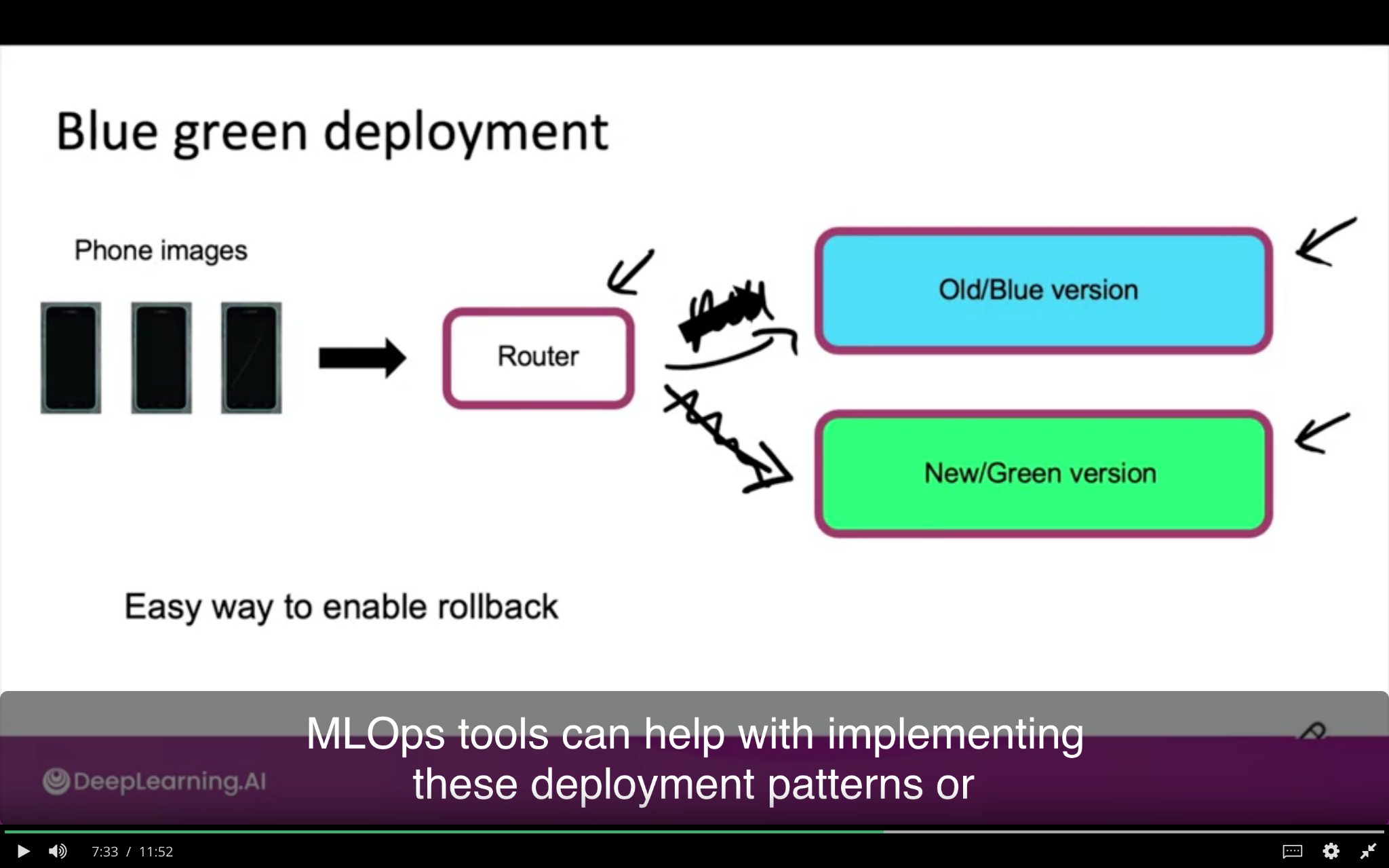
Select the traffic percentage may be 5 % and start let the algorithm make the real decisions. Running it on the small percentage of traffic. If algorithm makes mistake it will affect the small fraction

Give opportunity to monitor and ramp up traffic when you have confidence **Used when you have large consequences**



1. **Blue-Green Deployment**

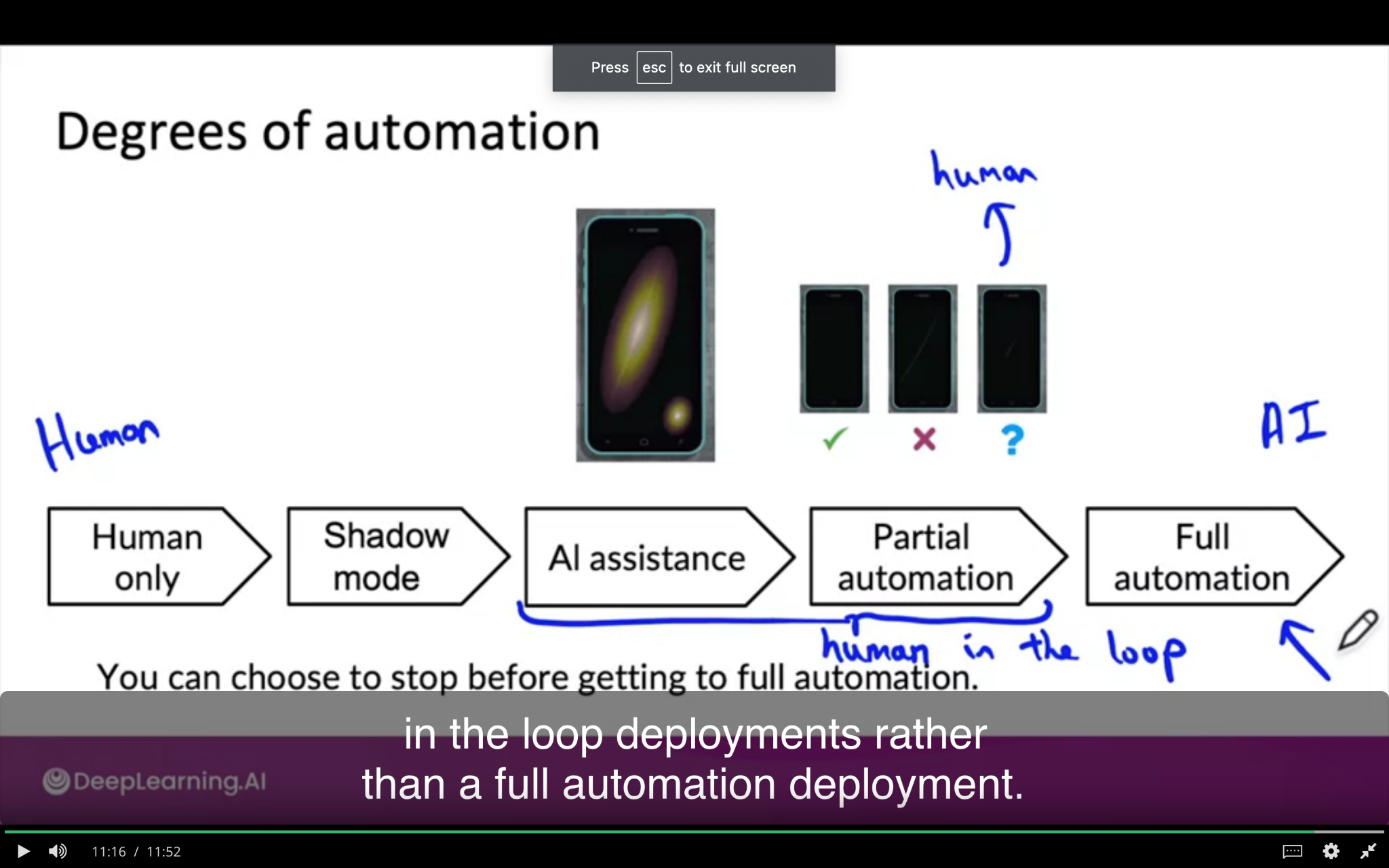
Camera software collecting phone picture - send to piece of software - old version is blue and new one is green. In blue green - the router switch to new version suddenly - old prediction to new prediction service . The advantage is **easy way to rollback (new to old one).**

****

**Note : ML Ops Tools help to Implement the patterns**

**How to Deploy:**

Think about degree of Automation - gradually increase the automation first it is human only and **shadow mode** then **AI assistance** and **Partial automation** where algo not sure about wether to say yes or not - so send to human

****