

1. What are 3 advantages of deploying using Model Serving methods Vs. deploying on GitHub Pages or HuggingFace for free?

- Simplifies integration with other technologies and organizational processes
- Model deployment and release is very flexible
- Model serving layer can be scaled out independently from the main application.
- Since models are being served in separate microservices, adding additional models is adding new containers/services that can be scaled up separately.
- Online monitoring and validation can happen per model in the rollout process.

2. What is ML model deployment?

- Machine learning model deployment is **the process of placing a finished machine learning model into a live environment where it can be used for its intended purpose**. Models can be deployed in a wide range of environments, and they are often integrated with apps through an API so they can be accessed by end users.

3. What is Causal Inference and How Does It Work?

- Despite the hype around AI, most Machine Learning (ML)-based projects focus on predicting outcomes rather than understanding causality. When humans rationalize the world, we often think in terms of cause and effect. If we understand why something happened, we can change our behavior to improve future outcomes. Causal inference is a statistical tool that enables our AI and machine learning algorithms to reason in similar ways.

How does Causal inference work?

- The gold standard for inferring causal effects is randomized controlled trials (RCTs) or A/B tests. In RCTs, we can split a population of individuals into two groups: treatment and control, administering treatment to one group and nothing (or a placebo) to the other and measuring the outcome of both groups. Assuming that the treatment and control groups aren't too dissimilar, we can infer whether the treatment was effective based on the difference in outcome between the two groups.

4. What is serverless deployment and how it's compared with deployment on server?

- Serverless is *a cloud-native development model that allows developers to build and run applications without having to manage servers*.