Week 13 - Model Serving Rubrics

What is the definition of a serverless application?

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

  There are still servers in serverless, but they are abstracted away from app development. A cloud provider handles the routine work of provisioning, maintaining, and scaling the server infrastructure. Developers can simply package their code in containers for deployment.

  Once deployed, serverless apps respond to demand and automatically scale up and down as needed. Serverless offerings from public cloud providers are usually metered on-demand through an event-driven execution model. As a result, when a serverless function is sitting idle, it doesn't cost anything.

What is the definition of model serving and what are the two types?

- Model serving is to host machine-learning models (on the cloud or on premises) and to make their functions available via API.
- The two types of model serving are online serving and scheduled batch serving. Online serving is when a model is hosted behind an API endpoint that can be called by other applications. Typically the API itself uses either REST or GRPC. Scheduled batch serving is a service which when called runs inference on a static set of data. Jobs can be scheduled on a recurring basis or on-demand.

What are 3 advantages of deploying using Model Serving methods vs. deploying on Github Pages or HuggingFace for free?

- Some of the advantages of deploying using Model Serving methods vs. deploying on free sites such as github or huggingface are security, hardware limitation, scalability. Free services might not have the best security for proprietary information, best hardware, and might not be suitable for scalability for large traffic.

What Is Machine Learning Inference? How Does Machine Learning Inference Work? Please walk us through an example?

- Machine learning inference is the process of running data points into a machine learning model to calculate an output such as a single numerical score. In machine learning inference, the data sources are typically a system that captures the live data from the mechanism that generates the data. The host system for the machine learning model accepts data from the data sources and inputs the data into the machine learning model. The data destinations are where the host system should deliver the output score from the machine learning model.