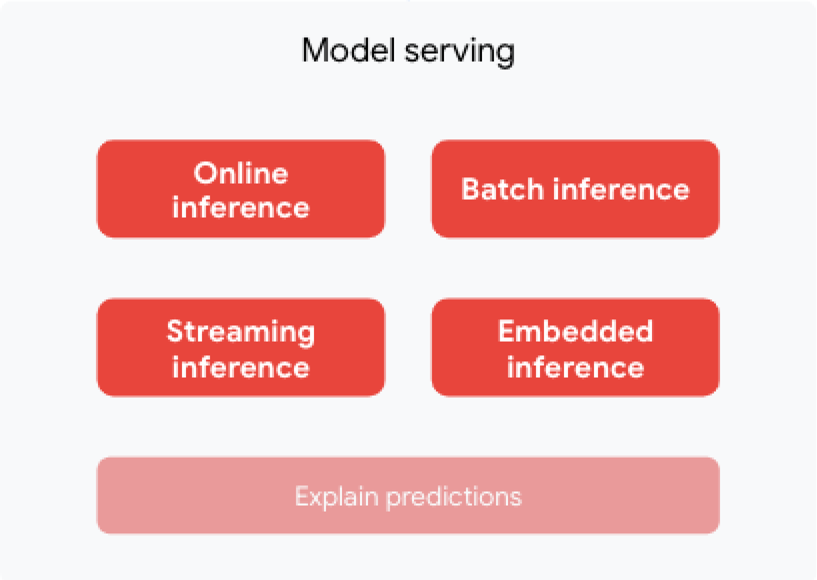
Prediction Serving Scenarios



In the upcoming few videos, we will be stepping through the various model serving scenarios. We'll learn about what each serving scenario is, what are some use cases of this scenario, what are the metrics that matter for such a scenario, as well as what are some pros and cons of the scenario. Briefly, here is a description of each scenario:

- **Online inference** generates machine learning predictions in real-time. It's sometimes called real-time or dynamic inference.
- **Batch inference** is the process of making predictions from a set of data. Batch jobs are usually generated on a regular basis (e.g. hourly, daily).
- **Streaming inference** is when you are working with data that is continuously flowing through, a good example of this is time-series data that
- **Embedded inference** is when you shrink wrap your model into an application. The application and the model are tightly coupled together for deployments in embedded devices like smartphones.