# Tools for distributed model serving

Model serving is making predictions using a ML model.

Some of the popular tools for that include:

* Ray
* Ray Cluster Autoscaler & Operator – Makes it easier to use Ray on Kubernetes