HTTP API and Model Serving

REST API: Standard way to serve ML models over HTTP

Client Request
POST /predict
JSON data

→

ML Model
Processing
Inference

→

Server Response
JSON
predictions

Popular Frameworks

- **Flask**
- **FastAPI**
- **TensorFlow Serving**

✓ Benefits

- Language-agnostic
- Easy integration
- Scalable with load balancers

⚠ Considerations

- Latency (~10-100ms)
- Throughput (requests/sec)
- Batching strategies

Production Tools
Docker + Kubernetes for deployment and scaling

Monitoring
Track API response time, error rates, model accuracy drift