

# Model Deployment Lab

---

**CS 203: Software Tools and Techniques for AI**

Duration: 3 hours

# Lab Overview

---

- Quantize models for efficiency
- Build FastAPI prediction service
- Dockerize the application
- Test performance and latency
- Deploy and monitor

## Structure:

- Part 1: Model Optimization (45 min)
- Part 2: API Development (60 min)
- Part 3: Docker & Deployment (60 min)
- Part 4: Monitoring (15 min)

# Setup

---

```
pip install fastapi uvicorn torch onnx onnxruntime scikit-learn joblib locust
```

# Exercise 1: Model Quantization

---

Create and quantize a model, then compare sizes and performance.

## Exercise 2: FastAPI Service

---

Build a complete prediction API with health checks, version management, and error handling.

# Exercise 3: Docker Deployment

---

Containerize the service and deploy it.

## Exercise 4: Load Testing

---

Test the API under load and optimize.

# Deliverables

---

- Working FastAPI service
- Docker setup
- Load test results
- Monitoring dashboard



# Resources

---

- FastAPI: <https://fastapi.tiangolo.com/>
- Docker: <https://docs.docker.com/>
- ONNX: <https://onnx.ai/>

# Course Complete!

---

Congratulations on completing the course!