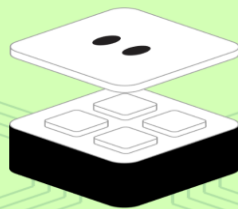


BentoML

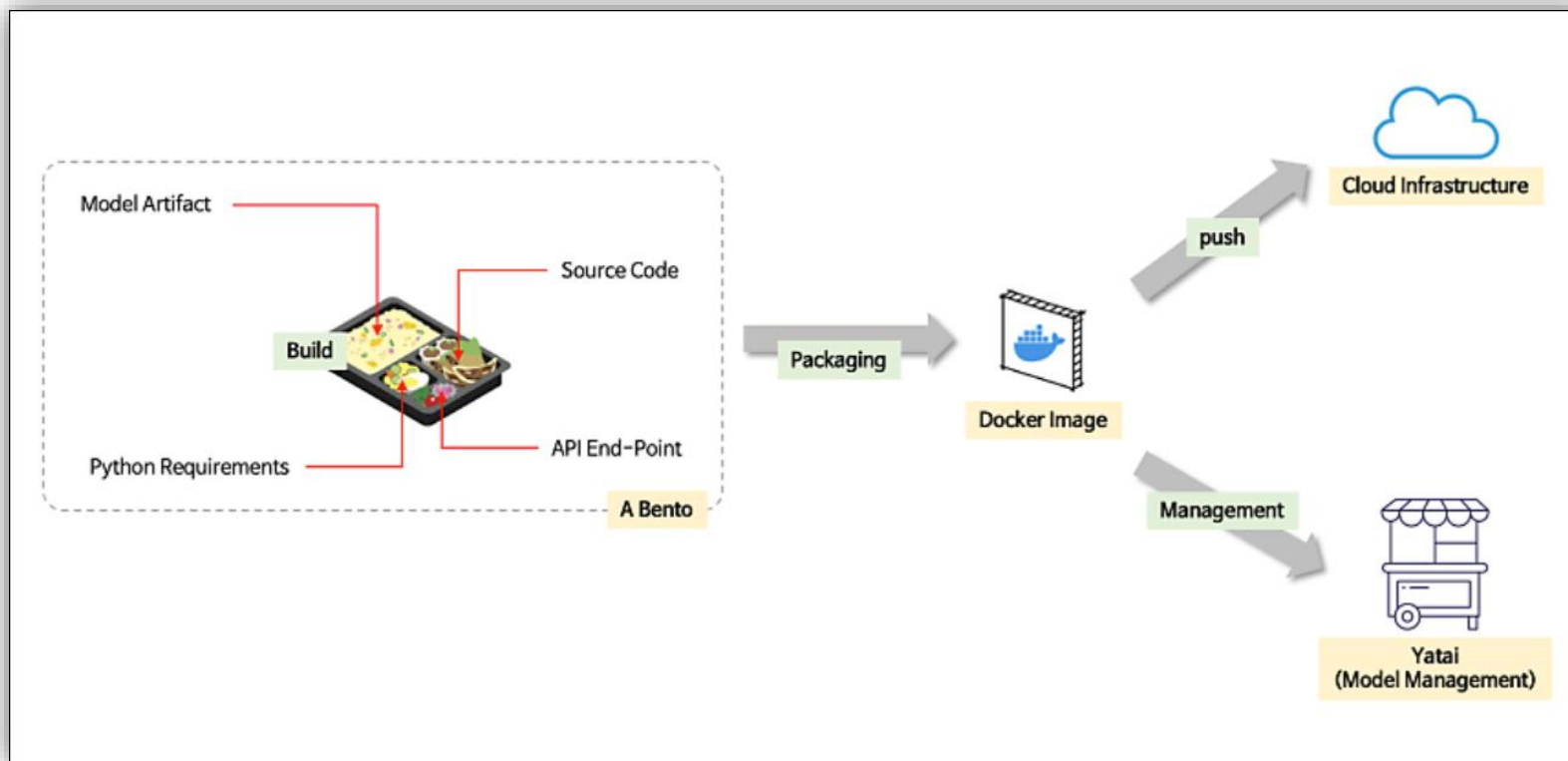
BentoML
Inference Platform for
Building Fast and
Scalable AI Systems



소프트웨어융합대학원
진혜진

■ BentoML

- AI/머신러닝 모델을 배포하고 서빙(serving)하기 위한 프레임워크



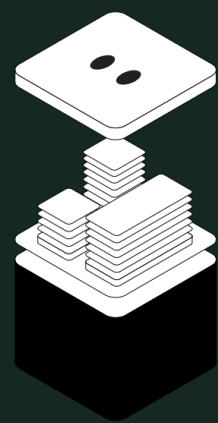
- <https://bentoml.com/>

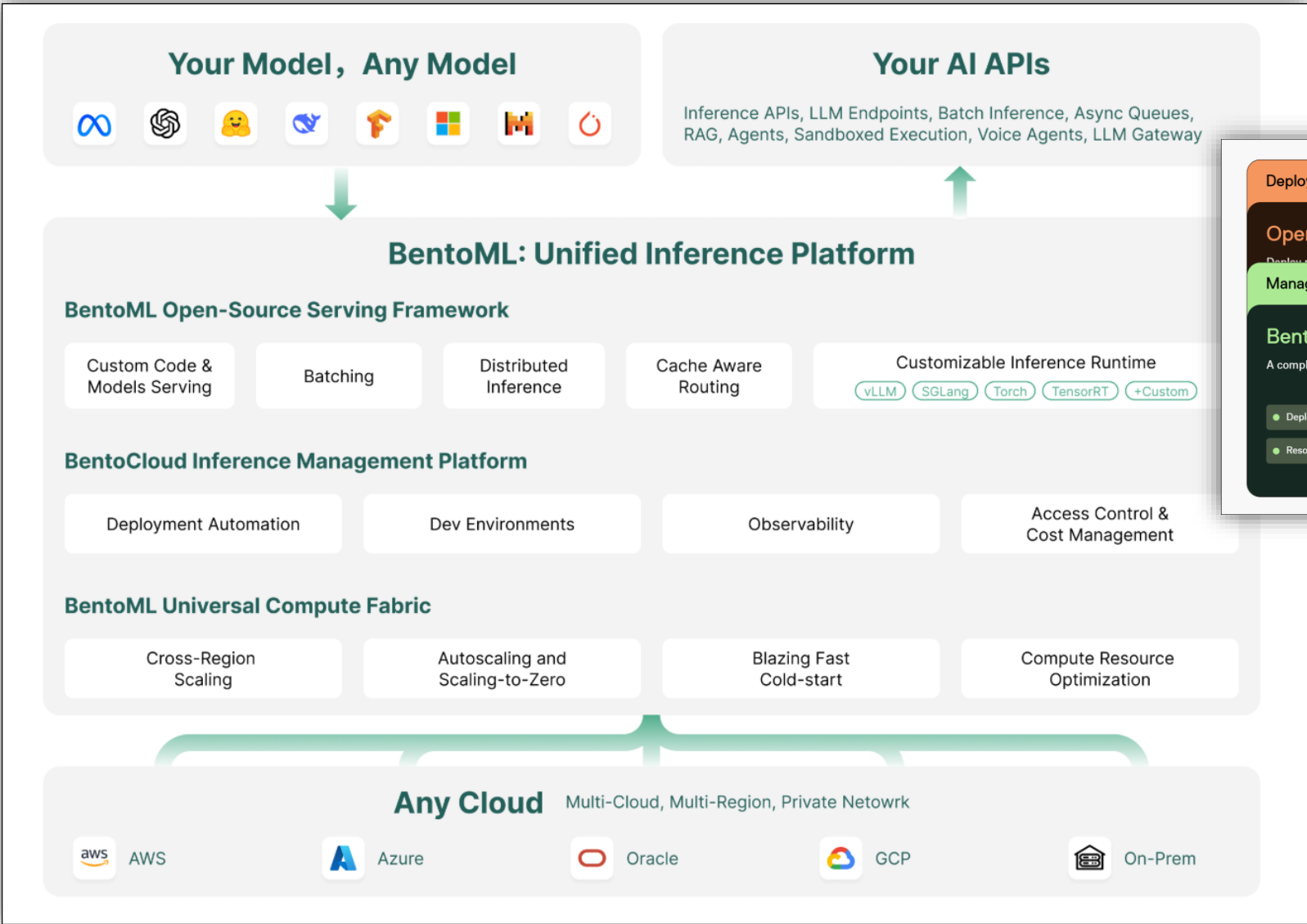
Inference **On Your Terms**

Inference Platform built for **speed and control**. Deploy any model anywhere, with tailored optimization, efficient scaling, and streamlined operations.

Start Building →

Get a Demo →





Deploy Any Model

Open Model Catalog

Deploy popular open-source models with a few clicks

Custom Models

Unified framework for packaging and deploying

Manage Inference

Bento Inference Platform

A complete platform for managing, monitoring, and optimizing AI model inference.

Deployment automation and CI/CD

Comprehensive observability

Fine-grained access control

Resource and quota tracking

Performance tuning

- Model Serving

- Batch Serving
- Online Serving

- <https://github.com/bentoml/>

- “BentoML enables users to create a machine learning powered prediction service in minutes and bridges the gap between data science and DevOps.”

■ Components of BentoML

- Model Artifacts
- BentoService
- API Functions and Adapters
- Model Management & Yatai

