

# Techtonic 2018

-  
Thu . Nov 15

-  
SAMSUNG SDS Tower  
West Campus B1F  
Magellan Hall /Pascal Hall



# 대규모 Machine Learning을 위한 Kubeflow 파헤치기

삼성SDS 이권호 프로, 이규성 프로



Techtonic 2018

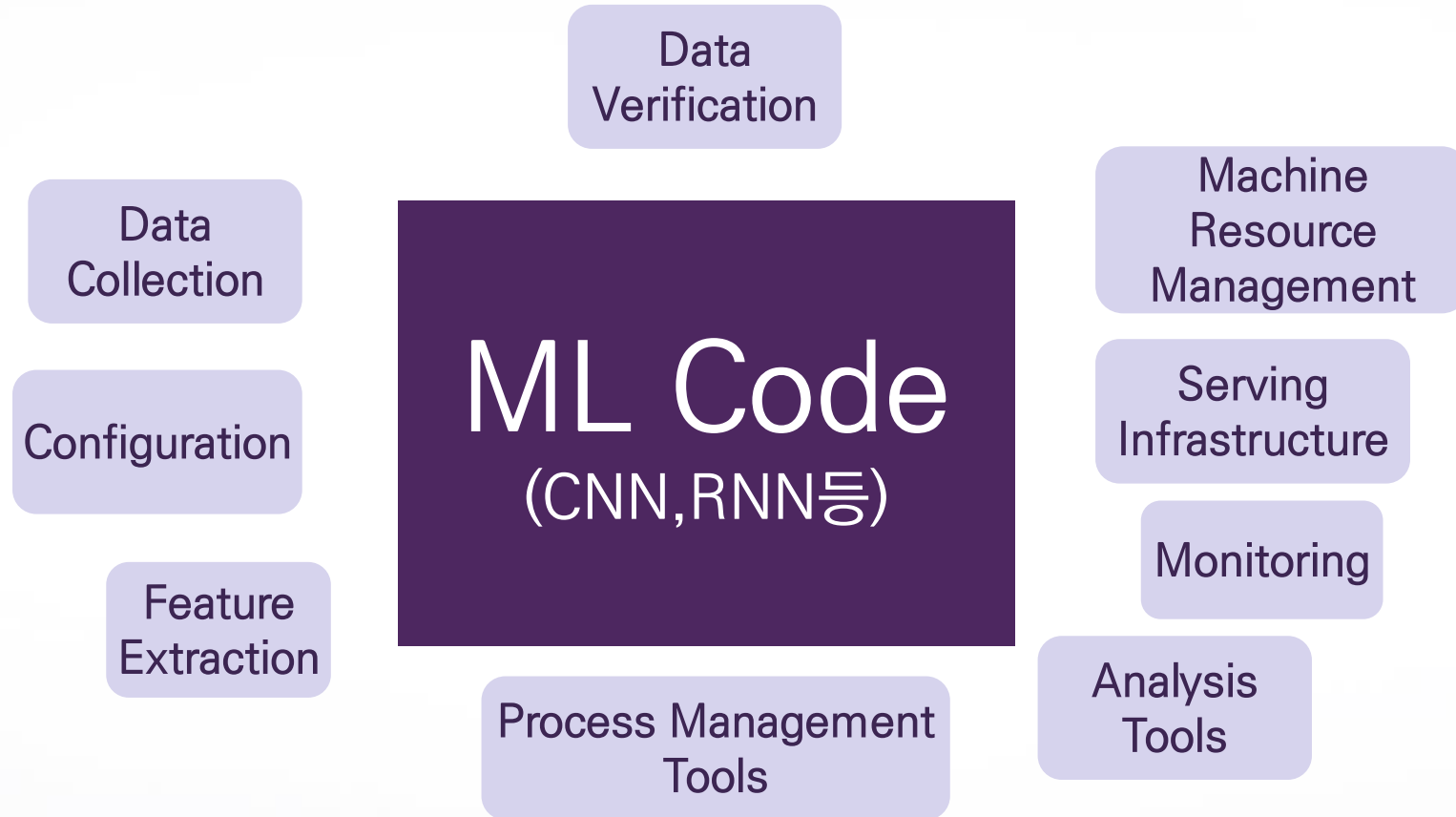
# Agenda

- Kubeflow 소개
- Demo

대규모 Machine Learning을 위한 Kubeflow 파헤치기

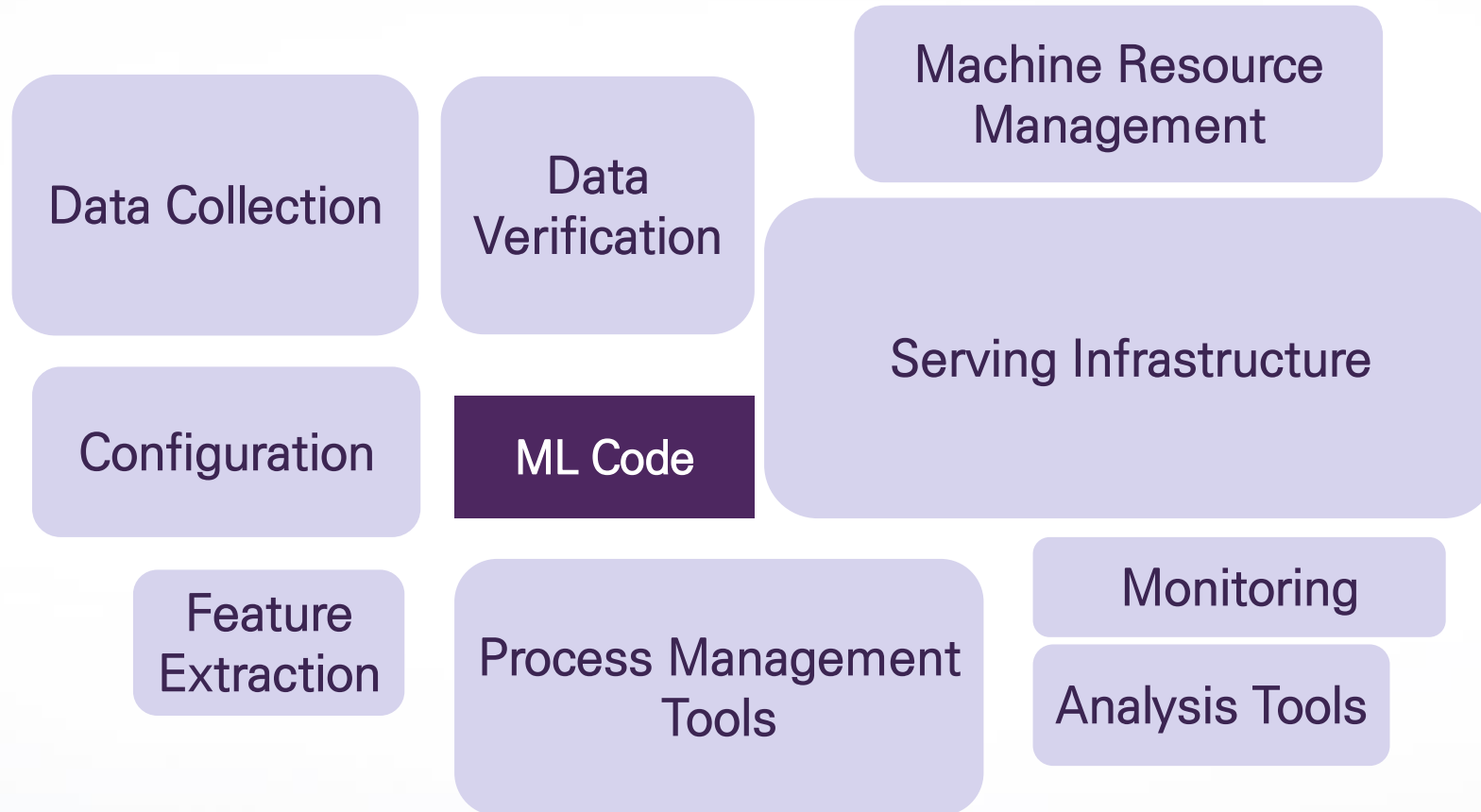
# Kubeflow 소개

# 머신러닝을 하기 전에...



<https://www.slideshare.net/matthiasfeys/running-tensorflow-in-production>

# 머신러닝을 해 본 뒤에...



<https://www.slideshare.net/matthiasfeys/running-tensorflow-in-production>

# 당면하는 3 가지 문제

유기적으로 연결이 어려워

- 데이터 준비/가공, Experiment, Training, Serving

반복되는 환경구성이 너무 많아

- Experiment
- Scalable Training
- Serving
- Multi Model / Team

시스템 확장은 어떻게 하지?

- Big Data for Training
- GPU/CPU Resource Pool
- Serving on Cloud



# Today's topic



Make it Easy for Everyone  
to Develop, Deploy and Manage Portable,  
distributed ML on Kubernetes

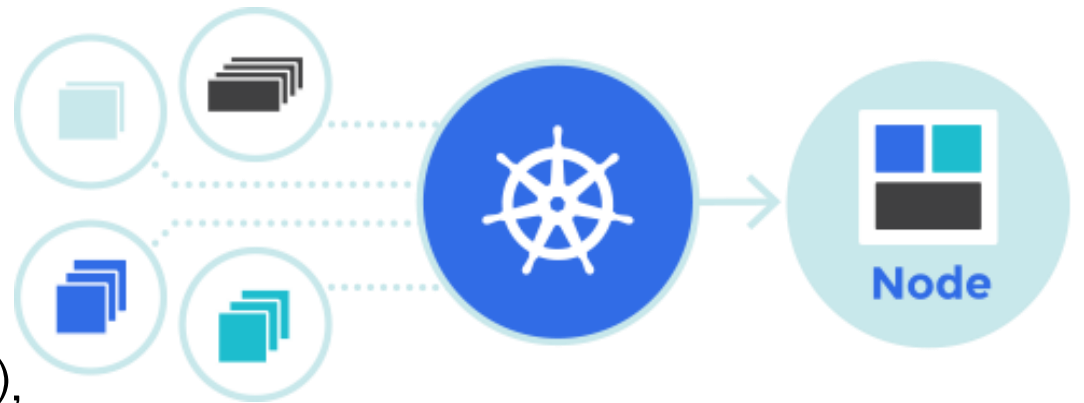
*<https://www.kubeflow.org>*



# Today's topic

## Kubernetes란?

- 여러 Node를 묶어 클러스터를 구성
- Container를 적절한 위치에 배포(Auto-placement)
- Container 자동 복구(Auto-restart)
- 필요에 따라 Container를 추가(scaling), 복제(replication), 업데이트(rolling-update), 롤백(rollback)
- <https://kubernetes.io>



# Kubeflow? 무엇인지 알아보자

## What is Kubeflow

- Kubernetes 기반으로 Machine Learning 작업을 쉽고, 확장성 있게 처리할 수 있도록 하는 프로젝트
- Bare-metal Server, Private Cloud, Public Cloud 등 다양한 환경에서 확장성 있는 머신 러닝 서비스 제공 가능

## Kubeflow mission

- 손쉽게, 지속적으로, 다양한 인프라 환경을 오가며 모델링/훈련/배포/실행 작업 가능
- 머신 러닝 구성 요소들을 Micro Service 형태로 관리 및 배포하는 환경
- 데이터, GPU, 서비스 노드 증가 등의 요청에 따른 확장성

**손이 많이 가는 반복적인 Machine Learning 환경 작업들을 시스템에게 위임!**

# Kubeflow의 구성 요소는?

- Jupyter Hub 기반 Jupyter Notebook 생성/관리
- Multi-architecture, 분산 Training 환경
- Model serving을 위한 Multi framework 지원
- AI Process 관리하기 위한 Integration 도구 제공
- Ksonnet 패키징 도구를 사용한 커스터마이징 가능



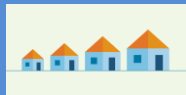
# 왜 Kubeflow를 사용해야 할까?



Composability



Portability



Scalability

# 왜 Kubeflow를 사용해야 할까?



Composability



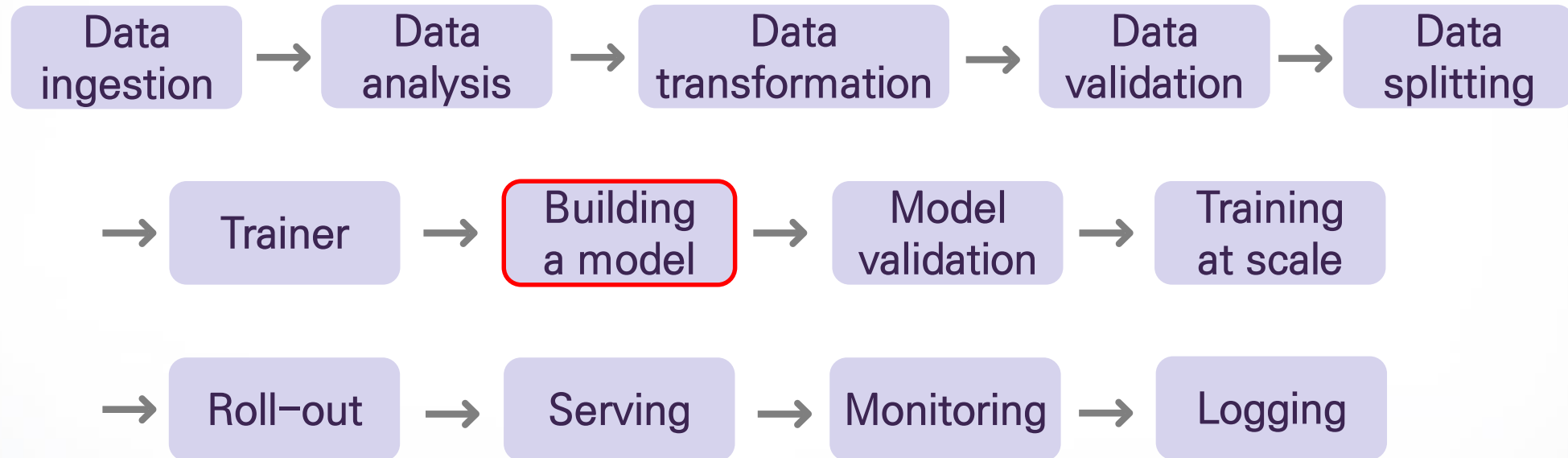
Portability



Scalability

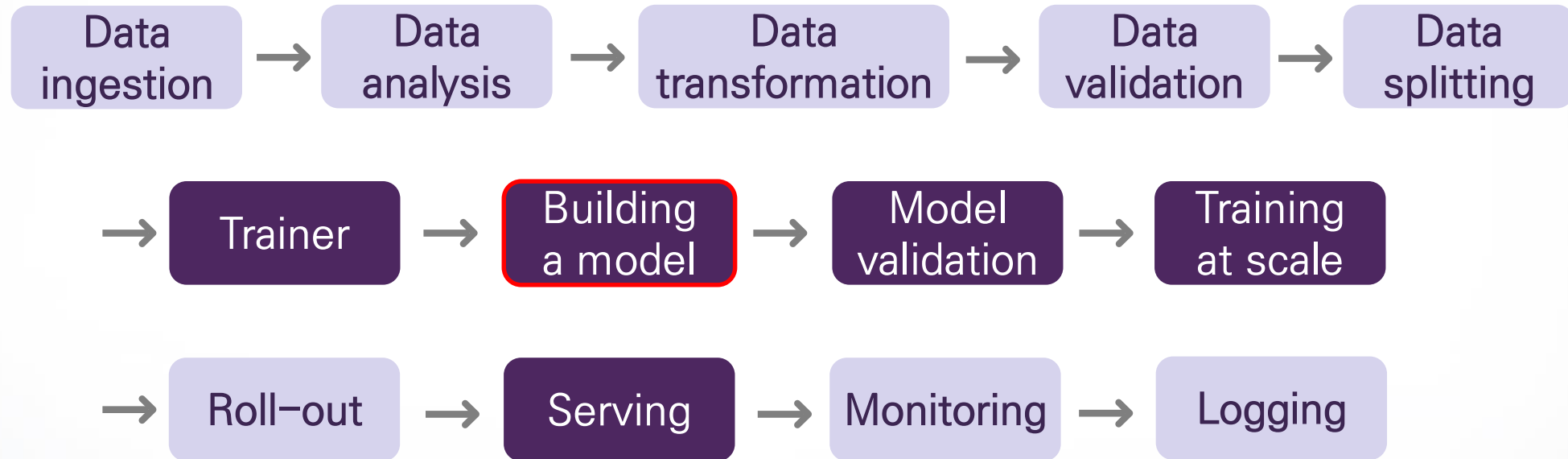
# Composability

Machine Learning 데이터 준비, Training, Serving 등의 서비스들이 결합되어 하나의 workflow로 생성되기 위한 통합 아키텍처



# Composability

Machine Learning 데이터 준비, Training, Serving 등의 서비스들이 결합되어 하나의 workflow로 생성되기 위한 통합 아키텍처



# 왜 Kubeflow를 사용해야 할까?



Composability



Portability

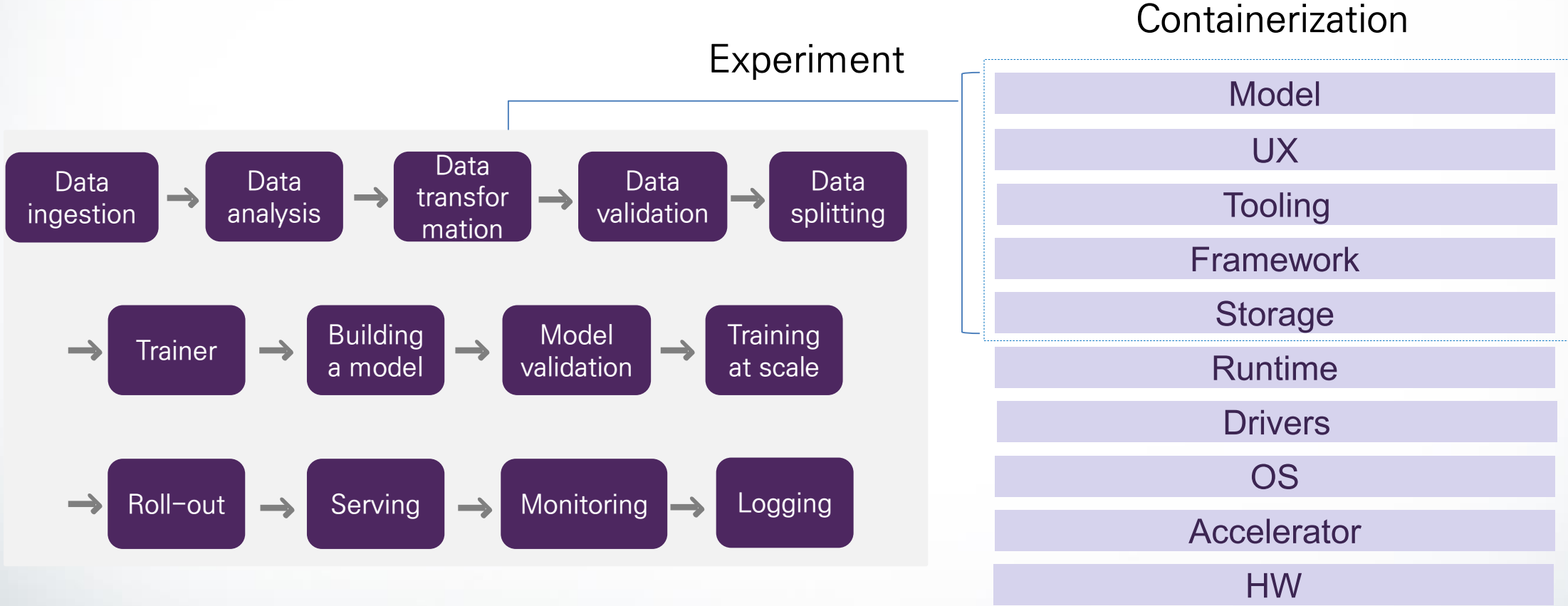


Scalability

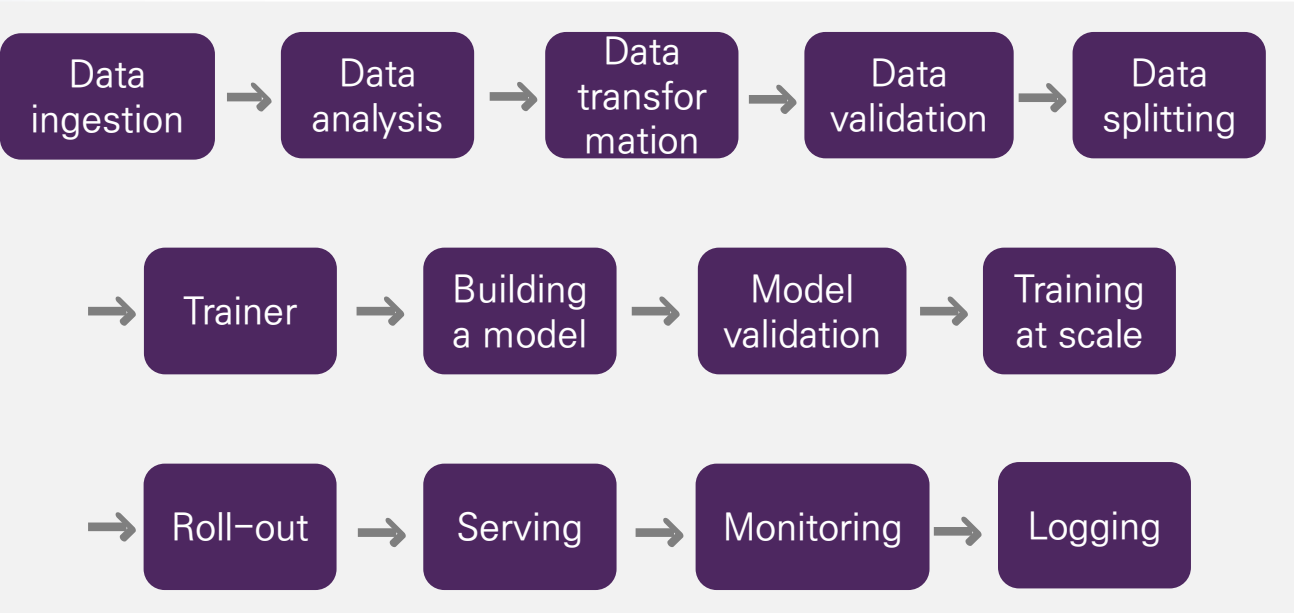


# Portability

Containerized된 Model을 사용하여 Machine Learning 수행 Workflow의 각 단계에 바로 적용 가능



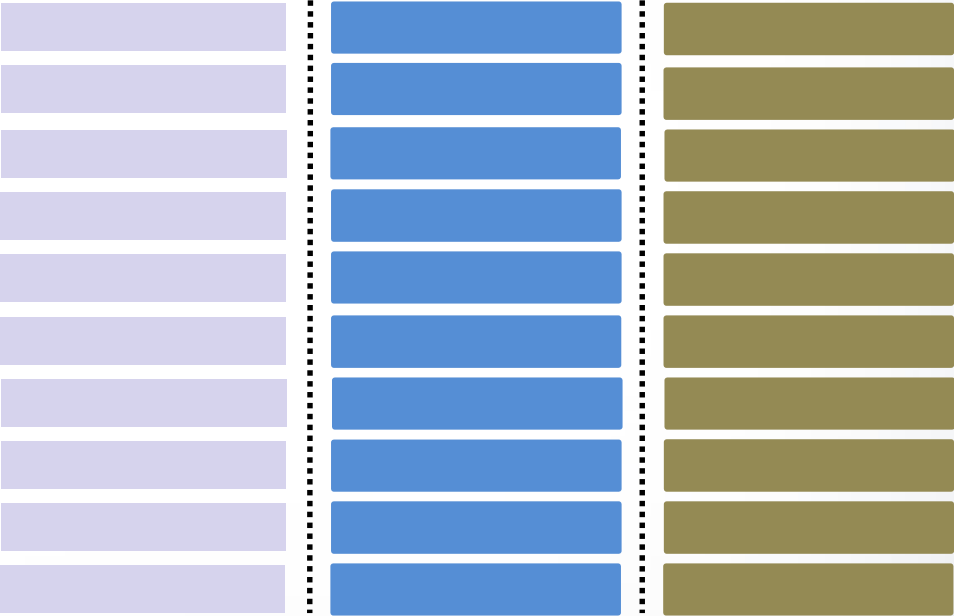
# Portability



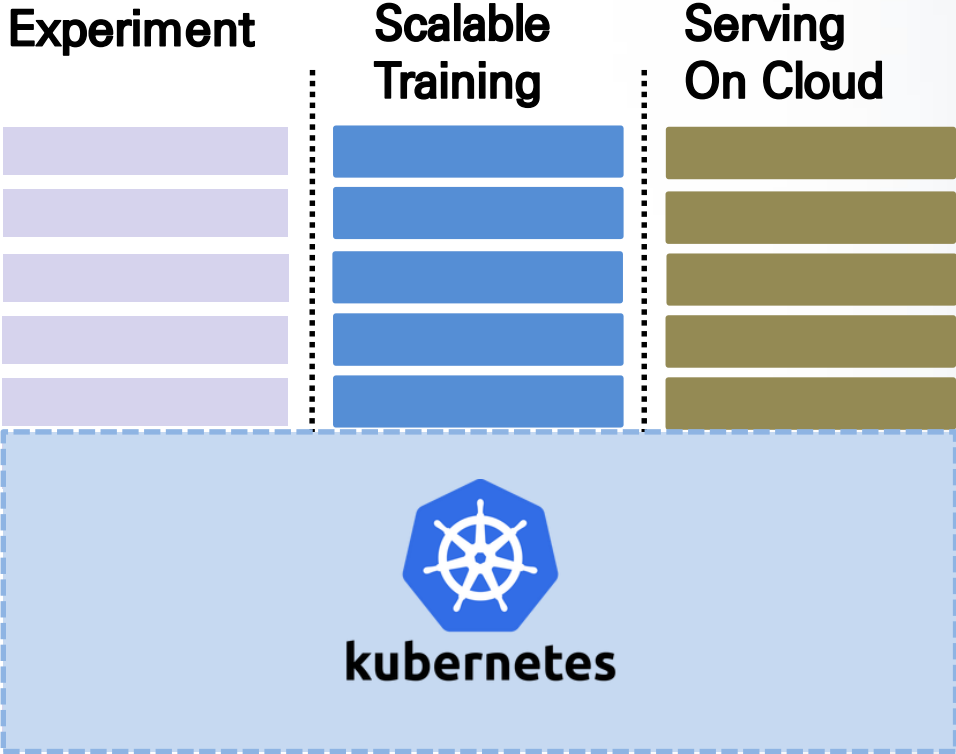
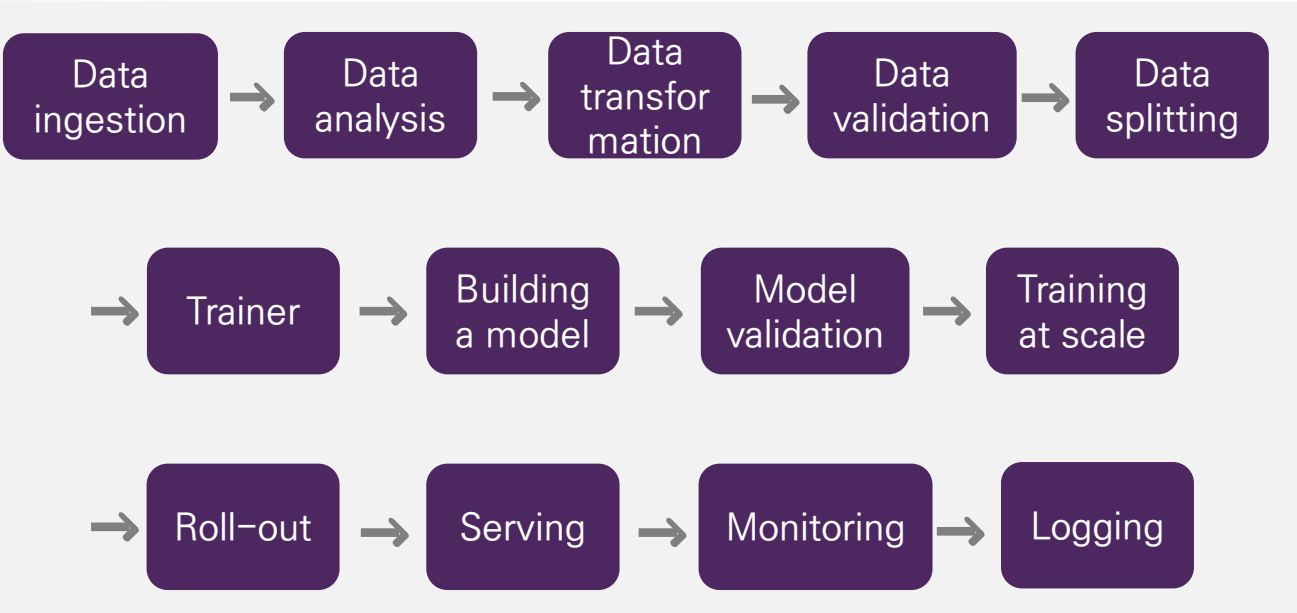
Experiment

Scalable Training

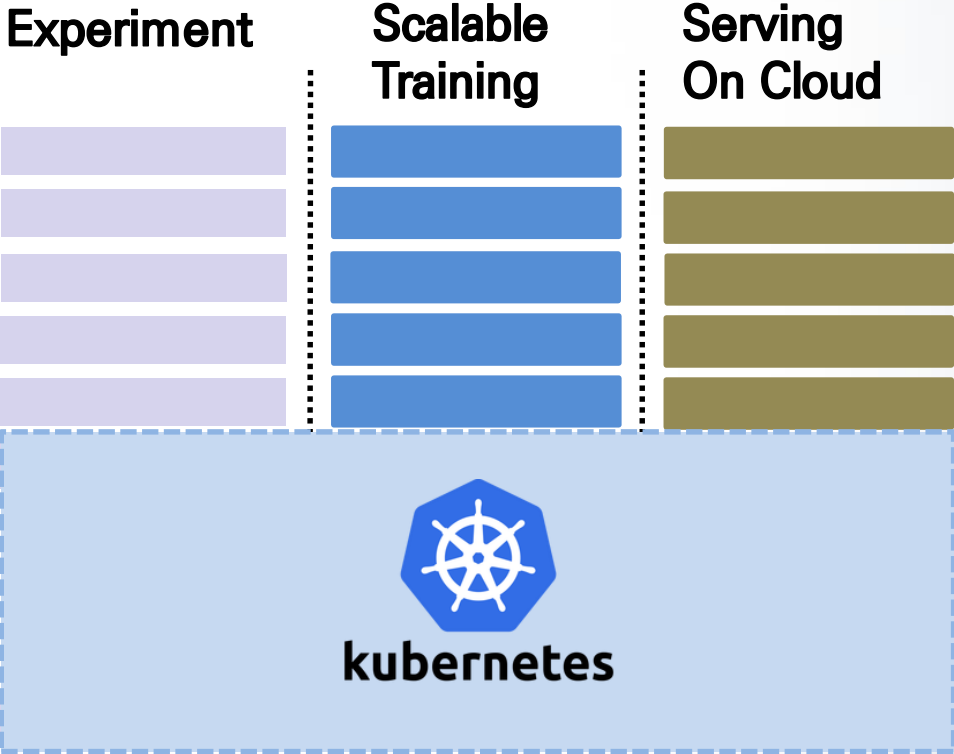
Serving On Cloud



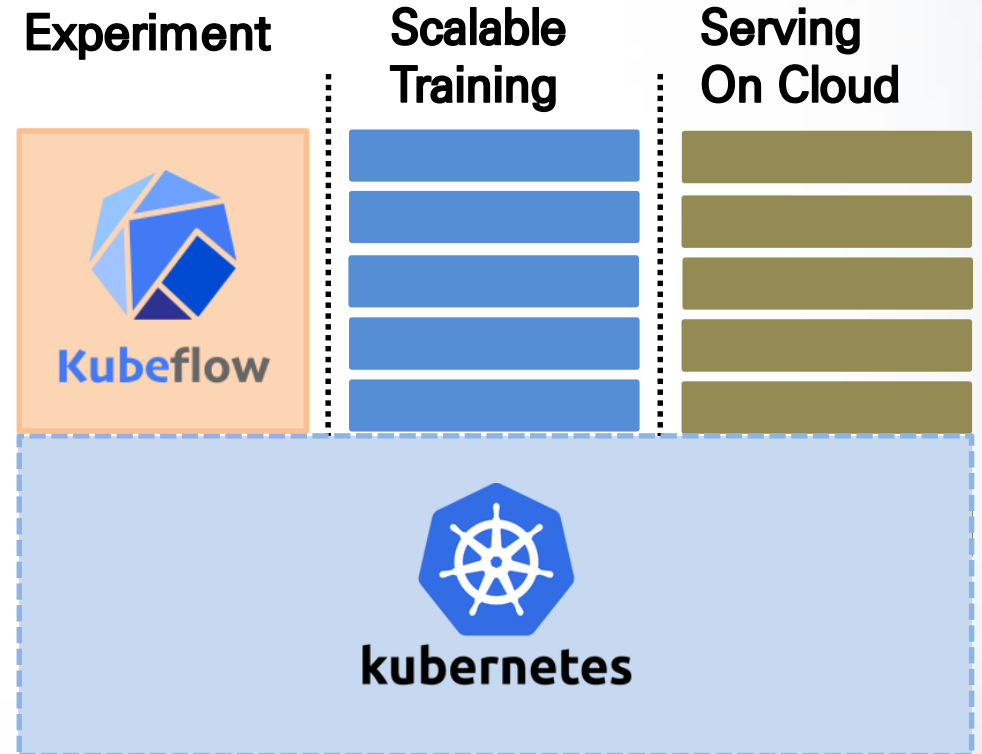
# Portability



# Portability



# Portability



# Portability



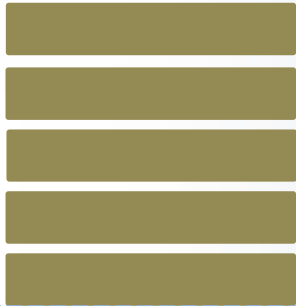
Experiment



Scalable Training



Serving On Cloud



# Portability



Experiment



Scalable  
Training



Serving  
On Cloud



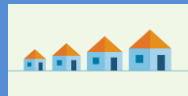
# 왜 Kubeflow를 사용해야 할까?



Composability



Portability



Scalability



# Scalability

Machine Learning의 사용이 확대 되고 점차 규모가 커지면서 Resource 사용량 증가 대응 필요

- ▶ More accelerators(GPU,TPU) & Servers
- ▶ More storage, faster networking
- ▶ More team & members
- ▶ More experiments & integrations

# SDS가 Kubeflow에 관심을 가지는 이유

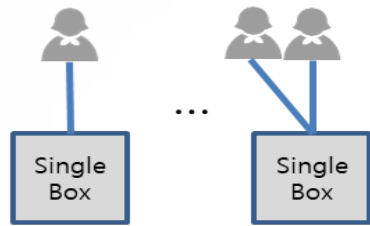
AI/ML 서비스 요구사항을 기반으로 Cloud 기반의 AI/ML 개발환경 구성 필요

Private / Public Cloud 에서 AI/ML 인프라를 제공하기 위한 요건

- ◊ AI/ML을 위한 Bigdata 저장/처리 필요
- ◊ Data process 자동화 필요
- ◊ 다수의 Data scientist를 위한 Tool 제공 필요
- ◊ Cloud Native 관리 필요
- ◊ 지속적인 AI/ML환경의 변화에 빠른 대응 필요

# AI Engineering Framework

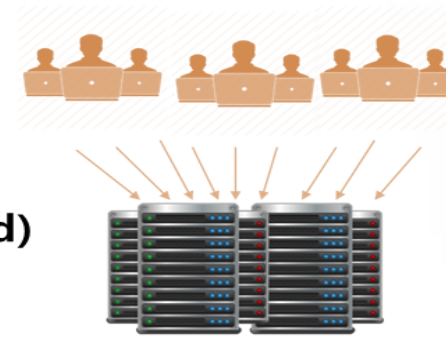
소규모 팀이  
진행하는 실험



AI 운영성



전사적 도입



- 1) 기계 학습 Life Cycle 지원 (End-to-End)
- 2) AI를 위한 Governance
- 3) Enterprise Scale의 실험/운영

## 전체 ML 라이프사이클 관리

- 개발 ~ 배포까지 One-stop AI 워크플로우 관리
- 협업환경
- Smart AI

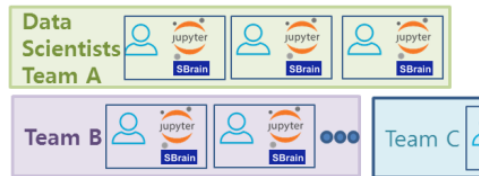
Semi-Auto Labelling

Hyper-parameter Optimization

Transfer Learning

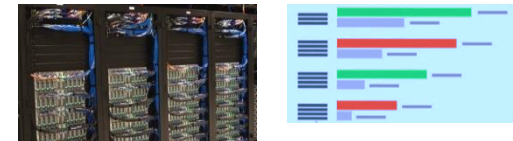
Continuous Learning

## AI Governance 확립



- AI 표준 개발환경
- 데이터/모델 버전관리

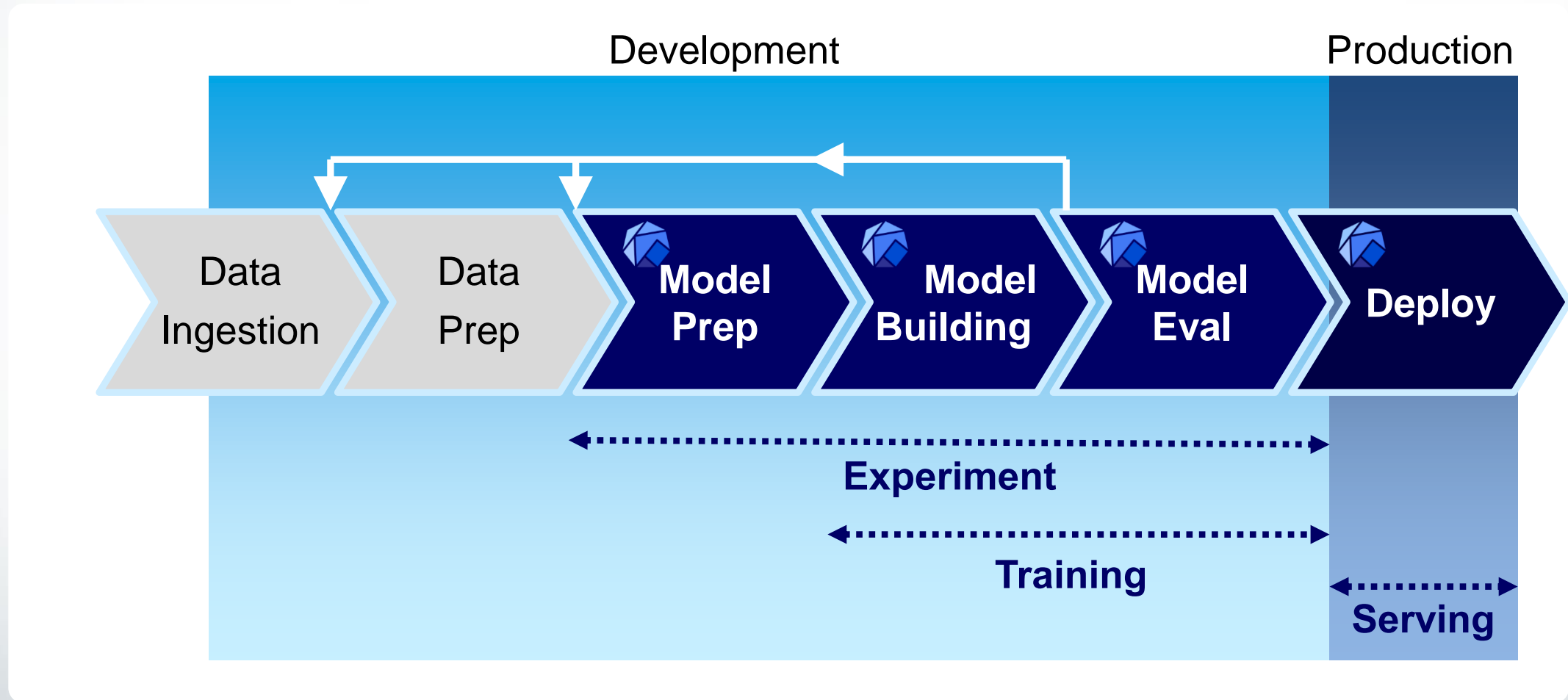
## 대규모 실험관리



- 쉽고 빠른 분산 모델 학습
- 지능형 병렬처리를 통한 실험 가속화

# AI Engineering Framework with KubeFlow

<< AI workflow >>



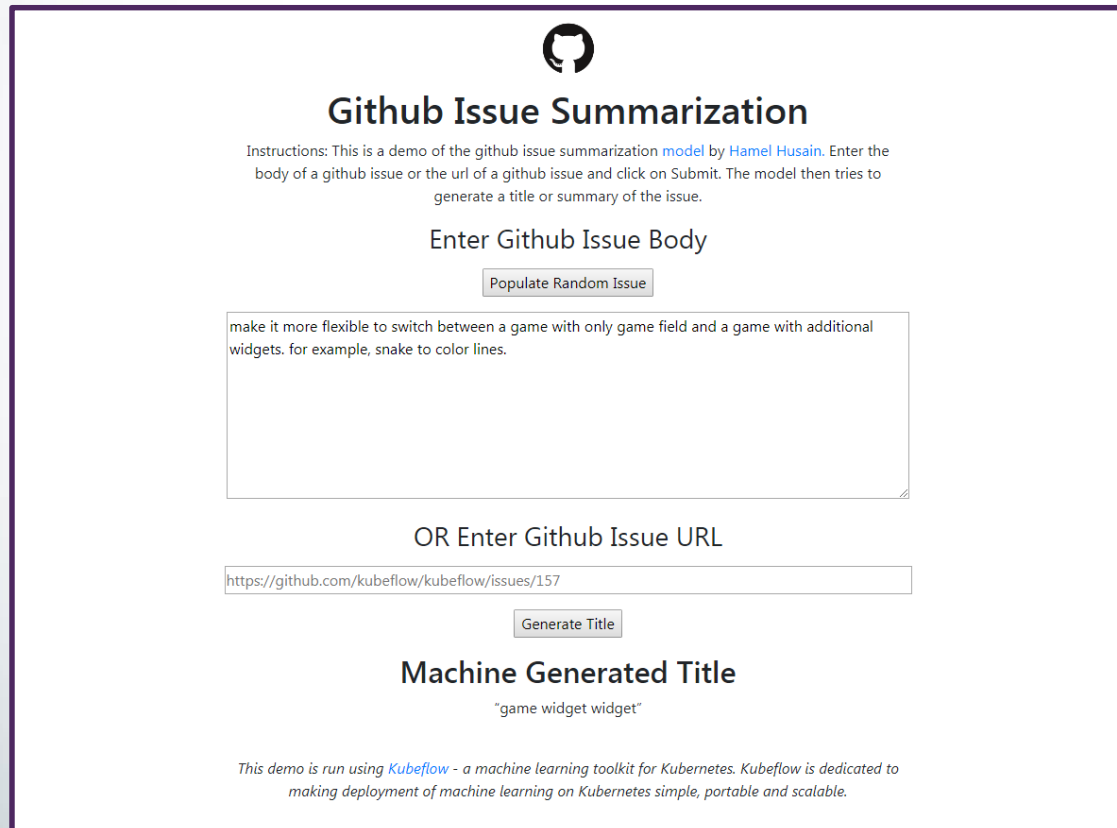
대규모 Machine Learning을 위한 Kubeflow 파헤치기

# Demo

# Github 이슈 타이틀 자동 생성 데모

Github issue에 대한 본문 내용 또는 URL을 입력 받아, 해당 이슈에 대한 타이틀을 생성하는 ML 서비스

## Demo Service UI

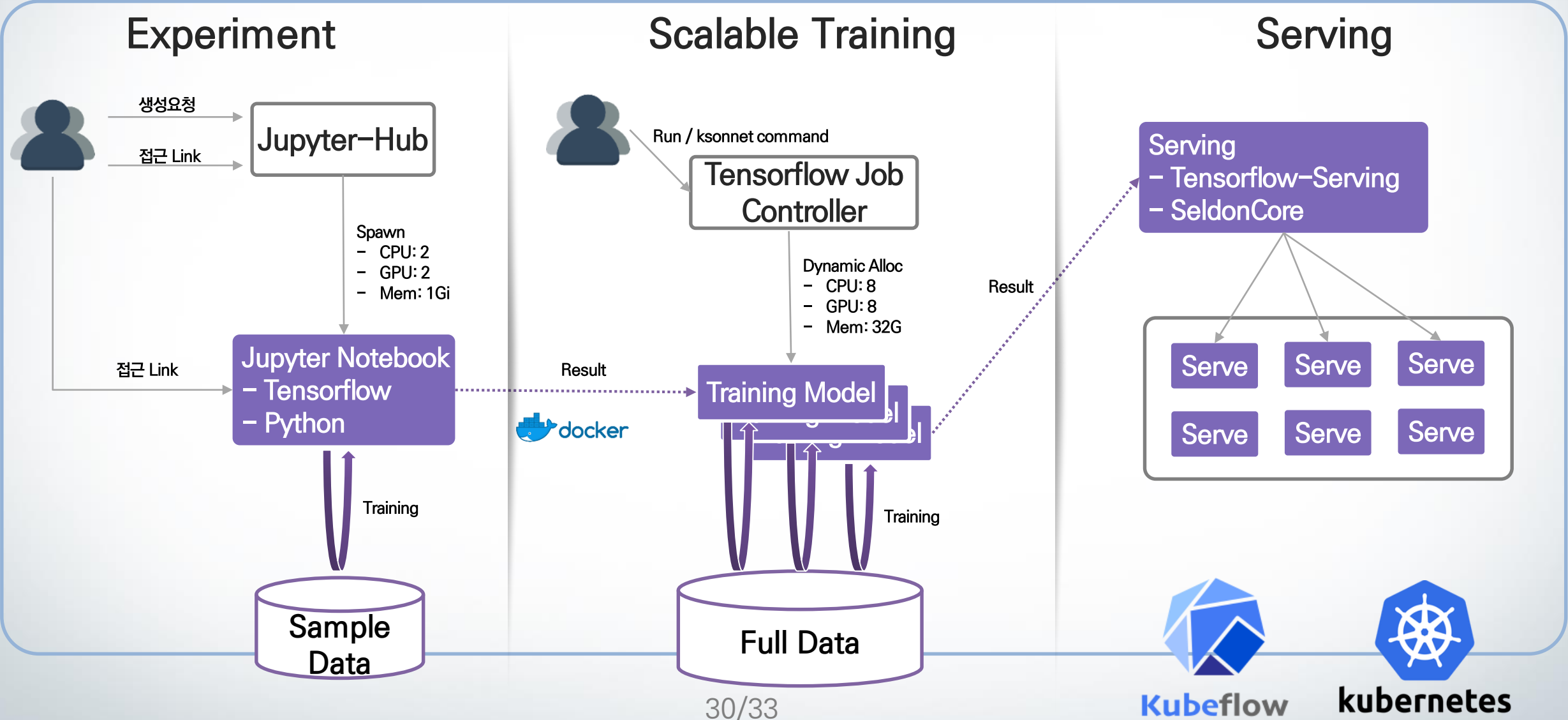


The screenshot shows a web interface for 'Github Issue Summarization'. At the top is the GitHub logo and the title 'Github Issue Summarization'. Below it are instructions: 'Instructions: This is a demo of the github issue summarization model by Hamel Husain. Enter the body of a github issue or the url of a github issue and click on Submit. The model then tries to generate a title or summary of the issue.' There are two input options: 'Enter Github Issue Body' with a 'Populate Random Issue' button, and 'OR Enter Github Issue URL' with a text input field containing 'https://github.com/kubeflow/kubeflow/issues/157' and a 'Generate Title' button. The output section shows 'Machine Generated Title' as '"game widget widget"'. At the bottom, a footer note states: 'This demo is run using Kubeflow - a machine learning toolkit for Kubernetes. Kubeflow is dedicated to making deployment of machine learning on Kubernetes simple, portable and scalable.'

## Demo Scenario

- ▶ Training the model using Jupyter notebook
  - Jupyter notebook creation
  - Model experiments
- ▶ Scalable Training the model using TF-Job Controller
- ▶ Deployment to Serving
- ▶ Querying the AI Service

# Demo Workflow



# Without Kubeflow vs With Kubeflow

## Without Kubeflow

### Setup infrastructure

- AI용 서버, 스토리지, 네트워크 구성
- 실험/대규모 훈련/서비스 환경 개별 구성

### Setup scheduling

- 구성된 환경 중 어떤 클러스터를 사용할지 결정
- Training Model을 각 서버에 배포

### Launch training

- 각각의 서버에 Training 실행

### Deploy model

### Setup load balancing

### Monitoring

## With Kubeflow

### Create a docker image

### Run training job

- Ksonnet 템플릿 생성
- Training 파라미터 설정
- 실행

### Deploy model

- Ksonnet 템플릿 생성
- Serving 파라미터 설정
- 실행



# 우리가 하고 있는 것들

## 만난 문제점들

- ▶ **Training과 Serving을 위한 Cloud 환경**
  - Training을 위한 GPU Resource 환경 필요
  - Data 처리를 위한 스토리지 필요
- ▶ **Kubeflow는 GCP기반으로 모든 세팅**
  - Jupyter Notebook을 위한 Block스토리지가 GCP의 Persistent Disk 와 연동됨
- ▶ **다른 서비스들과의 연동**
  - Jupyter Hub와 인증 시스템 연동
  - 머신러닝 수행과정의 모니터링
- ▶ **Kubeflow는 계속 버전업 中**

## 우리의 작업

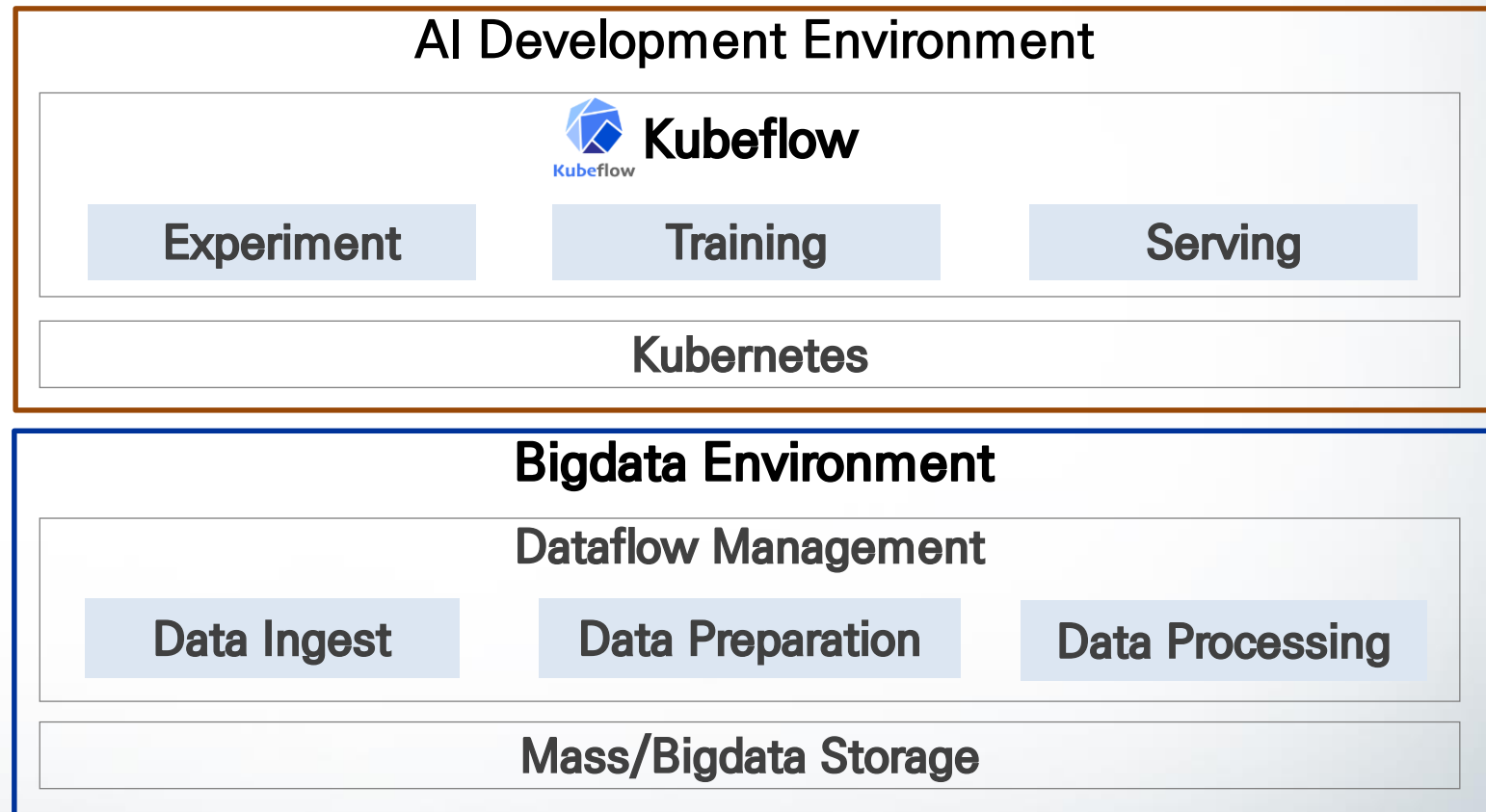
- ▶ **Training과 Serving을 위한 Private Cloud 구성**
  - GPU를 포함한 Kubernetes Cluster 구성
  - 각 단계별 NFS/GlusterFS등의 스토리지 구성
- ▶ **Private Cloud에서 동작하도록 변경**
  - Storage 바인딩을 위한 Kubernetes 설정 및 Kubeflow 배포/설치 코드 수정
- ▶ **Enterprise 서비스를 위한 오픈 소스 확인 中**
  - Jupyter Hub와 LDAP 연동 확인 中
  - TensorBoard를 비롯한 운영 관리 툴의 효율적인 사용
- ▶ **빠르게 따라가는 中**

# 향후 계획

## Cloud Infrastructure for AI HPC

### ◦ AI HPC 주요 특징

- Bigdata 수집/저장/처리 가능
- AI 데이터를 위한 Dataflow 관리
- GPU Pool을 관리하고,  
다수를 위한 AI 개발 Tool 제공
- Kubernetes 기반 환경 구성



# Q & A

Partner

Disrupt

Foresee



Thank you

