

MLOps 쉽게 구현하는 ML Model Serving A to Z

SKT 안승규

Who am I



- 안승규 (ahnsk@sk.com)
- SK텔레콤 Senior Software Engineer
- Java Enterprise Application 개발
- OpenStack 기반 Private Cloud 개발
- Kubernetes 기반 플랫폼 개발
- Istio 기반 MSA 플랫폼 개발
- ML Model Serving 플랫폼 개발

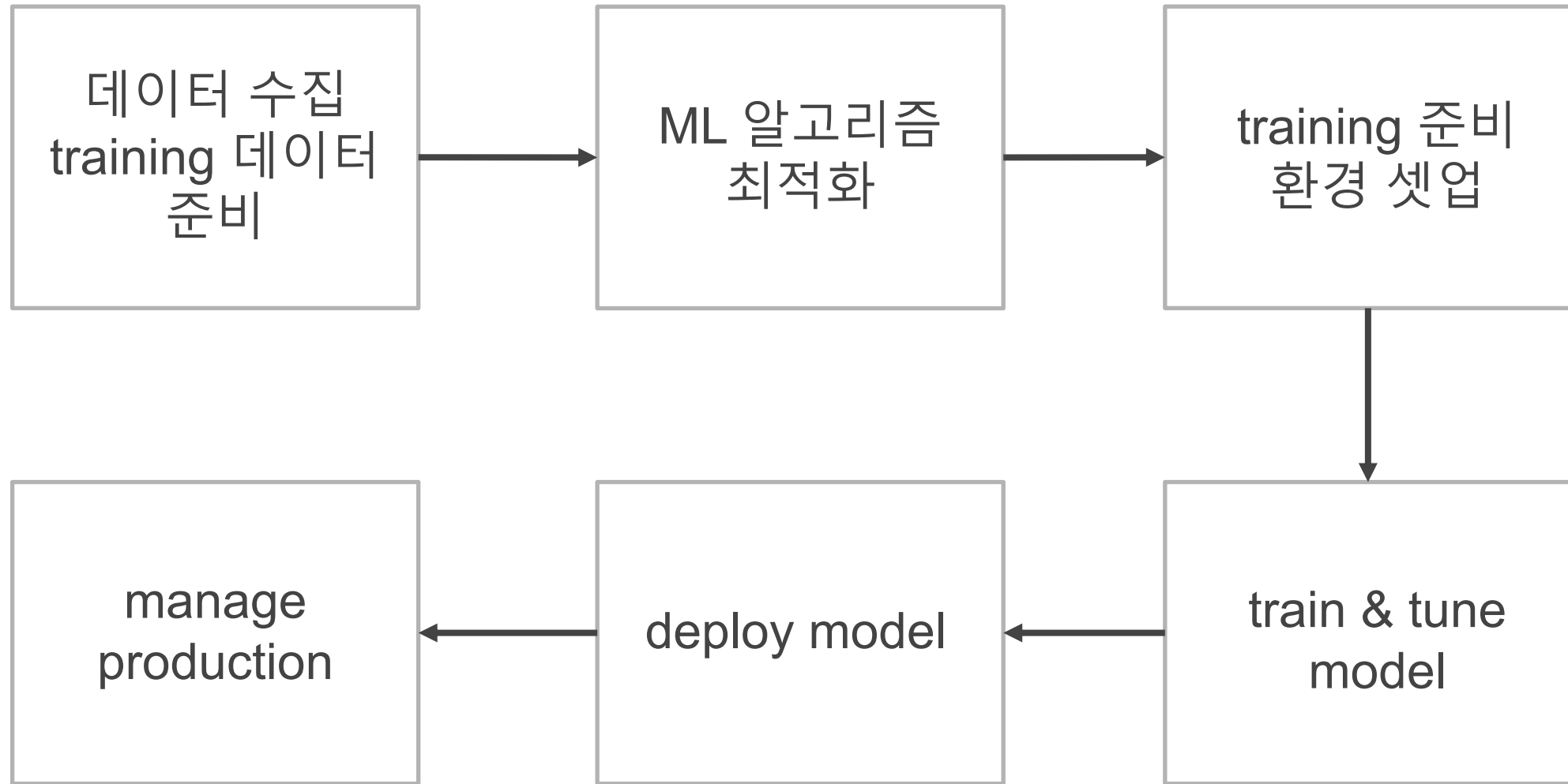
MLOps 의미

- ML 모델을 개발 / Training 하고, 이를 운영 환경으로 배포하는 프로세스를 자동화하여, 지속적으로 모델을 유지관리하고 모니터링할 수 있게 한 Machine Learning Engineering
- Data Scientist 와 ML Engineer 간 유기적 협력 지원

Why? MLOps

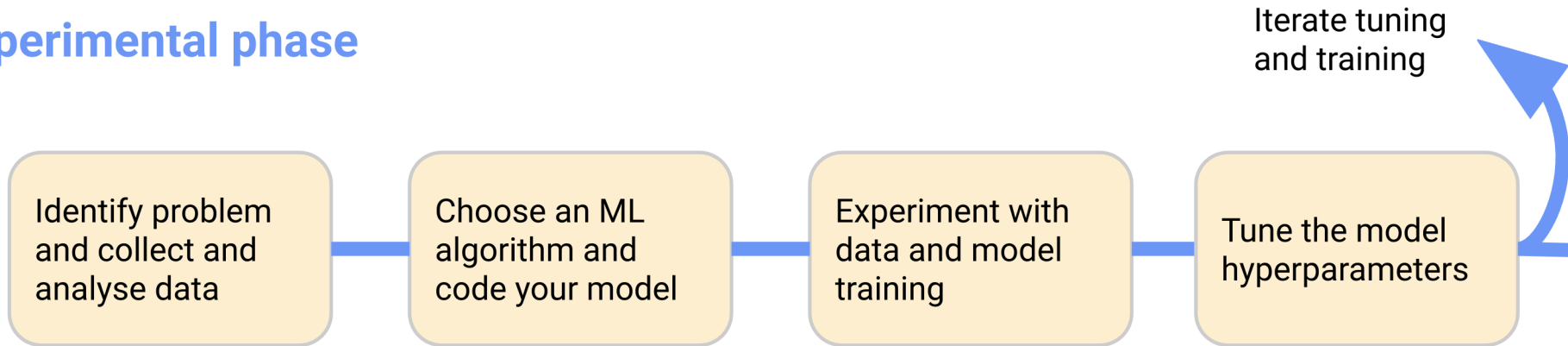
- 빠른 실험과 모델 개발
- 운영 환경에 신속/안정적 모델 배포 및 업데이트
- 정확한 검증

Machine Learning Pipeline

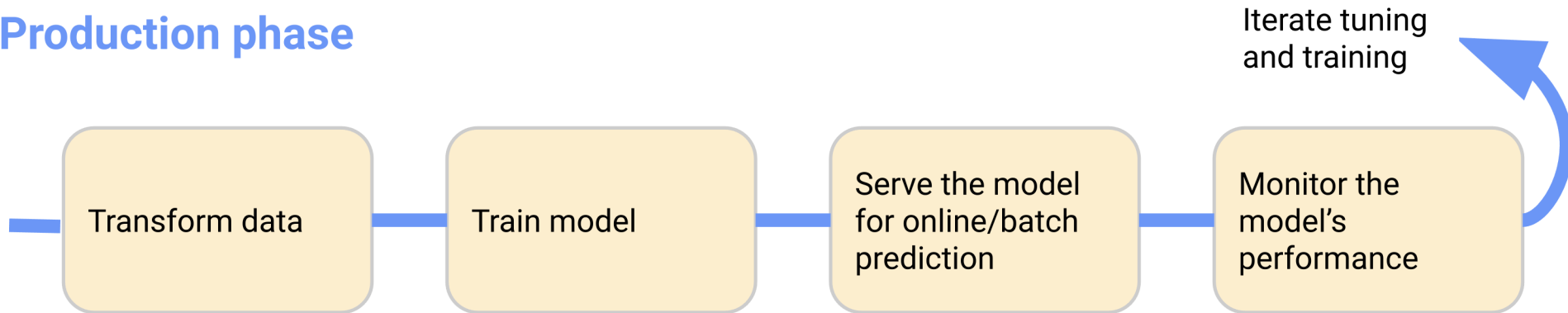


단계별 ML pipeline

Experimental phase



Production phase

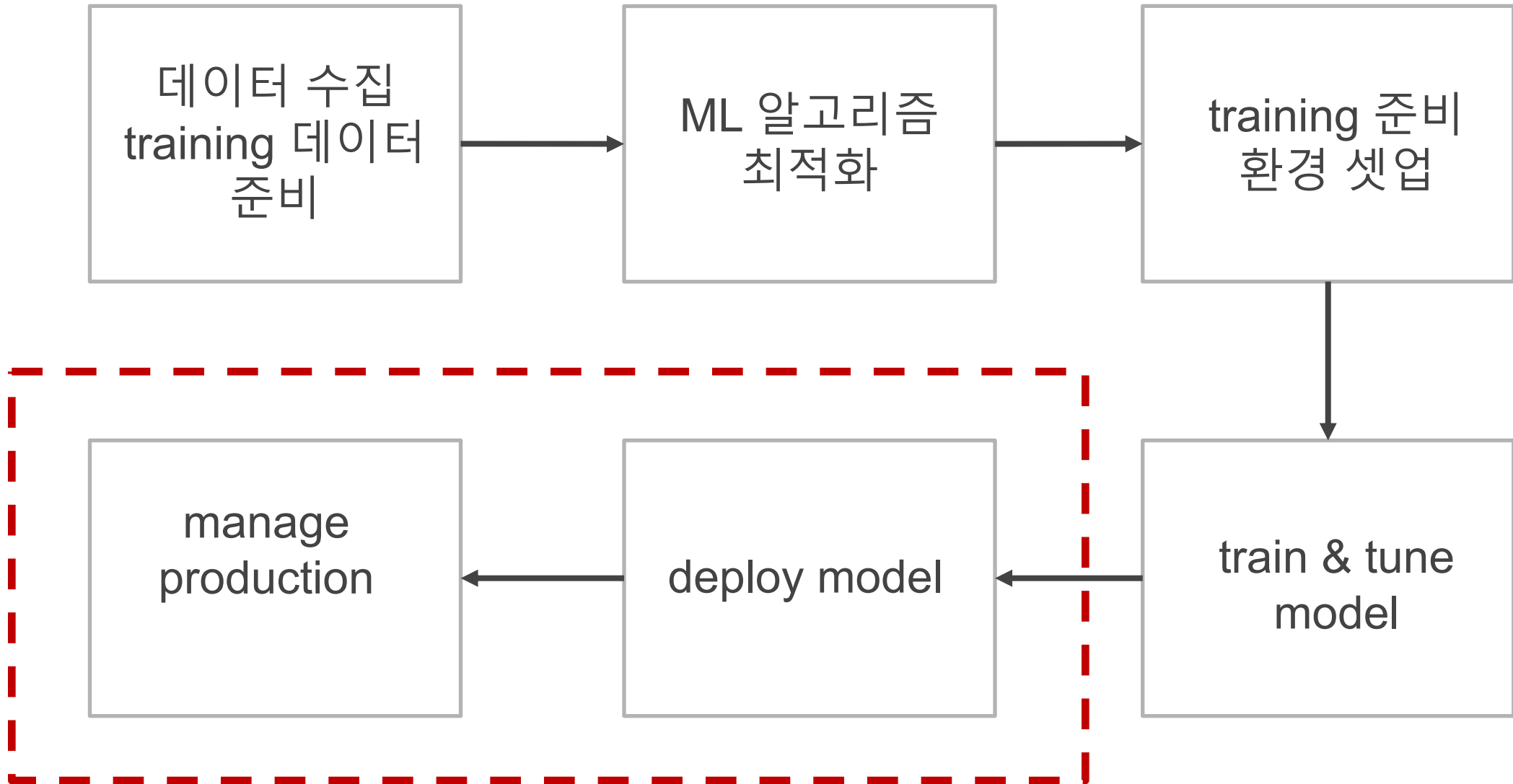


<https://www.kubeflow.org/docs/images/kubeflow-overview-workflow-diagram-1.svg>

What is Kubeflow ?

- Kubernetes 위에 simple, portable, scalable 하게 설치할 수 있는 MLOps Open Source S/W
- Google 이 내부적으로 Tensorflow Extended (pipeline)을 기반으로 Tensorflow를 지원
- 아래의 컴포넌트로 구성
 - Training Operators
 - KServing
 - Katib
 - Kubeflow Pipelines
 - Notebooks, Control Dashboard
- 최신 버전은 v1.6.1 -> Kubernetes 1.22 지원
- CNCF Project 로 관리하기 위해서 CNCF Incubating Project 로 지원 (Oct 24, 2022)

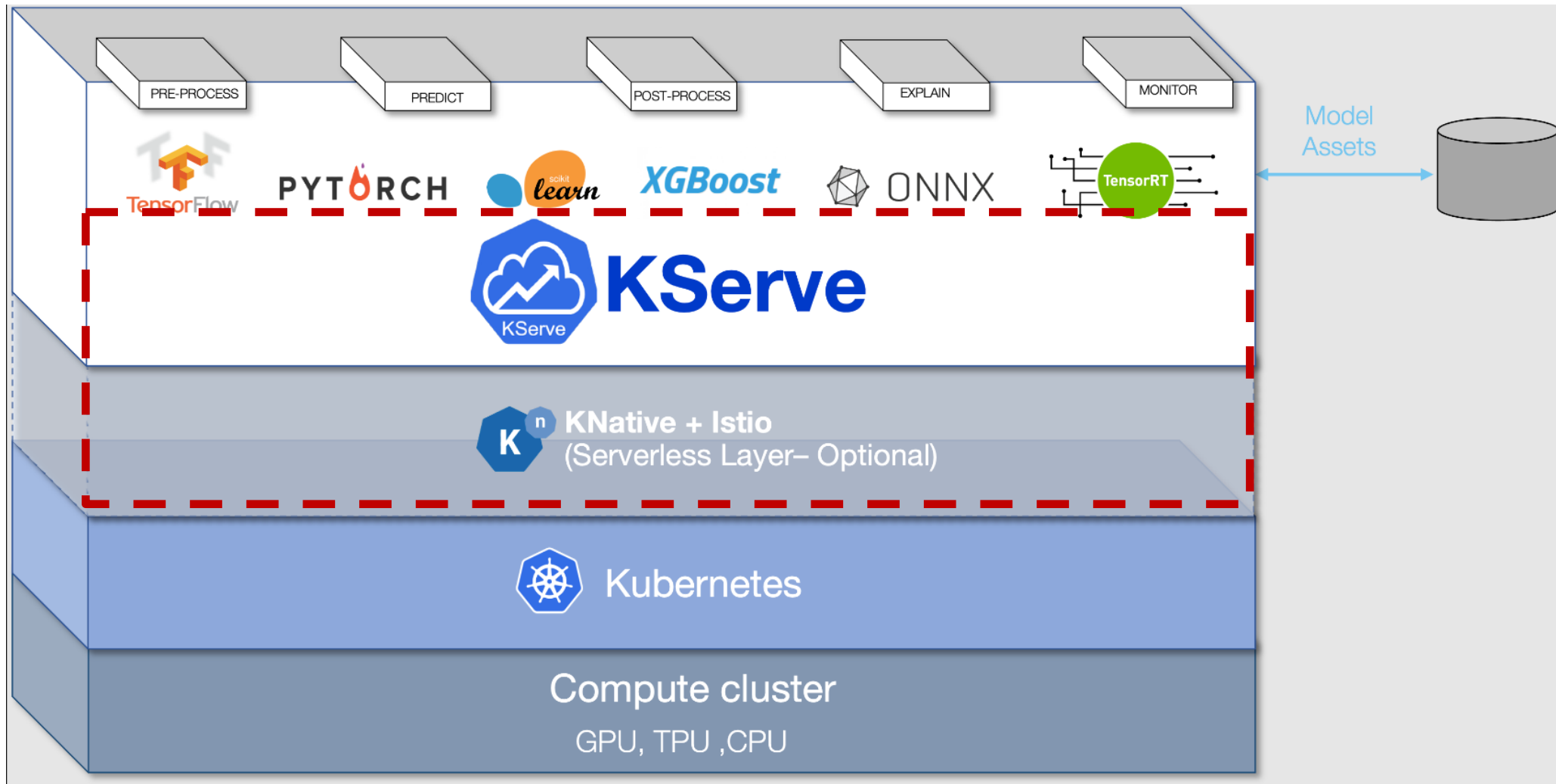
현재 우리가 집중하는 영역 (Stage 1)



Why Model Serving ?

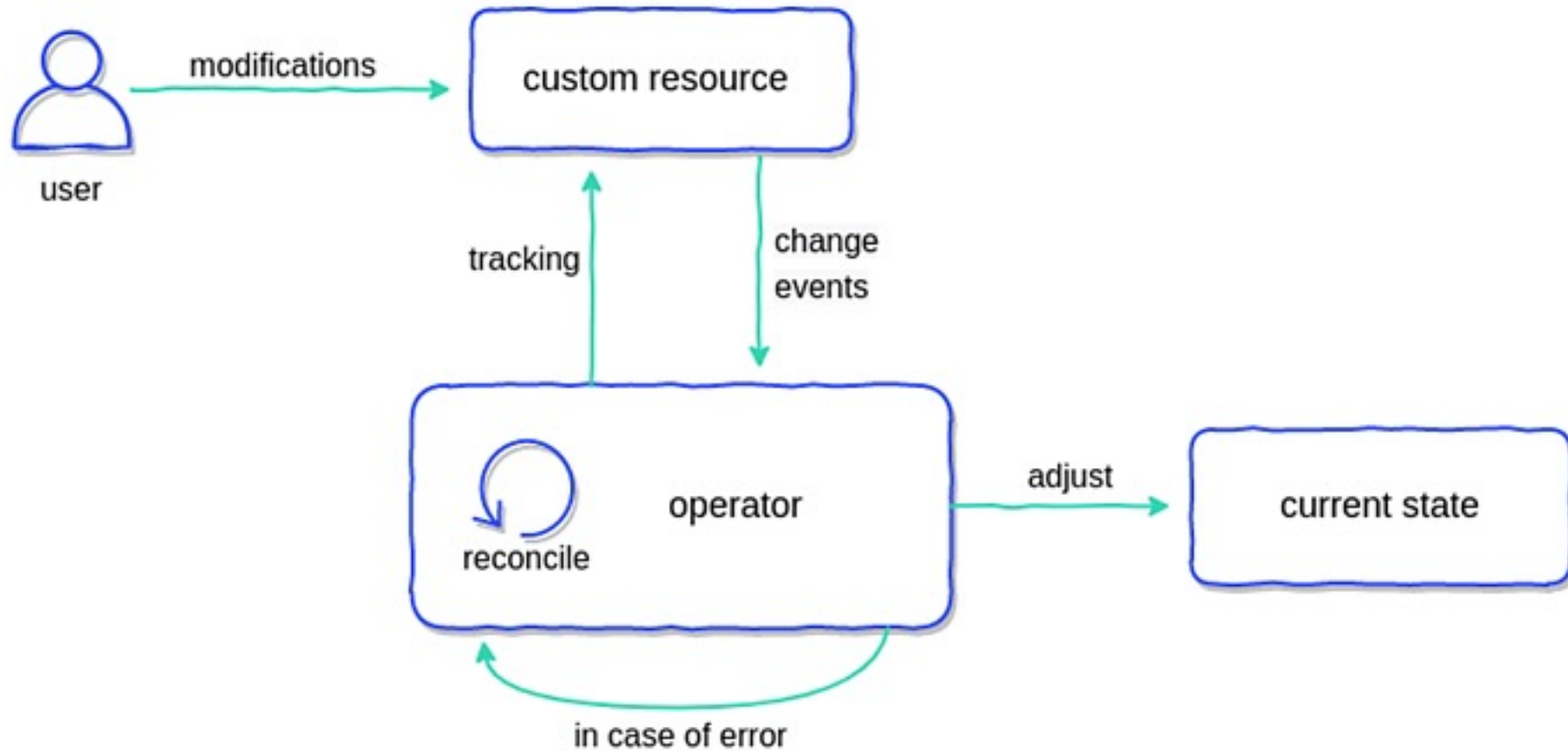
- Cost
인스턴스 갯수는 적절한가? 리소스는 효율적으로 할당되어 있나?
- Monitoring
서비스는 잘 동작하는가? 최적의 성능을 내는 설정은? Request Tracing 은?
- Rollout
배포는 안정적으로 배포되었나? 롤백은 어떻게? 트래픽 변경없이 변경사항 테스트 가능?
- Protocol Standard
Prediction 을 어떻게 만들까? gRPC? HTTP?
- Framework
Tensorflow 를 어떻게 Serving 할까? XGBoost? Scikit Learn? Pytorch? Custom Code?
- Features
Prediction 을 어떻게 설명하지? outlier, skew 탐지는?, custom pre/post processing 은 어떻게?

KServe Architecture



<https://www.kubeflow.org/docs/external-add-ons/kserve/pics/kserve.png>

What is Operator (Custom Controller) ?

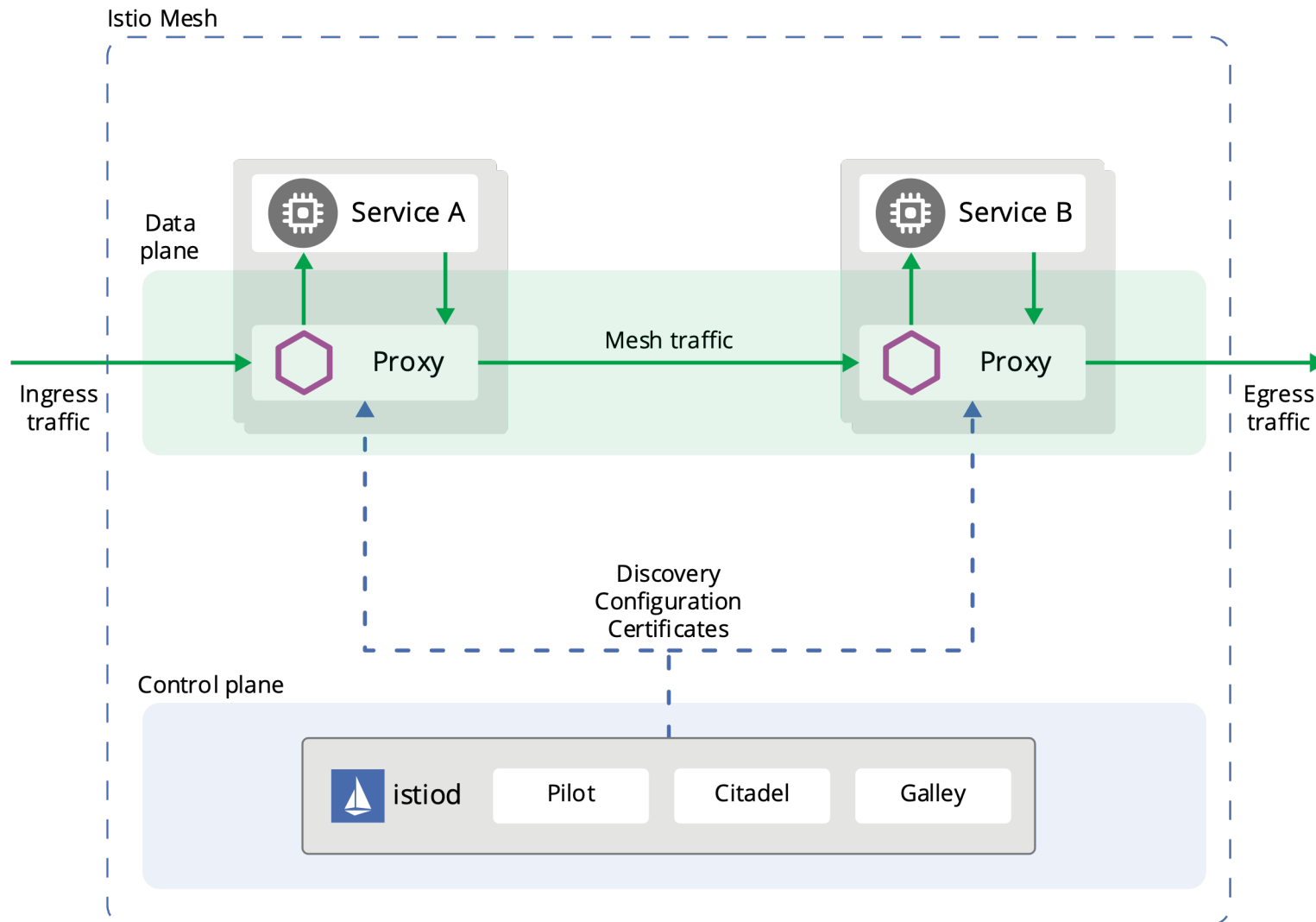


<https://docs.apimatic.io/assets/images/kubernetes-operator-diagram-ce95ced3bd0495aa4a5e4476dc17b5b8.png>

What is Istio ?

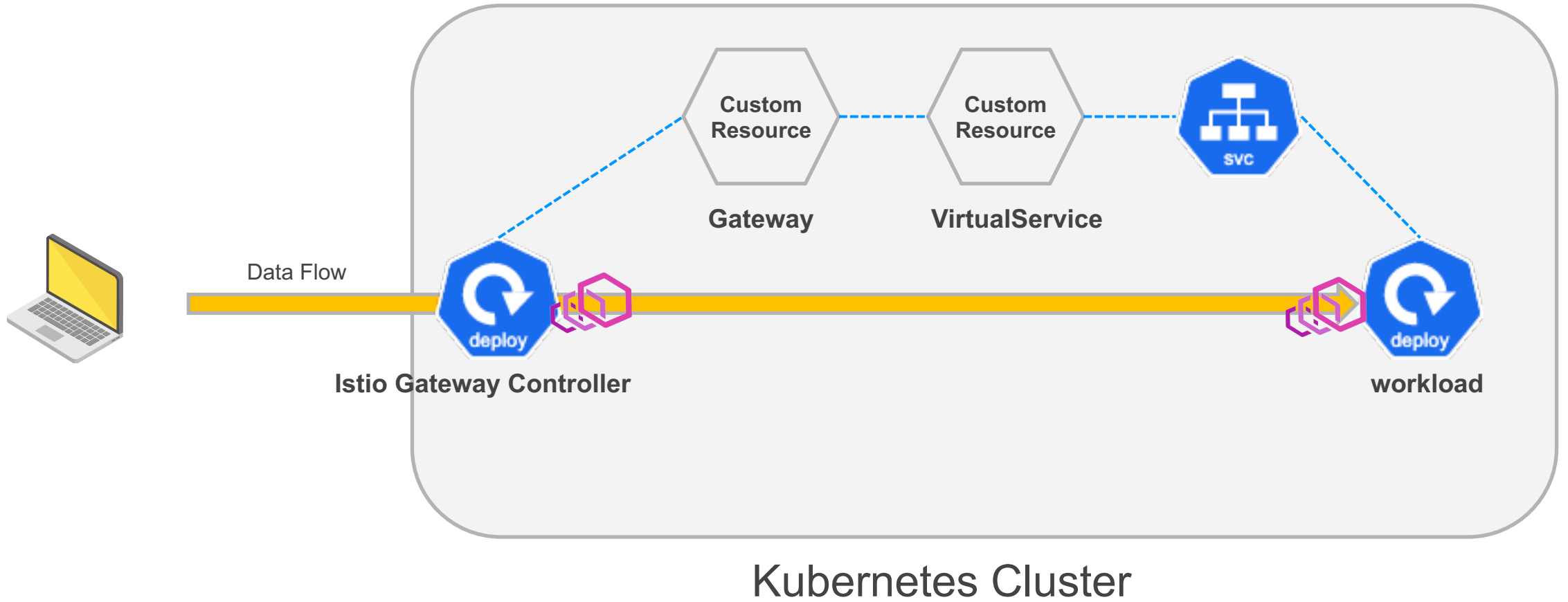
- Envoy Proxy (Side car) 기반의 Service Mesh
- CNCF Incubating Project 승인 (Sep 30, 2022)
- Operator 패턴으로 동작함
- 가장 큰 특징은 Traffic management

Istio Architecture



<https://istio.io/latest/docs/ops/deployment/architecture/arch.svg>

Istio Data Flow



Istio Custom Resource

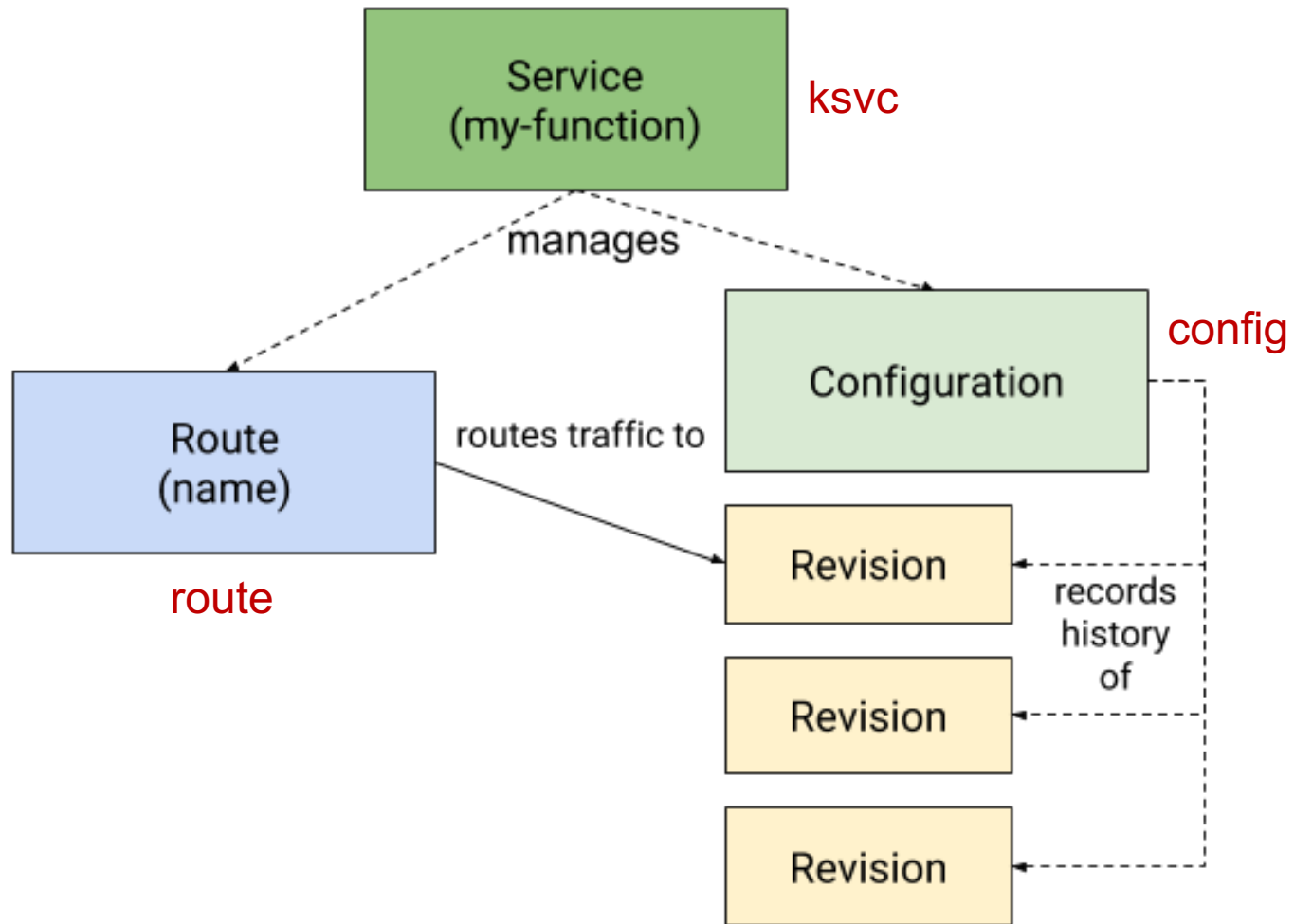
```
apiVersion: networking.istio.io/v1alpha3
kind: Gateway
metadata:
  name: bookinfo-gateway
spec:
  selector:
    istio: ingressgateway
  servers:
  - port:
      number: 80
      name: http
      protocol: HTTP
    hosts:
    - "*"
```

```
apiVersion: networking.istio.io/v1alpha3
kind: VirtualService
metadata:
  name: bookinfo
spec:
  hosts:
  - "*"
  gateways:
  - bookinfo-gateway
  http:
  - match:
    - uri:
        exact: /productpage
    - uri:
        exact: /login
    route:
    - destination:
        host: productpage.default.svc.cluster.local
        port:
          number: 9080
```

What is Knative ?

- Serverless and Event Driven Application
- CNCF Incubating Project 승인 (Mar 2, 2022)
- Operator 패턴으로 동작함

Knative Resource Types



https://github.com/knative/specs/raw/main/specs/serving/images/object_model.png

Knative Custom Resource

```
apiVersion: serving.knative.dev/v1
kind: Service
metadata:
  name: helloworld-go
  namespace: default
spec:
  template:
    spec:
      containers:
      - image: gcr.io/knative-samples/helloworld-go
        env:
        - name: TARGET
          value: "Go Sample v1"
```

```
$ kubectl get deploy
```

NAME	READY	UP-TO-DATE	AVAILABLE	AGE
helloworld-go-00001-deployment	0/0	0	0	5d15h

```
$ kubectl get kpa
```

NAME	DESIREDSCALE	ACTUALSCALE	READY	REASON
helloworld-go-00001	0	0	False	NoTraffic

What is KServe ?

- Model Inference Platform
- Operator 패턴으로 동작함
- 다양한 ML Framework 을 지원
 - Scikit-Learn, XGBoost, Tensorflow, PyTorch and Custom model runtime

KServe Custom Resource

```
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: mnist
spec:
  predictor:
    model:
      modelFormat:
        name: tensorflow
        version: "2"
      storageUri: "gs://kserve/models/mnist"
      runtime: kserve-tensorflow-serving
```

```
$ kubectl get isvc
```

NAME	URL	READY	PREV	LATEST	...
mnist	http://mnist-default.taco-cat.xyz	True		100	...

KServe Runtime

```
$ kubectl get clusterservingruntimes
```

NAME	DISABLED	MODELTYPE	CONTAINERS	AGE
kserve-lgbserver		lightgbm	kserve-container	5d12h
kserve-mlserver		sklearn	kserve-container	5d12h
kserve-paddleserver		paddle	kserve-container	5d12h
kserve-pmmlserver		pmml	kserve-container	5d12h
kserve-sklearnserver		sklearn	kserve-container	5d12h
kserve-tensorflow-serving		tensorflow	kserve-container	5d12h
kserve-torchserve		pytorch	kserve-container	5d12h
kserve-tritonserver		tensorrt	kserve-container	5d12h
kserve-xgbserver		xgboost	kserve-container	5d12h



감사합니다