

Still Burning GPUs On Debugging?

Scale AI In One Line



KubeCon



CloudNativeCon

Europe 2026



Anya Kramar

Software Engineer at Red Hat
Kubeflow SDK Maintainer

github.com/kramaranya



KubeCon



CloudNativeCon

Europe 2026

You write a training function



Now scale it



KubeCon



CloudNativeCon

Europe 2026

"Practitioners spend 80–90% of time debugging and tuning, not building."

Andrej Karpathy · Full Stack Deep Learning



KubeCon



CloudNativeCon

Europe 2026

Kubeflow SDK

Unified Pythonic APIs to run any AI workload at any scale

```
pip install kubeflow
```



KubeCon



CloudNativeCon

Europe 2026

Start local → Scale it

Local

```
client = TrainerClient(
    backend_config=LocalProcessBackendConfig()
)

client.train(
    trainer=CustomTrainer(
        func=train_fn,
    ),
)
```

Distributed

```
client = TrainerClient()

client.train(
    runtime="torch-distributed",
    trainer=CustomTrainer(
        func=train_fn,
        num_nodes=100,
        resources_per_node={"gpu": 8},
    ),
)
```

Same function → **100 × 8 = 800 GPUs**



KubeCon



CloudNativeCon

Europe 2026

One line decides where it runs

Quick prototype — subprocess

```
client = TrainerClient(backend_config=LocalProcessBackendConfig())
```

Local dev — Docker / Podman

```
client = TrainerClient(backend_config=ContainerBackendConfig())
```

Production — Kubernetes

```
client = TrainerClient()
```

Your training code doesn't change. Not a single line.



KubeCon



CloudNativeCon

Europe 2026

One SDK for the full AI workflow

Training

```
TrainerClient().train(  
    trainer=CustomTrainer(  
        func=train_fn,  
    ),  
)
```

Hyperparameter optimization

```
OptimizerClient().optimize(  
    trial_template=template,  
    search_space={  
        "lr": Search.loguniform(0.001, 0.1),  
    },  
)
```

Model registry

```
ModelRegistryClient().register_model(  
    name="my-model",  
    uri="s3://bucket/model",  
    version="v1.0",  
)
```

Data processing

```
SparkClient().connect(  
    num_executors=5,  
    resources_per_executor={  
        "cpu": "4",  
    },  
)
```



KubeCon



CloudNativeCon

Europe 2026

What's next

Pipelines (KFP)

Workflow orchestration

Feast

Feature store

OpenTelemetry

Tracing & metrics across the SDK

MLflow

Experiment tracking integration

MCP Server

AI agents can operate Kubeflow

Unified auth

One auth story across all clients

and more...



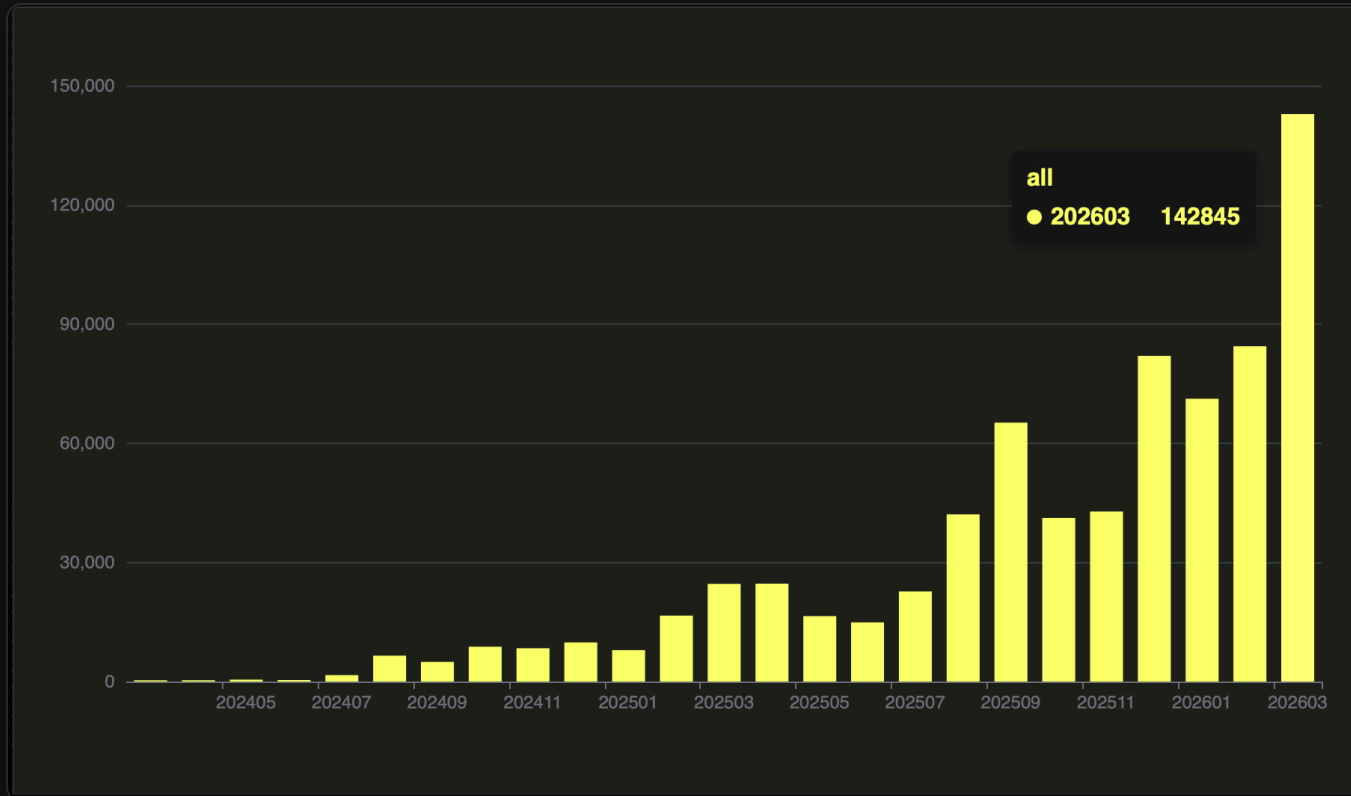
KubeCon



CloudNativeCon

Europe 2026

One more thing 🎉



KubeCon



CloudNativeCon

Europe 2026

Come find us

This week at KubeCon

Kubeflow Booth: P-8B, Thursday 10:00–14:00

Maintainer Session: Tuesday 15:15, E103-105

Get involved

Slack: [#kubeflow-ml-experience](#)

Meetings: Wednesdays 3pm UTC (bi-weekly)

GitHub: [kubeflow/sdk](#)



Kubeflow SDK Website



Blog post