Kubeflow Workshop

葉信和 / Hsin-Ho Yeh
Software Engineer / CEO @ 信誠金融科技
hsinho.yeh@footprint-ai.com

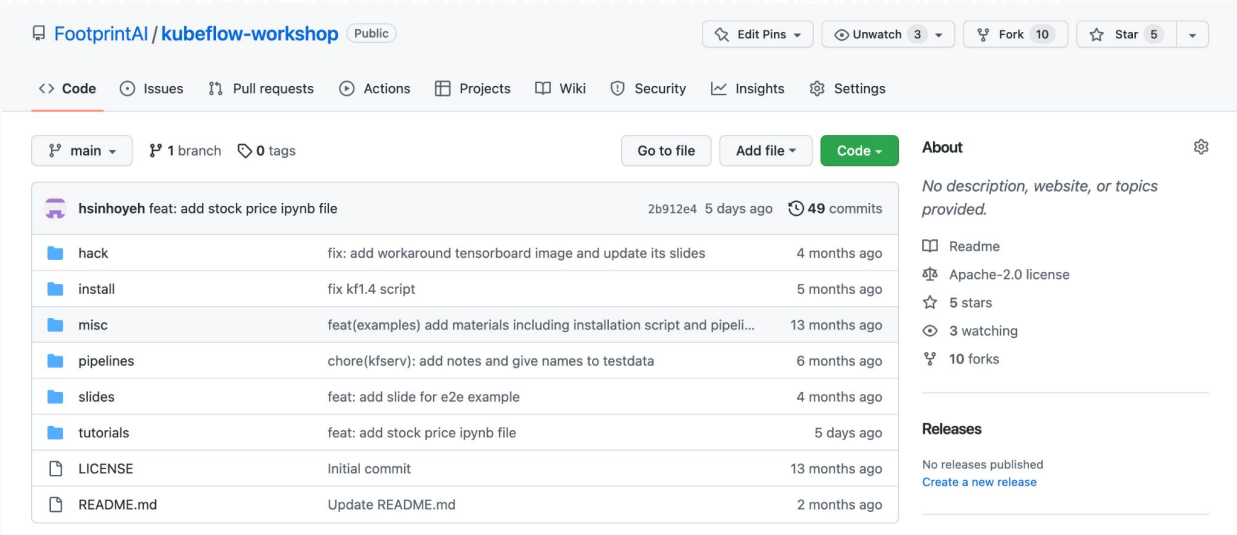
2021/04/26

Download Slides

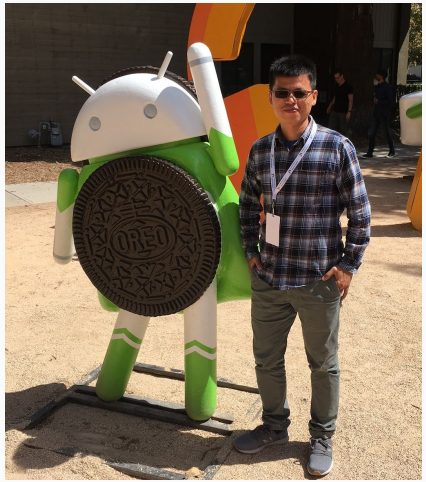
https://reurl.cc/Lm4Ab9

Workshop materials

```
git clone
https://github.com/FootprintAI/kubeflow-workshop
```



About me



- 2020 - Present at 信誠金融科技
  - Shrimping: A data-sharing platform
    - https://get-shrimping.footprint-ai.com
  - Tintin: a machine learning platform for everyone
    - https://get-tintin.footprint-ai.com
- 2016 - 2020 at IglooInsure (16M+ in series A+ 2020)
  - Provide digital insurance for e-conomic world
  - Funded in KUL, Headquartered in Singapore
  - First employee/ Engineering Lead / Regional Head/ Chief Engineer
- 2013 - 2016 at Studio Engineering @ hTC
  - Principal Engineer on Cloud Infrastructure Team
- 2009 - 2012 at IIS @ Academia Sinica
  - Computer vision, pattern recognition, and data mining
- CS@CCU, CS@NCKU alumni

Agenda

- Pre-requirement
- Why Kubeflow?
- What is Kubeflow?
- Kubeflow Architecture
- Hands-on time
- QA

Pre-requirement

- Be comfortable with UNIX command line
  - Navigating directories with `cd` or `tree`
  - Editing files, like `vim`, `nano`
  - Bash scripting, like env or looping
- Be an export with `Google`
  - https://letmegooglethat.com/?q=you+can+google+it

- It is totally OK if you don’t know what is Container and Kubernetes

荀子《儒效篇》

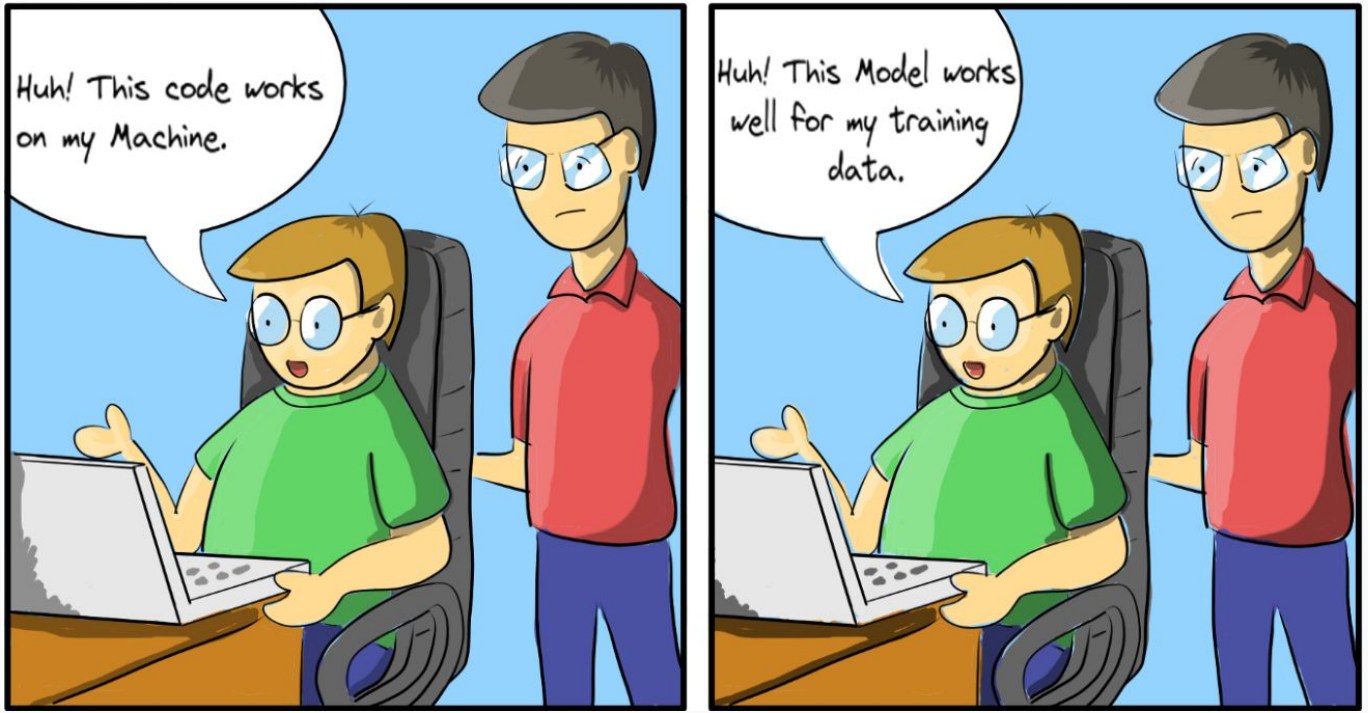
「不聞不若聞之，聞之不若見之，見之不若知之，知之不若行之；學至于行之而止矣。」

MLOps is the process of taking an experimental Machine Learning model into a production system by including continuous development practice of DevOps in the software field.

Ref: https://en.wikipedia.org/wiki/MLOps

Source: https://www.kubeflow.org/

A common scenario that we both experienced.



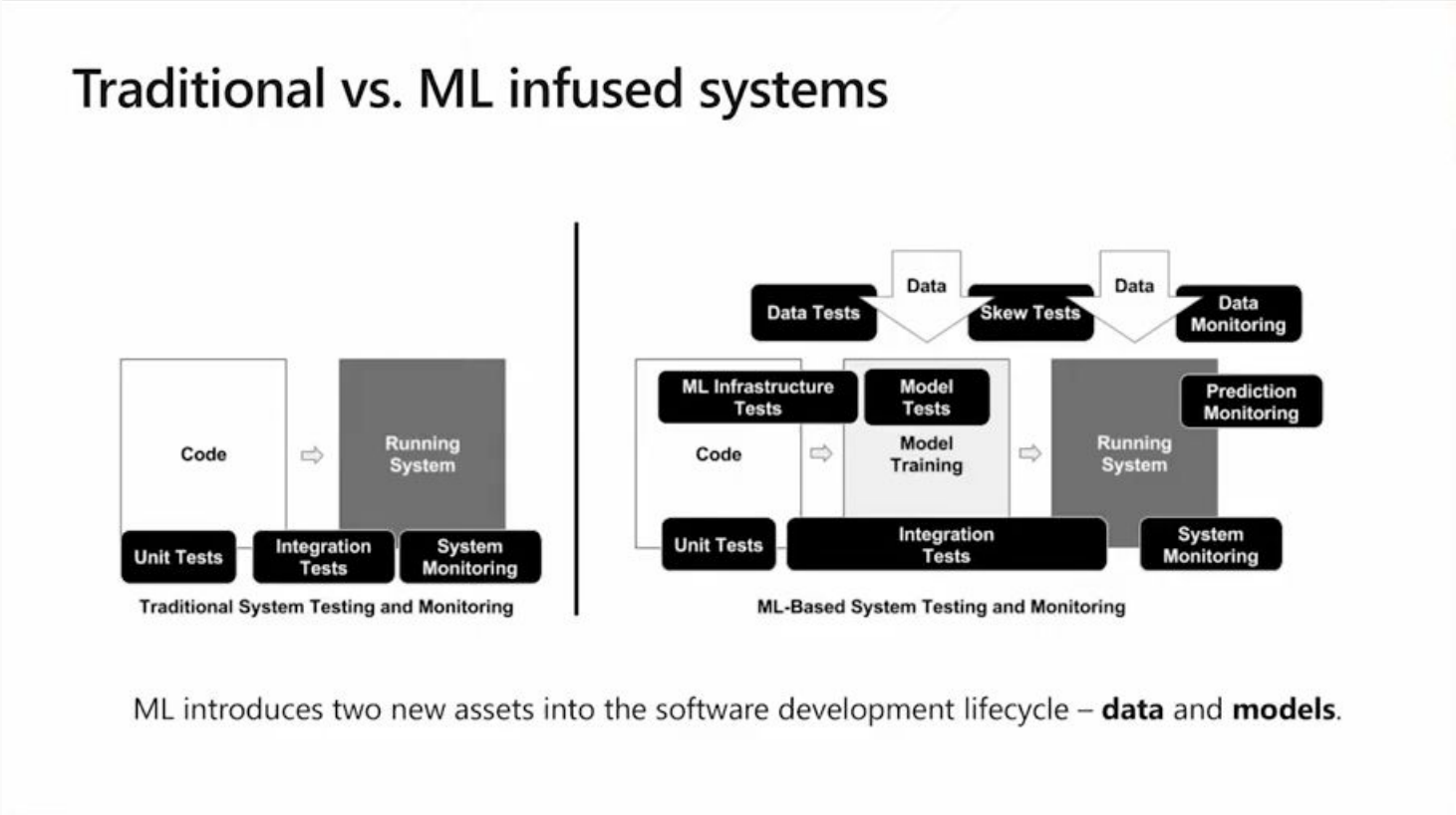
Ref: https://towardsdatascience.com/a-gentle-introduction-to-mlops-7d64a3e890ff

Building & deploying real-world ML application is *hard* and *costly* because of *lack of tooling* that covers end-to-end ML development & deployment

- CloudNext'19

Source CloudNext'19: https://www.youtube.com/watch?v=TZ1lGrJLEZ0

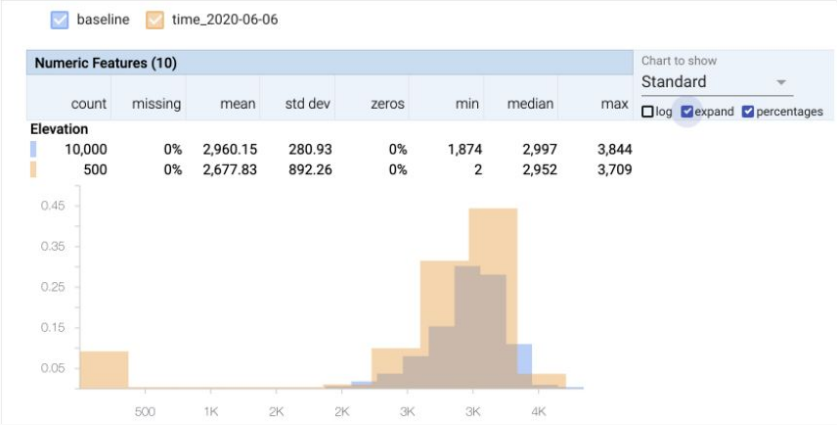
How Involving Machine Learning model could change the current software design?



Ref: DevOps Enterprise Summit' 19 https://www.youtube.com/watch?v=pqppGvTJm-A

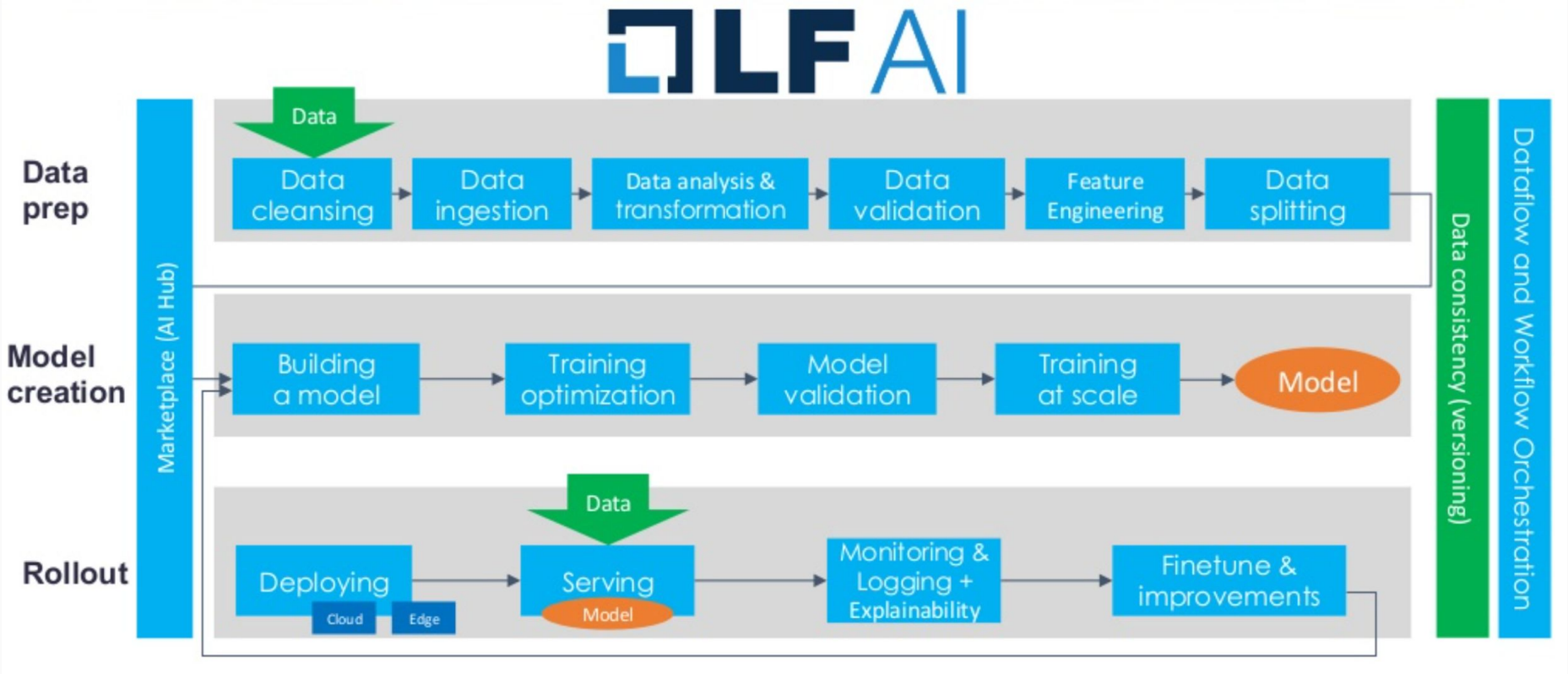
Why we should care about drifting?

- Data drifting
  - A skew grows between training data and serving data.
  - The discrepancies between training data and serving data can usually be classified as schema skews or distribution skews
- Concept drifting
  - The interpretation of the relationship between the input predictors and the target feature evolves



Ref: https://cloud.google.com/architecture/ml-modeling-monitoring-analyzing-training-server-skew-in-ai-platform-prediction-with-tfdv

Real-world Machine Learning Application - End-to-End ML LifeCycle



Source: https://www.slideshare.net/AnimeshSingh/advanced-model-inferencing-leveraging-kubeflow-serving-knative-and-istio-196096385

Why machine learning on Kubernetes?

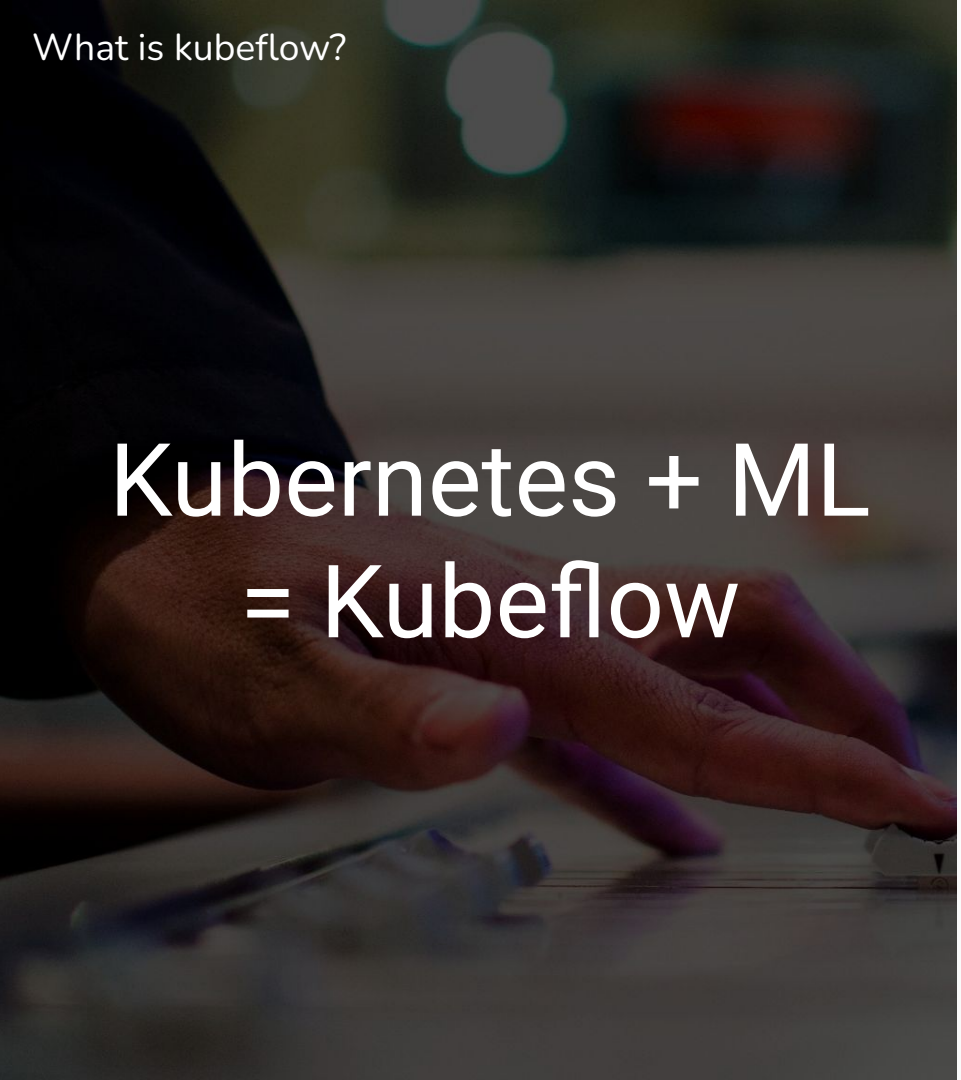
- Composability
  - Each stage are independent systems and are able to compose together
- Portability
  - Dev/Staging/Prod
  - Laptop/Edge/Cloud environment
- Scalability
  - Hyperparameter tuning, production workloads

Oh, you want to use ML on K8s?

Before that, can you become an expert in:

- Containers
- Packaging
- Kubernetes service endpoints
- Persistent volumes
- Scaling
- Immutable deployments
- GPUs, Drivers & the GPL
- Cloud APIs
- DevOps
- ...

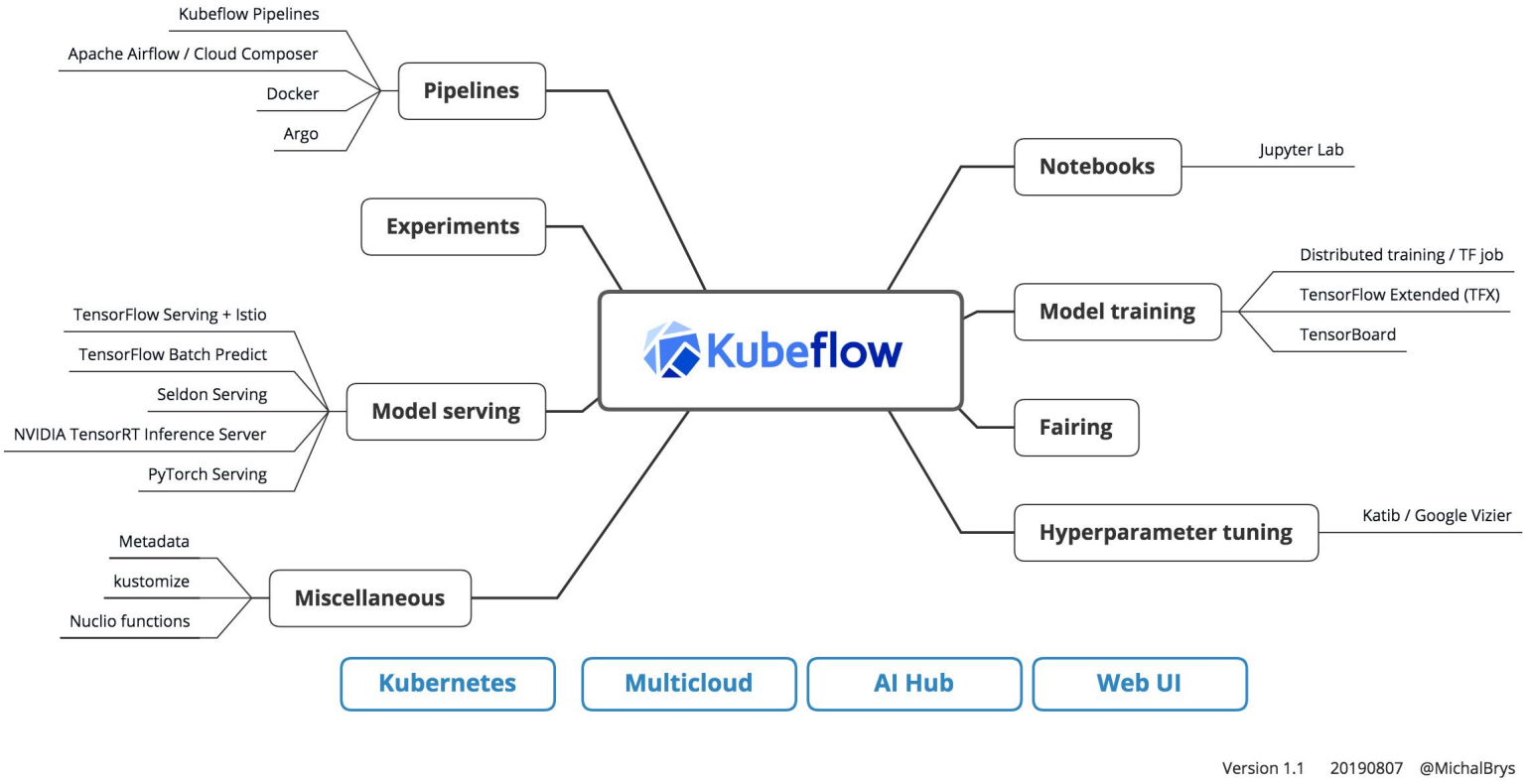
Source: @aronchick



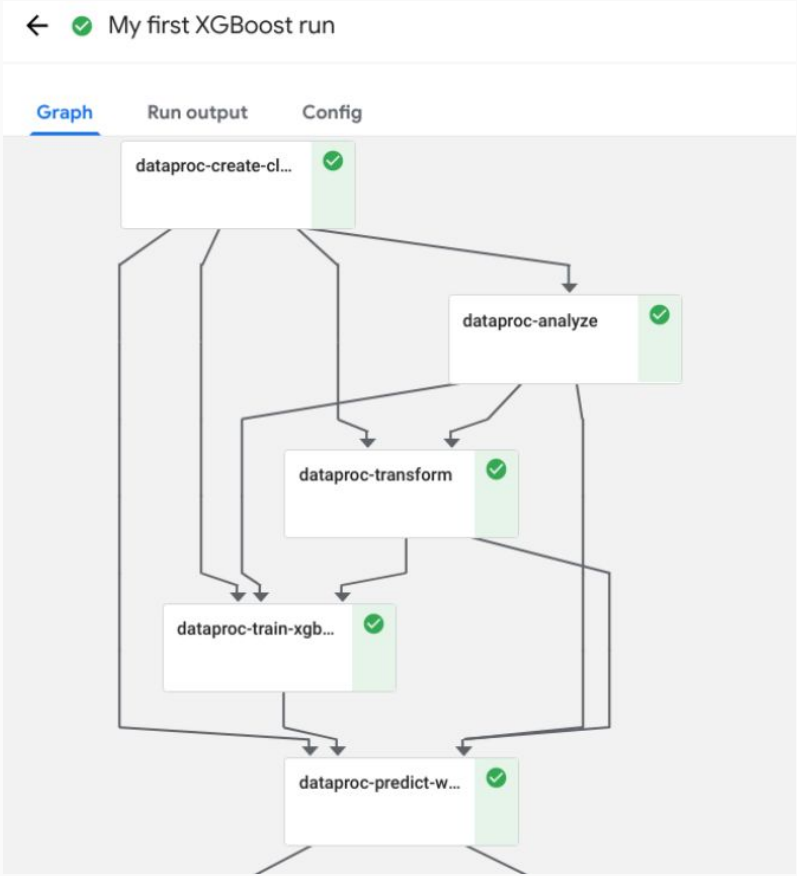
The Kubeflow project is dedicated to making deployments of machine learning (ML) workflows on Kubernetes simple, portable and scalable.

Source: https://www.kubeflow.org/

Architectures



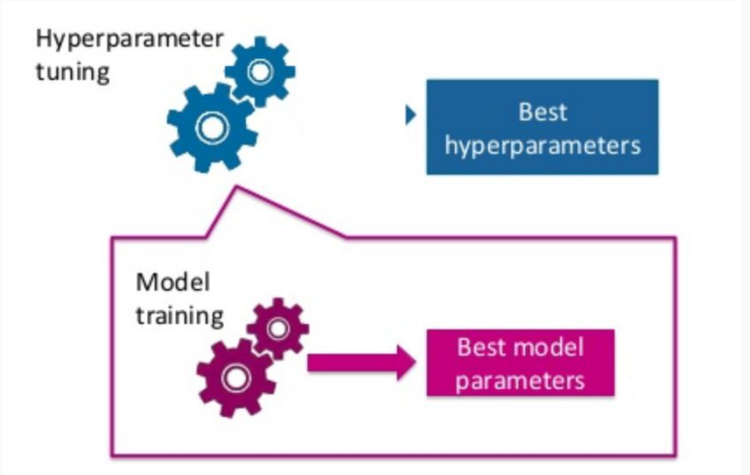
Kubeflow Pipelines



Source:
https://www.kubeflow.org/docs/pipelines/overview/pipelines-overview/

Hyperparameter tuning

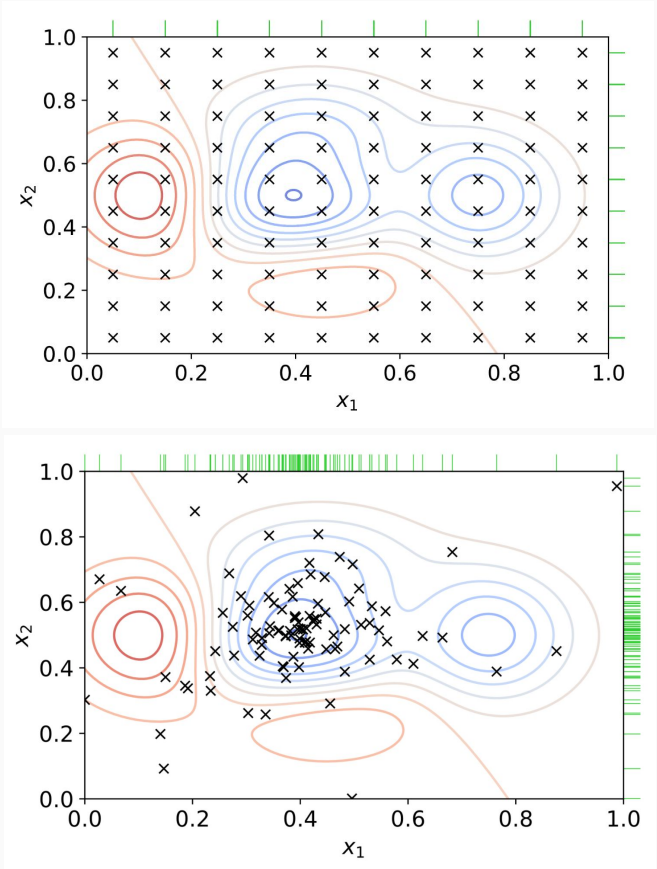
- Model Parameter vs Hyperparameter
  - Model parameters that will learn on its own during training process by the ML model, ex: weights and biases for a classifier.
  - Hyperparameter that directly control the behavior of training algorithm.



Source: https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

Hyperparameter tuning

- Algorithms
  - Grid Search (top)
  - Random Search
  - Bayesian Optimization
  - Gradient-based optimization (bottom)
  - … and more

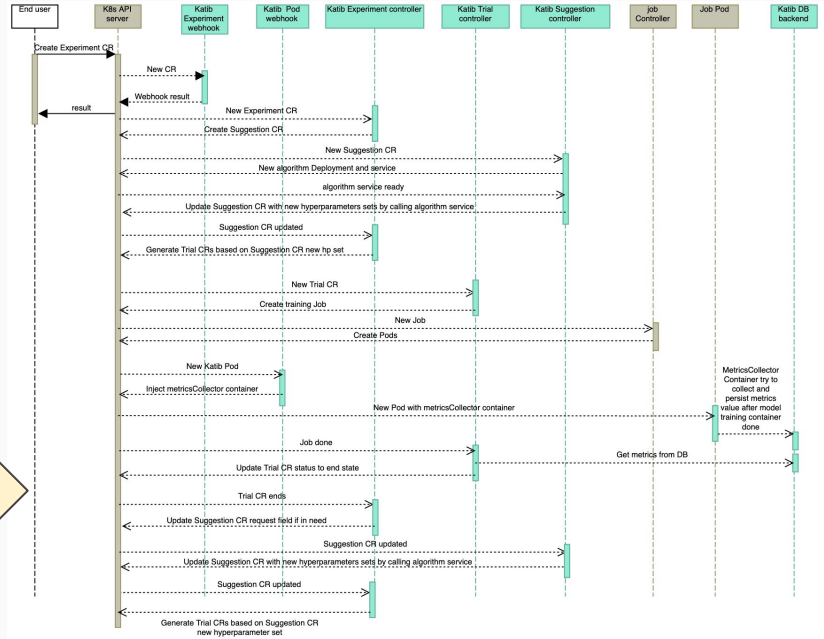


Source: https://en.wikipedia.org/wiki/Hyperparameter_optimization

Kubeflow Katib

- All works that katib has been doing can be described the following pseudocode and then, in reality, turn into the flow diagram below.

```
PROGRAM KATIB:
        exp = CreateExperiment();
        WHILE exp.objective is NOT reached:
            DO sugst = exp.CreateSuggestion();
                metrics = exp.CreateTrials(sugst.Assignments);
                exp.Report(metrics);
END.
```



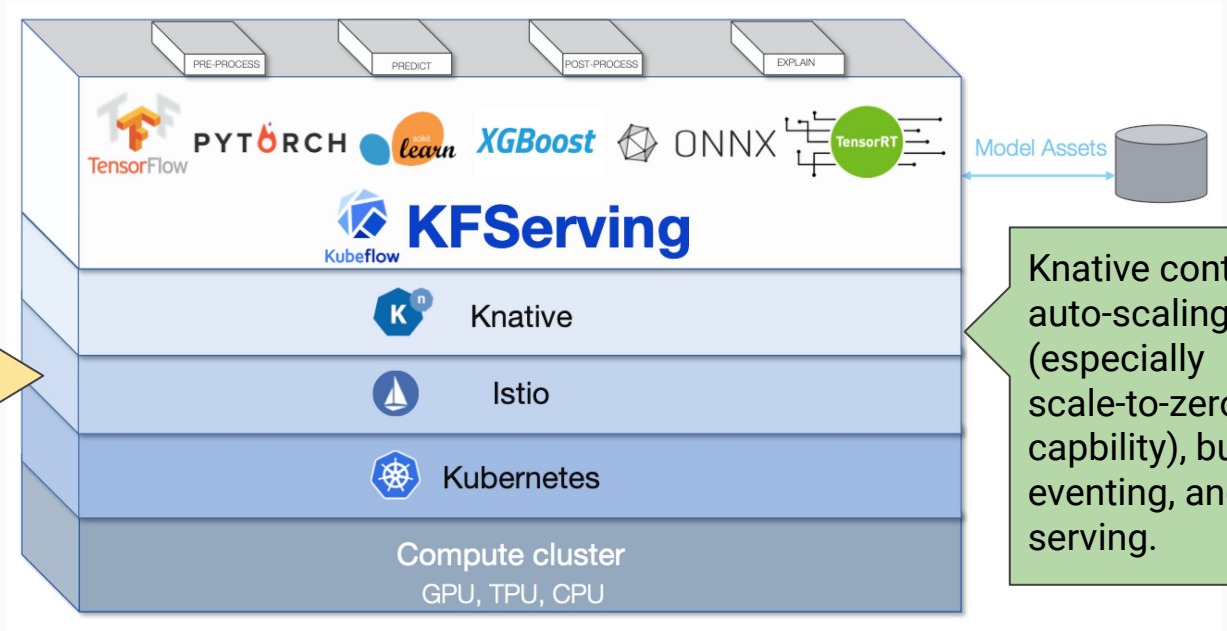
Source: https://github.com/kubeflow/katib/blob/master/docs/workflow-design.md

Model Serving

- How hard it could be to serve ML models in production scale?
  - Scale vs Cost
  - Seamless Rollout
  - Canary Rollouts
  - Service/Model monitoring

Source: https://unsplash.com/@srd844?utm_source=medium&utm_medium=referral

KServe



Istio is a service mesh, controlling who can access what services.

Knative controls auto-scaling (especially scale-to-zero capbility), building, eventing, and serving.

Source: https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

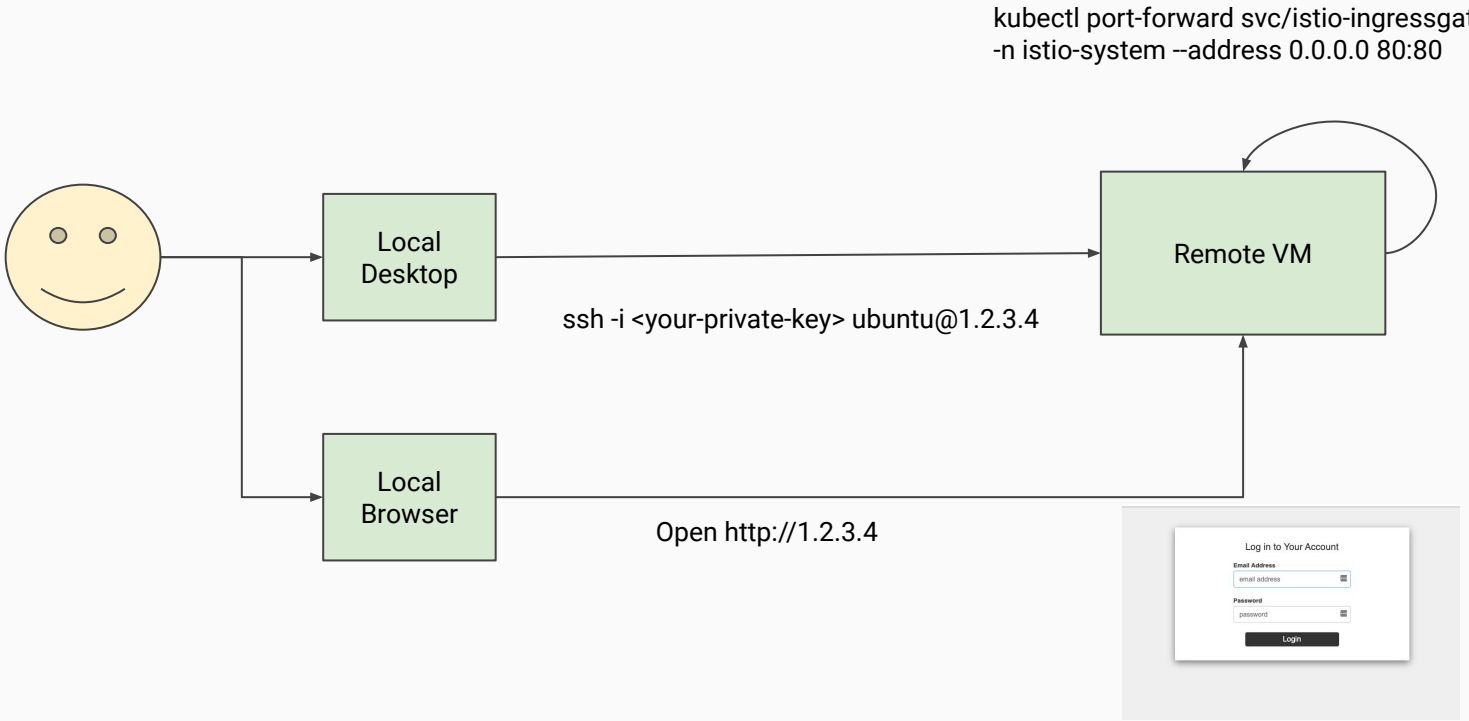
Environment Setup

Manual Installation

- Manual Installation
- https://github.com/FootprintAI/kubeflow-workshop/tree/main/install

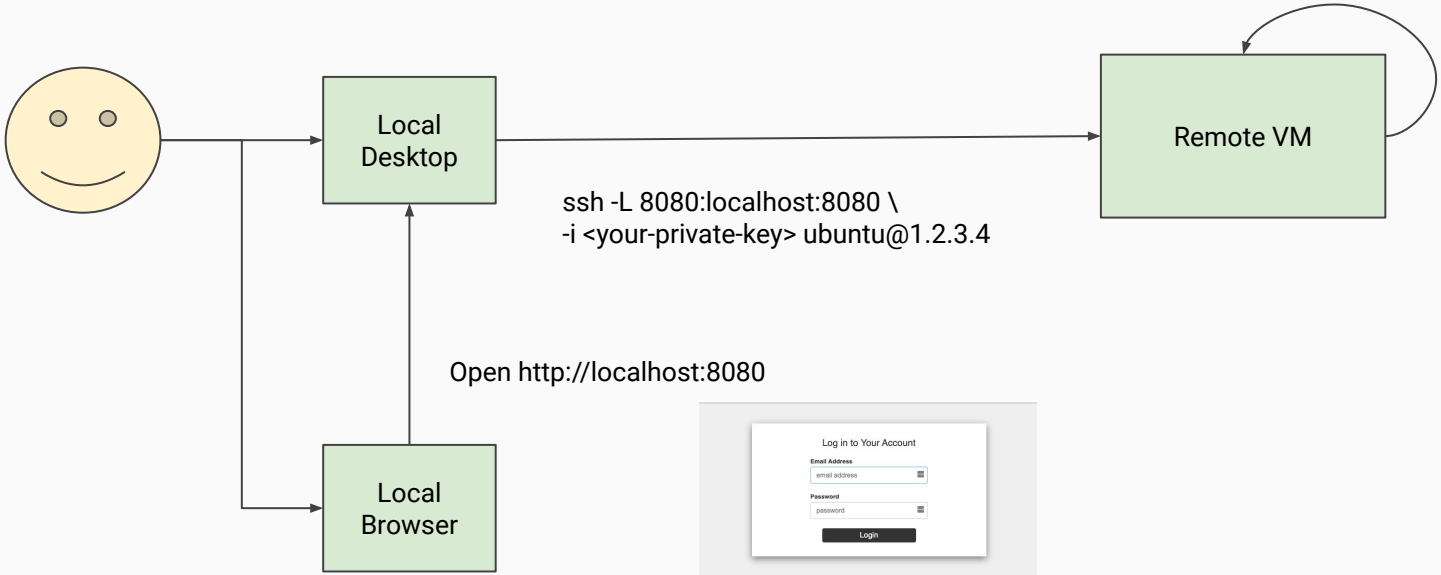
- Windows Installation
- https://github.com/FootprintAI/kubeflow-workshop/tree/main/install/windows

Manual Installation: Access Remote VM via Public IP



Manual Installation: Access Remote VM via SSH Local Forwarding

kubectl port-forward svc/istio-ingressgateway
-n istio-system 8080:80

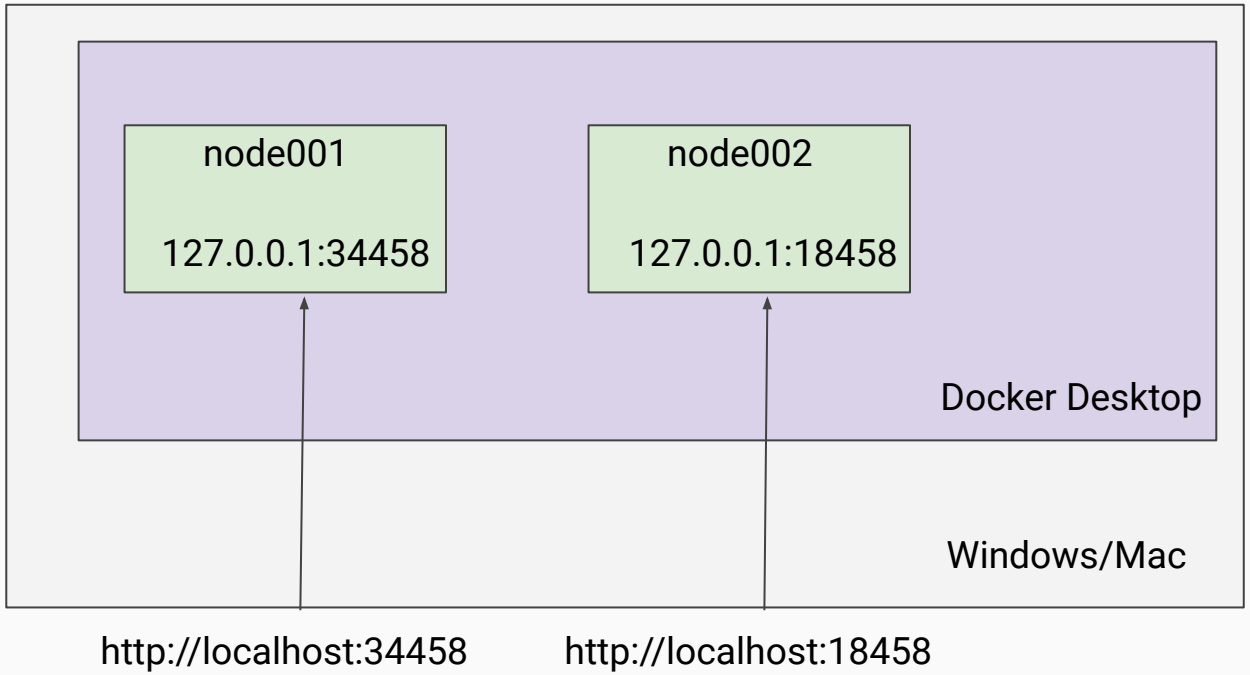


multikf: One-click Installation

```
// install dockerd
wget https://raw.githubusercontent.com/FootprintAI/multikf/main/docker/linux.sh
chmod +x linux.sh
sudo ./linux.sh

// install multikf
wget https://github.com/FootprintAI/multikf/raw/main/build/multikf.linux
chmod +x multikf.linux

// add an instances with port 80/443 exported
./multikf.linux add node002 --export_ports 80:80,443:443

// connect kubeflow
./multikf.linux connect kubefloe node002
```

multikf: One-click Installation

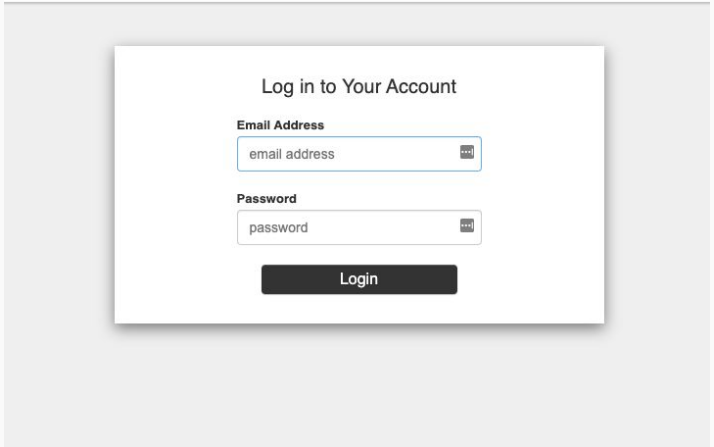
- Multikf: https://github.com/footprintai/multikf



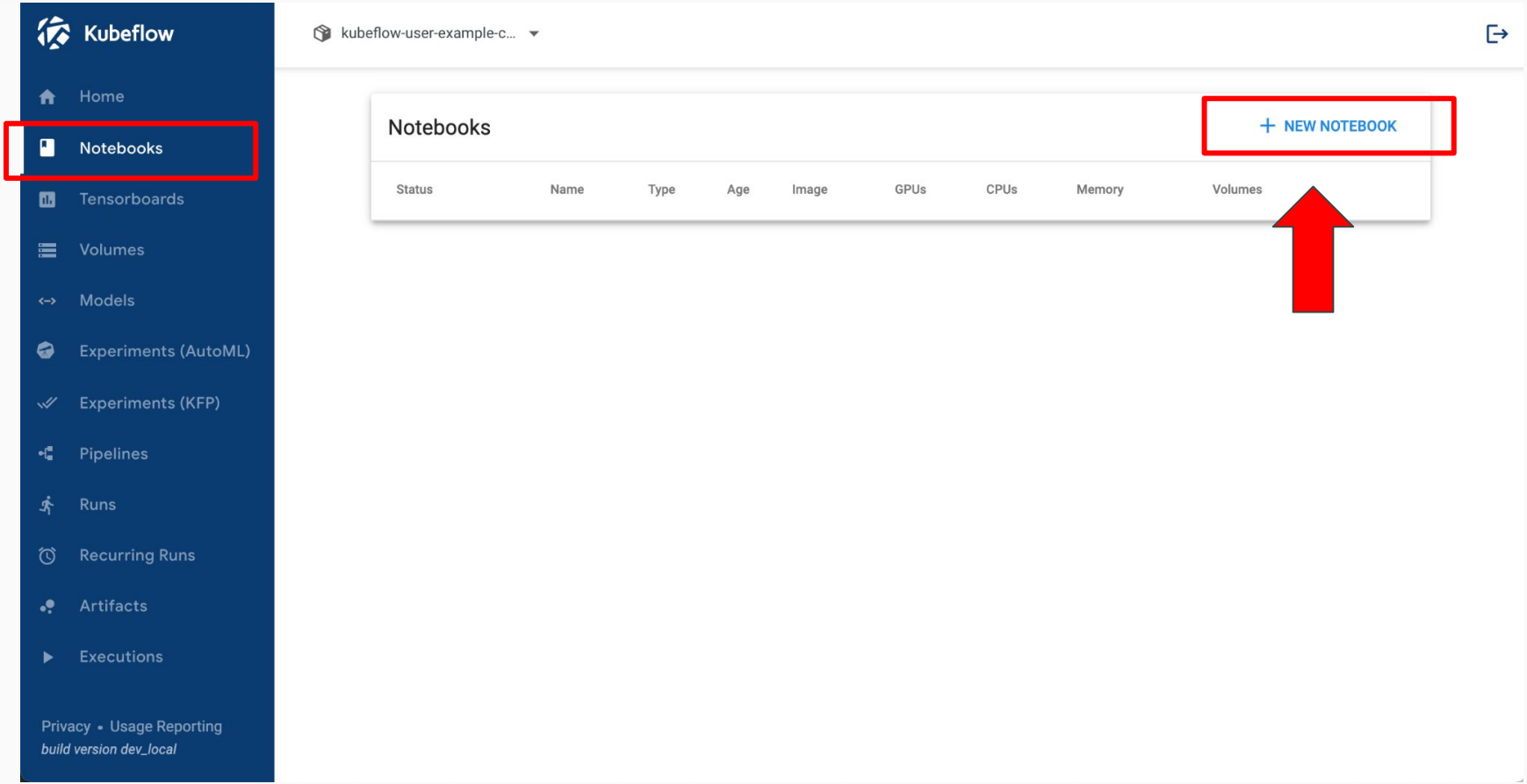
Hands-on Time

Wait! 所以我說那個帳號密碼呢?

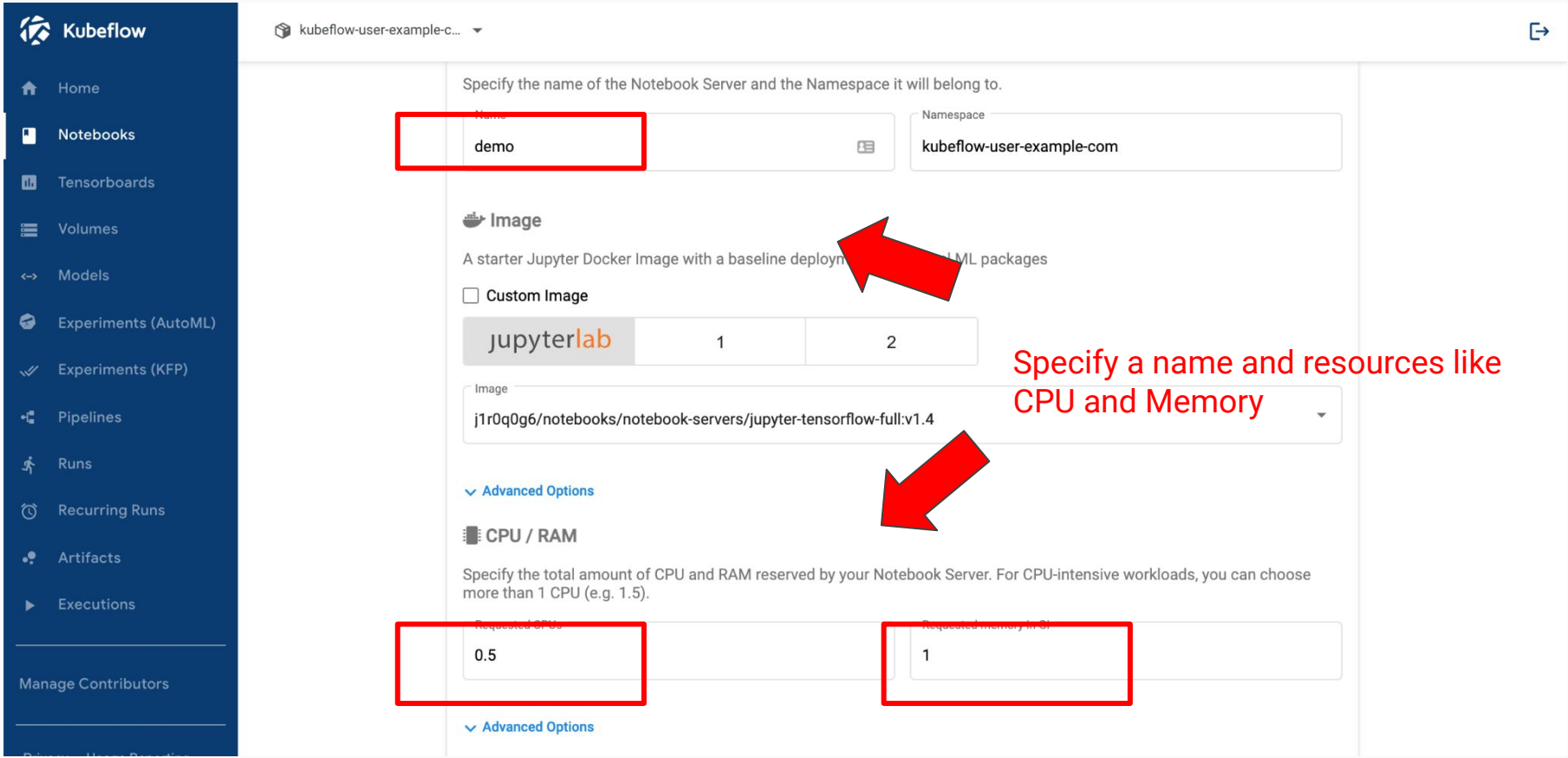
Account: user@example.com
Password: 12341234



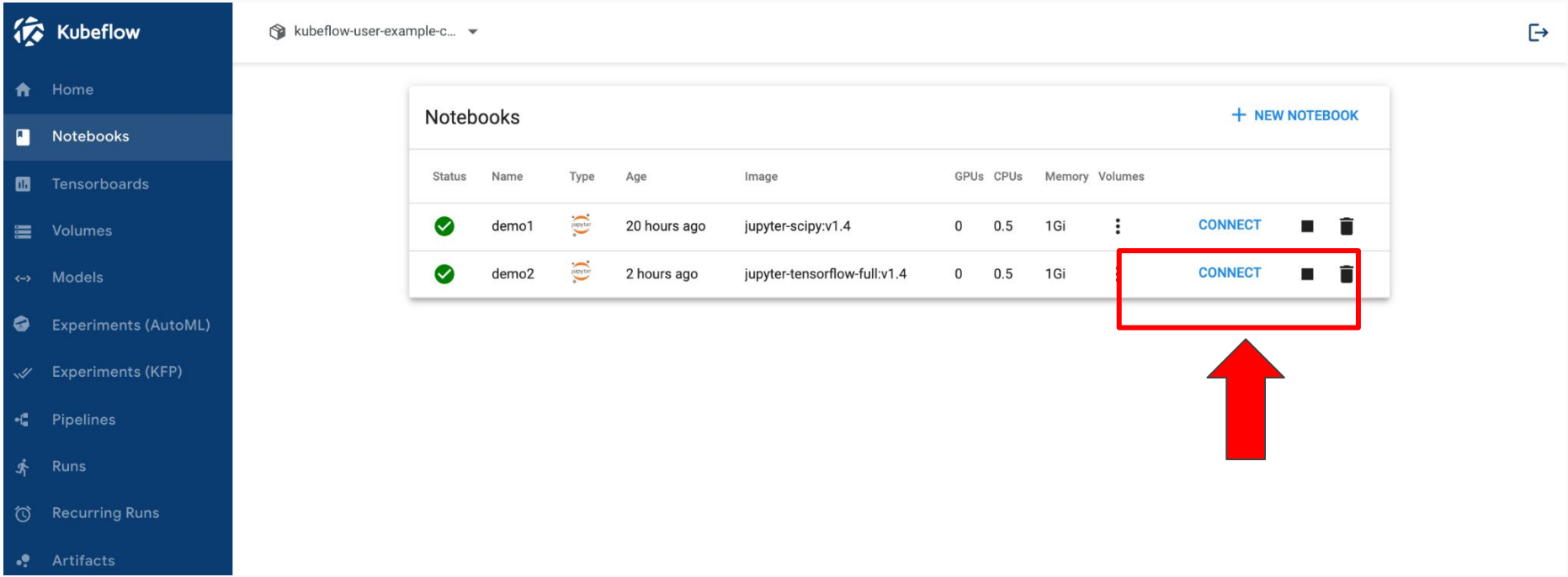
Step1: Use Notebook as Online IDE (1/3)



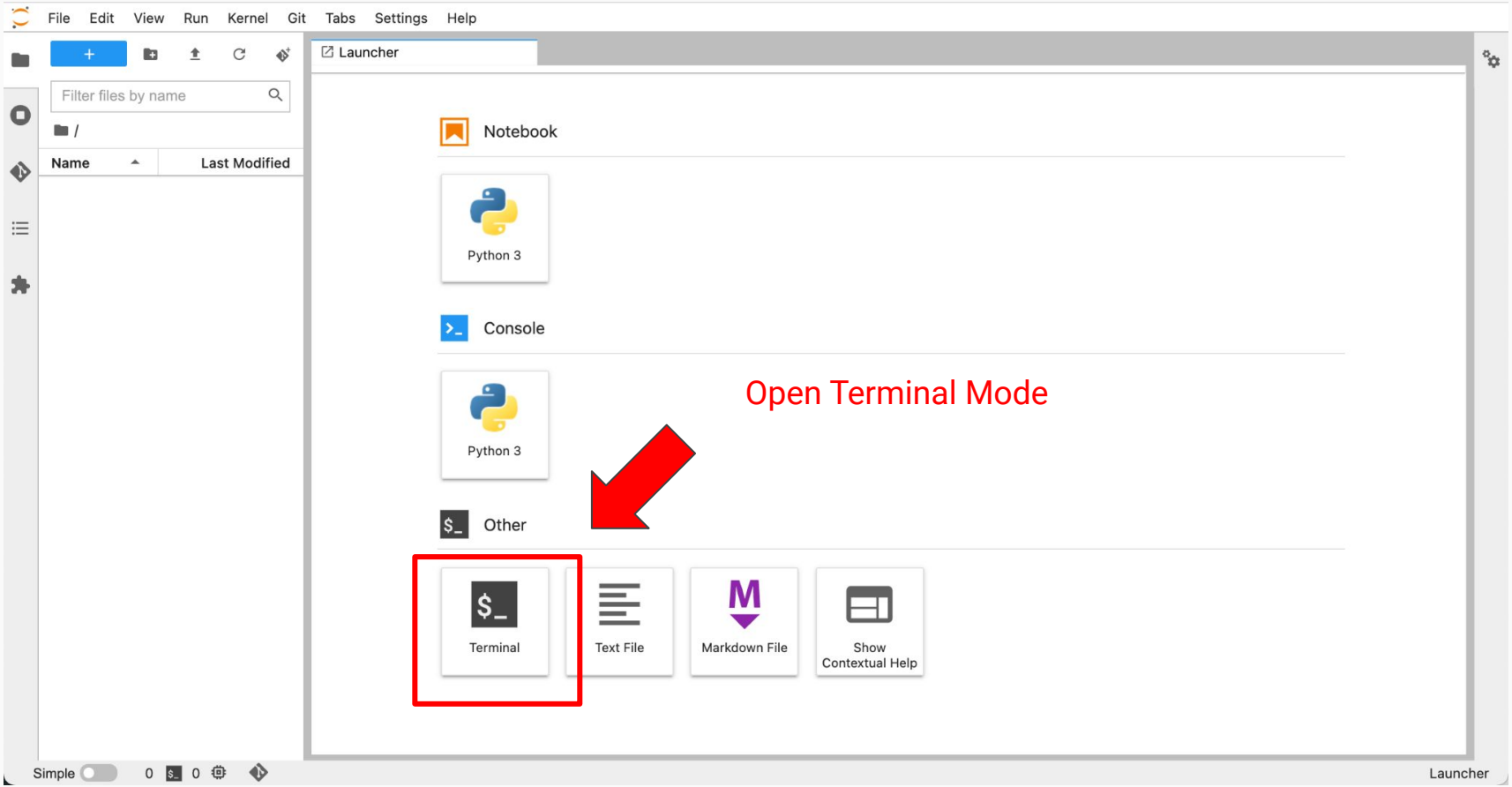
Step1: Use Notebook as Online IDE (2/3)



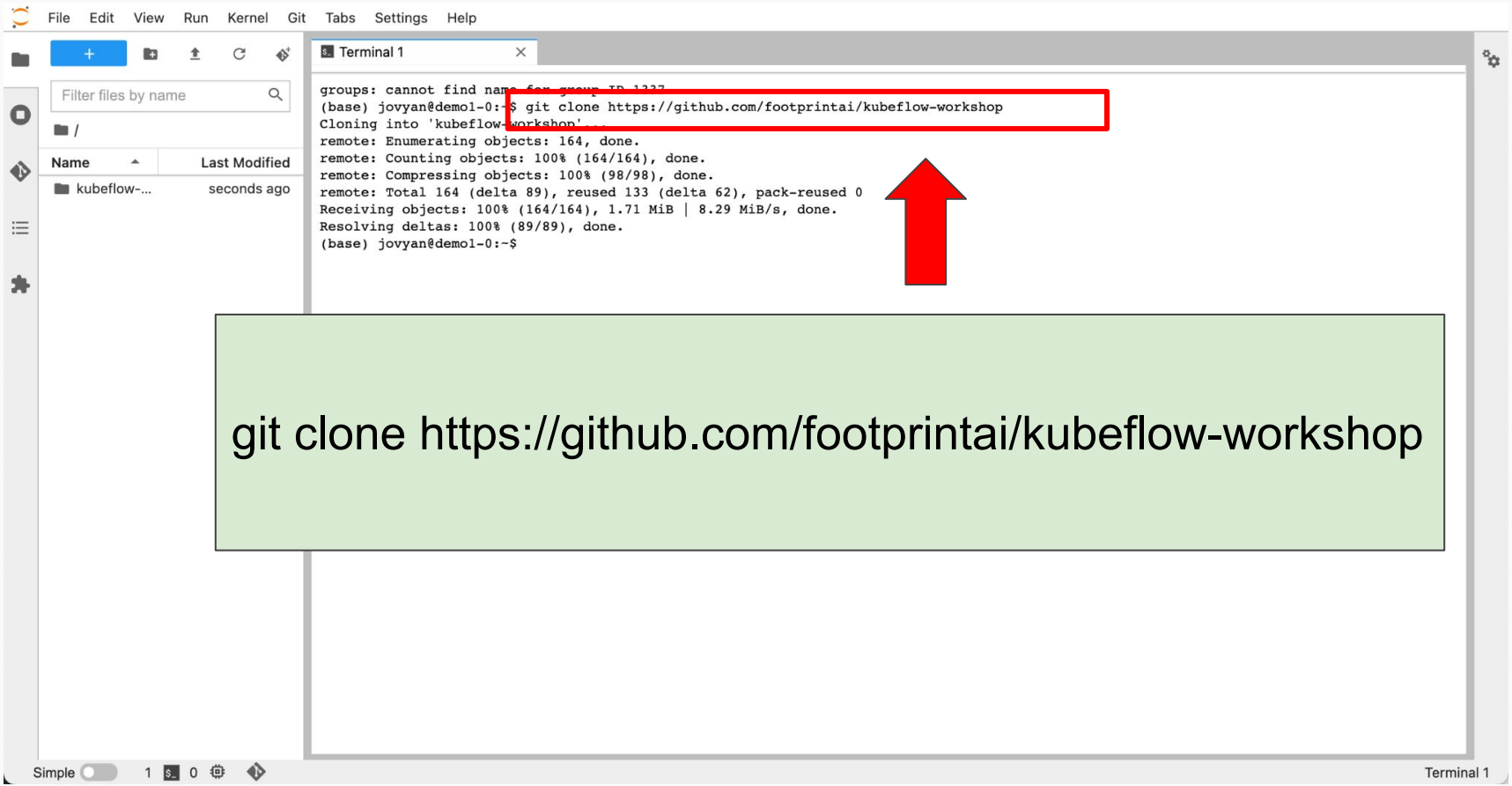
Step1: Use Notebook as Online IDE (3/3)



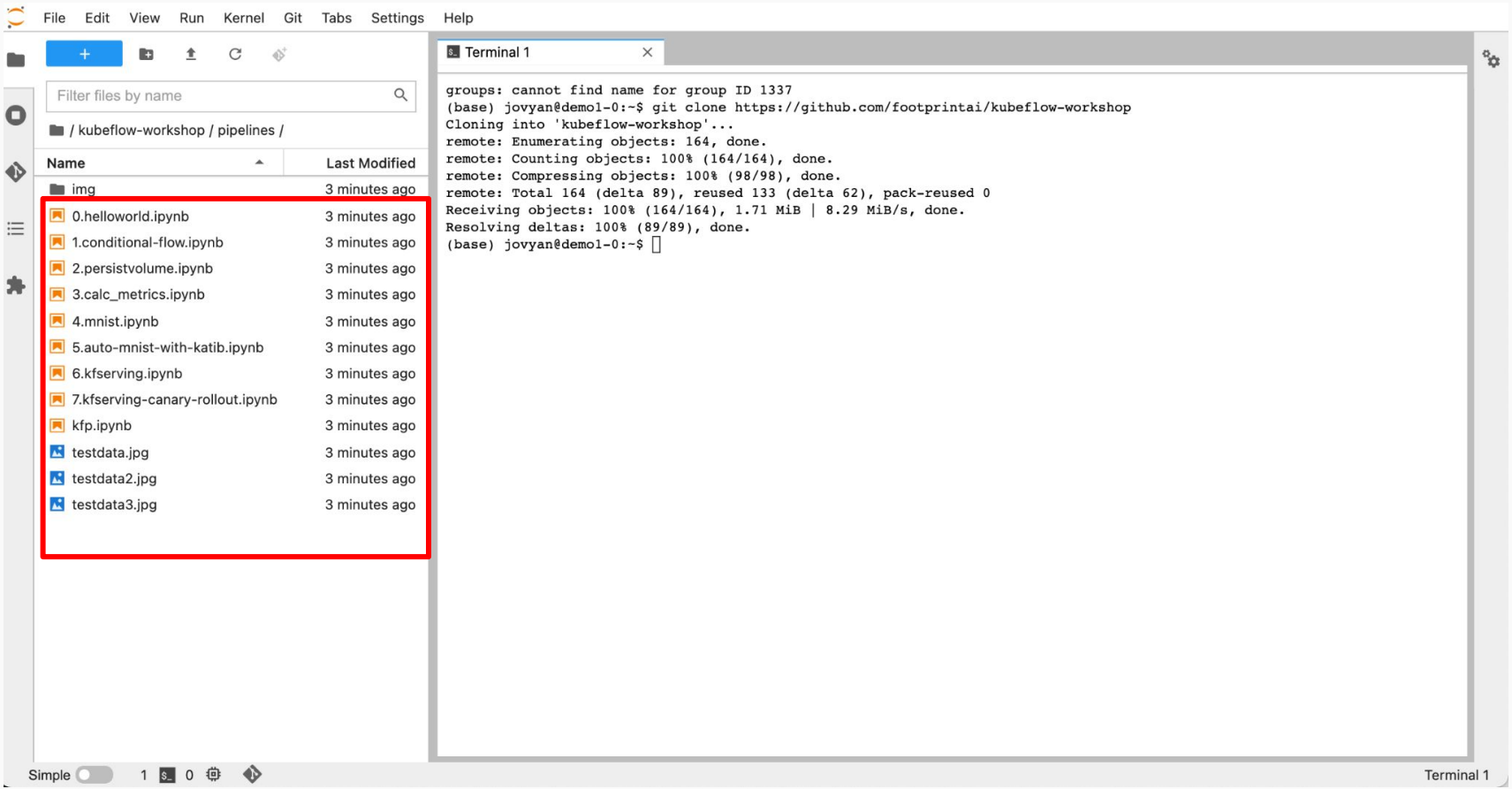
Step2: Use terminal to download the materials (1/3)



Step2: Use terminal to download the materials (2/3)

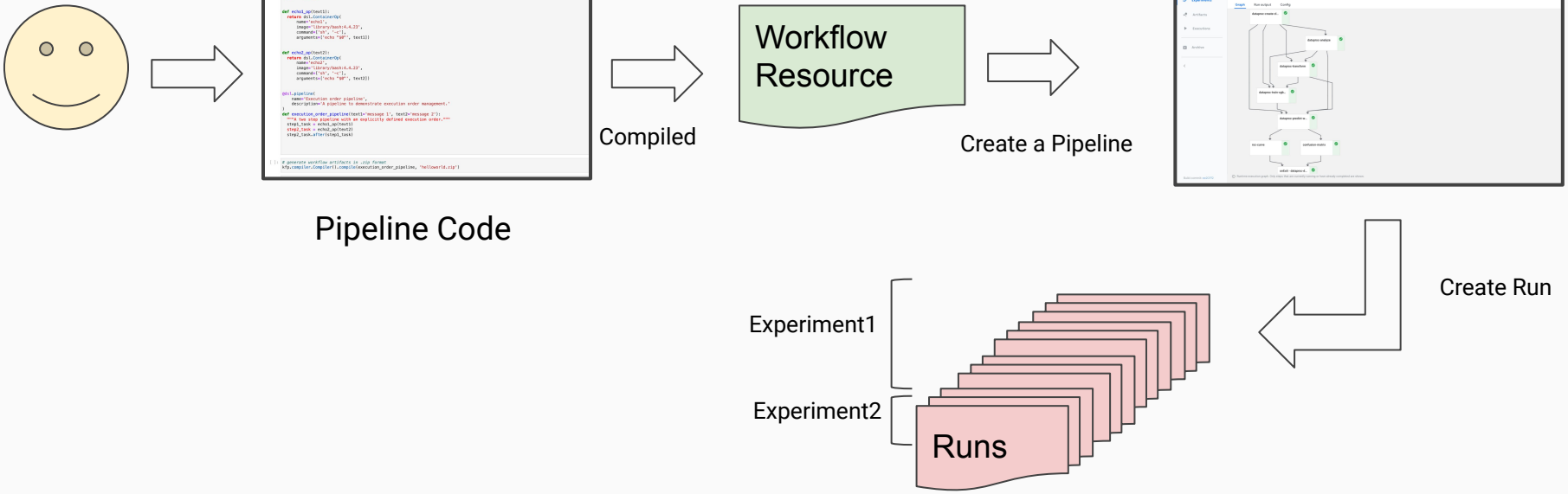


Step2: Use terminal to download the materials (3/3)



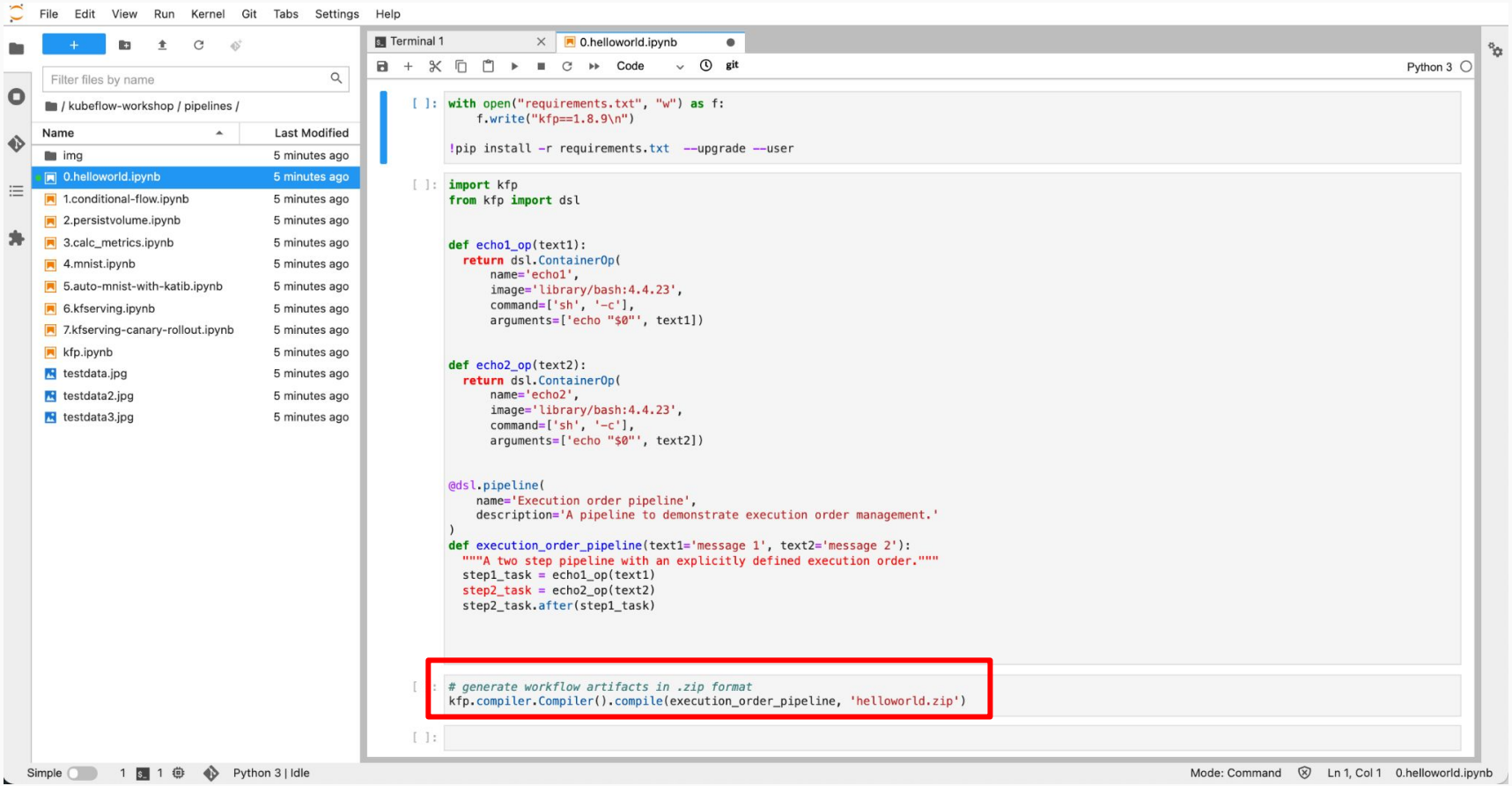
Kubeflow Terms

詞彙說明

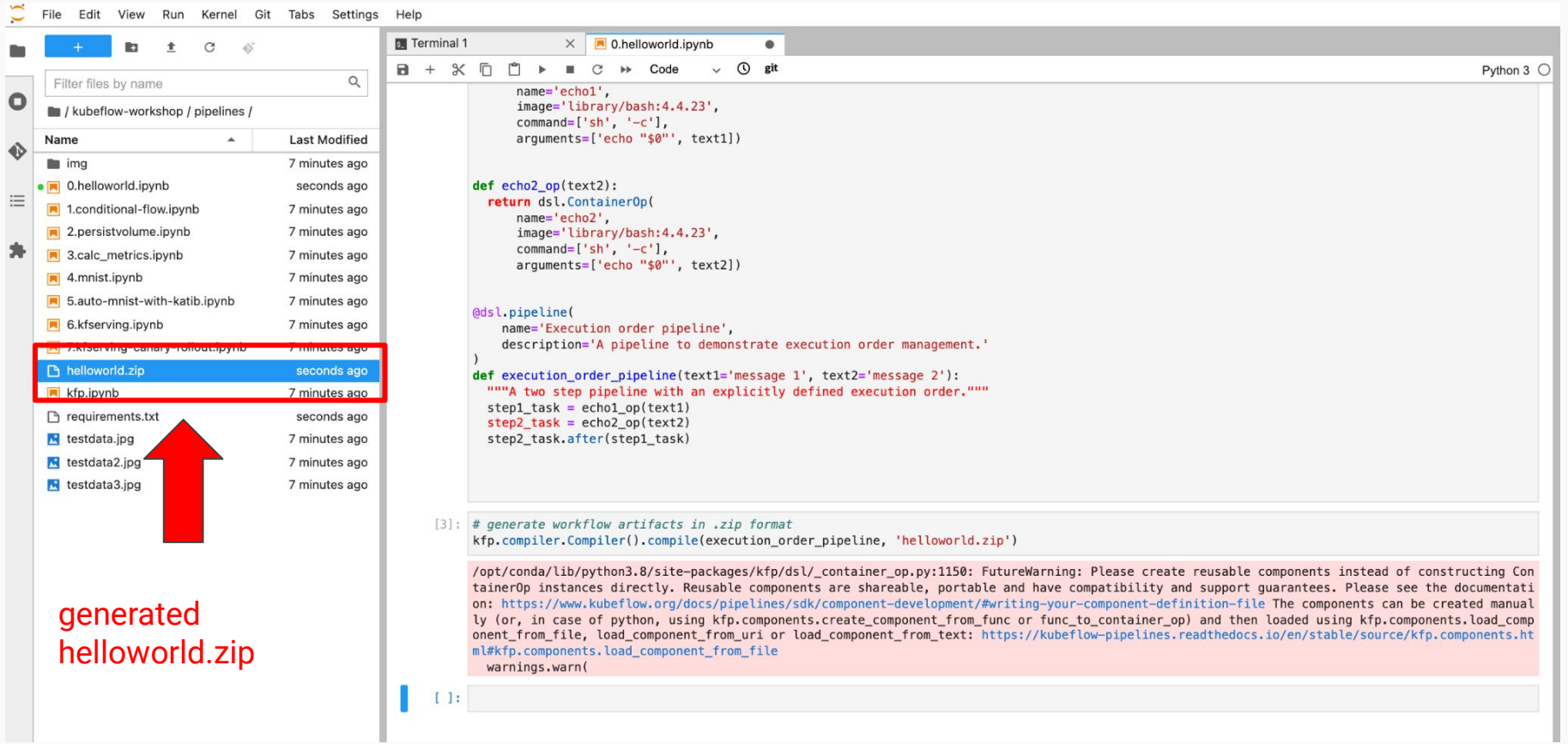

Pipeline Code
Compiled
Workflow Resource
Create a Pipeline
Create Run
Experiment1
Experiment2
Runs

Hello World Example

Step4: Compile helloworld.ipynb (1/2)

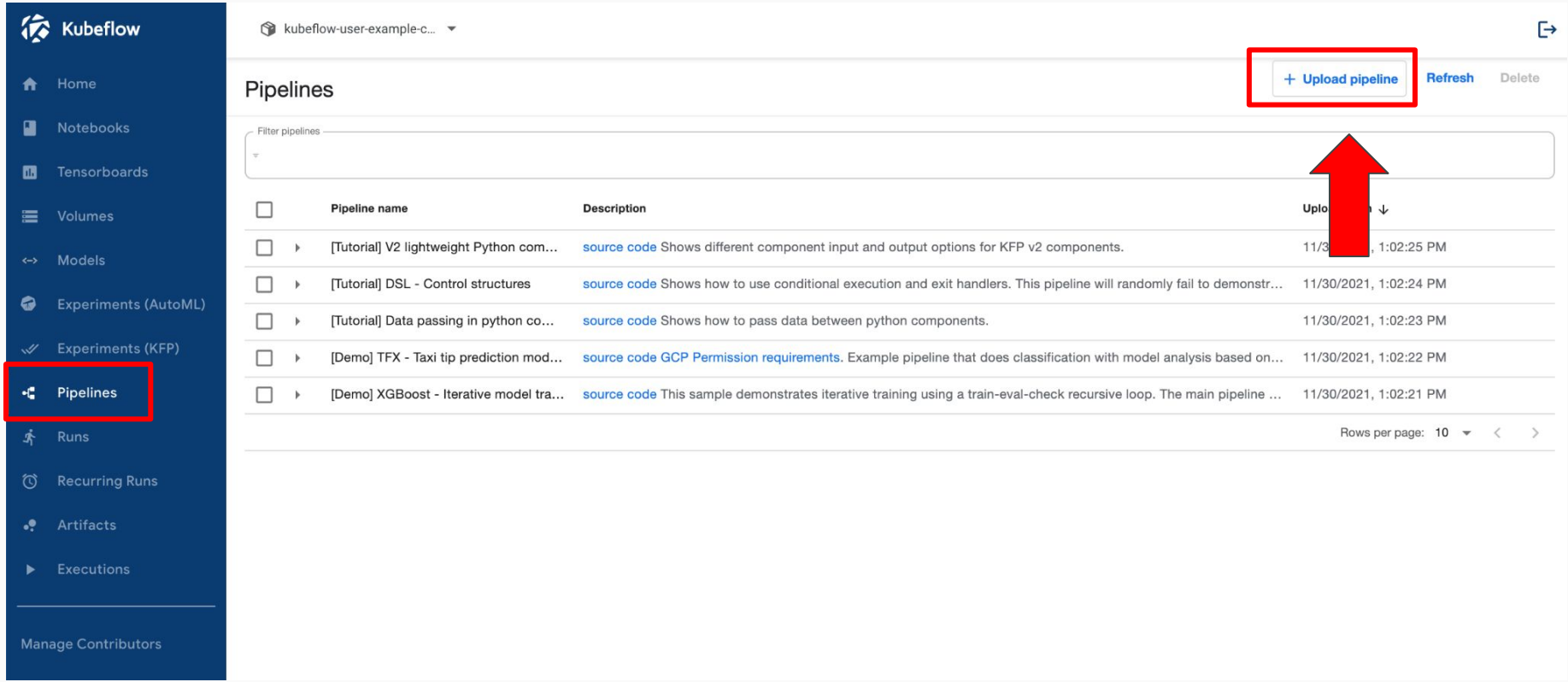


Step4: Compile helloworld.ipynb (2/2)

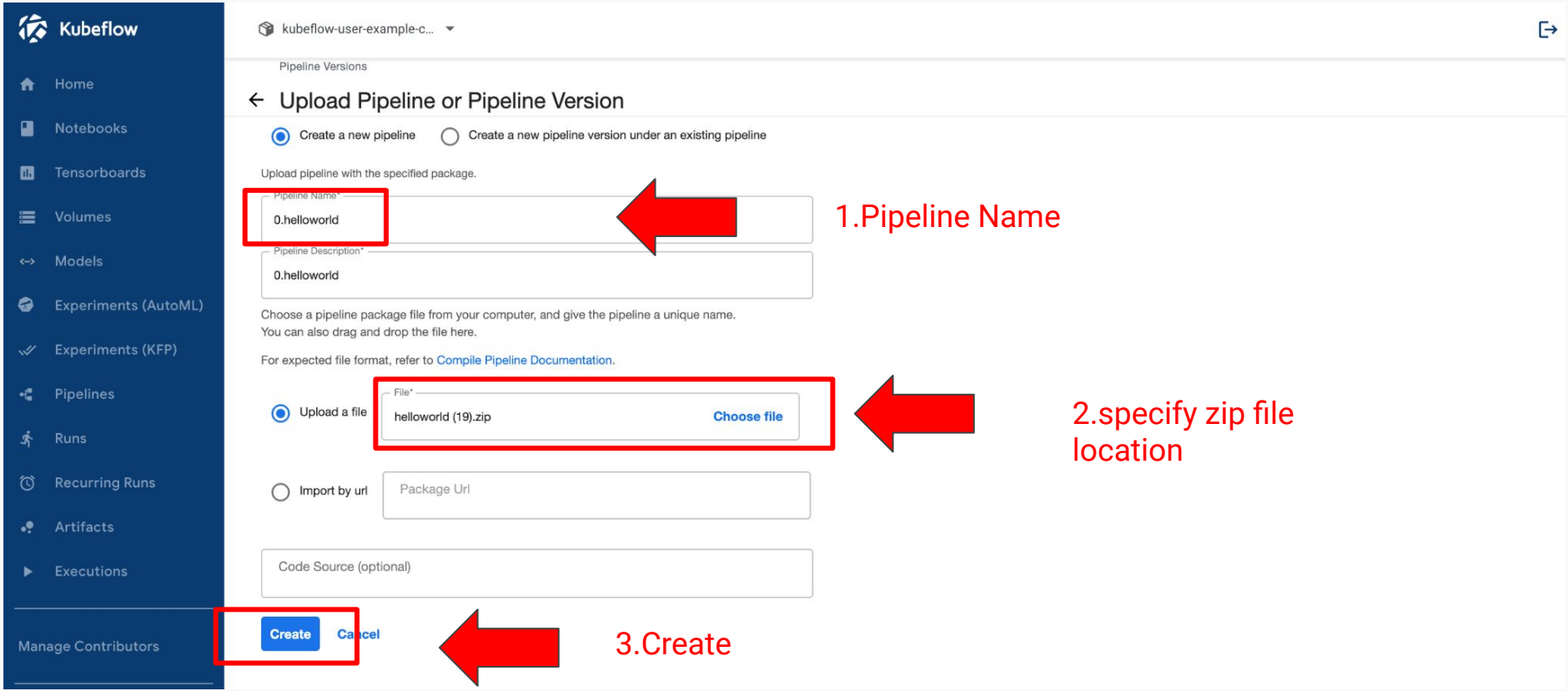


generated helloworld.zip

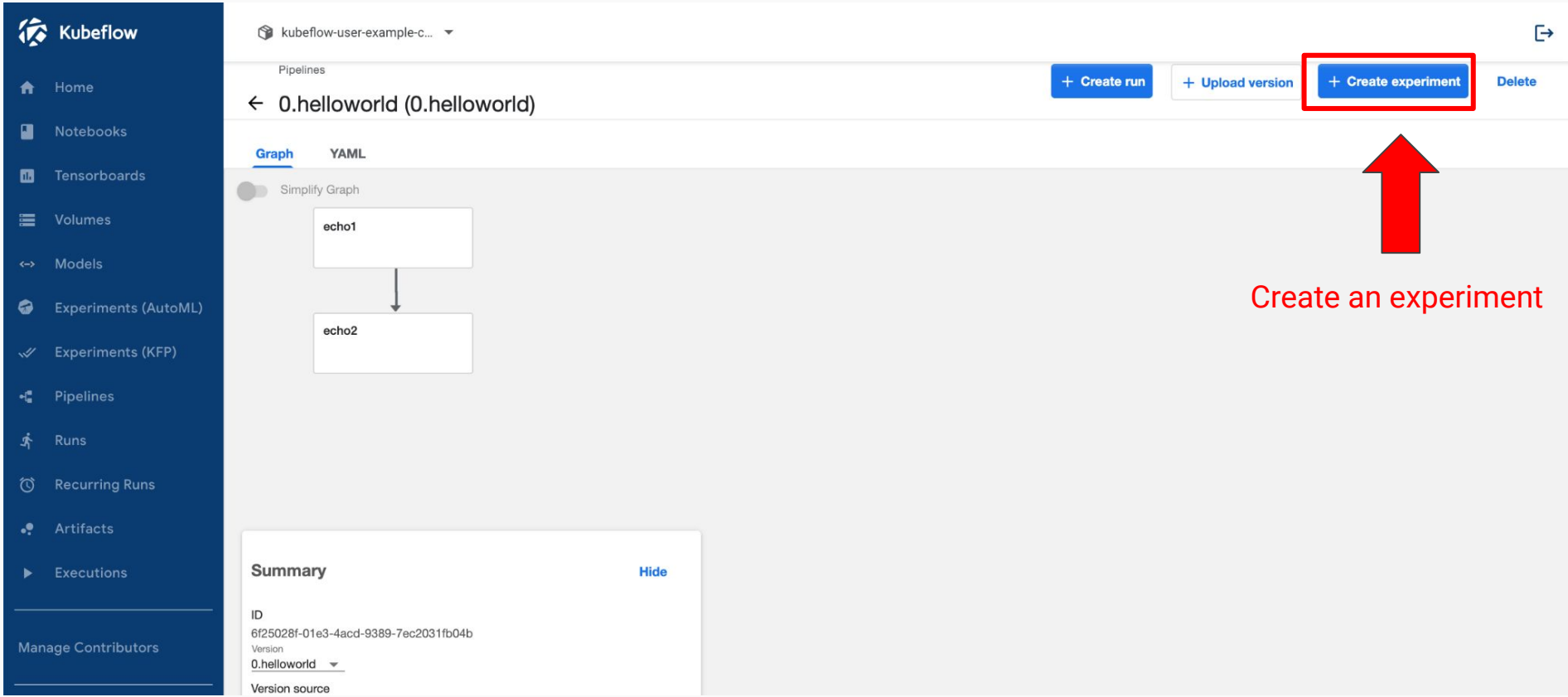
Step4: Create a Pipeline (1/7)



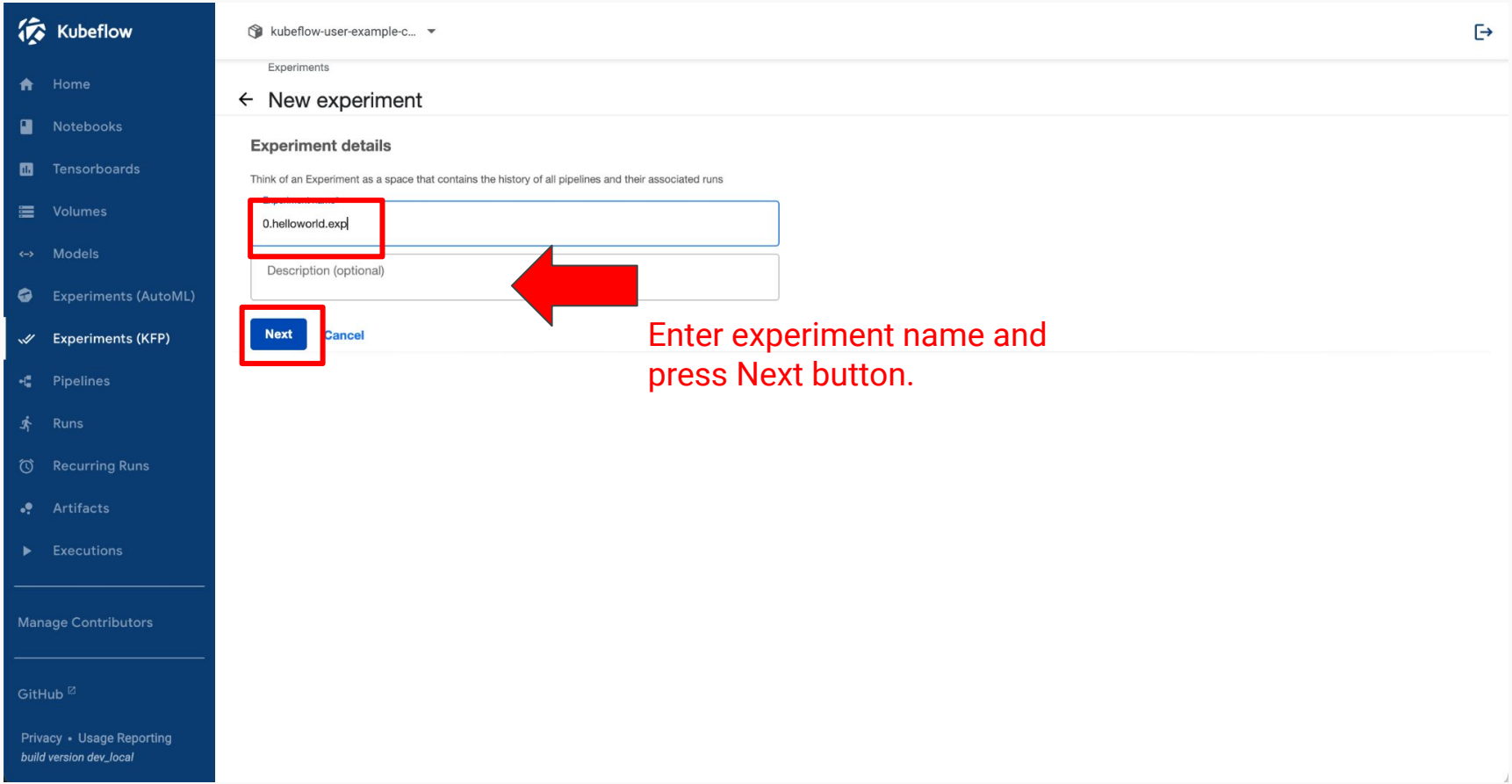
Step4: Create a Pipeline (2/7)



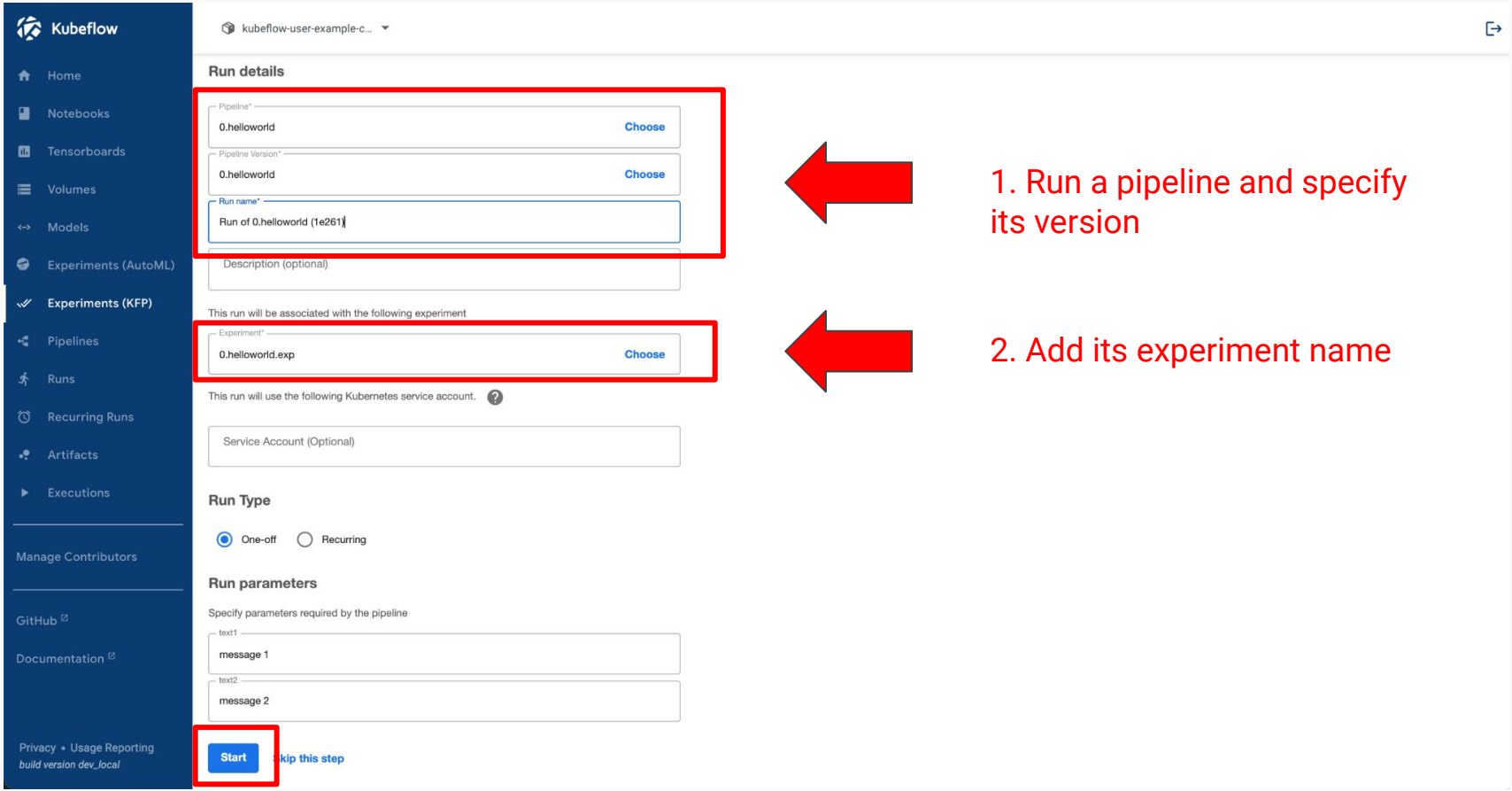
Step4: Create a Pipeline (3/7)



Step4: Create a Pipeline (4/7)



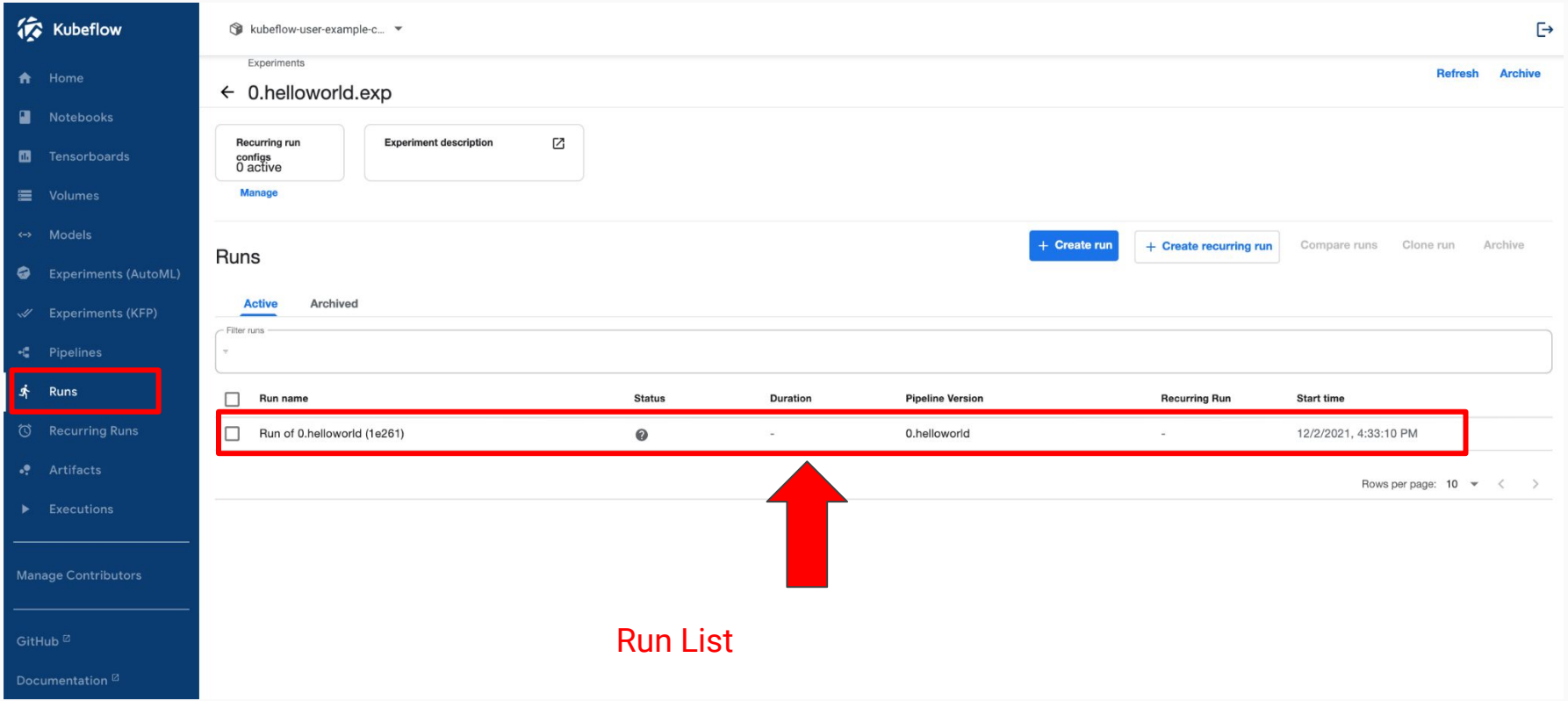
Step4: Create a Pipeline (5/7)



1. Run a pipeline and specify its version

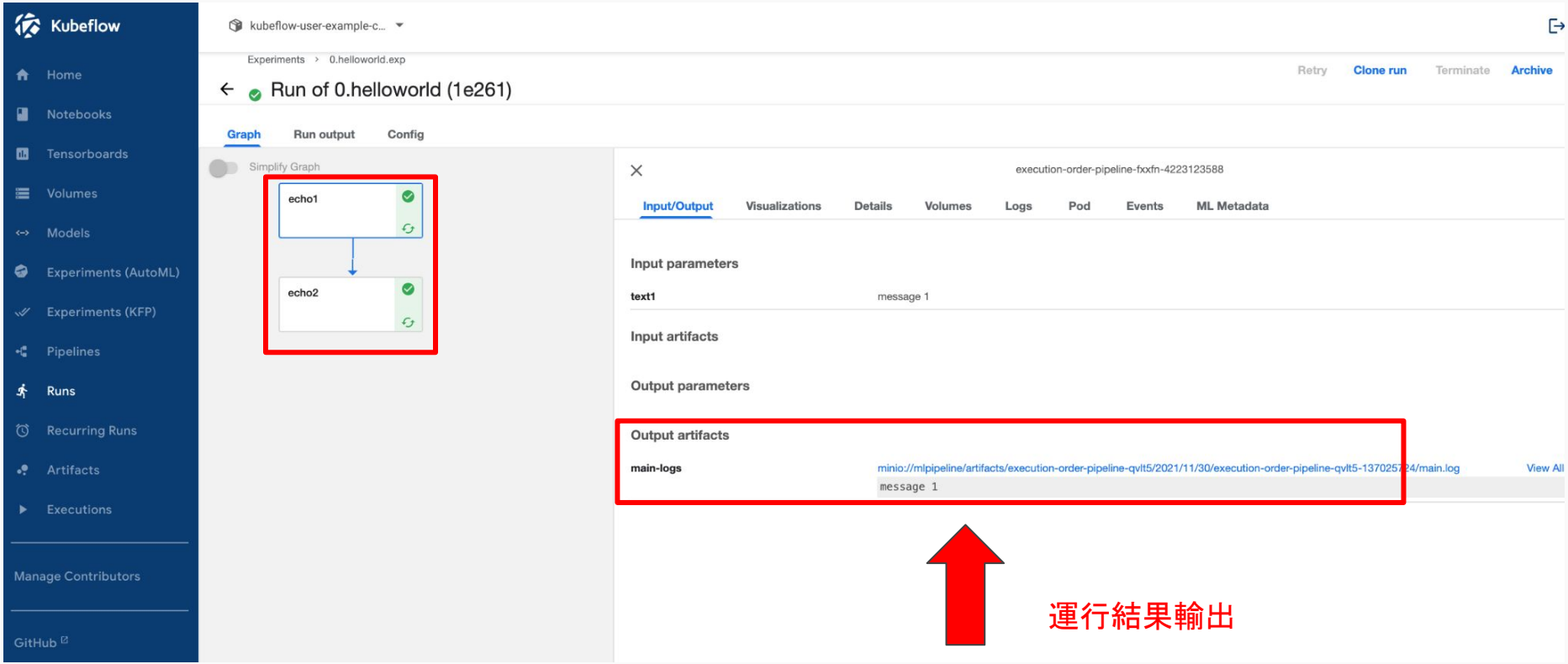
2. Add its experiment name

Step4: Create a Pipeline (6/7)



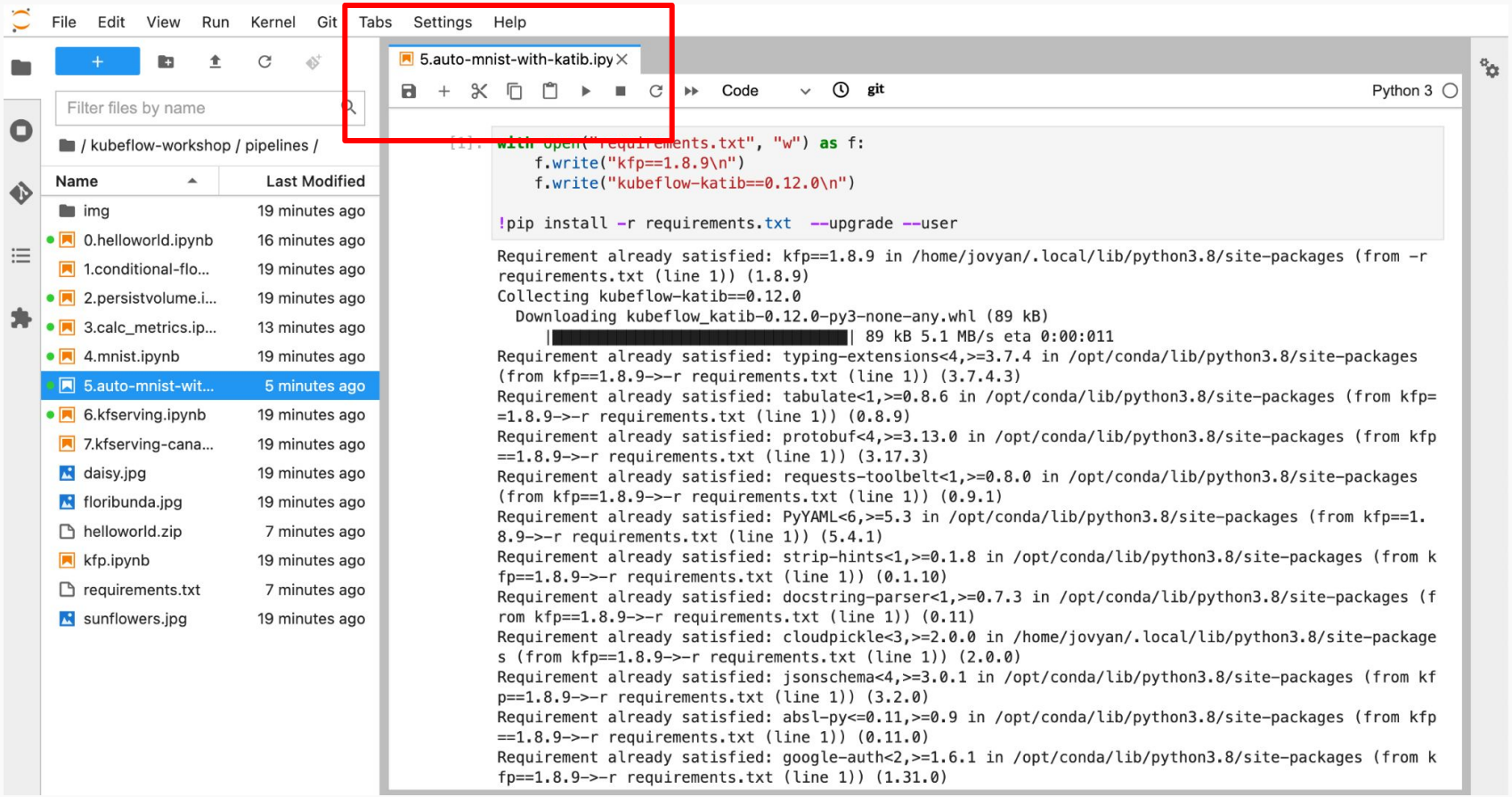
Run List

Step4: Create a Pipeline (7/7)

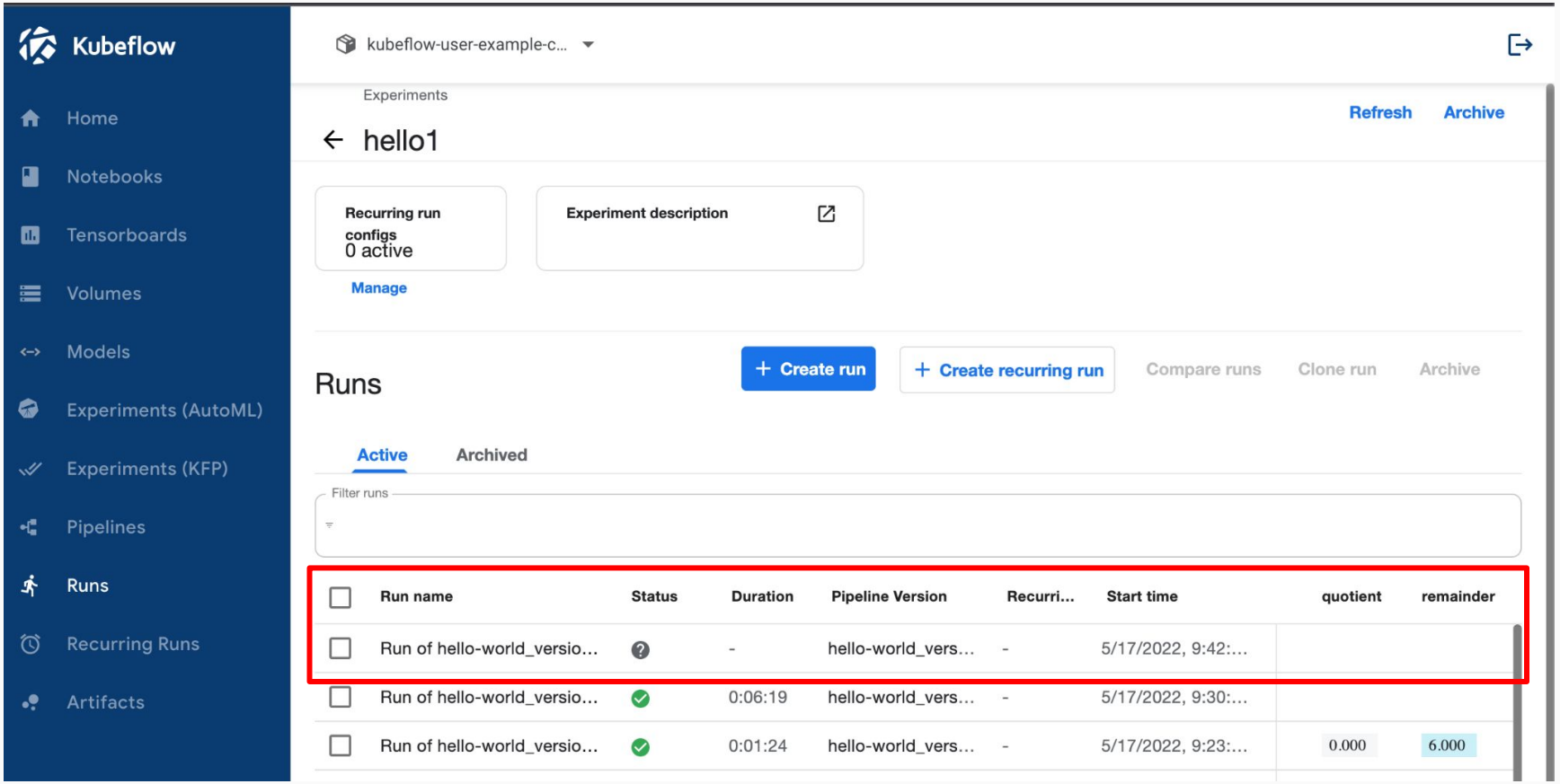


Hyperparameter Example

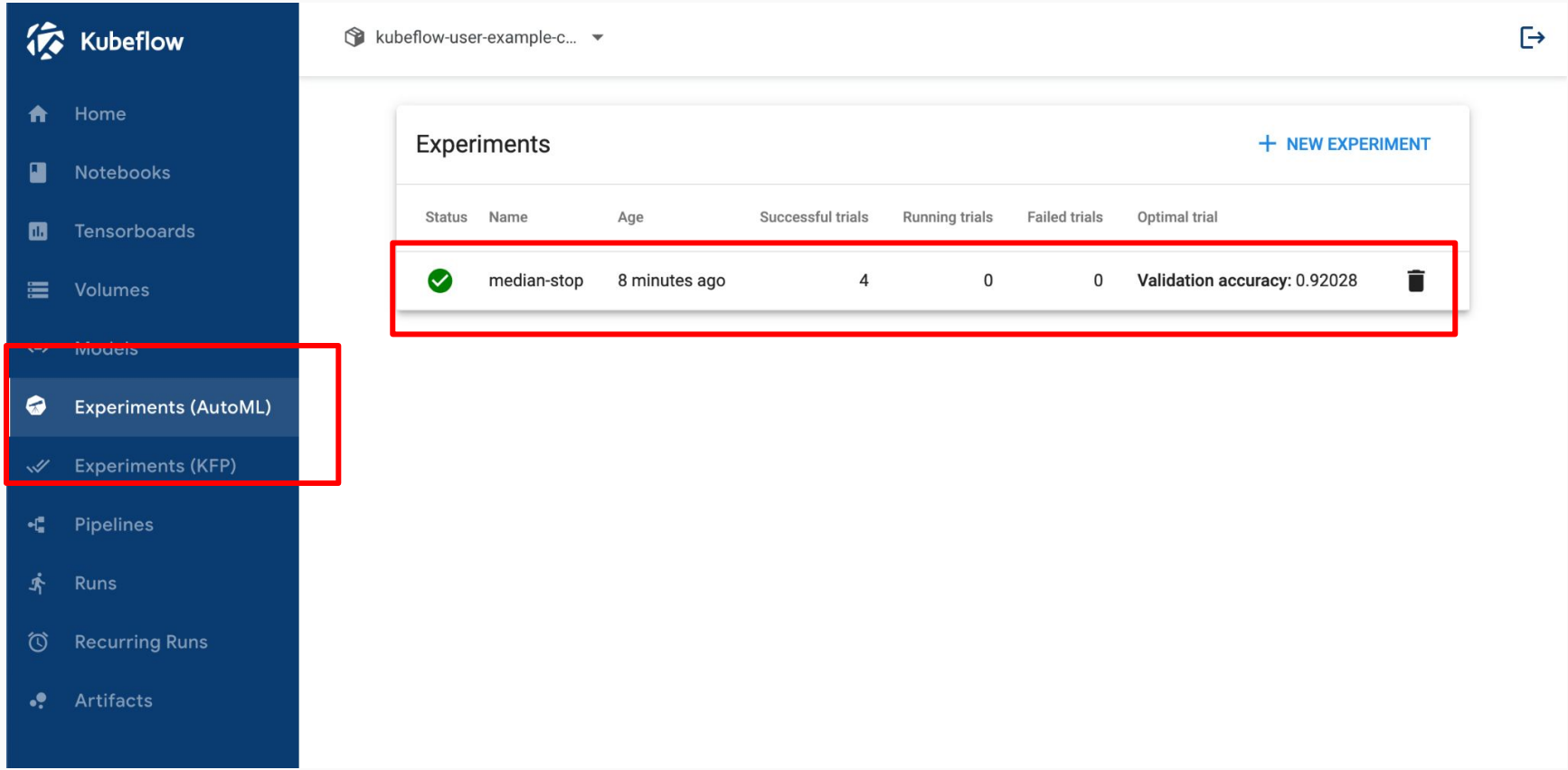
Step5: Hyperparameter tuning with katib (1/4)



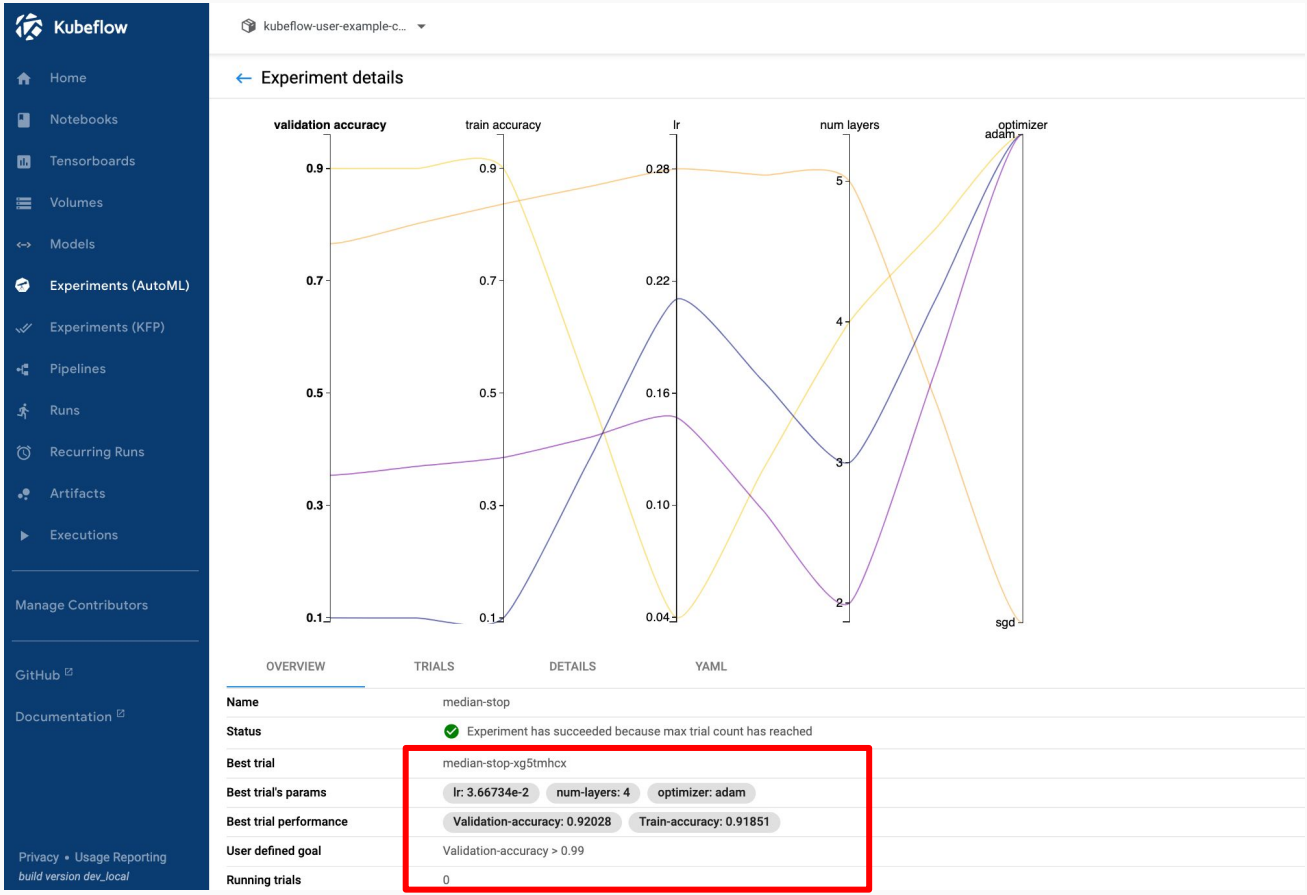
Step5: Hyperparameter tuning with katib (2/4)



Step5: Hyperparameter tuning with katib (3/4)

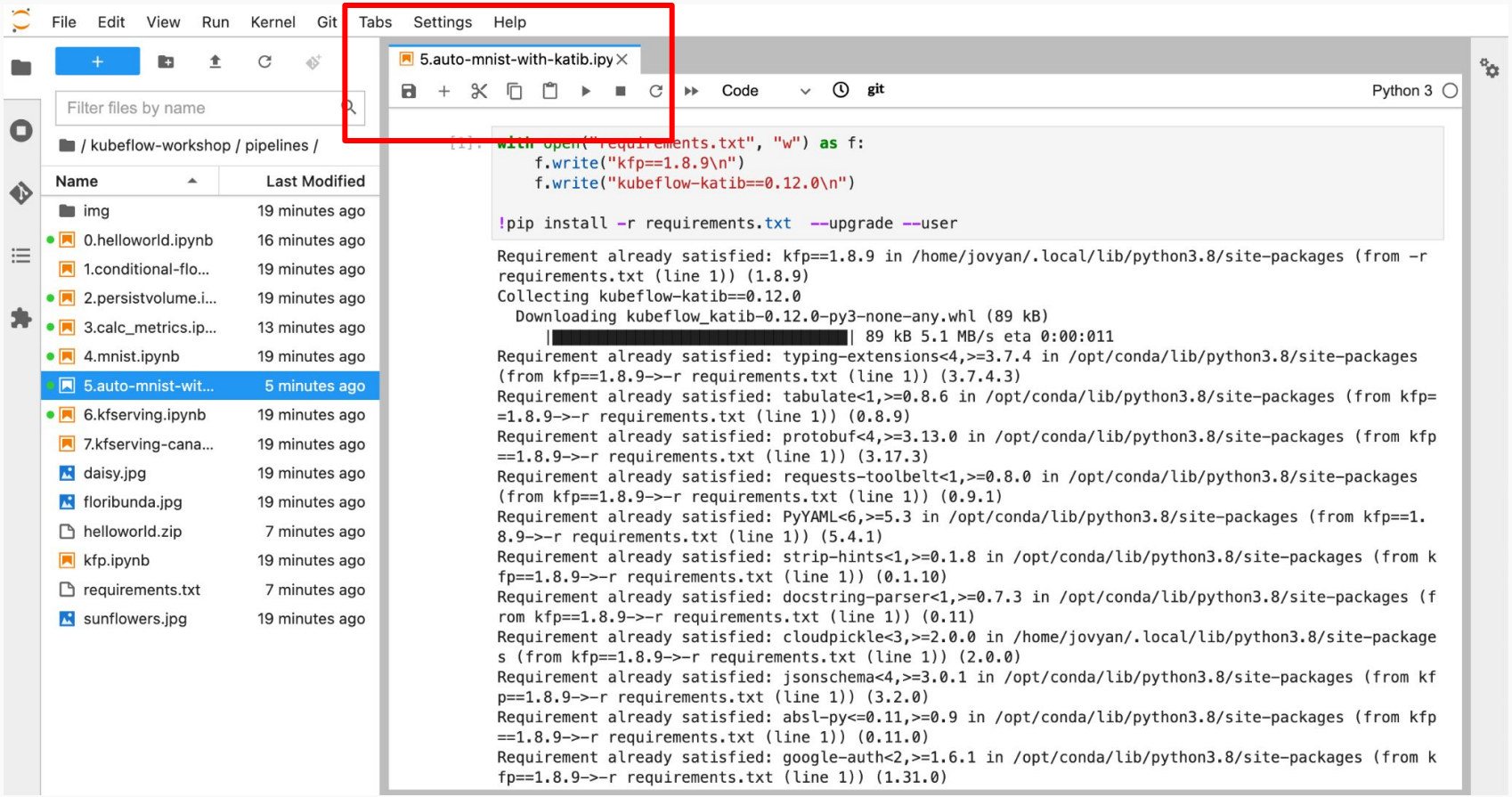


Step5: Hyperparameter tuning with katib (4/4)

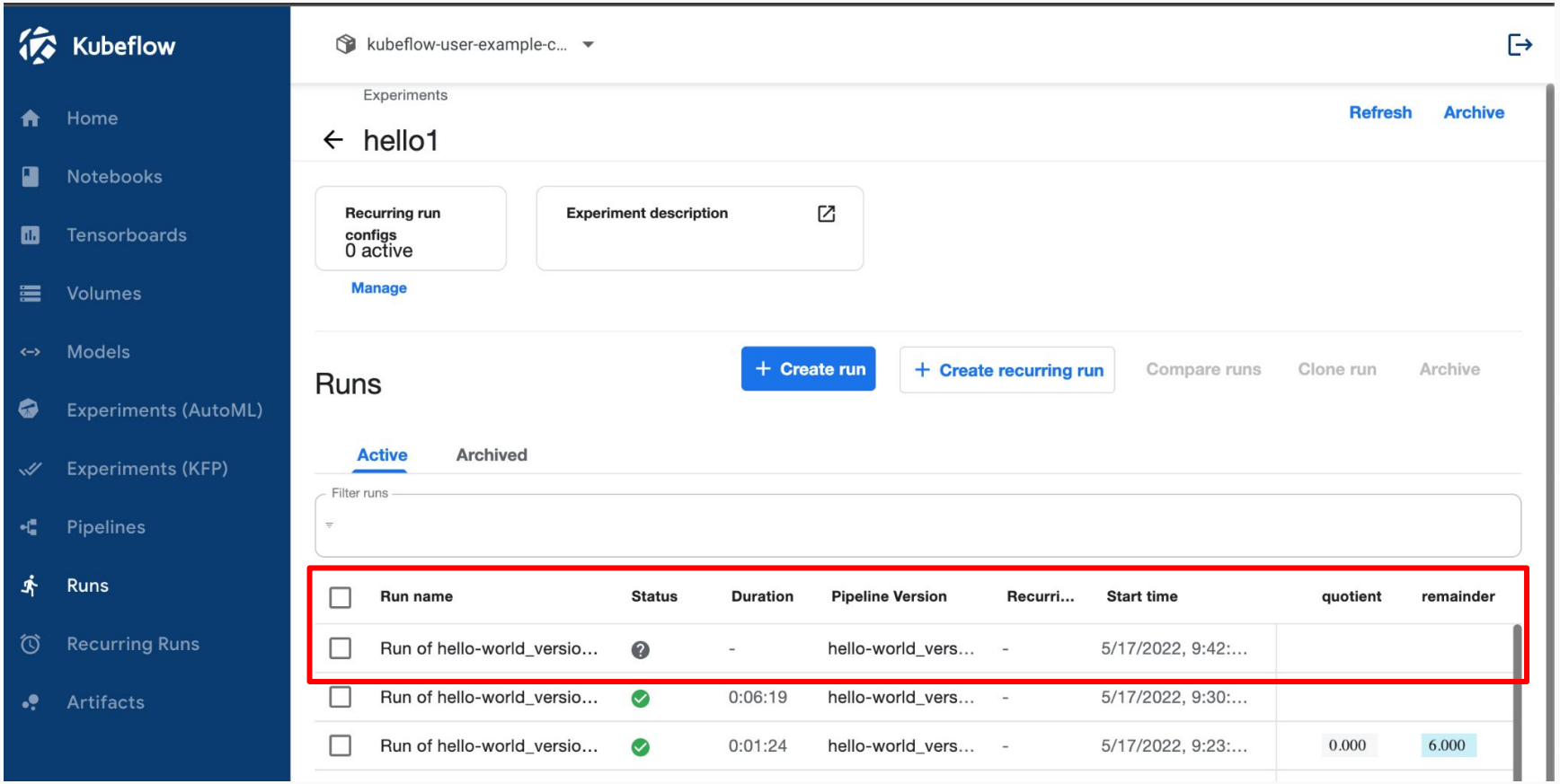


Kserve Example

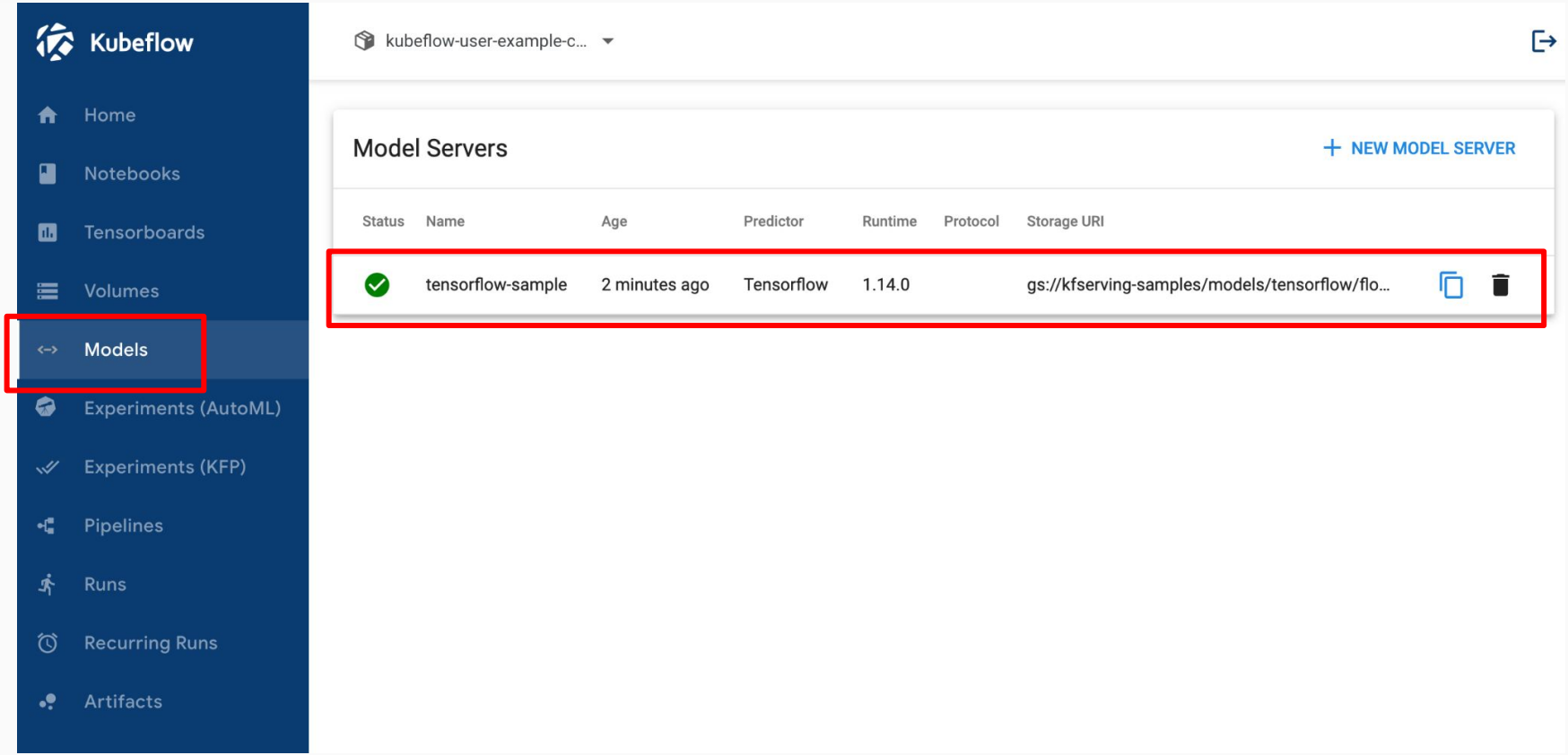
Step6: Serving model with Kserve (1/4)



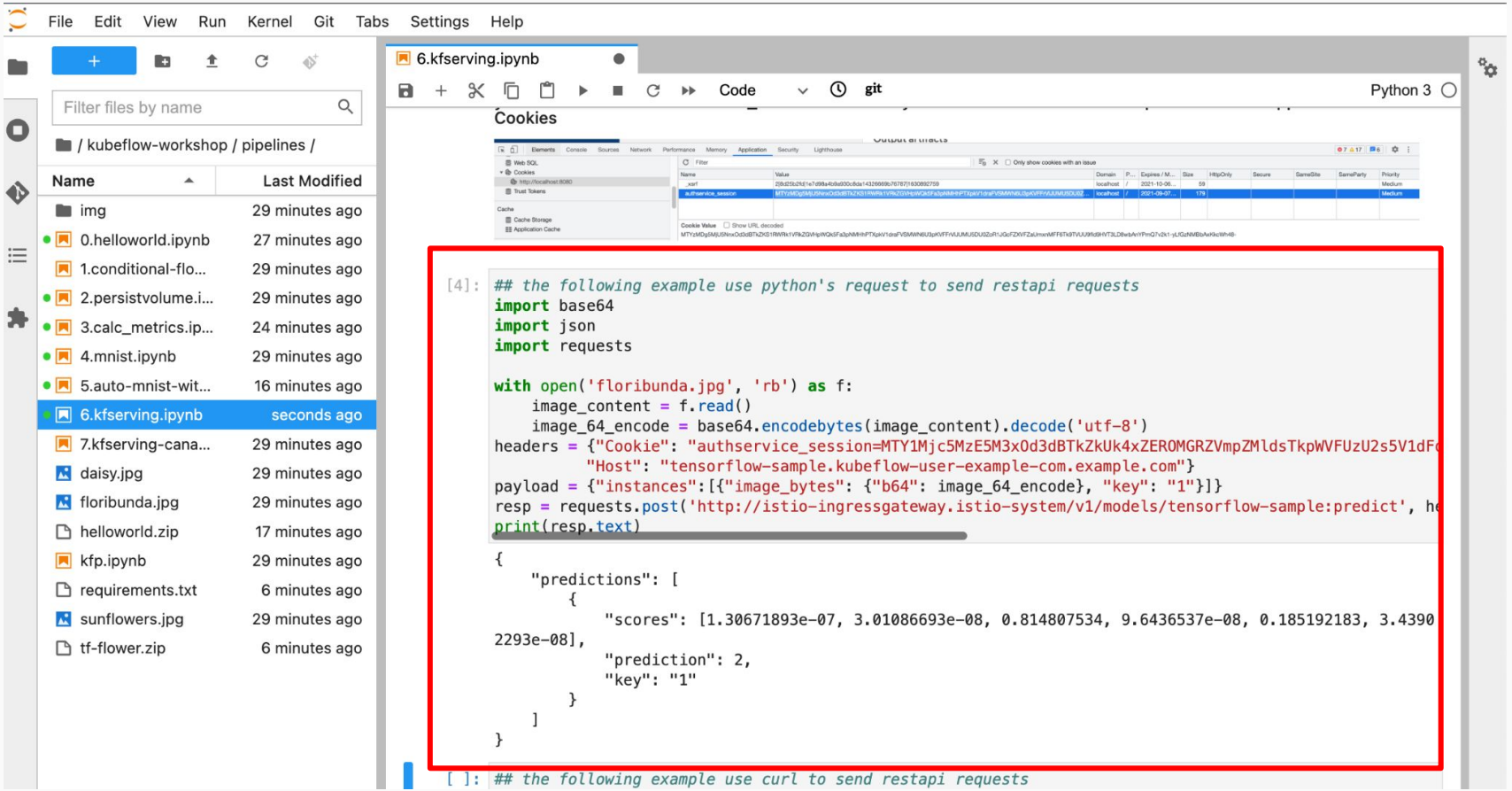
tep6: Serving model with Kserve (2/4)



Step5: Hyperparameter tuning with katib (3/4)



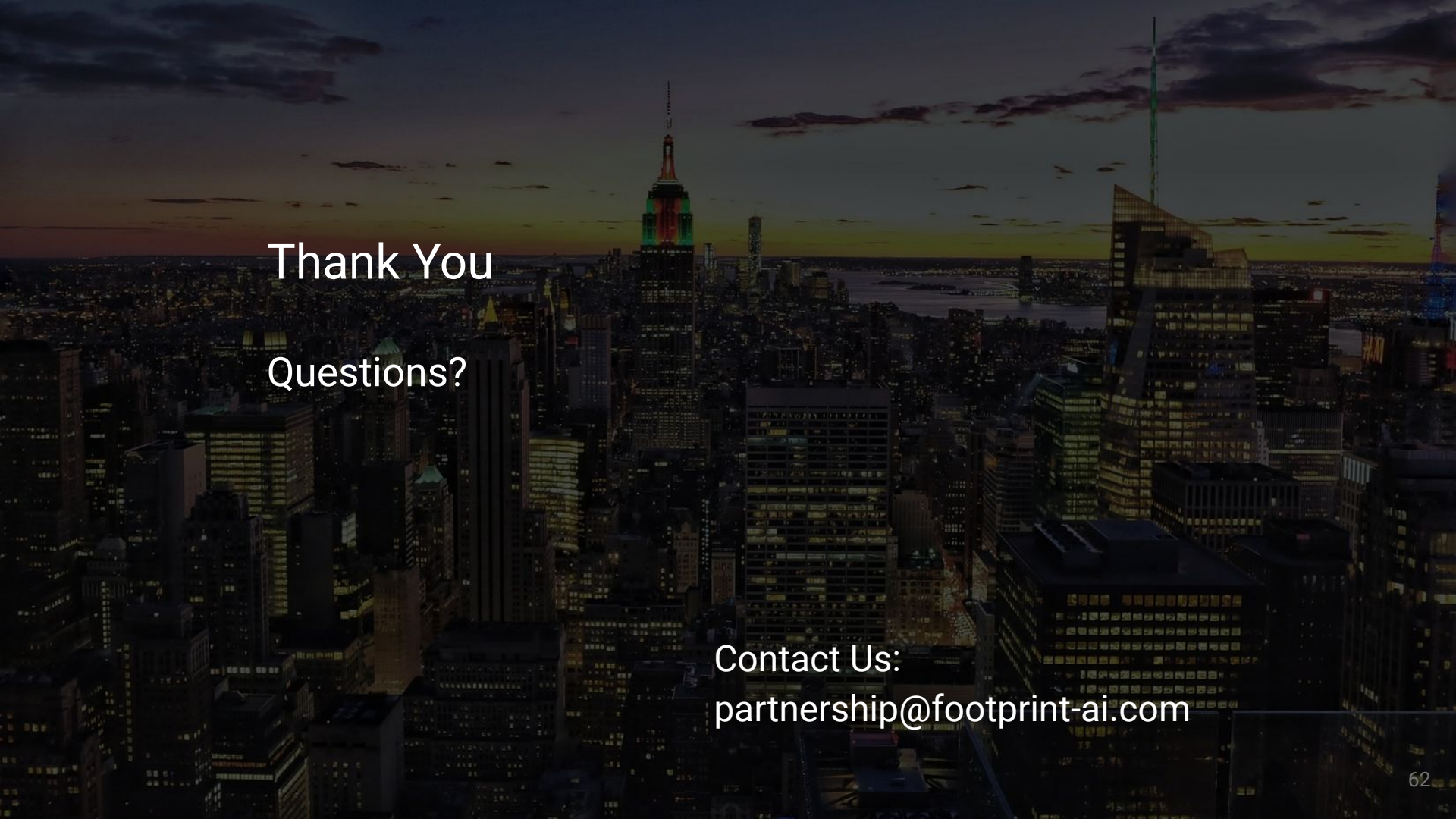
Step5: Hyperparameter tuning with katib (4/4)



Q&A

One minute takeaway

- Implemented remote development, training, and deployment.
- Pipeline-based development model
- Distributed model training and optimization
- Model deployment



Thank You

Questions?

Contact Us:
partnership@footprint-ai.com

Additional materials

- Documentations
  - https://www.kubeflow.org/
- Kubeflow materials
  - https://www.kubeflow.org/