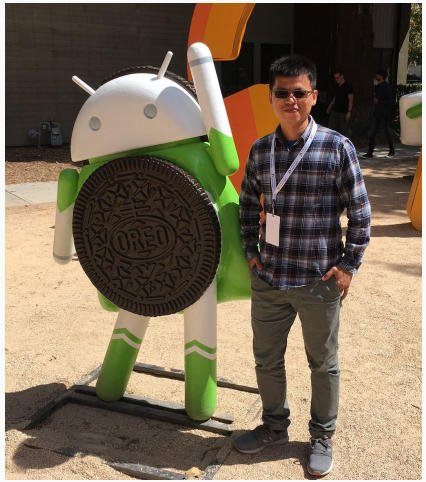
Fairness and Explainable AI

- An End-to-end example on Tensorboard with Kubeflow

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

2021/12/29

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

課程綱要

- 課前知識
- 概念簡介
- 環境介紹
- 詞彙定義
- 範例練習
- 問與答

課前知識

- 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》

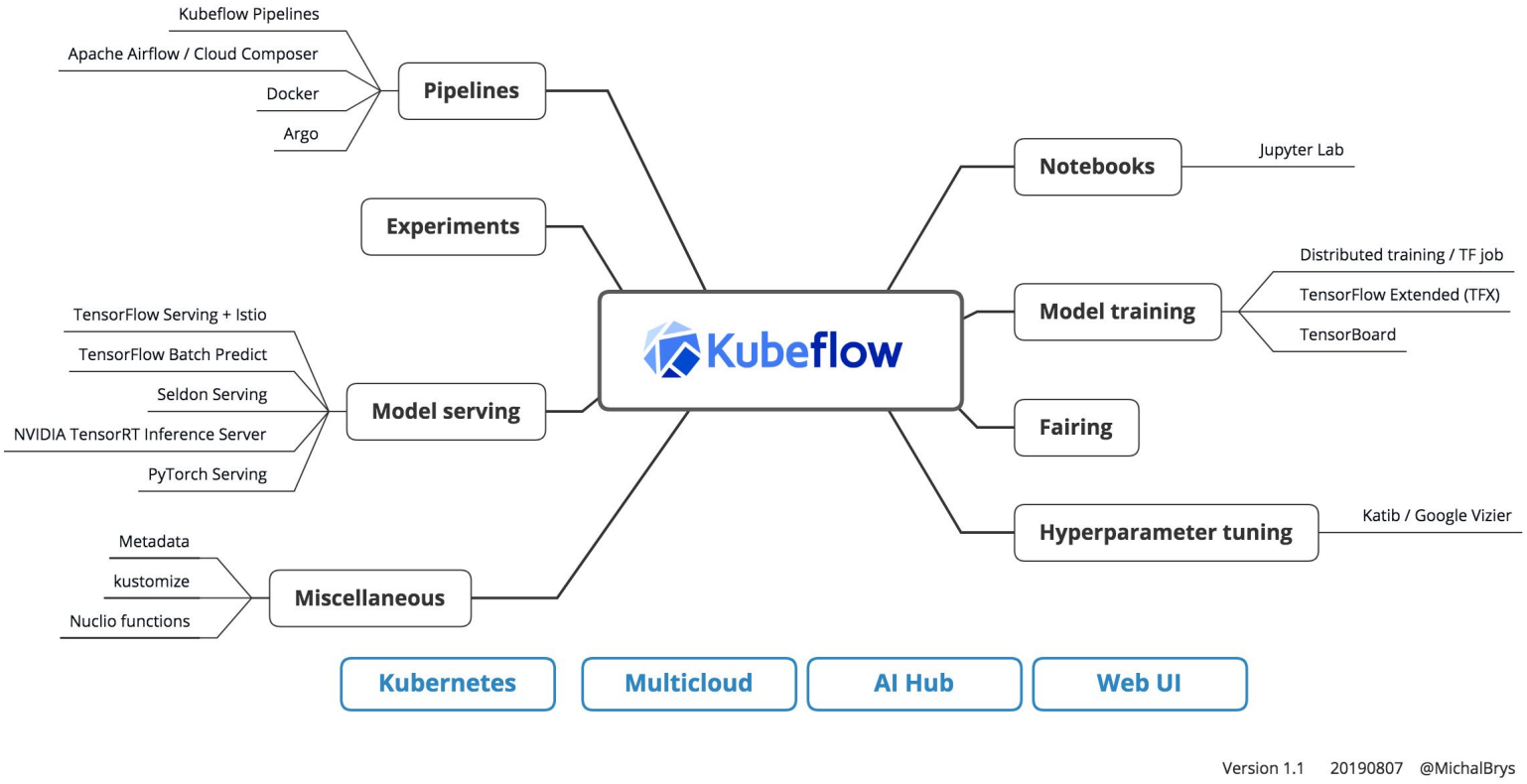
「不聞不若聞之，聞之不若見之，見之不若知之，知之不若行之；學至于行之而止矣。」

範例資源

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

Or Click Me

Kubeflow架構


Kubeflow Pipelines
Apache Airflow / Cloud Composer
Docker
Argo
Pipelines
Experiments
TensorFlow Serving + Istio
TensorFlow Batch Predict
Seldon Serving
NVIDIA TensorRT Inference Server
PyTorch Serving
Model serving
Metadata
kustomize
Nuclio functions
Miscellaneous
Kubeflow
Notebooks
Jupyter Lab
Model training
Distributed training / TF job
TensorFlow Extended (TFX)
TensorBoard
Fairing
Hyperparameter tuning
Katib / Google Vizier
Kubernetes
Multicloud
AI Hub
Web UI
Version 1.1 20190807 @MichalBrys

Fairness and Explainable AI

Why explainable AI? (1/2)

- Can we trust the model?
  - KDD 2008 Breast Cancer Identification Winner Report

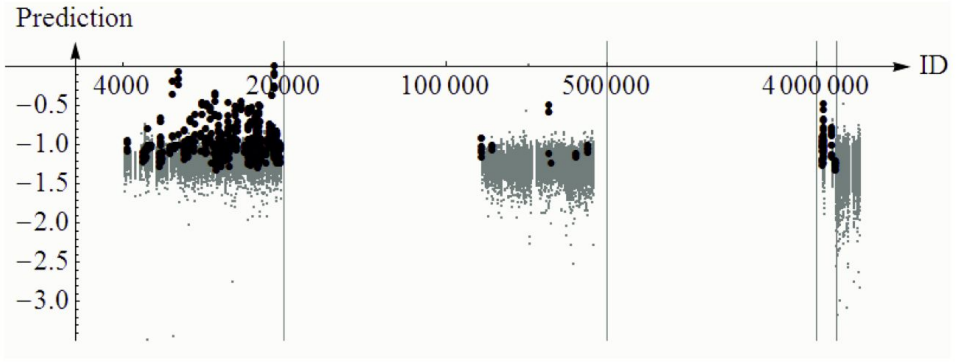


Figure 1: Distribution of malignant (black) and benign (gray) candidates depending on patient ID on the X-axis in log scale. The Y-axes is the score of a linear SVM model on the 117 features. Vertical lines show the boundaries of the identified ID bins.

Ref: http://kdd.org/exploration_files/KDDCup08-P1.pdf

Why explainable AI? (2/2)

- 是「神之一手」還是「無法接受的錯誤」？
  - “Although during the match it was unclear why the system played this move, it was the deciding move for AlphaGo to win the game.” [1]

> *“It’s not a human move. I’ve never seen a human play this move.”* (Fan Hui, 2016).

Fan Hui

From Wikipedia, the free encyclopedia

Fan Hui (Chinese: 樊麾; pinyin: *Fán Huī*; born 27 December 1981) is a Chinese-born French Go player.[2] Becoming a professional Go player in 1996, Fan moved to France in 2000 and became the coach of the French national Go team in 2005.[3] He was the champion of the European Go Championship in 2013, 2014[4] and 2015.[5] As of 2015, he is ranked as a 2 dan professional.[5] He additionally won the 2016 European Professional Go Championship.[6]

[1] https://arxiv.org/abs/1708.08296

What is explainable AI?

1. Verification of the system
    - A black box system is must not trust by default. For example, the use of models in health care should be interpreted and verified by medical experts in an absolute necessity.
2. Improvement of the system
    - If the model is explainable, it is easy to perform weakness analysis for better model improvement.
3. Learning from the system
    - AIs are trained with millions of examples while experts are trained with a limited number of examples. With explainable AI, we are able to acquire new insights from the distilled knowledge.
4. Compliance to legislation

Ref: https://arxiv.org/abs/1708.08296

From GDPR perspective

Art. 22 GDPR

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

   (c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

對個人有法律或重大影響的決定，不得基於（Article 9 規定的）個人種族、政治立場、宗教與哲學信仰、商業關係這些個人資料[2]

[1] https://gdpr-info.eu/art-22-gdpr/
[2] https://medium.com/trustableai/歐盟-gdpr-生效對機器學習應用的影響d153de909e4f

Explainable AI Examples (1/3)

- Highlight the symptoms that led to the prediction.

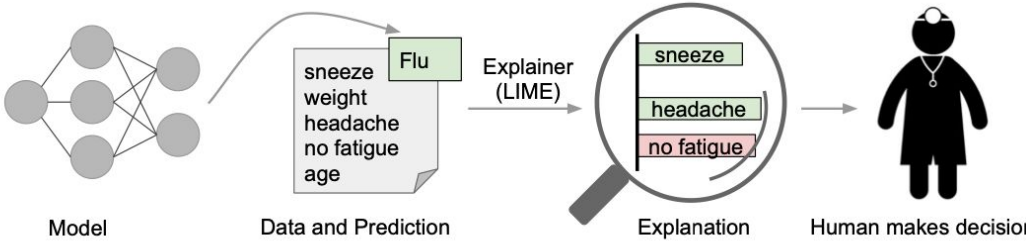
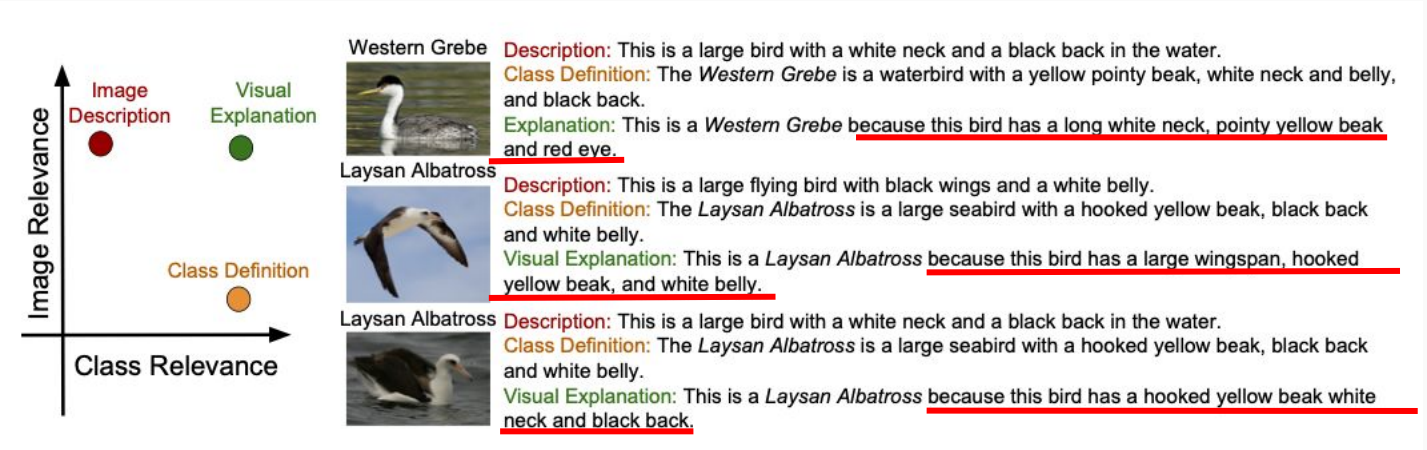


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

Ref: https://arxiv.org/pdf/1602.04938.pdf

Explainable AI Examples (2/3)

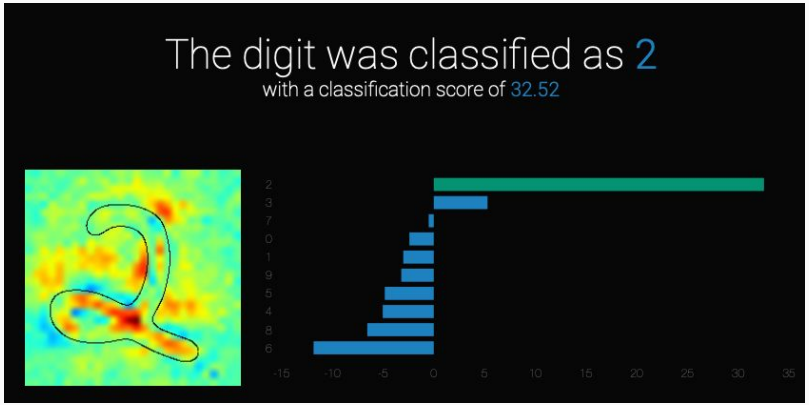
- Generating Visual Explanations

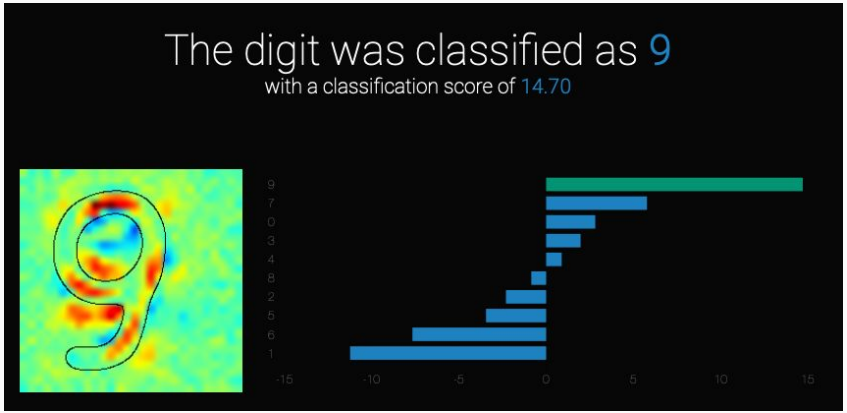


Ref: https://arxiv.org/pdf/1603.08507.pdf

Explainable AI Examples (3/3)

- Heatmap to support the decision in digital recognition.





Ref: http://heatmapping.org

What is Fairness?

Fairness has different definitions across disciplines too. This is captured in a paper by Deirdre Mulligan, Joshua Kroll, Nitin Kohli and Richmond Wong.[2] For example:

Law: fairness includes protecting individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories.

Social science: "often considers fairness in light of social relationships, power dynamics, institutions and markets."[3] Members of certain groups (or identities) that tend to experience advantages.

Quantitative fields (i.e. math, computer science, statistics, economics): questions of fairness are seen as mathematical problems. Fairness tends to match to some sort of criteria, such as equal or equitable allocation, representation, or error rates, for a particular task or problem.

Philosophy: ideas of fairness "rest on a sense that what is fair is also what is morally right."[4] Political philosophy connects fairness to notions of justice and equity.

Ref: https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf

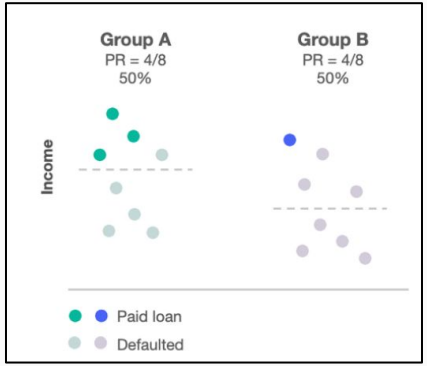
What is fairness AI?

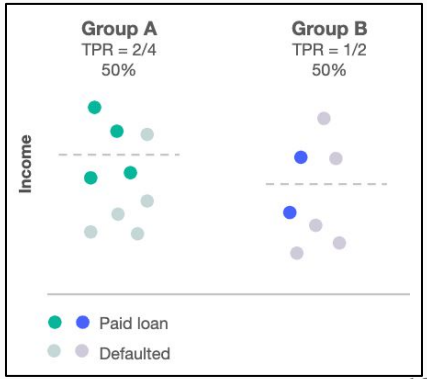
1. Fairness of a system
    - A particular problem is fit to be approximated or decided upon by an ML system
2. Fairness of a result
    - do the outputs of the algorithm actually correspond to what would constitute a fair response by a human.
3. Fairness of an approach
    - White box models
        i. Approaching fairness by developing training methods that aim to produce interpretable ML models.
        ii. Limitations:
            1. Not feasible for complex models (e.g. deep learning models)
            2. Fairness is bound by the ability for a subject to meaningfully understand.
    - Mathematical Fairness

Ref: https://medium.com/trustableai/機器學習公不公平51cc1bfab066
https://arxiv.org/pdf/1706.09976.pdf

What is fairness AI?

- Mathematical Fairness
  - Demographic parity (top)
    - The proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates.
    - Best for university admission.
  - Individual fairness / Group fairness
    - Similar individual (or groups) are treated similarly
  - Equal opportunity (bottom)
    - Each group should get the positive outcome at equal rates.
    - Best for fraud detection





- [Interactive website]

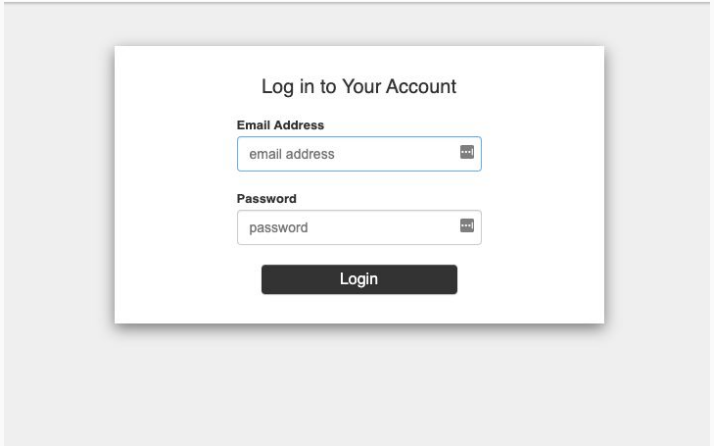
Ref: https://medium.com/trustableai/機器學習公不公平51cc1bfab066
https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2

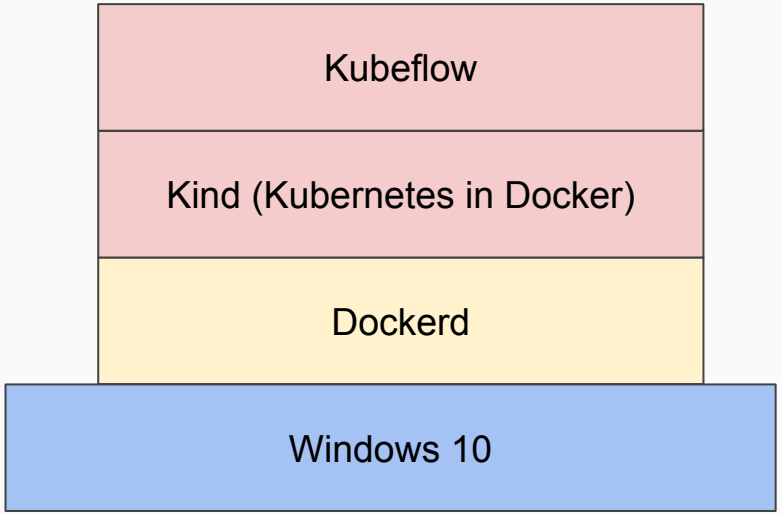
環境介紹

虛擬機環境

1.kubectl port-forward svc/istio-ingressgateway -n istio-system 8080:80 --address 0.0.0.0

2.Open http://localhost:8080





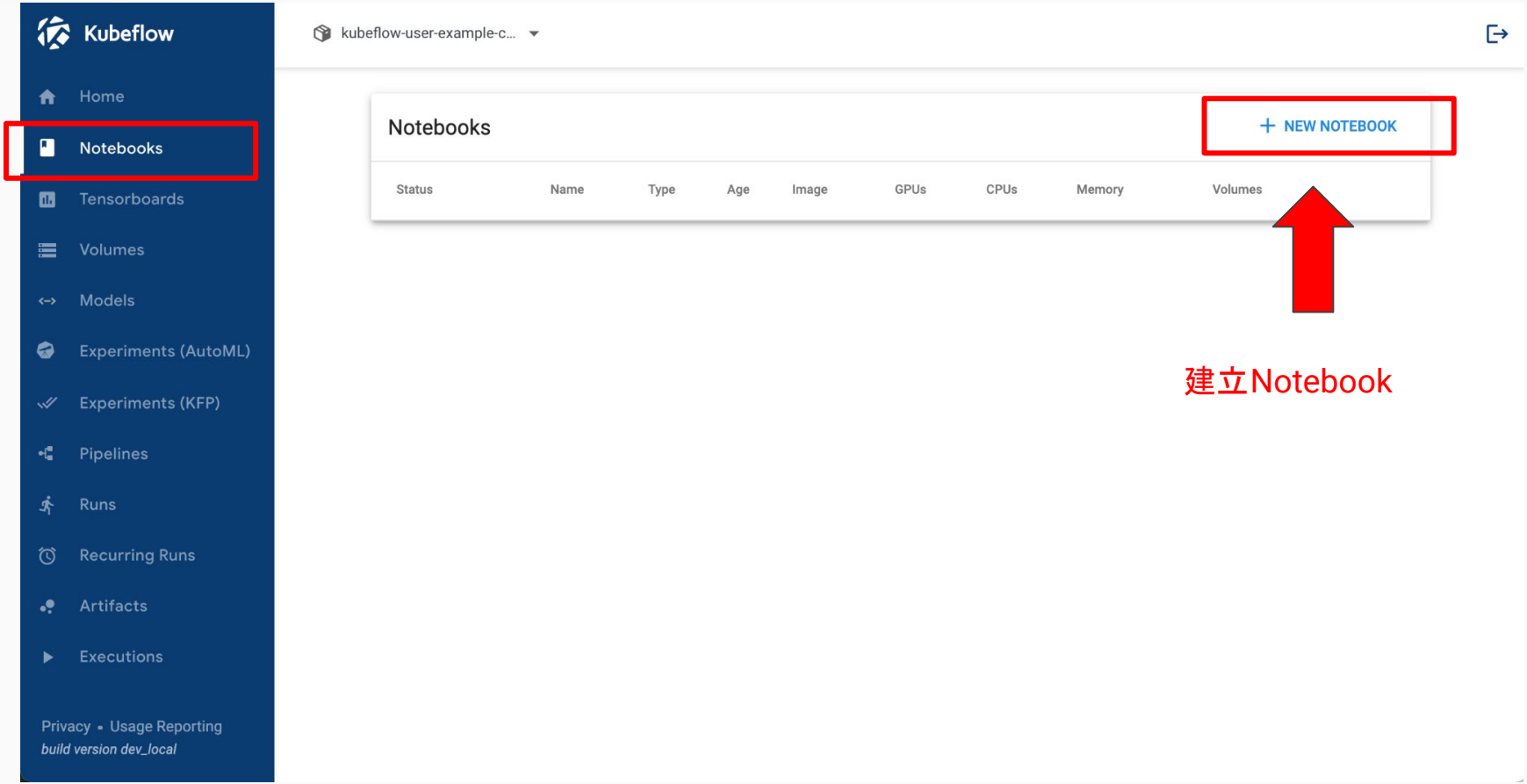
Wait! 所以我說那個帳號密碼呢?

Account: user@example.com
Password: 12341234

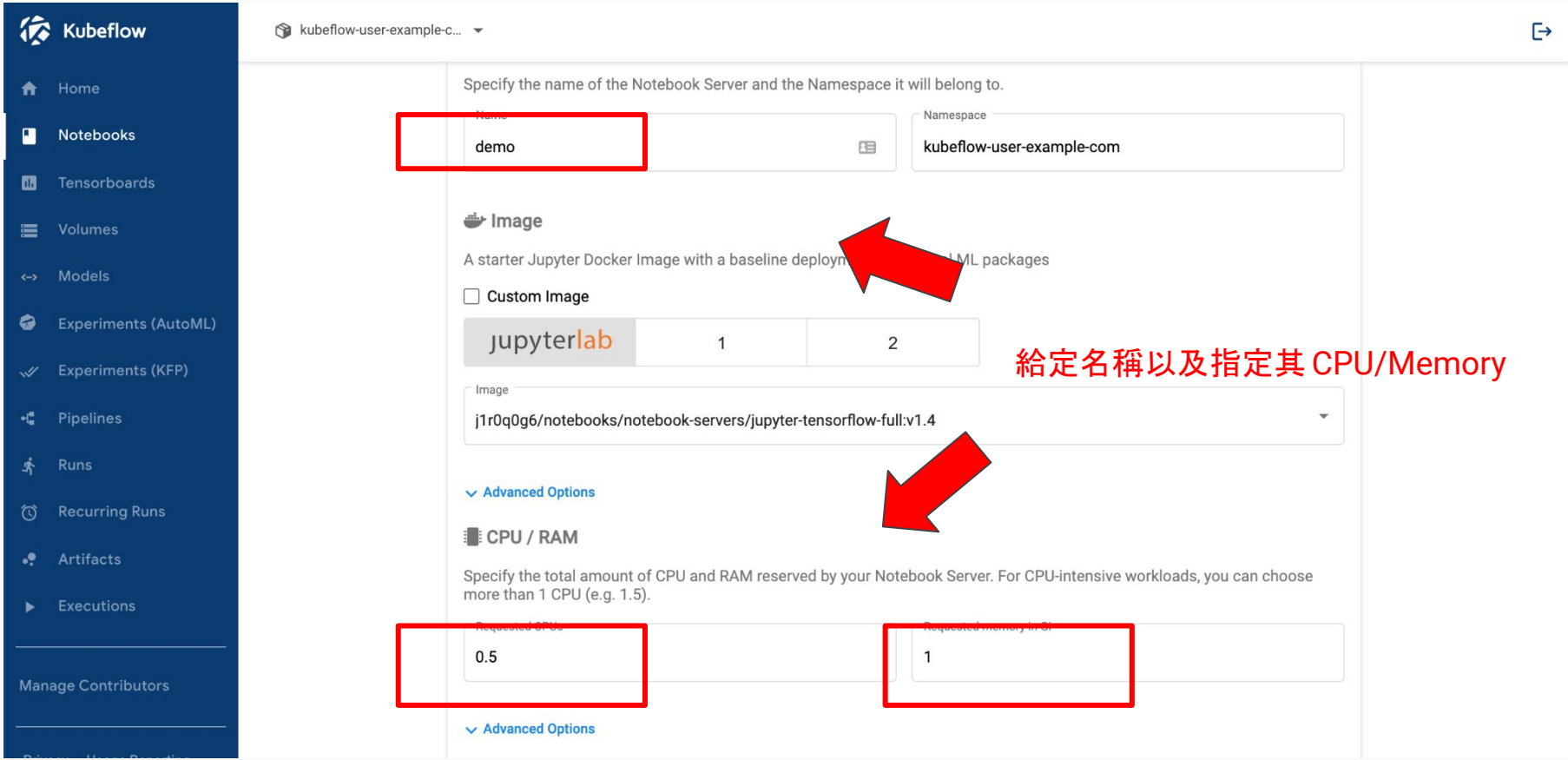
Kubectl Cli

```
// 查看所有namespace
Kubectl get namespaces


// 查看kubeflow中所有運行的Pod
kubectl get pods -n kubeflow


// 查看目前使用者運行的Pod
kubectl get pods -n kubeflow-user-example-com
```

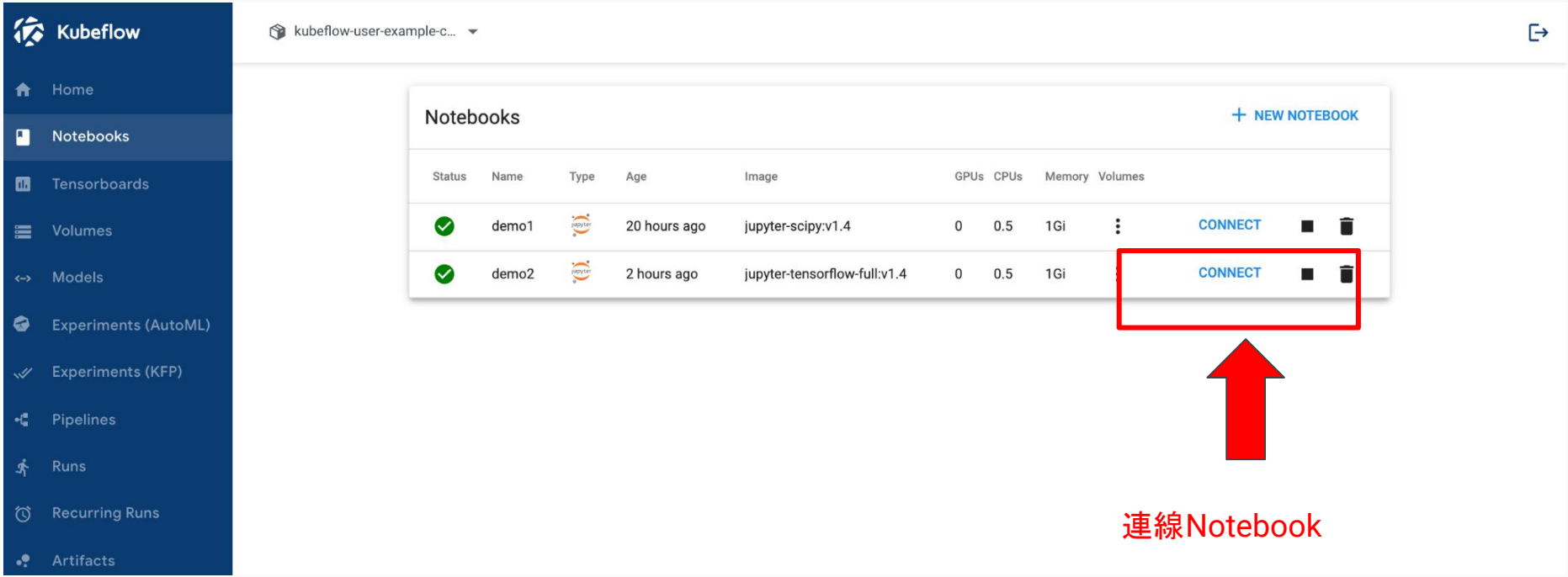
Step1: 開啟Notebook作為開發環境 (1/3)



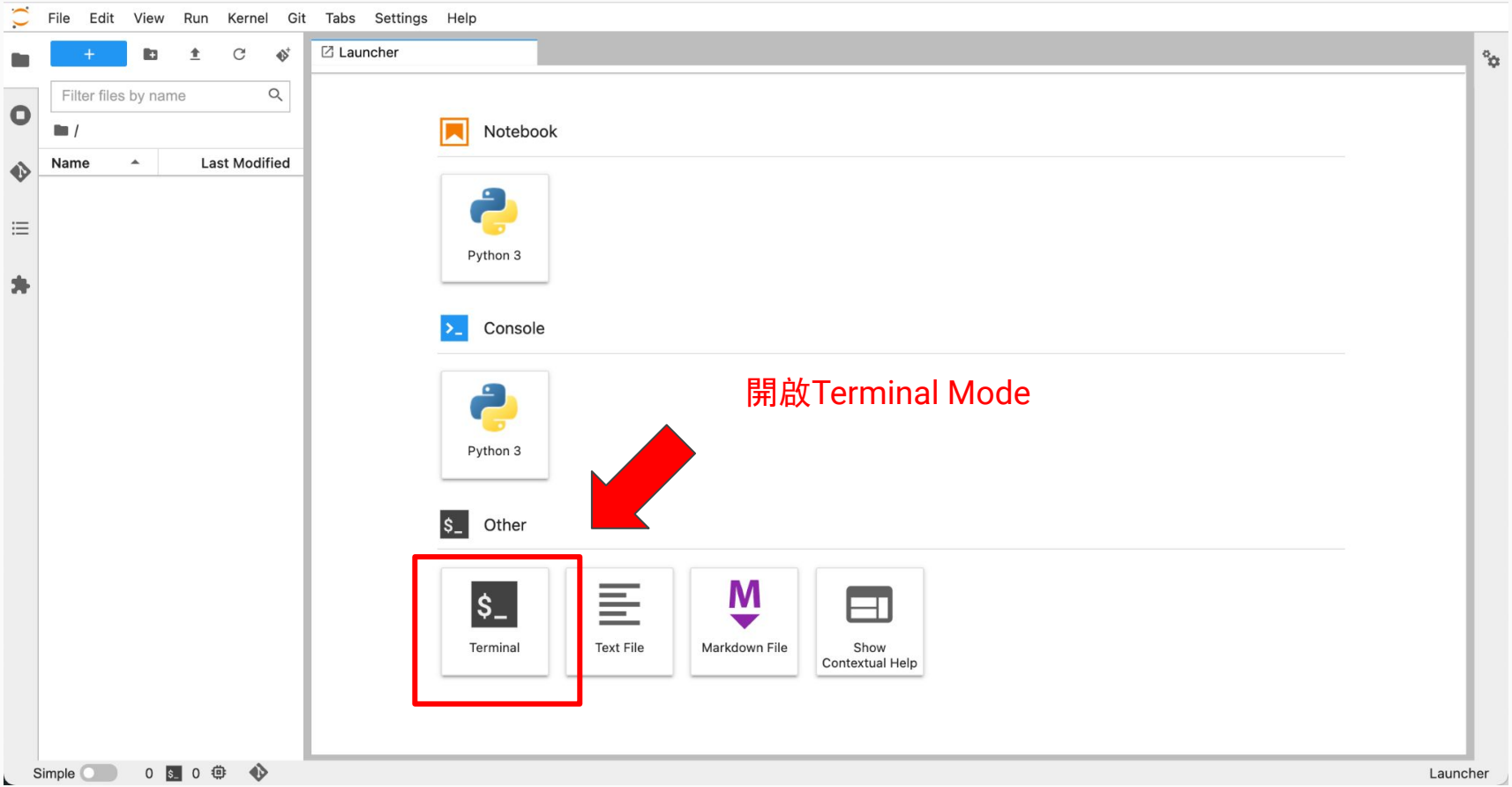
Step1: 開啟Notebook作為開發環境 (2/3)



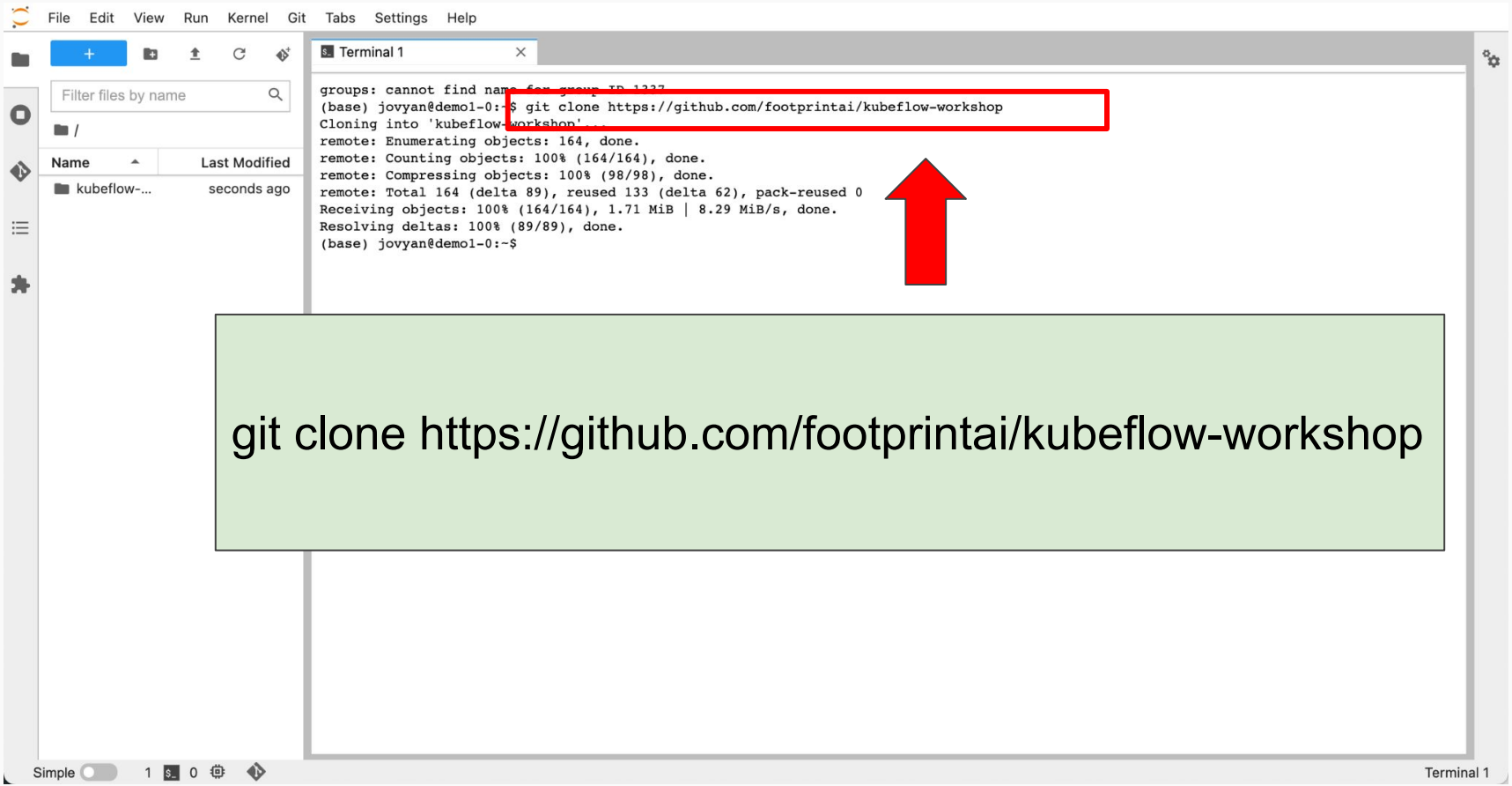
Step1: 開啟Notebook作為開發環境 (3/3)



Step2: 開啟Terminal下載範例程式(1/3)

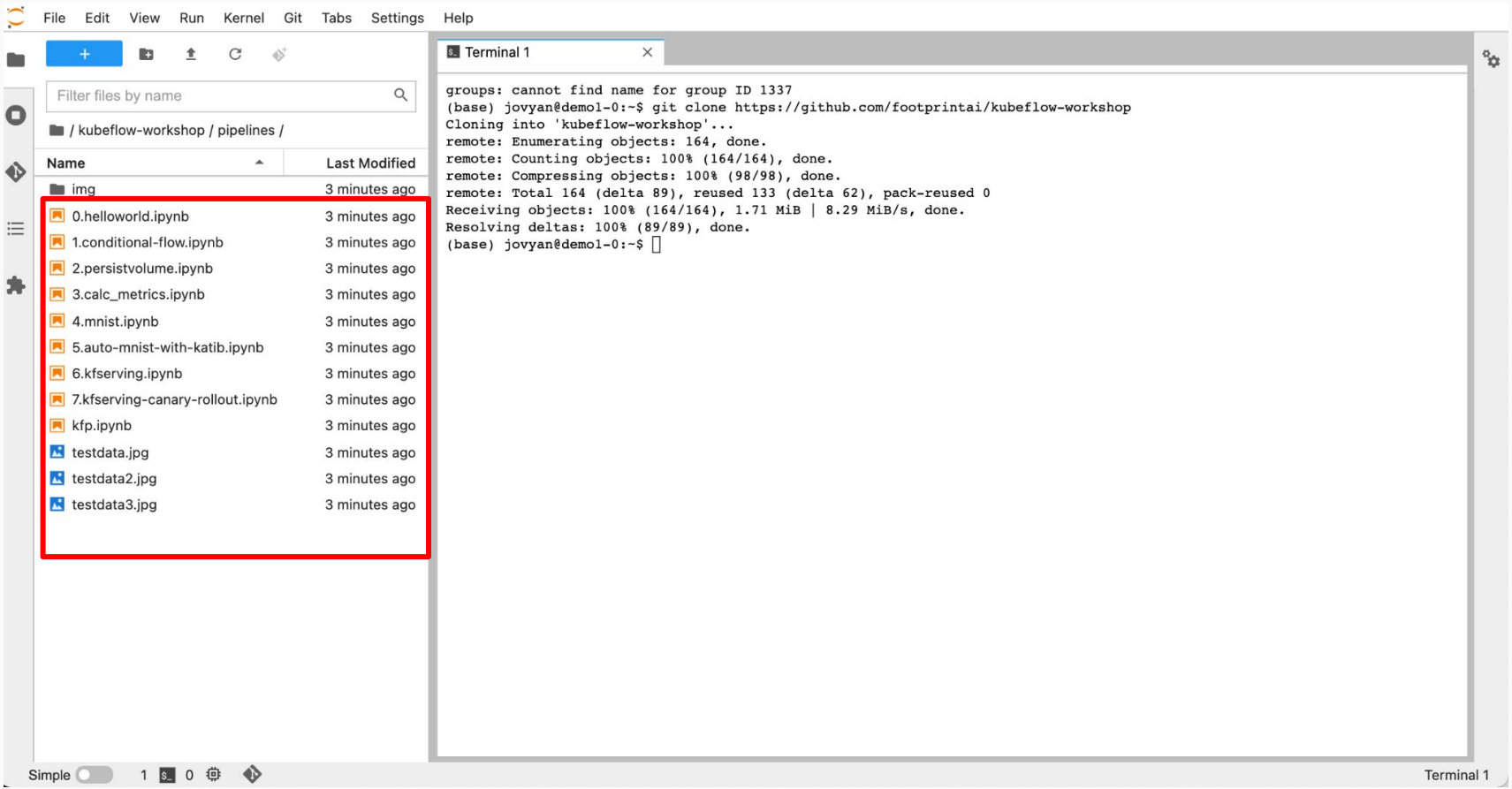


Step2: 開啟Terminal下載範例程式(2/3)



git clone https://github.com/footprintai/kubeflow-workshop

Step2: 開啟Terminal下載範例程式(3/3)



詞彙說明

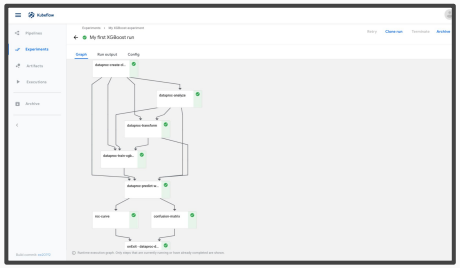
詞彙說明

Pipeline Code

編譯後

Workflow
Resource

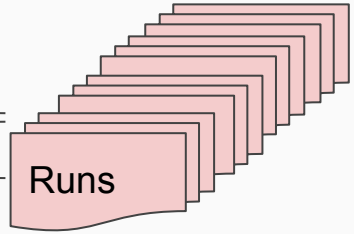
建立Pipeline



Create Run

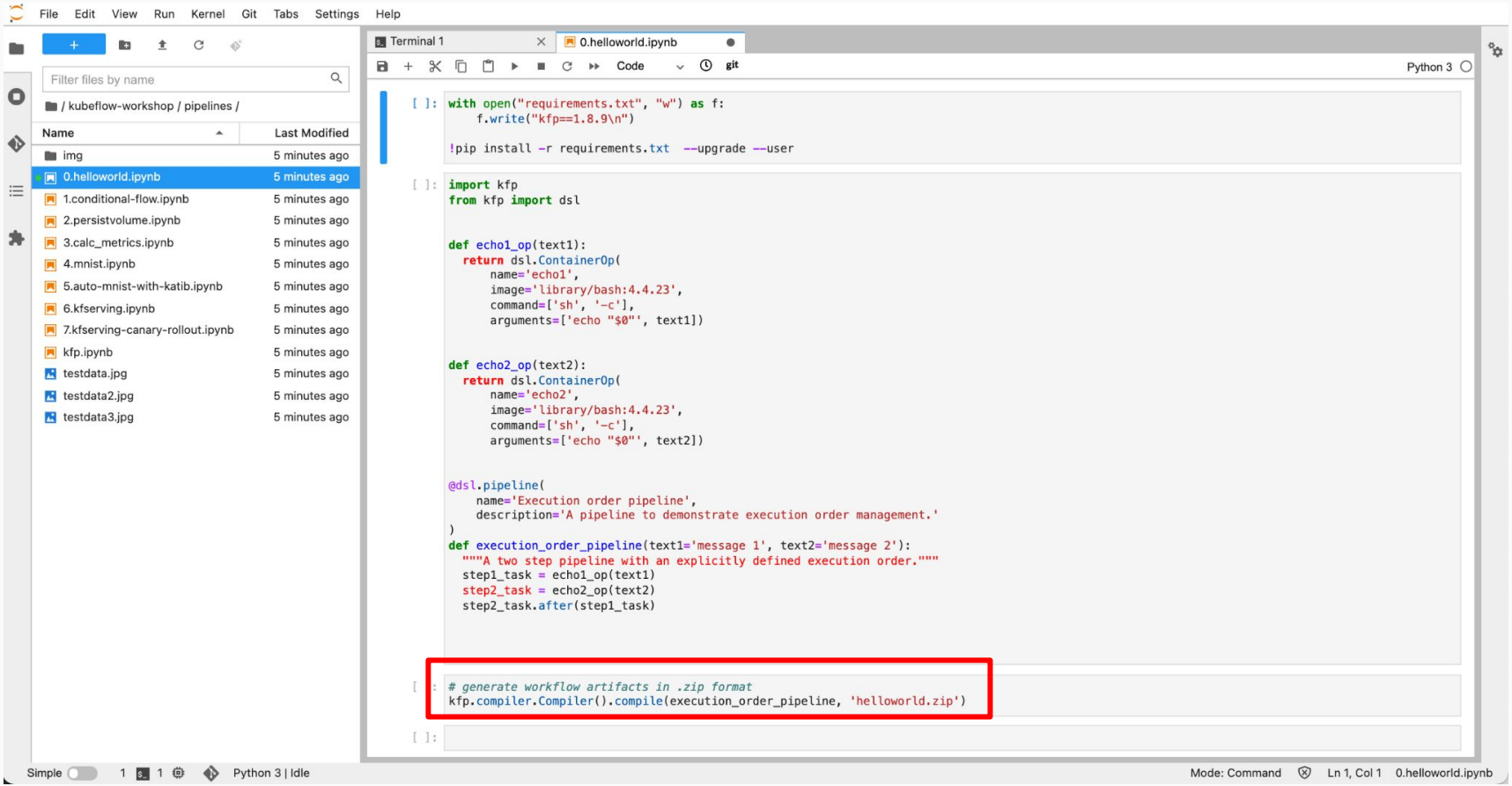
實驗1 (Experiment1)

實驗2 (Experiment2)

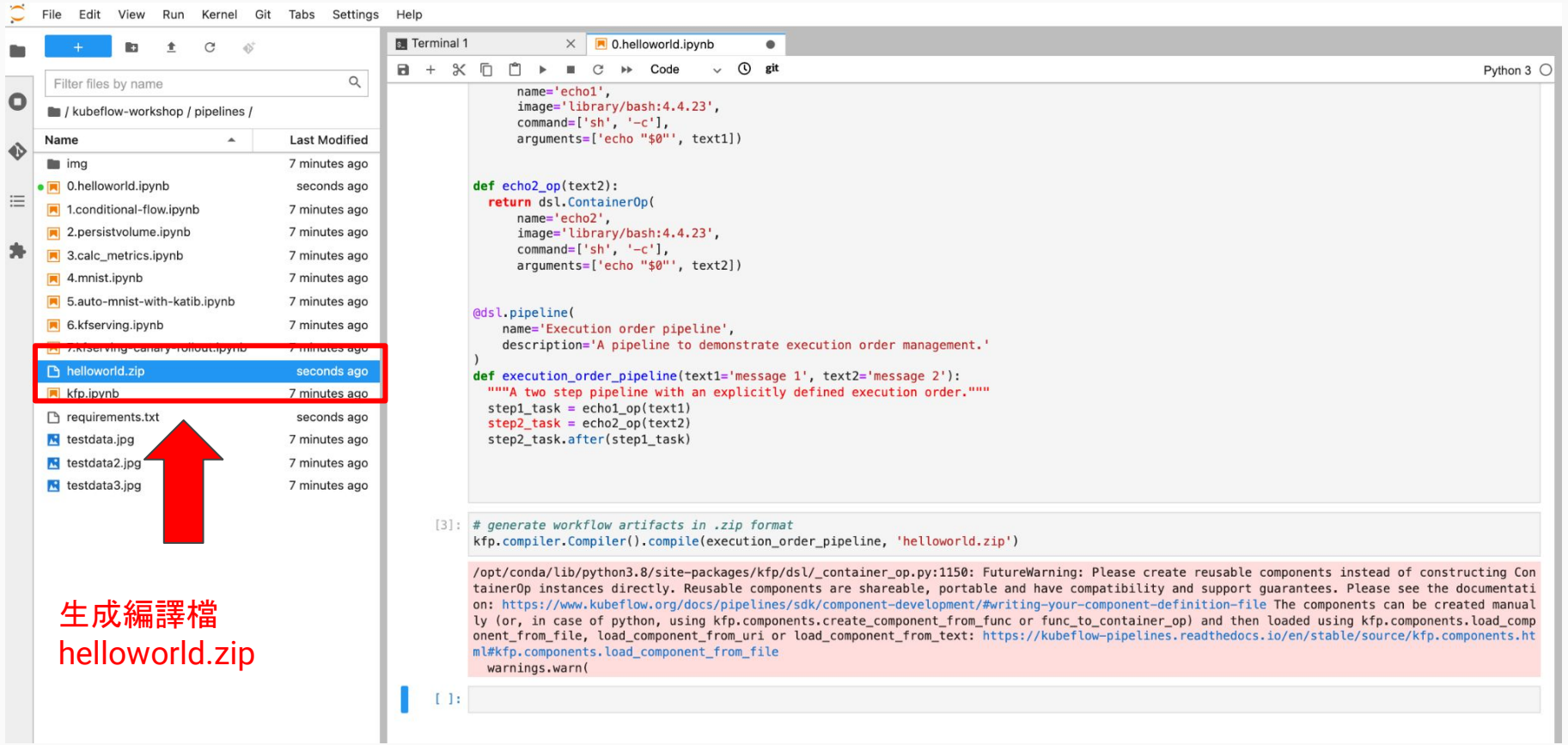


範例1 Hello World!

Step3: 編譯helloworld.ipynb (1/2)

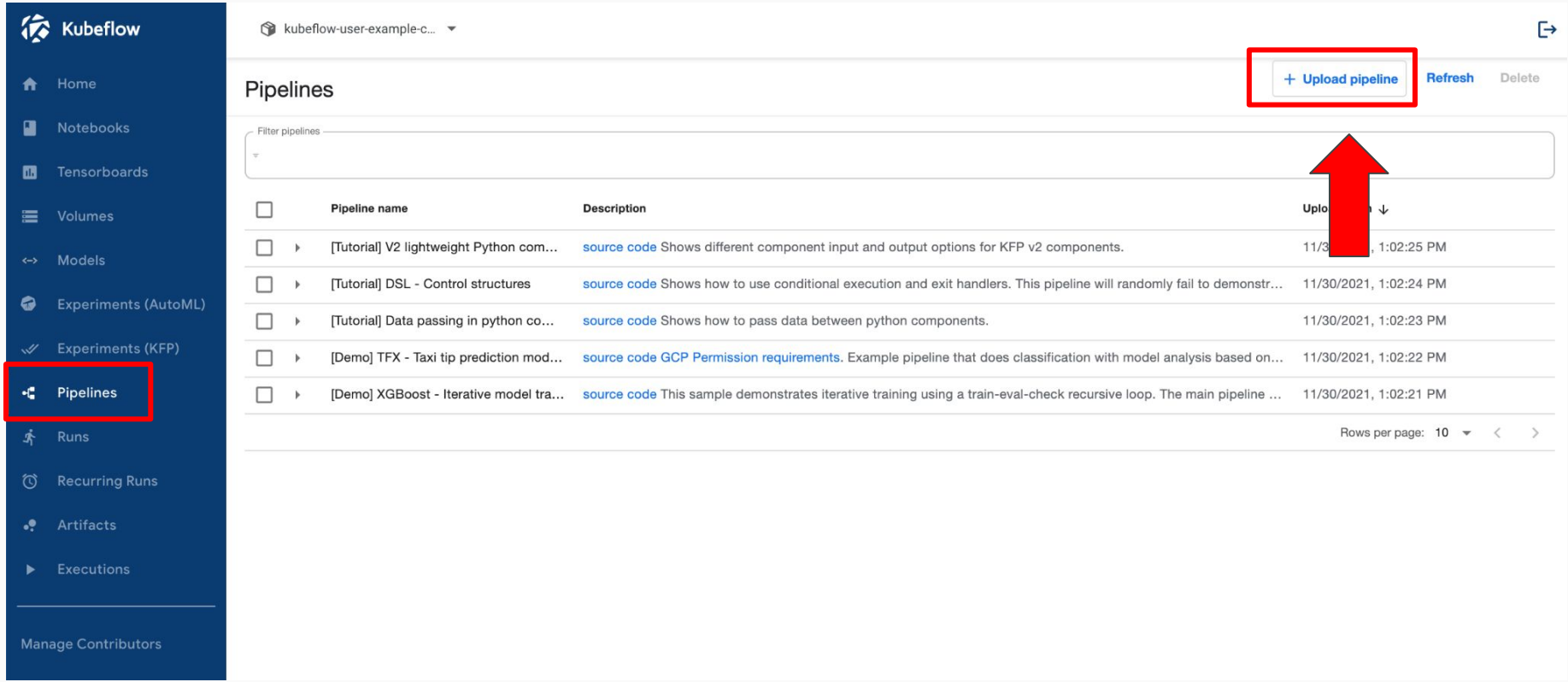


Step3: 編譯helloworld.ipynb (2/2)

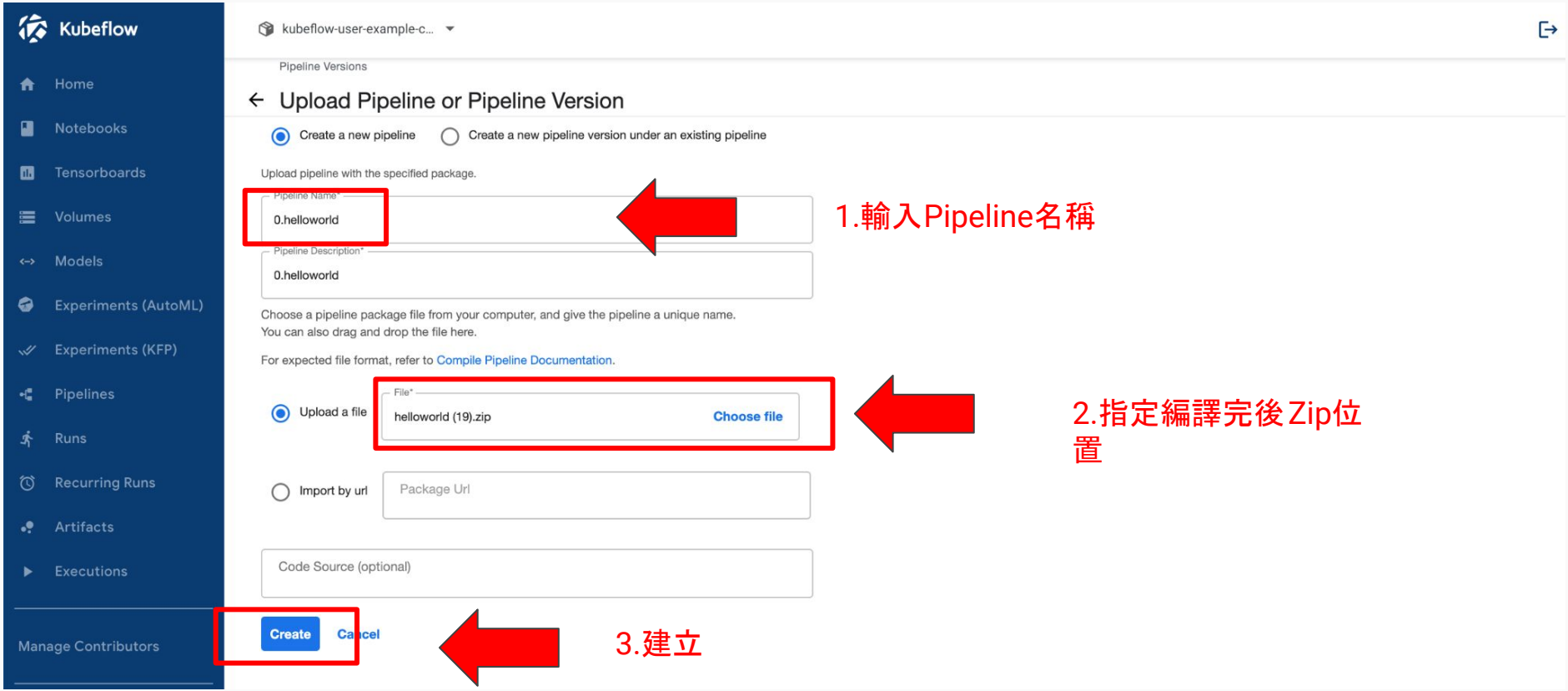


生成編譯檔
helloworld.zip

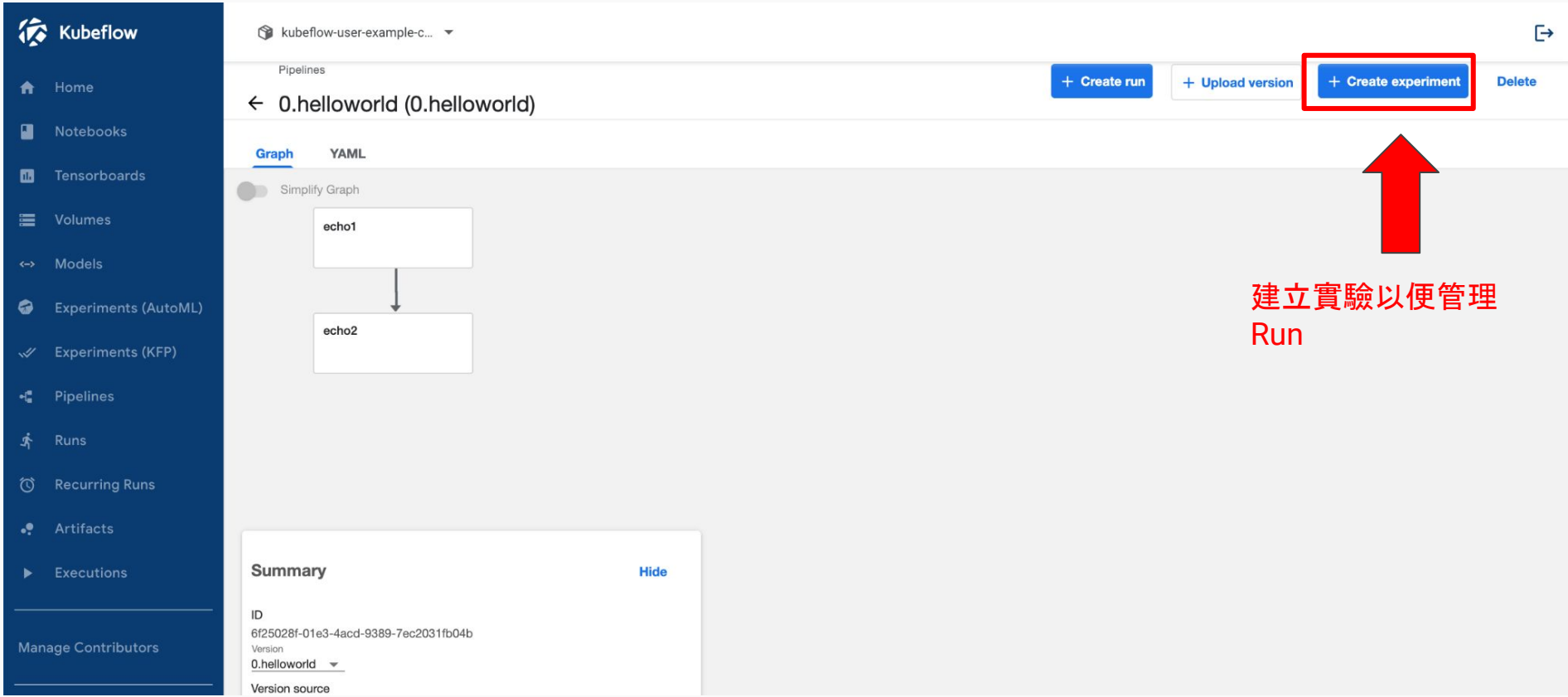
Step4: 建立Pipeline (1/7)



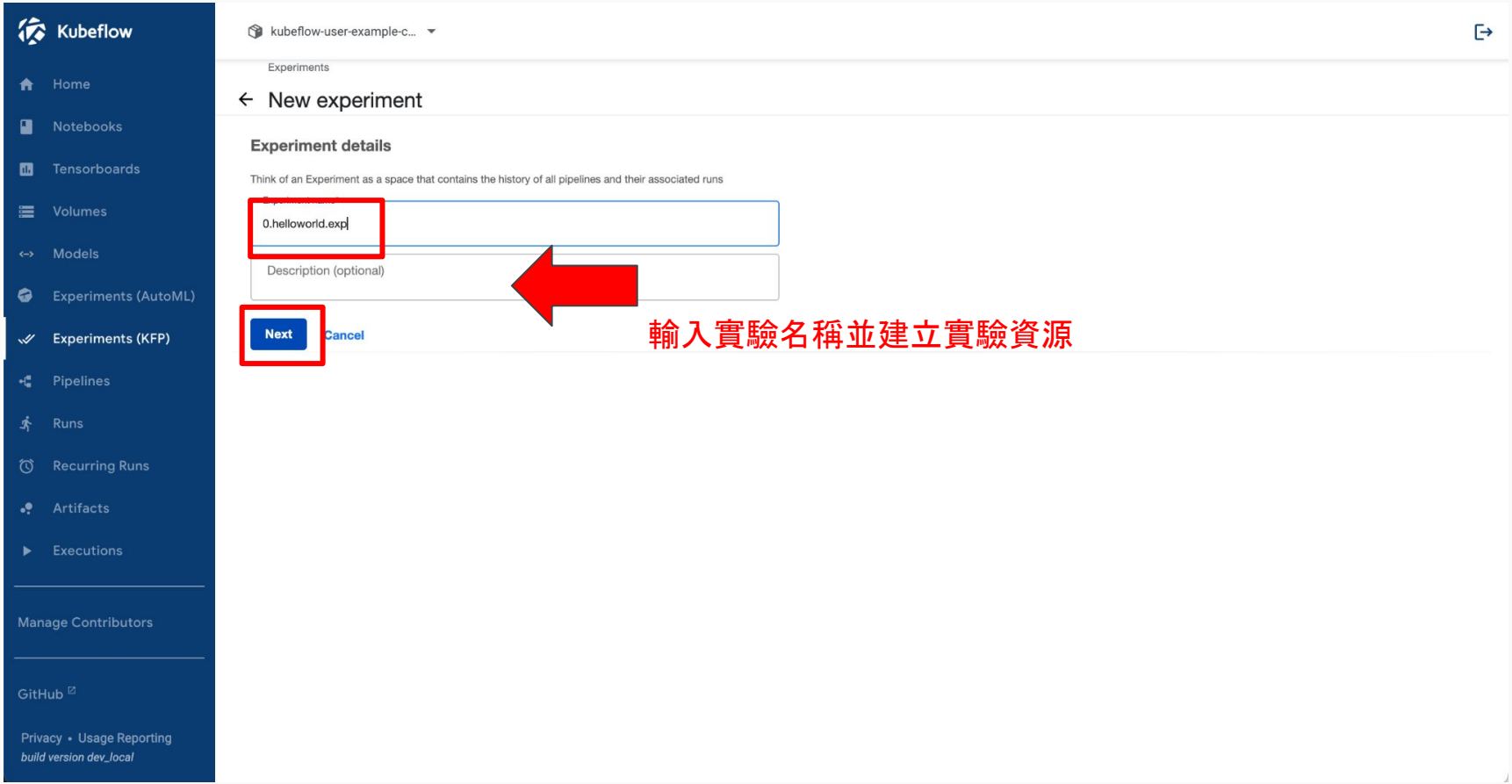
Step4: 建立Pipeline (2/7)



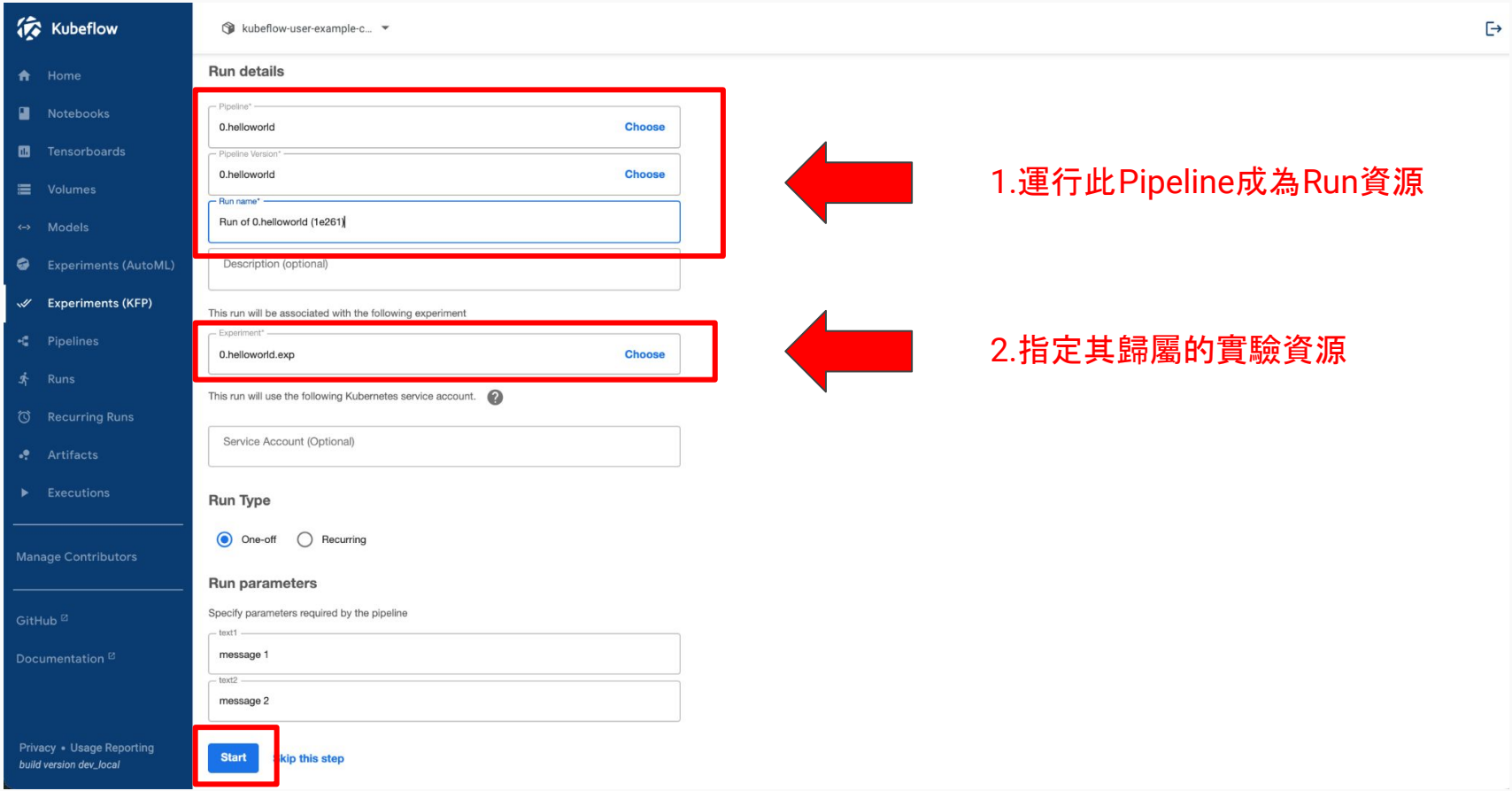
Step4: 建立Pipeline (3/7)



Step4: 建立Pipeline (4/7)



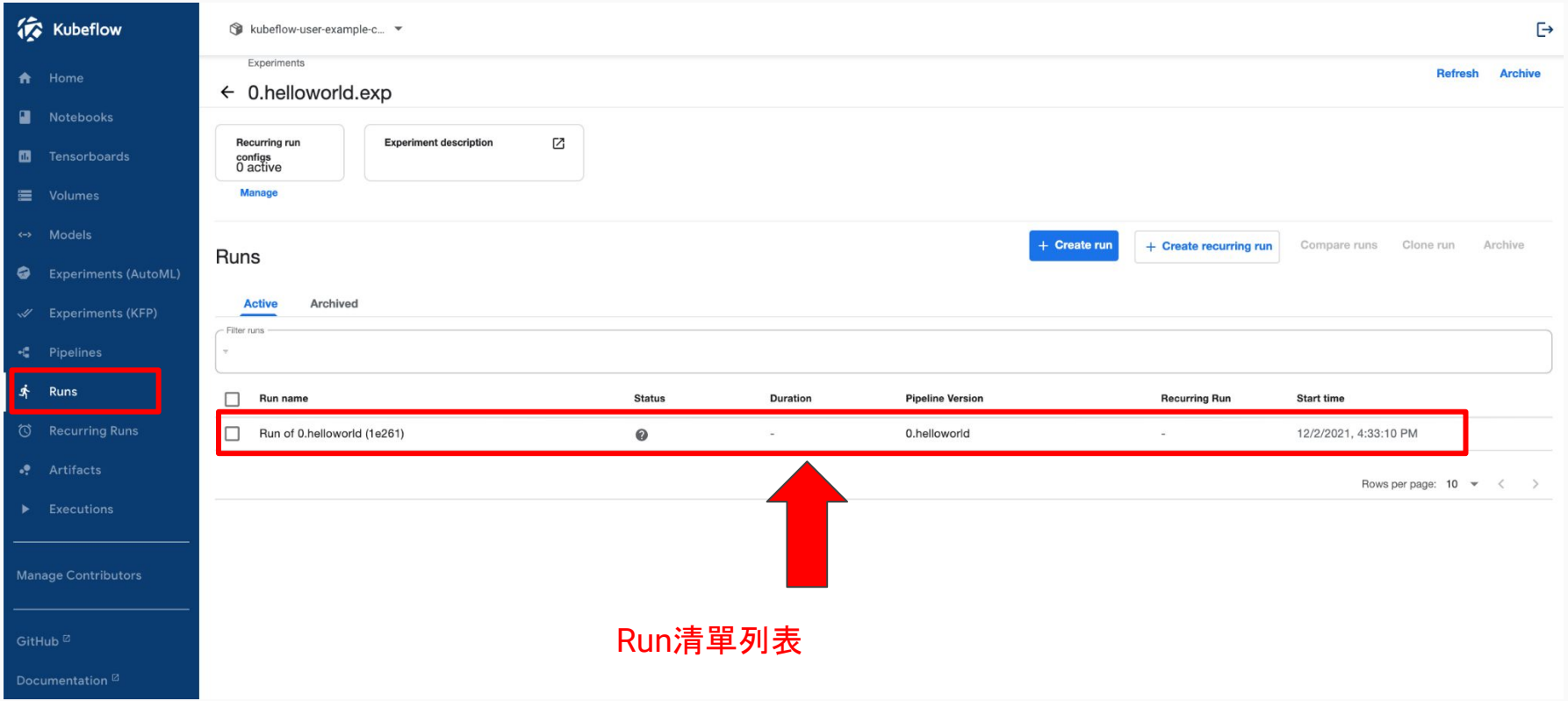
Step4: 建立Pipeline (5/7)



1.運行此Pipeline成為Run資源

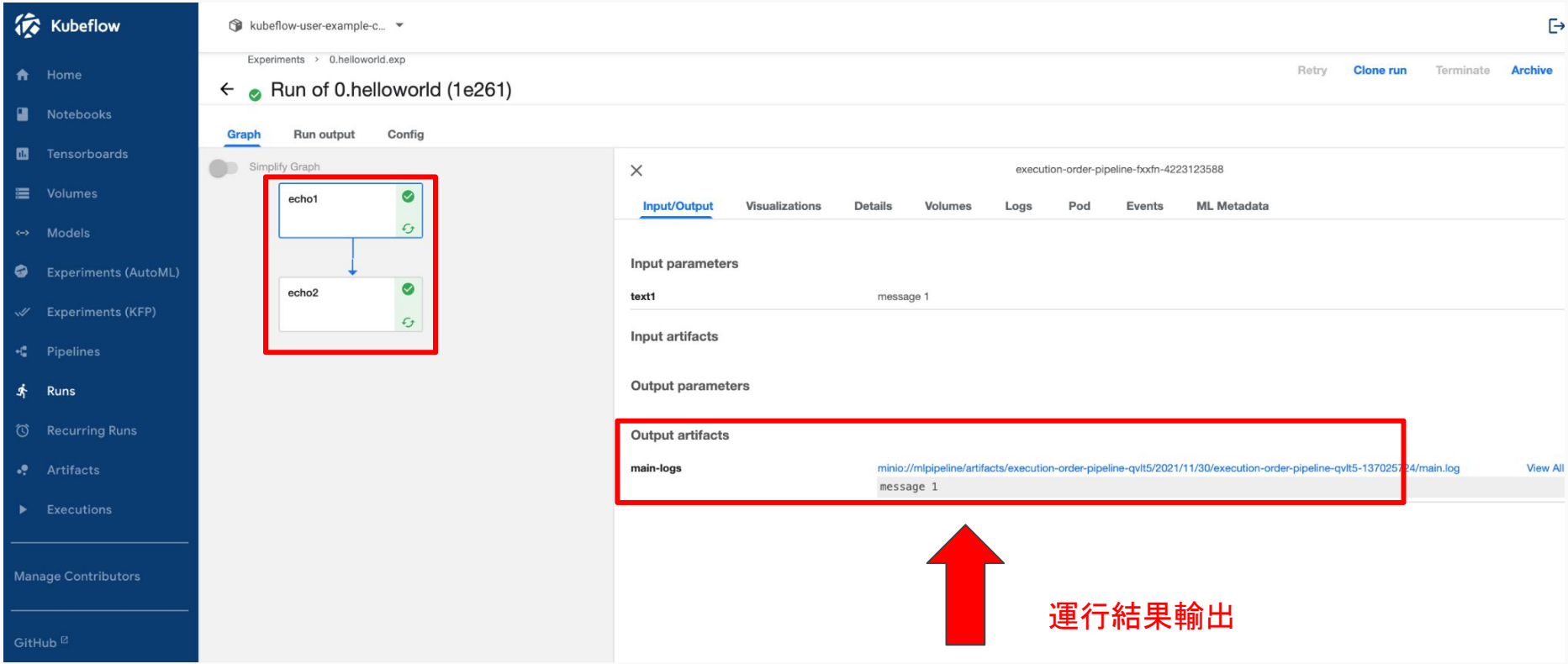
2.指定其歸屬的實驗資源

Step4: 建立Pipeline (6/7)



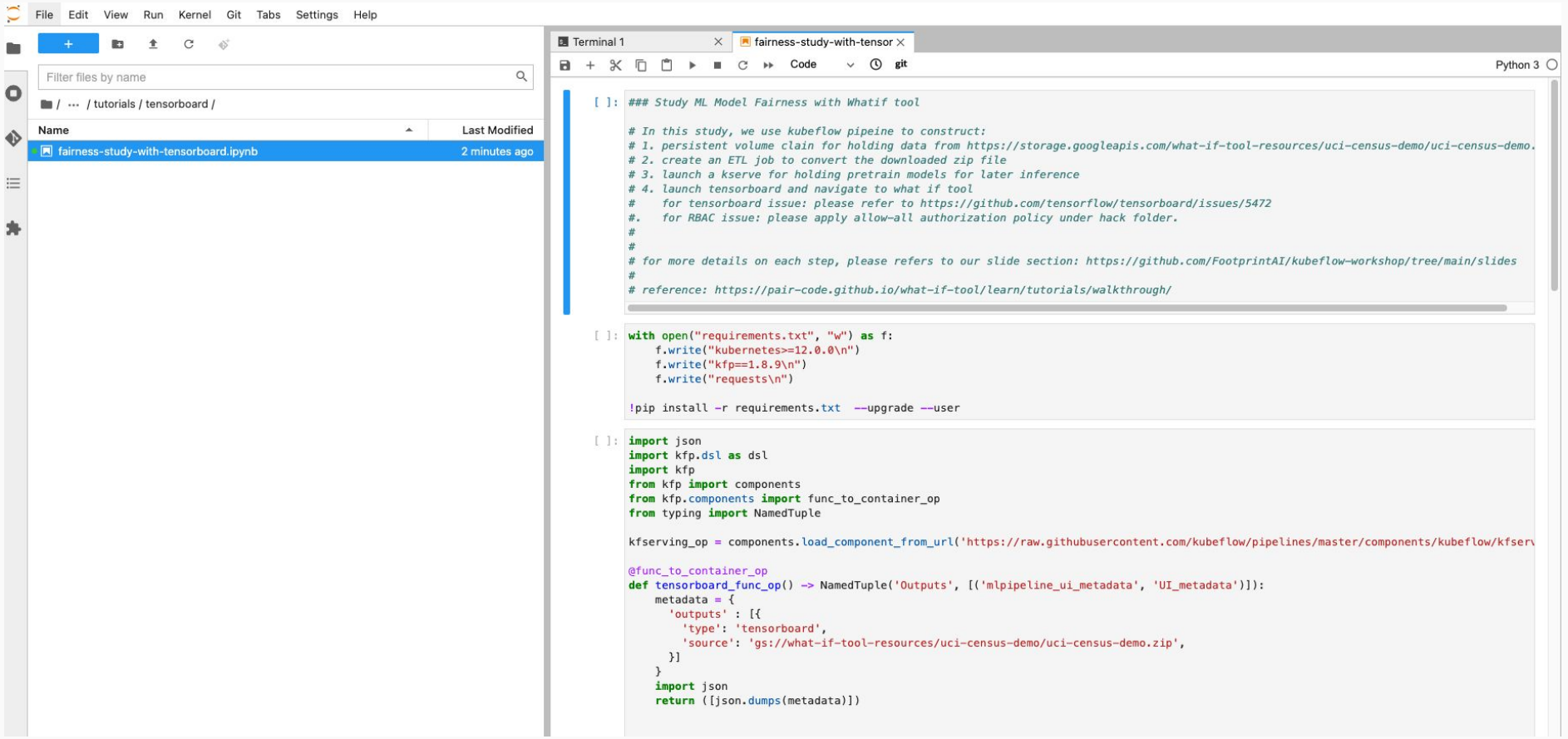
Run清單列表

Step4: 建立Pipeline (7/7)

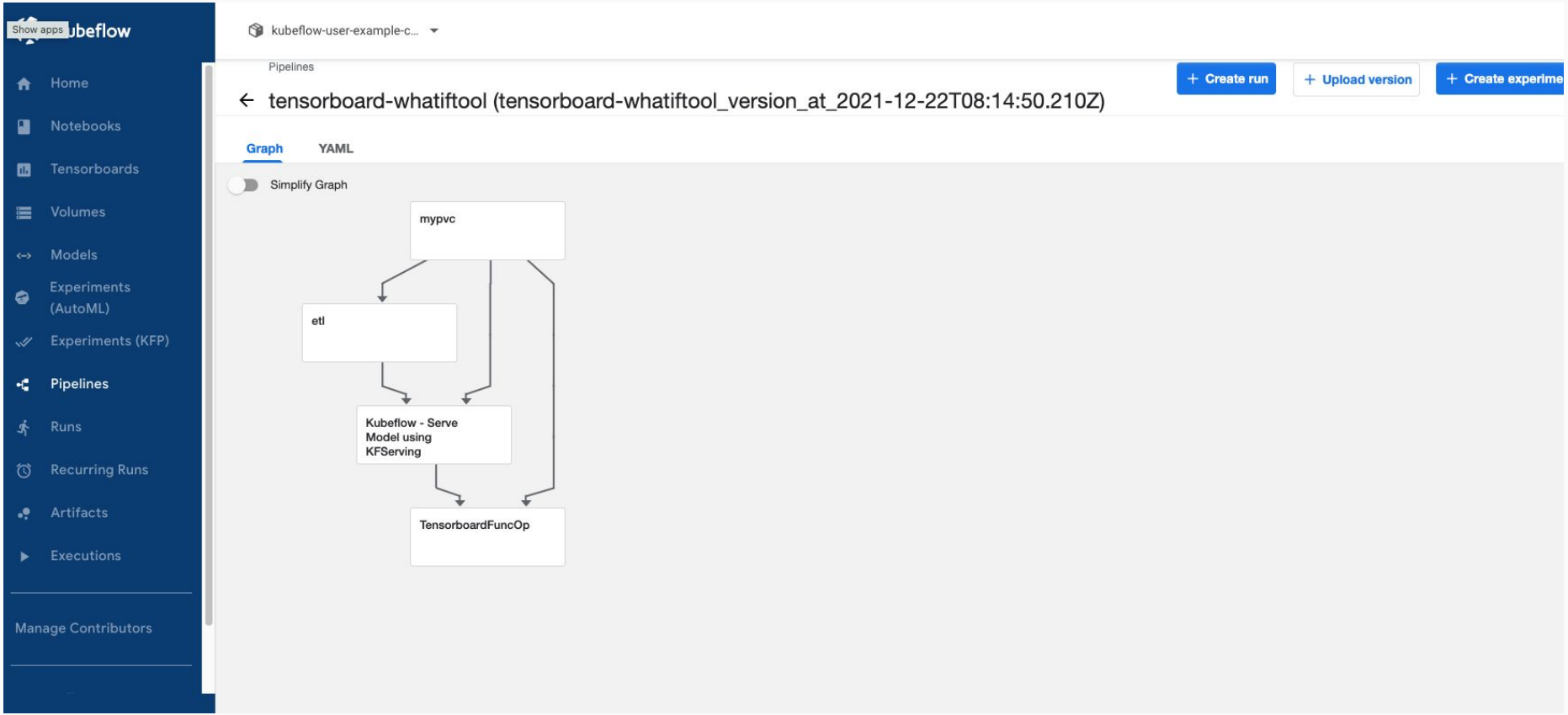


範例2 Fairness Study with What-If Tool and Tensorboard

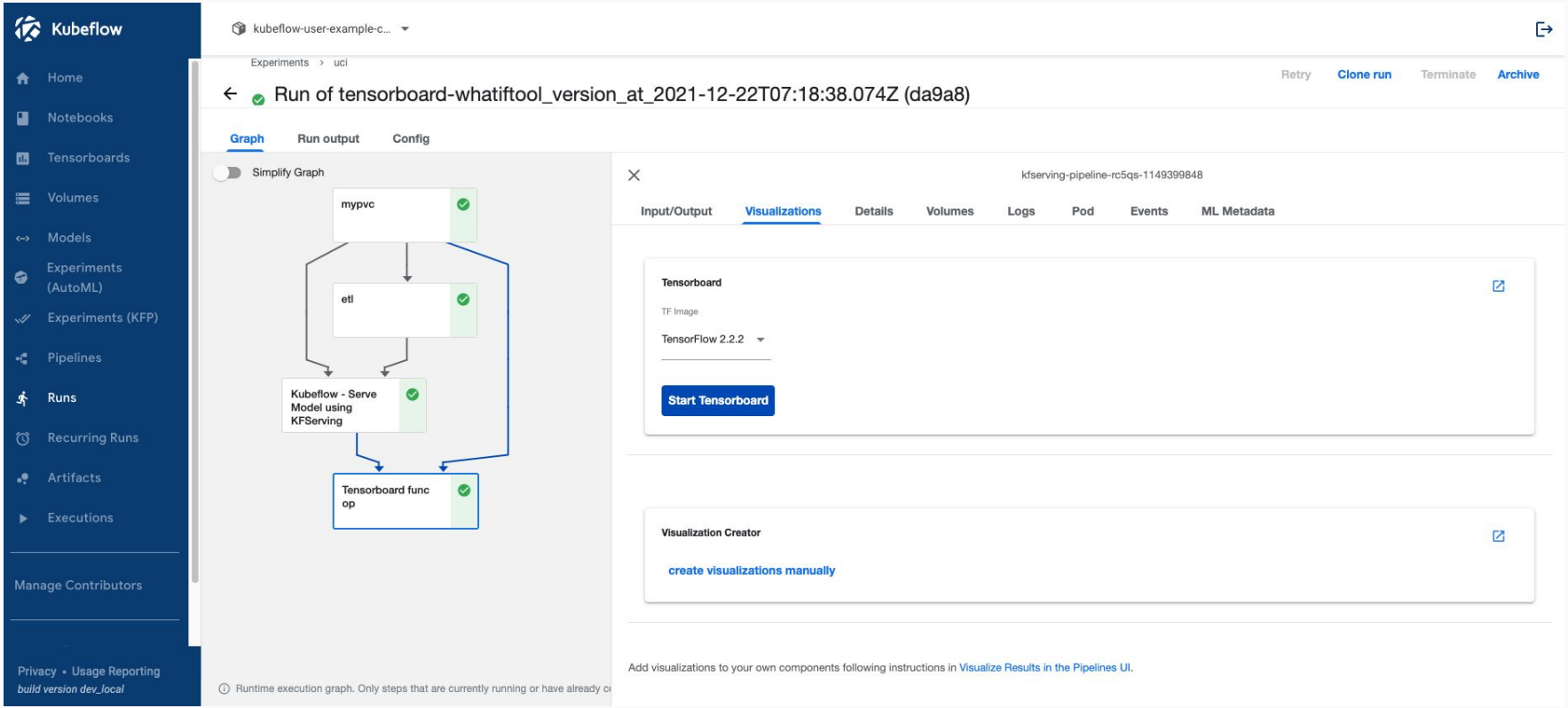
Step5: 建立fairness pipeline (1/3)



Step5: 建立fairness pipeline (2/3)



Step5: 建立fairness pipeline (3/3)



Step6: hack tensorboard (1/3)

(此步為進階功能，若無法實作也無仿)

為了讓tensorboard可以查看本地端的資料，我們得將其對應資源建立起：

1. 先尋找pvc的位置以及viewer id
2. 修改至hack/tensorboard-use-local-volume.yaml底下
3. 刪掉舊有的viewer resource (hack/delete-all-viewer.sh)

```
root@instance-x:/home/hsinhoyeh# kubectl delete viewers -n kubeflow-user-example-com --all
viewer.kubeflow.org "viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a" deleted
```

4. 然後kubectl apply -f hack/tensorboard-use-local-volume.yaml

```
root@instance-x:/home/hsinhoyeh# kubectl get viewer -n kubeflow-user-example-com
NAME                                                AGE
viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a     13s
```

```
root@instance-1:~# kubectl get pvc -n kubeflow-user-example-com
NAME                          STATUS   VOLUME                                     CAPACITY   ACCESS MODES   STORAGECLASS   AGE
volumeop-sequential-mv8sz-newpvc   Bound    pvc-17431607-83bd-44f3-abe9-79b0108ae1cc   1Gi        RWO            standard       2d21h
workspace-demo1               Bound    pvc-cf2d97d1-0202-4e2e-bafe-f398562627e2   5Gi        RWO            standard       19h
workspace-demo2               Bound    pvc-f1489494-999e-461c-8fe0-96654b625104   5Gi        RWO            standard       58m
workspace-tester1             Bound    pvc-c5fa720d-3ae7-4a2e-95d7-7edc4f668d13   5Gi        RWO            standard       2d22h
```

Step6: hack tensorboard (2/3)

~~(此步為進階功能，若無法實作也無仿)~~

~~修改tensorboard deployment launch script~~

1. ~~先找到deployment的名字 (kubectl get deployment -n kubeflow-user-example-com)~~

```
root@instance-x:/home/hsinhoyeh# kubectl get deployment -n kubeflow-user-example-com
NAME                                                    READY   UP-TO-DATE   AVAILABLE   AGE
ml-pipeline-ui-artifact                                 1/1     1            1           2d2h
ml-pipeline-visualizationserver                         1/1     1            1           2d2h
uci-census-predictor-default-00001-deployment           1/1     1            1           66m
viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a-deployment   1/1     1            1           2m11s
```

2. ~~修改deployment (kubectl edit deployment/<viewer id> -n kubeflow-user-example-com)~~

```
spec:
  containers:
  - args:
    - tensorboard
    - --logdir=/data
    - --path_prefix=/tensorboard/viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a/
    - --bind_all
```

3. ~~然後存檔~~

Step6: hack tensorboard (3/3)

(此步為進階功能，若無法實作也無仿)

1. 修改authorization policy, 改成不檢查權限 (kubectl apply -f hack/allow-all-authorization-policy.yaml)
2. ~~連線到Tensorboard via port forward~~

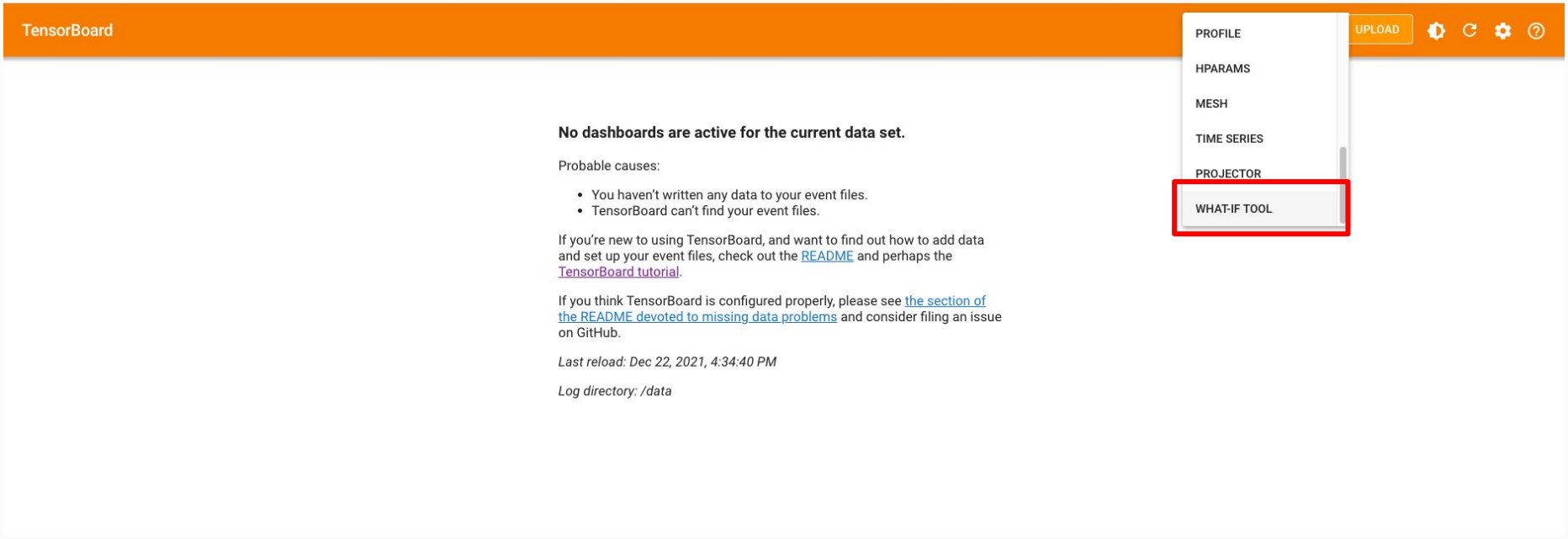
    1. ~~取得svc名稱 (kubectl get svc -n kubeflow-user-example-com)~~

```
root@instance-x:/home/hsinhoyeh# kubectl get svc -n kubeflow-user-example-com
NAME                                                TYPE           CLUSTER-IP      EXTERNAL-IP                                                PORT(S)                        AGE
ml-pipeline-ui-artifact                             ClusterIP      10.96.108.10    <none>                                                     80/TCP                         2d3h
ml-pipeline-visualizationserver                     ClusterIP      10.96.42.220    <none>                                                     8888/TCP                       2d3h
test1                                               ClusterIP      10.96.224.201   <none>                                                     80/TCP                         29h
uci-census                                          ExternalName   <none>          knative-local-gateway.istio-system.svc.cluster.local       <none>                         72m
uci-census-predictor-default                        ExternalName   <none>          knative-local-gateway.istio-system.svc.cluster.local       80/TCP                         72m
uci-census-predictor-default-00001                  ClusterIP      10.96.171.220   <none>                                                     81/TCP                         72m
uci-census-predictor-default-00001-private          ClusterIP      10.96.8.198     <none>                                                     80/TCP,9090/TCP,9091/TCP,8022/TCP   72m
viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a-service   ClusterIP      10.96.29.148    <none>                                                     80/TCP                         7m50s
```

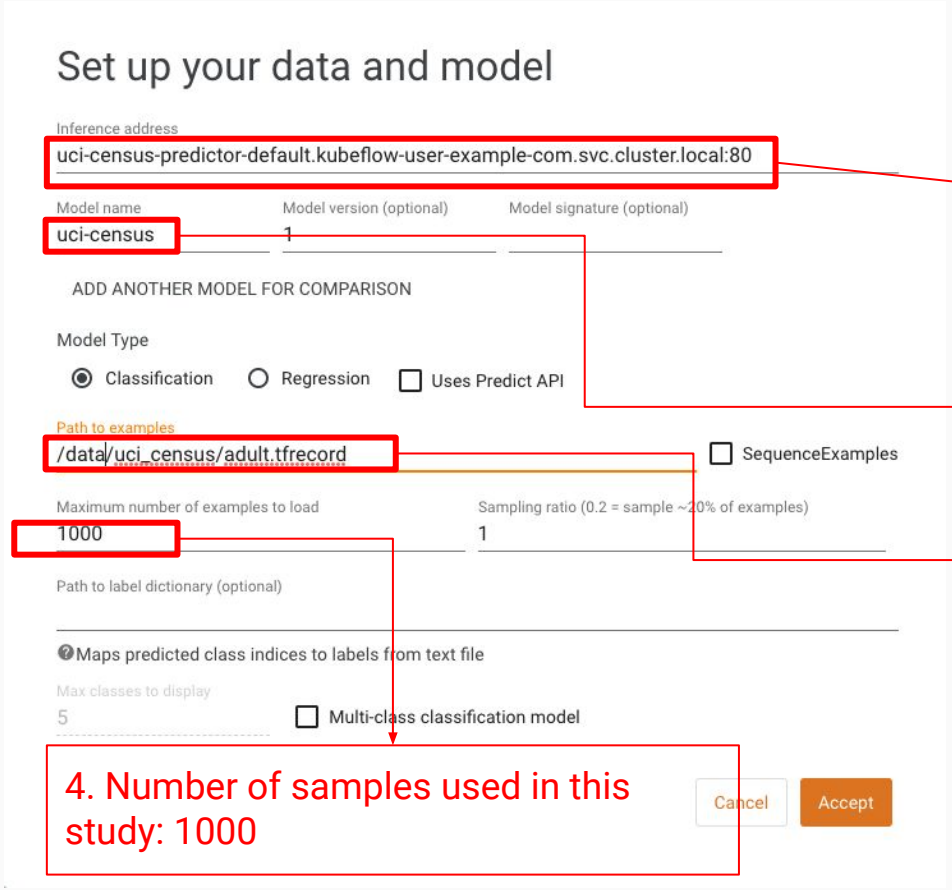
    2. ~~kubectl port-forward svc/viewer-a4e50b4f89b636c7e2d12bebd3da533901b3ea8a-service 8091:80 --address=0.0.0.0 -n kubeflow-user-example-com~~

    3. ~~Open http://localhost:8091~~

Step7: configure what-if tool with tensorboard (1/3)



Step7: configure what-if tool with tensorboard (2/3)



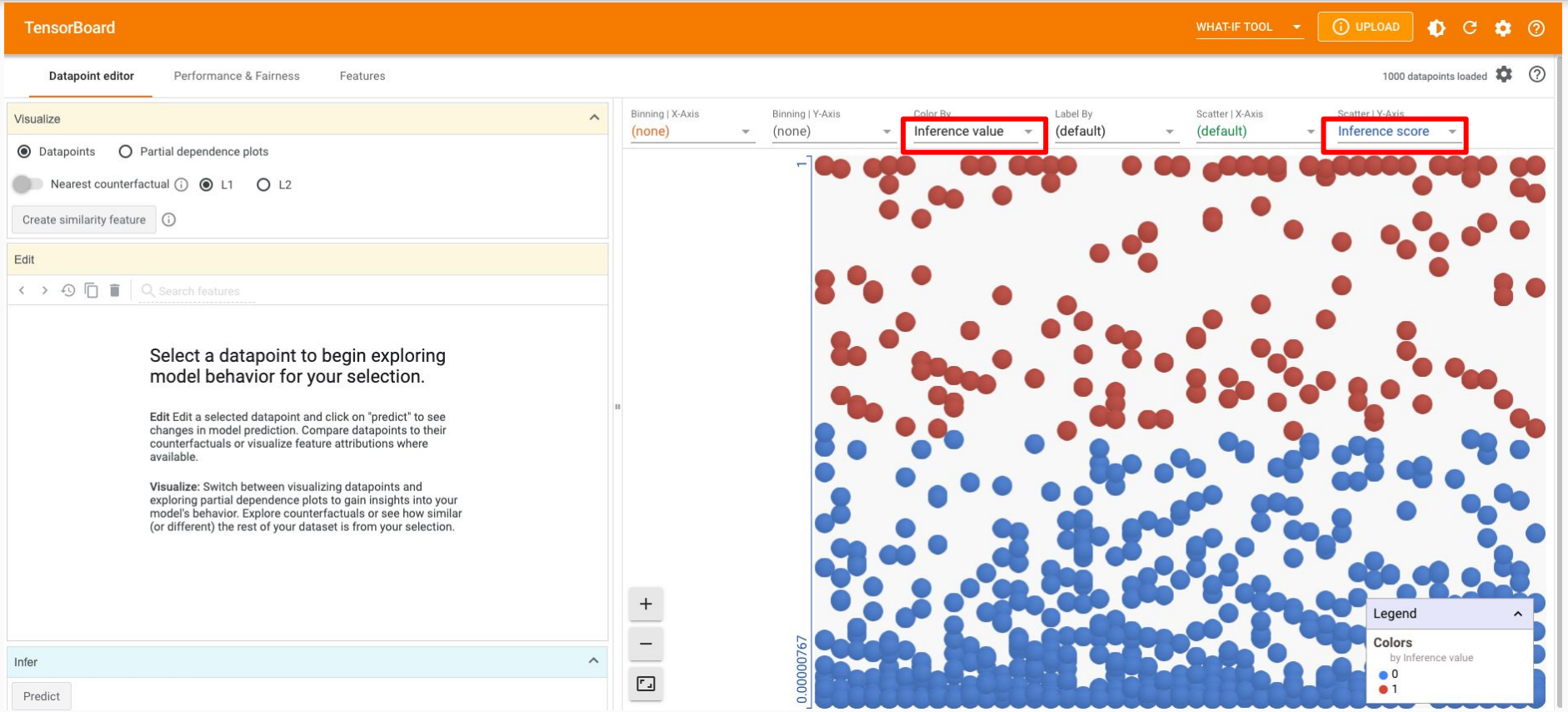
1. Our online model serving address: uci-census-predictor-default.kubeflow-user-example-com.svc.cluster.local:80

2. Model name: uci-census

3. Our test data used in this study: /data/uci_census/adult.tfrecord

4. Number of samples used in this study: 1000

Step7: configure what-if tool with tensorboard (3/3)



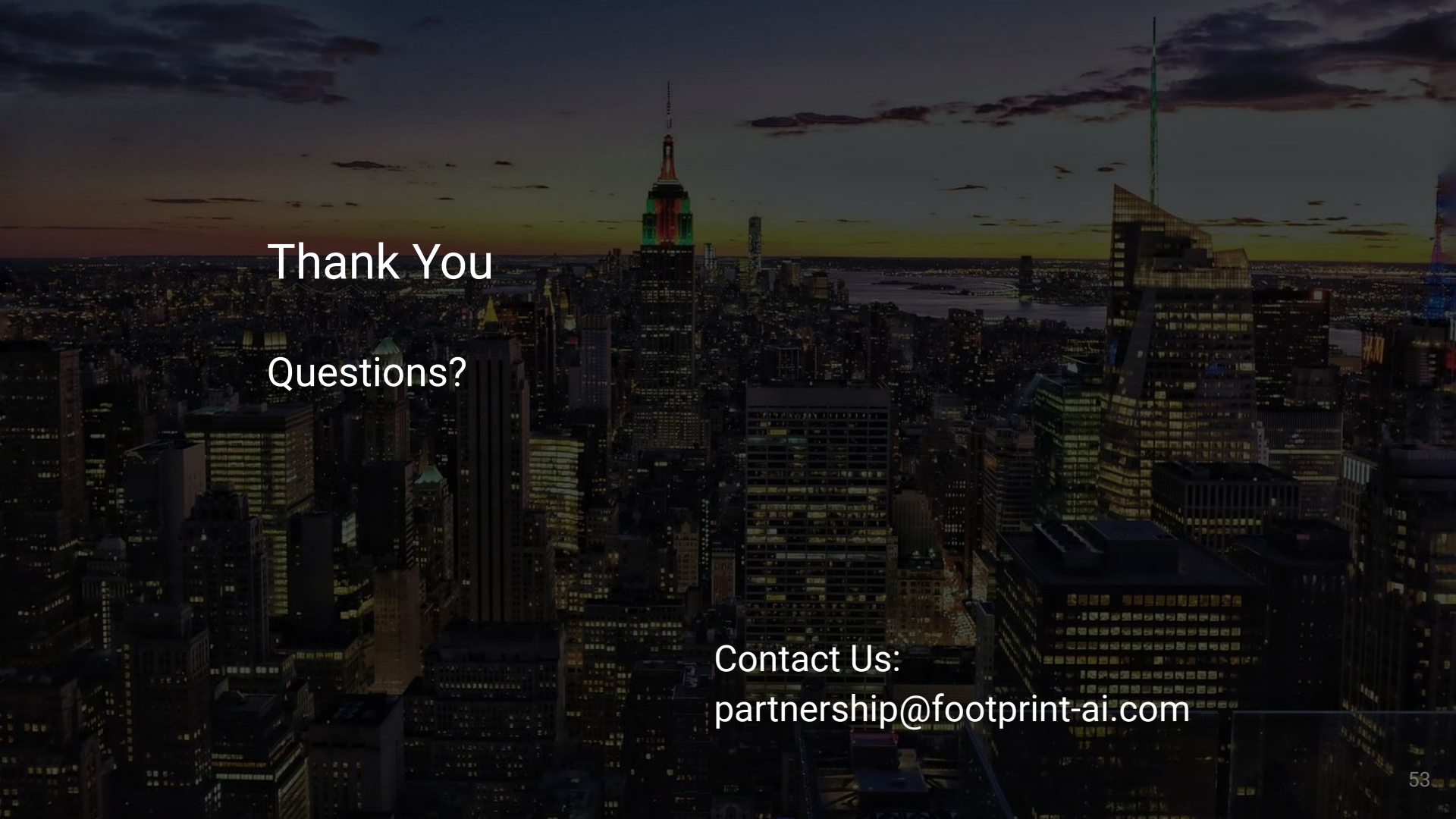
What-if Tool: Other Demos

- UCI Census Salary Prediction
- Smile Detection
- Flower Detection

Ref: https://github.com/pair-code/what-if-tool

What we have achieved...

- Understand fairness and explainable AI
- Build a Kubeflow pipeline for
  - Kserve for serving a pre-build machine learning models
  - Tensorboard for analyzing fairness and explainable on the given data
- Interactive with What-if Tool


Thank You
Questions?
Contact Us:
partnership@footprint-ai.com

Reference

- Documentations
  - https://www.kubeflow.org/
- Kubectl cheatsheet
  - https://kubernetes.io/docs/reference/kubectl/cheatsheet/