What’s News on Kubeflow V1.5

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

信誠金融科技股份有限公司
XINCHEN FINTECH CO., LTD.

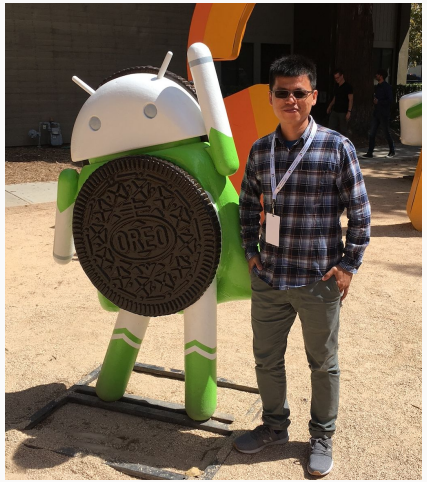
2022/05/25

Download Slides

https://reurl.cc/Lm5lO4

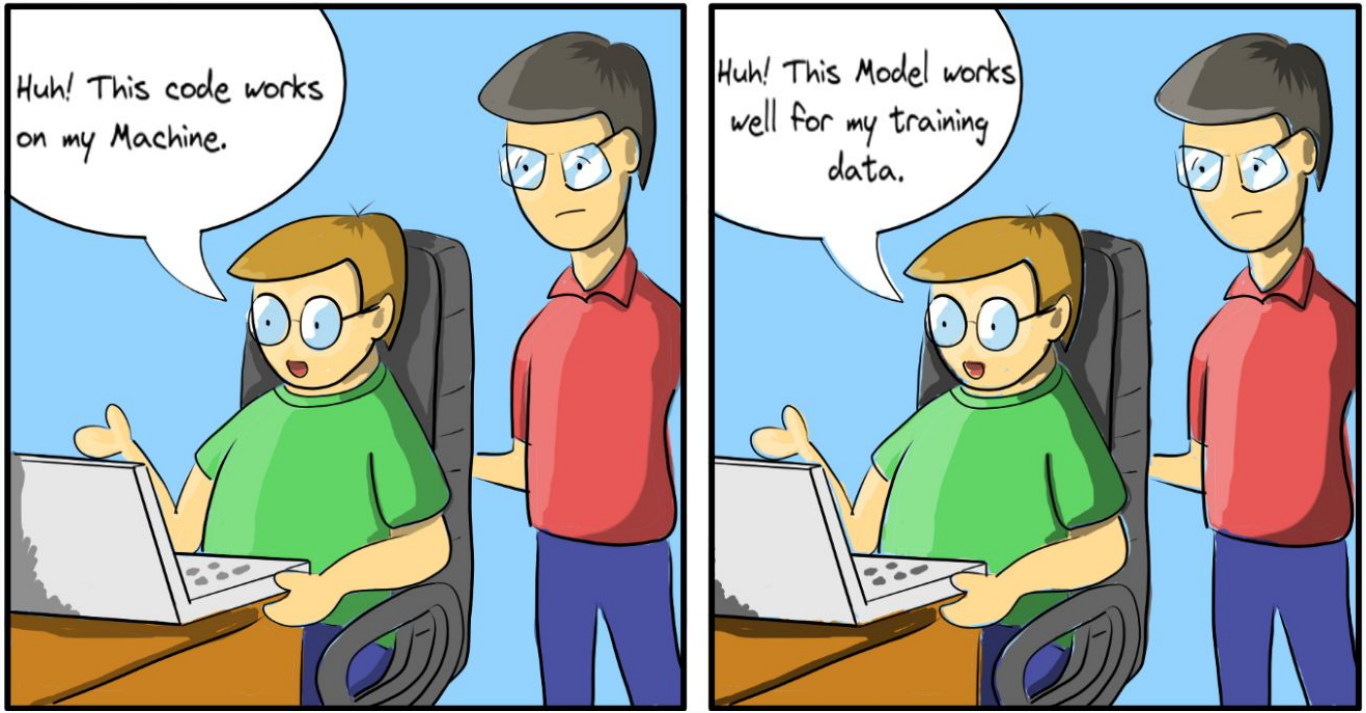
https://github.com/FootprintAI/talks/tree/main/slides

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

A common scenario that we both experienced.



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

What is deployment automation in Machine learning?

DevOps + ML
= MLOps

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/

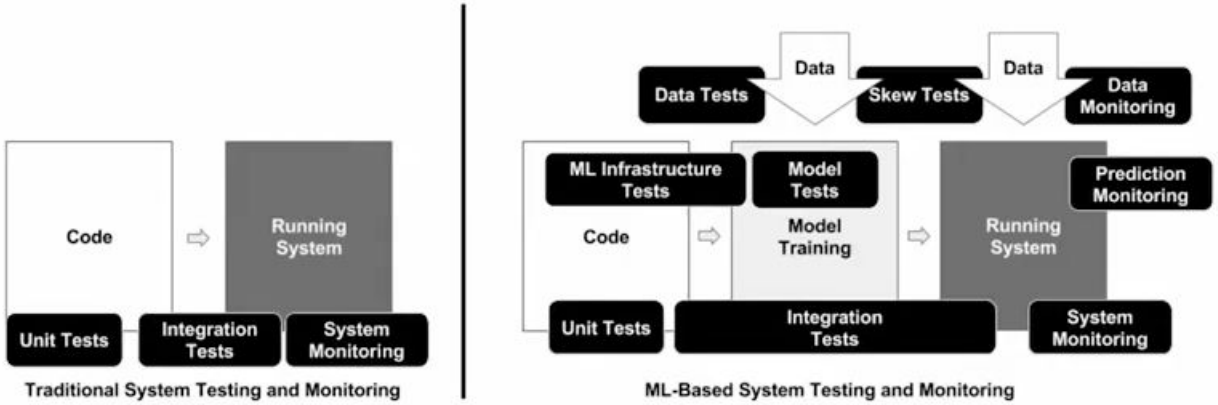
Building & deploying real-world ML application is *hard* and *costly* because of *lack of tooling* that covers end-to-end ML development & deployment

- CloudNext’19

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

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

Traditional vs. ML infused systems

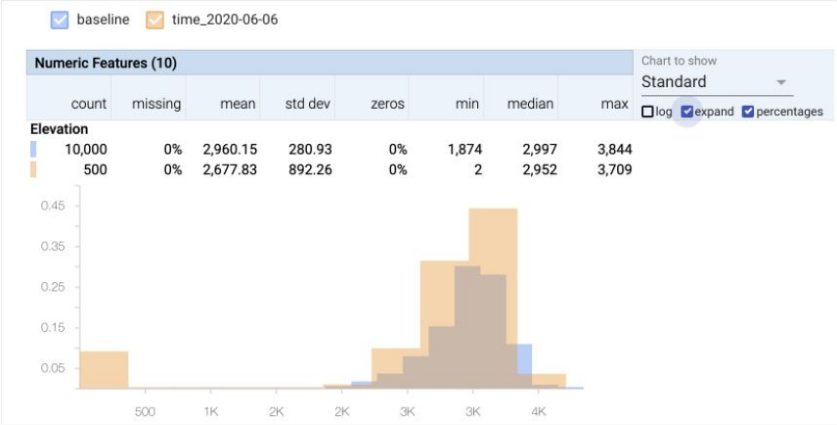


ML introduces two new assets into the software development lifecycle – **data** and **models**.

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

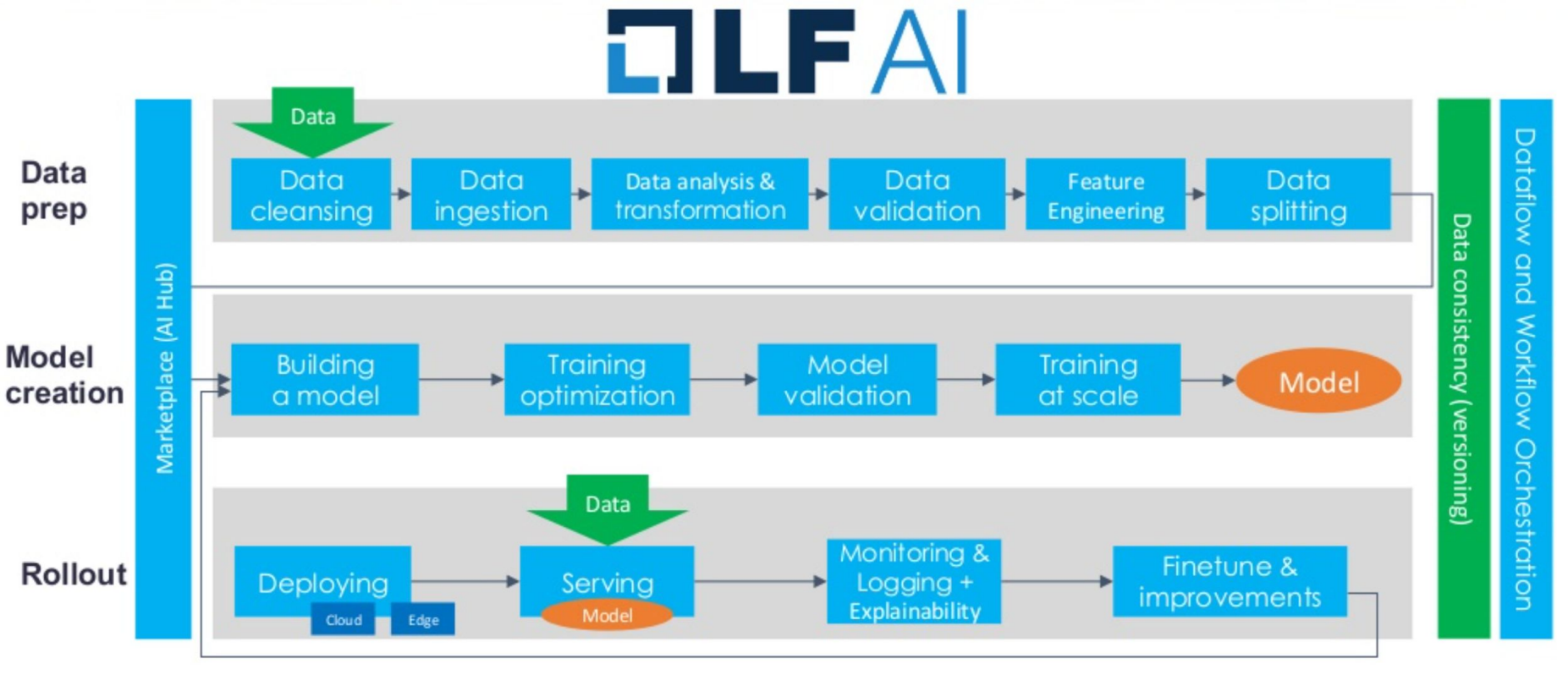
Why we should care about drifting?

- 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

History Of Kubernetes

- Borg: the predecessor to Kubernetes
  - Google revealed the first time of its detail in an academic research paper, describing a “cluster manager that runs hundreds of thousands of jobs, from many thousands of different applications, across a number of clusters each with up to tens of thousands of machines.”[1]
  - A in-house cluster manager system inside Google for running every google services including Gmail, Google Maps, Google Docs…[2]
  - In a scale with ‘over 2 billion containers per week` [3]
- The very first version of Kubernetes was released in 2015
- The latest version is v1.23, released at 2022.

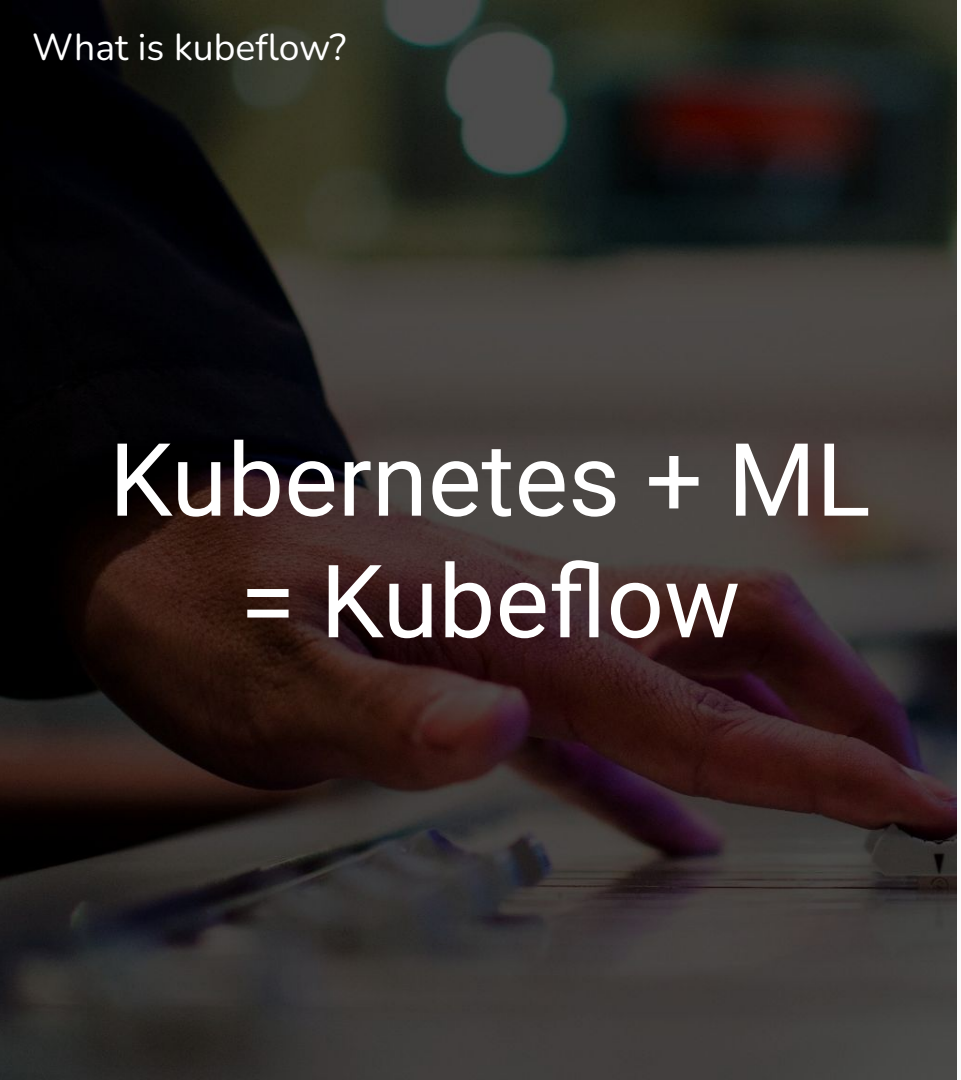
[1] https://research.google/pubs/pub43438/
[2] https://www.wired.com/2016/04/want-build-empire-like-googles-os/
[3] https://cloud.redhat.com/blog/building-kubernetes-bringing-google-scale-container-orchestration-to-the-enterprise

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

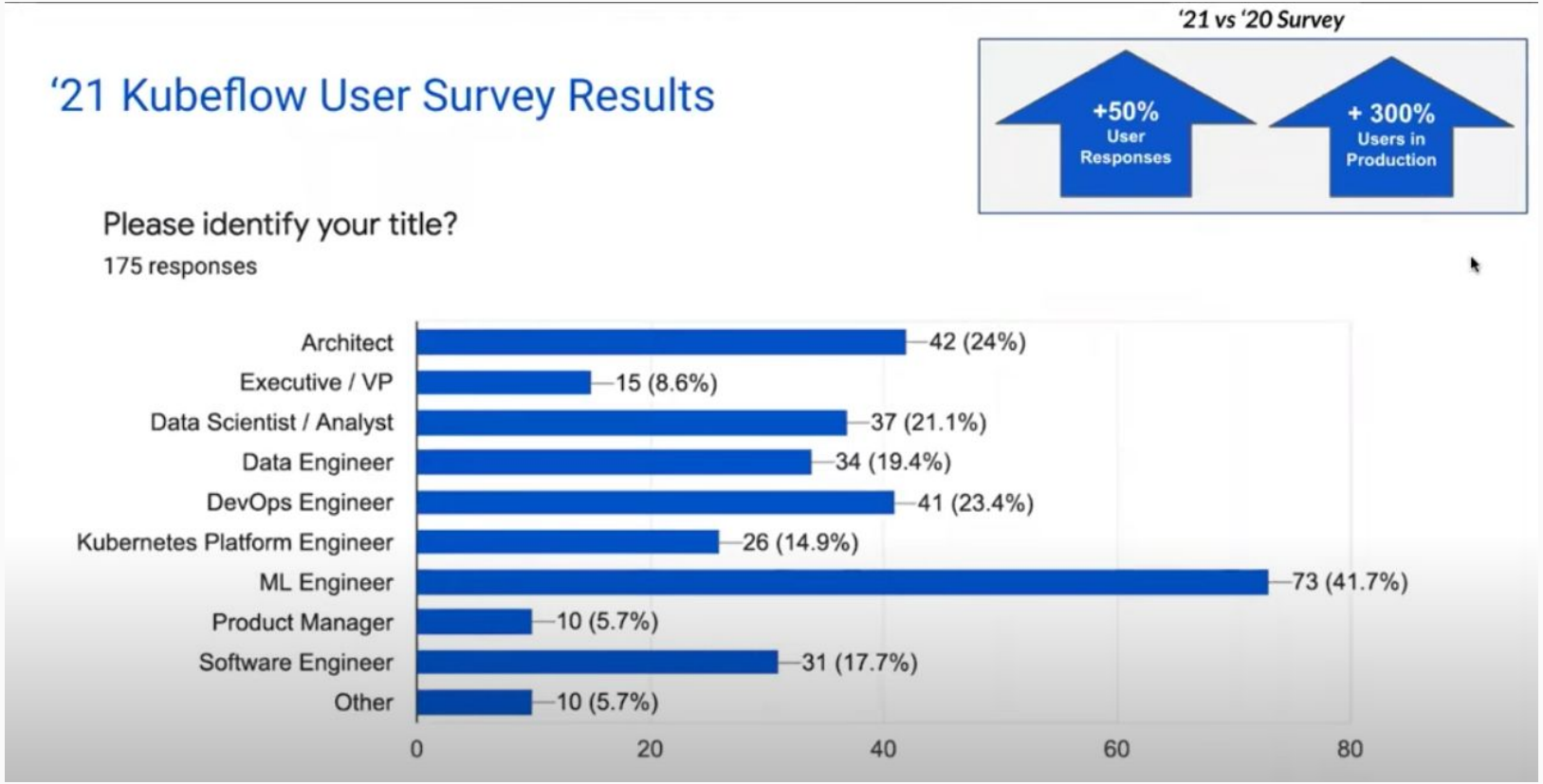


Kubernetes + ML = Kubeflow

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

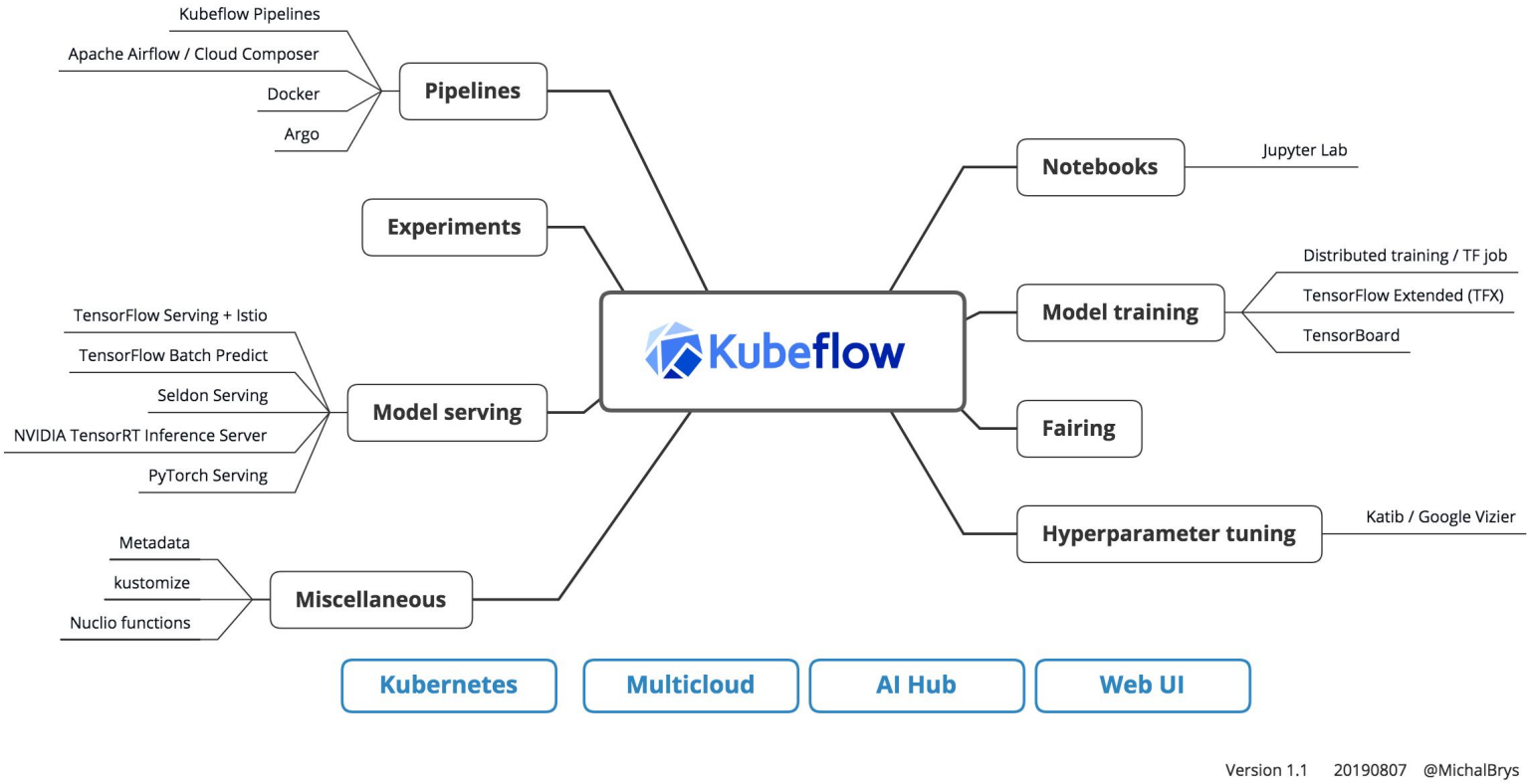
Source: https://www.kubeflow.org/

Kubeflow User Survey



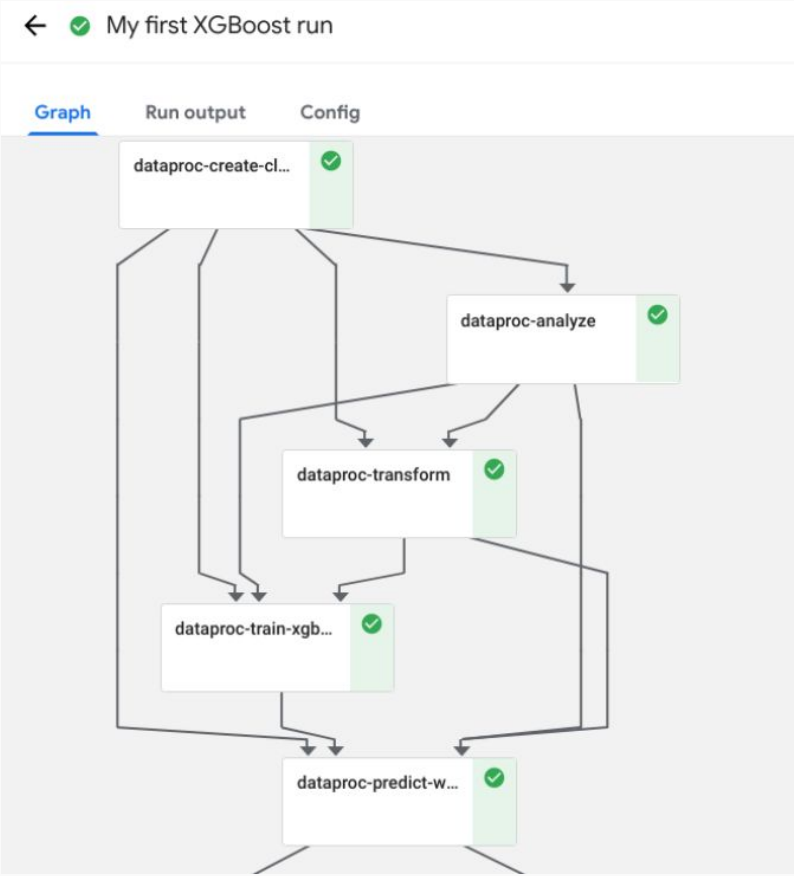
Ref: https://blog.kubeflow.org/kubeflow-continues-to-move-to-production

Architectures



Version 1.1 20190807 @MichalBrys

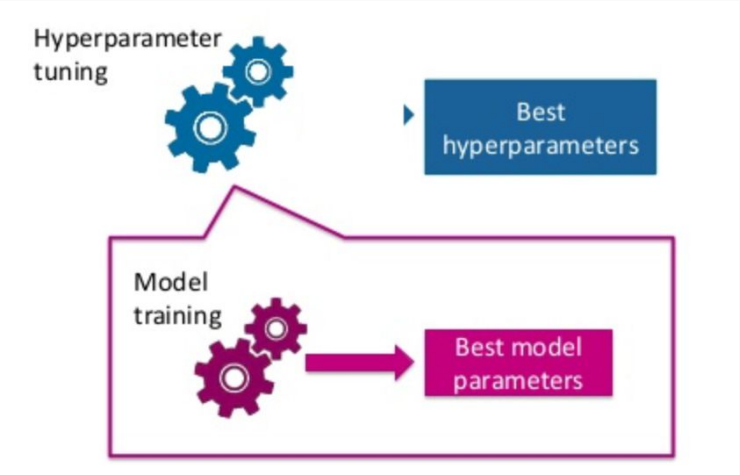
Kubeflow Pipelines



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

Hyperparameter tuning

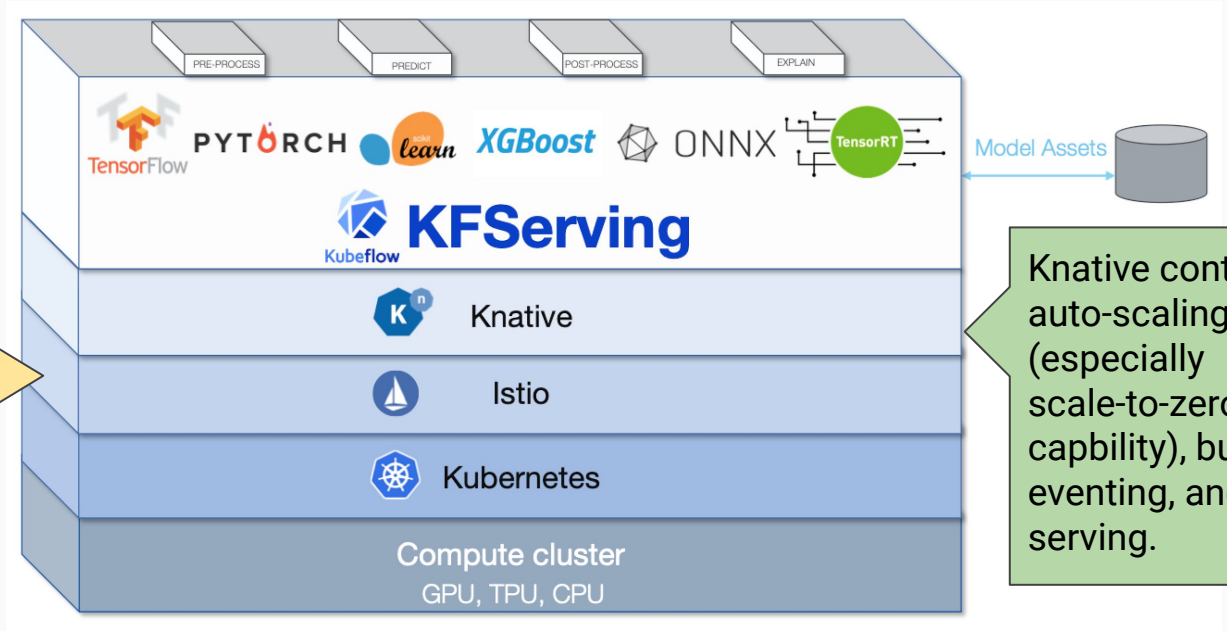
- Model Parameter vs Hyperparameter
  - Model parameters that will learn on its own during training process by the ML model, ex: weights and biases for a classifier.
  - Hyperparameter that directly control the behavior of training algorithm.



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

KServe

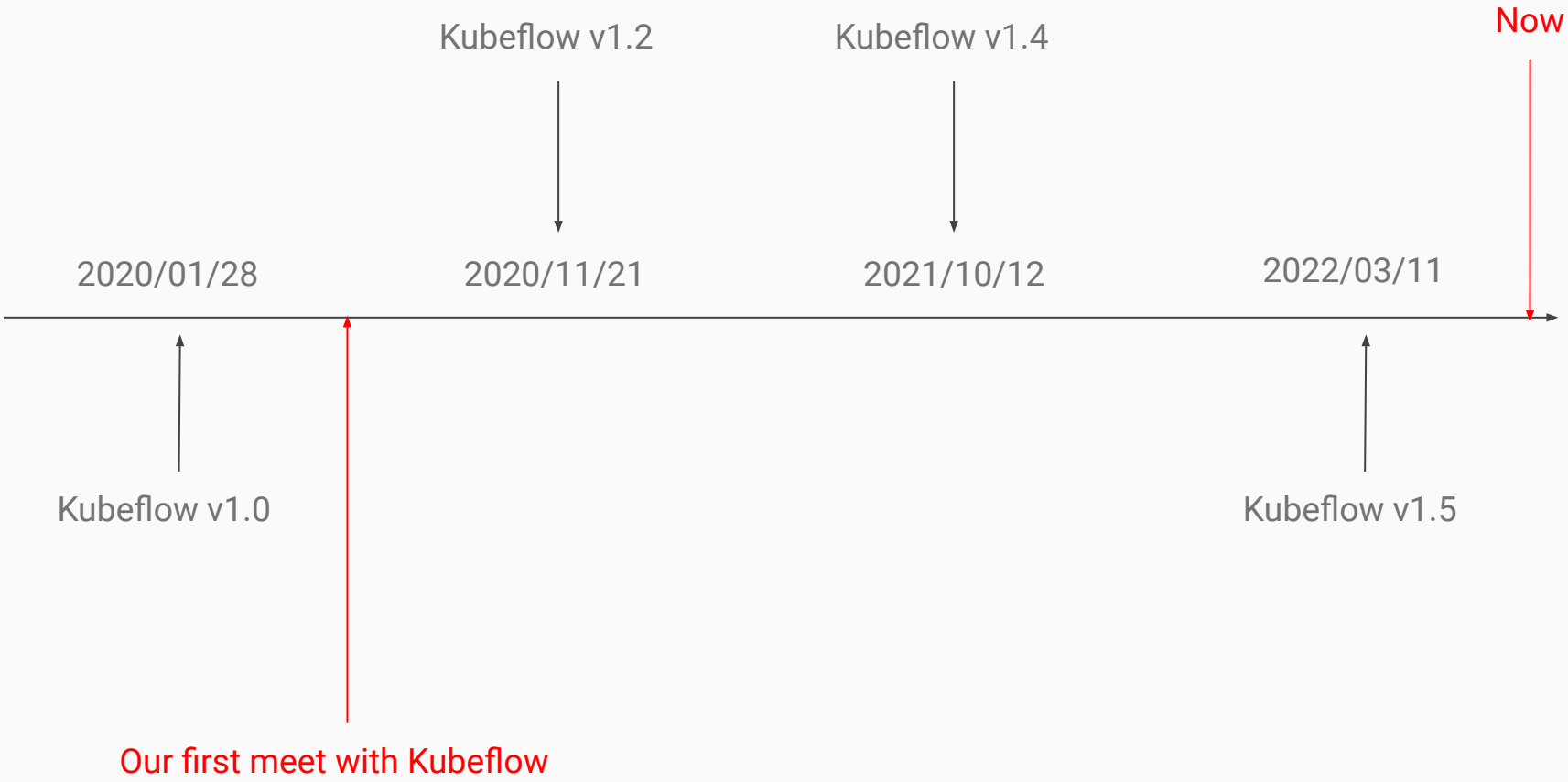
Istio is a service mesh, controlling who can access what services.



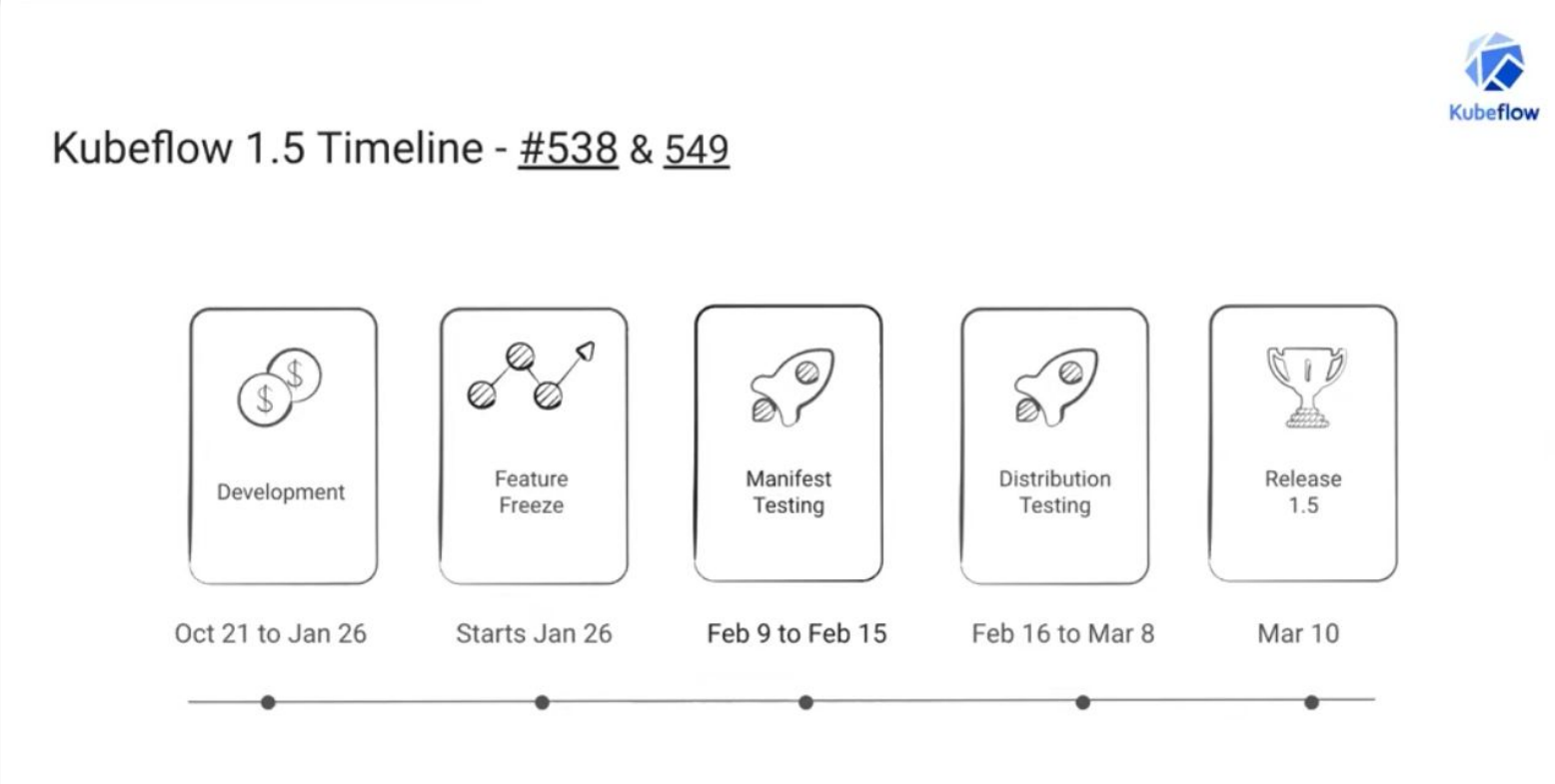
Knative controls auto-scaling (especially scale-to-zero capbility), building, eventing, and serving.

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

Kubeflow Release timeline



Kubeflow Release Cycle



Kubeflow 1.6 release: https://github.com/kubeflow/manifests/issues/2194

Ref: https://www.youtube.com/watch?v=QNNCM9Kq3Q0&t=1460s

Lessons learned in the wild - presented at 20210222 (v1.0.2)

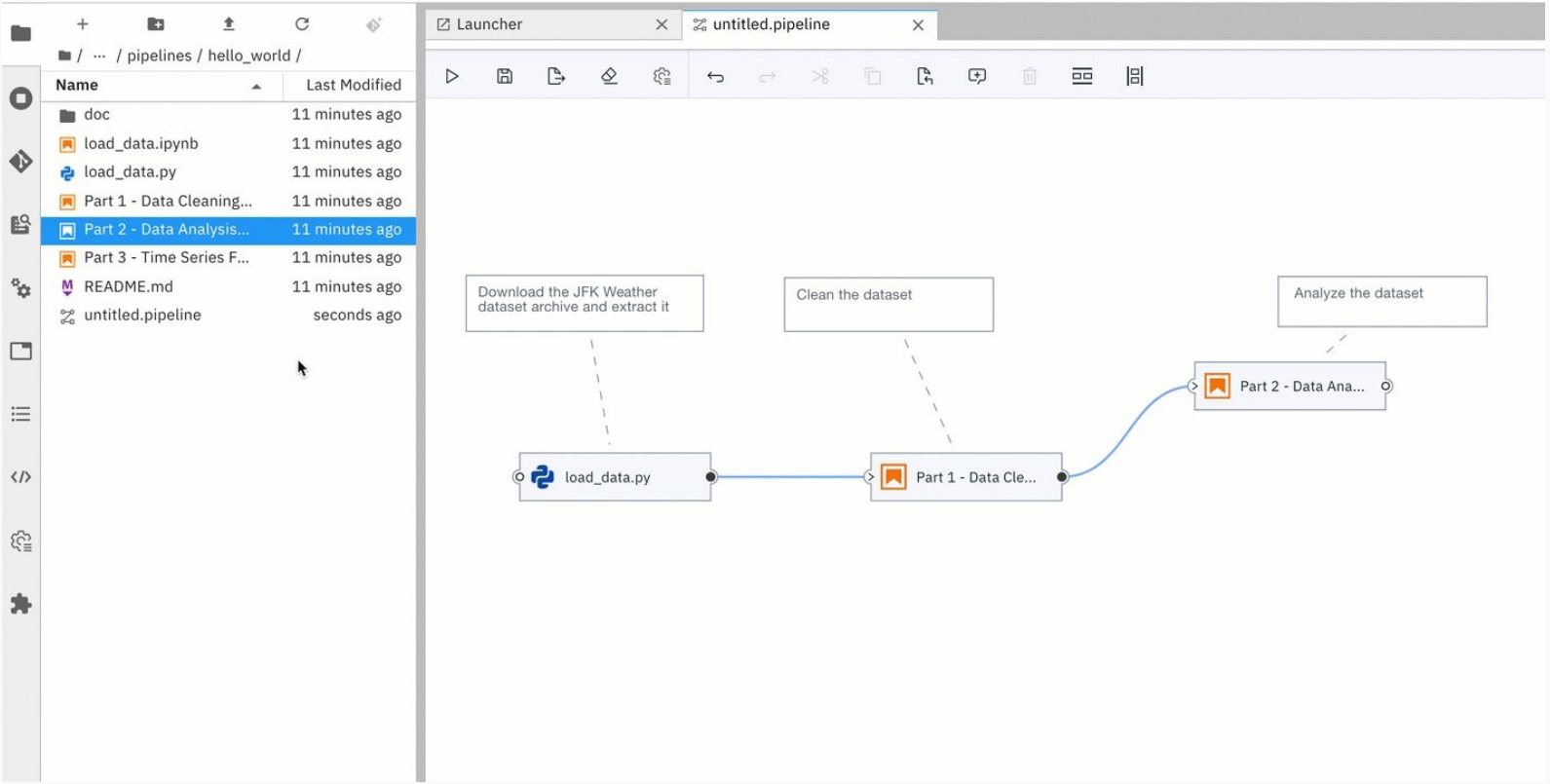
- Kubeflow favor GKE, not as friendly on EKS/AKS/On-Prem.
  - Most features/examples are built on top of GCP (Google Cloud Platform)
- ~~Buggy~~Unstable System (Yet, we knew it is in version 1.x now)
  - Early staged components
    - Katib (v0.10.0), kfserving(v0.5.0) are not yet production-ready release.
    - Multi-Tenancy: resource sharing and access-control isolation.
  - Out-of-date dependency
    - Istio: v1.3.1 is using in kubeflow v1.2 and v1.6 will be kubeflow v1.3
- Hidden Cost
  - ~~Linux~~Open source is free if you time have no value. - Jamie Zawinski
- Off-the-shelf software
  - No way to build better UI/flow to fulfill your business requirement without digging into components
  - Require more knowledge on k8s/devops expertise to be able to master it

What’s news?

- Elyra: Workflow Visualization
- Tekton Pipeline
- MultiModel Serving in Kserve
- ML UI
- Katib UI

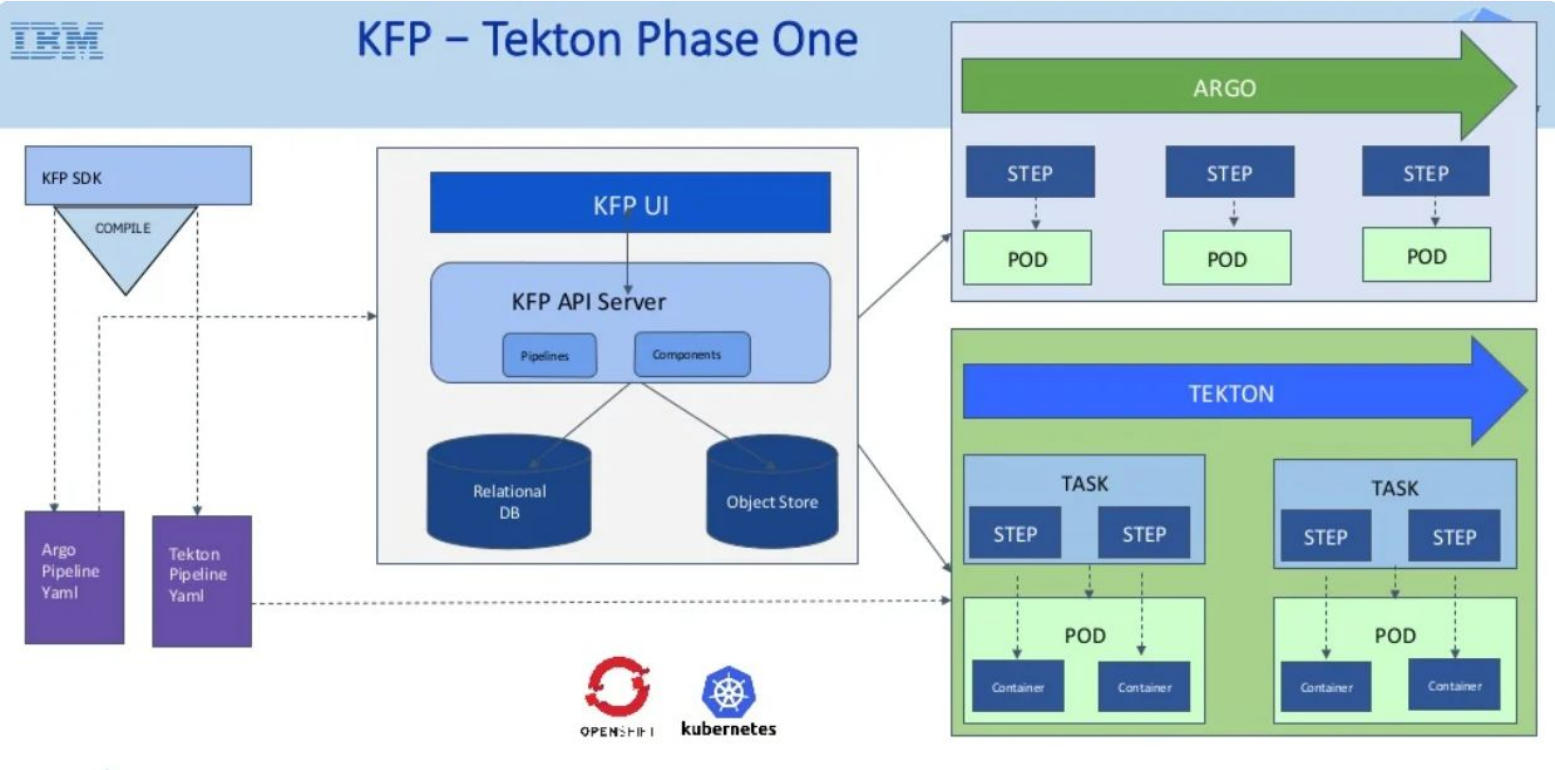
Source: https://www.kubeflow.org/

Elyra: Workflow Visualization



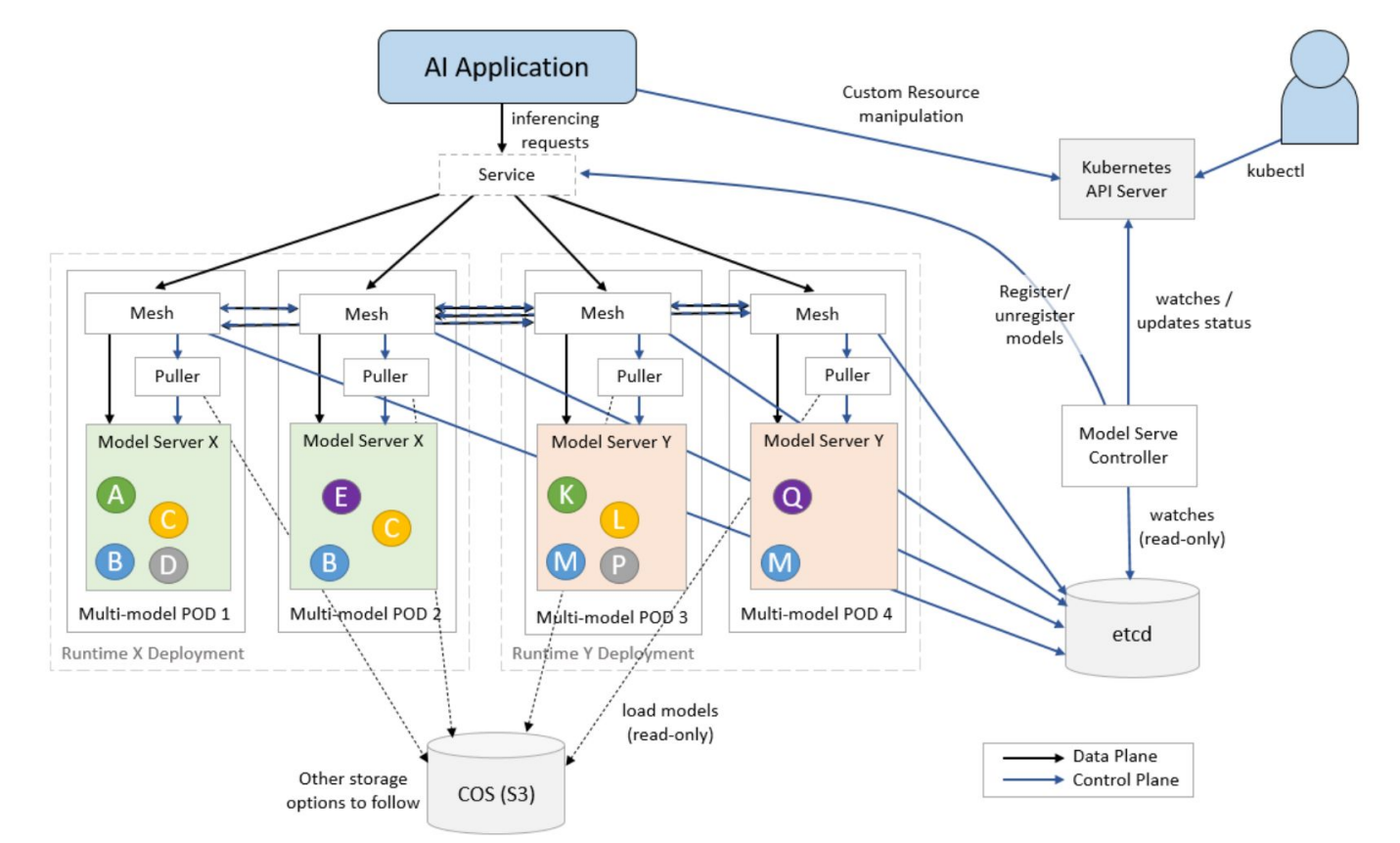
Ref: https://github.com/elyra-ai/elyra

Tekton Pipeline



Ref: https://pt.slideshare.net/AnimeshSingh/kubeflow-pipelines-with-tekton-236769976

Kserve: MulitModel Serving with Model Mesh



Ref: https://github.com/kserve/modelmesh-serving

Kserve: Model UI (1/2)

Kubeflow

kubeflow-user (Owner)

- Home
- Notebooks
- Tensorboards
- Models
- Snapshots
- Volumes
- Experiments (AutoML)
- Experiments (KFP)
- Pipelines
- Runs
- Recurring Runs
- Artifacts
- Executions

← Model server details

DELETE

flowers

OVERVIEW | DETAILS | METRICS | LOGS | YAML

Status	Ready
URL external	http://flowers.kubeflow-user.example.com
URL internal	http://flowers.kubeflow-user.svc.cluster.local/v1/models/flowers:predict
Component	predictor
Storage URI	gs://kfserving-samples/models/tensorflow/flowers
Runtime	Tensorflow 1.14.0
Protocol Version	

InferenceService Conditions

Status	Type	Last Transition Time	Reason	Message
✓	IngressReady	23 minutes ago		
✓	PredictorConfigurationReady	23 minutes ago		

Ref: https://github.com/kubeflow/website/blob/master/content/en/docs/external-add-ons/kserve/webapp.md

Kserve: Model UI (2/2)

Kubeflow

Home
Notebooks
Tensorboards
Models
Snapshots
Volumes
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions

kubeflow-user (Owner)

← Model server details

DELETE

flowers

OVERVIEW | DETAILS | METRICS | LOGS | YAML

Predictor: flowers-predictor-default-hb8r5

Total CPU Usage

60 ms
40 ms
20 ms
0 ns

15:48 15:50 15:52 15:54 15:56 15:58 16:00 16:02

istio-proxy kfserving-container queue-proxy

Total Memory Usage

300 MB
200 MB
100 MB
0 B

15:48 15:50 15:52 15:54 15:56 15:58 16:00 16:02

istio-proxy kfserving-container queue-proxy

Request Volume by Revision

1.0 ops
0.5 ops
0 ops
-0.5 ops
-1.0 ops

No data points

15:48 15:50 15:52 15:54 15:56 15:58 16:00 16:02

Response Time by Revision

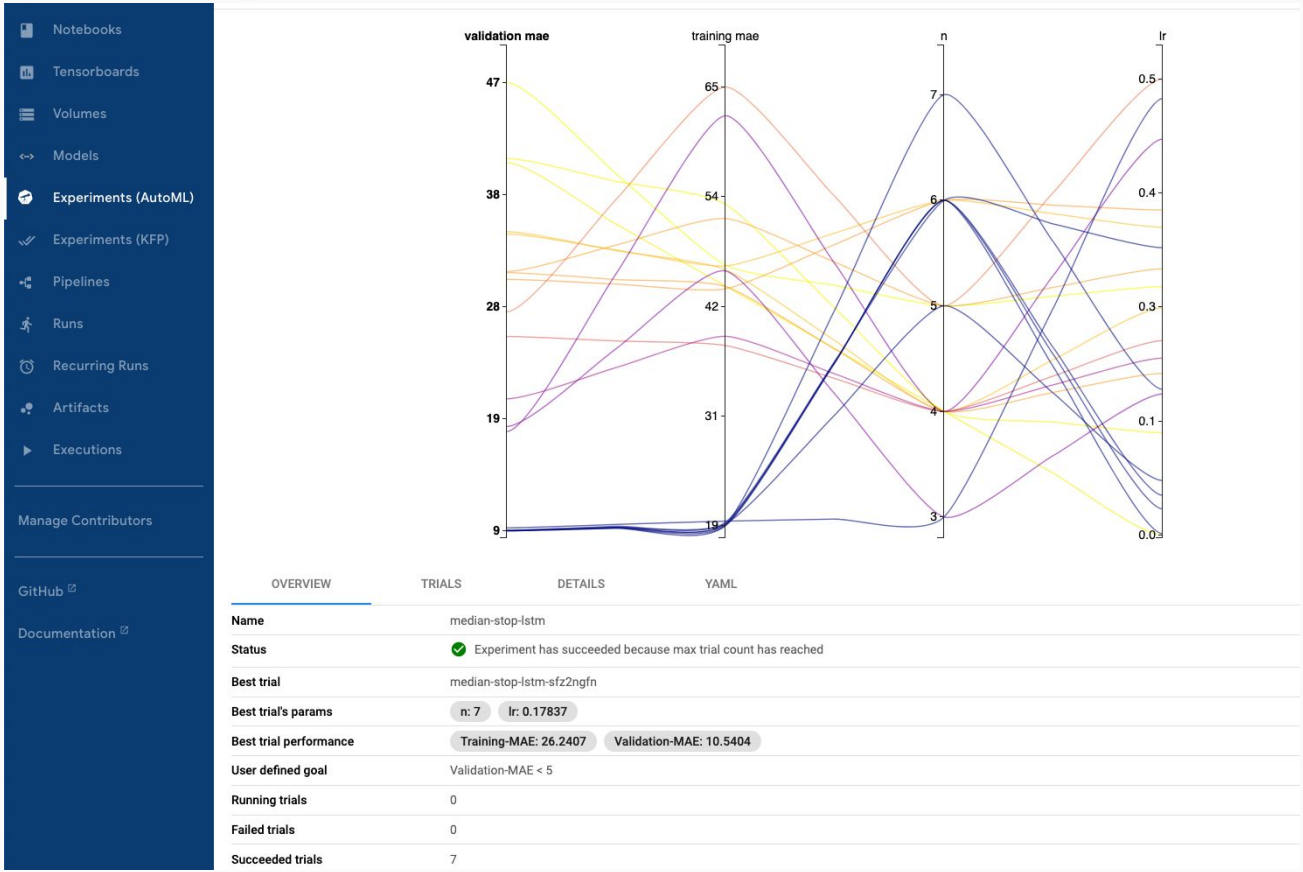
1.0 ms
0.5 ms
0 ms
-0.5 ms
-1.0 ms

No data points

15:48 15:50 15:52 15:54 15:56 15:58 16:00 16:02

Ref: https://github.com/kubeflow/website/blob/master/content/en/docs/external-add-ons/kserve/webapp.md

Katib: UI



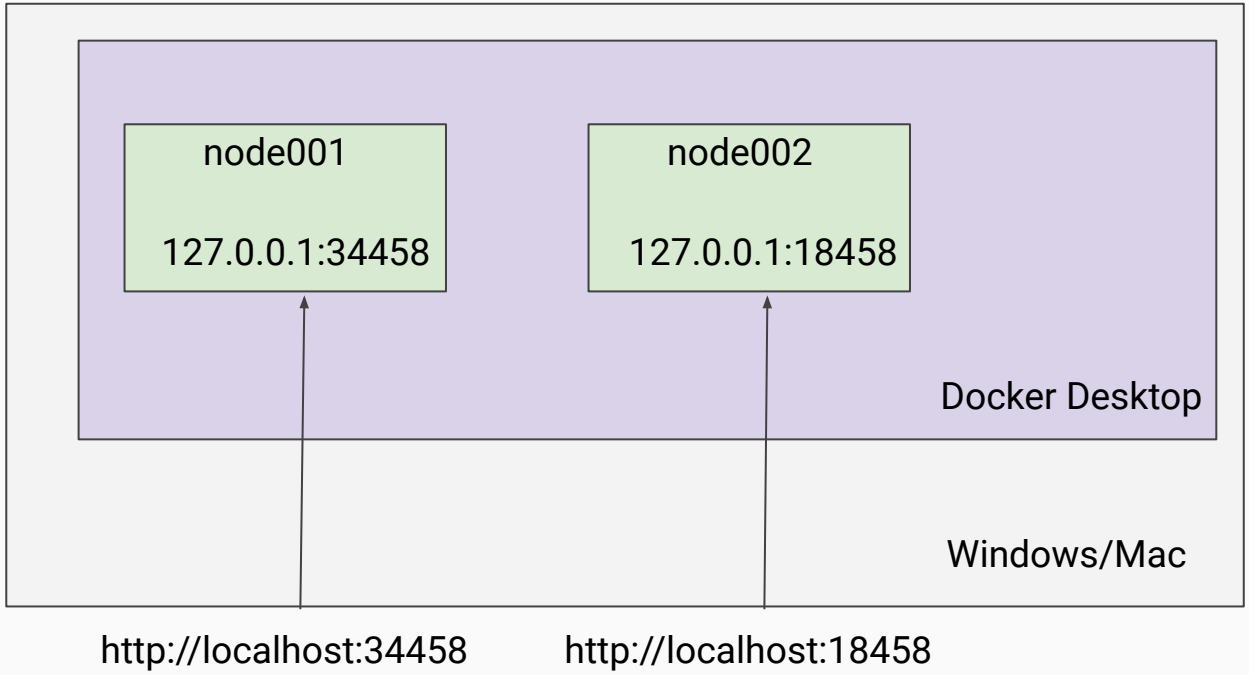
Ref: https://github.com/FootprintAI/kubeflow-workshop/blob/main/tutorials/stockprice-with-lstm/1.visualize-and-build-stockprice-model.ipynb

multikf(馬蒂庫夫)

Our open-sourced project for one-clicked running multiple Kubeflow instances within the single host machine.

multikf: One-click Installation

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



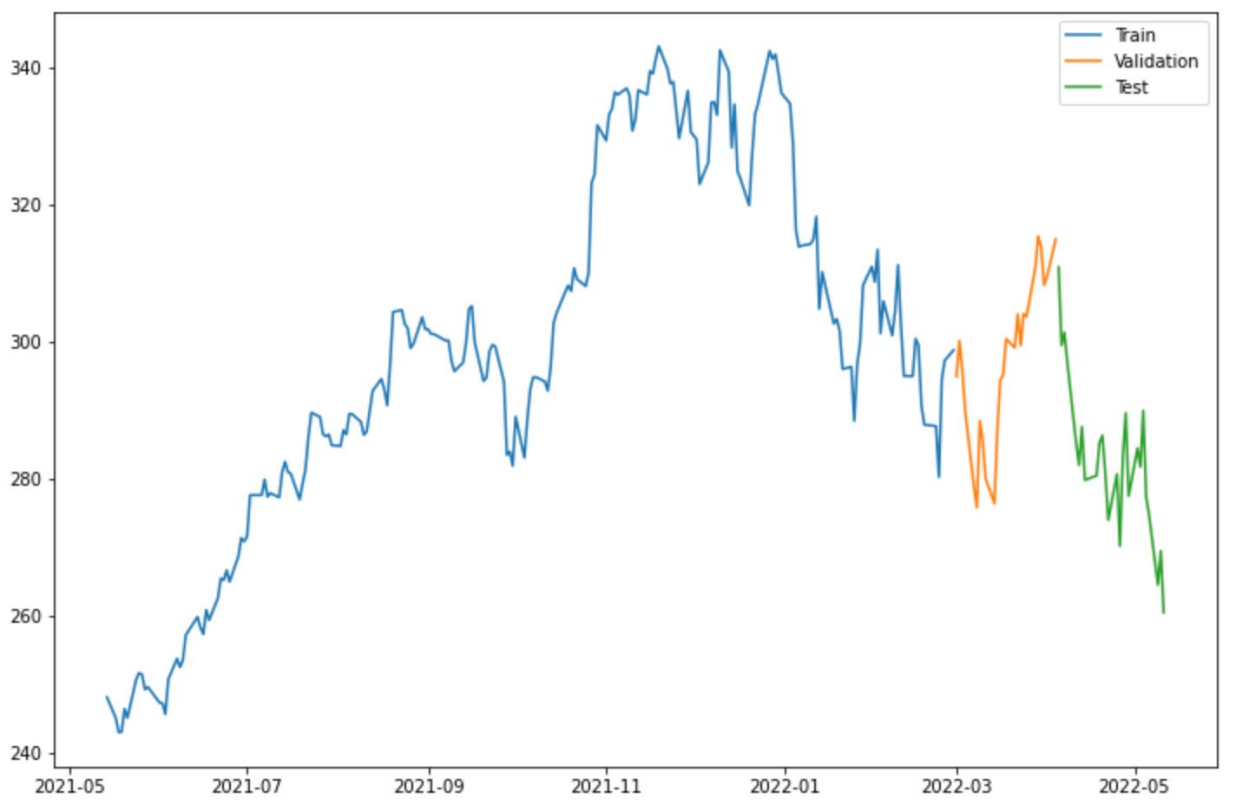
multikf: One-click Installation

```
// install dockerd (windows)
https://github.com/FootprintAI/kubeflow-workshop/blob/main/install/windows/dockerd.bat.md

// install multikf (windows)
wget https://github.com/FootprintAI/multikf/raw/main/build/multikf.windows.exe
chmod +x multikf.windows.exe

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

// connect kubeflow
./multikf.windows.exe connect kubeflow node002
```

Katib Case Study

- How to predict the stock market price of tomorrow?
- How to determine the parameter is good enough?

CastStudy: Data overview



Ref: https://github.com/FootprintAI/kubeflow-workshop/blob/main/tutorials/stockprice-with-lstm/1.visualize-and-build-stockprice-model.ipynb

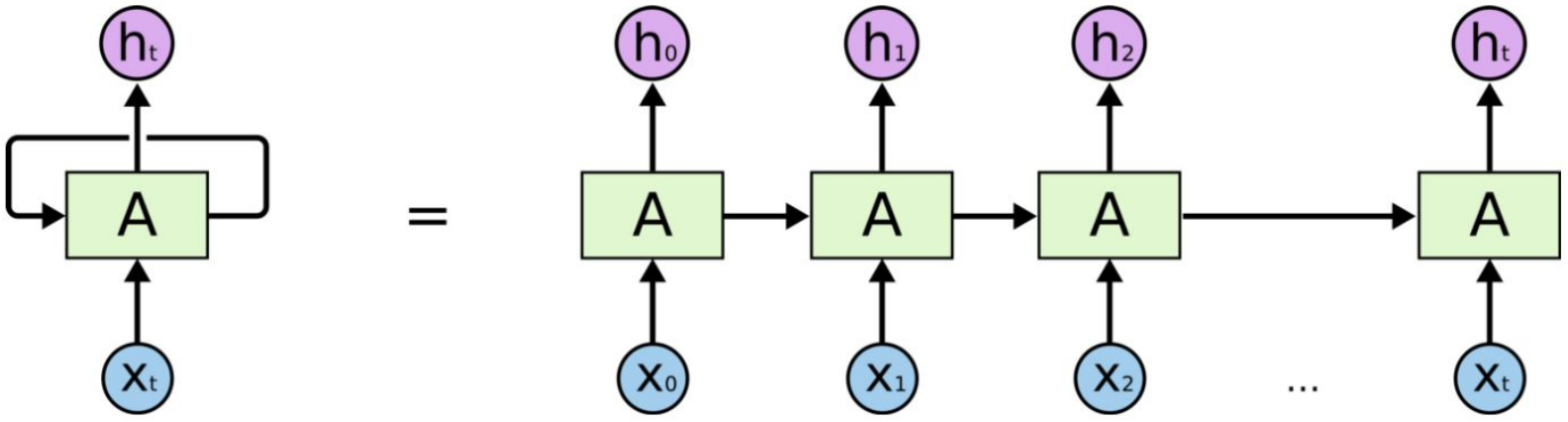
What sequences means in a sentence?

Input: I am a good guy

t0	I
t1	am
t2	a
t3	good
t4	guy
t5	?

What Is RNN (Recurrent Neural Network)?

Recurrent Neural Network (RNN) takes decisions on CURRENT (Xt) and PREVIOUS (Xt-1) inputs. Especially useful in topics including machine translation, speech recognition.



Recurrent Neural Network

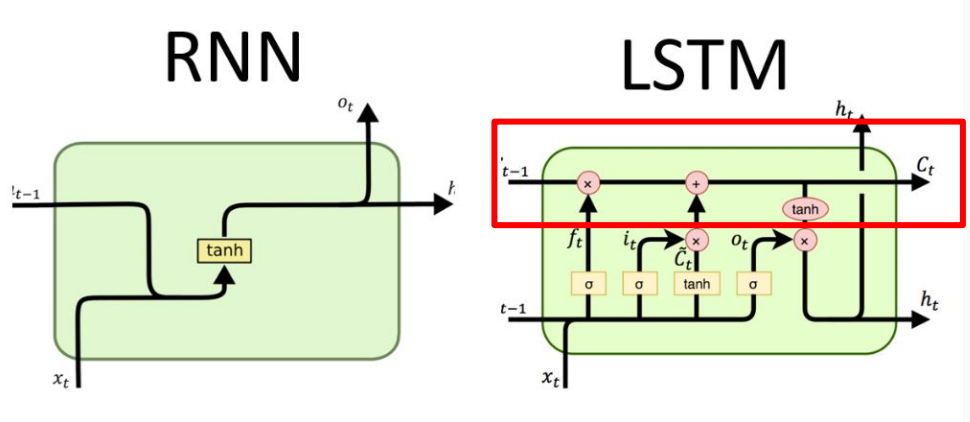
https://towardsdatascience.com/introduction-to-recurrent-neural-network-27202c3945f3

What Is LSTM (Long-short term memory)?

RNN only remember the latest things from X and it didn't remember(no memory) anything before at the beginning.

LSTM provides an information highway to let the neuron to selectively choose

1. forget from its memory (focus on the current inputs)
2. Listens to what information it added into memory (though information highway)



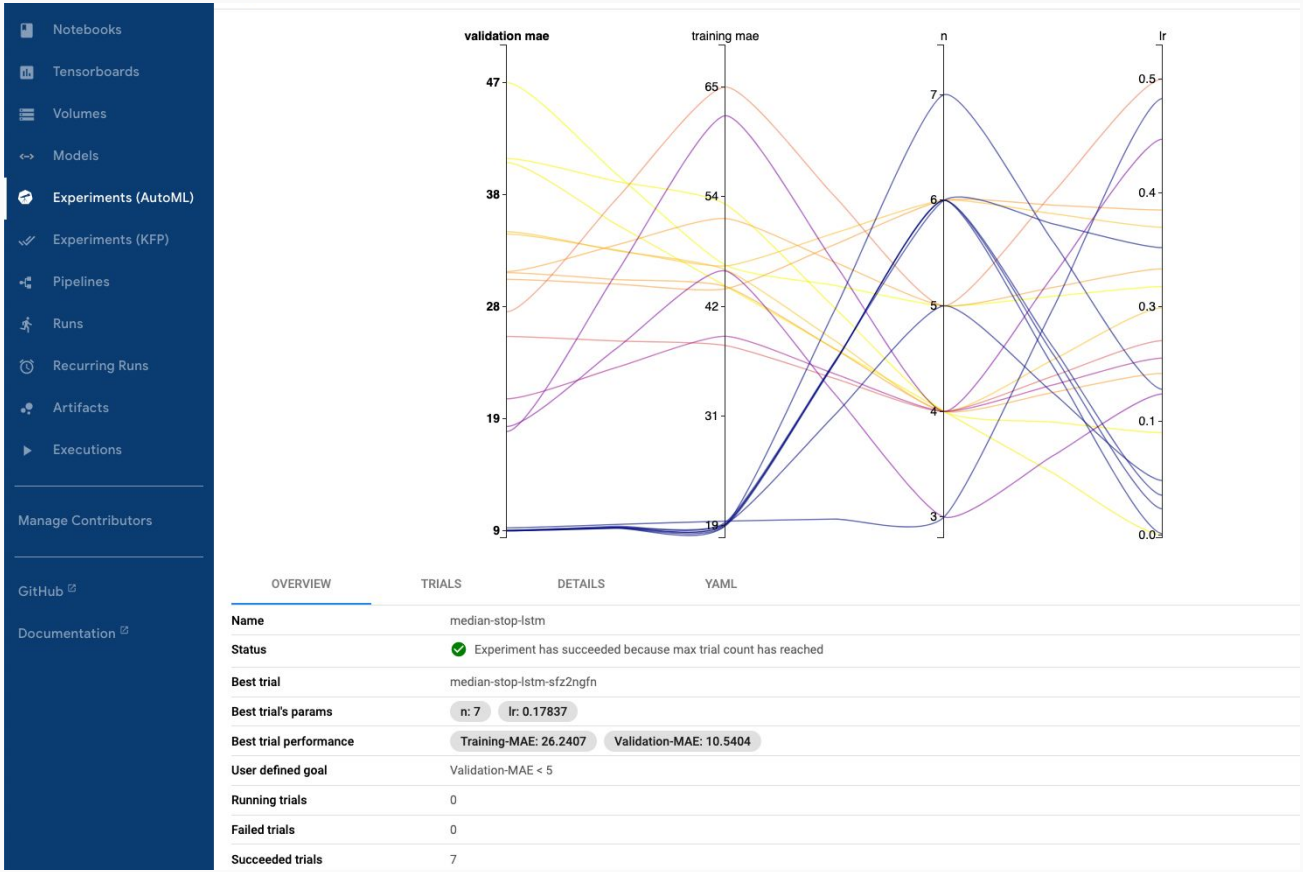
https://towardsdatascience.com/introduction-to-recurrent-neural-network-27202c3945f3
https://www.quora.com/How-is-LSTM-different-from-RNN-In-a-layman-explanation

CastStudy: Build a tensorflow model with LSTM

```
# Now we build a tensorflow model with LSTM
# the network is not something fancy, it is just a common way to build the model
# And you can also find a better model online.

# Note the input tensor size should have equal length with the windows size.
# In this example, we use windows size (n=3), so the input tensor is layers.Input((3, 1)

from tensorflow.keras.models import Sequential
from tensorflow.keras.optimizers import Adam
from tensorflow.keras import layers

model = Sequential([layers.Input((3, 1)),
                    layers.LSTM(64),
                    layers.Dense(32, activation='relu'),
                    layers.Dense(32, activation='relu'),
                    layers.Dense(1)])

model.compile(loss='mse',
              optimizer=Adam(learning_rate=0.001),
              metrics=['mean_absolute_error'])

model.fit(X_train, y_train, validation_data=(X_val, y_val), epochs=100)
```

Ref: https://github.com/FootprintAI/kubeflow-workshop/blob/main/tutorials/stockprice-with-lstm/1.visualize-and-build-stockprice-model.ipynb

CastStudy: Optimize model parameters with Katib



Ref: https://github.com/FootprintAI/kubeflow-workshop/blob/main/tutorials/stockprice-with-lstm/1.visualize-and-build-stockprice-model.ipynb

Conclusion

- The tooling and features on Kubeflow is more user-friendly compared to v1.0.2, all of this is contributed to the open source community and release automation.

Q&A

Thank You

Questions?

Contact Us:
partnership@footprint-ai.com