

# How cloud-native architectures can accelerate the adoption of machine learning

葉信和 / Hsin-Ho Yeh  
Software Engineer / Founder / CEO @ 信誠金融科技  
[hsinho.yeh@footprint-ai.com](mailto:hsinho.yeh@footprint-ai.com)

# Download Slides

<https://rb.gy/ku6jn0>



# 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>
  - **KaFeiDo**: machine learning platform for green economy
- 2016 - 2020 at IglooInsure (16M+ in series A+ 2020)
  - Provide digital insurance for e-economic 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

- Rapid growth of ML applications.
- What is cloud native architecture.
- How large scale model is trained.
- FAQ

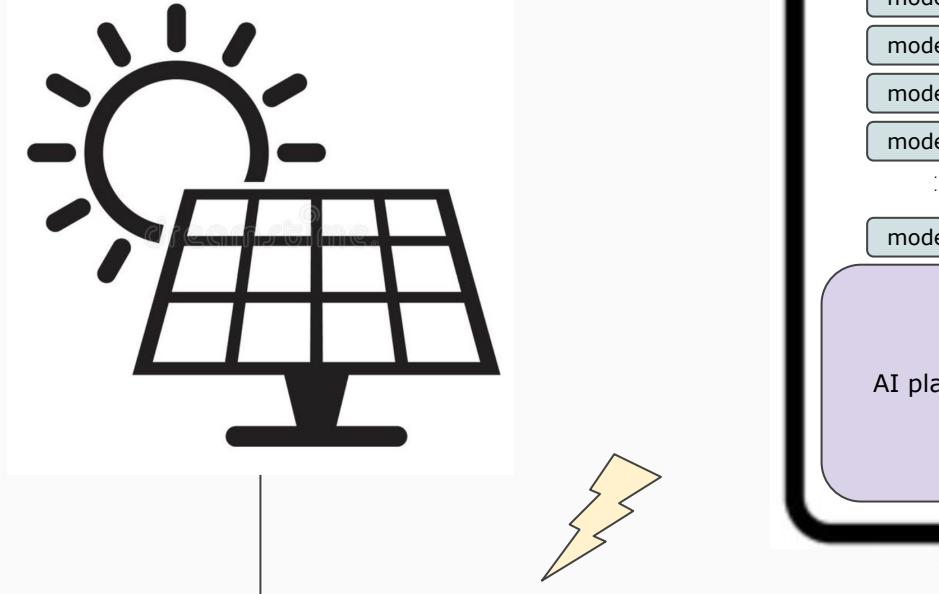
# Status and the challenge of AI/ML

- 無所不在的AI - 但需要把資料送回後端機房處理
  - Globally, data transmission networks consumed 260-340 TWh in 2020, or 1.1~1.4% of global electricity use. [1]
- MLDL模型效能特別好 - 但只適用於已知資料集
  - Is it reasonable to use copurs between 2010-2015 to predict what people is talking about in 2022?
  - Is it reasonable to train a car detector from 90s car dataset?
- 我們的電腦跑得很快 - 但需要機房的低溫設置避免熱當
  - In 2014, data centers in the U.S. consumed an estimated 70 billion kWh, representing about 1.8% of total U.S. electricity[2]

[1] <https://www.iea.org/reports/data-centres-and-data-transmission-networks>

[2] <https://www.techtarget.com/searchdatacenter/tip/How-much-energy-do-data-centers-consume>

# Ubiquitous AI platform (夢裡什麼都有?)



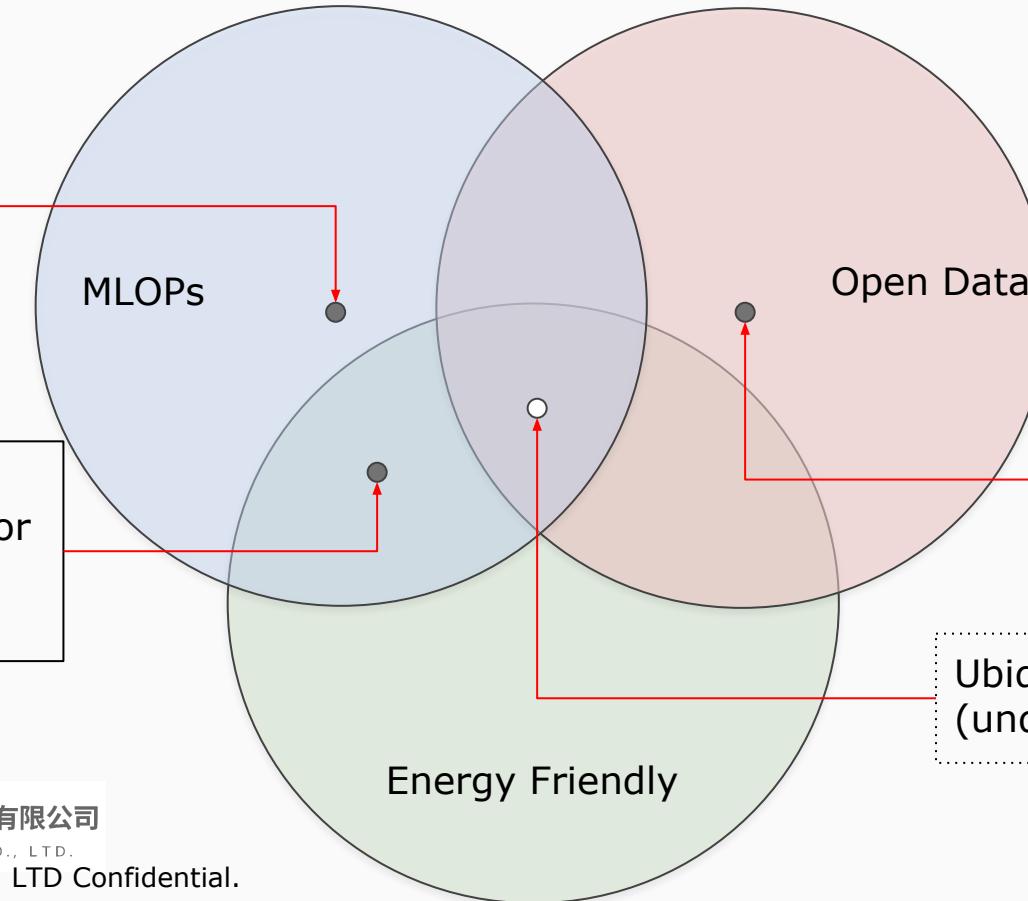
- Sustainability: Self-charging

- Energy-friendly: low power consumption
- Ubiquitous: device can be carried to anywhere, even network is not accessible.

# RoadMap

Tintin: Machine learning platform for everyone (2020-2021)

KaFeDo: Machine learning platform for green economy (2022)



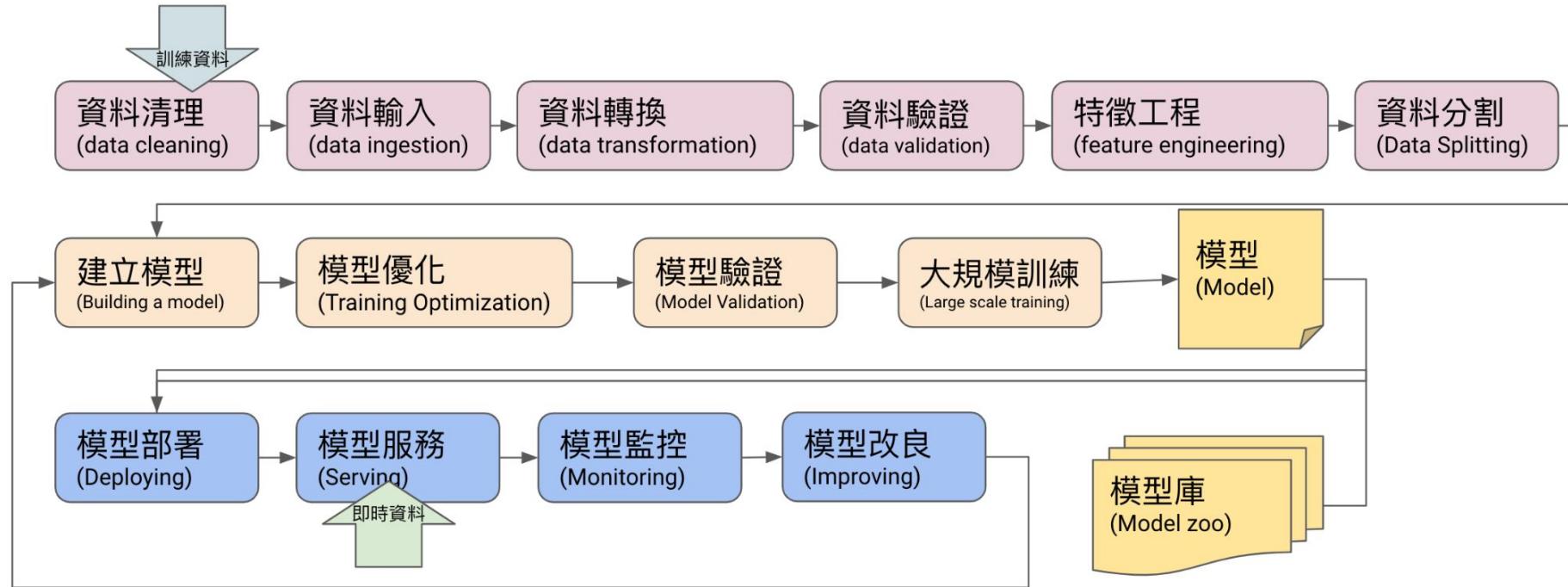
Shrimping: a data-sharing platform (2021)

Ubiquitous AI platform (under developing)

# Service Scope

- MLOps Professional Services
- MLOPs AI platform customization

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



# ChatGPT



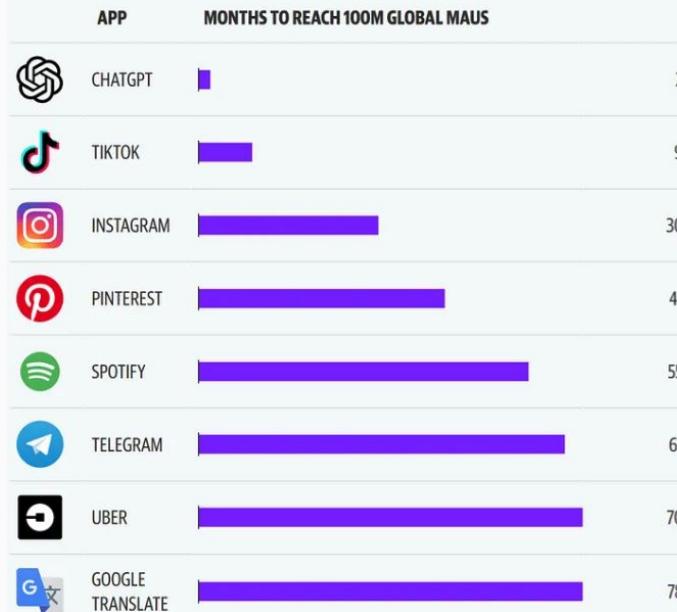
<https://chat.openai.com/chat>

# The golden age of NLP and **distributed** ML

## HOW LONG IT TOOK TOP APPS TO HIT 100M MONTHLY USERS

ChatGPT is estimated to have hit 100M users in January, 2 months after its launch.

Here's how long it took other top apps to reach that:



SOURCE: UBS



# What is standard language model?

Language models predict the next word(s) in a sequence

- Applications

- Machine Translation
- Text Classification
- News Article Generation
- Question Answering
- Auto Completion
- ...

$$P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1})$$

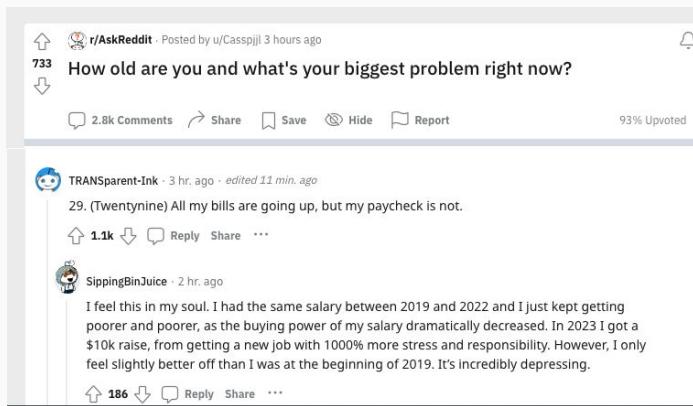
HS      The Highest mountain in \_\_\_\_?

 I would be happy to help you fill in the blank! Can you please tell me which region or country you would like me to find the highest mountain for?  

# What problem that ChatGPT wants to solve?

ChatGPT[1] is a Generative Pre-trained Transformers that generate human-like text from a large corpus of text data.

- Pre-training: Initial training process on a large text corpus where the model learns to predict the next word in a passage.
- Transformers [3]: attention mechanism to analyze input data, such as text, and extract important features or patterns



Q: How old are you and what's your biggest problem right now?

A: 29. (Twenty-nine) All my bills are going up, but my paycheck is not.

Web-text corpus

[1] [https://en.wikipedia.org/wiki/Generative\\_pre-trained\\_transformer](https://en.wikipedia.org/wiki/Generative_pre-trained_transformer)

[2] [https://www.reddit.com/r/AskReddit/comments/11jhj8o/how\\_old\\_are\\_you\\_and\\_whats\\_your\\_biggest\\_problem/](https://www.reddit.com/r/AskReddit/comments/11jhj8o/how_old_are_you_and_whats_your_biggest_problem/)

[3] <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>

### Self-supervision

1. The model's *only* objective is to learn co-occurrence patterns in the sequences it is trained on.
2. Alternatively: to assign high probability to attested sequences.
3. Generation then involves *sampling* from the model.
4. The sequences can contain anything.

# The evolution of GPT [1]

Name	Year	Data Size	Model Parameters
GPT	2018	4.5 GB of text, from 7000 unpublished books of various genres.	0.12B
GPT-2	2019	40 GB of text from 8M web pages	1.5 B
GPT-3	2020	45T data from <ul style="list-style-type: none"><li>• Common Crawl</li><li>• Web Text</li><li>• Books</li><li>• Wiki</li></ul>	175 B (i.e. 800GB storage)

# How ChatGPT is trained.

Step 1

Collect demonstration data and train a supervised policy.

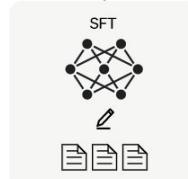
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

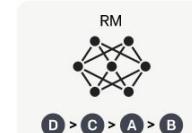
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



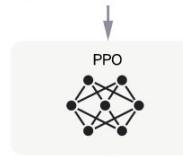
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



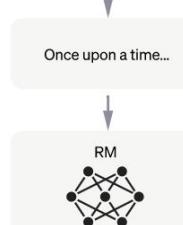
The PPO model is initialized from the supervised policy.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

$r_k$

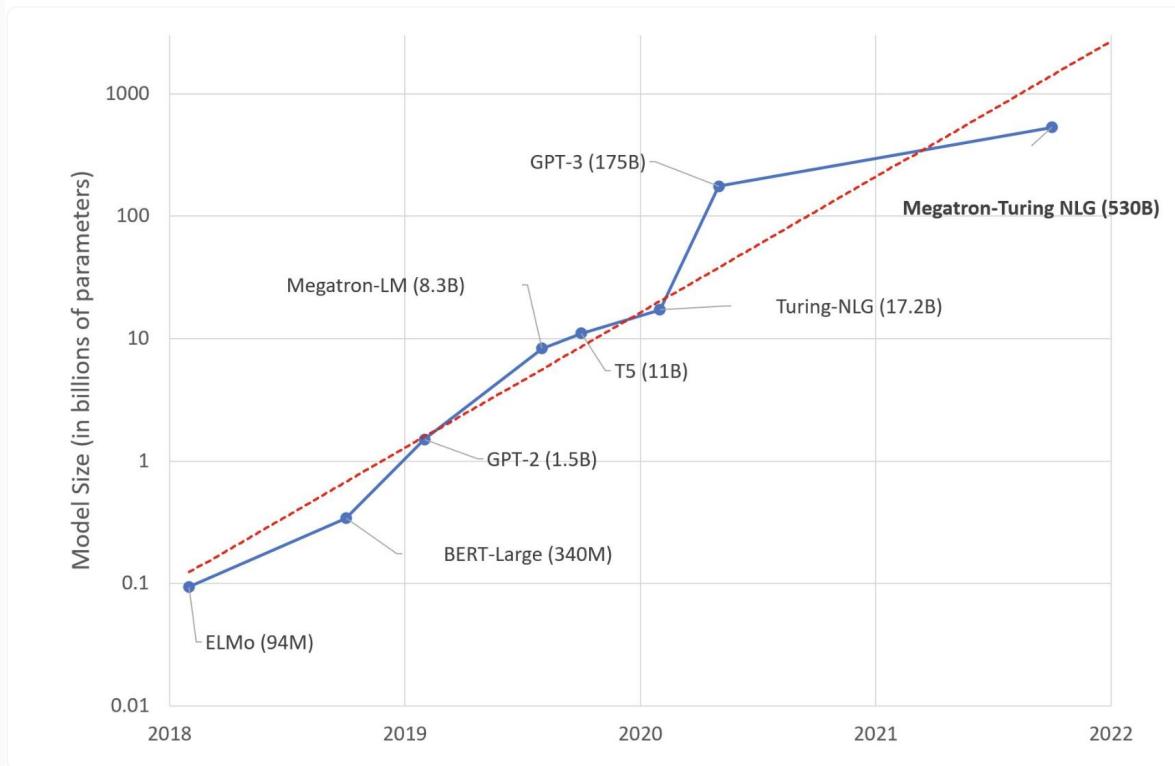
### (Recent) History: ChatGPT

(rumor) about 10x spend on human annotation budget

(rumor) modification of RLHF training

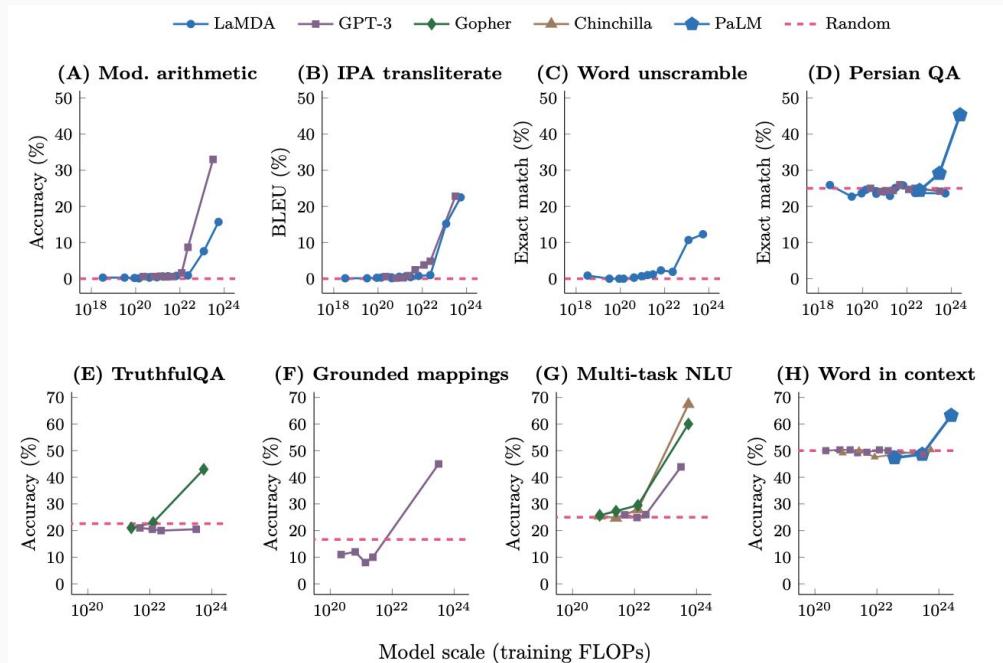
Huge impact!

# Large Language Models: A New Moore's Law?



# The trend of LLM: Emergent Ability

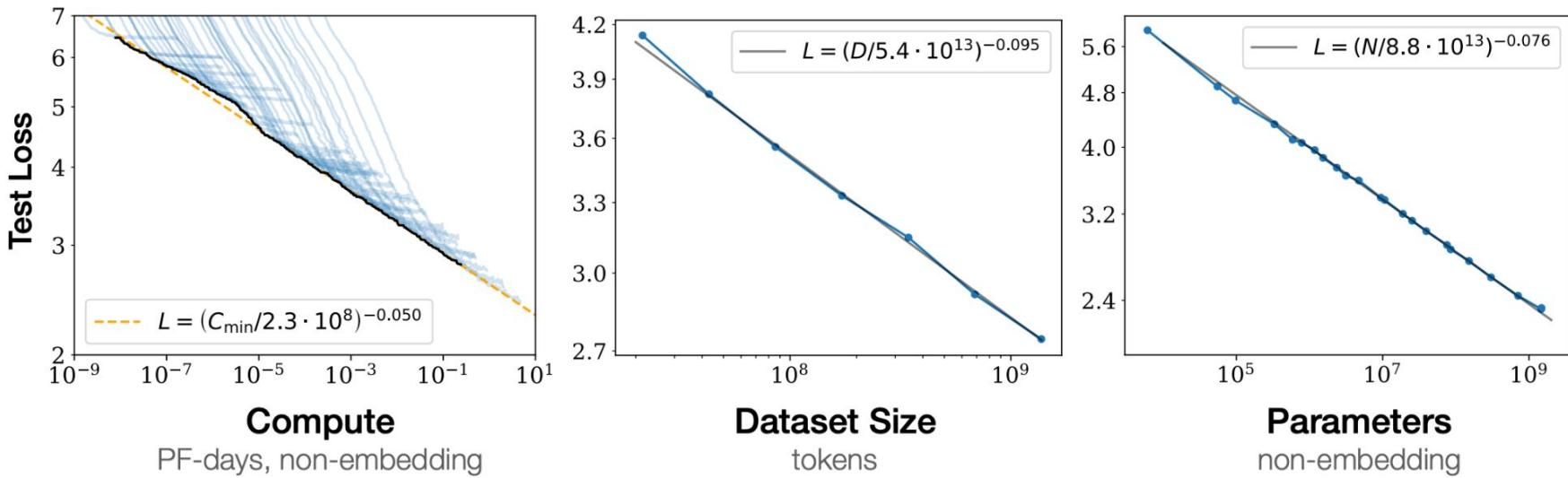
*An ability is emergent if it is not present in smaller models but is present in larger models.* - TMLR 2022



RTX 3090 is about  $35 * 10^{12}$  FLOPs.

For  $10^{24}$ , it needs  
~  $10^{19}$  FLOP per day  
And running for  $10^5$  days

# The trend of LLM: Scaling Law for Natural language model



## Standard supervision vs. in-context learning

Standard supervision for  
*nervous anticipation*

My palms started to sweat as the lotto numbers were read off.	<code>nervous anticipation = 1</code>
I took a deep breath as the curtain started to rise on my debut night.	<code>nervous anticipation = 1</code>
I couldn't shake a deep feeling of unease about the whole affair.	<code>nervous anticipation = 0</code>

“Few-shot in-context learning”

Hey model, here is an example of nervous anticipation: “My palms started to sweat as the lotto numbers were read off.”

Hey model, here’s an example without nervous anticipation: “...”

# In-context learning: use prompting, instead of fine-turning on each domain

## The three settings we explore for in-context learning

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



## Traditional fine-tuning (not used for GPT-3)

### Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



# An example on Few shots promoting

HS

please fill \_\_ with meaningful words. The Highest mountain in \_\_ is \_\_ .



Sure, here's an example sentence:



"The Highest mountain in South America is Aconcagua."

# In context learning: chain of thoughts

HS 10隻兔子有幾隻腳?

ChatGPT: 如果每只兔子都有4隻腳的話，那麼10隻兔子總共會有40隻腳。 ↑ ↓

HS 一隻兔子平均有兩隻腳，三隻兔子平均有兩隻腳。請問十隻兔子平均有幾隻腳?

ChatGPT: 這個問題有點棘手，因為兔子平均有兩隻腳和三隻兔子平均有兩隻腳這兩個條件似乎有些相互矛盾。  
如果我們假設每只兔子都有兩隻腳，那麼10隻兔子平均會有20隻腳。  
如果我們假設三隻兔子的總腳數是6，那麼一隻兔子的平均腳數就是2。因此，10隻兔子的總腳數將是10乘以2，等於20。  
但是，這個問題的問題陳述似乎暗示每只兔子的腳數可能會有所不同，因此我們無法確定答案。

HS Q: 一隻兔子平均有兩隻腳 A: 一隻兔子有兩隻腳，所以平均是2/1等於兩隻腳。Q: 三隻兔子平均有兩隻腳 A: 三隻兔子有六隻腳，所以平均是6/3等於兩隻腳。請問十隻兔子平均有幾隻腳?

ChatGPT: 根據這些條件，我們可以計算出每隻兔子平均有兩隻腳，無論是單獨的一隻兔子還是三隻兔子一起算。因此，根據平均數的性質，10隻兔子平均也會有兩隻腳。所以答案是2隻腳。 ↑ ↓

Teach ChatGPT how to think by prompting

# What is Cloud-native infrastructure?



*Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify this approach.*

*- Cloud Native Foundation.*

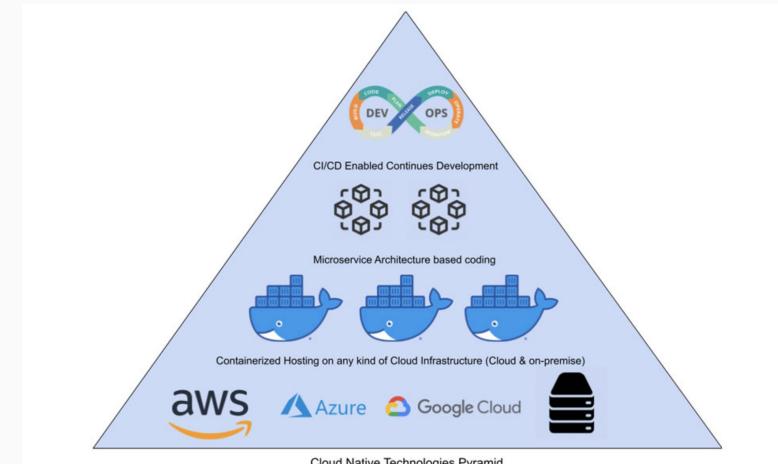
# Cloud Native Pillars

## Key Components include

- Container and its orchestration
- Serverless architecture
- Continuous integration and Continuous Deployment (CI/CD)
- ...

## Key Benefits:

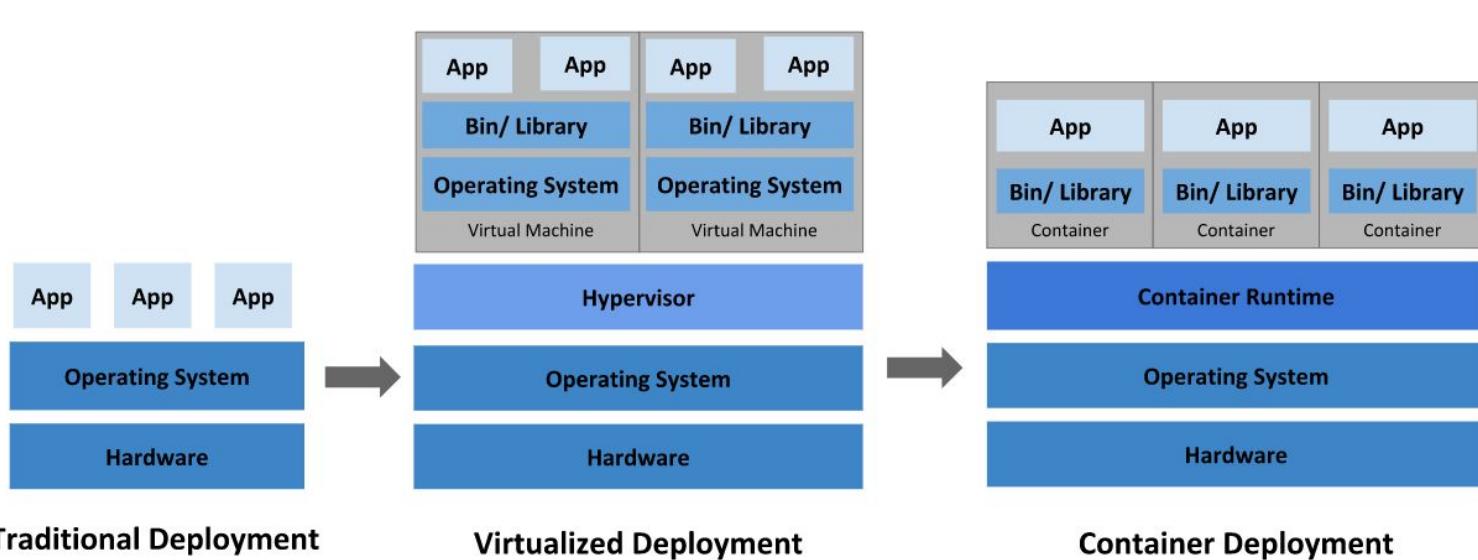
- Scalability
- Portability
- Cost-effective



# What is container?

- Container

- Container Image = Application code + dependencies
- Runtime environment (cgroups, namespaces, env vars)



# 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 v1.26, released at 2023.

[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>



# Kubernetes Distributions Evolution

To Cloud



To Edge



lightweight

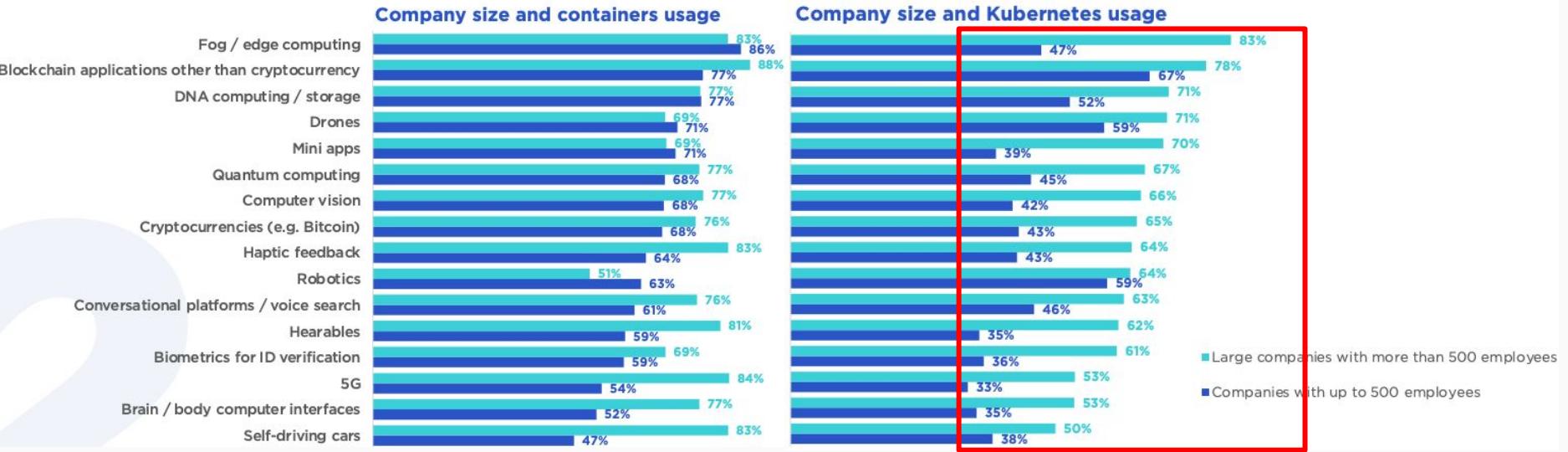


# What is Kubernetes adoption rate so far?

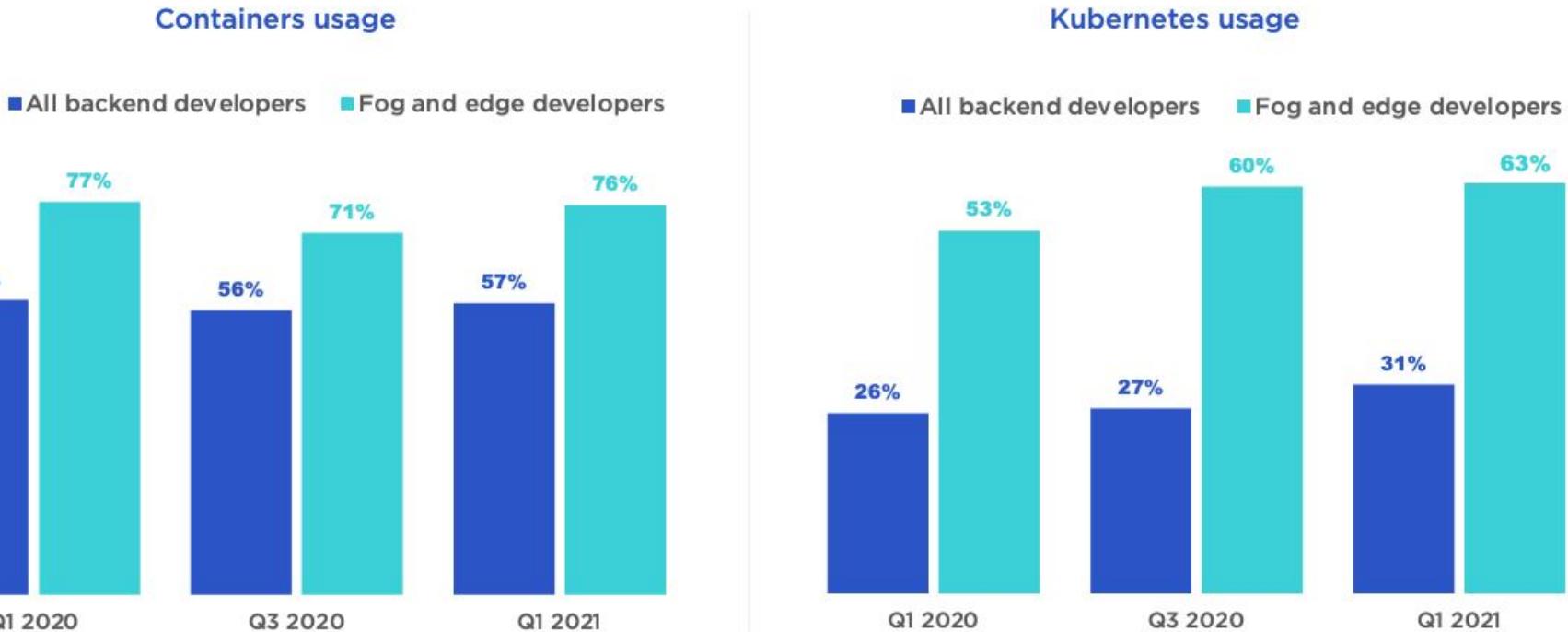
Usage of cloud native technologies across regions



# Container adoption rate vs Kubernetes among company size



# Container and Kubernetes adoption rate on edge computing



# Relevant Jobs In Taiwan

Setting Popular Recent

 Senior Site Reliability Engineer  
MaiCoin

Full time · Mid-Senior level  
📍 110 台灣台北市信義區 ▲ 2  
\$ 1.2M ~ 2.5M TWD/year

🕒 Updated 3 Months ago · 500+ · Unread

Save Follow

 Machine Learning Engineer (Ads) 🔥  
Dcard 狀卡科技股份有限公司

Full time · Entry level  
📍 台北 · 大安區  
\$ 900K ~ 2.2M TWD/year

🕒 Updated 18 days ago · 1,000+ · Unread

Save Follow

 (Sr.) DevOps Engineer 運維開發工程師  
OpenNet 開網有限公司

Full time · Mid-Senior level  
📍 台灣台北 ▲ 2  
\$ 800K ~ 1.8M TWD/year

🕒 Updated a month ago · 300+ · Unread

Save Follow

 【技術部】資深DevOps/SRE工程師  
聖霏有限公司

Full time · Mid-Senior level  
📍 大安區 · 台北  
\$ 40K ~ 100K TWD/month

🕒 Updated 3 Months ago · 50+ · Unread

Save Follow

 LINE TV 數據工程師 (Data Engineer) 🔥  
LINE TV\_巧克科技新媒體股份有限公司

Full time · Entry level  
📍 台灣 · 台北  
\$ 50K ~ 80K TWD/month

🕒 Updated 6 days ago · 1,000+ · Unread

Save Follow

 中階/資深後端工程師 Mid/Senior Backend Engineer (Python) 🔥  
Linker Networks Inc. 美商實益凌科網路科技有限公司台灣分公司

Full time · Mid-Senior level  
📍 Taipei · Kaohsiung ▲ 2  
\$ 50K ~ 150K TWD/month

🕒 Updated 2 months ago · 500+ · Unread

Save Follow

 Junior DevOps/SRE  
Splashtop Inc.

Full time · Entry level  
📍 105 台灣台北市松山區 ▲ 2  
\$ 650K ~ 1M TWD/year

🕒 Updated 4 days ago · Unread

Save Follow

 Senior DevOps / SRE  
Splashtop Inc.

Full time · Mid-Senior level  
📍 105 台灣台北市松山區 ▲ 2  
\$ 1M ~ 1.5M TWD/year

🕒 Updated 4 days ago · 50+ · Unread

Save Follow

 Site Reliability Engineer  
OneDegree

Full time · Mid-Senior level  
📍 台灣台北  
\$ 840K ~ 1.5M TWD/year

🕒 Updated 4 days ago · 100+ · Unread

Save Follow

 Machine Learning Engineer 🔥  
Dcard 狀卡科技股份有限公司

Full time · Mid-Senior level  
📍 台北 · 大安區  
\$ 900K ~ 2.2M TWD/year

🕒 Updated 18 days ago · 1,000+ · Unread

Save Follow

 Machine Learning Engineer (Search)  
Dcard 狀卡科技股份有限公司

Full time · Mid-Senior level  
📍 台灣台北市大安區  
\$ 900K ~ 2.2M TWD/year

🕒 Updated a month ago · 50+ · Unread

Save Follow

 Mid-Level Software Engineer, Cymetrics (Backend)  
OneDegree

Full time · Mid-Senior level  
📍 台北 · 信義區  
\$ 840K ~ 1.2M TWD/year

🕒 Updated 4 days ago · 100+ · Unread

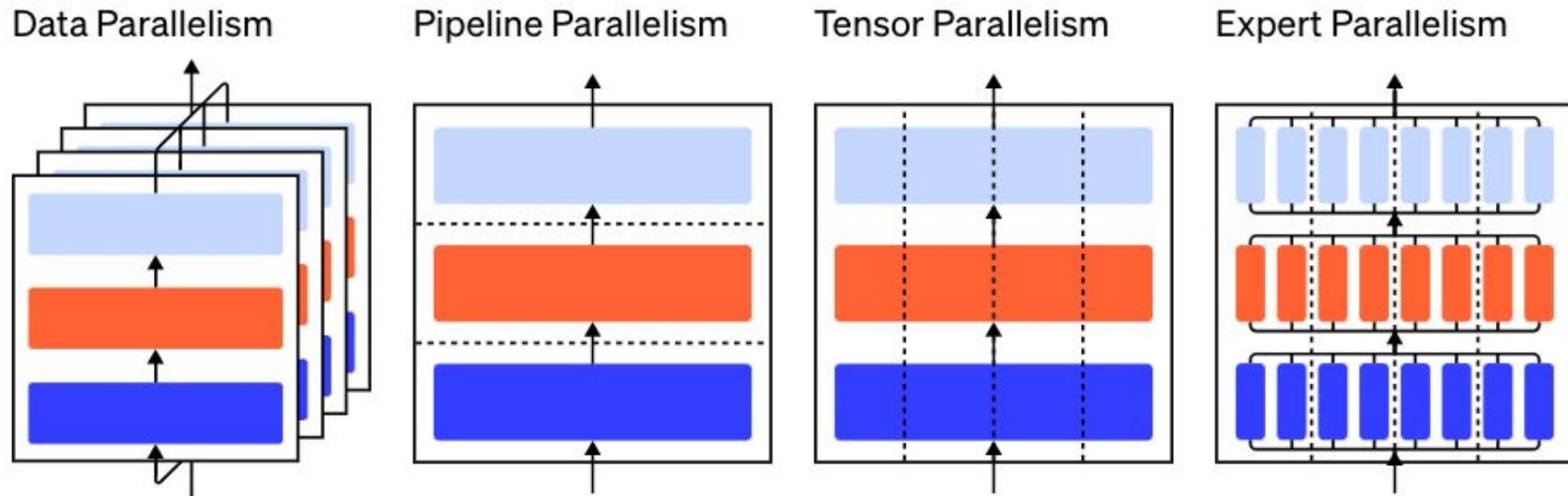
Save Follow

 【資訊研發部】PHP 後端資深工程師 (可遠端)  
聯合數位資通

Full time · Mid-Senior level  
📍 台灣台北 ▲ 5  
\$ 60K ~ 80K TWD/month

How large scale  
model is trained?

# Techniques for training large neural networks (1/2)



An illustration of various parallelism strategies on a three-layer model. Each color refers to one layer and dashed lines separate different GPUs.

## Other strategies:

- Mixed Precision Training
  - Use lower-precision numbers (most commonly FP16) during training phase.
- Data Offloading
  - Offloading data from GPU to CPU or amount different devices, and load it back when needed.
- Increase Memory Efficiency
  - Minimizing memory footprint

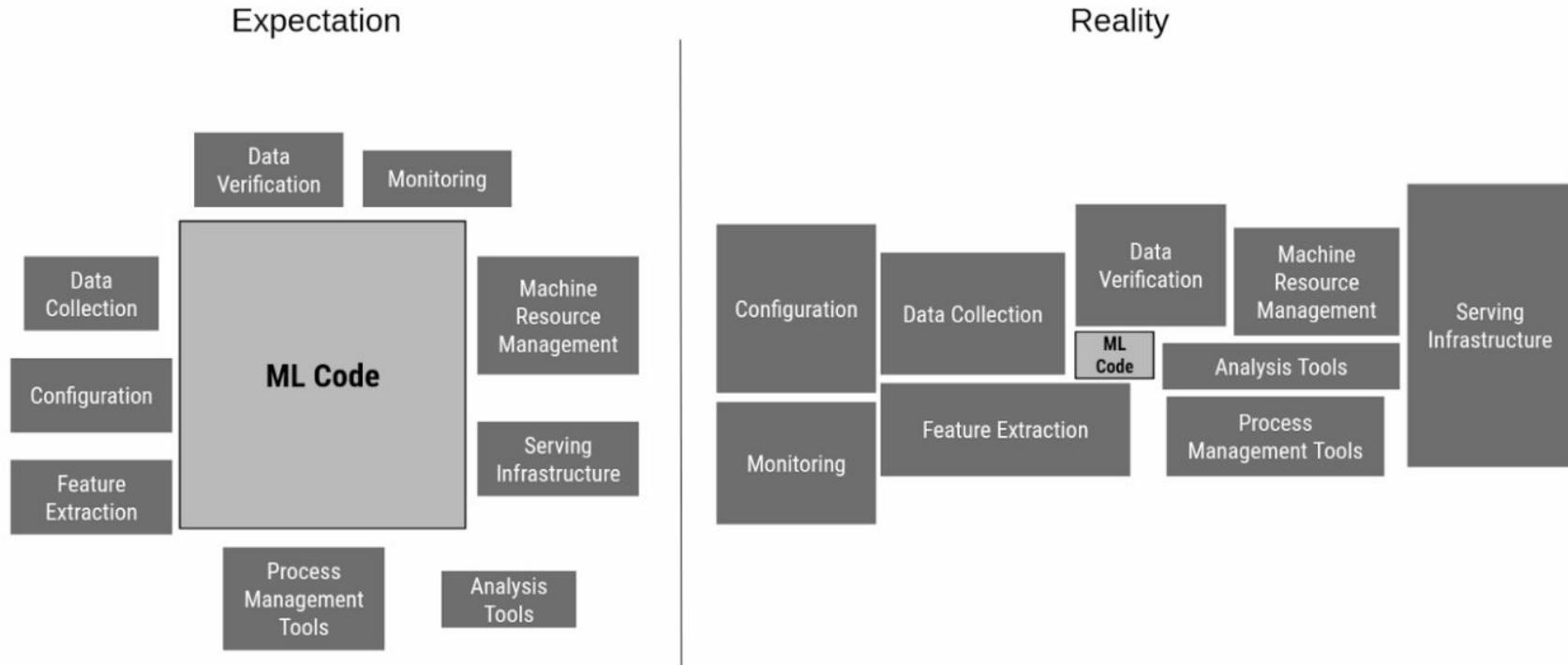
...

# Trade Off between budget, deadline and tech feasibility



# Real-world Machine Learning Application

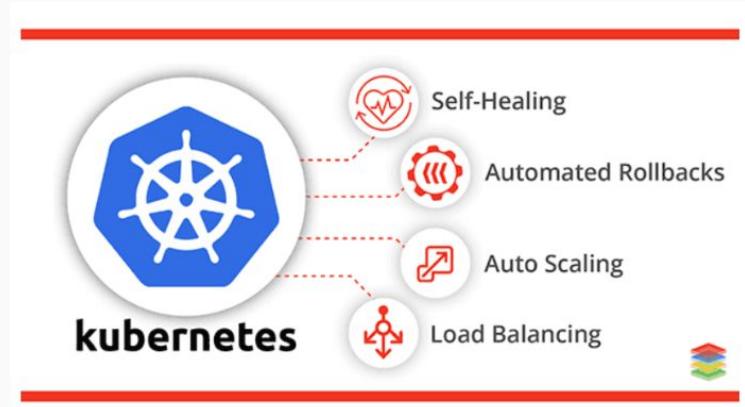
## Expectation vs Reality



Source: <https://medium.com/ci-t/e2e-kubeflow-pipeline-for-time-series-forecast-239019ad2e34>

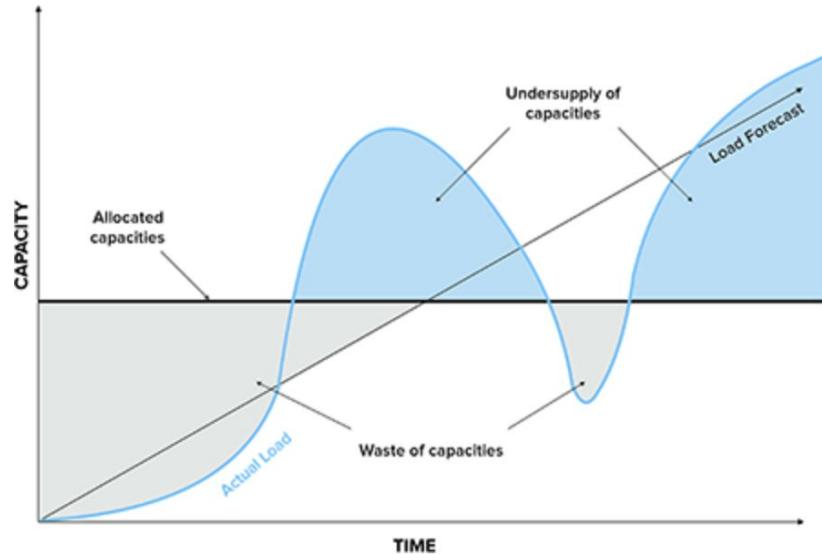
# Why machine learning on Kubernetes?

- Portability
  - Dev/Staging/Prod
  - Laptop/Edge/Cloud environment
- Scalability
  - Hyperparameter tuning, production workloads
- Isolation
  - Workloads from different experiments



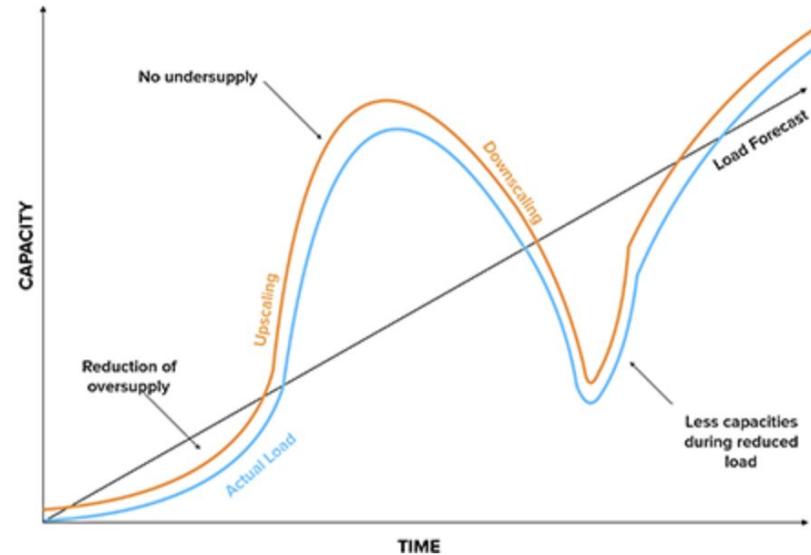
# Cost, Cost, and Cost

Static Architecture



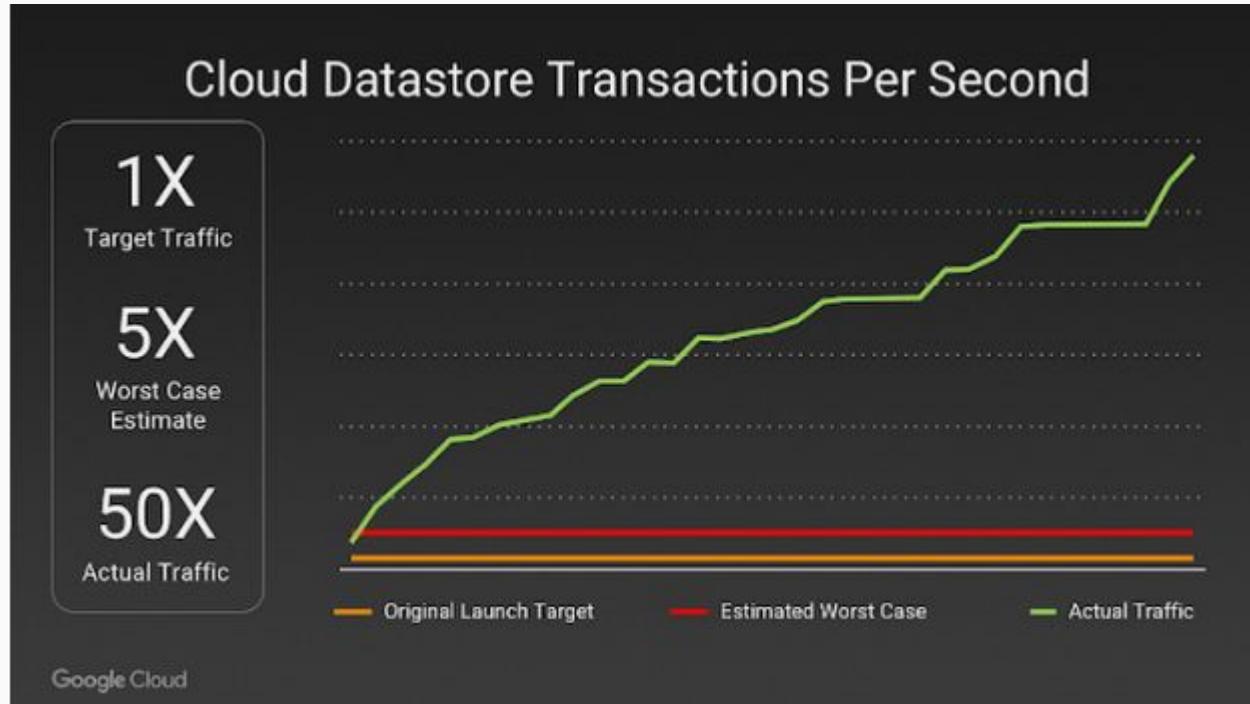
Static architectures are based on estimated load expectancy and are not flexible enough to adapt to unexpected load peaks or lulls.

Auto Scaling Architecture



Auto Scaling is the most cost-efficient solution for a fluctuating load. High performance and thereby user satisfaction are retained at all times.

# Case Study: Pokémon GO



<https://cloud.google.com/blog/products/containers-kubernetes/bringing-pokemon-go-to-life-on-google-cloud>  
<https://medium.com/nerd-for-tech/catch-em-all-with-kubernetes-pokemon-go-case-study-37a43c5dce1b>

# Case Study: Open AI

OpenAI adopted Kubernetes since 2016 for portability, cost saving, and improved efficiency[1,2].

Years	Nodes	Estimated Cost [3]
2018	2,500	= 3 * 2500 * 24 = US\$ 180,000 / day
2021	7,500	= 3 * 2500 * 24 = US\$ 540,000 / day
2023	?	

[1] <https://kubernetes.io/case-studies/openai/>  
<https://blog.pichuang.com.tw/20230214-openai-scaling-kubernetes-to-7500-nodes.html>

[2] [https://blog.pichuang.com.tw/20230214-openai-scaling-kubernetes-to-7500-nodes.html#\\_1](https://blog.pichuang.com.tw/20230214-openai-scaling-kubernetes-to-7500-nodes.html#_1)

[3] AWS P3.2xlarge equips 8 vCores, 61 GB Memory, 1 Core of V100-16Gb GPU, charges US\$ 3 per hour.

**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
- ...

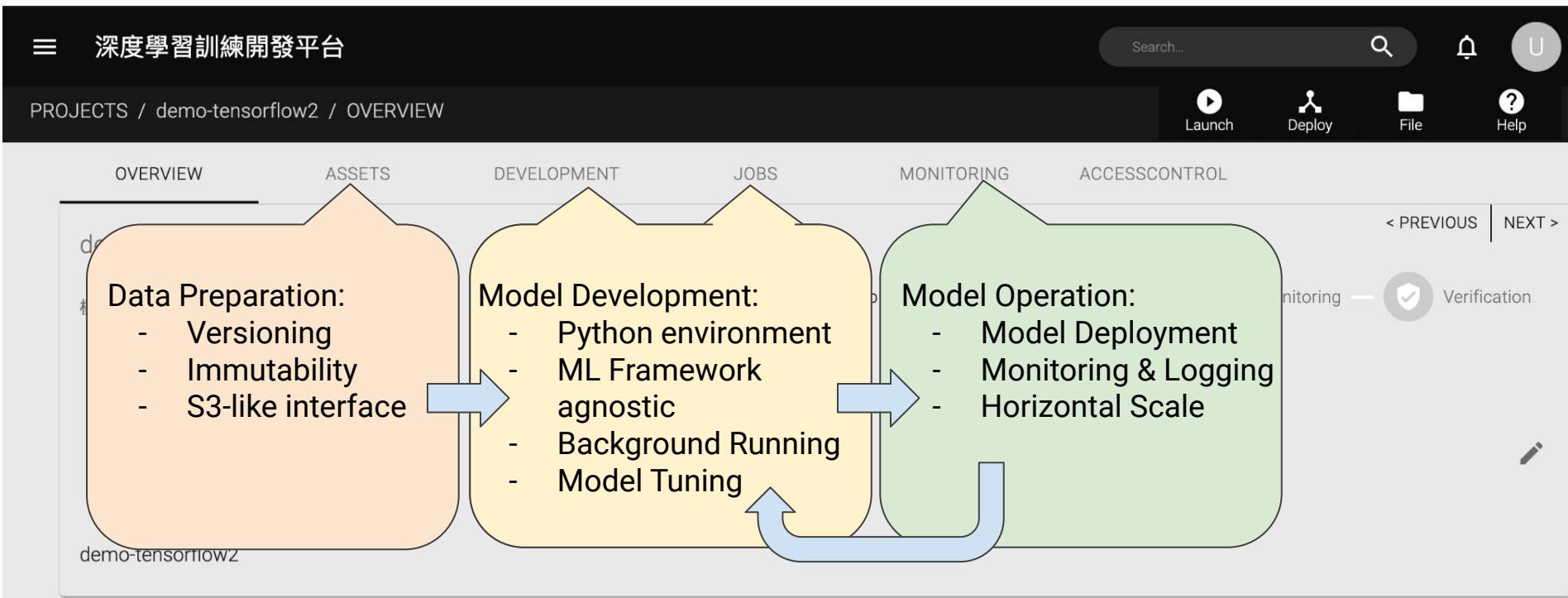


What is Tintin?

# Tintin: Machine learning platform for Everyone

<https://get-tintin.footprint-ai.com>

On top of Kubernetes, we built a machine learning development platform for easing the gap in using Kubernetes in scale.



# Tintin: Machine learning platform for everyone (2/5)

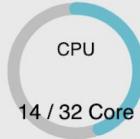


PROJECTS / OVERVIEW

歡迎您!

深度學習訓練開發平台!

## User Resource Overview



## Projects

New Project



英文語音識別

by [user2@footprint-ai.com](mailto:user2@footprint-ai.com)

最後編輯時間2021/1/26 14:59:31



Cifar-10 影像辨識

by [user2@footprint-ai.com](mailto:user2@footprint-ai.com)

最後編輯時間2021/1/18 15:52:02



# Tintin: Machine learning platform for everyone (3/5)



三 深度學習訓練開發平台

PROJECTS / Cifar-10 影像辨識 / JOBS

Search...



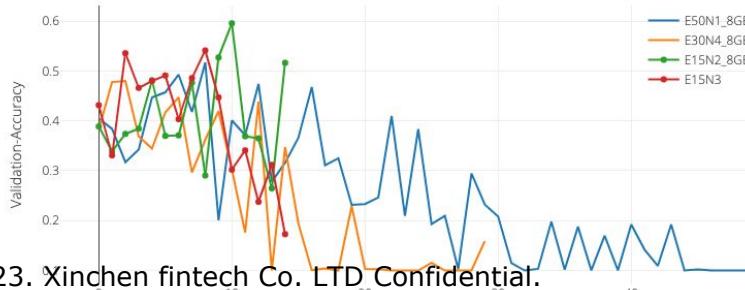
4 selected

Name	Job Status	Onlined	Duration	ScheduleAt	StartAt	EndAt	Metrics
E50N1_8GB	✓	🕒	1:43:30	-	2021/1/13 24:36:39	2021/1/13 02:20:09	▼
E30N4_8GB	✓	🕒	0:43:36	-	2021/1/13 24:26:46	2021/1/13 01:10:22	▼
E15N2_8GB	✓	🕒	0:43:23	-	2021/1/12 23:03:29	2021/1/12 23:46:52	▼
E15N3	✓	🕒	0:26:24	-	2021/1/12 22:56:24	2021/1/12 23:22:48	▼

Rows per page: 5 ▾ 1-4 of 4 < >

Comparison

Validation-Accuracy Comparison



Training-Accuracy Comparison



# Tintin: Machine learning platform for everyone (4/5)



Default Run at Wed Dec 09 2020 14:42:37 GMT+0800 (Taipei Standard Time)

Please use the following command to access inference server, or use customized UI below.

CURL:

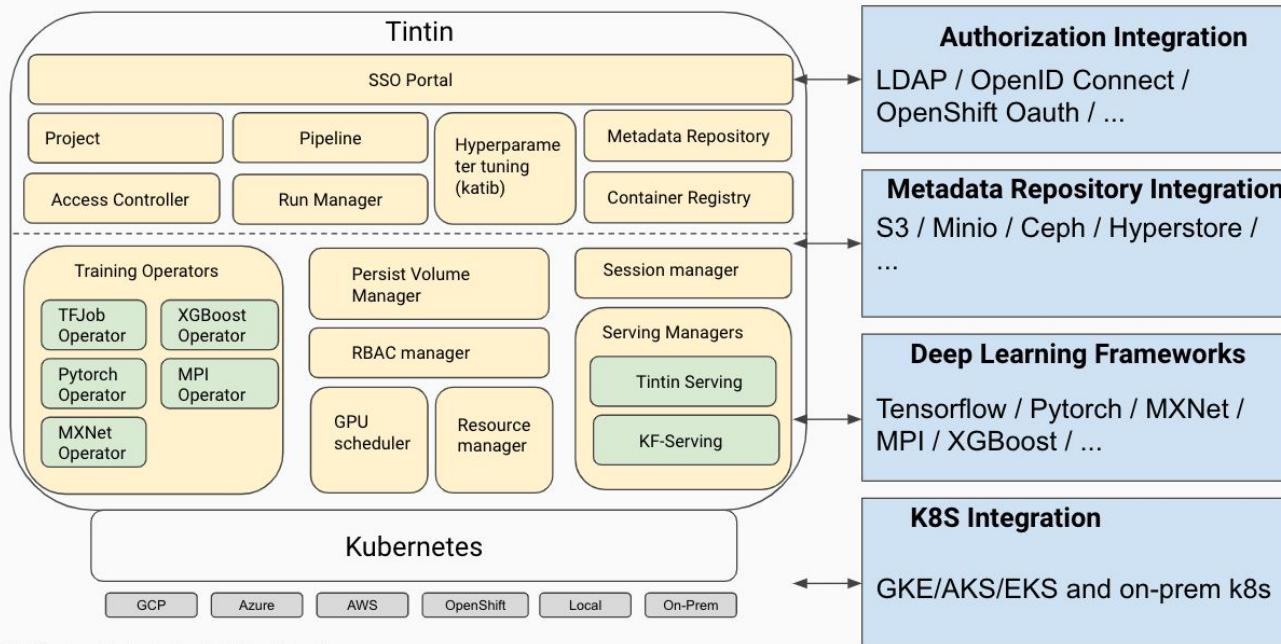
```
curl -X POST \
  https://api.wmlk.footprint-ai.com/api/v1/project/ypw17y259vk13jvwm6onzexrd048g3/onlineserving/run/a2b1b741-e2db-4f2b-8840-c3c69ca4
  -H "Content-Type: multipart/form-data" \
  -H "Authorization: Bearer
e3b0c44298fc1c67d1d01a0a3b1e139 \
  -F "file=@photo.png" -k
```

UI:



```
{
  "data": {
    "data": {
      "img_path": "/tmp/2973_airplane.png",
      "prediction": "horse",
      "scores": [
        {
          "label": "horse",
          "score": 0.999
        }
      ]
    }
  }
}
```

Empower EVERYONE to build and deploy model easily with Kubeflow and Kubernetes.

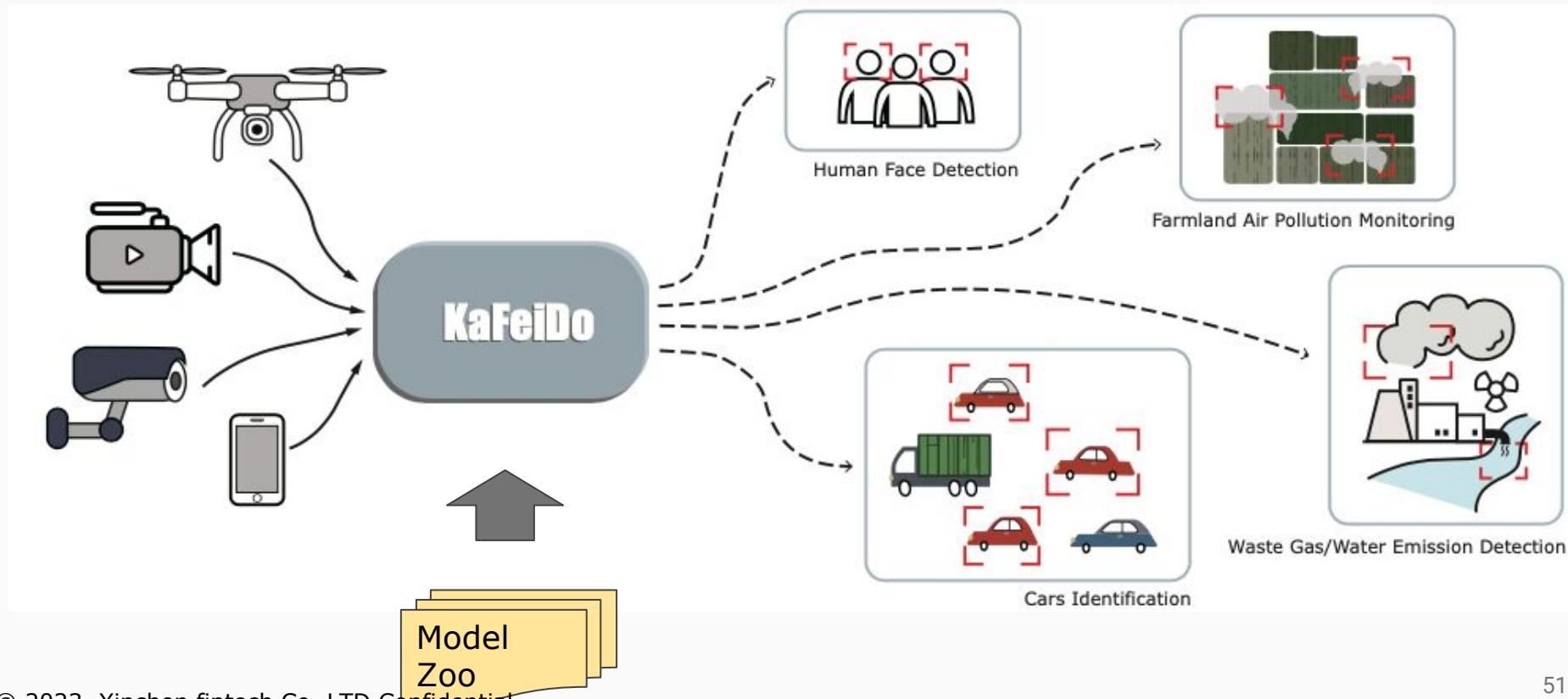


What is KaFeiDo?

# KaFeiDo: A Machine Learning Platform Built for Green Economy

KaFeiDo is a machine learning platform aiming at saving costs on hardware and energy while providing automation for machine learning models.

# Kafeido: instant model serving framework for green economy. (1/2)



# Kafeido: instant model serving framework for green economy. (2/2)

## How does **KaFeiDo** work?

Applicable places: community, school, hospital, shopping mall, factory, corporate  
Deployment plan: On-prem, SaaS

### Step 01

Select a model from our model zoo or upload your own model

### Step 02

Generate the model deployment

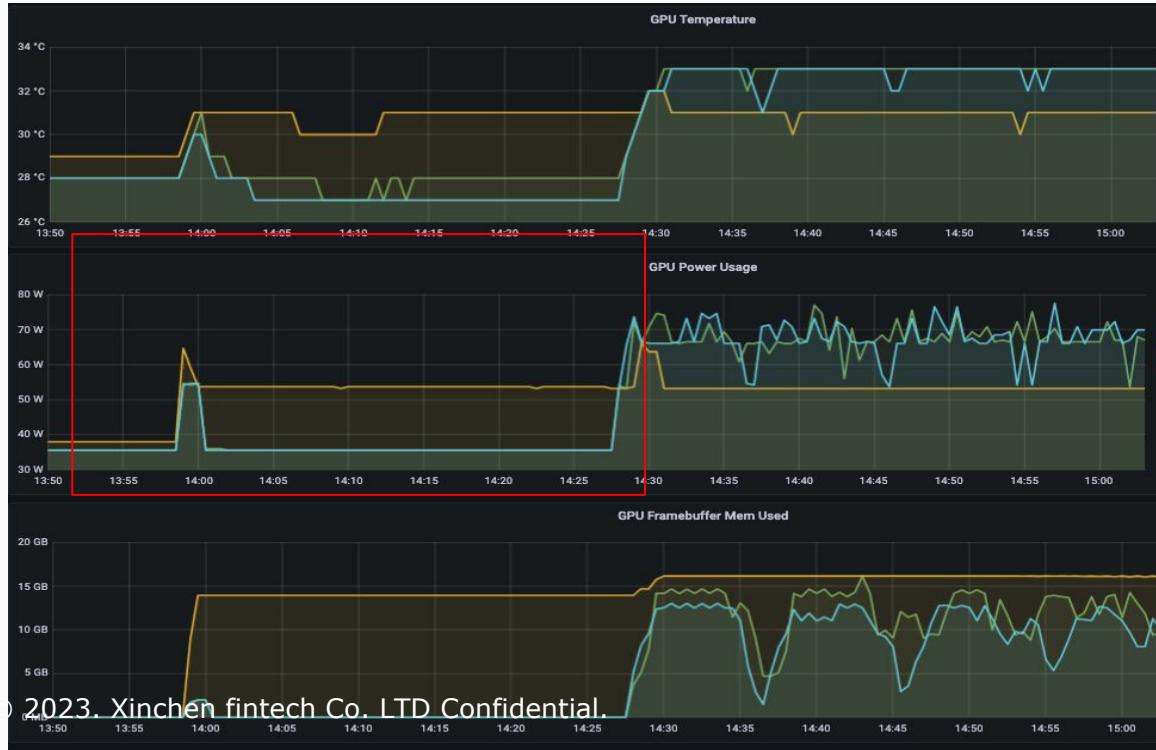
### Step 03

Select your data source

### Step 04

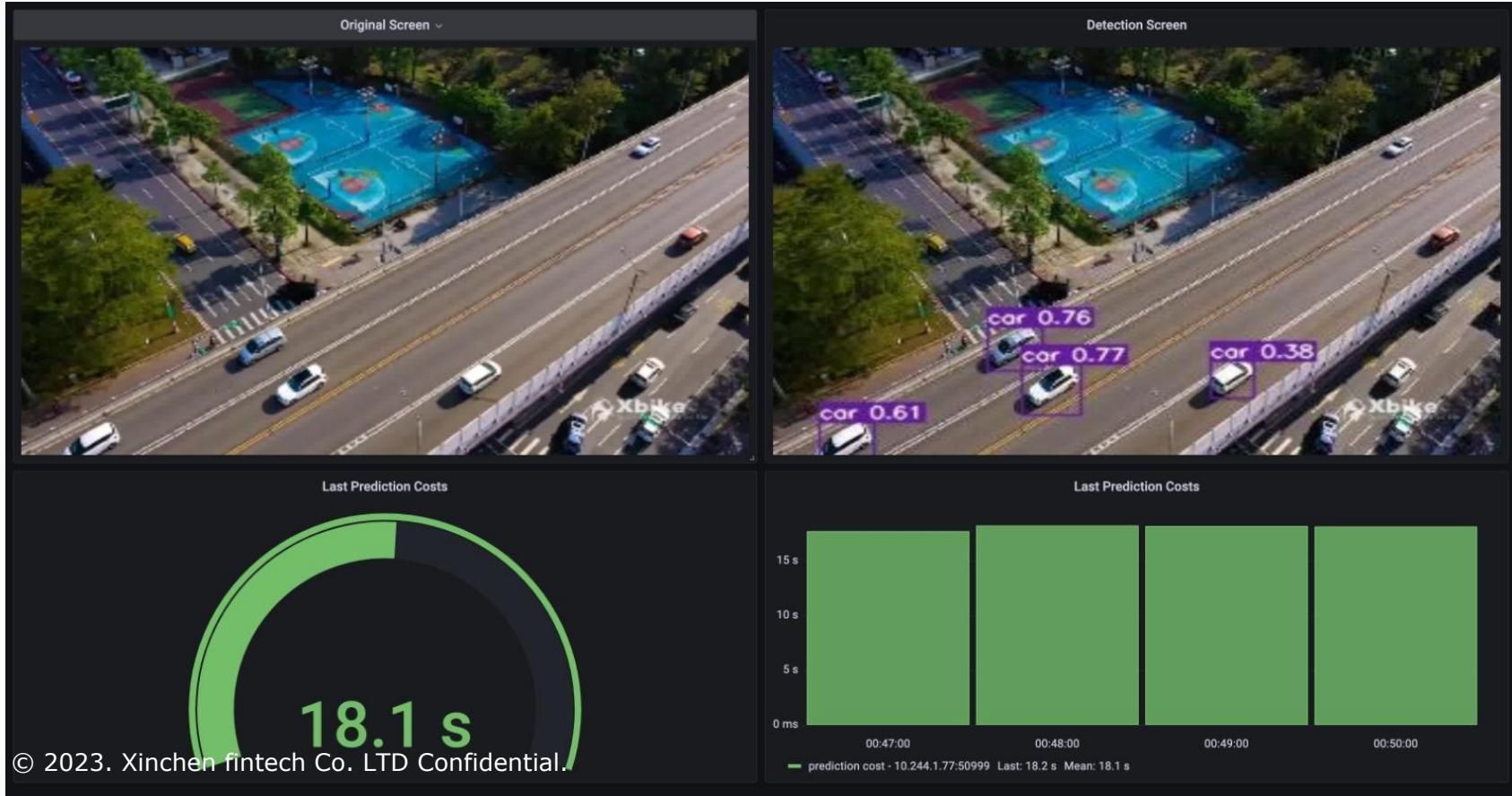
Automatic management model inference

# How we save electricity cost?



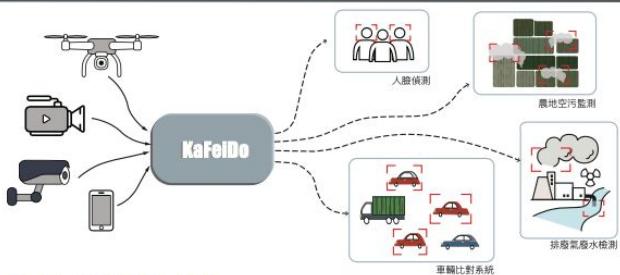
- MMA + serverless + autoscaling
- Monolithic + serverless + autoscaling
- Monolithic

# Inference Screenshot



## KaFeiDo：智慧節能即時多模型同步推論引擎

我們的一站式平台讓客戶可以選定既有的模型或自行上傳預先訓練的模型進行即時部署。兼具無運算架構與獨家微模型服務節能部署方案，不僅能替客戶省下在模型維運上的人事費用，更可以省下多餘的硬體與電力成本。



### Features highlighted

#### POINT 01

#### KaFeiDo 是您的模型部署好夥伴

KaFeiDo提供模型部署流水線模版將選定的模型數秒之內將模型服務化。由於模型服務化後與後續維護議題是軟體工程的問題而非資料科學家的專業範疇，KaFeiDo將整個過程專業且簡化，透過最佳的實作實例，讓資料科學家可以更專注在模型開發上，以提升客戶的核心價值。

#### POINT 02

#### 異質性多模型同步即時推論

KaFeiDo支援 Triton/Tensorflow/Pytorch 等主流推論框架，讓資料科學家可以使用自己熟悉的框架進行模型開發與部署，讓開發環境不再成為阻礙。

#### POINT 03

#### 無運算架構與水平擴展優勢

KaFeiDo提供無運算架構(Serverless architecture)與水平擴展模組，讓模型服務化不僅能以更低的成本運行，並隨時依據尖峰請求而提高服務量能。由於商品化機器學習模型的趨勢到來，模型數量增長的速度會遠大於硬體增長的速度，KaFeiDo的自動化與需求導向的硬體資源與模型管理機制能讓模型服務依據其請求量能提供適當的計算資源，並在模型服務閒置時將計算資源最小化。

#### POINT 04

#### 微模型服務節能部署架構

KaFeiDo獨家技術微模型服務架構(Micro-model architecture)，有別於傳統的集成式模型推論架構(Monolithic architecture)，微模型服務架構更能減少所需的硬體規格，還能有效降低電力成本。

## KaFeiDo如何運作

適用場景: 社區/學校/醫院/商場/工廠/企業場所  
部署方案: 落地部署(On-prem) / 虛擬服務(SaaS)



## KaFeiDo客戶案例

### 永續智慧城市監控中心

藉由匯集多個資料流與多種偵測模型進入即時推論框架，KaFeiDo依據其推論結果觸發警事件通知相關人員，來達成隨時(24/7/365)隨地(簡訊/電子郵件)的分散式監控模式。隨著永續環境概念意識逐漸抬頭，環境監控(如空汙監控、工廠排放廢氣等)更顯得其重要，而如何將智慧監控導入智慧城市變得是一個極嚴峻的問題。傳統的監控中心主要將各個(如攝影機)資訊匯集至單一控制台以便保全人員監控以及當事件發生時提供適當的協助，但隨著監控範圍逐漸拉大(如從閉路電視攝影機到無人機拍攝，從單點監控到場域監控等)，長期依賴保全人員的監控方式除了日益劇增的人力成本以外，也無法長期維持高標準監控。



專業的軟體架構與智慧的節能方案讓您的事業在導入AI上不僅容易且更負擔得起!

## 聯絡我們

本公司專營機器與深度學習平台 / 網路資料中台 / 多資料模型推論架構 / 客製化模型等服務。  
地址 : 103台北市大同區承德路三段287-2號 email : kafeido@footprint-ai.com



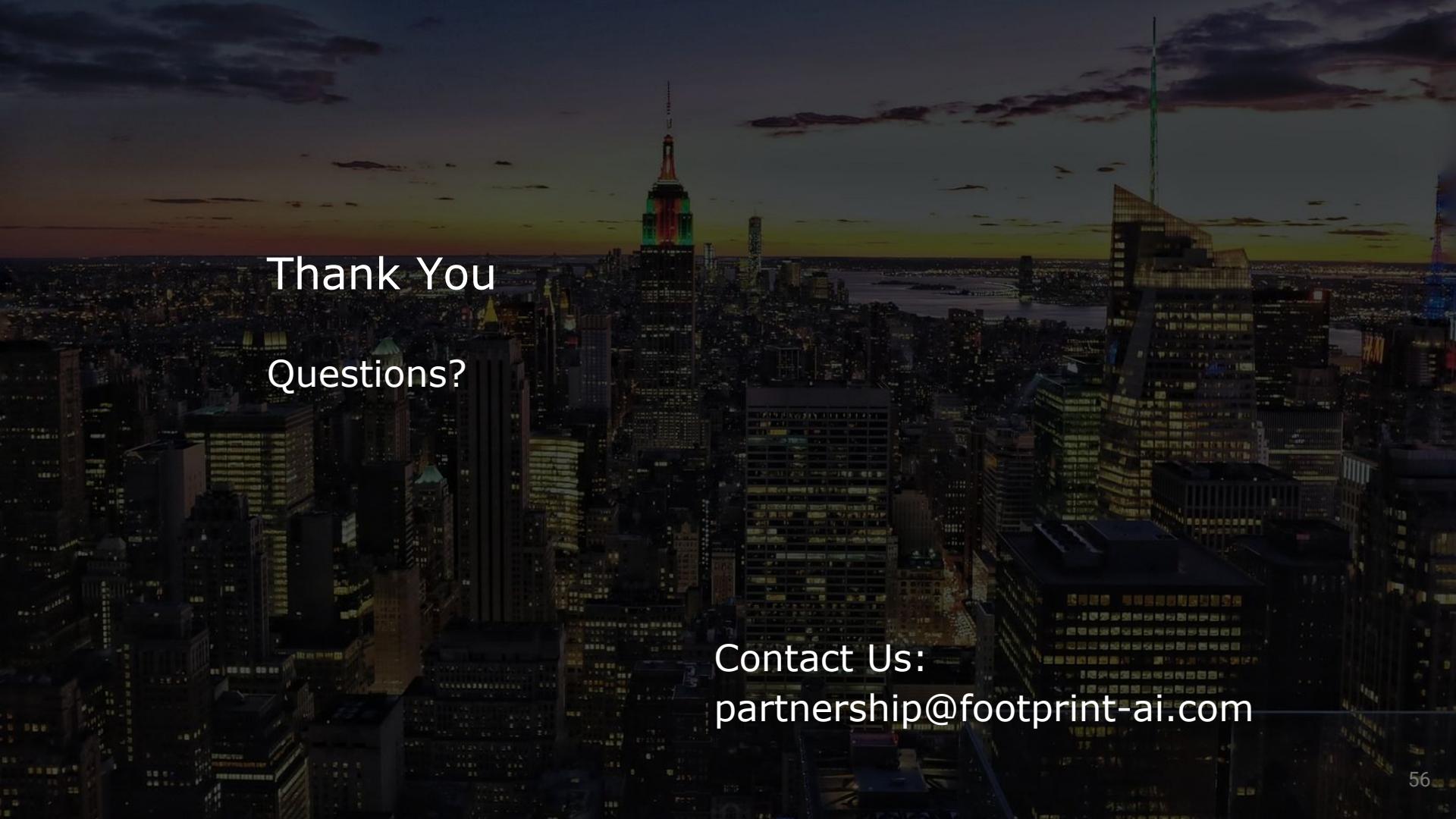
Tintin  
Machine Learning Platform for Everyone



NVIDIA.  
INCEPTION PROGRAM



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

The background image shows a panoramic view of a city skyline at dusk or night. The sky is filled with warm, orange and yellow hues from the setting sun. In the foreground, numerous skyscrapers are visible, their windows glowing with various shades of blue, green, and yellow. The Empire State Building stands prominently in the center-left, its Art Deco spire reaching towards the top of the frame. To its right, other recognizable buildings like the One World Trade Center are visible. A body of water is seen in the distance, reflecting the city lights.

Thank You  
Questions?

Contact Us:  
[partnership@footprint-ai.com](mailto:partnership@footprint-ai.com)



***“The Best Engineers  
Are Lazy”***

-Ancient Engineering Proverb

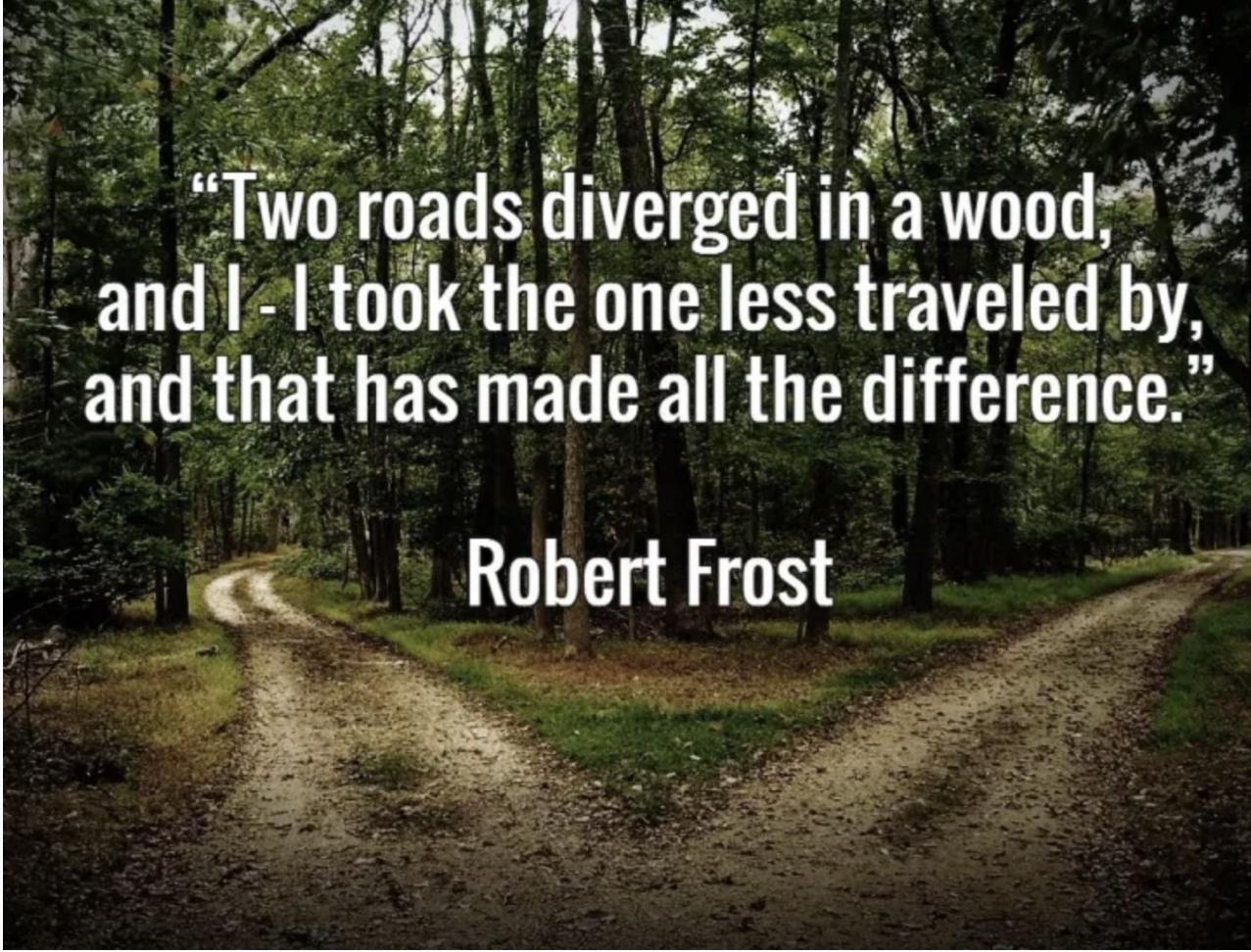
# What I have learned/gained during this endless journey (1/2)

- Some reflections on my career journey so far, including the challenges I've faced and the lessons I've learned along the way.
  - Life (or career) is an investment.
    - The first principle is never to lose your capital (time).
    - Always trade off between RISK and RETURN.
  - How to choose your first job?
    - High-salary job vs job with interest?
    - A big company vs a startup?
    - Local job vs oversea job?
    - How to access the job's risk & return?
    - Ephemeral vs Enduring?
- Define your career plan as early as possible
  - Define your ultimate destination could give you a clear picture of what you should do right now.



## What I have learned/gained during this endless journey (2/2)

- **Be proactive**
  - Would you ride a bike when you know how to ride it?
- **Be kind**
  - Friends are far better than enemies
- **Be globally**
  - English, english, english.
- **Be greedy**
  - Greedy for knowledge and anything that makes your feels rich.
- **Be responsible**
  - Every job is a self portrait of those who did it, Autograph your work with quality.

A photograph of a dirt road in a forest. The road is made of dirt and has some grass and fallen leaves on either side. It branches into two paths as it goes deeper into the woods. The background is filled with tall trees.

**“Two roads diverged in a wood,  
and I - I took the one less traveled by,  
and that has made all the difference.”**

**Robert Frost**