

End-to-end Industrial Practice on Data Engineering and Machine Learning

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

Download Slides

<https://reurl.cc/ymRMx0>



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



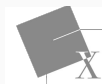
Agenda

- Who we are and what we do.
- Challenges from End-to-end data engineering and machine learning.
- FAQ

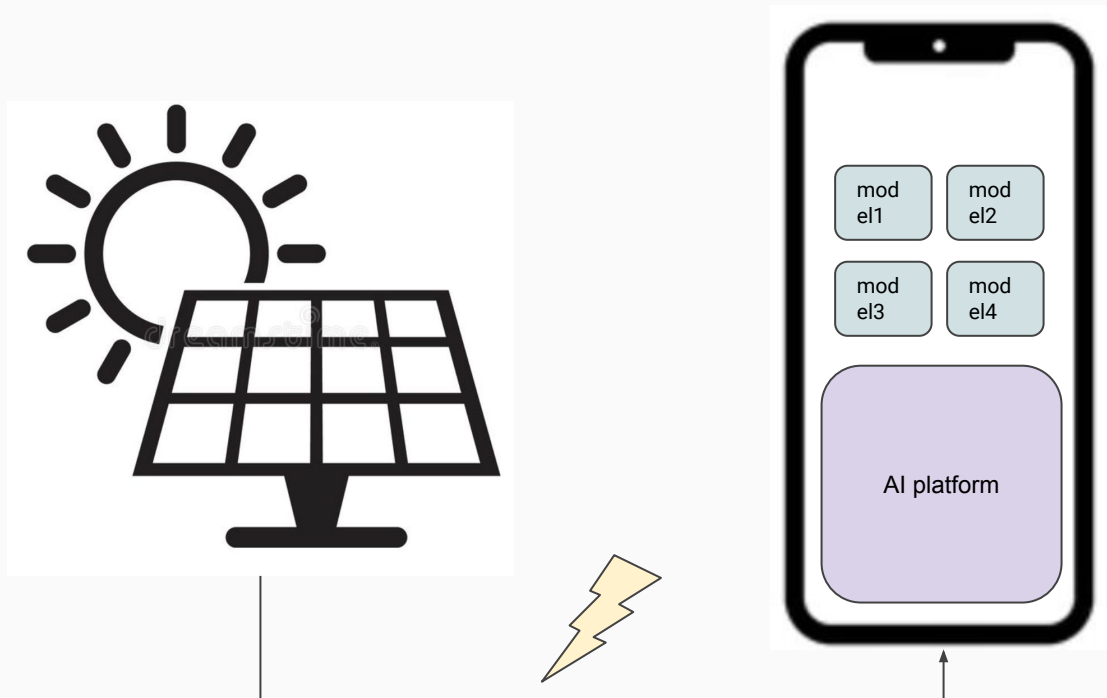
- 無所不在的AI - 但需要把資料送回後端機房處理
 - Globally, data transmission networks consumed 260-340 TWh in 2020, or 1.1-1.4% of global electricity use. [1]
- MLDL模型效能特別好 - 但只適用於已知資料集
 - Is it reasonable to use copurs between 2010-2015 to predict what people is talking about in 2022?
 - Is it reasonable to train a car 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?)



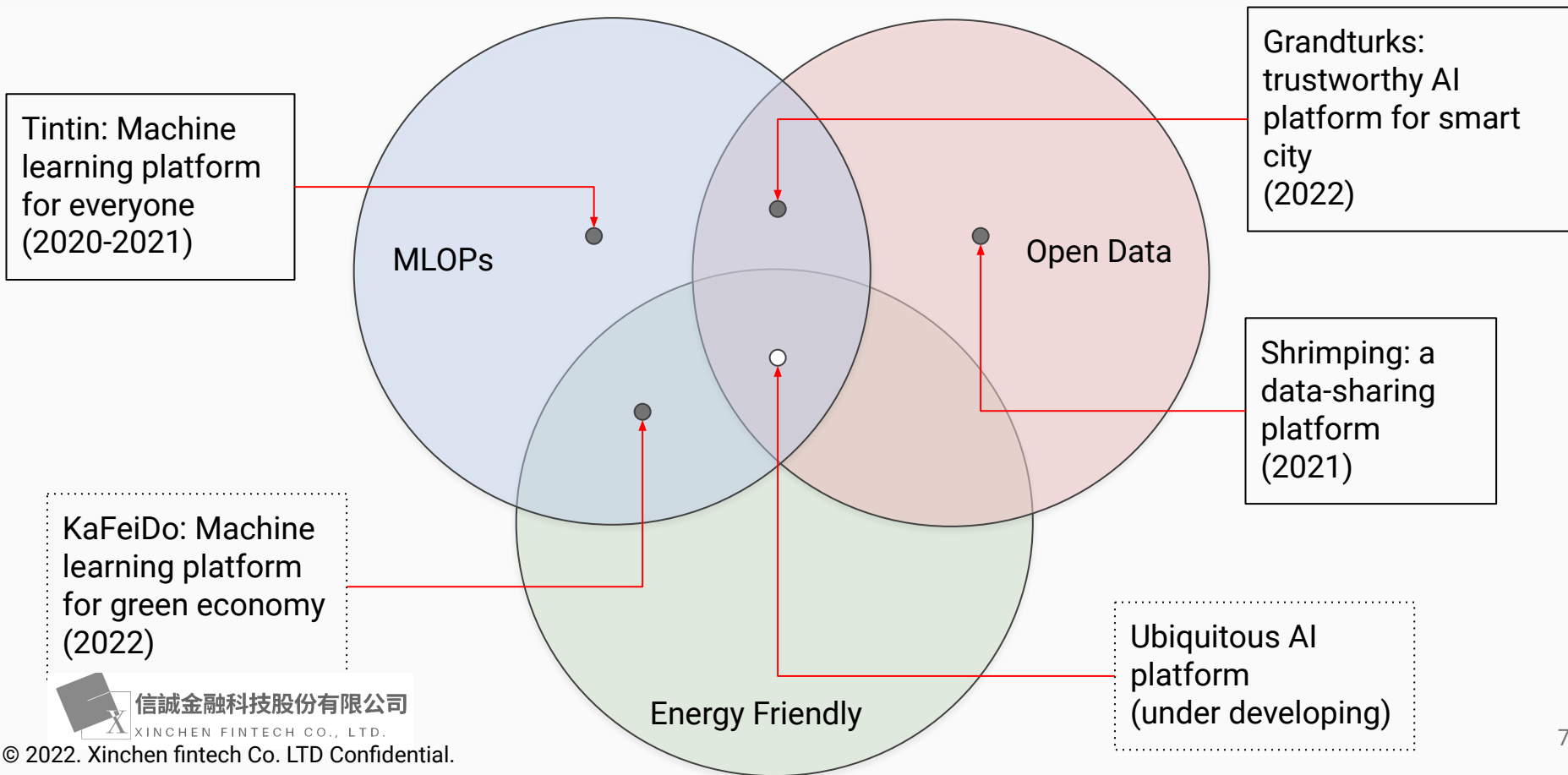
- Energy-friendly: low power consumption
- Ubiquitous: device can be carried to anywhere, even network is not accessible.

- Sustainability: Self-charging

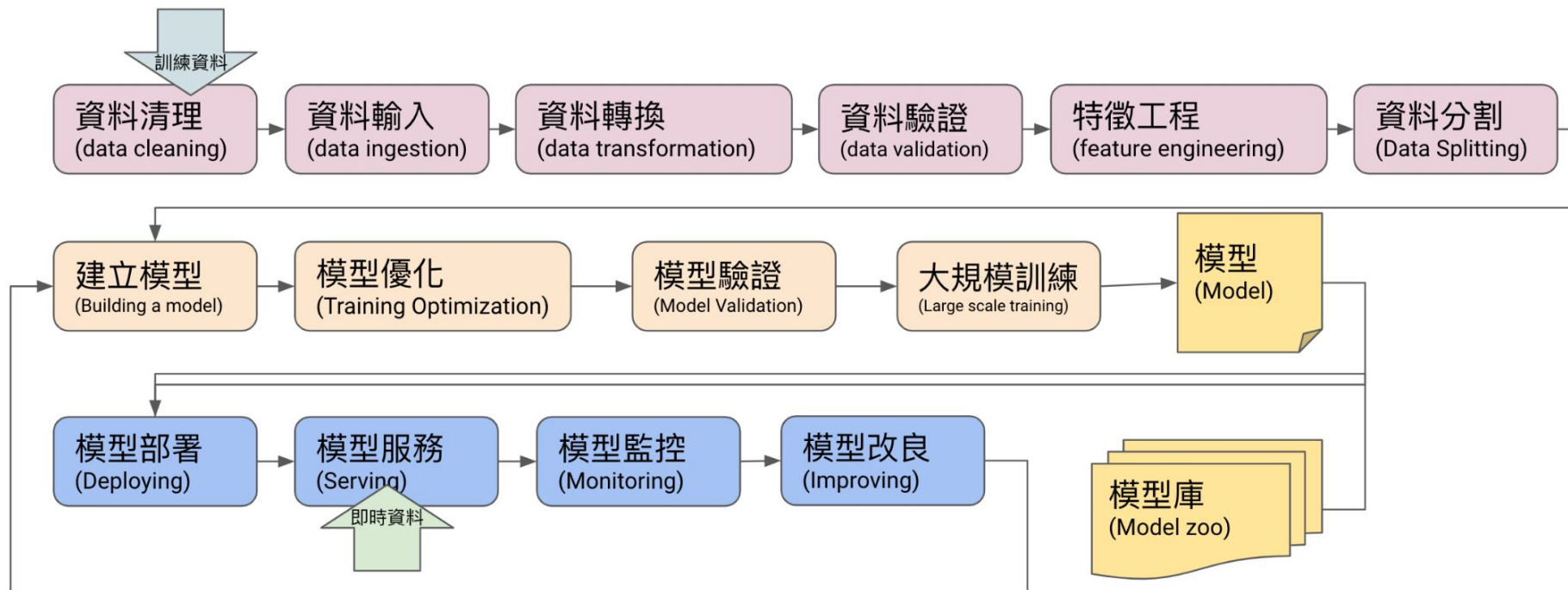


信誠金融科技股份有限公司
XINCHEN FINTECH CO., LTD.

RoadMap



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



信誠金融科技股份有限公司

XINCHEN FINTECH CO., LTD.

© 2022. Xinchen fintech Co. LTD Confidential.

■ Gathering
■ Modeling
■ Serving

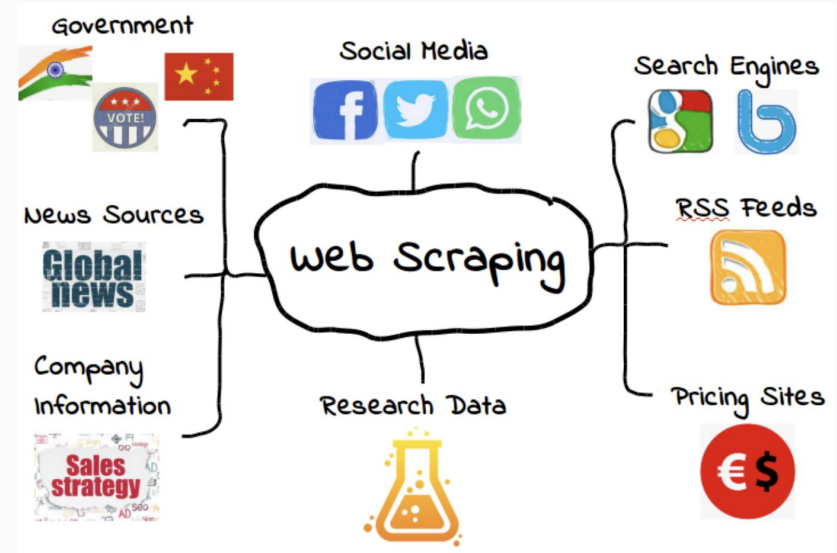
Gathering

Data Scraping With Selenium

Scraping is gathering information from public into your system for further analysis.

What is Web Scrapping?

- Web scrapping is an **automation** process of collecting **structure data** from public websites.
- Common Use Cases including
 - Competitor price monitoring on E-commerce
 - News monitoring from social network



Different ways of collecting data from websites (1/3)

- Manually (slow & slow & slow)



Different ways of collecting data from websites (2/3)

- API Requests with Python(Stable & Fast but not always works...Why?)
 - Not human-like interaction - anti-bot policy
 - Dynamic content
 - Too Frequently access

```
>>> import requests

>>> r = requests.get('https://www.google.com.tw/')

>>> print(r.status_code)
200

>>> print(r.text)
<!doctype html><html itemscope="" itemtype="http://schema.org/WebPage"
lang="zh-TW"><head><meta content="text/html; charset=UTF-8" http-equiv="Content-Type"><meta
content="/images/branding/googleg/1x/googleg_standa....
```

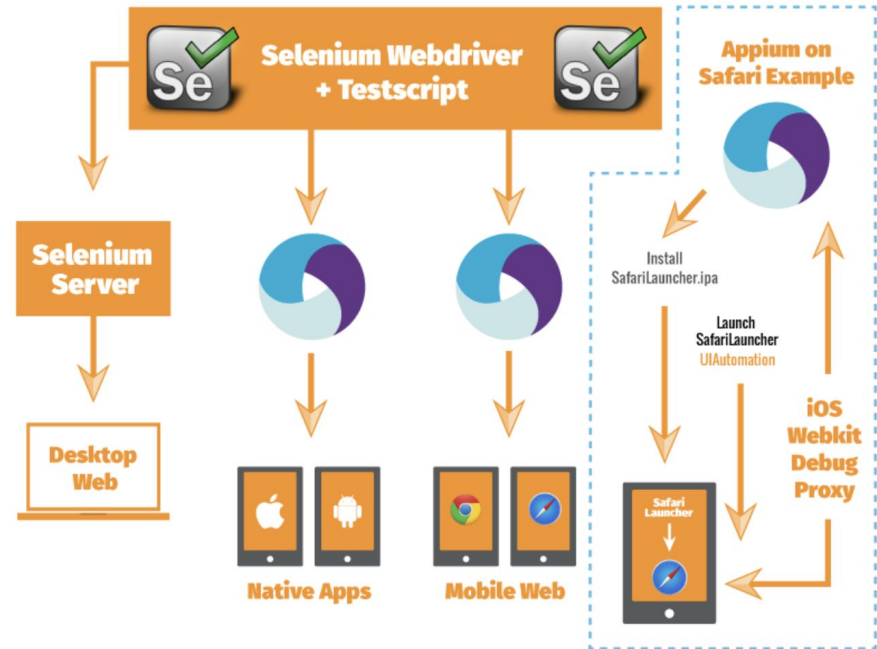
Key fields sent by Browser

- Without this information (including cookie, referrer, user-agent fields), your program is easy to detect as a bot and get banned when you are accessing restricted resources.

```
X Headers Payload Preview Response Initiator Timing Cookies
request headers
:authority: www.instagram.com
:method: POST
:path: /api/v1/feed/timeline/
:scheme: https
accept: */*
accept-encoding: gzip, deflate, br
accept-language: en-US,en;q=0.9
cache-control: no-cache
content-length: 153
content-type: application/x-www-form-urlencoded
cookie: mid=YpYwgAAEAAHaopTmodGr9VGA714p; ig_did=B8DC4143-652A-4F91-8CC7-7861EEE01533; ig_nrcb=1; ds_user_id=50697805654; csrftoken=sw4exlGoxUhwR2dK5yH; sessionId=50697805654%3ABXtwHsiI8CYm6Z%3A6%3AAYf-xPVGwmv536hXeB6WpCLJvBM5m65kDSHnTzHUqw
dnt: 1
origin: https://www.instagram.com
pragma: no-cache
referrer: https://www.instagram.com/
sec-ch-prefers-color-scheme: light
sec-ch-ua: "Google Chrome";v="107", "Chromium";v="107", "Not=A?Brand";v="24"
sec-ch-ua-mobile: ?0
sec-ch-ua-platform: "macOS"
sec-fetch-dest: empty
sec-fetch-mode: cors
sec-fetch-site: same-origin
sec-gpc: 1
user-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36
viewport-width: 1680
```

Different ways of collecting data from websites (3/3)

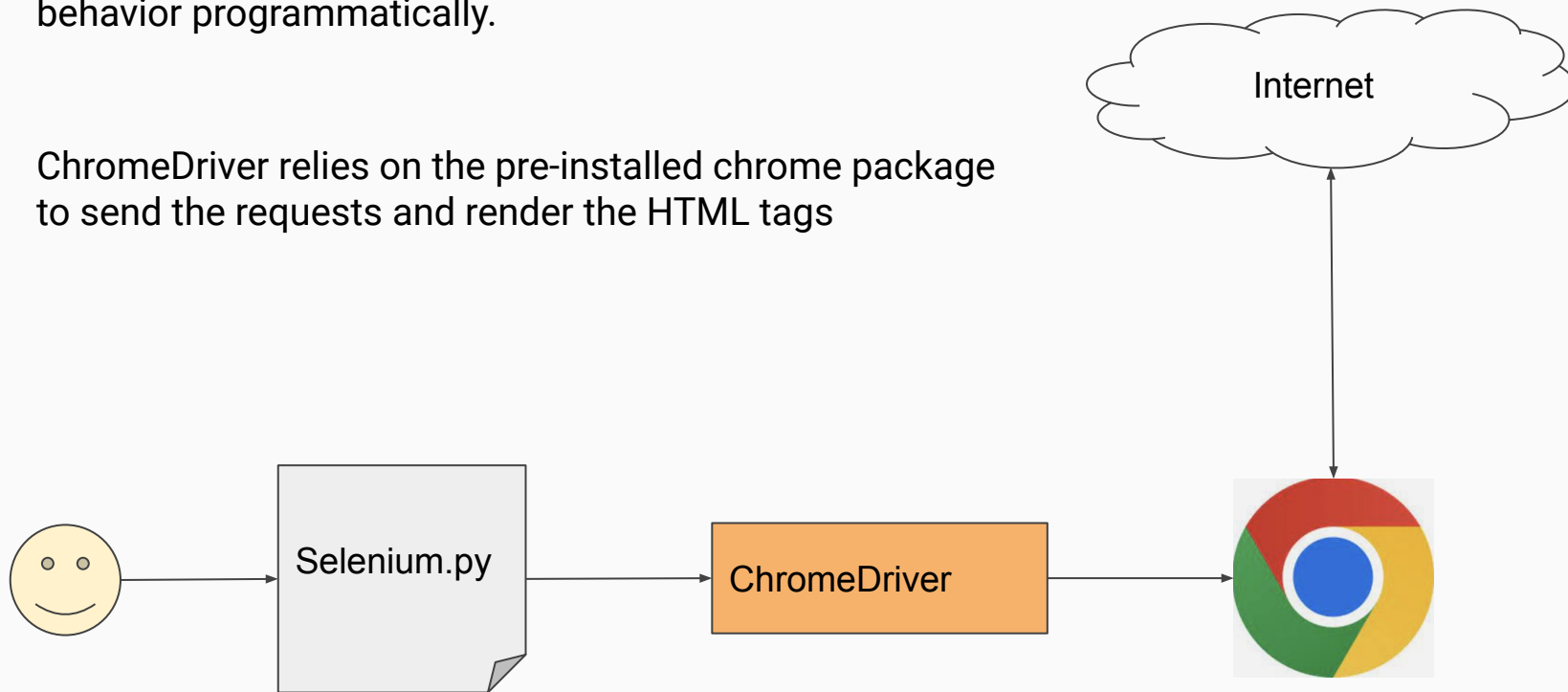
- **UI Automation with Selenium**
 - Selenium is an web automation tool which allows you to automate UI testing for different browsers via its webdriver.
 - Webdriver is used to control each browser's behavior making it more like real-human interaction.
- Because of its nature of automation and mimic human behavior, making it a good human-like web-scraping



Ref: <https://smartbear.com/blog/selenium-cross-browser-testing-on-mobile-devices/>

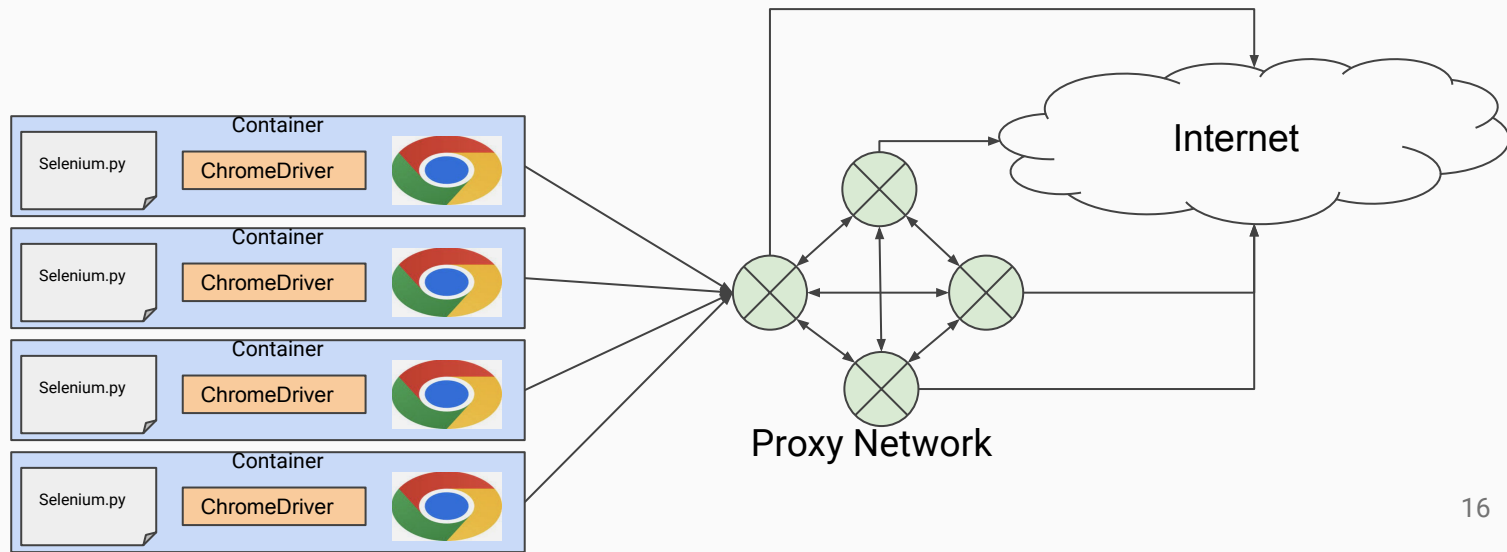
How Selenium WebDriver Works?

- ChromeDriver provides an interface to communicate with the underlying chrome browser, allows you to control its behavior programmatically.
- ChromeDriver relies on the pre-installed chrome package to send the requests and render the HTML tags



How Selenium WebDriver Works In Scale?

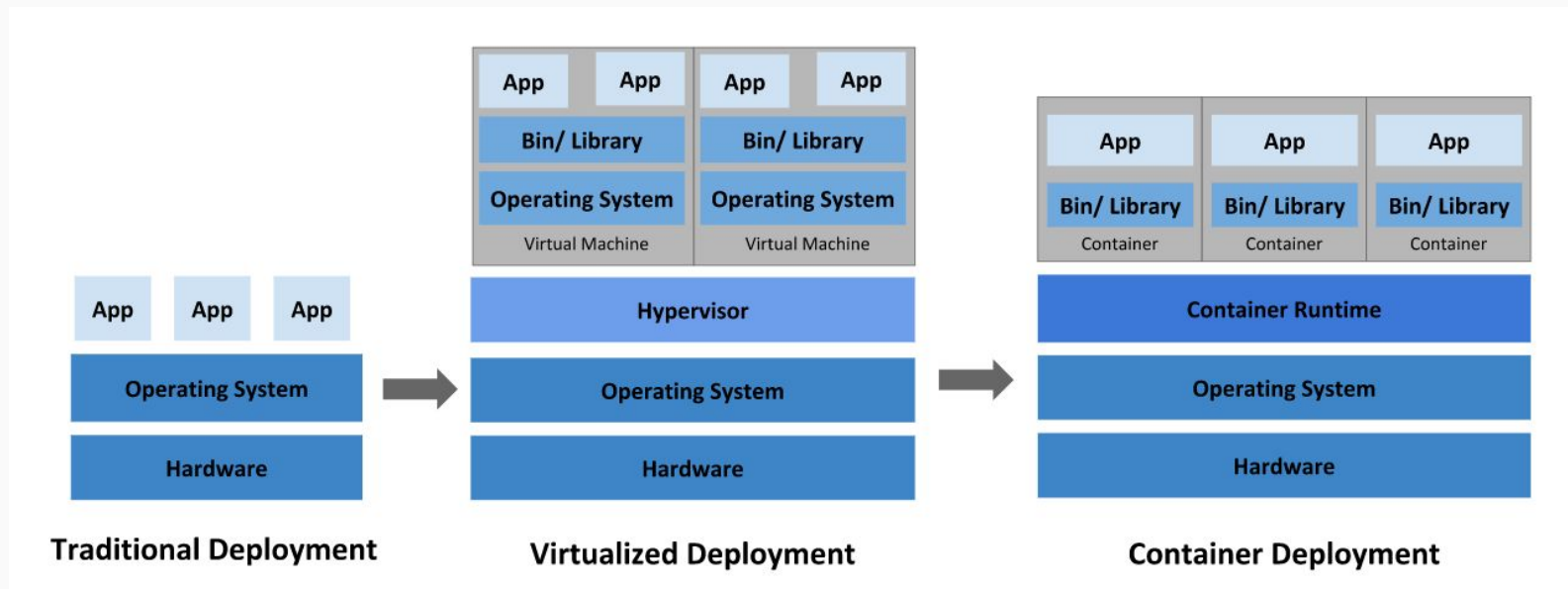
- Container provides a way to encapsulate your program into a single container image, allowing you to run it on any platform. (easy for scaling up/down)
- Proxy provides a good way to hide your source(IP / Identity) when you are accessing the public internet. The destination server won't know who is scraping their data.



What is container?

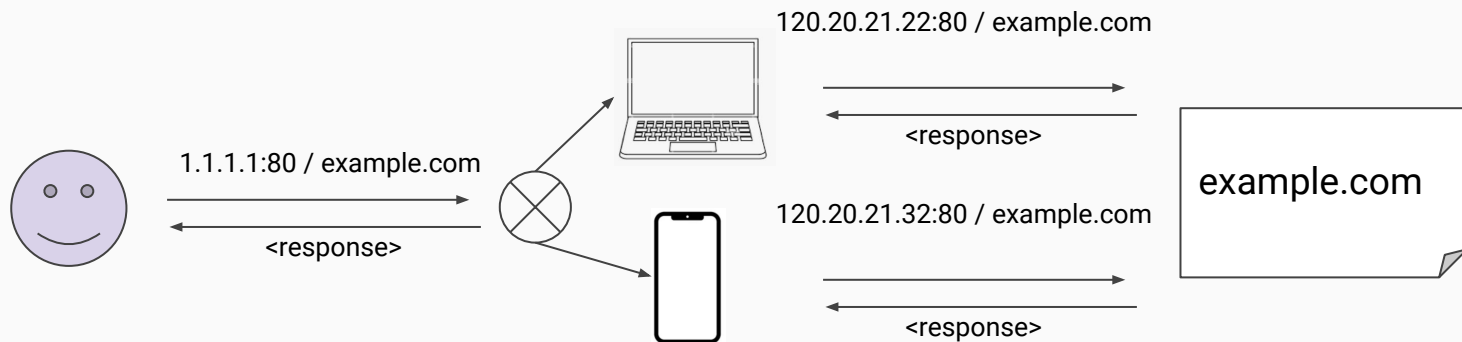
- **Container**

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



How Proxy is working?

- Hide your tail (your actual ip address)
- Improve speed with content caching



Scraping vs Anti-Scraping: A battle that never ended.

- How to bypass Anti-Scraping Tool
 - Keep Rotate Your IP Address
 - By a well-paid proxy service provider
 - By restricting a certain area of your location
 - Use Real User Agent and valid Referrer
 - UA(User-Agent) are http header which is used identify what browser type you used to visit this website.
 - Keep a list of valid UI and use them randomly.
 - Avoid Periodically Requests
 - Keep random intervals between requests, random delay are helpful.
 - Watch Exception and update your code
 - Anti-scraping tool would change DOM structure frequently, making your script failed to find target elements.
- However, Implementing these guidelines could costly if your application is still below a certain scale.

Shrimping: A data-sharing platform

Shrimping provides a unified way for clients to get human-centric information in a simple, easy, and low cost fashion.

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

Web scraping scenarios for Shrimping.

- **UI Validation**
 - When working with business partners, it is extremely important that your partner has interpreted your product correctly.
- **Collecting sell history on an ecommerce platform**
 - Track the sell volume of all products on an eCommerce platform.
 - As the number of products could be big (approximately 100M active products), how to get each product's sell records on daily basis is extremely challenging.
- **Scraping and analyzing KOL's feeds on social network platforms**
 - Find out a KOL and his/her fans preference for retargeting, reselling, or other marketing strategies.
 - Social network platform always implemented anti-bot mechanism, making it hard to collect in a large scale.

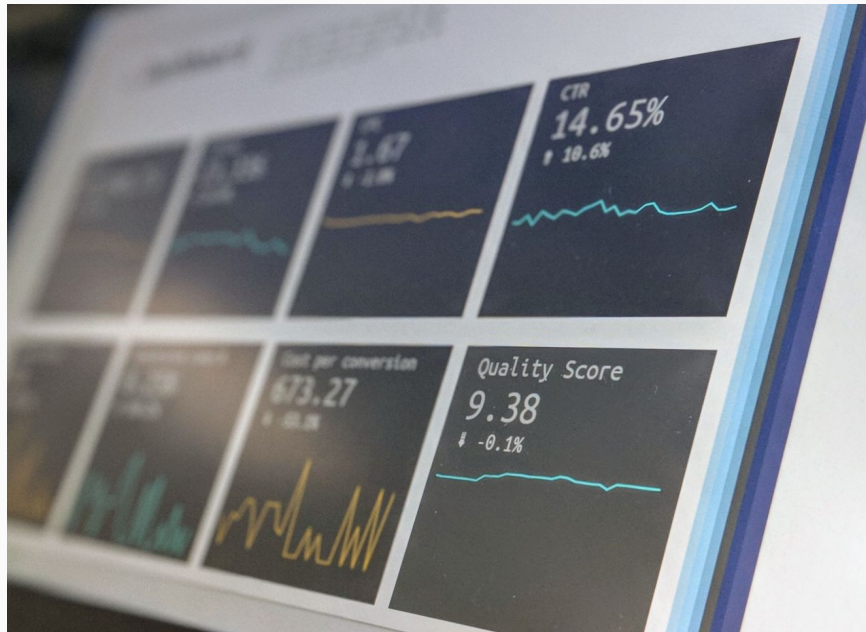
Serving

Serving machine
learning models with
low cost

Serving is the last mile to publish
your machine learning models to
the public.

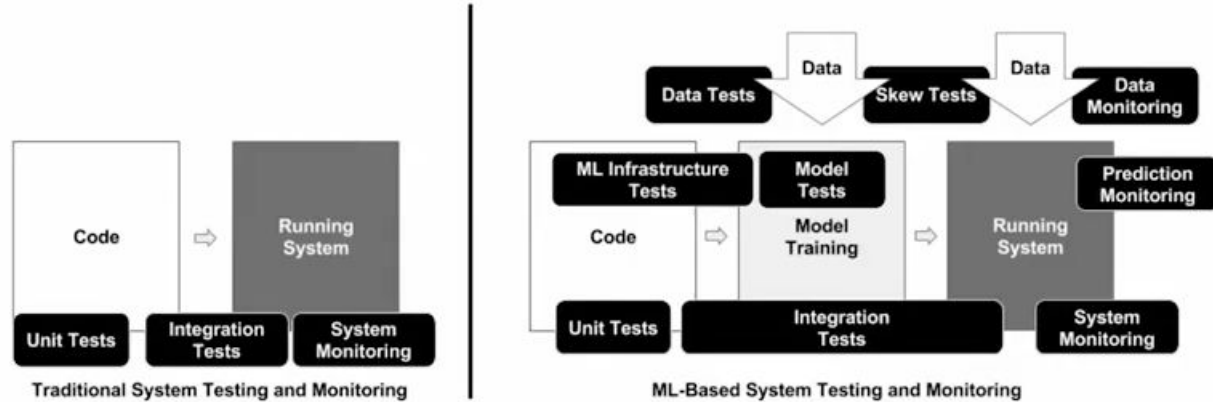
Model Serving

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



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

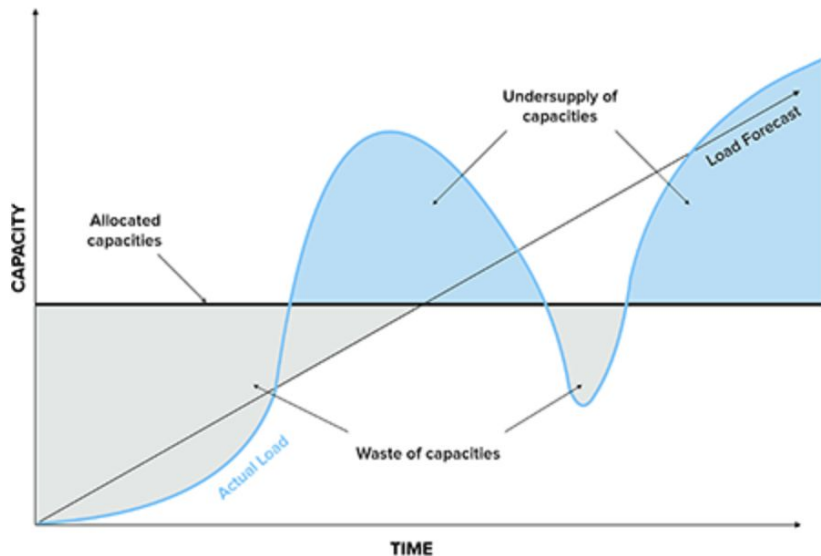
Traditional vs. ML infused systems



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

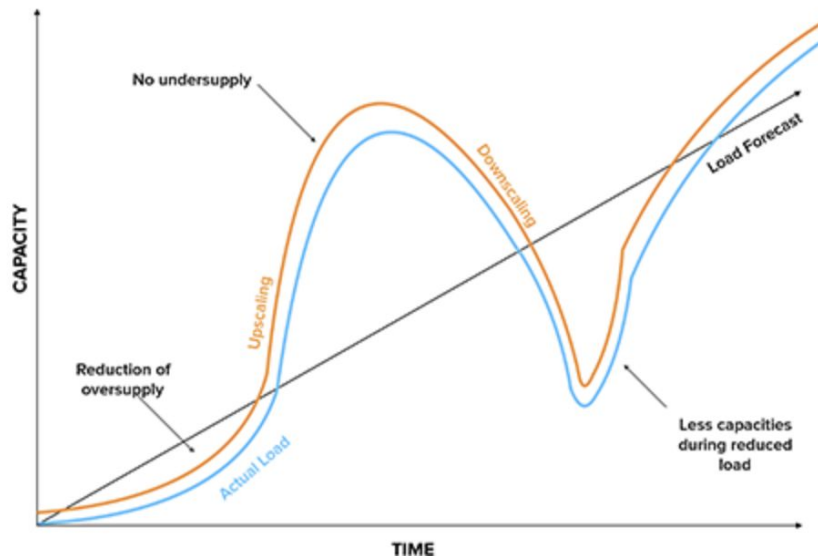
Cost, Cost, and Cost

Static Architecture



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

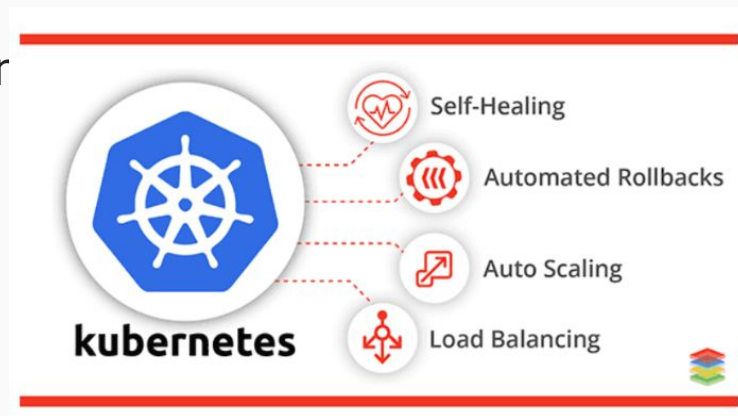
Auto Scaling Architecture



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

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
- ...



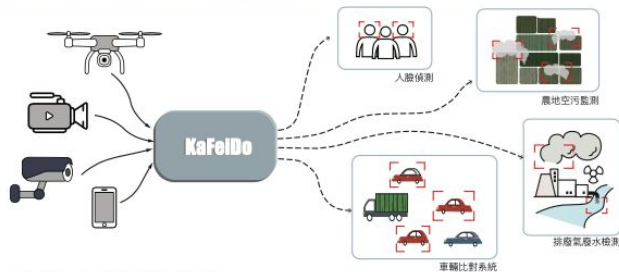
A hand is shown adjusting a potentiometer on a breadboard. The breadboard is populated with various electronic components, including resistors and integrated circuits. The background is blurred, showing more of the breadboard and components.

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：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護上的人事費用，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板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，並隨時依據尖峰需求而提高服務能力。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) / 雲端服務(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/電郵通知)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

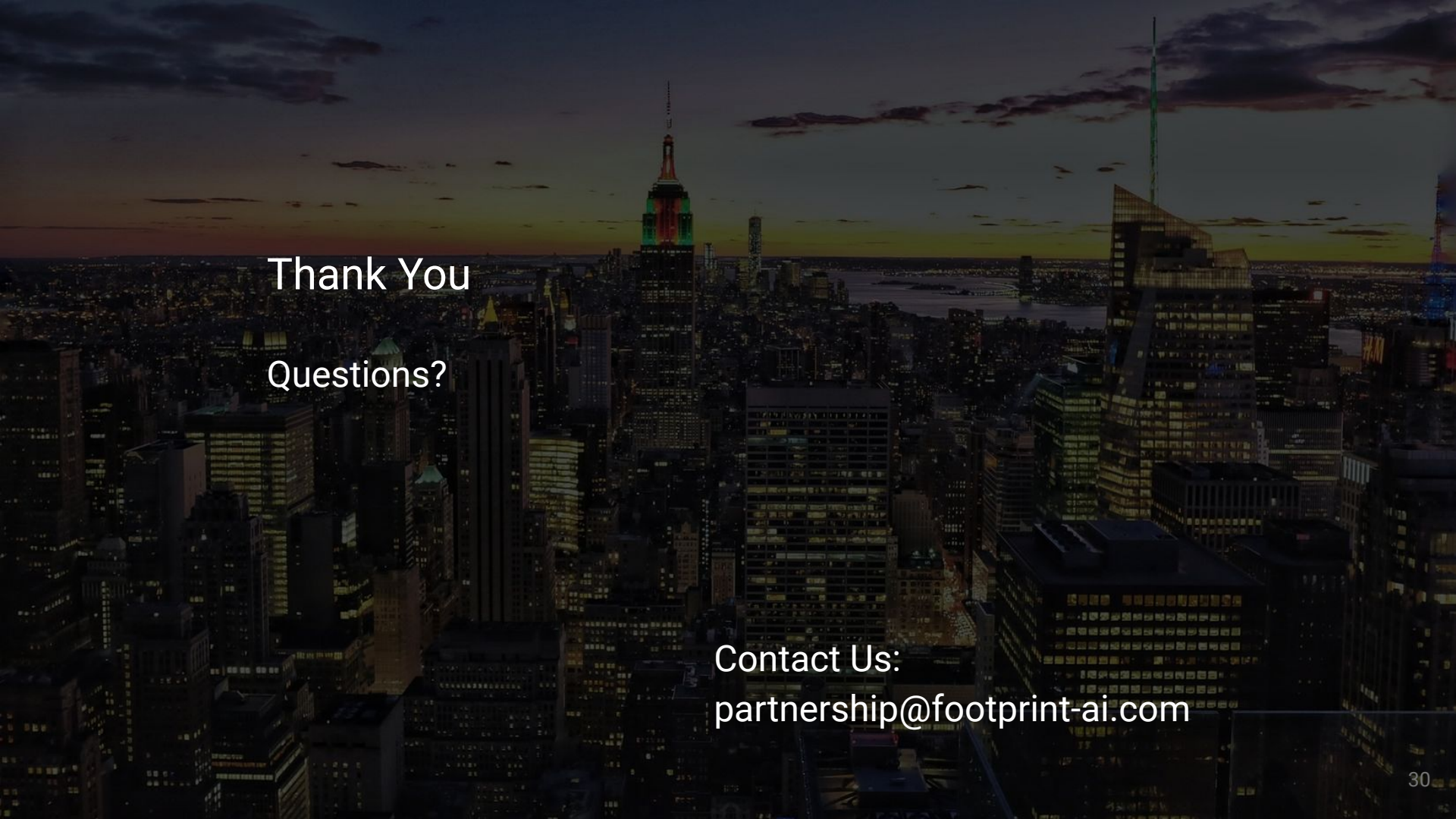
本公司專營機器與深度學習平台 / 網路資料中台 / 多資料模型推論架構 / 客製化模型等服務。
地址：103台北市大同區承德路三段287-2號 email: kafeido@footprint-ai.com



Tintin
Machine Learning Platform For Everyone




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

An aerial photograph of the New York City skyline at dusk. The Empire State Building is prominently featured in the center, illuminated with its characteristic red, white, and blue lights. The city is densely packed with skyscrapers, many of which have their lights on. The sky is a mix of dark blue and orange, indicating the time is either early morning or late evening. The water of the harbor is visible in the background.

Thank You
Questions?

Contact Us:
partnership@footprint-ai.com



***“The Best Engineers
Are Lazy”***

-Ancient Engineering Proverb

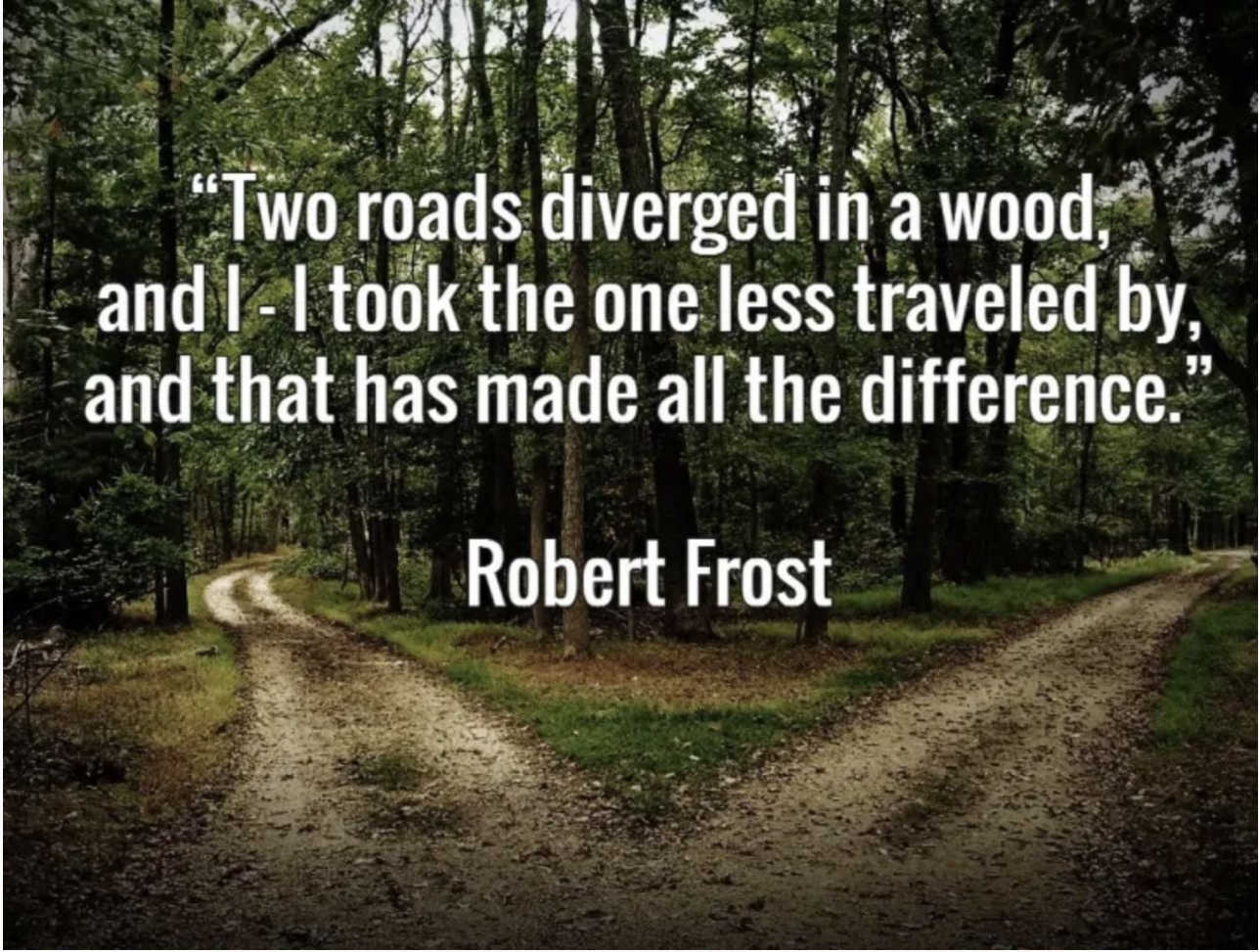
Some reflections on my career journey

- Life (or career) is an investment.
 - The first principle is never to lose your capital.
 - Always trade off between RISK and RETURN.
- How to choose your first job?
 - High-salary job vs job with interest?
 - Job in a big company vs job in 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

- **Be proactive**
 - Are you ready to play ball after you know how to play ball?
- **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.



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

Robert Frost

- Slides:
 - <https://github.com/FootprintAI/talks/tree/main/slides>
- Multikf
 - <https://github.com/FootprintAI/multikf>
- Kubeflow Workshop
 - <https://github.com/footprintai/kubeflow-workshop>
- Selenium example
 - <https://github.com/FootprintAI/selenium-example>