

devfest
2020

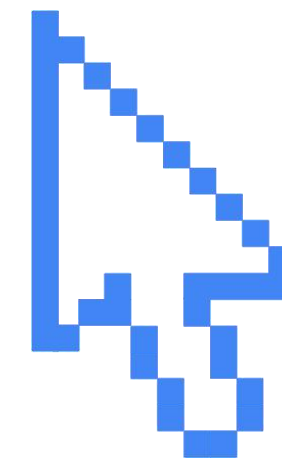
如何透過TensorFlow建構負責任的AI系統

Jerry老師
技術長, APMIC OpenTalk
Google機器學習開發專家





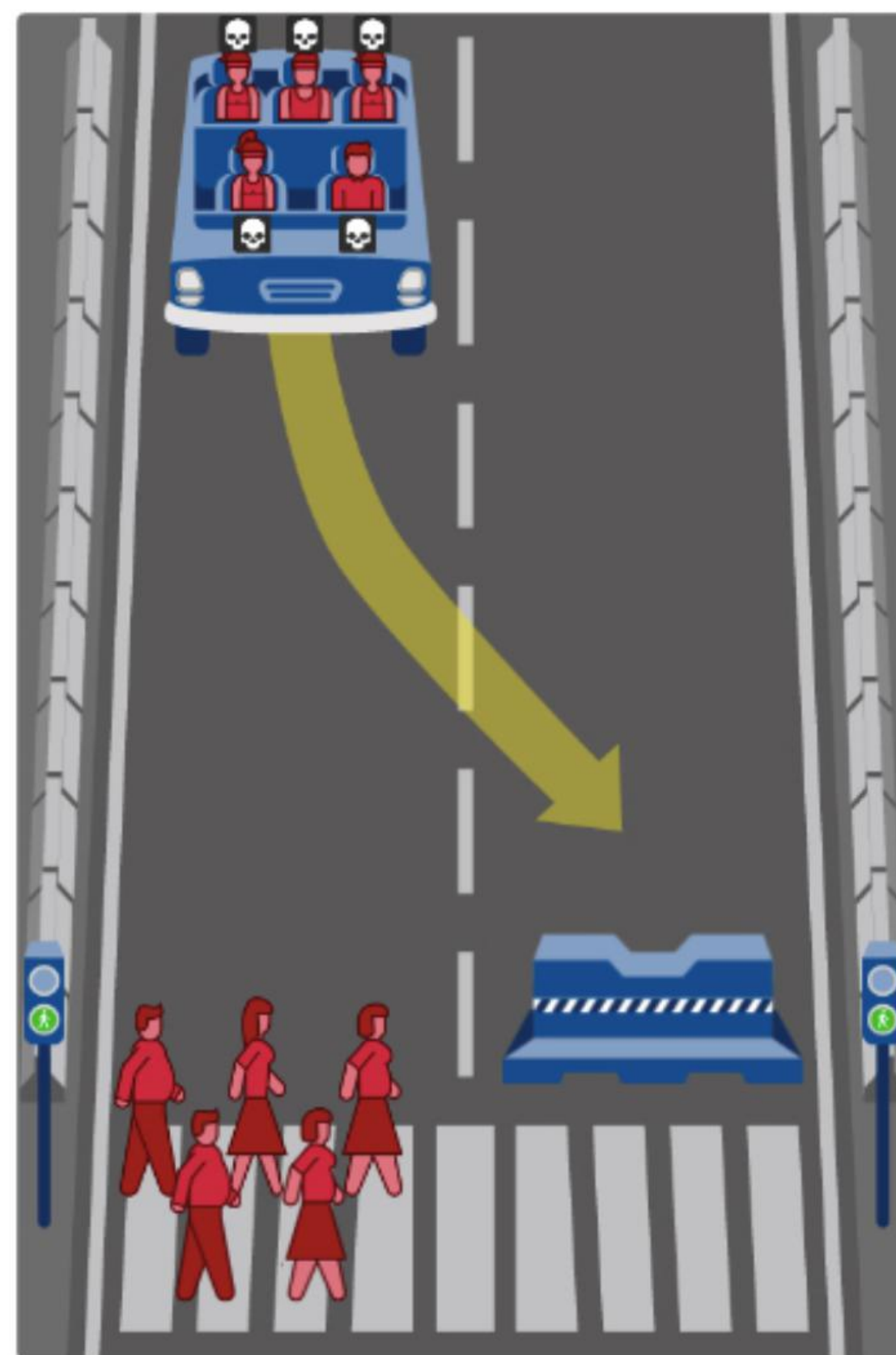
負責任的AI Responsible AI



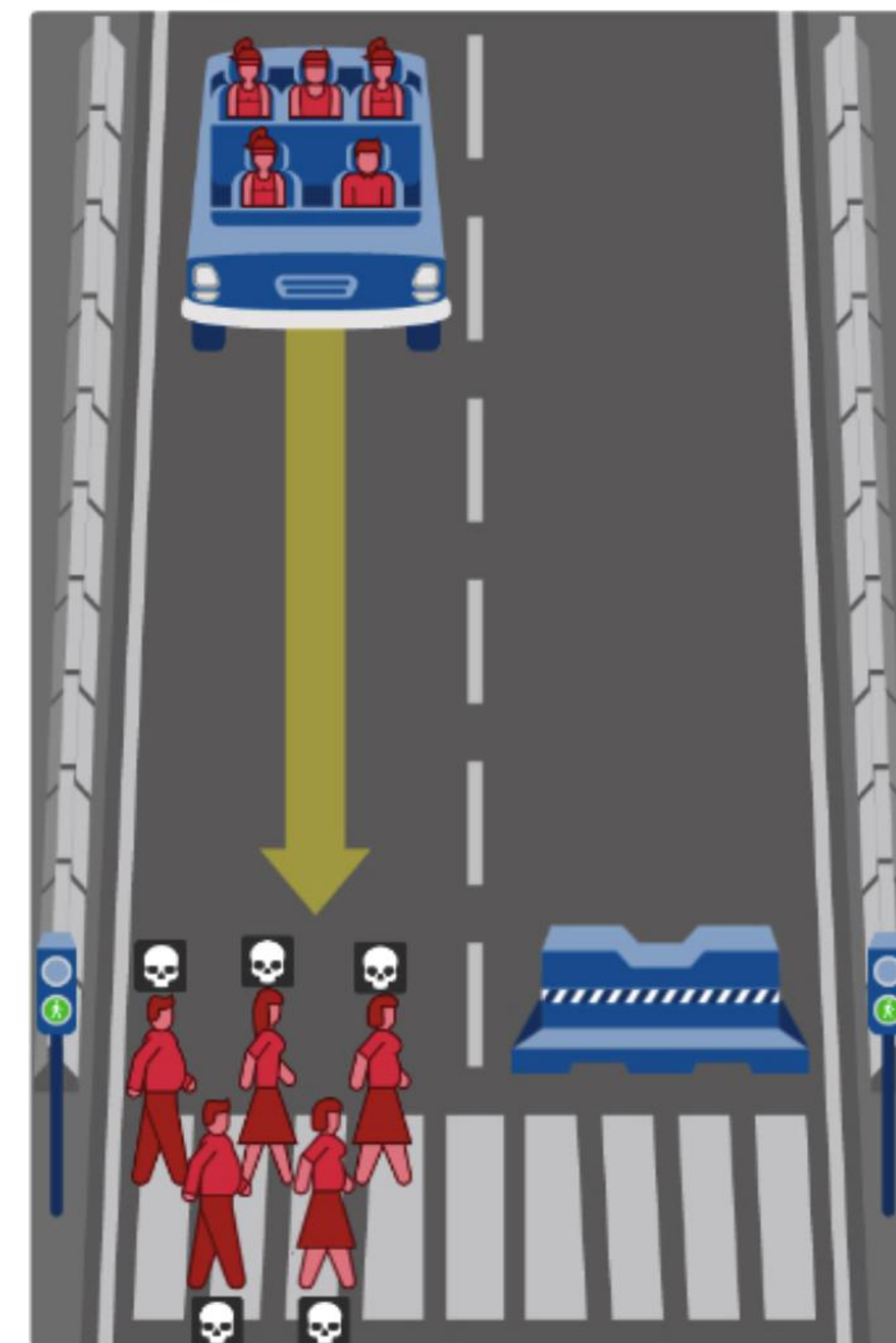


1. 何謂不負責任的AI? 如何給AI對的 價值觀?

What should the self-driving car do?

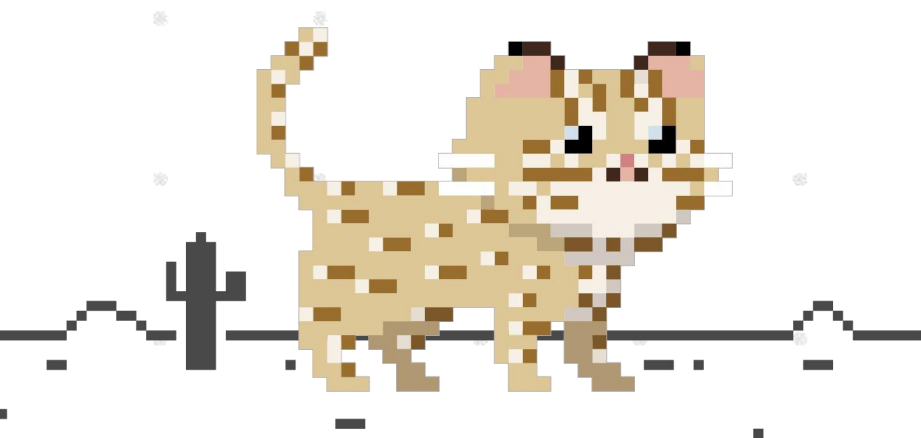


Show Description



Show Description

▲你到底會犧牲乘客還是路人？

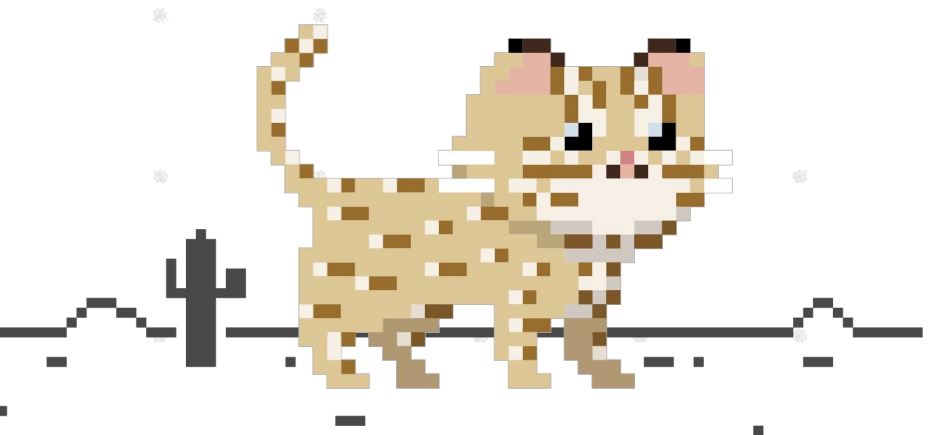




1. 何謂不負責任的AI?

常見機器學習系統不負責任的問題：

人臉辨識上因為沒有女性數據，導致無法辨識女性，就導致了數據選擇的偏見

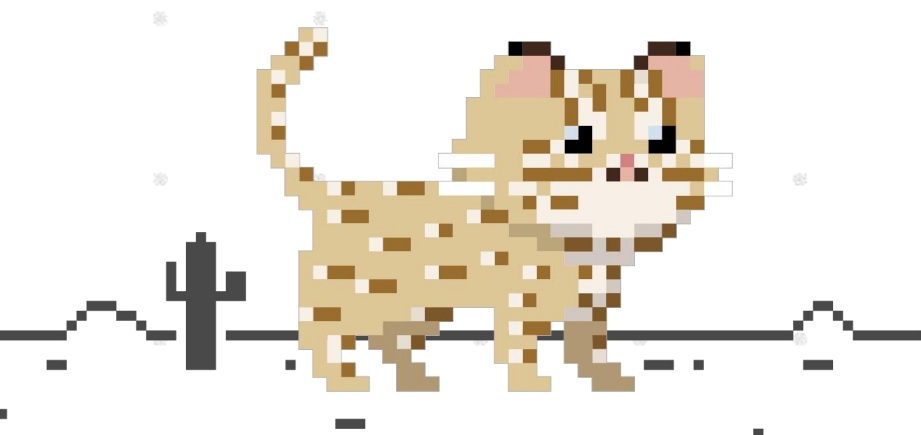
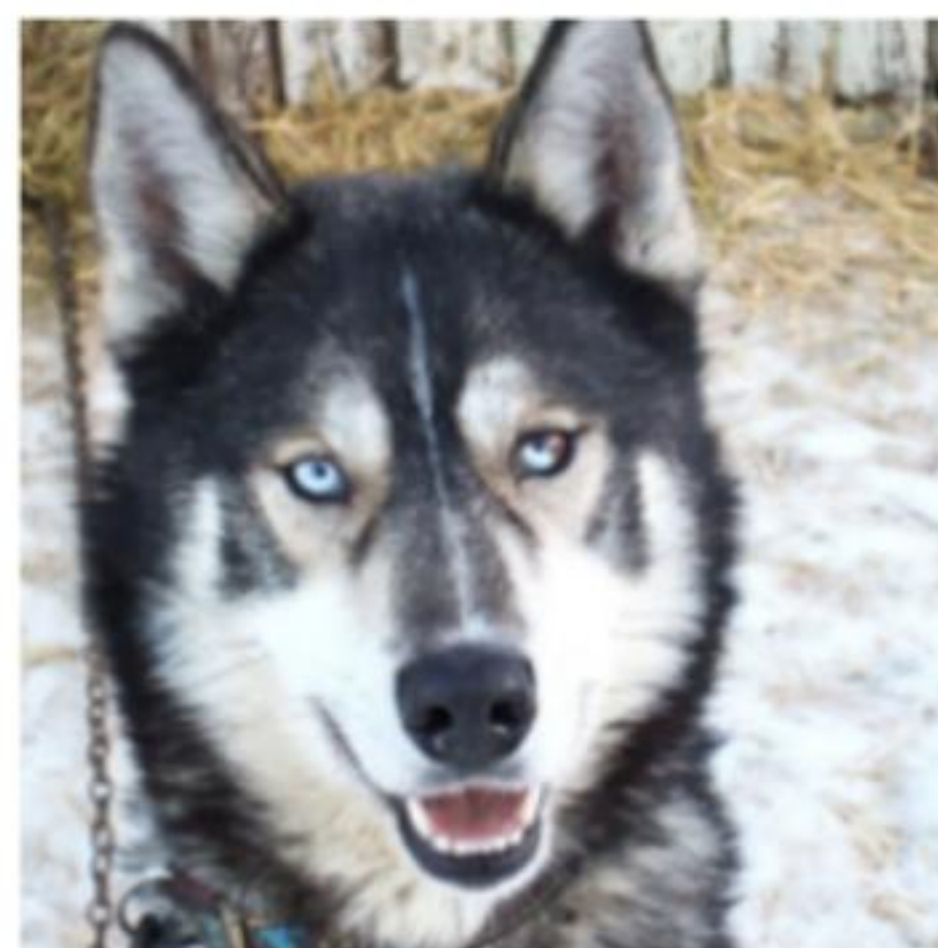




1. 何謂不負責任的AI?

常見機器學習系統不負責任的問題：

因為哈士奇的白色的臉與雪的訓練偏差，導致看到雪就辨識是哈士奇

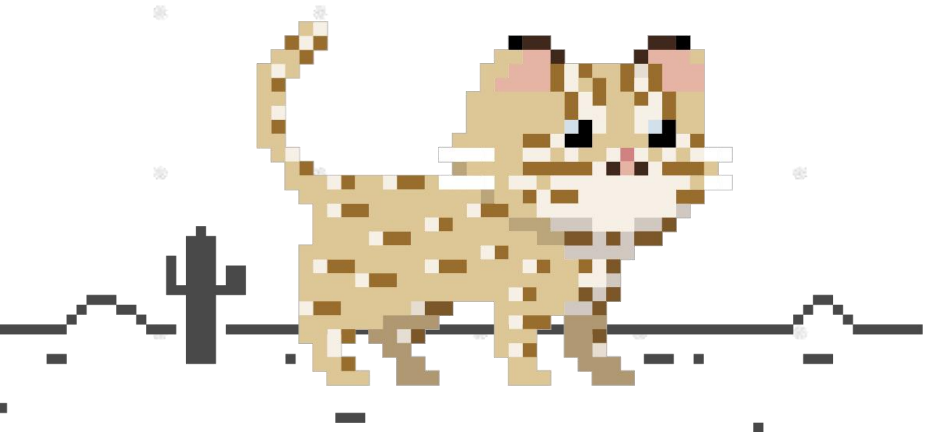




1. 何謂不負責任的AI?

常見機器學習系統不負責任的問題：

因為資料的偏誤，導致預測感染肺炎機率的模型認為患有「氣喘與心臟疾病」的人死於肺炎的機率要小於「一般健康」的人





人類添加偏差

人類添加偏差

人類添加偏差

人類添加偏差

人類添加偏差

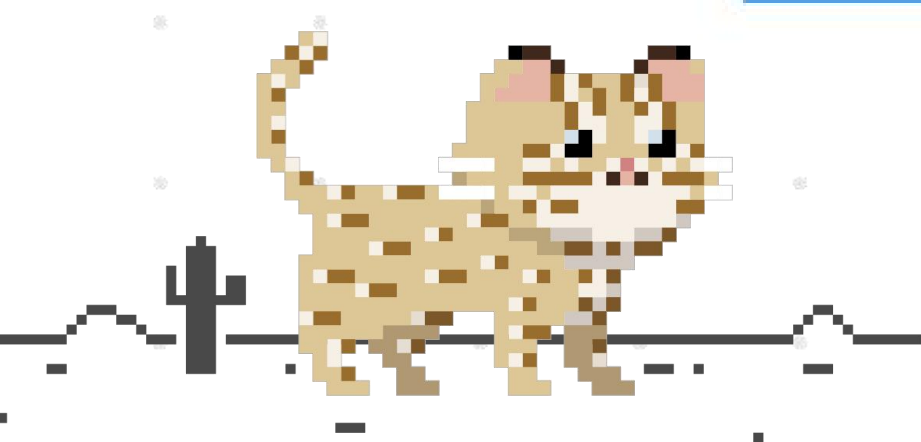
定義問題

準備數據

建立
與
訓練
模型

部署
模型

迭代
過程

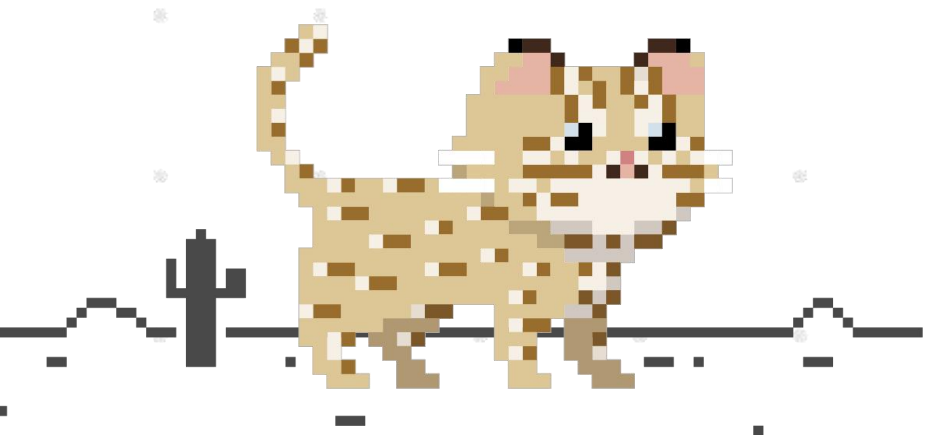




1.負責任的AI建構概念

簡稱RAI，透過一些方法建構出對所有人有利的 AI系統，
包含：

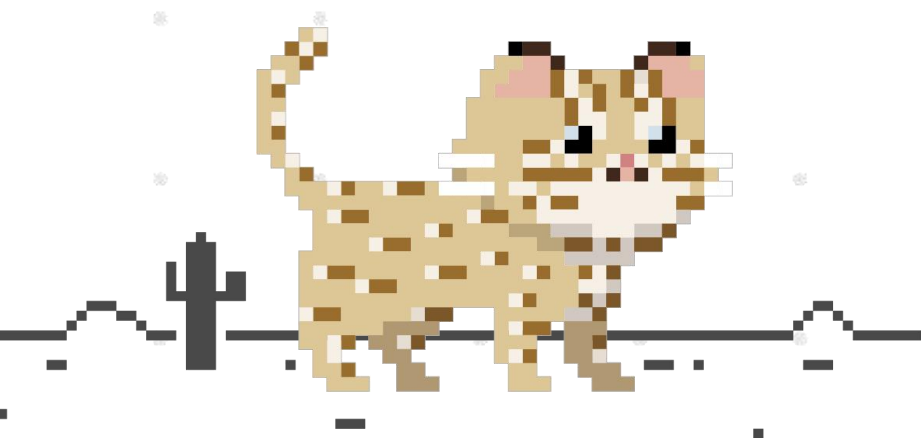
- 。採用以人為本的方法運用機器學習技術
- 。建立公平且可包容所有人的系統
- 。確保系統能夠如預期般運作並可解釋運算過程
- 。能妥善保護隱私權
- 。維持 AI 系統安全





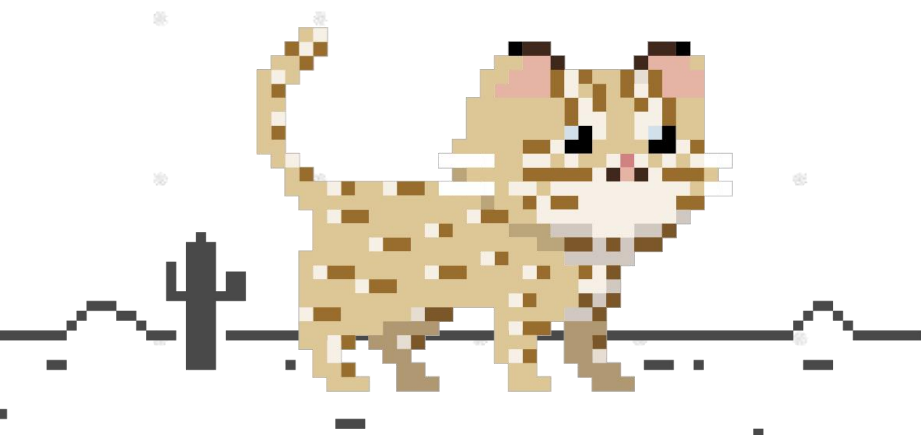
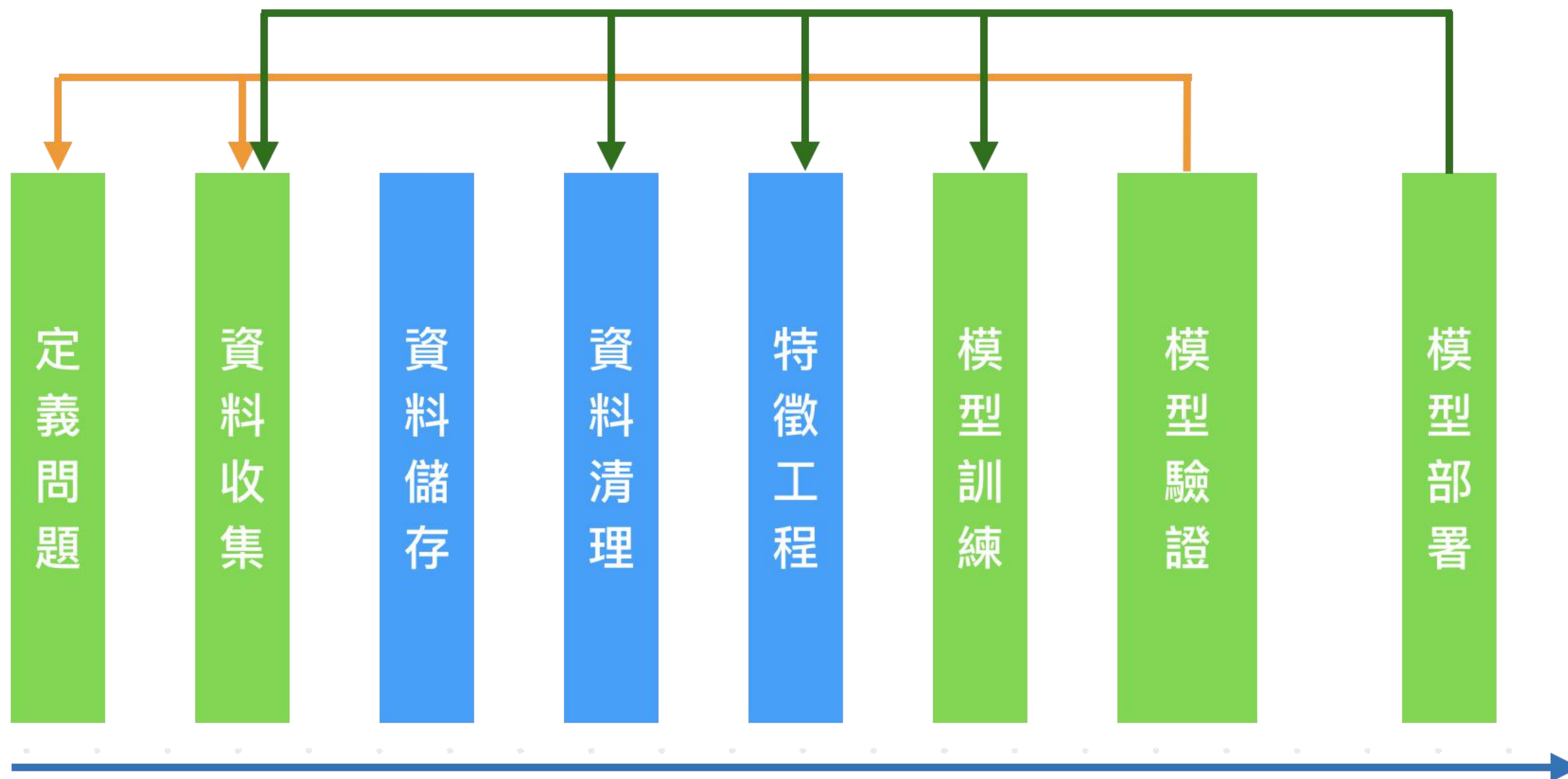
1.負責任的AI在企業的價值

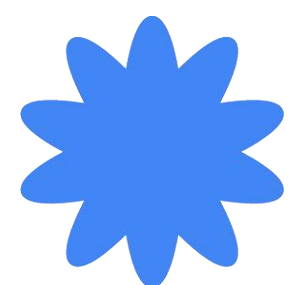
1. 確認如何透過AI轉換價值
2. 用於評估AI的公平性的指標
3. 在企業流程中分層與建構每一層的建議





1. 建構負責任的AI所顧及的流程





如何在機器學習 中導入RAI





界定問題

- 。我的機器學習系統是為誰而設計？
- 。確認整個機器學習的資料流程與目標

如：給品管部分掌握零件損壞的時間

步驟 1

界定問題

使用下列資源，設計出蘊含 Responsible AI 原則的模型。

People + AI Guidebook

人與 AI 研究 (PAIR) 指南

進一步瞭解 AI 開發流程和重要注意事項。

[瞭解詳情 ↗](#)

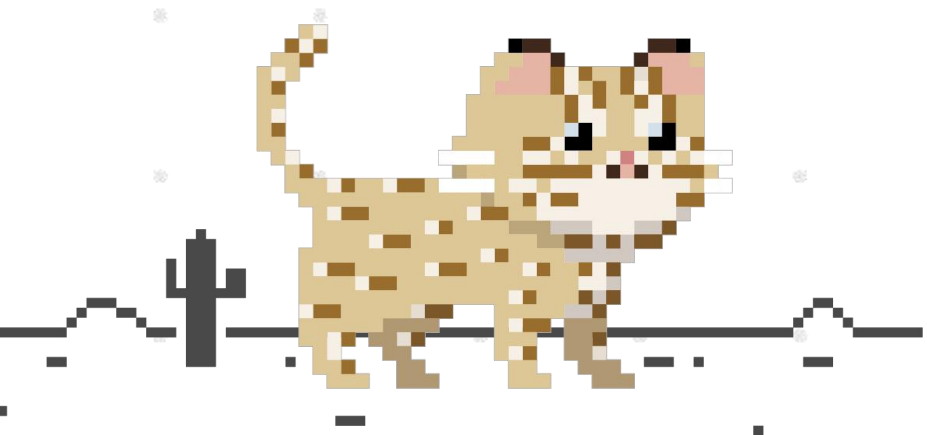


PAIR Explorables

透過互動式視覺化呈現的方式，探索 Responsible AI 的關鍵問題與概念。

[瞭解詳情 ↗](#)

<https://pair.withgoogle.com/explorables/>





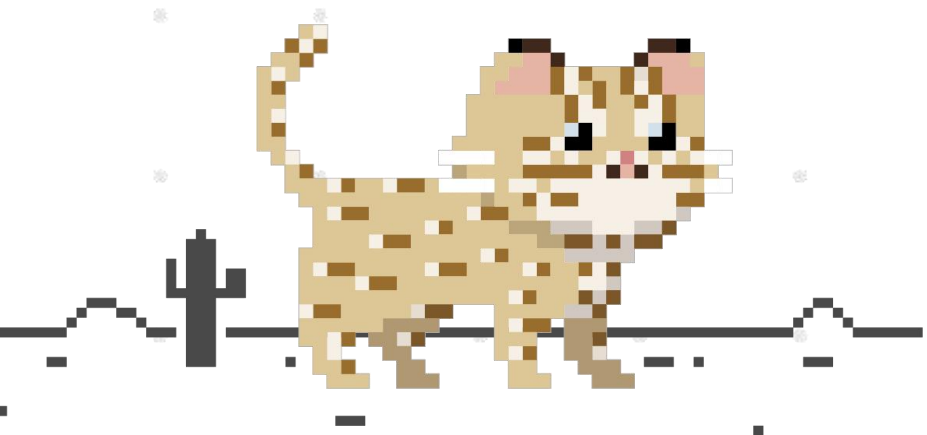
準備資料

- 。資料是否代表問題本身、具備真實環境？
- 。我的資料是否存在偏誤？
- 。我個人是否存在偏誤？

如：零件損壞的因子是否掌握？ 資料是否有問題？



<https://pair-code.github.io/facets/>





- 數據有出現大量缺少資料的特徵，表示有可能某些關鍵結果無法正確體現

缺失特徵值 Missing Feature Values

| 位置 | 均收 | 戶數 | 設施 | 位置 | 均收 | 戶數 | 設施 |
|-------|------|-----|----|-------|------|-----|----|
| 111.6 | 3000 | 200 | 2 | 111.6 | 3000 | 200 | 2 |
| 218.6 | 3000 | 200 | 2 | 218.6 | ? | 200 | 2 |
| 567.9 | 3000 | 200 | 2 | 567.9 | ? | 200 | 2 |

- 數據有出現突兀異常的特徵狀況，導致機器學習模型產生偏差

突兀特徵值 Unexpected Feature Values

| 戶長 年齡 | 均收 | 戶數 | 設施 |
|----------|------|-----|----|
| 70 | 3000 | 200 | 2 |
| 150 | 3000 | 200 | 2 |
| 30 | 3000 | 200 | 2 |

資料歪斜 Data Skew

- 數據中任何形式的資料歪斜，導致模型出現偏見

| 戶長 年齡 | 均收 | 戶數 | 設施 | 購買 |
|----------|------|-----|----|----|
| 70 | 3000 | 200 | 2 | 1 |
| 150 | 3000 | 200 | 2 | 1 |
| 30 | 3000 | 200 | 2 | 1 |



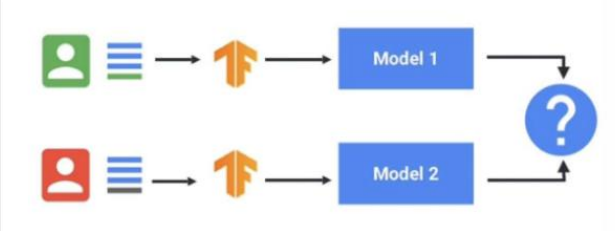


訓練模型

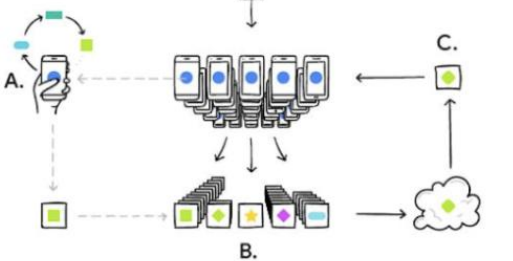
使用可在模型中具備下列條件的訓練方法，包含

- 。訓練時的隱私權保護
- 。聯合學習（Federated Learning）
- 。資料平衡
- 。靈活、受控制與可解釋的網路模型

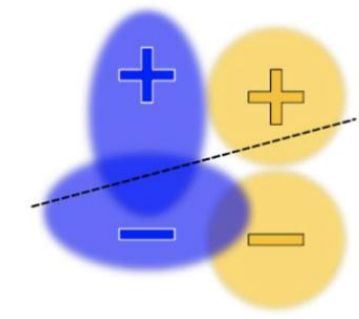
如：確保資料訓練過程的隱私性



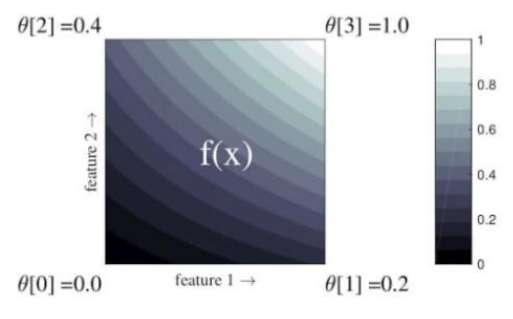
TF Privacy
訓練機器學習模型時兼顧隱私權保護。
[瞭解詳情 ↗](#)



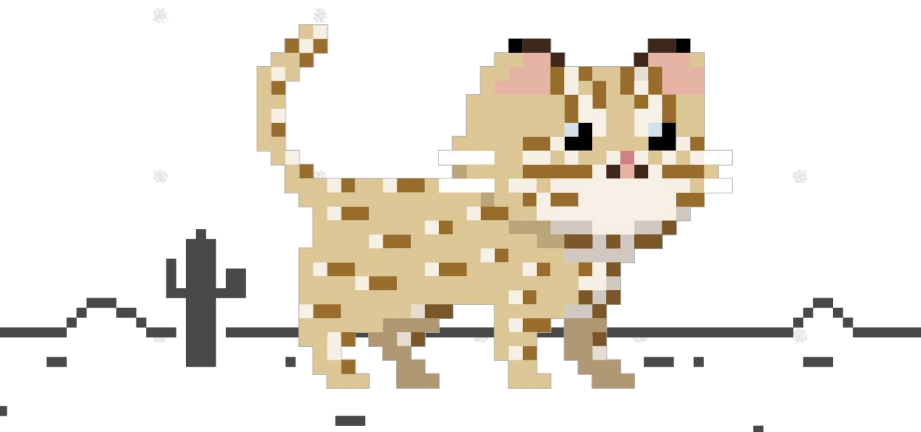
TF Federated
使用聯合學習技巧訓練機器學習模型。
[瞭解詳情 →](#)



TF Constrained Optimization
最佳化受不平等限制的問題。
[瞭解詳情 ↗](#)



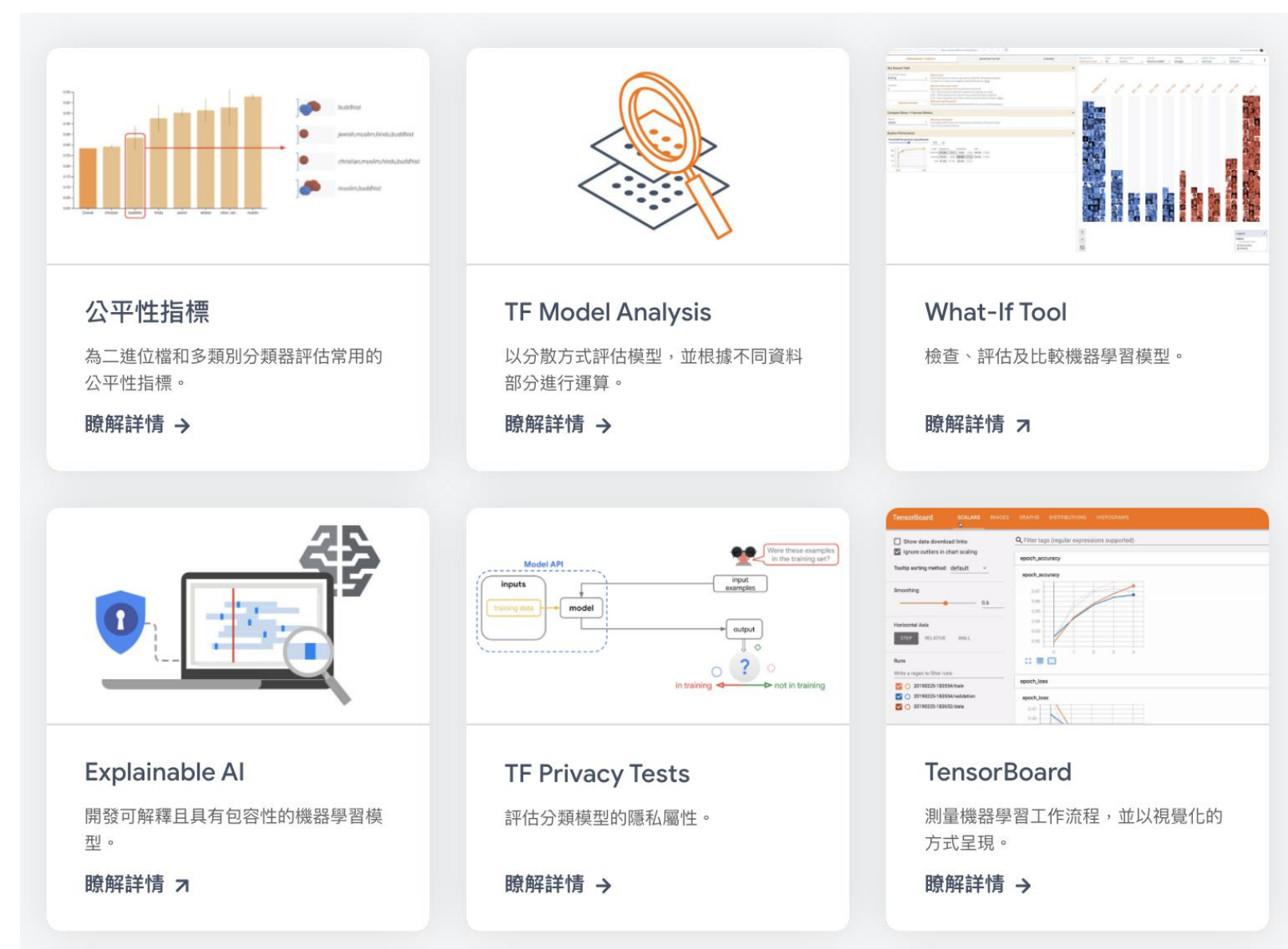
TF Lattice
實作具有彈性、受控制、可解釋且以 Lattice 為基礎的模型。
[瞭解詳情 →](#)



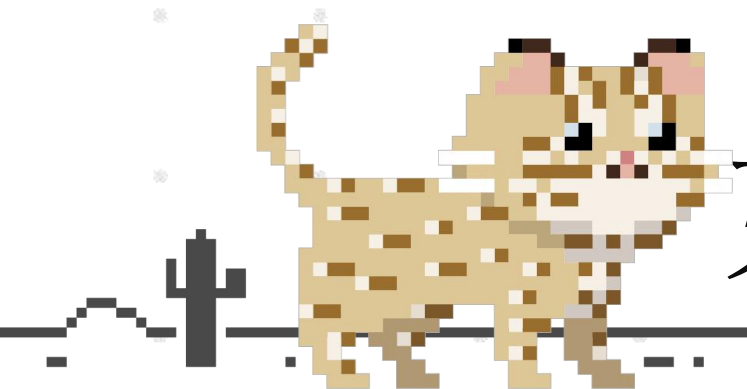


評估模型

- 。針對真實世界的不同使用者、情境做評估體驗
- 。從Dogfood測試中不斷迭代
- 。持續發佈與測試作業
- 。公平性測試
- 。可解釋性
- 。隱私測試
- 。安全性測試



如：預測出來的時間與自己使用零件的損壞比較

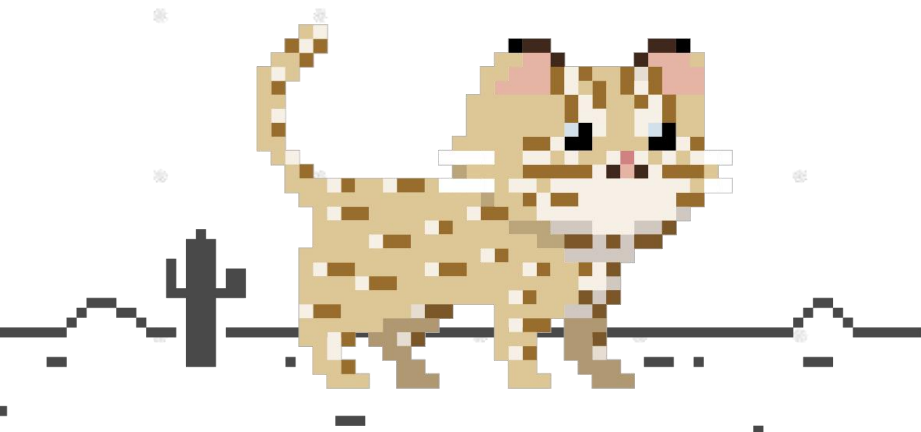




部署與監控模型

。持續監控模型在真實環境的運作狀況
。並持續確認關注問題狀況與使用者使用情形

如：持續透過工具追蹤及溝通模型的情境與資料





Model Card 工具包

輕鬆使用 Model Card 工具包建立模型卡。

[瞭解詳情 ↗](#)



機器學習中繼資料

記錄和擷取有關機器學習開發人員和數據科學家工作流程的中繼資料。

[瞭解詳情 →](#)



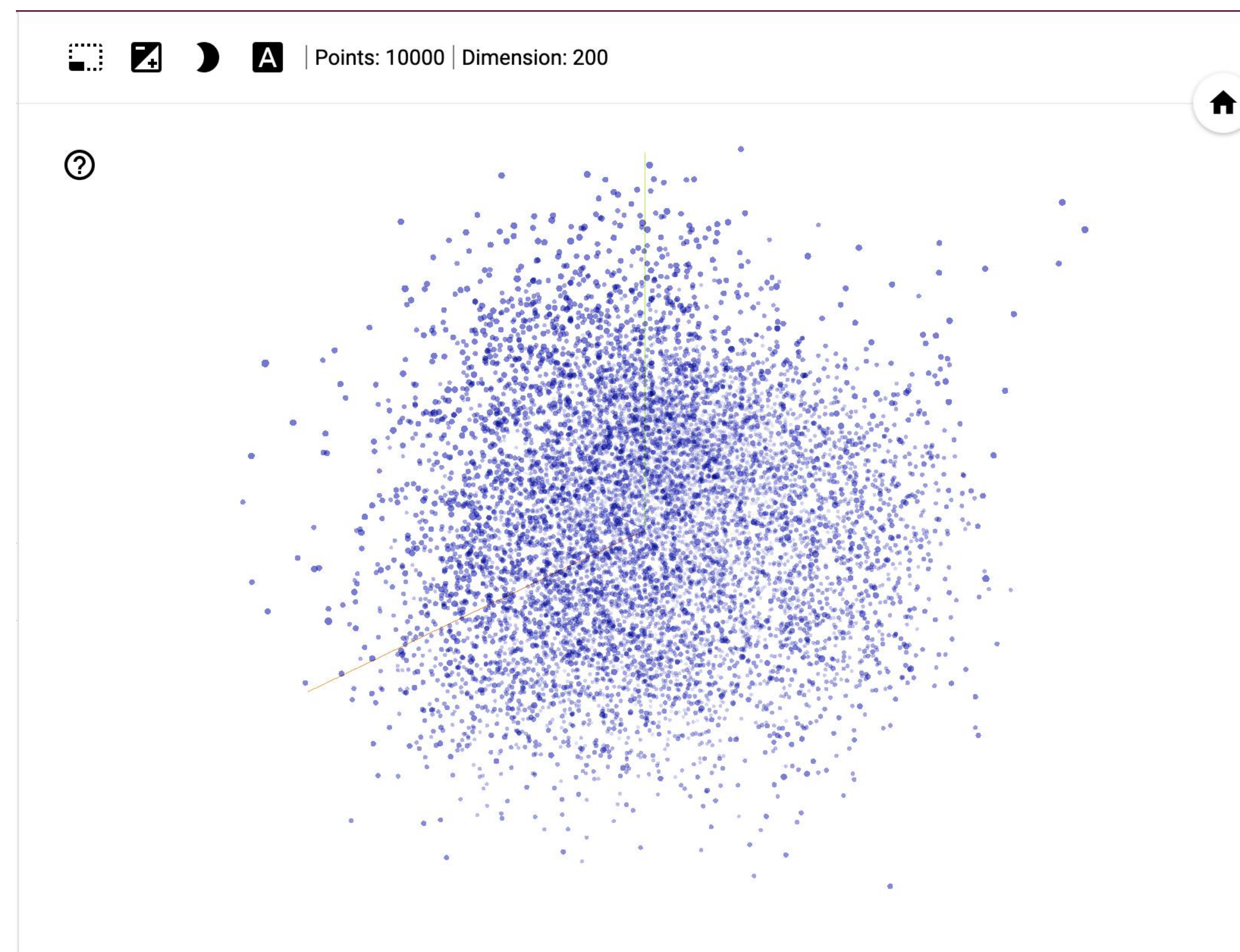
Model Card

採用結構化方法整理機器學習的重要內容。

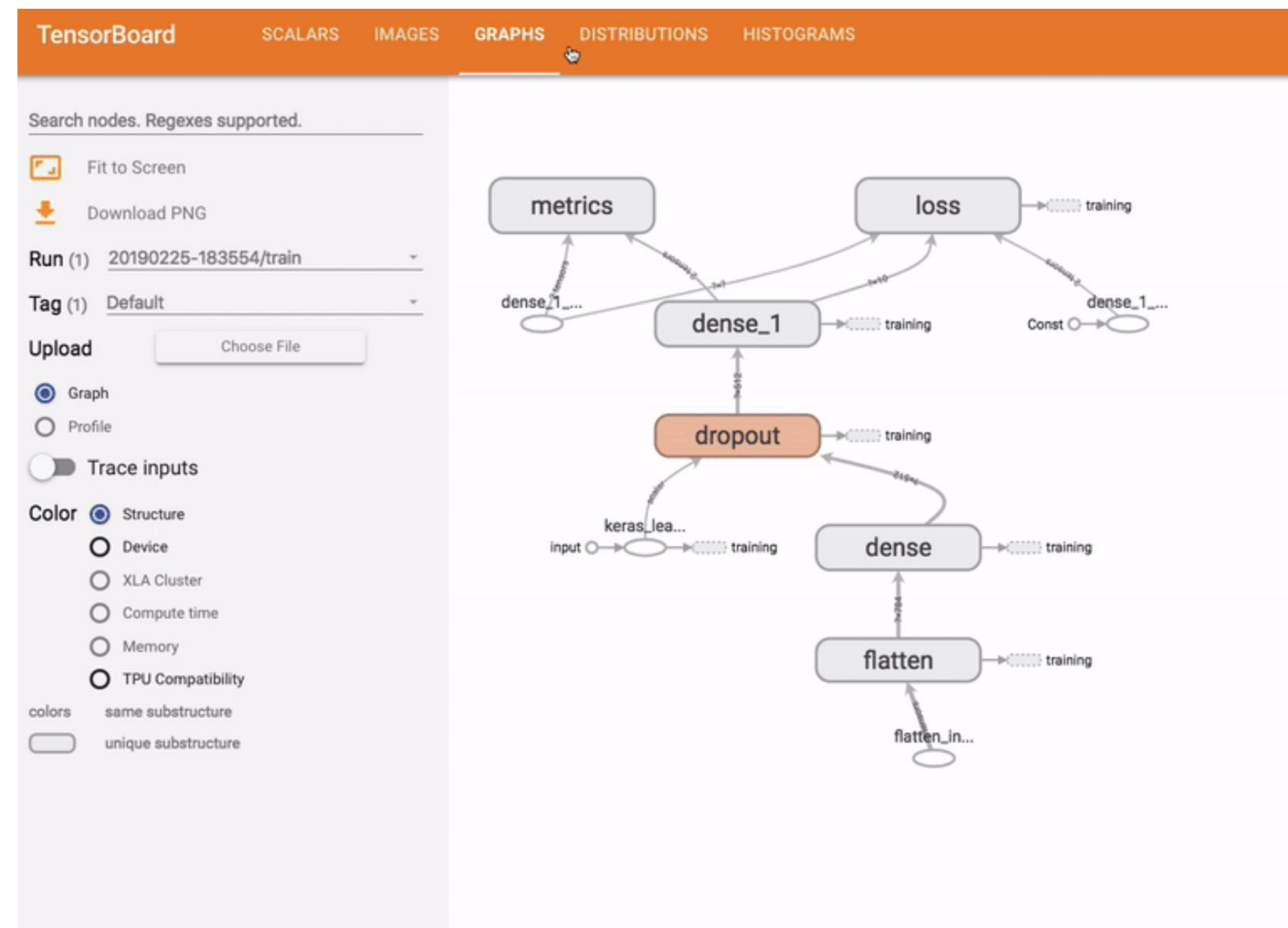
[瞭解詳情 ↗](#)



1. Embedding Projector



1.TensorBoard

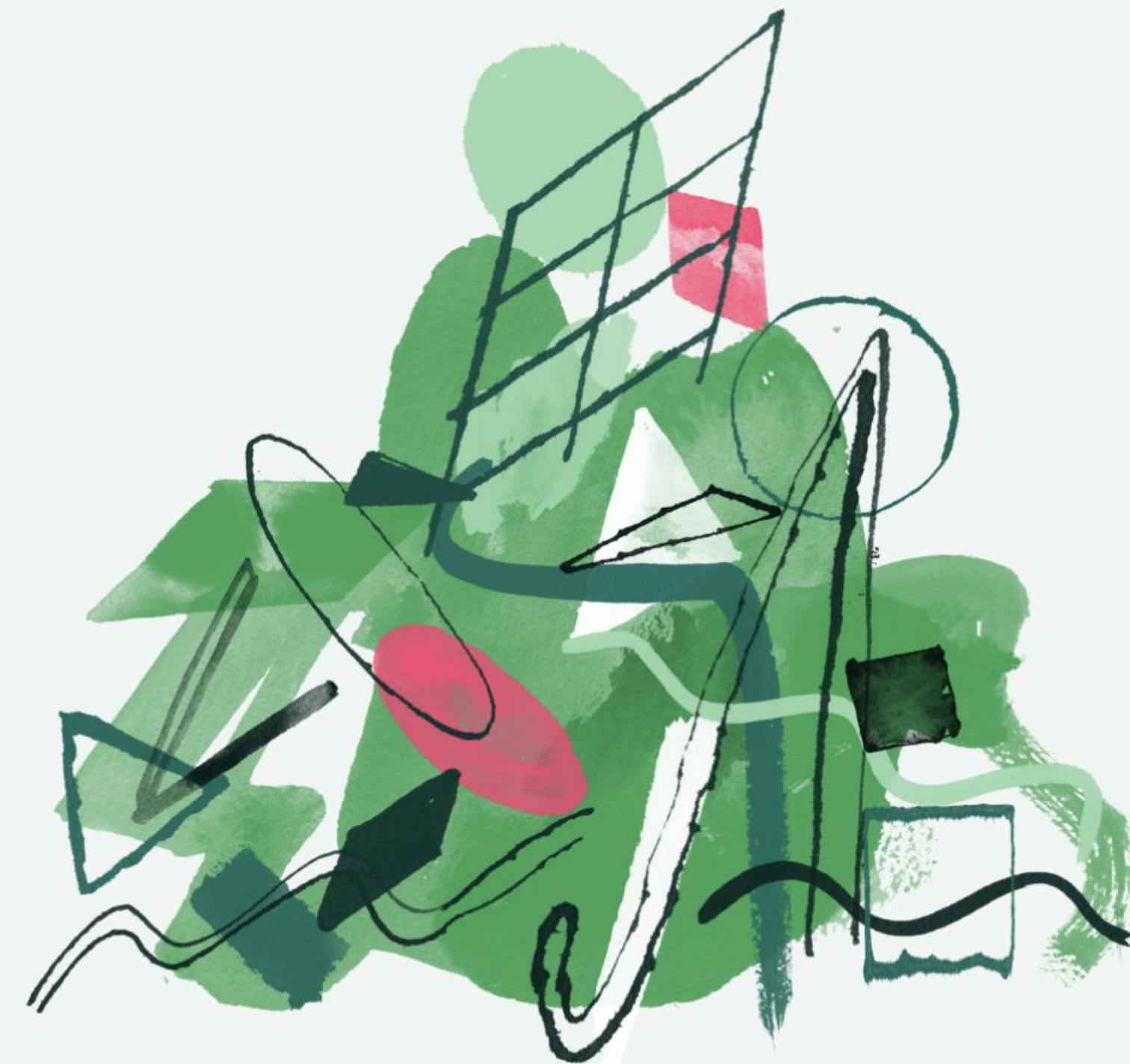


AI Explorables

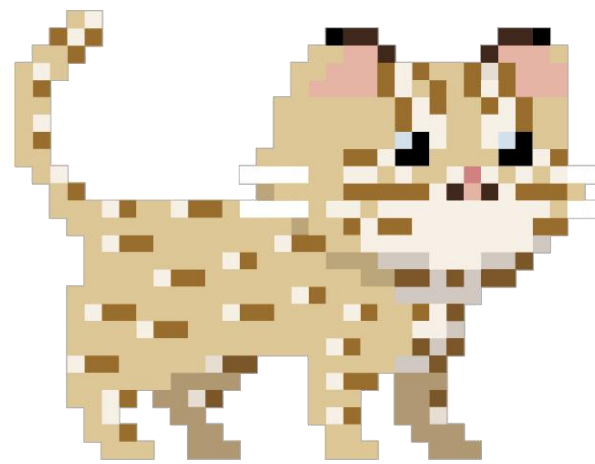
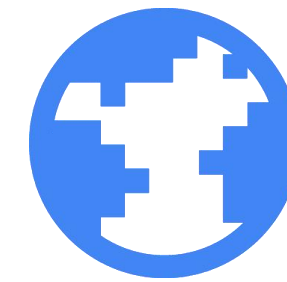
Big ideas in machine learning, simply explained

The rapidly increasing usage of machine learning raises complicated questions: How can we tell if models are fair? Why do models make the predictions that they do? What are the privacy implications of feeding enormous amounts of data into models?

This ongoing series of interactive, formula-free essays will walk you through these important concepts.



評估模型 What-If Tool





想更多了解機器學習與應用：

https://hiskio.com/courses/413/about?promo_code=LG282JG

想更多了解TensorFlow：

<https://tf.wiki/>

RAI：

<https://www.tensorflow.org/resources/responsible-ai?hl=zh-tw>



Jerry老師的Line群組

感謝聆聽

jerry@ap-mic.com

