LIONBRIDGE

AI 資料收集個案研究

如何在一週內收集 28,000 筆資料

1 週	8,000	28,000
交付時間	額外交付的資料集數目	交付的影像數目

客戶簡介

Lionbridge AI™ 協助的這個客戶，是間媒合品牌與其專案所需之創意專業人才的公司。

該公司設置並營運一個創新的平台，協助創意人才與需要創意人才執行專案的品牌搭上線。創意人才可以刊登他們的作品集並設定條件以供搜尋；品牌方則可張貼工作或專案摘要。總的來說，這個 Lionbridge 客戶致力於促進品牌與創意人才間的聯繫與交流，雙方可以透過他們的平台聯絡，為目前或日後的專案建立合作關係。

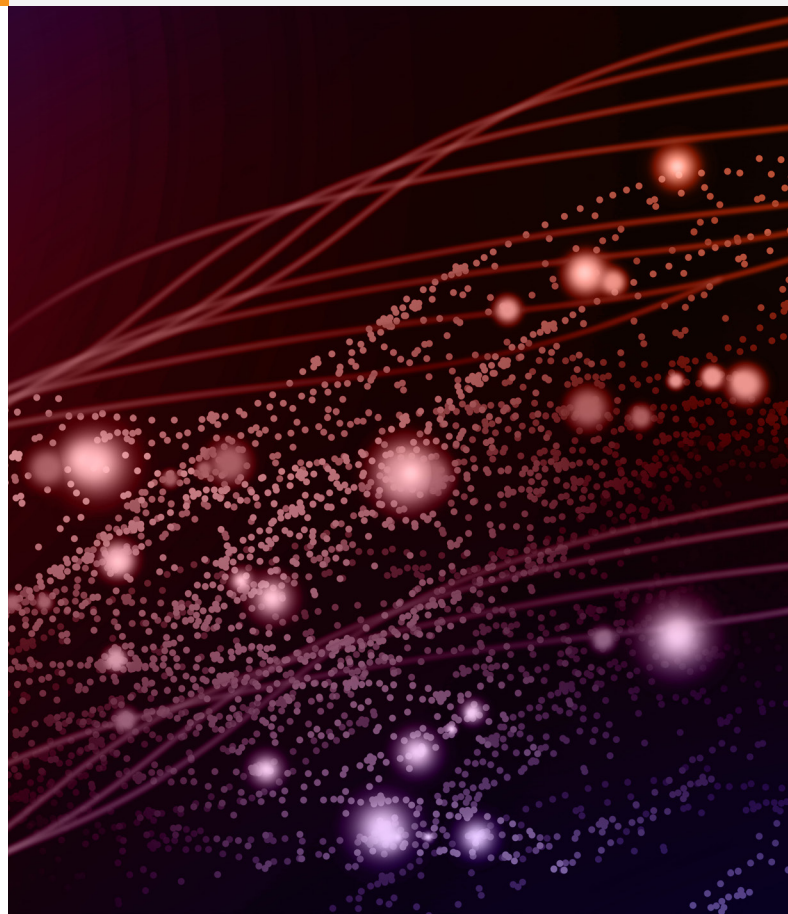
面臨的挑戰

Lionbridge 的客戶需要在非常緊迫的時程內，取得大量的資料為他們的 AI 模型進行物件辨識訓練。客戶要求要在數天內取得約 20,000 個影像，而且是高品質的人類資料，而非合成影像(後者的品質往往太低，可能會導致模型效能明顯低落)。客戶尤其指定他們需要的是物件中還有物件的影像，這會有助 AI 模型學習辨識那些呈現位於其他物件中的物件影像，像是籃子、包包、碗，諸如此類。他們要求一個資料集內要有四種影像：

- 各個物件各自的影像
- 一個物件在另一個物件中的影像
- 不含任何物件的空背景影像

另外值得注意的是，客戶也以來自不同觀點的資料為優先。以同一類物品的 20,000 張影像訓練模型，當然比不上以 20,000 張不同影像進行訓練的結果。從全球各地許多不同人口組成的參加者取得的影像，亦會有助模型理解資料，進而為所有人口組成的受眾提供更好的服務。

來自不同國家/地區、年齡層、性別等的人們，往往能為客戶網站帶來許多不同種類的影像。而以更多元的資料集來訓練網站的模型，也就能為更廣泛的客群服務。



我們眾包服務的一些統計數據

使用的語言超過 **350** 種

居住在遍佈全球超過 **450** 個地點

- 北美：**26%**
- 歐洲：**29.5%**
- 亞洲：**27%**
- 非洲：**6.5%**
- 南美：**8.5%**
- 大洋洲/澳洲：**2.5%**

- 女性：**52%**
- 男性：**48%**

年齡層：**18 到 60 歲以上**

從事的產業：

- **16%** 商業
- **16%** 工程
- **15%** 人文
- **13%** 科學
- **9%** 技術
- **4%** 電子商務
- **2%** 人力資源 (HR)
- **2%** 醫療保健
- **2%** 藝術
- **2%** 法律
- **10%** 其他

解決方案

為了服務這個客戶，我們使用了自有的 Lionbridge Aurora AI Studio™ 平台。透過這個平台，我們能與五十多萬遍佈全球各地、組成多元的眾包人才聯繫。這些人才無論在地點、使用的語言、年齡、性別以及從事的產業等方面都非常多元。我們會邀請他們一同合作，進而以前所未見的速度完成包括資料收集等許多工作。

針對這個專案，我們在 Aurora AI Studio 中為客戶建立了工作，並在數小時內吸引了超過 2,000 名參加者加入。參加者們有八個小時的時間完成資料收集工作。同一時間，我們亦建置了規模較大的內部品管團隊來審閱這些資料。之所以要建立較為大型的團隊，是因為資料量龐大同時專案時程又非常緊迫。這個團隊審閱了我們收集到的資料，確認它們正確無誤並符合客戶的需求。完　　成品質審閱後，我們也為每個影像命名其中繼資料，替客戶節省了許多小時的人力。

Lionbridge 將收集到的資料分三個批次交付給客戶。我們提早將第一批資料集交付給客戶，以便他們能盡快開始訓練模型。這一批資料含有三天內收集到的 12,000 個影像。第二批則是在五天內交付，內含剩餘的 8,000 個影像。

由於客戶對收集到的這兩批 AI 訓練資料非常滿意，因此又要求我們再以兩天的時間額外提供 8,000 個影像。我們欣然從命，總共交付了 28,000 個影像給客戶，而且在短短一週內全數完成。最初所要求的 20,000 個影像，則只花了四天就完成交付。

方法

Lionbridge 採用了由多個步驟組成的流程來支援客戶：

步驟 1：在 Aurora AI Studio 中建立工作

步驟 2：吸引參加者加入

步驟 3：接受收集到的資料

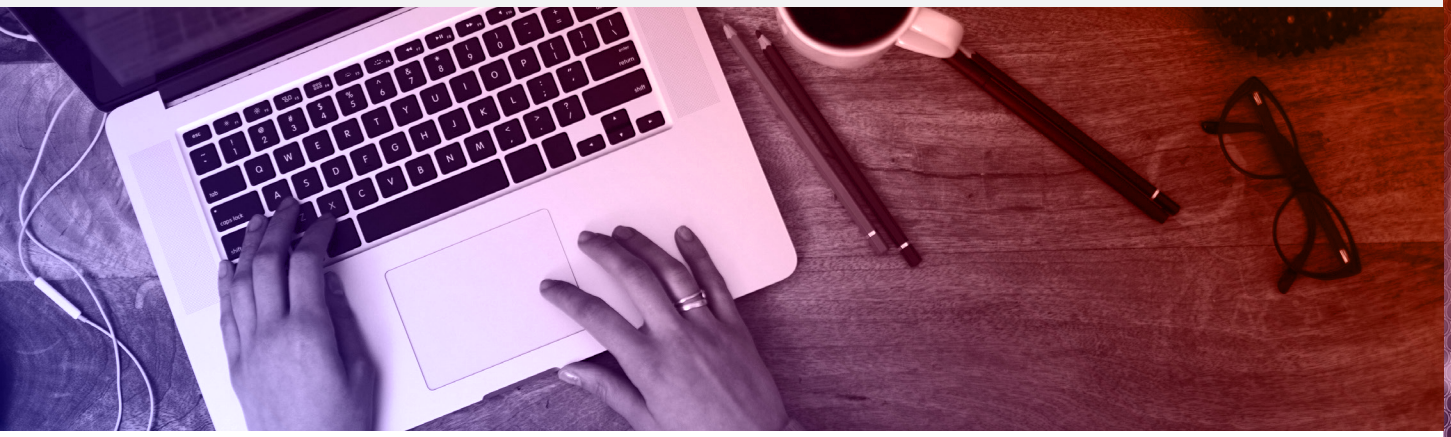
步驟 4：審閱資料並命名其中繼資料

步驟 5：將收集到的資料交付給客戶

步驟 6：根據客戶的要求，收集並交付更多資料

跨領域團隊

方案主管 | 技術 AI 工程師 | QA 組長 | 專案經理 | 10 名內部 QA 專業人員



成果

Lionbridge AI 主要是以三個方法來協助這個客戶。首先，運用我們的 Aurora AI Studio 平台以及極為多元的眾包人才，我們累積了非常大量由人類收集、能反映全球各地之觀點和影像的高品質資料。這麼做可確保客戶能避免完全仰賴合成資料所導致的後果，因為這可能會進一步造成模型訓練成效不佳。Lionbridge 甚至籌組了規模較大的品管團隊，驗證過資料後再進行交付。這些無可挑剔的資料將可協助客戶訓練其模型，進而能熟練順暢地提供品牌及創意人才所需的支援，不會受限於所在地點、人口組成或作品集的內容。

我們協助客戶的第二個方法，是在一週內便交付了非常大量的資料 (事實上甚至比客戶原本要求的數量還多)。快速的交付時程對訓練模型而言非常重要，因為每多耽誤一天都得付出高昂的代價。Lionbridge 運用自有的 Aurora AI Studio 平台極其快速地完成這個專案，使客戶得以在人力以及業務損失成本上省下數千美元。他們亦可確保自己平台的模型能在短短一週內便獲得足夠的訓練。

我們讓客戶大感驚艷的第三個方法，就是提供超越他們最初要求的服務。客戶原本要求 20,000 個影像，我們最終取得了 28,000 個。Lionbridge 對此深感自豪，因為我們不但提供了客戶所需的資料，更交付了他們原本沒意識到會想要的資料。能夠在短短一週內，滿足客戶所需的一切甚至更多，著實是很大的成就。

如需深入了解，歡迎造訪
LIONBRIDGE.COM