前言

任何足夠先進的技術都與魔法無異。

—亞瑟 • 克拉克（Arthur C. Clarke）

您的工作、職業生涯或日常生活可能已經或即將受到人工智慧（artificial intelligence，AI）的影響。這本書將幫助您建立並提升對 AI 的關鍵部分：訓練資料的概念和對機制的理解。

您的生活真的會受到影響嗎？先來測試一下。您是否從事科技行業，正在開發軟體產品？您的工作或公司的產品是否有任何形式的重複性任務？即您或您的產品使用者會在一定週期內做的事情？只要這些問題中有一個答案是肯定的，就表示 AI 和機器學習（machine learning，ML）有潛力承擔更多的工作量，使您或您的使用者能夠專注於更具大方向的工作，從而影響到您。如果您想跟上這波新的 AI 浪潮，本書將介紹許多讓 AI 在實務上真正可行的細節，將有助於讓您目前的工作更為成功，並讓您得以扮演以 AI 為中心的新角色。

說到工作，您也知道剛開始新工作的頭幾天或幾週情況：壓力大、瘋狂、不可預測，對吧？然後突然間，工作還有所有日常事情物進入正軌並有其意義，再難以想像的事情都一成不變，因為您已學會融入及適應。在還算短的時間內，您就從那個把咖啡灑在老闆襯衫上的人，成為整個系統有具有生產力的一部分。

AI 的工作方式也類似。不同的是，AI 的老闆就是您！從一開始和之後的訓練都是您的責任。就像新團隊成員一樣，首次訓練 AI 時，會有不可預測的結果。隨著時間過去和更頻繁的訓練與監督，它就會越來越好。事實上，這種變化發生得如此之快，以至於顛覆一切可行和不可行的自動化假設。所有 AI 系統，從自駕車到農業雜草偵測、醫學診斷、安全、運動分析等等，都需要這種概論式的監督。

我將揭開 AI 最基本層面的神祕面紗：將對人類有意義之事，轉化為 AI 可讀的形式，也就是所謂的訓練資料（training data），從生成式 AI 到完全監督系統的一切，這都非常重要。我會幫助您理解圍繞著訓練資料的各種表達法和概念，涵蓋它在實務上的運作方式，包括運算、工具、自動化和系統設計；同時，也會將這一切與實用案例研究和建議結合起來。

您的知識是使 AI 能夠運作的魔法，AI 則能擴展您的影響力，做出更多創造性工作，增強知識效能。在學會訓練 AI 後，您就是受益者。

本書目標讀者

這本書是訓練資料的基礎概覽，非常適合剛接觸或剛開始瞭解訓練資料的人閱讀。

至於中階實務工作者，後面章節會提供他處所沒有的獨特價值和見解；簡而言之，就是內行人才知道的知識。我將強調主題專家、工作流程管理者、訓練資料負責人、資料工程師和資料科學家感興趣的特定領域。

此處不需要電腦科學（computer science，CS）知識，但瞭解 CS、機器學習或資料科學，會更易於理解大部分章節。我也會盡最大努力，讓這本書對資料標註者，包括那些主題專家來說易於理解，因為他們在訓練資料中方面扮演著關鍵角色，這也包括監督系統。

對技術專業人士和工程師來說

您可能一直在尋找發布或改善系統的方法，或者已經找到新的 AI 功能，且希望將其應用於您的領域。這本書將引導您完成這些過程，並解答更細的問題，例如應該使用的媒體類型、配置系統方式，以及重要的自動化措施等。

作法百百種，本書旨在提供平衡的涵蓋範圍、凸顯權衡取捨、並成為您訓練資料需求的主要參考資源。一直在更新的新概念有的時候會讓人覺得制式且嚴肅，因此，可以的話，我會盡最大的努力，讓這本書的風格帶點一般公開文件少有的隨意與輕鬆。

如果您是一位專家，這本書可以作為參考、知識更新的來源、並幫助您向團隊中的新成員傳達核心概念。如果您已在這一領域已擁有一些知識，但仍想通盤性的確認，這本書也能擴展您的方法工具包，並為常見觀點提供新的視角。如果您是十足的菜鳥，這將是入門的最佳資源。

對經理和主管來說

簡單說，這本書會提供您無法從其他地方獲得的內容，它增加新穎的獨特性和密集的背景知識，將幫助您及團隊獲得洞察，可能讓您提前數月甚至數年達成目標。

此外，這本書的一大重點在致力於人員和流程。訓練資料提出了新穎的人機互動概念，並涉及跨學科互動的各種層次，而它們將為您在這個令人興奮的 AI 領域成功，提供寶貴的新見解。

第 6 章〈理論、概念與維護〉、第 7 章〈人工智慧轉型與應用案例〉和第 9 章〈案例研究和故事〉對您來說尤其重要。其餘章節則將幫助您熟悉能夠識別成功和失敗的細節，並有助於路徑修正。

對主題專家和資料標註專家來說

標註者（annotator）是訓練資料的日常生產中最關鍵角色之一。2020 年世界經濟論壇報告指出，需求增加的前三大工作角色都涉及資料分析和 AI[1]。找到處理訓練資料的方法，是您現在應該增加的一種有價值技能，也是一個新的職涯機會。

雇主要求所有員工瞭解 AI 甚至是訓練資料的基礎知識已越來越常見。例如，一家大型汽車製造商要求應徵資料標註者的申請者必須：「瞭解我們的學習演算法使用標籤方式，以便更能判斷那些艱澀的邊緣案例。」[2] 不管您的行業或背景如何，若能將您和團隊的知識奠基在訓練資料上，都將有很大的機會來擴展知識範圍，和提升公司生產力。

雖然任何人都可以監督他們熟悉的領域，但像醫生、律師和工程師這樣的主題專家（subject matter expert，SME）尤其具有價值。主題專家既可以直接監督 AI，也可以提供詳細的指令和訓練，以達成更有效的資源利用。如果您是一位主題專家，閱讀這本書至關重要，甚至更要仔細地瞭解您的工作融入 AI 大局的辦法、可供使用的工具和槓桿，以及設置其他人可以遵循的流程方法。

這本書除了提供標準素材，例如詳細指令之外，同時也會提供對經過測試機制的洞察，例如一個稱為綱要（*schema*）的概念。透過閱讀這本書，您將深入理解建立和維護有效 AI 系統所需的一切知識。

1 〈The Future of Jobs Report 2020〉（*https://oreil.ly/m6uXd*），《World Economic Forum》，2020 年 10 月（第 30 頁，圖 22）。

2 「Data Annotation Specialist」，特斯拉（Tesla）網站，2020 年 11 月 5 日存取。

對資料科學家來說

作為資料科學家，您扮演著擔任他人顧問的重要角色：幫助他們理解實際使用資料的方式。即使是最先進且整合的 AutoML 系統，通常也需要有人來解釋和理解它們的輸出意義，並能夠在出現問題時除錯。這本書將幫助您與不同的標註和技術合作夥伴有更好的互動。

任何資料都可以訓練或視為訓練資料。「apple」指的是水果，但「Apple」是一家公司，就跟許多名詞一樣，訓練資料也有多重意義。這本書聚焦於監督式（supervised）訓練資料，指的是由人類直接參與豐富化資料。雖然標註的細節可能並不總是與您的日常工作相關，但更廣泛的理解，可以進一步確保達到最佳的最終結果。

為了設下期望，這本書會聚焦於現代訓練資料，特別是至少涉及人類角色的監督式系統。即使一般認為非監督式（unsupervised）的生成式 AI 背景下，人類對齊（alignment）也扮演著關鍵角色。雖然關於監督式、自監督式（self-supervised）、半監督式（semi-supervised）、非監督式等概念的界限或實用性仍不斷改變，但似乎很多實際應用案例可以透過一定程度的監督而達成，而且某種形式的監督可能還會存在很長一段時間。

閱讀時，請考慮以下主題。如何更深入地與標註和技術合作夥伴互動？如何參與包括建立和維護的資料蒐集過程？如何協助將建模需求與綱要對齊，反之亦然？如何幫忙確保對模型來說，這是最好的訓練資料？如果要從這本書獲得一個重要的收穫，我希望它能讓您以新視角來看待「資料標註」，也就是在其自身技術領域中，所謂的訓練資料。

我寫這本書的原因

在與 Diffgram 合作時，我注意到那些「懂了」和不懂的人之間，有著非常大的差距，這就好像觀察到某人在不知道數字系統存在的情況下，仍試圖學習乘法一樣。他們搞不懂訓練資料最基本的基礎；但糟糕的是，他們常常不知道自己搞不懂！

一開始，我只是寫一些相對簡短、最多只有幾頁的短文，主題也很局限。但這些文章能幫忙填補知識缺口，儘管我只是在自己的小領域，分享我剛好知道的事物，但仍然感覺有許多部分缺漏。我需要寫些更全面的東西，一本書聽起來是個合理的選擇；但是，我有什麼資格寫書？

開始寫這本書時，我有很多疑慮。我已經在這個領域工作了大約 3 年，但我仍然覺得自己計畫要寫的某些素材是「理想化」的目標，而不僅僅是總結已經知道的內容。回顧這 5 年，我今天寫這小節時，仍然覺得只是略微觸及這個領域的表面而已。

然而，此時此刻，我不得不回顧並明白，在我所知道的人中，很少人在擴大業務時，仍然能像我這樣深入理解技術。這表示我是少數具有這種特性的人：深入理解這個領域的技術、瞭解其進步歷程、能夠用非工程師的術語解釋這些主題，並且有意願花時間記錄，和將這些知識分享給他人。

我真的相信，訓練資料是存在已久的技術領域中，最重大的概念性轉變之一。監督式訓練資料橫跨每個行業和幾乎每個產品，在接下來的幾十年裡，我相信它將以今日幾乎無法想像的方式塑造我們的生活，希望這本書能在您的旅程中提供幫助。

本書架構

首先，我將介紹訓練資料的用途、使用機會、它很重要的原因，以及實際應用中的訓練資料，即第 1 章〈訓練資料導論〉。實際專案需要訓練資料工具，而實際使用它們時將有助於理解概念，為了開始，第 2 章〈快速上手〉將提供一個動手和開始工作的框架。

一旦掌握大方向概念和工具，就該談談綱要，也就是對所有商業知識編碼的範式。綱要是訓練資料最重要的概念之一，因此詳細的處理方式，能真正有助於建立這種理解，見第 3 章〈綱要〉。接下來是第 4 章〈資料工程〉和第 5 章〈工作流程〉，這些關鍵的工程概念能將系統建立起來，並投入生產。

接著過渡到概念和理論：第 6 章〈理論、概念與維護〉、AI 轉型：第 7 章，〈人工智慧轉型與應用案例〉，和第 8 章〈自動化〉，並以實際案例研究作為結束：第 9 章〈案例研究和故事〉。

主題

本書分為 3 個主題，如下所示。

基礎和入門

瞭解訓練資料的重要之因及其內容，掌握基本術語、概念和各種表達法形式。我從監督式與傳統機器學習方法之間的相似性和差異開始設置背景，然後解析所有關於抽象化、人員、流程等方面的內容。這是最基本的基礎。

概念和理論

此處會更具體地研究系統和使用者操作，以及流行的自動化方法。這裡稍微偏離基礎，以擴展到不同的觀點。

綜合應用

把基礎性和理論的需求放在心上，探索具體實作，進一步擴展趨勢來涵蓋尖端研究主題和方向。

關於本書的術語，您偶爾會看到同時使用訓練資料和 AI 資料這兩個詞。AI 資料是指 AI 所使用的任何類型資料，所有訓練資料也都是 AI 資料。

我比較偏向以舉例方式讓內容更容易理解和記憶，除非有其必要性，我都會刻意避免技術性術語。如果您是專家，請忽略已經熟悉的任何內容；至於非專家，請考慮到許多技術細節就只是細節，能有助於理解，但並非必需。

我的目標是盡可能專注於監督式訓練資料，這也包括會稍微涉獵到深度學習和機器學習知識，但通常這些都超出本書範圍。訓練資料是一個跨行業的通用概念，同樣適用於許多行業。所呈現的大多數概念同樣適用於多個領域。

儘管我有親身經歷 ML 和 AI 的演進，不過這不是一本歷史書；我只會引用一些發展過程來解釋當前主題。

圍繞訓練資料建構的軟體引入各種假設和限制。我試圖挖掘隱藏的假設，並強調在特定圈子中廣為人知，但對大多數人來說卻是全新的那些概念。

本書使用慣例

本書使用以下排版慣例：

斜體字（*Italic*）

表示新的術語、URL、電子郵件地址、檔名和延伸檔名。

定寬字（`Constant width`）

用於程式列表，以及在段落中參照的程式元素，例如變數或函數名稱、資料庫、資料型別、環境變數、敘述和關鍵字。

第一章

訓練資料導論

資料無處不在，包括影片、影像、文字、文件，以及地理空間和多維資料等。然而，原始資料對於監督式機器學習（ML）和人工智慧（AI）來說幾乎沒有多大的用處。我們要如何利用這些資料？如何記錄智慧，使其能透過 ML 和 AI 複製？答案就是訓練資料的藝術，也就是讓原始資料變成有用的一門學問。

您將在本書中學到：

- 全新的訓練資料（AI 資料）概念
- 訓練資料的日常實務
- 如何提高訓練資料效率的方法
- 轉變團隊，使其更加專注於 AI ／ ML 之辦法
- 實際案例研究

在探討這些概念之前，首先需要瞭解基礎知識，本章將就此解析。

訓練資料涉及將原始資料塑形、改造、整理和消化成新的形式：從原始資料中創造全新意義以解決問題。這些創造和破壞的行為位於專業知識、商業需求和技術要求的交叉點，是橫跨多個領域的多元化活動。

這些活動的核心是標註（annotation）。標註會產生機器學習模型所需耗用的結構化資料，沒有標註，就等於這些原始資料是非結構化的，通常價值較低，且經常不適用於監督式學習。這就是為什麼訓練資料對於現代機器學習的應用場景，包括電腦視覺（computer vision）、自然語言處理（natural language processing）和語音辨識（speech recognition）等來說，是必不可少的。

為了具體說明這個觀點，以下將詳細審視標註過程。標註資料時，實際上是在捕捉人類的知識，過程通常如下：一個媒體元素，例如影像、文字、影片、3D 設計或音訊，會伴隨一組預定義的選項，即標籤（label）而呈現。人類將審視這些媒體並決定最適當的答案，例如，指出影像中的某個區域是「好」或「不好」的，這個標籤提供了應用機器學習概念所需的背景（如圖 1-1）。

但要如何達到這個階段呢？如何做到在正確時間，向正確的人展示具有正確預定義選項集的正確媒體元素？有許多概念會引領我們直達標註或知識捕捉實際發生的那一刻，而正是這些概念整體構成訓練資料的藝術。

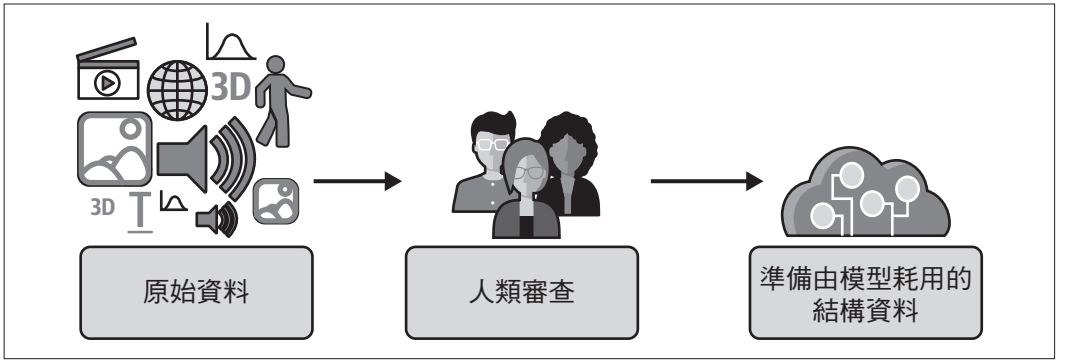

圖 1-1　訓練資料過程

本章將介紹訓練資料的內容、重要之因，並深入探討許多關鍵概念，好為本書其他部分打下基礎。

訓練資料的意圖

訓練資料的目的因不同的應用案例、問題和情境而異。以下來探索一些最常見的問題，例如您可以用訓練資料做什麼事？它和什麼息息相關？一般人用訓練資料是想達成什麼目標？

可以用訓練資料做什麼事？

訓練資料是 AI ／ ML 系統的基礎，是讓這些系統運作的支柱。

透過訓練資料，可以建立和維護現代 ML 系統，例如用來建立下一代的自動化程序、改善現有產品、甚至創造全新產品。

人類角色

人類可以透過控制訓練資料來影響資料程式。這包括決定到目前為止討論過的層面：綱要、原始資料、品質，以及與其他系統的整合。當然，當人類一一審視個別樣本時，也就參與了標註本身。

從建立初始訓練資料，到執行對資料科學輸出的人工評估，以及驗證資料科學結果，許多經手的人和階段都可見到這種控制的實現。所涉及的大量人員與傳統機器學習截然不同。

會有新的度量（metric），比如接受多少樣本、每項任務花費的時間、資料集的生命週期、原始資料的準確度或綱要的分布情況等等。這些層面可能與資料科學術語，如類別分布重疊，但也可視為單獨的概念來思考。例如，模型度量的基準是訓練資料的基本真實情況，所以如果資料錯誤，度量也會錯誤。正如第 255 頁「品質保證自動化」小節的討論，圍繞像標註者一致性這類事情的度量，可能會忽略綱要和原始資料範疇的更大問題。

人類監督（human supervision）所牽涉的遠遠超過量化度量，也關於質性理解。人類觀察、還有對綱要、原始資料、個別樣本等的人類理解非常重要，這種質性視角會延伸到商業和使用案例概念上。此外，這些驗證和控制，很快就會從容易定義的形式，轉變為偏向藝術形式或創作行為，更不用說那些圍繞著系統效能和輸出所可能產生的複雜政治和社會期望了。

處理訓練資料是一個創造的機會：以新穎的方式捕捉人類智慧和洞見，在新的訓練資料情境中定義問題，建立新的綱要、蒐集新的原始資料、並運用其他針對訓練資料的特定方法。

這樣創造與控制都前所未見，雖然已經建立各種人機互動樣式，但相較之下，人類與機器學習程式的互動則建立得較少，也就是對人類監督而言為資料驅動（data-driven）系統，讓人類可以直接修正資料和對資料進行程式設計。

例如說，一般都預期辦公室工作人員會使用文字處理軟體，但不會期望他們使用影片編輯工具。訓練資料需要專業知識，因此，就像今日的醫生也要會用電腦看診一樣，他們現在也必須學習使用標準標註樣式。而這隨著由人類控制的資料驅動程式出現會更為普遍，將持續增加這些互動的重要性和變化性。

流程改善機會

考慮一些人們想要改善的常見領域，例如：

- 標註品質差、成本過高、過於手動、錯誤率高
- 重複工作
- 主題專家的勞動成本過高
- 過多例行或乏味的工作
- 幾乎不可能獲得足夠的原始資料
- 原始資料量明顯超過任何合理的手動審查能力

您可能希望有更廣泛的業務轉型、學習新工具，或優化特定專案或流程。這自然會產生一個問題，接下來應該採取哪項最佳步驟，又為什麼要採取這一步？為了回答這個問題，先來談談訓練資料的重要性。

訓練資料的重要性

本節將介紹訓練資料對您的組織重要性，以及強大的訓練資料實務之必要性。這些是貫穿全書的核心主題，會一再出現。

首先，訓練資料決定了您的人工智慧程式和系統能力，沒有訓練資料，就沒有系統；有了訓練資料，機會只受限於您的想像力！好吧，實際上還有預算、硬體等資源，以及團隊專業知識的限制。但理論上，任何能形成綱要並用以記錄原始資料的事物，系統都可以重複。概念上，模型可以學習任何事物，這表示系統的智慧和能力取決於綱要的品質，還有您能教給它的資料數量及多樣性。實際上，有效的訓練資料在預算、資源等其他方面都平等的情況下，可以為您提供關鍵的優勢。

其次，訓練資料工作是位於資料科學工作之前的上游階段，這意味著資料科學依賴於訓練資料。訓練資料中的錯誤會流向資料科學，或者套句俗話，資料爛，結果就爛。圖 1-3 展示了這種資料流在實務工作中的樣子。

第三，訓練資料的藝術指的是建構人工智慧系統的思維轉變方式。與其過於關注改善數學演算法，不如與它們平行工作，繼續優化訓練資料以更匹配我們的需求。這是正在發生的人工智慧轉型的核心，也是現代自動化的核心。這是第一次，知識工作正在自動化中。

通常，原始資料會比可以標註的資料還多，因此，基本選擇過程的一部分是選擇將標註的原始樣本，通常有多個原始資料集；實際上，許多專案很快就會擁有數百個這樣的原始資料集。更一般地說，這也是整體組織和資料集結構的關注點，包括選擇哪些樣本該納入哪些集合。

樣本建立

現在會從頭開始探索，建立一個訓練資料的單一樣本方式，這將有助於建立對訓練資料核心機制的理解。

系統將會擁有一個已經在一組訓練資料上訓練過的深度學習模型，這些訓練資料主要包括兩個組成部分：

- 原始影像（或影片）
- 標籤

以下將討論幾種不同的方法。

用於草莓採摘系統的簡單綱要

想像一下正在開發一個草莓採摘系統，需要知道草莓的樣子、它在哪裡以及成熟度，這裡引入一些新術語，好更有效率地工作：

標籤

> 標籤（label），也稱為類別（*class*），[7] 代表最具大方向的意義，例如，一個標籤可以是「草莓」或「葉子」。對於技術人員來說，可以將其想像為資料庫中的一個表。通常附加到特定的標註（*annotation*）（實例）上。[8]

標註（實例）

> 圖 6-7 中顯示一個單一範例，與一個標籤相連，以定義它的內容；標籤通常還包含位置或空間資訊，定義某物的位置。沿用之前的技術舉例，就像資料庫中的一列。

7 其他名稱包括：標籤模板（label template）、標註名稱（annotation name）、類別（class）。

8 通常只透過參照 ID。

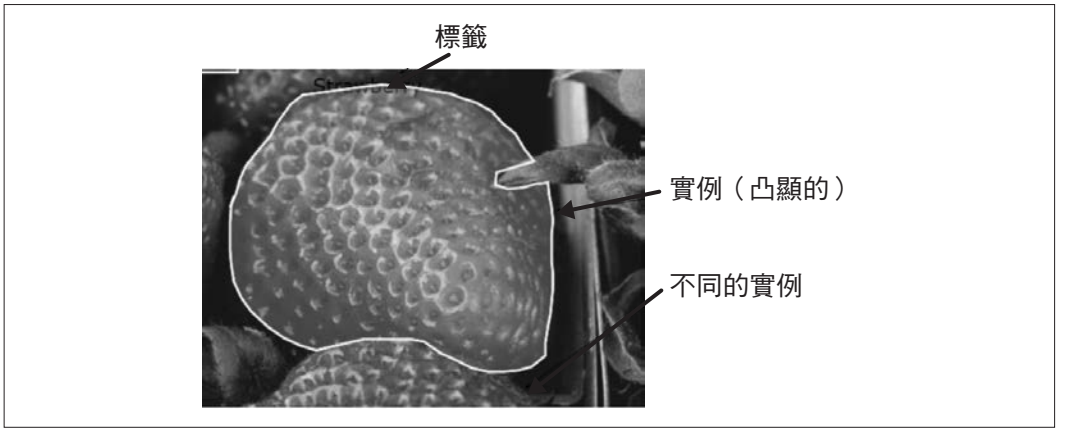

圖 6-7　標記和未標記的實例

屬性

一個實例可能有許多屬性（attribute），指特定實例獨有的特徵。通常，屬性代表物件本身的可選特性，而不是其空間位置。

這個選擇將影響監督的速度。單一實例可能有許多獨特的屬性；例如，除了成熟度，還可能有疾病識別、產品品質等級等。圖 6-8 就是一個範例。

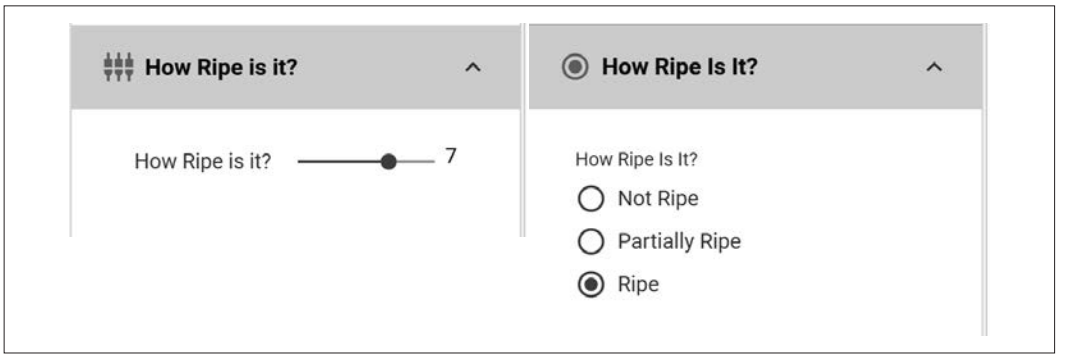

圖 6-8　顯示屬性選擇的使用者介面範例

想像一下，您只希望系統採摘特定成熟度的草莓，就可以將成熟度表達為一個滑塊，或者也可以有一個關於成熟度的多選擇選項，如圖 6-8 所示。從資料庫的角度來看，這有點類似於一個行（column）。

幾何表達法

也要選擇要使用的幾何表達法類型。這裡的選擇就像實際監督一樣，是訓練資料的一部分。

幾何形狀可以用來表達物體，例如，可以將圖 6-9 中顯示的草莓表達為方塊、多邊形和其他選項，其他章節有相關討論。

從系統設計的角度來看，可以選擇使用哪種類型的原始資料，如影像、音訊、文本或影片；有時甚至可以將多種模態結合在一起。

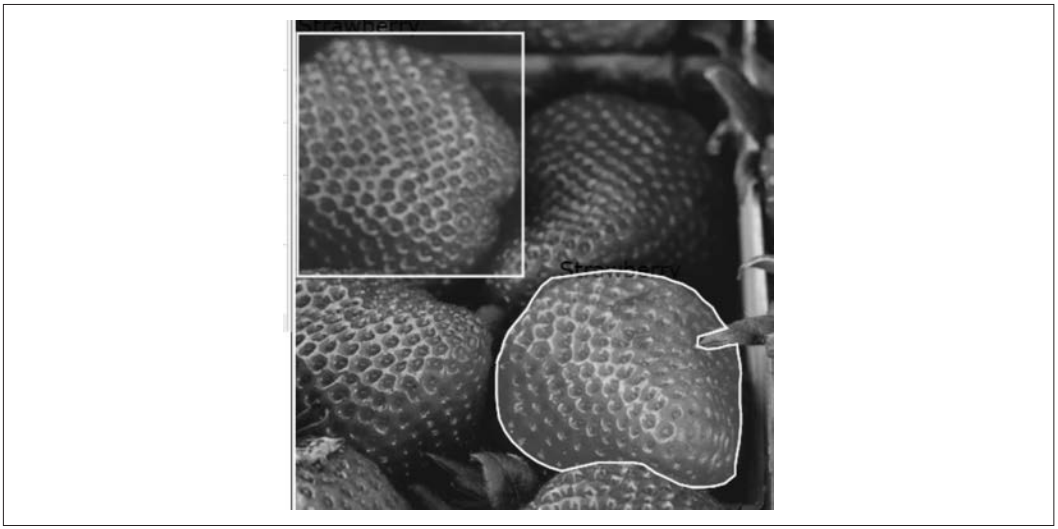

圖 6-9　帶有方塊和多邊形幾何形狀的標註範例

角度、大小和其他屬性在這裡也可能適用。

空間位置通常是單一的。現在，雖然物件在給定時刻一般只會在一個地方[9]，但在某些情況下，單一圖框可能擁有多個空間標註。想像一下，一個橫梁擋住對汽車的視線：為了準確標記汽車的區域，就可能需要使用兩個以上與同一標註相關的封閉多邊形。

9　這裡說的是在一個參考圖框中。多模態標註可以表達為一組相關實例，或給定參考圖框的空間位置，例如攝影機 ID x。

二元分類

一種基本的作法是偵測某物存在與不存在之間的差異。做到這一點的其中一種方式是二元分類（binary classification），有點像是問「照片中是否存在紅綠燈？」

例如，該集合中的兩張影像可能看起來會像圖 6-10：

圖 6-10　左圖，資料集中顯示紅綠燈以及背景中的樹木影像；右圖，資料集只顯示樹木的影像

要監督第一個範例，只需要兩件事：

- 捕捉與檔案本身的關聯：例如，檔案名稱是「*sensor_front_2020_10_10_01_000*」，這是與原始像素的「連接」。最後將讀取這個檔案，並將值轉換為張量，例如，位置為 0,0，指派 RGB 值。
- 以對我們有意義的方式來宣告它，例如：「Traffic_light」或「1」。這類似於說檔案名稱「*sensor_front_2020_10_10_01_000*」中存在一個「Traffic_light」。

第二個例子，可以宣告為：

- 「sensor_front_2020_10_10_01_001」
- 「No」或「0」

實務上，通常會有更多影像集合，不僅僅是兩個。

升級後的分類

為了從「存在」擴展到「有內容」，將需要多個類別，表 6-2 顯示這種布局的一個例子。

表 6-2　原始資料與相應標籤的視覺比較

樣本編號	原始媒體	標籤名稱	整數 ID
樣本 1		紅色	1
樣本 2		綠色	2
樣本 3		無	0

關於要使用字串還是整數，一般來說，大多數實際訓練將使用整數值；然而，這些整數值通常只有在附加到某種字串標籤時，才會產生意義：

```
{ 0 : "None",
  1 : "Red",
  2 : "Green"}
```

我會在這裡介紹標籤映射的概念，雖然這種類型的映射對所有系統來說都很常見，但這些標籤映射可能會變得更加複雜。同樣要記住的是，一般來說，「標籤」一詞對系統並沒有意義，但它會把 ID 映射到原始資料，如果這些值錯誤，可能就會嚴重失敗；更糟糕的是，想像一下，如果測試也依賴於同樣映射的話會怎樣！

這也就是有可能的話，最好「列印輸出」；這樣可以視覺上檢查標籤是否與所需的 ID 匹配。有一個簡單的例子，一個測試案例，可能會對一個匹配了 `string` 的已知 ID 進行 `assert`。

紅綠燈在哪裡？

繼續紅綠燈範例，先前方法的問題是不知道紅綠燈的位置。在機器學習建模中有一個常見的概念：*物件性分數*（*objectness score*），前文有提到過，還有其他更複雜的方法來識別位置。從訓練資料的角度來看，只要有一個定界框存在，就確定了*哪裡*，即空間位置，學習這一點的演算法實作則取決於資料科學。

維護

現在已經講解建立單一樣本的基礎知識方式，並介紹一些關鍵術語，可回到大方向的過程觀察。第 2 章曾講解設置訓練資料軟體、綱要和任務等基本事項，但是持續的維護行動真實面貌為何？

行動

這些是在從模型訓練過程傳回某種形式的資訊之後，通常可以採取的行動。

增加綱要深度以提高效能

提高效能的最常見方法之一是增加綱要的深度，一個例子是將標籤類別各個擊破，特別是表現不佳的類別；本質上，這既是要識別也是要改善最弱的特定類別。借用之前的例子，這裡使用「紅綠燈」標籤，當效能表現參差不齊時，可能不清楚哪些範例需要改善效能。

在審查結果時，會注意到「綠色」似乎在失敗案例中出現得比較頻繁，一種選擇是嘗試將更多綠色加入普通紅綠燈集合中；或者更好的方法是，將「紅綠燈」類別分割為「紅色」和「綠色」，這樣可以非常清楚看出哪一個表現比較好，可以重複這個過程，直到達到期望的效能，可透過如圖 6-12 所示分割大小來做到這一點。要實作這一點有一些細微之處和方法，但它們大抵不脫這個想法，關鍵在於，每次「分割」時，在效能方面就會更清楚該怎麼走下一步了。

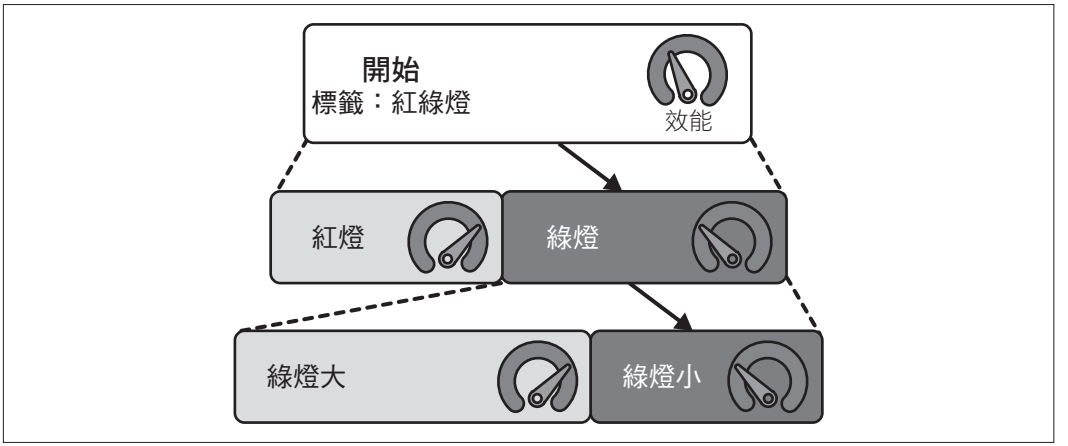

圖 6-12　將單一標籤基於效能需求而擴展為多個更具體屬性的改善路徑範例

進一步對齊空間類型與原始資料

想像一下，您一開始選擇了影像分割，然後，在意識到模型沒有按預期訓練時，可能可以簡單地切換到「更容易」的任務，如物件偵測，甚至是全影像分類。

或者，可能物件偵測產生了一堆重疊的方框，它們沒什麼用處，而您需要切換到分割以準確捕捉意義，如圖 6-13 所示。

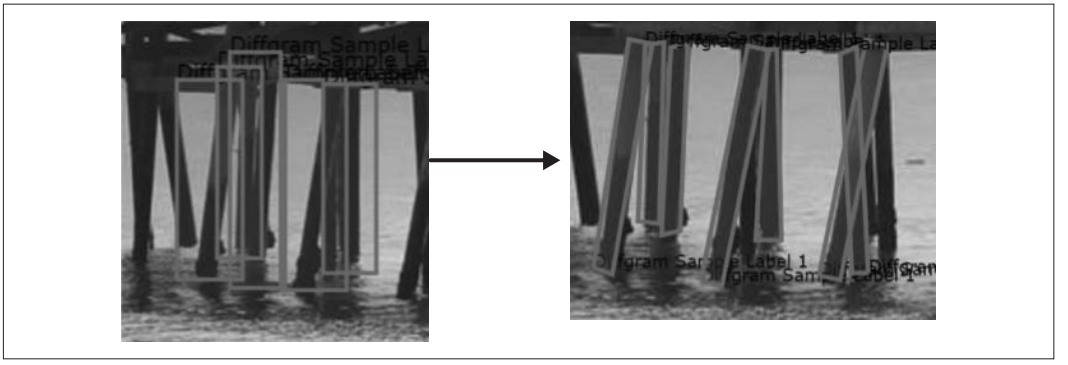

圖 6-13　方框無法提供有用資訊的範例，故切換到分割以獲得更好的空間結果

圖的左側從方框開始，導致了重疊的偵測結果；透過轉移到分割，可以獲得更清晰的分割，如右圖所示。儘管在某些情況下，有些方法較不理想似乎很明顯，但最佳方法往往沒有那麼明確。