天文

中華民國107年　秋季號

機器學習

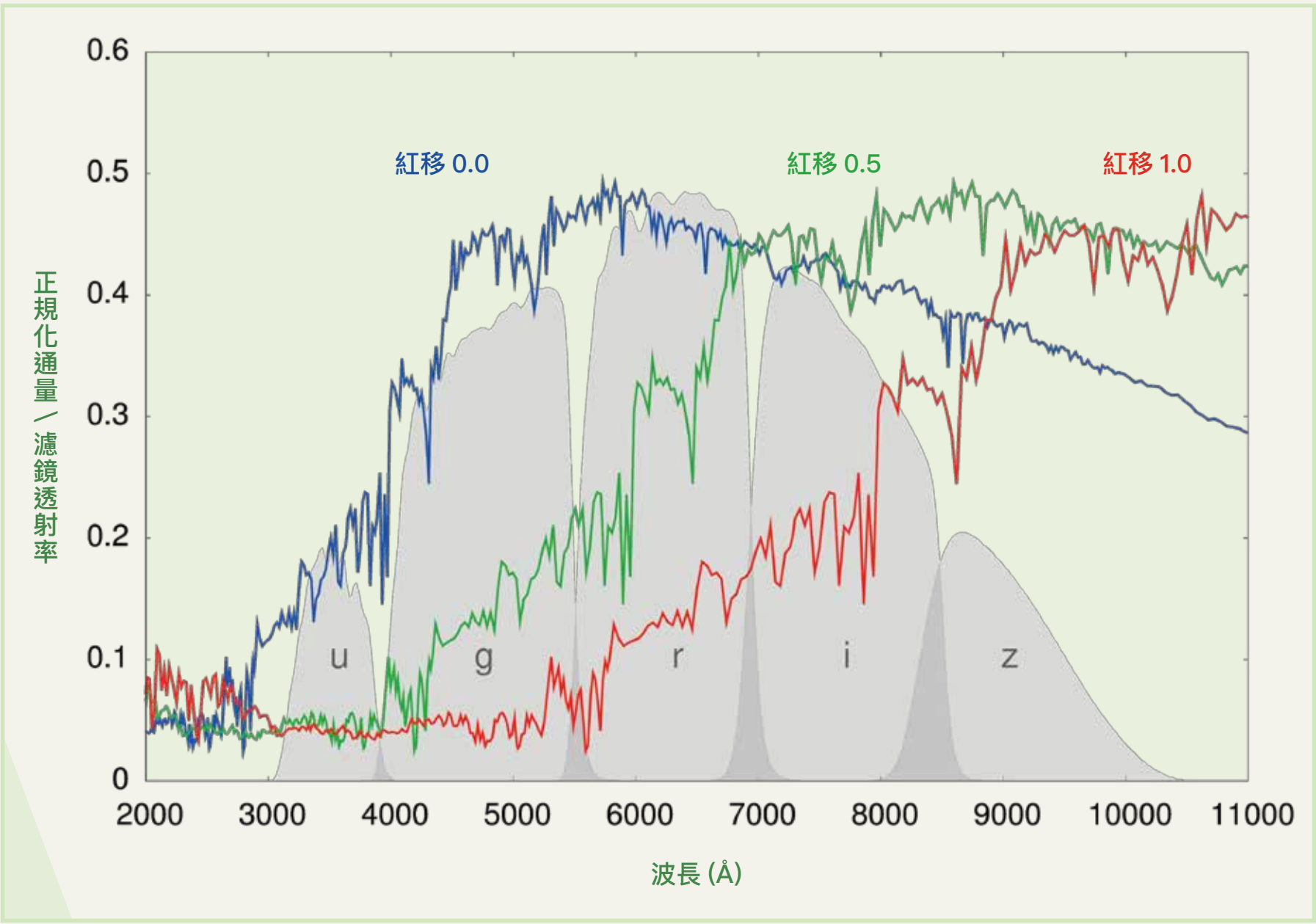
中研院天文所季報
ASIAA Quarterly Press
http://www.asiaa.sinica.edu.tw

機器學習
在光度紅移測量的應用

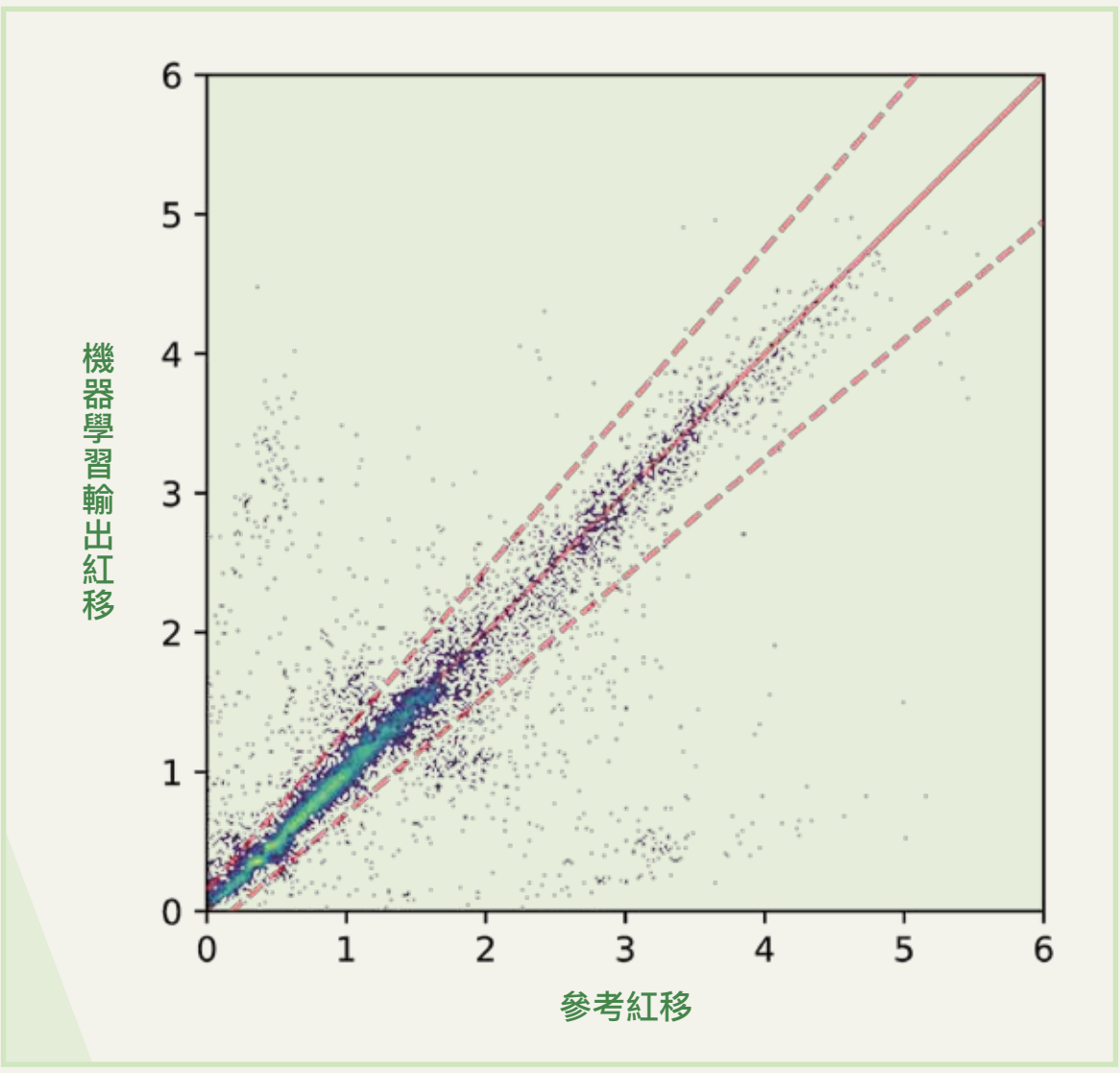
作者 / 謝寶慶

大數據和機器學習是近年來非常熱門的名詞。但因為從大約二十年前開始，就已經有大型的巡天觀測計畫，數據量已經龐大到無法用人工一一處理。所以，雖然當時「大數據」和「機器學習」等名詞還沒流行，但其實科學界就已經開始利用「機器學習」在處理「大數據」了。

「機器學習」這個名詞聽起來雖然似乎很深奧，但在大部分的應用上，其實非常接近統計學和數據分析裡的「統計模型分析」。接下來就舉個實際的應用來說明：在星系研究的領域裡，星系離地球的距離是一項非常重要的數據。測量星系與地球之間的距離的方法有很多種，在這裡我們是以「紅移」（redshift）為主。由於宇宙膨脹，所以距離地球越遠的天體，會以越快的速度遠離地球。因為都卜勒效應的影響，這些遠離地球的天體所發出的光線，從地球上來觀測，光譜會往紅端移動（也就是往長波長移動）；遠離速度越快，光譜往紅端移動的幅度就越大。這個移動的



圖中為標準橢圓星系在三個不同紅移的光譜。藍色光譜位於紅移 0，綠色光譜位於紅移 0.5，而紅色光譜位於紅移 1.0。灰色曲線代表史隆數位巡天（SDSS）五個濾鏡 （u、g、r、i、z）的透射曲線。可以很清楚的看到，相同的天體在不同的紅移，會產生不同的相對光度。



這是實際將機器學習應用在 Hyper Suprime-Cam Subaru Strategic Program 計畫的資料上，所求得的光度紅移的準確度。可以看出大部分的機器學習光度紅移，跟參考紅移（正確答案）非常接近，都落在兩條虛線之間，表示利用機器學習可以得到極為準確的紅移值。

幅度，便叫做紅移。雖然不是精確的距離單位，但卻是對中遠距離天體最具普適性的距離代表。

一般來說，測量天體的紅移，是藉由觀測天體的光譜，從光譜中的發射譜線和吸收譜線，計算出紅移的值。但是光譜觀測的效率非常低，需要很長的時間才能完成少量天體的觀測，對近年來快速發展的星系研究領域，有點緩不濟急。於是，天文學家開始嘗試使用多波段濾鏡的光度測量，當做一個超低解析度的光譜，來求出天體的紅移。跟光譜從發射譜線和吸收譜線得到紅移不同，由於從多波段濾鏡的光度測量，無法得知譜線的正確位置，所以是從連續光譜的特徵來估計紅移值。以橢圓星系來說，其連續譜線在波長 4000Å 的地方有個明顯落差，而多波段濾鏡的光度測量便可以判斷這個原來位於波長 4000Å 的光譜落差，大約移動到什麼波長，進而估計出紅移的值。這種利用多波段光度觀測來估計紅移的方法，被稱為「光度紅移」（photometric redshift）。

估計光度紅移的方式，主要分為兩大類。第一大類為「樣板擬合」（template fitting），第二大類為「經驗擬合」（empirical fitting）。第一大類的方法，是將天體的多波段光度，與預先準備好的樣板（template）做比對。這些樣板通常是利用天體理論模型，將計算出來的多種天體光譜，套用不同的紅移值，並求出相對應的多波段光度值之後，儲存起來。舉例說明，如果有 5 種天體，放置在 100 個不同的紅移，就會有 500 種不同的光譜。假設多波段光度有 5 個濾鏡，那麼將這 500 個光譜套用在這五個濾鏡上，就可以得出 500 組，每組 5 個光度值（共 2500 個光度值）。之後只要在樣板裡搜尋哪一組光度與未知紅移天體的光度最接近，那麼這組光度所對應的紅移值，便是這個天體的光度紅移。而第二大類，便是應用機器學習的方法，估計出紅移。在估計紅移之前，必須先「訓練」電腦，讓電腦知道不同紅移的天體，在多波段光度上的差異。因此，要先準備一組稱為「訓練集」（training set）的資料來讓電腦學習。這個訓練集裡包含了許多位於不同紅移的天體的多波段光度資料，而電腦必須利用這個訓練集，來找出多波段光度和紅移值的關係。我們將紅移值定義為「z」，而多波段光度有 5 個濾鏡的資料，定義為「M」、「N」、「O」、「P」和「Q」，則電腦要做的事情就是求出 z = f(M,N,O,P,Q) 這個函式。一旦求出這個函式之後（也就是電腦已經學會了），所有未知紅移的天體，就可以套用這個函式，直接從多波段光度 M、N、O、P 和 Q 求出紅移值 z 了。至於如何求出這個函式，就分成非常多種方式，從最基本的高次多項式曲線擬合，到類神經網路等等，依不同的應用有不同的適用性。更重要的是，經驗擬合等機器學習的方法，不但可以用來估計天體的紅移，還可以用來估計天體的其他參數，如質量、絕對亮度等等，彈性非常大，而且精度也非常高。

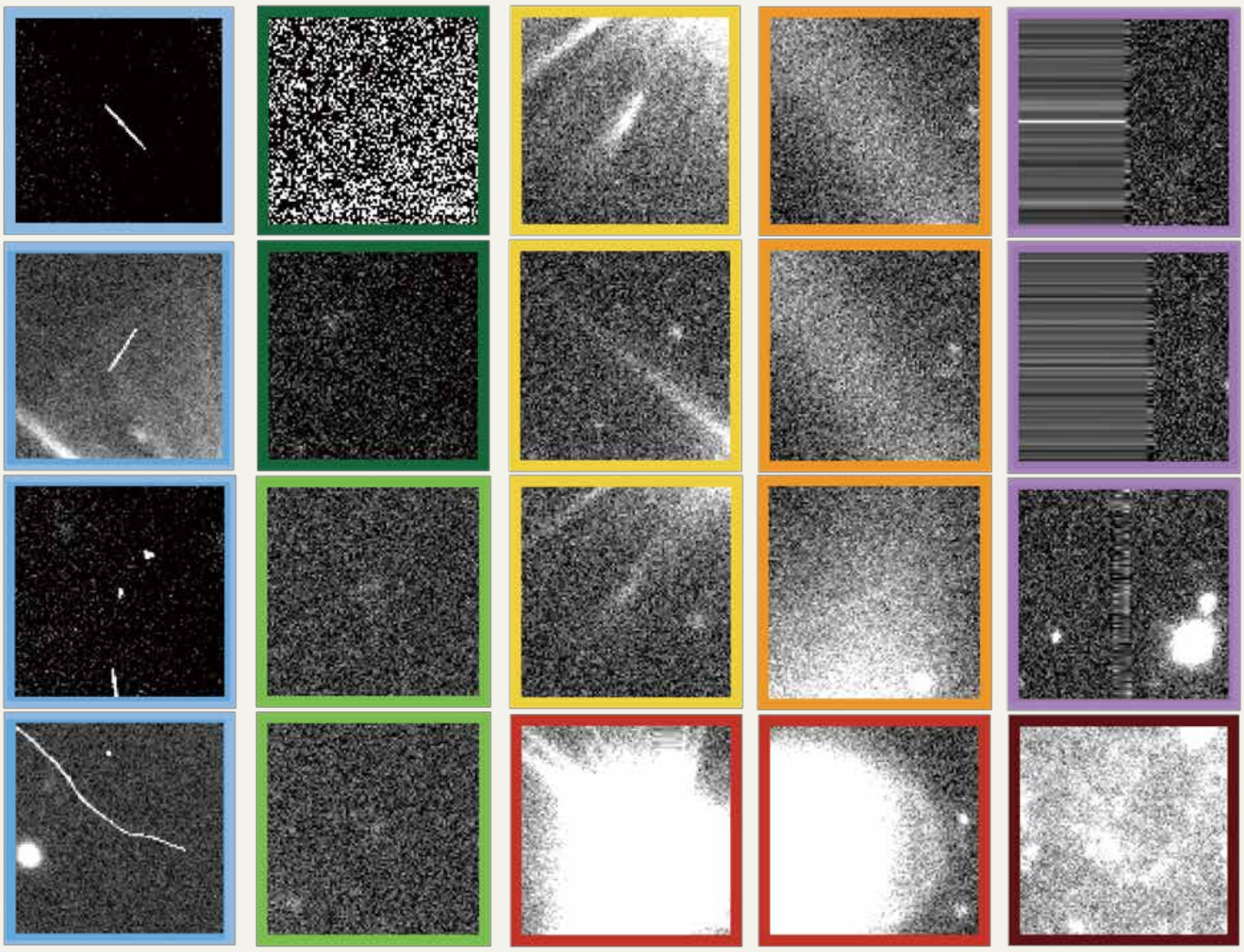
上面所提的「經驗擬合」方法，正被套用在臺灣也參與的 Hyper Suprime-Cam Subaru Strategic Program 大型國際合作計畫的資料上。目前的資料量已經高達 500TB，被觀測的天體數量以數十億計。在觀測完成後，總資料量會超過 1PB（1000TB）。為了利用機器學習估計如此龐大數量天體的各項參數，本所已經建構了一個計算叢集，專門用來處理此計畫的資料；對於推動計劃的進行，有莫大的助益。

夏威夷毛納基峰的夜晚，照片下方從左至右分別為 JCMT 望遠鏡、次毫米波陣列望遠鏡（SMA）及 Subaru 望遠鏡。

利用機器學習
尋找太陽系小天體

作者／陳英同

機器學習除了在產業界上的應用外，在科學上的應用也很多，尤其是天文。由於各種天文的巡天計劃在這十幾年如雨後春筍般的成立，大量資料處理變成一項瓶頸，所以好的資料處理方法就順理成章成了非常重要的關鍵技術。天文上，要從巨量資料裡找出有趣的天體，真的就像是在大海裡撈針，所以機器學習基本上就是要先把海水濾過一次，讓研究人員比較能快速的找到那根針，或是確定這根針的大小或材質。目前機器學習在天文上的應用主要在：（1）找出或確認有趣的天體，（2）



各種在天文影像中的假天體。藍框是宇宙射線，綠框是雜訊或過暗的背景，黃框是星星周圍的散射光，紅框是亮星的過度曝光區域，橘框是望遠鏡結構所造成的散射光、紫框是資料處理過程中重新修正過的區塊、棕框是密集星場或大型星雲的複雜結構。

判斷可能的紅移。本文著重在前者，也就是找出我們有興趣的太陽系小天體。

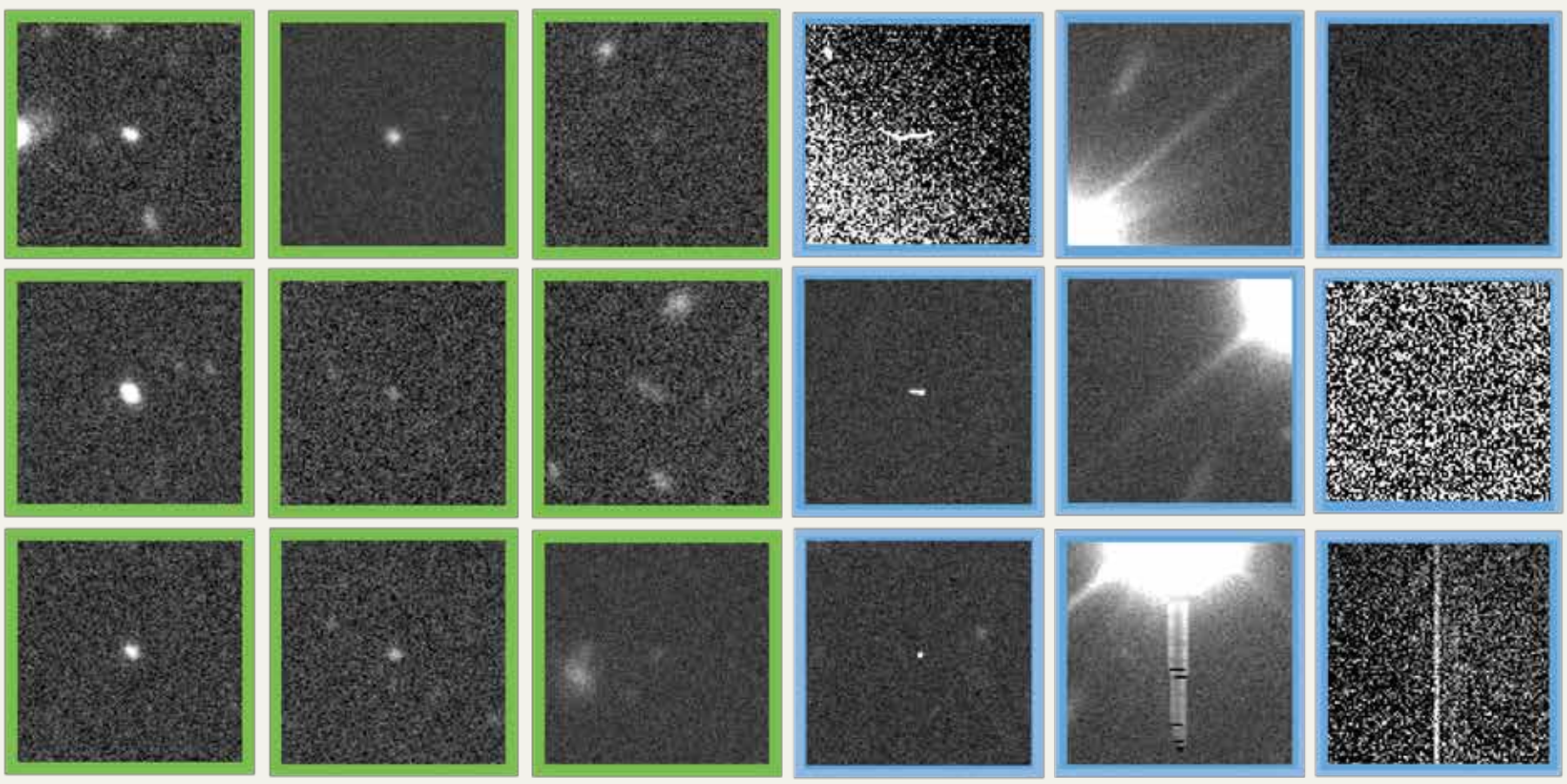
太陽系裡的小天體（包含主帶小行星及古柏帶天體）其實都是不斷的重覆出現在這些巨量的天文資料裡，但是我們非常難判斷它是小行星還是一般的背景天體（恆星或是星系）。除非是在長時間曝光的資料，這些天體會變成長條狀，否則它們其實看起來跟一般的恆星一樣都是點光源。但是太陽系小天體會在不同時間相對於背景天體改變位置，成為另一種判別小天體和恆星的方法。所以要應用機器學習於找尋太陽系小行星的話，需要有不一樣的靈感方向。就如同剛剛說的，由於小行星在一般的天文影像中，跟常見的背景天體都一樣是點光源，唯一的確認方法就是利用可能天體在不同時間的觀測影像，再根據軌道力學計算是否有合理的軌道。但是萬一要檢查的對象很多的話，就會有電腦計算能力上的瓶頸。要解決這一個問題，預先把不可能的對象先行過濾掉就變的非常重要。

在真實的天文資料庫裡，其實假的資料佔了不小的一部分。而大部分的「假」天體都是影像處理過程中產生的。由於偵測演算法的限制，像是宇宙射線、亮星周圍或是感光元件上的異常結構都有可能被誤判出不少的假天體。此外，光學系統所產生的散色光也容易被誤判成星系。

這些假天體其實人眼一看就能立即下判斷是否為真正的天體，但是在大型巡天計劃裡，資料實在太過於龐大，一天就有可能產生出超過 500GB 的資料，即使動員所有計劃的成員也只能判斷極少一部分的資料而已。這個時候機器學習就是一個非常有用的工具了。

機器學習這項技術就如同它的名稱，學習是一個最重要的過程。在讓機器開始學習之前，人要先學習；也就是說，人需要先自己大概判斷哪些參數可能對於機器判斷會最有幫助。如圖所示，這些假的天體都有非常奇怪的形狀，或是亮度很暗，但是沒有單一個絕對的判斷準則。所以我們要先行把這些「一定是假的天體」選出，當做「假範本」。另外我們要選出不同時間都待在同一個坐標的星體，當做「真範本」。這個時候再把「真範本」跟「假範本」一起拿給電腦做機器學習，也就是跟電腦說哪一些是真的，哪一些是假的天體。而學習的內容就是我們剛剛所提到的：形狀、亮度。在反覆測試之後，我們就能夠選出哪一批真假範本是真好的訓練樣本，哪些參數是判斷真假中比較有效的，最後我們就能夠訓練出我們可接受的機器學習模型。

結合上述的方法與工具，讓我們在找尋太陽系的移動天體過程簡化了許多：（1）先用上述挑出真範本的方法選出背景星與星系，然後把它去掉。（2）利用機器學習把假的天體再進一步去除。（3）最後再試著把不同時間可能的移動天體試著做軌道的擬合，如果誤差很小，就列為可能的移動天體。以上就是目前使用機器學習，尋找太陽系小天體的方法。



左邊綠框為被找出來的真正太陽系小天體，右邊藍框是被排除掉的假天體。

機器學習與
大型巡天計畫

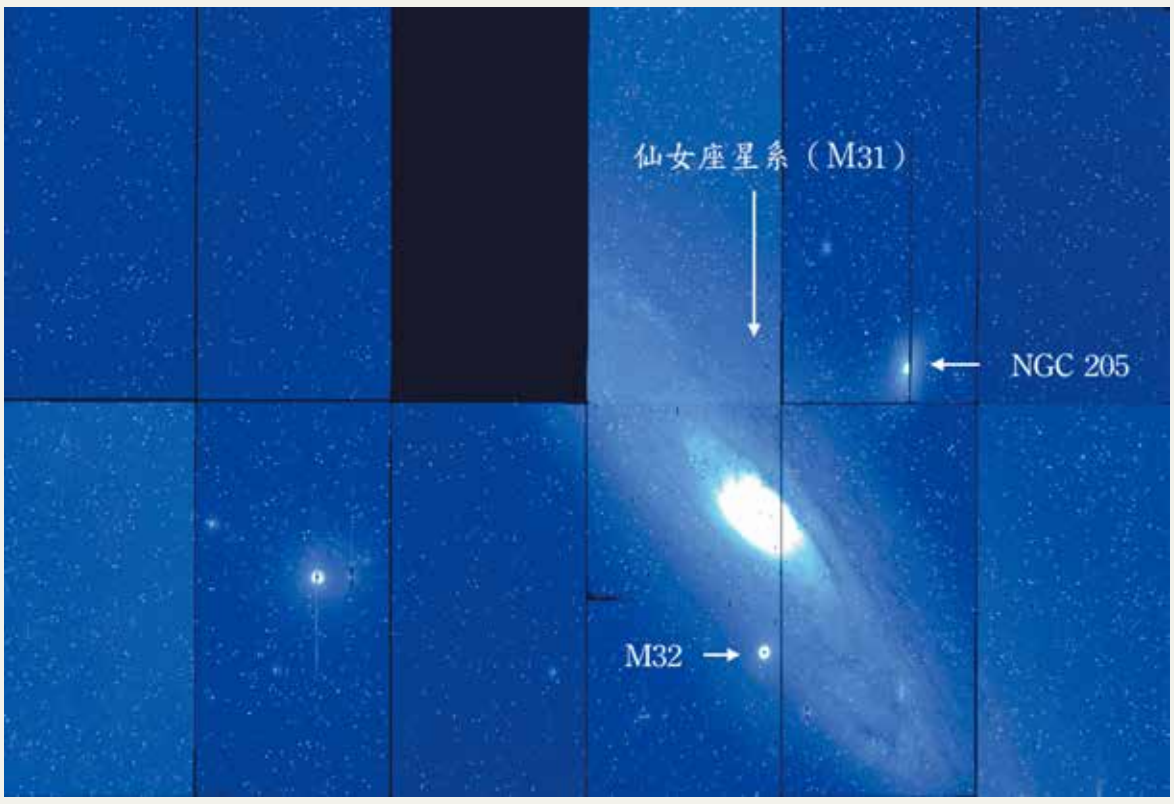
作者 / 李見修（美國國家光學天文台）

在單純的天體分類之餘，機器學習在巡天計畫後續觀測上，更扮演了關鍵角色。最有名的例子便是由加州理工學院主導，許多國際機構（包括臺灣的中央大學與清華大學）所參與的帕洛瑪瞬變天體巡天計畫（Palomar Transient Factory, PTF）。這個計畫的主力是帕洛瑪天文台裡兩座瀕臨汰舊的一米廣角施密特望遠鏡。加州理工學院從加法夏天文台購入汰換掉的舊相機，安裝在口徑一米二的 P48 望遠鏡上，將其廣角視野的特性發揮得淋漓盡致，單幅影像可達到 7.2 平方度，相當於 40 倍滿月大小，連離地球最近的仙女座星系都能完全入鏡。廣角相機雖然可以快速巡天，但缺點是解析度低，像素約一平方角秒（一般天文相機為 0.25 至 0.01 平方角秒），因此需要高解析度的相機進行後續觀測。這時帕洛瑪山頭上另一座老舊的 1.5 米望遠鏡—— P60 便派上用場了。加州理工大學為此特別設計了一具多波段測光暨光譜儀（軟體部分由中央大學主導），可以一次取得 griz 四個濾鏡的光度，同時拍攝低解析度的光譜，快速且準確地分類新的瞬變天體。

但問題來了，巡天相機一晚偵測到的瞬變天體數量極多，一個一個目視分類，即便有一組天文學家日以繼夜投入也分類不完，更遑論瞬變天體稍縱即逝，如何能在極短的時間內挑出有趣的天體，並啟動後續觀測呢？聰明的天文學家們想到一個方法，那就是利用機器學習自動分類，並依重要性排序。這麼一來，便能達到幾乎即時（在同一個晚上）啟動後續觀測，重要的是天文學家們還不用熬夜！

有許多人害怕機器學習，尤其許多天文台的工作人員，認為這會奪走大家的工作機會。其實恰恰相反，機器學習幫我們免去無聊的繁瑣工作，讓天文學家們能集中心力，專注在更重要也更有趣的科學研究上。

（本文改寫自科學月刊第 578 期）



帕洛瑪瞬變天體巡天計畫相機視野極廣，不僅仙女座星系，連附近的矮星系 M32 以及 NGC 205 都能一併入鏡，綽綽有餘！

天聞季報編輯群感謝各位閱讀本期內容。本季報由中央研究院天文所發行，旨在報導本所相關研究成果、天文動態及發表於國際的天文新知等，提供中學以上師生及一般民眾作為天文教學參考資源。歡迎各界來信提供您的迴響、讀後心得、天文問題或是建議指教。來信請寄至：「臺北市羅斯福路四段 1 號 中央研究院 / 臺灣大學天文數學館 11 樓 中央研究院天文所天聞季報編輯小組收」。歡迎各級學校師生提供天文相關活動訊息，有機會在天聞季報上刊登喔！

發行人｜朱有花。執行主編｜周美吟。美術編輯｜王韻青、楊翔伊。執行編輯｜曾耀寰、劉君帆、蔣龍毅。
發行單位｜中央研究院天文及天文物理研究所。天聞季報版權所有｜中研院天文所。ISSN 2311–7281。GPN 2009905151。
地址｜中央研究院 / 臺灣大學天文數學館 11 樓。（臺北市羅斯福路四段 1 號）。電話｜(02)2366–5415。電子信箱｜ epo@asiaa.sinica.edu.tw。