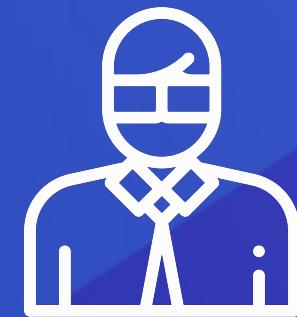




Day 59 非監督式機器學習

降維方法 - 主成分分析

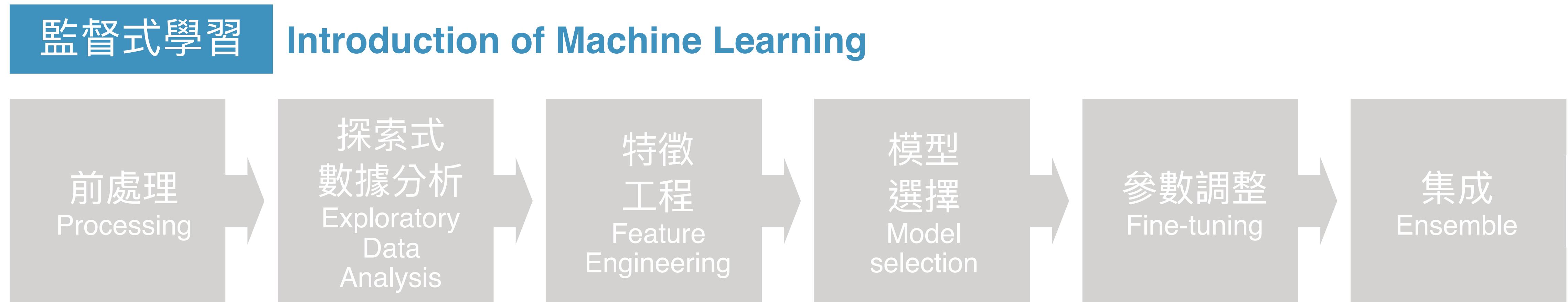


周俊川 / 聯成數網

出題教練

知識地圖 非監督學習

- 機器學習概論 Introduction of Machine Learning



非監督式學習 Unsupervised Learning



本日知識點目標

- 降低維度的好處，及其應用領域
- 主成分分析 (PCA) 概念簡介

為何需要降維 (Dimensionlit Reduction)

- **數據維度過大** 會提高模型的複雜度，特別對於一些樣本數據不足的情況，使得模型訓練的結果泛化較性差

維度目的

- 減少特徵的個數、去除特徵間的共線性問題
- 降低模型的計算量，減少模型執行時間
- 減少雜訊對於模型的影響
- 確保特徵間相互獨立

常見的降維方法

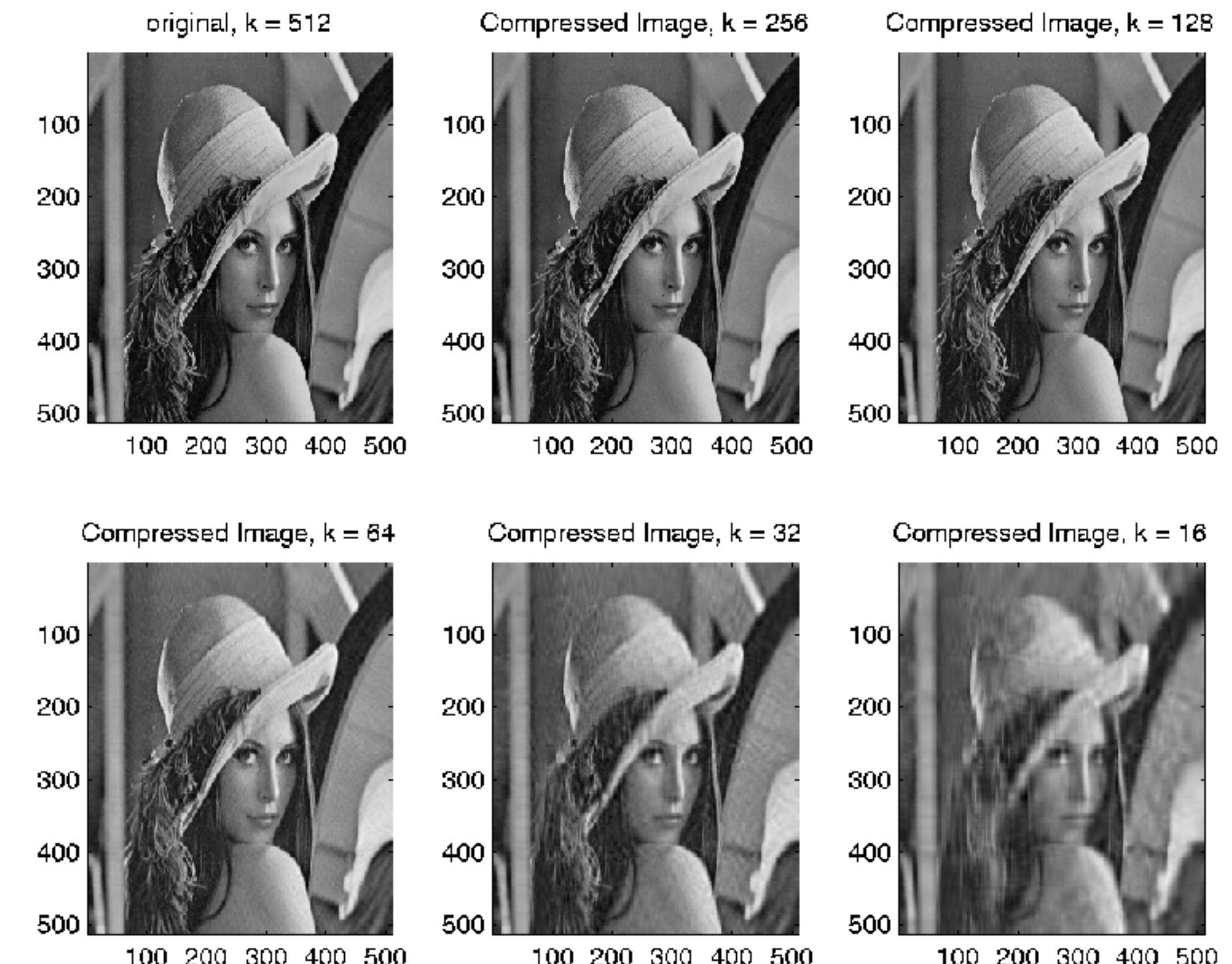
- PCA
- t-SNE

為什麼需要降低維度？壓縮資料

- 有助於使用較少的 RAM 或 disk space，也助於加速 learning algorithms

影像壓縮

原始影像維度為 512，在降低維度到 16 的情況下，圖片雖然有些許模糊，但依然 **保有明顯的輪廓和特徵！**

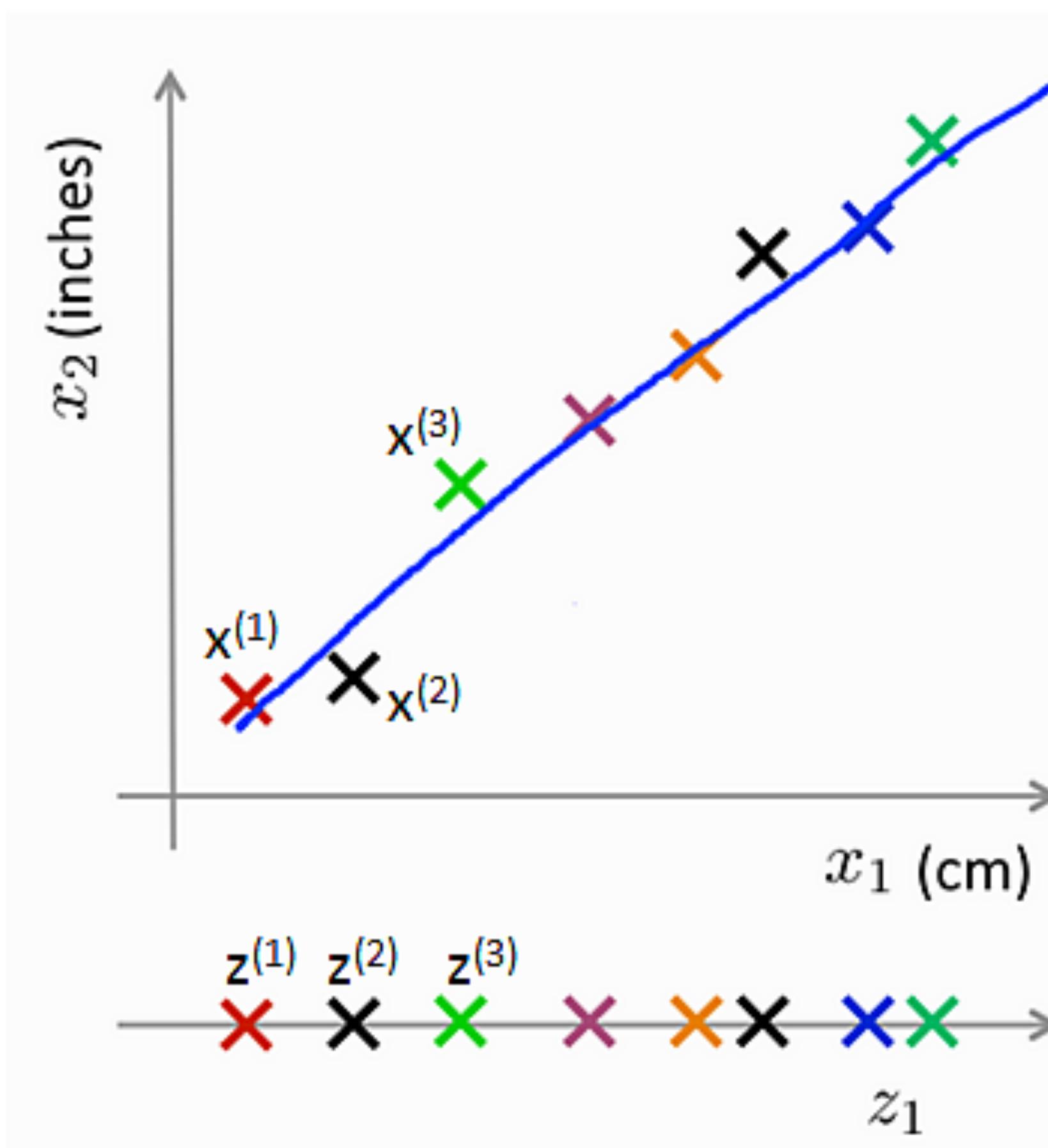
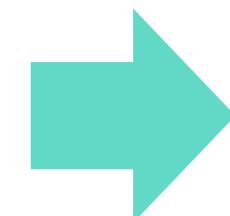
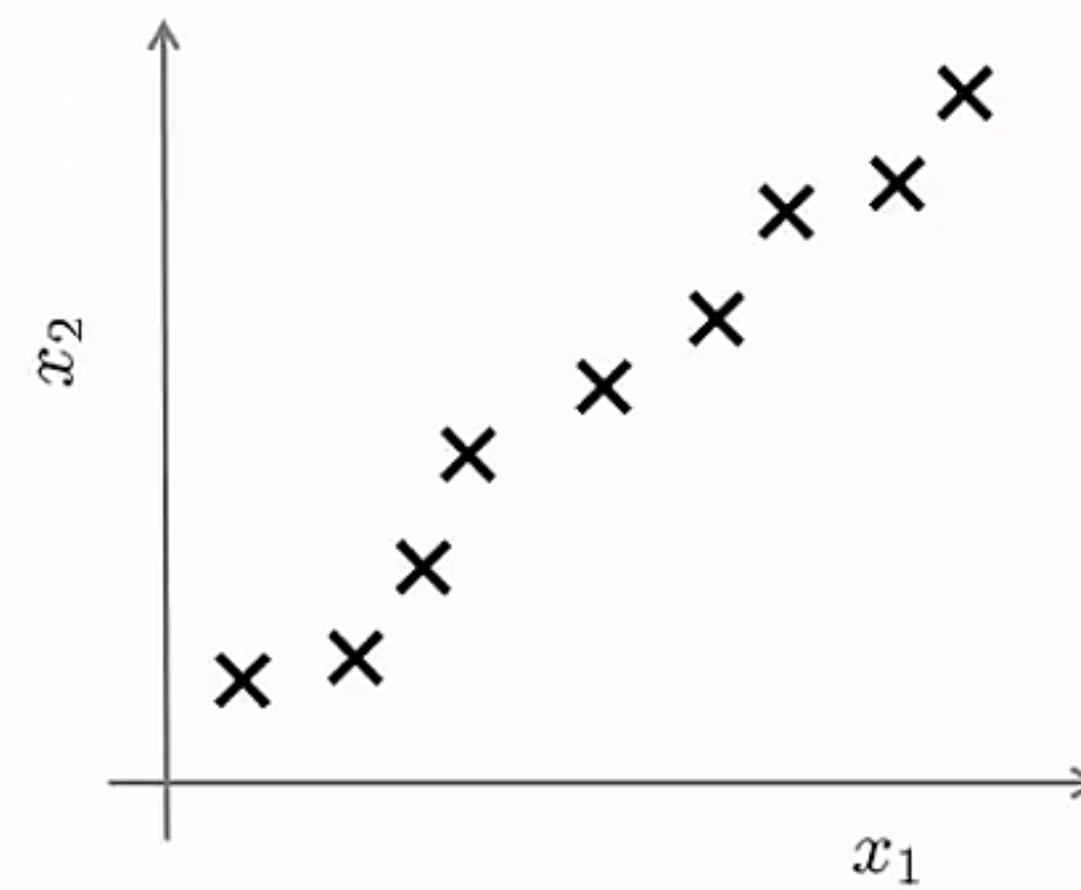


為什麼需要降低維度？特徵組合及抽象化

- 壓縮資料可進而組合出新的、抽象化的特徵，減少冗餘的資訊

合併成1個特徵

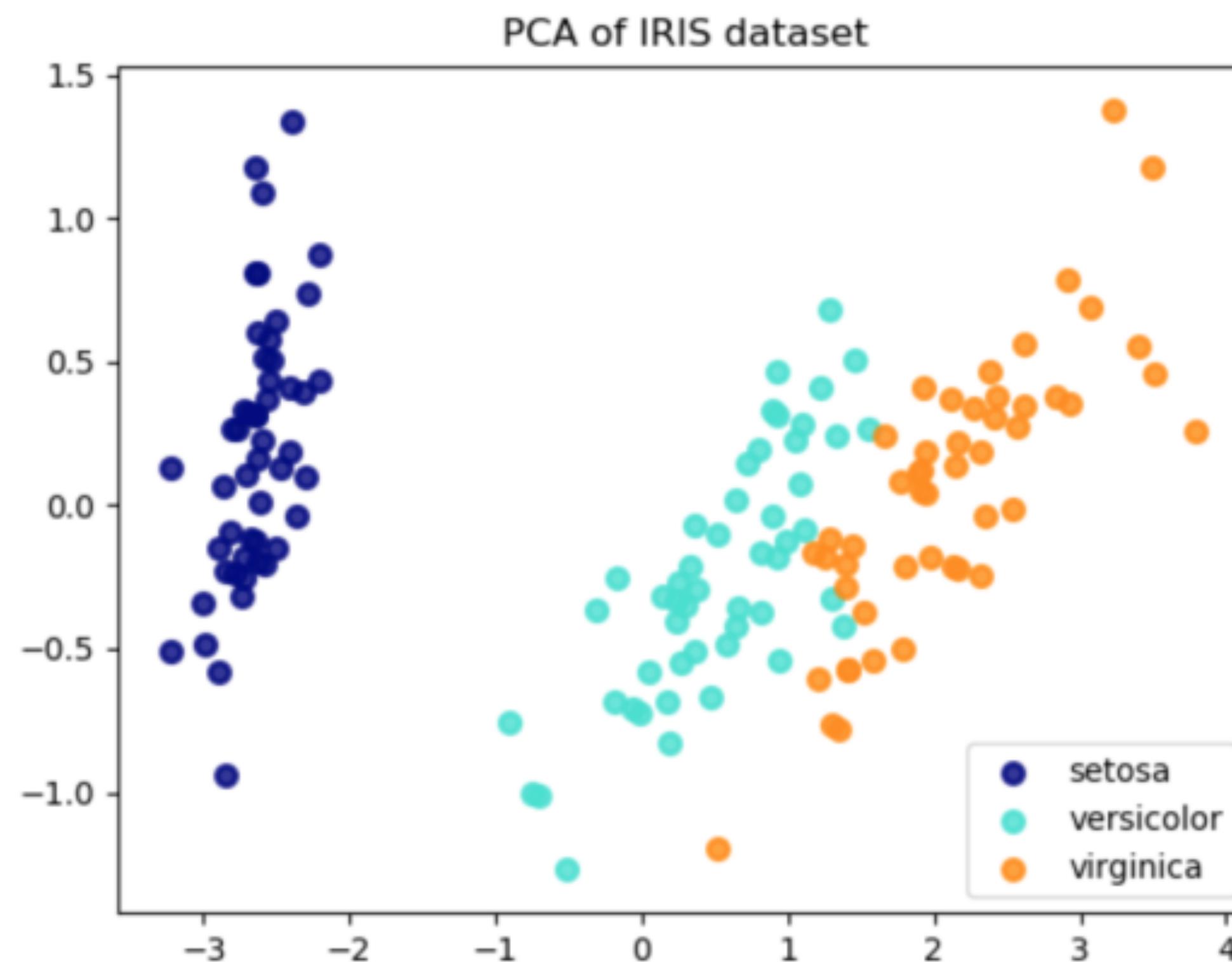
x_1 和 x_2 高度相關



把 $x(i)$ 投影到 藍色線
從 2 維 降低為 1 維！

為什麼需要降低維度？資料視覺化

- 特徵太多時，很難 visualize data，不容易觀察資料
- 把資料維度 (特徵) 降到 2 到 3 個，則能夠用一般的 2D 或 3D 圖表呈現資料

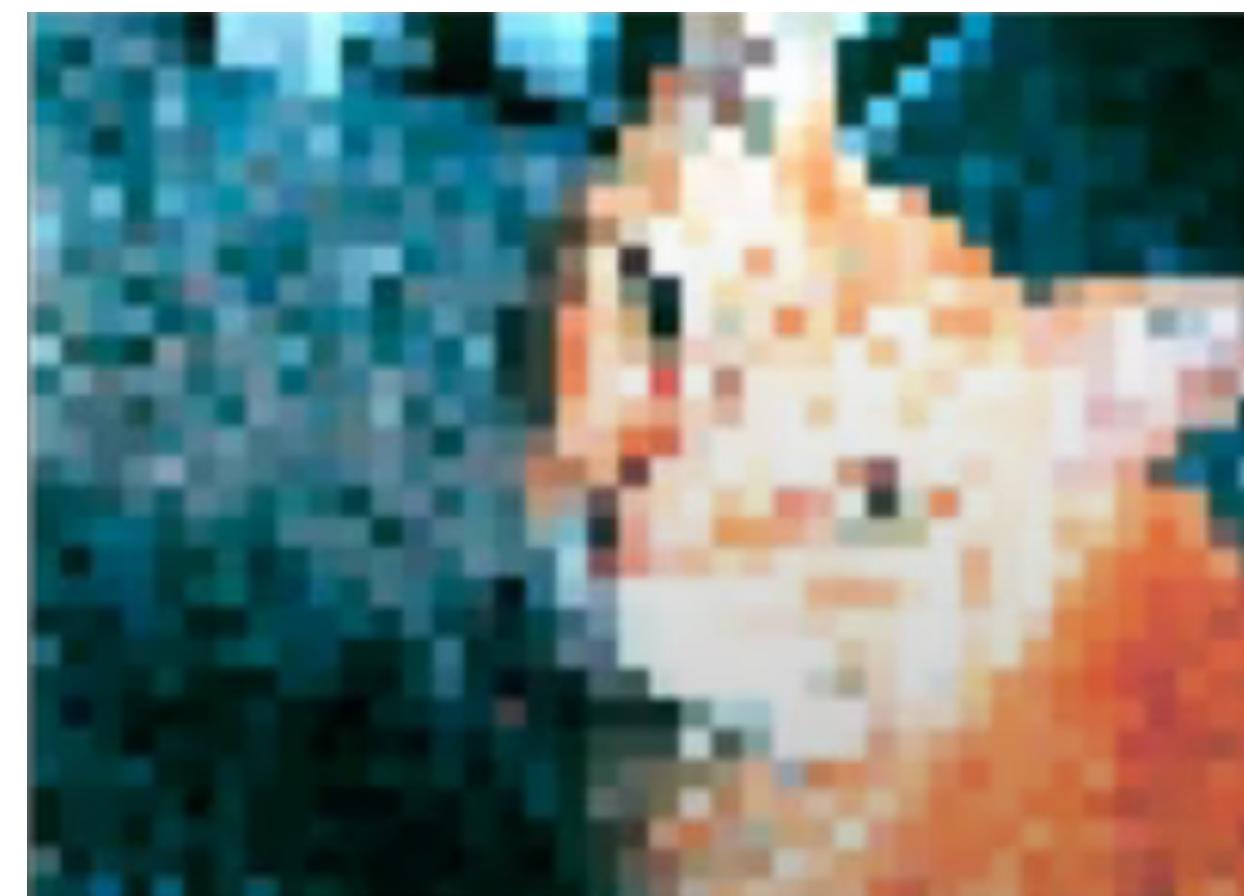


(Principal components analysis , PCA 主成分分析)

- PCA原理:利用線性轉換，將資料從高維(k 維)映射到低維(m 維)空間，並提取(m 維)資料的主要特徵，保有原始資料的重要資訊



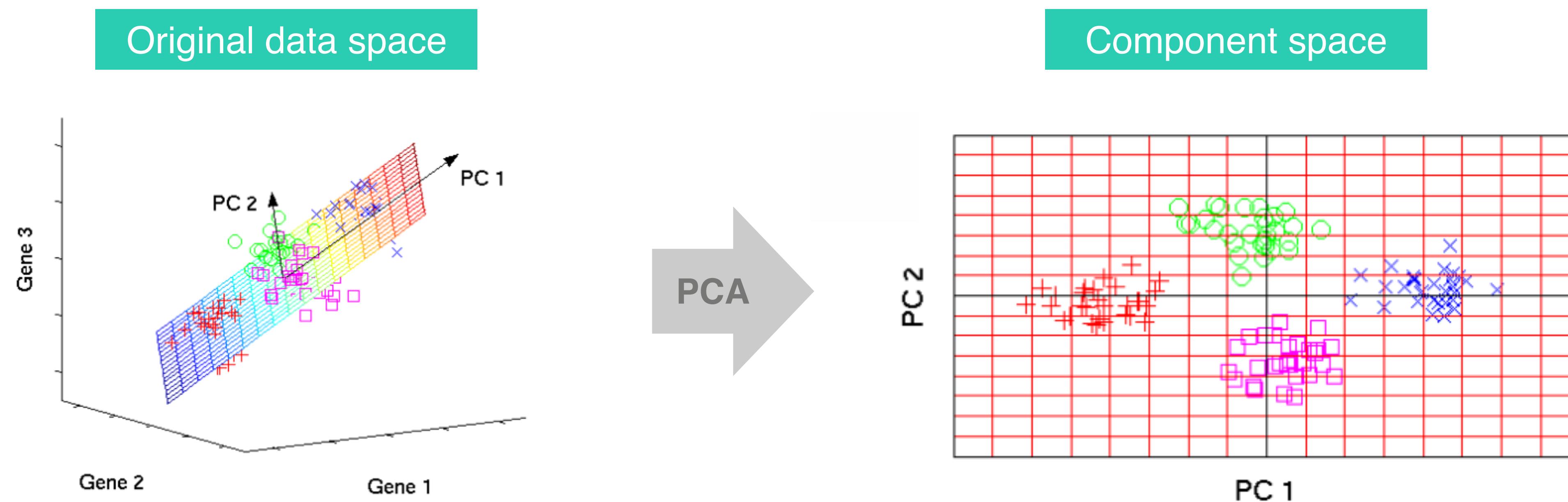
縮減



原始資料特徵之間表現出較強的相關性，若相關性較弱，降維效果較差。

(Principal components analysis , PCA 主成分分析)

- **PCA** 不是從原始資料中捨棄不重要的特徵來降維，而是由這些特徵與其向量(eigenvector)的線性組合，**降維至二維平面上**，所產生的新特徵來代表原始資料



(Principal components analysis , PCA 主成分分析)

- 實務上我們經常遇到資料有非常多的 features, 有些 features 可能高度相關，有什麼方法能夠把高度相關的 features 去除？
- PCA 透過計算 eigen value, eigen vector, 可以將原本的 features 降維至特定的維度
 - 原本資料有 100 個 features，透過 PCA，可以將這 100 個 features 降成 2 個 features !
 - 新 features 為舊 features 的線性組合

PCA演算法流程

- ① 標準化 d 維原資料集
- ② 建立共變異數矩陣 (covariance matrix)
- ③ 將共變異數矩陣 (covariance matrix) 分解為特徵向量 (eigenvector) 與特徵值 (eigenvalues)。
- ④ 選取 k 個最大特徵值 (eigenvalues) 相對應 k 個的特徵向量 (eigenvector)，其中 k 即為新特徵子空間的維度。
- ⑤ 使用排序最上面的 k 個的特徵向量 (eigenvector)，建立投影矩陣 (project matrix) W 。
- ⑥ 使用投影矩陣 (project matrix) W 轉換原本 d 維的原數據至新的 k 維特徵子空間。

PCA流程



PCA流程 (1)- 標準化資料集

- PCA 對於特徵值的範圍高度敏感，若沒有進行標準化之動作會導致在計算時，很容易造成 PCA 偏向數值較大的特徵
- 如果特徵值不同維度應該針對此項目進行 **特徵標準化處理**，讓特徵值縮放至同一個範圍，使得各特徵皆具有相同的重要性

PCA流程 (2)- 建立共變異係數 (covariance , 協方差)

- 假設有 n 個 2 維資料 $(x,y)=\{(x_i,y_i)\}$ ($i=1 \dots, n$)
 - 利用共變異數來衡量 2 個變數之間的相關性程度，判斷 x 、 y 軸的資料是否相關
 - 衡量 2 個變數的相關程度

$E(X)$ 和 $E(Y)$ 為 X 和 Y 的樣本平均數

$$E(X) = \frac{1}{n} \sum_{i=1}^n x_i, \quad E(Y) = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

PCA流程 (2)- 利用共變異係數判斷相關性

- 如 (x : 身高 , y : 體重) , 比較這 2 個變數間是否存在相關性
- 當 $\text{cov}(X, Y) > 0$ 時，表明 X 與 Y 正相關
- 當 $\text{cov}(X, Y) < 0$ 時，表明 X 與 Y 負相關
- 當 $\text{cov}(X, Y) = 0$ 時，表明 X 與 Y 不相關

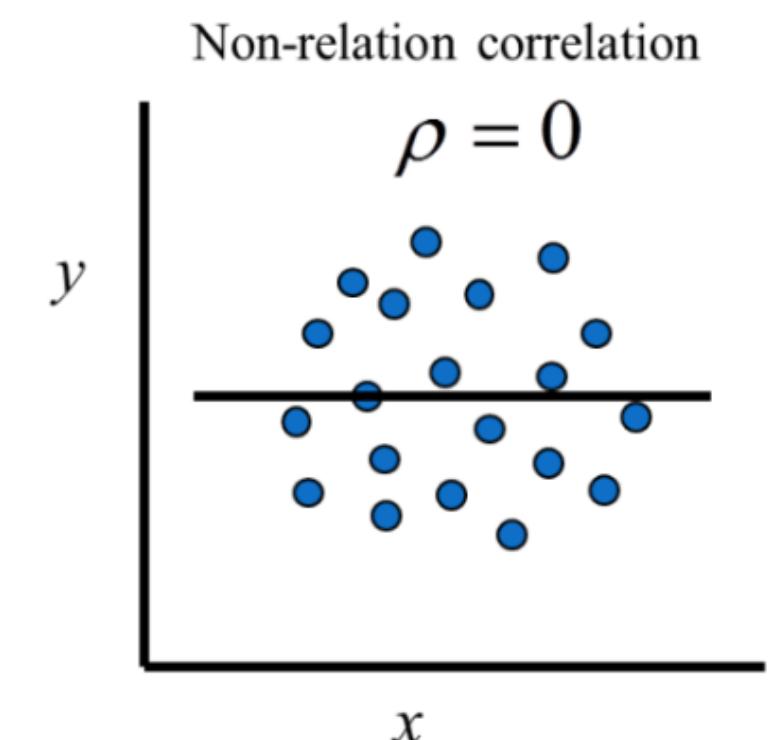
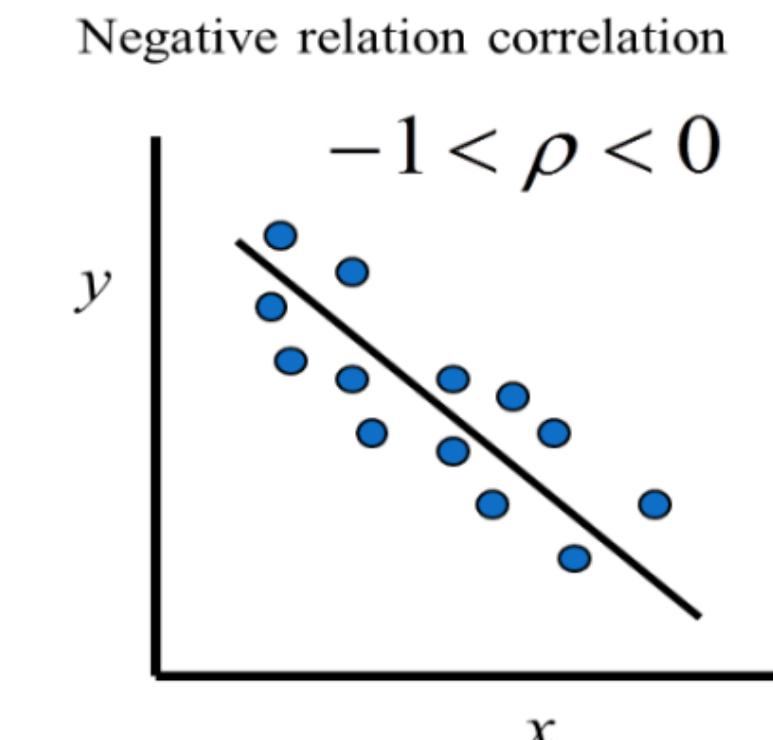
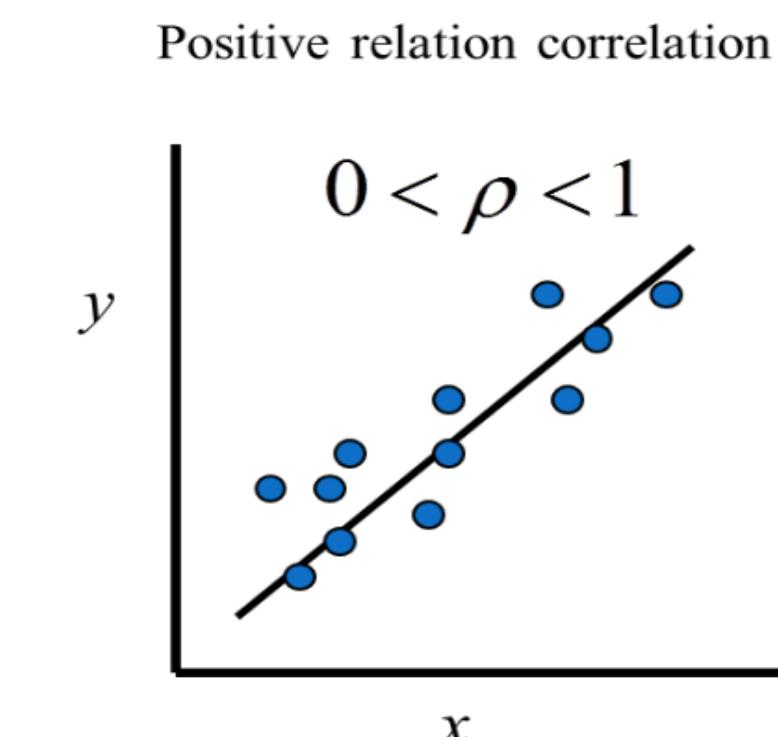
PCA流程 (2)- 相關性的定義 (皮爾森相關係數，Pearson correlation)

- 主要探討各變數之間的線性關係，其值介於 -1~1 之間

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

若身高體重的相關性結果為：0.8

則表示身高體重具有線性相關，
符合皮爾森相關係數定義

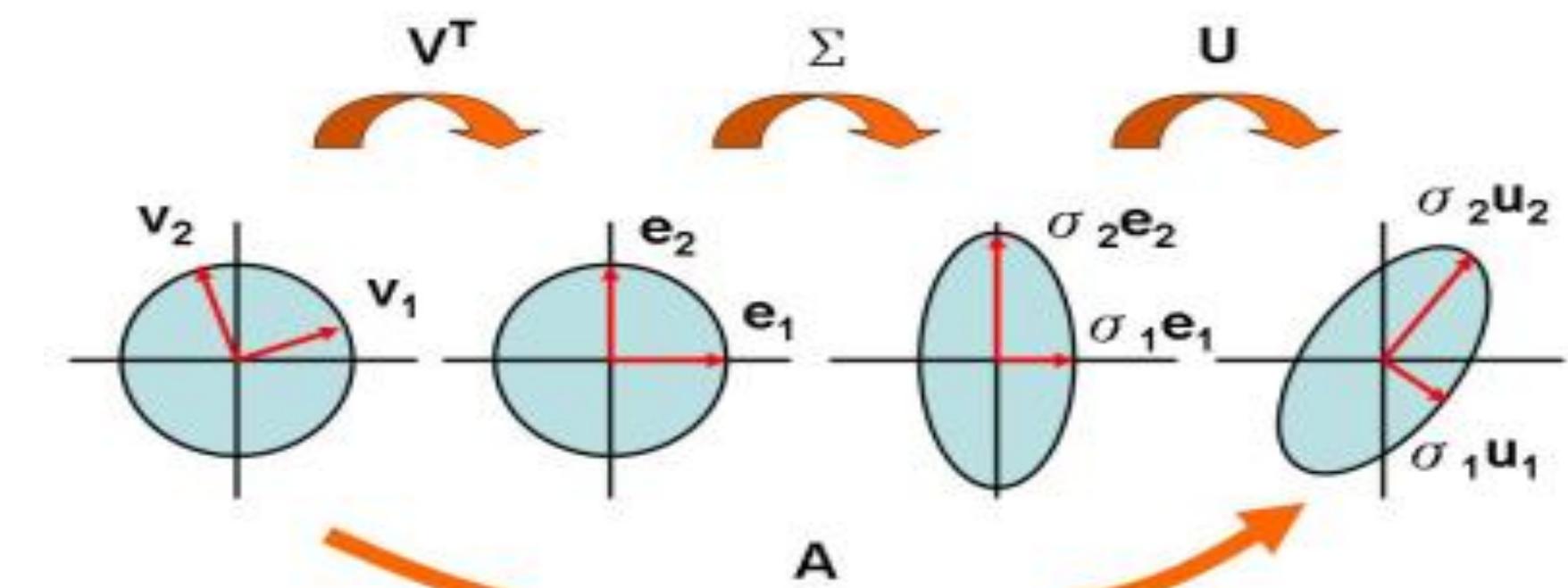


PCA 流程 (3)- 分解共變異數矩陣 (4)- 取出最大特徵值

- 經由計算所得到的共變異數矩陣，所取得之特徵值向量，並從中選取特徵值最大的 K 個特徵，並將特徵矩陣轉換到新的特徵空間中進行降維

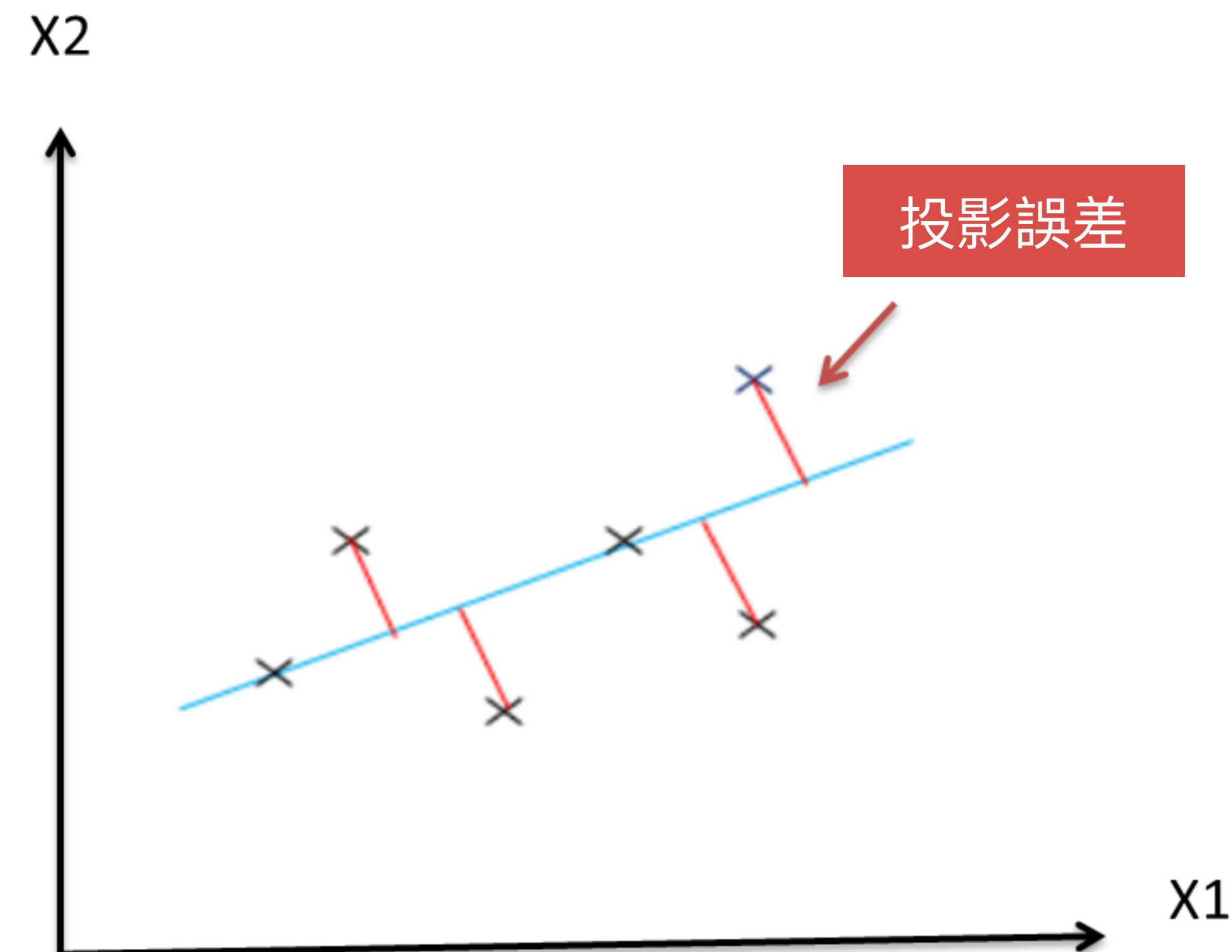
“
詳細特徵值 (EVD) 與奇異值 (SVD) 分解推導及運算公式如下參考附件：

- 特征值(EVD)分解
- 主成分分析 (PCA) 原理詳解-特徵值(EVD)分解&奇異值分解(SVD)分解
- 特徵值分解與奇異值分解原理與計算



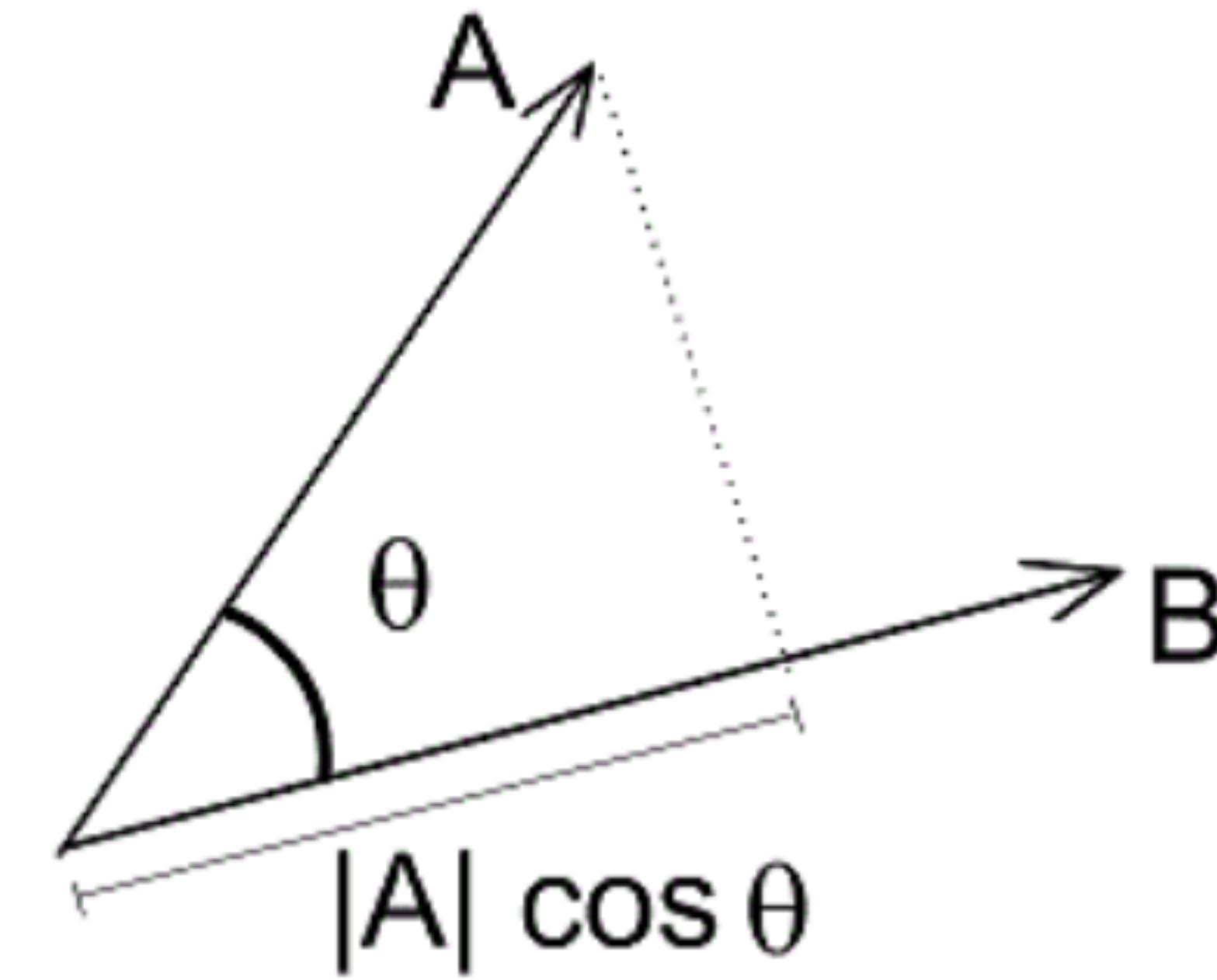
PCA 流程 (5)-建立PCA投影

- 將資料從 n 維降低到 k 維，找到 k 個向量，用於投影原始資料，使投影誤差（投影距離）最小，並投影到該直線上表示二維的資料。



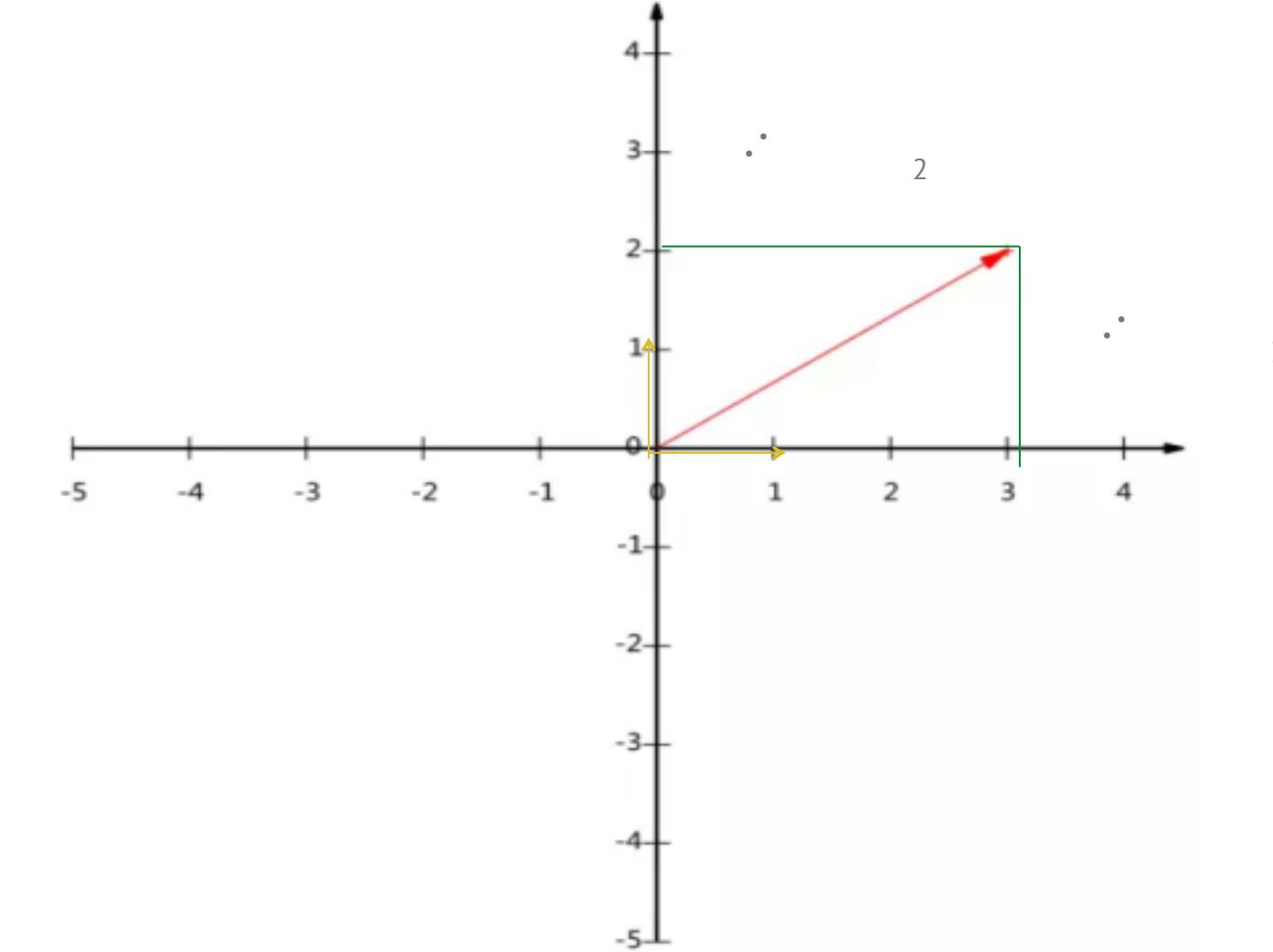
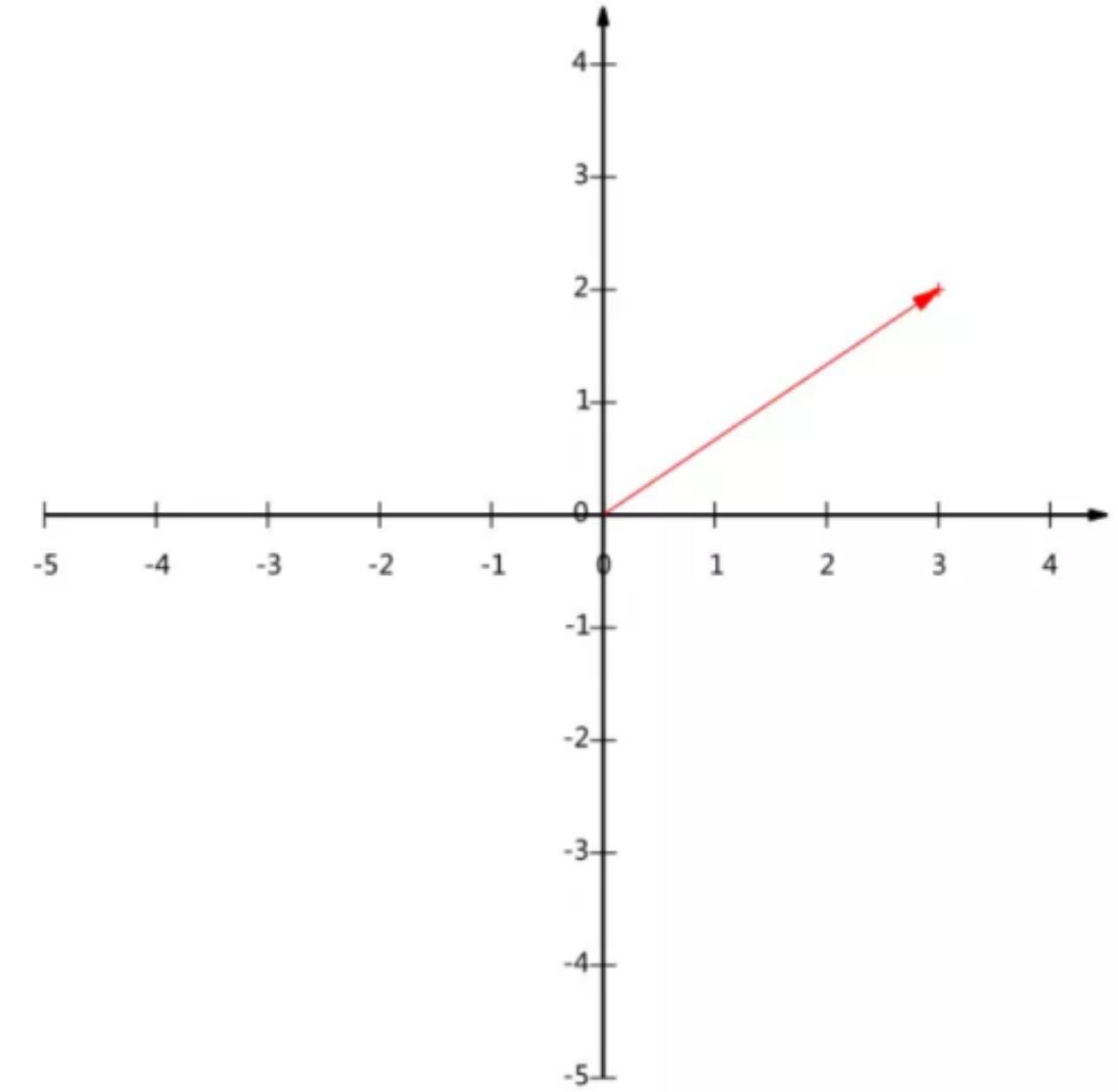
PCA 流程 (6)-向量投影方式

- 向量 a 長度乘上 $\cos\theta_a$ ，得到鄰邊，而鄰邊為 a 投影在 b 的長度。



PCA 流程 (6)-基的概念

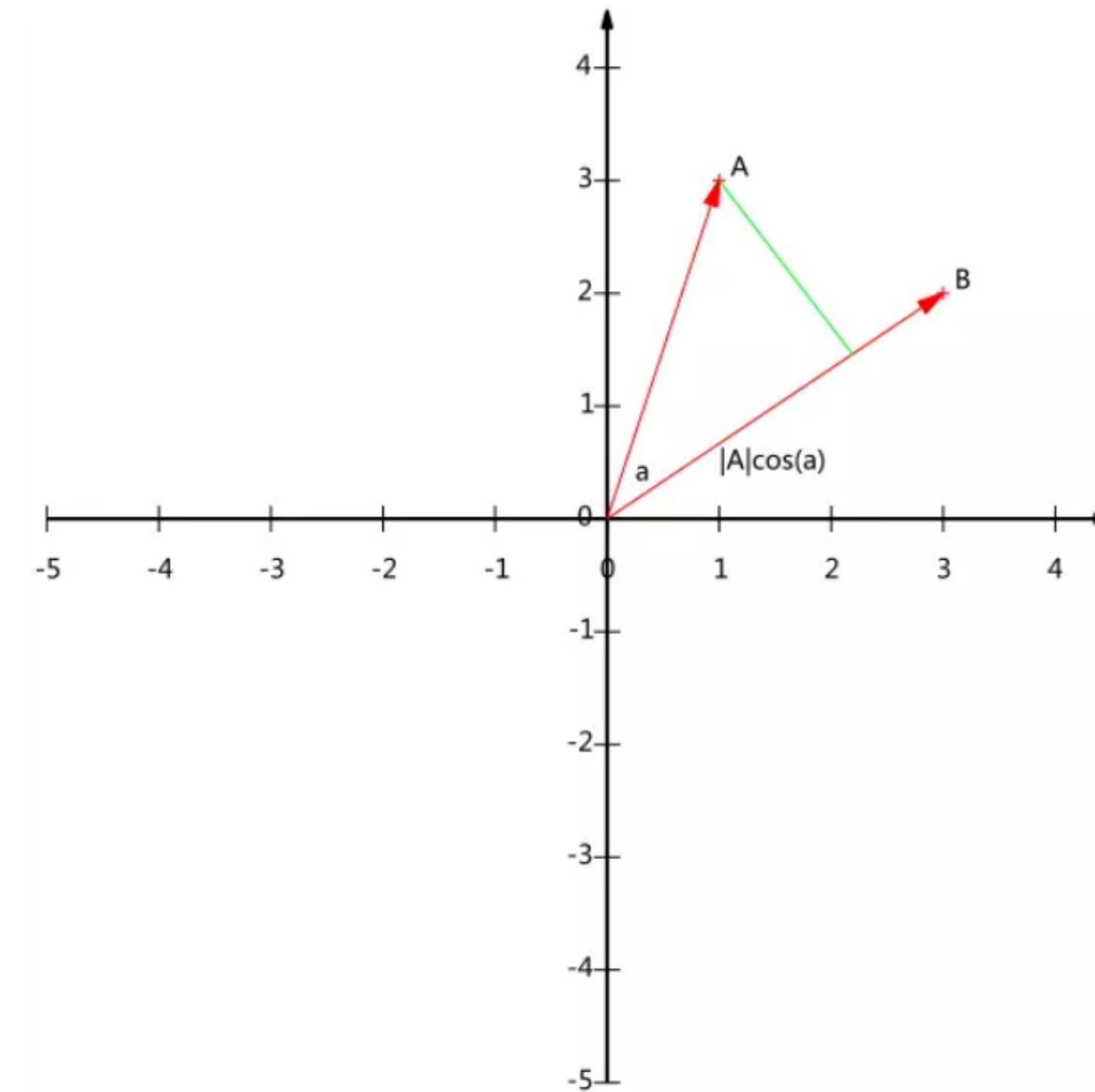
- 以 A(3,2) 座標表示為空間中的向量
- 向量 (3,2) 可以說是在 X 軸投影長度為 3 , y 軸投影長度為 2
- 向量 (x,y) 表示的線性組合 $x(1,0)^T + y(0,1)^T$
而單位向量 $(1,0)$ 與 $(0,1)$ 在二維空間中可稱做為一組基
- 二個線性相關的向量皆可以組成一個基



PCA 流程 (6)-向量與基底轉換表示

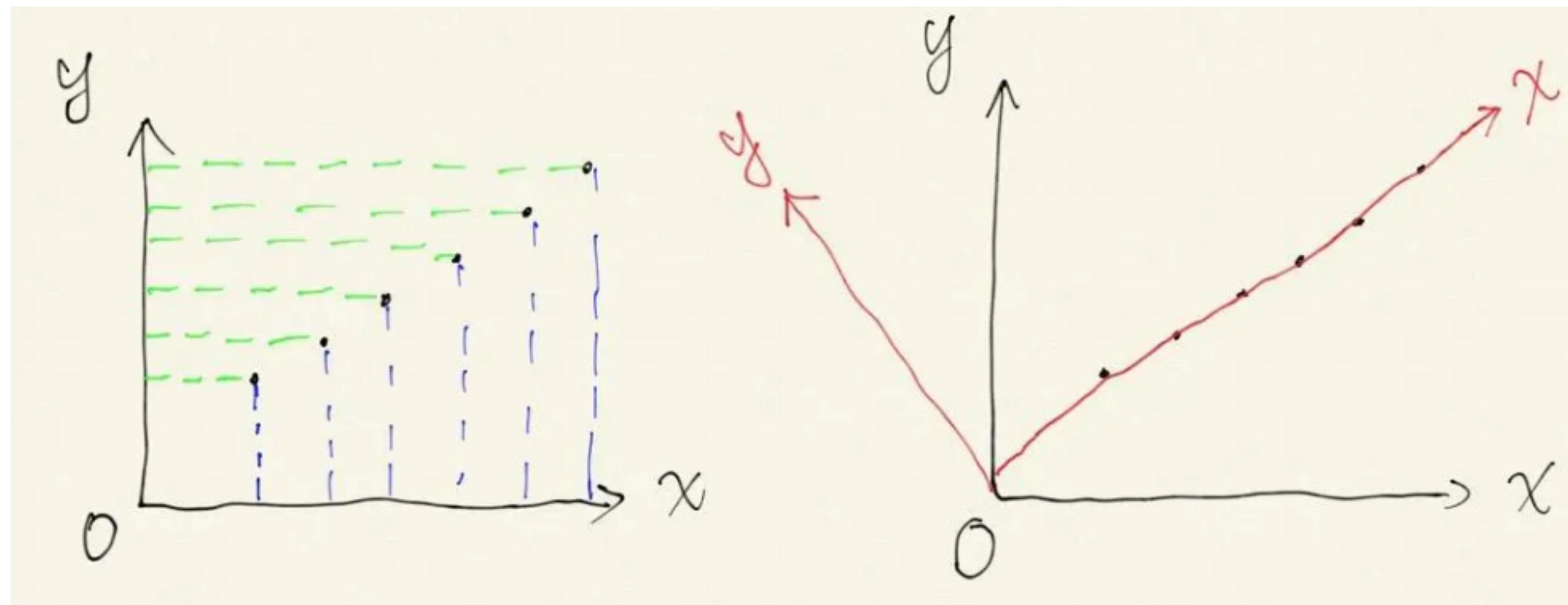
- 設 $A(x_1,y_1)$, $B(x_2,y_2)$,在 xy 座標上呈現 2 個向量，並在 A 與 B 的交點中畫上垂直線
- 向量內積公式： $A \cdot B = |A| * |B| * \cos\theta$
- A與B的交點：投影
- $|A|\cos(a)$ ：投影向量長度
- $|A|$ 為A向量的長度
- 若 $|B|$ 向量的長度為 1
- 則A與B內積值等於A至B間的投影長度

$$A \cdot B = |A| * |B| * \cos\theta$$



PCA 流程 (6)- (基底轉換，base conversion) 概念

- 藉由基底轉換將 xy 軸進行旋轉，讓 x 軸向量能夠擬和原始的資料，並將原始資料投影至 x 軸上
- 藉由二維降至一維，使得資料損失的訊息降低

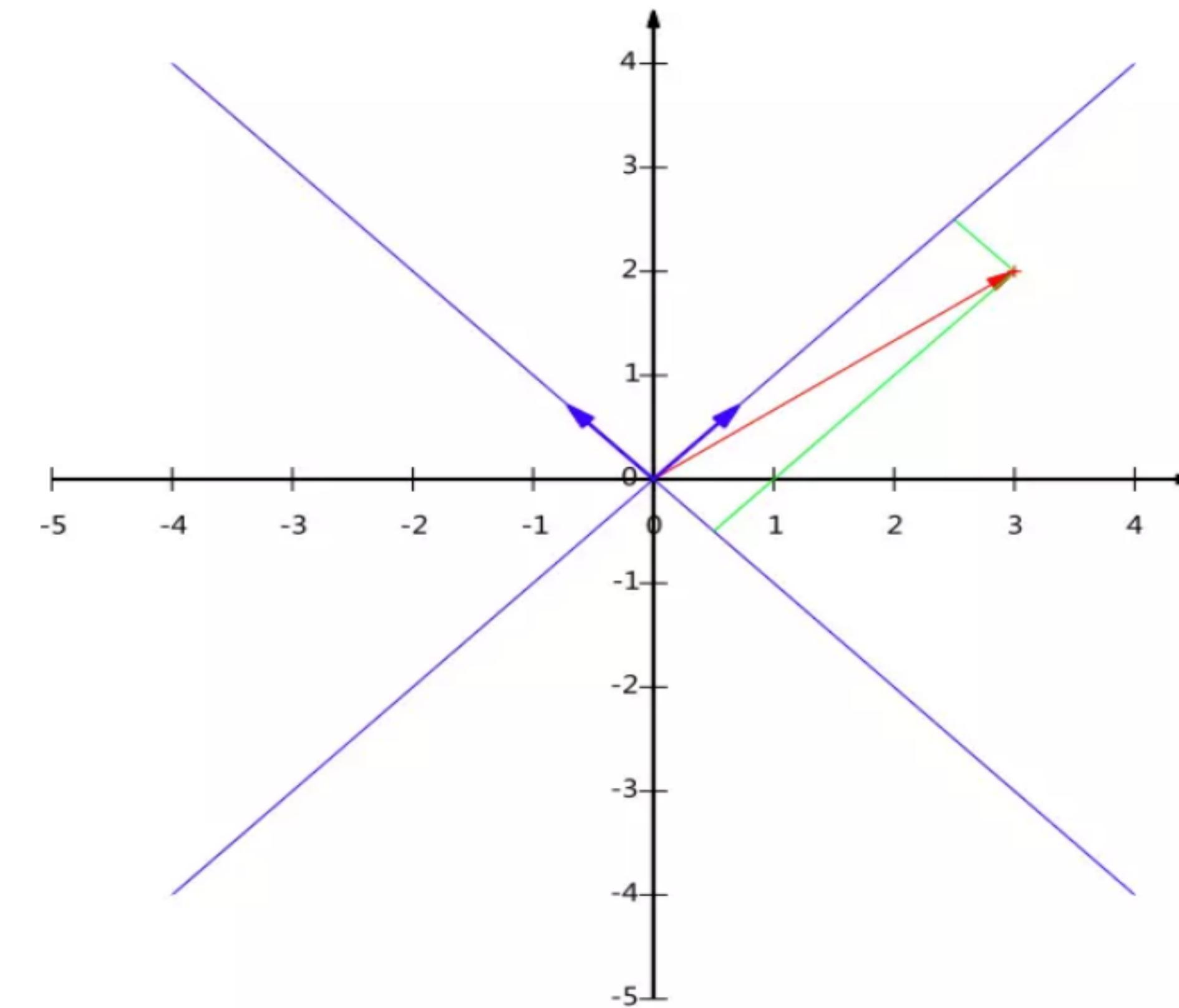


PCA 流程 (6)-基底轉換

- 2 個向量內積為 0(互相垂直) 且 xy 軸上 2 個向量的長度為線性不相關，資料不會互相影響

基底轉換方式

數值與第一個基做內積運算，並將結果做為新的座標向量，再與第二個內積做運算，得到第二個座標向量



PCA 完整流程

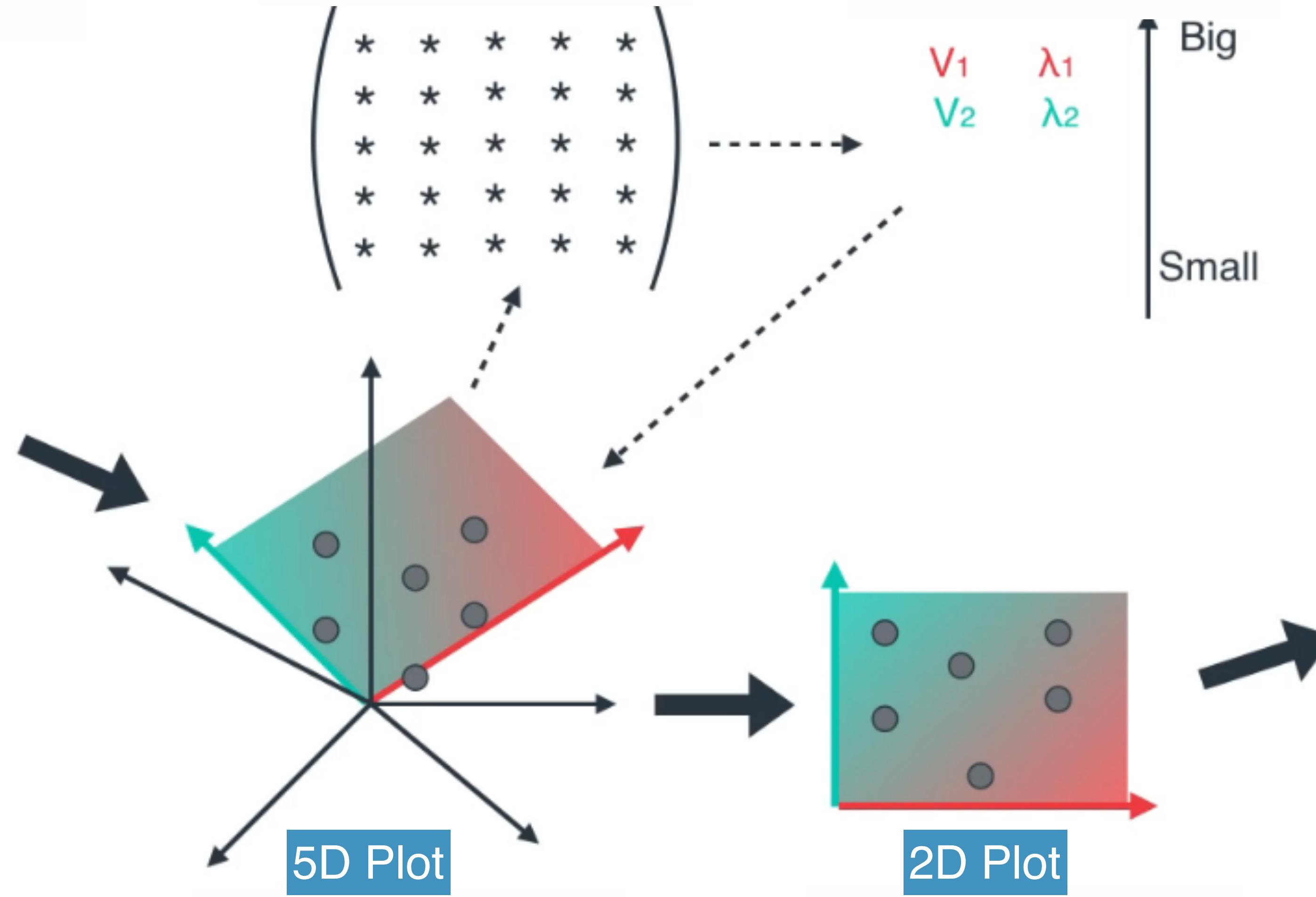
PCA

Large Table

標準化資料集

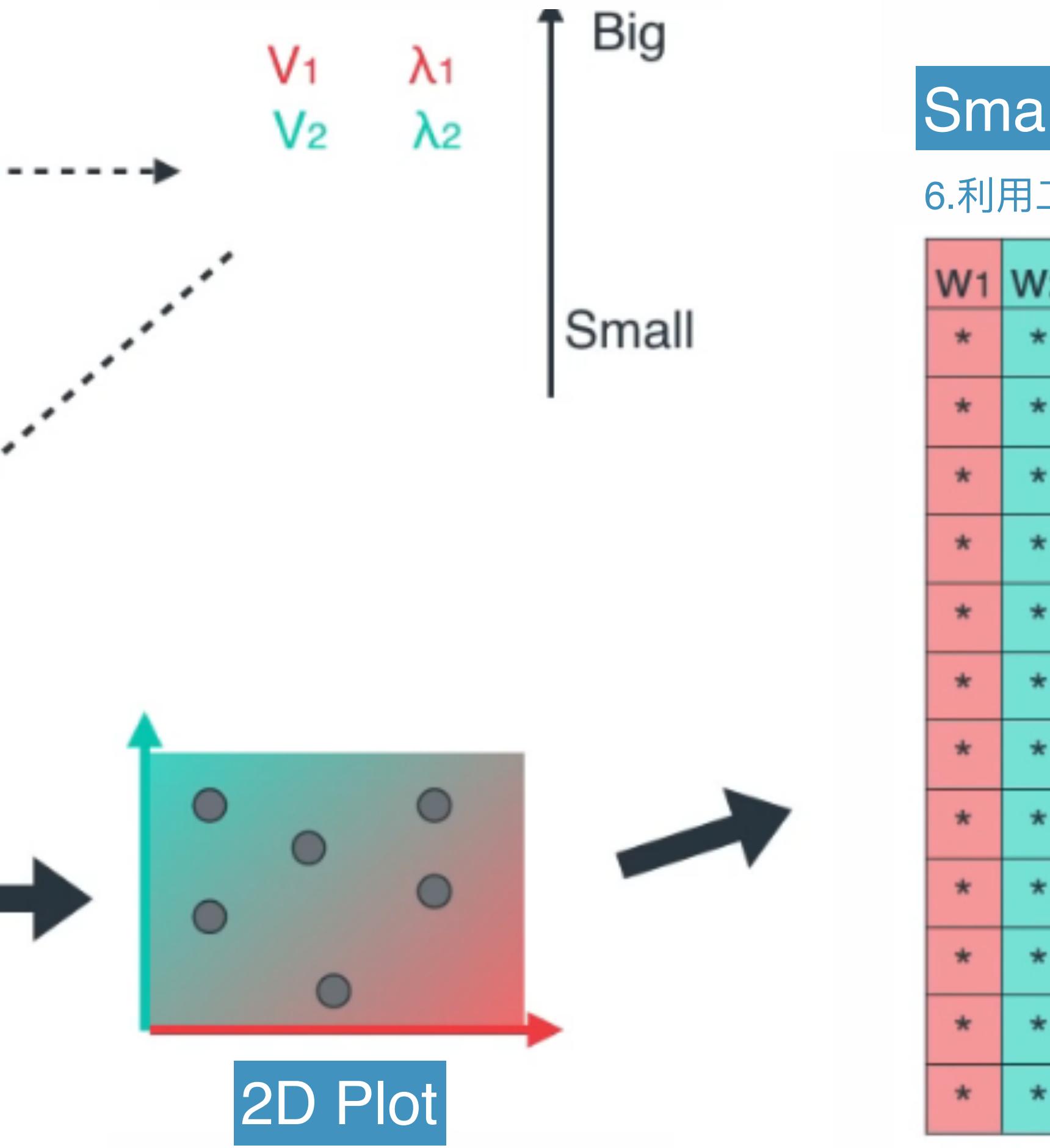
Covariance matrix

2. 將資料映射到5維空間



Eigenstuf

3. 建立共變異數矩陣，並分解特徵向量與特徵值



Small Table

6.利用二維平面顯示降維後的結果

應用：加速監督式學習

- 組合出來的這些新的 features 可以進而用來做 supervised learning 預測模型
- 以判斷人臉為例，最重要的特徵是眼睛、鼻子、嘴巴，膚色和頭髮等都可以捨棄，將這些不必要的資訊捨棄除了可以加速 learning，也可以避免一點 overfitting

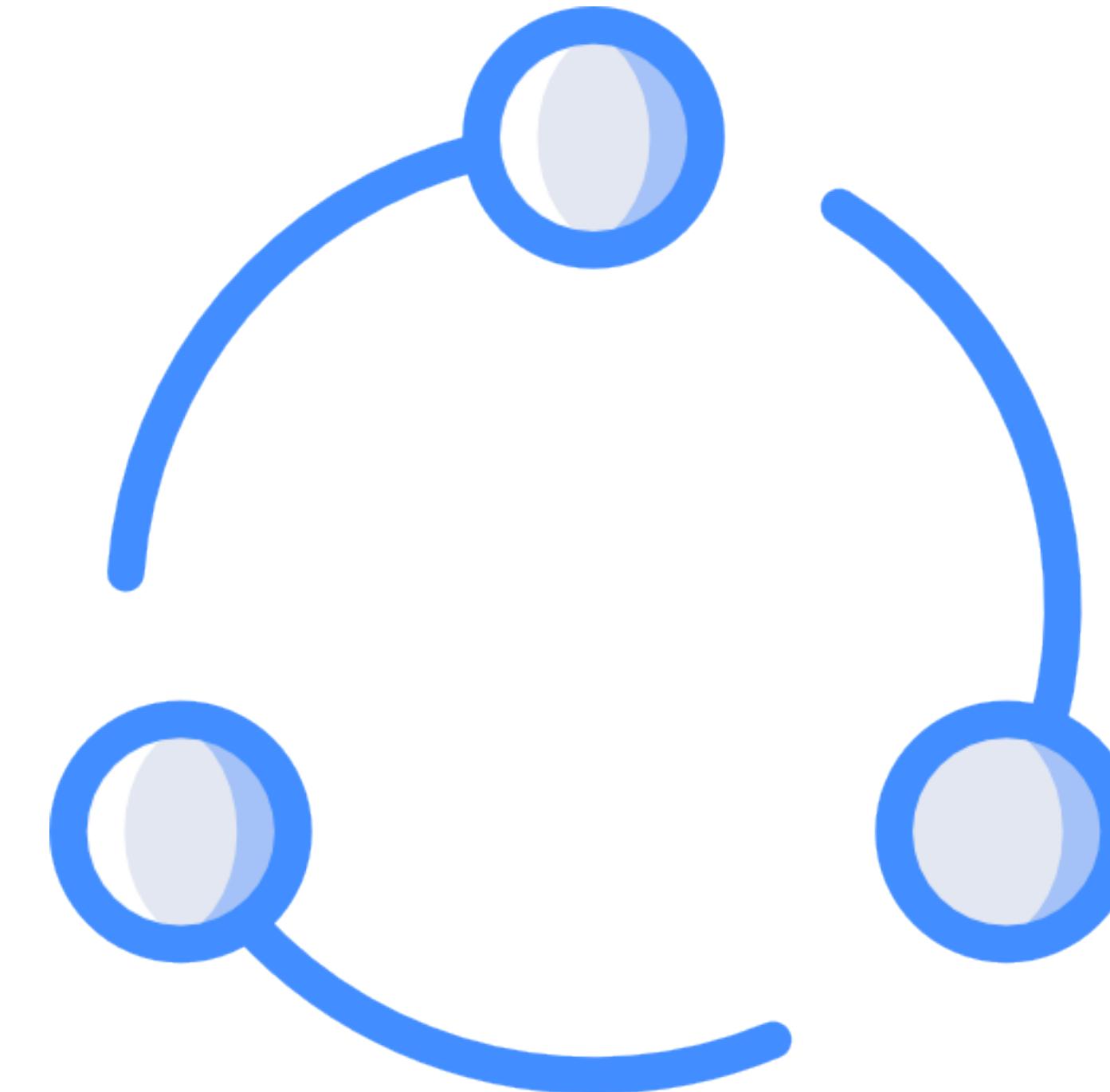
PCA 應用在監督式學習的注意事項

- 不建議在早期時做，否則可能會丟失重要的 features 而 underfitting
- 可以在 optimization 階段時，考慮 PCA，並觀察運用了 PCA 後對準確度的影響



重要知識點複習

- 降低維度可以幫助我們壓縮及丟棄無用資訊、抽象化及組合新特徵、呈現高維數據
- 常用的算法為**主成分分析**
- 在維度太大發生 overfitting 的情況下，可以嘗試用 PCA 組成的特徵來做監督式學習，但不建議一開始就做



PCA參考影片

- 主成份分析(作者:中國大學MOOC-慕課)
<https://www.youtube.com/watch?v=5MbJOnUZDc>
- PCA降維概述(作者:小文)
<https://www.youtube.com/watch?v=lvgYo1qeGZc>
- PCA要優化的目標(作者:小文)
<https://www.youtube.com/watch?v=HU8WuvMdTVE>
- PCA求解(作者:小文)
<https://www.youtube.com/watch?v=lsJmWBZvzf0>
- Visual Explanation of Principal Component Analysis, Covariance, SVD (作者:mm Freedman)
https://www.youtube.com/watch?v=5HNr_j6LmPc
- Principal Component Analysis (PCA) (作者:Luis Serrano)
<https://www.youtube.com/watch?v=g-Hb26agBFg>
- Eigenvectors and eigenvalues I Essence of linear algebra, chapter 14 (特徵轉換) (作者:3Blue1Brown)
<https://www.youtube.com/watch?v=PFDu9oVAE-g>
- Principal component analysis (PCA)(降維動畫)(作者: Paulo Ricardo Gherardi Hein)
<https://www.youtube.com/watch?v=4pnQd6jnCWk>



解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題