

Dimension Reduction

PCA 、 t-SNE 、 Autoencoder

Outline

PCA

Autoencoder

t-SNE

如果要預測

顧客信用卡消費明細

> 15,000 間店家

	國泰人壽	福容大飯店	台塑加油站	大潤發	高鐵	HOLA	...
顧客 1	1	0	0	0	0	0	...
顧客 2	0	0	0	0	1	1	...
顧客 3	0	0	1	0	0	0	...
...
顧客 n	0	1	0	1	0	1	

> 640,000 位顧客

預測變數

- 男性 / 女性
- 公教人員
- 薪轉戶
- 理財 VIP 會員
- 世界卡顧客
- 行員
- ...



如果資料量更大、記憶體更小呢？
n 個 feature 轉換為 n 個 component
但全部都用的到嗎？



能不能只用部分的特徵數，去

完美詮釋所有的資料，這就是降維

那什麼是 PCA ?

(Principal Component Analysis)

將具有 N 個特徵空間的樣本，轉換為具有 K 個特徵空間
的樣本，其中 $K < N$ (通常 $K=2$)，並找到一個特徵
空間，使得某點與投影在特徵空間上的點距離最小

聽不懂 沒關係

林俊傑



OFFICIAL
Audio

Taihe Music



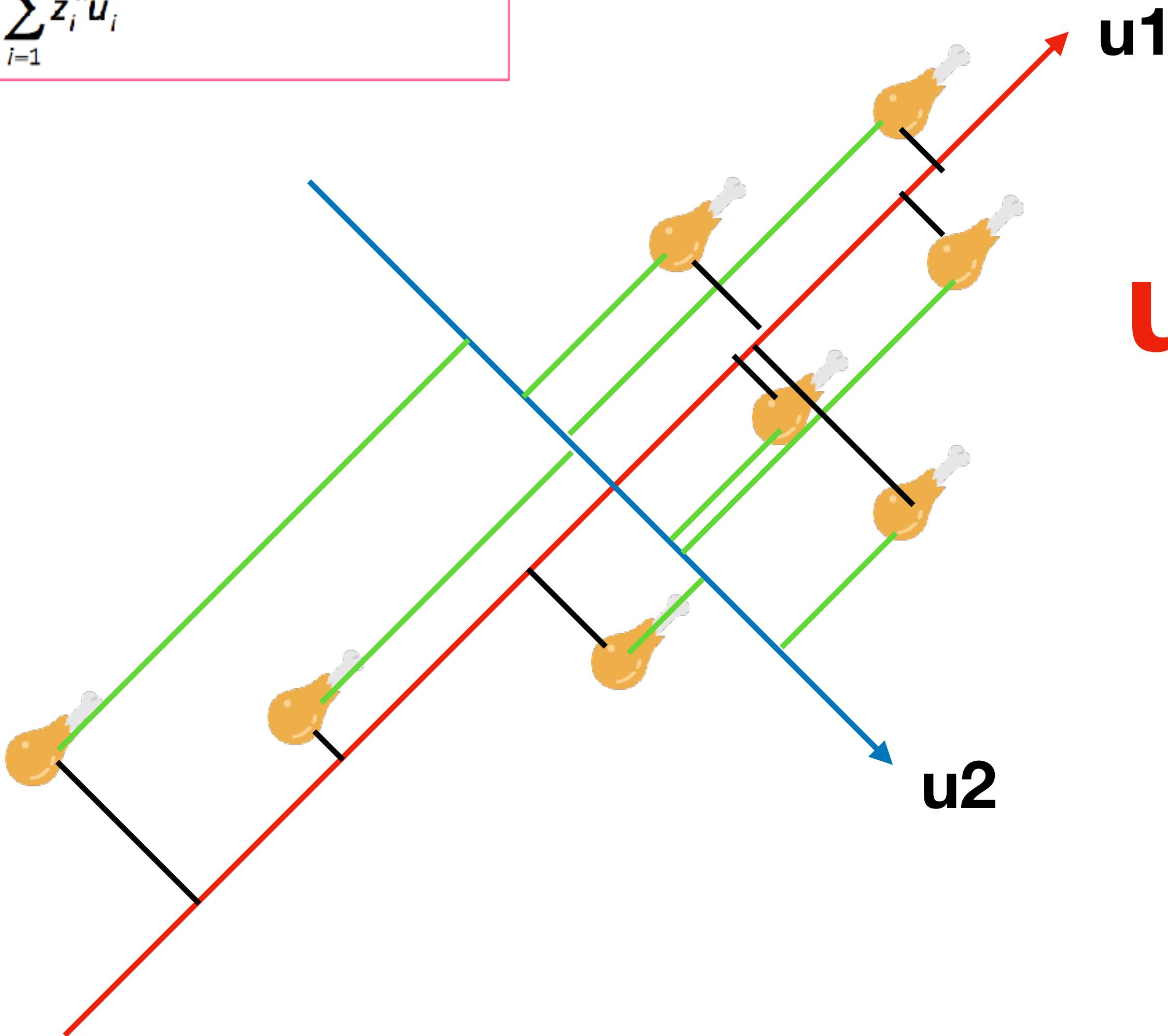
把一個勁辣雞腿堡完全壓扁，在這個
壓扁的平面上以剩下的兩軸(主成份)，重
建一個平面座標系來表達雞腿的位置



PCA: given $M < d$. Find $(u_1 \dots u_M)$

that minimizes $E_M = \sum_{k=1}^d \|x_k - \hat{x}_k\|_2^2$

where $\hat{x}_k = \bar{x} + \sum_{i=1}^M z_i^k u_i$

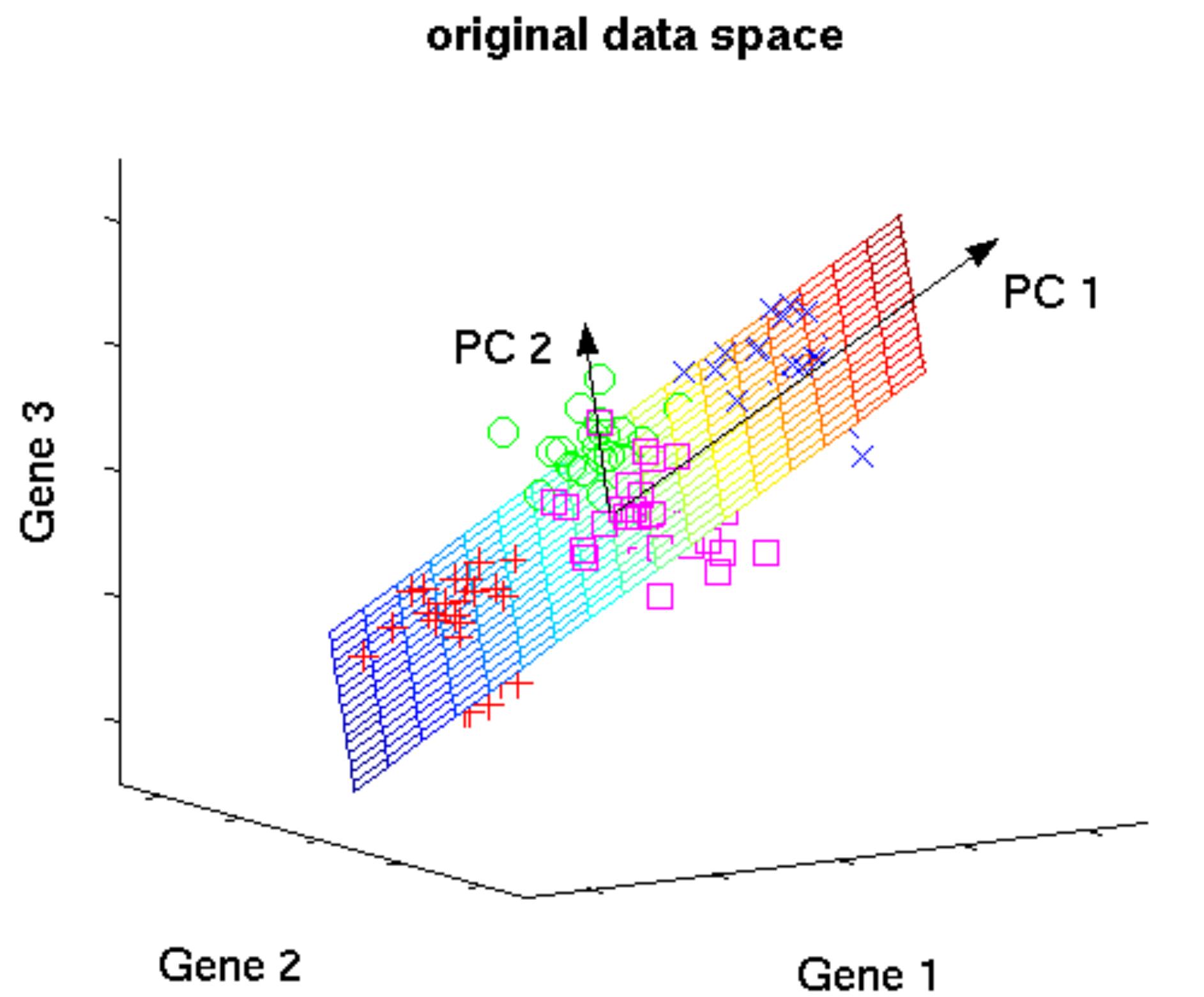


u1 作為主成份
好過 u2

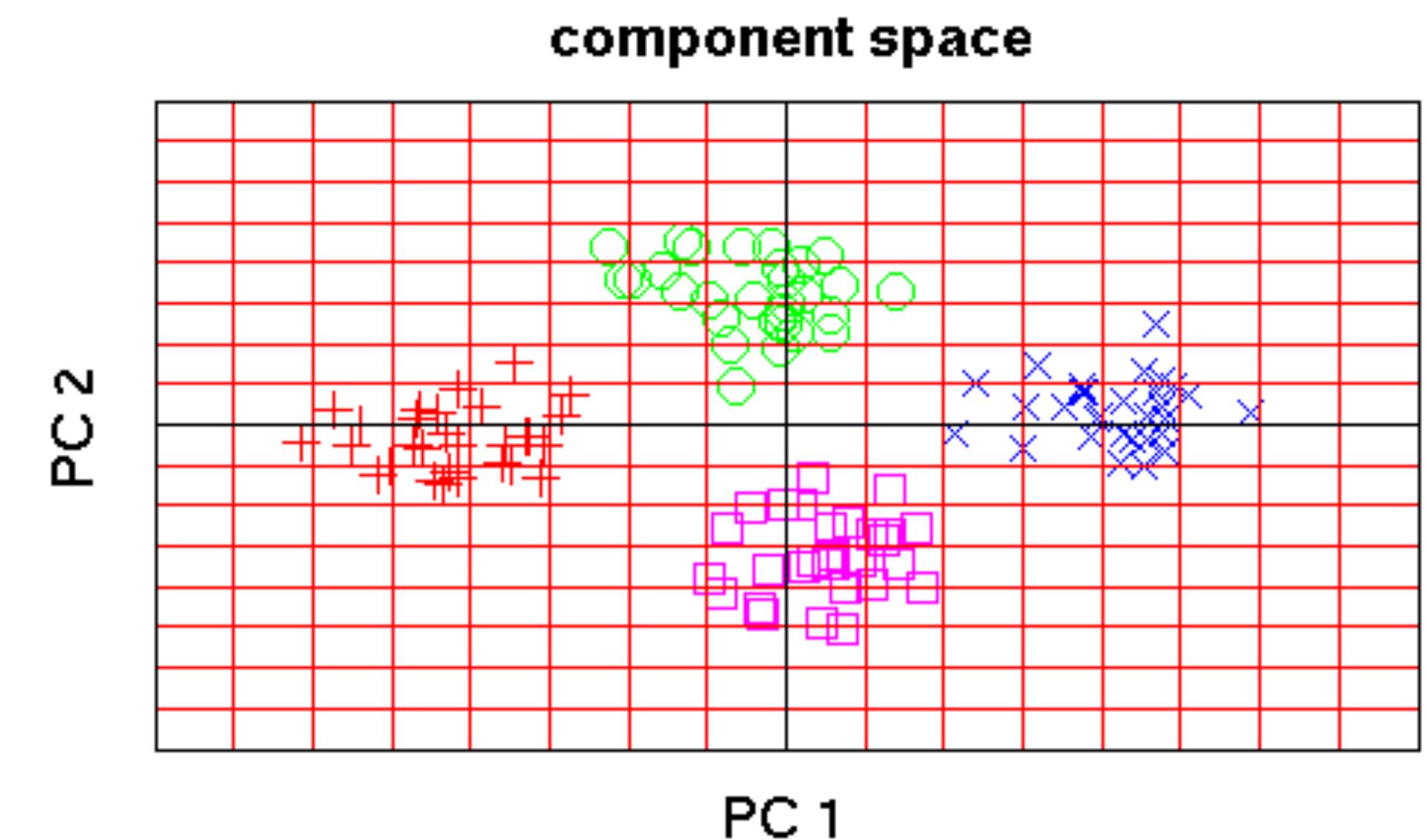


PCA

通常取前2名主成份作降維代表



PCA



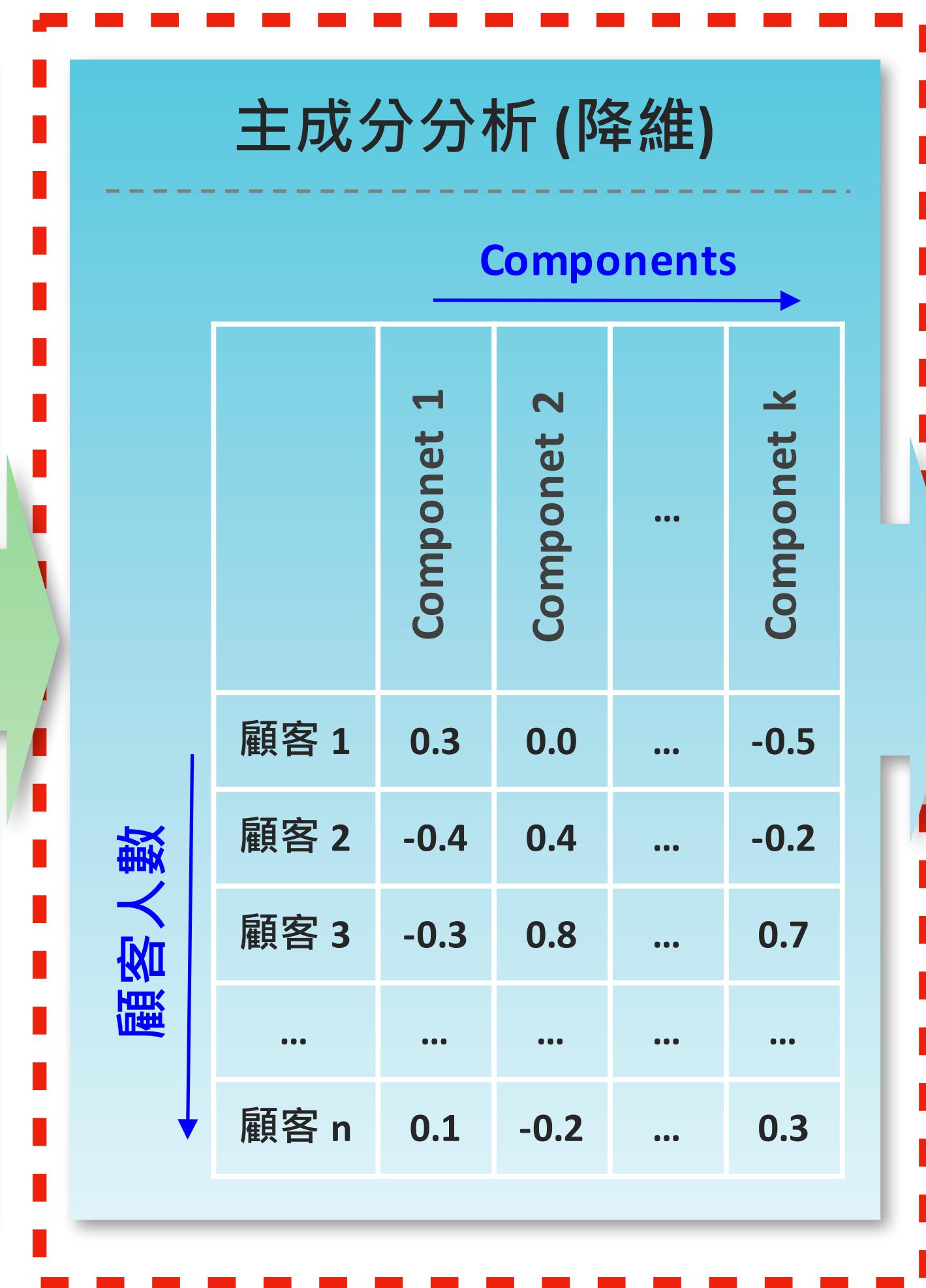
透過PCA

顧客信用卡消費明細

> 15,000 間店家

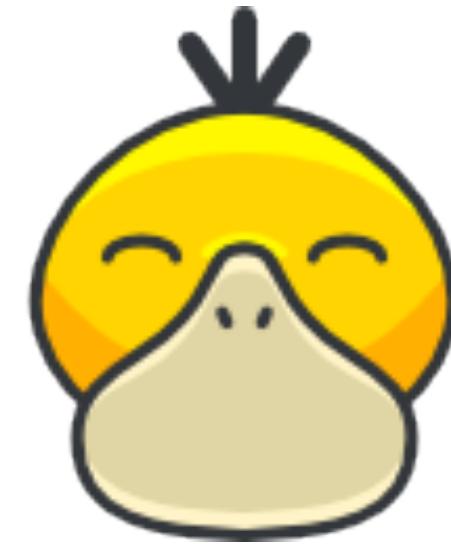
> 640,000 位顧客

	國泰人壽	福容大飯店	...
顧客 1	1	0	
顧客 2	0	0	
顧客 3	0	0	...
...	
顧客 n	0	1	



PCA的限制

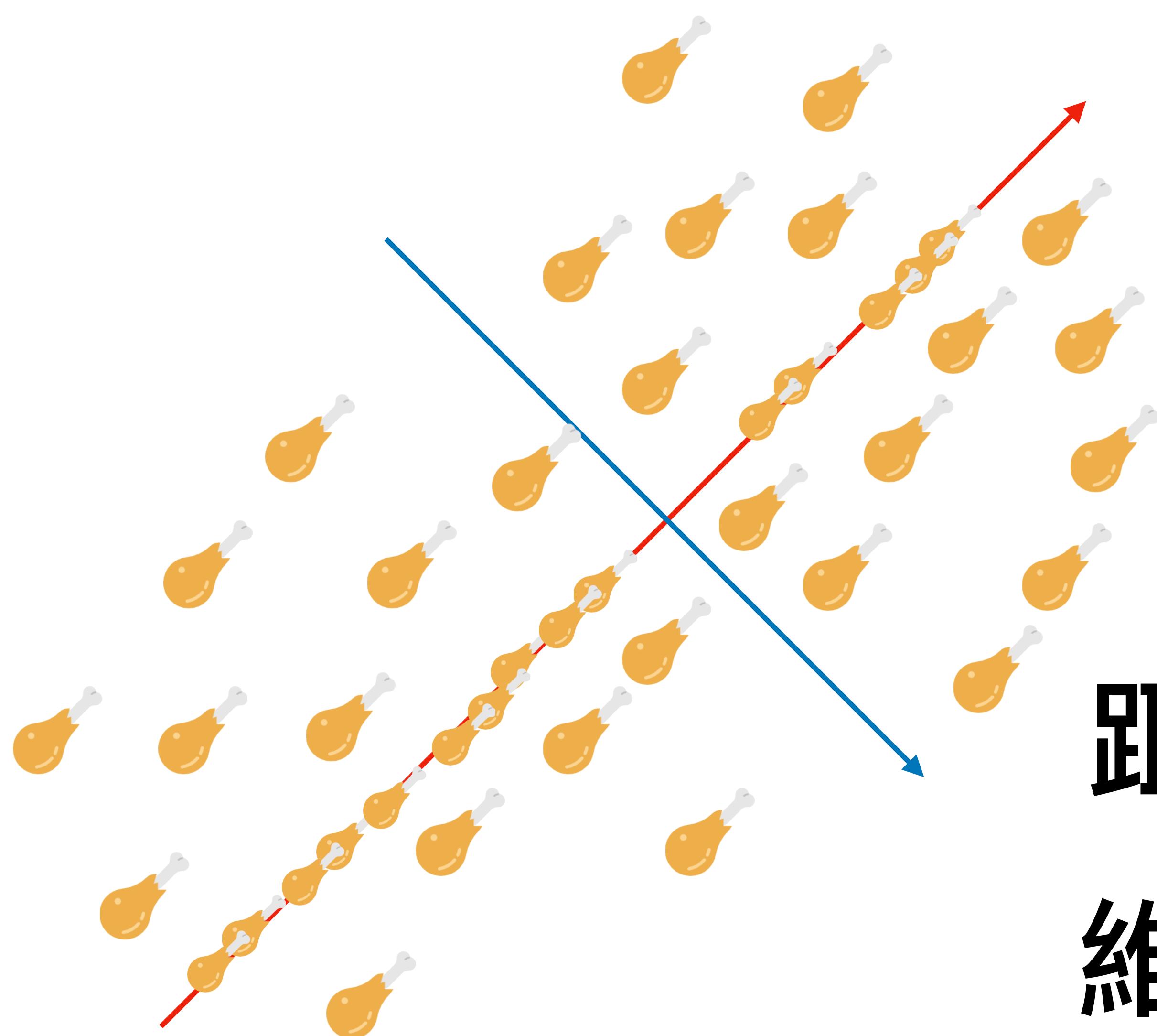




是一種線性降維的方式，若特徵間為

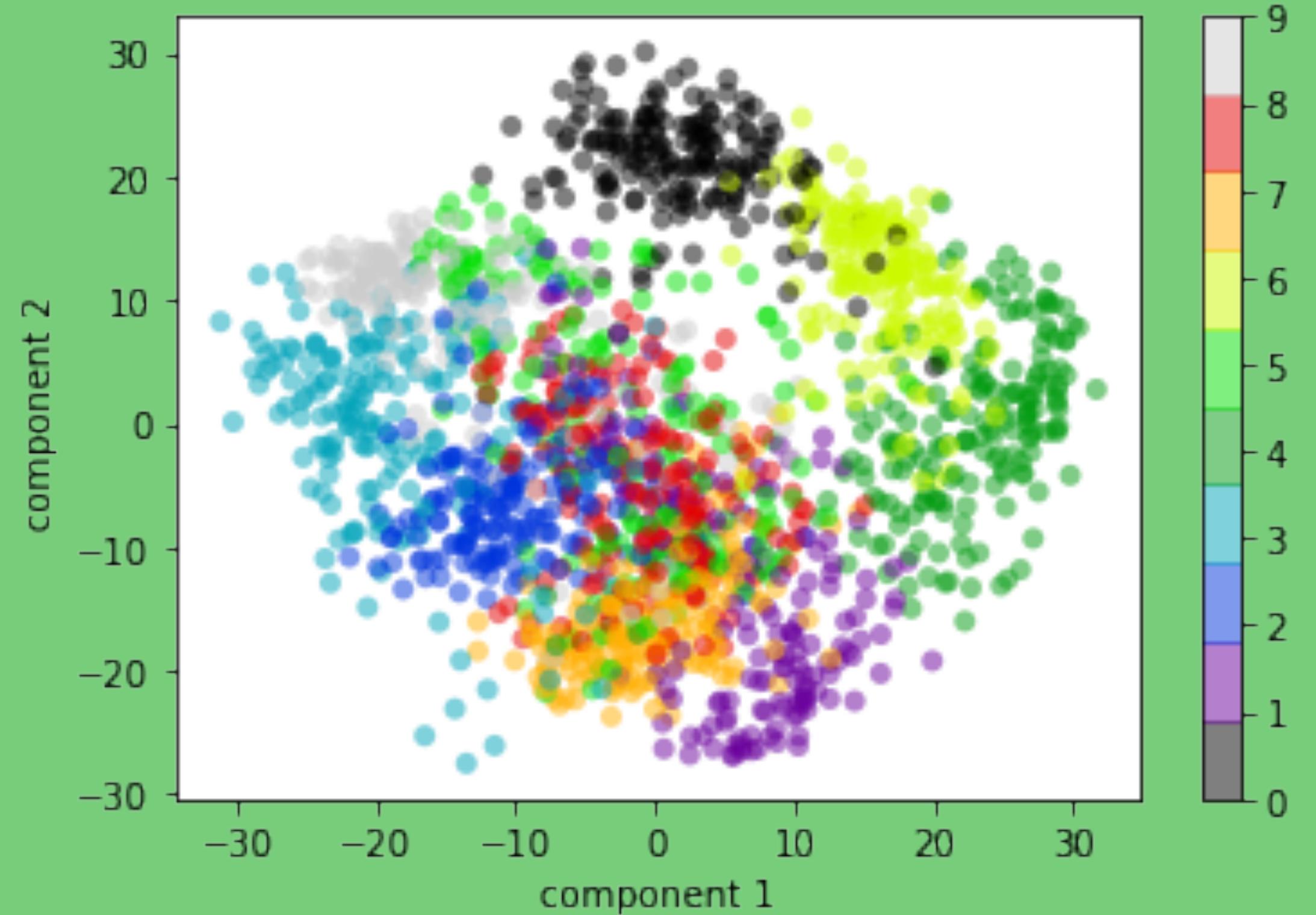
非線性關係，容易造成Underfitting



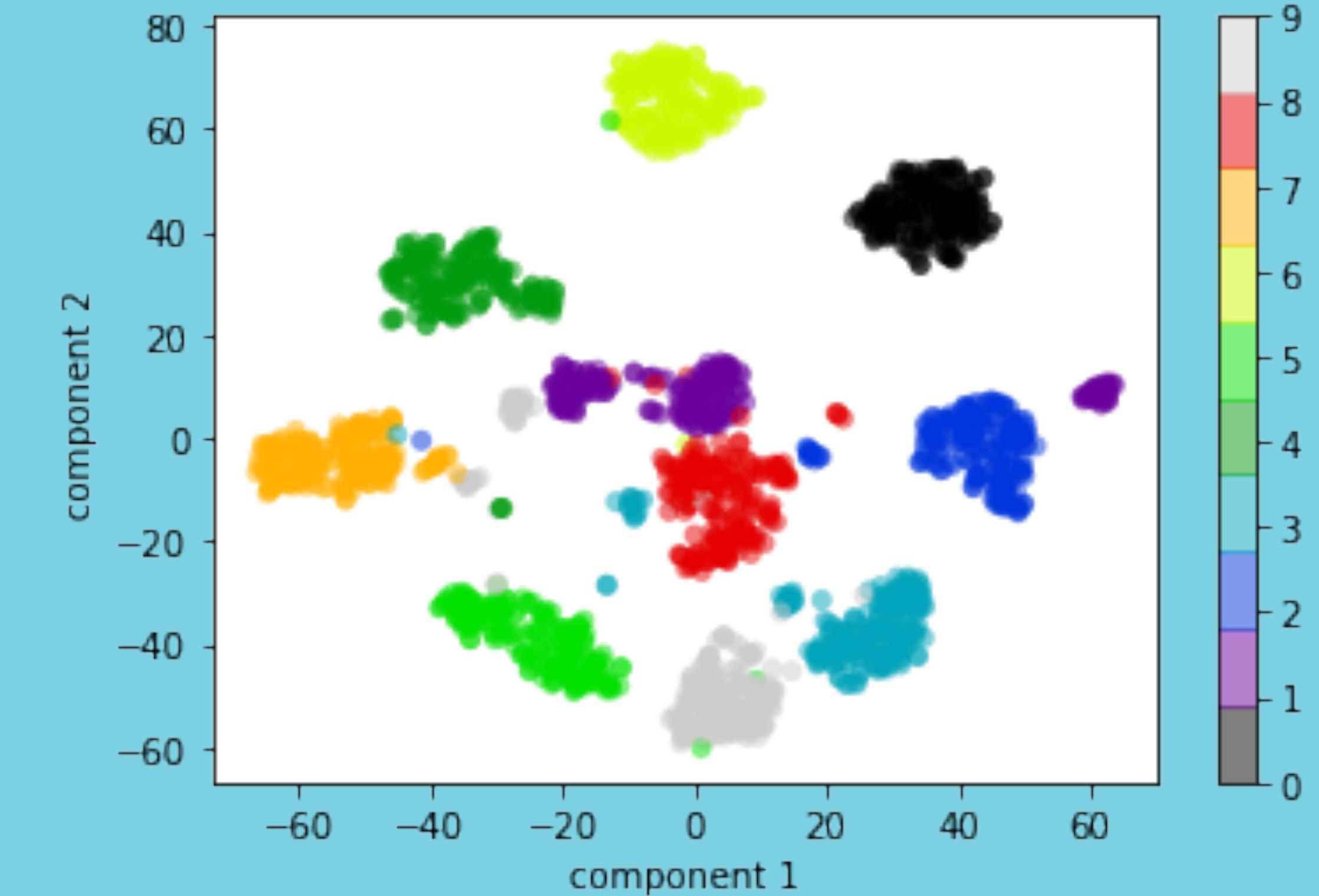


距離近的雞腿，降到低
維會重疊無法辨識，
有沒有其他映射方式

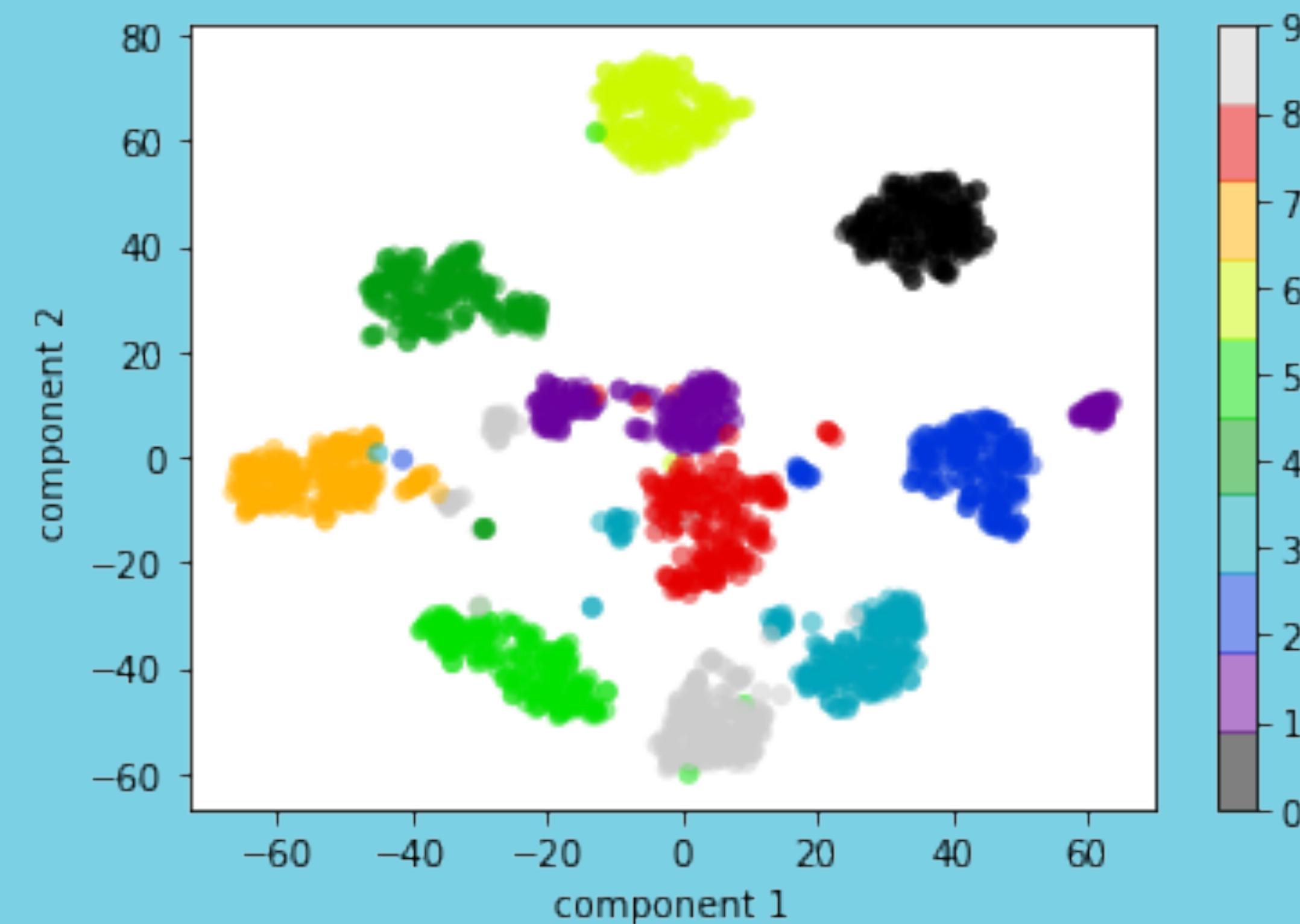
PCA



t-SNE



t-SNE



那什麼是 t-SNE ?
(t-distributed stochastic
neighbor embedding)

在講t-SNE之前，
請容許我先介紹SNE

以往表示相似性的做法，都是用歐式距離，而SNE的概念是，把這種距離關係轉換為以條件機率來表示。在高維空間相似的數據點，映射到低維空間距離也是相似的。

高維

X_i



X_j



越靠近



, $P_{j|i}$ 越大

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

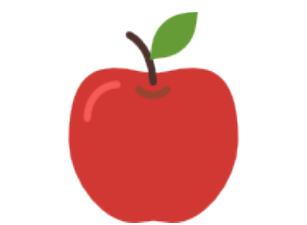
$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

低維

Y_i



Y_j



越靠近



, $Q_{j|i}$ 越大

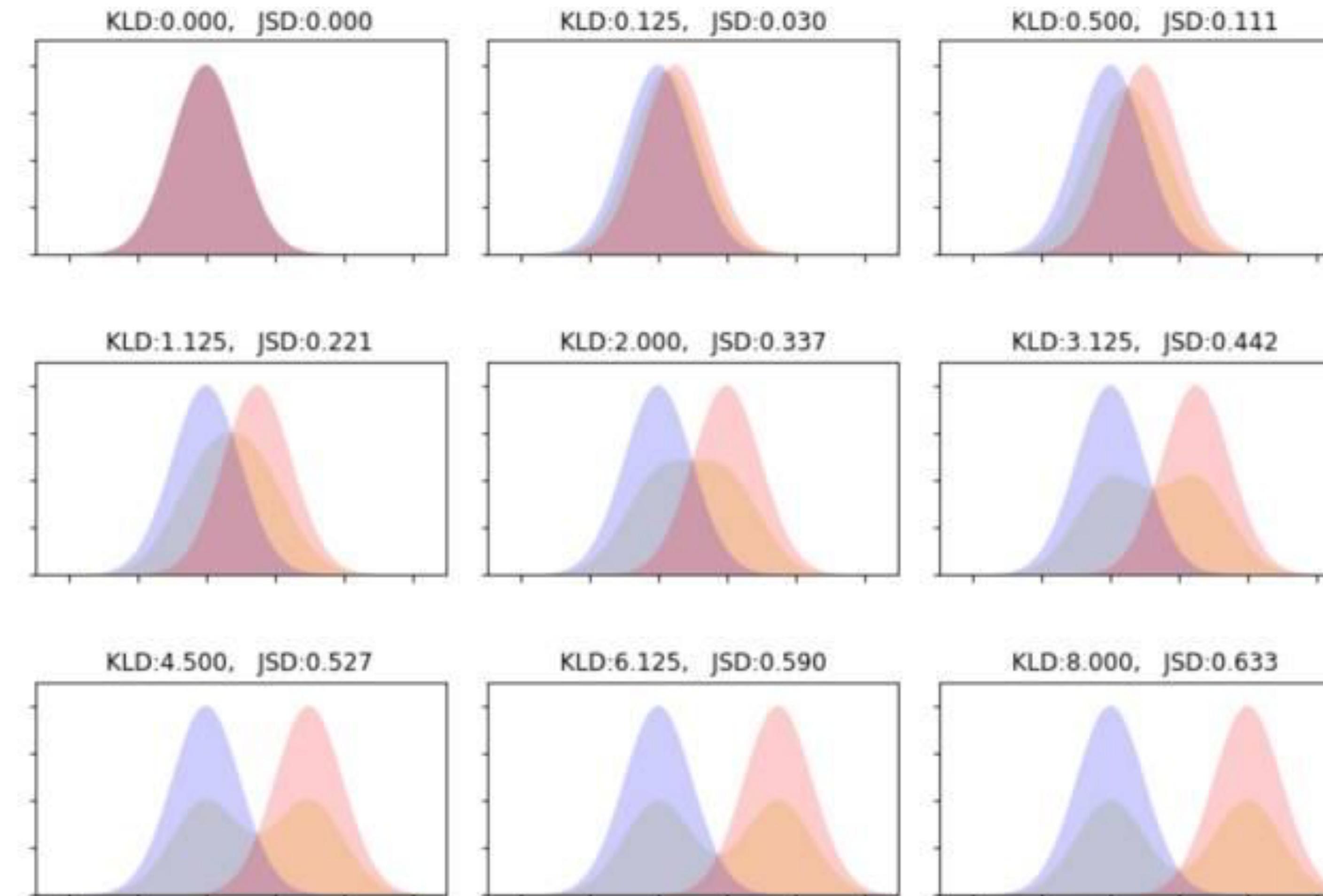
目標 : $P_{j|i}$ 、 $Q_{j|i}$ 分佈越相近越好

怎麼做 : KL 散度

KL 散度

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

KL散度越大，分布相似度越低



問題1



越靠近



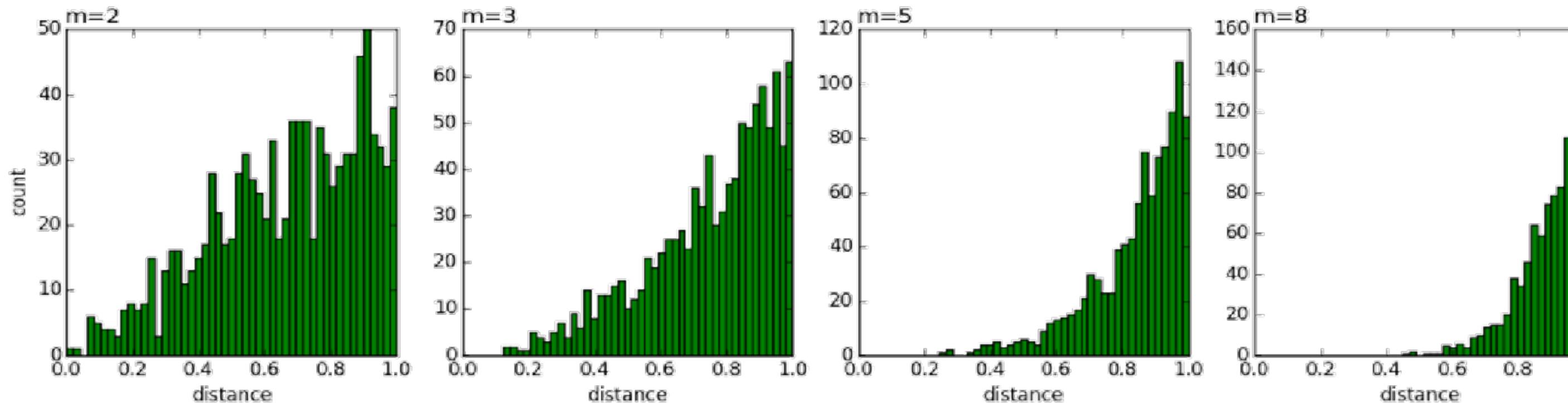
， $P_{j|i}$ 越大

條件機率非對稱性：

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

$P_{j|i} \neq P_{i|j}$ 且 $Q_{i|j} \neq Q_{j|i}$ 且高維空間中兩點相距較遠時，
若映射到低維空間距離較近，反而得到很低的懲罰，**有問題！**

問題2



Crowding Problem(低維空間問題)：

想像一顆以點 X_i 為中心，半徑 r 的 m 維球，其他點與 X_i 隨維度增
大距離越不均衡，若壓縮到**低維**，便會出現擠壓問題

問題怎麼解決？

t-SNE

問題1 $P_{j|i} \neq P_{i|j}$ 且 $Q_{i|j} \neq Q_{j|i}$

解法

以聯合機率分佈得到 $P_{ij} = P_{ji}$

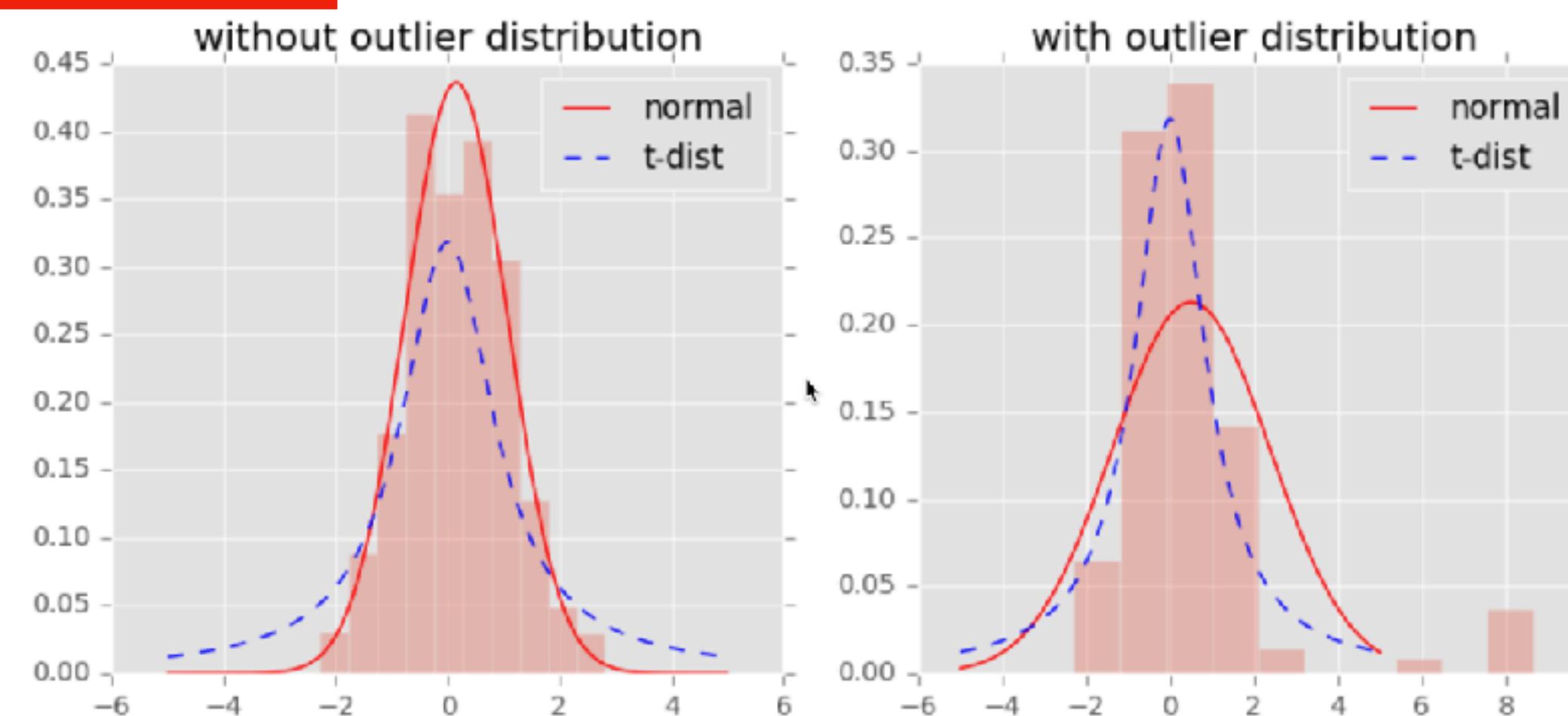
$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}$$

問題2

Crowding Problem：壓縮到低維，會出現擠壓問題

解法

t就是T分佈，對異常點較不敏感，低維空間時取代



高斯分佈，自由度越大，
趨近於常態分佈



不如一槍打死我



t-SNE 將高維的數據用高斯分佈的機率密度函數近似，
再將低維數據的部分使用t分佈的方式來近似，接著使用
KL距離計算相似度，最後再以梯度下降求最佳解。



t-SNE的限制

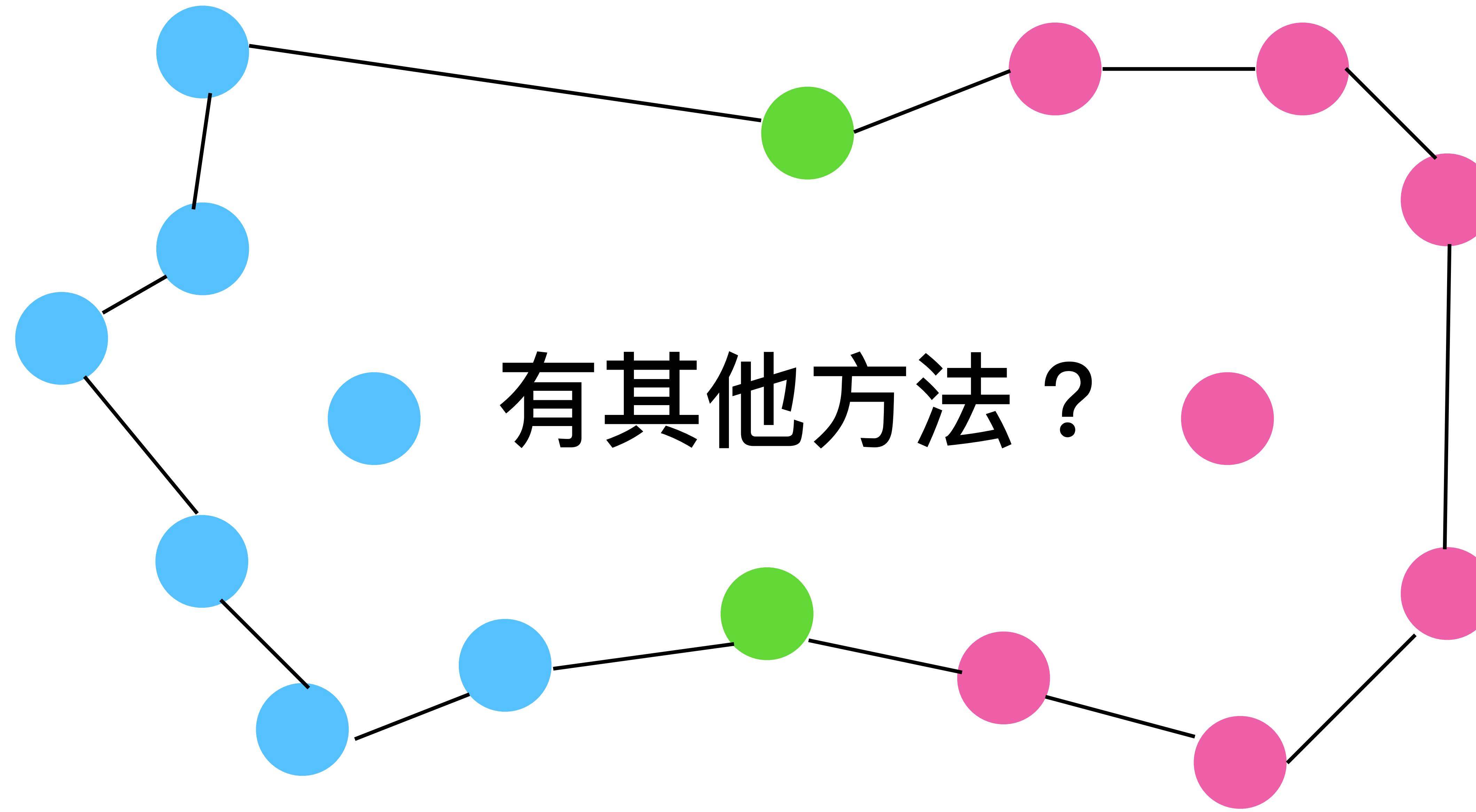


降維是透過機率分佈的方式，無法應用
在全新資料集，通常僅拿來做成可視化

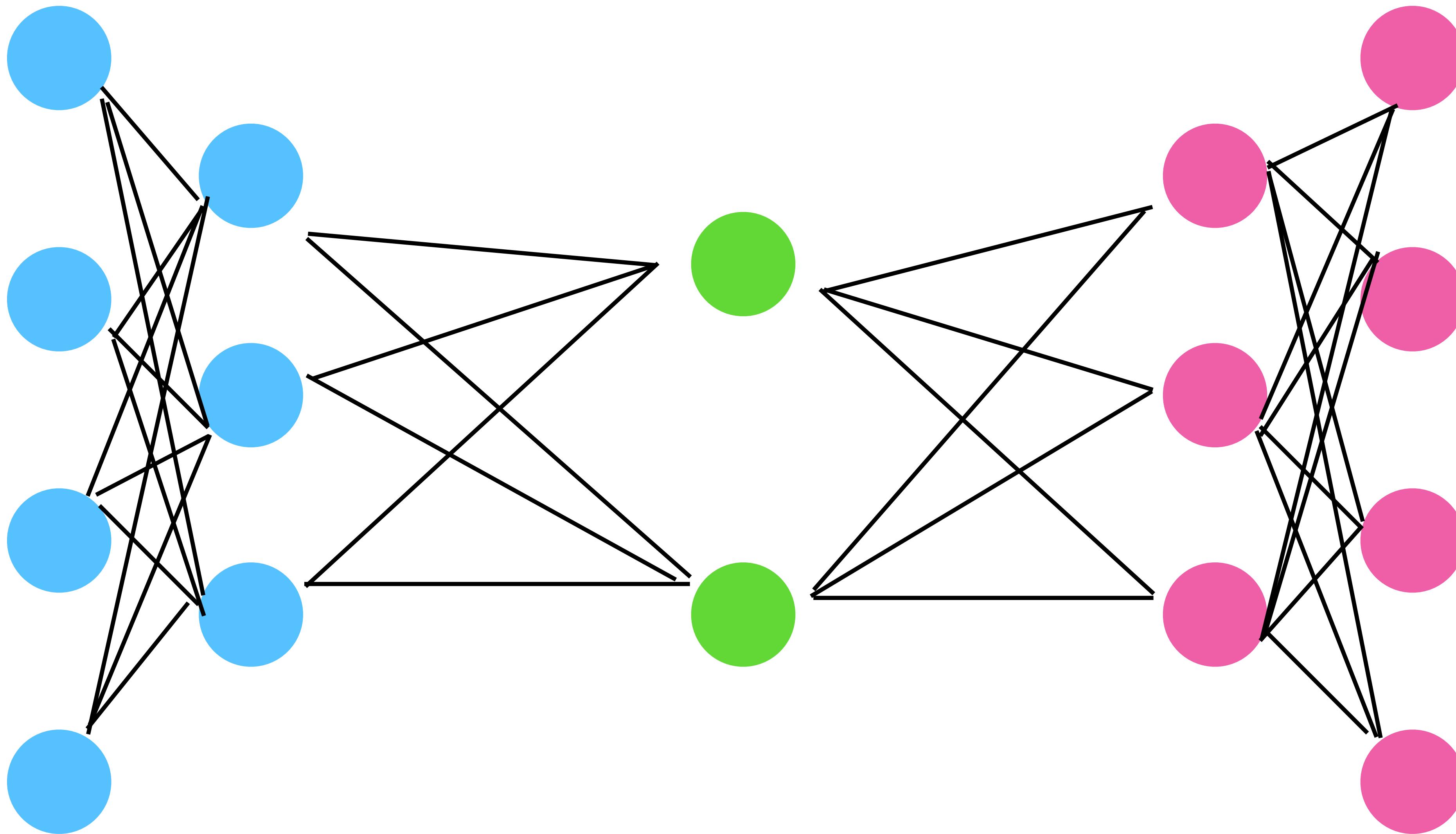


	PCA	t-SNE
Definitions	<ul style="list-style-type: none"> • Calculated through eigenvalues and eigenvectors. • <u>Linear reduction.</u> 	<ul style="list-style-type: none"> • Calculated through probability distribution. • Also for <u>non-linear reduction.</u>
Stochasticity	<ul style="list-style-type: none"> • PCA is <u>deterministic</u> 	<ul style="list-style-type: none"> • t-SNE is <u>not</u>.
Application to new Data	<ul style="list-style-type: none"> • Eigenvectors offer a new axes system what can be used to project new data. 	<ul style="list-style-type: none"> • Learned by directly moving the data across the low dimensional space.
Time	<ul style="list-style-type: none"> • Fast. 	<ul style="list-style-type: none"> • Usually need more time.

有其他方法？



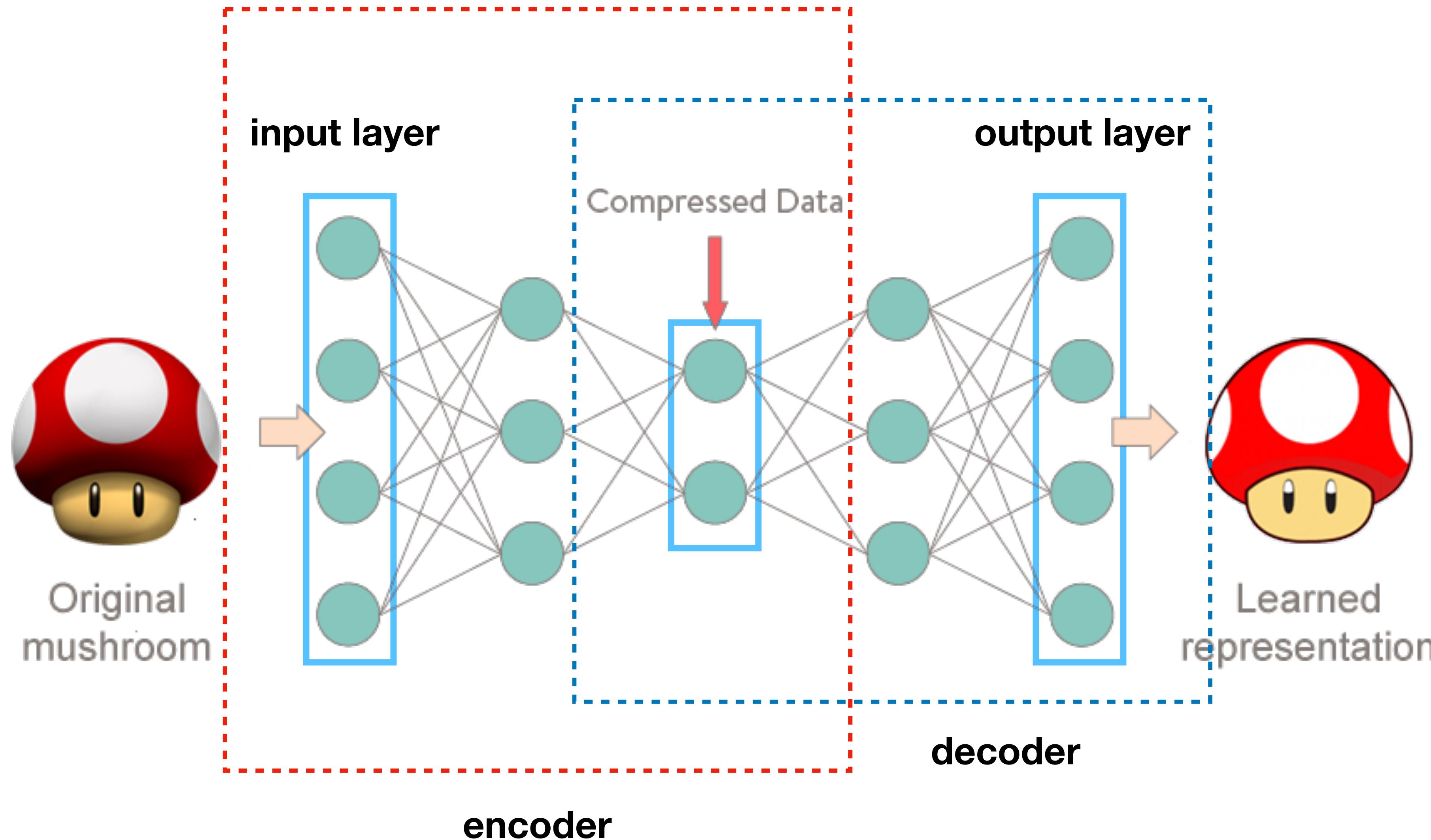
Autoencoder



什麼是 (AE)
Autoencoder ?

一種無監督式算法，利用反向傳播算法，使得目標值 = 輸入值

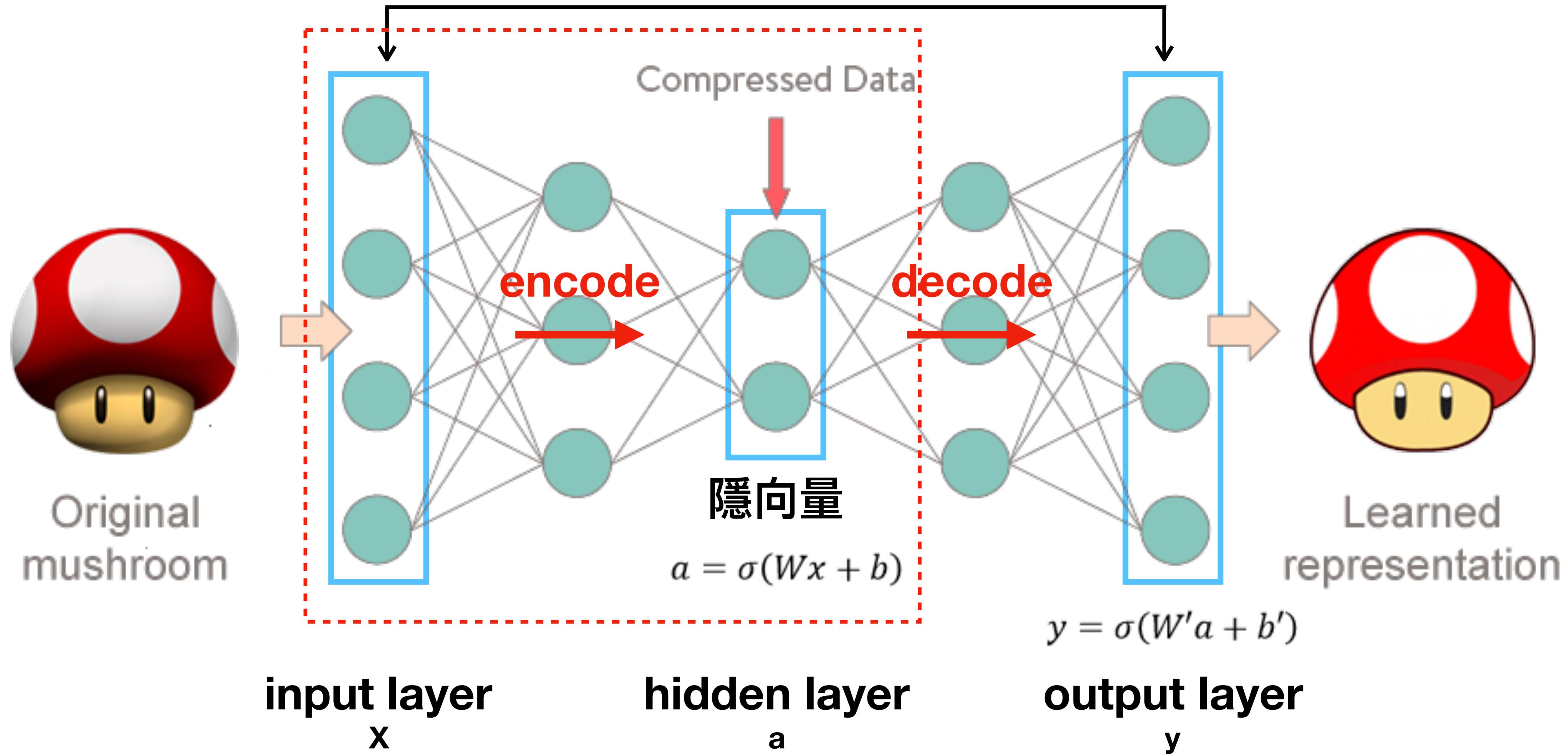
Autoencoder = encoder(降維) + decoder

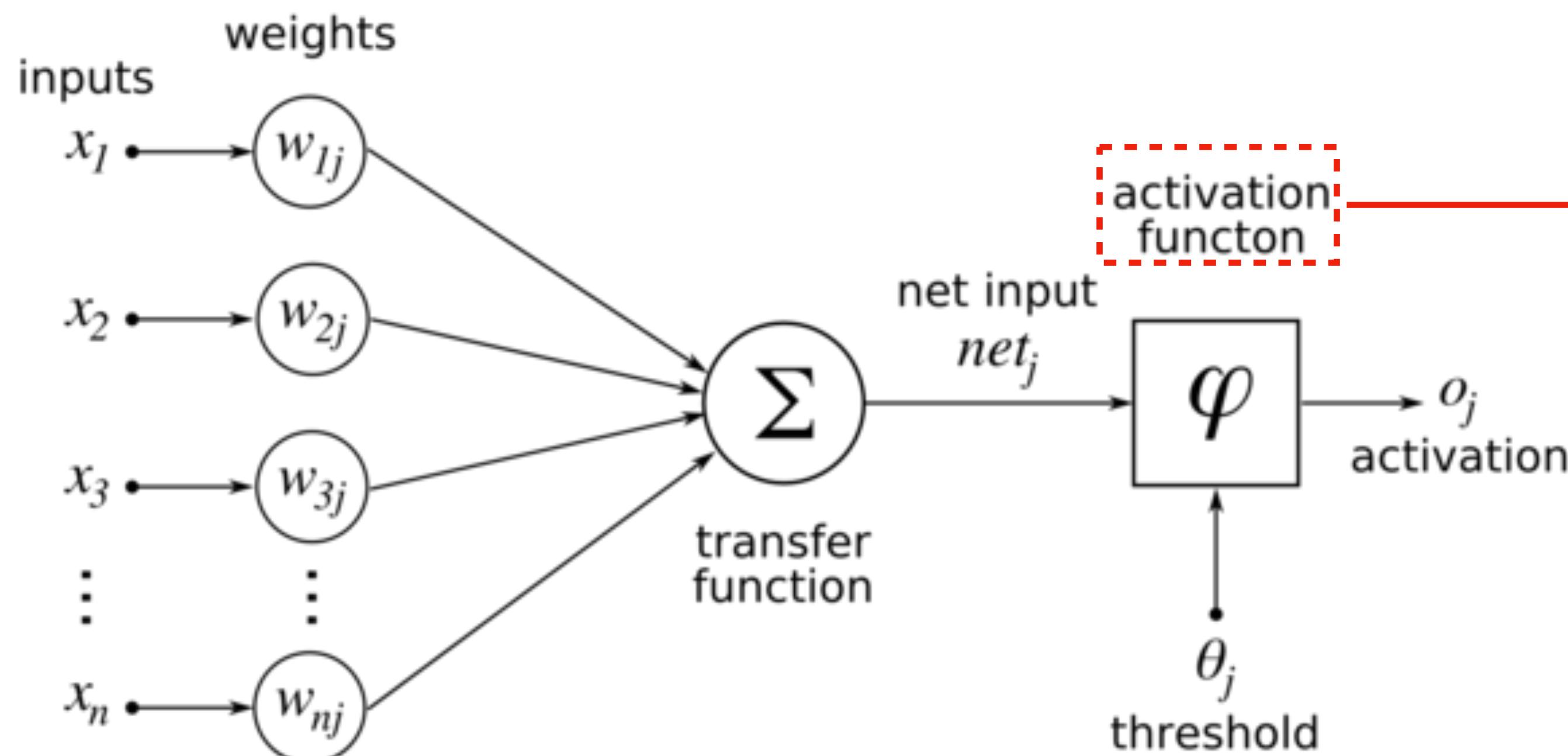
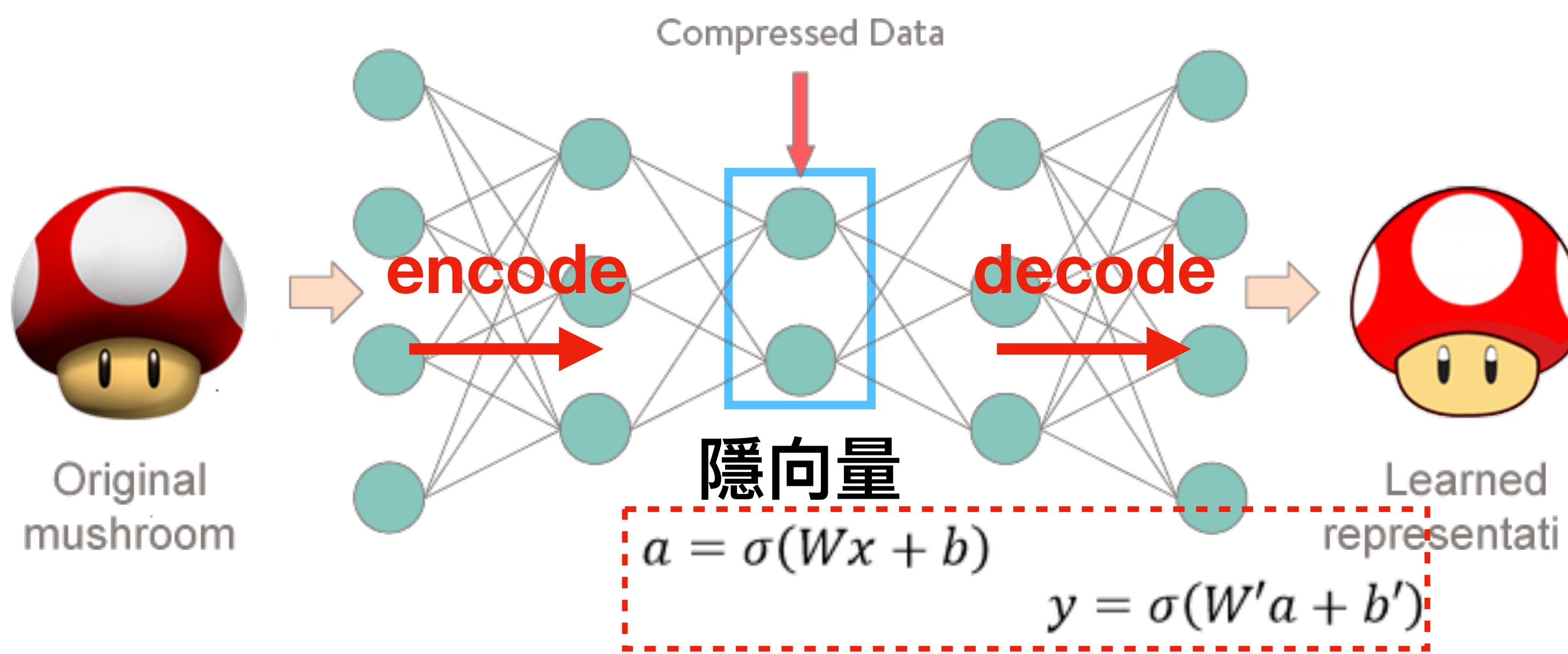


從有碼到無碼的還原過程

隱向量 = 精華層

Reconstruction error = minimize $(x-y)^2$



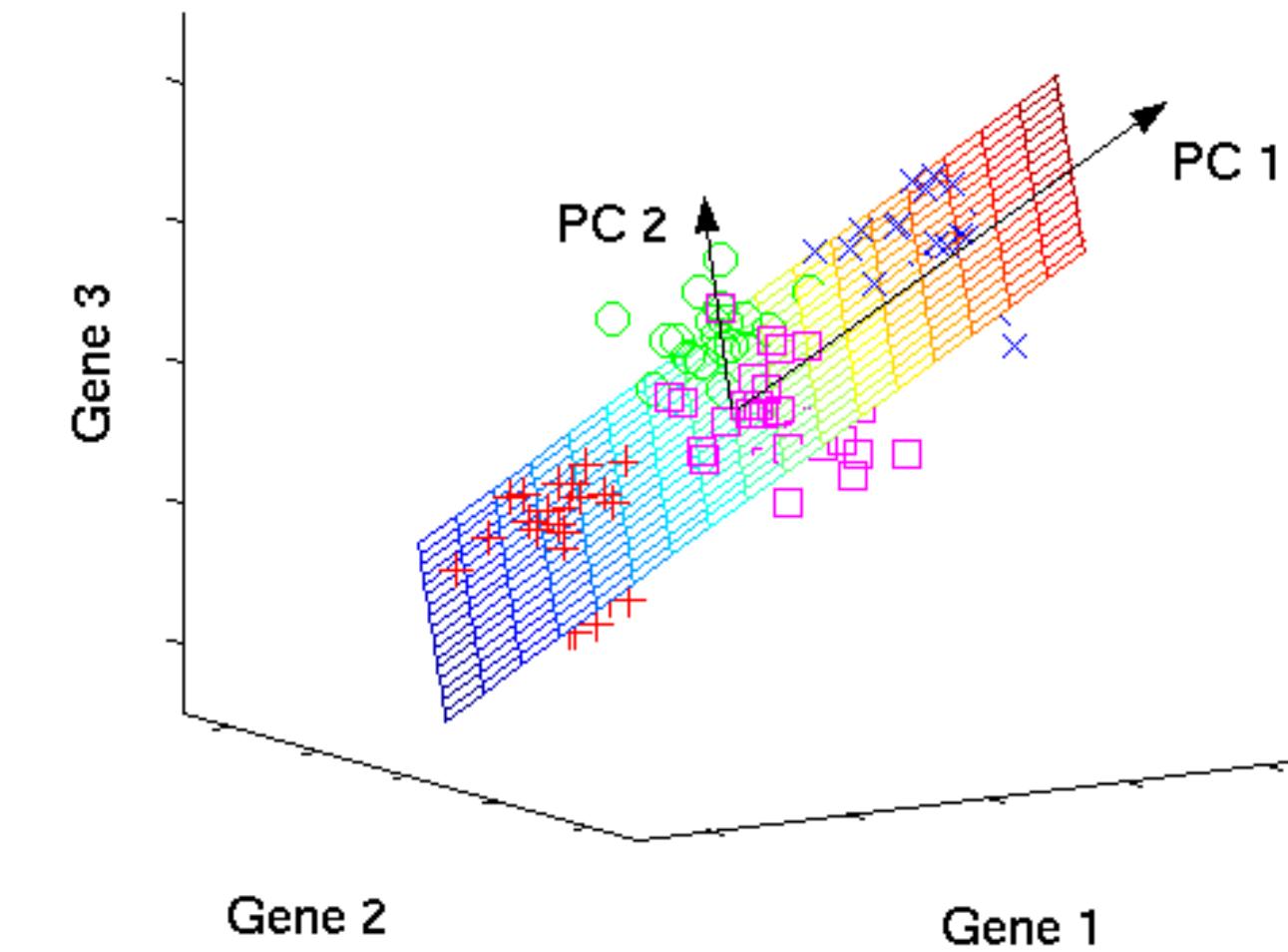


Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}. \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = \max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^z)$	Multi-layer Neural Networks	

Copyright © Sebastian Raschka 2016 (<http://sebastianraschka.com>)

你發現什麼了嗎？

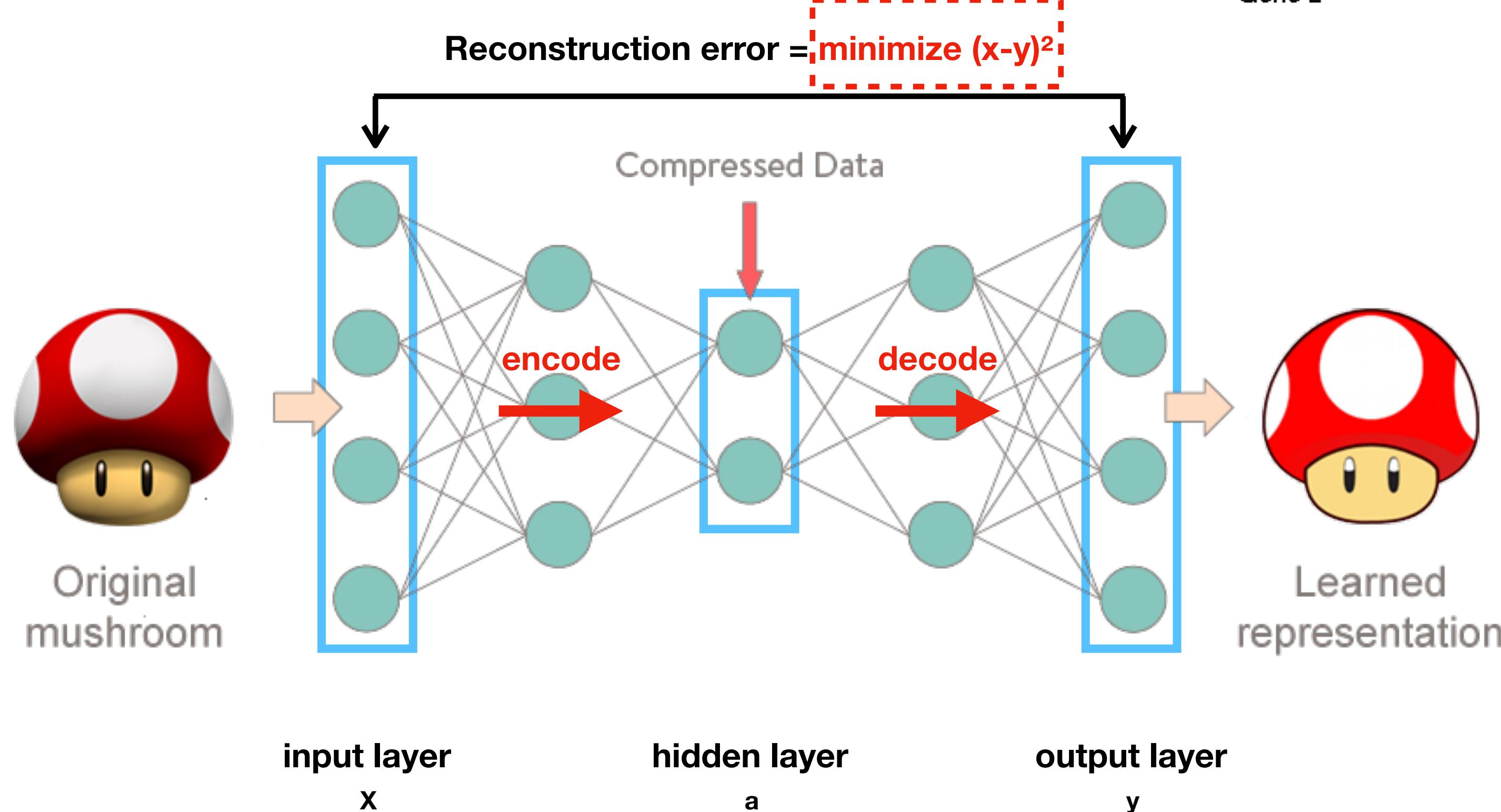
original data space



PCA: given $M < d$. Find $(u_1 \dots u_M)$

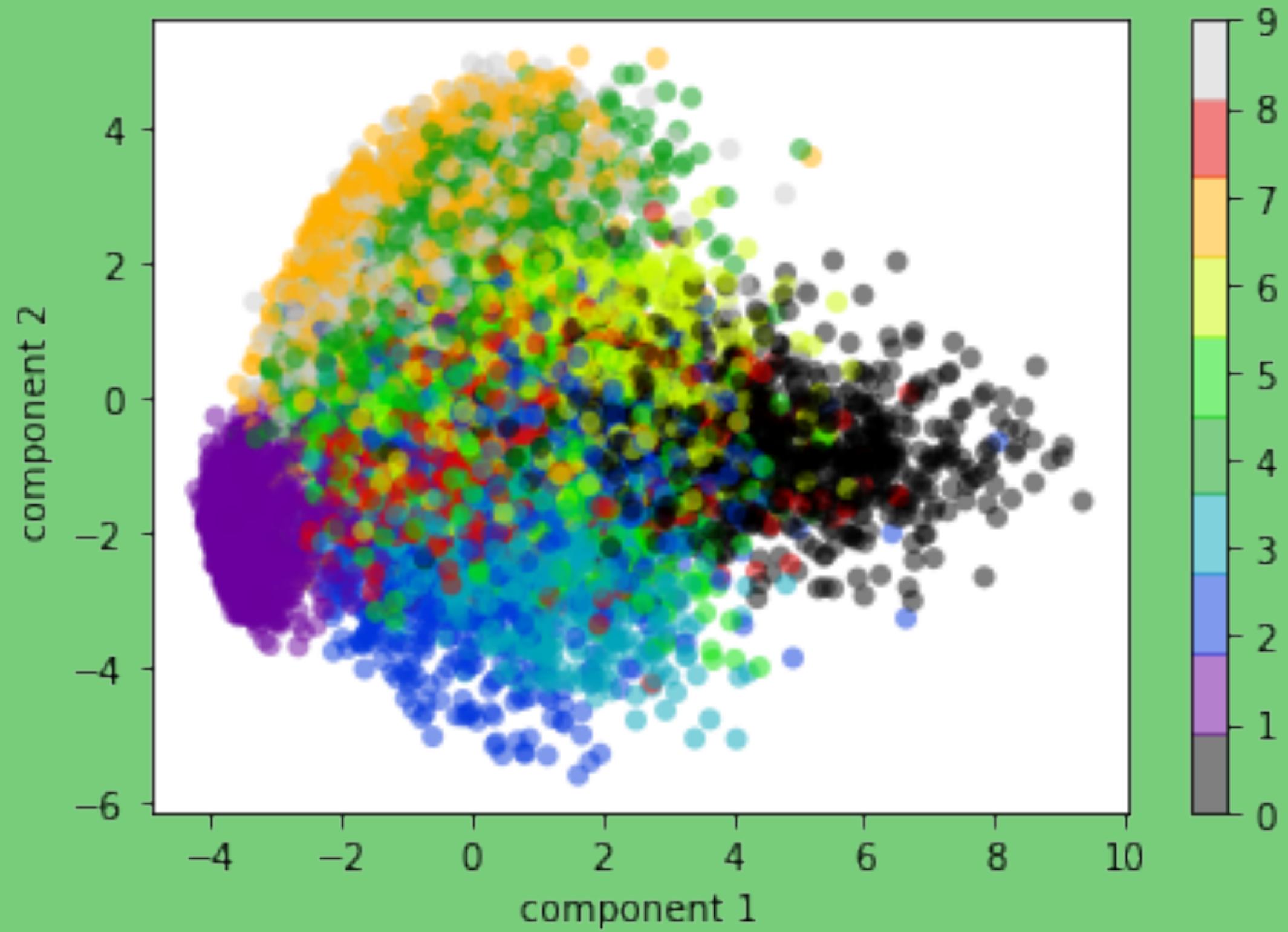
that minimizes $E_M = \sum_{k=1}^d \|x_k - \hat{x}_k\|_2^2$

$$\text{where } \hat{x}_k = \bar{x} + \sum_{i=1}^M z_i^k u_i$$

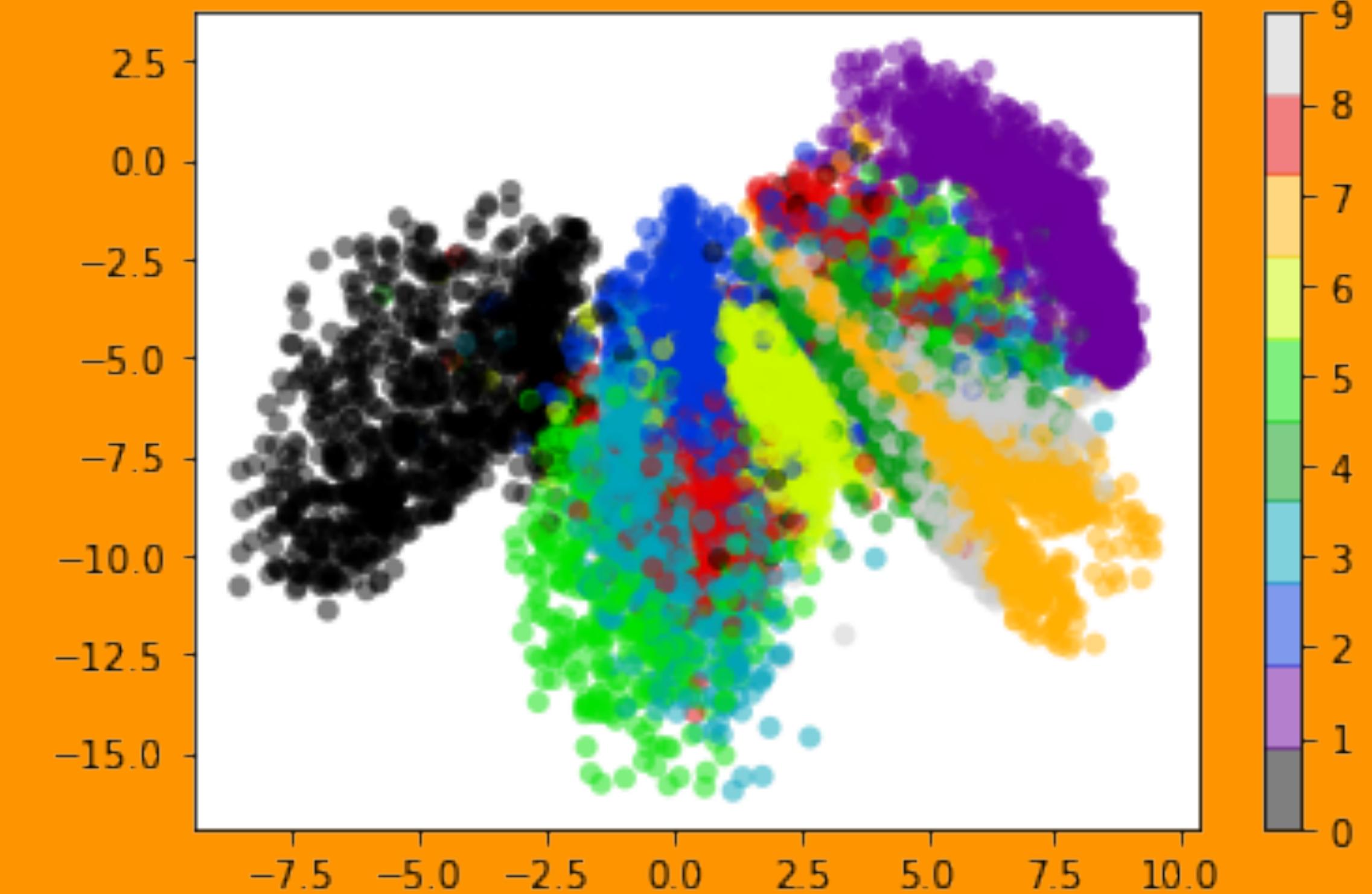


當autoencoder只有一個隱含層且
為線性的時候，其原理相當於PCA

PCA



Autoencoder



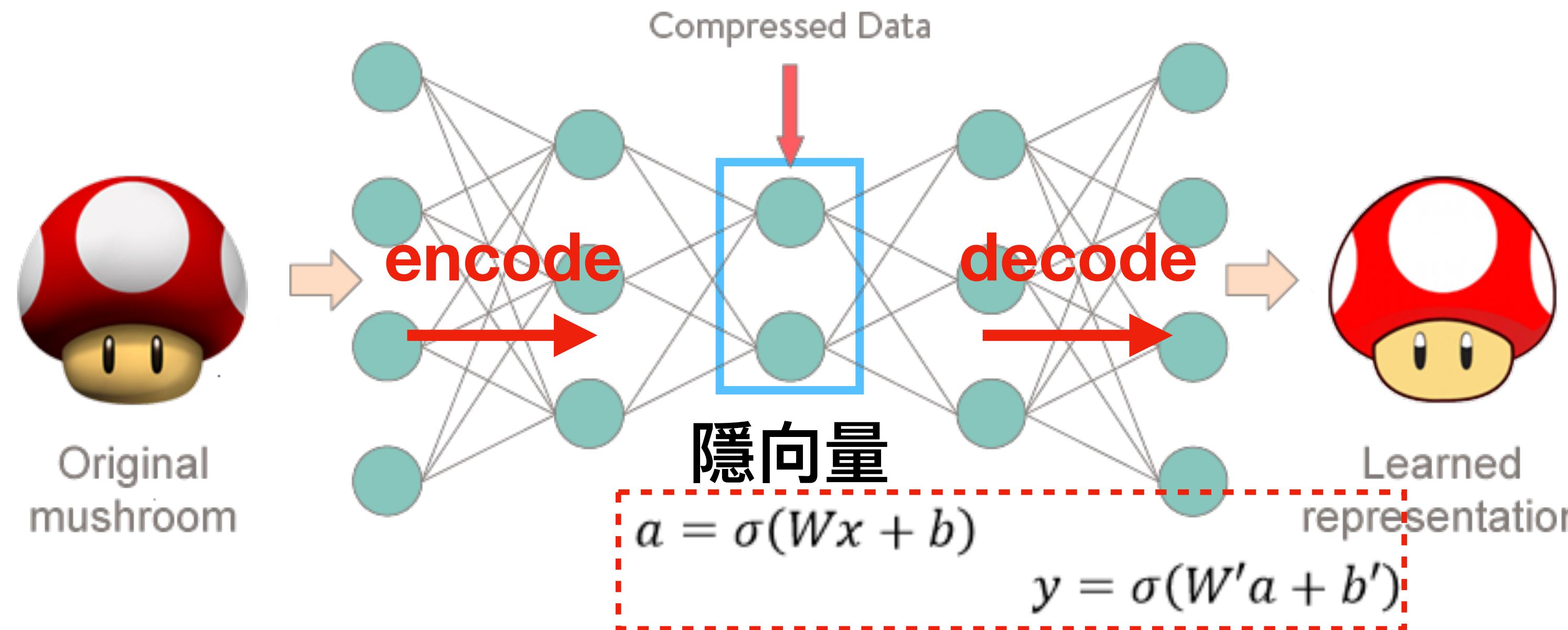
Autoencoder應用

Feature extraction & reduction with AutoEncoder



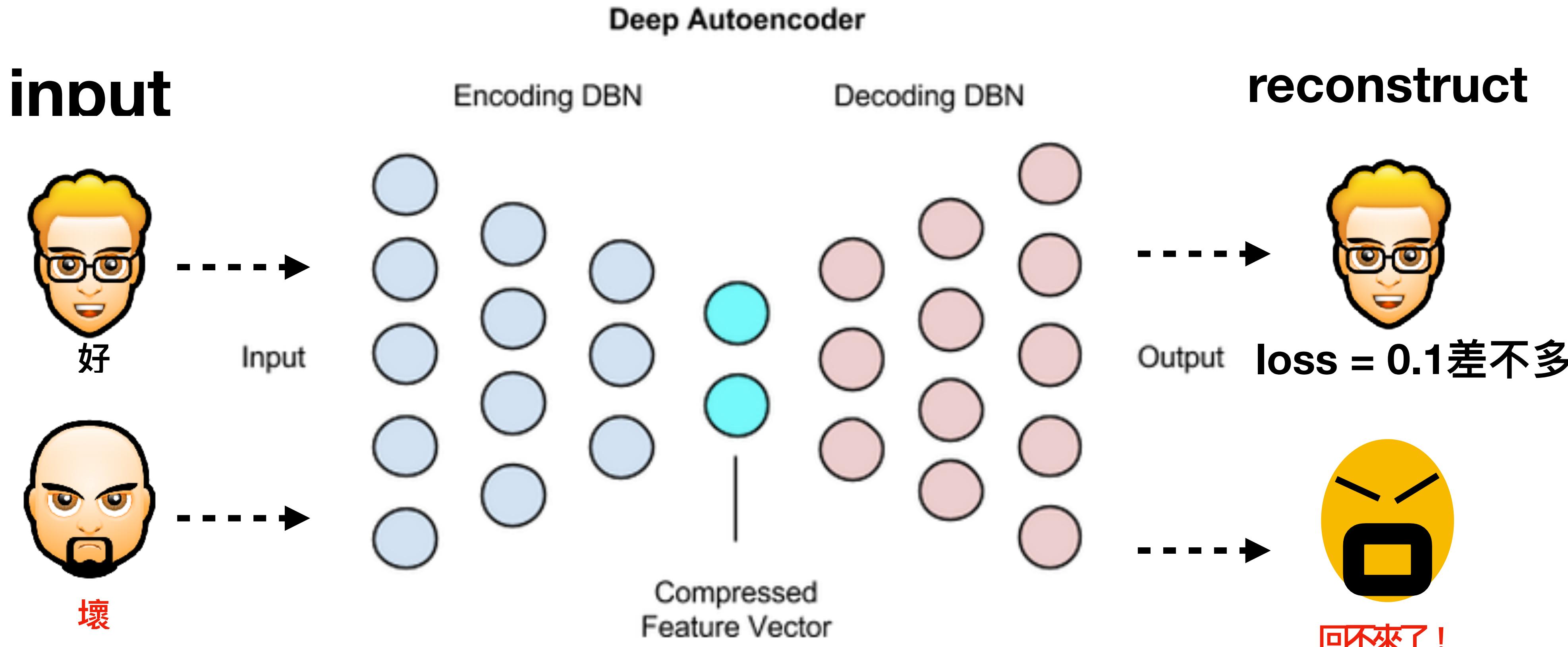
從高維度訊息量，提取最具信息量的關鍵
特徵，達到降維的效果

Pre-Training with AutoEncoder



Pre-Training 對初始化網絡訓練出
更好的權重(W)，方便後續模型訓練

Anomaly Detection using AutoEncoder

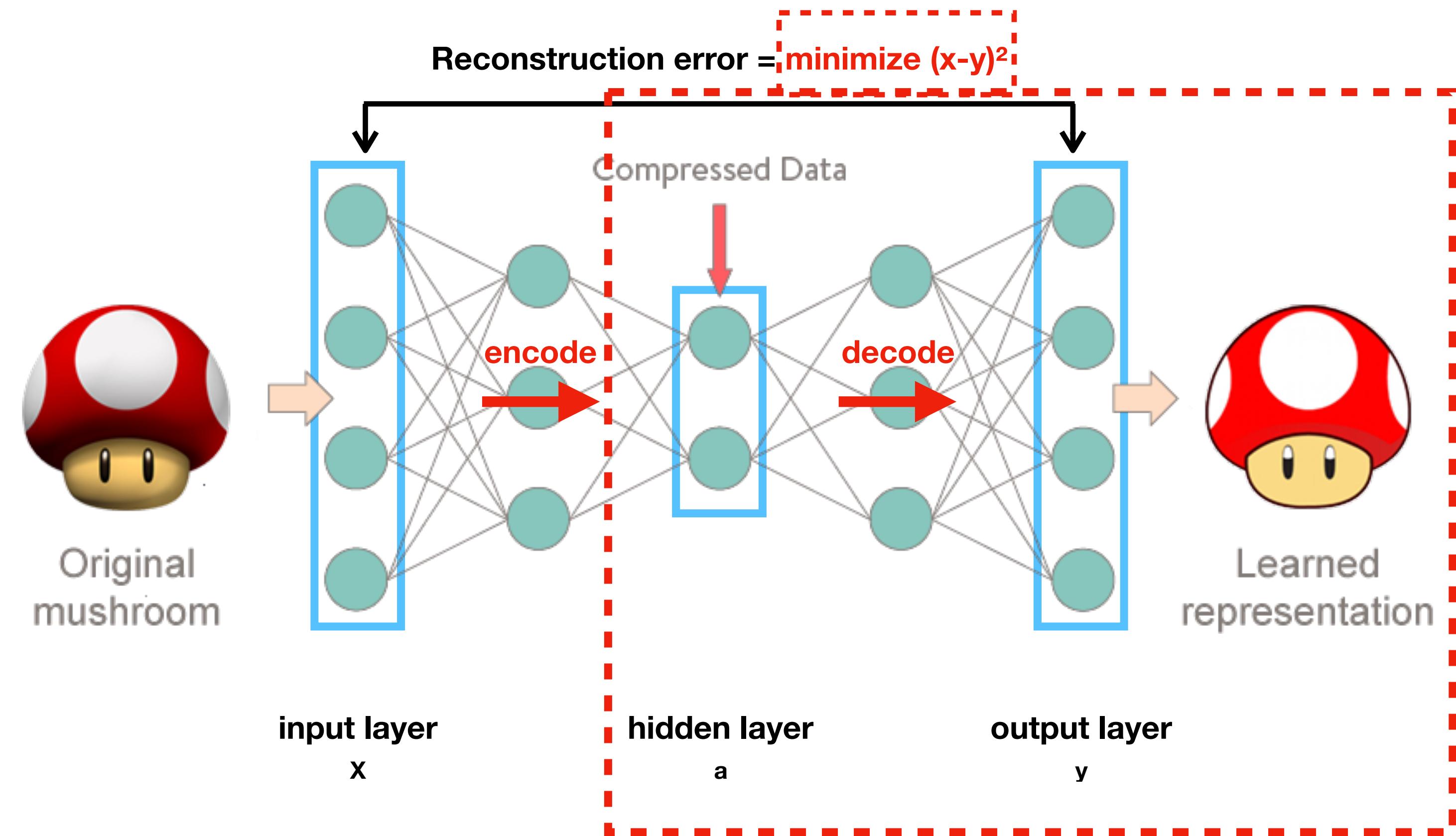


用好人訓練模型，
——
loss大大代表壞人



你真的以為我那麼好騙嗎

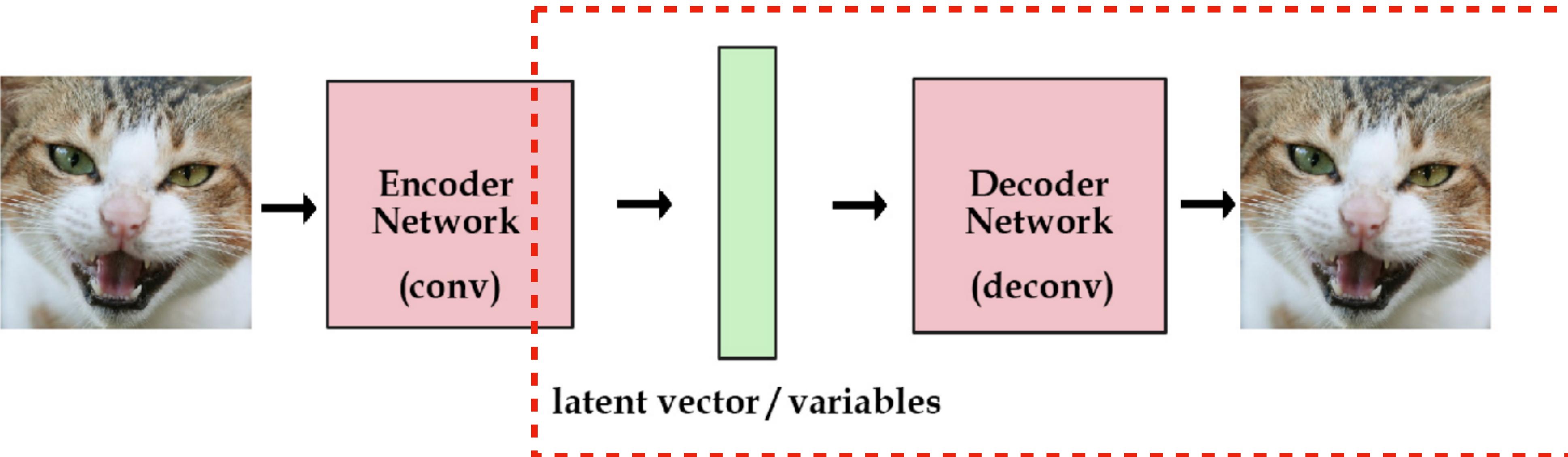
回憶



Decoder是一種還原過程

?

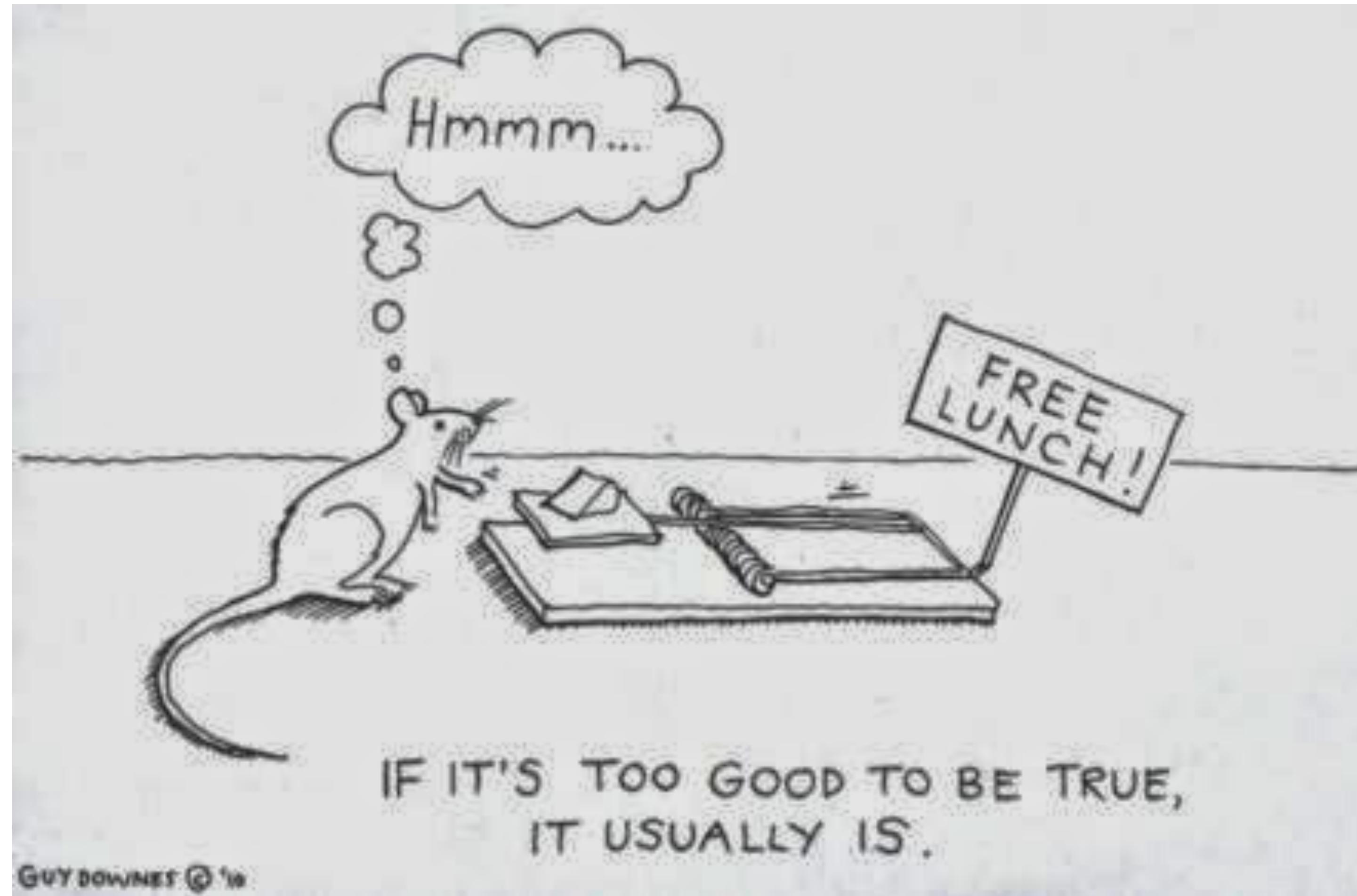
Decode新圖片很糟



Autoencoder訓練隱向量還原效果有時**很差**
(因為學不到資料的機率分佈)

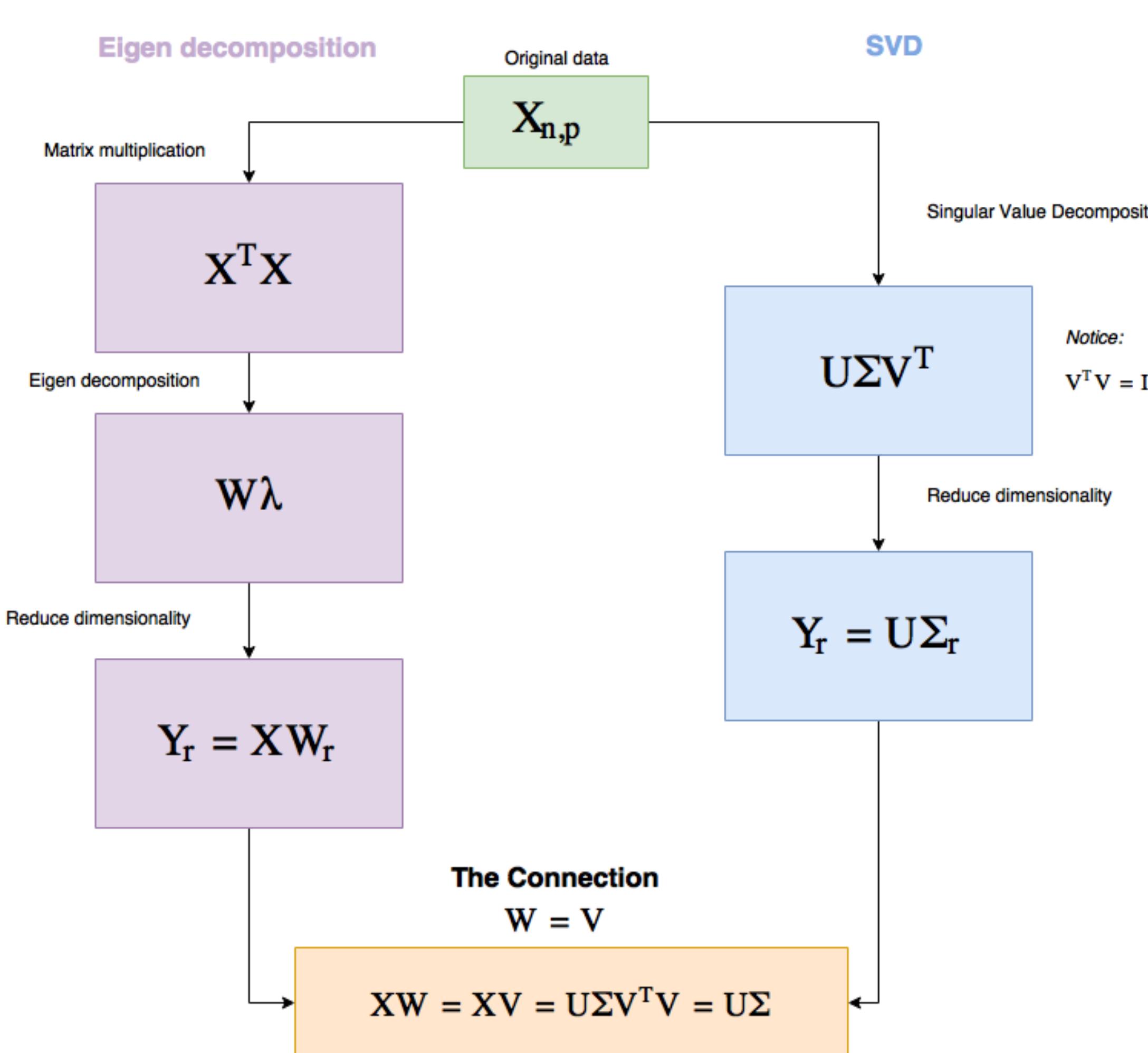
?

數據缺失怎麼辦啊



且待下回揭曉

PCA



t-SNE

