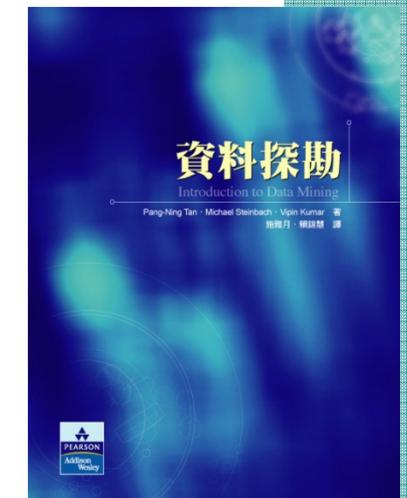


# 第 3 章

# 資料的探索



© 2008 台灣培生教育出版 (Pearson Education Taiwan)

# 什麼是資料的探索？

---

## 資料的初步探索，以更瞭解資料的特性

- 資料探索的主要目的在於
  - 幫助選擇適當的前處理方式以及資料探勘的技術
  - 協助解決資料探勘的問題
    - ◆ 人們可以利用視覺化的方式來發現樣式以及進行結果的解釋
- 探索性的資料分析（Exploratory Data Analysis，EDA）領域相關
  - 由John Tukey所發展的統計方法

# 資料探索使用的技術

---

- Tukey 所發展的EDA
  - 強調視覺化
  - 將**分群分析**和**異常偵測**皆視為資料探索技術
  - 分群分析和異常偵測是資料探勘的重要研究領域，而非僅視為**資料探索**
- 本章將資料探索的重點放在
  - 統計彙總
  - 視覺化
  - 線上分析處理（OLAP）

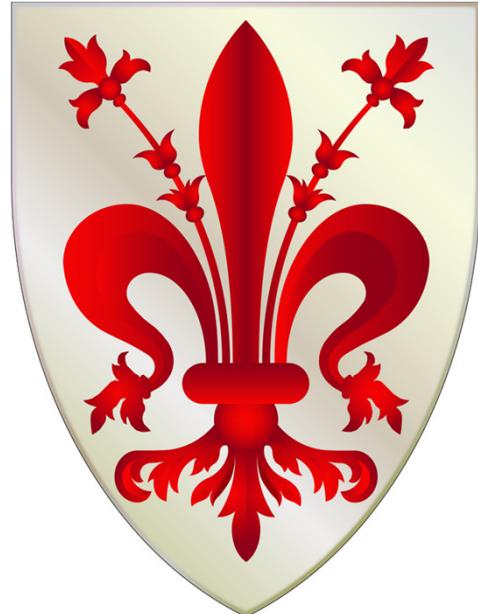
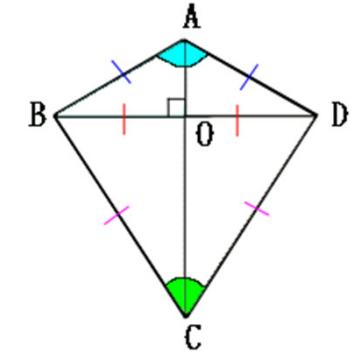
# 鳶尾花 (Iris) 資料集

---

- 有許多的資料探索技術是利用鳶尾花資料集進行說明
  - 有三種主要的類別：
    - ◆ Setosa
    - ◆ Virginica
    - ◆ Versicolour
  - 有四種屬性
    - ◆ 莖長
    - ◆ 莖寬
    - ◆ 花瓣長
    - ◆ 花瓣寬



Virginica Iris的圖片（Robert H. Mohlenbrock @USDA-NRCS PLANTS Database/USDA NRCS. 1995）。東北濕地植物：野外辦公室植物物種指南。東北國家技術中心，Chester，賓州（刪除了背景）



# 彙總統計

---

- 彙總統計屬於量化的資料
  - 包括資料的**次數**、資料的**落點**和資料的**分佈**
    - ◆ 範例：資料的落點 — 平均數
    - 資料的分佈 — 標準差
- 次數和眾數
  - 屬性值的**次數**是指在資料集中該值所發生**次數的百分比**
  - **眾數**是指具有**最高次數的屬性值**
- 百分位數
  - 具順序性的資料，其百分位數（percentiles）是很有用的資訊
  - 對**順序性屬性**或是**連續值屬性**  $x$  而言，其值  $p$  介於0到100之間， $x$  的第  $p$  個百分位數為  $x_p$ 。如第50個百分位數是  $x_{50\%}$ ，表示其所有  $x$  值的50%小於  $x_{50\%}$

# 彙總統計

萼寬、萼長、花瓣長及花瓣寬的百分位數(所有單位為公分)

百分位數	萼長	萼寬	花瓣長	花瓣寬
0	4.3	2.0	1.0	0.1
10	4.8	2.5	1.4	0.2
20	5.0	2.7	1.5	0.2
30	5.2	2.8	1.7	0.4
40	5.6	3.0	3.9	1.2
50	5.8	3.0	4.4	1.3
60	6.1	3.1	4.6	1.5
70	6.3	3.2	5.0	1.8
80	6.6	3.4	5.4	1.9
90	6.9	3.6	5.8	2.2
100	7.9	4.4	6.9	2.5

理論上， $\min(x) = x_{0\%}$ ;  $\max(x) = x_{100\%}$

# 資料的落點：平均數及中位數

---

- 對連續型資料，平均數和中位數是兩個最常見的彙總統計公式
- 平均數對於具有離群值的資料很敏感

$$\text{平均數 } (\bar{x}) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{中位數 } (x) = \begin{cases} x_{(r+1)} & \text{若 } m \text{ 為奇數，即 } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{若 } m \text{ 為偶數，即 } m = 2r \end{cases}$$

一組資料  $X = \{3, 5, 2, 1, 4, 90\}$ ，其平均數為何？中位數為何？

# 資料的分佈：全距及變異數

- 全距是指**最大值**和**最小值**間的**差距**
- **變異數**和**標準差**是最常用來衡量一組資料分佈的方式
- **平均數**易受**離群值**所影響，但因**變異數**也要用到**平均數**，所以也不適用於具有離群值的資料上
- 可用下列計算改善
  - Absolute average deviation, AAD (絕對平均離差)
  - Median absolute deviation, MAD (中位數絕對離差)
  - Interquartile range, IQR (四分位差)

$$AAD(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

$$MAD(x) = \text{中位數} \left( \{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\} \right)$$

$$\text{四分位差}(x) = x_{75\%} - x_{25\%}$$

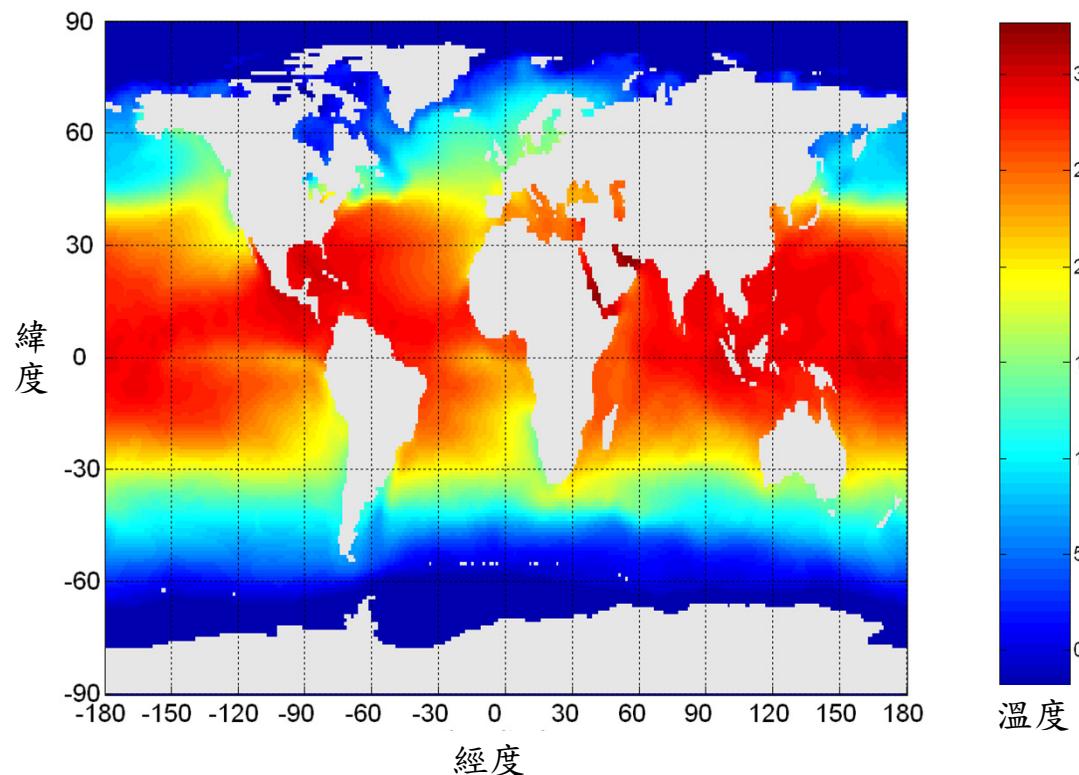
# 視覺化

---

- 資料的視覺化是要用**圖形**或是**表格**的方式來呈現資料，一個成功的視覺化圖表就是能夠清楚的呈現資料的**特性**、以及資料間或是屬性間的**關係**，而且可以輕易的讓人看圖釋義
- 視覺化的圖表通常可以用來解釋氣象、經濟及選舉的預測結果，就是可以利用圖形來解釋資料
- 資料探勘的視覺化技術有時稱為視覺化的資料探勘（visual data mining）

# 範例：海平面溫度

- 下圖為 1982 年海平面的溫度
  - 圖形彙整了 25 萬筆資料



# 表示法

---

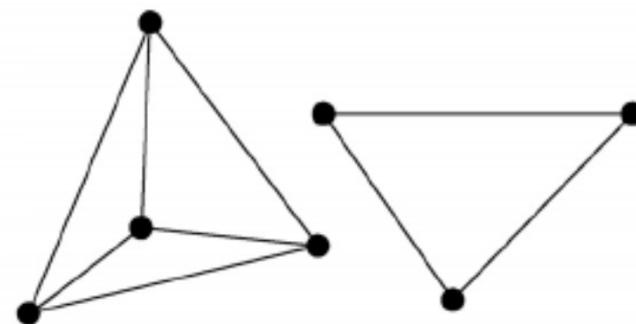
- 將資訊映射至視覺化圖形
- 將物件、屬性及關係映射至視覺化物件、屬性及關係，也就是分別對應至圖形上的點、線、面
- 物件的三種表示方法
  - 若物件只有一個類別屬性，則其物件通常是根據屬性值歸成一個區塊，而這些類別將用表單或是一個區域來表示
  - 若物件有很多屬性，那麼其物件將用表單的行、列或者是線來表示
  - 物件通常是用二維或三維空間來呈現，而其樣本點通常會用圓圈、方形符號來表示

# 圖形的安排

- 以視覺化方式安排
- 可以很清楚的呈現物件型態
- 範例：



(a) 原始的圖形

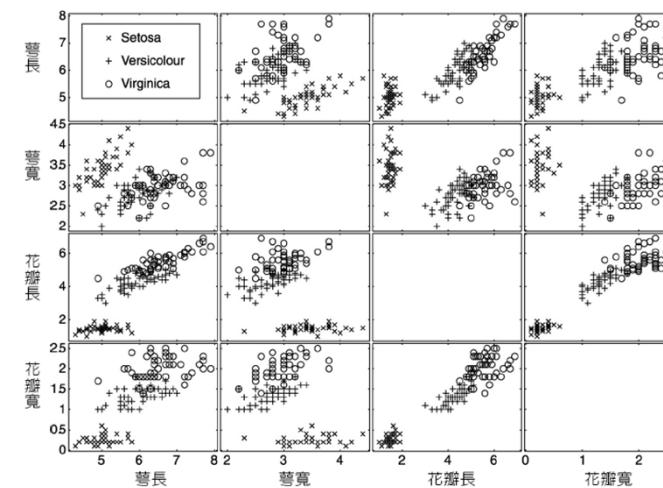


(b) 將圖形分開

將圖上的連接元件分開，如圖3.3(b)，則其節點與圖形的關係將變得更簡單易懂

# 選取

- 選取某些要刪除或是不重要的物件或屬性
- 如何選，是一個重要的議題
- 選取的方法
  - 選取屬性的子集合
    - ◆ 通常是兩個，維度太高可用散佈圖，或以一系列的二維圖形顯示，呈現直覺的圖形
  - 選取物件的子集合
    - ◆ 資料點過多不易呈現所有物件，可以不同顏色凸顯或刪除部份特性較不明顯之物件
- 視覺化技術
  - 以屬性個數為主
  - 以屬性型態為主
  - 以應用問題為分類依據



# 莖葉圖(Stem and leaf plot)

- 莖葉圖的用法，以下表資料為例。50位哈斯肯斯公司(Haskens Manufacturing)的應徵者參加能力測驗的結果，這項測驗共有150道題目，這些資料代表應徵者答對的題數
- 首先將每一個資料的十位數安排到垂直線的左邊且由小至大依序排列，垂直線的右邊則記錄每一個資料的個位數，所放的位置須對應十位數的位置。

表 2.9 能力測驗答對題數

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

資料來源：統計學，滄海書局

# 莖葉圖實例

6	9	8									
7	2	3	6	3	6	5					
8	6	2	3	1	1	0	4	5			
9	7	2	2	6	2	1	5	8	8	5	4
10	7	4	8	0	2	6	6	0	6		
11	2	8	5	9	3	5	9				
12	6	8	7	4							
13	2	4									
14	1										

- 將資料重新安排如上述的形式後，資料排序就非常簡單。排序完成後，即完成莖葉圖如下。

資料來源：統計學，滄海書局

# 莖葉圖實例

<b>6</b>	8	9									
<b>7</b>	2	3	3	5	6	6					
<b>8</b>	0	1	1	2	3	4	5	6			
<b>9</b>	1	2	2	2	4	5	5	6	7	8	8
<b>10</b>	0	0	2	4	6	6	6	7	8		
<b>11</b>	2	3	5	5	8	9	9				
<b>12</b>	4	6	7	8							
<b>13</b>	2	4									
<b>14</b>	1										

- 直線左邊的數字(6, 7, 8, 9, 10, 11, 12, 13與14)是莖(stem)，線右邊每一個數字是葉(leaf)，例如，第一列的6是莖，8, 9是葉。

6 | 8 9

資料來源：統計學，滄海書局

# 莖葉圖實例

---

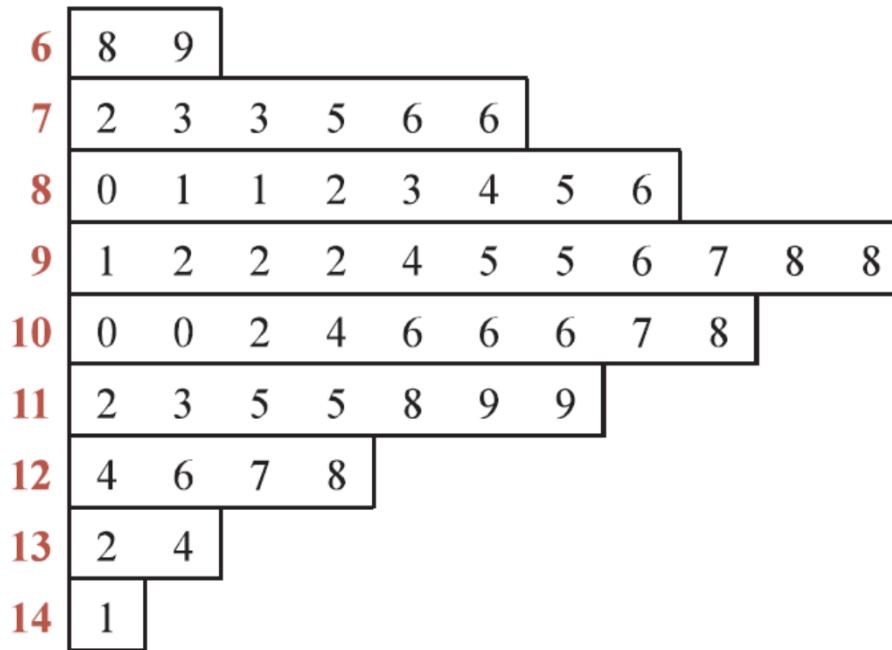
- 這表示有兩個資料值的第一位數字是6，葉的數值顯示兩個資料是68與69。同理，第二列是

7 | 2 3 3 5 6 6

表示第一位數是7的資料有6筆：72, 73, 73, 75, 76以及76。

- 為了強調莖葉圖的形狀，我們利用長方形將每一個莖的葉之部分框起來。如此一來，我們便可以得到以下的表示圖。

# 莖葉圖實例



- 將上面的圖形依逆時針方向旋轉90度，則得到一個組界為60-69, 70-79, 80-89等的直方圖。

資料來源：統計學，滄海書局

# 莖葉圖

---

- 優點
  1. 莖葉圖容易繪製。
  2. 在一個分類組別區間內，由於莖葉圖列出所有實際資料值，故能提供比直方圖更詳細的資訊。
- 莖葉圖沒有絕對的列或莖的數目
- 可將原始資料的第一個數字再分成兩個或兩個以上的莖，輕易地擴充莖葉圖。
- 莖葉圖以單一個數字來定義葉的值，葉單位顯示莖葉圖的數字應乘上的適當倍數，如此一來莖葉圖即可以近似原始資料。葉單位可以是100, 10, 1, 0.1等等。

# 視覺化技術：直方圖

- 直方圖：通常用來顯示單一屬性的分佈情形
- 範例：鳶尾花屬性的直方圖（分別有10和20個箱子）

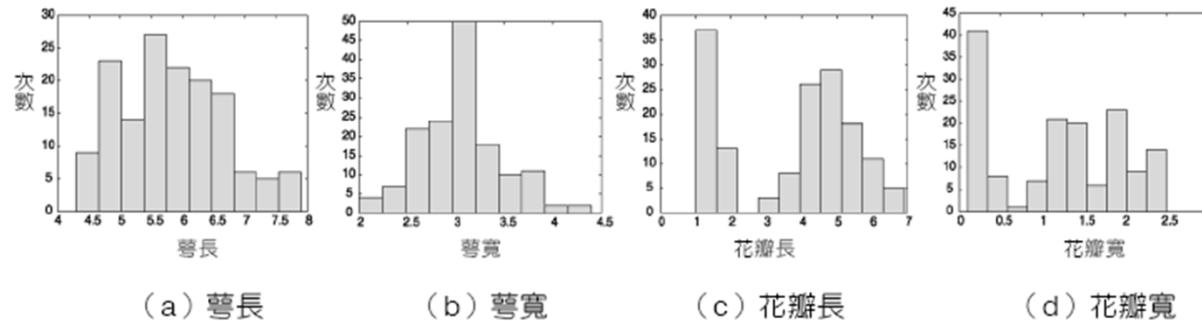


圖 3.7 ▶ 四個鳶尾花屬性的直方圖（10 個箱子）

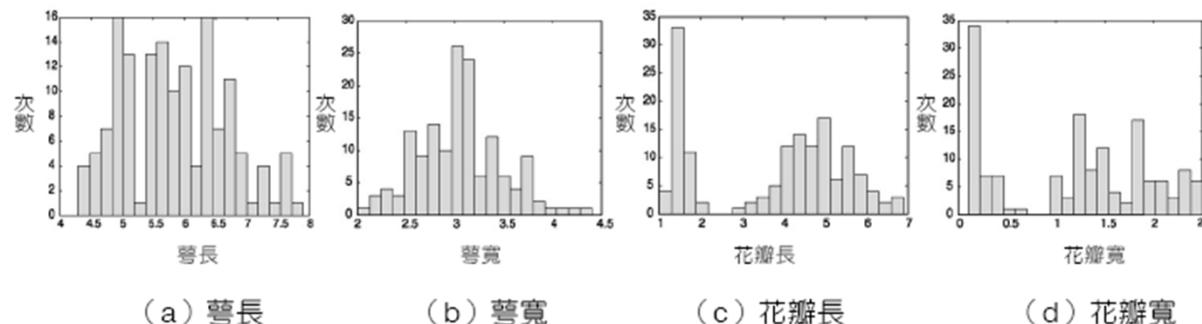
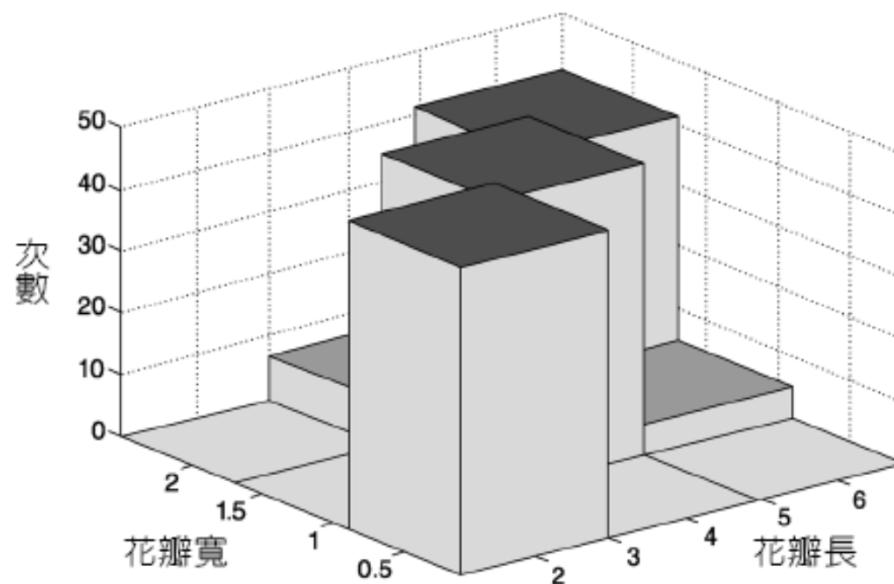


圖 3.8 ▶ 四個鳶尾花屬性的直方圖（20 個箱子）

## 二維直方圖

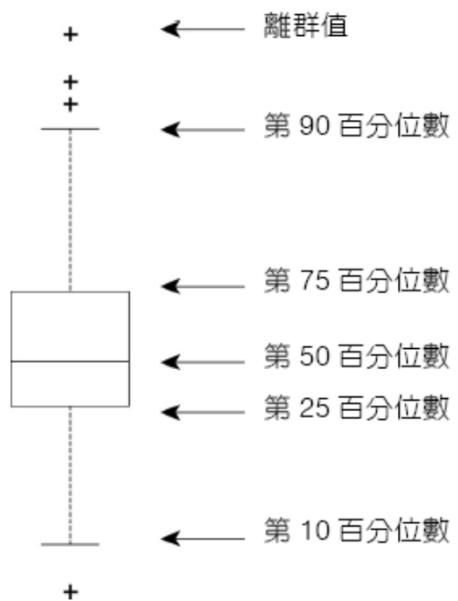
- 將每個屬性分成兩個區間，再將其區間的資料視為兩個維度
- 範例：顯示花瓣長和花瓣寬的二維直方圖
  - 從中可以發現什麼？



# 視覺化技術：盒狀圖

- 盒狀圖

- J. Tukey 發明
- 另一種呈現單一數值屬性分佈的作法
- 下圖是萼長的盒狀圖，在箱子的最底層及最上層分別為第25及第75個百分位數，而中間的線則為第50個百分位數。上方及下方的線分別為第10及第90個百分位數。離群值則用「+」來表示



# 盒狀圖的範例

- 盒狀圖可以用來比較不同物件類別間屬性的差異

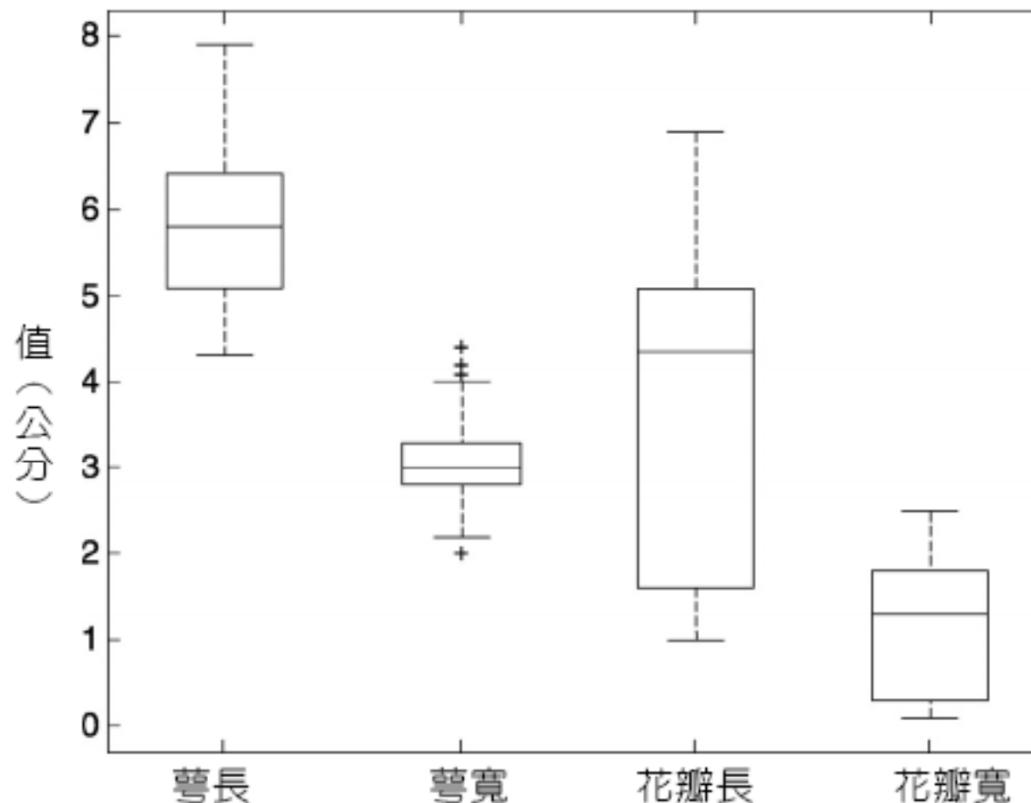
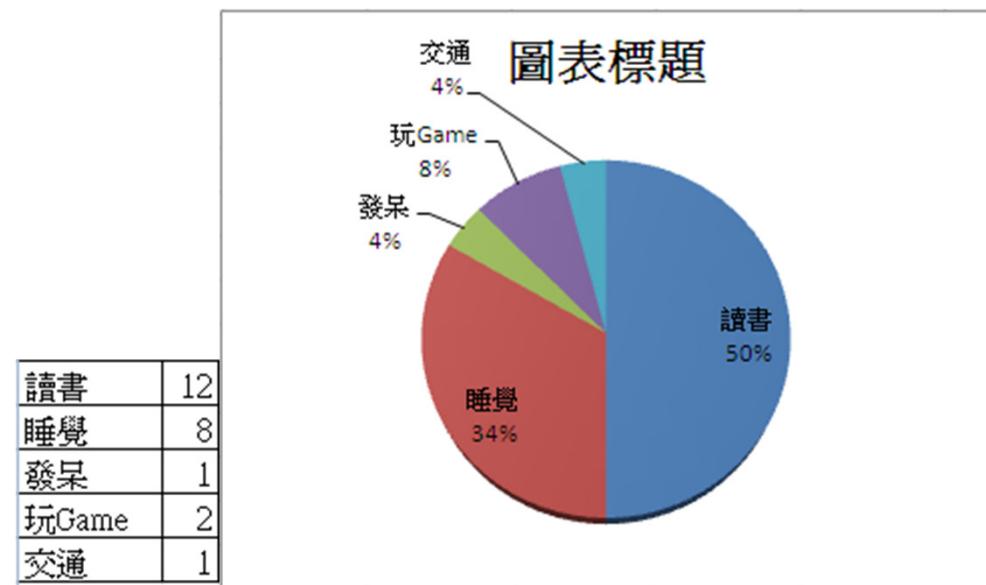


圖 3.11 ▶ 翠尾花的盒狀圖

# 圖餅圖

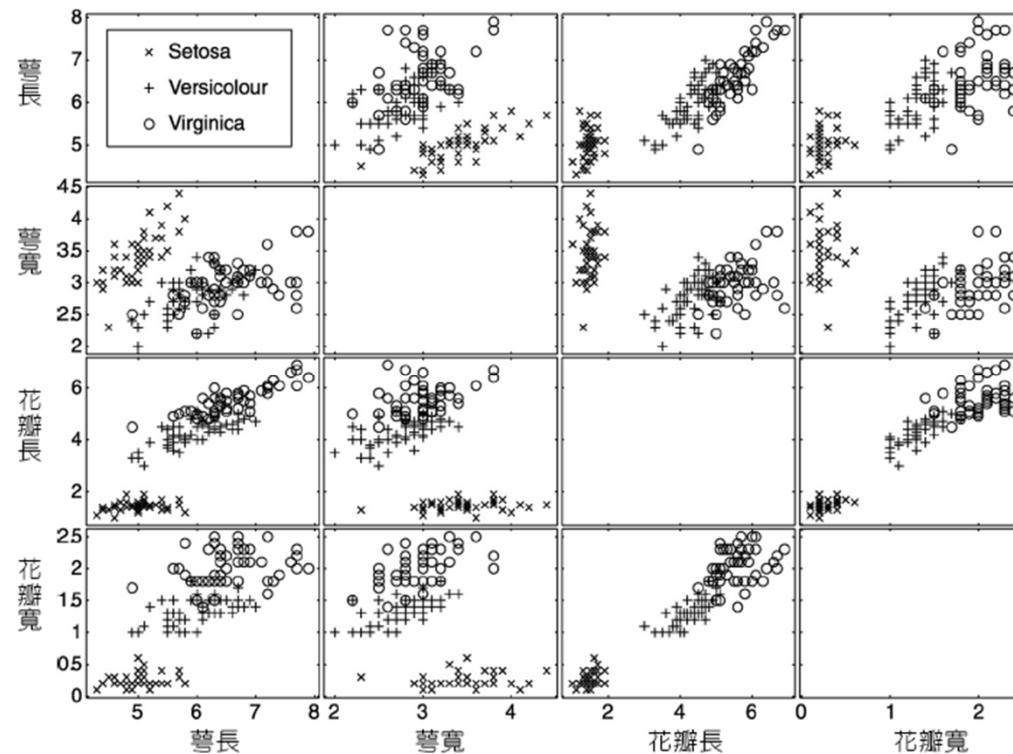
- 通常用於類別屬性
- 值相對較小
- 不易依圖判斷其相對的面積程度，故技術文獻不常見



# 視覺化技術：散佈圖

- 散佈圖

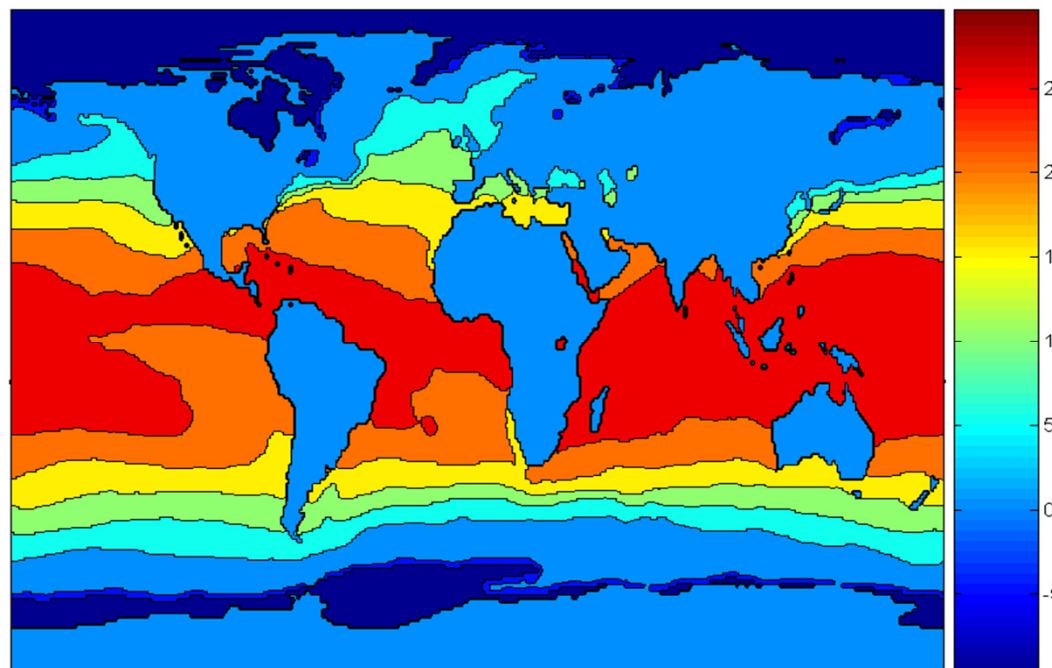
- 用來呈現**兩個屬性間的關係**
- 可用來偵測**非線性關係**
- **二維或三維圖形**可用來表示額外的屬性，但資料愈多，視覺化圖形會愈複雜，更不易解釋



# 視覺化技術：等高線圖

## ● 等高線

- 對於**三維度**資料而言，**二維度**屬性是指平面上的位置，第三維是**連續值**，像是氣溫等，這時就可以用等高線圖來將平面分成不同區域，三個屬性的值（如溫度或海拔高度）大都相等。常見的等高線圖範例是顯示陸地的海拔高度



1998年12月平均海平面溫度的等高線圖

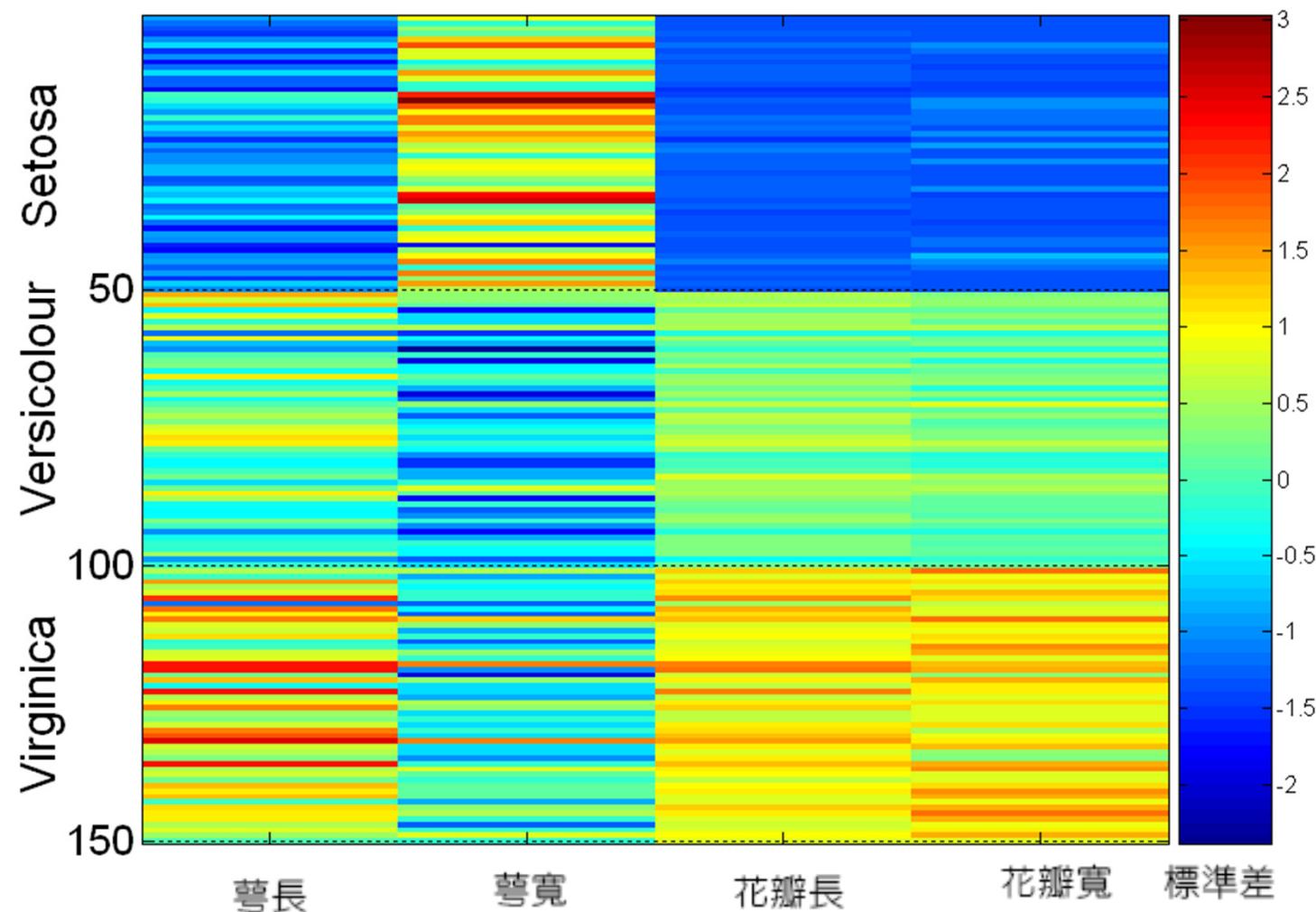
# 視覺化技術：矩陣（高維度資料）

---

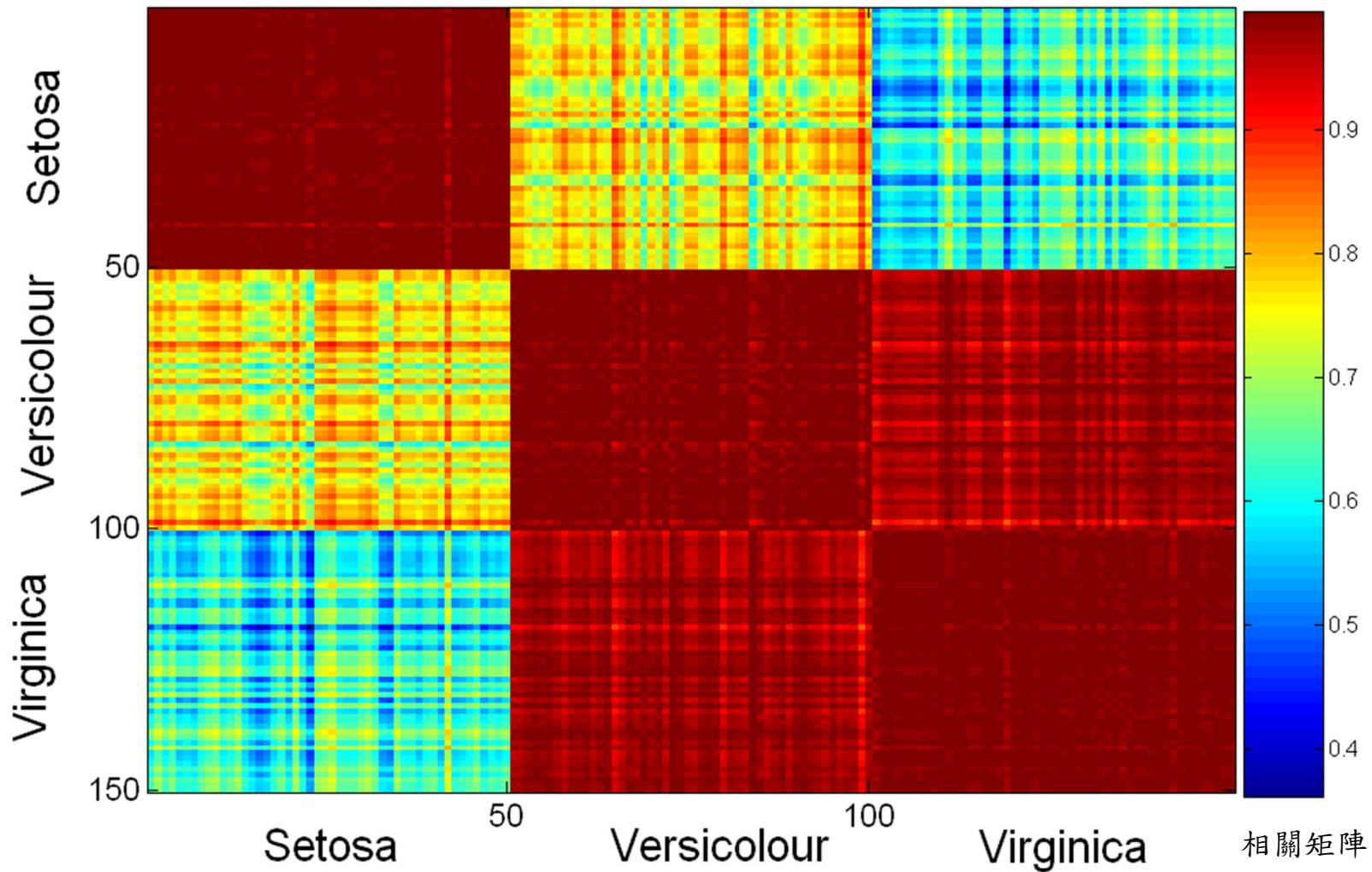
- 矩陣

- 其圖形可以視為一個**點矩陣**，每個陣列上的元素都是一個顏色或是亮度，所以資料矩陣上的元素都是圖形上的像素資訊
- 若其類別標記已知，重新排序資料矩陣是有必要的，如此一來可以將類別中的物件聚類在一起
- 若要偵測是否所有類別中的物件具有相似的屬性，如其屬性有不同的全距，則其屬性通常可以將其標準化，使其**平均數**為0且**標準差**為1。這可以避免屬性的值過大或太小而決定了圖形
- 可以用來觀察圖形中物件鄰近矩陣的結構，也就是在當類別標記已知的情形下，可以將相似矩陣的列及行進行排序，如此一來，其類別中的物件就會聚類在一起，這也可以用來評估每個類別的**聚合力**及和其他類別的**差異性**

# 範例：鳶尾花資料的資料矩陣



# 範例：鳶尾花資料的相關矩陣



# 視覺化技術：平行座標

## ● 平行座標

- 平行座標中，每個**屬性**有一個**座標軸**，但是不同的座標軸是互相平行的。每個物件屬性的值會對應至座標軸上的一點，而且點與點之間將連成線，以表示一個物件

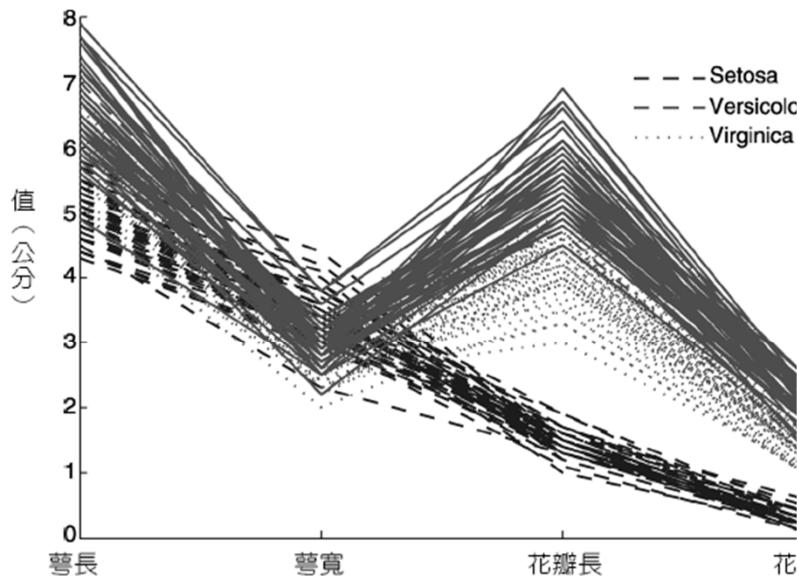


圖 3.25 ▶ 翁尾花資料中，四個屬性的水平座標軸

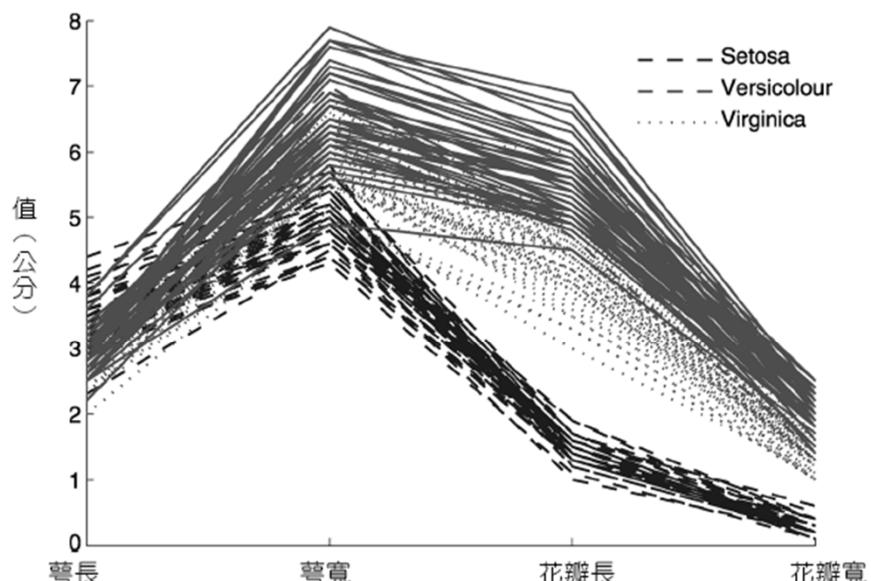


圖 3.26 ▶ 重新排序後的四個屬性之水平座標軸，其目的在於強調群內物件的相似情形

# 其他視覺化技術

---

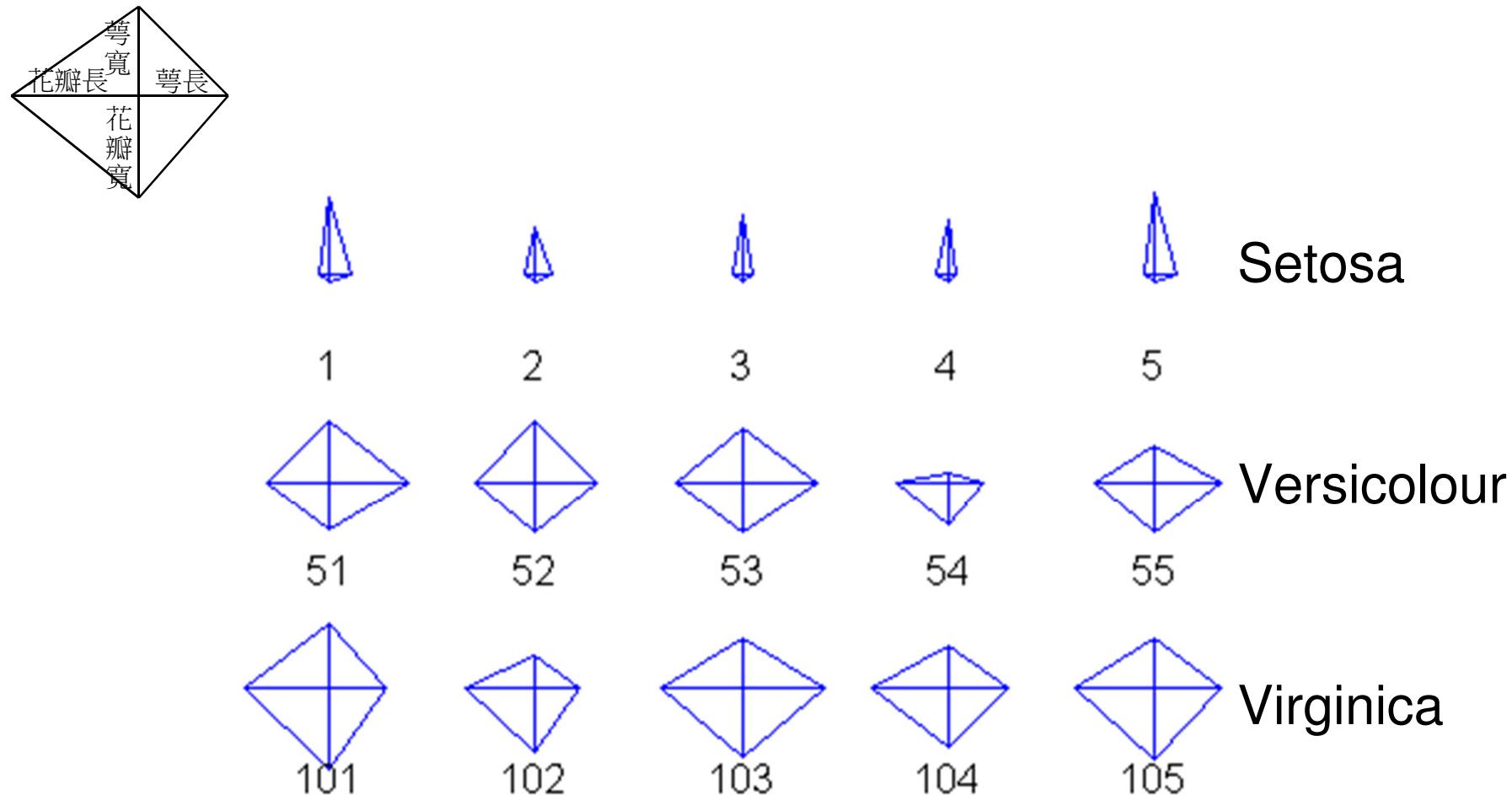
- 星狀座標軸

- 類似平行座標，但軸是從中心點開始放射
- 物件將用以下步驟進行對應：首先將物件的每個屬性值，轉換成屬性間的最小及最大值間的距離。其距離將對應至屬性座標軸上的一點；每個點將連成一線段，最後形成一個多邊形

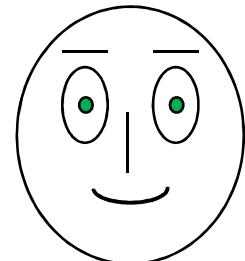
- 臉譜

- 由 Herman Chernoff 提出的技術
- 在這個技術中，每個屬性都是臉譜上的一個特徵
- 每個屬性值決定對應的臉譜上的特徵
- 每個物件變成一個臉譜

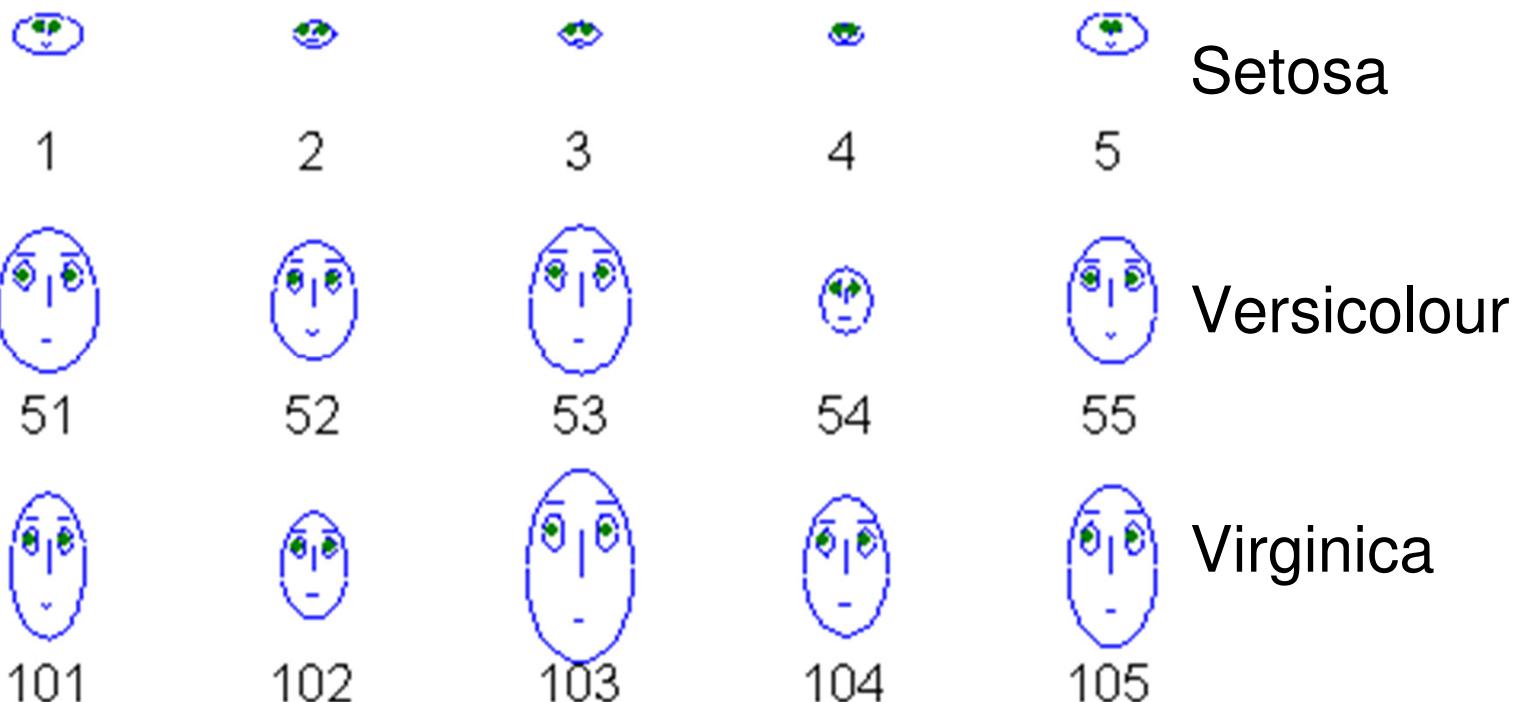
# 15筆鳶尾花資料的星狀座標軸



# 15筆鳶尾花資料的臉譜



資料特徵	臉部特徵
萼長	臉部大小
萼寬	前額/下巴的比例相對較長
花瓣長	前額形狀
花瓣寬	下巴形狀



# OLAP

---

- 線上分析處理（On-Line Analytical Processing，OLAP）是由關聯式資料庫之父E. F. Codd 提出
- 線上分析處理強調**互動性**的分析資料，並且提供視覺化資料的能力及**產生彙總統計**的資訊，因此OLAP系統適合做為**多維度資料分析**的主要分析方法
- 大部分的資料都可視為一個**表單**，而表單中的**每一列**為一個**物件**，同時**每一行**為一個**屬性**，在很多情形下，也可以將資料視多維度的陣列

# 建立多維度陣列

---

- 將資料視為多維度資料的步驟有二個：
  - 維度的識別
  - 屬性識別
- 維度是類別屬性、或者是從連續屬性轉換而來的類別屬性，將屬性的值視為一個陣列的索引，而屬性的個數就是維度的個數

# 範例：將鳶尾花資料視為多維度陣列

- 將花瓣長、寬等轉換成多維度的陣列
  - 首先，將花瓣的長和寬的屬性分割成低、中及高三類，並且計算每一類所包含的個數

花瓣長	花瓣寬	花型	個數
低	低	Setosa	46
低	中	Setosa	2
中	低	Setosa	2
中	中	Versicolour	43
中	高	Versicolour	3
中	高	Virginica	3
高	中	Versicolour	2
高	中	Virginica	3
高	高	Versicolour	2
高	高	Virginica	44

# 範例：將鳶尾花資料視為多維度陣列（續）

- 下圖是將三個維度切成三個二維表單
- 從表中可知：*Setosa*花型有低的寬及長；*Versicolour*花型有中等的寬及長；而*Virginica*花型的寬及長則較高

表 3.8 *Setosa* 花型根據花瓣長及寬所建立的交叉列聯表

		寬		
		低	中	高
長	低	46	2	0
	中	2	0	0
	高	0	0	0

表 3.9 *Versicolour* 花型根據花瓣長及寬所建立的交叉列聯表

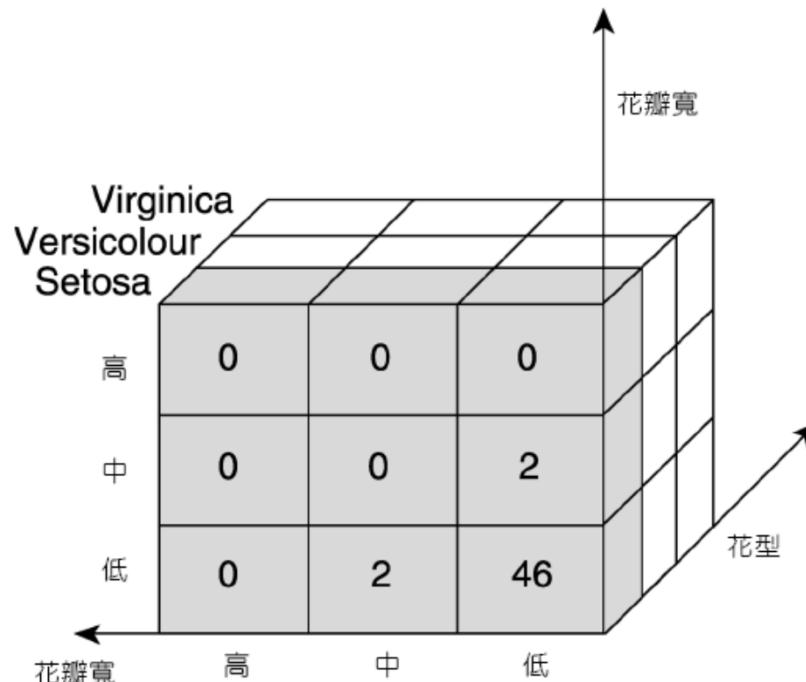
		寬		
		低	中	高
長	低	0	0	0
	中	0	43	3
	高	0	2	2

表 3.10 *Virginica* 花型根據花瓣長及寬所建立的交叉列聯表

		寬		
		低	中	高
長	低	0	0	0
	中	0	0	3
	高	0	3	44

## 範例：將鳶尾花資料視為多維度陣列（續）

- 每個屬性的組合，都是多維陣列中的一個元素
- 這個元素會被指定一個數值
- 下方的圖是以多維度表示鳶尾花資料



# 資料立體方塊：聚集總和

---

- 多維度分析的主要動機在於用不同的方法來進行資料的彙總，一般我們所討論的彙總資料是計算其聚合的值
- 所有可能的多維度聚集總和，稱為資料立方體，其名稱、每個維度大小並非相等，而且資料立方體可以超過三個維度，更重要的是，資料立方體其實是統計方法中的交叉列聯表

# 切片及切塊

---

- 切片（slicing）是指選定一個或是多個維度上的特定值所產生的立方體
- 切塊（dicing）是指選擇某個範圍的屬性值所形成的立方體，其做法相當於從整個陣列中定義一個子陣列

# 上捲及下鑽

---

- 屬性值通常具有層級結構
  - 日期是由月和週所組成
  - 地點是由國家和城市組合而成
  - 產品可分為衣飾、電子產品和家電等類別
- 這些類別都可以建構成一個層級樹狀結構或是晶格
  - 年是由月組成，月是由天組成
  - 國家是由州組成，州是由城市所組成
- 這些層級樹狀結構可以經由上捲（roll-up）或是下鑽（drill-down）的操作來進一步瀏覽彙總資料
  - 以銷售資料為例，可以上捲日期的維度而得知每月的銷售額；或是利用下鑽月的銷售額而得到每日銷售額