Python 科學計算與統計分析介紹

目次

Python 語言進階_Scipy 科學計算與統計分析 1
1 前言 3
1.1 本書適用對象 3
1.2 本書特點 3
2 Python 語言 4
2.1 Python 簡介 4
2.2 Python 優點 4
2.3 為何選擇 Python 4
3 IPython Notebook(Jupyter) 6
4 Python 科學計算及統計分析套件 Scipy 8
4.1 Scipy Library 9
4.2 Numpy 9
4.3 Matplotlib 10
4.4 Pandas 10
5 準備 Python 科學計算的使用環境 Anaconda 12
5.1 Anaconda3 科學計算套件 12
5.2 Anaconda3 下載網址 12
5.3 Anaconda3 安裝 12

1 前言

1.1 本書適用對象

針對從未學習過程式語言的入門者,介紹Python應用於科學計算與統計分析領域的概念課程。本書以程式入門必須建立的基礎概念為目標，將內容集中介紹及了解使用的觀念，讓學習者快速進入Python科學計算與統計分析的世界。

1.2 本書特點

1.2.1 介紹Python應用於科學計算與統計分析領域的相關知識，及使用到的Python套件。

1.2.2 介紹適合非程式專業入門者能夠輕鬆入門的Anaconda3套件。

1.2.3 加註相關網頁資源，使學習者熟悉如何使用開源環境，建立使用網路資源的能力。

2 Python 語言

2.1 Python 簡介

Python 是程式語言，具有基本的迴圈控制、判斷式和一般用途資料型別，如 List、Tuple、Dict(Dictionary)、Set 等，但沒有陣列型別。Python 能夠使用程式邏輯根據特定的條件進行資料的過濾及處理，並呼叫內建及特定的函式進行各種運算，最終根據特定需求輸出結果或圖表。

Python 內建有許多標準函式庫，如一般資料型別的處理及數學計算、文字處理、檔案資料輸入輸出等。

2.2 Python 優點

2.2.1 語法簡潔，容易理解，且容易偵錯，適合初學程式語言者。

2.2.2 支援各種網路、資料庫、科學計算、圖形介面套件，容易上手及使用。

2.2.3 支援網路資料存取，可輕易處理網路數據資料。

2.2.4 支援多種資料庫存取，可輕易處理大數據資料。

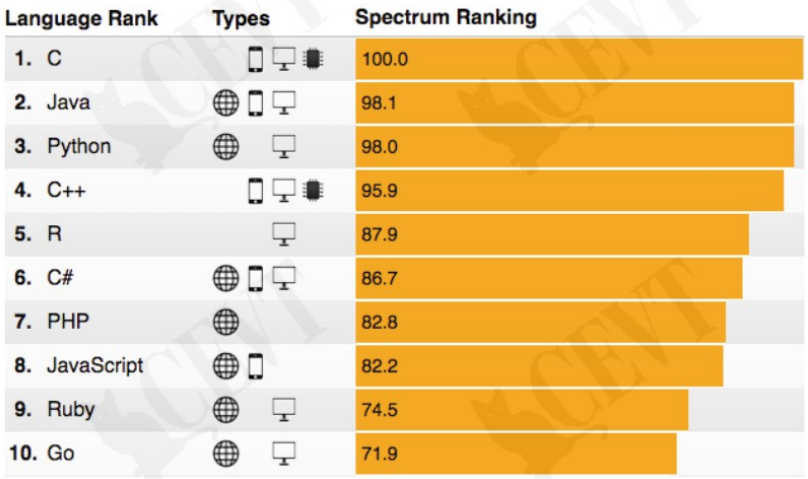
2.2.5 免費、開源，容易尋求大量教學及技術支援。

2.3 為何選擇 Python

2.3.1 IEEE Spectrum 最熱門語言排名

IEEE Spectrum 發布了 2016 年熱門程式語言排名，標題為『C is No.1, but big data is still the big winner』(C 語言是第一名，但大數據仍是大贏家)」，隨著大數據時代來臨的需求，大數據處理相關的程式語言最相關的兩個語言 Python 和 R 語言已經分居第三名和第五名。

（圖片來源：http://spectrum.ieee.org/…/the-2016-top-programming-languag…）



2.3.2 學術界 TIOBE 排名最熱門語言排名

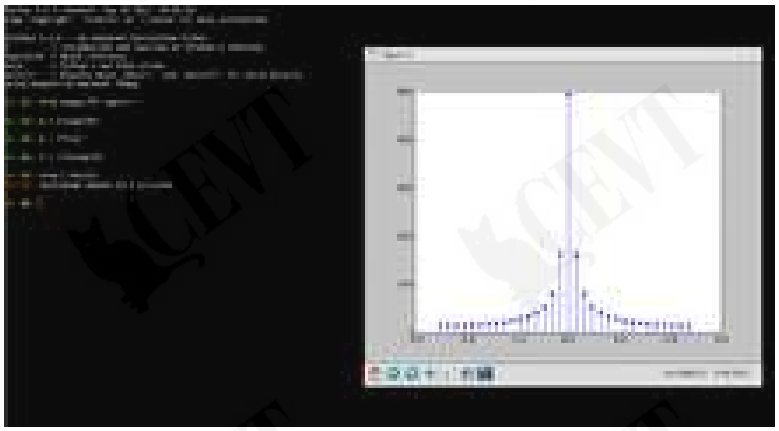
TIOBE 發布的 2017 年最熱門語言排名，Python 佔據第 4 名。其中前三名均為程式專業科系選擇的語言，Python 則為非程式專業科系的第一選擇。

（圖片來源：http://www.tiobe.com/tiobe-index/）

Feb 2017	Feb 2016	Change	Programming Language	Ratings	Change
1	1		Java	16.676%	-4.47%
2	2		C	8.445%	-7.15%
3	3		C++	5.429%	-1.48%
4	4		C#	4.902%	+0.50%
5	5		Python	4.043%	-0.14%
6	6		PHP	3.072%	+0.30%
7	9	^	JavaScript	2.872%	+0.67%
8	7	v	Visual Basic .NET	2.824%	+0.37%
9	10	^	Delphi/Object Pascal	2.479%	+0.32%
10	8	v	Perl	2.171%	-0.08%
11	11		Ruby	2.153%	+0.10%
12	16	^^	Swift	2.125%	+0.75%
13	13		Assembly language	2.107%	+0.28%
14	38	^^	Go	2.105%	+1.81%
15	17	^	R	1.922%	+0.73%
16	12	vv	Visual Basic	1.875%	+0.02%
17	18	^	MATLAB	1.723%	+0.63%
18	19	^	PL/SQL	1.549%	+0.49%
19	14	vv	Objective-C	1.536%	+0.13%
20	23	^	Scratch	1.500%	+0.71%

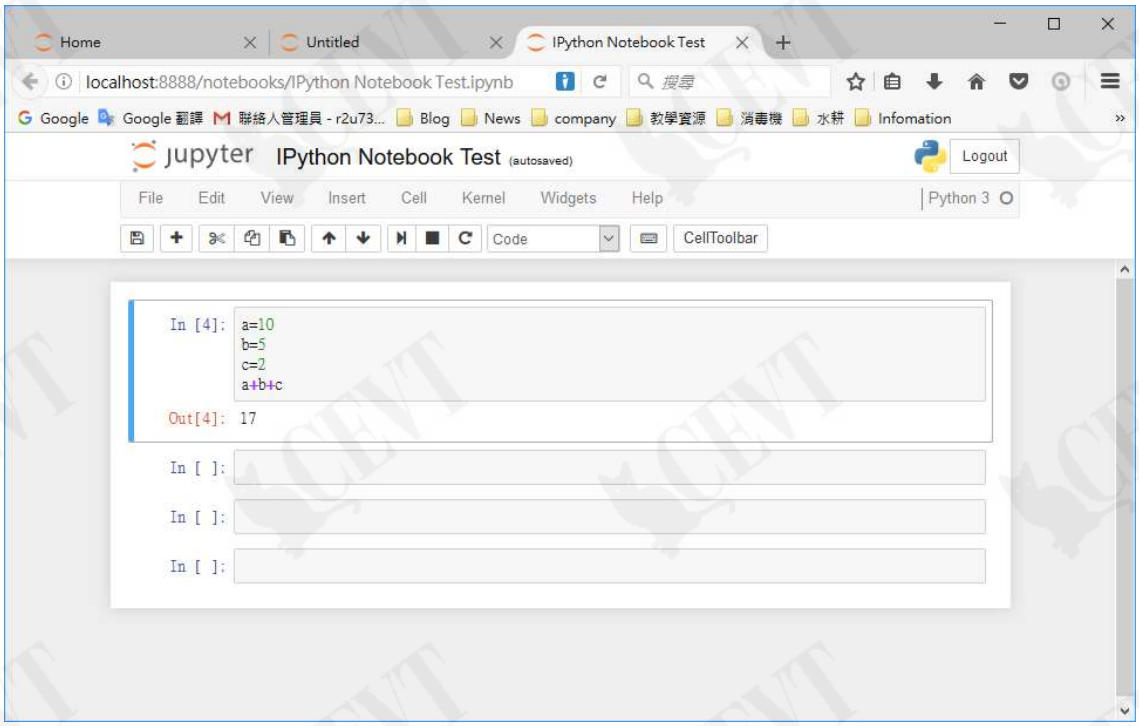
3 IPython Notebook(Jupyter)

IPython 是一個支援 Python 環境的交互式使用者介面(IDE)，相較於 Python Shell 的 IDE 環境，具備更強大的編輯及交互功能。

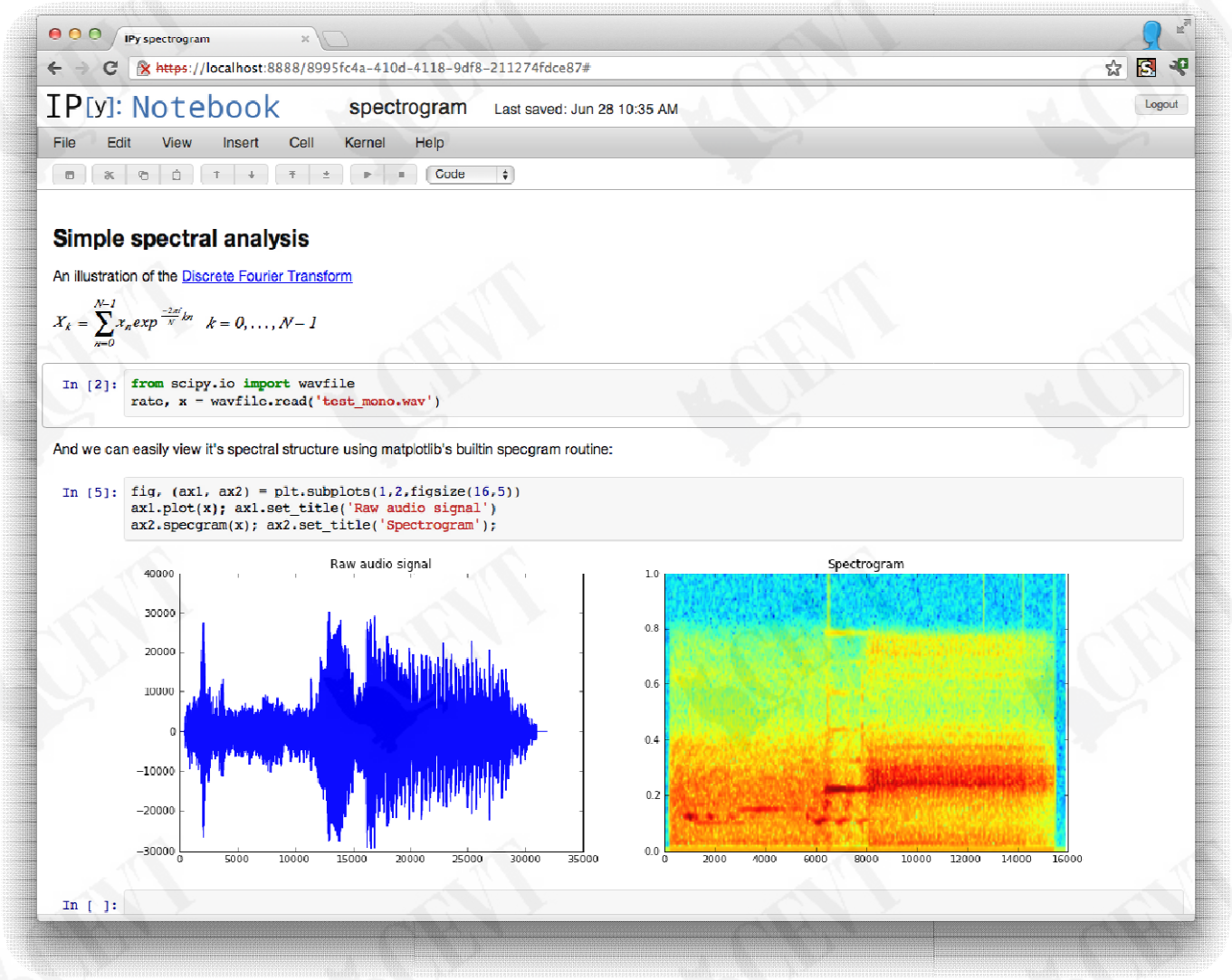


基於 IPython 發展出的 IPython Notebook 更進一步使用 Web 介面的 IDE，整合了網路資源，更方便分享程式碼及結果，也整合了圖表輸出的圖形化介面，對於非程式專業領域的程式撰寫人員更具有親和力，使 Python 成為科學人員使用來處理實驗數據的首選。

在 Python3.4 版之後，IPython NoteBook 更名為 Jupyter 以紀念偉大的科學家伽利略，以承襲伽利略分享實驗結果及數據的無私精神。



IPython Notebook 開始的 Web 畫面



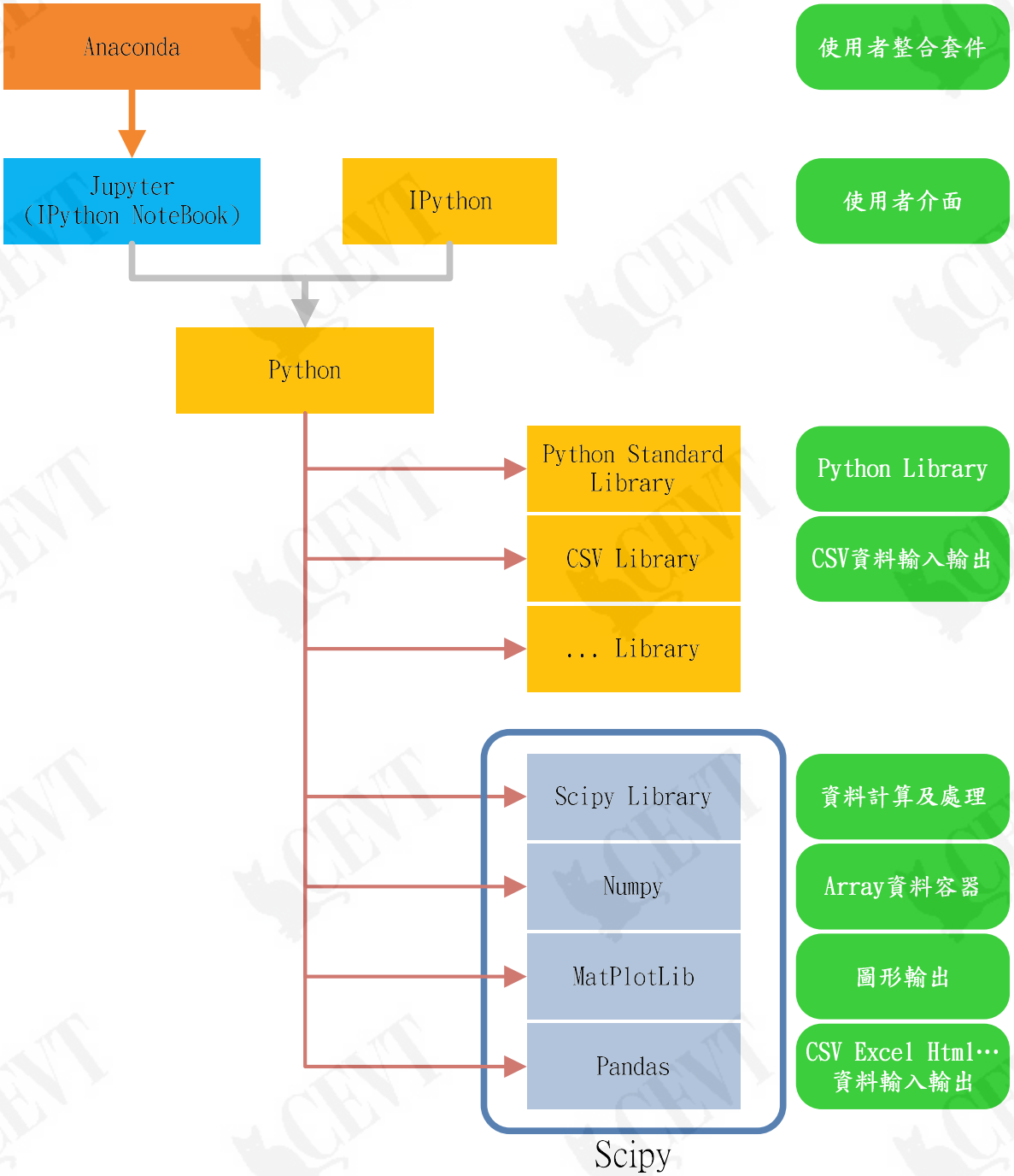
進行數學式視覺化輸出及圖表輸出的 IPython Notebook 畫面

4 Python 科學計算及統計分析套件 Scipy

Scipy 是 Python 的支援套件，其中包含了許多支援各種用途的模組，包括 Scipy Library、Numpy、Matplotlib、Panda 等，其後章節針對各模組做進一步的介紹。

Scipy Library 則提供了許多計算用函數，如傅立葉轉換、線性傳換，以及統計相關的各種函數。

Scipy 主要由 Numpy 模組提供了支援科學計算的數值容器，能夠支援向量或是矩陣的運算，已進行科學需要的數學運算。



4.1 Scipy Library

Scipy Library 是 Numpy 的擴充函式庫，具有許多進階的數學函式，可以進行許多常見的統計計算，如計算 PDF 值（Probability Density Function），CDF 值(Cumulative Distribution Function)。
Scipy 也具備信號處理函式，可以做迴旋(convolution)、傅立葉轉換(Fourier Transform)，及各種線性轉換。

常見的統計功能：

4.1.1 統計函式：算術平均、幾何平均、調和平均、標準差、變異數、四分位距、峰度、偏態、1-way 和 2-way ANOVA、p-value、F-value、T-test、相關係數...

4.1.2 間斷機率分佈：白努力、二項式、幾何、Poisson、Uniform...

4.1.3 連續機率分佈：Alpha、Beta、F、t、常態、卡方...

4.1.4 假設檢定：Z 檢定、t 檢定...

4.1.5 其他還有多種迴歸分析、最小平方法、估計方程式、信賴區間...等等，還有更多更複雜的統計函式。

4.2 Numpy

Numpy 是 Scipy 的數學計算擴充程式集，Numpy 的核心是 Numpy Array 陣列物件，Numpy Array 是一個數學物件如一個向量或一個矩陣，能夠以矩陣或是向量形式進行各種運算操作，如陣列搜尋和排序。
Numpy Array 最重要的部分是針對速度進行了優化。所以使用 Numpy 能夠高速進行數學運算。適合使用於複雜的矩陣運算，如乘積、反轉、計算因數、線性系統求解等。

常見的計算功能：

4.2.1 一維陣列及多維矩陣的建立、索引（Indexing）、切片（Slicing）

4.2.2 基本統計運算：總和、平均、標準差、最大值

4.2.3 基本數值運算：絕對值、平方根、log

4.2.4 矩陣運算：轉置矩陣（Transpose）、矩陣相乘

4.2.5 數學式視覺化呈現：

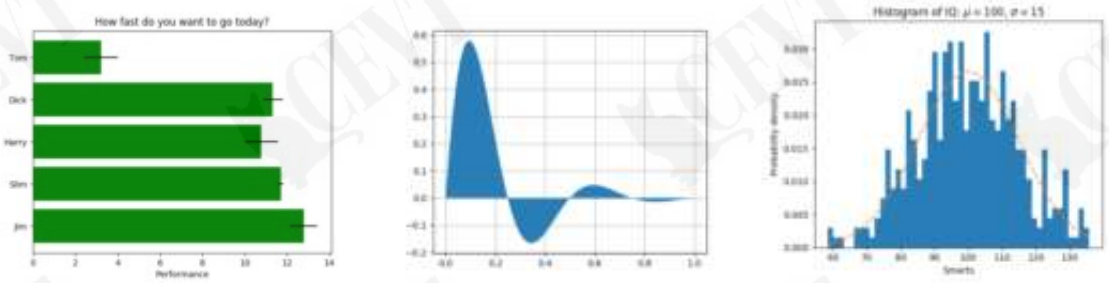
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

4.3 Matplotlib

Matplotlib 是視覺化繪圖的套件，不論是各種統計作圖、熱度圖、直條圖、地形圖、幾何圖、文字圖、雷達圖、3D 圖等，有多種作圖風格可供選擇，功能非常完整，是使用 Python 實作資料的必學套件。

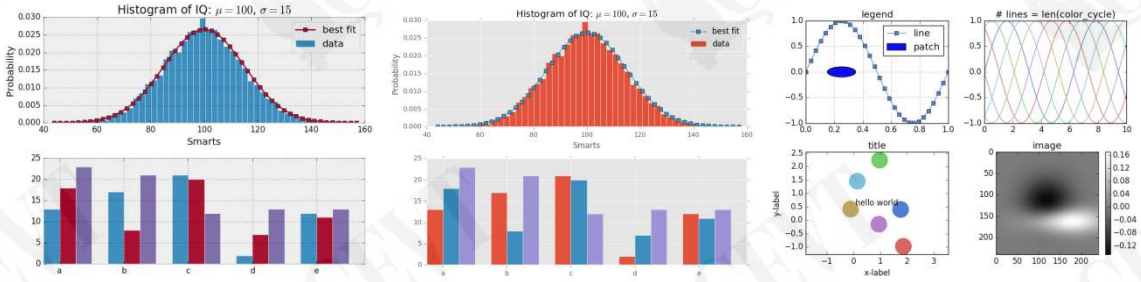
4.3.1 Matplotlib 圖集：

http://matplotlib.org/gallery.html



4.3.2 自選 Matplotlib 圖集：

https://tonysyu.github.io/…/matplotlib-style-g…/gallery.html



4.4 Pandas

Pandas 是資料表處理及視覺化圖表最重要的套件，因為 Pandas 提供 DataFrame 的資料結構，能夠以類似 SQL 的功能處理資料表，能夠進行資料庫的輸入輸出，將資料表整理為視覺化圖表需要的資料列，只需要一行程式碼就可以直接輸出視覺化圖表。

4.4.1 常使用的資料庫輸入輸出

4.4.1.1 CSV 資料格式檔案

4.4.1.2 Excel 資料格式檔案

4.4.1.3 HTML 資料格式檔案

4.4.2 常見的支援圖表種類：

4.4.2.1 ’line’：(default) 折線圖

4.4.2.2 ’bar’：垂直長條圖

4.4.2.3 ’barh’：水平長條圖

4.4.2.4 ’hist’：直方圖

4.4.2.5 ’box’：箱型圖

4.4.2.6 ’kde’：機率密度圖

4.4.2.7 ’area’：面積圖

4.4.2.8 ’pie’：圓餅圖

4.4.2.9 ’scatter’：散佈圖

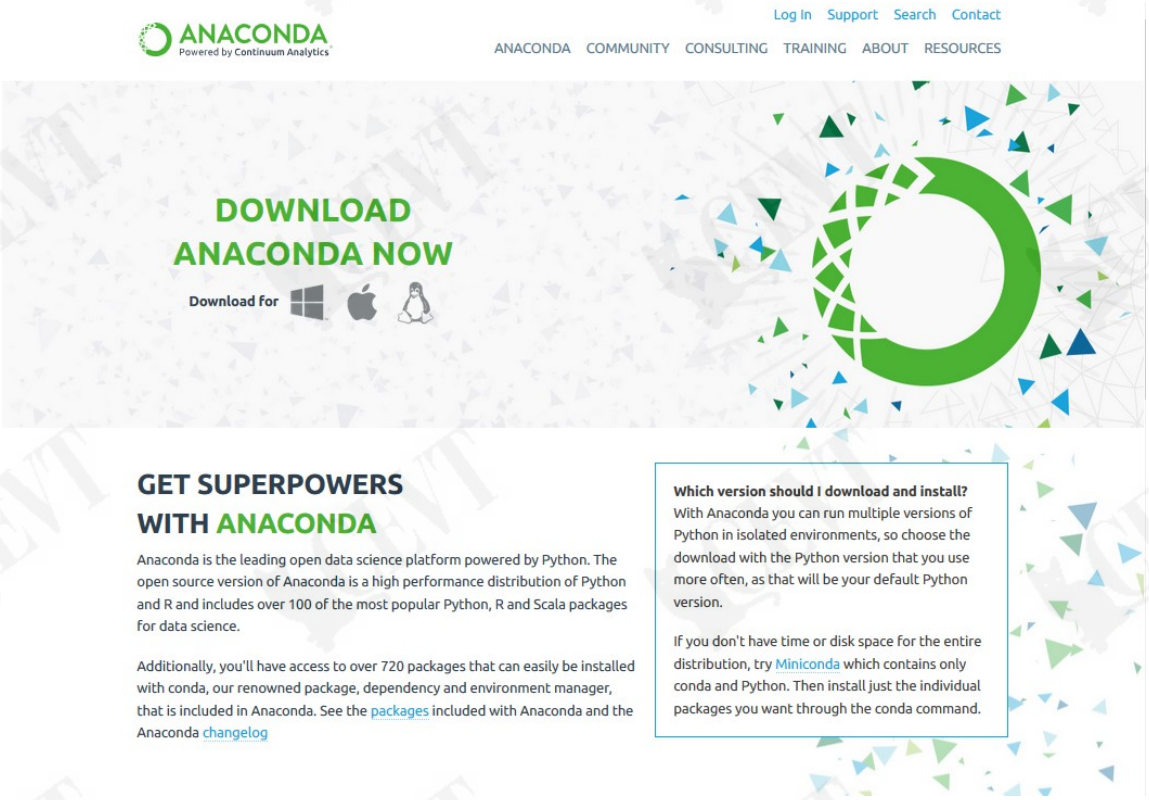
4.4.2.10 ’hexbin’：蜂窩圖

5 準備 Python 科學計算的使用環境 Anaconda

5.1 Anaconda3 科學計算套件

Anaconda 是一套將本文前述各項 Python 及支援函式庫完整打包的套件，一次下載及安裝就能夠建立起 Python 科學計算需要的使用環境，相較於以前需要各自下載並且在 Prompt 環境下安裝，Anaconda 能夠在 Windows 或 MacOS 下直接安裝，對初次進行 Python 程式領域的使用者來說，真是一大福音。

在 Python3.X 版後，Anaconda 也升級至 Anaconda3 以支援 Python3.x，請注意 Python2.x 及 3.x 版的差異，以避免程式無法順利編譯執行的問題。



Anaconda 包含的函式庫及各項套件，請參閱 Anaconda_CheatSheet.pdf 文件說明。

5.2 Anaconda3 下載網址

官網下載網址：https://www.continuum.io/downloads

5.3 Anaconda3 安裝

請參閱 Anaconda3 科學計算套件_介紹及安裝.pdf 文件說明。