



# Python for Data Analysis

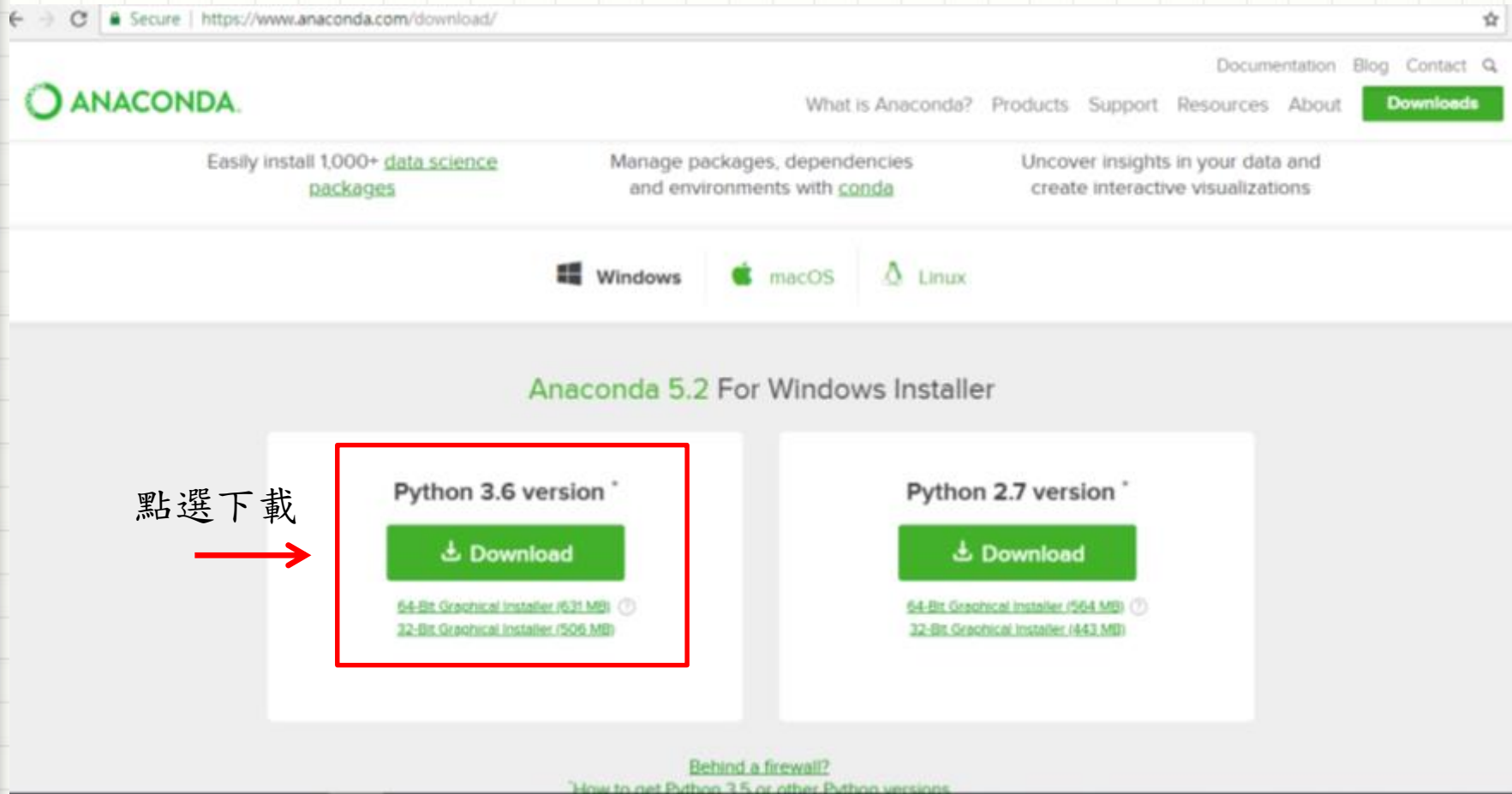
## 基礎知識

講者: 楊翔斌  
n07061033@mail.ncku.edu.tw

# 大綱

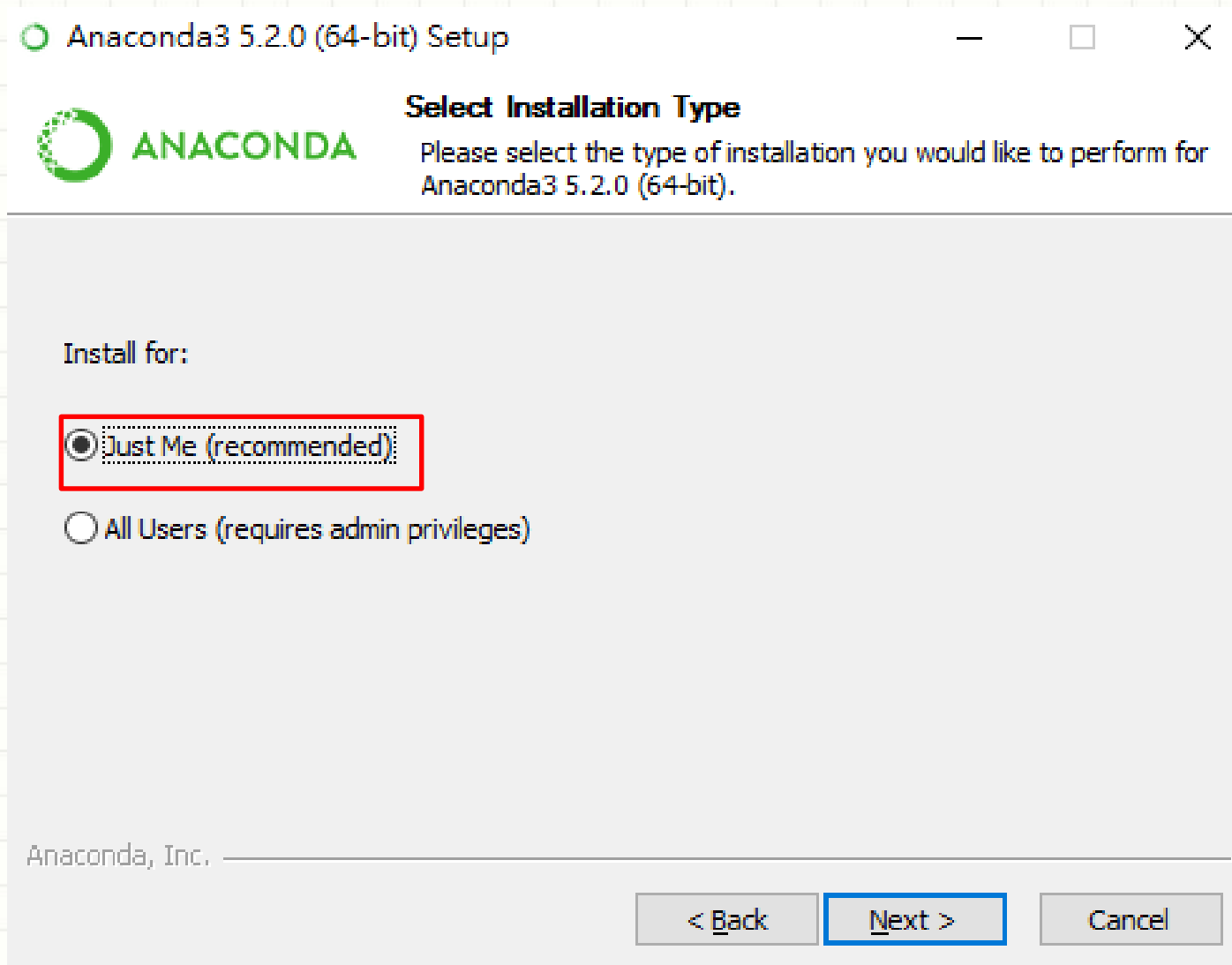
1. python(anaconda)安裝介紹
2. 套件使用
3. 介面使用
4. 小結
5. 相關資源介紹

# python(anaconda)安裝介紹

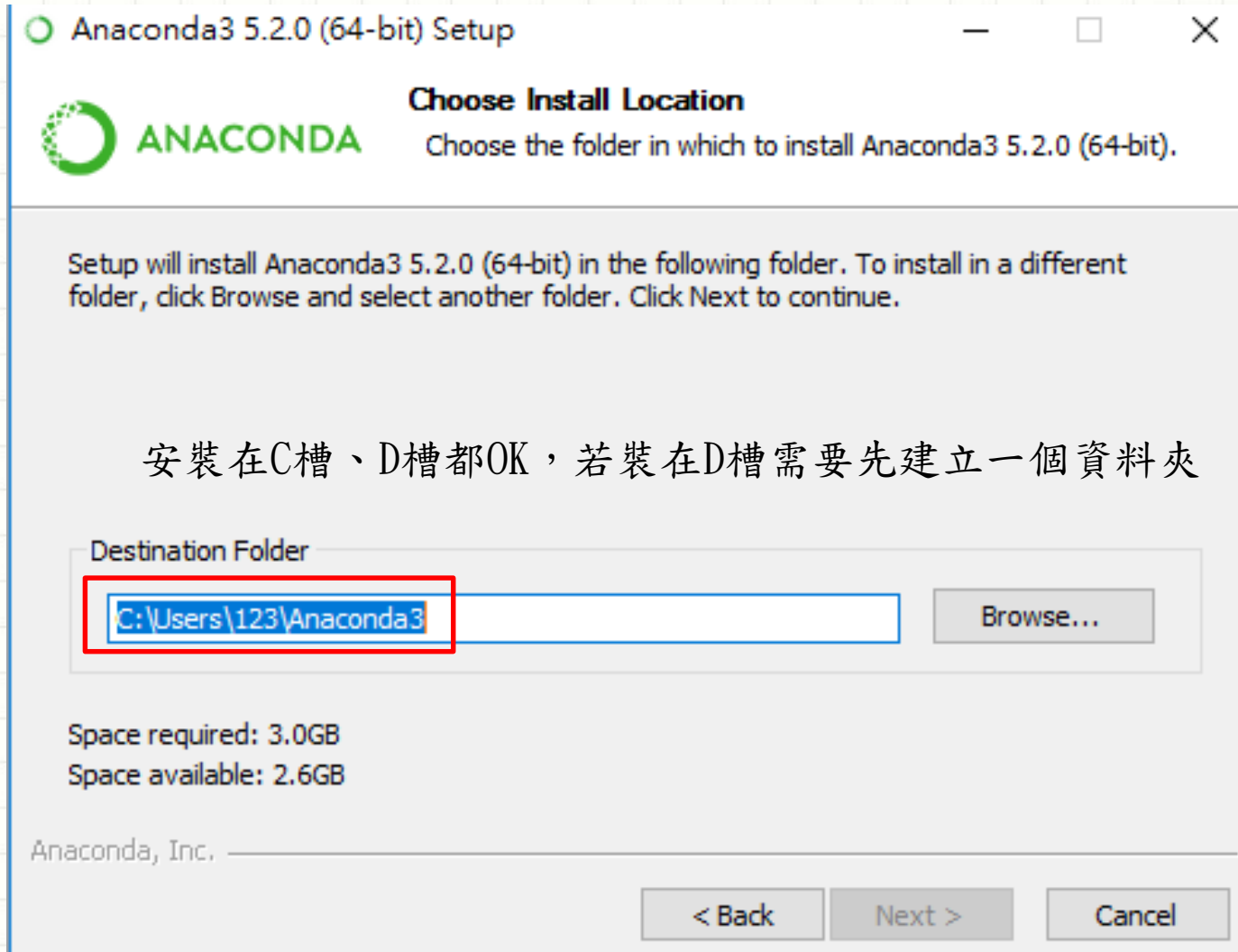


<https://www.anaconda.com/download/>

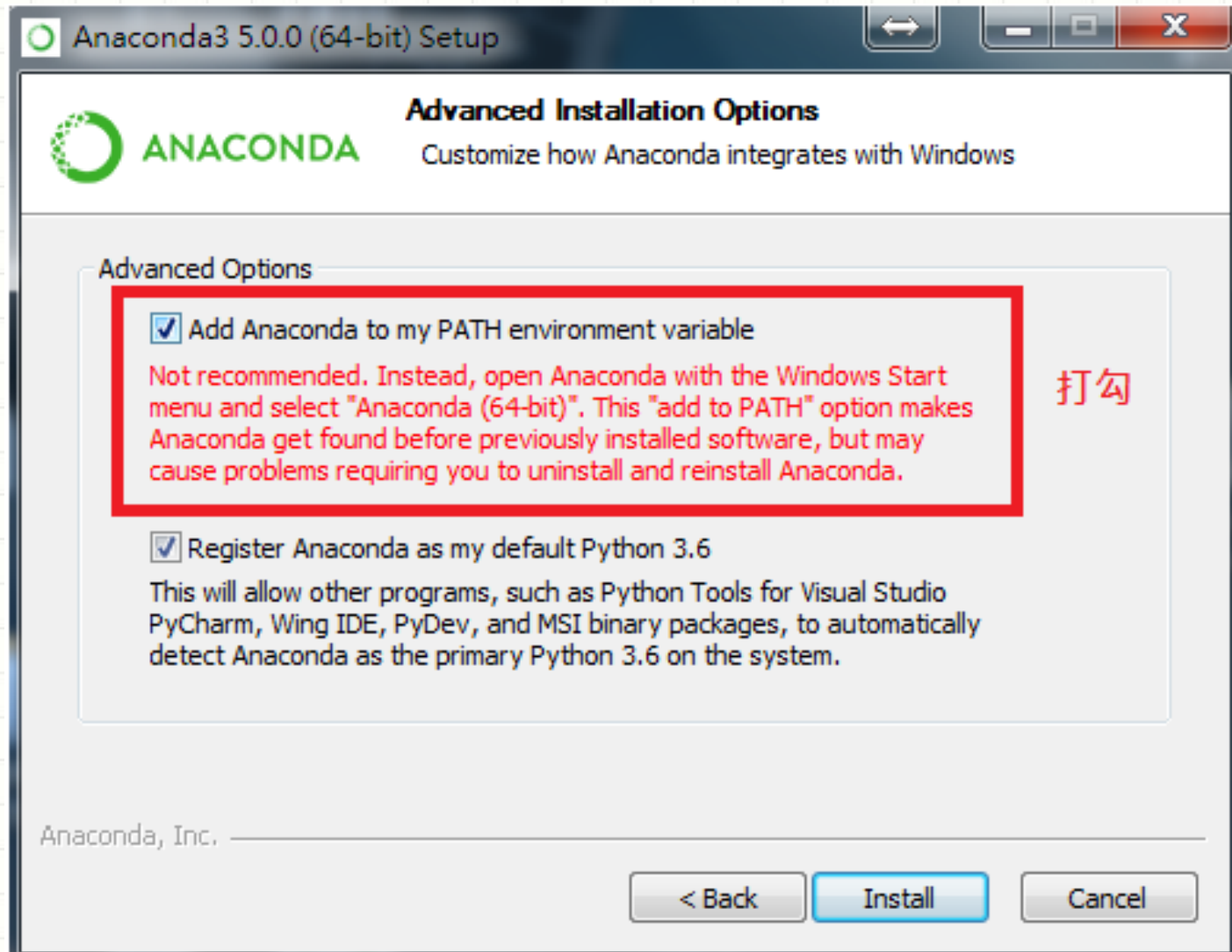
# python(anaconda)安裝介紹



# python(anaconda)安裝介紹



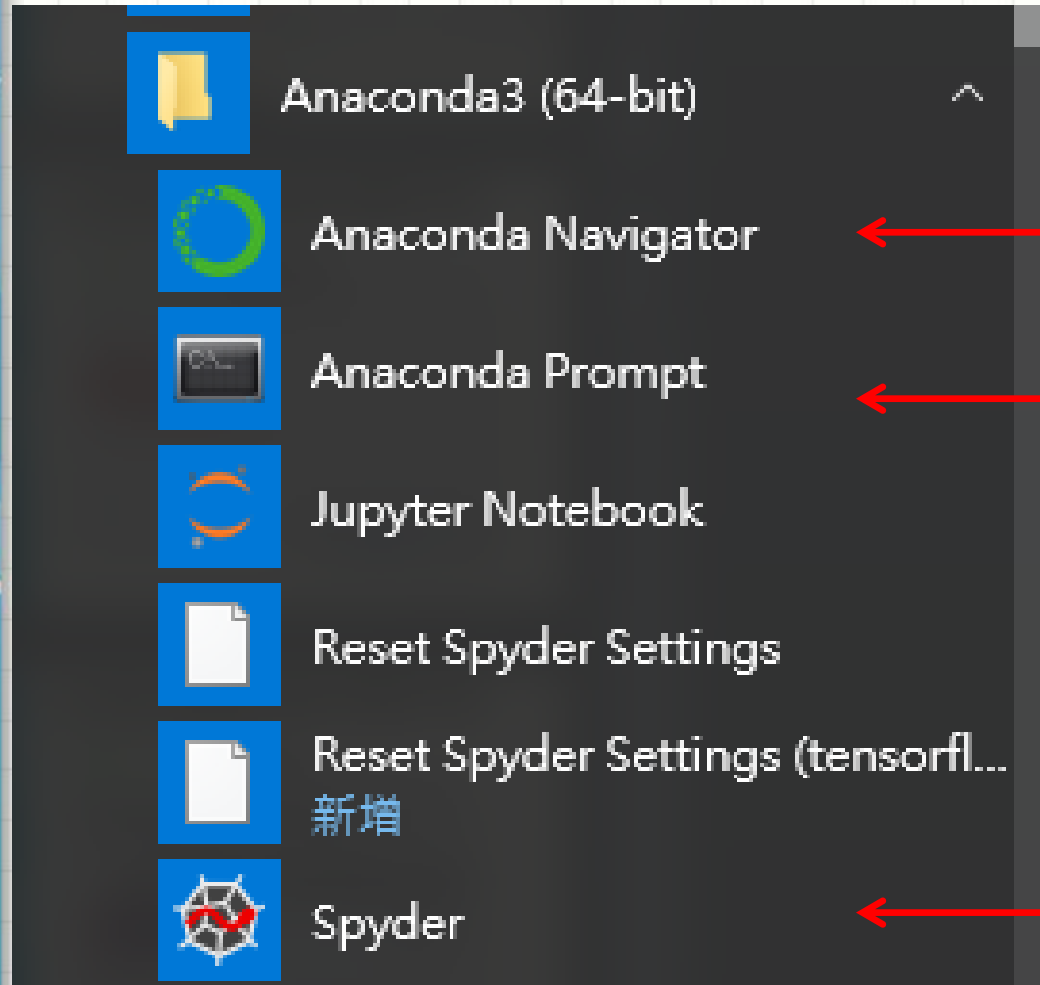
# python(anaconda)安裝介紹



此步驟時打勾，可節省之後環境變數設定python/pip指令的時間



# python(anaconda)安裝介紹



此為管理python所用到的  
套件和環境之介面

此介面可用於  
安裝新的套件

此為python  
編譯CODE的利器

# python(anaconda)安裝介紹

The screenshot displays the Spyder Python IDE interface. The main editor window on the left contains a Python script for training a Keras model. The script imports necessary libraries, defines a model, and includes training and evaluation logic. A red box highlights the code section, with a red text overlay indicating where to write code and press F9 to execute. The variable explorer on the right shows the current state of variables, with a green box highlighting the 'age\_df' and 'all\_xray\_df' DataFrames, and a green text overlay indicating the variable display area. The IPython console at the bottom shows the execution output, with a purple box highlighting the console area and a purple text overlay indicating the display results.

```
1 from keras.layers import GlobalAveragePooling2D, Dense, Dropout, Flatten, BatchNormalization
2 from keras.models import Sequential
3 base_mobilenet_model = MobileNet(input_shape = t_x.shape[1:],
4                                   include_top = False,
5                                   weights = None)
6 bone_age_model = Sequential()
7 bone_age_model.add(BatchNormalization(input_shape = t_x.shape[1:]))
8 bone_age_model.add(base_mobilenet_model)
9 bone_age_model.add(BatchNormalization())
10 bone_age_model.add(GlobalAveragePooling2D())
11 bone_age_model.add(Dropout(0.5))
12 bone_age_model.add(Dense(1, activation = 'linear')) # Linear is what 16bit did
13 from keras.metrics import mean_absolute_error
14 def mae_months(in_gt, in_pred):
15     return mean_absolute_error(boneage_div*in_gt, boneage_div*in_pred)
16
17 bone_age_model.compile(optimizer = 'adam', loss = 'mse',
18                       metrics = [mae_months])
19 from keras.callbacks import ModelCheckpoint, LearningRateScheduler, EarlyStopping, ReduceLROnPlateau
20 weight_path = "{}_weights.best.hdf5".format('bone_age')
21
22 checkpoint = ModelCheckpoint(weight_path, monitor='val_loss', verbose=1,
23                             save_best_only=True, mode='min', save_weights_only = True)
24
25 reduceLROnPlate = ReduceLROnPlateau(monitor='val_loss', factor=0.8, patience=10, verbose=1, mode='auto', epsilon=0.0001, cooldown=
26                                     patience=5) # probably needs to be more patient, but kaggle time is limited
27 callbacks_list = [checkpoint, early, reduceLROnPlate]
28
29 train_gen.batch_size = 16
30 bone_age_model.fit_generator(train_gen,
31                             validation_data = (test_X, test_Y),
32                             epochs = 2,
33                             callbacks = callbacks_list)
34
35 bone_age_model.load_weights(weight_path)
36 pred_Y = boneage_div*bone_age_model.predict(test_X, batch_size = 16, verbose = True)+boneage_mean
37 test_Y_months = boneage_div*test_Y+boneage_mean
38
39 fig, ax1 = plt.subplots(1,1, figsize = (6,6))
40 ax1.plot(test_Y_months, pred_Y, 'r.', label = 'predictions')
41 ax1.plot(test_Y_months, test_Y_months, 'b-', label = 'actual')
42 ax1.legend()
43 ax1.set_xlabel('Actual Age (Months)')
44 ax1.set_ylabel('Predicted Age (Months)')
45
46 rand_idx = np.random.choice(range(test_X.shape[0]), 8)
47 fig, m_axs = plt.subplots(4, 2, figsize = (16, 32))
48 for (idx, c_ax) in zip(rand_idx, m_axs.flatten()):
49     c_ax.imshow(test_X[idx, :, :, 0], cmap = 'bone')
50
51     c_ax.set_title('Age: %2.1f\nPredicted Age: %2.1f' % (test_Y_months[idx], pred_Y[idx]))
52     c_ax.axis('off')
53
54 fig.savefig('trained_img_predictions.png', dpi = 300)
```

顯示變數的欄位

Name	Type	Size	Value
age_df	DataFrame	(12611, 4)	Column names: id, bone_age, bone_age_path
all_xray_df	DataFrame	(12611, 3)	Column names: id, bone_age
base_bone_dir	str	1	E://temp

寫code的地方  
寫完按F9執行

顯示結果

```
In [85]:
In [85]: bone_age_model.fit_generator(train_gen,
...:                                 validation_data = (test_X, test_Y),
...:                                 epochs = 2,
...:                                 callbacks = callbacks_list)
Epoch 1/2
102/469 [=====] - ETA: 5:27:05 - loss: 0.6658 - mae_months: 51.7537
```



# 套件使用

## ANACONDA NAVIGATOR的介面

File Help

ANACONDA NAVIGATOR

Sign in to Anaconda Cloud

此處顯示有安裝及未安裝的套件

Home

Environments

Learning

Community

Documentation

Developer Blog

Feedback

Search Environments

base (root)

Not installed

Channels

Update index...

Search Packages

Name	T	Description	Version
<input type="checkbox"/> vs2015_win-32			14.0.25123
<input type="checkbox"/> vs2015_win-64			14.0.25123
<input type="checkbox"/> vs2017_win-32			15.5.2
<input type="checkbox"/> vs2017_win-64			15.5.2
<input type="checkbox"/> vtk		3d computer graphics, image processing, and visualization	8.1.0
<input type="checkbox"/> w3lib		Library of web-related functions	1.8.1
<input type="checkbox"/> waitress		Production-quality wsgi server with very acceptable performance	1.1.0
<input type="checkbox"/> webkitgtk-cos6-i686			1.4.3
<input type="checkbox"/> webkitgtk-cos6-x86_64			1.4.3
<input type="checkbox"/> webkitgtk-devel-cos6-i686			1.4.3
<input type="checkbox"/> webkitgtk-devel-cos6-x86_64			1.4.3
<input type="checkbox"/> webob		Wsgi request and response object	1.8.1
<input type="checkbox"/> websocket		Websocket implementation for gevent	0.2.1
<input type="checkbox"/> webtest		Helper to test wsgi applications	2.0.29
<input type="checkbox"/> whoosh		Full-text indexing and searching library	2.7.4
<input type="checkbox"/> winkerberos		High level interface to sspi for kerberos client auth	0.7.0
<input type="checkbox"/> word2vec		Python interface to google word2vec	0.9.2
<input type="checkbox"/> workerpool		Module for distributing jobs to a pool of worker threads	0.9.4
<input type="checkbox"/> ws4py			0.5.1

1500 packages available

Create

Clone

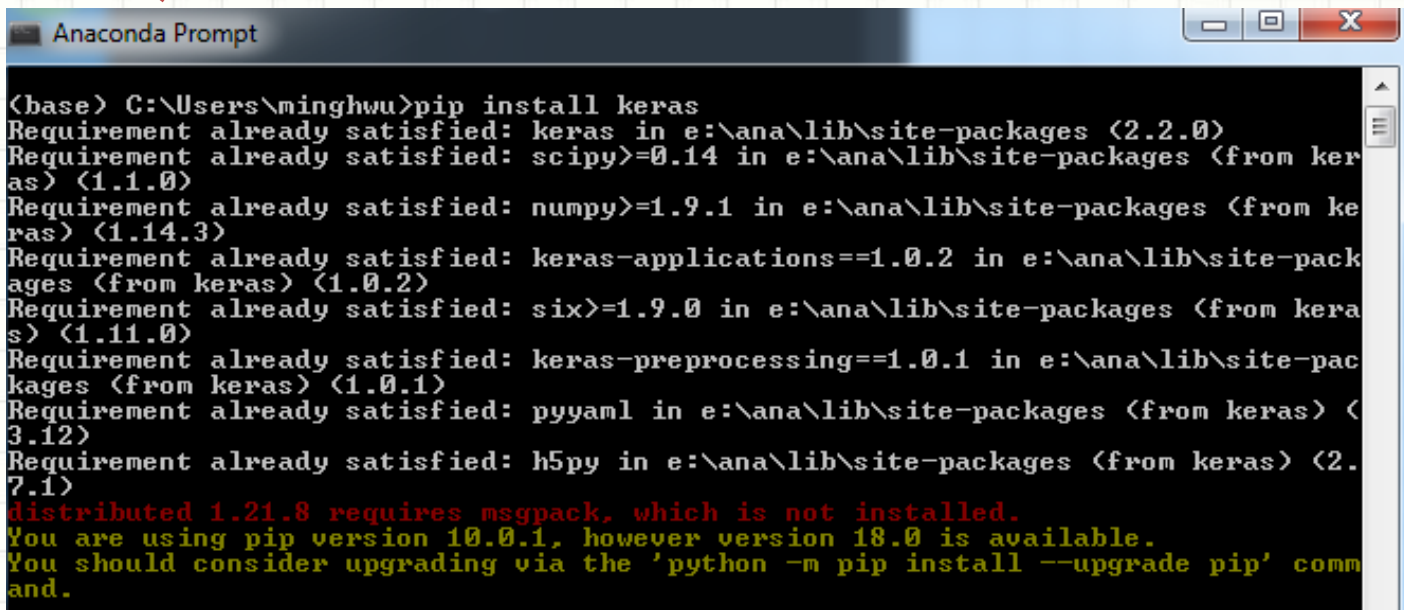
Import

Remove

# 套件使用

若要安裝全新的套件？

進入Anaconda Prompt，輸入`pip install XXX`，XXX為套件名



```
Anaconda Prompt

(base) C:\Users\minghwu>pip install keras
Requirement already satisfied: keras in e:\ana\lib\site-packages (2.2.0)
Requirement already satisfied: scipy>=0.14 in e:\ana\lib\site-packages (from ke
ras) (1.1.0)
Requirement already satisfied: numpy>=1.9.1 in e:\ana\lib\site-packages (from ke
ras) (1.14.3)
Requirement already satisfied: keras-applications==1.0.2 in e:\ana\lib\site-pack
ages (from keras) (1.0.2)
Requirement already satisfied: six>=1.9.0 in e:\ana\lib\site-packages (from kera
s) (1.11.0)
Requirement already satisfied: keras-preprocessing==1.0.1 in e:\ana\lib\site-pac
kages (from keras) (1.0.1)
Requirement already satisfied: pyyaml in e:\ana\lib\site-packages (from keras) (
3.12)
Requirement already satisfied: h5py in e:\ana\lib\site-packages (from keras) (2.
7.1)
distributed 1.21.8 requires msgpack, which is not installed.
You are using pip version 10.0.1, however version 18.0 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' comm
and.
```

# 套件使用

常用的套件：

- 圖片相關
  - ❖ PIL(可視為套件的套件)
- 科學計算&資料分析
  - ❖ numpy(太常用  $\Sigma$  ☆)
  - ❖ matplotlib( $\Sigma$  ☆)
  - ❖ pandas( $\Sigma$  ☆)
  - ❖ scikit-learn( $\Sigma$  ☆)
  - ❖ sklearn
- 深度學習
  - ❖ tensorflow
  - ❖ keras

.....too much

# 套件使用

## 如何使用？

法一：

```
import PIL
```

```
im = PIL.Image.open('p7(1).jpg')
```

.表示  
PIL再呼  
叫套件  
Image

.表示Image呼  
叫函式open

法二：

```
from PIL import Image
```

```
im = Image.open('p7(1).jpg')
```

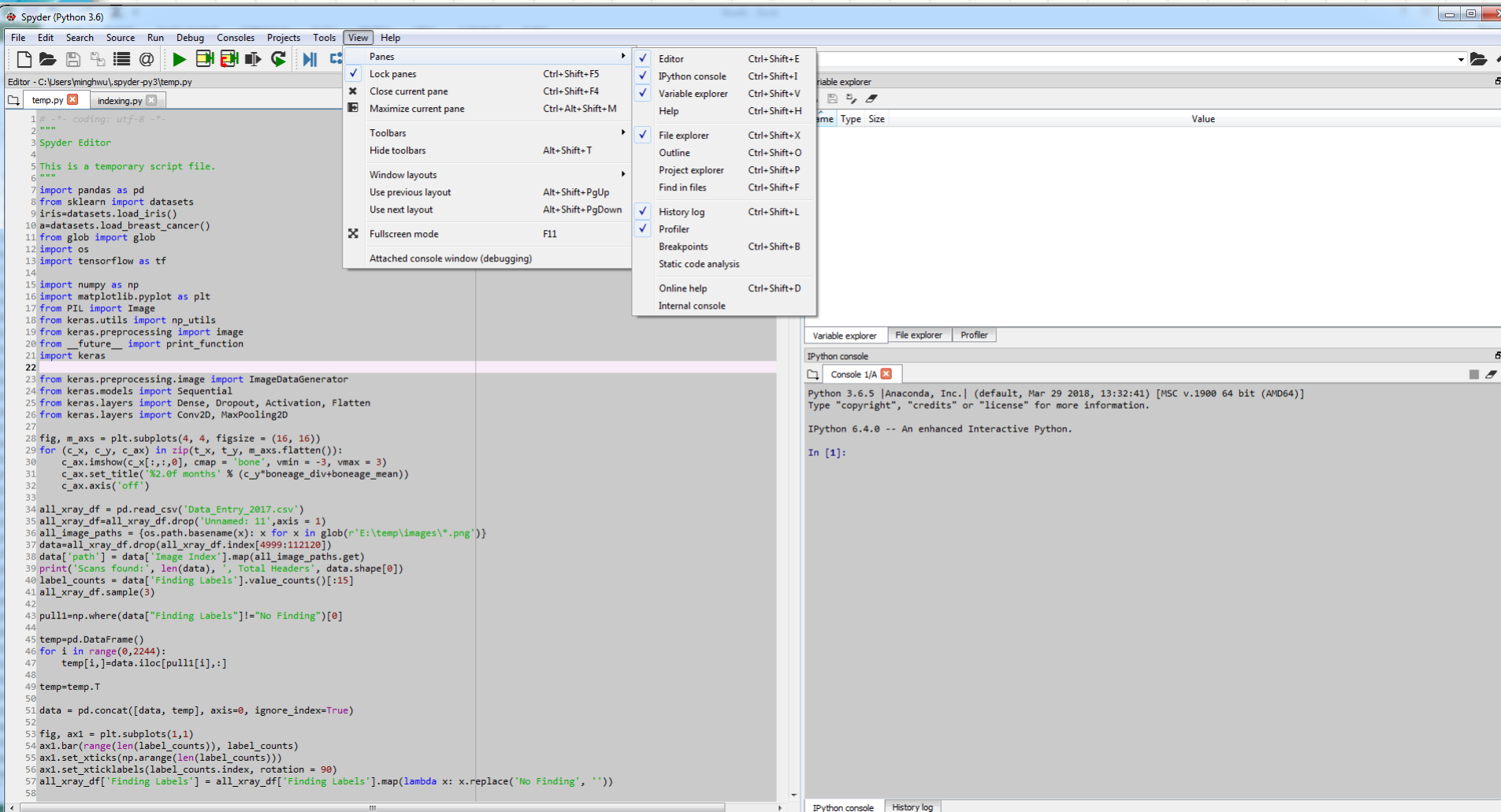
```
import pandas as pd
```

—— 載入pandas套件並簡寫成pd

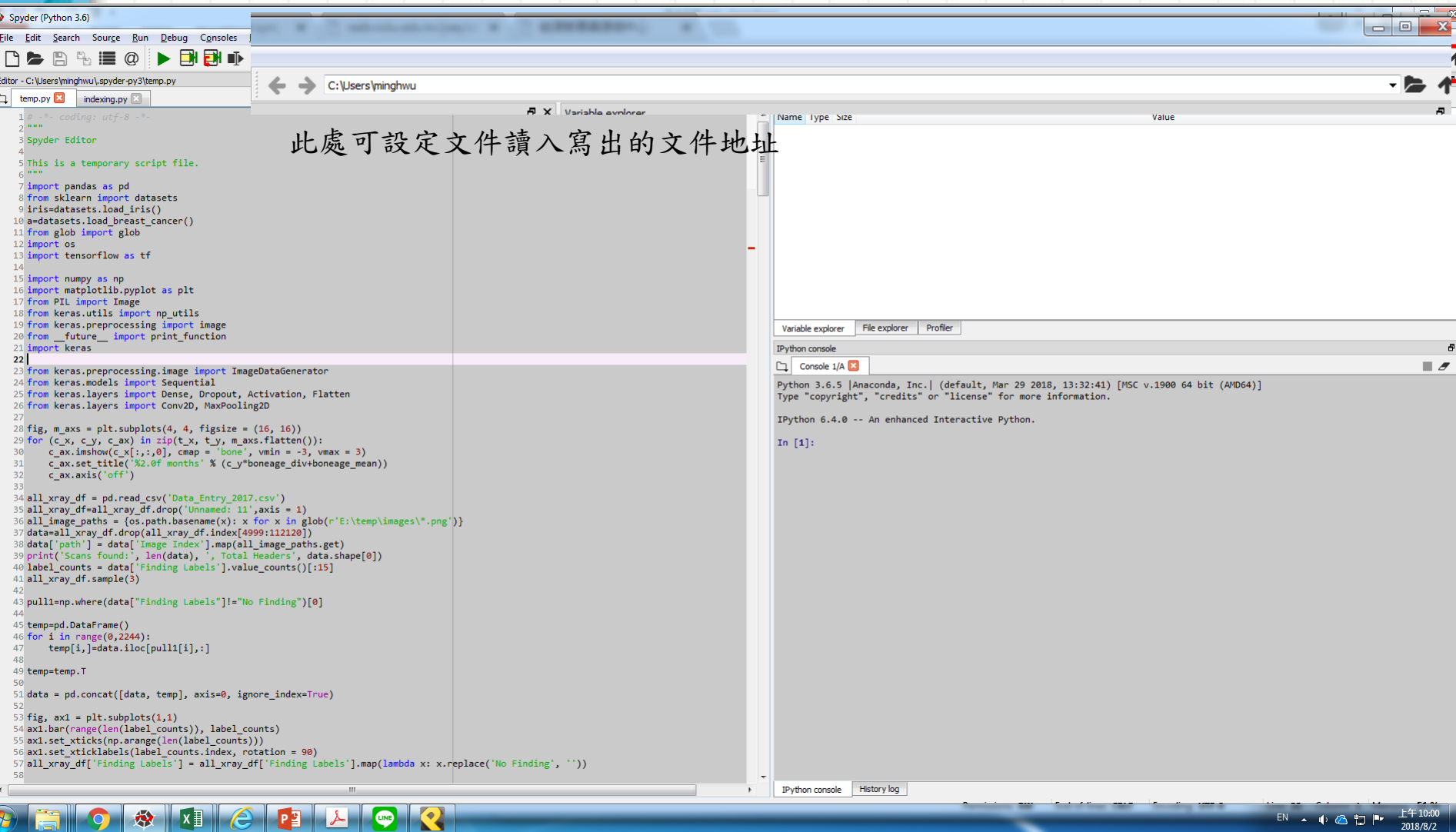
```
import numpy as np
```

# 介面使用

從View → Panes 可選取各種介面

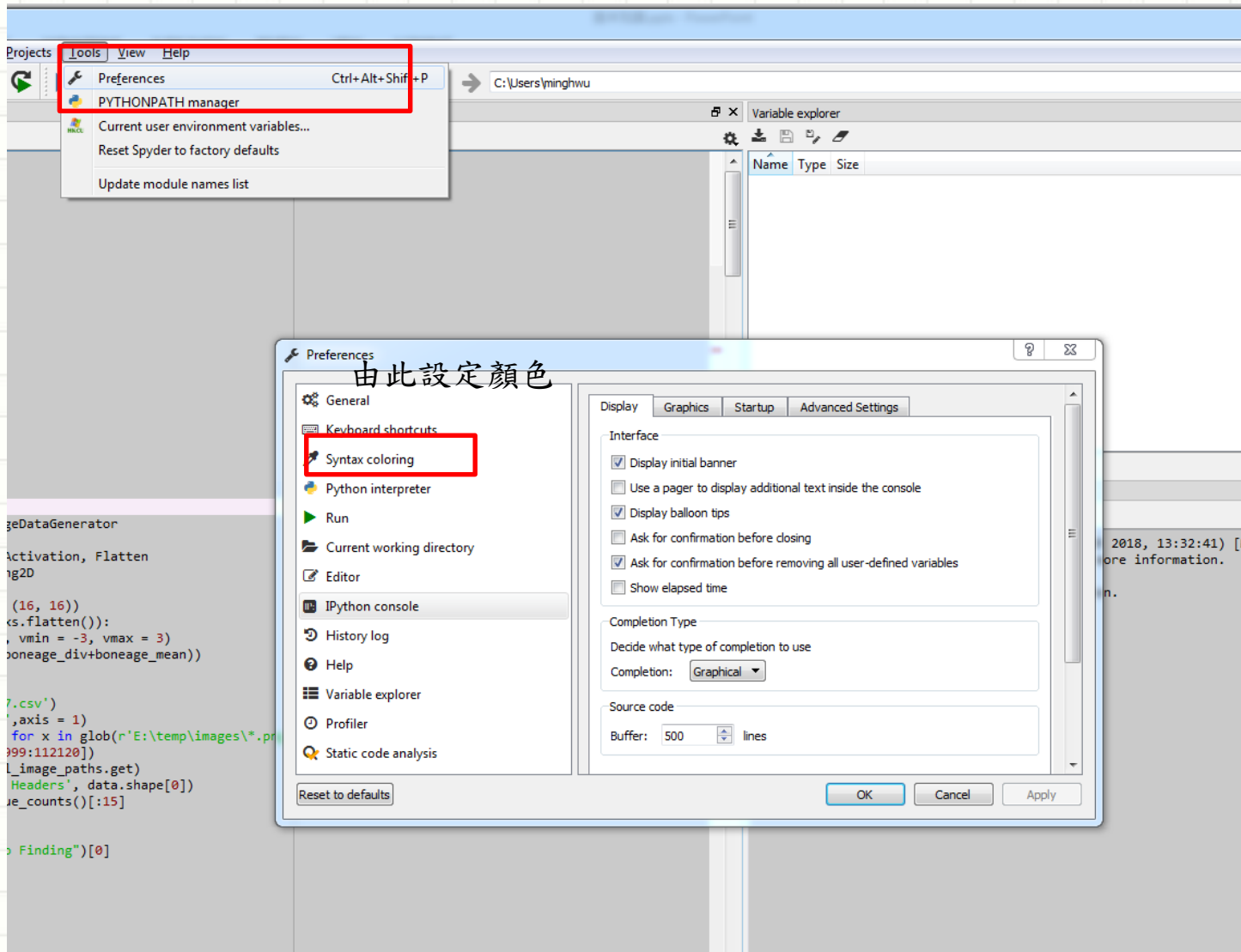


# 介面使用





# 介面使用



# 小結

## Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.



Note: Data do not include managers Source: Burtch Works

The Wall Street Journal

# 小結

## About data scientist

### Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY  SHARE  COMMENT  TEXT SIZE  PRINT \$8.95 BUY COPIES

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early." Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would land in. He could imagine that new features capitalizing on the heuristics he was developing might provide value to users. But LinkedIn's engineering team, caught up in the challenges of scaling up the site, seemed uninterested. Some colleagues were openly dismissive of Goldman's ideas. Why would users need LinkedIn to figure out their networks for them? The site already had an address book importer that could pull in all a member's connections.

Luckily, Reid Hoffman, LinkedIn's cofounder and CEO at the time (now its executive chairman), had faith in the power of analytics because of his experiences at PayPal, and he had granted Goldman a high degree of autonomy. For one thing, he had given Goldman a way to circumvent the traditional product release cycle by publishing small modules in the form of ads on the site's most popular pages.

# 小結

但日理萬機的我們要從何開始？

Ctrl C

Ctrl V

# 相關資源介紹

## Kaggle

Use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies. [Got it](#) [Learn more](#)

[kaggle](#)  [Competitions](#) [Datasets](#) [Kernels](#) [Discussion](#) [Learn](#) [Sign In](#)

**Datasets** [Documentation](#) [New Dataset](#)

Join Kaggle's newest Data Science for Good challenge with PASSNYC. Click to learn more and participate to win from \$15,000 in prizes.

Public Sort by: Hotness

9,132 Datasets Sizes File types Licenses Tags Search datasets

33		<b>Air Quality in Madrid (2001-2018)</b> Different pollution levels in Madrid from 2001 to 2018 Decide Soluciones updated a month ago	time series pollution	Other 150.6 MB Other	</> 5 2 6k
87		<b>120 years of Olympic history: athletes and results</b> basic bio data on athletes and medal results from Athens 1896 to Rio 2016 Randi H Griffin updated 2 months ago	olympic ga... sports history	CSV 5.4 MB CC0	</> 17 1 11k
124		<b>Mobile App Store ( 7200 apps)</b> Analytics for Mobile Apps Ramanathan updated 2 months ago	business internet mobile web	CSV 5.6 MB GPL	</> 33 7 20k
507		<b>PASSNYC: Data Science for Good Challenge</b> Help PASSNYC determine which schools need their services the most PASSNYC updated a month ago	education demograph... data visuali... recommen...	CSV 163.8 KB CC0	</> 106 21 57k
18		<b>LA County Restaurant Inspections and Violations</b> Environmental health inspections and violations in LA County restaurants	food and dr... public health	CSV 18.7 MB CC0	</> 6 2 3k

搜尋有興趣的  
資料集



# 相關資源介紹

Featured Dataset

## Flowers Recognition

This dataset contains labeled 4242 images of flowers.

Alexander Mamaev · last updated a month ago

148 voters

Data Overview **Kernels** Discussion Activity

Download (225 MB) New Kernel

Public Your Work Favorites

Sort by Hotness

Outputs Languages Types Search kernels

- 32 Flowers are mesmerizing  
7mo ago tutorial, beginner, advanced
- 14 Flower Recognition - FastAI 94% Accuracy  
19d ago
- 6 Introduction to Deep Learning  
2mo ago
- 4 Flow(ers) Data Preparation  
1mo ago tutorial, data cleaning
- 2 Using InceptionV3  
1mo ago
- 2 Recognize my flower  
4mo ago image processing, classification, feature engineering, object recognition
- 1 Introduction to Transfer Learning  
1mo ago
- 1 kernel17c732dbd1  
1mo ago

有大量  
高人寫的  
CODE提供參  
考



# 相關資源介紹

## Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 440 data sets as a service to the machine learning community. You may [view all data sets](#) through our searchable interface. Our [old web site](#) is still available, for those who prefer the old format. For a general overview of the Repository, please visit our [About page](#). For information about citing data sets in publications, please read our [citation policy](#). If you wish to donate a data set, please consult our [donation policy](#). For any other questions, feel free to [contact the Repository librarians](#). We have also set up a [mirror site](#) for the Repository.

Supported By:



In Collaboration With:



### Latest News:

04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!  
03-01-2010: [Note](#) from donor regarding Netflix data  
10-16-2009: Two new data sets have been added.  
09-14-2009: Several data sets have been added.  
07-23-2008: [Repository mirror](#) has been set up.  
03-24-2008: New data sets have been added!  
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope












### Featured Data Set: [ICU](#)



Data Type: Multivariate, Time-Series

Data set prepared for the use of participants for the 1994 AAAI Spring Symposium on Artificial Intelligence in Medicine.

### Newest Data Sets:

- 07-13-2018:  [EEG Steady-State Visual Evoked Potential Signals](#)
- 06-06-2018:  [Simulated Falls and Daily Living Activities Data Set](#)
- 06-01-2018:  [Multimodal Damage Identification for Humanitarian Computing](#)
- 05-31-2018:  [Victorian Era Authorship Attribution](#)
- 05-06-2018:  [GNFUV Unmanned Surface Vehicles Sensor Data](#)
- 04-26-2018:  [Condition monitoring of hydraulic systems](#)
- 04-14-2018:  [SCADI](#)
- 04-09-2018:  [Sports articles for objectivity analysis](#)
- 04-05-2018:  [Absenteeism at work](#)
- 04-05-2018:  [Carbon Nanotubes](#)
- 03-29-2018:  [Optical Interconnection Network](#)

### Most Popular Data Sets (hits since 2007):

- 2022128:  [Iris](#)
- 1228193:  [Adult](#)
- 939548:  [Wine](#)
- 807873:  [Car Evaluation](#)
- 739484:  [Breast Cancer Wisconsin \(Diagnostic\)](#)
- 703957:  [Wine Quality](#)
- 703112:  [Heart Disease](#)
- 675006:  [Human Activity Recognition Using Smartphones](#)
- 656745:  [Forest Fires](#)
- 652419:  [Bank Marketing](#)
- 640218:  [Abalone](#)

# 相關資源介紹

輸入您想要搜尋的關鍵字 (資料集)



還在爬蟲嗎？本平臺35192筆資料集全都放在這裡

## 資料集服務分類



生育保健(344)



出生及收養(46)



求學及進修(561)



服兵役(181)



求職及就業(498)



開創事業(427)



婚姻(4)



投資理財(1532)



休閒旅遊(844)



交通及通訊(1588)



就醫(845)



購屋及遷徙(577)