你今天表特了嗎

part2 - python & crawler

NC

Beauty Pool

minstrel.tw/beauty/

你今天表特了嗎

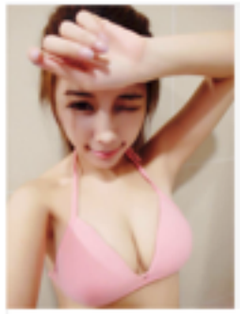
想看什麼妹

搜尋條件

推文數: 文章數: submit

爆。正妹



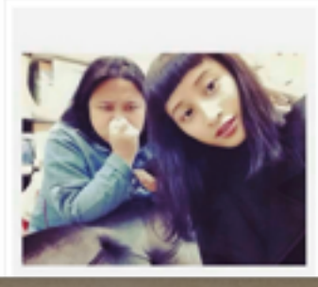







最新情報





使用到的技術

- 網站前端：html, css, jQuery, bootstrap
- 網站後端：Django
- 爬蟲程式：python (requests, pyquery)
- 資料庫：mongoDB (pymongo)
- 伺服器：apache

今天分享的部分

- 爬蟲程式：python (requests, pyquery)
- 資料庫：mongoDB (pymongo)

Python

Outline

- Why Python?
- Install
- hello world
- Property
- Type
- Control Flow
- Function
- File I/O
- Object and Class
- Module
- Built-in function
- Reference

Why Python?

- PEP 20

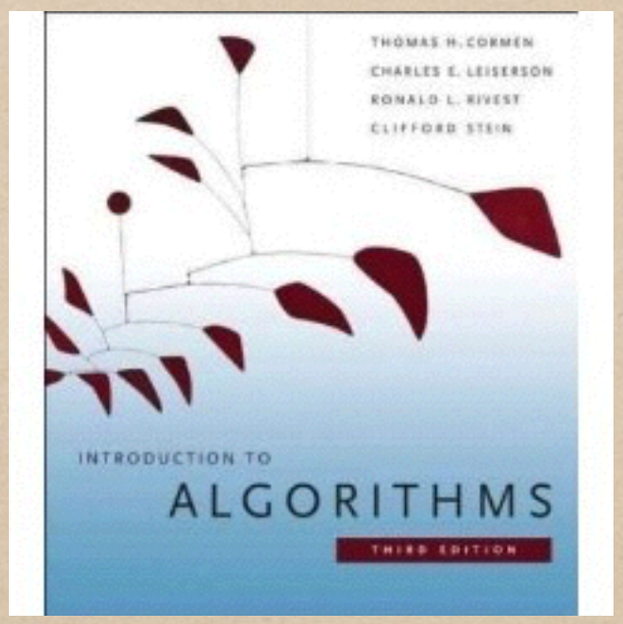
```
>>> import this
The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
```

Why Python?

- 好寫、好讀、好學
- 不用加分號「;」（想想看你被它陰過幾次）
- 利用縮行取代括號
- 好用的內建型別與內建函式

“I was translating pseudocode into Python.

It got smaller and more readable.”



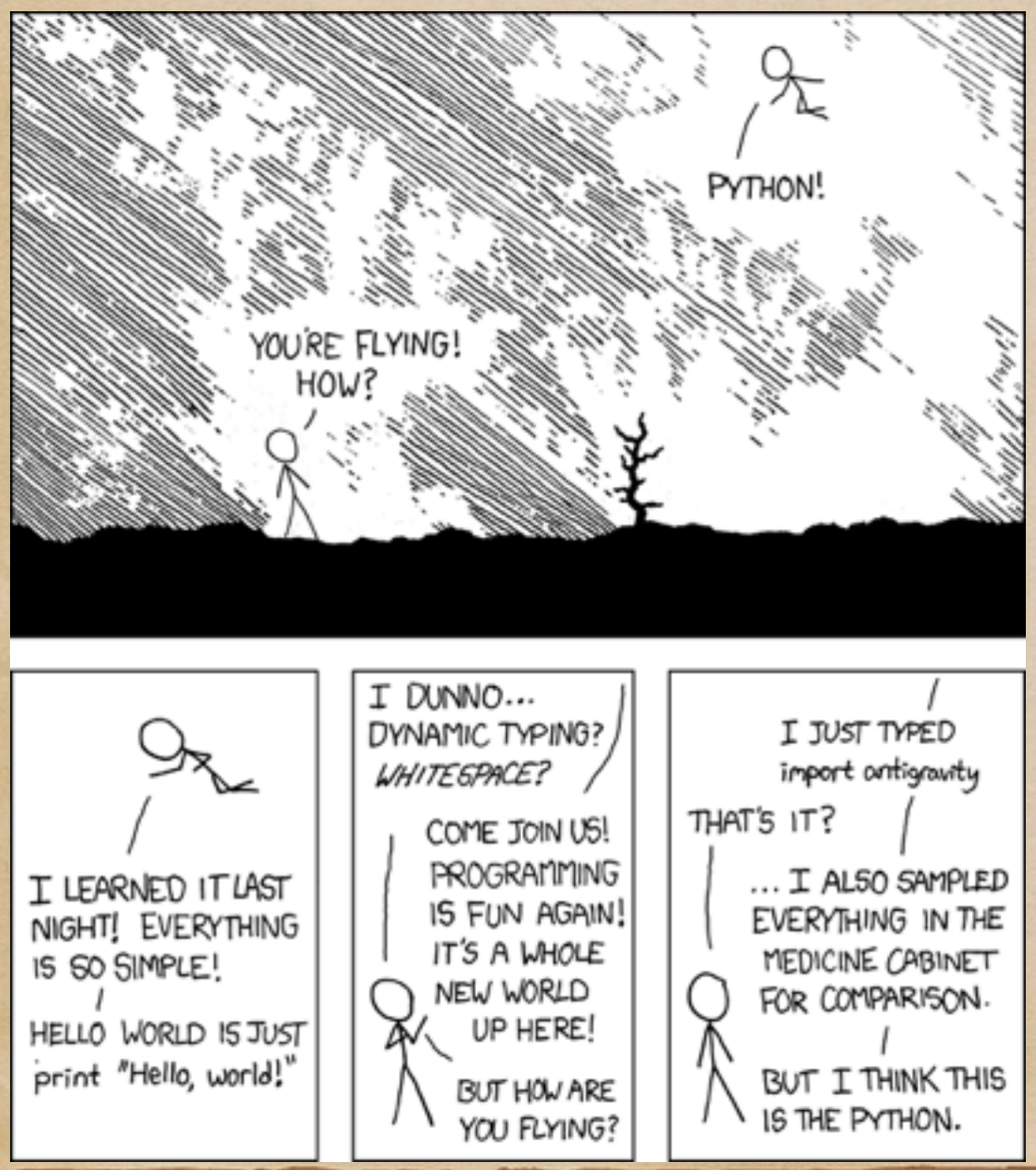
– Allen B. Downey (author of Think Python)

Why Python?

```
$ python -c "import antigravity"
```

Why Python?

```
$ python -c "import antigravity"
```



Only Python 2 today…

and something important related to Crawler / Django

Install

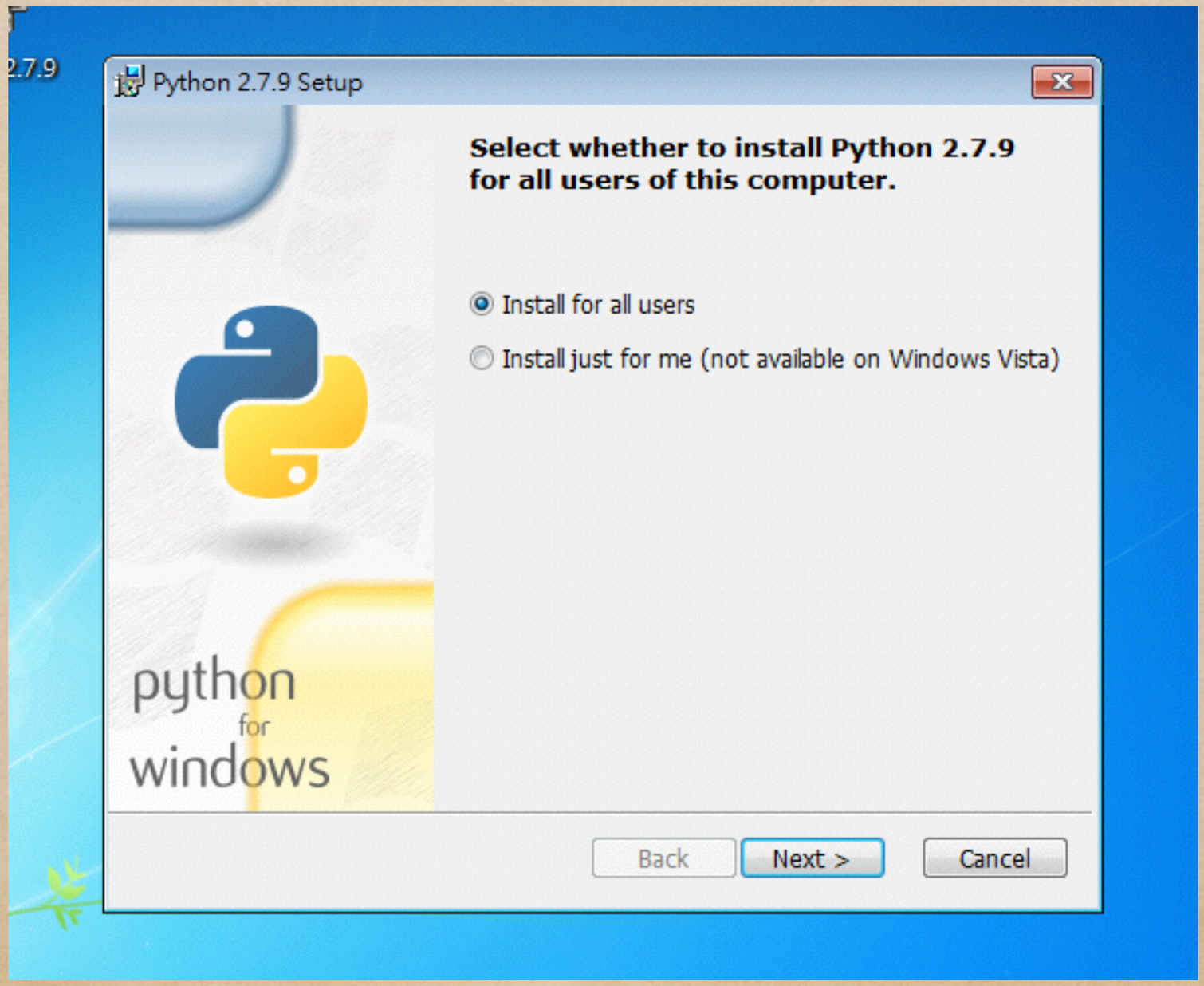
Install

- Windows (win7)
  - https://www.python.org/downloads/
  - 下載2.7.9

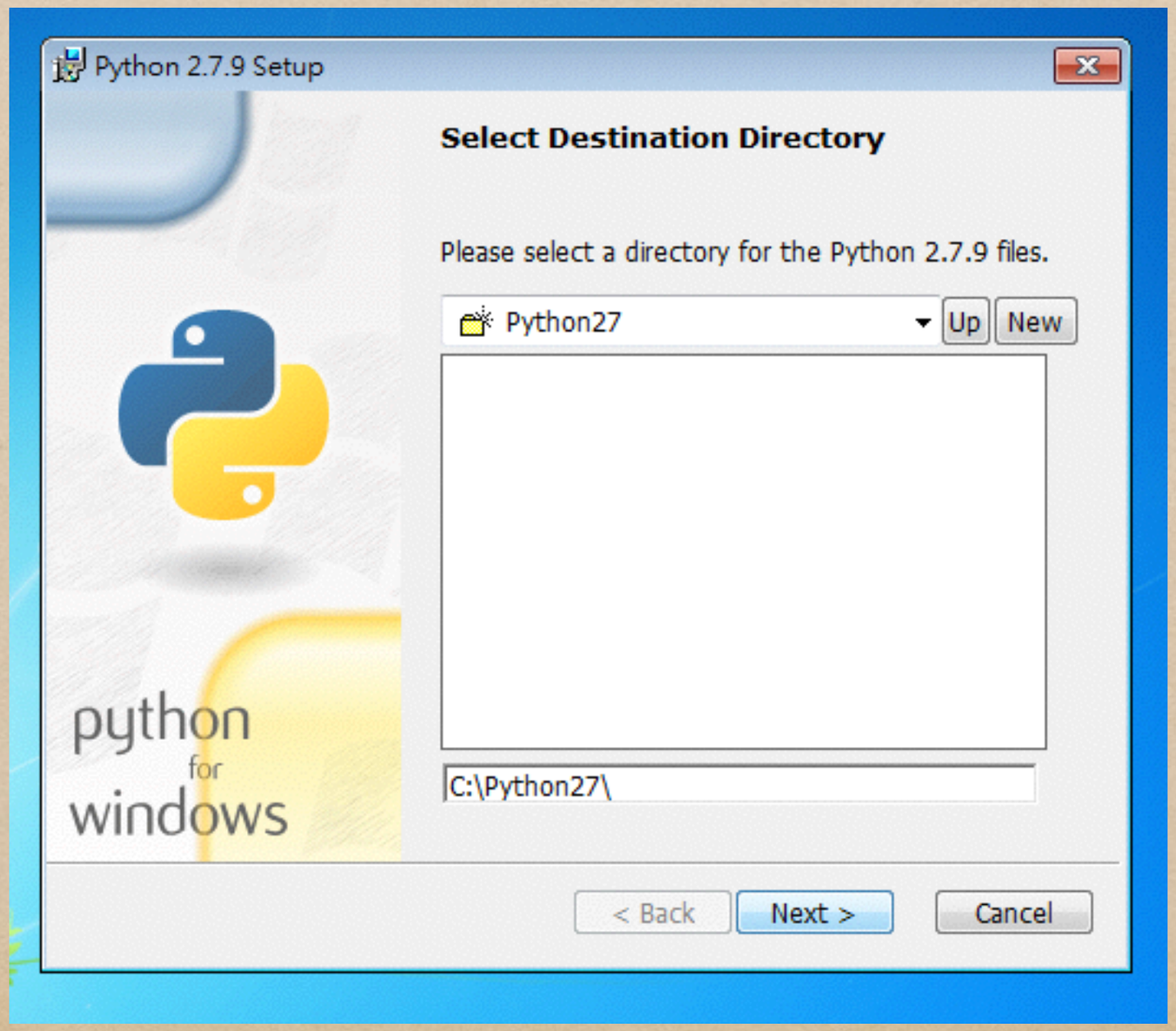
Files

Version	Operating System	Description	MD5 Sum	File Size	GPG
Gzipped source tarball	Source release		5eebcaa0030dc4061156d3429657fb83	16657930	SIG
XZ compressed source tarball	Source release		38d530f7efc373d64a8fb1637e3baaa7	12164712	SIG
Mac OS X 32-bit i386/PPC installer	Mac OS X	for Mac OS X 10.5 and later	8d8a26fed767302ff38bc5056612c73a	23759976	SIG
Mac OS X 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.6 and later	307c2b99a212204e7a1182a354328e94	22006891	SIG
Windows debug information files	Windows		c5838ec1cdd529a7065902c7573d1607	25969730	
Windows debug information files for 64-bit binaries	Windows		544e1137e8ecdce4f4cd2954ea520fa7	23979074	
Windows help file	Windows		dd438e999824c48001e54a2138c4f455	6120616	
Windows x86-64 MSI installer	Windows	for AMD64/EM64T/x64, not Itanium processors	21ee51a9f44b7160cb6fc68e29a1ddd0	18833408	
Windows x86 MSI installer	Windows		3ed20d8b06dcd339f814b38861f88fc9	18309120	

設定使用者
按「next」

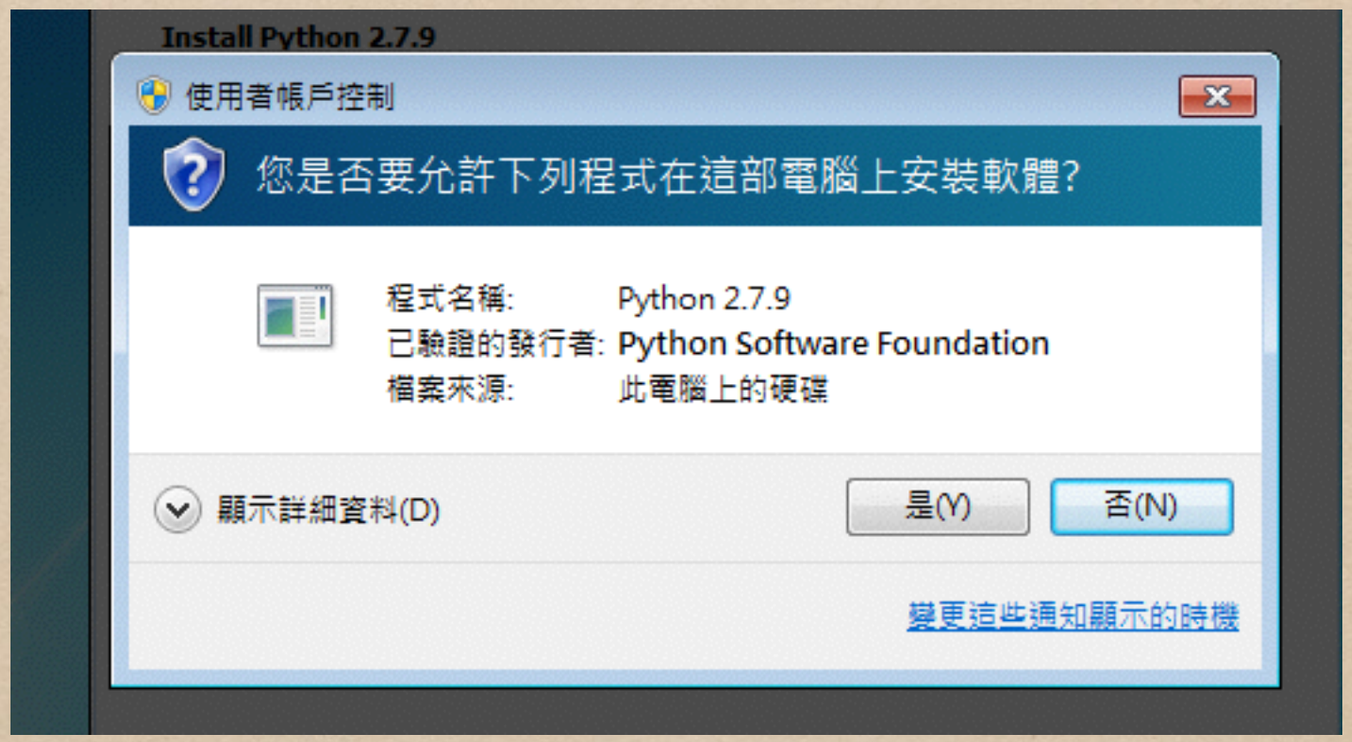

2.7.9
Python 2.7.9 Setup
Select whether to install Python 2.7.9 for all users of this computer.
Install for all users
Install just for me (not available on Windows Vista)
python
for
windows
Back
Next >
Cancel

設定安裝路徑
再按「next」



設定安裝項目
再按「next」

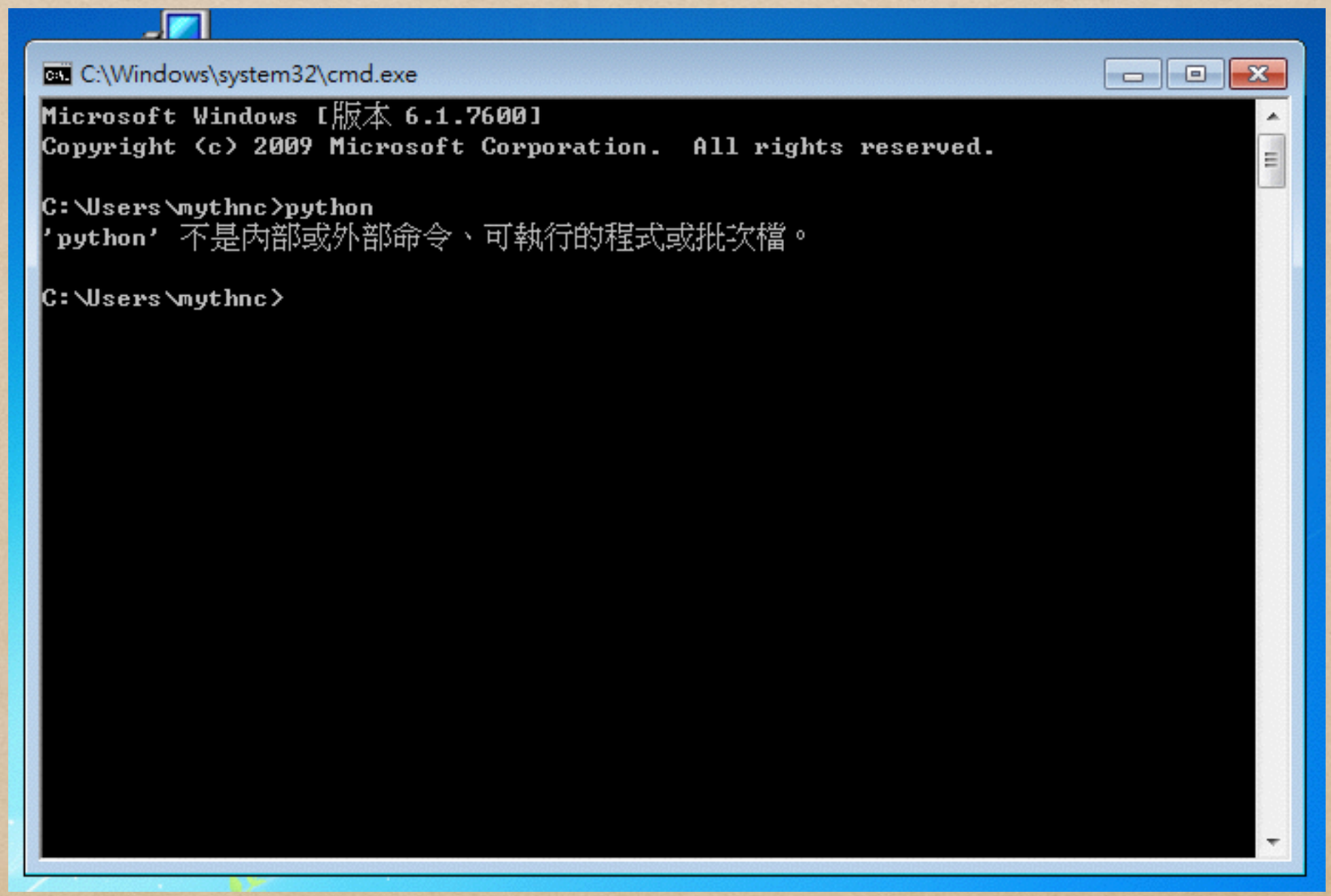
按「是」


Install Python 2.7.9
使用者帳戶控制
您是否要允許下列程式在這部電腦上安裝軟體?
程式名稱: Python 2.7.9
已驗證的發行者: Python Software Foundation
檔案來源: 此電腦上的硬碟
顯示詳細資料(D)
是(Y)
否(N)
變更這些通知顯示的時機

安裝完成

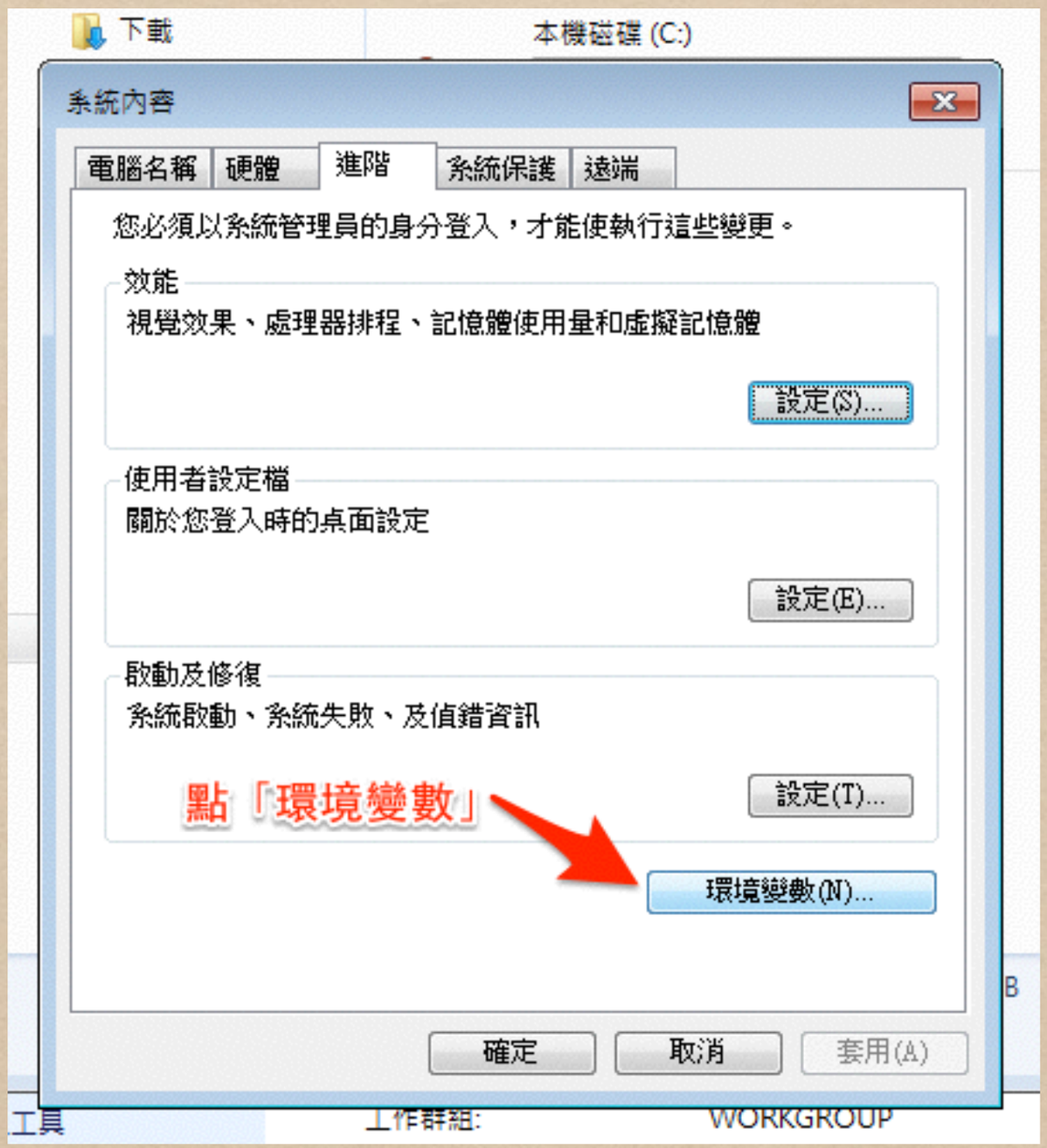

Python 2.7.9 Setup
Complete the Python 2.7.9 Installer
Special Windows thanks to:
Mark Hammond, without whose years of freely shared Windows expertise, Python for Windows would still be Python for DOS.
python
for
windows
Click the Finish button to exit the Installer.
< Back
Finish
Cancel

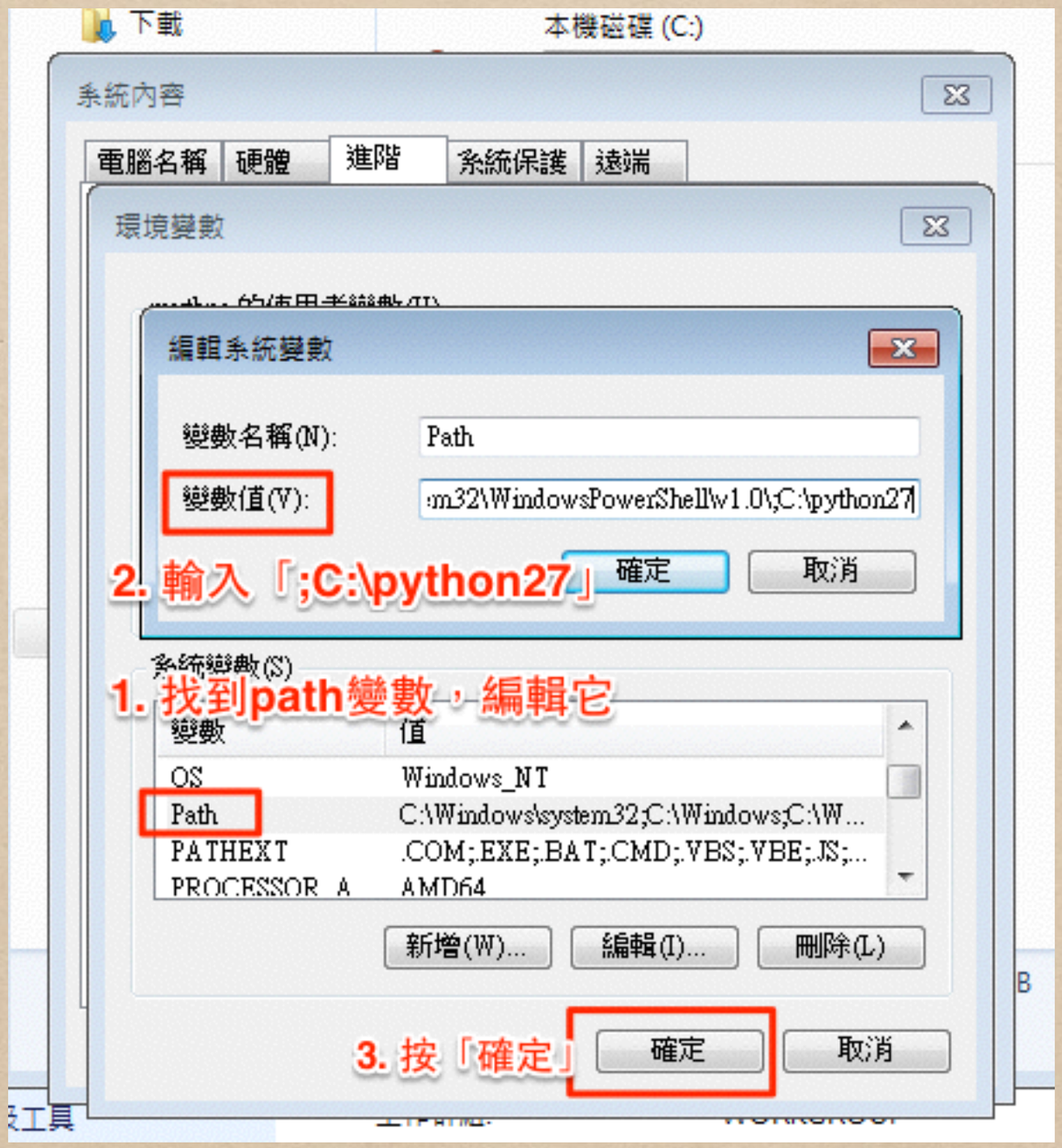
沒設定PATH，錯誤示範



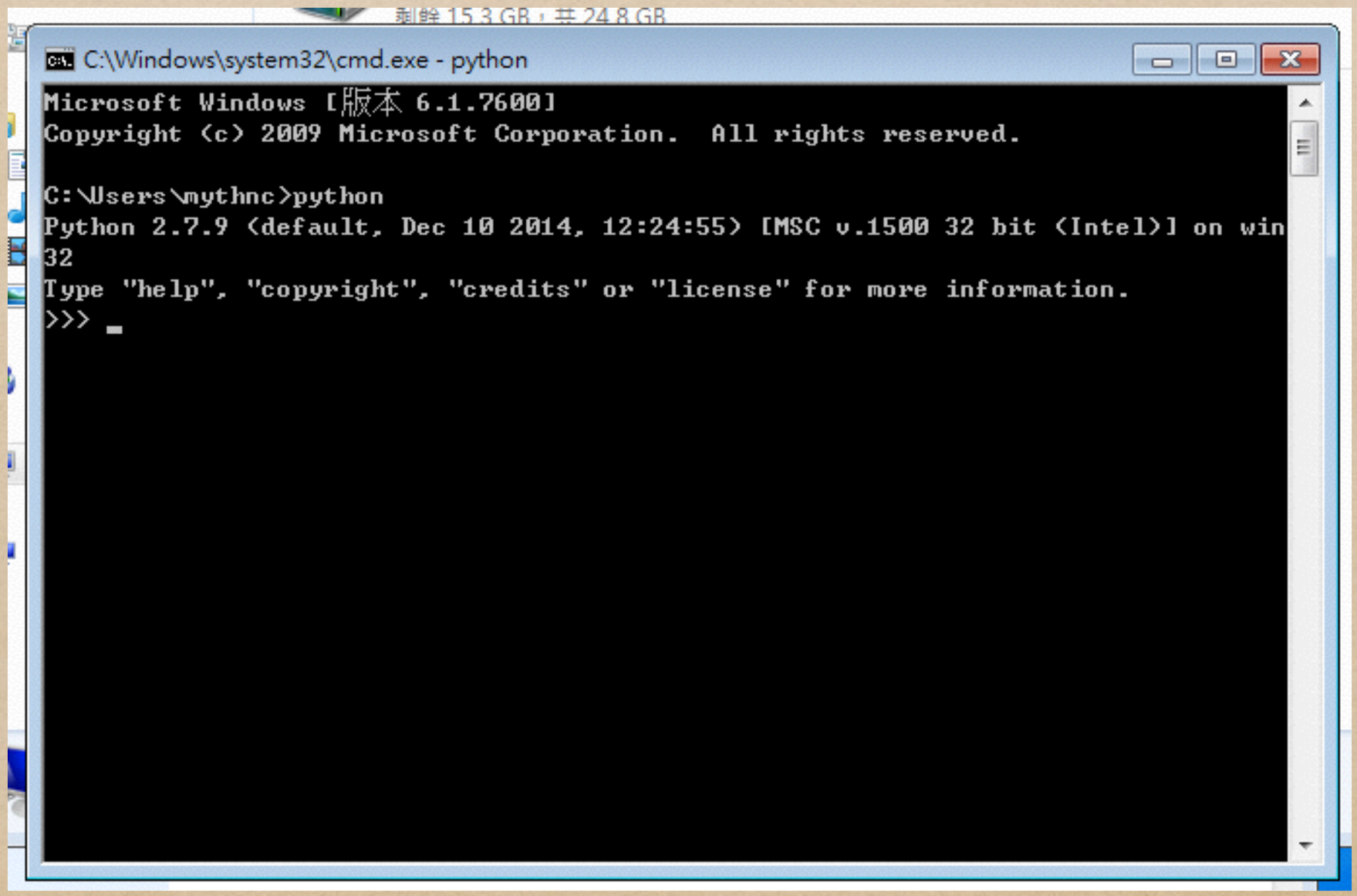

電腦
搜尋 電腦
組合管理
系統內容
解除安裝或變更程式
連線網路磁碟機
開啟控制台
我的最愛
下載
桌面
最近的位置
媒體櫃
文件
音樂
視訊
圖片
家用群組
電腦
網路
硬碟 (1)
本機磁碟 (C:)
剩餘 15.3 GB，共 24.8 GB
裝置中含有卸除式存放裝置 (1)
CD 光碟機 (D:)
1. 開啟「電腦」
檢視(V)
排序方式(O)
分組方式(P)
重新整理(E)
貼上(P)
貼上捷徑(S)
復原 複製(U) Ctrl+Z
新增一個網路位置(L)
內容(R)
2. 點內容
MYTHNC-PC 工作群組: WORKGROUP
記憶體: 2.00 GB
處理器: Intel(R) Core(TM) i5-3...


控制台 ▸ 系統及安全性 ▸ 系統
搜尋控制台
控制台首頁
裝置管理員
遠端設定
系統保護
進階系統設定
點「進階系統設定」
檢視電腦的基本資訊
Windows 版本
Windows 7 企業版
Copyright © 2009 Microsoft Corporation. All rights reserved.
系統
分級: 系統分級無法使用
處理器: Intel(R) Core(TM) i5-3210M CPU @ 2.50GHz 2.50 GHz
安裝的記憶體 (RAM): 2.00 GB
系統類型: 64 位元作業系統
手寫筆與觸控: 此顯示器不提供手寫筆或觸控式輸入功能。
電腦名稱、網域及工作群組設定
電腦名稱: mythnc-PC
變更設定
完整電腦名稱: mythnc-PC
電腦描述:
工作群組: WORKGROUP
請參閱
行動作業中心
Windows Update
效能資訊及工具


下載
本機磁碟 (C:)
系統內容
電腦名稱
硬體
進階
系統保護
遠端
您必須以系統管理員的身分登入，才能使執行這些變更。
效能
視覺效果、處理器排程、記憶體使用量和虛擬記憶體
設定(S)...
使用者設定檔
關於您登入時的桌面設定
設定(E)...
啟動及修復
系統啟動、系統失敗、及偵錯資訊
設定(T)...
點「環境變數」
環境變數(N)...
確定
取消
套用(A)
工具
工作群組:
WORKGROUP


下載
本機磁碟 (C:)
系統內容
電腦名稱
硬體
進階
系統保護
遠端
環境變數
編輯系統變數
變數名稱(N):
Path
變數值(V):
m32\WindowsPowerShell\v1.0\;C:\python27
確定
取消
2. 輸入「;C:\python27」
系統變數(S)
1. 找到path變數，編輯它
變數
值
OS
Windows_NT
Path
C:\Windows\system32;C:\Windows;C:\W...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;...
PROCESSOR_A
AMD64
新增(W)...
編輯(I)...
刪除(L)
3. 按「確定」
確定
取消

成功！


剩餘 15.3 GB，共 24.8 GB
C:\Windows\system32\cmd.exe - python
Microsoft Windows [版本 6.1.7600]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.
C:\Users\mythnc>python
Python 2.7.9 (default, Dec 10 2014, 12:24:55) [MSC v.1500 32 bit (Intel)] on win
32
Type "help", "copyright", "credits" or "license" for more information.
>>>

Install

- Mac
  - http://brew.sh
  - 需下載「命令列開發者工具」（系統會提示）
  - $ brew install python
- 可裝可不裝

Install

- Linux (Debian / Ubuntu)
  - $ sudo apt-get install python
- 可裝可不裝

Install

- How to check?
- type “python” in terminal / command line

```
Python 2.7.9 (default, Jan  7 2015, 11:49:12)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.56)]
on darwin
Type "help", "copyright", "credits" or "license" for
more information.
>>>
```

- Interactive mode (very useful!!)

Install

- 如果安裝遇到問題，試著自行解決。
  （善用google）

- 因為你會一直遇到安裝的問題
  （安裝ruby...安裝java...安裝eclipse...安裝...）

pip path setting

- A tool for installing python packages
- windows練習題：試著加入pip的路徑於path中
  - 一樣的做法
  - 在path中加入「;C:\Python27\Scripts」
  - 如此，即可在cmd中使用pip

IPython

- 神兵利器
- 易查詢，提供補齊

- `$ pip install ipython`

- windows使用者請再裝pyreadline才有補齊功能
- `$ pip install pyreadline # windows`

- `$ ipython`

IPython

- 練習題
  1. 試著在ipython中查詢函式，並閱讀說明
     輸入"help(str.find)"與"str.find?"，
     並比較之間的差異

  2. 試著查詢module
     輸入"import csv"，接著輸入"csv?"與"csv??"，"csv."按<tab>，並比較之間的差異

IPython

- 練習題
  3. 試著隨便輸入變數(a = 3, b = 's')
     接著輸入'who', 'whos'，你有什麼發現?

4. 試著交替使用python與ipython，
   你覺得哪個比較好用？

Editor / IDE

- Pick whatever you like
  - VIM, Emacs, Sublime, Notepad++…etc
  - 如果可以，選個可以寫任何程式語言的編輯器
- https://wiki.python.org/moin/PythonEditors

Hello World

hello world!

1. in Interactive mode

```
In [1]: print 'hello world!'
hello world!
```

2. in py file (hello.py) (在windows上免寫第一行)

```
#!/usr/bin/env python

print 'hello world!'
```

```
$ python hello.py
hello world!
```

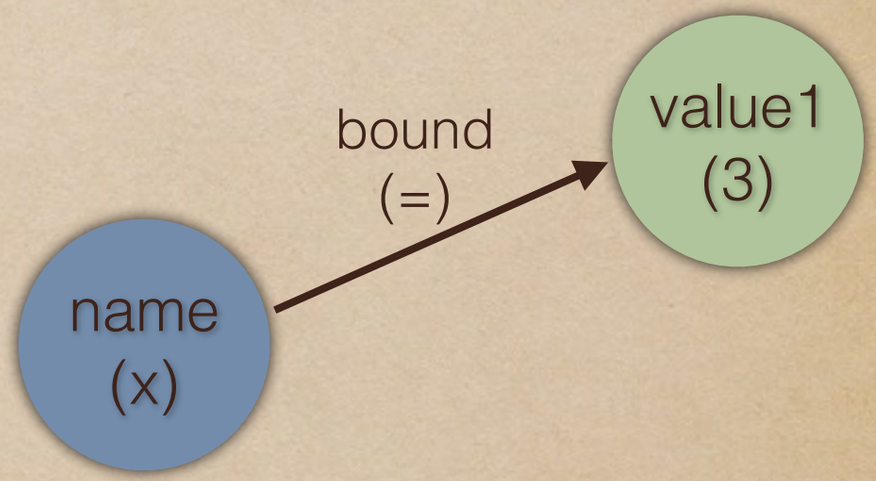
Property

- Dynamic Typing
  - checking types in run time
  - 變數不需要事先宣告
  - a name bound to a value

Property

- Dynamic Typing
  - checking types in run time
  - 變數不需要事先宣告
  - a name bound to a value

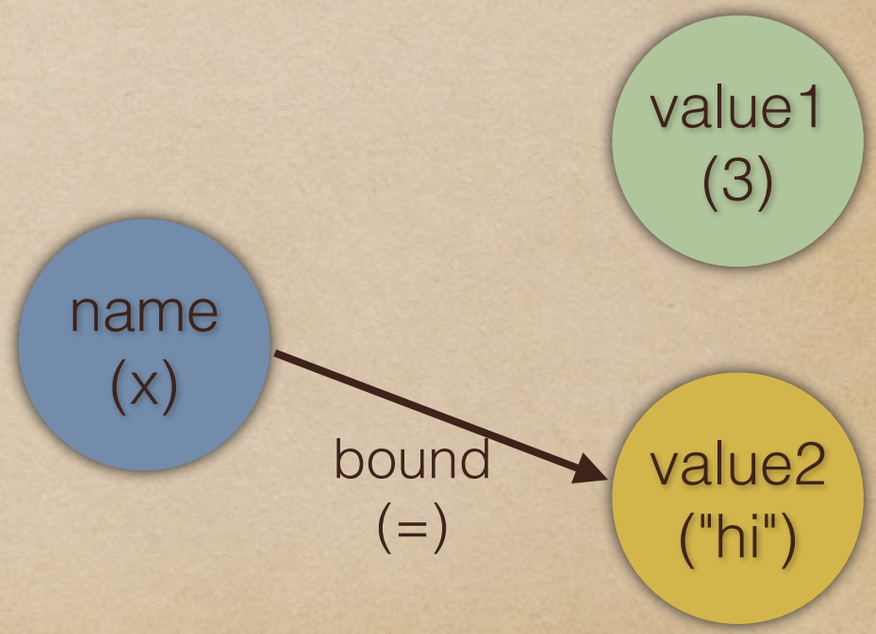
```
>>> x = 3
```



Property

- Dynamic Typing
  - checking types in run time
  - 變數不需要事先宣告
  - a name bound to a value

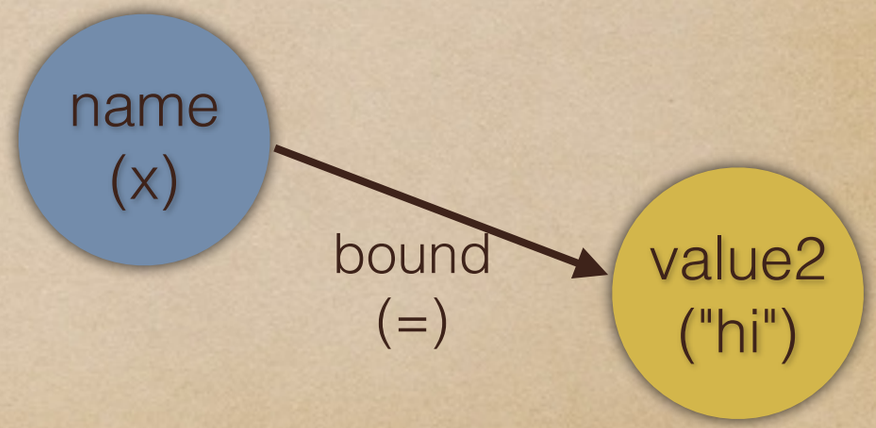
```
>>> x = 3
>>> x = "hi"
```



Property

- Dynamic Typing
  - checking types in run time
  - 變數不需要事先宣告
  - a name bound to a value

```
>>> x = 3
>>> x = "hi"
```



Property

- Strong Typing
  - 不同型別無法相互運算

```
>>>'hello' + 3  # TypeError
```

```
>>>'hello' + str(3)  # work
'hello3'
```

Property

- mutable
  - value can be modified after created.
  - dict, list, set
- immutable
  - value can’t be modified after created.
  - number, str, tuple

Type

- None
- Numbers: int, long, float, complex, bool
- Sequences: str, unicode, list, tuple, xrange
- Mapping: dict
- Sets: set, frozenset

Type - 今天會講到的

-
- Numbers: int, , float, , bool
- Sequences: str, , list, tuple,
- Mapping: dict
- Sets: set,

Please check the python_tutorial.ipynb file

Built-in Function

Built-in Functions				
abs()	divmod()	input()	open()	staticmethod()
all()	enumerate()	int()	ord()	str()
any()	eval()	isinstance()	pow()	sum()
basestring()	execfile()	issubclass()	print()	super()
bin()	file()	iter()	property()	tuple()
bool()	filter()	len()	range()	type()
bytearray()	float()	list()	raw_input()	unichr()
callable()	format()	locals()	reduce()	unicode()
chr()	frozenset()	long()	reload()	vars()
classmethod()	getattr()	map()	repr()	xrange()
cmp()	globals()	max()	reversed()	zip()
compile()	hasattr()	memoryview()	round()	__import__()
complex()	hash()	min()	set()	apply()
delattr()	help()	next()	setattr()	buffer()
dict()	hex()	object()	slice()	coerce()
dir()	id()	oct()	sorted()	intern()

don't reinvent the wheel!!

https://docs.python.org/2.7/library/functions.html

3rd party library

- 主要還是看你想做什麼(寫網頁, 寫系統, 文字處理, 資料分析, 寫database...etc)

- 幾乎都有相對應的第三方函式庫跟教學文件 - PyPI

- 善用pip管理python套件

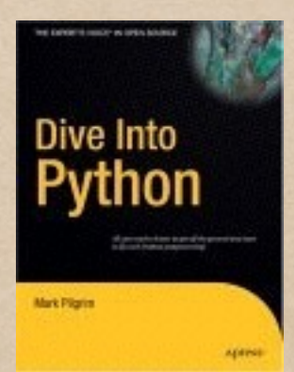
Reference

Resources

- Python / 第一次用就上手
  - http://wiki.python.org.tw/Python/第一次用就上手
- ptt@python板
- 官方文件很完整
- 電子週報 - PythonWeekly

E-books

1. Dive Into Python (free)



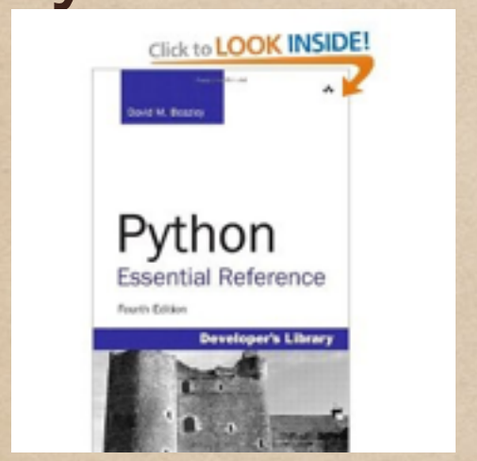
   - http://www.diveintopython.net

2. Think Python (free)

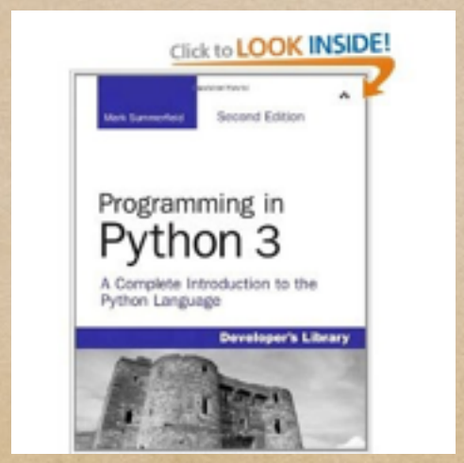
   - http://www.greenteapress.com/thinkpython/thinkpython.html

Books

3. Python Essential Reference



4. Programming in Python 3



Community

- Taipei.py
- PyHUG (新竹)
- Taichung.py
- Tainan.py
- PyLadies @ Taiwan (女生專屬)
- 花蓮.py
- … …

Conference

- 2012首次舉辦
- PyCon APAC 2015
  - 6/5 ~ 6/7 @中研院
    - 學生票很便宜！
    - 研究生沒有的福利
- 好吃又好玩
- 錄影檔會釋出放在youtube

Question?

Crawler

預先安裝套件

- 依序裝入下列套件
- $ pip install requests
- $ pip install pyquery
- $ 安裝mongodb
  - Install MongoDB on Windows
- $ pip install pymongo

下載對應的安裝檔

Production Release (3.0.2)

4/9/2015 Release Notes Changelog

Download Source: tgz | zip

Windows | Linux | Mac OS X | Solaris

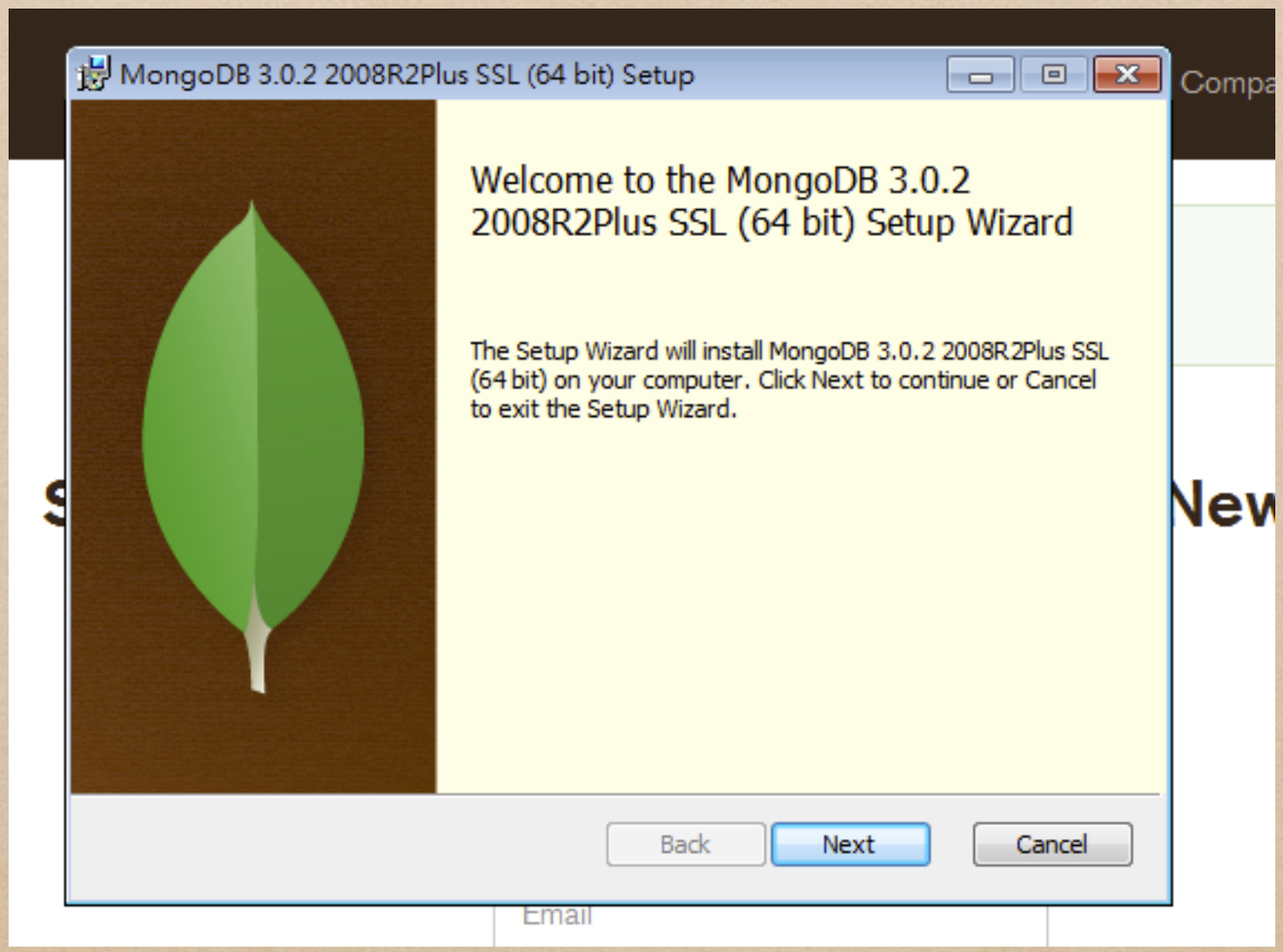
VERSION:

Windows 64-bit 2008 R2+

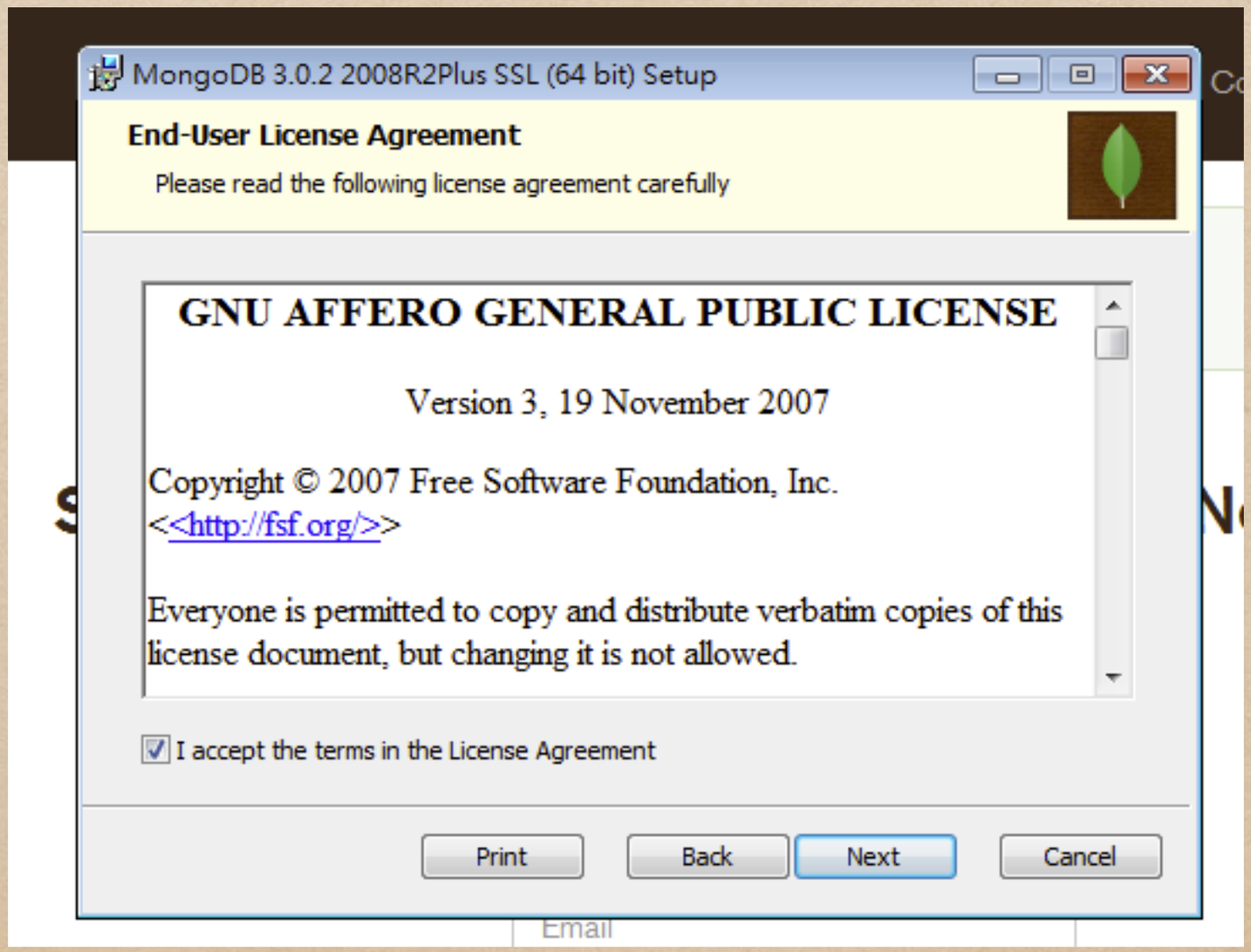
INSTALLATION PACKAGE: Installation Instructions View Build Archive

DOWNLOAD (MSI)

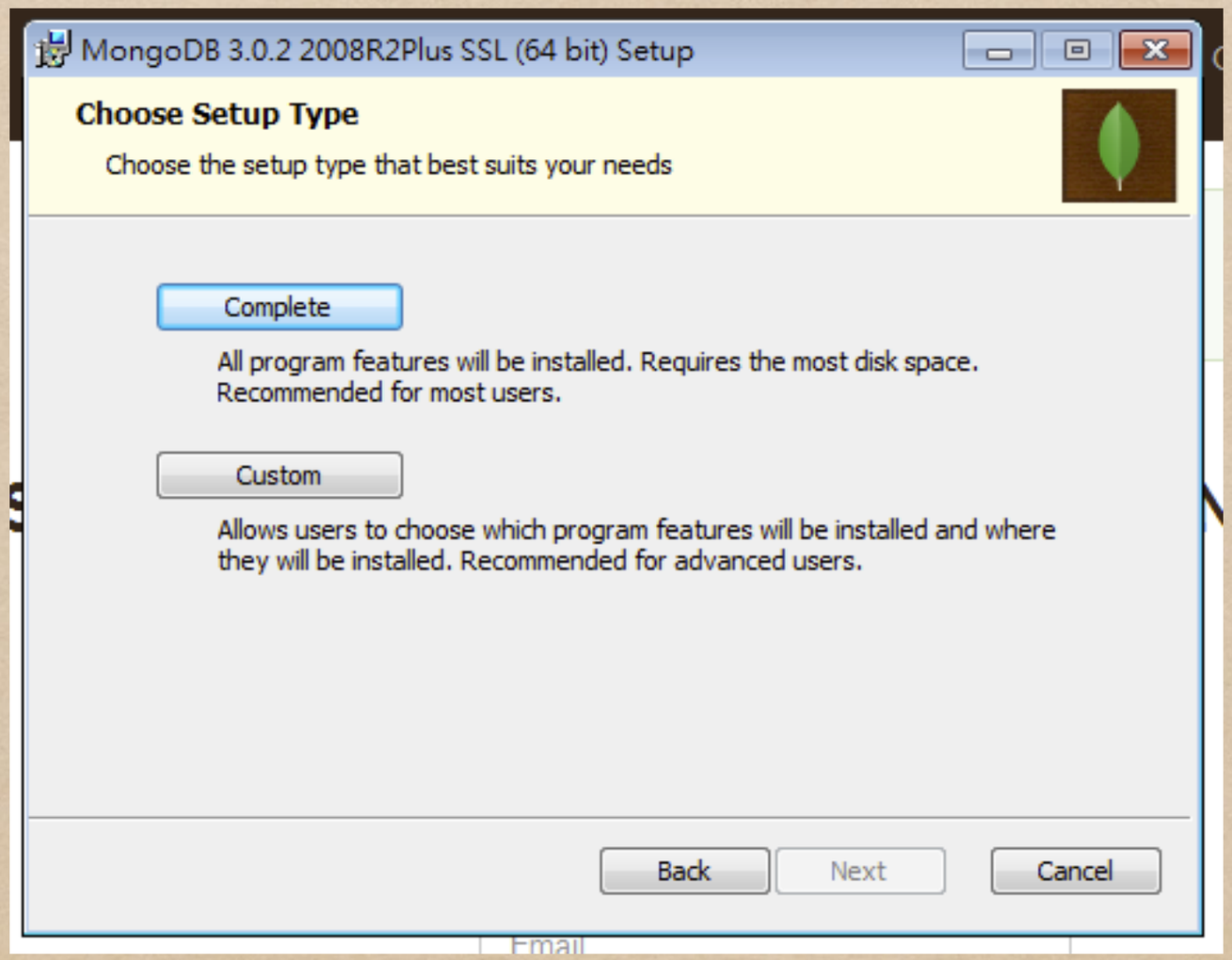
按next


MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup
Welcome to the MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup Wizard
The Setup Wizard will install MongoDB 3.0.2 2008R2Plus SSL (64 bit) on your computer. Click Next to continue or Cancel to exit the Setup Wizard.
Back
Next
Cancel

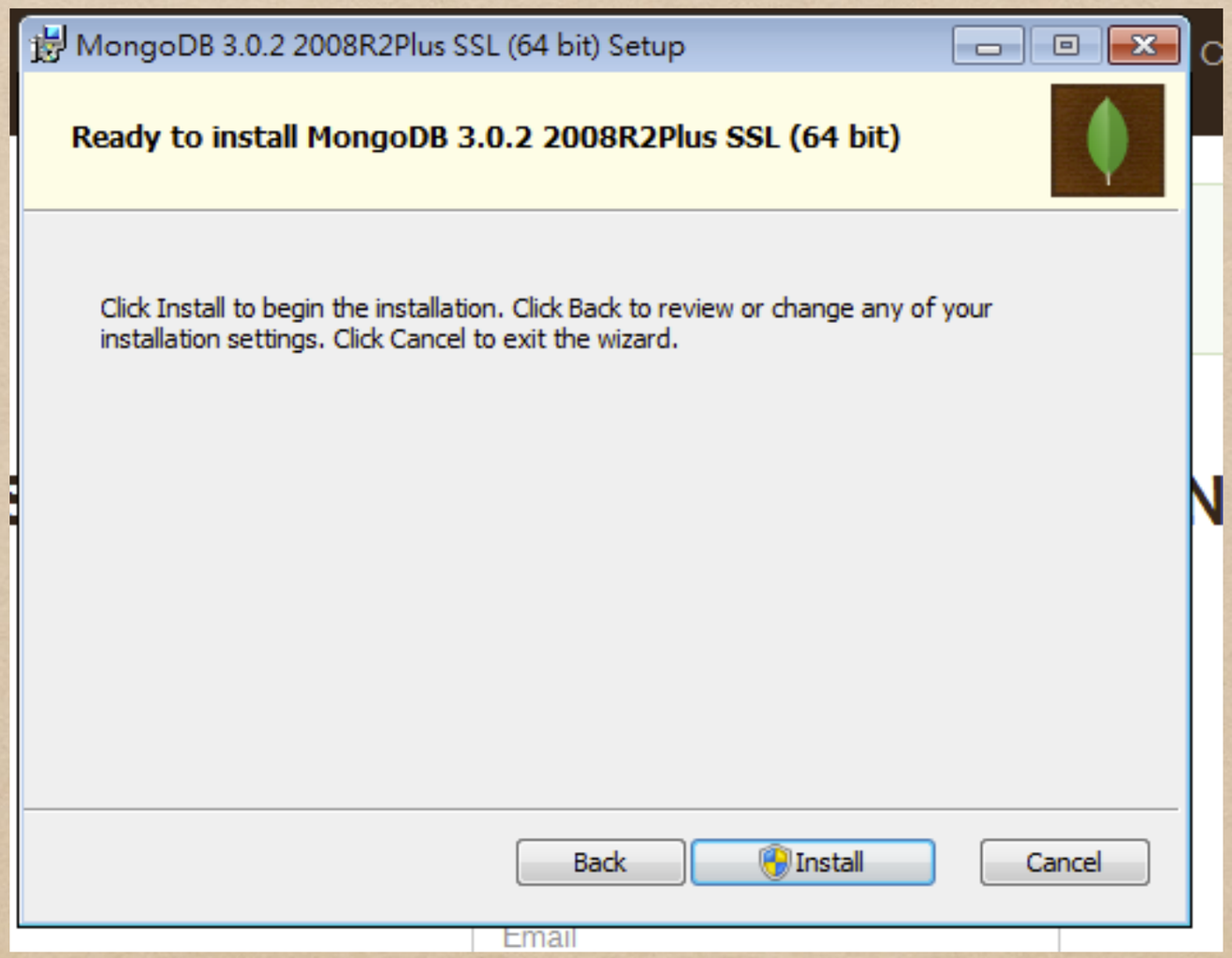
打勾，按next


MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup
End-User License Agreement
Please read the following license agreement carefully
GNU AFFERO GENERAL PUBLIC LICENSE
Version 3, 19 November 2007
Copyright © 2007 Free Software Foundation, Inc.
<<http://fsf.org/>>
Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.
I accept the terms in the License Agreement
Print
Back
Next
Cancel

選complete，按next



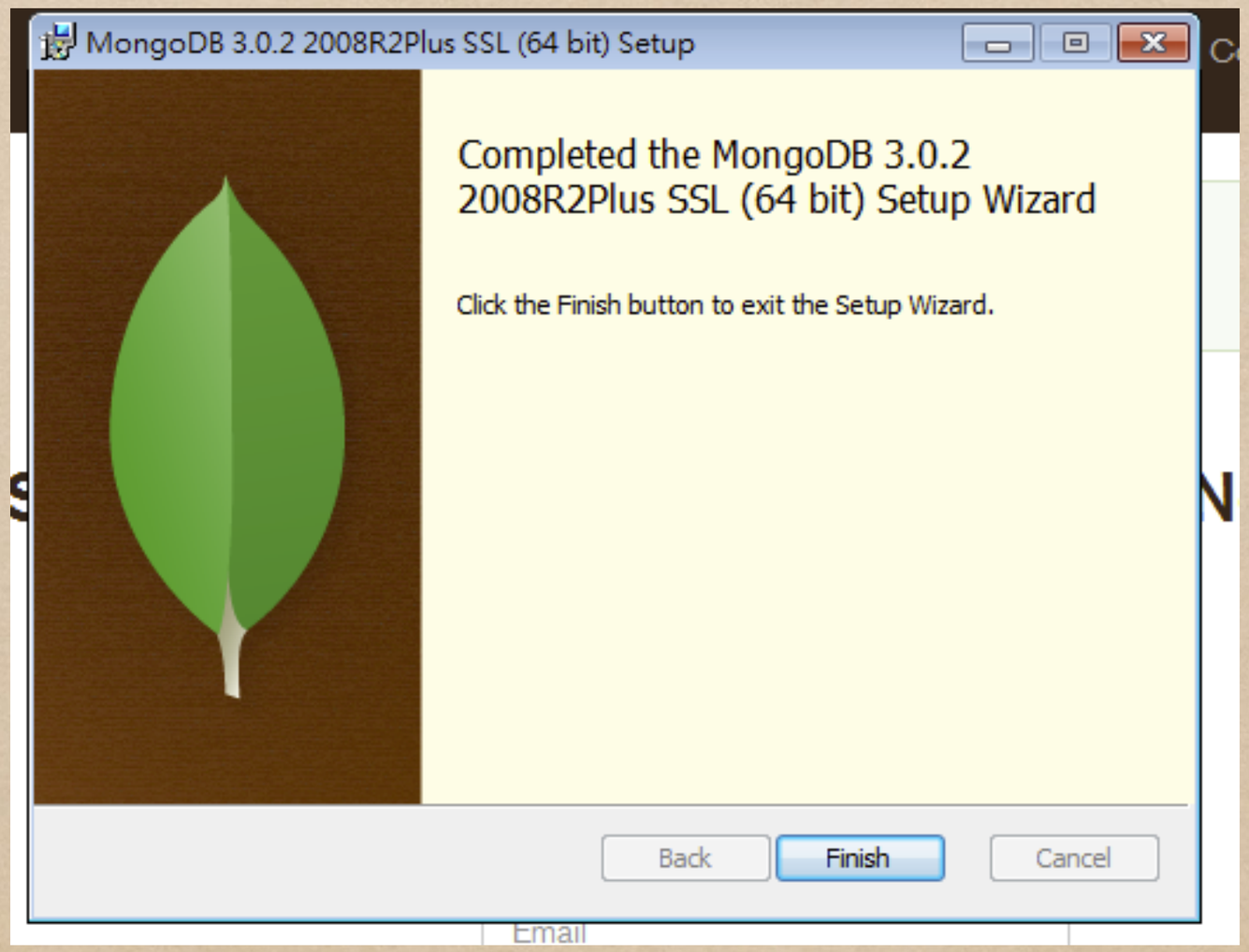
按install


MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup
Ready to install MongoDB 3.0.2 2008R2Plus SSL (64 bit)
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel

選「是」


使用者帳戶控制
您是否要允許下列程式在這部電腦上安裝軟體?
程式名稱: 29f0dbb.msi
已驗證的發行者: MongoDB, Inc.
檔案來源: 從網際網路下載
顯示詳細資料(D)
是(Y)
否(N)
變更這些通知顯示的時機

安裝完成


MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup
Completed the MongoDB 3.0.2 2008R2Plus SSL (64 bit) Setup Wizard
Click the Finish button to exit the Setup Wizard.
Back
Finish
Cancel

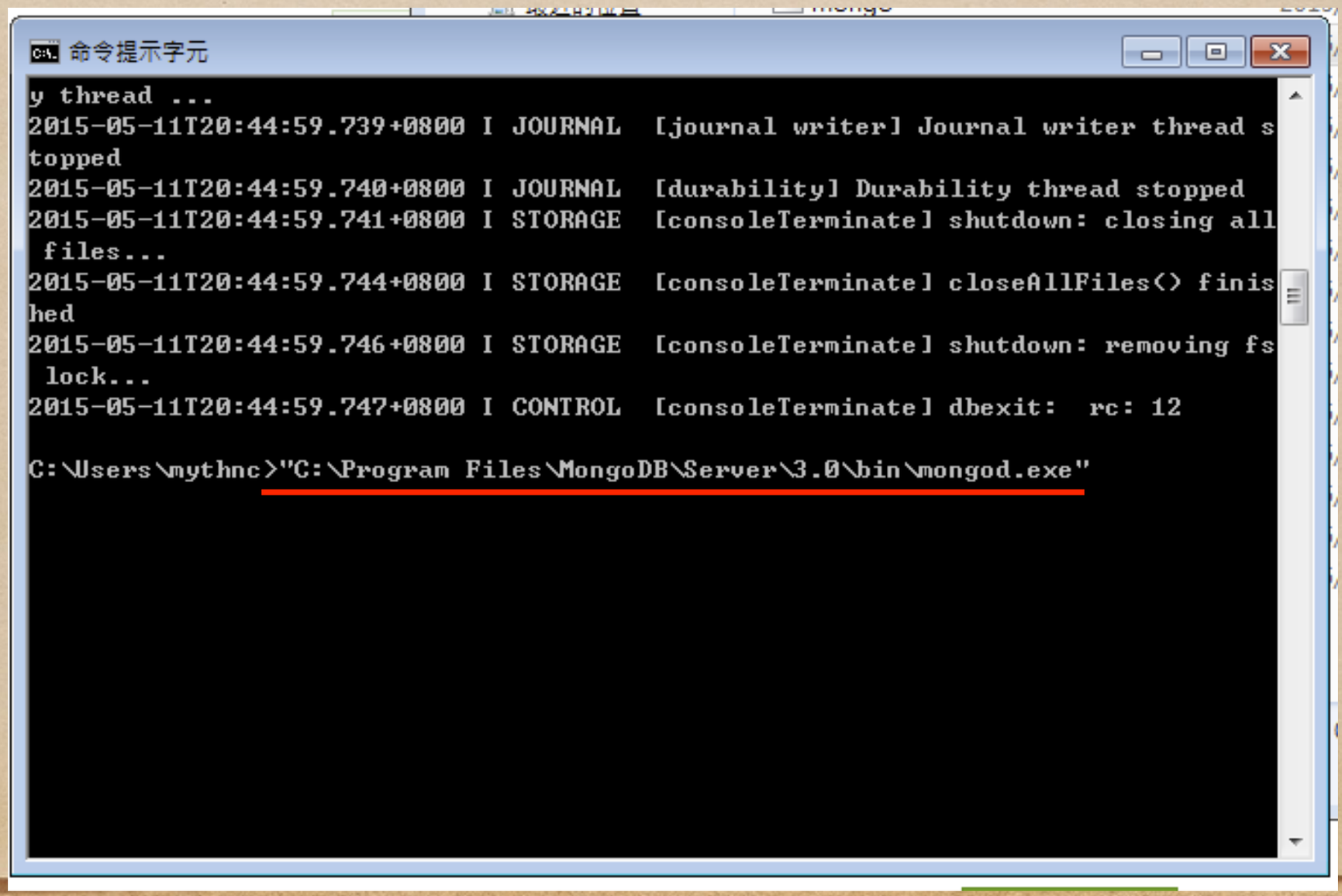
接著是測試

設置資料庫

- $ md \data\db

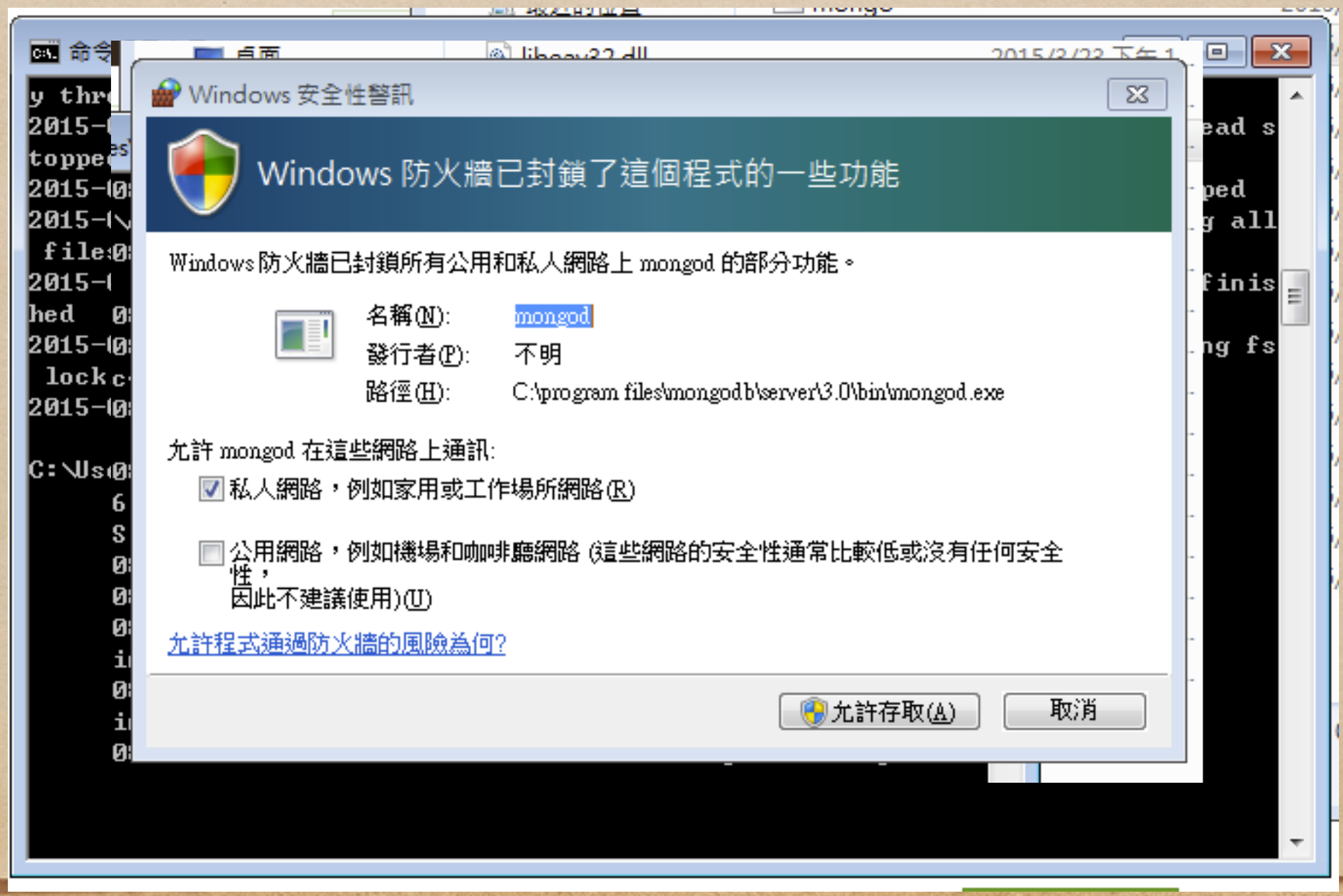
啟動mongod(一定要先開)

- 把mongod拖曳到cmd中，按enter



點「允許存取」

- 把mongod拖曳到cmd中，按enter

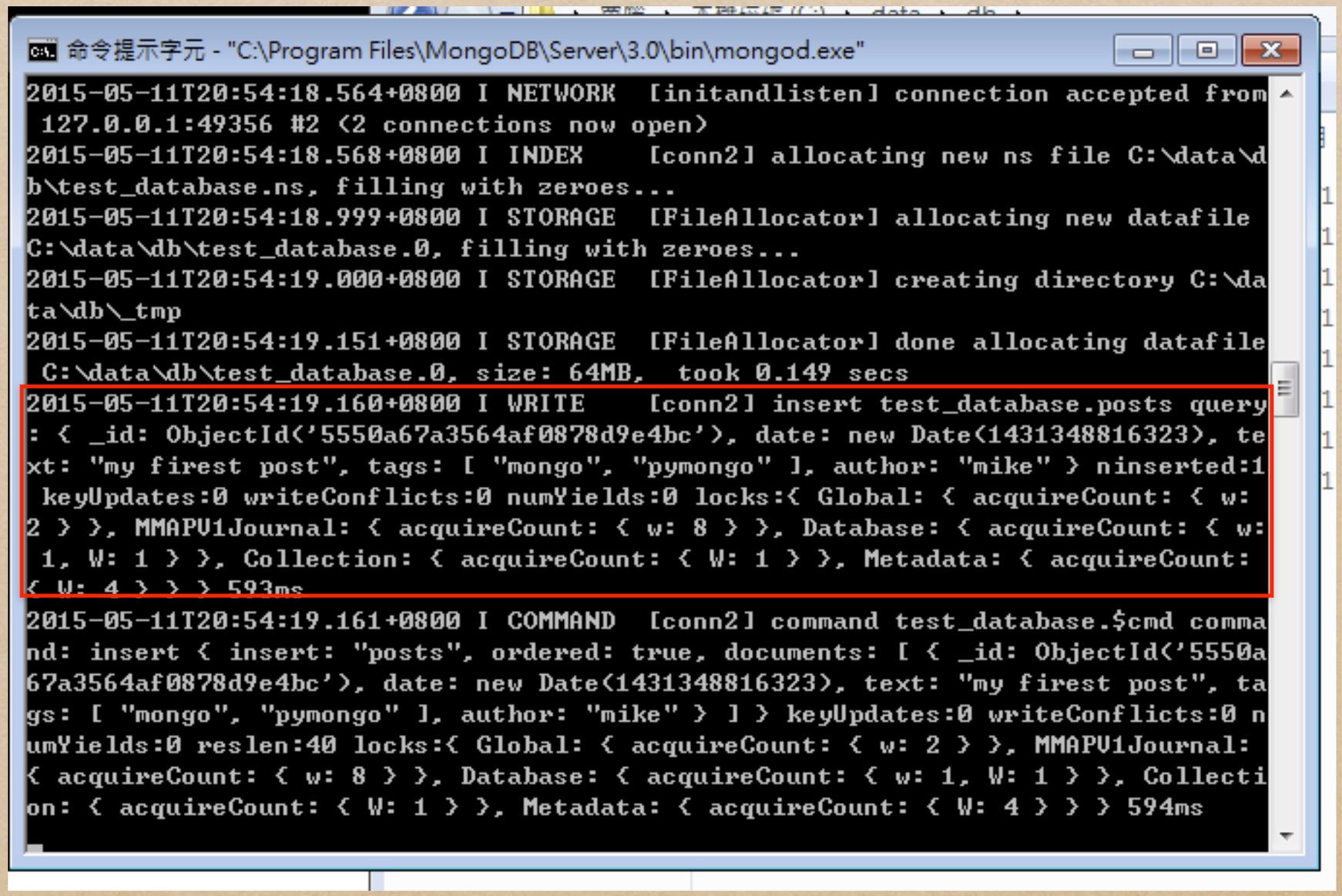


寫入資料至mongodb

- 參考教學（連結）
- mongod記得要先開

```
from pymongo import MongoClient
import datetime

client = MongoClient()
db = client.test_database
collection = db.test_collection

post = {"author": "Mike",
        "text": "My first blog post!",
        "tags": ["mongodb", "python", "pymongo"],
        "date": datetime.datetime.utcnow()}

posts = db.posts
post_id = posts.insert_one(post).inserted_id
print post_id
```

寫入資訊


命令提示字元 - "C:\Program Files\MongoDB\Server\3.0\bin\mongod.exe"
2015-05-11T20:54:18.564+0800 I NETWORK [initandlisten] connection accepted from 127.0.0.1:49356 #2 (2 connections now open)
2015-05-11T20:54:18.568+0800 I INDEX [conn2] allocating new ns file C:\data\db\test_database.ns, filling with zeroes...
2015-05-11T20:54:18.999+0800 I STORAGE [FileAllocator] allocating new datafile C:\data\db\test_database.0, filling with zeroes...
2015-05-11T20:54:19.000+0800 I STORAGE [FileAllocator] creating directory C:\data\db_tmp
2015-05-11T20:54:19.151+0800 I STORAGE [FileAllocator] done allocating datafile C:\data\db\test_database.0, size: 64MB, took 0.149 secs
2015-05-11T20:54:19.160+0800 I WRITE [conn2] insert test_database.posts query: { _id: ObjectId('5550a67a3564af0878d9e4bc'), date: new Date(1431348816323), text: "my firest post", tags: ["mongo", "pymongo"], author: "mike" } ninserted:1 keyUpdates:0 writeConflicts:0 numYields:0 locks:{ Global: { acquireCount: { w: 2 } }, MMAPV1Journal: { acquireCount: { w: 8 } }, Database: { acquireCount: { w: 1, W: 1 } }, Collection: { acquireCount: { W: 1 } }, Metadata: { acquireCount: { W: 4 } } } 593ms
2015-05-11T20:54:19.161+0800 I COMMAND [conn2] command test_database.$cmd command: insert { insert: "posts", ordered: true, documents: [{ _id: ObjectId('5550a67a3564af0878d9e4bc'), date: new Date(1431348816323), text: "my firest post", tags: ["mongo", "pymongo"], author: "mike" }] } keyUpdates:0 writeConflicts:0 numYields:0 reslen:40 locks:{ Global: { acquireCount: { w: 2 } }, MMAPV1Journal: { acquireCount: { w: 8 } }, Database: { acquireCount: { w: 1, W: 1 } }, Collection: { acquireCount: { W: 1 } }, Metadata: { acquireCount: { W: 4 } } } 594ms

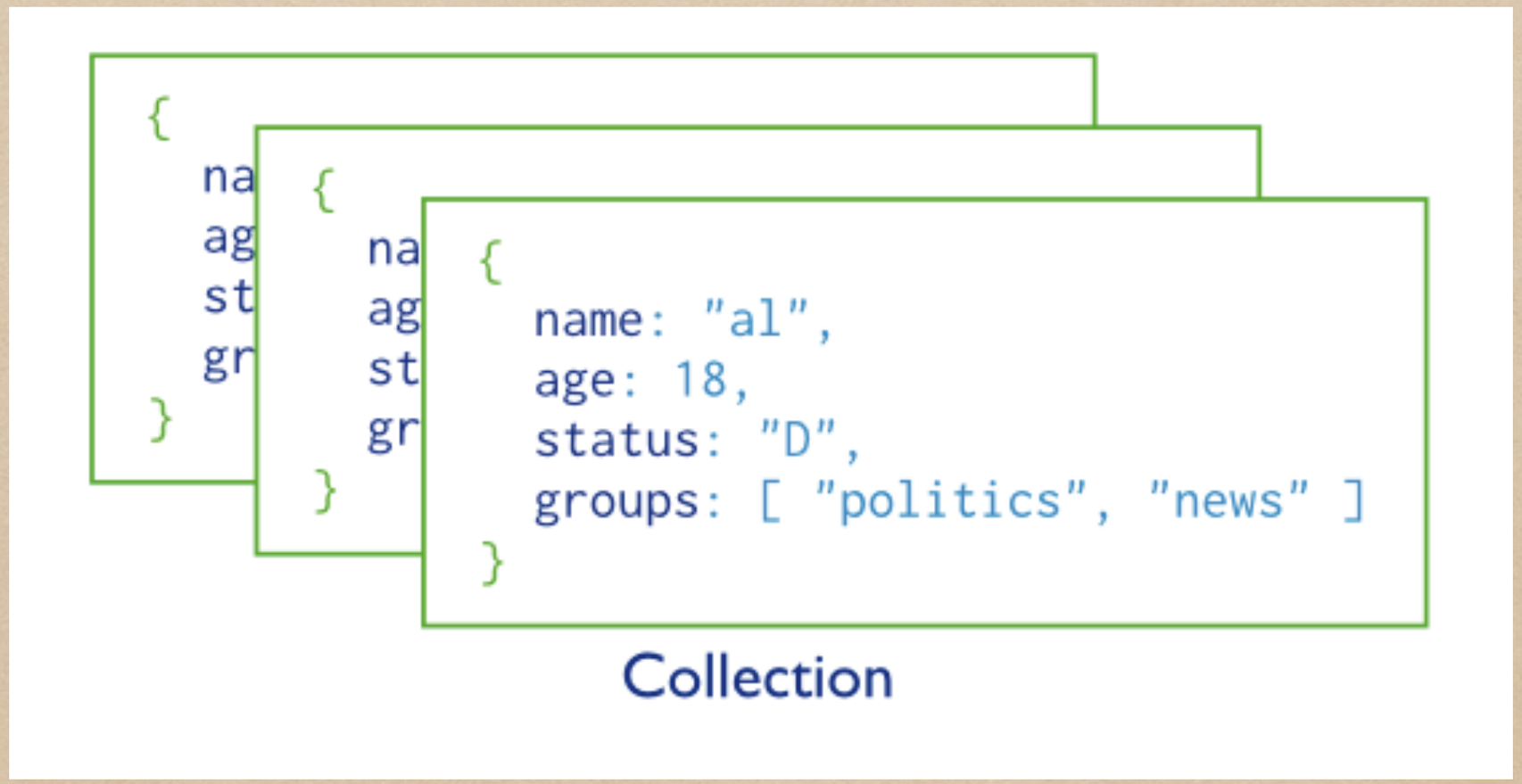
簡易解說

- MongoDB is No SQL Database
- collections / databases in MongoDB is that they are created lazily
- Database - Collection - Document (NoSQL)
  Database - Table - Row/Record (RDBMS)

Document

- documents are similar to JSON objects

Collection


{
na
ag
st
gr
}
{
na
ag
st
gr
}
{
name: "al",
age: 18,
status: "D",
groups: ["politics", "news"]
}
Collection

Query

- for post in posts.find():
    post
- 結果會以dict形式顯示，每個post是一筆結果
- 搜尋條件放在find()中
- for post in posts.find({'author': 'Mike'}):
    post

Index

- To make query faster

```
from pymongo import MongoClient
import datetime

client = MongoClient()
db = client.test_database
collection = db.test_collection

post = {"author": "Mike",
        "text": "My first blog post!",
        "tags": ["mongodb", "python", "pymongo"],
        "date": datetime.datetime.utcnow()}

posts = db.posts
post_id = posts.insert_one(post).inserted_id
print post_id
```

```
from pymongo import ASCENDING, DESCENDING
posts.create_index([("date", DESCENDING), ("author", ASCENDING)])
```

Please check the crawler_tutorial.ipynb file

API

- 爬蟲之餘的選擇
- facebook API, twitter API, plurk API, flickr API

Reference

- 爬蟲相關
  - Web Crawler教學 (c3h3)
  - Web Cralwer工具箱 (joe)
  - kimono (turn websites into APIs)
- MongoDB相關
  - MongoDB Tutorial (CodeData)
  - mongoDB doc
- Library相關
  - PyMongo Tutorial
  - Requests
  - pyquery

Question?