Organizing Your Code

Demitri Muna
OSU

9 July 2013

Common Code Pattern

A common pattern in people's code goes something like this:

```
[set some parameters for the script]
[read "datafile.dat"]
[loop over the data file]
[read into some structure]
[analyze code for x]
[analyze code for y]
[write results to output file]
```

Then another script...

```
[set other parameters for the script]
[read "datafile.dat"]
[loop over the data file]
[read into some structure]
[analyze code for x]
[analyze code for z]
[write results to output file]
```

- Much of code becomes a template
- Code is repeated
- Changes/fixes to one script need to be made to the others
- Code can't be reused in new scripts or by others
- Scripts are essentially one-offs
- Analysis code mixed in with bookkeeping code

Example: Analyzing SDSS Data

```
[go to data directory]
[loop over plates (one per directory)]
   [open FITS file]
   [read header]
   [read table]
   [select spectra that match “galaxy” and “0.1 < z < 0.2”]
   [“bookmark” that spectrum]
[loop over found spectra]
   [open FITS file]
   [read spectrum from correct HDU]
   [read ra/dec from correct header]
   [analyze spectrum]
   [write results out to file]
[generate plots]
```

mostly bookkeeping, would be duplicated for other scripts

parameters buried in script

code depends on file format – if this changes, many scripts need to be updated

Aims of Code

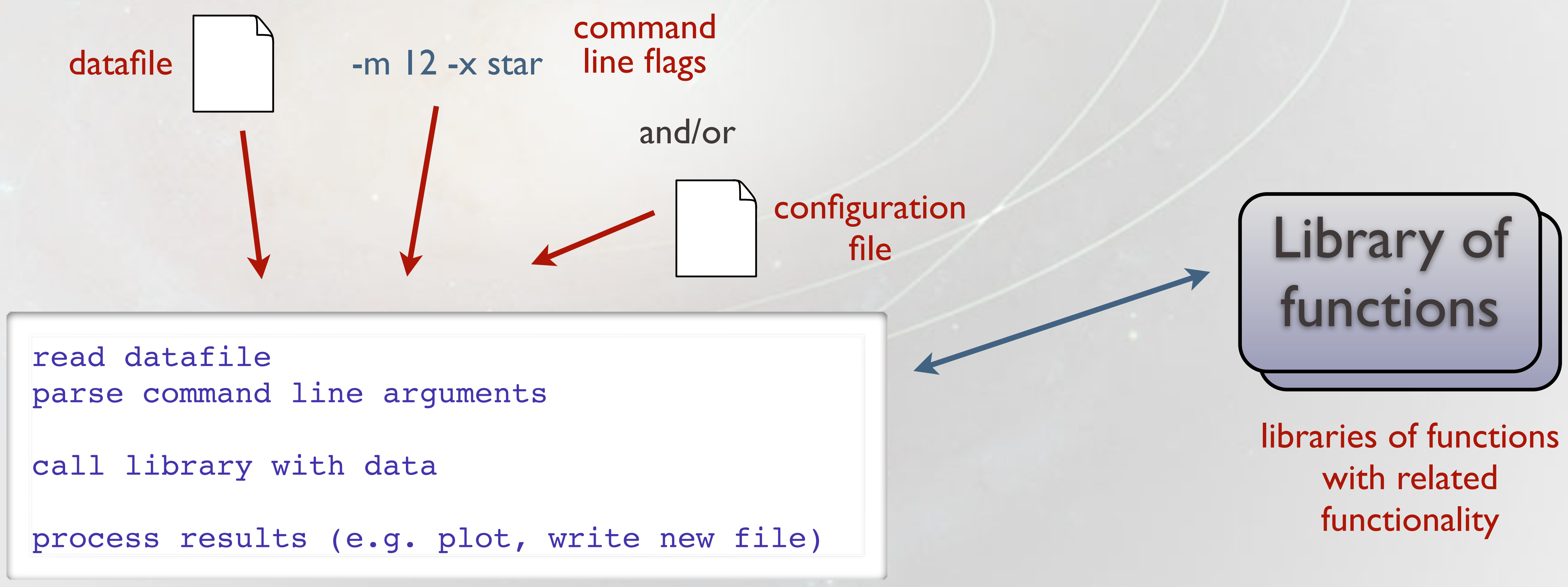
- Functions that do a particular job should be written once and used by other pieces of code.
- Data files should be separate from functionality.
- Input parameters should be separate from functionality.

Unix Model of Code

Unix has a particular design philosophy for code.

- All functionality is written in a library.
- The library can be called by programs or other libraries
- Programs call the library to perform tasks.
- Programs are "thin" – they:
    - read files
    - set parameters
    - produce output

Unix Model of Code



```
read datafile
parse command line arguments

call library with data

process results (e.g. plot, write new file)
```

libraries of functions with related functionality

Calling method examples:

```
my_script -m 12 -x star datafile

my_script configfile.cfg
```

sample config file

```
datafile = mydata.fits
m = 12
iterations = 100
type = star
```

Example Revisited: Analyzing SDSS Data

```
import SDSSData

[argparse config file]

spectra = SDSSData.SpectrumSearch(z=[zmin, zmax],
                                  kind='galaxy',
                                  mag=[man_min, mag_max])
for spectrum in spectra:
ra = spectrum.ra
dec = spectrum.dec
sigma = spectrum.standard_deviation
x = SDSSData.find_lyman_alpha_break(spectrum)
```

all analysis in SDSSData module

parameters read from configuration file

Extracting the "ra" requires detailed knowledge of the file, but it's basically bookkeeping (open FITS file, read HDU x, extract header with keyword "y", convert to float, etc.).

Create an object for each "thing" you work with – a spectrum, a data file, etc. Put functions into those objects that do the bookkeeping, hiding it from your analysis code.

Application Programming Interface (API)

- An application programming interface defines functions for certain tasks or data.
- The implementation is hidden in the library.
- Changes to how the task is performed can be implemented, but the API stays the same.

“ra” is an API call

```
from SDSSData import Spectrum

s = Spectrum(file="spectrum_file.fits")
print s.ra
```

The details of how ra is looked up don’t matter – they are in the library.
If the file format changes, change the library. Any code that uses the API still works.

Our Own Python Modules

```
from SDSSData import Spectrum

s = Spectrum(file="spectrum_file.fits")
print s.ra
```

"SDSSData" is a module we write

directory structure for our module

```
SDSSData/
    __init__.py
    Galaxy.py
    Spectrum.py
    Star.py
    Telescope.py
    utilities.py
    errors.py
    tests/
        sample_spectrum.fits
        test_spectrum.py
```

- If a file called `__init__.py` is present in a folder, Python will see the folder as a module.
- The file can be empty, or can contain initialization for the module.
- `import SDSSData` executes `__init__.py`.

Example Spectrum.py File

"." says to look for "errors.py" in this directory.
Our custom errors are defined there.

```python
from .errors import SDSSFileNotSpecified

class Spectrum(object):

    def __init__(self, file=None):
        if file == None:
            raise SDSSFileNotSpecified("A spectrum file must "
                                       "be specified to create a spectrum.")
        self.filename = file
        self.datafile = open(datafile)

    def ra(self):
        ''' Returns the RA of this spectrum in degrees. '''
        # open the FITS file
        # read the right HDU
        return hdu.header["ra"]
```

A Note on Imports

```
import math
print math.pi

# 3.14159265359
```

“pi” is defined in the “math” module. Access it by specifying the module, then the value (or function).

“pi” is not defined by calling import alone

```
import math
print pi

# Traceback (most recent call last):
#   File "untitled text 54", line 2, in
<module>
#     print pi
# NameError: name 'pi' is not defined
```

```
from math import *

d = 89
e = 20
f = 297
# ... lots of code
print e**2
# 400, not 2.71828182846^2
```

bring everything in math into our namespace

overwrites the ‘e’ variable from math

The *namespace* is the context where variables are defined. Your script has a namespace. Each module has an independent namespace.

```
import math

pi = 3 # exactly 3
print pi
print math.pi

# 3
# 3.14159265359
```

our namespace

the “math” module namespace

```
from math import pi

print pi
# 3.14159265359
```

bring “pi” into our namespace - no “math.” prefix needed

“`import *` ” is bad form and can easily lead to errors. Don’t use it unless you really know what you’re doing.

Testing Your Code

- How do we know our code works?
- We check the code as we're writing it.
- As code base grows, how do we know all of it is being run?
- When we make changes, how do we make sure we haven't broken anything that worked before? (This is called *regression*.)
- Testing my code takes time – I'm trying to publish!

Unit Testing

- The `pytest` module provides a framework to make testing easy.
- Write small tests to check each piece of functionality, not a huge program to test everything.
- Run tests regularly as you write code to avoid regression.
- Best practice: write the tests *before* you write your code, then write your code until the tests pass.

Sample Test File

known file kept in testing directory to check against

```python
import numpy as np
from ..Spectrum import Spectrum

sample_spectrum_filename = "sample_spectrum.fits"

def test_spectrum_read():
    ''' Check that we can read a spectrum file. '''
    s = Spectrum("sample_spectrum_filename")
    assert len(s.hdu_list) == 3, "Unexpected no of HDUs found in file."

def test_ra():
    ''' Check that we can read the RA from an SDSS file. '''
    s = Spectrum("sample_spectrum_filename")
    assert numpy.testing.assert_approx_equal(s.ra, 12.3432)
```

Use multiple asserts to check any functionality to be tested. Use them liberally!

The numpy testing.assert_approx_equal method is useful to check floating point values.

Running Tests

On the command line, go to the testing directory in your module and run "py.test" (installed when you install the pytest module).

```
% py.test
============================== test session starts ==============================
platform darwin -- Python 2.7.5 -- pytest-2.3.4
collected 14 items

test_spectrum.py ..
test_galaxy.py ..FF
```

two tests found, two tests passed

four tests found; two passed, two failed

When tests fail, py.test produces output to tell you which ones failed and where.

Write your tests first and keep writing code until they all pass. Run tests regularly to make sure nothing you do breaks existing code.