

# SCIENTIFIC COMPUTING IN PYTHON

KONRAD HINSEN

CENTRE DE BIOPHYSIQUE MOLÉCULAIRE (ORLÉANS)

AND

SYNCHROTRON SOLEIL (ST AUBIN)



# ARRAY COMPUTING

## NUMPY



# NUMPY

Basic functionality for scientific computing:

- ▷ Multidimensional arrays
  - ▷ Arithmetic and mathematical functions on arrays
  - ▷ Linear algebra (LAPACK)
  - ▷ Fourier transforms (FFTPACK)
  - ▷ Random numbers
- ➡ Efficient implementation that makes handling large data possible using pure Python code.

Documentation: <http://numpy.scipy.org/>



# ARRAYS

- ▷ multidimensional rectangular data container
- ▷ all elements have the same type
- ▷ compact data layout, compatible with C/Fortran
- ▷ efficient operations
- ▷ arithmetic
- ▷ flexible indexing



# WHY ARRAYS?

Arrays are the most “natural” data structure for many types of scientific data:

- ▷ Matrices
- ▷ Time series
- ▷ Images
- ▷ Functions sampled on a grid
- ▷ Tables of data
- ▷ ... many more ...

Python lists can handle this, right?



# WHY ARRAYS?

Python lists are nice, but...

- ▷ They are slow to process
- ▷ They use a lot of memory
- ▷ For tables, matrices, or volumetric data, you need lists of lists of lists... which becomes messy to program.

```
from random import random
from operator import add
import numpy as np
n = 1000000
l1 = [random() for i in range(n)]
l2 = [random() for i in range(n)]
a1 = np.array(l1)
a2 = np.array(l2)

%timeit l3 = map(add, l1, l2)
10 loops, best of 3: 147 ms per loop
```

```
%timeit a3 = a1+a2
100 loops, best of 3: 8 ms per loop
```

Bytes per element in a list of floats: 32

Bytes per element in an array of floats: 8



# CHARACTER ARRAYS

Character arrays vs. strings:

- ▷ More efficient slicing
- ▷ Can be modified in-place
- ▷ Can be memory-mapped to files (but see also module [mmap](#))
- ▷ More efficient for many operations, in particular comparisons
- ▷ Downside: arrays lack many useful string operations

I will use number arrays in all examples, but remember that characters are just numbers in the range 0-255.



**NEVER FORGET:**

```
import numpy as np
```

(I won't repeat this on every slide!)



# ARRAY CREATION

▷ `np.array([ [1, 2], np.array([3, 4]) ])`  
`array([[1, 2],`  
 `[3, 4]])`

▷ `np.zeros((2, 3), dtype=np.float)`  
`array([[ 0., 0., 0.],`  
 `[ 0., 0., 0.]])`

▷ `np.arange(0, 10, 2)`  
`array([0, 2, 4, 6, 8])`

▷ `np.arange(0., 0.5, 0.1)`  
`array([ 0. , 0.1, 0.2, 0.3, 0.4])`

Optional dtype=...  
everywhere:

`dtype=np.int`

`dtype=np.int16`

`dtype=np.float32`

...

Watch out for round-off problems!

You may prefer `0.1*np.arange(5)`



# ARRAY CREATION

▷ `np.linspace(0., 1., 6)`

```
array([ 0. ,  0.2,  0.4,  0.6,  0.8,  1. ])
```

▷ `np.eye(3)`

```
array([[ 1.,  0.,  0.],  
       [ 0.,  1.,  0.],  
       [ 0.,  0.,  1.]])
```

▷ `np.diag([1., 2., 3.])`

```
array([[ 1.,  0.,  0.],  
       [ 0.,  2.,  0.],  
       [ 0.,  0.,  3.]])
```



# INDEXING

```
a = np.arange(6)  
    array([0, 1, 2, 3, 4, 5])
```

▷ a[2]  
 2

▷ a[2:4]  
 array([2, 3])

▷ a[1:-1]  
 array([1, 2, 3, 4])

▷ a[:4]  
 array([0, 1, 2, 3])

▷ a[1:4:2]  
 array([1, 3])

▷ a[::-1]  
 array([5, 4, 3, 2, 1, 0])

This works  
exactly like for  
lists!



# INDEXING

```
a = np.array([ [1, 2], [3, 4] ])  
array([[1, 2],  
       [3, 4]])
```

▷ a[1, 0]  
3

▷ a[1, :]      a[1]  
array([3, 4])

▷ a[:, 1]  
array([2, 4])

▷ a[:, :, np.newaxis]  
array([[[1],  
 [2]],  
 [[3],  
 [4]]])



# ARITHMETIC

```
a = np.array([ [1, 2], [3, 4] ])    a.shape = (2, 2)
      array([[1, 2],
            [3, 4]])
```

▷ `a + a`  
`array([[2, 4],
 [6, 8]])`

▷ `a + 1`  
`array([[2, 3],
 [4, 5]])`

▷ `a + np.array([10, 20])`      `array([10, 20]).shape = (2,)`  
`array([[11, 22],
 [13, 24]])`

▷ `a + np.array([[10], [20]])`      `array([[10], [20]]).shape = (2, 1)`  
`array([[11, 12],
 [23, 24]])`



# BROADCASTING RULES

$c = a + b$  with  $a.shape == (2, 3, 1)$  and  $b.shape == (3, 2)$

1)  $\text{len}(a.shape) > \text{len}(b.shape)$

➡  $b \rightarrow b[\text{newaxis}, :, :]$ ,  $b.shape \rightarrow (1, 3, 2)$

2) Compare  $a.shape$  and  $b.shape$  element by element:

- $a.shape[i] == b.shape[i]$ : *easy*
- $a.shape[i] == 1$ : *repeat a  $b.shape[i]$  times*
- $b.shape[i] == 1$ : *repeat b  $a.shape[i]$  times*
- otherwise : **error**

3) Calculate the sum element by element

4)  $c.shape == (2, 3, 2)$



# STRUCTURAL OPERATIONS

```
a = (1 + np.arange(4))**2
```

```
array([ 1,  4,  9, 16])
```

▷ `np.take(a, [2, 2, 0, 1])`

```
array([9, 9, 1, 4])
```

or `a.take([2, 2, 0, 1])`

▷ `np.where(a >= 2, a, -1)`

```
array([-1,  4,  9, 16])
```

▷ `np.reshape(a, (2, 2))`

```
array([[ 1,  4],  
       [ 9, 16]])
```

or `a.reshape((2, 2))`

▷ `np.resize(a, (3, 5))`

```
array([[ 1,  4,  9, 16,  1],  
       [ 4,  9, 16,  1,  4],  
       [ 9, 16,  1,  4,  9]])
```

or `a.reshape((3, 5))`

▷ `np.repeat(a, [2, 0, 2, 1])`

```
array([ 1,  1,  9,  9, 16])
```

or `a.repeat([2, 0, 2, 1])`



# “FANCY” INDEXING

```
a = np.arange(6)**2  
array([0, 1, 4, 9, 16, 25])
```

▷ `a[a % 2 == 0]`      same as `a.repeat(a % 2 == 0)`  
`array([0, 4, 16])`

▷ `a[[3, 0, 2]]`      same as `a.take([3, 0, 2])`  
`array([9, 4])`

these are safer

Watch out:

▷ `a[np.array([True, False, False, True, False, True])]`  
`array([ 0, 9, 25])`

▷ `a[[True, False, False, True, False, True]]`  
`array([1, 0, 0, 1, 0, 1])`



# FUNCTIONS

arccos, arcsin, arctan, arctan2, ceil, cos, cosh, exp, fabs, floor, fmod, hypot, log, log10, sin, sinh, sqrt, tan, tanh

Constants :  $\pi$ , e

Three sources:

- ▷ Module math : only for real arguments
- ▷ Module cmath : real and complex
- ▷ Module numpy :  
real, complex, **arrays**, and more..

**Always use module numpy !**

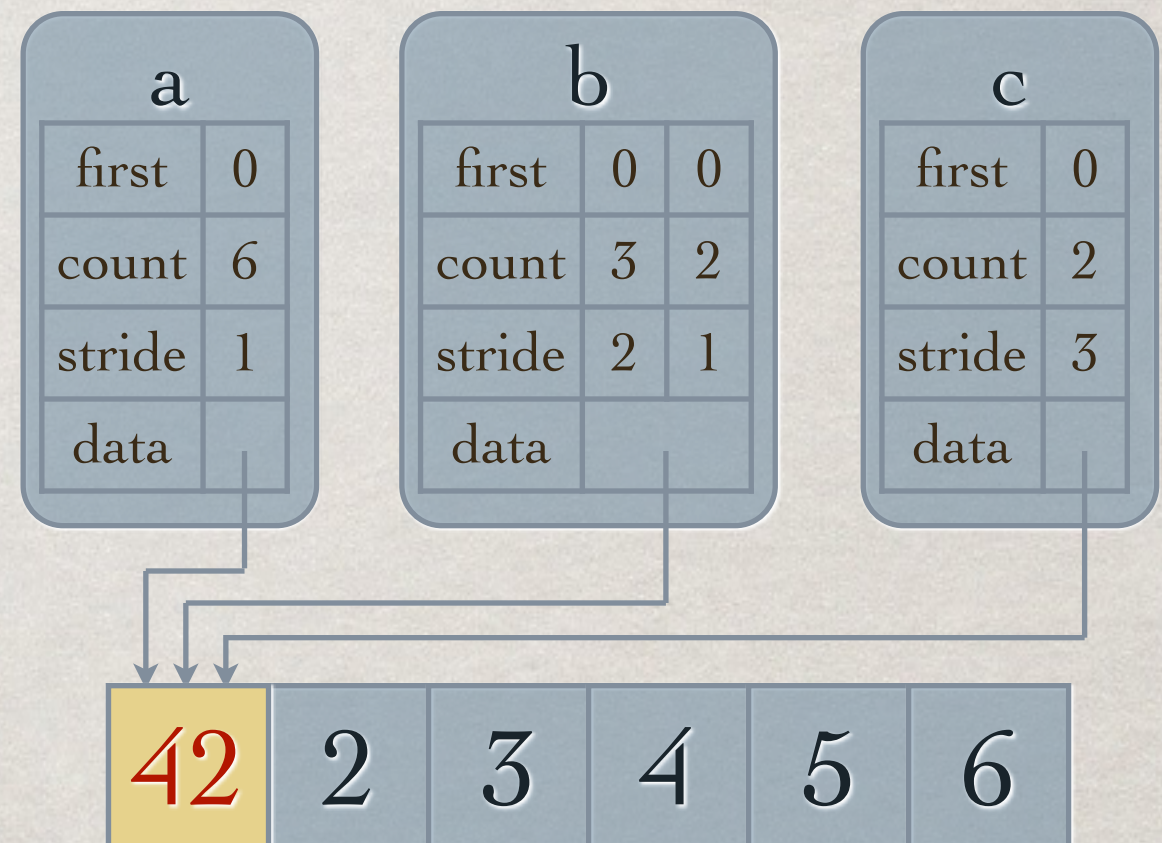


# ARRAY STRUCTURE

```
a = np.arange(1, 7)  
array([1, 2, 3, 4, 5, 6])
```

```
b = np.reshape(a, (3, 2))  
array([[1, 2],  
       [3, 4],  
       [5, 6]])
```

```
c = a[::3]  
array([1, 4])
```



Watch out:

```
a[0] = 42  
print c
```

```
array([42, 4])
```



# VIEWS

A view is a new array (i.e. a new Python object) that references that storage space of the array from which it was created.

If you modify array elements in the original array or in the view, they also change on the other side!

**The big question:** which operations return views, and which fresh arrays with independent storage areas?

**Rule of thumb:** An operation creates a view if this is possible for all its allowed arguments. Otherwise it returns a fresh array.

So... how do you find out if an array is a view on another arrays storage space? Check the attribute `base`.

```
a = np.arange(10)      assert a.base is None
b = a[::2]              assert b.base is a
```



# ARRAY PROGRAMMING

- ▷ Array operations are fast, Python loops are slow.  
(array operation = everything from module `numpy`)
- ▷ Top priority: avoid loops
- ▷ It's better to do the work three times with array operations than once with a loop.
- ▷ This does require a change of habits.
- ▷ This does require some experience.
- ▷ NumPy's array operations are designed to make this possible.

Get started with today's exercises!



# ARRAY PROGRAMMING STRATEGY

- ▷ Identify the kind of operation you want to do (applying a function, filtering, rearranging, ...)
- ▷ Go through the list of array operations and check if they do that kind of operation
- ▷ Use a mixture of thinking and trying out to get the job done.
- ▷ There is often more than one way to do it.

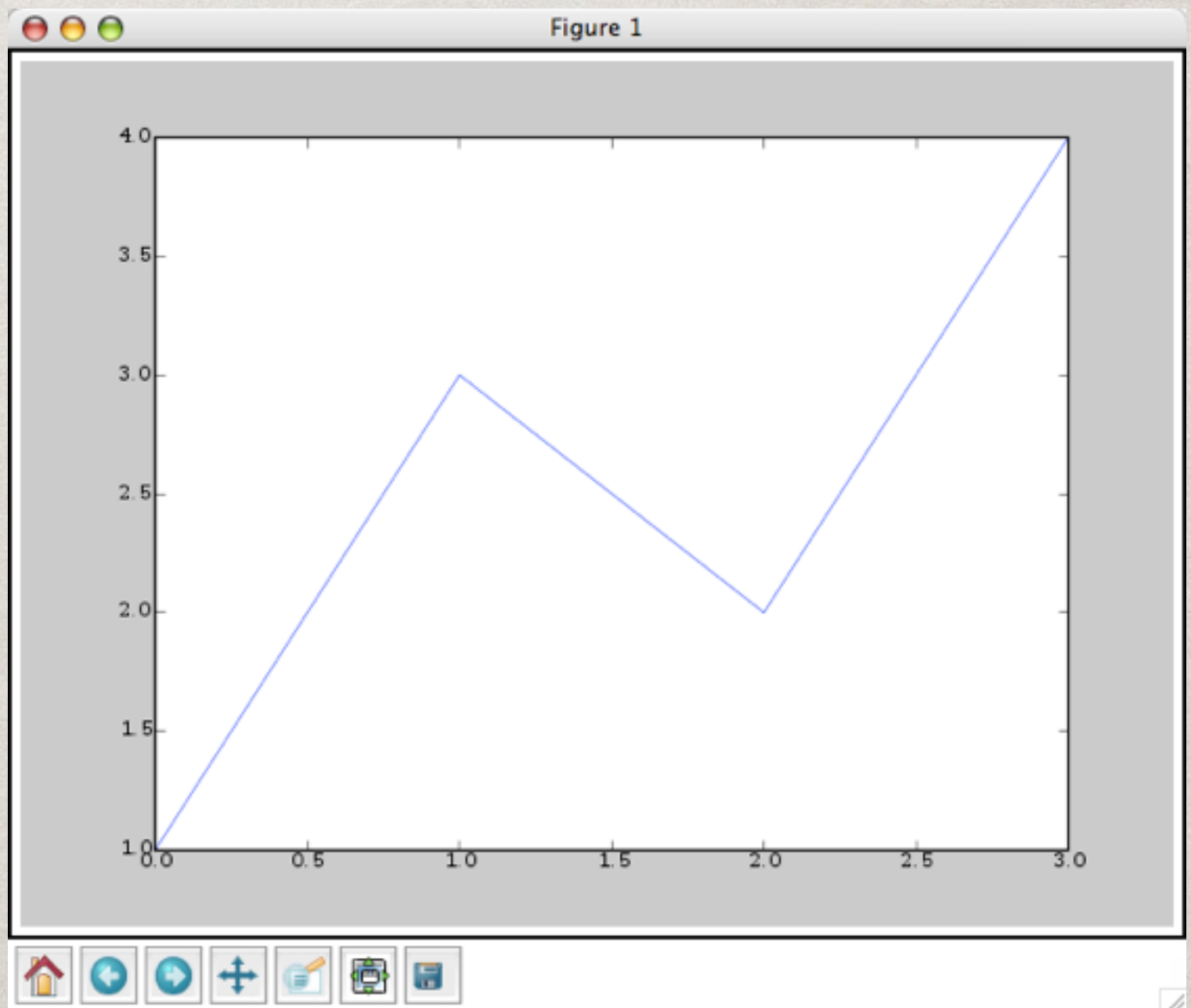


# PLOTTING: MATPLOTLIB



# BASIC PLOTTING

```
import matplotlib.pyplot as plt  
plt.plot([1,3,2,4])  
plt.show()
```





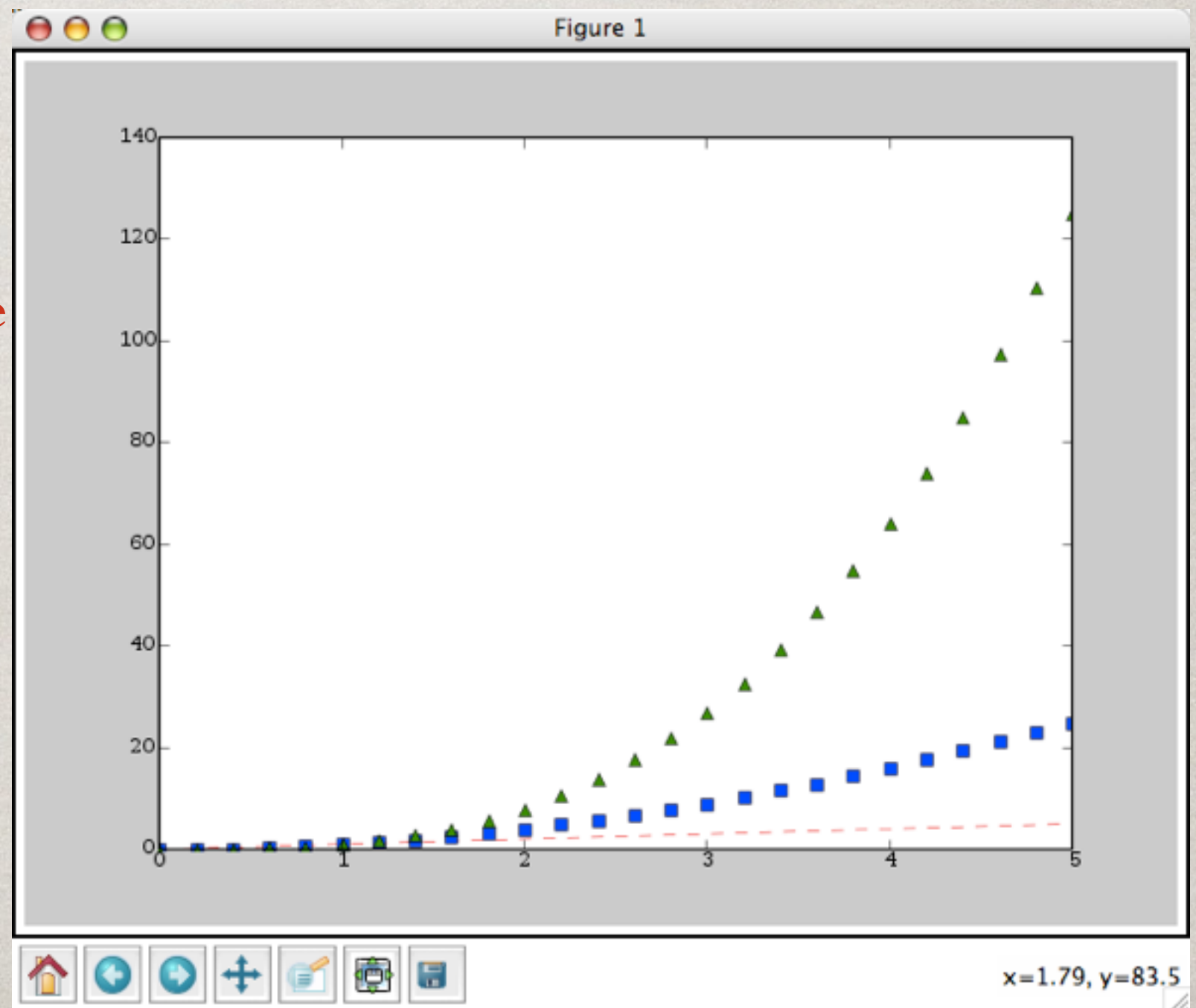
# LESS BASIC PLOTTING

```
import matplotlib.pyplot as plt
t = np.arange(0.0, 5.2, 0.2)
plt.plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^')
plt.show()
```

x y style:  
red  
dashed line

blue squares

green triangles

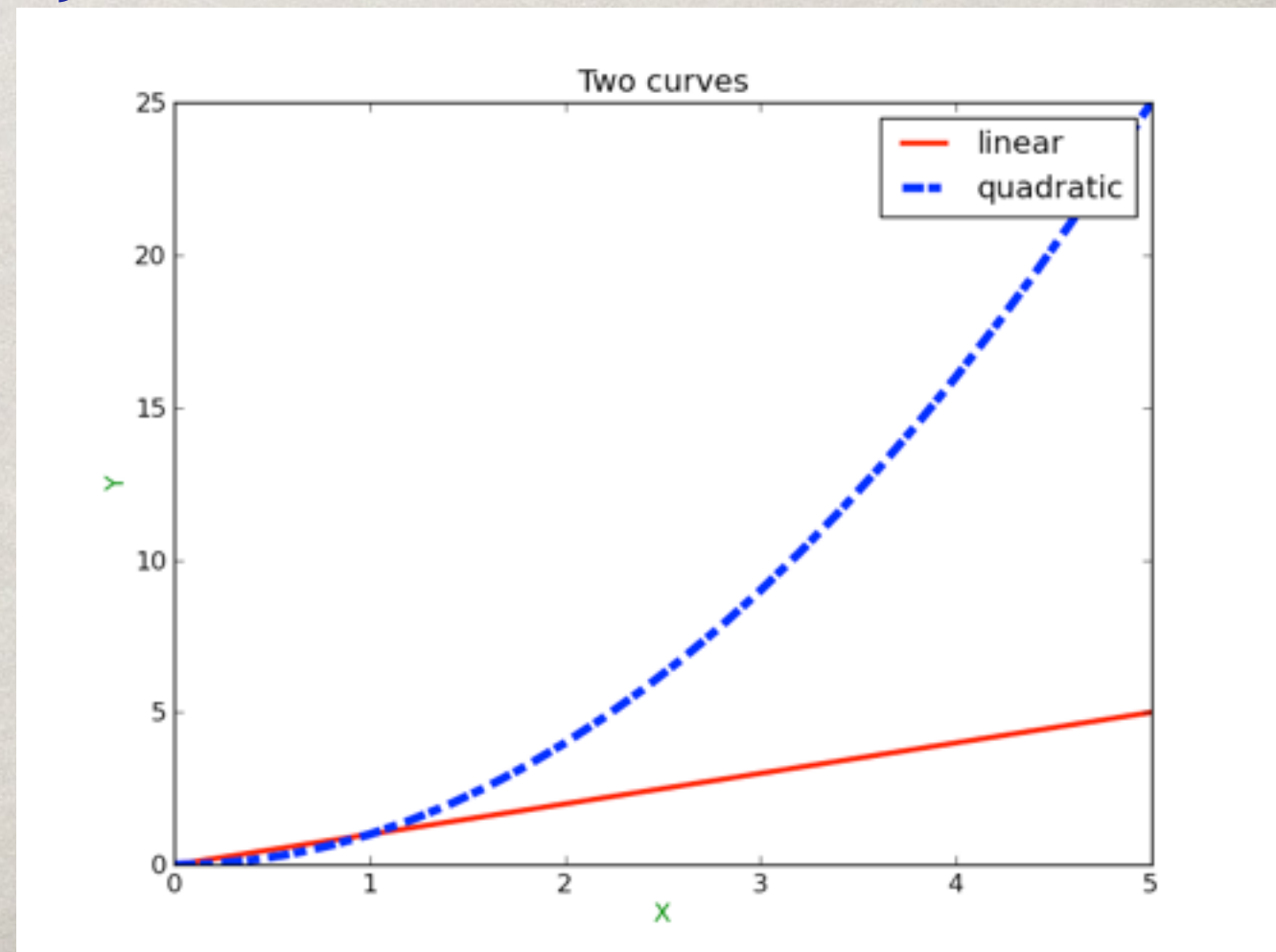




# LINE PROPERTIES AND LABELS

```
import matplotlib.pyplot as plt
import numpy as np
t = np.arange(0.0, 5.2, 0.2)
lines = plt.plot(t, t, 'r-', t, t**2, 'b--')
plt.setp(lines[0], linewidth=2.5,
          label="linear")
plt.setp(lines[1], linewidth=4.,
          dashes=[10, 2, 6, 4],
          label="quadratic")

plt.legend()
plt.title("Two curves")
plt.xlabel("X", color="green")
plt.ylabel("Y", color="green")
plt.show()
```





# MULTIPLE PLOTS

```
import matplotlib.pyplot as plt
import numpy as np
t = np.arange(-2.*np.pi, 2.*np.pi, 0.2)
```

```
fig = plt.figure()
```

```
axes_1 = fig.add_subplot(2, 2, 1)
axes_1.plot(t, np.sin(t))
```

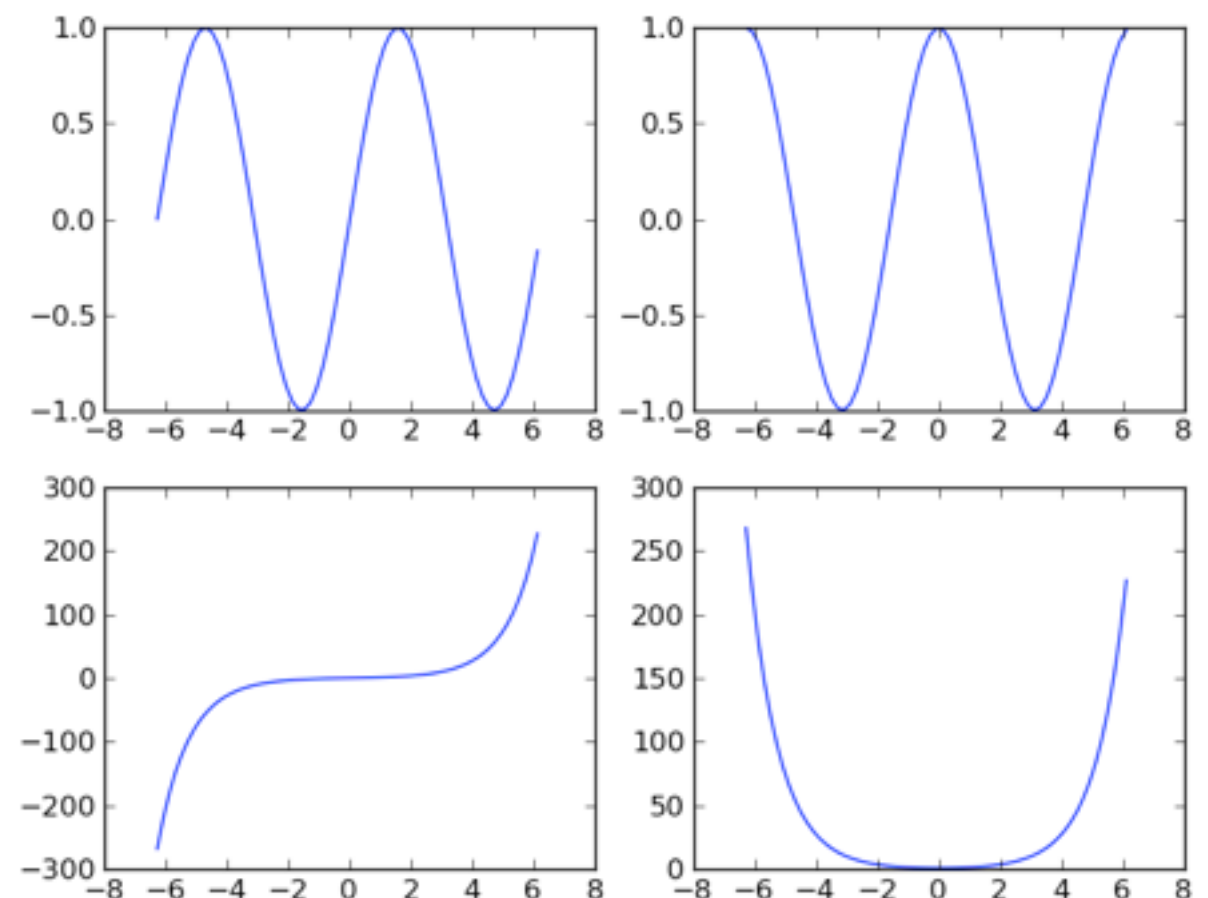
```
axes_2 = fig.add_subplot(2, 2, 2)
axes_2.plot(t, np.cos(t))
```

```
axes_3 = fig.add_subplot(2, 2, 3)
axes_3.plot(t, np.sinh(t))
```

```
axes_4 = fig.add_subplot(2, 2, 4)
axes_4.plot(t, np.cosh(t))
```

```
plt.show()
```

number of rows,  
number of columns,  
subplot number



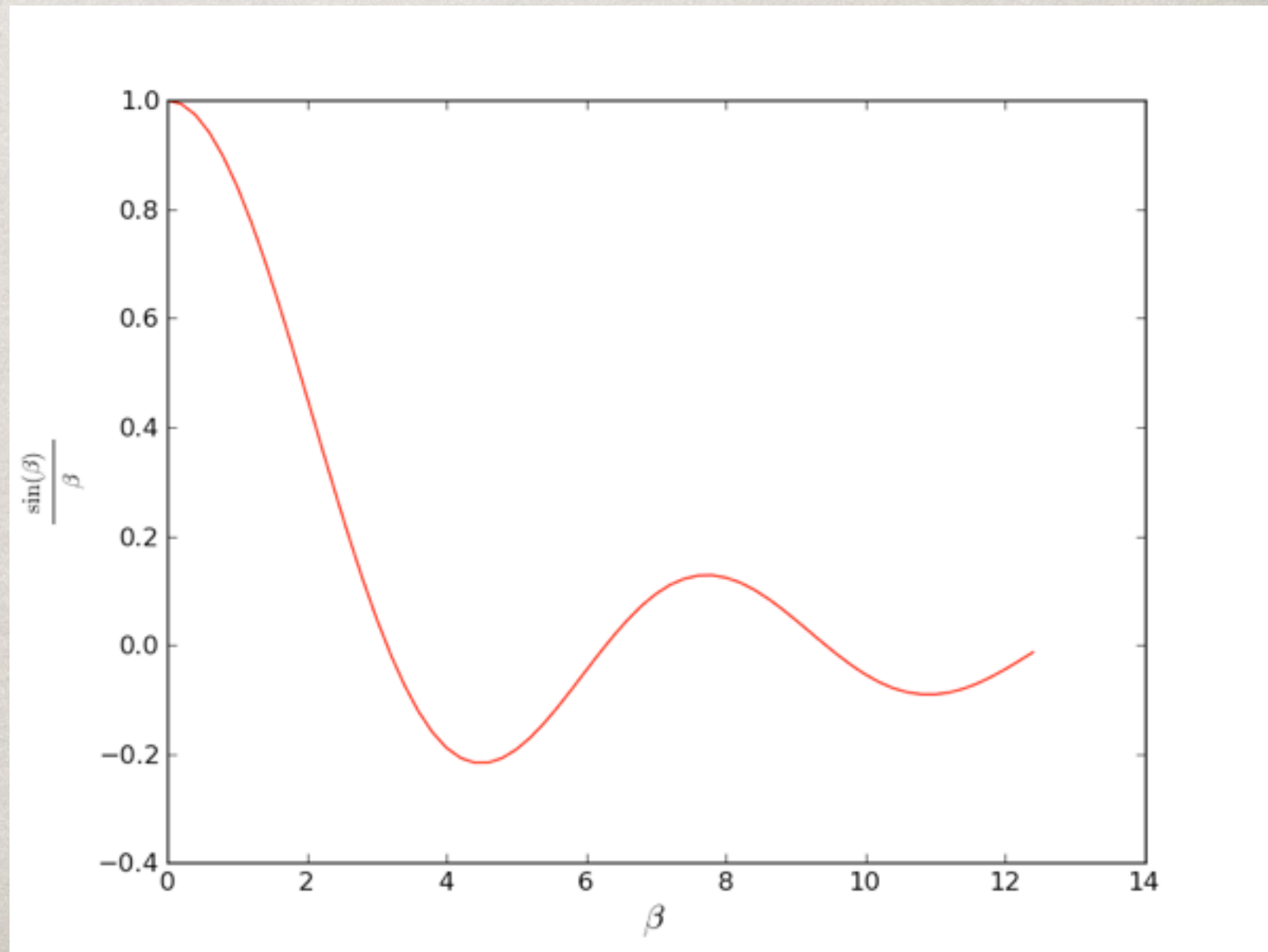


# FORMULAS IN TEXT

```
import matplotlib.pyplot as plt
import numpy as np
t = np.arange(0.0001, 4.*np.pi, 0.2)
plt.plot(t, np.sin(t)/t, 'r-')
plt.xlabel(r"$\beta$", fontsize=16)
plt.ylabel(r"$\frac{\sin(\beta)}{\beta}$", fontsize=16)
plt.show()
```

“raw” string:  
no interpretation  
of special codes

formula in  
TeX notation





# OTHER PLOT TYPES

- ✻ Bar charts
- ✻ Pie charts
- ✻ Polar plots
- ✻ 3D plots
- ✻ ...

Matplotlib has lots of options, lots of functions, lots of... everything.

Tutorial:

<http://www.loria.fr/~rougier/teaching/matplotlib/>

Lots of examples:

<http://matplotlib.org/gallery.html>



**MORE**



# SCIPY

Interfaces to many scientific libraries (Fortran/C):

- ▷ Statistics
- ▷ Optimization
- ▷ Numerical integration
- ▷ Linear algebra
- ▷ Fourier transforms
- ▷ Signal processing
- ▷ Image processing
- ▷ Genetic algorithms
- ▷ Partial differential equations
- ▷ Special functions

<http://scipy.org/scipylib/>



# SCIENTIFICPYTHON

## Python modules for scientific computing:

- ▷ **Geometry**  
vectors, tensors, linear transformations
- ▷ **Functions**  
interpolation, automatic derivatives, polynomials, rational functions
- ▷ **Statistics**  
moments of a distribution, histograms
- ▷ **Fitting**  
linear and non-linear least-squares
- ▷ **Physical units**  
conversion, arithmetic
- ▷ **Visualization**  
VRML, VPython, VMD
- ▷ **Parallelism**  
distributed computing, BSP, MPI interface
- ▷ **Input/Output**  
Fortran-compatible formatting, arrays, netCDF, PDB

<http://dirac.cnrs-orleans.fr/ScientificPython/>



# PANDAS

Handles data from observations and experiments

- ▷ Data frame objects
- ▷ I/O for CSV, Excel, SQL databases, HDF5
- ▷ Good support for missing data items
- ▷ Select and combine data from different sets
- ▷ Efficient for large data sets

<http://pandas.pydata.org/>



# OTHER

**BioPython:** bioinformatics

<http://biopython.org/>

**scikit-learn:** machine-learning and statistics

<http://scikit-learn.org/>

**statsmodels:** statistical analysis

<http://statsmodels.sourceforge.net/>

**MMTK:** (bio)molecular simulation

<http://dirac.cnrs-orleans.fr/MMTK/>



# EXERCISES

REMEMBER: NO LOOPS!



# ARRAY CREATION

Create these two arrays::

```
[[0. 0. 0. 0. 0.]  
 [2. 0. 0. 0. 0.]  
 [0. 3. 0. 0. 0.]  
 [0. 0. 4. 0. 0.]  
 [0. 0. 0. 5. 0.]  
 [0. 0. 0. 0. 6.]]
```

```
[[ 1  1  1  1]  
 [ 1  1  1  1]  
 [ 1  1  1  2]  
 [ 1  6  1  1]]
```



# POSITIVE ELEMENTS OF AN ARRAY

Write a function that takes a one-dimensional array argument and returns another one-dimensional array containing the positive elements of the input array.

An example of how your function should behave:

```
import numpy as np
x = np.arange(10)-5
print x
pos_x = positive_elements(x)
print pos_x
```

prints

```
[-5 -4 -3 -2 -1  0  1  2  3  4]
[1  2  3  4]
```



# MULTIPLICATION TABLE

Write a function that takes two one-dimensional array arguments and returns a two-dimensional array containing the products of each element of the first input array with each element of the second input array.

An example of how your function should behave:

```
import numpy as np
a = np.arange(3)
b = np.array([-1., 1., 2.])
print multiplication_table(a, b)
```

prints

```
[[ -0.   0.   0.]
 [ -1.   1.   2.]
 [ -2.   2.   4.]
```

Hint: have another look at the indexing options, in particular `numpy.newaxis`!



# DIFFERENCE ARRAYS

Write a function that takes a one-dimensional array argument and returns another one-dimensional array containing the differences between neighbouring points in the input array

An example of how your function should behave:

```
import numpy as np
x = np.array([1., 2., -3., 0.])
print differences(x)
```

prints

```
[1.  -5.  3.]
```

Hint: the simplest solution uses little more than clever indexing.



# REPEATING ARRAY ELEMENTS

Write a function that takes a two-dimensional array argument and returns another two-dimensional array of twice the size of the input array along each dimension. Each element of the input array is copied to four adjacent elements of the output array.

An example of how your function should behave:

```
import numpy as np
a = np.array([[1, 2], [3, 4]])
print repeat_twice(a)
```

prints

```
[[1 1 2 2]
 [1 1 2 2]
 [3 3 4 4]
 [3 3 4 4]]
```



# FITTING POLYNOMIALS

Write a function that fits a set of data points  $(x, y)$  to a polynomial of a given order  $N$ ,

$$P_N(x) = \sum_{i=0}^N a_i x^i$$

and returns the fitted coefficients  $a_i$ .

Don't forget error checking: the number of data points must be greater than the number of polynomial coefficients!

Hint: Write the fitting problem as a linear least-squares fit problem of the form

$$\min_{a_j} \sum_{j=0}^N (M_{ij} a_j - y_i)^2$$

where the elements of  $M_{ij}$  are powers of the  $x_i$ . Use `numpy.linalg.lstsq` to solve this least-squares problem.



# PLOTTING FUNCTIONS

Write a program that produces a plot of the exponential function in the interval  $[0 .. 3]$ .

Write a program that produces a plot of the sine and cosine functions in the interval  $[0 .. 2\pi]$ . Both functions should be in the same plot.

Write a program that produces a plot of the sine and cosine functions in the interval  $[0 .. 2\pi]$ . Put each function in its own sub-plot.