Numpy

Thomas Schwarz

Goals

- Overview of the Numpy and Pandas module
- Emphasis on how to process and present data

Free ebook

- https://riptutorial.com/ebook/numpy

NumPy Fundamentals

- Numpy is a module for faster vector processing with numerous other routines
- Scipy is a more extensive module that also includes many other functionalities such as machine learning and statistics

NumPy Fundamentals

- Why Numpy?
  - Remember that Python does not limit lists to just elements of a single class
  - If we have a large list $[a_1, a_2, a_3, \ldots, a_n]$ and we want to add a number to all of the elements, then Python will asks for each element:
    - What is the type of the element
    - Does the type support the + operation
    - Look up the code for the + and execute
  - This is slow

NumPy Fundamentals

- Why Numpy?
  - Primary feature of Numpy are arrays:
    - List like structure where all the elements have the same type
      - Usually a floating point type
    - Can calculate with arrays much faster than with list
    - Implemented in C / Java for Cython or Jython

Getting Numpy

- The new versions of Python (3.9.1) will
  - install pip (the python installer package)
  - put python and pip on the path automatically
- MacOS has Python 2.7 installed
  - python and pip defaults to Python 2.7
  - To install packages use
    - `pip3 install numpy`
    - Then install scipy, matplotlib, pandas

NumPy Arrays

- NumPy Arrays are containers for numerical values
- Numpy arrays have dimensions
  - Vectors: one-dimensional
  - Matrices: two-dimensional
  - Tensors: more dimensions, but much more rarely used
- Nota bene: A matrix can have a single row and a single column, but has still two dimensions

NumPy Arrays

- After installing, try out `import numpy as np`
- Making arrays:
  - Can use lists

```
import numpy as np
my_list = [1,5,4,2]
my_vec = np.array(my_list)
my_list = [[1,2],[4,3]]
my_mat = np.array(my_list)
```

Numpy Data Types

- Aside:
  - Because numpy is based on arrays, numpy has many more data types
    - int_ intc intp int8 int16 int32 int64
    - uint8 uint16 uint32 uint64
    - float_ float16 float32 float64
    - complex_ complex64 complex128
    - bool
    - <U8   (strings)

Numpy Data Types

- Example:

```
example = np.array([1,2,'three'])
```

  - Creates an np.array of strings as the most generic type

```
array(['1', '2', 'three'], dtype='<U21')
```

- Example:

```
example = np.array([1,2,3.1])
```

  - Creates an np.array of 64 bits floats

```
>>> example.dtype
dtype('float64')
```

Array Creation

- Numpy can generate arrays:
  - From disks or the net,
    - using various libraries
    - using loadtxt and similar functions
  - From lists and similar data structures
  - Generate them natively

Array Creation

- Numpy has a number of ways to create an array
  - Import numpy as np
  - `np.zeroes((2,3))`
    - `array([[ 0., 0., 0.], [ 0., 0., 0.]])`
  - `np.ones(5)`
    - `array([1., 1., 1., 1. , 1.])`
  - `np.eye(3)` generates the identity matrix
    - `array([[1., 0., 0.], [0., 1., 0.],[0., 0., 1.]])])`

Array Creation

- Numpy has a number of ways to create arrays
  - `np.linspace(1., 4., 6)` creates an array of 6 elements between 1.0 and 4.0 evenly spaced out
    - `array([ 1. ,  1.6,  2.2,  2.8,  3.4,  4. ])`
  - `np.arange(2, 3, 0.1)` a more generalized version of Python's range function (with float step)
    - `array([ 2. , 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9])`
  - `np.arange(2,5)`
    - `array([2, 3, 4])`

Array Creation

- Can generate using lists, tuples, etc. even with a mix of types
  - `np.array([[1,2,3], (1,0,0.5)])`
    - `array([[1. , 2. , 3. ], [1. , 0. , 0.5]])`

Array Creation

- Creating arrays:
  - np.full to fill in with a given value

```
np.full(5, 3.141)

array([3.141, 3.141, 3.141, 3.141, 3.141])
```

Array Creation

- For multi-dimensional arrays:
  - replace the length with a tuple of dimensions
    - This is called the **shape**
  - Example:

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

       [[1., 1.],
        [1., 1.],
        [1., 1.],
        [1., 1.]],

       [[1., 1.],
        [1., 1.],
        [1., 1.],
        [1., 1.]]])
```

Array Creation

- Can also use random values.
  - Uniform distribution between 0 and 1

```
>>> np.random.random((3,2))
array([[0.39211415, 0.50264835],
       [0.95824337, 0.58949256],
       [0.59318281, 0.05752833]])
```

Array Creation

- Or random integers

```
>>> np.random.randint(0,20,(2,4))

array([[ 5,  7,  2, 10],
       [19,  7,  1, 10]])
```

Array Creation

- Or other distributions, e.g. normal distribution with mean 2 and standard deviation 0.5

```
>>> np.random.normal(2,0.5, (2,3))
array([[1.34857621, 1.34419178, 1.977698  ],
       [1.31054068, 2.35126538, 3.25903903]])
```

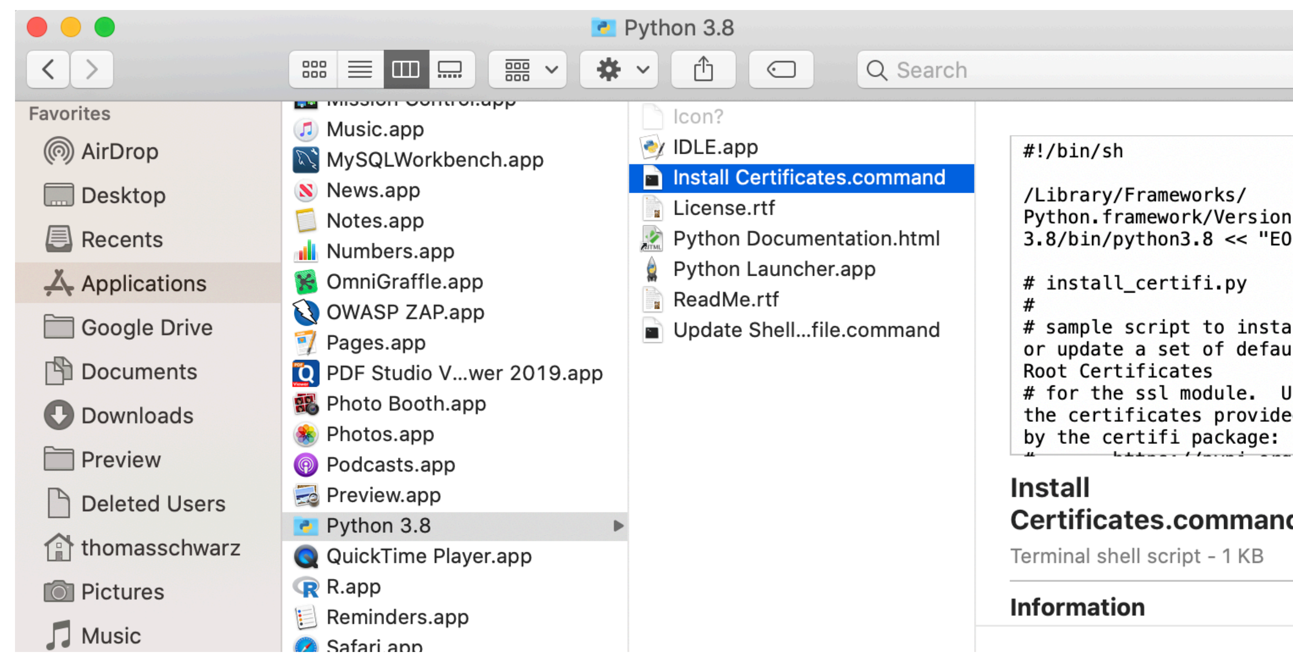
Array Creation

- fromfunction

```
>>> x = np.fromfunction(lambda i,j: (i**2+j**2)//2, (4,5) )

>>> x.astype(int)

array([[ 0,  0,  2,  4,  8],
       [ 0,  1,  2,  5,  8],
       [ 2,  2,  4,  6, 10],
       [ 4,  5,  6,  9, 12]])

>>> x.shape
(4,5)
```

Array Creation

- Creating from download / file
  - We use urllib.request module
    - If you are on Mac, you need to have Python certificates installed
      - Go to your Python installation in Applications and click on "Install Certificates command"



Array Creation

- Use urllib.request.urlretrieve with website and file name
  - Remember: A file will be created, but the directory needs to exist

```
import urllib.request
urllib.request.urlretrieve(
    url = "https://ndownloader.figshare.com/files/12565
    filename = "avg-monthly-precip.txt"
)
```

- This is a text file, with one numerical value per line
- Then create the numpy array using

```
avgmp = np.loadtxt(fname = 'avg-monthly-precip.txt')
print(avgmp)
```

Array Creation

- Example: Get an account at openweathermap.org/appid
- Install requests and import json
  - Use the openweathermap.org api to request data on a certain town
  - Result is send as a JSON dictionary

Array Creation

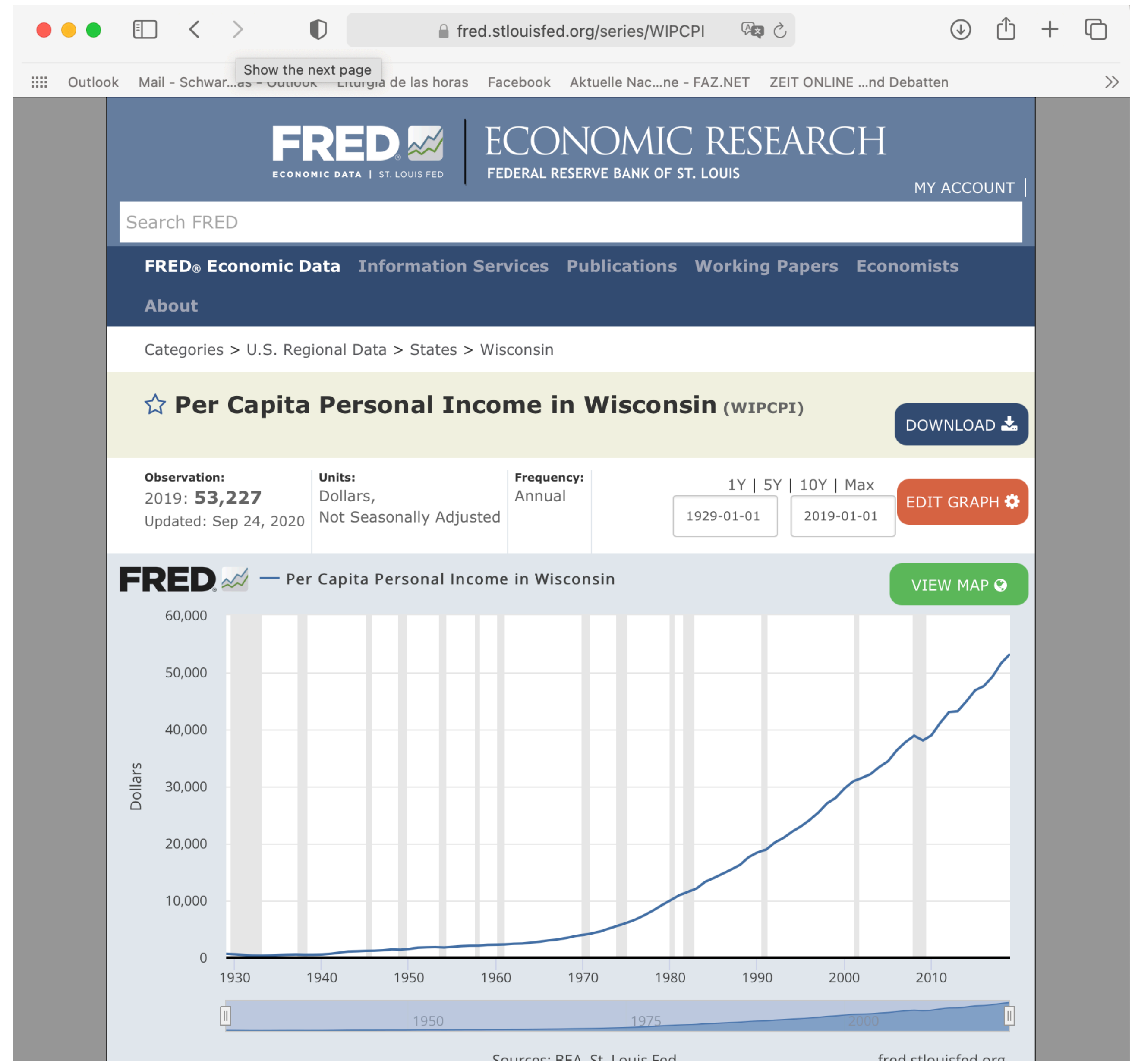
```
import numpy as np
import requests
import json

mumbai=json.loads(requests.get('http://api.openweathermap.org/data/
2.5/weather?
q=mumbai,india&APPID=4561e0cd15ec2ee307bdcfe19ec22ab9').text)
vasai = json.loads(requests.get('http://api.openweathermap.org/data/
2.5/weather?
q=vasai,india&APPID=4561e0cd15ec2ee307bdcfe19ec22ab9').text)
navi_mumbai = json.loads(requests.get('http://api.openweathermap.org/
data/2.5/weather?
q=navi%20mumbai,india&APPID=4561e0cd15ec2ee307bdcfe19ec22ab9').text)
chalco = json.loads(requests.get('http://api.openweathermap.org/data/
2.5/weather?q=Chalco,MX&APPID=4561e0cd15ec2ee307bdcfe19ec22ab9').text)
milwaukee = json.loads(requests.get('http://api.openweathermap.org/
data/2.5/weather?
q=Milwaukee,USA&APPID=4561e0cd15ec2ee307bdcfe19ec22ab9').text)
```

Array Creation

- Can use np.genfromtext
  - Very powerful and complicated function with many different options

Array Creation Example

- Getting data:
  - Standard format for processing is csv
  - Federal Reserve Bank St. Louis: Median annual household income in Wisconsin
  - https://fred.stlouisfed.org



Array Creation Example

- Look at the data:

```
with open('wis.csv') as infile:
   for line in infile:
      print(line.strip())
```

- First line has headings
- All other lines have comma separated two entries
- First is the year, second the income

```
DATE,MEHOINUSWIA672N
1984-01-01,48750
1985-01-01,52819
1986-01-01,59021
1987-01-01,56926
1988-01-01,61621
1989-01-01,58138
1990-01-01,58427
1991-01-01,57151
1992-01-01,59660
...
```

Array Creation Example

- Getting only the income (last column) as int values

```
import numpy as np
import matplotlib.pyplot as plt


with open('wis.csv') as infile:
    data = np.genfromtxt( infile,
                          delimiter = ',',
                          usecols = -1,
                          dtype = int,
                          skip_header = 1)
```

Array Creation Example

- Visualization with matplotlib.pyplot:
  - Import `matplotlib.pyplot as plt`
  - Needs an X and a Y value
    - X-value `np.arange(1984,2020)`
    - Set a number of arguments
    - At the end: use plt.show( ) to actually display the object generated

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

`plt.style.available` displays possible styles

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

`np.arange`
is similar to range,
this are the X-values

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

Y-values

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

Many different ways to specify colors

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

This label is used in the legend

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

Chart title

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

x and y axes get legends

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

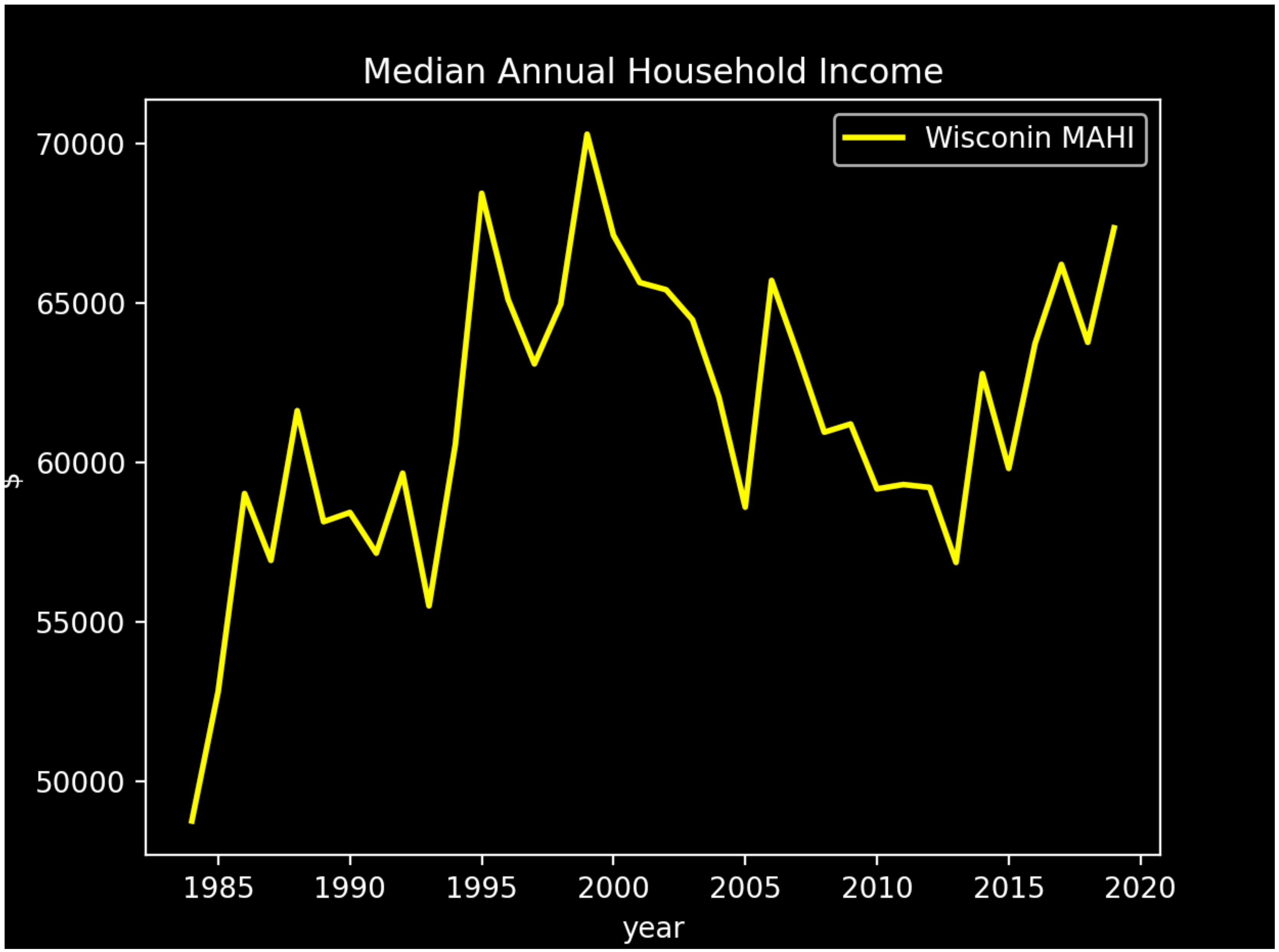
Need to explicitly create the legend

Array Creation Example

```
plt.style.use('dark_background')
plt.plot(np.arange(1984, 2020),
         data,
         color = 'yellow',
         linestyle = '-',
         linewidth = 2,
         label = 'Wisconin MAHI')
plt.title('Median Annual Household Income')
plt.xlabel('year')
plt.ylabel('$')
plt.legend( )
plt.show( )
```

Use plt.show to display the pyplot object

Array Creation Example



Array Creation

- genfromtxt is a complicated function
  - Converters can be used to interpret certain columns

```
converters = {5: lambda x: int(0 if 'Iris-setosa' else 1 if 'Iris-
virginica' else 2) }
my_array = np.genfromtxt('../Classes2/Iris.csv',
                         usecols=(1,2,3,4,5),
                         dtype =[float, float, float, float, int],
                         delimiter = ',',
                         converters = converters,
                         skip_header=1)
```

Array Creation

- genfromtxt is a complicated function
  - dtype with a list will create a structured array with default names
    - This can be useful or problematic

```
converters = {5: lambda x: int(0 if 'Iris-setosa' else 1 if 'Iris-virginica' else 2) }
my_array = np.genfromtxt('Iris.csv',
                         usecols=(1,2,3,4,5),
                         dtype =[float, float, float, float, int],
                         delimiter = ',',
                         converters = converters,
                         skip_header=1)
```

Array Creation

- This is an array of 150 tuples
- Use comprehension to convert to a two-dimensional array

```
m = np.array( [ [row[0], row[1], row[2], row[3], row[4]]
                 for row in my_array ])
```

Array Creation

- Alternatively:
  - Set dtype to a single type

```
converters = {5: lambda x: float(0 if 'Iris-setosa' else 1
if 'Iris-virginica' else 2) }
with open('Iris.csv') as infile:
    my_array = np.genfromtxt(infile,
                             usecols=(1,2,3,4,5),
                             dtype = float,
                             delimiter = ',',
                             converters = converters,
                             skip_header=1)
```

NumPy Array Attributes

- The number of dimensions: ndim
- The values of the dimensions as a tuple: shape
- The size (number of elements)

```
>>> tensor
array([[[2.11208424, 2.01510638, 2.03126777, 1.89670846],
        [1.94359036, 2.02299445, 2.08515919, 2.05402626],
        [1.8853457 , 2.01236192, 2.07019962, 1.93713157]],

       [[1.84275427, 1.99537922, 1.96060154, 1.90020305],
        [2.00270166, 2.11286224, 2.03144254, 2.06924855],
        [1.95375653, 2.0612986 , 1.82571628, 1.86067971]]])
>>> tensor.ndim
3
>>> tensor.shape
(2, 3, 4)
>>> tensor.size
24
```

NumPy Array Attributes

- The data type: dtype
  - can be bool, int, int64, uint, uint64, float, float64, complex ...
- The size of a single element in bytes: itemsize
- The size of the total array: nbytes

NumPy Array Indexing

- Single elements
  - Use the bracket notation [ ]
    - Single array: Same as in standard python

```
>>> vector = np.random.normal(10,1,(5))
>>> print(vector)
[10.25056641 11.37079651 10.44719557 10.54447875 10.43634562]
>>> vector[4]
10.436345621654919
>>> vector[-2]
10.544478746079845
```

NumPy Arrays Indexing

- Matrix and tensor elements: Use a single bracket and a comma separated tuple

```
>>> tensor
array([[[2.11208424, 2.01510638, 2.03126777, 1.89670846],
        [1.94359036, 2.02299445, 2.08515919, 2.05402626],
        [1.8853457 , 2.01236192, 2.07019962, 1.93713157]],

       [[1.84275427, 1.99537922, 1.96060154, 1.90020305],
        [2.00270166, 2.11286224, 2.03144254, 2.06924855],
        [1.95375653, 2.0612986 , 1.82571628, 1.86067971]]])
>>> tensor[0,0,1]
2.0151063761913
```

NumPy Arrays Indexing

- Multiple bracket notation
  - We can also use the Python indexing of multi-dimensional lists using several brackets

```
>>> tensor[0][1][2]
2.0851591502853
```

  - It is more writing and more error prone than the single bracket version

NumPy Arrays Indexing

- We can also define slices

```
>>> vector = np.random.normal(10,1,(3))
>>> vector
array([10.61948855,  7.99635252,  9.05538706])
>>> vector[1:3]
array([7.99635252, 9.05538706])
```

NumPy Arrays Indexing

- In Python, slices are new lists
- In NumPy, slices are **not** copies
  - Changing a slice changes the original

NumPy Arrays Indexing

- Example:
  - Create an array

```
>>> vector = np.random.normal(10,1,(3))
>>> vector
array([10.61948855,  7.99635252,  9.05538706])
```

  - Define a slice

```
>>> x = vector[1:3]
```

NumPy Arrays Indexing

- Example (cont.)
  - Change the first element in the slice

```
>>> x[0] = 5.0
```

  - Verify that the change has happened

```
>>> x
array([5.        , 9.05538706])
```

  - But the original has also changed:

```
>>> vector
array([10.61948855,  5.        ,  9.05538706])
```

NumPy Arrays Indexing

- Slicing does **not** makes copies
  - This is done in order to be efficient
    - Numerical calculations with a large amount of data get slowed down by unnecessary copies

NumPy Arrays Indexing

- If we want a copy, we need to make one with the copy method
- Example:
  - Make an array

```
>>> vector = np.random.randint(0,10,5)
>>> vector
array([0, 9, 5, 7, 8])
```

  - Make a copy of the array

```
>>> my_vector_copy = vector.copy()
```

NumPy Arrays Indexing

- Example (continued)
  - Change the middle elements in the copy

```
>>> my_vector_copy[1:-2]=100
```

  - Check the change

```
>>> my_vector_copy
array([  0, 100, 100,   7,   8])
```

  - Check the original

```
>>> vector
array([0, 9, 5, 7, 8])
```

  - No change!

NumPy Arrays Indexing

- Multi-dimensional slicing
  - Combines the slicing operation for each dimension

```
>>> slice = tensor[1:, :2, :1]
>>> slice
array([[[1.84275427],
        [2.00270166]]])
```

NumPy Arrays Conditional Selection

- We can create an array of Boolean values using comparisons on the array

```
>>> array = np.random.randint(0,10,8)
>>> array
array([2, 4, 4, 0, 0, 4, 8, 4])
>>> bool_array = array > 5
>>> bool_array
array([False, False, False, False, False,
       False,  True, False])
```

NumPy Arrays
Conditional Selection

- We can then use the Boolean array to create a selection from the original array

```
>>> selection=array[bool_array]
>>> selection
array([8])
```

- The new array only has one element!

Selftest

- Can you do this in one step?
  - Create a random array of 10 elements between 0 and 10
  - Then select the ones larger than 5

Selftest Solution

- Solution:
  - Looks a bit cryptic
    - First, we create an array

```
>>> arr = np.random.randint(0,10,10)
>>> arr
array([3, 2, 7, 8, 7, 2, 1, 0, 4, 8])
```

    - Then we select in a single step

```
>>> sel = arr[arr>5]
>>> sel
array([7, 8, 7, 8])
```

NumPy Arrays Conditional Selection

- Let's try this out with a matrix
  - We create a vector, then use **reshape** to make the array into a vector
    - Recall: the number of elements needs to be the same

```
>>> mat = np.arange(1,13).reshape(3,4)
>>> mat
array([[ 1,  2,  3,  4],
       [ 5,  6,  7,  8],
       [ 9, 10, 11, 12]])
```

NumPy Arrays Conditional Selection

- Now let's select:

```
>>> mat1 = mat[mat>6]
>>> mat1
array([ 7,  8,  9, 10, 11, 12])
```

- This is no longer a matrix, which makes sense

Slicing

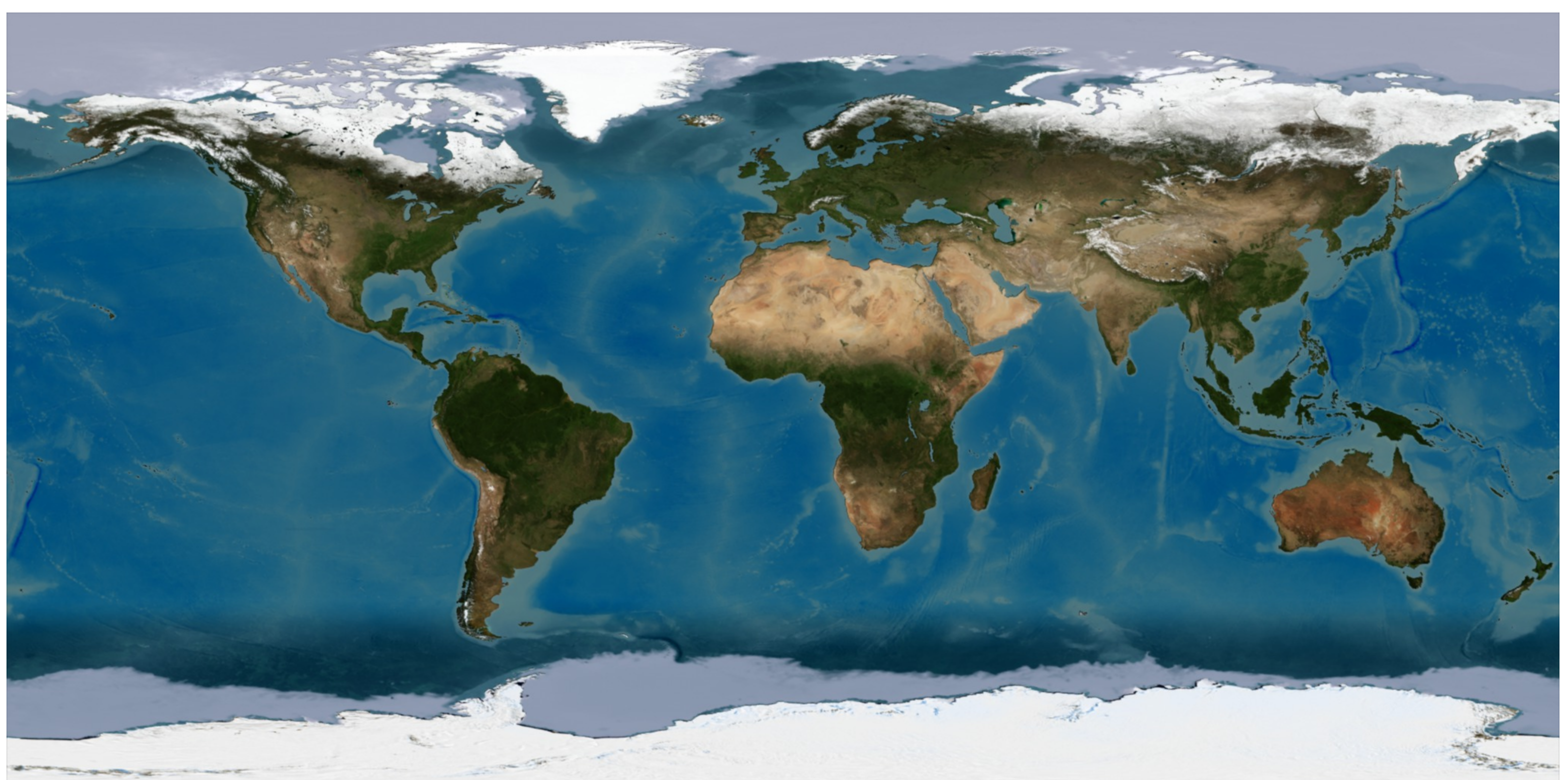
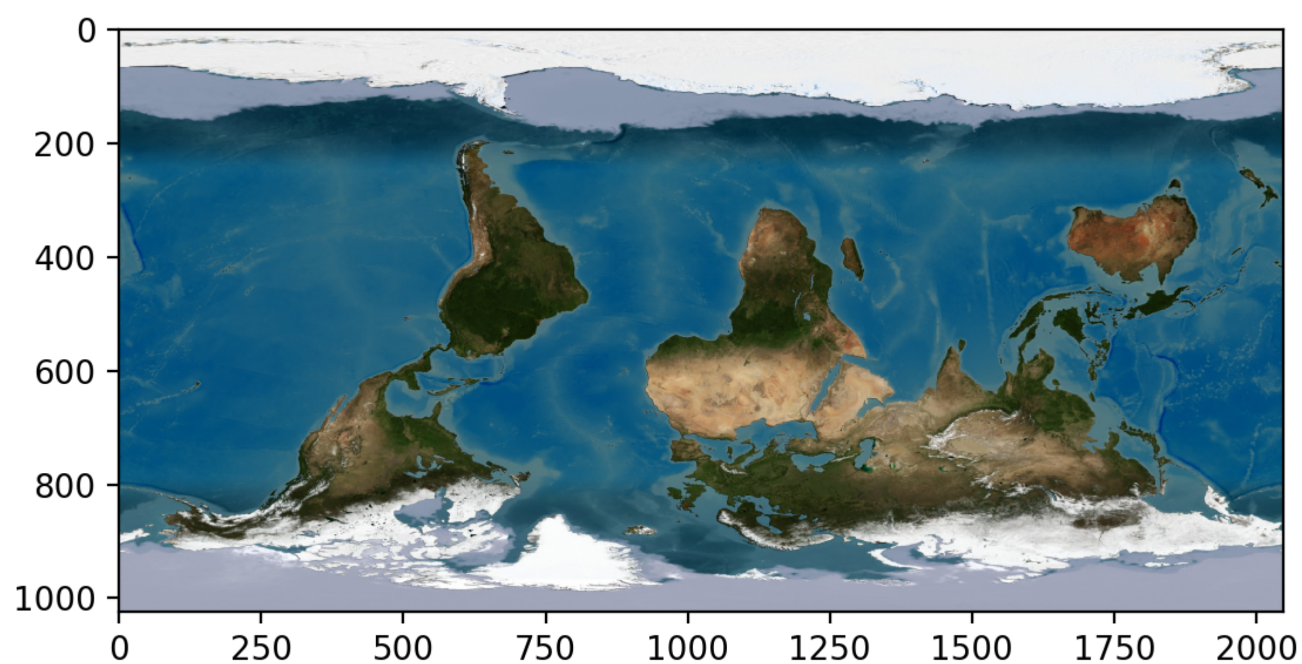
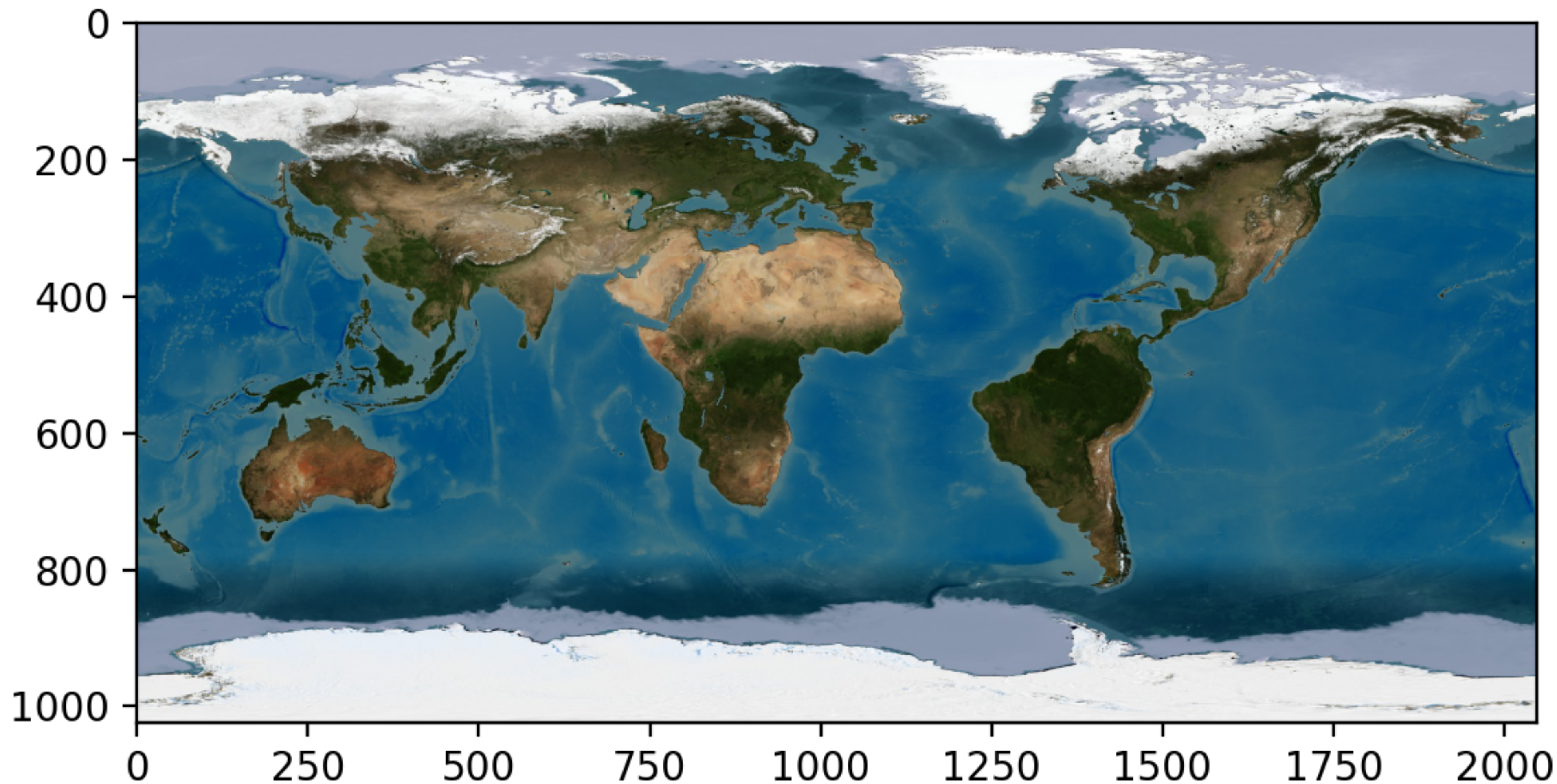
- Photo Manipulation
  - Need to install imageio and matplotlib

```
import imageio
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
```

  - Get a photo as a 3-dimensional array
    -
```
im = mpimg.imread('earth.jpg')
print(im.shape)
```

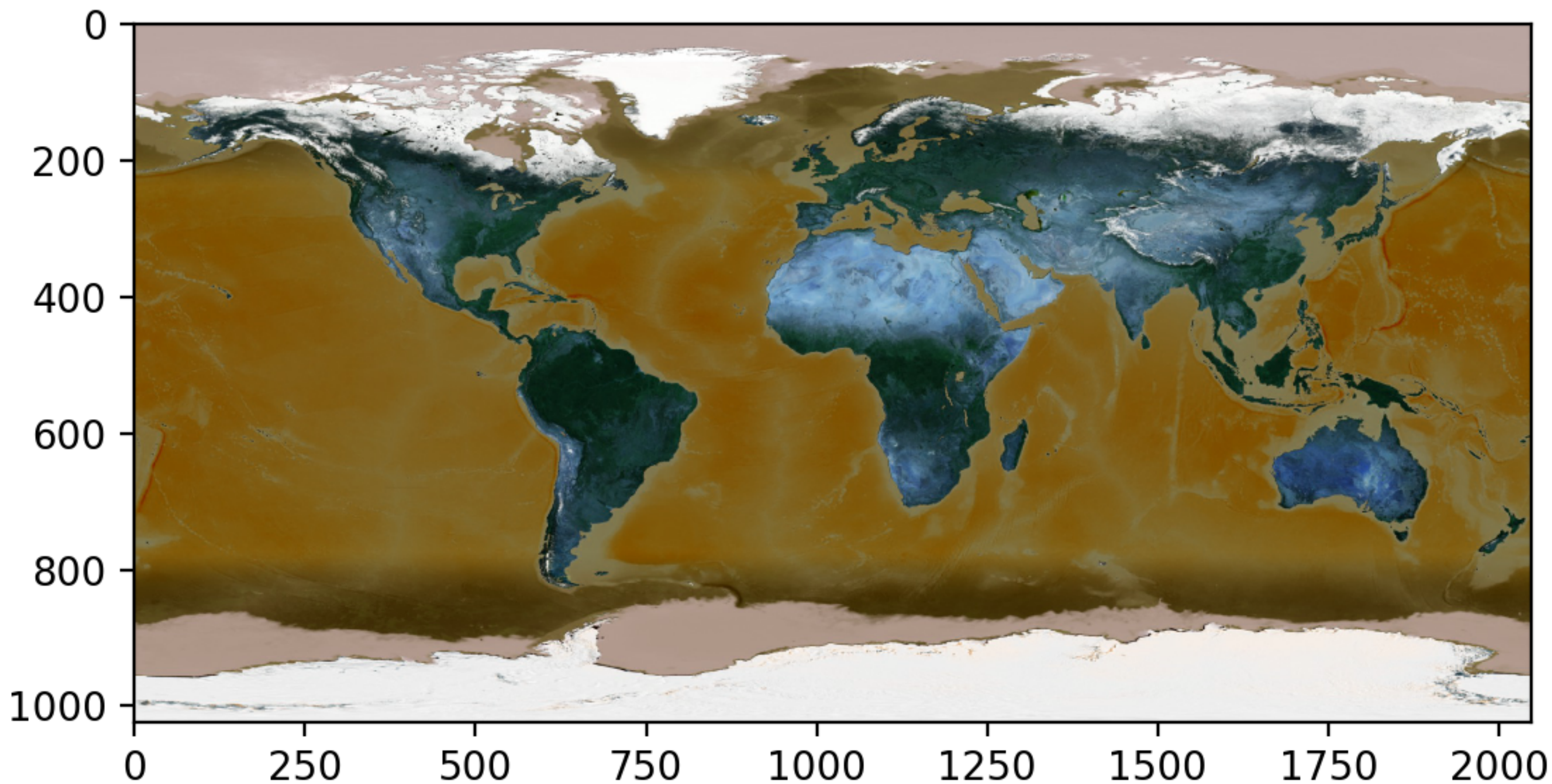
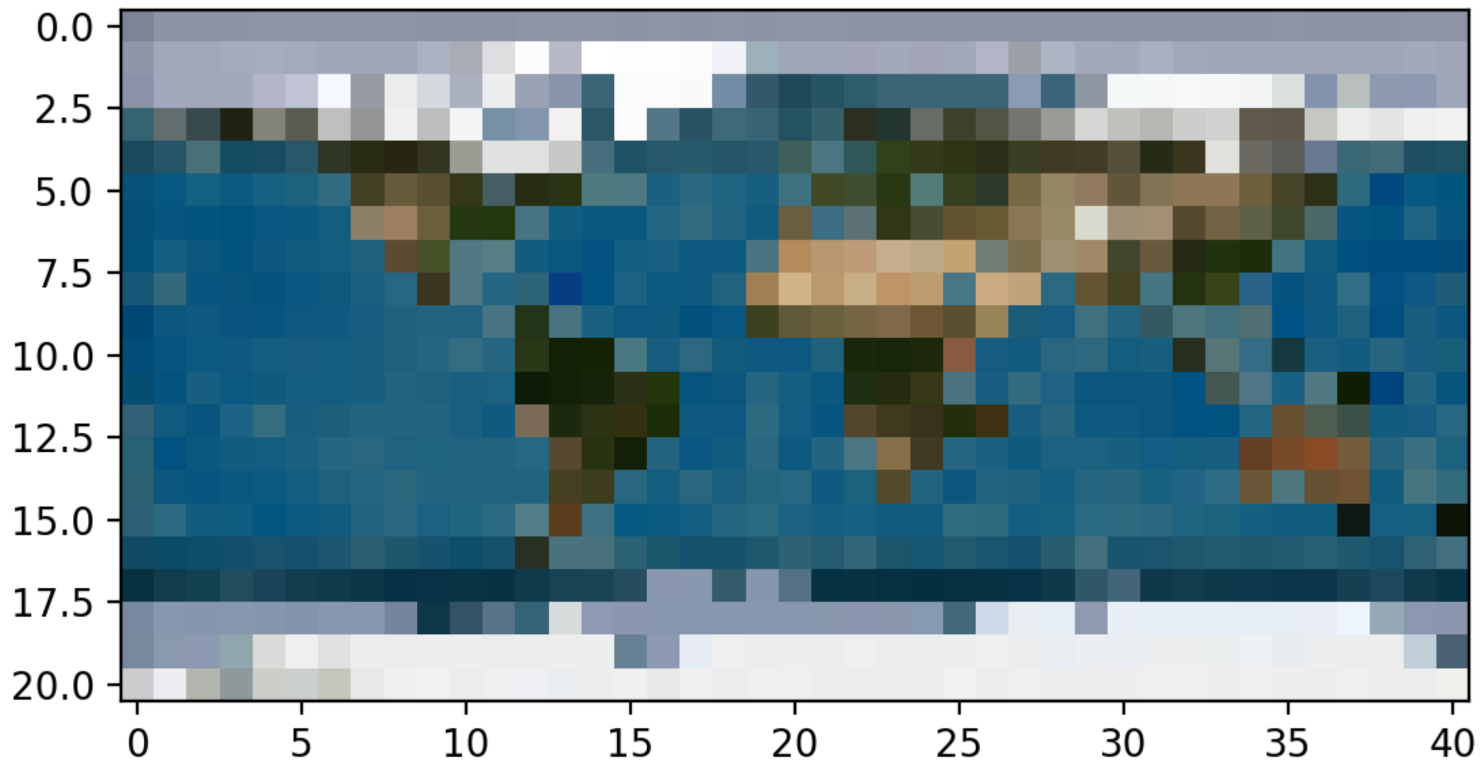
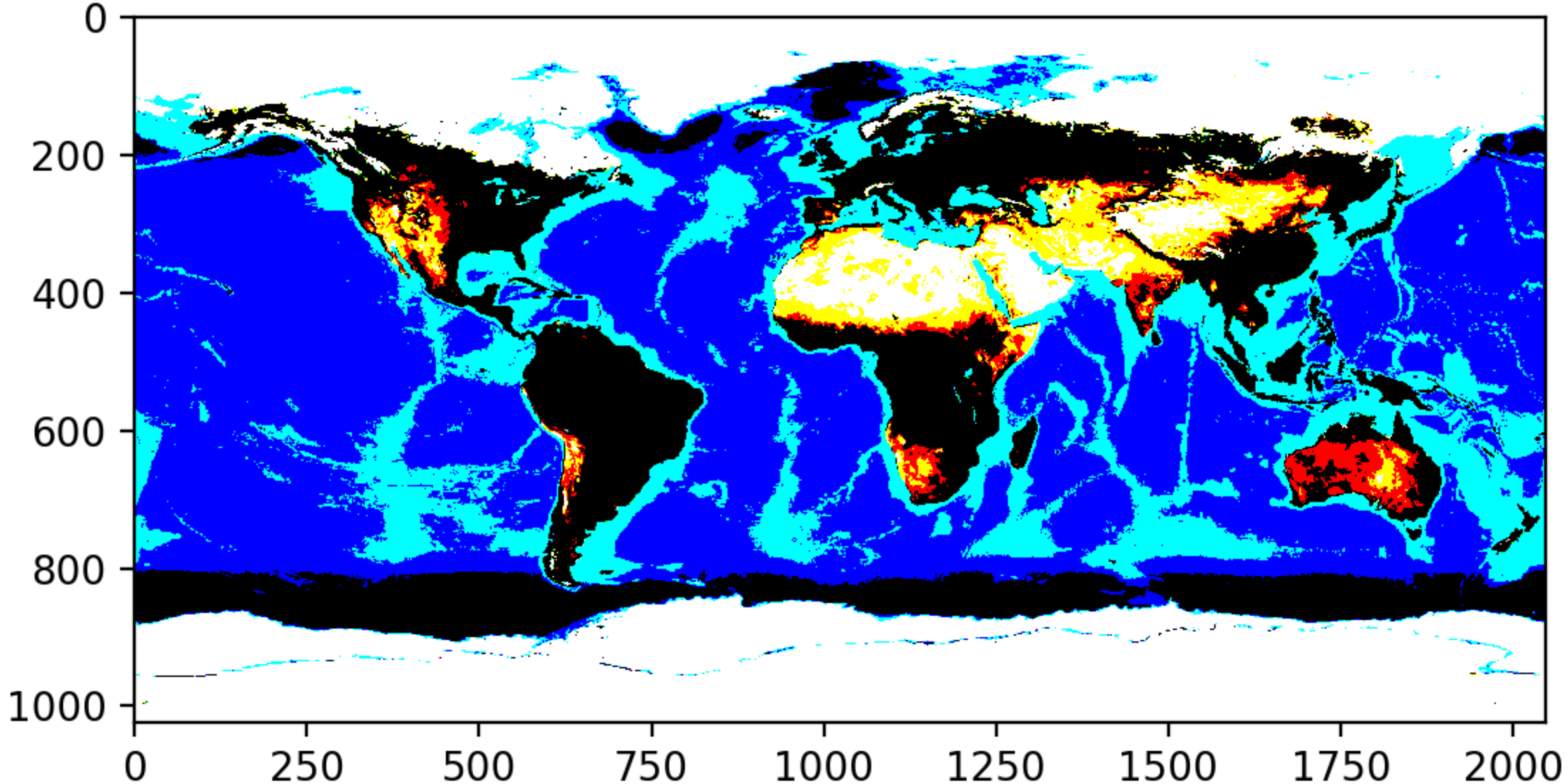
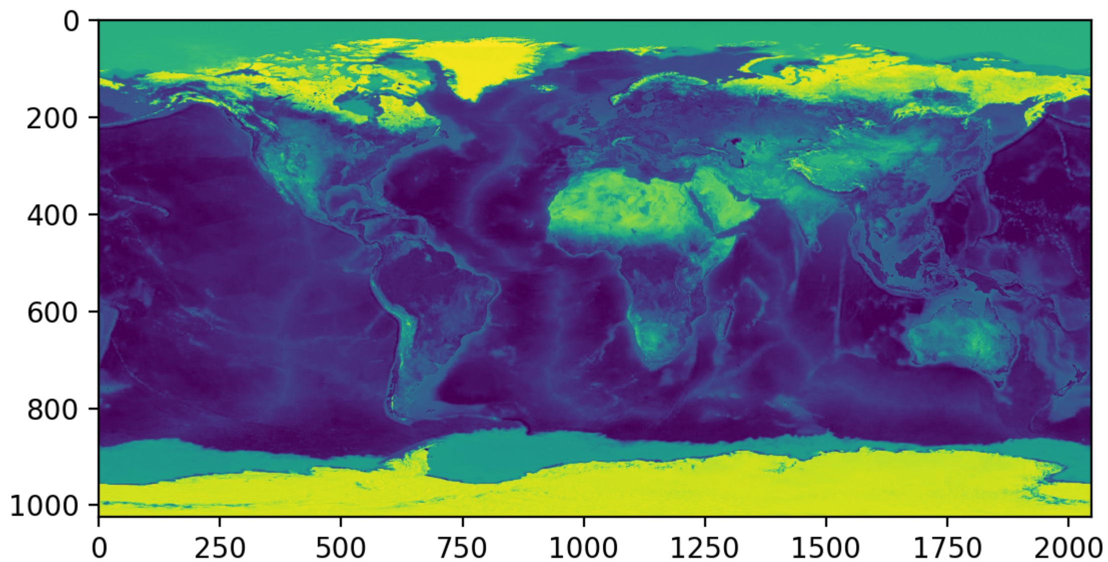
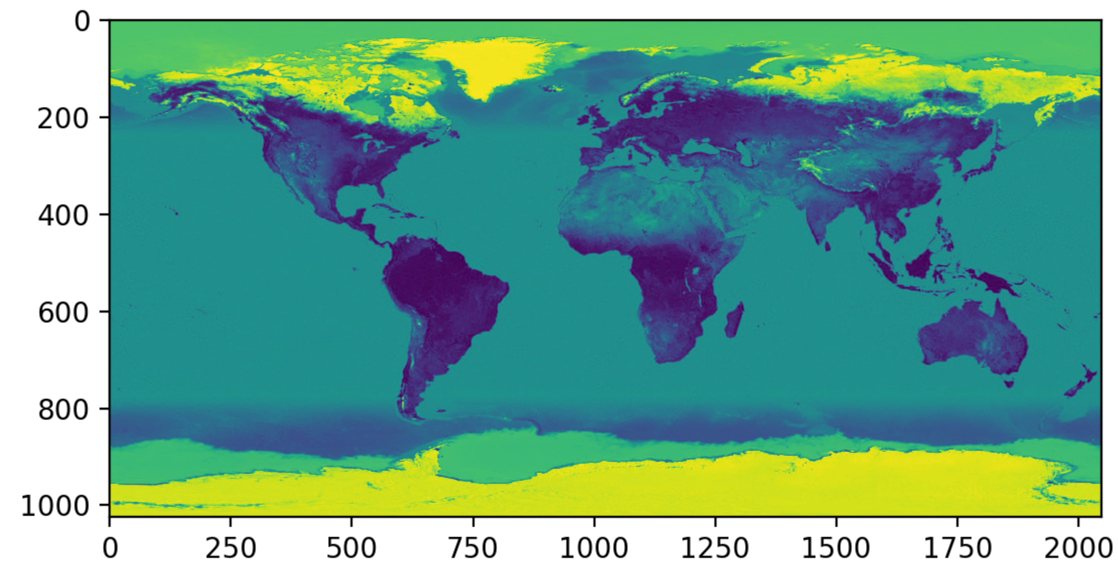
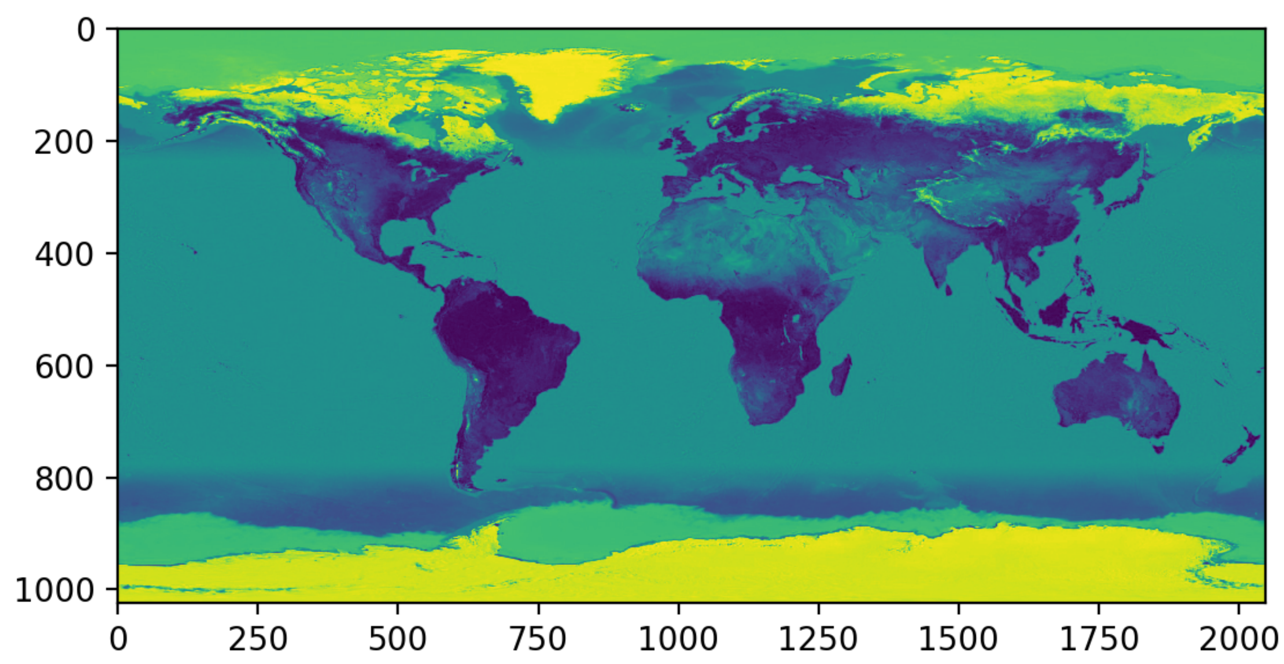
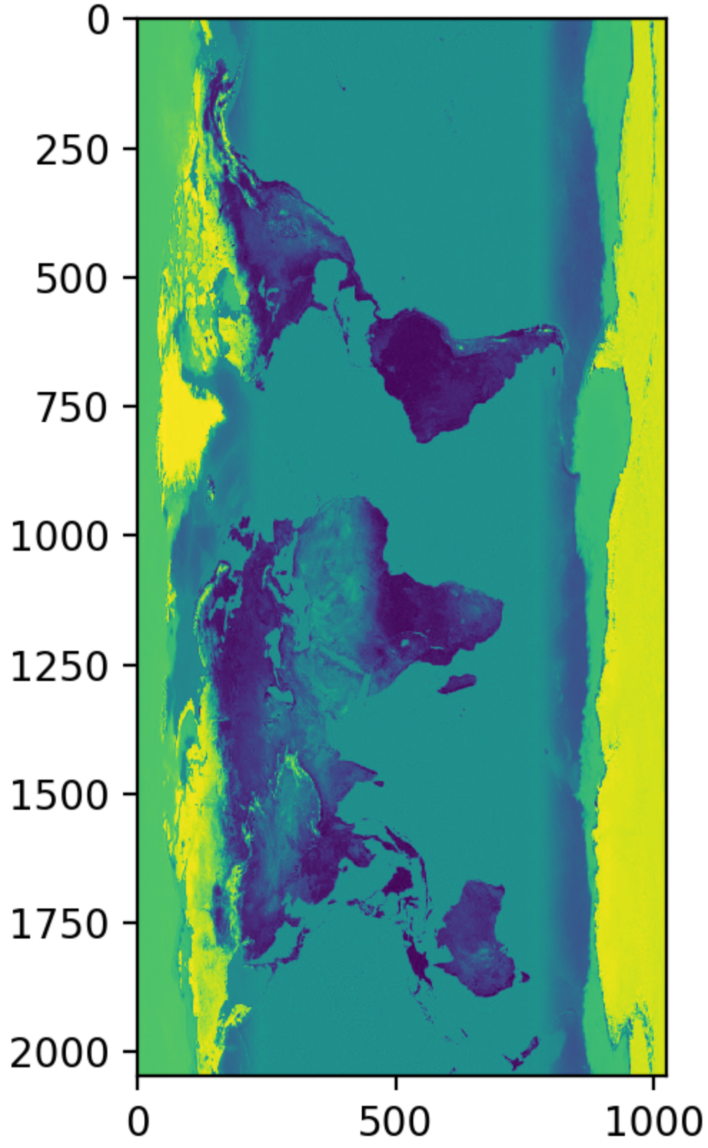
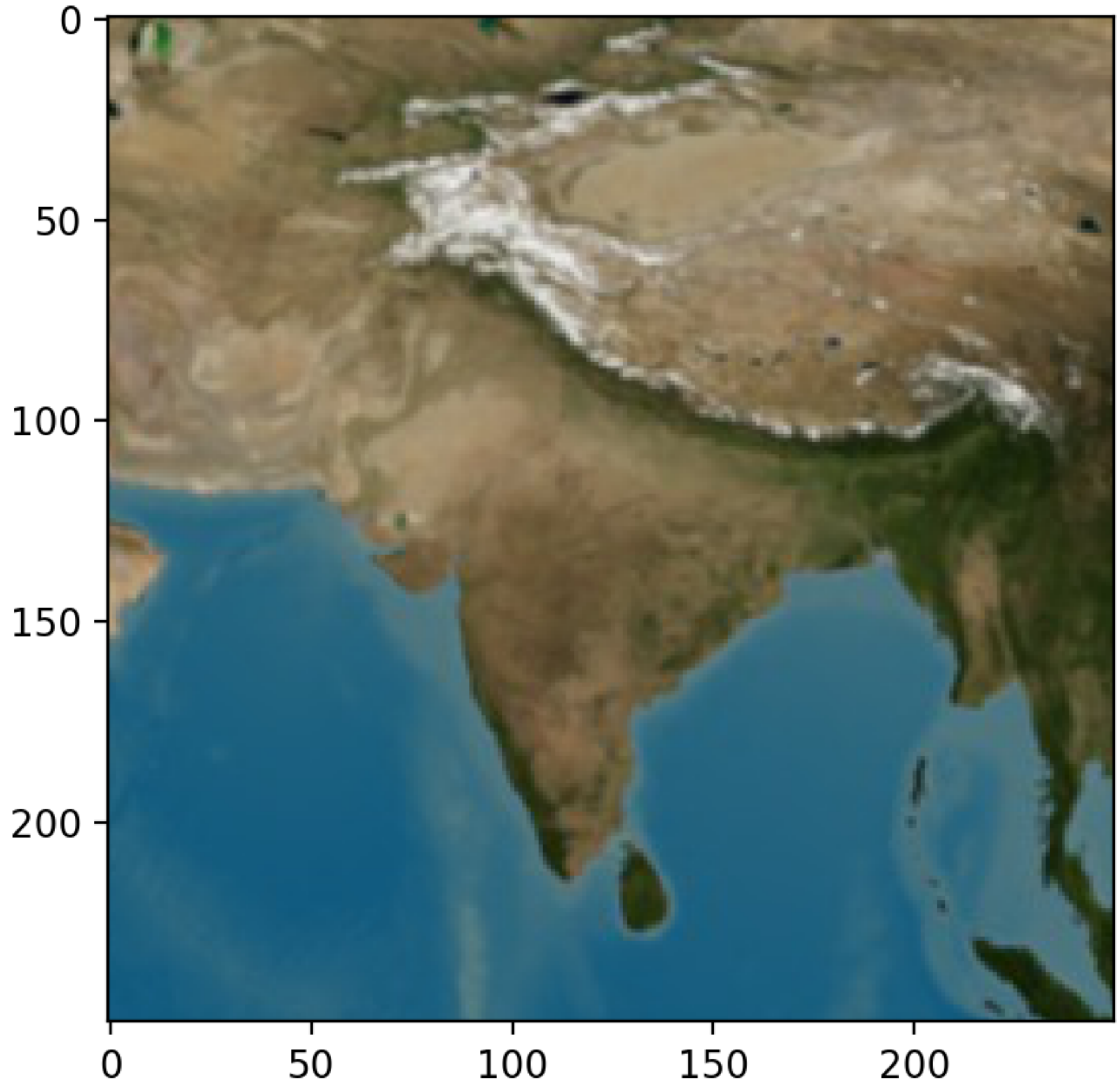
Slicing

- Display the photo
- ```
  plt.imshow(im)
  plt.show()
  ```


  ```

Slicing

- Swap first coordinate

```
plt.imshow(im[::-1,])
plt.show()
```



Slicing

- Swap second coordinate
- ```
  plt.imshow(im[:,::-1,])
  plt.show()
  ```


  ```

Slicing

- Swap third coordinate (color coordinate)
  - ```
    plt.imshow(im[:,:,::-1])
    plt.show()
    ```


    ```

Slicing

- Take every 50th line and column

```
plt.imshow(im[::50, ::50,])
plt.show()
```



Slicing

- We can also apply functions
  - np.where allows us to replace values
    - `image = np.where(im>100, 255, 0)`
      - Where-ever the value of the image is less than 100, replace it with 0
      - Otherwise, replace it with 255

Slicing



Slicing

- Can use a sub-image

```
plt.imshow(im[:,:,0])
plt.show()
```

```
plt.imshow(im[:,:,1])
plt.show()
```

```
plt.imshow(im[:,:,2])
plt.show()
```

Slicing


0
200
400
600
800
1000
0 250 500 750 1000 1250 1500 1750 2000


0
200
400
600
800
1000
0 250 500 750 1000 1250 1500 1750 2000


0
200
400
600
800
1000
0 250 500 750 1000 1250 1500 1750 2000

Slicing

- Can use the transpose

```
plt.imshow(im[:,:,2].T)
plt.show()
```



Slicing

- Or just concentrate on the essential

```
plt.imshow(im[250:500,1350:1600, : ])
```

Slicing



NumPy Operations

- Numpy allows fast operations on array elements
- We can simply add, subtract, multiply or divide by a scalar

```
>>> vector = np.arange(20).reshape(4,5)
>>> vector
array([[ 0,  1,  2,  3,  4],
       [ 5,  6,  7,  8,  9],
       [10, 11, 12, 13, 14],
       [15, 16, 17, 18, 19]])
>>> vector += 1
>>> vector
array([[ 1,  2,  3,  4,  5],
       [ 6,  7,  8,  9, 10],
       [11, 12, 13, 14, 15],
       [16, 17, 18, 19, 20]])
```

NumPy Operations

- Numpy also allows operations between arrays

```
>>> mat = np.random.normal(0,1,(4,5))
>>> mat
array([[ 0.04646031, -1.32970787,  1.16764921, -0.48342653,  0.42295389],
       [ 0.70547825,  1.51980589,  1.46902433, -0.46742839,  1.42472386],
       [ 0.78756679, -0.39975927,  1.24411043, -0.67336526, -0.92416835],
       [ 0.4708628 , -0.29419976, -0.58634161,  0.29038393, -0.78814955]])
>>> vector + mat
array([[ 1.04646031,  0.67029213,  4.16764921,  3.51657347,  5.42295389],
       [ 6.70547825,  8.51980589,  9.46902433,  8.53257161, 11.42472386],
       [11.78756679, 11.60024073, 14.24411043, 13.32663474, 14.07583165],
       [16.4708628 , 16.70580024, 17.41365839, 19.29038393, 19.21185045]])
```

NumPy Operations

- What happens if there is an error?
  - Python would throw an exception, but not so NumPy
    - Example: Create two vectors, one with a zero

```
>>> vector = np.arange(5)
>>> vector2 = np.arange(2,7)
```

    - If we divide, we get a warning
    - But the result exists, with an inf value for infinity

```
>>> vec = vector2/vector
Warning (from warnings module):
  File "<pyshell#11>", line 1
RuntimeWarning: divide by zero encountered in true_divide
>>> vec
array([        inf, 3.        , 2.        , 1.66666667, 1.5       ])
```

NumPy Operations

- If we divide 0 by 0, we get an nan -- not a value

```
>>> vec=np.arange(4)
>>> vec
array([0, 1, 2, 3])
>>> vec/vec

Warning (from warnings module):
  File "<pyshell#15>", line 1
RuntimeWarning: invalid value encountered in
true_divide
array([nan,  1.,  1.,  1.])
```

NumPy Operations

- There are rules for how to define operations with nan and inf, that make intuitive sense
  - IEEE Standard for Binary Floating-Point Arithmetic (IEEE 754)
- We can create inf directly by saying `np.inf`
  - Example: Infinity divided by infinity is not defined

```
>>> np.inf/np.inf
nan
```

Operations between Vectors and Matrices

- Adding two vectors:

```
>>> v1 = np.array([1,2,3])
>>> v2 = np.array([5,4,3])
>>> v1 + v2
array([6, 6, 6])
```

Operations between Vectors and Matrices

- Adding two matrices

```
>>> m1 = np.array([[1,2,3],[4,5,6],[9,10,0]])
>>> m1
array([[ 1,  2,  3],
       [ 4,  5,  6],
       [ 9, 10,  0]])
>>> m2 = np.array([[4,2,0],[7,3,1],[5,1,2]])
>>> m2
array([[4, 2, 0],
       [7, 3, 1],
       [5, 1, 2]])
>>> m1+m2
array([[ 5,  4,  3],
       [11,  8,  7],
       [14, 11,  2]])
```

Operations between Vectors and Matrices

- Scalar multiplication

```
>>> v = np.array([5,3,-2,4])
>>> 5*v
array([ 25,  15, -10,  20])
```

Operations between Vectors and Matrices

- Scalar multiplication

```
>>> m1
array([[ 1,  2,  3],
       [ 4,  5,  6],
       [ 9, 10,  0]])

>>> 3*m1
array([[ 3,  6,  9],
       [12, 15, 18],
       [27, 30,  0]])
```

Operations between Vectors and Matrices

- Element-wise multiplication **is not matrix multiplication**

```
>>> m1
array([[ 1,  2,  3],
       [ 4,  5,  6],
       [ 9, 10,  0]])
>>> m2
array([[4, 2, 0],
       [7, 3, 1],
       [5, 1, 2]])
>>> m1*m2
array([[ 4,  4,  0],
       [28, 15,  6],
       [45, 10,  0]])
```

Operations between Vectors and Matrices

- **Matrix multiplication uses the (new) @ operator**
    - Python 3.5 and later

```
>>> m1
array([[ 1,  2,  3],
       [ 4,  5,  6],
       [ 9, 10,  0]])
>>> m2
array([[4, 2, 0],
       [7, 3, 1],
       [5, 1, 2]])
>>> m1@m2
array([[ 33,  11,   8],
       [ 81,  29,  17],
       [106,  48,  10]])
```

Operations between Vectors and Matrices

- Can be used to multiply matrix and vector

```
>>> m = np.array([[2,3],[1,-1]])
>>> v = np.array([1,2])
>>> m@v
array([ 8, -1])
```

- Notice that the vector is in row form
- $\begin{pmatrix} 2 & 3 \\ 1 & -1 \end{pmatrix} \cdot (1, \quad 2) = (8, \quad -1)$
- Follows usage of matlab and Mathematica

Operations between Vectors and Matrices

- Transpose with np.transpose or the .T operator

```
>>> m
array([[ 2,  3],
       [ 1, -1]])
>>> m.T
array([[ 2,  1],
       [ 3, -1]])
```

Operations between Vectors and Matrices

- Thus, could have used

```
>>> m@ v.T
array([ 8, -1])
```

Operations between Vectors and Matrices

- We can use this to make a linear transform of a data set

```
def transform(matrix, dataset):
    return (matrix@ dataset.T).T
```

```
mat = np.array([[.1, .2, .3, .4],
                [.2, .2, .3, .4],
                [.1, -.1, .2, 3],
                [3, 2, 1, -2]
               ])
print(transform(mat, iris))
```

Operations between Vectors and Matrices

- Dot-product of two vectors:
  - ```
    v = np.array([1, 2, 3, 4, 5])
    >>> v@v.T
    55
    >>> np.vdot(v,v)55
    ```
    ```

Operations between Vectors and Matrices

- Can use linear algebra package in numpy
  - numpy.linalg
  - $$\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}^{10} = \begin{pmatrix} 243 & 0 \\ 0 & 243 \end{pmatrix}$$

```
np.linalg.matrix_power(np.array([[1,2],[1,-1]]),10)
array([[243,   0],
       [  0, 243]])
```

Operations between Vectors and Matrices

- Can calculate matrix inverses
  - Throws LinAlgError if singular

```
>>> np.linalg.inv( np.array([1,-2],[-2,4]) )
Traceback (most recent call last):
…
numpy.linalg.LinAlgError: Singular matrix
```

Operations between Vectors and Matrices

- Can directly solve linear equations
  - Solving $x + 2y = 2, x - y = 3$
    - With solution $x = 8/9, y = -1/3$
  - Gives an error if matrix is not square or singular

```
>>> np.linalg.solve( np.array([[1,2],[1,-1]]) ,
      np.array([2,3]) )
array([ 2.66666667, -0.33333333])
```

NumPy:
Universal Array Functions

- There is a plethora of functions that can be applied to a numpy array.
- These are much faster than the corresponding Python functions
- You can find a list in the numpy u-function manual
  - https://docs.scipy.org/doc/numpy/reference/ufuncs.html

NumPy:
Universal Array Functions

- There are universal functions around which the operations are wrapped
  - np.add, np.subtract, np.negative, np.multiply, np.divide, np.floor_divide, np.power, np.mod
- The absolute function is
  - abs
  - np.absolute

NumPy: Universal Array Functions

- Trigonometric functions
  - np.sin, np.cos, np.tan, np.arcsin, np.arccos, np.arctan
- Exponents and logarithms
  - np.log, np.log2 (base 2), np.log10 (base 10)
  - np.expm1 (more exact for small arguments)
  - np.log1p (more exact for small arguments)

NumPy: Universal Array Functions

- Special u-functions:
  - In addition, the submodule scipy.special contains many more specialized functions

NumPy: Universal Array Functions

- Avoid creating temporary arrays
  - If they are large, too much time spent on moving data
  - Specify the array using the 'out' parameter

```
>>> y = np.empty(10)
>>> x = np.arange(1,11)
>>> np.exp(x, out = y)
array([2.71828183e+00, 7.38905610e+00, 2.00855369e+01, 5.45981500e+01,
       1.48413159e+02, 4.03428793e+02, 1.09663316e+03, 2.98095799e+03,
       8.10308393e+03, 2.20264658e+04])
>>> y
array([2.71828183e+00, 7.38905610e+00, 2.00855369e+01, 5.45981500e+01,
       1.48413159e+02, 4.03428793e+02, 1.09663316e+03, 2.98095799e+03,
       8.10308393e+03, 2.20264658e+04])
```

NumPy: Universal Array Functions

- Can use np.min, np.max, sum
- Use np.argmin, np.argmax to find the index of the maximum / minimum element
- Can use np.mean, np.std, np.var, np.median, mp.percentile to get statistics
  - Not the only way, see the scipy module

NumPy: Broadcasting

- Operations can be also made between arrays of different sizes
    - Example 1: adding a scalar (zero-dimensional) to a vector

```
>>> x = np.full(5,1)
>>> x+1
array([2, 2, 2, 2, 2])
```

NumPy: Broadcasting

- Adding a vector to a matrix:
  - Create a matrix

```
>>> matrix = np.arange(1,11).resha
>>> matrix
array([[ 1,  2,  3,  4,  5],
       [ 6,  7,  8,  9, 10]])
```

  - Create a vector

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

  - Add them together: The vector has been broadcast to a 2 by 5 matrix by doubling the single row

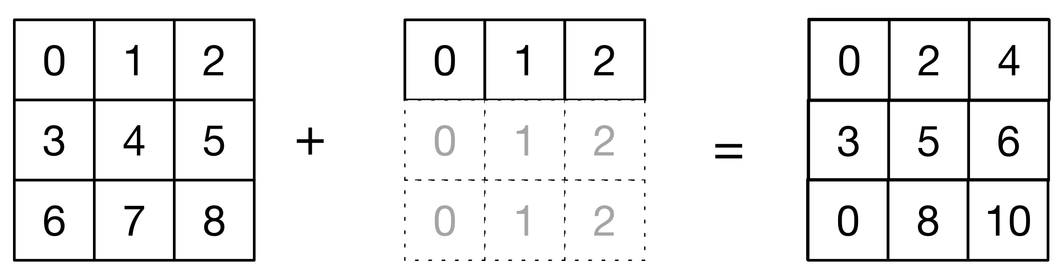
```
>>> matrix+x
array([[ 2,  4,  6,  8, 10],
       [ 7,  9, 11, 13, 15]])
```

NumPy: Broadcasting

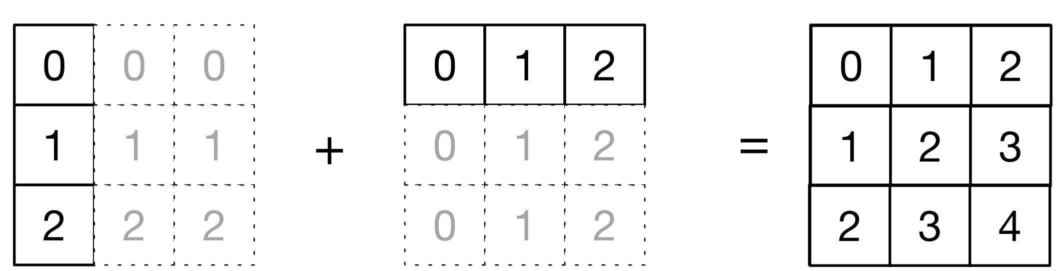
- The broadcast rules: Expand a single coordinate in a dimension in one operand to the value in the other

np.arrange(3) + 5

np.arrange(9).reshape((3,3)) + np.arrange(3)



np.arrange(3).reshape((3,1)) + np.arrange(3)



NumPy: Broadcasting

- Rule 1: If the two arrays differ in their number of dimensions, the shape of the one with fewer dimensions is padded with ones on its leading site
- Rule 2: If the shape of two arrays does not match in any dimension, the array with shape equal to 1 in that dimension is stretched to match the other shape
- Rule 3: If in any dimensions the sizes disagree and neither is equal to 1, an error is raised

Neat Example

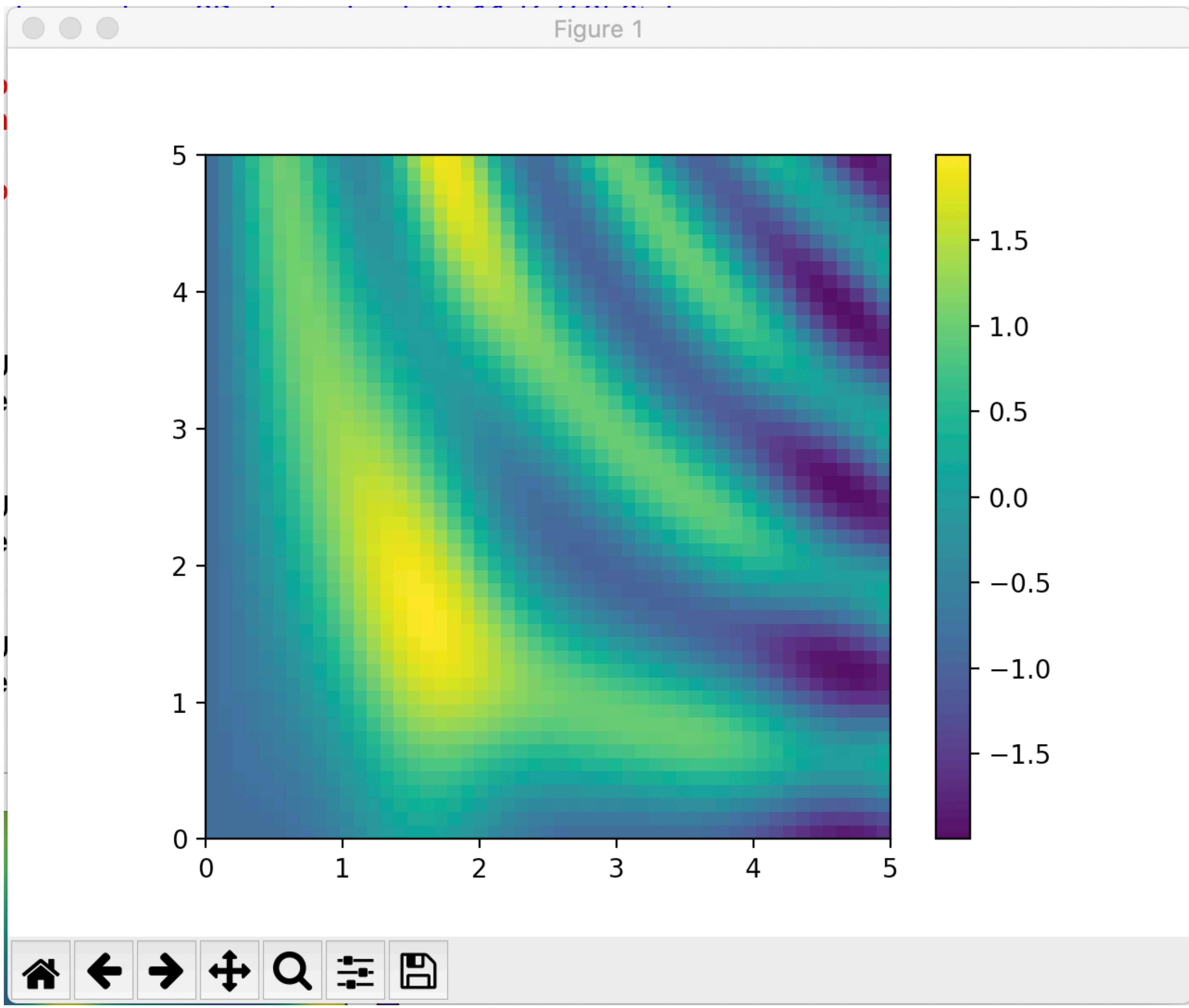
- We combine broadcasting with mathplotlib
    - Using IDLE, we need to call the show function at the end.

NumPy: Broadcasting

- Create a row and a column vector x and y
- Then use broadcasting to combine them for something two-dimensional
- This will get displayed

```
import matplotlib.pyplot as plt
def prob7():
    x = np.linspace(0,5,51)
    y = np.linspace(0,5,51).reshape(51,1)
    z = np.sin(x)**5+np.cos(10+x*y)
    plt.imshow(z, origin='lower', extent=[0, 5, 0, 5],
           cmap='viridis')
    plt.colorbar()
    plt.show()
```

NumPy: Broadcasting



NumPy: Fancy Indexing

- Fancy indexing:
  - Use an array of indices in order to access a number of array elements at once

NumPy: Fancy Indexing

- Example:
  - Create matrix

```
>>> mat = np.random.randint(0,10,(3,5))
>>> mat
array([[3, 2, 3, 3, 0],
       [9, 5, 8, 3, 4],
       [7, 5, 2, 4, 6]])
```

  - Fancy Indexing:

```
>>> mat[(1,2),(2,3)]
array([8, 4])
```

NumPy: Fancy Indexing

- Application:
  - Creating a sample of a number of points
- Create a large random array representing data points

```
>>> mat = np.random.normal(100,20, (200,2))
```

- Select the x and y coordinates by slicing

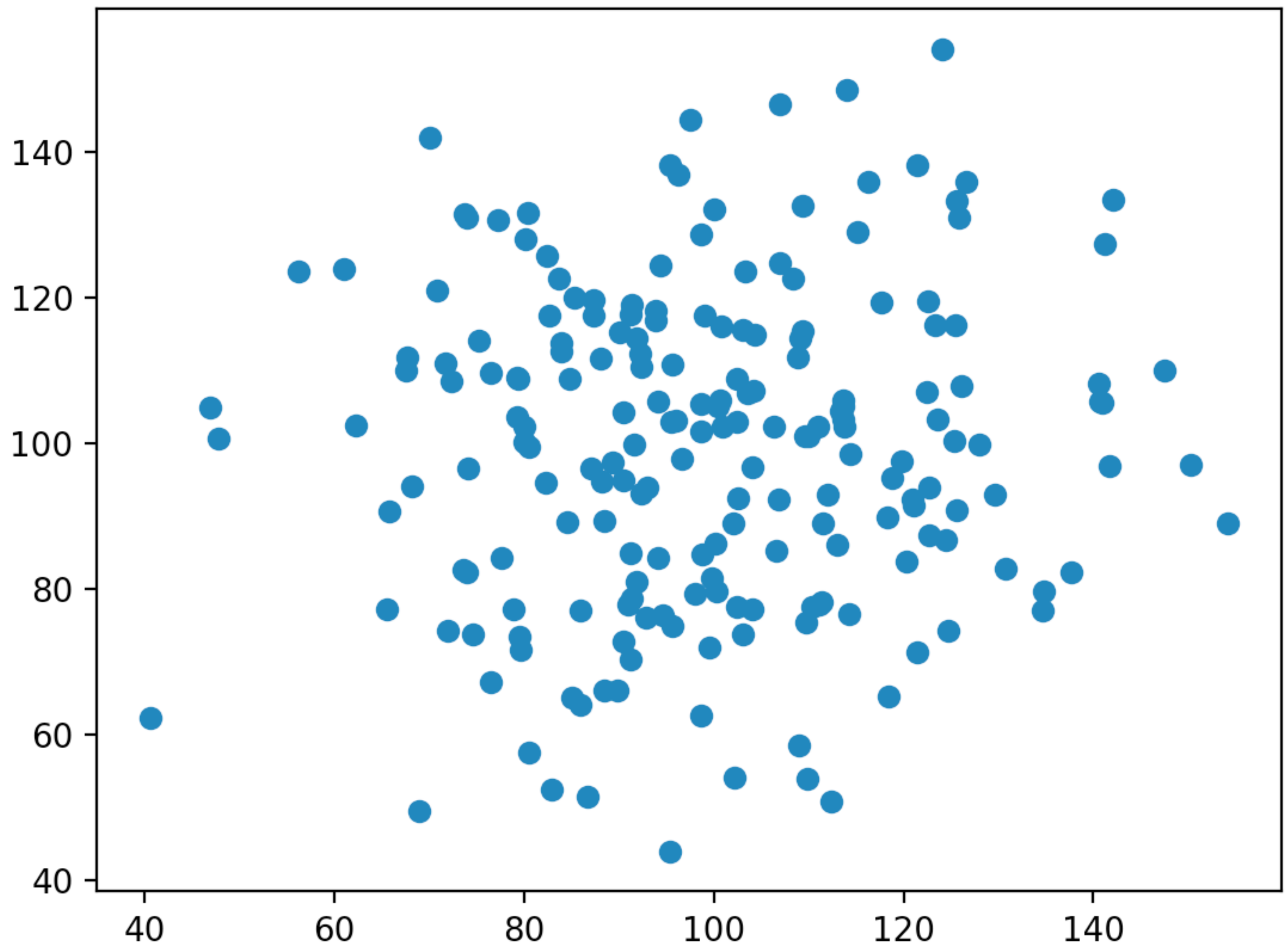
```
>>> x=mat[:,0]
>>> y=mat[:,1]
```

NumPy: Fancy Indexing

- Create a matplotlib figure with a plot inside it

```
>>> fig = plt.figure()
>>> ax = fig.add_subplot(1,1,1)
>>> ax.scatter(x,y)
>>> plt.show()
```

NumPy: Fancy Indexing



NumPy: Fancy Indexing

- Create a list of potential indices

```
>>> indices = np.random.choice(np.arange(0,200,1),10)
>>> indices
array([ 32,  93, 172, 134,  90,  66, 109, 158, 188,
30])
```

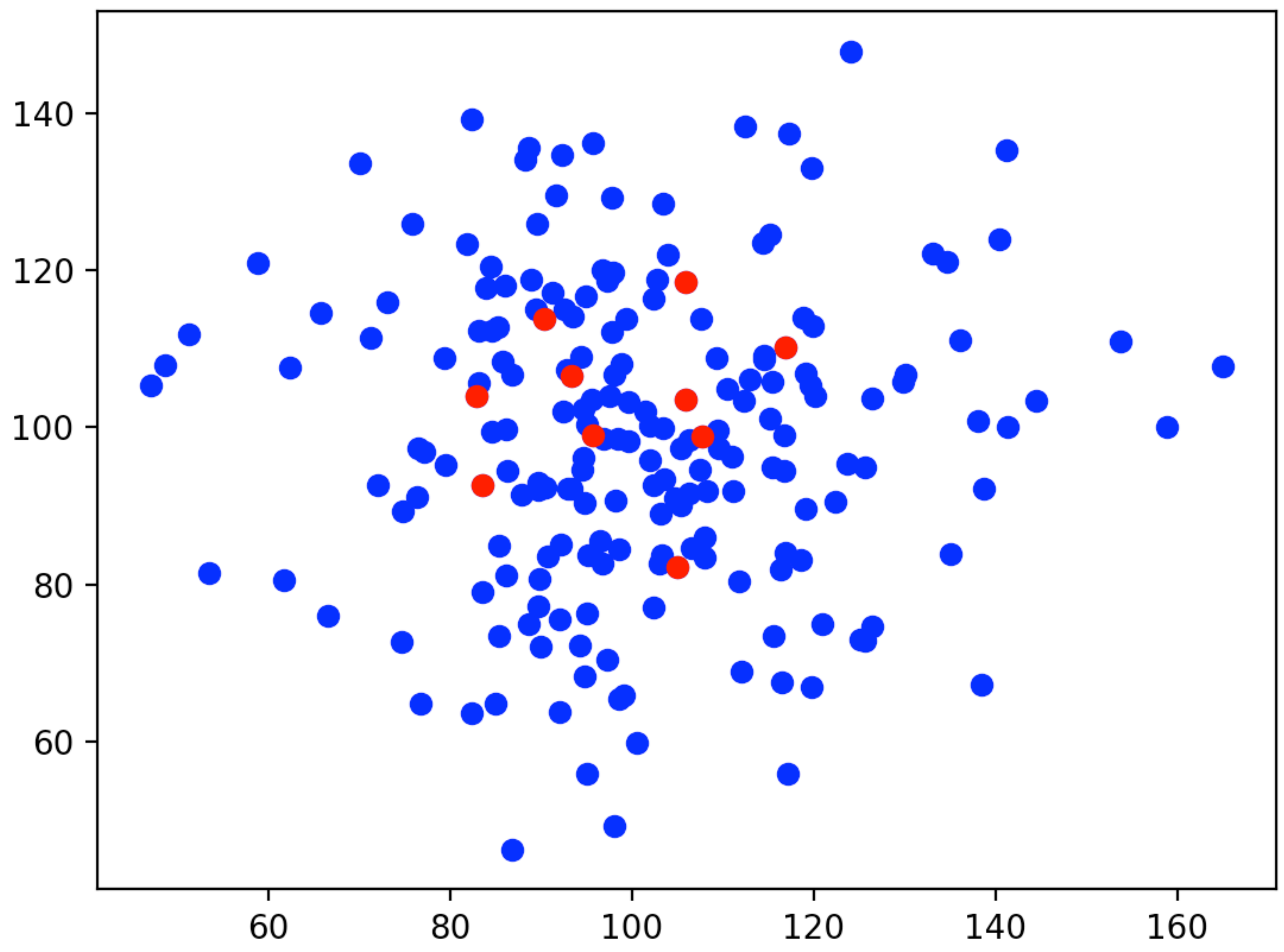
- Use fancy indexing to create the subset of points

```
>>> subset = mat[indices]
```

NumPy: Fancy Indexing

```
import numpy as np
import matplotlib.pyplot as plt

mat = np.random.normal(100, 20, (200,2))
x = mat[:,0]
y = mat[:,1]

indices = np.random.choice(np.arange(0, 200),10)
subset = mat[indices]

fig = plt.figure( )
ax = fig.add_subplot(1,1,1)
ax.scatter(x,y, color='blue')
ax.scatter(subset[:,0], subset[:,1], color = 'red')
plt.show()
```

NumPy: Fancy Indexing



Iris data set

- Remember the Iris data set?

```
def change(x):
    x=x.strip()
    if x==b'Iris-setosa':
        return 0.0
    if x==b'Iris-versicolor':
        return 1.0
    return 2.0

convert = {5: change }
with open('Iris.csv') as infile:
    iris = np.genfromtxt(infile,
                         usecols=(1,2,3,4,5),
                         dtype = float,
                         delimiter = ',',
                         converters = convert,
                         skip_header=1)
```

Iris data set

- We normalize it by dividing all but the last column by the maximum
    - This is taking the maximum along axis 0

```
total = np.max(iris[:,0:4], axis=0)
iris[:,0:4] = iris[:,0:4]/total
```

Simple Stats

- Calculate average along of all values

```
>>> np.mean(iris[:,0:4])
0.6144885806329 67
```

- Much more important: calculate average **along an axis**

```
>>> np.mean(iris[:,0:4], axis=0)
array([0.73966245, 0.69409091, 0.5447343 ,
0.47946667])
```

Simple Stats

- Simularly: np.min, np.max, np.median
  - With version in case nan (not a value) is present
- Example: Normalizing the iris data set
-

```
def normalize(array):
    maxs = np.max(array, axis = 0)
    mins = np.min(array, axis = 0)
    return (array-mins)/(maxs-mins)


iris[:,0:4] = normalize(iris[:,0:4])
```

Simple Stats

- Or normalize to have mean 0 and standard deviation 1

```
def normalizeS(array):
    means = np.mean(array, axis = 0)
    stdevs = np.std(array, axis = 0)
    return (array - means)/stdevs
```

Simple Stats

- Can determine percentiles and quantiles

```
>>> np.percentile(iris[:,0:4], 5, axis=0)
array([0.08333333, 0.14375   , 0.05084746, 0.04166667])
>>> np.percentile(iris[:,0:4], 95, axis=0)
array([0.82083333, 0.75      , 0.86440678, 0.91666667])
```

Broadcast Application

- Getting the difference matrix of a vector
  $(v_0, \quad v_1, \quad , \ldots, \quad v_{n-1})$

$$\begin{pmatrix} v_0 - v_0 & v_0 - v_1 & \cdots & v_0 - v_{n-1} \\ v_1 - v_0 & v_1 - v_1 & \cdots & v_1 - v_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n-1} - v_0 & v_{n-1} - v_1 & \cdots & v_{n-1} - v_{n-1} \end{pmatrix}$$

Broadcast Application

- Because of broadcast rules, this will not work

```
>>> v = np.array([1,2,3,4,5,6,7])
>>> v - v.T
array([0, 0, 0, 0, 0, 0, 0])
```

Broadcast Application

- But we can embed the vector into a two-dimensional vector in two different ways

```
>>> v[None,:]
array([[1, 2, 3, 4, 5, 6, 7]])
>>> v[:,None]
array([[1],
       [2],
       [3],
       [4],
       [5],
       [6],
       [7]])
```

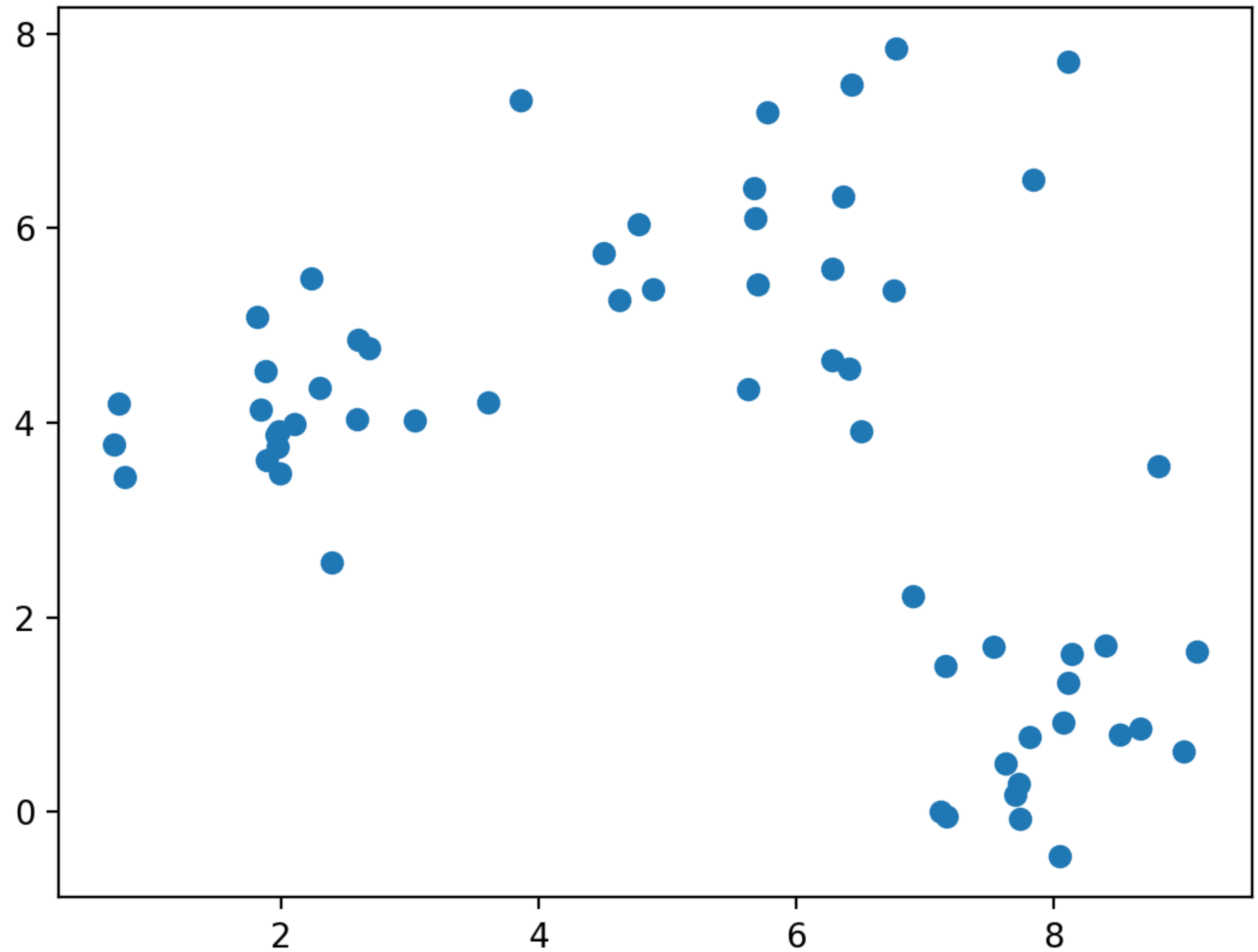
Broadcast Application

- Now we can use broadcasting

```
>>> v[:,None]-v[None,:]
array([[ 0, -1, -2, -3, -4, -5, -6],
       [ 1,  0, -1, -2, -3, -4, -5],
       [ 2,  1,  0, -1, -2, -3, -4],
       [ 3,  2,  1,  0, -1, -2, -3],
       [ 4,  3,  2,  1,  0, -1, -2],
       [ 5,  4,  3,  2,  1,  0, -1],
       [ 6,  5,  4,  3,  2,  1,  0]])
```

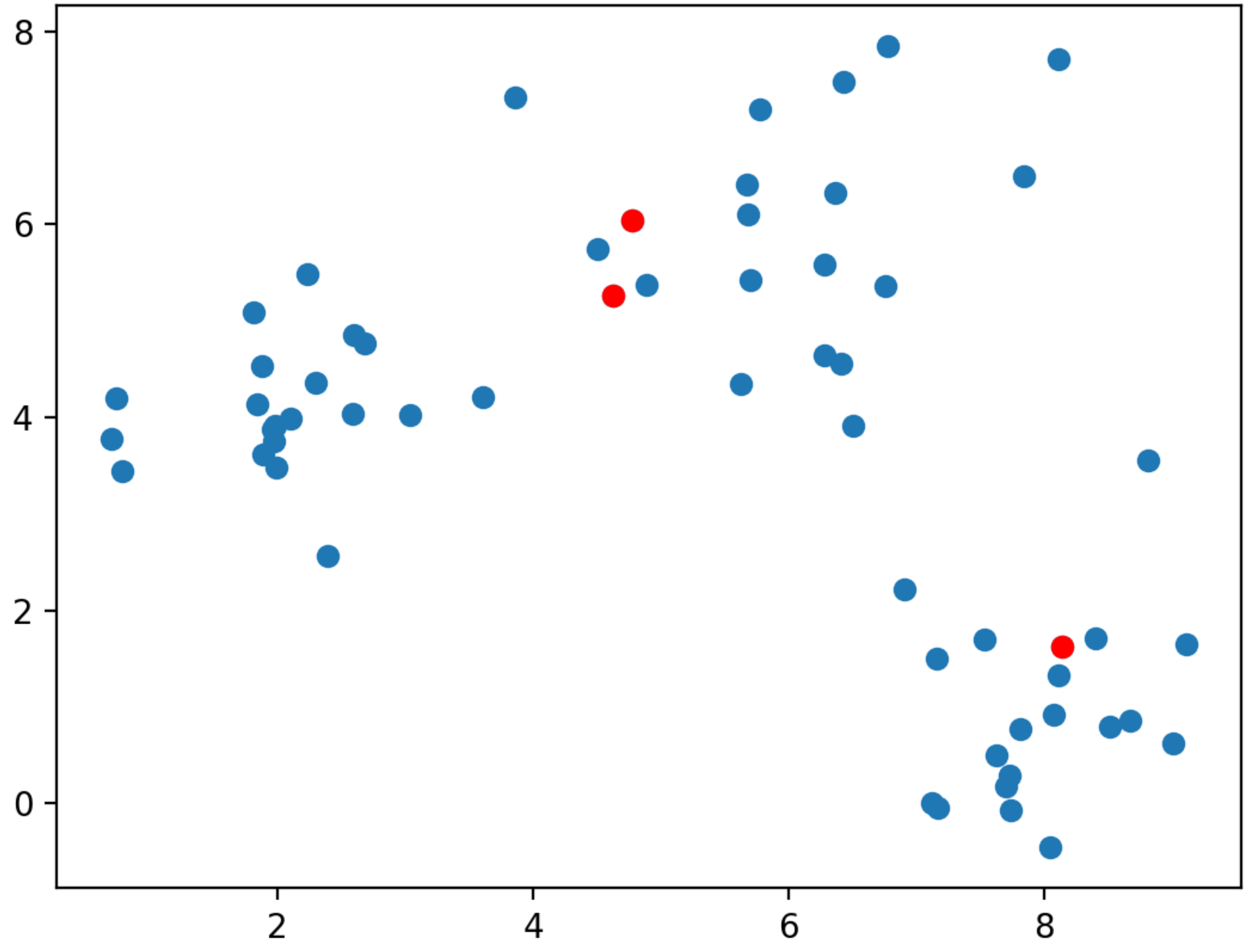
k-means clustering

- Given a set of data, can we cluster it even if we do not know its structure?



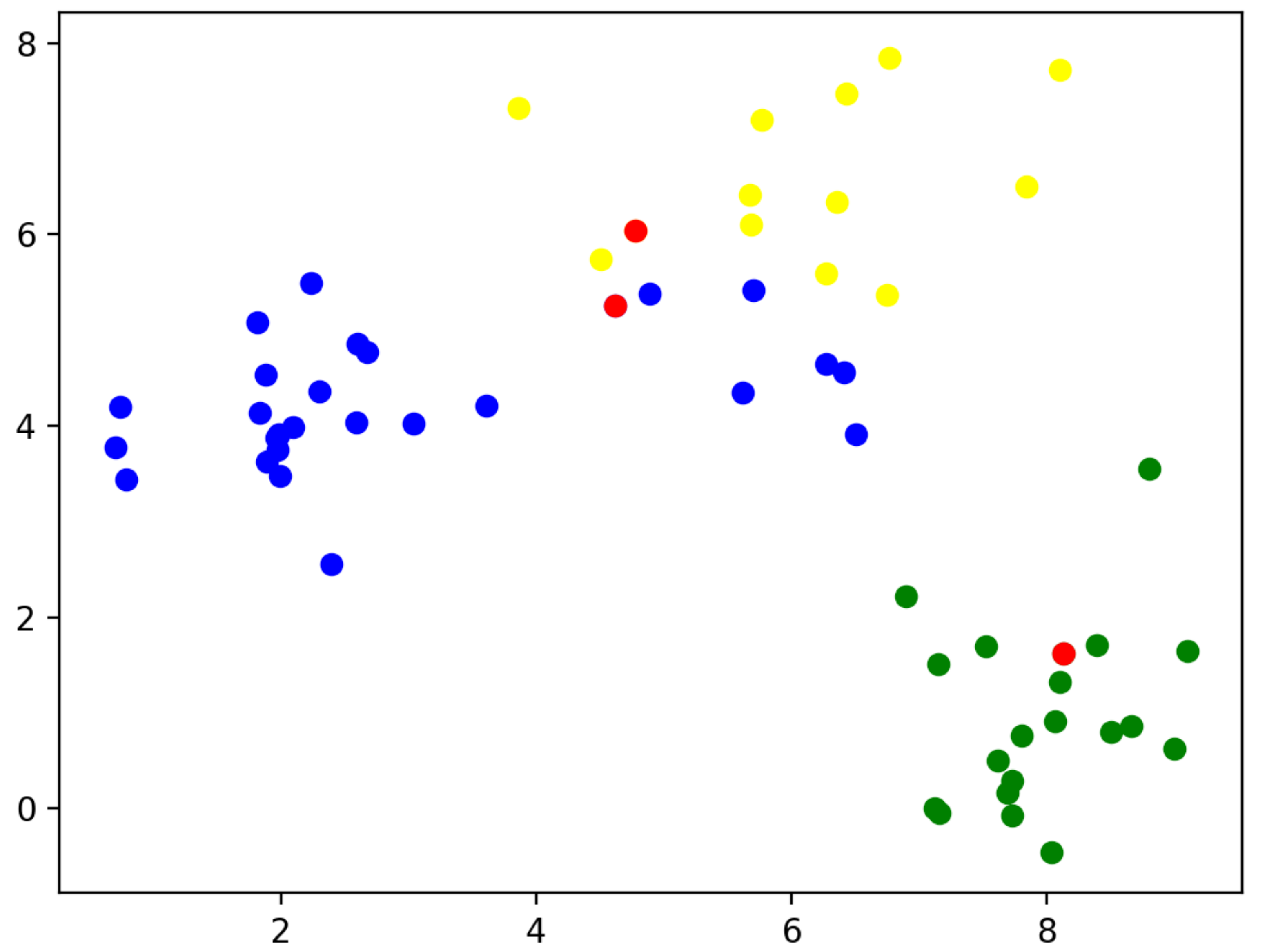
k-means clustering

- Guess a number of clusters and pick *k* arbitrary points



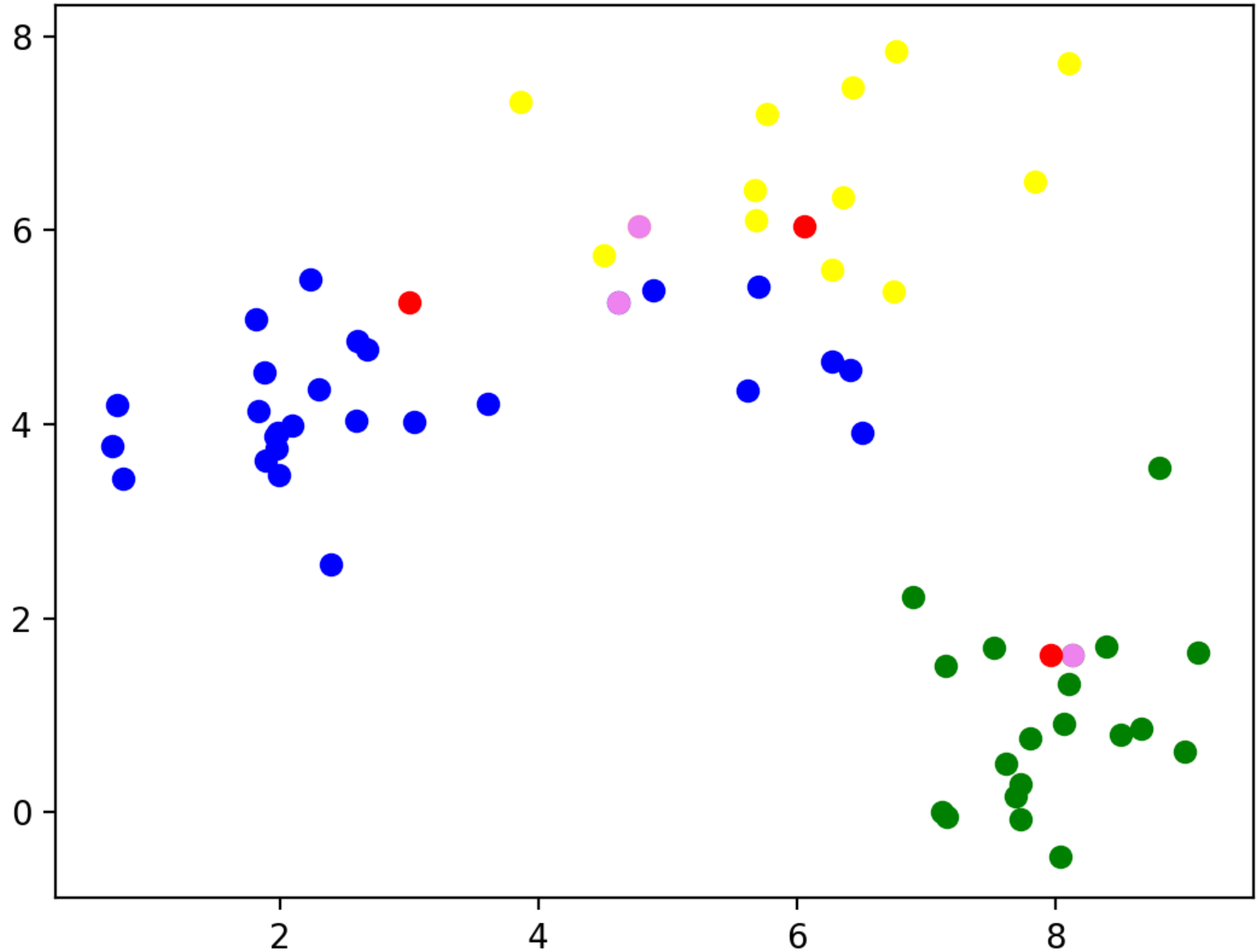
k-means clustering

- Classify all points according to which of the points they are closest



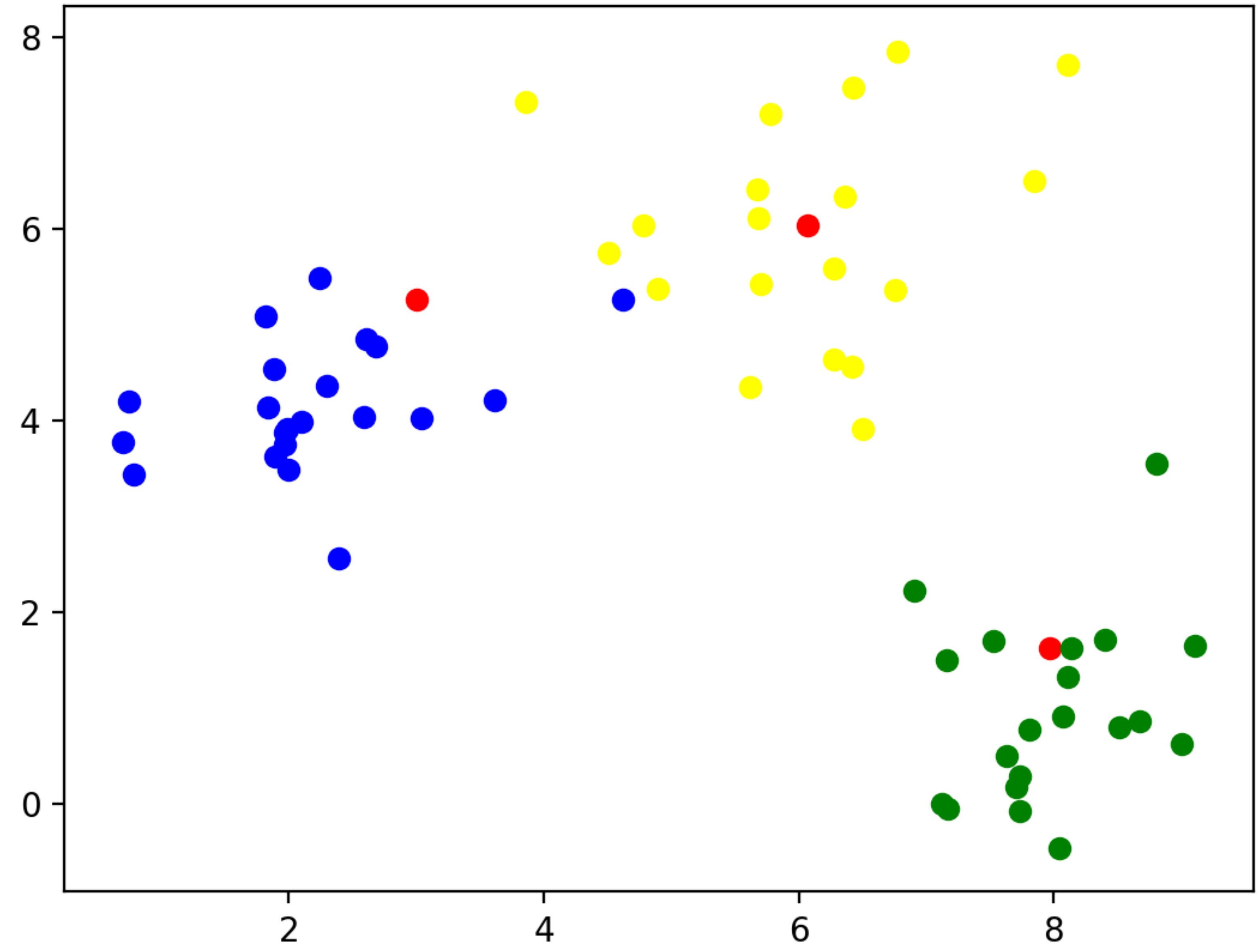
k-means clustering

- Calculate the mean of all the data points and set it as the new center



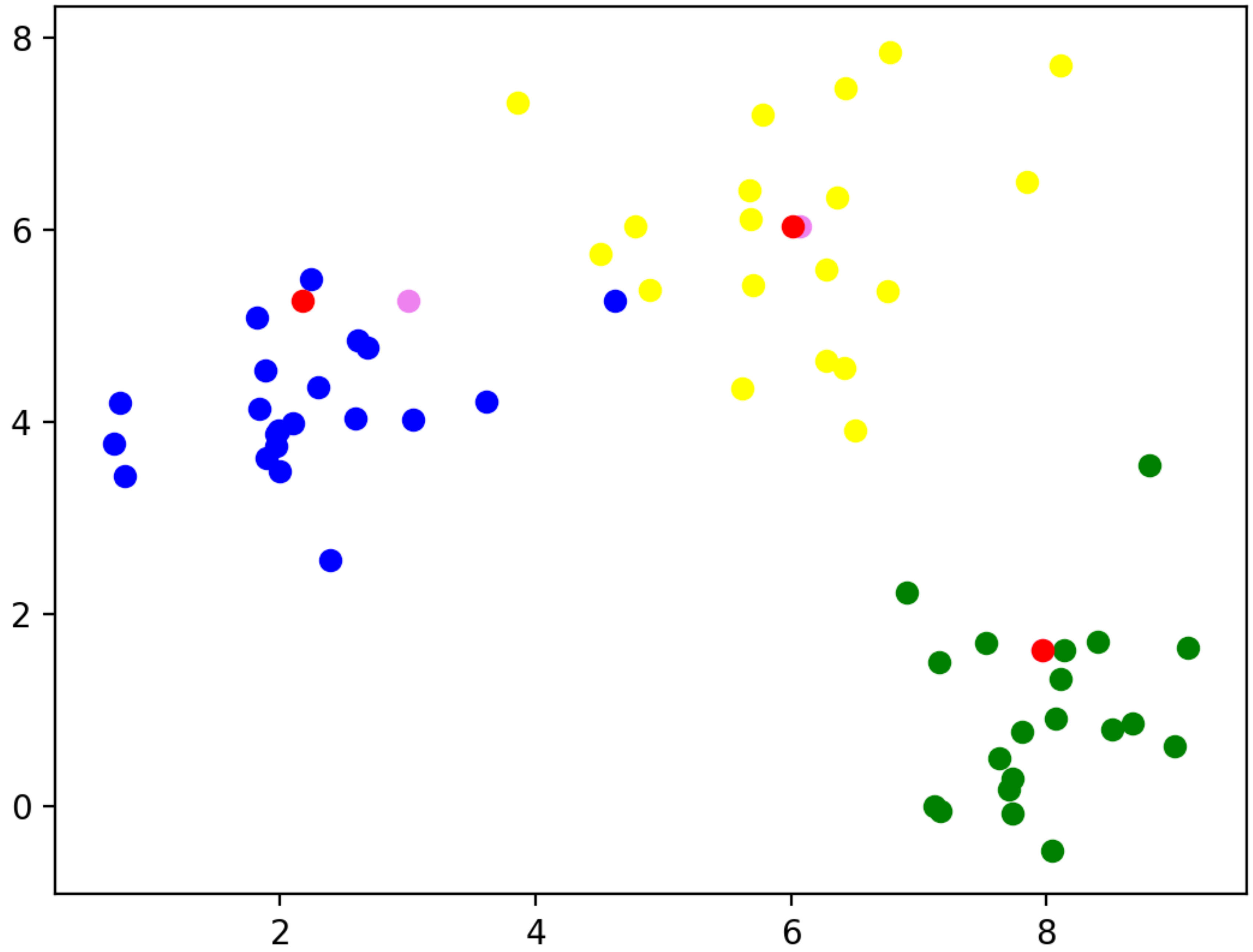
k-means clustering

- Reclassify all the points according to their closeness to the new centers



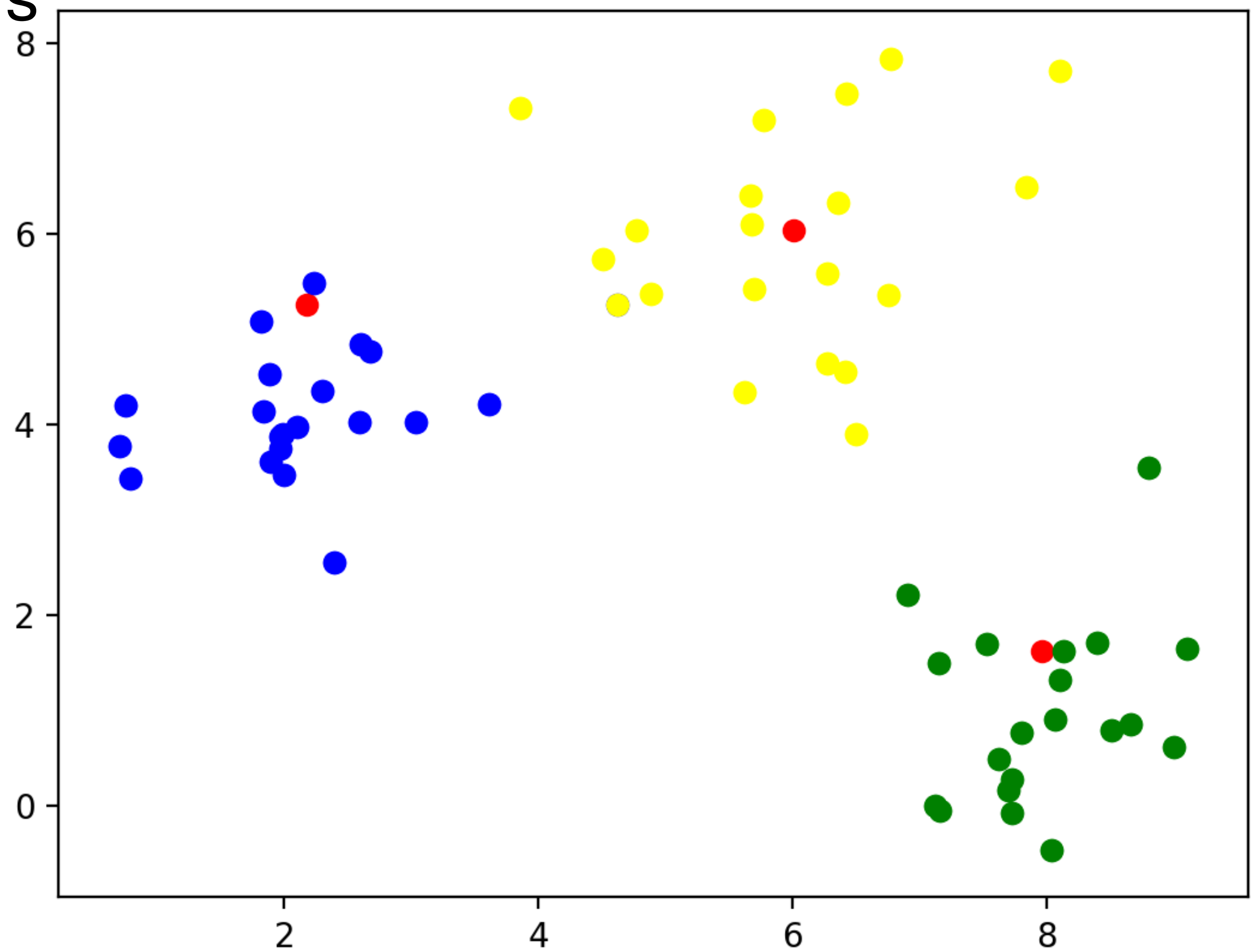
k-means clustering

- Now calculate the new centers of the groups



k-means clustering

- Repeat: Classify according to closeness to the new centers



k-means clustering

- Continue
  - The centers no longer move when points are no longer moved between different categories

k-means clustering

- Implementation
  - Find starting points by random selection

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) for j in 
range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

k-means clustering

- Enter a limited loop:

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) for j in
range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- Use the previous trick to calculate the difference between all points and the centers

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) for j in 
range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- For each point, find the closest distance

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k,
replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] -
centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0)
for j in range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- The new centers are obtained by taking the mean of the points with a given classification

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, 
replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - 
centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) 
for j in range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- If the centers do not move, we are done

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, 
replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - 
centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) 
for j in range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- Possible to not have convergence
    - For production quality code: consider raising an exception

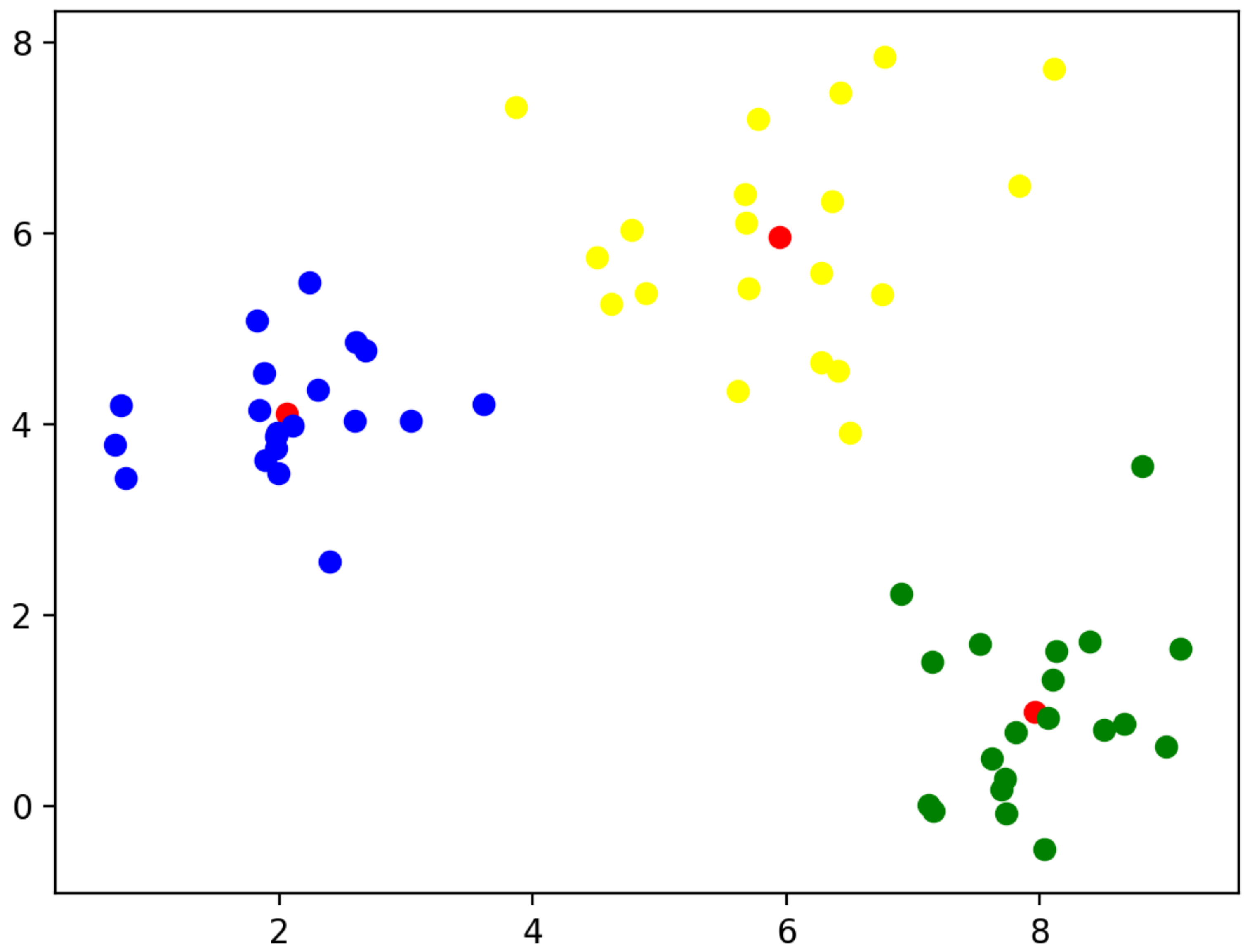
```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k,
replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] -
centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0)
for j in range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

- The loop stabilized, we are done

```
def cluster(data, k, limit):
    centers = data[ np.random.choice(np.arange(data.shape[0]), k, 
replace=False),: ]
    for _ in range(limit):
        distances = ((data[:,:,None] - 
centers.T[None, :, :])**2).sum(axis=1)
        classification = np.argmin( distances, axis=1)
        new_centers = np.array([data[classification==j,:].mean(axis=0) 
for j in range(k)])
        if np.max(np.abs(new_centers - centers)) < 0.01:
            break
        else:
            centers = new_centers
    else:  #loop did not end
        print('No convergence')
    return centers
```

k-means clustering

- Final result

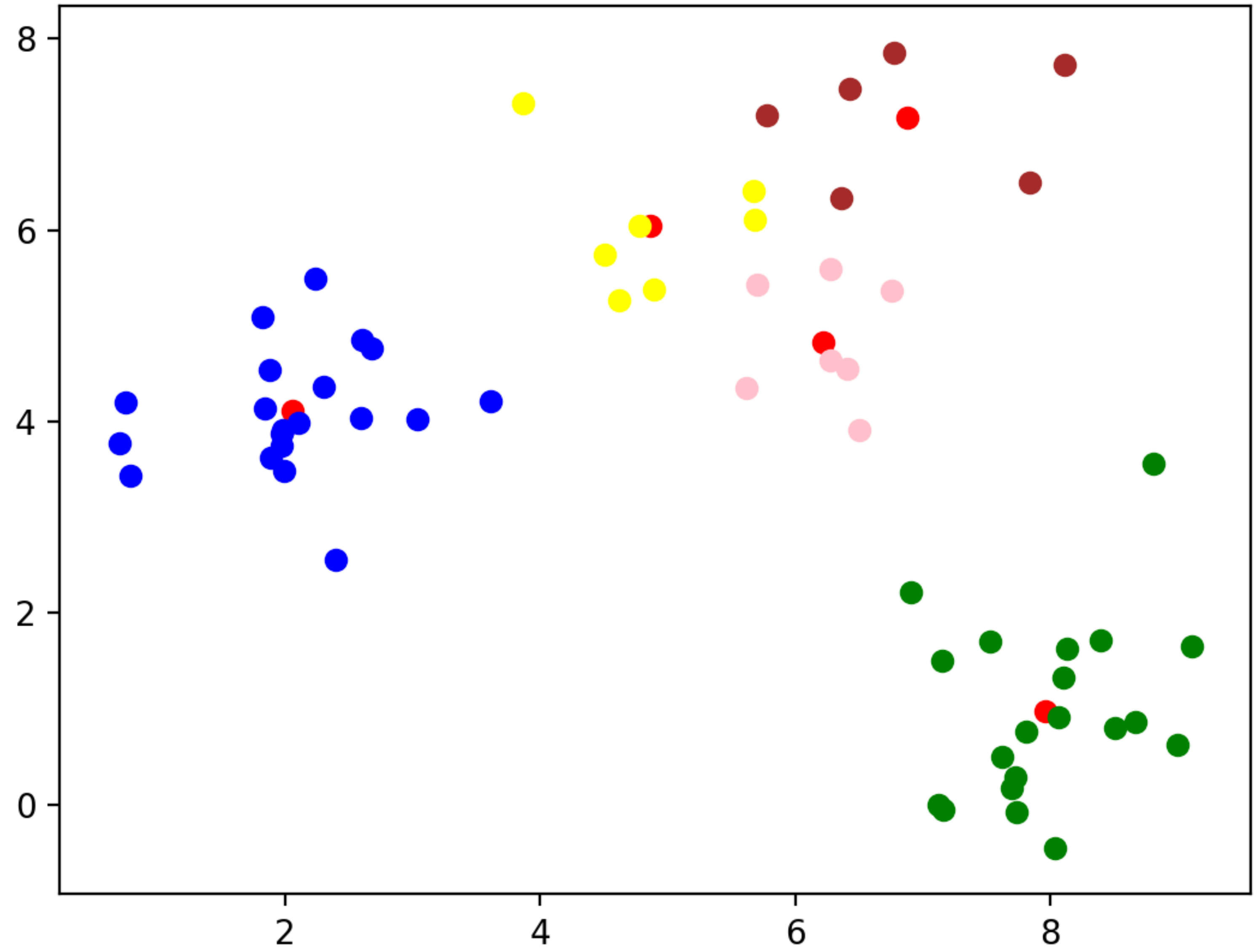


k-means clustering

- This worked because I used normalvariate to generate points around (2,4), (8,1), and (6,6)

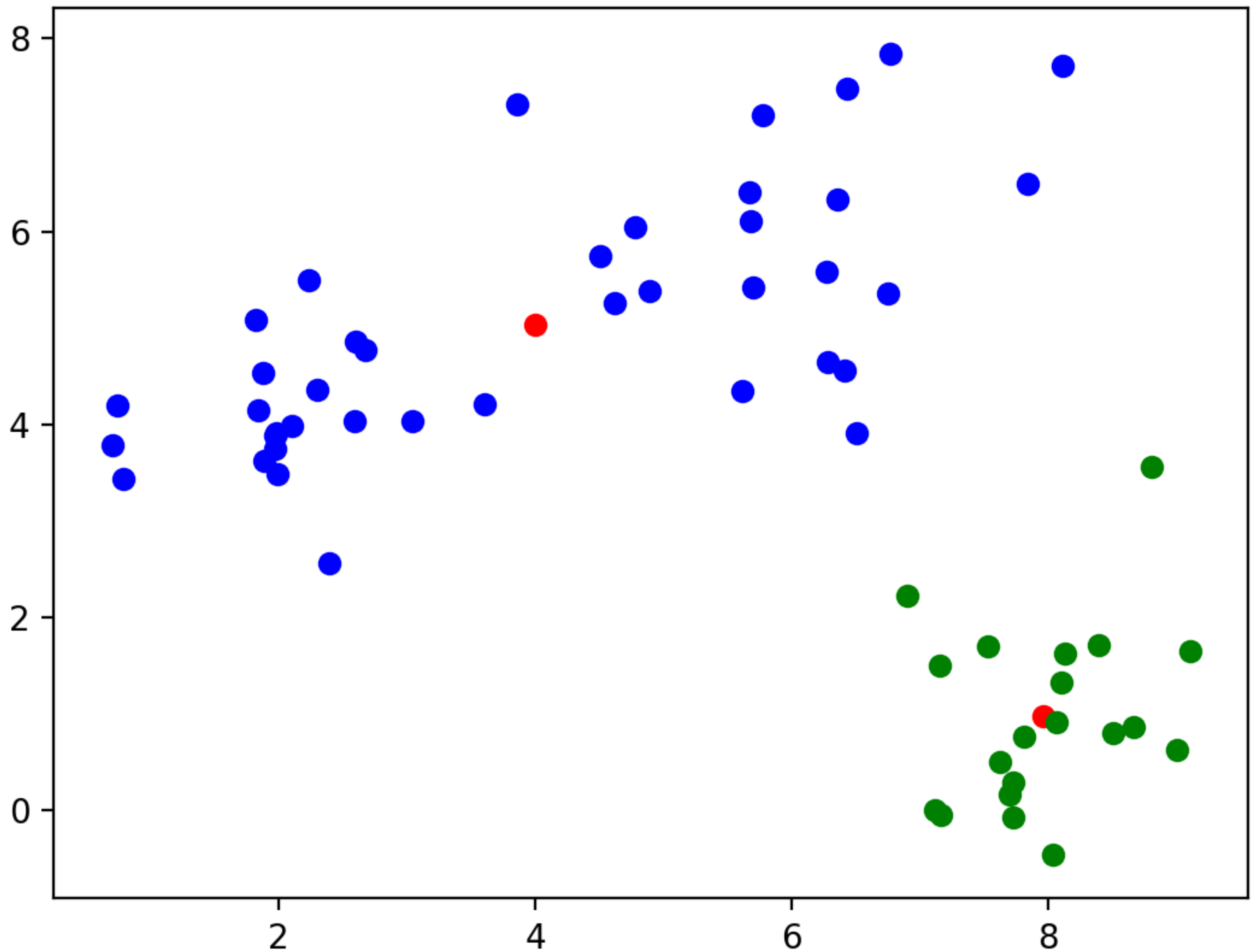
k-means clustering

- What happens if we use a different *k*?
- *k*=5: A cluster gets arbitrarily split



k-means clustering

- *k*=2 Two clusters get merged



Iris Data Set

- Let's try this out on the Iris data set
  - We only keep the measurements
  - We can normalize data using the min-max method

```
def normalize(array):
    maxs = np.max(array, axis = 0)
    mins = np.min(array, axis = 0)
    return (array-mins)/(maxs-mins)
```

Iris Data Set

- We use clustering in order to find the centers of the clusters

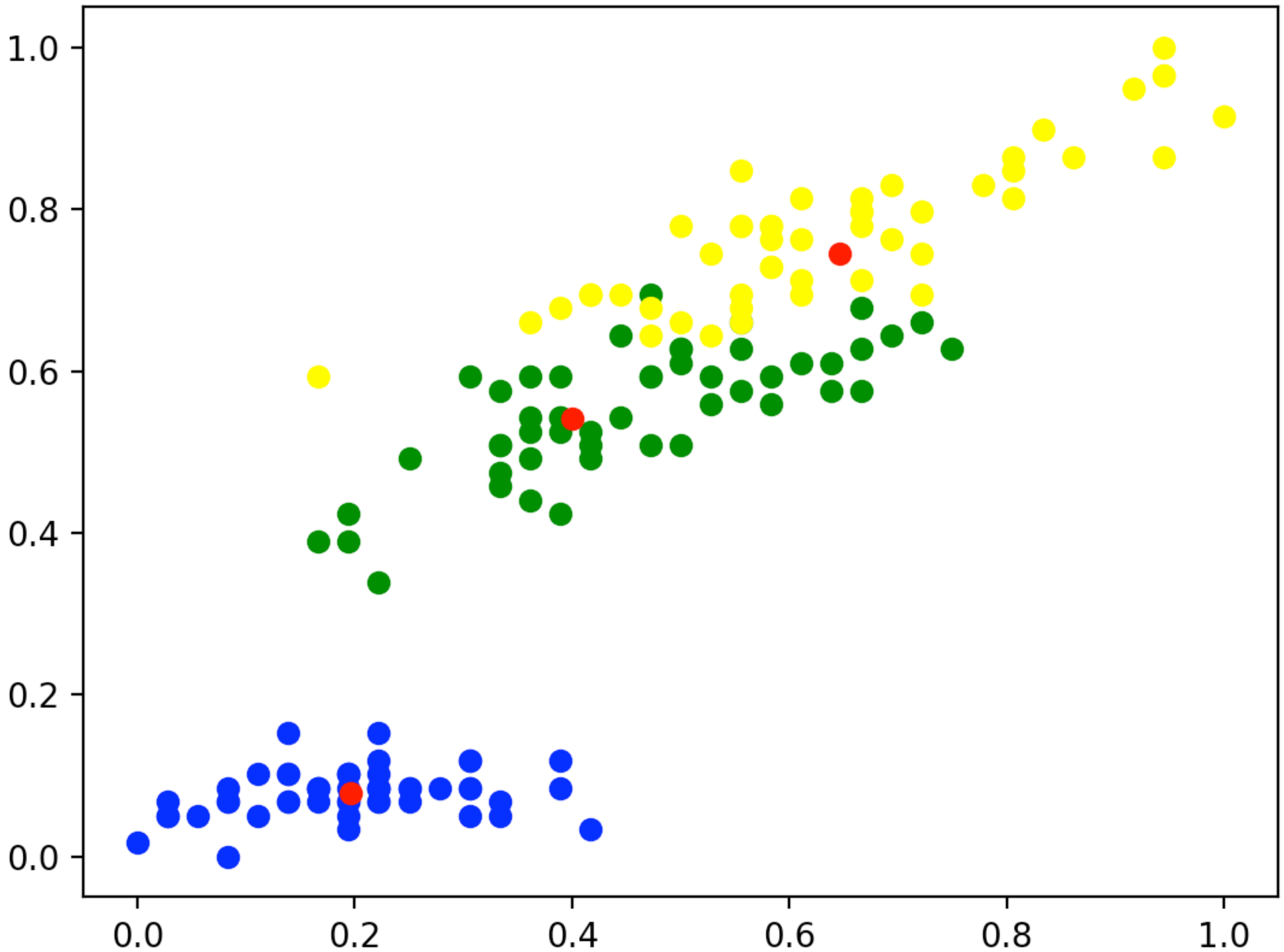
```
centers = cluster( iris[:,0:4], 3, 1000)
```

Iris Data Set

- Now we display the results
- Pick two dimensions:

```
plt.scatter(iris[0:50,0], iris[0:50,2], c='blue')
plt.scatter(iris[50:100,0], iris[50:100,2], c='green')
plt.scatter(iris[100:,0], iris[100:,2], c='yellow')
plt.scatter(centers[:,0], centers[:,2], c='red')
plt.show()
```

Iris Data Set



Iris Data Set

- To see whether classification works, we calculate the distances of all points to the three centers
-

```
data = iris[:,0:4]
distances = ((data[:,:,None] -
centers.T[None, :, :])**2).sum(axis=1)
classification = np.argmin( distances, axis=1)
print(classification)
```

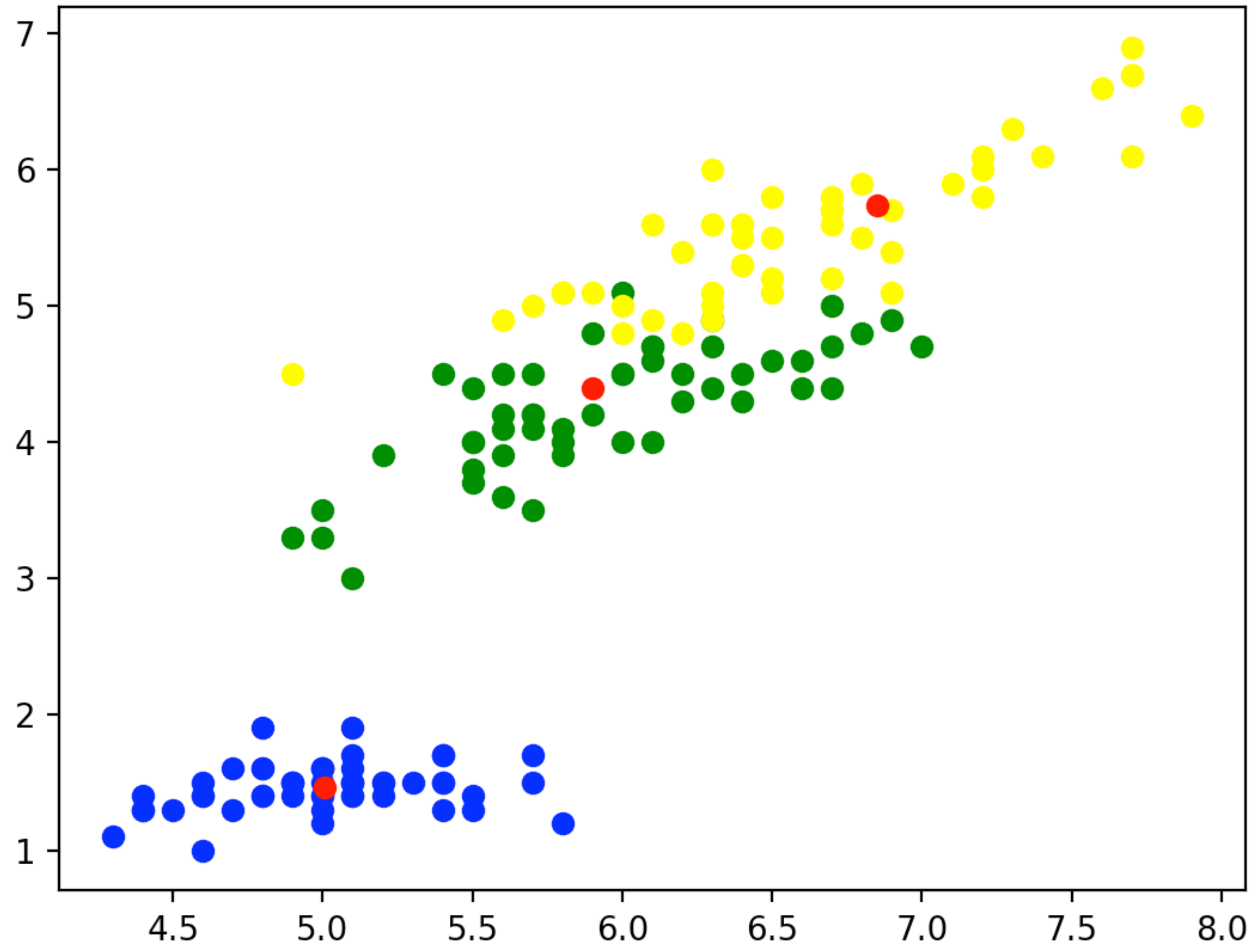
Iris Data Set

- The result is not overwhelming

```
[2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1
 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 0 1 0 0 0 0
 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 0 0
 0 0]
```

Iris Data Set

- Without normalization, we get for columns 0 and 2



Iris Data Set

- Result is still not very good

```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 0 2 2 2 2 2
 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2 2
 2 0]
```

Iris Data Set

- Summary:
  - k-means clustering is a relatively primitive way of finding clusters
  - With Iris data set:
    - The last two clusters are difficult