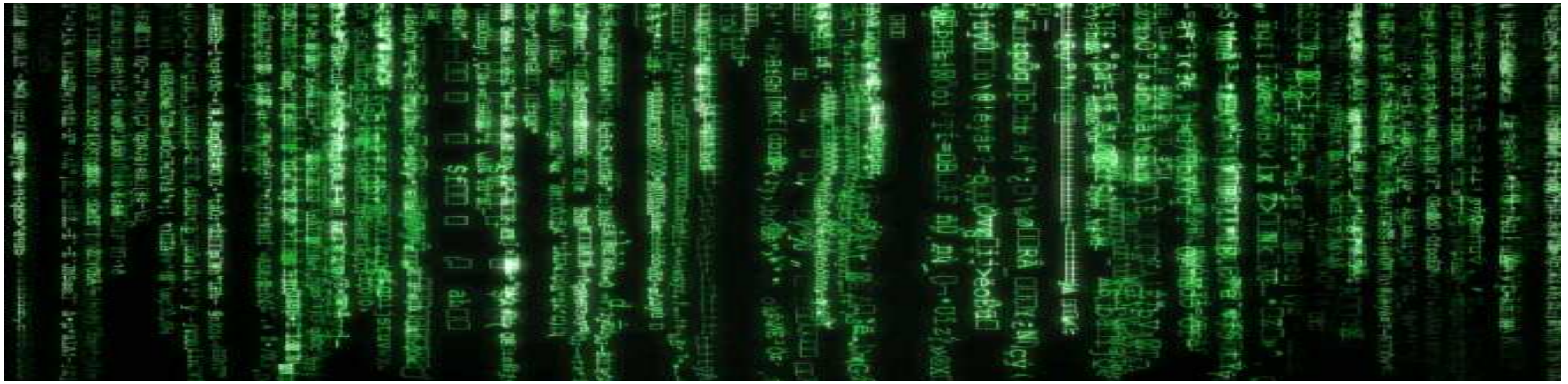

INF 110 Discovering Informatics

NumPy and Data Science

NAU
NORTHERN
ARIZONA
UNIVERSITY

A Language for Big Data

- Doing math and logic quickly is critical for big data problems
- Fortran and C have historically been the languages of choice for computationally challenging problems because they are highly optimized and work very close to the “metal”

- But Fortran and C are relatively hard languages to use – especially for non-Computer Scientists; Python, for example, is much simpler

Problem: Python is actually pretty slow..

- Python isn't as fast as languages like C or Fortran; Why?
  - Bytecode interpreted
  - Garbage collection
  - Duck typing
  - High abstraction level
  - Basically all the things that make it pleasant to use also make it slow

- Every operation is relatively expensive (time, memory)!

How slow?

- This C code took 0.109 seconds
- This Python code took 8.657 seconds
- That's 80 times slower! For complicated operations it's worse!

Code in C

```c
#include <stdio.h>
int main() {
  int i; double s=0;
  for (i=1; i<=100000000; i++) s+=i;
  printf("%.0f\n",s);
}
```

Code in Python

```python
s=0.
for i in xrange(1,100000001):
    s+=i
print s
```

Both of the codes compute the sum of integers from 1 to 100,000,000.

NumPy's Trick

- Methods are written in optimized C (or assembly)
- Methods work on lots of data at once – minimizing how much Python code gets executed

- Other languages use the same trick
  (e.g. MatLab, Mathematica, Maple)

- Python is particularly easy to extend in this way

An Example

```
x = np.arange(1, 100000001)
np.sum(x)
```

- Only a few Python methods get called.
- The 100,000,000 addition operations happen in highly optimized C code.
- That's why array values have to be the same type.
- But you have to think about problems in a certain way..

Naive Calculations

If you wanted to take each element in an array add 1 and then scale by 5 many computer scientists would write something like this:

```
for i in len(values):
    values[i] = (values[i] + 1) * 5
```

But this is SLOW!

Performing Calculations the NumPy way

Avoid using loops and instead do everything with method calls.

```
values = np.add(values, 1)
values = np.multiply(values, 5)
```

This is FAST!

NumPy Cheat Sheet

- There are **hundreds** of NumPy methods
- Only a few dozen are commonly used
- Check out the NumPy Cheat Sheet at dataquest.io

https://www.dataquest.io/blog/numpy-cheat-sheet/

NumPy – A Brief History

- Different array standards competed from 1995-2005
- NumPy 1.0 was released in 2006
- Quickly became the dominant numeric array class for Python
- Still evolving!

Some methods

- `np.prod` Multiply all elements together
- `np.sum` Add all elements together
- `np.all` Test if all elements are true values (non-zero numbers are true)
- `np.any` Test if any elements are true values (non-zero numbers are true)
- `np.count_nonzero` Count the number of non-zero elements

Some methods

- `np.add` Add a scalar or array to an array
- `np.subtract` Subtract a scalar or array from an array
- `np.multiply` Multiply a scalar or array by an array
- `np.divide` Divide a scalar or array by an array

Live Code It's Cold Outside

Task: Change an array of Celsius temperature values to Fahrenheit. Hint: $T_{(°F)} = T_{(°C)} \times 9/5 + 32$

Learning Outcomes

- Creating NumPy arrays
- Looking up NumPy documentation in Jupyter
- Adding and multiplying values in NumPy

Live Code It's Cold Outside Redux

Task: Do the same thing but with a table! Add a column with your temperature conversion

Learning Outcomes

- Loading tables
- Sorting columns
- Adding new columns


end