Numpy

Oliver W. Layton

CS251/2: Data analysis and visualization

Lecture 3, Spring 2021

Monday February 15

What is Numpy?

- Numpy is a library that allows us to perform computations quickly on lots of data with little code.
- Virtually all data science work related to Python uses Numpy.
- Numpy supports one main data structure: **ndarray** (any dimensional array).
  - e.g. `[1, 2, 3]` (vector) or `[[1,2], [3,4], [5,6]]` (matrix) or `[[[1,2], [3,4]], [[5,6], [7,8]], ...]` (3D array), etc
- Numpy ndarrays work a bit like Python lists, but using Numpy is **MUCH** more efficient for storing and performing computations on data.

What makes Numpy more efficient than Python lists?

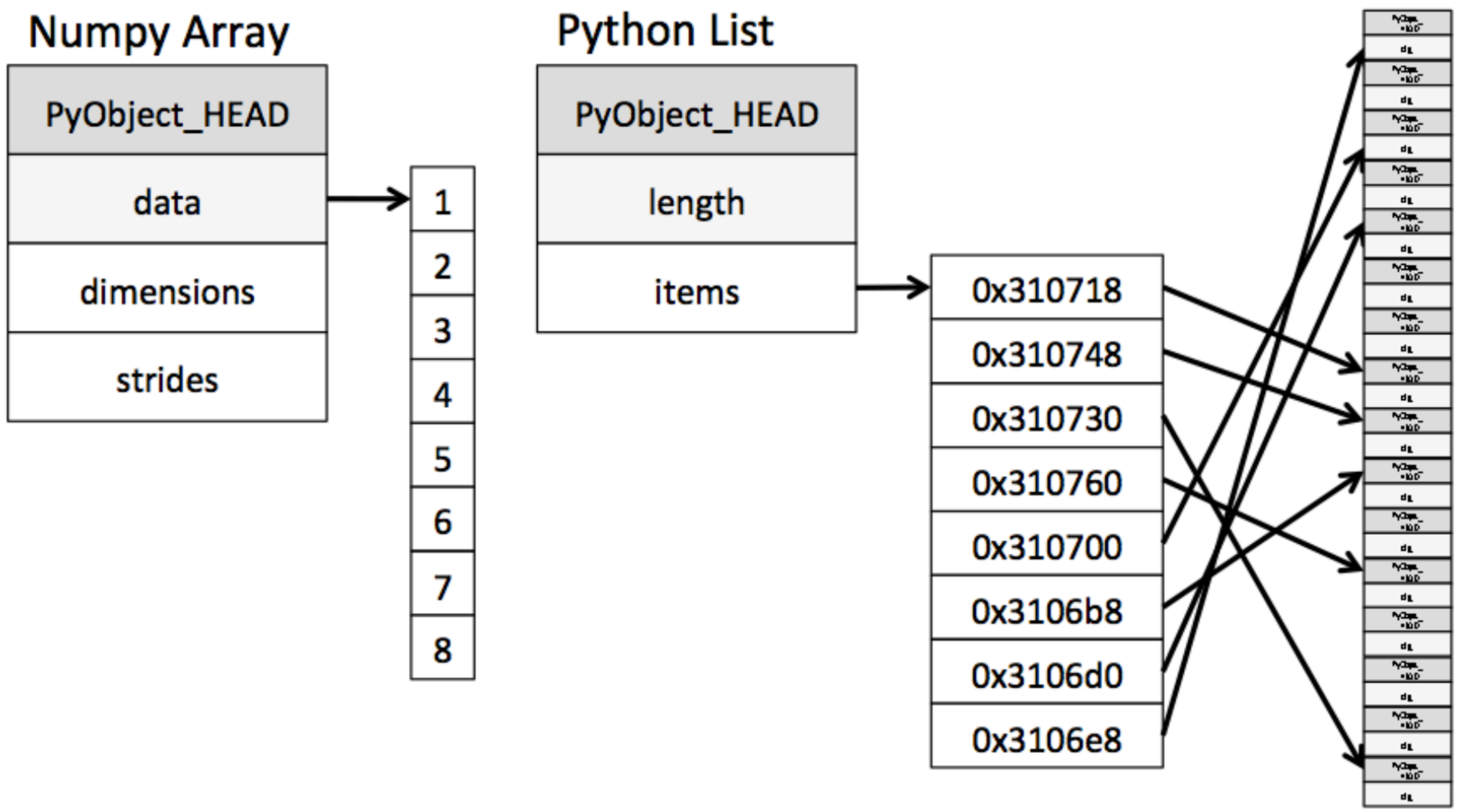
A lot of Python is written in C. Python stores much more in memory than an a single int with a simple assignment like `x = 1000`. In the underlying C, this is (`struct` is like a baby class):

```
struct _longobject
{
    long ob_refcnt;
    PyTypeObject *ob_type;
    size_t ob_size;
    long ob_digit[1];
};
```

- In C, an int assignment like `x = 1000` is literally just 4 bytes stored in memory...no overhead. The above is the cost of Python's dynamic typing.

Numpy vs. Python lists

- Numpy arrays are contiguous blocks of memory (like several ints in C chained together).
- Python lists hold many references to the struct objects, which is a collection of references to other data (VanderPlas, 2016).



Let's spend the rest of our time diving into Numpy!