Data Structures @ PrincetonPy!

February 2025 Meeting

collections.deque

double **e**nded **que**ue - great for insert/remove but slow for accessing a given element

collections.defaultdict

Can set a default value that is set whenever a previously nonexistent element is accessed (in an ordinary dictionary this would be an error)

Valid types: int, float, list, set, dict, a function, a lambda, a callable class…

collections.Counter

Quick and easy counts and categorizations

heapq (vs numpy.sort or pandas.largest)

This can be handy because when a heap is created it is done with O(n) complexity and you can then easily obtain the smallest or largest element

In fact, you can actually use a heap sort algorithm with np.sort too (kind can be “quicksort”, “mergesort”, “heapsort”, and “stable”)

…but I don’t want to “rehash” my data structures course in undergrad 😉

bisect (vs numpy.searchsorted)

This one is a tool more than a data structure

Can be faster for small to medium datasets (has functions like “insort” to insert while keeping the sorting accurate and “bisect” to return the position that an element would be at if added)

dataclasses.dataclass

You could make your own dataclass instead of pulling in the entirety of a library like a Pandas DataFrame!

But this is tricky because this use case sits between a dictionary (JSON-like) and DataFrame (lots of out-of-the-box functionality). Probably for very specific use cases or when having a bunch of complex computed fields would be helpful.

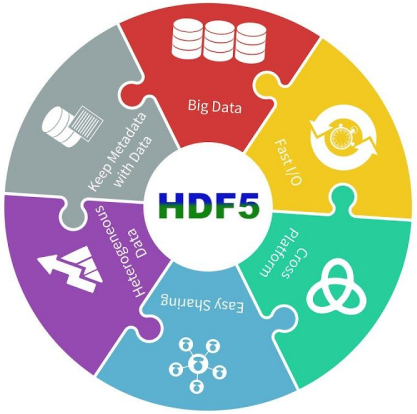
Rationals (GitHub)

As previously mentioned, Jonathan has created a rational number library that extends NumPy: https://github.com/jpalafou/rationalpy

Extending numpy with new datatype

This is probably more complicated than we have time for but we can briefly discuss the process

This can be an easy way to create labels for your data en mass without needing to restructure much



HDF5/h5py (Hierarchical Data Format)

- Easy to store (or read into) numpy arrays of most data types
- Performant and efficient binary data storage
- Parallel I/O (with MPI)
- Used frequently by hydro sims