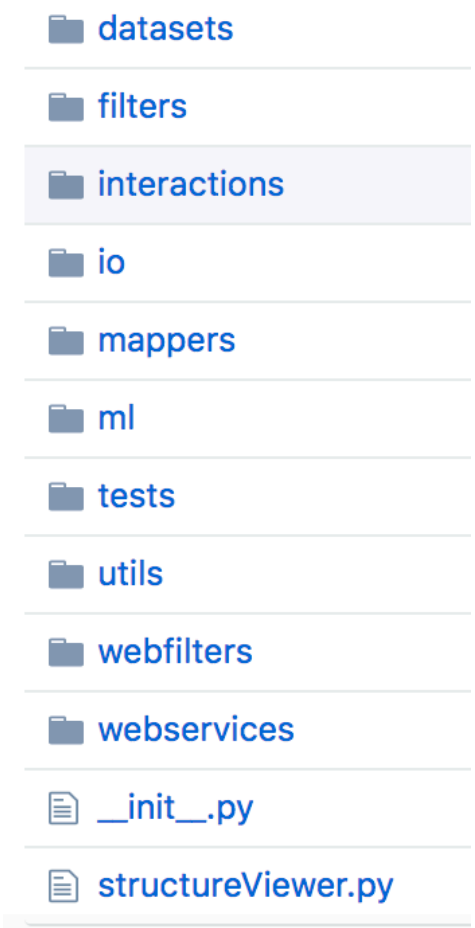
Structural Bioinformatics Training Workshop & Hackathon 2018

mmtfPyspark Datastructures

Peter Rose
Director, Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego

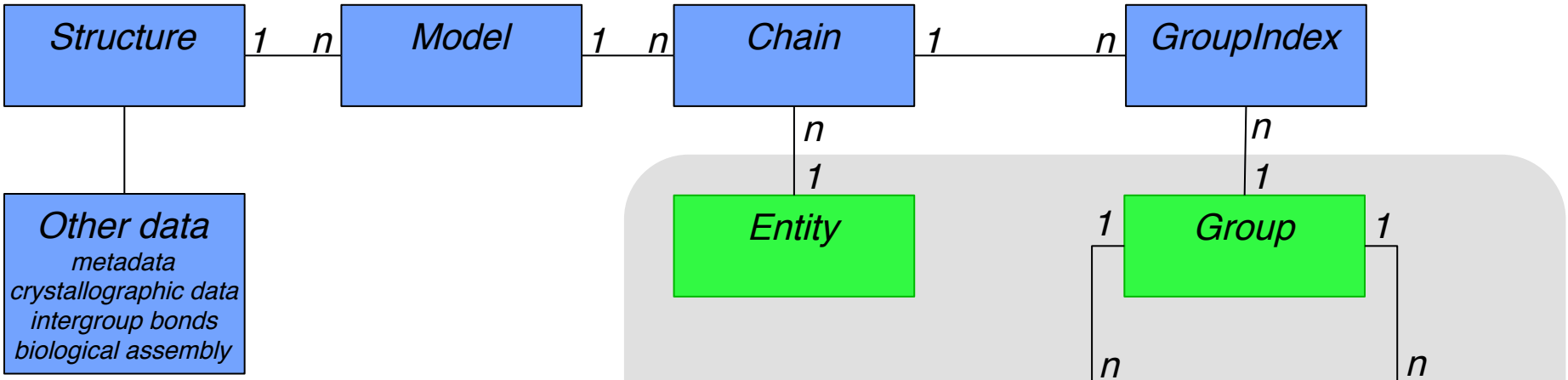
mmtfPyspark Modules Covered



- **utils**
  - Core mmtf data structures
  - Misc. utility methods

MMTF Datastructure

Flat (columnar encoded) data structure with an implicit hierarchy



Entity
A unique molecule in a structure, e.g., polymer chain, ligand, or water

Group
A unique chemical group (residue)

unique entities and groups are stored only once
e.g., 20 natural amino acids, water

Columnar Datastructure

- Atom-array-based datastructure as numpy arrays
- Enables efficient operations, including boolean indexing
- Chain and group indices are available
  - Chain to atom indices
    - Index to first atom in chain
    - Index to last atom in chain + 1
  - Chain to group indices
    - Index to first group in chain
    - Index to last group in chain + 1
  - Groups to atom indices
    - Index to first atom in group
    - Index to last atom of group + 1

Jupyter Notebook Tutorials

- 1-MMTF-Datastructure
- 2-ColumnarDatastructure
- 3-ColumnarStructureIndexing
- Problem-1 (Solution-1)
- Problem-2 (Solution-2)

Funding

This workshop was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.