Introduction to HDF5

Dr. Shelley L. Knuth
Research Computing, CU-Boulder
December 11, 2014

http://researchcomputing.github.io/meetup_fall_2014/

Download data used today from: http://neondataskills.org/HDF5/Exploring-Data-HDFView/

Download HDF5 from: http://www.hdfgroup.org/products/java/release/download.html

Outline

- What is HDF5?
- Data Model and Structure
- Example HDF5 file
- How can you view HDF5 data?
- Data subsetting
- How do you create an HDF5 file?

What is HDF5?

- Hierarchical Data Format version 5 (HDF5)
  - A set of file formats with libraries and tools for storing and managing large and complex scientific datasets
  - Supported by HDF Group
- Open source
- Can house different types of data in one HDF5 file
  - Data for different sites
  - Text and image data

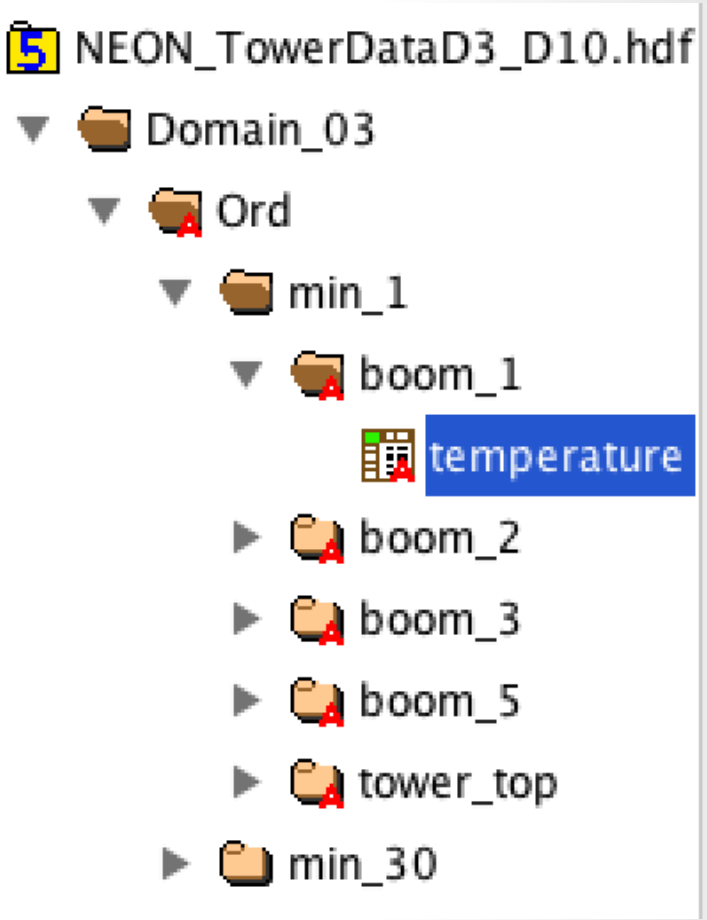
What is HDF5?

- Self-describing
  - Metadata embedded within the HDF5 file
  - Describes exactly what the data is
    - Units, location, site description, sensor information
- Files are compressed in such a way that it makes it easy to extract portions of a dataset without reading everything into memory
- Wide support by multiple languages

Hierarchical Structure

- Hierarchical structure
  - Like a directory structure you might have on your computer
  - For example, you are collecting one minute average temperature data at Site X
  - Your folder structure might look like:

    Site X → Temperature → 1-Min-Avg
  - This can exist in one HDF file



http://neondataskills.org/HDF5/Exploring-Data-HDFView/

Data Model and Structure

- Data model consists of two primary structures
  - Directories: "groups"
    - Provide structure to the data
    - Contains instances of zero or more groups or datasets
    - Has metadata
  - Files: "datasets"
    - Holds the actual data
    - Multi-dimensional array of data elements
    - Also has metadata
- Very similar to working with directories and files in Unix

Metadata/Attributes

- Information about your dataset
- Includes:
  - Dimensions
  - Datatype
  - How data is stored and organized
  - List of attributes

http://wenku.baidu.com/view/60ab43cb102de2bd960588c0.html

Attributes

- Attributes are something you attach to a dataset that provides extra information
  - Describes the intended use of the dataset or group
  - User defined
- Optional; can be overwritten, deleted, etc
- Example: laboratory readings collected at a constant temperature of 20C and pressure of 980 mb
- Then attributes would be:

temp=20

pressure=980

http://www.hdfgroup.org/HDF5/doc/UG/UG_frame13Attributes.html

Example HDF5 File

```
HDF5 "dset.h5" {
GROUP "/" {
  DATASET "dset" {
    DATATYPE  H5T_STD_I32BE
    DATASPACE  SIMPLE { ( 4, 6 ) / ( 4, 6 ) }
    DATA {
      1, 2, 3, 4, 5, 6,
      7, 8, 9, 10, 11, 12,
      13, 14, 15, 16, 17, 18,
      19, 20, 21, 22, 23, 24
    }
    ATTRIBUTE "Winds_ms" {
      DATATYPE  H5T_STD_I32BE
      DATASPACE  SIMPLE { ( 2 ) / ( 2 ) }
      DATA {
        20, 60
      }
    }
  }
}
}
```

Filename

Group top directory (no subgroups)

One dataset named dset

1) Portable, 32 bit big-endian integer

2) Size of data space: 4x6 matrix, one item per slot

The dataset consists of four items:

3) Data

4) Attributes
- Small dataset structured similar to dataset it describes
- Dataset consists here of 2 integers

http://beige.ucs.indiana.edu/I590/node120.html

Second Example - Groups

```
HDF5 "groups.h5" {
GROUP "/" {
  GROUP "MyGroup" {
    GROUP "Group_A" {
      DATASET "dset2" {
        DATATYPE  H5T_STD_I32BE
        DATASPACE  SIMPLE { ( 2, 10 ) / ( 2, 10 ) }
        DATA {
          1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
          1, 2, 3, 4, 5, 6, 7, 8, 9, 10
        }
      }
    }
    GROUP "Group_B" {
    }
    DATASET "dset1" {
      DATATYPE  H5T_STD_I32BE
      DATASPACE  SIMPLE { ( 3, 3 ) / ( 3, 3 ) }
      DATA {
        1, 2, 3,
        1, 2, 3,
        1, 2, 3
      }
    }
  }
}
}
```

The following groups exist in this file:

- /
- /MyGroup, which contains the dataset /MyGroup/dset1
- /MyGroup/GroupA, which contains the dataset /MyGroup/GroupA/dset2
- /MyGroup/GroupB, which is empty

http://beige.ucs.indiana.edu/I590/node120.html

Viewing the Contents of a HDF5 File

- HDFView
  - Visual tool for browsing, generating, and editing HDF5 files
  - With HDFView, you can:
    - View file hierarchy
    - Create new files
    - View and modify dataset content
    - Add, delete and modify attributes

- Several useful tools at the command line:
  - h5import: import text files into an HDF5 file without writing a program
  - h5dump: examine contents of HDF5 file and dump to ASCII text

h5dump Tool

- Can run this utility at the command line to get information about the contents of an HDF5 file
- Displays the contents as text
- Be default, displays entire contents of file
- Common flags:
  - -H
    - Displays header information only (no data)
  - -n
    - Displays list of objects in file

Displaying File Content and Structure

http://www.hdfgroup.org/HDF5/Tutor/cmdtoolview.html#h5ls

h5dump:

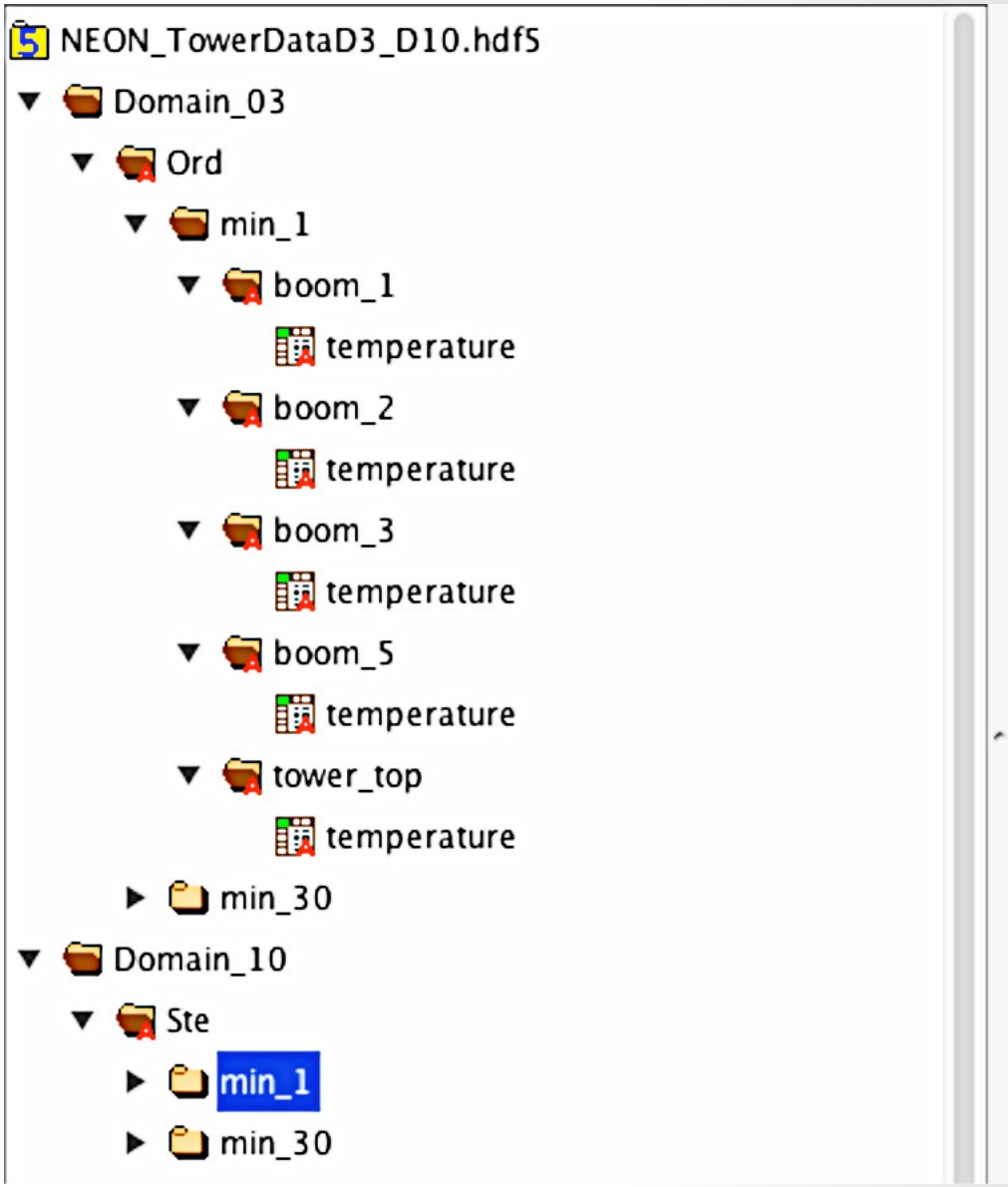
h5dump -n NEON_TowerDataD3_D10.hdf5

- Output below (subsection)

```
HDF5 "NEON_TowerDataD3_D10.hdf5" {
FILE_CONTENTS {
 group      /
 group      /Domain_03
 group      /Domain_03/Ord
 group      /Domain_03/Ord/min_1
 group      /Domain_03/Ord/min_1/boom_1
 dataset    /Domain_03/Ord/min_1/boom_1/temperature
 group      /Domain_10
 group      /Domain_10/Ste
 group      /Domain_10/Ste/min_1
 group      /Domain_10/Ste/min_1/boom_1
 dataset    /Domain_10/Ste/min_1/boom_1/temperature
}
}
```

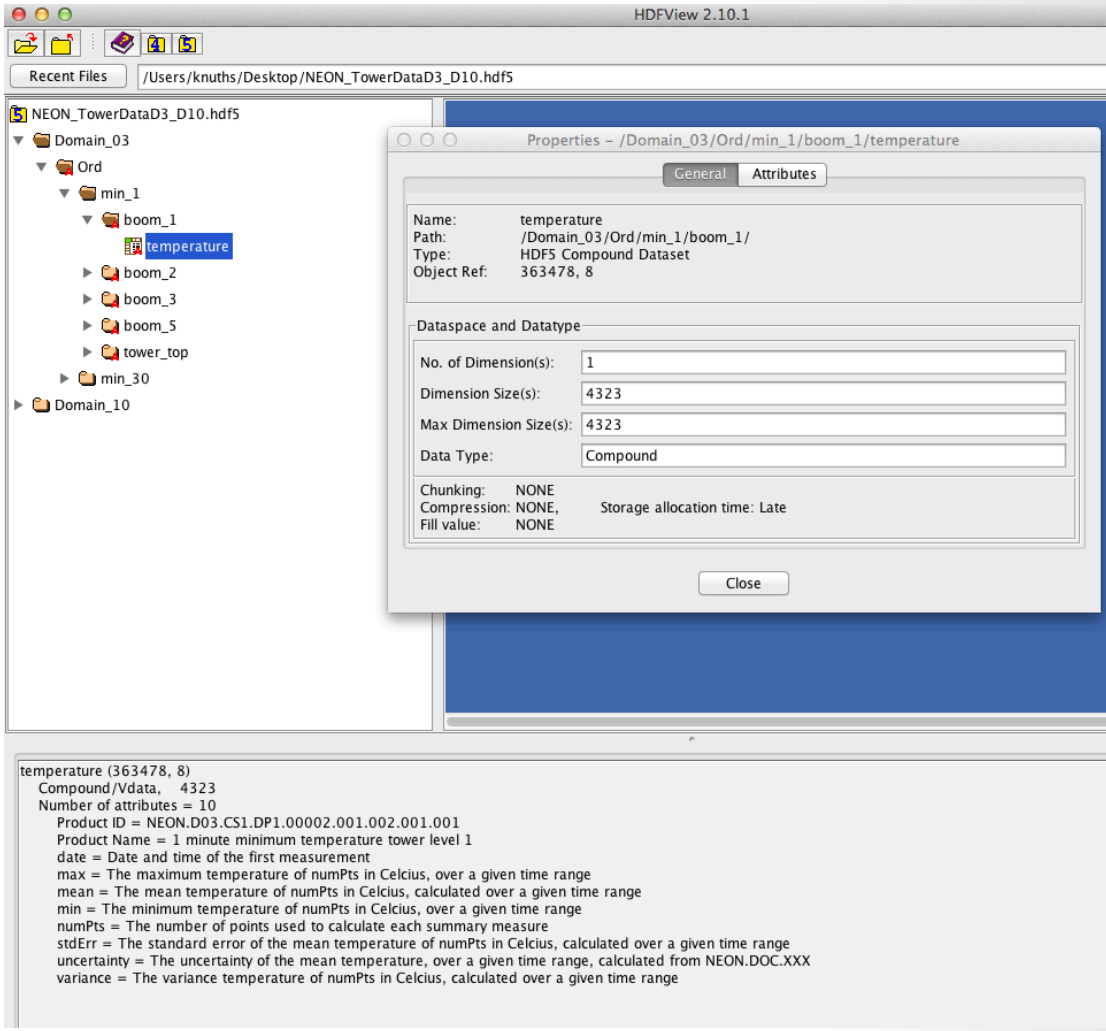
HDFView

- Displays the file structure in a series of drop down menus
- Data objects are icons
- Group objects are folders



HDFView

- Can view information regarding size, attributes, etc by clicking on each object
- Right click or just clicking on the file
  - Metadata, attributes, etc



Dataset and Dataset Properties

http://www.hdfgroup.org/HDF5/Tutor/cmdtoolview.html#h5ls

h5dump

- To view the dataset, make sure you specify the entire path to the data
- For example, */Domain_10/Ste/min_1/boom_1/temperature*
- This is because there might be multiple datasets with the same name within the tree structure

h5dump -d: View data contents

h5dump -d /Domain_10/Ste/min_1/boom_1/temperature NEON_TowerDataD3_D10.hdf5

```
HDF5 "NEON_TowerDataD3_D10.hdf5" {
DATASET "/Domain_10/Ste/min_1/boom_1/temperature" {
  DATATYPE  H5T_COMPOUND {
    H5T_STRING {
      STRSIZE 30;
      STRPAD H5T_STR_NULLPAD;
      CSET H5T_CSET_ASCII;
      CTYPE H5T_C_S1;
    } "date";
    H5T_STD_I32LE "numPts";
    H5T_IEEE_F64LE "mean";
    H5T_IEEE_F64LE "min";
    H5T_IEEE_F64LE "max";
    H5T_IEEE_F64LE "variance";
    H5T_IEEE_F64LE "stdErr";
    H5T_IEEE_F64LE "uncertainty";
  }
  DATASPACE  SIMPLE { ( 4323 ) / ( 4323 ) }
  DATA {
  (0): {
        "2014-04-01 00:00:00.0\000\000\000\000\000\000\000\000\000",
        60,
        6.72064,
        6.66785,
        6.77449,
        0.00127469,
        0.00460922,
        0.0129818
    },
```

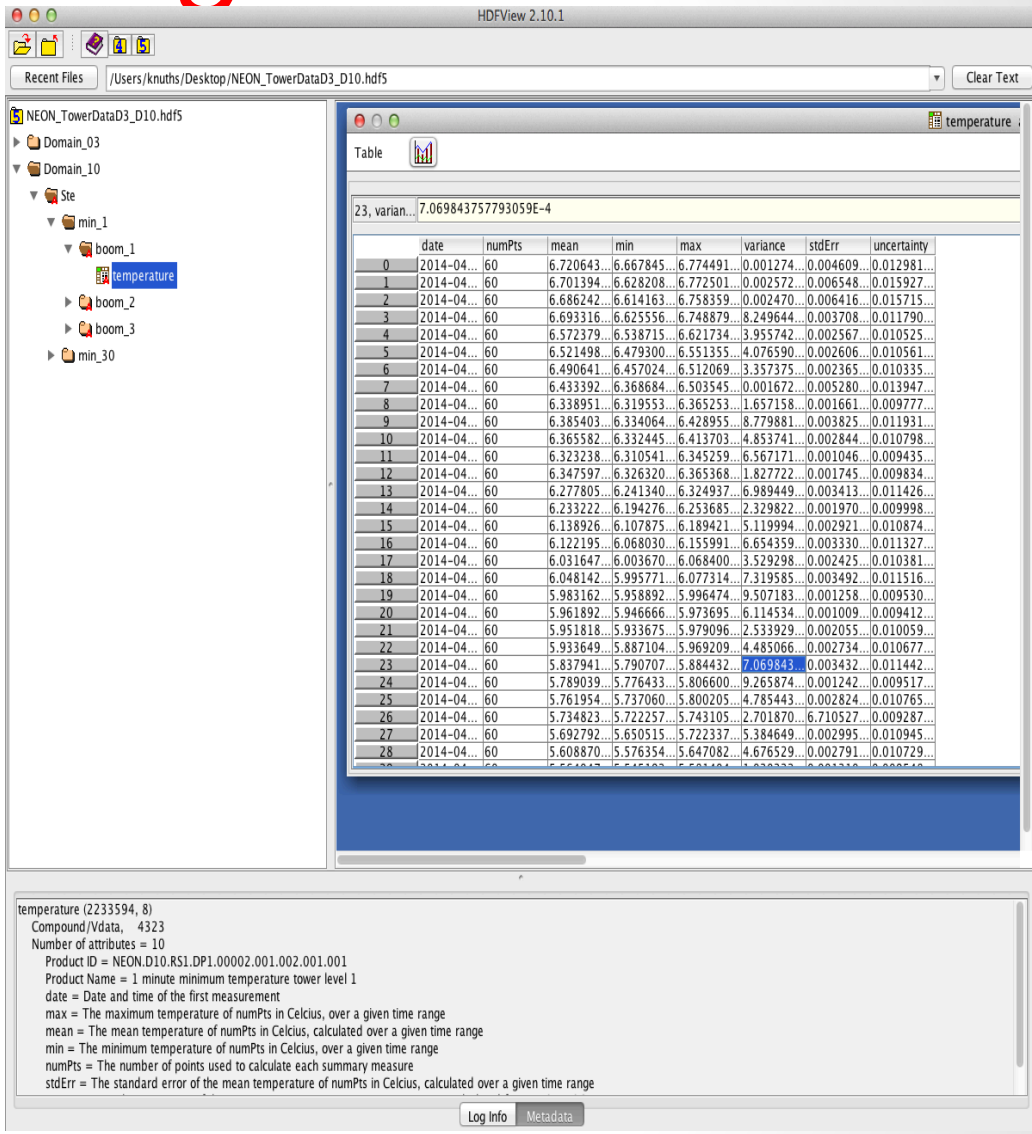
h5dump -Hd: View header information only

h5dump -Hd /Domain_10/Ste/min_1/boom_1/temperature NEON_TowerDataD3_D10.hdf5

```
HDF5 "NEON_TowerDataD3_D10.hdf5" {
DATASET "/Domain_10/Ste/min_1/boom_1/temperature" {
  DATATYPE  H5T_COMPOUND {
    H5T_STRING {
      STRSIZE 30;
      STRPAD H5T_STR_NULLPAD;
      CSET H5T_CSET_ASCII;
      CTYPE H5T_C_S1;
    } "date";
    H5T_STD_I32LE "numPts";
    H5T_IEEE_F64LE "mean";
    H5T_IEEE_F64LE "min";
    H5T_IEEE_F64LE "max";
    H5T_IEEE_F64LE "variance";
    H5T_IEEE_F64LE "stdErr";
    H5T_IEEE_F64LE "uncertainty";
  }
  DATASPACE  SIMPLE { ( 4323 ) / ( 4323 ) }
  ATTRIBUTE "Product ID" {
    DATATYPE  H5T_STRING {
      STRSIZE H5T_VARIABLE;
      STRPAD H5T_STR_NULLTERM;
      CSET H5T_CSET_ASCII;
      CTYPE H5T_C_S1;
    }
    DATASPACE  SCALAR
  }
```

HDFView – Viewing File Contents

- Can view file contents by simply double clicking on data
- Can also the data graphically by clicking on “table”



Dataset Subset

http://www.hdfgroup.org/HDF5/Tutor/cmdtoolview.html#h5ls

Subsetting Data

- With large datasets, it might be useful to only visualize part of the dataset
- You can do this with h5dump or HDView

h5dump -d: Subset Data

- Flags to use when subsetting data with h5dump:
  - -d: dataset
  - -s: start of subsetting selection (can use H,W)
  - -S: stride (default=1)
  - -c: number of blocks to include
  - -k: size of block (default=1)

h5dump -d: Subset Data

h5dump -A 0 -d /Domain_10/Ste/min_1/boom_1/temperature -s "0" -c "2" NEON_TowerDataD3_D10.hdf5

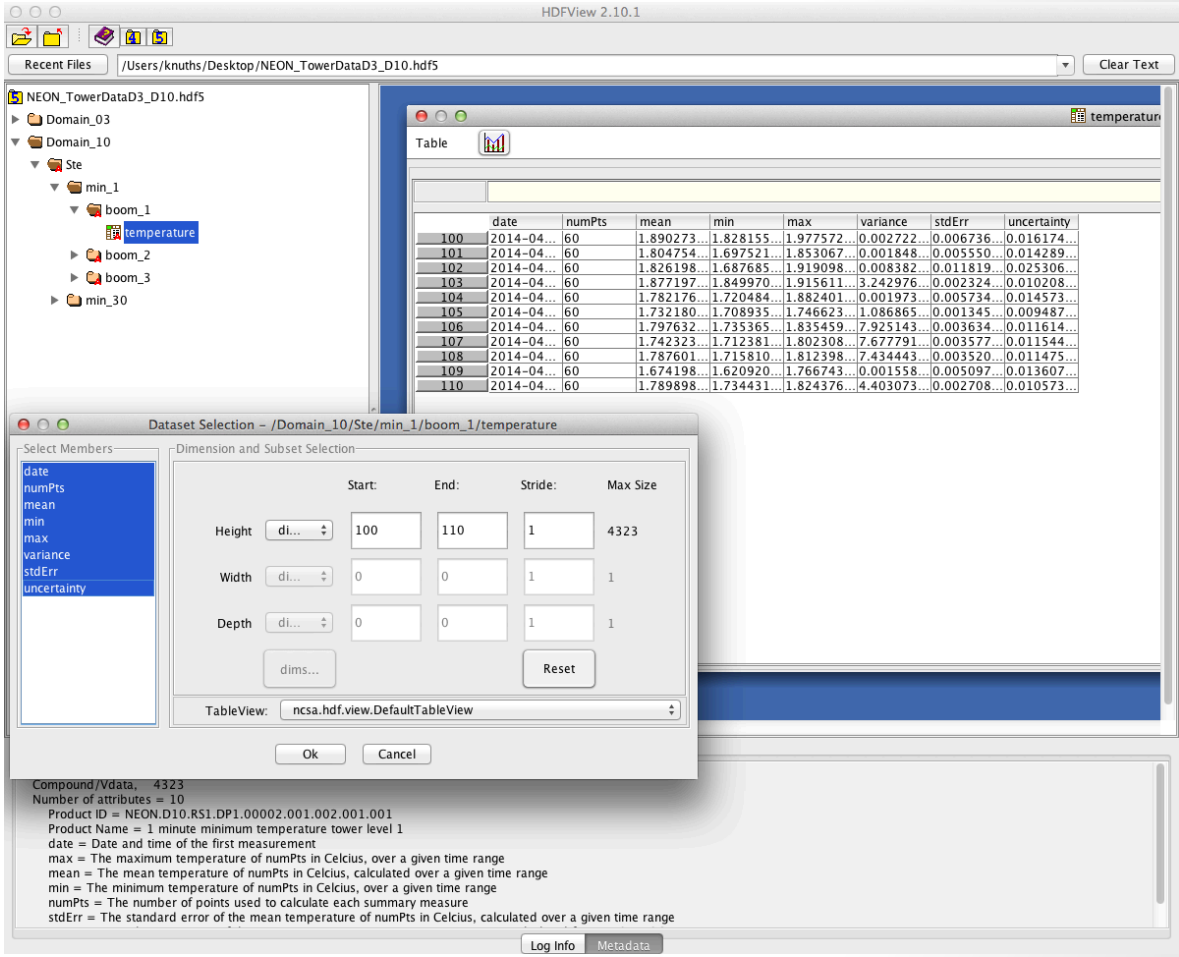
- This command says to sample the first 2 elements beginning with position 0
- The –A 0 flag simply suppresses the attribute output

Same as before, but only shows the first two data values:

```
    DATA {
    (0): {
        "2014-04-01 00:00:00.0\000\000\000\000\000\000\000\000\000",
        60,
        6.72064,
        6.66785,
        6.77449,
        0.00127469,
        0.00460922,
        0.0129818
      },
    (1): {
        "2014-04-01 00:01:00.0\000\000\000\000\000\000\000\000\000",
        60,
        6.70139,
        6.62821,
        6.7725,
        0.00257267,
        0.00654811,
        0.0159273
      }
    }
  }
}
}
```

HDFView – Subset Data

- Open a very large dataset in HDFView could cause an out of memory error
- To view a portion of the data click on the data and select “open as”
- Make selection by entering start, end, and stride



http://www.hdfgroup.org/products/java/hdfview/UsersGuide/ug05spreadsheet.html#ug05subset

Creating an HDF5 File

http://beige.ucs.indiana.edu/I590/node121.html

How to Create an HDF5 File

- You can create a new HDF5 file or convert an existing file to HDF5 file format
- Create or/convert with language/software that can work with HDF5 files
  - C, Fortran, R, Matlab, Python, Java, HDFView
  - http://www.hdfgroup.org/tools5desc.html
  - Conventions are similar in most languages

General Procedure for HDF5 File Creation

- Objects are opened or created
- Objects are accessed
- Objects are closed

- When creating an HDF5 file, must specify:
  - File name
  - File access mode (if file exists, should current contents be truncated or not allowed to be created?)
  - File creation property list (controls the file metadata – size of data structures, etc)
  - File access property list (controls I/O methods – parallel, etc)

Sample Matlab code

```
% Creating and closing a file (no data added).

% Create a new file using default properties.
  file_id = H5F.create('newfile.hdf', 'H5F_ACC_TRUNC', 'H5P_DEFAULT',
'H5P_DEFAULT');

  % Terminate access to the file.
  H5F.close(file_id);
```

Output: Creating an HDF5 File in Matlab

- Once you run the program, your file has been created
- Then if you do an h5dump, you will see:

```
HDF5 ”newfile.hdf" {
GROUP "/" {
}
}
```

- There is only a top level group

Creating a Dataset

1. Obtain location ID where dataset is to be created
   - File or group identifier
2. Define dataset characteristics
   - Datatype (integer)
     - Predefined: `H5T_IEEE_F64LE`, etc
   - Dataspace (# of dimensions, size, etc)
   - Dataset storage (chunked, compressed, etc)
3. Create the datasest
4. Close the datatype, dataspace, and property list
5. Close the dataset

Adding a Dataset to an HDF5 File

```
% Creating and closing a file.

% Create a new file using default properties.
  file_id = H5F.create('newfile.hdf', 'H5F_ACC_TRUNC', 'H5P_DEFAULT',
'H5P_DEFAULT');

% Create the dataset.
   h5create('newfile.hdf', '/mydata', [4 6]);

% Close file.
   H5F.close(file_id);
```

Output: Creating an HDF5 Dataset in Matlab

- If you do an h5dump, you will see:

```
HDF5 "newfile.hdf" {
GROUP "/" {
  DATASET "mydata" {
    DATATYPE  H5T_IEEE_F64LE
    DATASPACE  SIMPLE { ( 6, 4 ) / ( 6, 4 ) }
    DATA {
    (0,0): 0, 0, 0, 0,
    (1,0): 0, 0, 0, 0,
    (2,0): 0, 0, 0, 0,
    (3,0): 0, 0, 0, 0,
    (4,0): 0, 0, 0, 0,
    (5,0): 0, 0, 0, 0
    }
  }
}
}
```

Adding Data to a Dataset

```
% Create random data matrix
  randomData=rand(4,6)

% Write data to file
   h5write('newfile.hdf', '/mydata', randomData);
```

Output: Writing to Dataset in Matlab

- If you do an h5dump, you will see:

```
HDF5 "newfile.hdf" {
GROUP "/" {
  DATASET "mydata" {
    DATATYPE  H5T_IEEE_F64LE
    DATASPACE  SIMPLE { ( 6, 4 ) / ( 6, 4 ) }
    DATA {
    (0,0): 0.814724, 0.905792, 0.126987, 0.913376,
    (1,0): 0.632359, 0.0975404, 0.278498, 0.546882,
    (2,0): 0.957507, 0.964889, 0.157613, 0.970593,
    (3,0): 0.957167, 0.485376, 0.80028, 0.141886,
    (4,0): 0.421761, 0.915736, 0.792207, 0.959492,
    (5,0): 0.655741, 0.0357117, 0.849129, 0.933993
    }
  }
}
}
```

Writing Attributes

- The command: `h5writeatt('newfile.hdf', '/mydata', 'temp', 20)` gives the output

```
HDF5 "newfile.hdf" {
GROUP "/" {
  DATASET "mydata" {
    DATATYPE  H5T_IEEE_F64LE
    DATASPACE  SIMPLE { ( 6, 4 ) / ( 6, 4 ) }
    DATA {
    (0,0): 0.0495265, 0.303303, 0.735416, 0.30518,
…
    }
    ATTRIBUTE "temp" {
      DATATYPE  H5T_IEEE_F64LE
      DATASPACE  SIMPLE { ( 1 ) / ( 1 ) }
      DATA {
      (0): 20
      }
    }
  }
}
}
```

Converting Data to an HDF5 File

In Matlab and in an HDF5 Utility

- Matlab
  - Pretty easy! Just read in your text data, then use h5write as in previous example to convert to HDF5 dataset
- HDF5 utility: `h5import`
  - Converts data from one or more ASCII or binary files (infile) into the same number of datasets in an existing or new HDF5 file (outfile)
  - Syntax:

```
h5import infile OPTIONS –o outfile
```

http://www.hdfgroup.org/HDF5/doc1.6/Tools.html#Tools-Import

Questions?

Shelley.knuth@colorado.edu
@Cu_data
@shelley_knuth

- References in addition to those already mentioned
  - http://neondataskills.org/HDF5/About/
  - http://www.hdfgroup.org/ (Download HDF5 here)
- Content in this talk is taken liberally from all mentioned sources