Store & manage data *effortlessly* with HDF5

Margaret Mahan
Twitter: @mymahan

PyData Amsterdam 2016

Presentation Materials: <INSERT LINK>

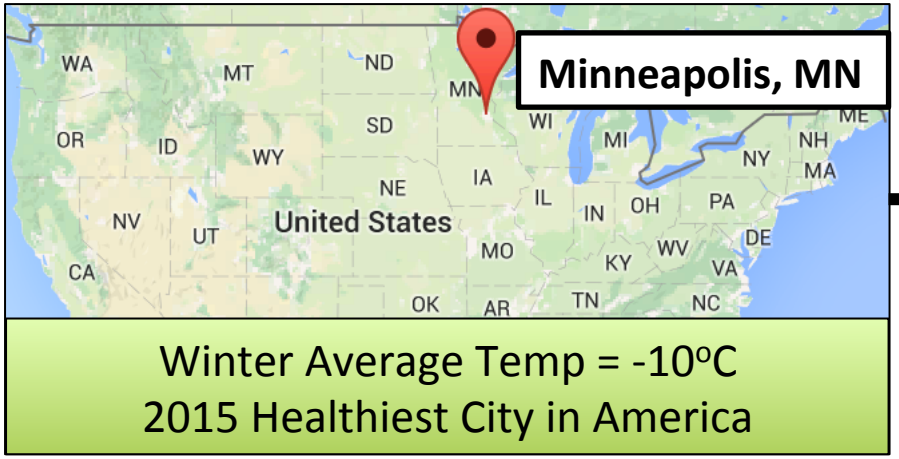
Are you …

- a Pyentist[1]?
- drowning in text files?
- frequently ‘grep’-ing?
- extending filenames for each processing step?
- looking for accessible, compressed, organized data?

Then HDF5 might be the solution you’re looking for!

[1]A Pyentist is a Python programming scientist

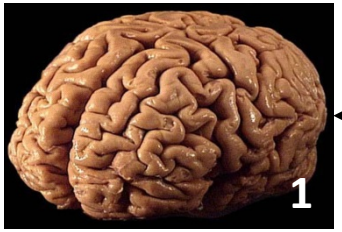
Who am I



UNIVERSITY OF MINNESOTA

BICB

Scientist
(data, computer, neuro, network)



BRAIN | SCIENCES | CENTER

BrainInjuryResearchLab

http://samadanilab.com/

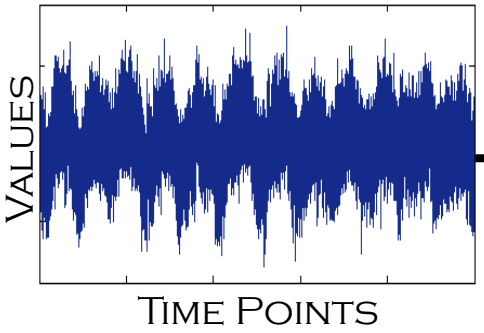
1 - Brain Image: https://commons.wikimedia.org/wiki/File:Human_brain_01.jpg

Why I use HDF5

Magnetoencephalography

high temporal resolution brain function

For each MEG sensor:



- Raw: scan-type, comments
- Processed: residuals, model order, statistics
- Analyzed: correlations, standard errors

- # subjects x # files --> unmanageable
  for storage, search, & filename
  - HDF5 presented as most optimal solution
- Pipeline rewritten in Python & HDF5 interfaced well

What is HDF5

- HDF5 => **H**ierarchical **D**ata **F**ormat, v**5**
  - binary file type specification
- Data model, library, & file format
  for storing & managing data
- File system within a file
- Designed for
  - flexible & efficient storage & I/O
  - high volume & complex data
    - supports unlimited variety of datatypes

Why you'd use HDF5

- **Processing**
  - *Supports large, complex, heterogeneous data & data slicing*
  - Get more done by accessing data quickly and easily
- **Sharing**
  - *Platform independent, self-describing, open format*
  - Is caring
- **Archiving**
  - *Compression, self-describing*
  - Data will last forever
- **b/c you can with python**

Image: https://www.sketchport.com/drawing/1782016/camera

HDF5 Structure

Groups – think "Python dictionaries" or directories
Datasets – think "NumPy array" or files
Attribute – think "metadata" or labels

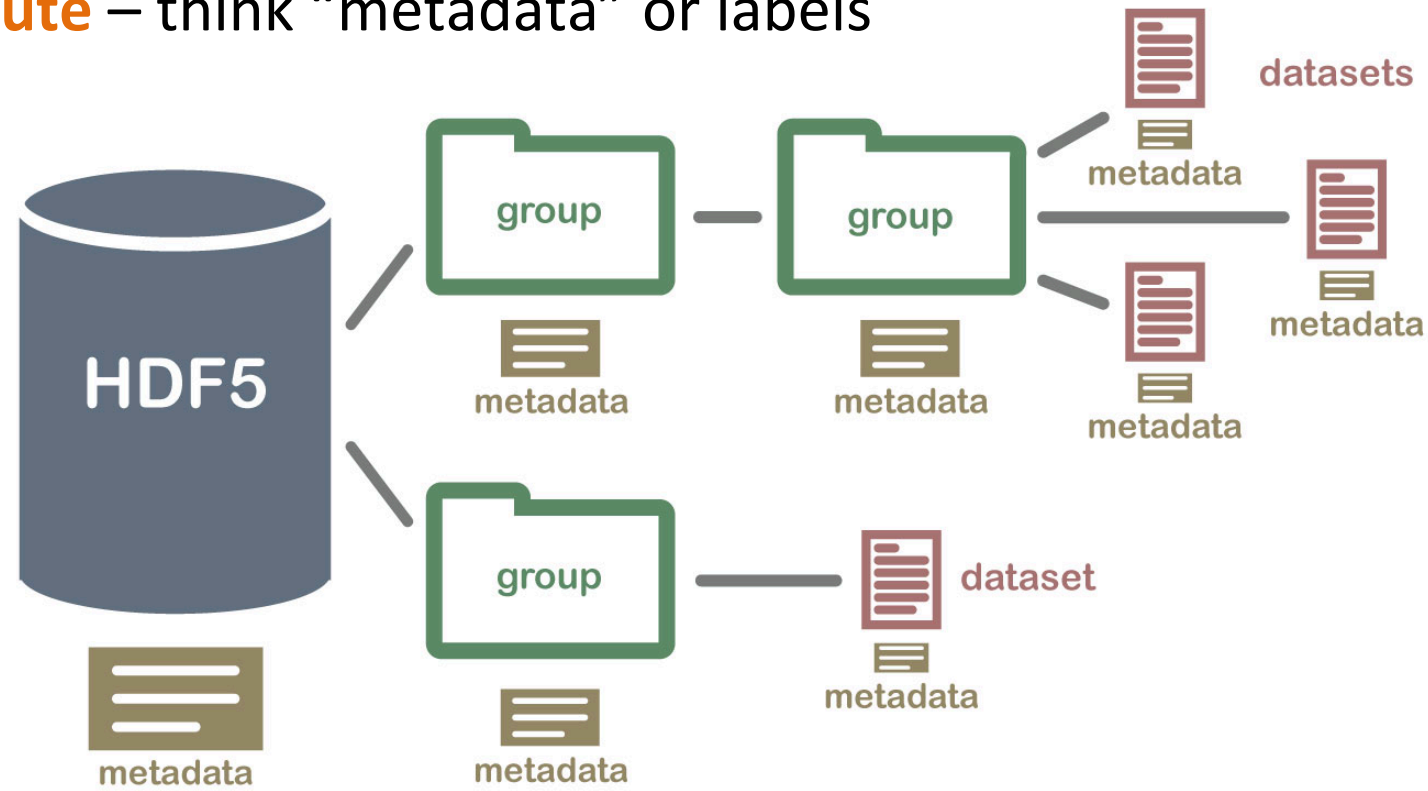
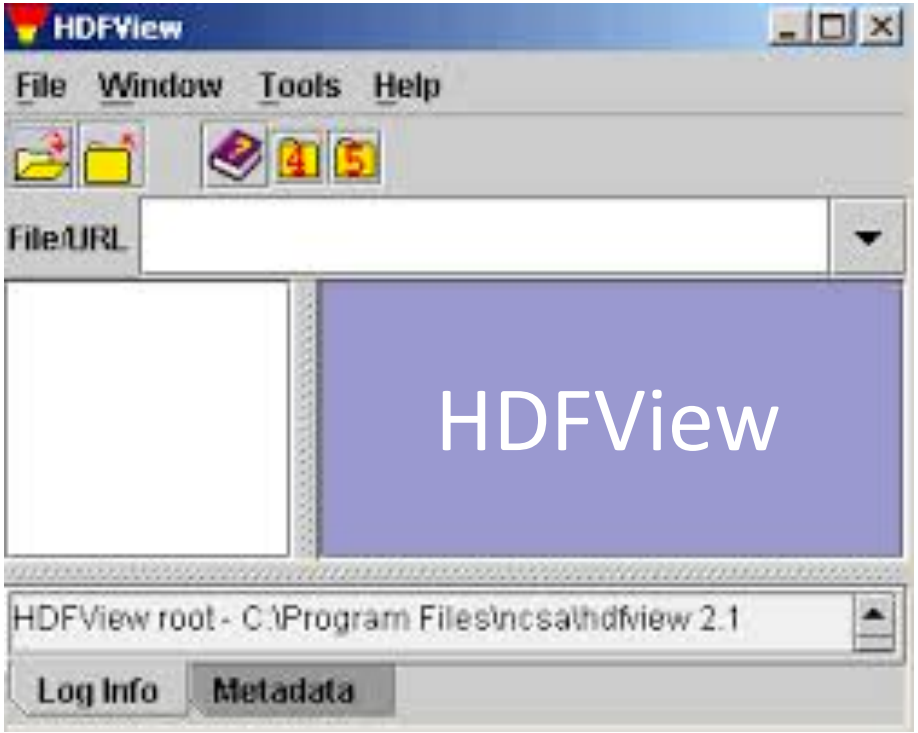


Image: http://neondataskills.org/HDF5/About

Getting started with HDF5 in Python using h5py

Viewers



https://www.hdfgroup.org

HDF5 Take Home Messages

- File system within a file
  - Groups (folders), datasets (files), attributes (labels)
- **Processing** *(work smarter, not harder)*
  - Accessible; supports large data, complex, heterogeneous data; supports data slicing
- **Sharing** *(is caring)*
  - Platform independent; self-describing; open format
- **Archive** *(data in perpetuity)*
  - Compressed; organized; self-describing

thank you

- Former BSC Colleagues
  - Specifically Chelley Chorn
- Andrew Collette
  - Lead Author h5py
- The HDF Group

History of HDF5

- 80's - National Center for Supercomputing Applications, University of Illinois
  - Create a file format & library for moving scientific data among platforms
- 90's - Adoption by organizations (e.g., NASA, DOE)
- 98 - HDF5 released
- 00's - Adoption by more organizations in government, academia, & industry
- 06 - The HDF Group officially began full operations independent of the University as a non-profit
- 15 - Latest Release: HDF5-1.8.16 (November)

subject.hdf5 example