



Alcator *C-Mod*

AN AUTOMATED PROCESS
FOR GENERATING ARCHIVAL
DATA FILES FROM MATLAB
FIGURES

G.M. Wallace, M. Greenwald, and J. Stillerman
MIT Plasma Science and Fusion Center


New federal directive requires archival datasets for all published figures

- Directive from White House Office of Science and Technology Policy (OSTP) to federal funding agencies requires storage and access of digital data from all non-classified sponsored research
- Department of Energy has developed a Public Access Plan, as have other federal agencies such as NSF
- **I won't argue about the merits of this directive. If you want to gripe about it, then go elsewhere**

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren 
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

4. Objectives for Public Access to Scientific Data in Digital Formats

To the extent feasible and consistent with applicable law and policy²; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, **digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze. For** purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens. Each agency's public access plan shall:

Public Access Plan



U.S. Department of Energy
July 24, 2014

[ENERGY.GOV](http://www.energy.gov)

DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved and, at a minimum, describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the principles stated above. The published article should indicate how these data can be accessed. Individual research offices will encourage researchers to deposit data in existing community or institutional repositories or to submit these data to the article publisher as supplemental information.

HDF5 file format chosen for data archive format at MIT PSFC

5

- HDF5 file format is an open, machine readable, cross platform standard
- HDF5 supported by many commonly used computing environments (MATLAB, IDL, Python, FORTRAN, C...)
- Format allows storage of data (e.g. x,y points) and attributes (e.g. labels, colors) within a single file
- Manually creating HDF5 file for each figure will change established user workflow dramatically

Problem: How to create HDF5 file without changing workflow?

- MATLAB *.fig file contains all the data and metadata to reproduce the figure
- Users often save a *.fig file so it can be modified later (e.g. change symbol shapes)
- If you have access to a *.fig file, it would be much more convenient to export the *.fig file to HDF5 rather than access all the raw data from storage, re-analyze it, then manually save each data trace to HDF5 file after the fact

Solution: export_fig.m script converts MATLAB *.fig file to HDF5

- Simple, easy to operate script
- Less than 1 second to process most figures
- No need to find your old code, re-run analysis, etc. Just locate the *.fig file and go
- The script creates HDF5 file without any user input
- Example usage for file 'example1.fig':

```
>> export_fig('example1')
```
- Stores attributes: line style, line width, color, marker shape, legend name, x-label, y-label, z-label, title

Example: simple figure

8

HDF5 example9.h5

Group '/'

Group '/Plot1'

Attributes:

'XLabel1': '\theta [rad]'

'YLabel1': 'Voltage [V]'

Group '/Plot1/Data1'

Attributes:

'StructPath': 's.hgS_070000.children(1).children(1).properties'

'Color [r g b]': 0.000000 0.447000 0.741000

Dataset 'XData'

Size: 1x100

MaxSize: 1x100

Datatype: H5T_IEEE_F64LE (double)

ChunkSize: []

Filters: none

FillValue: 0.000000

Dataset 'YData'

Size: 1x100

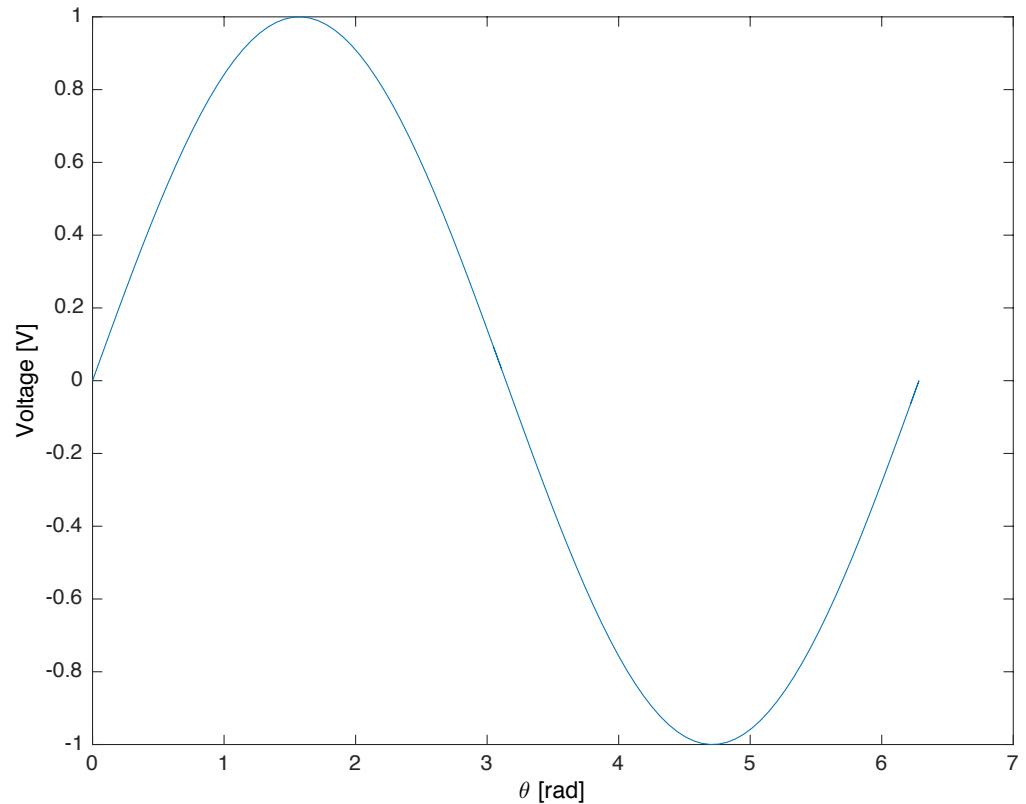
MaxSize: 1x100

Datatype: H5T_IEEE_F64LE (double)

ChunkSize: []

Filters: none

FillValue: 0.000000



Example: 2-D color plot

HDF5 example3.h5

Group '/'

Group '/Plot1'

Attributes:

'XLabel1': '\theta'

'YLabel1': '\phi'

Group '/Plot1/Data1'

Attributes:

'StructPath': 's.hgS_070000.children.children(1),properties'

Dataset 'XData'

Size: 1x100

MaxSize: 1x100

Datatype: H5T_IEEE_F64LE (double)

ChunkSize: []

Filters: none

FillValue: 0.000000

Dataset 'YData'

Size: 1x101

MaxSize: 1x101

Datatype: H5T_IEEE_F64LE (double)

ChunkSize: []

Filters: none

FillValue: 0.000000

Dataset 'ZData'

Size: 101x100

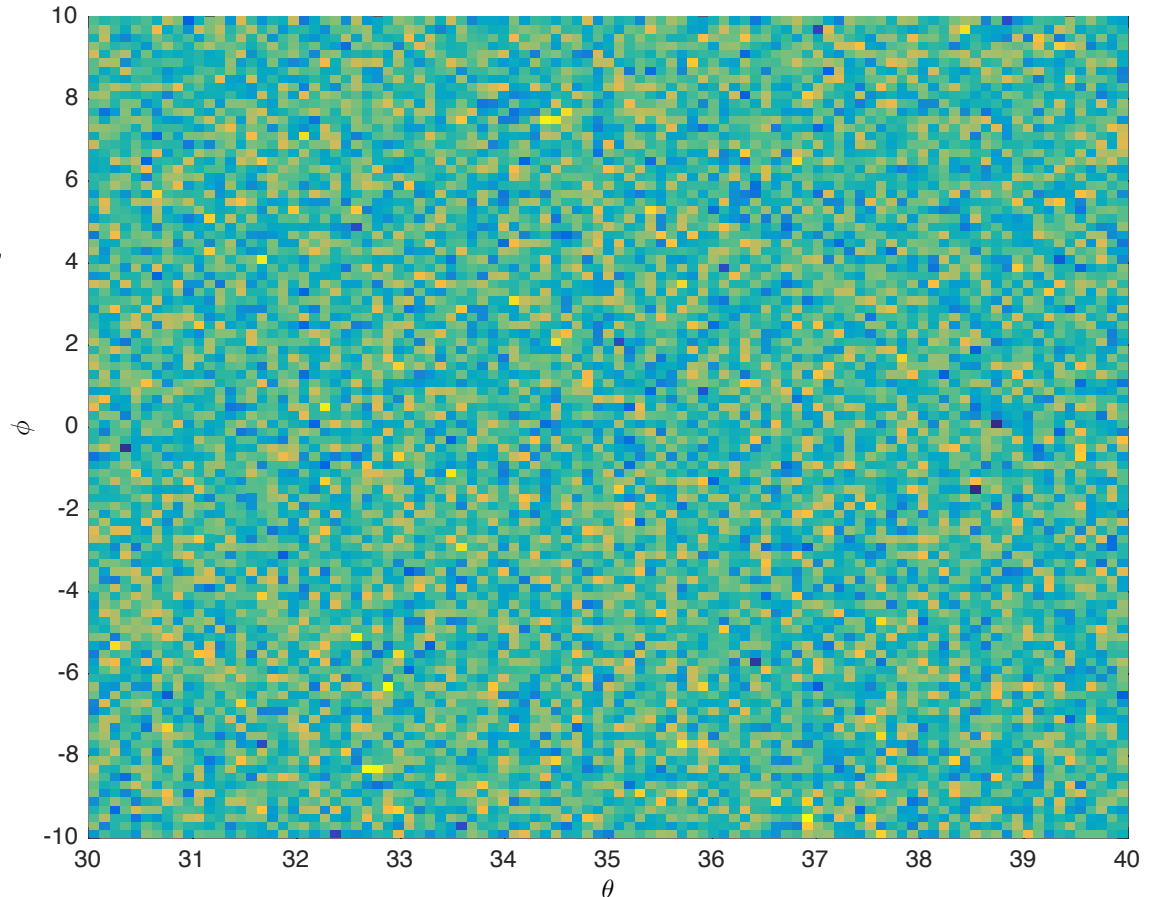
MaxSize: 101x100

Datatype: H5T_IEEE_F64LE (double)

ChunkSize: []

Filters: none

FillValue: 0.000000



Example: multiple curves on plot

10

HDF5 example4.h5

Group '/'

Group '/Plot1'

Attributes:

'XLabel1': 'Phase [$^{\circ}$]

'YLabel1': 'Volts [V]'

Group '/Plot1/Data1'

'DisplayName': 'tan'

...

Group '/Plot1/Data2'

'DisplayName': 'cot'

...

Group '/Plot1/Data3'

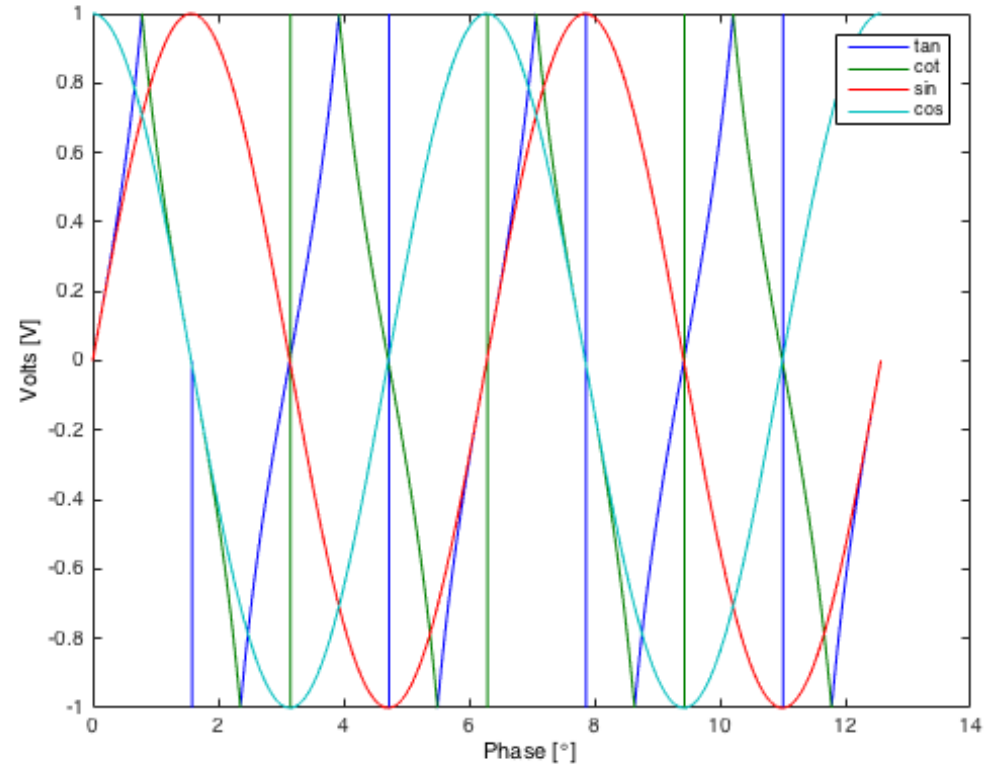
'DisplayName': 'sin'

...

Group '/Plot1/Data4'

'DisplayName': 'cos'

...



Example: multiple subplots

11

HDF5 example6.h5

Group '/'

Group '/Plot1'

Attributes:

'YLabel1': 'P_{LH} [kW]'

'Title1': '1140225019'

Group '/Plot1/Data1'

...

Group '/Plot2'

Attributes:

'YLabel1': '[10^{20} m^{-3}]'

Group '/Plot2/Data1'

...

Group '/Plot3'

Attributes:

'YLabel1': 'V_{loop} [V]'

Group '/Plot3/Data1'

...

Group '/Plot4'

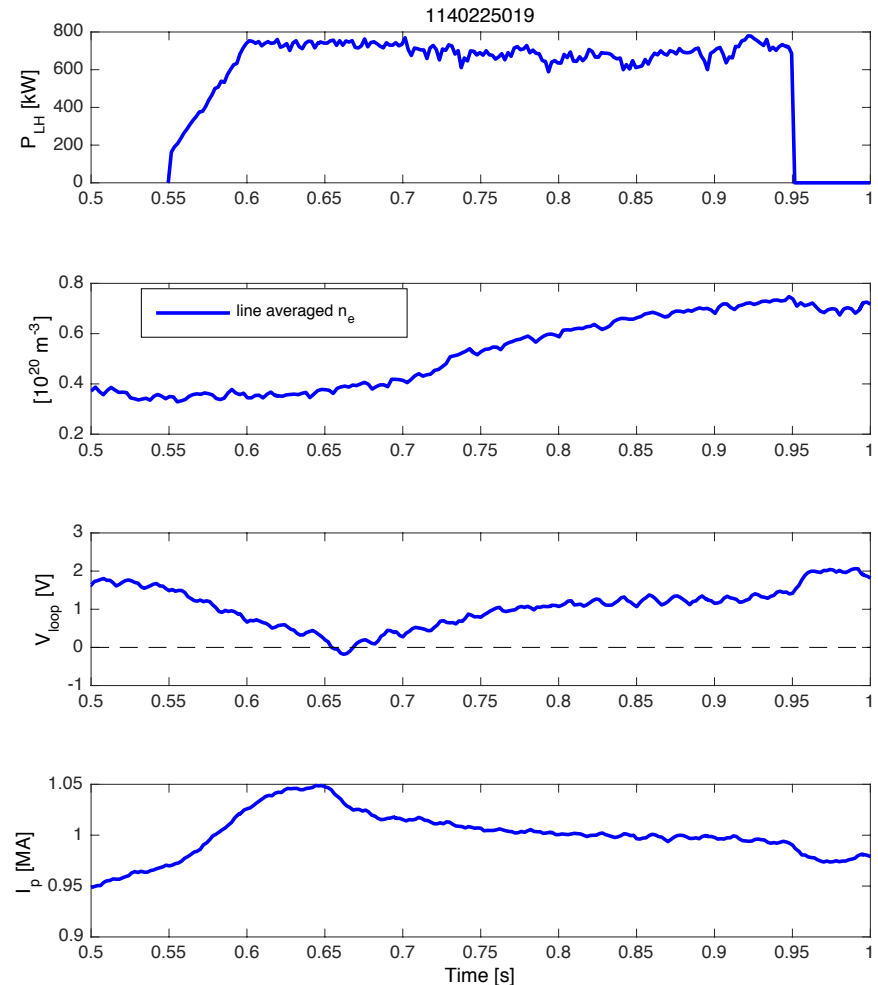
Attributes:

'YLabel1': 'I_p [MA]'

'XLabel1': 'Time [s]'

Group '/Plot4/Data1'

...



Alcator C-Mod Example: two y-axes

12

HDF5 example12.h5

Group '/'

Group '/Plot1'

Attributes:

'YLabel1': 'Slow Decay'

'XLabel1': 'Time (\musec)'

'Title1': 'Multiple Decay Rates'

Group '/Plot1/Data1'

Attributes:

'StructPath': 's.hgS_070000.children(1).children(1).properties'

'LineStyle': '--'

'Color [r g b]': 0.000000 0.000000 1.000000

...

Group '/Plot2'

Attributes:

'YLabel1': 'Fast Decay'

Group '/Plot2/Data1'

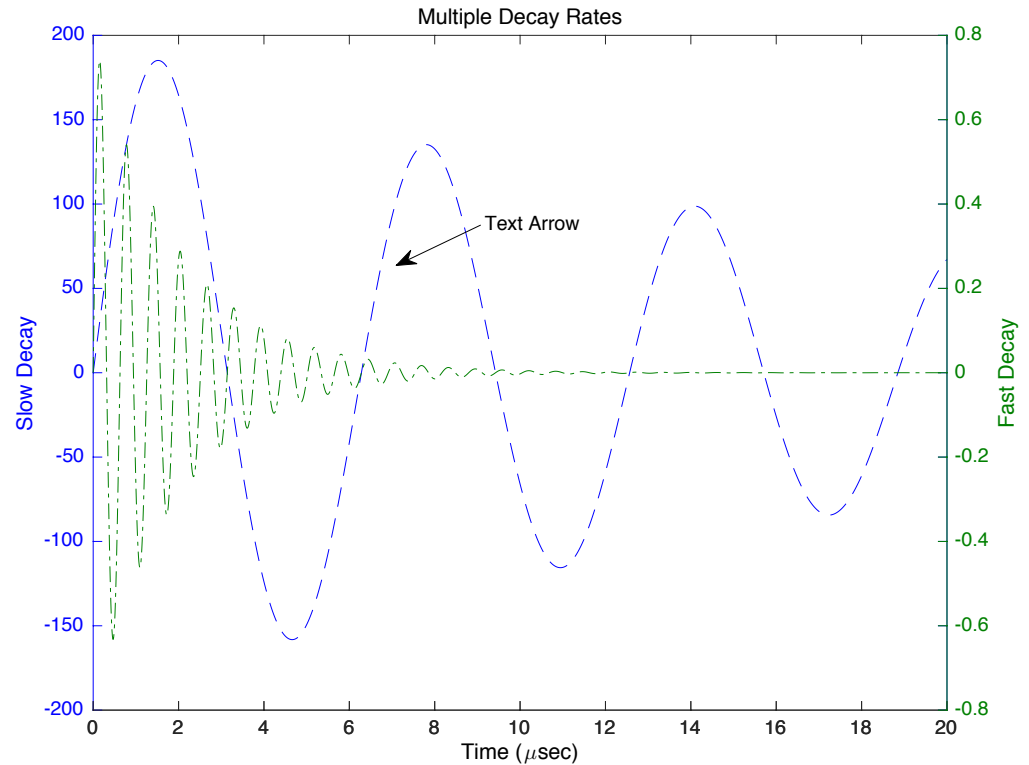
Attributes:

'StructPath': 's.hgS_070000.children(2).children(1).properties'

'LineStyle': '-.'

'Color [r g b]': 0.000000 0.500000 0.000000

...



HDF5 example2.h5

Group '/'

Group '/Plot1'

Attributes:

'XLabel1': 'time [s]'

'YLabel1': 'position [mm]'

Group '/Plot1/Data1'

Attributes:

...

'Color [r g b]': 0.000000 0.000000 1.000000

Dataset 'LData'

...

Dataset 'UData'

...

Dataset 'XData'

...

Dataset 'YData'

...

Group '/Plot1/Data2'

Attributes:

'...

'Color [r g b]': 0.000000 0.500000 0.000000

Dataset 'LData'

...

Dataset 'UData'

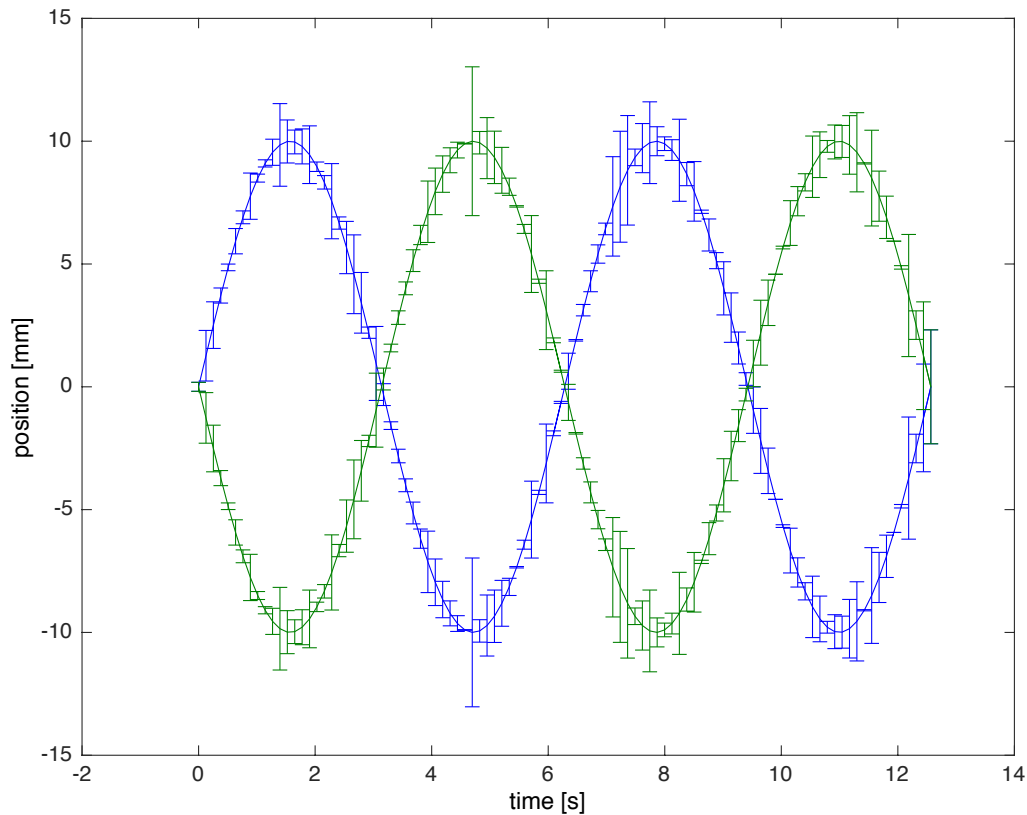
...

Dataset 'XData'

...

Dataset 'YData'

...



Example: quiver plot

14

HDF5 example13.h5

Group '/'

Group '/Plot1'

Group '/Plot1/Data1'

Attributes:

'StructPath': 's.hgS_070000.children.children.properties'

'Color [r g b]': 0.000000 0.447000 0.741000

Dataset 'UData'

...

Dataset 'VData'

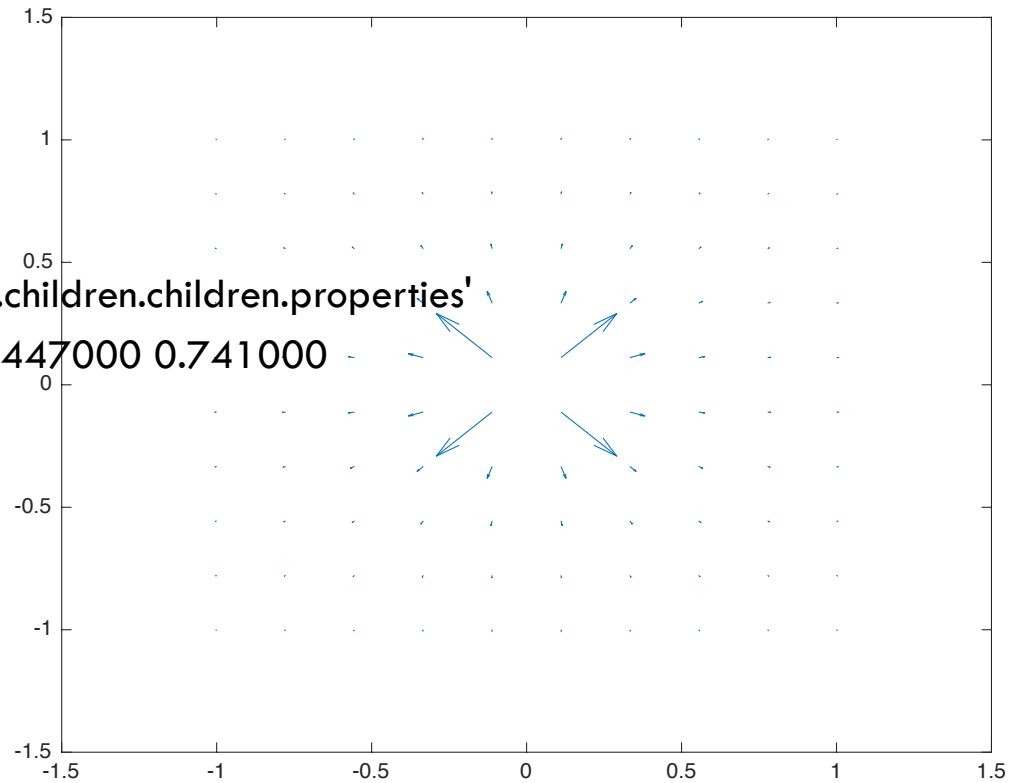
...

Dataset 'XData'

...

Dataset 'YData'

...



Changes to figure outside of MATLAB not reflected in HDF5 file

- Attributes (labels, units, colors, symbols, etc) come directly from the *.fig file, not the final *.pdf
- Users sometimes modify figures in other software (e.g. Adobe Illustrator)
- Any changes made outside MATLAB will not be reflected in the HDF5 file
- HDF5 file can be edited manually in this case, but this is tedious and time consuming
- Best usage is to generate publication quality figure fully within MATLAB

Technique should be applicable to Python, perhaps other languages

- Storage of data within figure data structure similar between MATLAB and Python matplotlib
- Should be possible to write a similar script for Python, **but I'm not doing it**
- Happy to share my source code and discuss approach with whomever wants to work on the Python implementation

Script available to download from MATLAB file exchange

- File ID: #59937
- <https://www.mathworks.com/matlabcentral/fileexchange/59937-export-fig-filename->

