



# Introduction to other file types



# Other file types

- Excel spreadsheets
- MATLAB files
- SAS files
- Stata files
- HDF5 files



#### **Pickled files**

- File type native to Python
- Motivation: many datatypes for which it isn't obvious how to store them
- Pickled files are serialized
- Serialize = convert object to bytestream



#### Pickled files

```
In [1]: import pickle
In [2]: with open('pickled_fruit.pkl', 'rb') as file:
    ...: data = pickle.load(file)

In [3]: print(data)
{'peaches': 13, 'apples': 4, 'oranges': 11}
```



### Importing Excel spreadsheets



#### You'll learn:

- How to customize your import
  - Skip rows
  - Import certain columns
  - Change column names





# Let's practice!





# Importing SAS/Stata files using pandas



#### **SAS** and **Stata** files

- SAS: Statistical Analysis System
- Stata: "Statistics" + "data"
- SAS: business analytics and biostatistics
- Stata: academic social sciences research



#### **SAS** files

- Used for:
  - Advanced analytics
  - Multivariate analysis
  - Business intelligence
  - Data management
  - Predictive analytics
- Standard for computational analysis



# **Importing SAS files**



# **Importing Stata files**

```
In [1]: import pandas as pd
In [2]: data = pd.read_stata('urbanpop.dta')
```





# Let's practice!





# Importing HDF5 files



#### **HDF5** files

- Hierarchical Data Format version 5
- Standard for storing large quantities of numerical data
- Datasets can be hundreds of gigabytes or terabytes
- HDF5 can scale to exabytes



# **Importing HDF5 files**

```
In [1]: import h5py
In [2]: filename = 'H-H1_LOSC_4_V1-815411200-4096.hdf5'
In [3]: data = h5py.File(filename, 'r') # 'r' is to read
In [4]: print(type(data))
<class 'h5py._hl.files.File'>
```



#### The structure of HDF5 files

This gives a high level picture of what's contained in a LSGO data file. There are 3 types of information:

• sens: Meta-data for the file. This is book information such as the GPS times covered, which instrument, etc.

• quality: Refers to data quality. The main item here is a 1 Hz time series describing the data quality for each second of data. This is an important topic, and we'll devote a whole step of the tutorial to working with data quality information.

• storage Strain data from the interferometer. In some sense, this is "the data", the main measurement performed by LIGO.



#### The structure of HDF5 files



# The HDF Project

Actively maintained by the HDF Group



Based in Champaign, Illinois





# Let's practice!





# **Importing MATLAB files**



#### **MATLAB**

- "Matrix Laboratory"
- Industry standard in engineering and science
- Data saved as .mat files



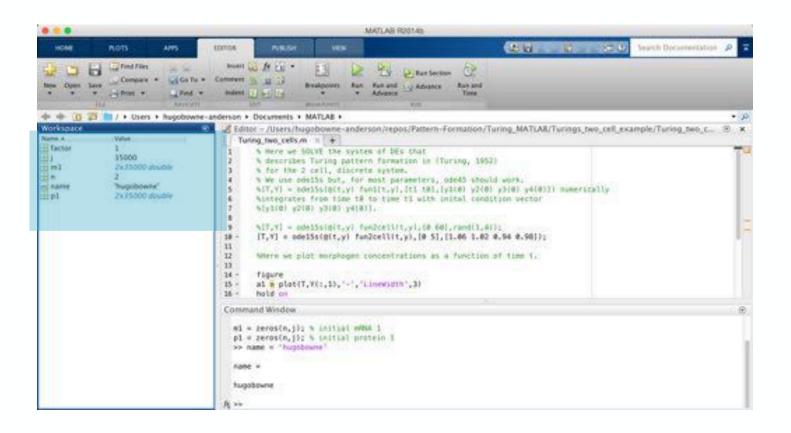


# SciPy to the rescue!

- scipy.io.loadmat() read.matfiles
- scipy.io.savemat() write.mat files



#### What is a .mat file?





# Importing a .mat file

```
In [1]: import scipy.io
In [2]: filename = 'workspace.mat'
In [3]: mat = scipy.io.loadmat(filename)
In [4]: print(type(mat))
<class 'dict'>
In [5]: print(type(mat['x']))
<class 'numpy.ndarray'>
```

- keys = MATLAB variable names
- values = objects assigned to variables





# Let's practice!