Introduction to Data Science Libraries: Pandas, Seaborn, and Matplotlib

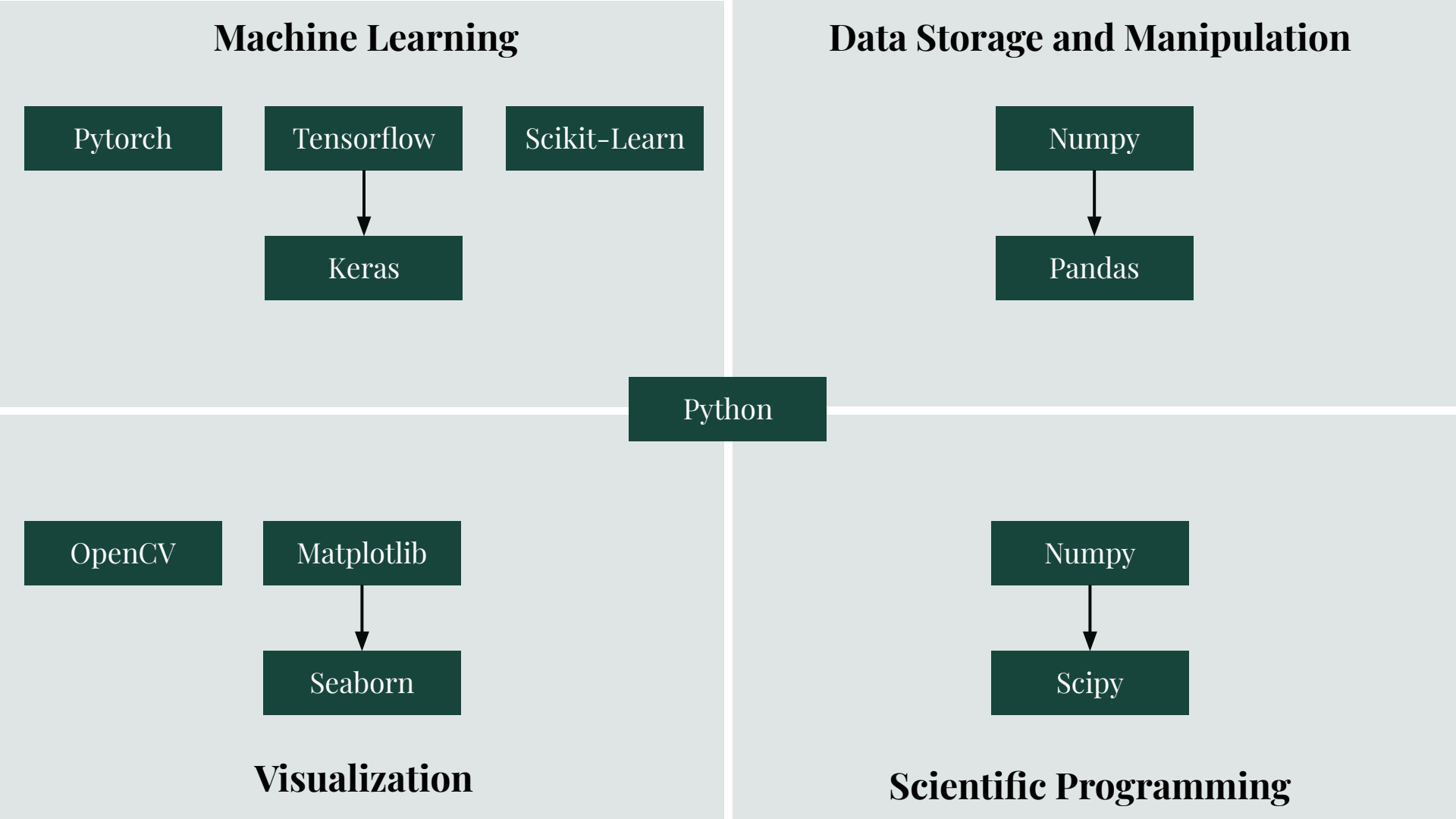
Julie Butler

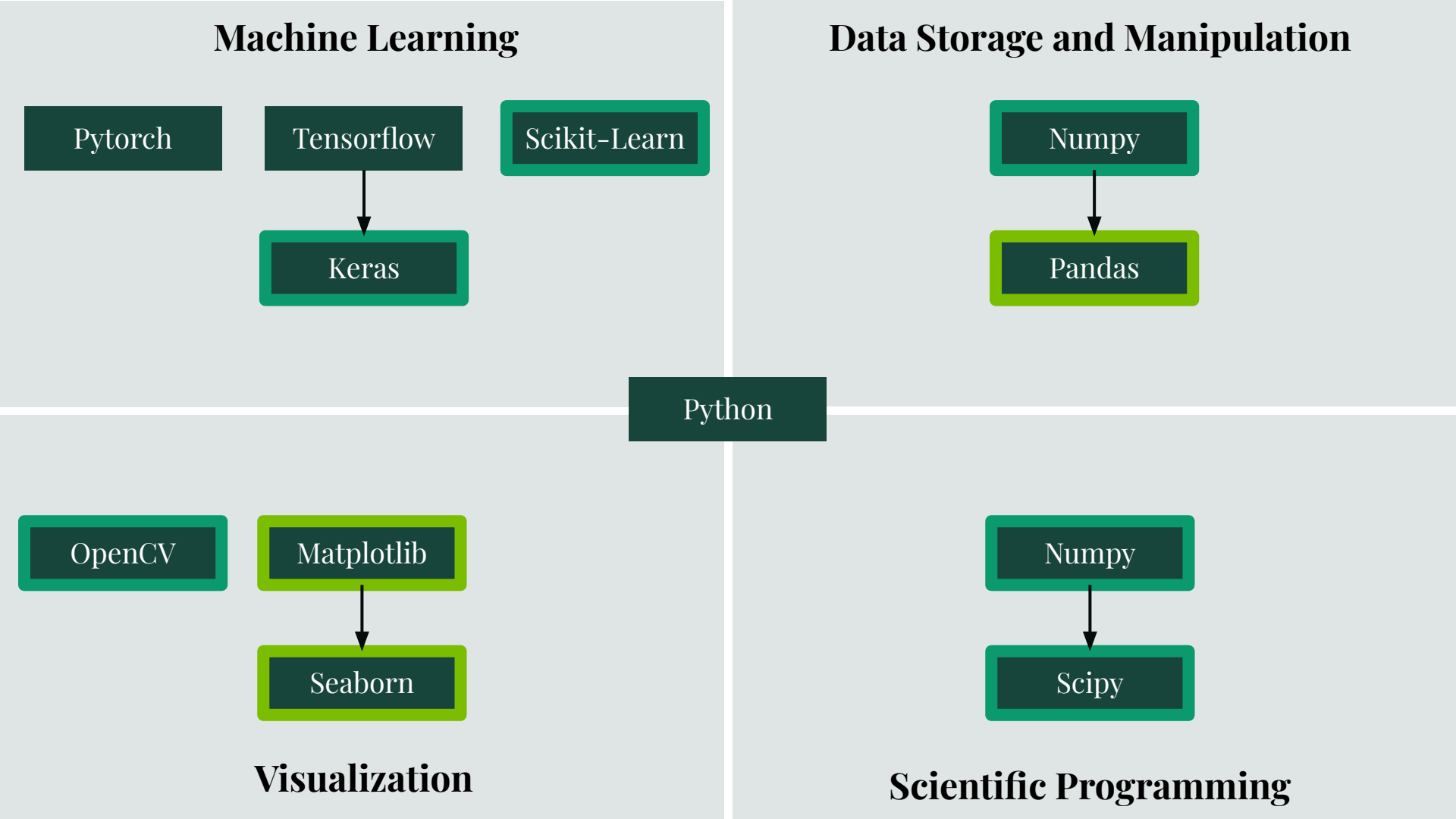
June 26, 2023

Michigan State University

University of Mount Union

MICHIGAN STATE UNIVERSITY


Machine Learning
Pytorch
Tensorflow
Scikit-Learn
Keras
Data Storage and Manipulation
Numpy
Pandas
Python
OpenCV
Matplotlib
Seaborn
Visualization
Numpy
Scipy
Scientific Programming


Machine Learning
Pytorch
Tensorflow
Scikit-Learn
Keras
Data Storage and Manipulation
Numpy
Pandas
Python
OpenCV
Matplotlib
Seaborn
Visualization
Numpy
Scipy
Scientific Programming

The Data Science Workflow

What is the process of analyzing a data set?

N	Z	A	EL	BE	BEd	MASS	MASSd
1	0	1	n	0.0	0.0	1008664.91582	0.00049
0	1	1	H	0.0	0.0	1007825.03224	0.00009
1	1	2	H	1112.283	0.000	2014101.77811	0.00012
2	1	3	H	2827.265	0.000	3016049.28199	0.00023
1	2	3	He	2572.680	0.000	3016029.32265	0.00022
0	3	3	Li	-2267#	667#	3030775#	2147#
3	1	4	H	1720.449	25.000	4026431.868	107.354
2	2	4	He	7073.915	0.000	4002603.25413	0.00006
1	3	4	Li	1153.760	53.033	4027185.562	227.733
4	1	5	H	1336.359	17.889	5035311.493	96.020
3	2	5	He	5512.132	4.000	5012057.224	21.470
2	3	5	Li	5266.132	10.000	5012537.800	53.677
1	4	5	Be	18#	401#	5039870#	2150#
5	1	6	H	961.639	42.354	6044955.437	272.816
4	2	6	He	4878.519	0.009	6018885.891	0.057
3	3	6	Li	5332.331	0.000	6015122.88742	0.00155
2	4	6	Be	4487.247	0.908	6019726.409	5.848
1	5	6	B	-467#	334#	6050800#	2150#
6	1	7	H	940#	143#	7052749#	1078#
5	2	7	He	4123.057	1.080	7027990.654	8.115
4	3	7	Li	5606.439	0.001	7016003.43666	0.00454

1. Determine what the data is

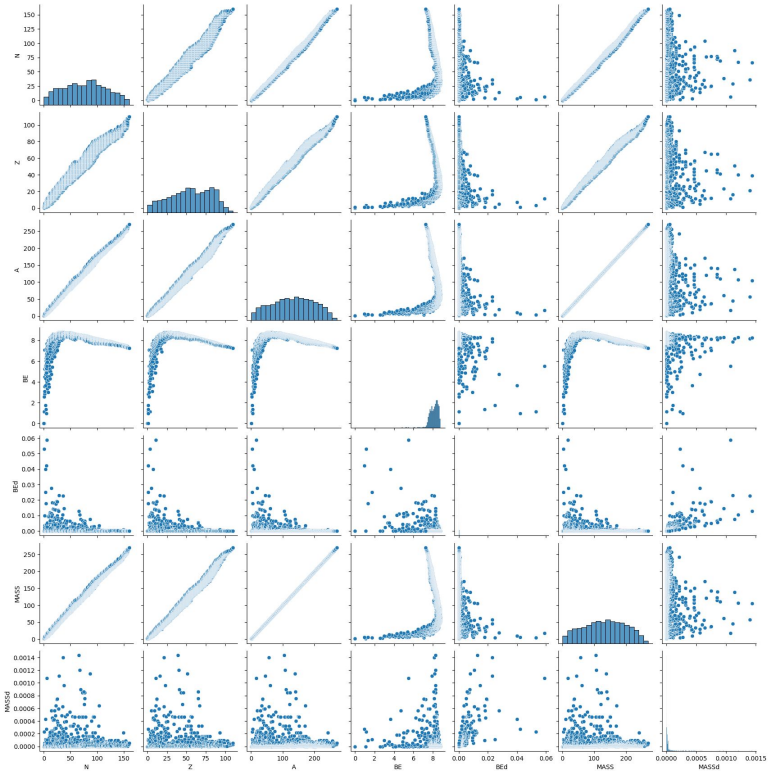
N	Z	A	EL	BE	BEd	MASS	MASSd
1	0	1	n	0.0	0.0	1008664.91582	0.00049
0	1	1	H	0.0	0.0	1007825.03224	0.00009
1	1	2	H	1112.283	0.000	2014101.77811	0.00012
2	1	3	H	2827.265	0.000	3016049.28199	0.00023
1	2	3	He	2572.680	0.000	3016029.32265	0.00022
0	3	3	Li	-2267#	667#	3030775#	2147#
3	1	4	H	1720.449	25.000	4026431.868	107.354
2	2	4	He	7073.915	0.000	4002603.25413	0.00006
1	3	4	Li	1153.760	53.033	4027185.562	227.733
4	1	5	H	1336.359	17.889	5035311.493	96.020
3	2	5	He	5512.132	4.000	5012057.224	21.470
2	3	5	Li	5266.132	10.000	5012537.800	53.677
1	4	5	Be	18#	401#	5039870#	2150#
5	1	6	H	961.639	42.354	6044955.437	272.816
4	2	6	He	4878.519	0.009	6018885.891	0.057
3	3	6	Li	5332.331	0.000	6015122.88742	0.00155
2	4	6	Be	4487.247	0.908	6019726.409	5.848
1	5	6	B	-467#	334#	6050800#	2150#
6	1	7	H	940#	143#	7052749#	1078#
5	2	7	He	4123.057	1.080	7027990.654	8.115
4	3	7	Li	5606.439	0.001	7016003.43666	0.00454

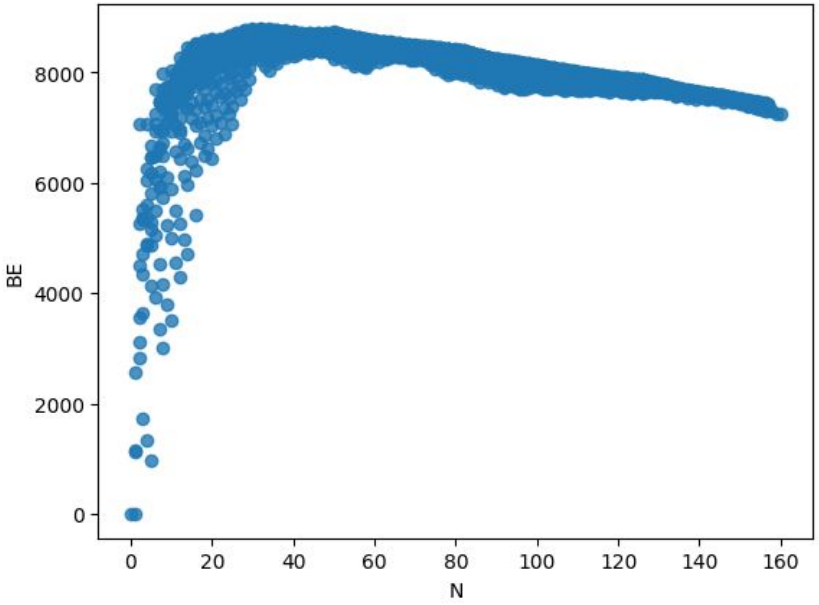
2. Determine if any of the data needs to be removed or reformatted

N	Z	A	EL	BE	BEd	MASS	MASSd
1	0	1	n	0.0	0.0	1008664.91582	0.00049
0	1	1	H	0.0	0.0	1007825.03224	0.00009
1	1	2	H	1112.283	0.000	2014101.77811	0.00012
2	1	3	H	2827.265	0.000	3016049.28199	0.00023
1	2	3	He	2572.680	0.000	3016029.32265	0.00022
0	3	3	Li	-2267#	667#	3030775#	2147#
3	1	4	H	1720.449	25.000	4026431.868	107.354
2	2	4	He	7073.915	0.000	4002603.25413	0.00006
1	3	4	Li	1153.760	53.033	4027185.562	227.733
4	1	5	H	1336.359	17.889	5035311.493	96.020
3	2	5	He	5512.132	4.000	5012057.224	21.470
2	3	5	Li	5266.132	10.000	5012537.800	53.677
1	4	5	Be	18#	401#	5039870#	2150#
5	1	6	H	961.639	42.354	6044955.437	272.816
4	2	6	He	4878.519	0.009	6018885.891	0.057
3	3	6	Li	5332.331	0.000	6015122.88742	0.00155
2	4	6	Be	4487.247	0.908	6019726.409	5.848
1	5	6	B	-467#	334#	6050800#	2150#
6	1	7	H	940#	143#	7052749#	1078#
5	2	7	He	4123.057	1.080	7027990.654	8.115
4	3	7	Li	5606.439	0.001	7016003.43666	0.00454

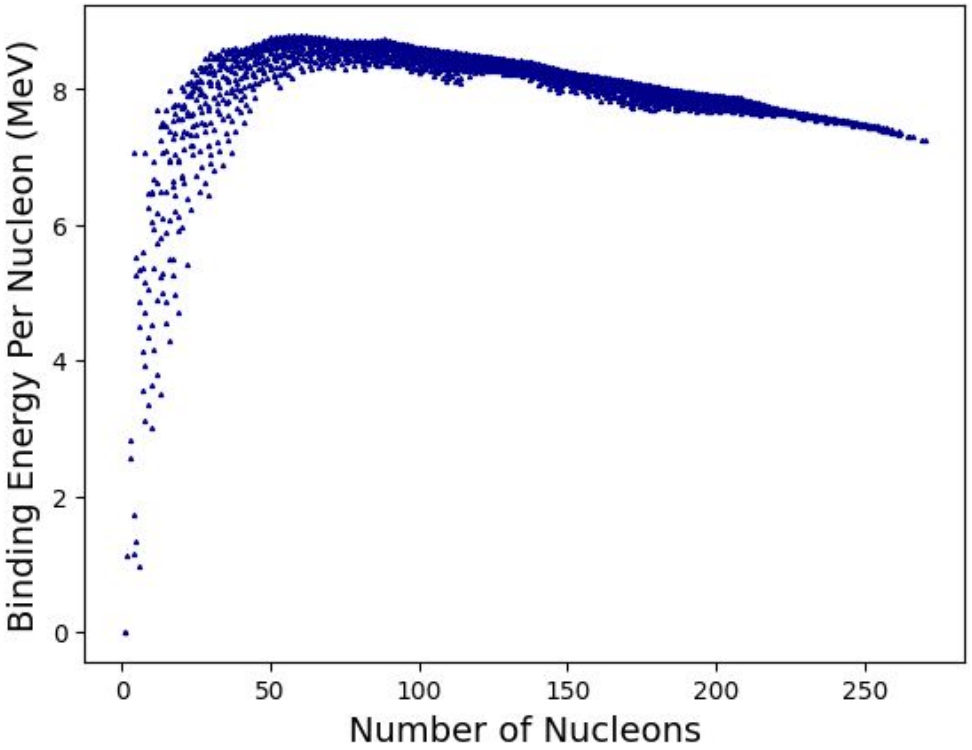
- BE and BEd are in units of keV
- MASS and MASSd are in units of μu

3. Make some initial graphs
4. Determine which graphs are best

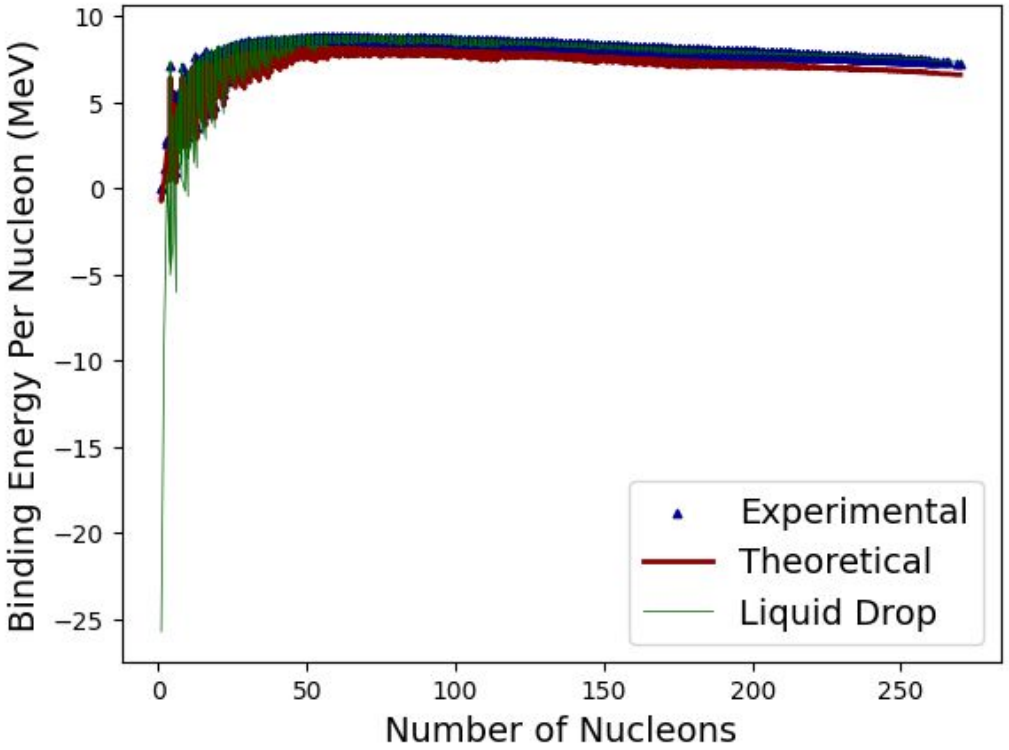




5. Create a plot that can be shared or published



6. Compare the data to known results



How can these 6 steps be accomplished with Python?

1. Determine what the data is
2. Determine if any of the data needs to be removed or reformatted
3. Make some initial graphs
4. Determine which graphs are best
5. Create a plot that can be shared or published
6. Compare the data to known results

How can these 6 steps be accomplished with Python?

Pandas

1. Determine what the data is
2. Determine if any of the data needs to be removed or reformatted

Seaborn

3. Make some initial graphs
4. Determine which graphs are best

Matplotlib

5. Create a plot that can be shared or published
6. Compare the data to known results

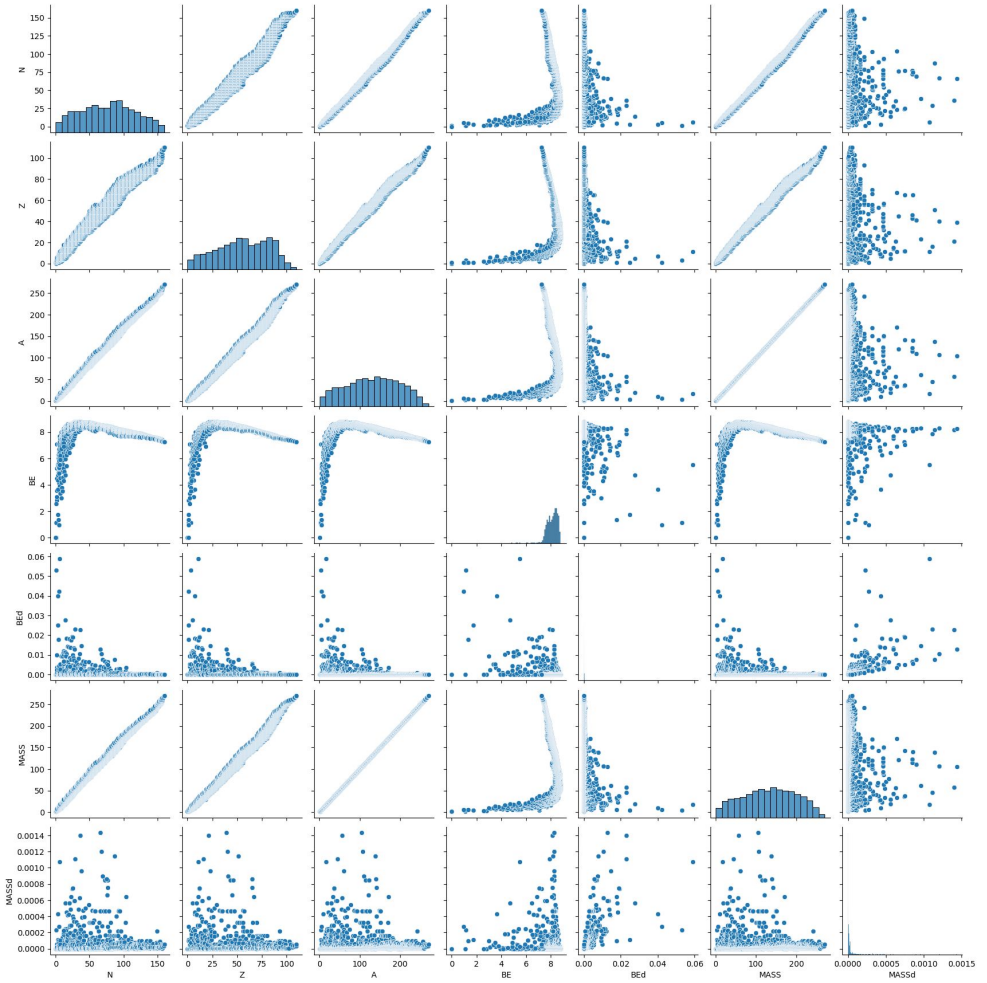
What this Module Covers

Pandas: Initial Data Analysis

- Importing data from a file
- Printing the Pandas Dataframe and just the “head”
- Accessing the different data columns and manipulating the values
- Removing unwanted data
- Creating masks (sub-Dataframes)

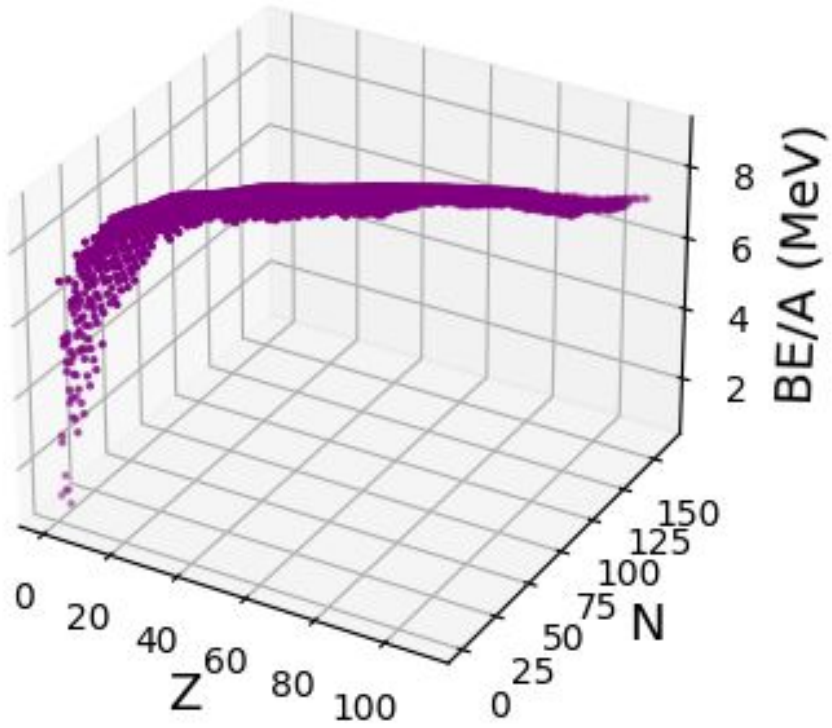
Seaborn: Initial Graphical Analysis

- Use pairplot to determine which plots may be of interest
- Use regplot to create a larger version of the plots of interest



Matplotlib: Creating Custom Graphs

- Creating 2D line plots and scatter plots
  - Customizing line widths, marker styles, colors, etc.
  - Adding axes labels
- Plotting multiple data sets on one plot
  - Creating a legend
- Creating three dimensional plots
- Creating plots with error bars



Conclusion

- The module cannot cover every feature of these libraries
- Conclusion provides links on where to learn more

Extra Exercise/Extended Practice

Main Section of the Notebook

- Create a scatter plot with these variables
- Change the color
- Change the marker style
- Change the marker size
- Add an x label
- Add a y label

Extended Practice

- Create a masks the extracts all elements with atomic number greater than 103 (super heavy elements).
- Create a pairplot with the hue being the proton number.
- Create a formatted two dimensional plot for atomic mass with error bars and axes labels.
- Create a formatted three dimensional plot for atomic mass with axes labels.

Details of the Module

Assumed Knowledge

- There is no assumed physics knowledge, not even of nuclear physics
- It is assumed the students know basic Python and Numpy BUT the exercises are very simple

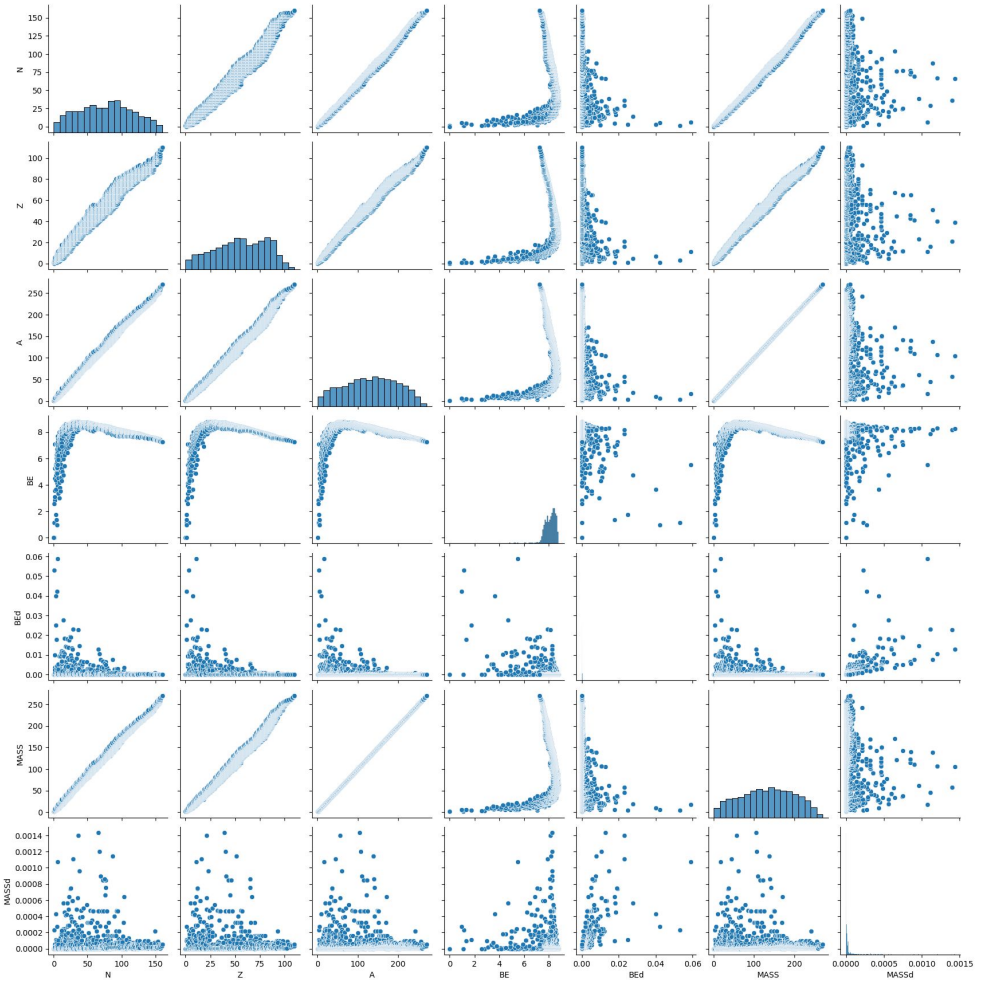
EXERCISE 6: Set the variable `color` in the below cell to your chosen Matplotlib color. Make sure you make the color name a string (enclosed in `"` or `'`), and spell it exactly as it is found in the documentation. Next, run the updated plot statement and make sure the color of the plot has changed.

```
[ ]  1 color =
```

```
[ ]  1 plt.plot(A,BE,color=color)
```

Physics Learning Goals

- Be able to create plots that display physics data in a meaningful way



Data Science Learning Goals

- Be able to read documentation for functions from the various data science libraries
- Be able to use Pandas to import a data file, format it as a Pandas Dataframe, and begin exploring and analyzing the data
- Be able to use Matplotlib and Seaborn to further your analysis of the data set
- Be able to use Matplotlib and Seaborn to make physically relevant plots of the data set both in two dimensions and in three dimensions

Class and Time Estimates

- This could be plugged into any class at any level as long as the students have basic Python knowledge
  - Lab course to learn how to graph data
- Time Estimates
  - Could be completed in a standard 3 hour lab period
  - Could be started in a standard 1 hour class and then completed as homework
- The module can be broken up
  - Pandas section can function independently
  - Seaborn section requires the Dataframe from Pandas
  - Matplotlib requires the extracted columns from the Dataframe

Conclusion

Conclusion

- This module is designed to teach students of any physics level and a basic Python level the three important data science libraries: Pandas, Seaborn, and Matplotlib
- Though the module only introduces these libraries, it gives the students the skills and resources to continue to learn to use the libraries
- It also teaches them the basic workflow that comes with analyzing a new data set

Questions?