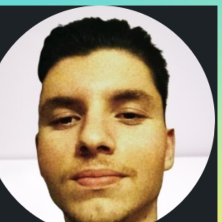
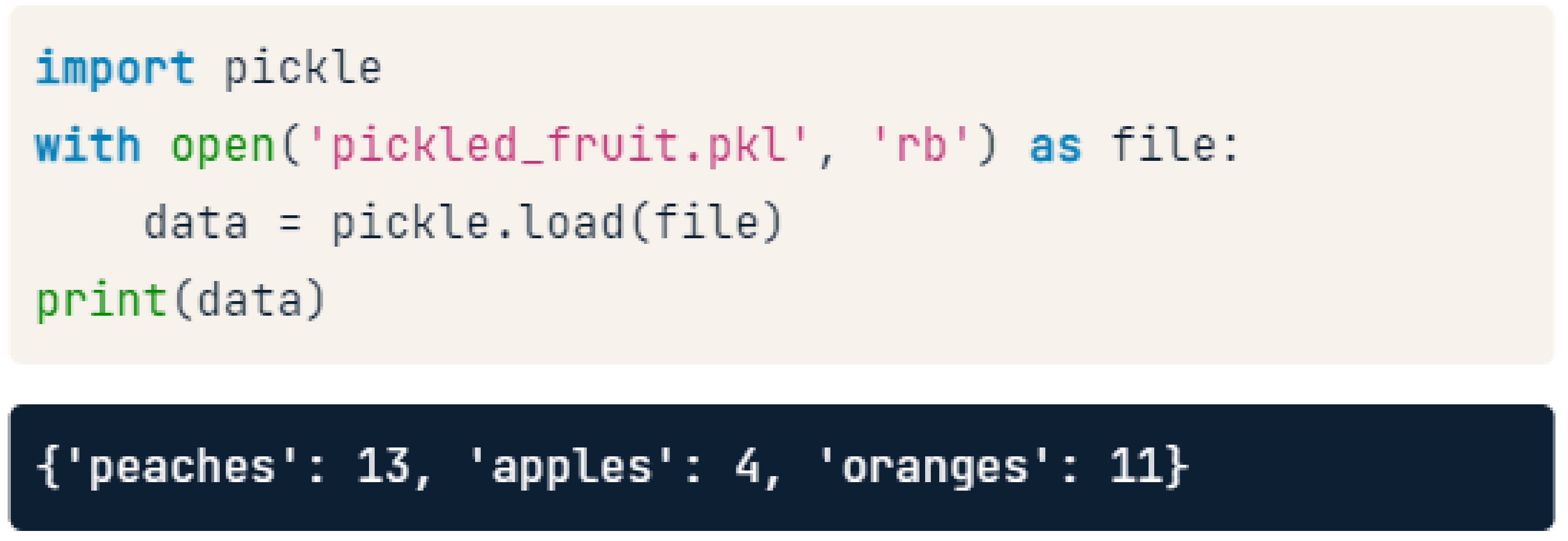
Importing Data in Python



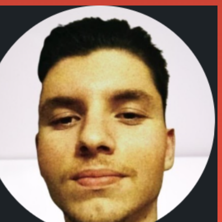
Pickled Files

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



```
import pickle
with open('pickled_fruit.pkl', 'rb') as file:
    data = pickle.load(file)
print(data)
```

```
{'peaches': 13, 'apples': 4, 'oranges': 11}
```



SAS Files

- Used for:

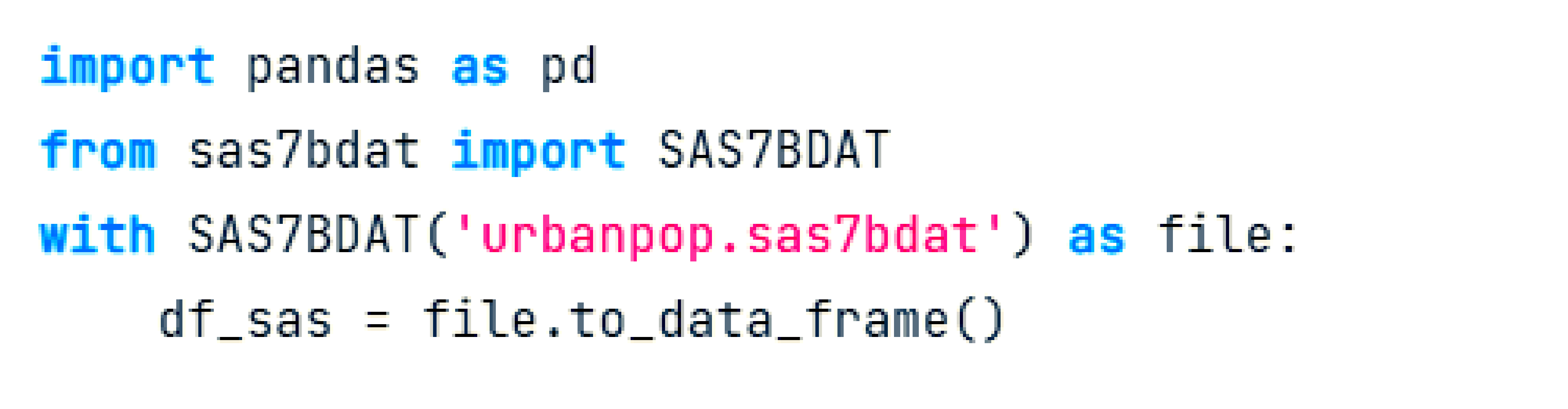
  Advance Analytics

  Multivariate Analysis

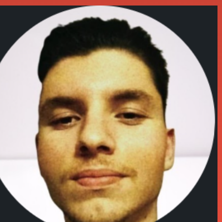
  Buisness Intelligence

  Data Management

  Predictive Analytics

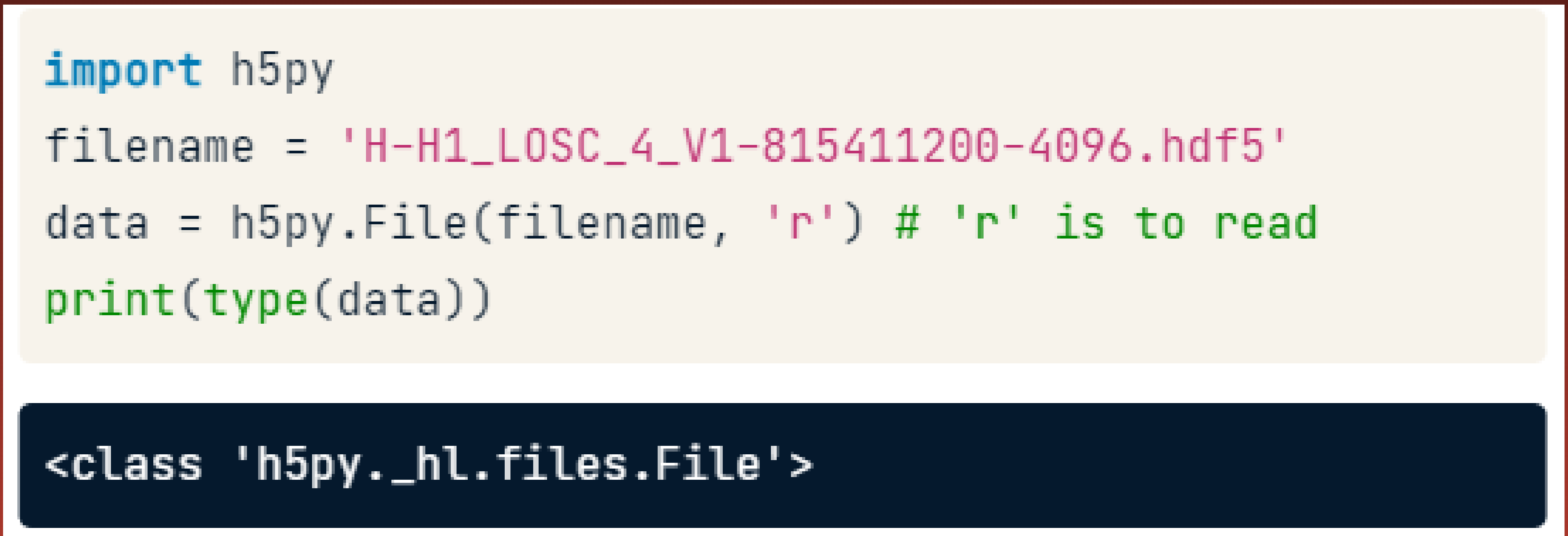


```
import pandas as pd
from sas7bdat import SAS7BDAT
with SAS7BDAT('urbanpop.sas7bdat') as file:
    df_sas = file.to_data_frame()
```



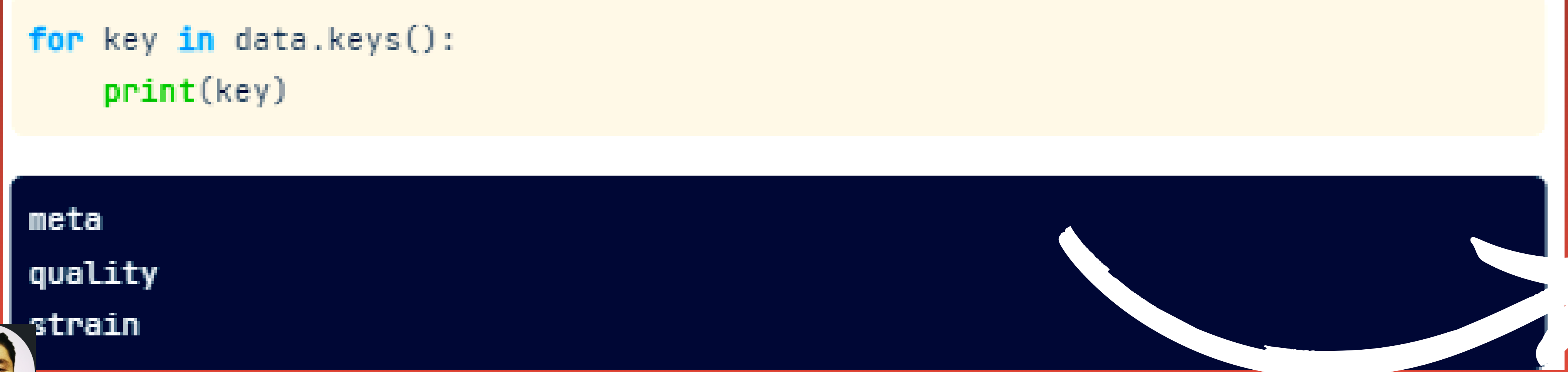
HDF5 Files

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



```python
import h5py
filename = 'H-H1_LOSC_4_V1-815411200-4096.hdf5'
data = h5py.File(filename, 'r') # 'r' is to read
print(type(data))
```

```
<class 'h5py._hl.files.File'>
```

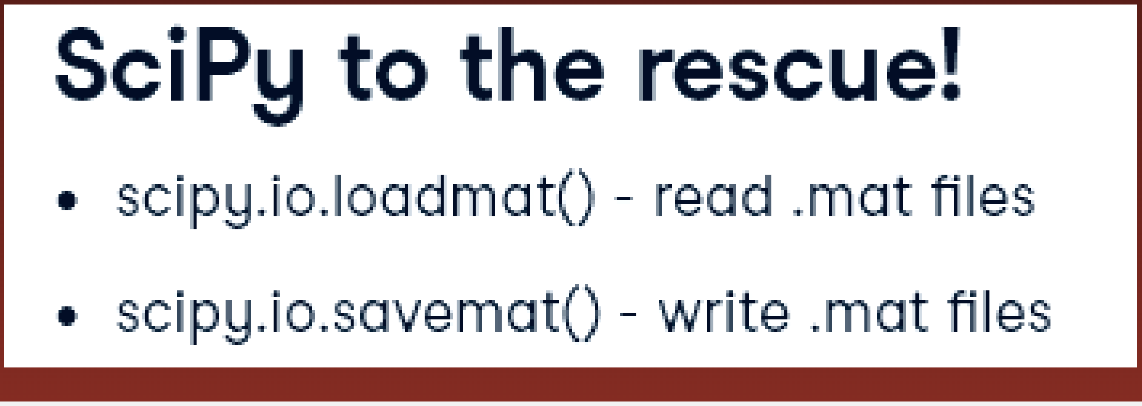


```python
for key in data.keys():
    print(key)
```

```
meta
quality
strain
```

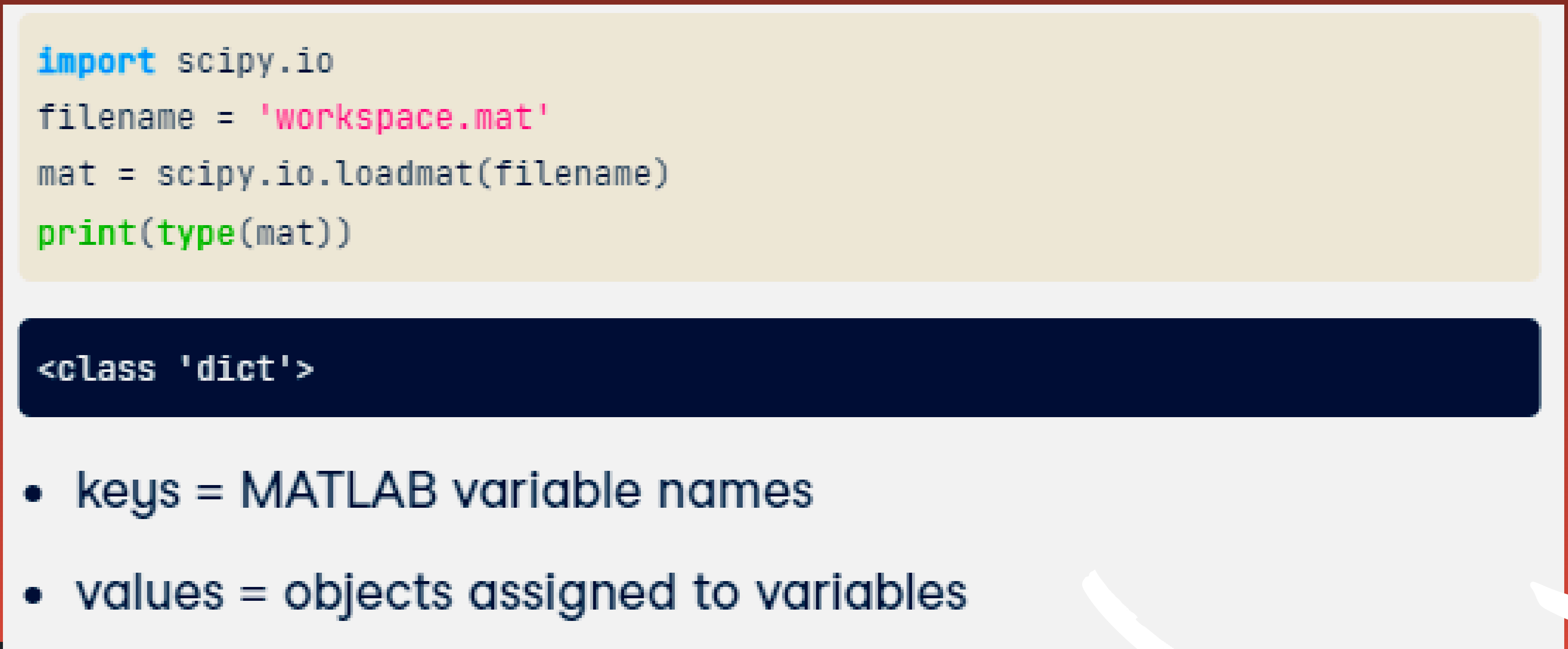
MATLAB Files

- MATLAB = "Matrix Laboratory"
- Industry Standard in Engineering and Science
- Data saved as .mat files



SciPy to the rescue!

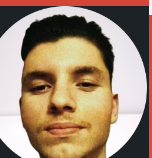
- scipy.io.loadmat() - read .mat files
- scipy.io.savemat() - write .mat files



```python
import scipy.io
filename = 'workspace.mat'
mat = scipy.io.loadmat(filename)
print(type(mat))
```

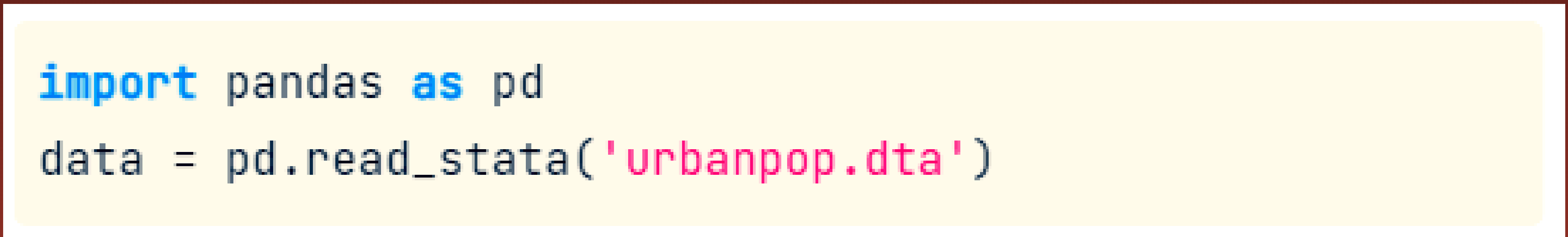
```
<class 'dict'>
```

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

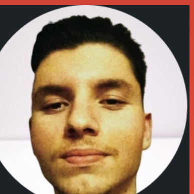


Stata Files

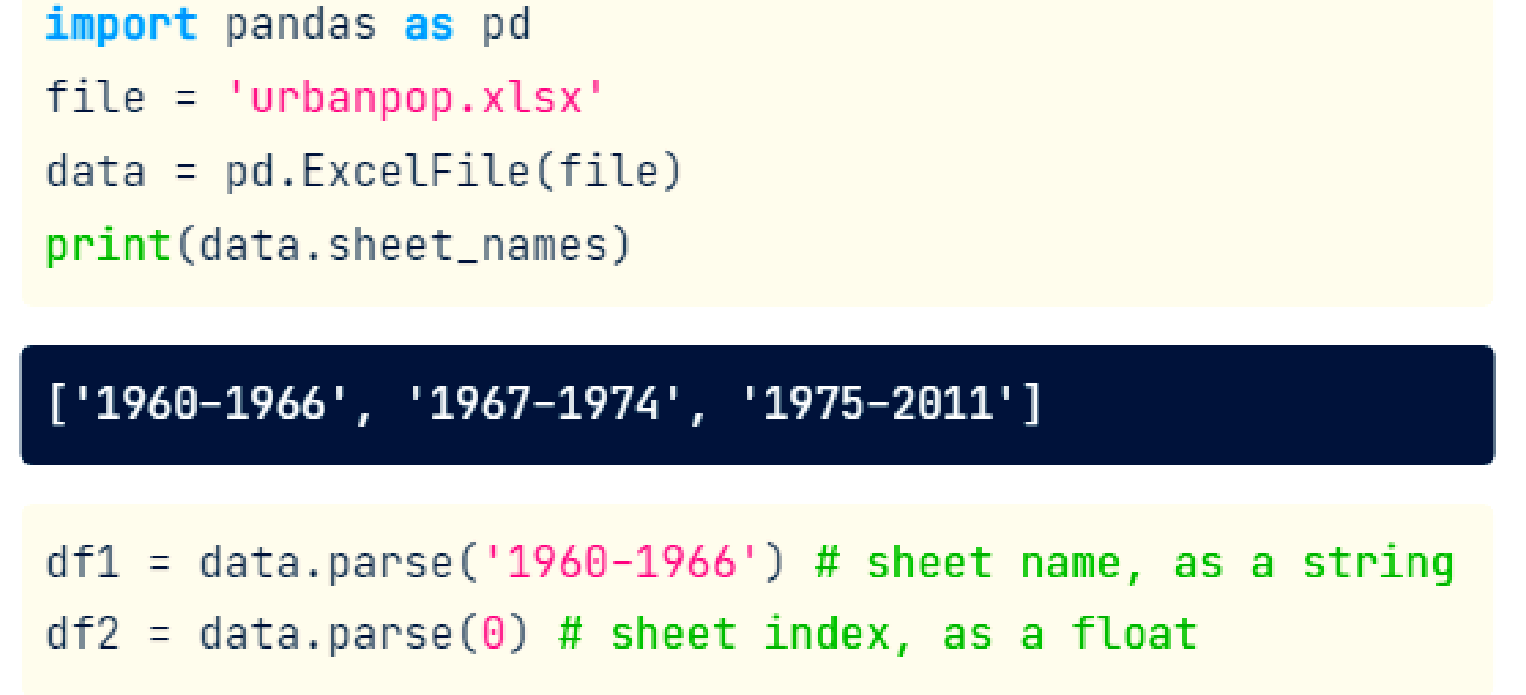
- Stata: "Statistics" + "Data"
- Academic Social Sciences Research



```python
import pandas as pd
data = pd.read_stata('urbanpop.dta')
```



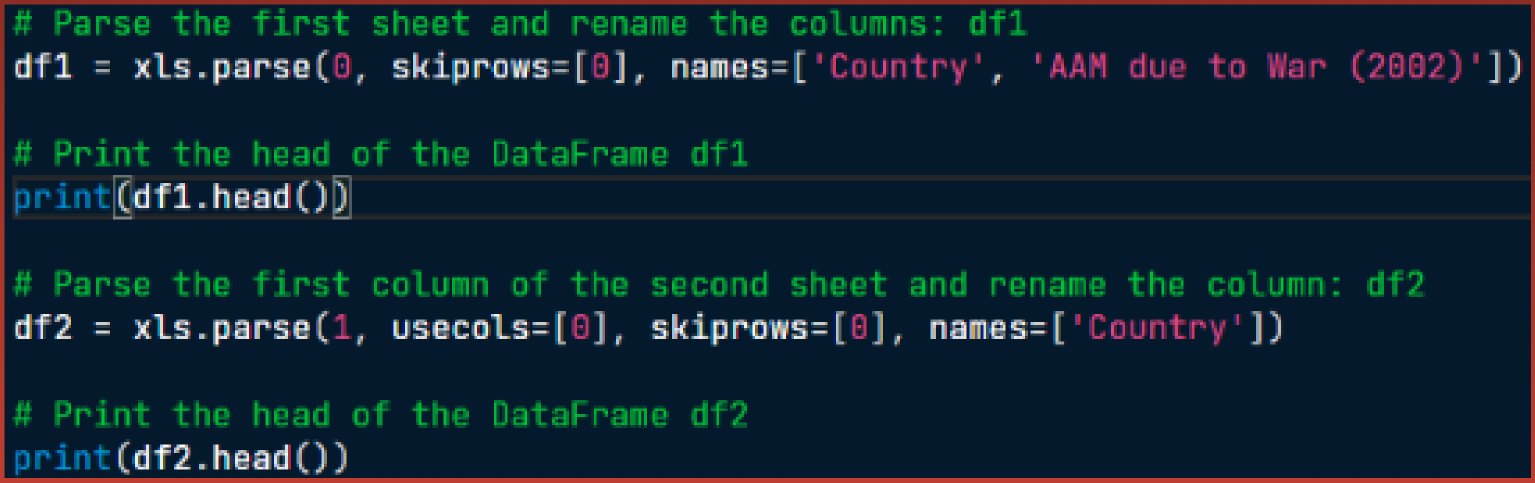
Excel Files



```
import pandas as pd
file = 'urbanpop.xlsx'
data = pd.ExcelFile(file)
print(data.sheet_names)
```

```
['1960-1966', '1967-1974', '1975-2011']
```

```
df1 = data.parse('1960-1966') # sheet name, as a string
df2 = data.parse(0) # sheet index, as a float
```

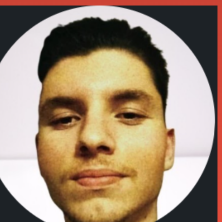


```
# Parse the first sheet and rename the columns: df1
df1 = xls.parse(0, skiprows=[0], names=['Country', 'AAM due to War (2002)'])

# Print the head of the DataFrame df1
print(df1.head())

# Parse the first column of the second sheet and rename the column: df2
df2 = xls.parse(1, usecols=[0], skiprows=[0], names=['Country'])

# Print the head of the DataFrame df2
print(df2.head())
```



Would be Glad to Connect