Centre for
Data Analytics

Insight

Research Toolbox - Data Analysis with Python

A Waternomics Case Study

Umair ul Hassan

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi

Agenda

- An overview of Python ecosystem
- Waternomics case study
- Data Access
- Data Manipulation
- Data Visualization
- Tips & Tricks
- Advanced Libraries
- Q & A

The Python Language

- According to Wikipedia

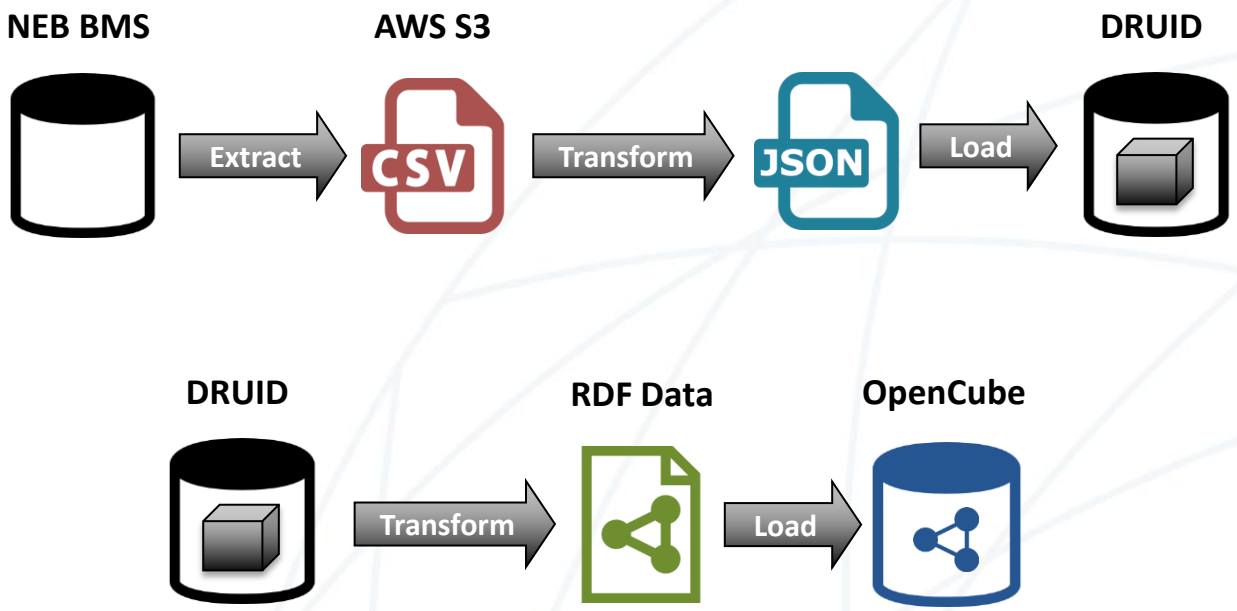
> *a widely used high-level, general-purpose, **interpreted, dynamic programming language**. Its design philosophy emphasizes code **readability**, and its syntax allows programmers to express concepts in **fewer lines of code***

Python Distribution

- Official open source interpreter is **CPython** available at www.python.org
- A distribution packages a set of python tools, modules and libraries to simplify setup and installation

Waternomics Case Study

- Linked Water Dataspace



Data Access

- **Simple file IO functions**
  - *open, read, write*
- **Pandas**
  - *read_csv, read_excel, read_hdf, read_sql, read_json, read_msgpack, read_html, read_gbq, read_stata, read_sas, read_clipboard, read_pickle*
  - For writing replace “read” with “to” e.g. *to_csv*
- **RDFlib**
  - *parse, serialize*
- **Requests (for HTTP/HTTPS)**
  - *get, post, put, delete, head, options*
- **json**
  - *dumps, loads*

Data Manipulation

- **Numpy**
  - Base N-dimensional array package
- **Pandas**
  - Data structures & analysis
  - Allows multi-dimensional OLAP like operations
- **Scipy**
  - Set of package for mathematics, science, and engineering
  - Integration, optimization, signal processing, linear algebra, image processing, spatial data analysis, etc
- **Statsmodels**
  - Statistical models, tests, and analysis

Data visualization

- **Matplotlib**
  - Library for 2D Plotting
  - Allows export to images
- **Seaborn**
  - Attractive visualization using matplotlib
  - Use themes for appealing graphs
- **Bokeh**
  - Interactive visualizations for web browsers
  - Deploy visualization of as part of a webside

Tips & Tricks

- **Running a IPython/Jupyter server on Virtual Machine**
  - Allows remote access and data analysis
  - Always password protect the server
  - Do not print or view large datasets in browser

- **Figures and tables for Latex**
  - Generate Latex code for DataFrames using *to_latex*
  - Save matplotlib plots as *.pgf* for inclusion in Latex

- **Package/module management**
  - pip - The Python package and dependency manager
  - conda - Cross-platform, Python-agnostic binary package manager
  - setuptools – Python project packaging, testing, installation, etc

Advanced Libraries

- **scikt-learn**
  - Python library for machine learning
- **Pyomo**
  - Library for optimization modelling
  - Use in conjuction with glpk, grobi, CPLEX, etc
- **NLTK**
  - Natural language toolkit for
- **RDFLib**
  - Set of libraries for RDF and OWL processing
- **Tweepy**
  - Library to access Twitter API

Other resources

- Conferences (SciPy, EuroSciPy, PyData)

- Web frameworks (Django, Flask, CherryPy, Bottle)
- Cross platform GUI frameworks (PyQT, Kivy)

- Awesome Python List https://github.com/vinta/awesome-python

- MOOCs
  - Introduction to Python for Data Science https://www.edx.org/course/introduction-python-data-science-microsoft-dat208x-1
  - Python for Everybody https://www.coursera.org/specializations/python